Comparative Analysis of Interprocess Communication on the X1, XD1, and XT3

Patrick H. Worley
Oak Ridge National Laboratory

CUG 2005
May 18, 2005
Albuquerque Marriott Pyramid North
Albuquerque, New Mexico

Acknowledgements

- Research 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.

- Oak Ridge National Laboratory is managed by UT-Battelle, LLC for the United States Department of Energy under Contract No. DE-AC05-00OR22725.

Paper Co-authors

- Sadaf Alam
- Tom Dunigan
- Mark Fahey
- Jeff Vetter

all at Oak Ridge National Laboratory (ORNL)

Outline

- System Descriptions
- Technical Specifications Summary
- Topology
  - Distance
  - Contention
- Collectives, including
  - HALO
  - Allreduce
- Standard communication microbenchmarks
- Optimal communication protocols
- Applications
  - POP (latency-sensitive)
  - GYRO (bandwidth-sensitive)

* Only in Paper

Caveats

- Much of the data were collected over the last few weeks, and we don't yet understand it all.
- The systems being measured and compared will be changing dramatically in the next few months.
  - X1 => X1e
  - XT3 portals implementation updates
  - XD1 reconfiguration(s): two 72 processor systems combined into a single 144 processor system, replacing direct connect topology with fat tree?

  Some aspects of performance described here will continue to be accurate qualitatively, but some may not.

X1 at ORNL (Phoenix)

Cray X1 with 128 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 512 processors (MSPs), 1024 GB of memory , and ~ 6500 GF/s peak performance.

Custom interconnect providing distributed shared memory access through entire system:

- 4-D hypercube
- 51 GB/s per SMP peak bandwidth
- 5 μs MPI latency between nodes (lower for SHMEM and Co-Array Fortran)

XD1 at ORNL (Tiger)

Twelve chassis Cray XD1:

- Six SMP nodes per chassis
- Two 2.2 GHz 64-bit AMD Opteron 2000 series (single core) processors per node
- 8 GB of memory per node
- 12.8 GB/s memory bandwidth per node

for a total of 144 processors,

576 GB of memory, and

~ 633 GF/s peak performance

Experiments conducted on a six chassis system with:

- 2 or 4 Cray RapidArray links per node (4 or 8 GB/s per node)
- Fully nonblocking Cray RapidArray switch fabric (48 or 96 GB/s)
- 12 or 24 external Cray RapidArray interchassis links - 24 or 48 GB/s aggregate
- 1.6 µs latency between nodes
- Direct Connect Topology
- Linux version 2.4.21 with synchronized Linux scheduler
- MPICH 1.2.5-based communication library

XT3 at ORNL (Jaguar)

40 cabinet Cray XT3 with 3748 compute nodes

- One 2.4 GHz 64-bit AMD Opteron model 150 processor per node
- 2 GB of memory per node
- 6.4 GB/s memory bandwidth per processor

for a total of 3724 processors, 7448 GB of memory ,and ~ 17875 GF/s peak performance.

(System growing to 5212 compute nodes later this year.)

Nodes connected in a 10 x 16 x 24 configuration (X x Y x Z) with a torus in the X and Z directions and a mesh in the Y direction. Cray SeaStar communications and routing chip provides:

- Six links (to six neighbors) in 3D torus/mesh configuration
- Each link has peak bidirectional BW of 7.6 GB/s, with a sustained BW of 4 GB/s
- Linux on service nodes; Catamount v. 1.15 on compute nodes
- MPICH 1.2.5-based communication library

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 a HPS interconnect with two 2-port 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.

Technical Specifications

Specs:	X1	XD1	XT3	Altix 3700	p690 cluster with HFS network
Processor	Cray	AMD Opteron	AMD Opteron	Intel Itanium2	IBM Power4
MHz	800	2200	2400	1500	1300
L1	16K	64K	64K	32K	32K
L2	2MB	1MB	1MB	256K	1.5MB
L3				6MB	128MB/node
peak Gflop/s	12.8	4.4	4.8	6.0	5.2
proc./node	4	2	1	2	32
memory/node	16GB	4GB	2GB	2GB	32-128GB

Technical Specifications

Specs:	X1	XD1	XT3	Altix	p690 cluster with HFS
Latency (8 byte msg., 1 way, usec, measured*)					
MPI (intra-node)	7.3	1.7	-	1.1	3
MPI (inter-node)	7.3	1.7	29	1.1	6
SHMEM	3.8				
Co-Array Fortran	3.9				
Bandwidth (1MB msg., unidirectional, MB/s, measured^)					
MPI (intra-node)	9503	1087	-	1595	1580
MPI (inter-node)	9364	1342	1111	1397	936
Bandwidth (1MB msgs., bidirectional, MB/s, measured^)					
MPI (intra-node)	17145	1095	-	2286	2402
MPI (inter-node)	16936	2173	2150	2561	985

* Dunigan custom benchmarks ^ Worley custom benchmarks

Topology Experiments

Distance:

i-j

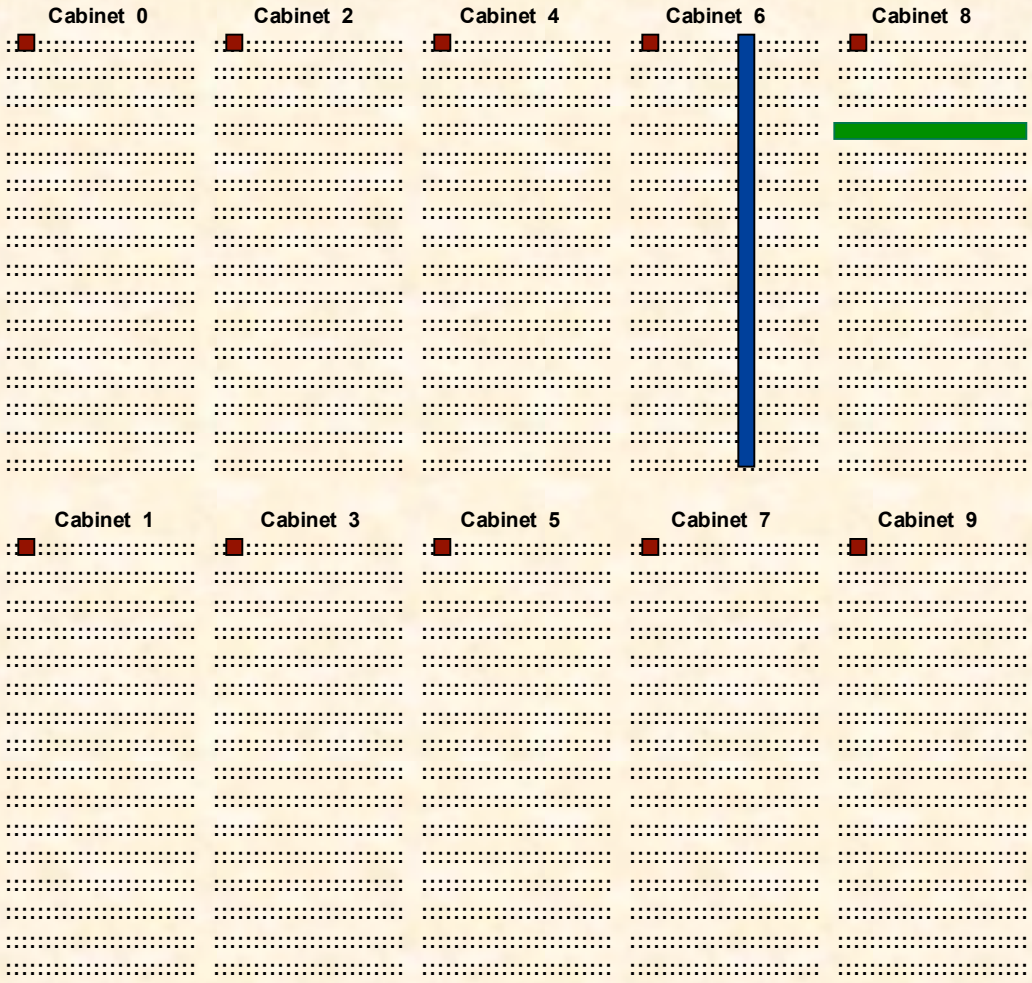
processor i exchanges data with processor j, either simultaneously or one at a time. Depending on i and j, this can be within an SMP node or between SMP nodes.

Contention

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

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

Topology: XT3 at ORNL



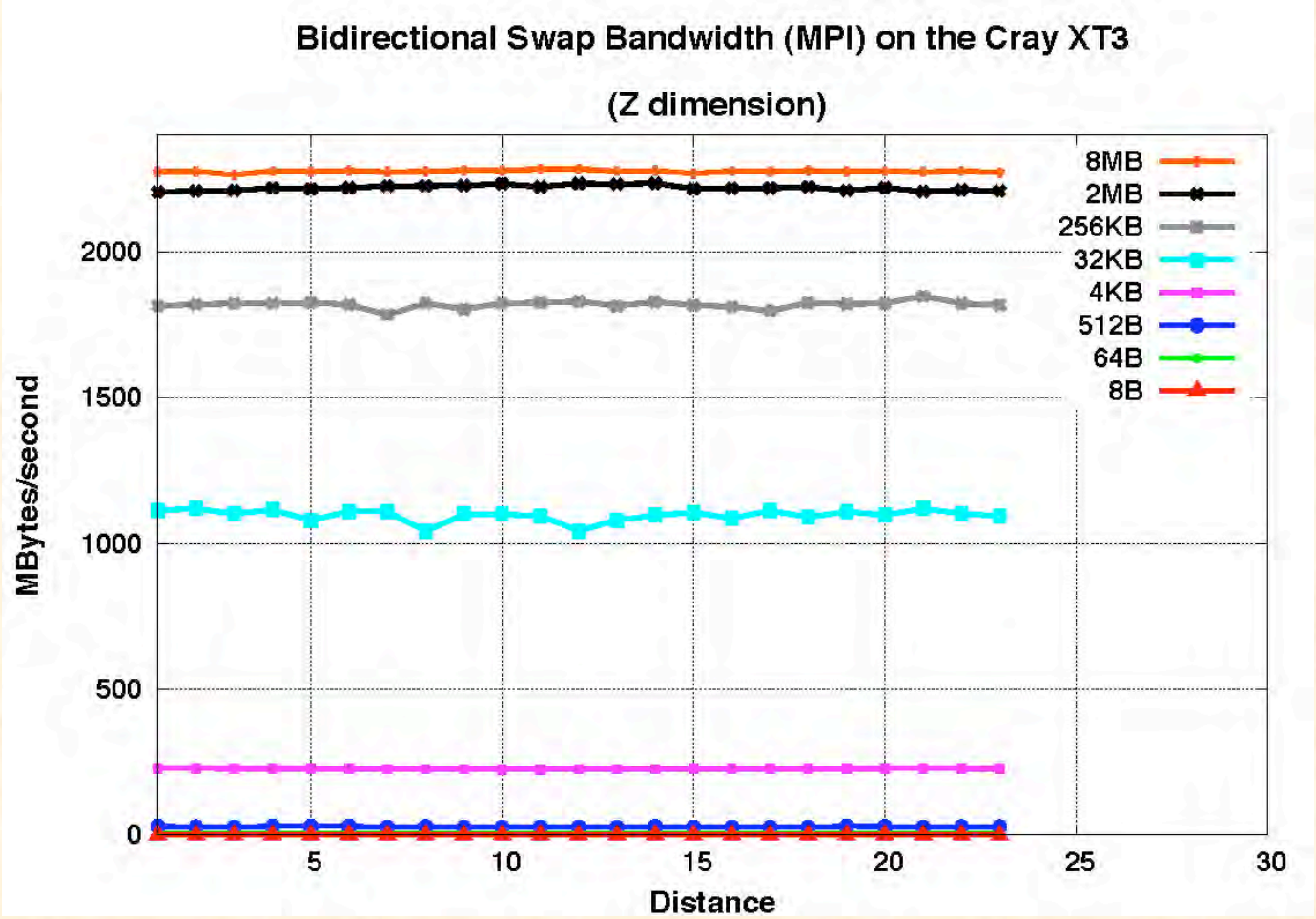
X-dimension (torus):
10 cabinets*

Y-dimension (mesh):
16 rows per cabinet

Z-dimension (torus):
24 columns per cabinet

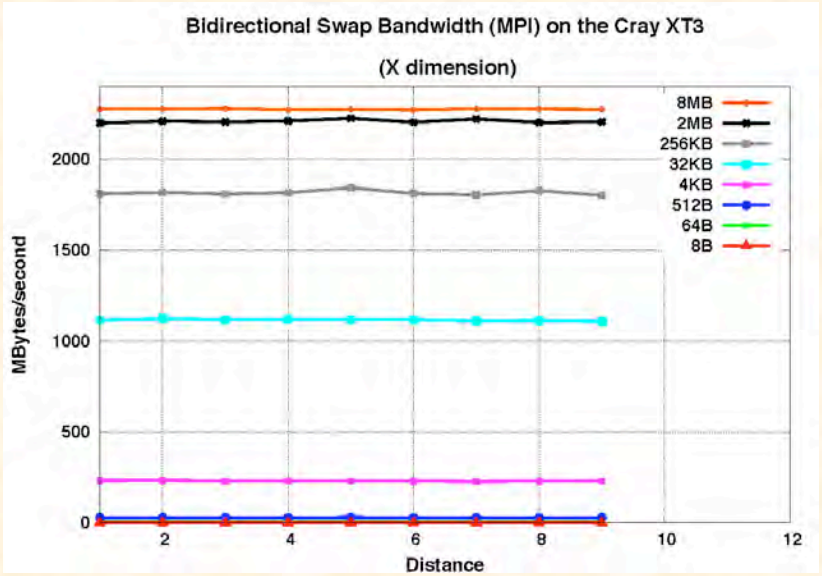
* A physical cabinet is 4 rows by 24 columns. A logical 16x24 cabinet is made up of 4 physical cabinets.

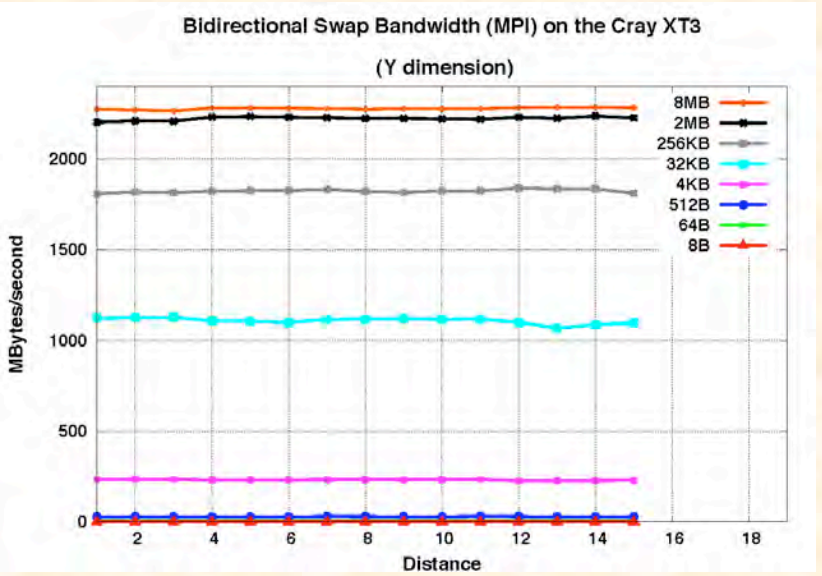
Distance: XT3 (Z dimension)

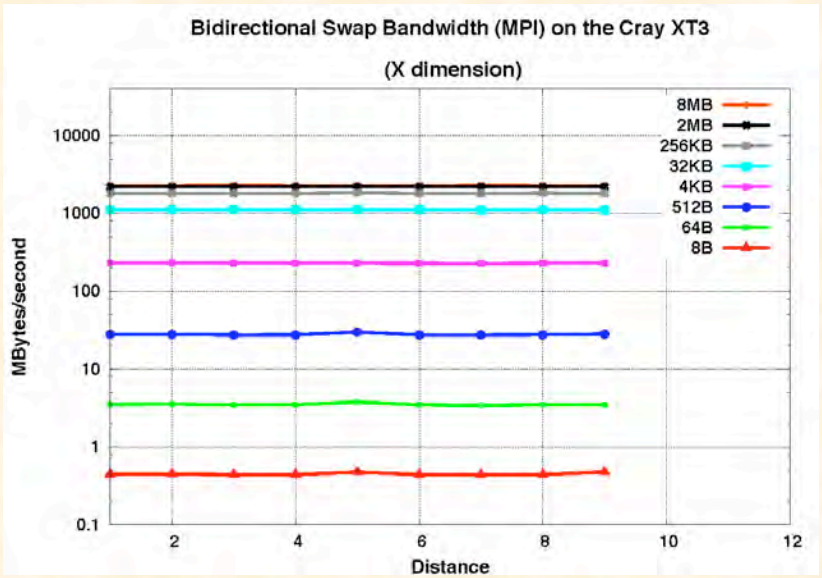


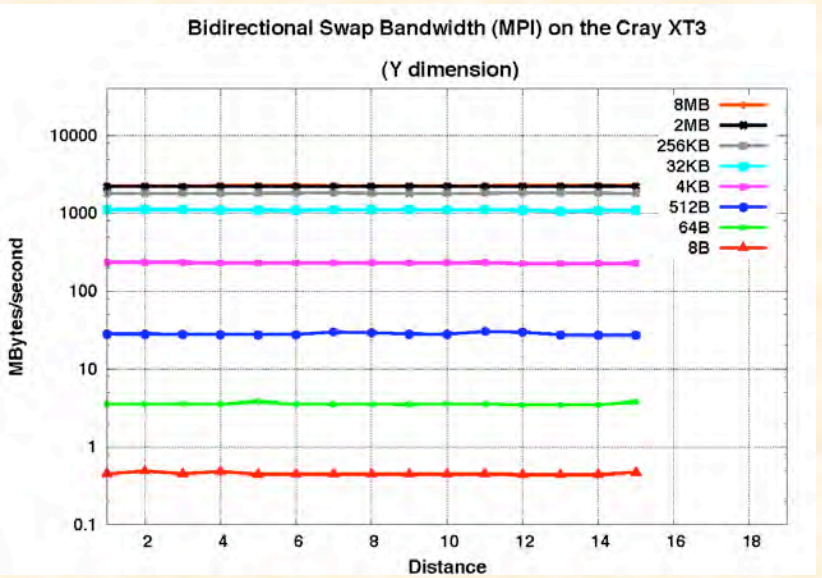
A horizontal curve indicates no performance dependence on distance. These experiments look at distance along the Z dimension for different message sizes. The next two slides look at data for X, Y, and Z dimensions, and a mixed 4x24 YxZ processor subset, using both linear-linear and linear-log plots.

Distance: XT3 (X and Y)

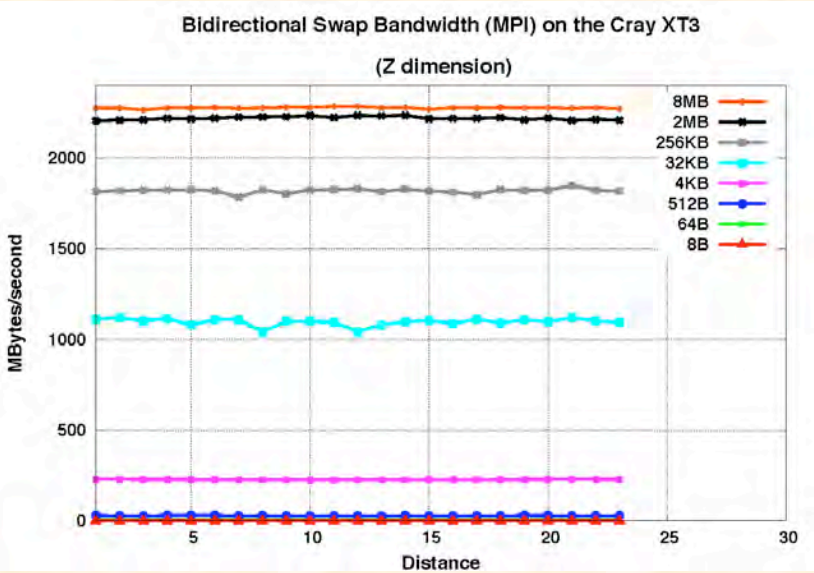


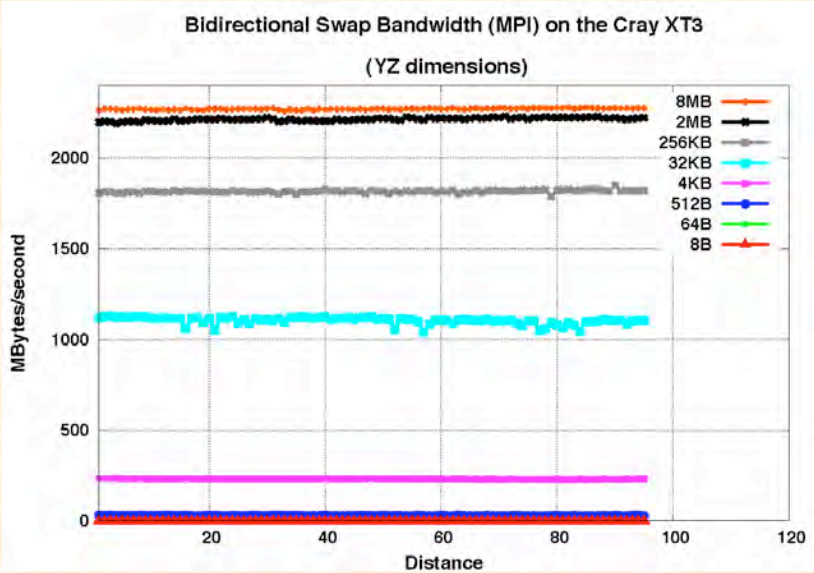


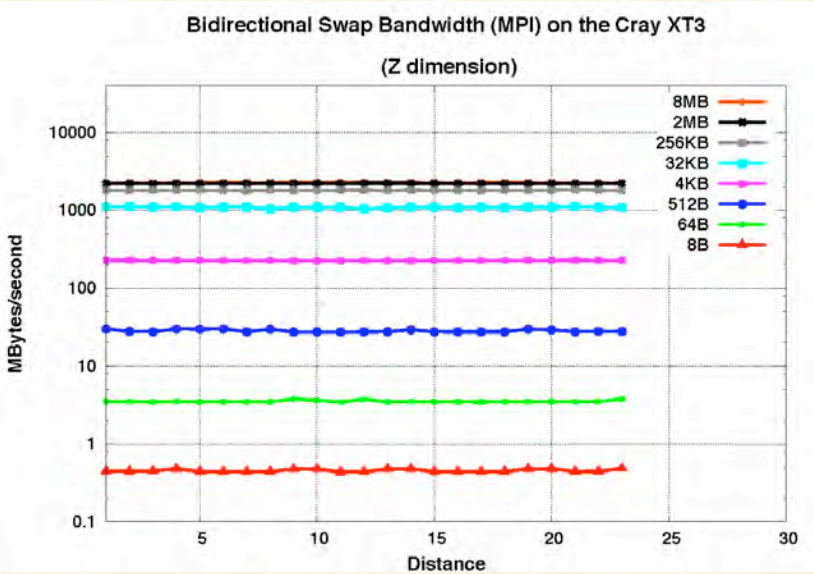


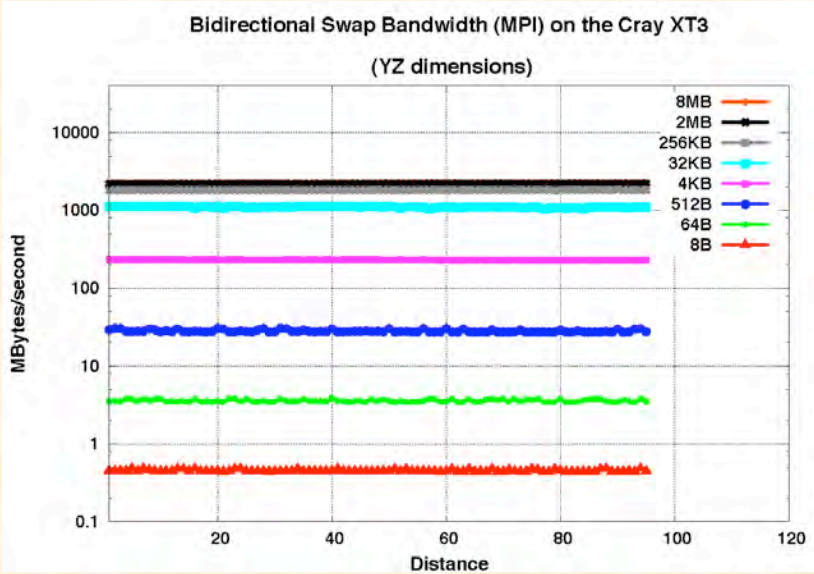


Distance: XT3 (Z and YxZ)

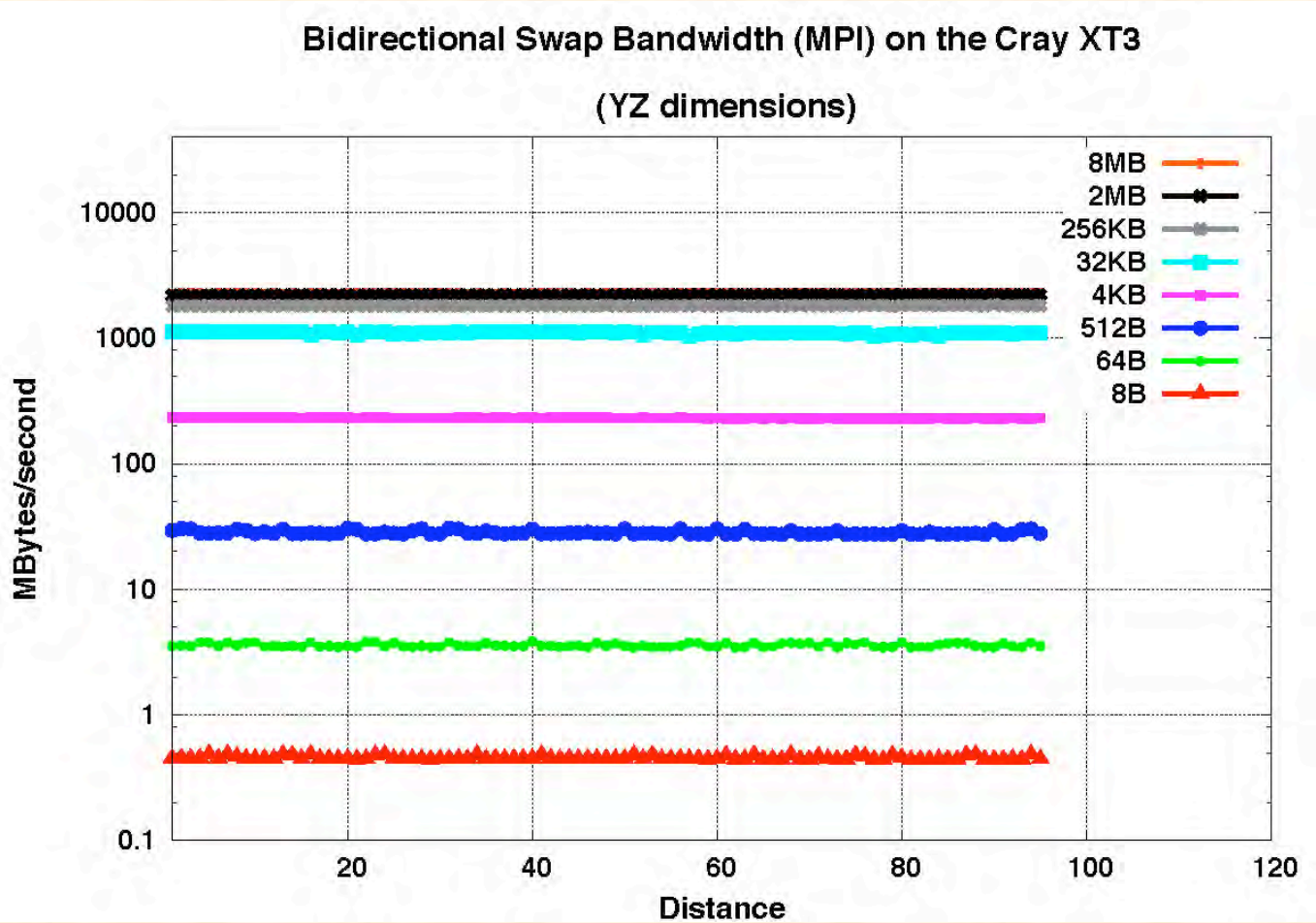






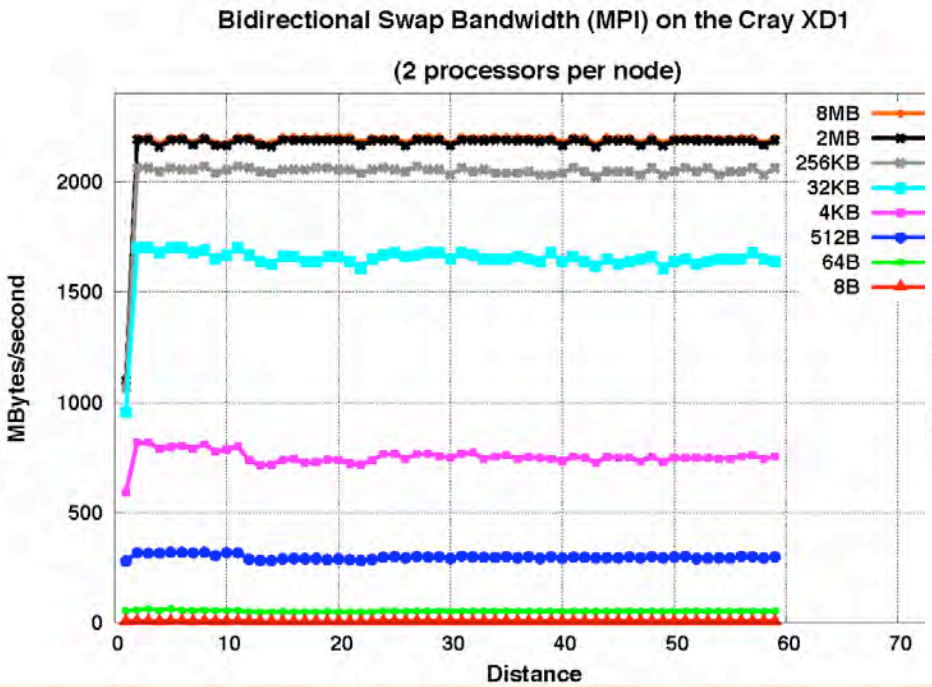


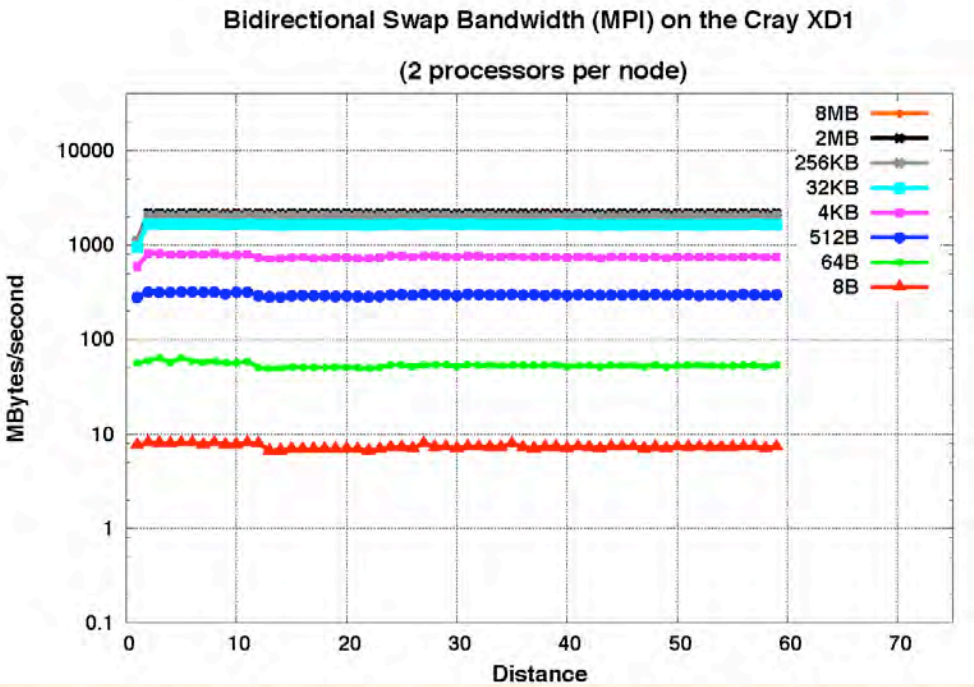
Distance: XT3 (YxZ)



This plot is a repeat of the YxZ data. Note that all curves in the previous two slides are essential horizontal, and performance is the same for a given message size, independent of the processor subset.

Topology and Distance: XD1



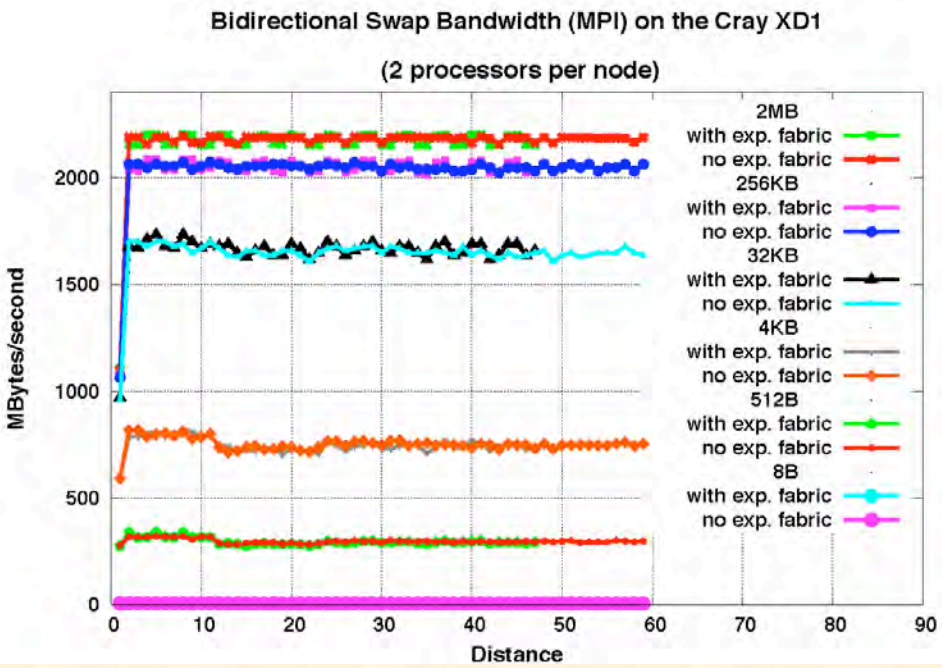


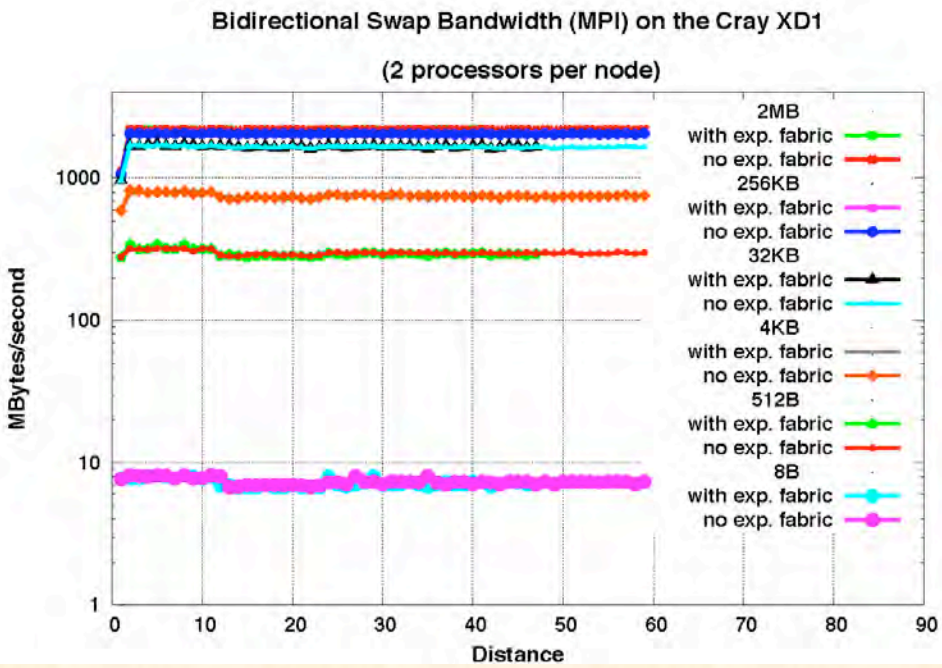
Direct Connect Topology

- Intranode: distance 1
- Intra-chassis: distance 2-11
- Inter-chassis: distance 12-59

Note that intranode performance is half that of performance between nodes for large messages. Other than that, performance is not sensitive to distance.

Distance: XD1 (w/expansion fabric)





Main Fabric

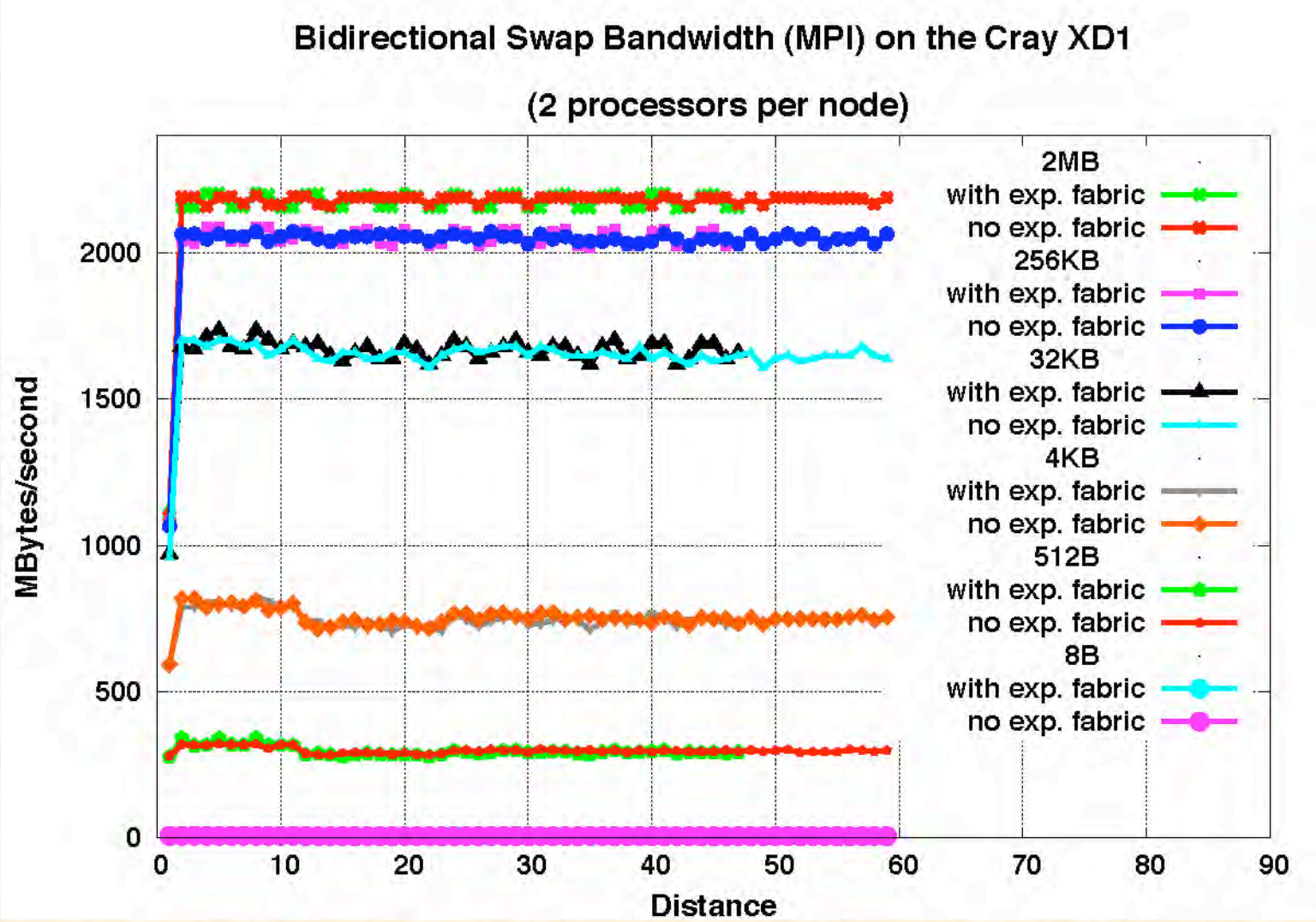
- 2 RapidArray links per node, 12 external RapidArray interchassis links

+ Expansion Fabric

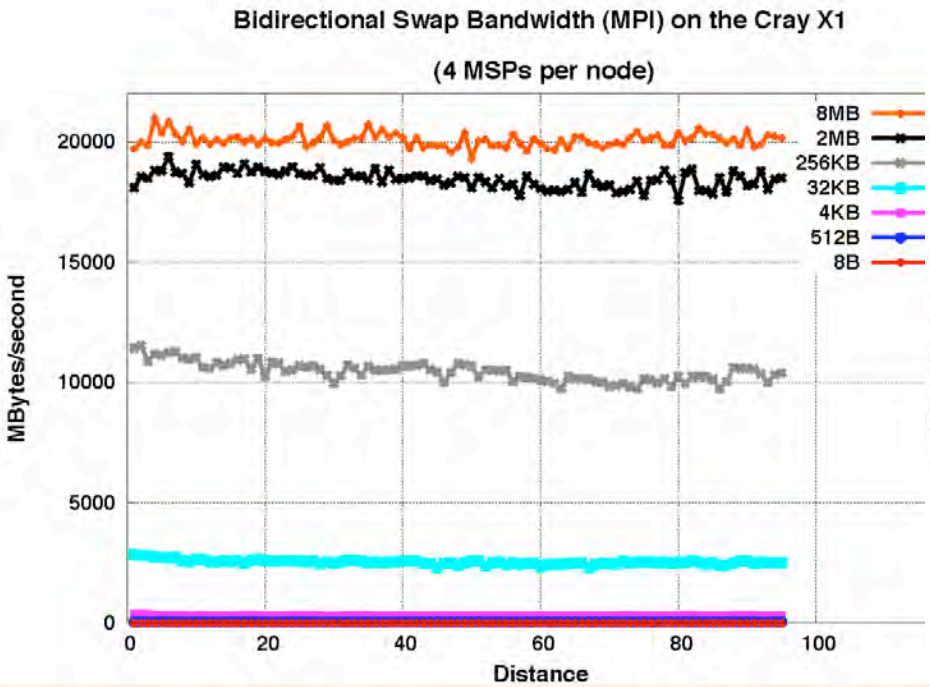
- 4 RapidArray links per node, 24 external RapidArray interchassis links

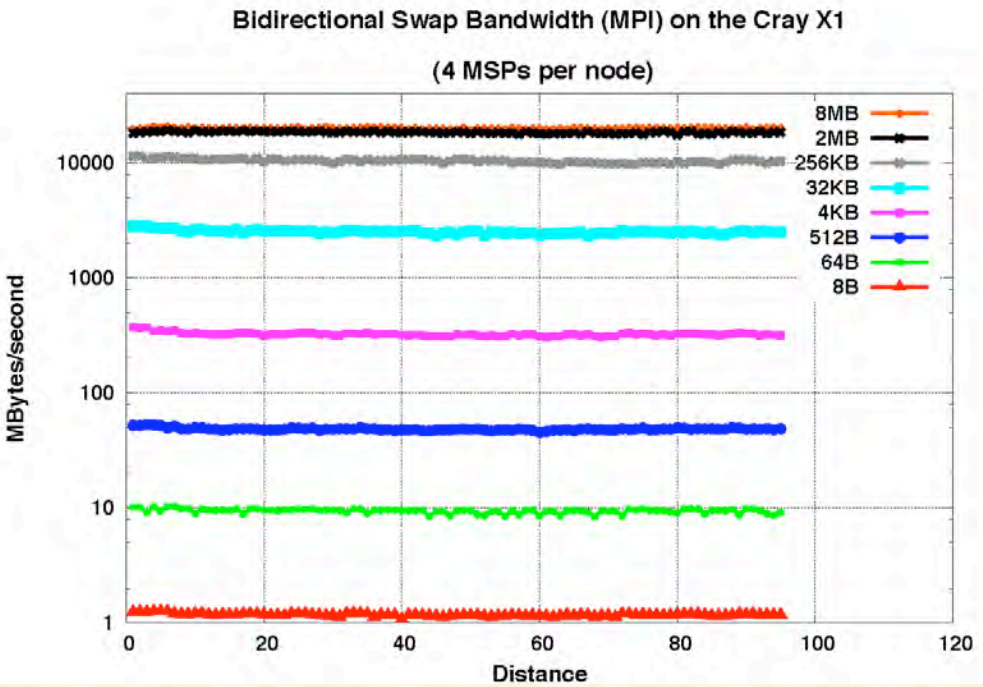
Expansion fabric does not improve performance in these experiments.

Distance: XD1 (w/expansion fabric)



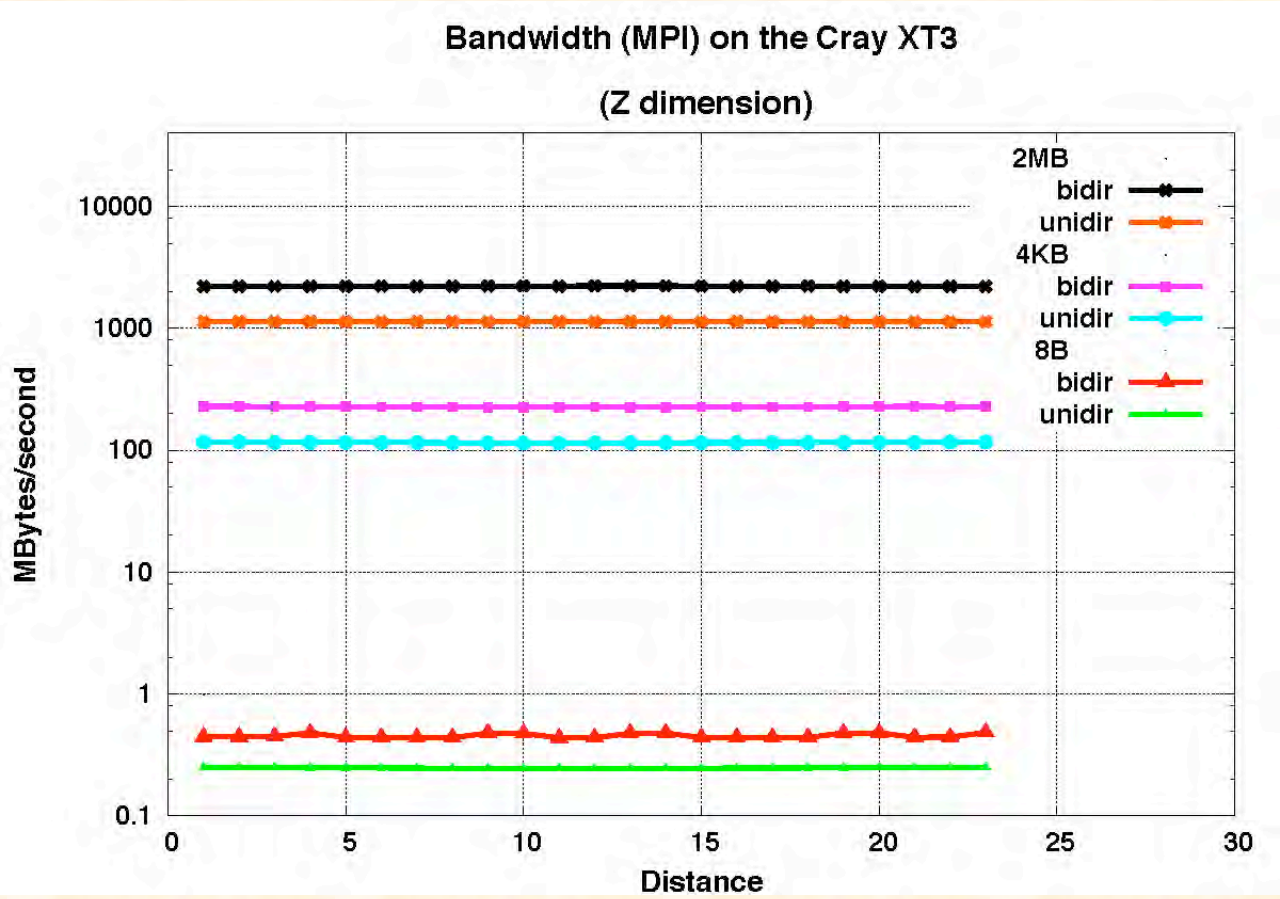
Topology and Distance: X1





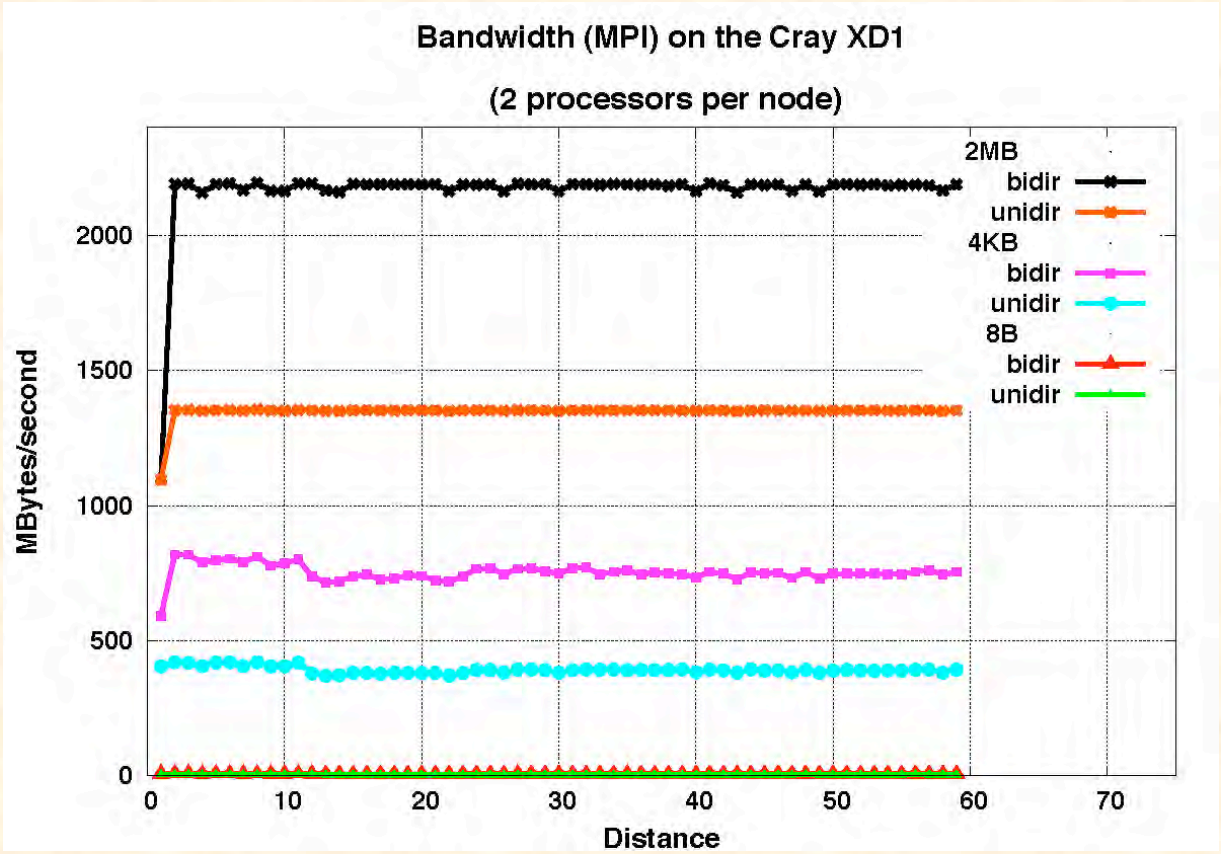
- Systems of size up to 512 MSPs have a 4-D hypercube interconnect.
- "Contiguous" MSPs used in experiments, but system was not dedicated.
- While bandwidth curves are somewhat noisy, there is no practical performance difference due to distance observable in these plots.

Uni- vs. Bidirectional BW: XT3



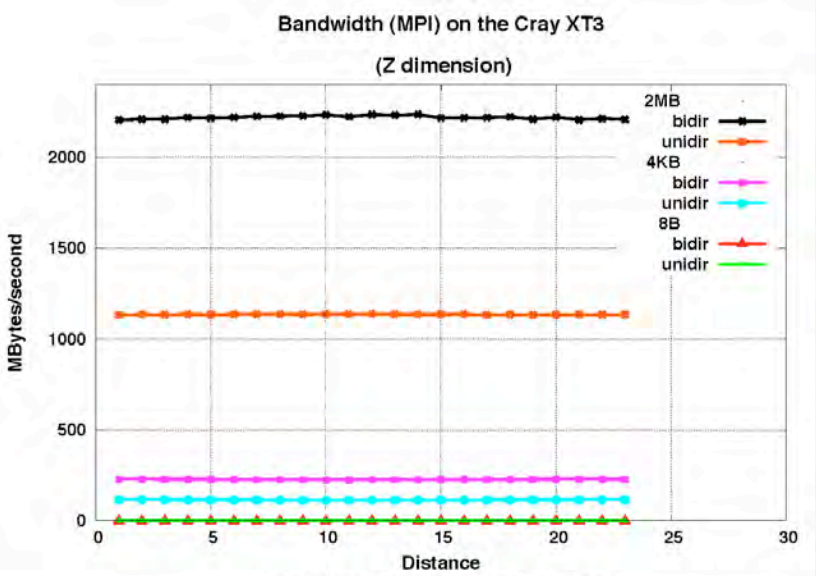
- Comparing unidirectional and bidirectional bandwidth for different message sizes. On the XT3 (and other systems), performance continues to be insensitive to distance.
- Unidirectional bandwidth is half that of bidirectional bandwidth on the XT3.

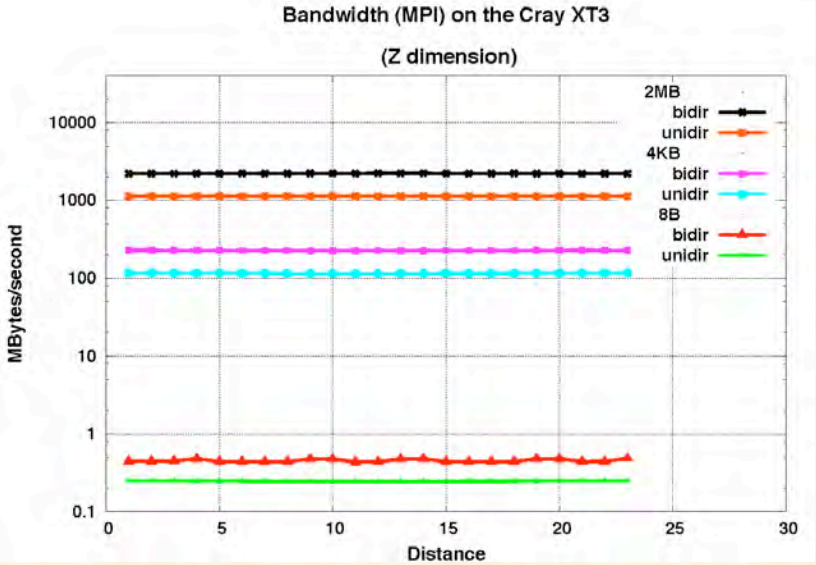
Uni- vs. Bidirectional BW: XD1

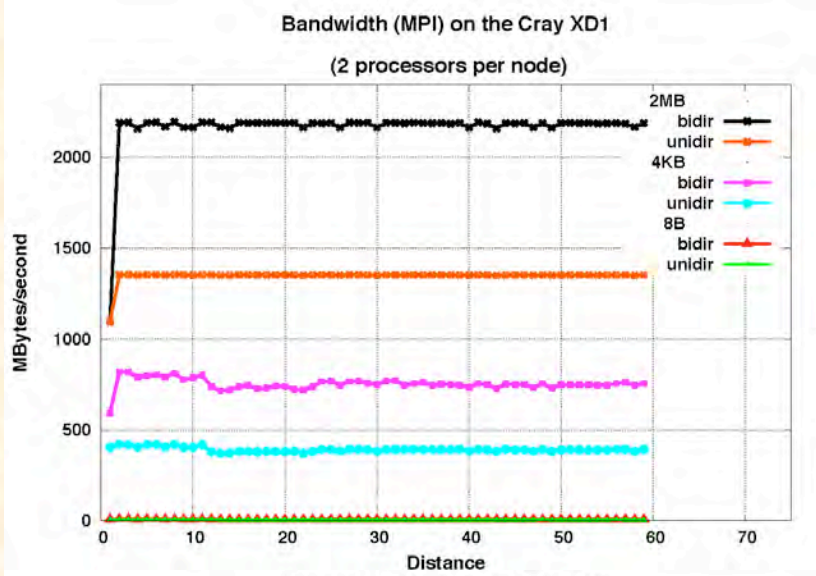


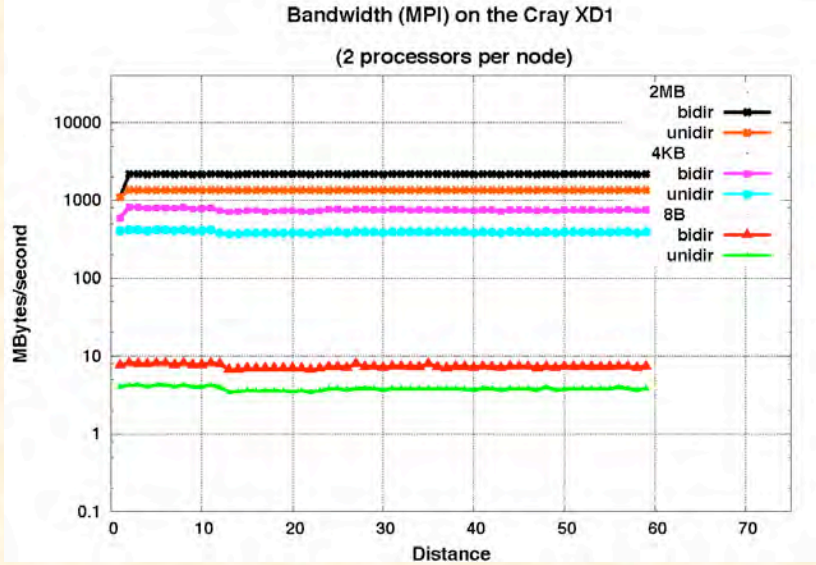
- Unidirectional bandwidth is approx. 60% that of bidirectional bandwidth on the XD1 for larger message sizes (>= 32KB), but is approximately 50% for smaller messages sizes.
- Next slide compares XT3 and XD1 performance, showing similar bandwidth for largest message sizes, but superior performance on the XD1 for smaller messages (due to lower latency) and for unidirectional bandwidth.

Uni- vs. Bidirectional BW: XT3 and XD1

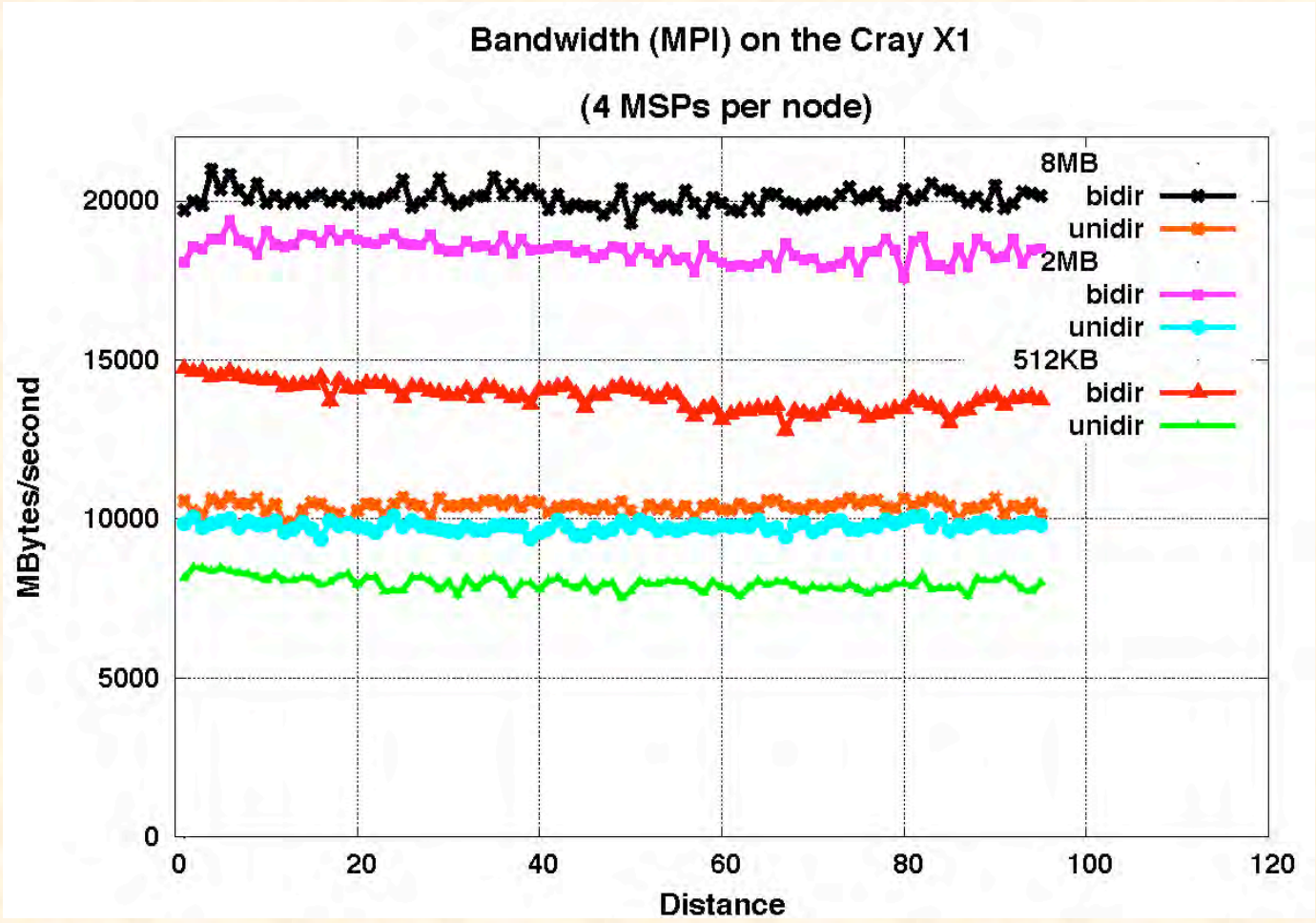






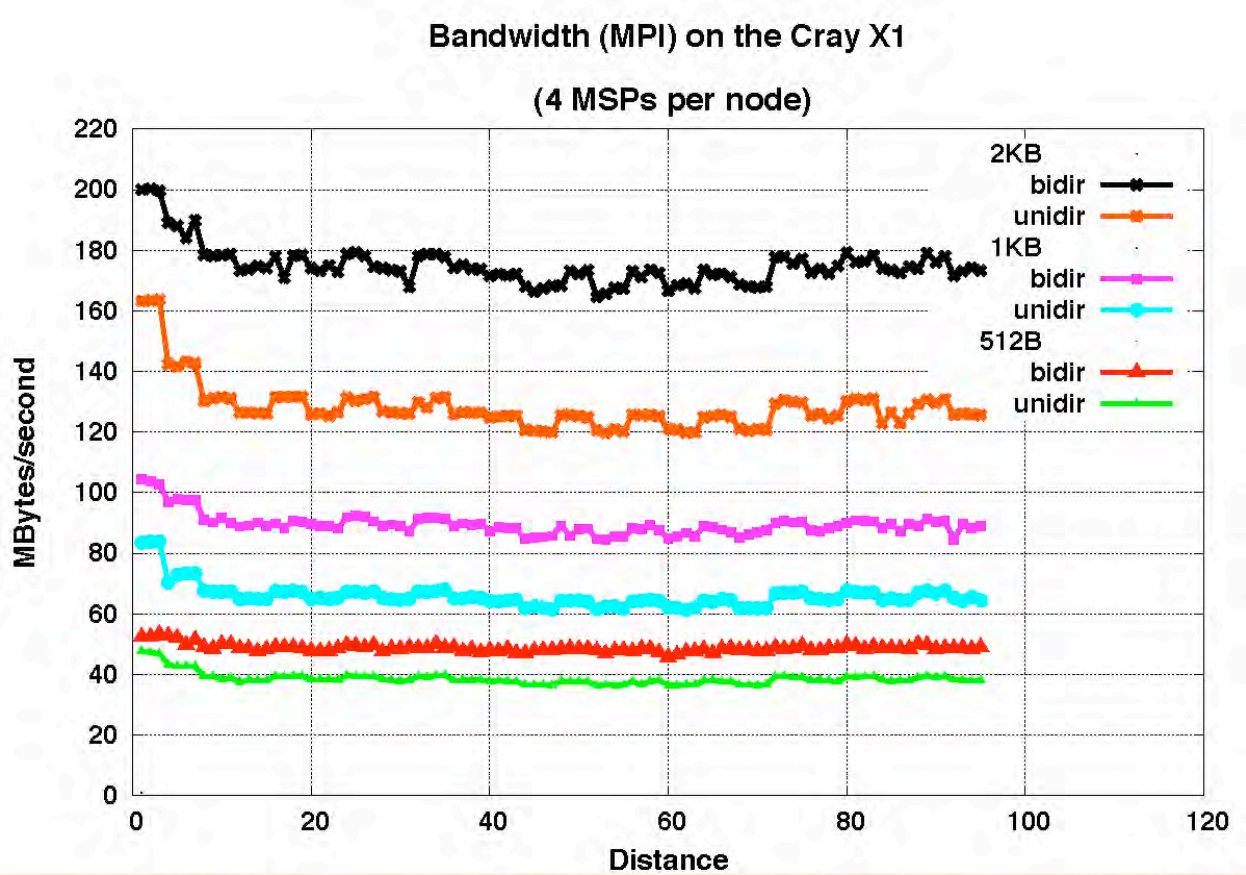


Uni- vs. Bidirectional BW: X1



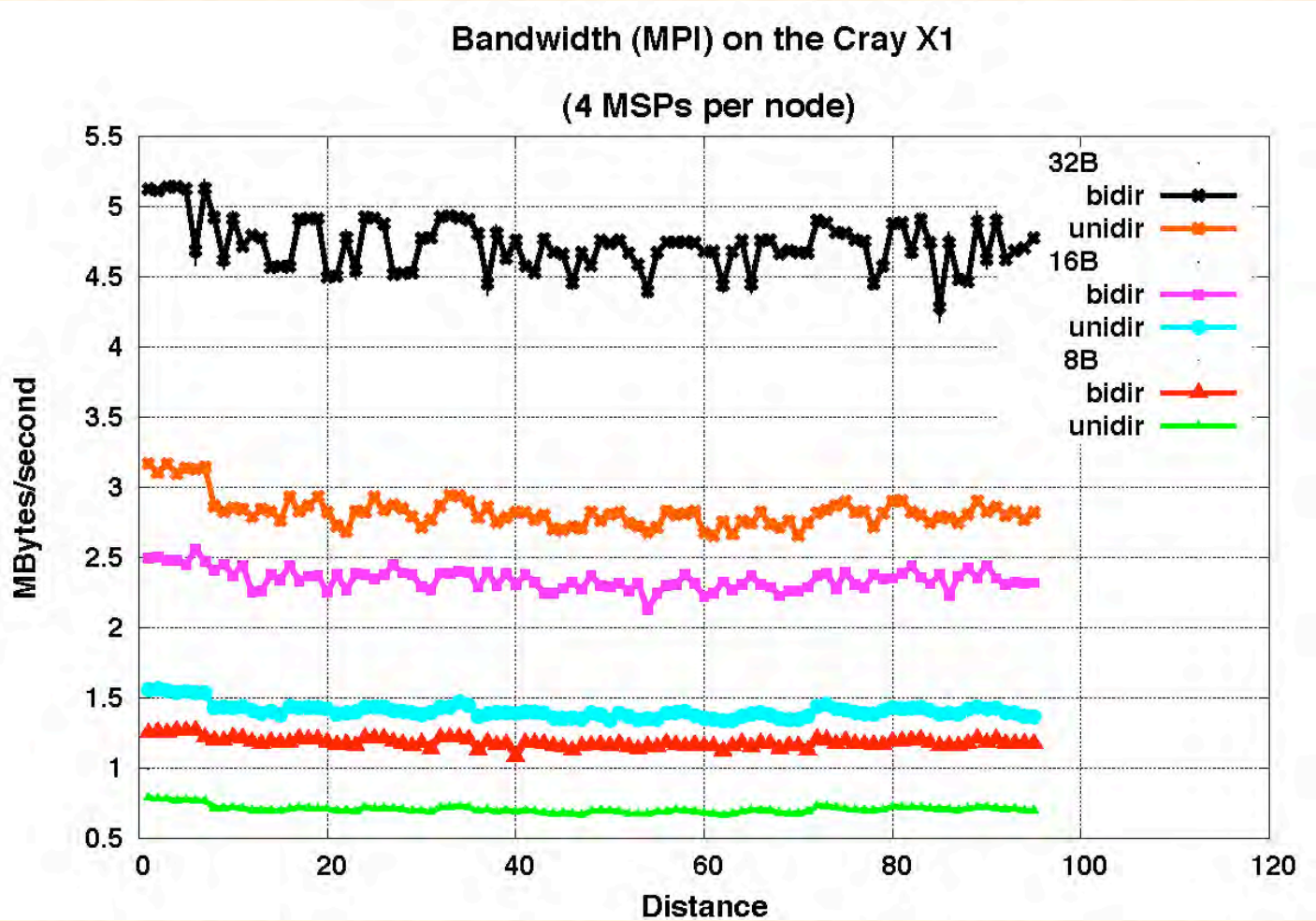
- Unidirectional vs. bidirectional bandwidth comparison is more complicated on the X1. For the largest message sizes unidirectional bandwidth is approximately half that of the bidirectional bandwidth.

Uni- vs. Bidirectional BW: X1



- For the 512B - 2048B message sizes unidirectional bandwidth is approximately 75% that of the bidirectional bandwidth. Also note that bandwidth is higher when communicating between processors in the same SMP node.

Uni- vs. Bidirectional BW: X1

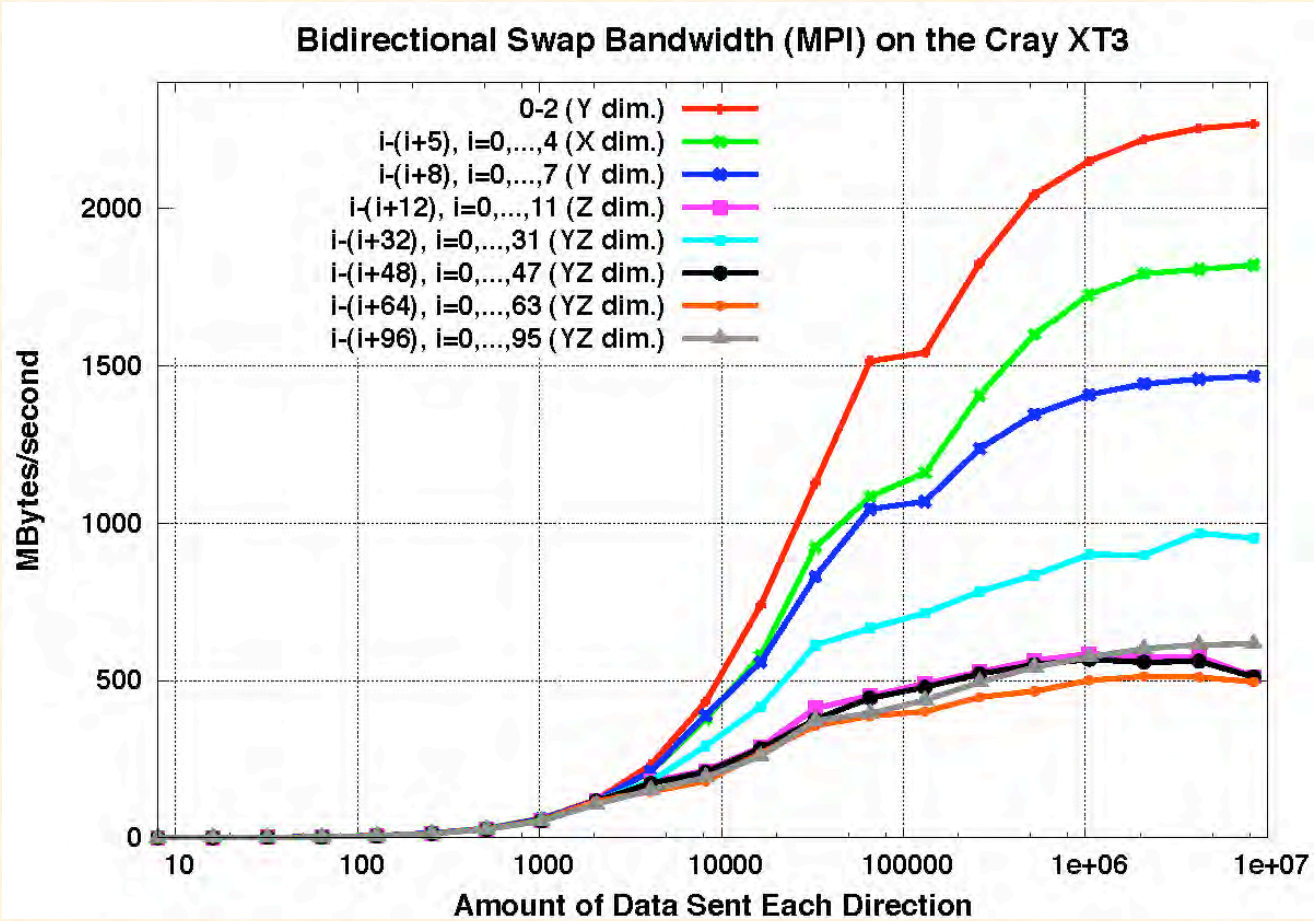


- For small message sizes unidirectional bandwidth is 50-60% that of the bidirectional bandwidth.

Distance Summary

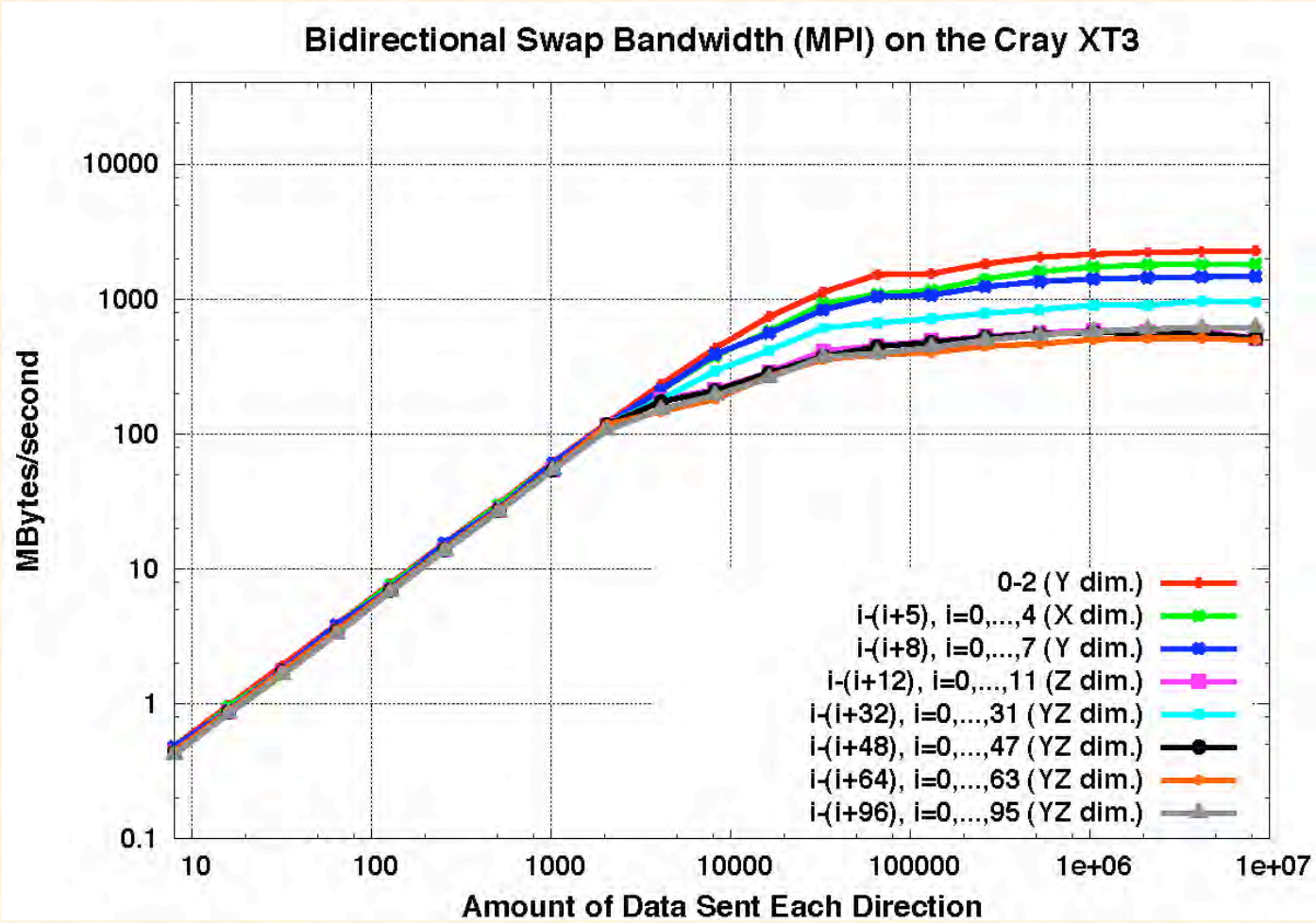
- MPI bandwidth (and latency) is not sensitive to distance between communicating processors on any of the systems (for current configurations and system software) when all other processors are idle, except for intranode communication on the XD1 and X1.
- MPI unidirectional bandwidth is 50%-60% that of bidirectional bandwidth, except on the X1 for 512B-2048B message sizes where it is 75%.
- X1 MPI peak bandwidth is much higher than that on the other Cray systems (as per specifications).
- XD1 MPI latency is much lower than that on the other Cray systems (as per specifications).
- XT3 MPI performance is hurt by current high latency (as expected), but peak bandwidth is comparable to that on XD1.
- XD1 expansion fabric did not enhance communication performance in these experiments

Contention: XT3



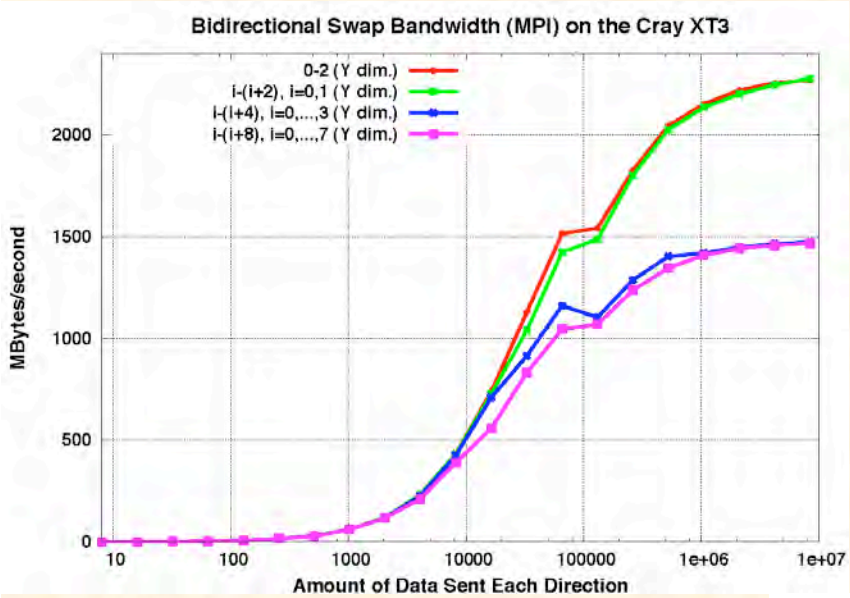
- Examining bandwidth achieved for a single processor pair when multiple pairs are communicating simultaneously. For the XT3 (and other systems) aggregate bandwidth limitations and contention impact single pair performance. On XT3, maximum single pair performance drops from 2.2 GB/s to 500 MB/s.

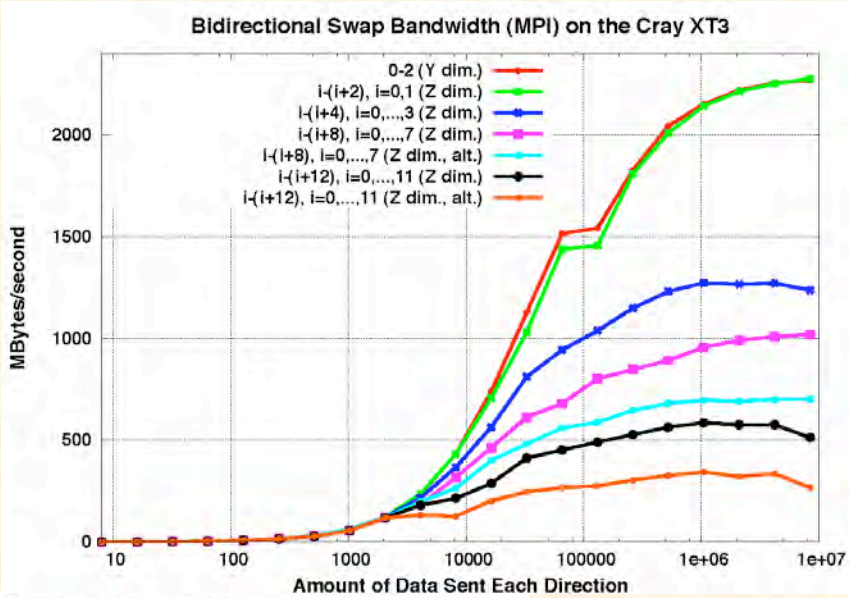
Contention: XT3

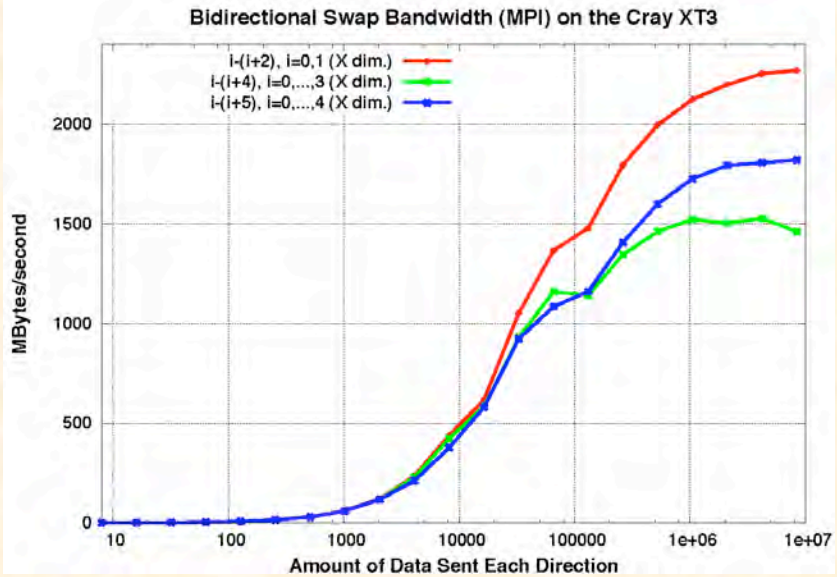


- Linear-log plot shows no contention for small messages (<= 2KB).
- Next slide compares contention in each coordinate direction. Details do vary with direction.

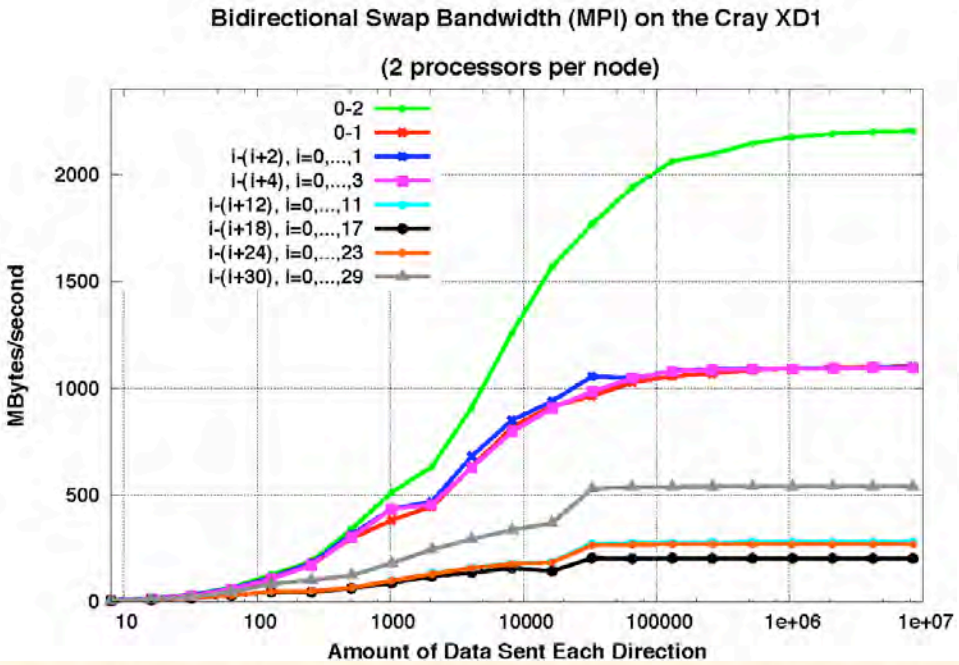
Contention: XT3 (Y vs. Z vs. X)

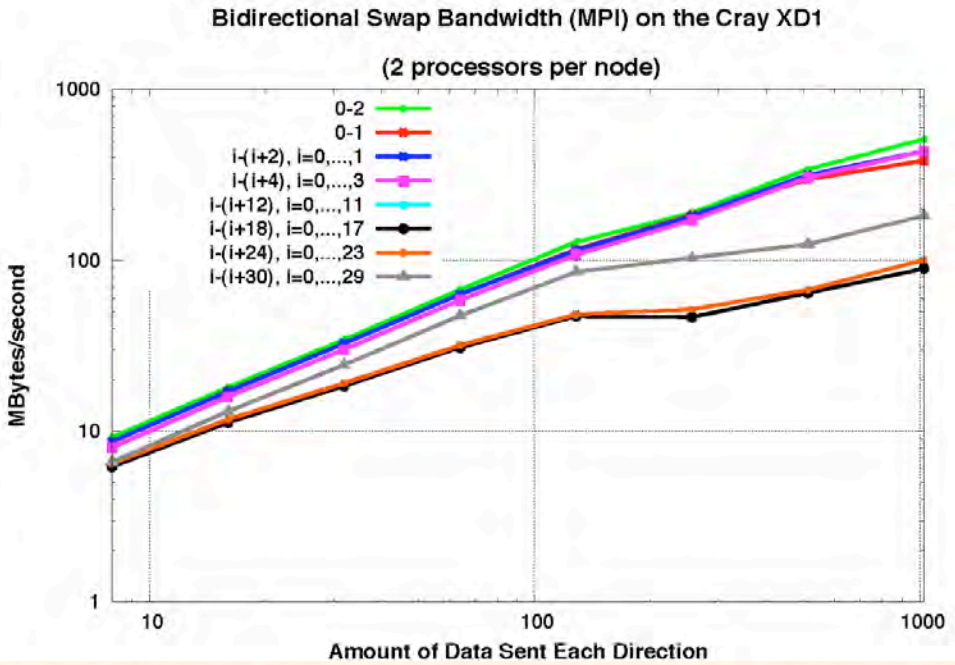






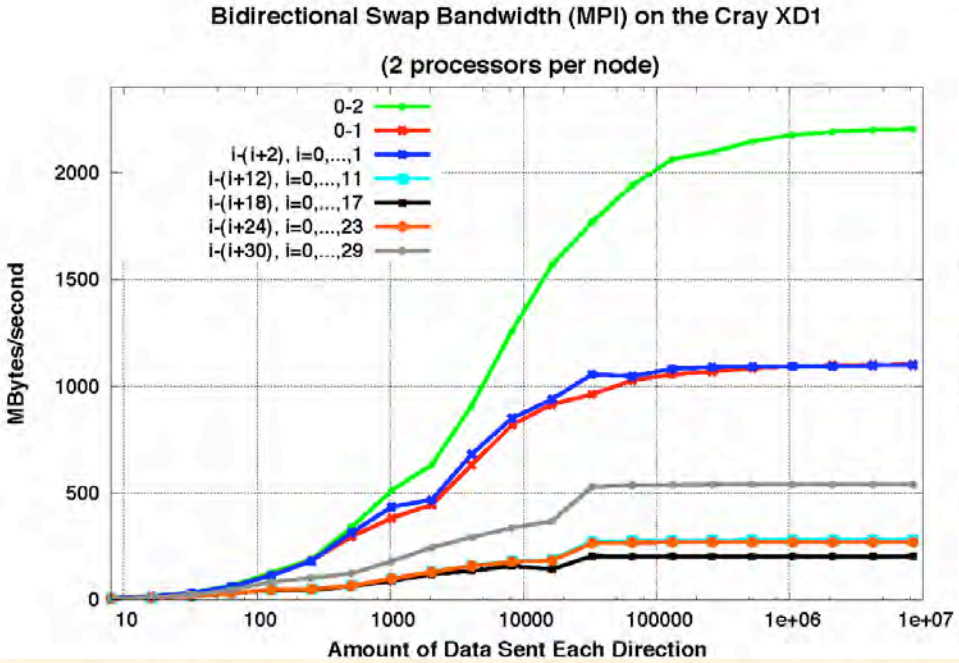
Contention: XD1

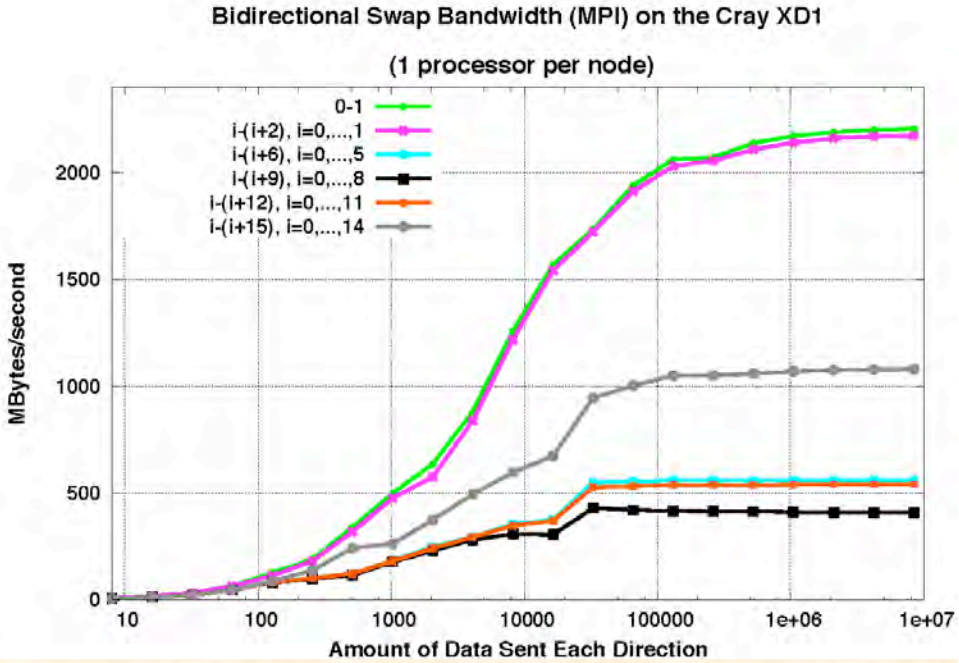




- Contention (decrease in single pair bandwidth) is worse on the XD1 than on the XT3 when using both processors on an XD1 node. Performance is also more “erratic” as a function of the number of communicating pairs, and the communication pattern between chassis appears to be important. Note that contention is apparent for even the smallest message sizes.

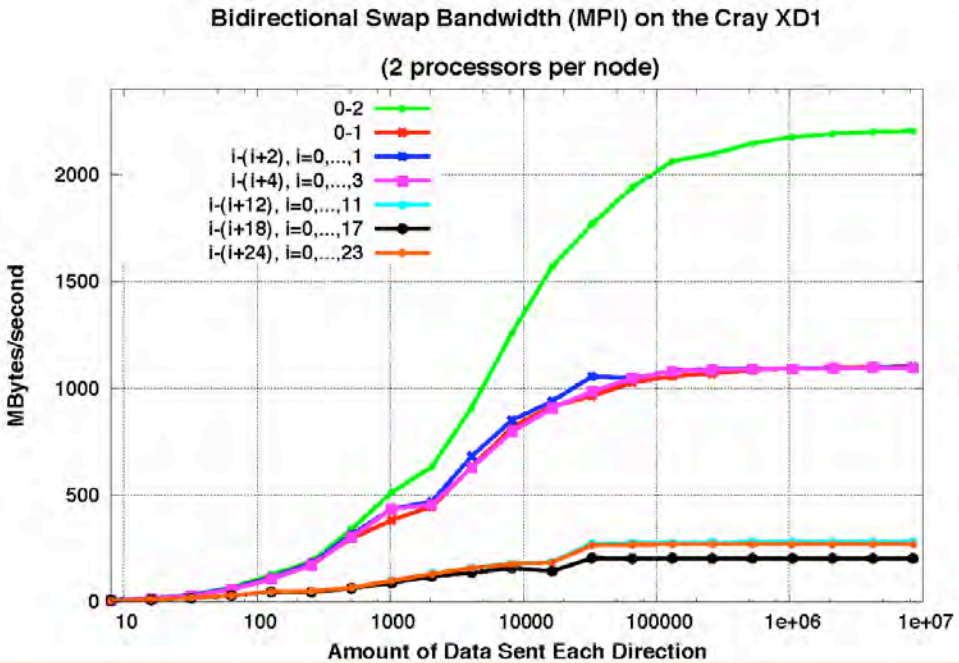
Contention: XD1 (2 proc. vs. 1 proc.)

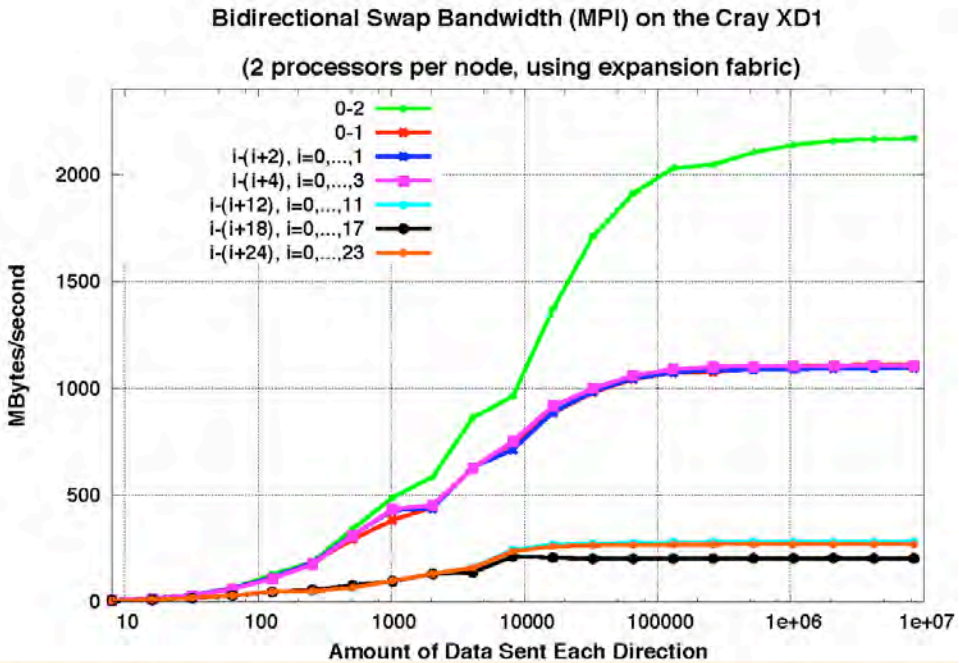




- Using only one processor per node doubles bandwidth (under contention) as a function of the number of nodes.

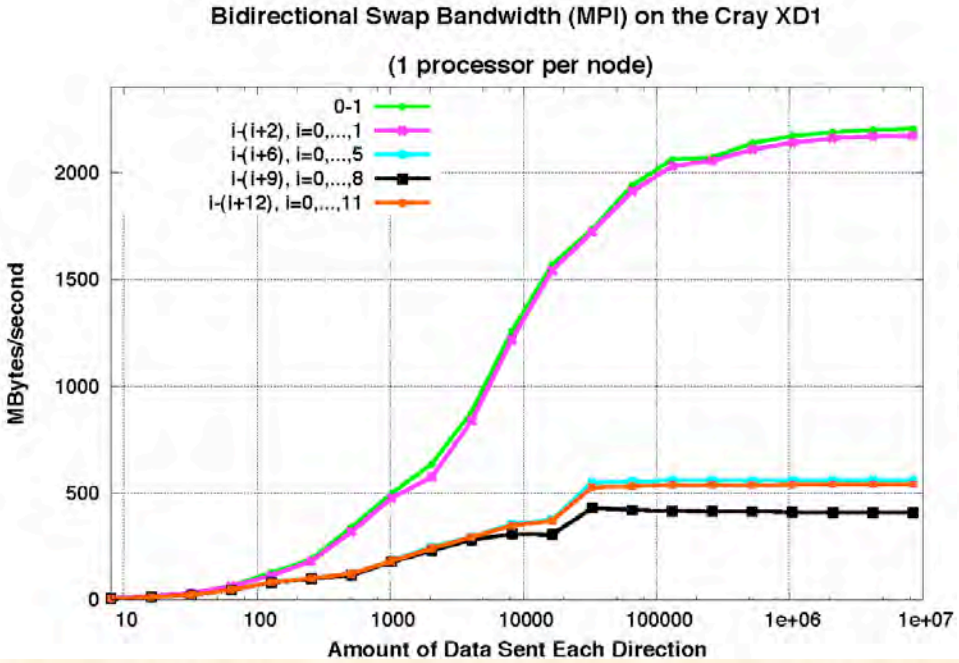
Contention: XD1 (w/expansion fabric)

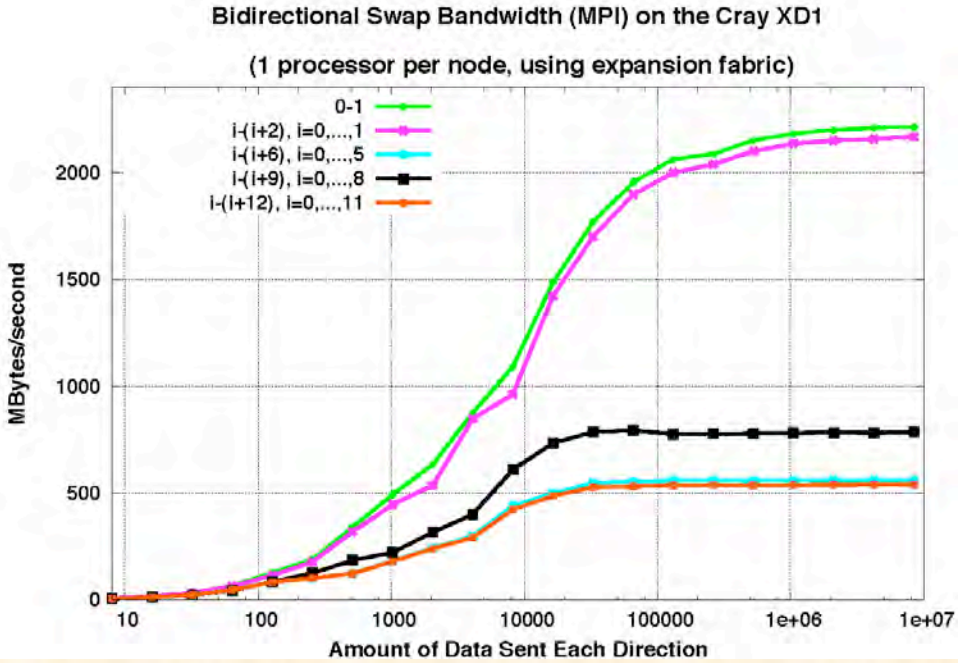




- When using 2 processors per node, using both the main and the expansion fabric achieves the same performance as when using only the main fabric.

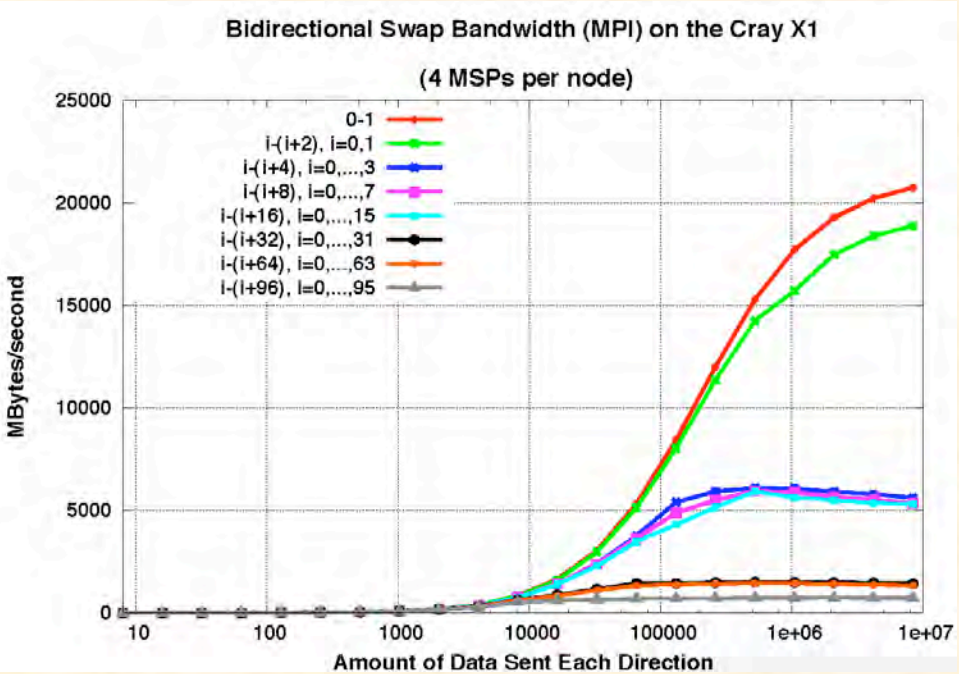
Contention: XD1 (w/expansion fabric, 1proc.)

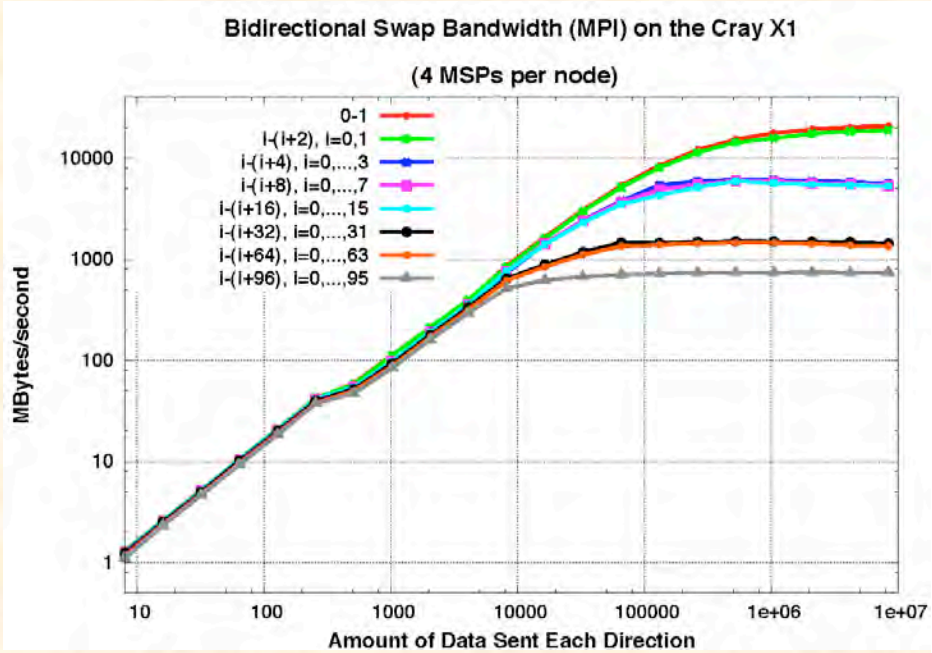




- When using 1 processor per node, using both the main and the expansion fabric achieves the same performance as when using only the main fabric in most contention experiments. However, using the expansion fabric doubles the performance for the 8-pair contention experiment for large message sizes.

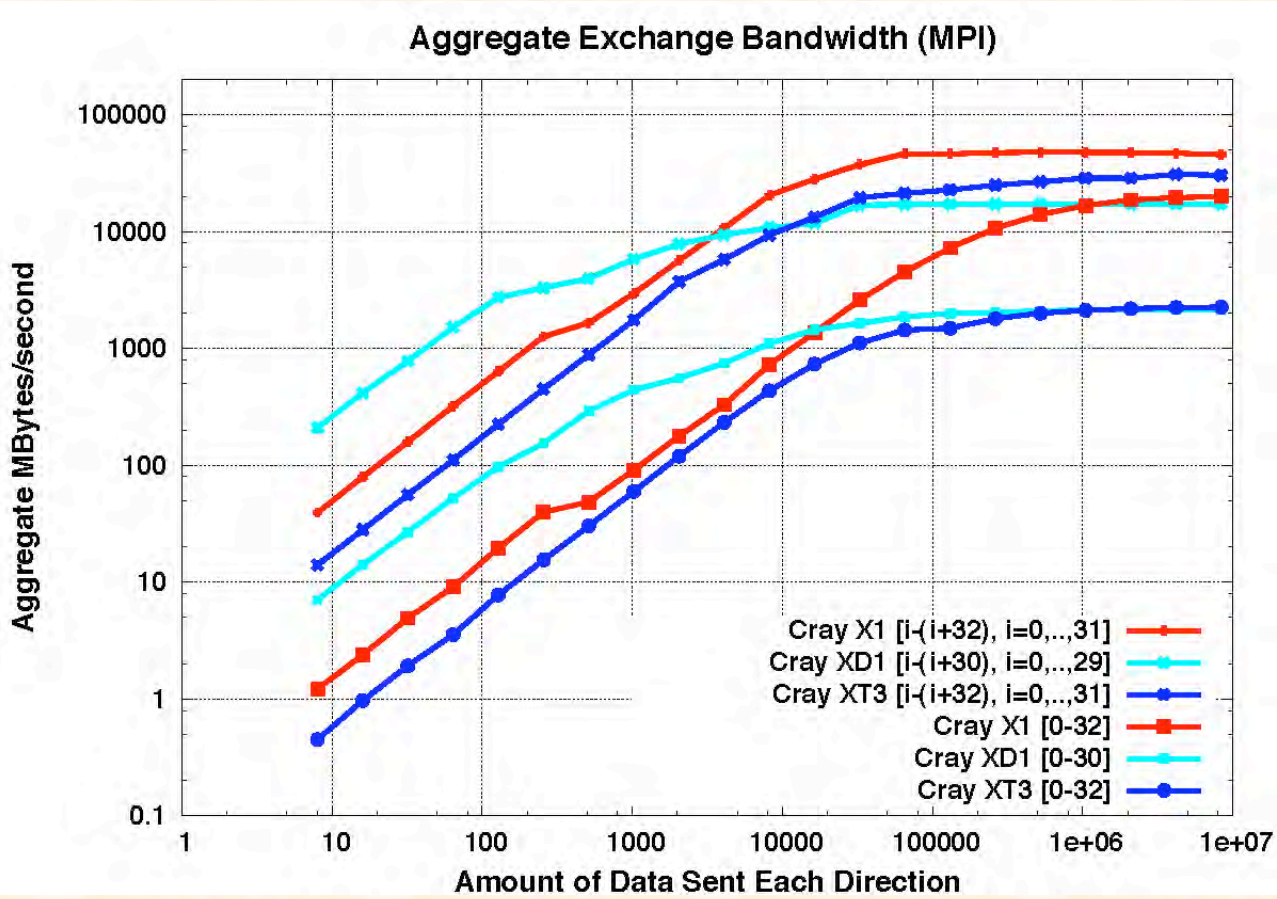
Contention: X1





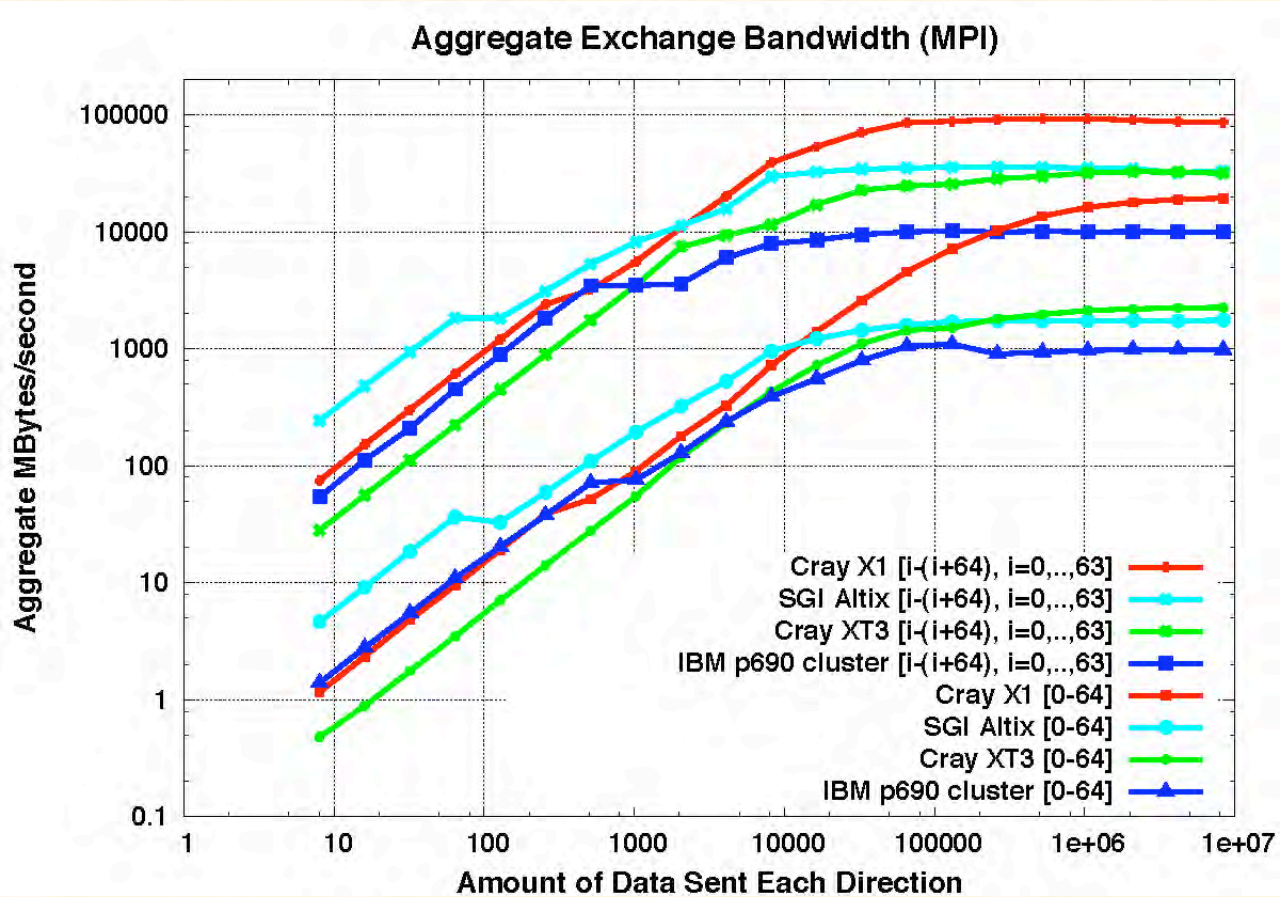
- Performance degradation due to contention is larger on the X1 than on the XT3, but the absolute performance is still better on the X1. Additional experiments are needed, but the conjecture, based on the coordinate direction experiments, is that the maximum XT3 contention has already been observed. This is not clear from the X1 data.

Contention: Platform Comparisons



- Comparing X1, XD1, and XT3 performance, for single pair and 32 simultaneous pairs experiments. (Used 30 pairs experiments for the XD1.) Contention experiments are plotted as aggregate bandwidth (i.e., 32 times worst case single pair performance). XD1 latency and X1 large message bandwidth advantages are easily observed. Also, XT3 bandwidth under contention is better than that for the XD1.

Contention: Platform Comparisons



- Comparing X1, XT3, Altix, and p690 cluster performance, for single pair and 64 simultaneous pairs experiments. Altix demonstrates best latency. (Compared to XD1 results on previous slide, XD1 latency is slightly better.) XT3 large message performance in similar to that for the Altix.

Contention Summary

- MPI bandwidth is affected by contention on all of the systems when multiple processors are communicating simultaneously.
- For XD1, using one processor per node achieves twice the bandwidth of using two processors for the same number of nodes.
- For XT3, performance details depend on contention “direction”.
- X1 MPI peak bandwidth is much higher than other Cray systems (as per specifications).
- XT3 MPI peak bandwidth is comparable to Altix for examined processor count.
- XD1 expansion fabric only enhanced communication performance in one of these experiments (one processor per node, particular number of nodes)
- Need to look at fat tree topology for XD1, to see whether it improves aggregate bandwidth.

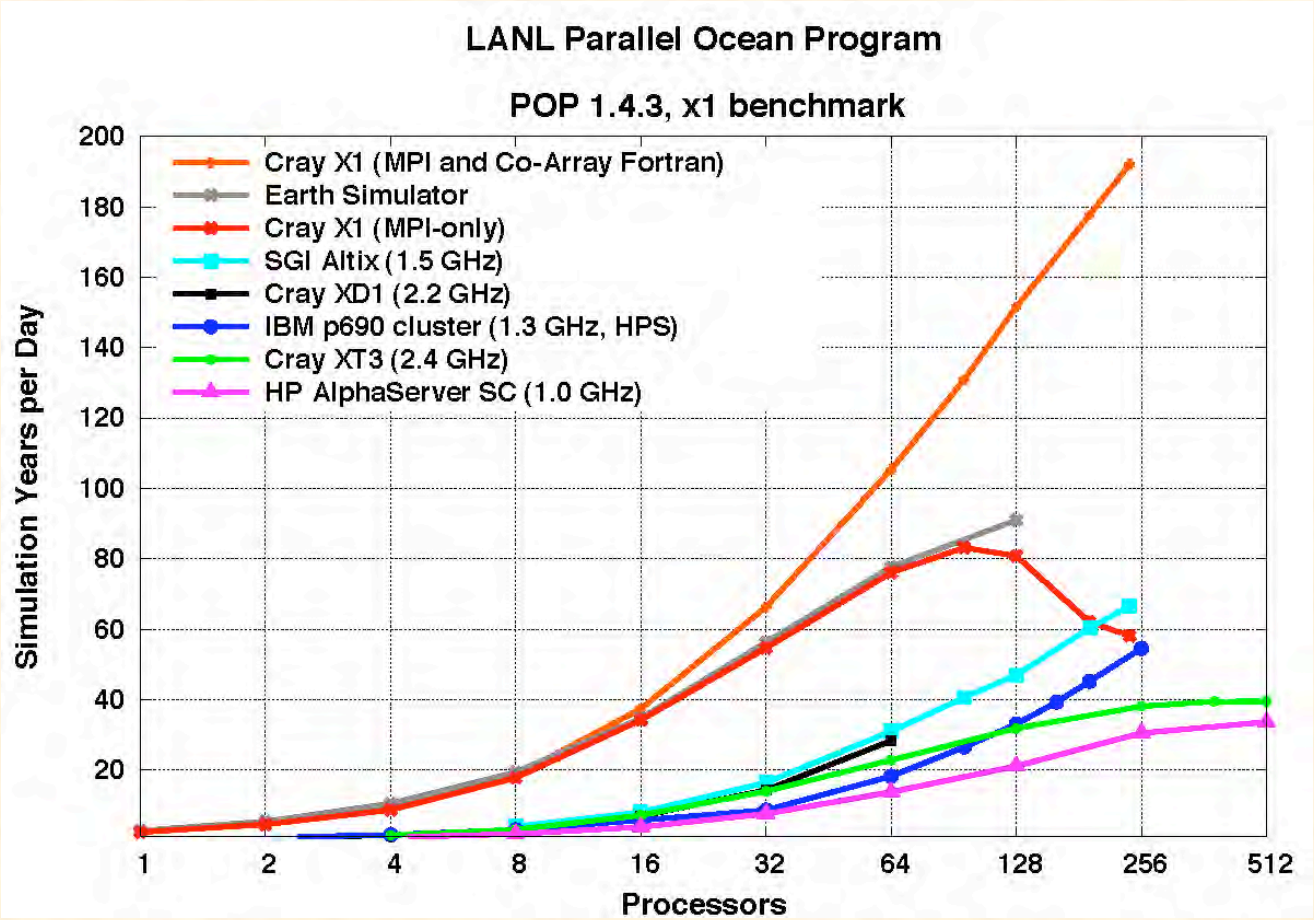
Parallel Ocean Program (POP)

- 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.
- The version of POP used in these experiments is a pure MPI code (i.e., does not use SMP parallelism). In the Cray X1 experiments POP is run with one process per MSP.

POP Experiment Particulars

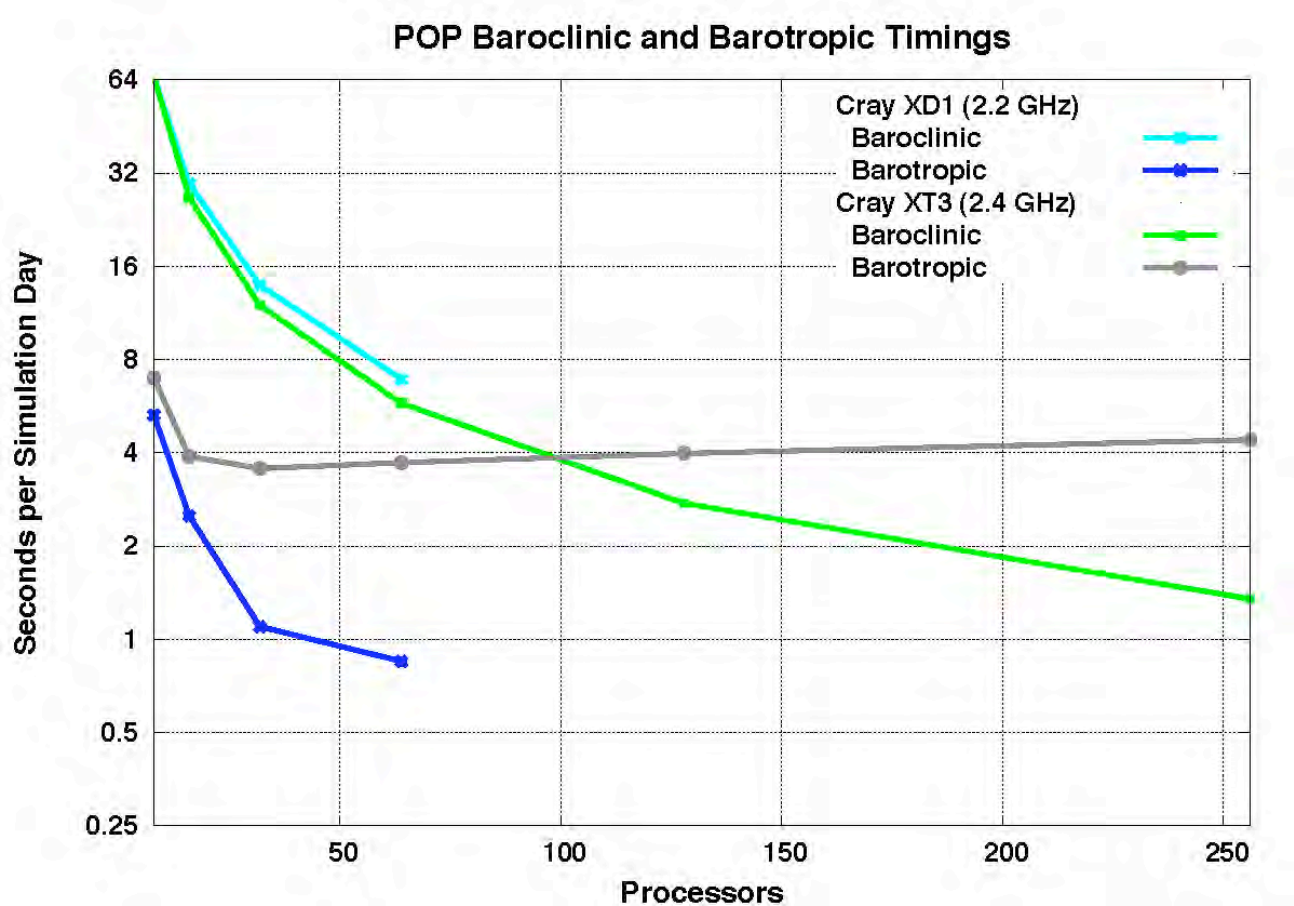
- Two primary computational phases
  - Baroclinic: 3D with limited nearest-neighbor communication; scales well.
  - Barotropic: dominated by solution of 2D implicit system using conjugate gradient solves; scales poorly due to communication overhead. Communication is dominated by residual calculations (halo updates) and inner product calculations (single word allreduce), so is primarily latency sensitive at scale.
- 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



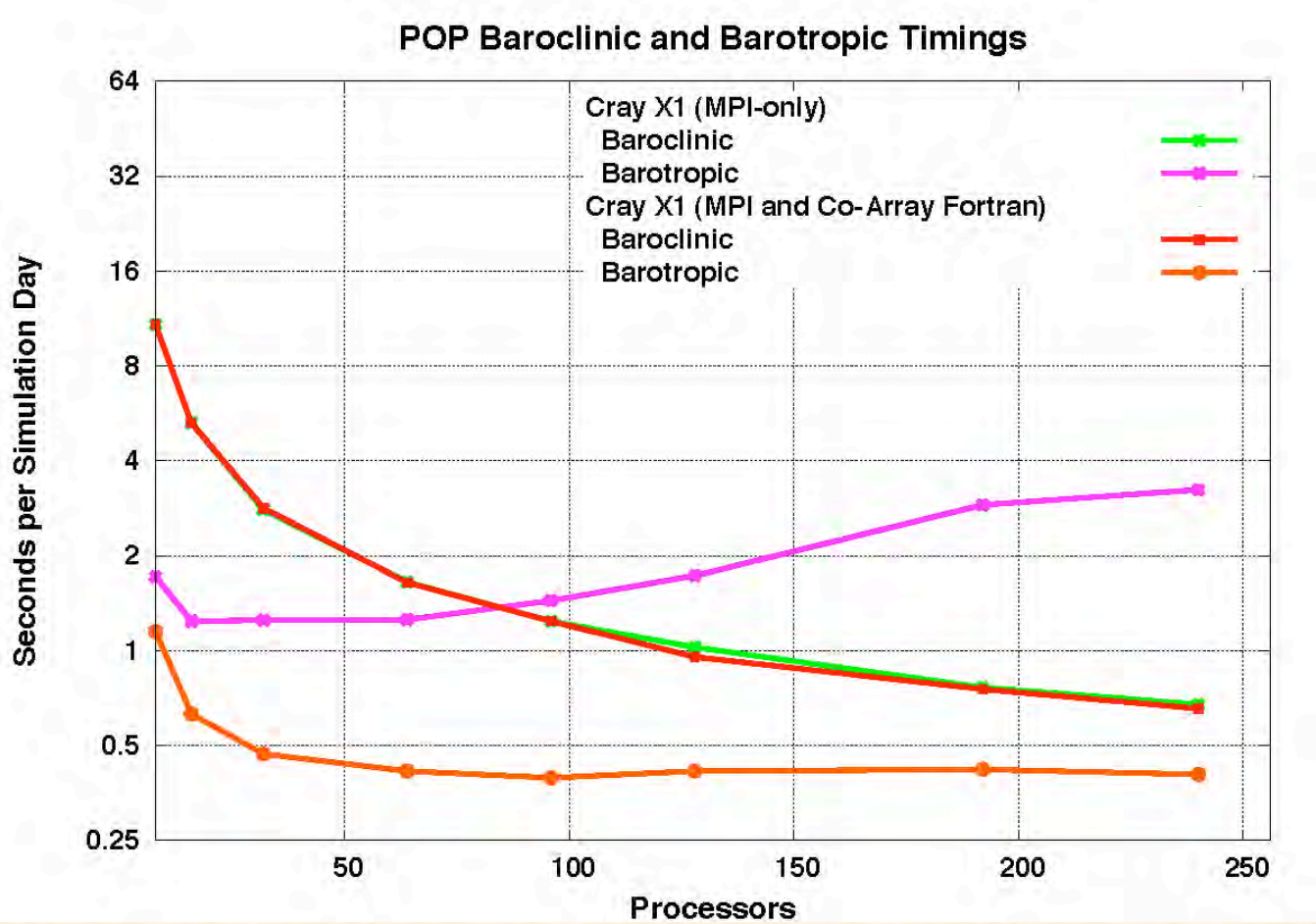
- Earth Simulator results courtesy of Dr. Y. Yoshida. X1 performance is excellent when using Co-Array Fortan, but scalability is limited when using only MPI. XD1 performance is very similar to Altix performance up to 64 processors. XT3 performance is severely limited by the current high latency.

POP Performance Diagnoses



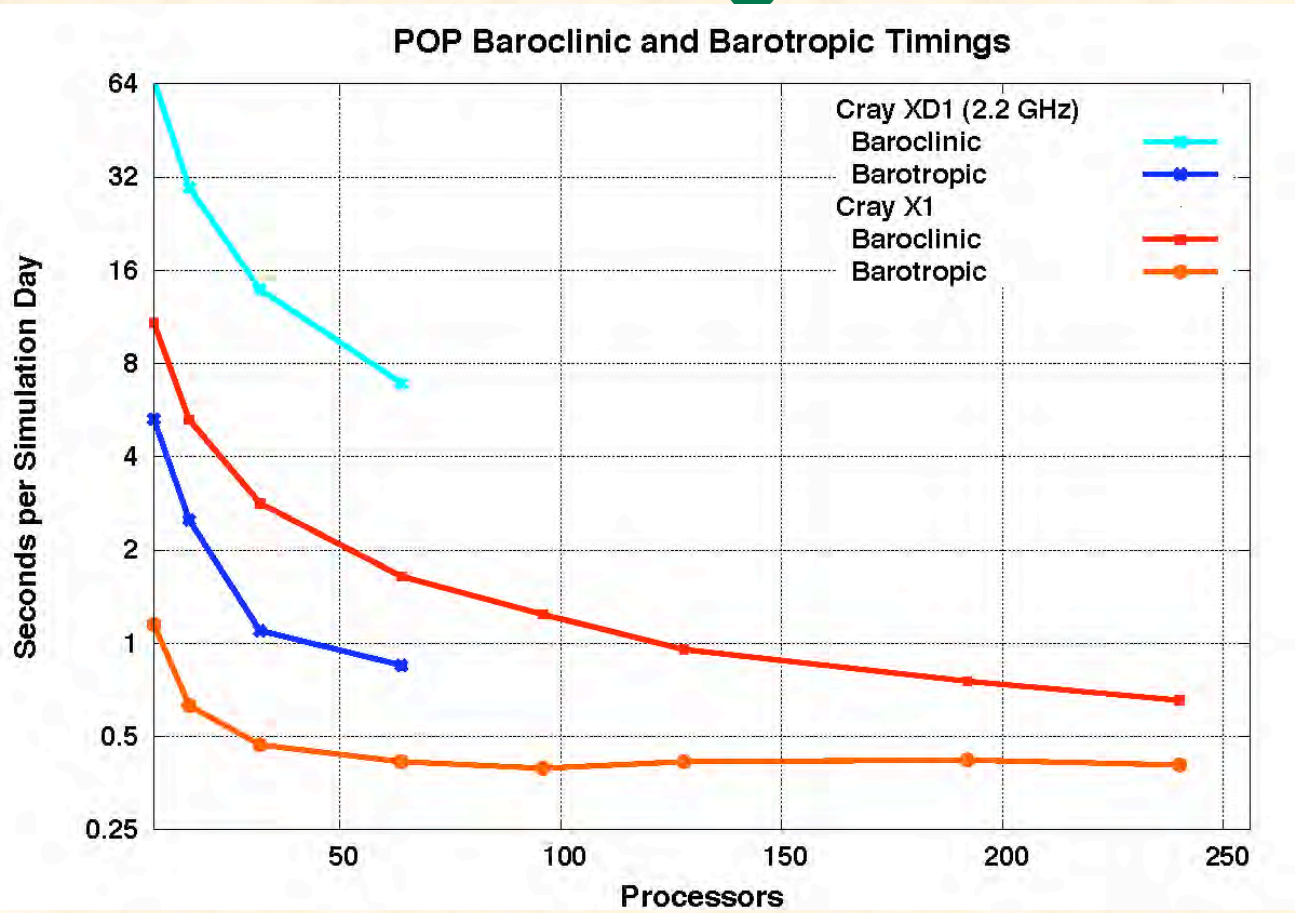
- Examining time spent in baroclinic and barotopic phases for the XD1 and XT3. The higher performance processor gives the advantage to the XT3 for the baroclinic. Lower latency gives the advantage to the XD1 for the barotropic. Note the crossover on the XT3 at approx. 100 processors, indicating that POP is communication bound for >= 128 processors on the XT3.

POP Performance Diagnoses



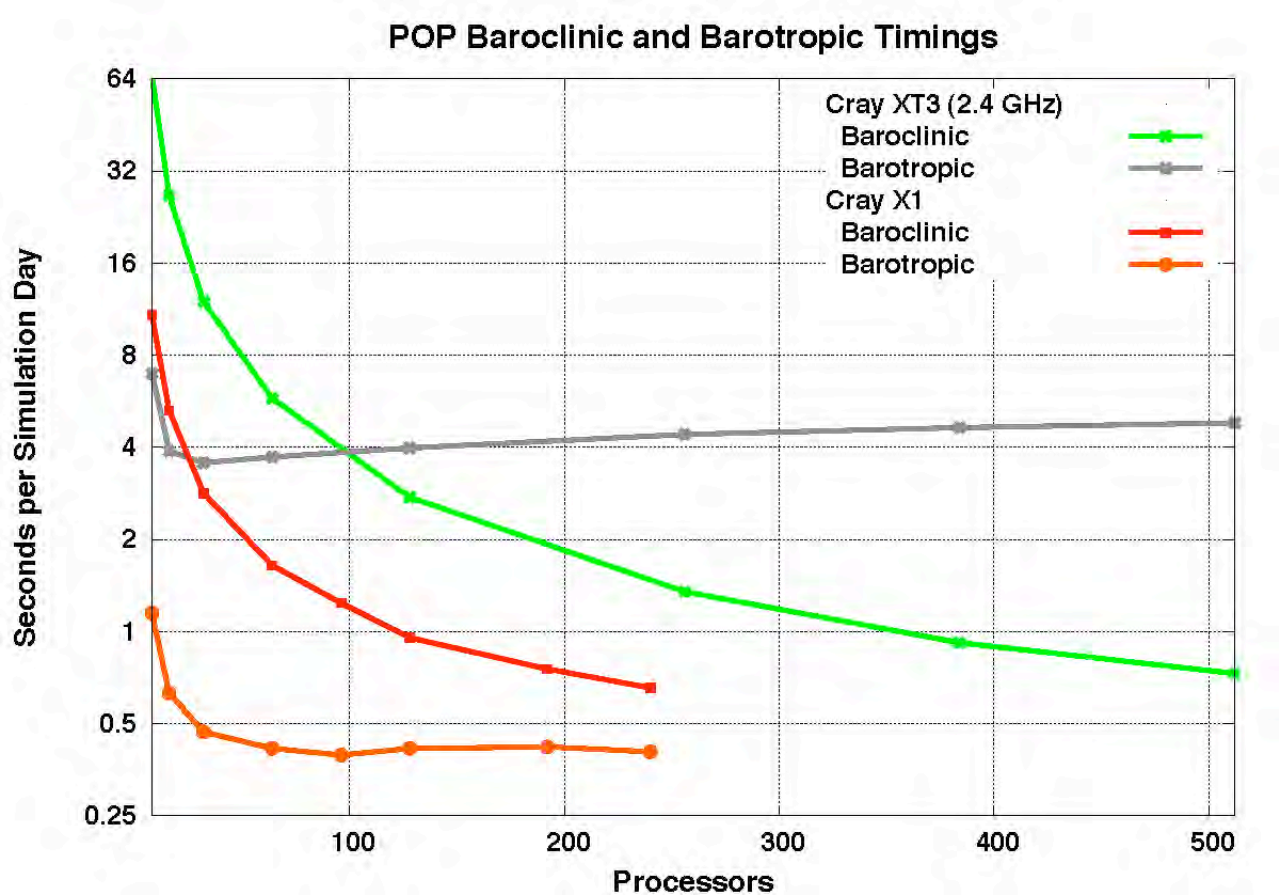
- Examining time spent in baroclinic and barotopic phases for the X1, with and without Co-Array Fortran. Note that Co-Array Fortran is used only to reimplement halo update and allreduce in the barotropic conjugate gradient solver. The Co-Array Fortran implementation is more than 3 times faster than the MPI implementation for 64 processors, and scales much better.

POP Performance Diagnoses



- Examining time spent in baroclinic and barotopic phases for the XD1 and X1 (with Co-Array Fortran). Barotropic is faster on the X1 than on the XD1. For the X1, the barotropic is dominated by communication overhead at 64 processors (MSPs). However, on the XD1 computation is still a significant part of barotropic time, so part of this performance difference is due to the difference in processor performance.

POP Performance Diagnoses



- Examining time spent in baroclinic and barotopic phases for the XT3 and X1 (with Co-Array Fortran). Baroclinic performance scales well on the XT3, and 512 processor performance on the XT3 is approaching that of the 256 processor performance on the X1. The problem size is fixed and relatively small and the vector length is becoming small on the X1 for large processor counts. This limits X1 processor performance at scale.

POP Performance Summary

- Small communication latency is required to achieve good scalability for the POP “x1” benchmark.
- Good performance on the X1 was achieved by using Co-Array Fortran to implement two collectives: allreduce and halo update.
- Good performance on the XT3 will not be possible until MPI (or SHMEM) latency is decreased.
- Performance of the barotropic phase on the XD1 is good, but is not scaling as well as expected. The performance of the allreduce and halo update need to be examined.

GYRO

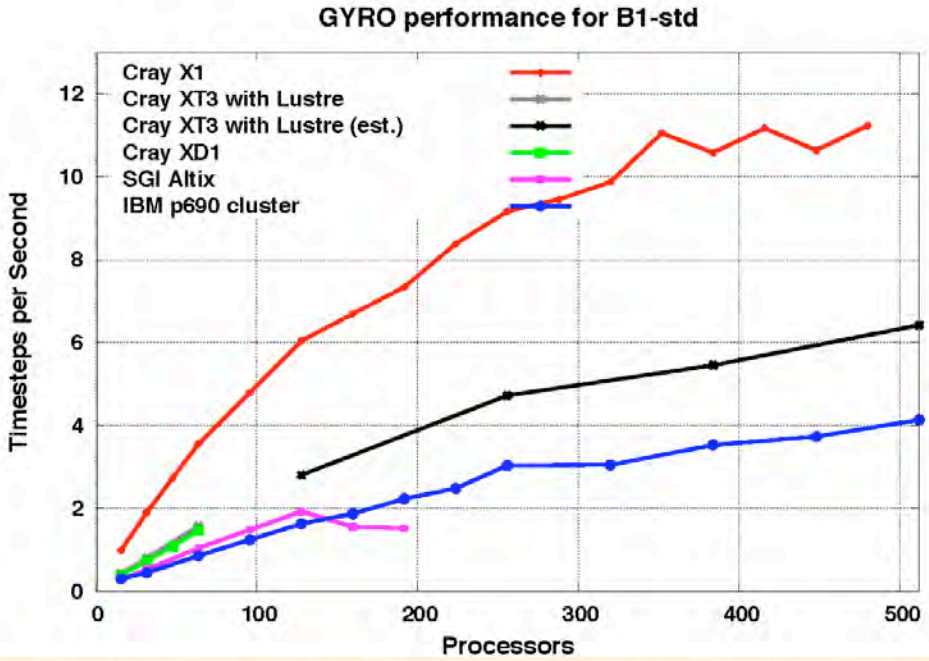
- GYRO is an Eulerian gyrokinetic-Maxwell solver developed by R.E. Walsh and J. Candy at General Atomics. It is used to study plasma microturbulence in fusion research.
- GYRO comes with ports to a number of different platforms. The port and optimization on the Cray X1 is primarily due to Mark Fahey of ORNL.
- GYRO is a pure MPI code (i.e., does not use SMP parallelism). In the Cray X1 experiments GYRO is run with one process per MSP.

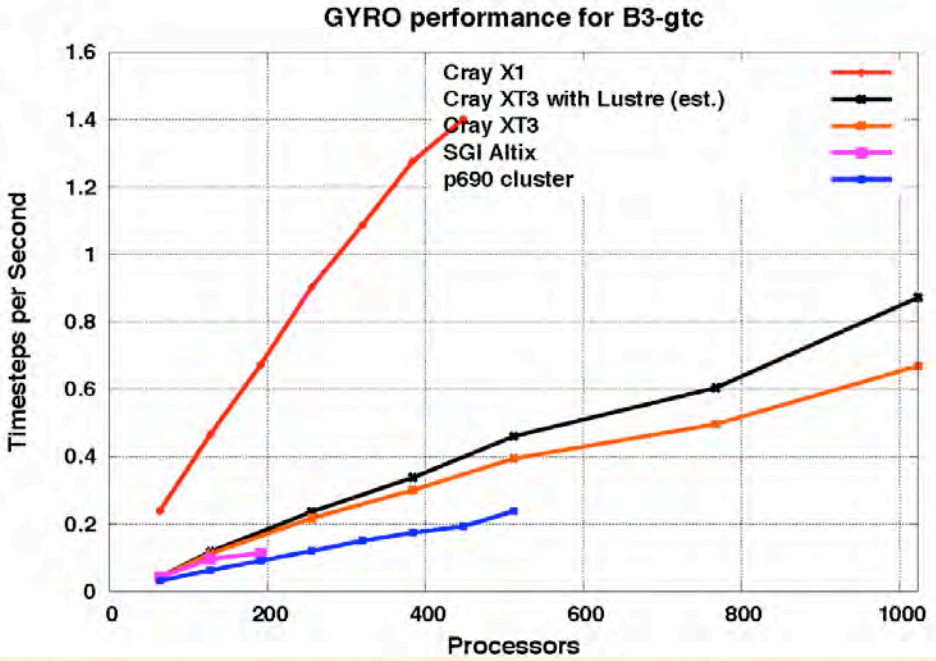
GYRO Experiment Particulars

Two benchmark problems, both time dependent:

1. B1-std
   - 16-mode simulation of electrostatic turbulence using kinetic electrons and ions and electron collisions. Duration is 500 timesteps.
2. B3-gtc
   - 64-mode adiabatic electron case. It is run on multiples of 64 processors. Duration is 100 timesteps.

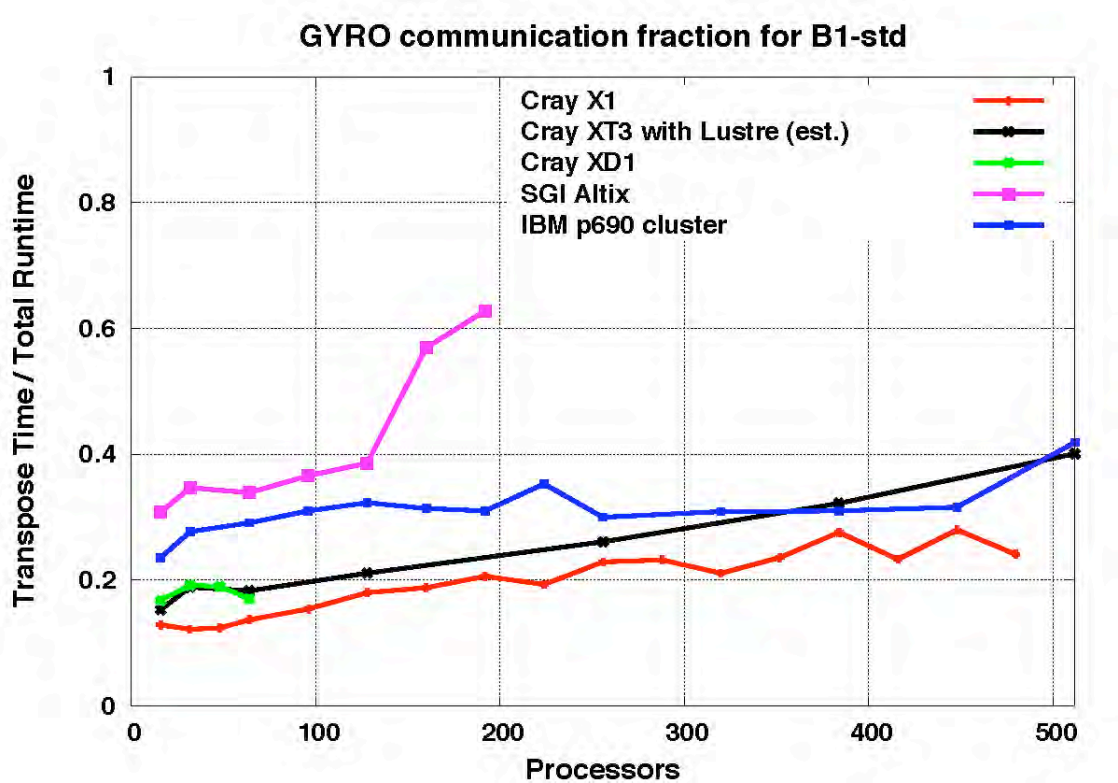
GYRO Platform Throughput Comparison





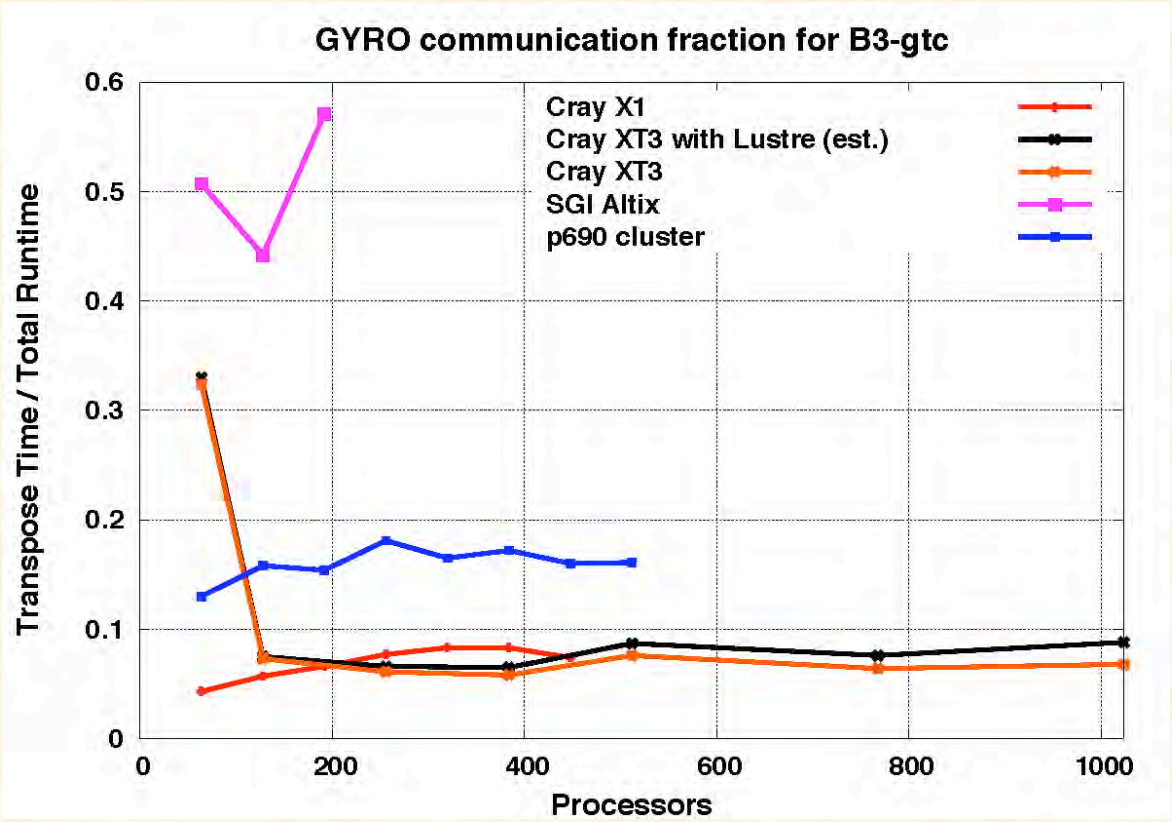
- Examining GYRO performance for both benchmarks on the Cray, IBM, and SGI systems. I/O on the XT3 is extremely slow currently. However, a 96 processor development system has a Lustre file system. Data collected on this system is also included in the B1-std figure, and the difference between Lustre and non-Lustre performance is used to predict the performance on the large XT3 system with a Lustre file system. While X1 performance is the best, the XT3 performance scales very well, especially with a Lustre file system. Note that I/O overhead is insignificant on all of the platforms except for the XT3 without Lustre. Using a higher performance file system on the other systems would not change the comparisons.

GYRO Communication Analysis: B1-std



- Examining the fraction of time spent in the transposes used in the parallel implementation of GYRO for the different platforms and for B1-std. As the entire computational domain is being remapped during the transposes, the communication is bandwidth limited for all but the largest processor counts. The transposes are implemented using MPI_Allreduce, so the efficiency of the collective implementation is also important. All of the systems are scaling well except the Altix. The advantage of the high bandwidth performance on the X1 is especially evident.

GYRO Communication Analysis: B3-gtc



- Examining the fraction of time spent in the transposes used in the parallel implementation of GYRO for the different platforms and for B3-gtc. As before, performance is scaling well on all systems except the Altix. The smallest processor count, representing the largest message sizes, causes a performance problem on the XT3, but it scales extremely well for larger processor counts, doing as well as the X1 (relative to the processor speed). Note that the current slow I/O on the XT3 is not as important to performance for this benchmark.

GYRO Performance Summary

- High bandwidth is required to achieve good scalability for the GYRO benchmarks.
- All three Cray systems provide sufficient bandwidth to allow good scalability, especially for the large B3-gtc benchmark. In particular, communication/computation ratio is similar for all Cray systems for the GYRO benchmarks.
- Communication performance behavior is very similar for the XD1 and XT3 for GYRO for B1-std. There is insufficient information to estimate scalability on larger XD1 systems.
- XT3 transpose fraction grows with processor count for B1-std somewhat faster than on the other systems, possibly representing an increasing sensitivity to latency as granularity decreases.
- 64 processor count performance on XT3 demonstrated unexpectedly large communication overhead for B3-gtc.

Talk Summary

- Simple technical specifications were relatively accurate predictors of relative communication performance between the X1, XD1, and XT3.
- Performance scalability aspects of all three networks are good, especially with regard to distance.
- XT3 latency needs to be improved (and Sandia and internal Cray results indicate that it will be much better within the next few months).
- XD1 bandwidth under contention and scalability are suspect for the direct connect topology. This may be improved by use of a fat tree topology. The expansion fabric improved performance in only a few instances in these experiments.