NAS Technical Report; NAS-06-012
August 2006

Interconnect Performance Evaluation of SGI Altix 3700 BX2, Cray X1, Cray Opteron Cluster, and Dell PowerEdge

Rod Fatoohi[1], Subhash Saini[2], and Robert Ciotti[2]

[1]Computer Engineering Department
San Jose State Univeristy
One Washington Square
San Jose, California 95192
rfatoohi@sjsu.edu

[2]Advanced Supercomputing Division
NASA Ames Research Center
Moffett Field, California 94035
Subhash.Saini@nasa.gov,
ciotti@nas.nasa.gov

Abstract

We study the performance of inter-process communication on four high-speed multiprocessor systems using a set of communication benchmarks. The goal is to identify certain limiting factors and bottlenecks with the interconnect of these systems as well as to compare these interconnects. We measured network bandwidth using different numbers of communicating processors and communication patterns - such as point-to-point communication, collective communication, and dense communication patterns. The four platforms are: a 512-processor SGI Altix 3700 BX2 shared-memory machine with 3.2 GB/s links; a 64-processor (single-streaming) Cray X1 shared-memory machine with 32 1.6 GB/s links; a 128-processor Cray Opteron cluster using a Myrinet network; and a 1280-node Dell PowerEdge cluster with an InfiniBand network. Our results show the impact of the network bandwidth and topology on the overall performance of each interconnect.

1. Introduction

The message passing paradigm has become the de facto standard in programming high-end parallel computers. The performance of real-world applications depends on the performance of the Message Passing Interface (MPI) functions implemented on these systems. Bandwidth and latency have traditionally been used as two metrics in the assessing the performance of the interconnect fabric of the system. However, these two metrics are not adequate to determine the performance of real-world applications. Computer vendors highlight the performance of network by latency using zero byte message sizes and peak bandwidth for very large message sizes ranging from 2 MB to 4 MB for a very small system typically 32 to 64 processors. Real world applications tend to send messages ranging from 10 KB to 2 MB using not only point-to-point communication but using all possible communications patterns, including collective and reduction patterns.

In this paper, we focus on the communication network of four state-of-the-art high-speed multiprocessors with different network speeds and topologies. Two of these systems (SGI Altix BX2 and Cray X1) are shared memory machines while the other two (Cray Opteron and Dell PowerEdge) are distributed-memory machines – clusters of dual-processor computers. Two of these platforms use custom networks (SGI Altix 3700 and Cray X1) while the other two platforms employ commercial networks (Cray Opteron and Dell PowerEdge). We used three different benchmarks to get a better insight into the performance of four different networks. Our benchmarks measure the unidirectional and bidirectional bandwidth of communication links, collective communication and dense communication patterns.

Recently, several performance evaluation studies have been conducted of Cray X1 and SGI Altix 3700/BX2 supercomputers, mainly at NASA Ames Research Center and Oak Ridge National Laboratory (ORNL) where some of these machines are located. The focus of most of these studies has been on the overall performance of these systems including floating point operations, memory bandwidth, message passing and using several kernels as well as scientific applications. The results of two studies conducted at NASA Ames by Biswas, et al. [1], [2] indicate close performance between the SGI Altix 3700 BX2 and the Cray X1 for several micro-benchmarks, kernels, and applications. Among several performance studies conducted at ORNL, Dunigan, et al. [4] found that the Altix 3700 is competitive with the Cray X1 on a number of kernels and applications. Another study at ORNL by Worley, et al. [11] focuses on recent Cray products: X1/X1E, XD1,

and XT3 with an emphasis on the inter-process communication. Their study shows that the X1 communication bandwidth is significantly better than that of the other two systems while MPI latency is unimpressive on the X1 and very low on the XD1.

2. Interconnect Networks

The SGI Altix BX2 system [10] is a 512-processor global shared memory architecture with one Tbytes of memory, a peak performance of 3.28 Tflops, running the Linux operating system. It is one node of the 20-node super-cluster (with a total of 10,240 processors), called Columbia, located at NASA Ames. The Altix 3700 BX2 is essentially a double-density version of the 3700 – doubling the number of processors, memory size, and link bandwidth. Each processor is an Intel Itanium-2 64-bit processor and runs at 1.6 GHz clock. The Altix 3700 system is built from a number of component modules called bricks. The compute brick (called C-brick) on the Altix BX2 system contains 8 processors, 16 Gbytes of local memory, and 4 ASICs called Scalable Hub (SHUB). Each SHUB interfaces with the processors, memory, I/O devices, other SHUBs, and an interconnection network called NUMAlink4. The NUMAlink4 interconnect is a custom SGI network with a fat-tree topology and a peak bandwidth of 6.4 Gbytes/s. Within a C-brick, the SHUBs and each pair of processors are connected internally by a 6.4 Gbytes/s bus. In addition to the C-bricks, the BX2 system has I/O modules (called IX-bricks) and router modules (called R-bricks). The R-bricks are used to build the interconnect fabric between the C-bricks. There are 48 R-bricks in the 512-processor BX2 system with two levels: 32 R-bricks in level 1, which are directly connected to the 64 C-bricks, and 16 R-bricks at level 2, which are connected to the R-bricks of level 1.

The Cray X1 at NASA Ames contains 64 single streaming processors (SSPs) configured into four separate nodes and 64 Gbytes of memory with a peak performance of 204.8 Gflops and running the UNICOS operating system. Each node has four multi-streaming processors (MSPs) sharing a flat memory through 16 memory controllers, called MChips. Each MSP has four SSPs sharing a 2 Mbyte cache. The machine at NASA Ames is configured with one node used for system purposes while the remaining three nodes are available for computing. Each node has 32 network ports; each port supports 1.6 Gbytes full duplex links. A 4-node system can be connected directly through the MChips while larger systems use a 4-D hypercube or a modified 2-D torus. An X1 application can run in either SSP mode or MSP mode, through a compiler directive. In SSP mode, each SSP runs independently of the others, executing its own stream of instructions; in MSP mode, each MSP closely couples the interactions of its four SSPs and distributes the parallel parts of an application to its SSPs.

The Cray Opteron cluster at NASA Ames has 64 nodes with 130 Gbytes of memory, a peak performance of 512 Gflops, and runs on the Linux operating system. Each node has two AMD Opteron 246 series processors running at 2.0 GHz. The machine is configured with one node used as the server node and the remaining 63 nodes (126 processors) used as compute nodes. The nodes are interconnected via a Myrinet network. Myrinet [3] uses cut-through routing and remote memory direct access to write to/read from the remote memory of other host adapter cards, called Lanai cards. The Myrinet PCIXD cards run at 225 MHz with 2 MB memory plugged into the Newisys 133 MHz PCI-X slot. The limit of the PCI-X bus is 1067 Mbytes/s, and the NIC cards are capable of sustained PCI data rates.

The Dell PowerEdge 1850 cluster at the National Center for Supercomputer Applications (NCSA), called Tungsten 2, has 1280 nodes with 7.68 Tbytes of memory, a peak performance of 9.2 Tflops/s and running the Linux operating system. Each node has two Intel Xeon EM64T 3.6 GHz processors, 6 Gbytes of memory, and PCI-X IB card in a 133 MHz slot. The nodes are interconnected with a high-speed InfiniBand fabric. InfiniBand [8] is a bit-serial switched network with a raw data rate of 250 Mbytes/s in each direction per serial link. The nodes in the cluster use four serial links run in parallel giving a peak data rate of 1 Gbytes/s in each direction. The InfiniBand adapters are connected to the system through a PCI-Express X8 slot which has a theoretical bandwidth of 2 Gbytes/s. The InfiniBand fabric design is based on a two-tier network with edge switches connecting the hosts and core switches comprising the backbone of the fabric.

3. Results

3.1. Effective Bandwidth Benchmark

We used the effective bandwidth benchmark [9] to measure the accumulated bandwidth of our networks. The benchmark employs several message sizes, communication patterns and methods where the result is a single number, called the effective bandwidth (b_eff). It is defined as: a) a logarithmic average over several ring patterns (a total of 6) and random patterns, b) using the average of different message sizes (a total of 21 sizes ranging from 1 byte to 1/128 of the memory of each processor), and c) the maximum over three communication methods (MPI_Sendrecv; MPI_Alltoallv; and non-blocking with MPI_Irecv, MPI_Isend, and MPI_Waitall). A fundamental

difference between this benchmark and the PingPong benchmark is that here all processes are sending messages to neighbors in parallel.

Table 1 shows the b_eff benchmark results on the four platforms using different numbers of processors. In addition to reporting the measured b_eff using different patterns and message sizes (3rd column), the benchmark measures b_eff at the maximum message size L_{max} (1 Mbytes for all cases) using ring and random patterns (5th column), b_eff at L_{max} using ring patterns only (7th column), the point-to-point bandwidth (ping-pong) measurement (9th column) and the latency measurement (10th column). The b_eff per processor results (4th, 6th and 8th columns) extrapolate to the network performance if all processors are communicating to their neighbors.

One way to interpret the results of Table 1 is a comparison across platforms for a specific measurement (horizontally). The latency results (last column) show that the Cray Opteron has the lowest latency (about 0.7 μsec) while the Cray X1 (in both modes) has the highest latency (about 10 μsec) among the four platforms - a similar observation about the Cray X1 latency was reported in [11]. The PingPong results (9th column) show that the Cray X1 has the highest link bandwidth (over 9 GB/sec in MSP mode and over 4 GB/sec in SSP mode). In the MSP mode, it outperformed the Altix 3700 BX2, the Cray Opteron, and the Dell PowerEdge by factors of about 9, 13, and 23, respectively. The b_eff results (3rd column) show that with respect to the effective bandwidth of the whole system the 512-processor Altix 3700 BX2 outperformed a 48-processor Cray X1, the 128-processor Dell PowerEdge, and the 64-processor Cray Opteron by factors of about 9, 10, and 26, respectively.

Another way to interpret the results of Table 1 is a comparison across different measurements for a specific platform (vertically). Comparing the ping-pong results (9th column) with the b_eff at L_{max} per processor using ring patterns only (8th column), we observe the impact of communicating in parallel on each processor which is quite significant on the Cray X1 in SSP mode (a factor of over six using 32 processors) while it is only 64% on the Altix 3700 BX2 (for both configurations). Another comparison is between b_eff at L_{max} per processor using ring patterns only (8th column) and its value using rings and random patterns (6th column) to show the effect of random neighbor locations where we notice a drop of about 50% on the 64 processor Cray Opteron while the Cray X1 in the SSP mode shows no degradation. Yet another comparison is between b_eff at L_{max} using ring and random patterns (6th column) and the overall b_eff per processor (4th column) to show the impact of different message sizes, where we notice significant drops for all systems, as the overall b_eff is an average over several message sizes. These drops range between a factor of 4.6 for the Cray X1 in the MSP mode and 2 for the 64-processor Cray Opteron.

As the number of processors increases for the same platform, the b_eff per processor decreases but by different factors. It decreases by 20% as the number of processors doubled on the Altix 3700 BX2, while it decreases by 30% as the number of processors increased by a factor of eight on the Cray Opteron.

3.2. Intel MPI Benchmark

We used the Intel MPI Benchmarks (IMB) suite [7] for both point-to-point communication and collective communication. We employed five IMB benchmarks: *PingPong, PingPing, Barrier, Reduce,* and *Alltoall*. The PingPong benchmark measures the point-to-point bandwidth of a message sent between two processes using MPI_Send and MPI_Recv. The PingPing benchmark also measures the point-to-point bandwidth of a single message but when the message is obstructed by oncoming messages. Here the two processes communicate with each other using MPI_Isend, MPI_Recv, and MPI_Wait with the two MPI_Isend functions issued simultaneously. The expected number of the latter is between half and the whole amount of the former. We call the former the unidirectional bandwidth and the later the bidirectional bandwidth (our bidirectional bandwidth is about ½ of the aggregate bidirectional bandwidth normally reported by vendors). The Barrier, Reduce, and Alltoall benchmarks measure the MPI_Barrier, MPI_Reduce, and MPI_Alltoall functions, respectively.

Figure 1 shows the unidirectional and bidirectional benchmark results for different message sizes on the four platforms. The Cray X1 in the MSP mode achieved a rate of over 13 Gbytes/s using PingPong and a drop of less than 30% due to oncoming messages. On the other hand, the Dell PowerEdge with the InfiniBand network achieved a rate of about 400 Mbytes/s using PingPong with a drop of 50% due to oncoming messages. The Altix 3700 BX2 achieved a rate of about 1800 Mbytes/s using ping-pong with a drop of about 50% due to oncoming messages mainly for large messages. In comparing the modes of the Cray X1, we noticed a three-plus factor difference between MSP and SSP modes, because in MSP mode the Cray X1 can use four times the number of ports than in SSP mode. We also noticed that the best performance on the Dell PowerEdge and Cray Opteron was achieved with messages of sizes 16K and 128K bytes, respectively, due to switch buffering.

We measured the point-to-point data rate as we varied the distance between the two communicating processors on both the SGI Altix 3700 and Cray Opteron. Figure 2 shows the unidirectional bandwidth

(using PingPong) results measured on the 512-processor Altix 3700 for nine cases ranging from a distance, between the communicating processors, of 1 to 256 (the farthest two communicating processors). As mentioned earlier, the 512-processor BX2 consists of 64 C-bricks with each C-brick containing four nodes and each node with two Itanium-2 processors. Figure 2 shows the differences in transfer rate whether communication is between processors on the same node (distance of one), on the same C-brick (distances of two and four), or between C-bricks (distances of 8, 16, 32, 64, 128, and 256). Clearly, the highest rate achieved is between processors on the same node. Interestingly, the highest rates achieved are for messages of size either 1 or 2 Mbytes while it drops (by as much as 1/3 for a distance of one) for the 4 Mbytes message. The highest measured rates are: 1762, 1264, 1191, 1097, 1016, and 917 Mbytes/s for distances of 1, 2 or 4, 8 or 16, 32 or 64, 128, and 256, respectively. The rate drops for longer distances (more than 4) can be attributed to the number of the R-bricks (routers) that the message has to travel between C-bricks.

Figure 3 shows the results of distance sensitivity on the Cray Opteron for distances between communicating processors of 1, 2, 4, 8, and 16. Similar to the SGI Altix 3700, each node has two processors (AMD Opteron 246 series) so communication of distance one stays within the node. The results show that a rate of about 900 Mbytes/s is achieved with a distance of one for the 128 Kbytes message. This rate drops to 670 Mbytes/s (by about 25%) for the 4 Mbytes message with the same distance (one). For all other distances (2 to 16) the rate is about 234 Mbytes for large messages – a drop of 2/3 from distance one rate. Interestingly, the measured results for all messages with a distance of more than one are about the same which is an indication of distance insensitivity for the Myrinet network.

In comparing the PingPong results of the b_eff benchmark (8^{th} column of Table 1) with the IMB results (Figures 1 through 3), we noticed some differences largely due to the message size and location of the communicating processes. In Table 1, we reported a single value for PingPong, which is the measured bandwidth between the processes with rank 0 and 1 in MPI_COMM_WORLD using a 1 Mbyte message, while Figures 1 through 3 show a range of values for different messages and communicating partners.

The three collective operation functions measured (MPI_Barrier, MPI_Reduce and MPI_Alltoall) are used extensively in many applications [6]. The MPI_Reduce function implements an all-to-one reduction operation, where each process sends a message of size M to a single process and data from all processes are combined through an associative operator at the single destination process into a buffer of size M; this function is used in many parallel algorithms such as matrix-vector multiplication, vector-inner product, and shortest paths. The MPI_Alltoall function implements all-to-all personalized communication (also called total exchange) operation, where each process sends a distinct message to every other process; it is used many parallel algorithms such as fast Fourier transform, matrix transpose, sample sort, and some parallel database join operations. The MPI_Barrier function implements a synchronization point, where each process is held until all other participating processes have reached the barrier; it is heavily used in parallel algorithms as well as in debugging. The performance of these functions reflects not only the richness of the network (in latency, bandwidth and topology) but also the efficient implementation, by the vendor, of optimized communication libraries.

Figures 4 through 6 show the measured timings of these functions on the four platforms for three message sizes 8, 1K, and 1M bytes (for the last two functions only). The results for MPI_Barrier (Figure 4) show that the shared memory systems (SGI Altix 3700 and Cray X1) perform much better than the distributed memory systems (Dell PowerEdge and Cray Opteron), even though the Cray Opteron has a very low latency. For example, on 64 processors, the Altix 3700 BX2 runs more than six times faster than the Dell PowerEdge and more than 13 times faster than the Cray Opteron using MPI_Barrier. The results for MPI_Reduce (Figure 5) show the Cray X1 outperforming the other three platforms for the three message sizes, even in SSP mode. Using 32 processors and one Mbytes message, for example, the Cray X1 in the SSP outperformed the Altix 3700 BX2, Dell PowerEdge, and Cray Opteron by factors of 10, 8.6, and 20, respectively. The Cray X1 also outperformed the other platforms using MPI_Alltoall (Figure 6), but the performance gap between the X1 and the Altix 3700 BX2 is narrower than for MPI_Reduce, especially for the large message.

3.3. Dense Communication Benchmark

We used the dense communication benchmark [5] to evaluate our networks with multiple processors communicating in parallel using four different intense communication algorithms (also referred consecutively as algorithms 1 through 4): congested-controlled all-to-all personalized communication (AAPC); simple pair-wise; cumulative pair-wise; and random pair-wise. In algorithm 1, each process sends data to its next higher neighbor (in rank) and receives data from its next lower neighbor. The algorithm proceeds in phases such that the distance between the communicating processes increases in each phase till the last phase where every process sends data to its lower neighbor and receives

data from its higher neighbor. In algorithm 2, a set of processes communicates in pairs and all pairs send and receive data in parallel and at full duplex. The algorithm proceeds in phases, as in the first algorithm, with the distance between the communicating processes increases in each phase until it reaches its maximum (the total number of processes minus one). Algorithm 3 is similar to algorithm 2 except that the number of the communicating pair is increased during successive phases of communication with only one pair communicating in the first phase and all pairs communicating in the last phase. Finally, in algorithm 4, all processes communicate in pairs as in algorithm 2 but the processes are shuffled for the next phase so as different pairs are formed in each phase. Here the number of phases is chosen at run time.

We implemented the four algorithms on the four platforms using different number of processors. Figures 7 through 10 present some of these results with the emphasis on the cases that have some significance for the sake of brevity. Several observations can be drawn from these results. First, algorithms 1 and 2 demonstrated similar behavior on all platforms with drops in the middle phases (farthest communication distances) compared to the first and last phases (shortest communication distances). These drops range from more than a factor of 5 (Altix 3700 BX2 and Cray Opteron) to a factor of 2 (Cray X1). In some cases, the Cray Opteron, for example, there is a drop of about 40% between phase one and phase two, because after the first phase, all communications are through the Myrinet network. Second, in many cases the highest obtained rates are not for the largest messages, such as the 3Kbyte message on the SGI Altix 3700 using algorithm 1, mainly related to message buffering. Third, for algorithm 3 both the Altix 3700 BX2 and Dell PowerEdge showed small drops of up to 20% as the number of communicating pairs increased to 256, especially for large messages while the Cray Opteron showed no drops for all messages. On the other hand, the Cray X1 in SSP mode showed a drop of up to a factor of 3 when the number of communicating processors increased to 24 pairs, which shows a typical bottleneck for many shared-memory architectures. Finally, all platforms showed the impact of randomness of communicating pairs on the measured bandwidth as demonstrated in algorithm 4. The impact of randomness was also noticeable in the b_eff benchmark (Table 1).

4. Conclusions

Our study provides a better understanding of certain limitation of interconnects on high-speed computers. The study has shown the relative speed of network links and how it is impacted under different circumstances. For example, we noticed that the Cray Opteron has the lowest latency, the Cray X1 has the highest link bandwidth while the effective bandwidth of the Cray X1 per processor is much lower than its link bandwidth (by a factor of over 6).

In studying the impact of oncoming message on the link bandwidth, we noticed that all systems (Altix 3700 BX2, Cray X1 in SSP mode, Cray Opteron, and Dell PowerEdge) experienced a drop of about 50% for large messages except the Cray X1 in MSP mode. The study has also demonstrated the distance sensitivity of point-to-point communication. It showed a drop in bandwidth as distance increases between the two communicating processors. For example, a drop of almost 50% was observed on the Altix 3700 BX2 when the distance between communicating processors is increased from one to 256. An even larger drop (2/3) was noticed on the Cray Opteron when the distance increased from one to 16. The results of three widely used MPI collective communication functions showed that the shared-memory machines (Cray X1 and Altix 3700 BX2) outperformed the distributed-memory machines (Cray Opteron and Dell PowerEdge) especially for MPI_Barrier. For MPI_Reduce and MPI_Alltoall, the Cray X1 outperformed the other platforms significantly.

Finally, our study reports the impact of intense communication patterns on these interconnects. In all cases, there were significant drops in performance as all processors communicated in parallel and away from each other. Another significant drop was observed on the Cray X1 when the number of communicating processors increased from one pair to 24 pairs.

Acknowledgements

We would like to thank Mindy Wilson and Johnny Chang of NASA Ames for accessing and running code on NASA machines. We are also grateful to NCSA for providing us access to the Dell PowerEdge cluster.

References

[1] R. Biswas, J. Djomehri, R. Hood, H. Jin, C. Kiris, and S. Saini, An Application-Based Performance Characterization of the Columbia Supercluster, In *SC|05*, (Seattle, Washington), November 2005.

[2] R. Biswas, S. Saini, S. Gavali, H. Jin, D. Jesperson, M. Djomehri, N. Madavan, and C. Kiris, NAS Experience with the Cray X1, In *47th Cray User Group Conf.* (Albuquerque, New Mexico), May 2005.

[3] N. Boden, D. Cohen, R. Felderman, A. Kulawik, C. Seitz, J. Seizovic, and W. Su, Myrinet: A Gigabit-persecond Local Area Network, *IEEE Micro*, Vol. 15, No. 1, February 1995, pp. 29 – 36.

[4] T. Dunigan, J. Vetter, and P. Worley, Performance Evaluation of the SGI Altix 3700, In *2005 Int. Conf. on Parallel Processing* (Oslo, Norway), June 2005.

[5] R. Fatoohi, K. Kardys, S. Koshy, S. Sivaramakrishnan, and J. Vetter, Performance Evaluation of High-Speed Interconnects using Dense Communication Patterns, In *1st workshop on Performance Evaluation of Networks for Parallel, Cluster and Grid Computing Systems*, (Oslo, Norway), June 2005, pp. 554 - 561.

[6] A. Grama, A. Gupta, G. Karypis, and V. Kumar, *Introduction to Parallel Computing, 2nd ed.,* Addison-Wesley, 2003.

[7] Intel MPI Benchmarks: Users Guide & Methodology Description, Intel GmbH, Germany, 2004.

[8] G. Pfister, Aspects of the InfiniBand Architecture, In *2001 IEEE Int. Conf. on Cluster Computing*, (Newport Beach, California), October 2001, pp. 369 – 371.

[9] R. Rabenseifner and A. E. Koniges, The Parallel Communication and I/O Bandwidth Benchmarks: b_eff and b_eff_io. In *43rd Cray User Group Conference,* (Indian Wells, California), May 2001.

[10] M. Woodacre, D. Robb, D. Roe, and K. Feind, The SGI Altix 3000 Global Shared-Memory Architecture – White paper, Silicon Graphics, Inc., 2003.

[11] P. Worley, S. Alam, T. Dunigan, M. Fahey, and J. Vetter, Comparative Analysis of Interprocess Communication on the X1, XD1, and XT3, In *47th Cray User Group Conf.*, (Albuquerque, New Mexico), May 2005.

System	# of proc	b_eff (MB/s)	b_eff per proc (MB/s)	b_eff at L_{max} rings & random (MB/s)	B_eff at L_{max} per proc rings & random (MB/s)	B_eff at L_{max} rings only (MB/s)	B_eff at L_{max} per proc rings only (MB/s)	BW ping-pong (MB/s)	Latency ping-pong (µsec)
SGI Altix 3700	256	47166	184	123579	483	167071	653	1069	1.267
SGI Altix 3700	512	75726	148	202946	396	315591	616	1012	1.249
Cray X1 (SSP)	8	1858	232	5742	718	5838	730	4231	9.044
Cray X1 (SSP)	32	5907	185	20838	651	20288	634	4070	10.330
Cray X1 (SSP)	48	8479	177	30752	641	30137	628	4021	10.365
Cray X1 (MSP)	8	7686	961	35089	4386	45049	5631	9400	10.559
Dell PowerEdge	128	7202	56	21444	168	24713	193	399	2.000
Cray Opteron	8	530	66	1203	150	1745	218	711	0.718
Cray Opteron	64	2922	46	5935	93	12271	192	704	0.709

Table 1. Effective bandwidth benchmark results.

Fig. 1. Unidirectional and Bidirectional Bandwidth

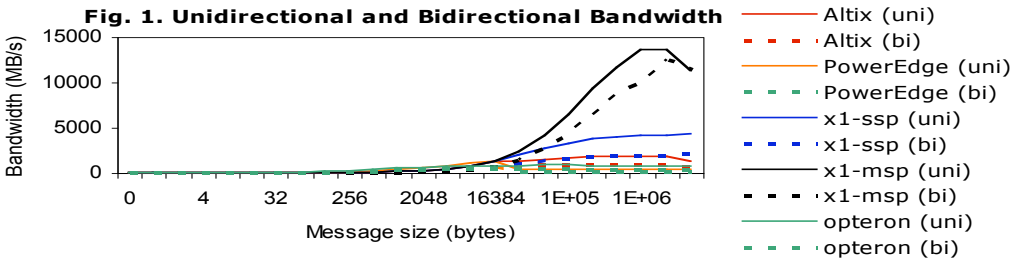


Fig. 2. Unidirectional bandwidth on Altix

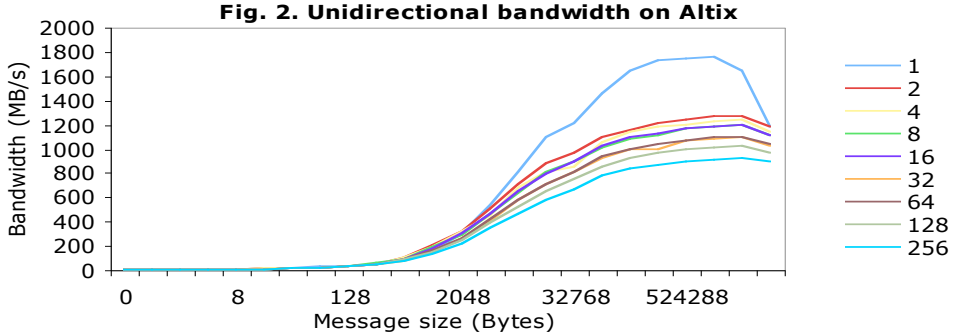


Fig. 3. Unidirectional bandwidth on Cray Opteron

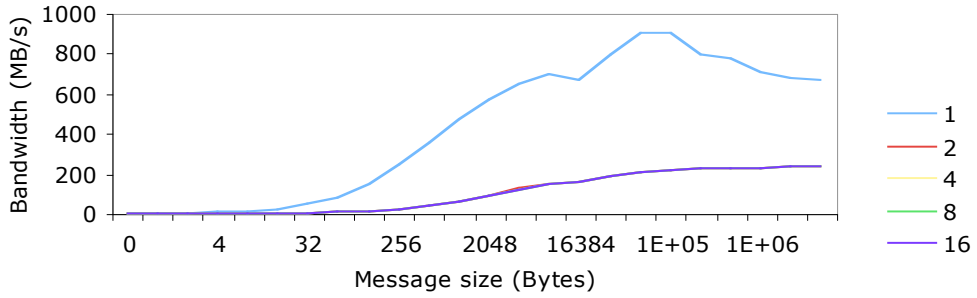


Fig. 4. MPI_Barrier

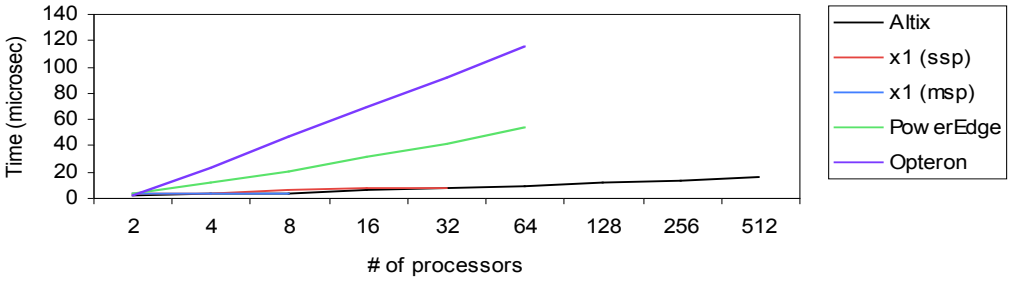


Fig.5. MPI_Reduce

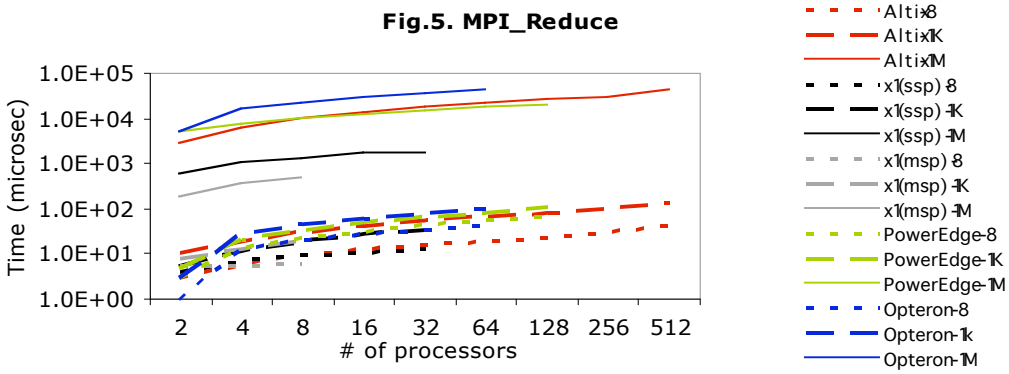


Fig.6. MPI_Alltoall

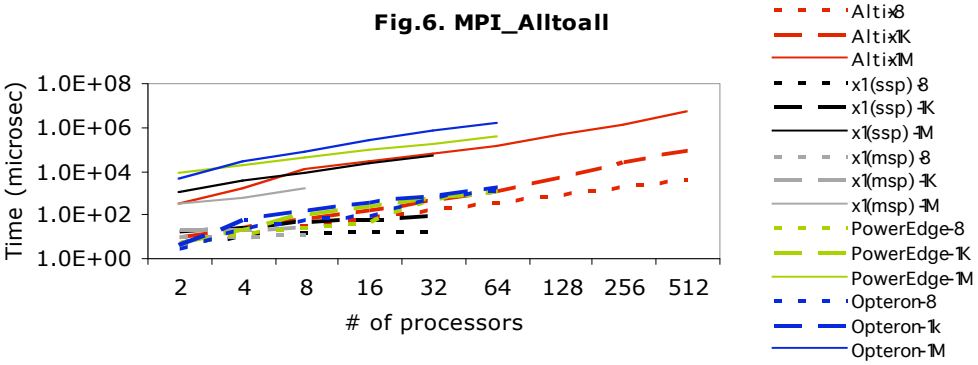


Fig. 7. Congested-controlled AAPC on 512-processor Altix

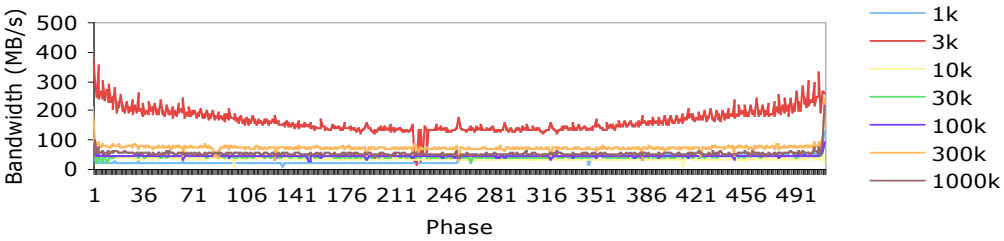


Fig. 8. Cumulative pairwise on 48-processor Cray X1

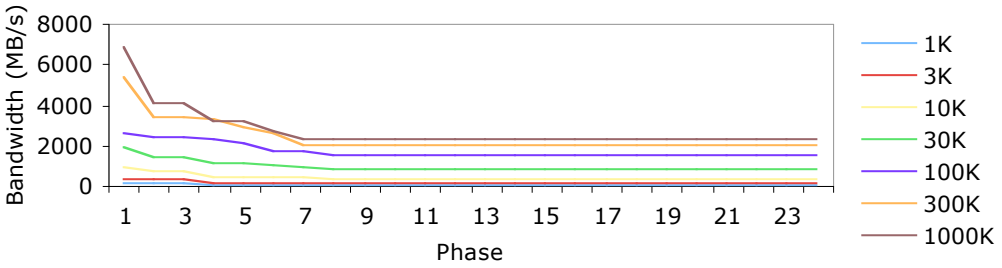


Fig. 9. Simple Pairwise on 512-processr PowerEdge

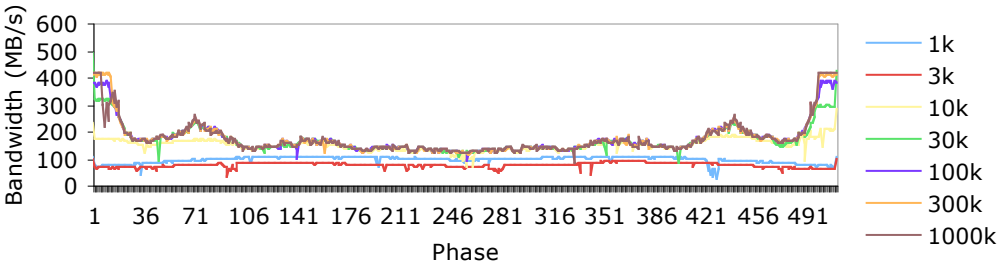


Fig. 10. Simple pairwise on 64-processor Cray Opteron

