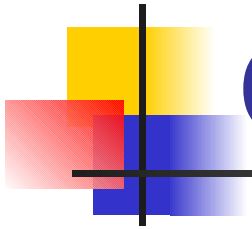
# Interconnect Performance Evaluation of SGI Altix 3700 BX2, Cray X1, Cray Opteron Cluster, and Dell PowerEdge

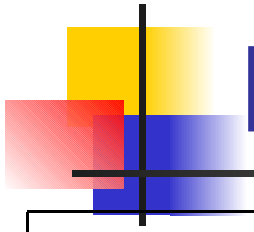
**Panos Adamidis**
**University of Stuttgart, HLRS, Germany**
**S. Saini, R. Fatoohi & R. Ciotti**
NASA Ames Research Center, Moffett Field, California, USA

Cray User Group Meeting
May 8-11, 2006



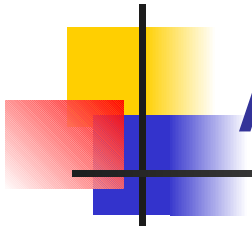
# Objectives

- Identify limiting factors & bottleneck w/ high-speed interconnects
- Compare performance of interconnects



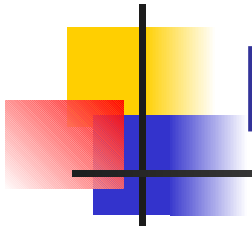
# Platforms

| Platform | # of procs | Procs/ node | Clock (GHz) | Peak (Gflop/s) | Network | Link BW (GB/s) |
|---|---|---|---|---|---|---|
| SGI Altix 3700 BX2 | 512 | 2 | 1.6 | 3280 | NUMAlink4 | 6.4 |
| Cray X1 | 64 | 4 | 0.8 | 205 | Custom | 51.2 |
| Cray Opteron Cluster | 128 | 2 | 2.0 | 512 | Myrinet | 1.067 |
| Dell PowerEdge | 2560 | 2 | 3.6 | 9200 | InfiniBand | 1 |



# Approach

- Using 3 benchmarks:
  - Effective Bandwidth Benchmark (b_eff)
  - Intel MPI Benchmarks (IMB)
  - Dense Communication Benchmarks
- Measuring: unidirectional BW, bidirectional BW, latency, collective communication & dense communication
- Employing: different # of processors, different topologies & different message sizes



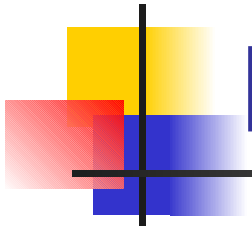
# Results: b_eff

- Measure accumulated BW of network
- B_eff:

a) Log avg over 6 ring patterns & random patterns
b) Avg of 21 message sizes (1 – 1M bytes)
c) Max over 3 communication methods: MPI_Sendrecv, MPI_Alltoallv & non-blocking w/ MPI_Irecv, MPI_Isend & MPI_Waitall

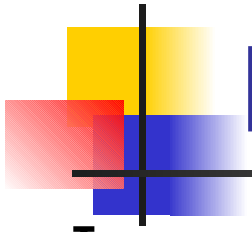
# Results: b_eff

| 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 *(MB/s)* | b_eff at $L_{max}$ per proc rings *(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 |



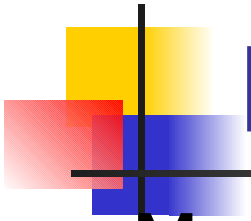
# Results: b_eff

- Latency: lowest w/ Opteron (0.7 µsec), highest w/ X1 (10 µsec)
- Link BW (ping-pong): highest w/ X1 (9.4 GB/s in MSP), lowest w/ PowerEdge (0.4 GB/s)
- b_eff: highest w/ 512-proc Altix (75.7 GB/s), lowest w/ 64-proc Opteron (2.9 GB/s)



# Results: b_eff

- Impact of communication in parallel (comparing ping-pong w/ b_eff at $L_{max}$ per proc using rings only): significant on X1 (SSP), less significant on Altix
- Impact of random neighbor locations (comparing b_eff at $L_{max}$ per proc using rings w/ the one using rings & random patterns): 50% drop on 64-proc Opteron, no drop on X1 (SSP)
- Impact of message size (comparing b_eff at $L_{max}$ per proc using rings & random patterns w/ b_eff per proc): significant drops on all systems

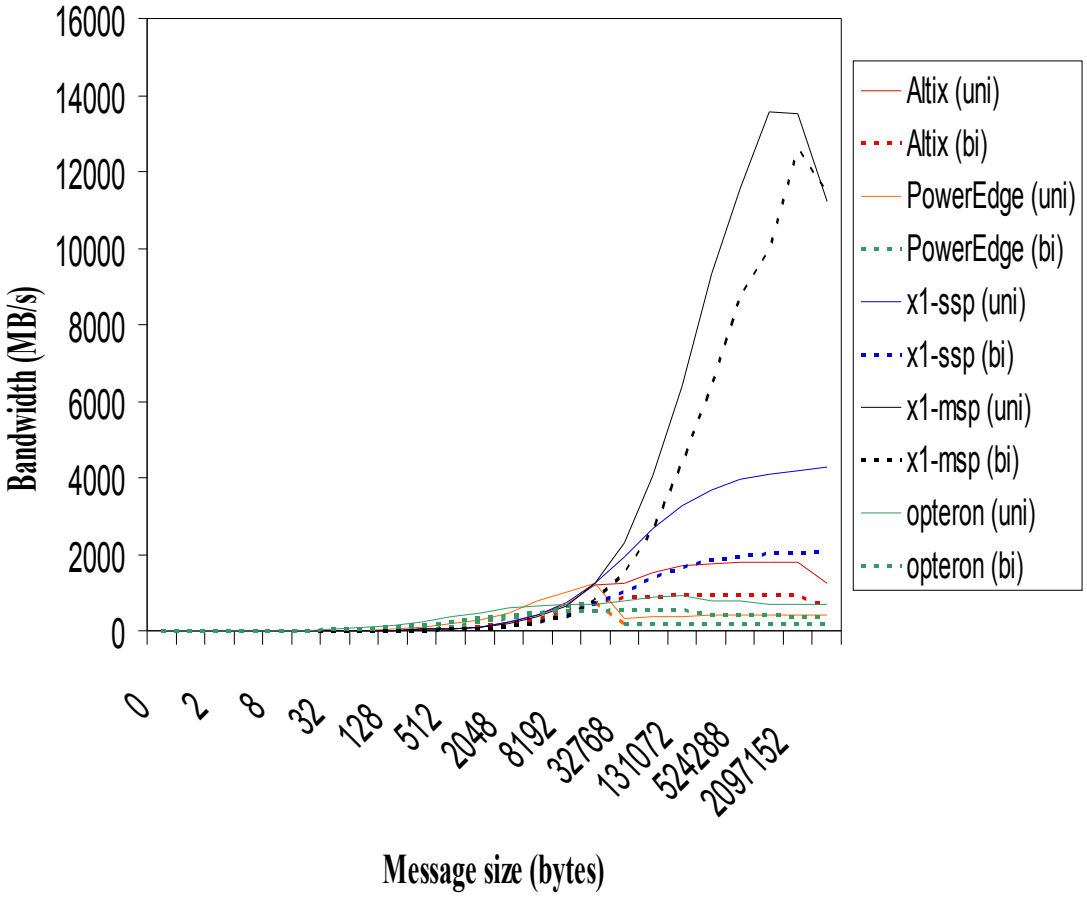


# Results: IMB

- Measure point-point communication:
  - Unidirectional (PingPong)
  - Bidirectional (PingPing): message obstructed by oncoming message
  - Unidirectional w/ varying distance between communicating processors
- Measure collective communication: Barrier, Reduce & AlltoAll

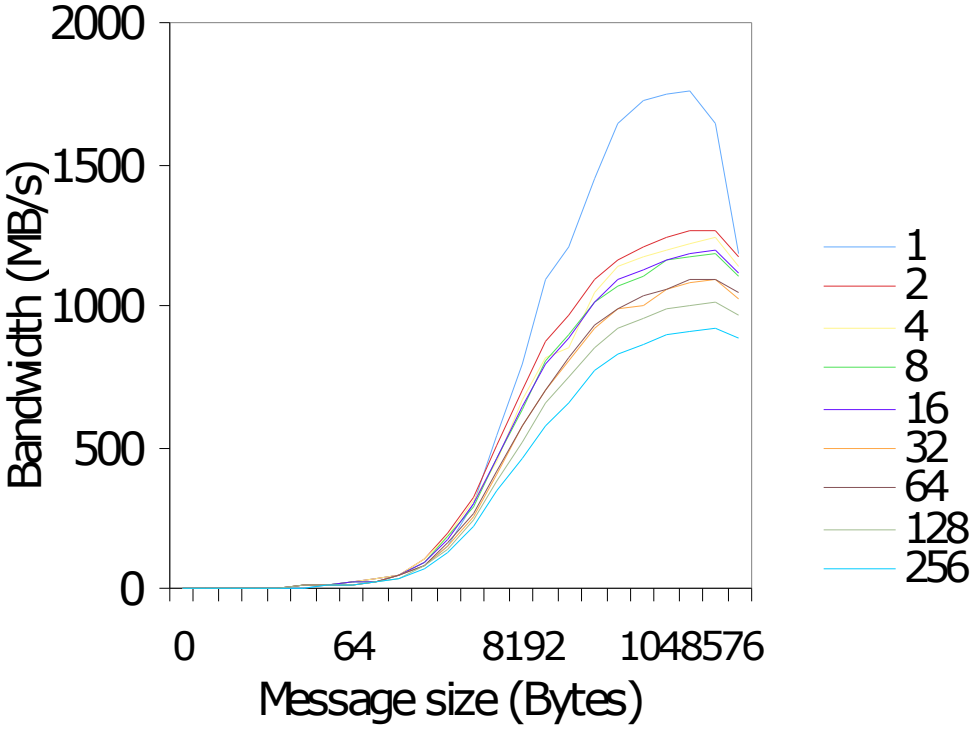
# Results: IMB - Unidirectional vs bidirectional

- X1 (MSP) has highest rate (13 GB/s)
- Drop of 50% for bidirectional on most systems except on X1 (MSP)
- Factor of over 3 between X1 SSP & MSP modes

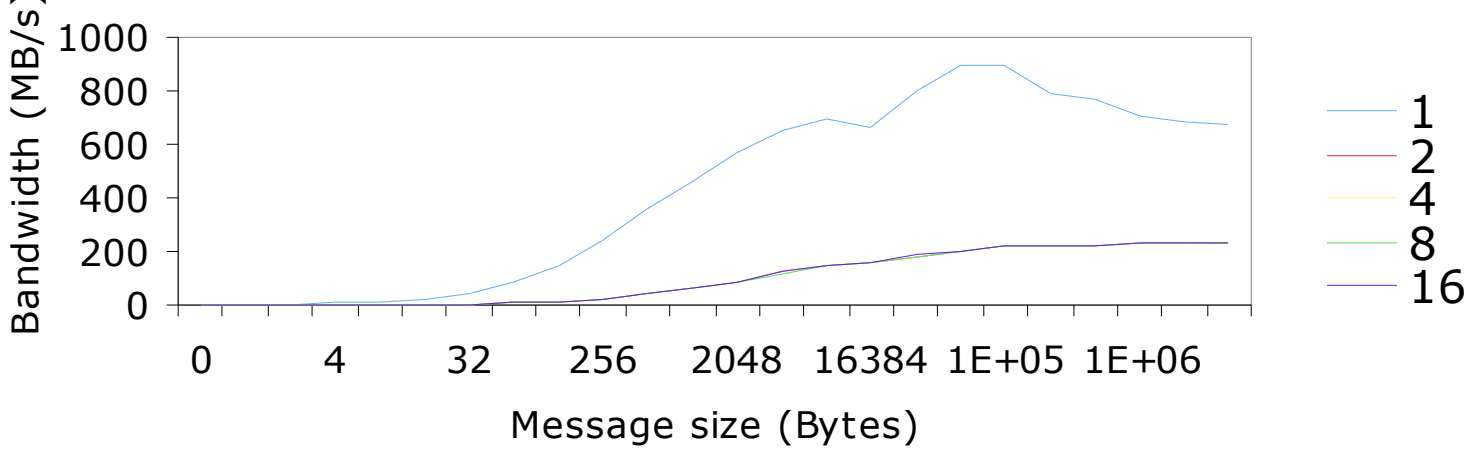


# Results: IMB - Unidirectional BW on Altix (varying distance)

- Rates (GB/s):
  - 1 hop (on node): 1.76
  - 2 or 4 hops (on C-brick): 1.26
  - 8 - 128 hops: 1.19 – 1.02
  - 256 hops: 0.92
- Over 4 hops: # of R-bricks

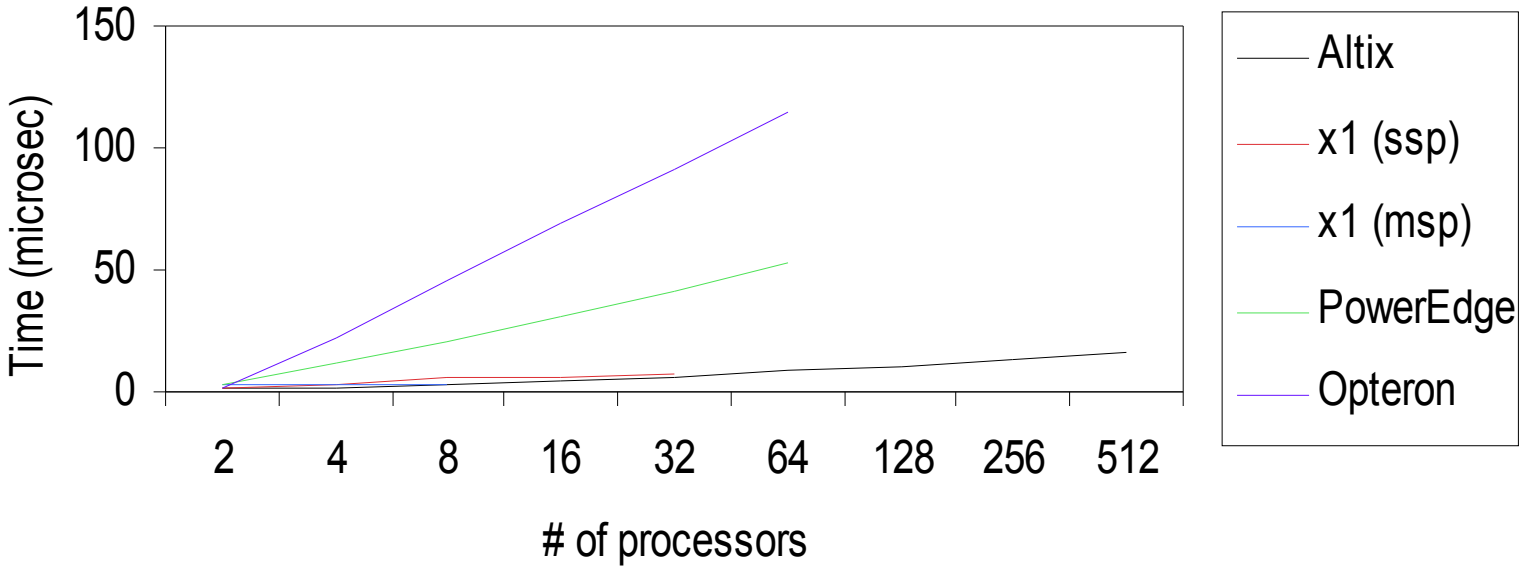


# Results: IMB - Unidirectional BW on Cray Opteron (varying distance)



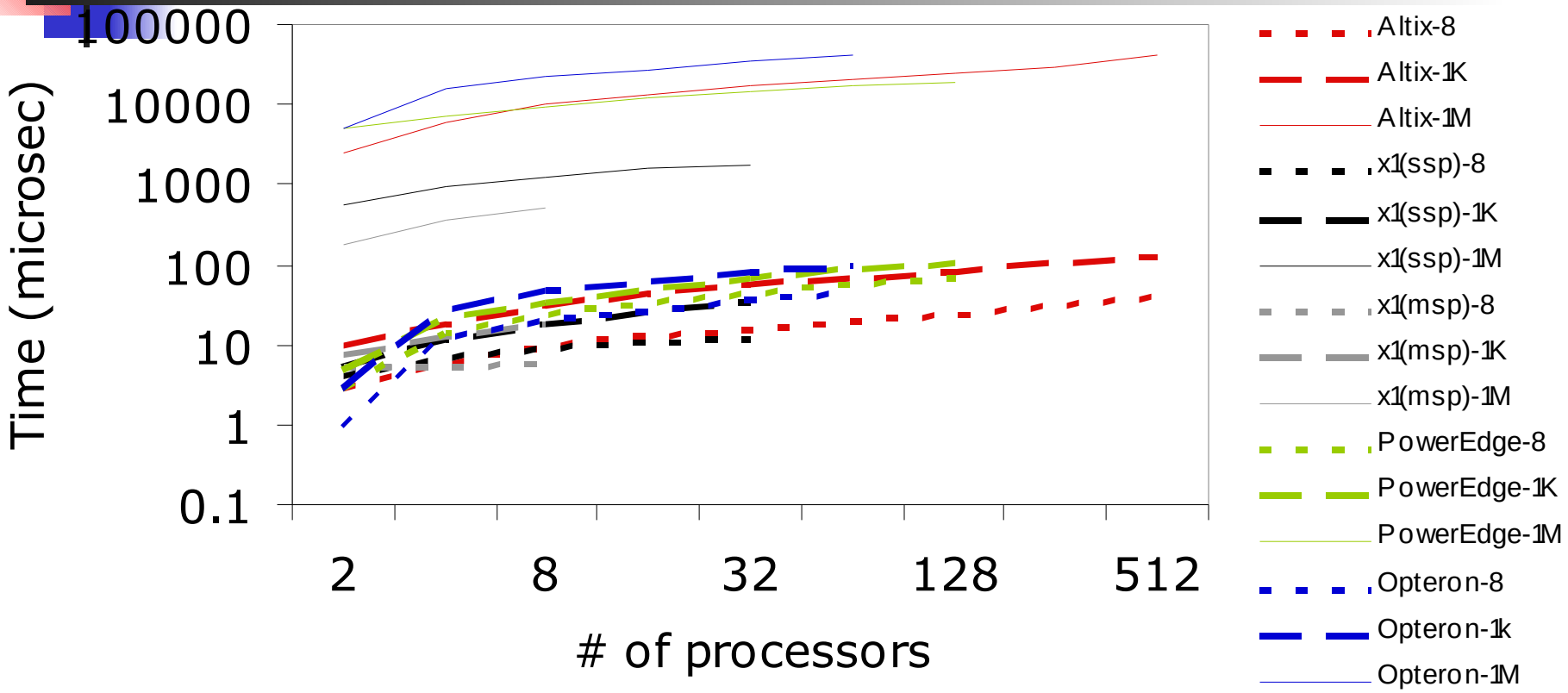

- Rates (MB/s):
  - 1 hop (on node): 900
  - 2 - 16 hops (between nodes): 234
- Distance insensitivity between nodes for Myrinet

# Results: IMB – MPI_Barrier



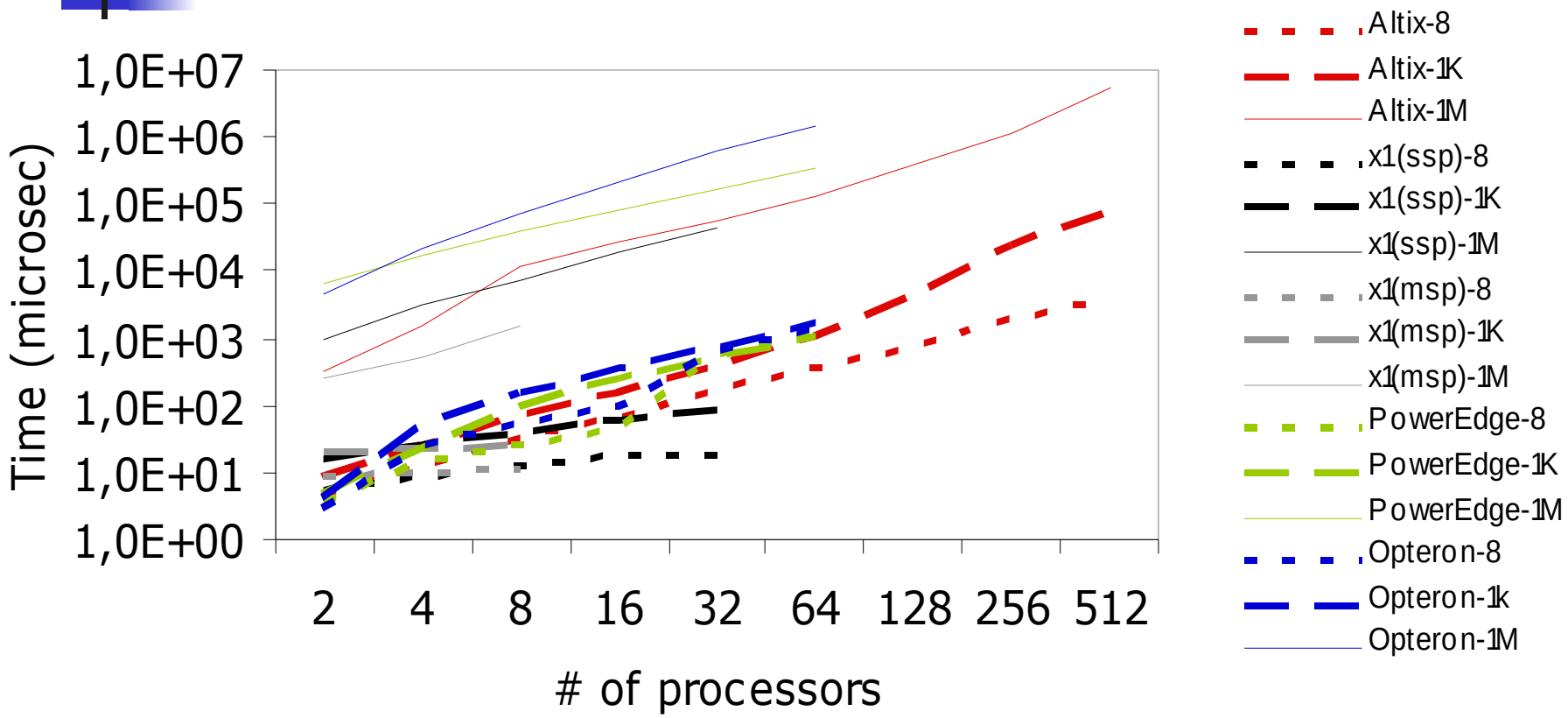

- Shared-memory systems (Altix, X1) outperformed distributed-memory systems (PowerEdge, Opteron)

# Results: IMB – MPI_Reduce



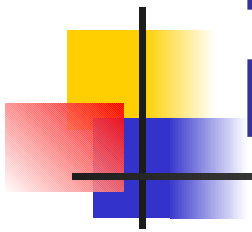

- X1 in both modes outperformed other systems

# Results: IMB – MPI_Alltoall



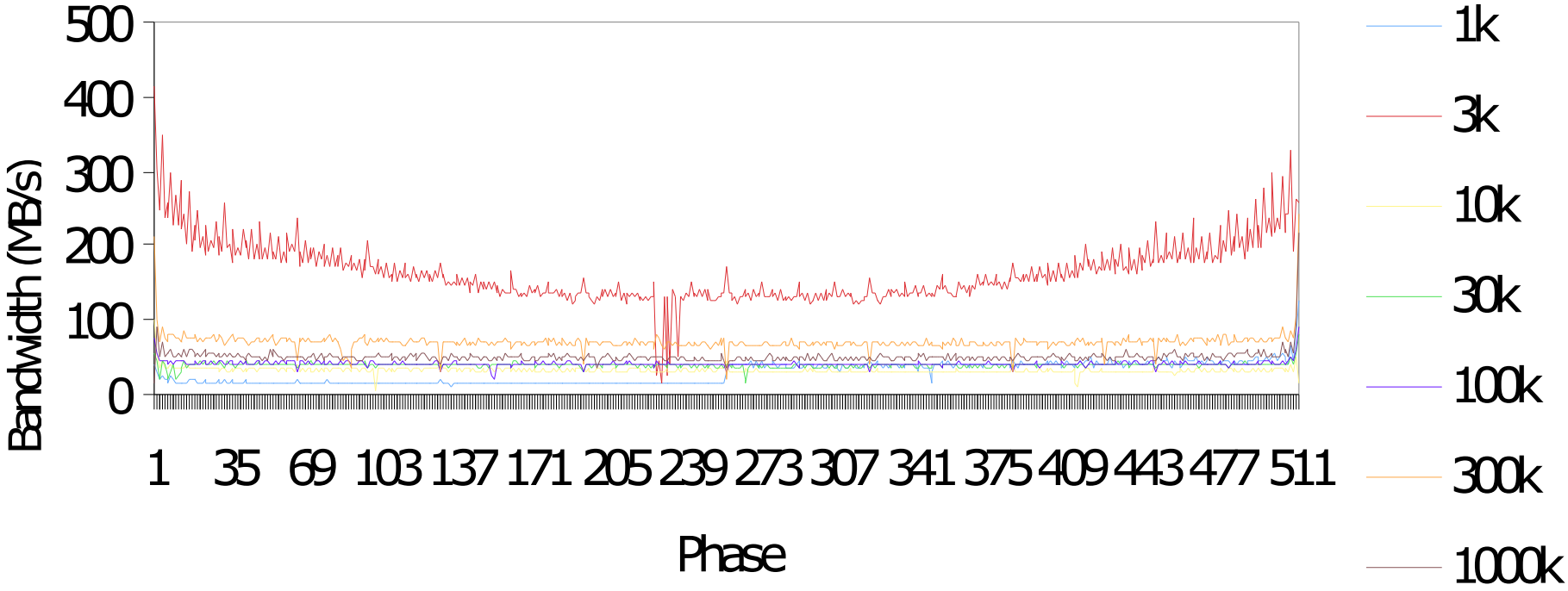

- X1 especially in MSP mode outperformed other systems

# Results: Dense Communication Benchmarks



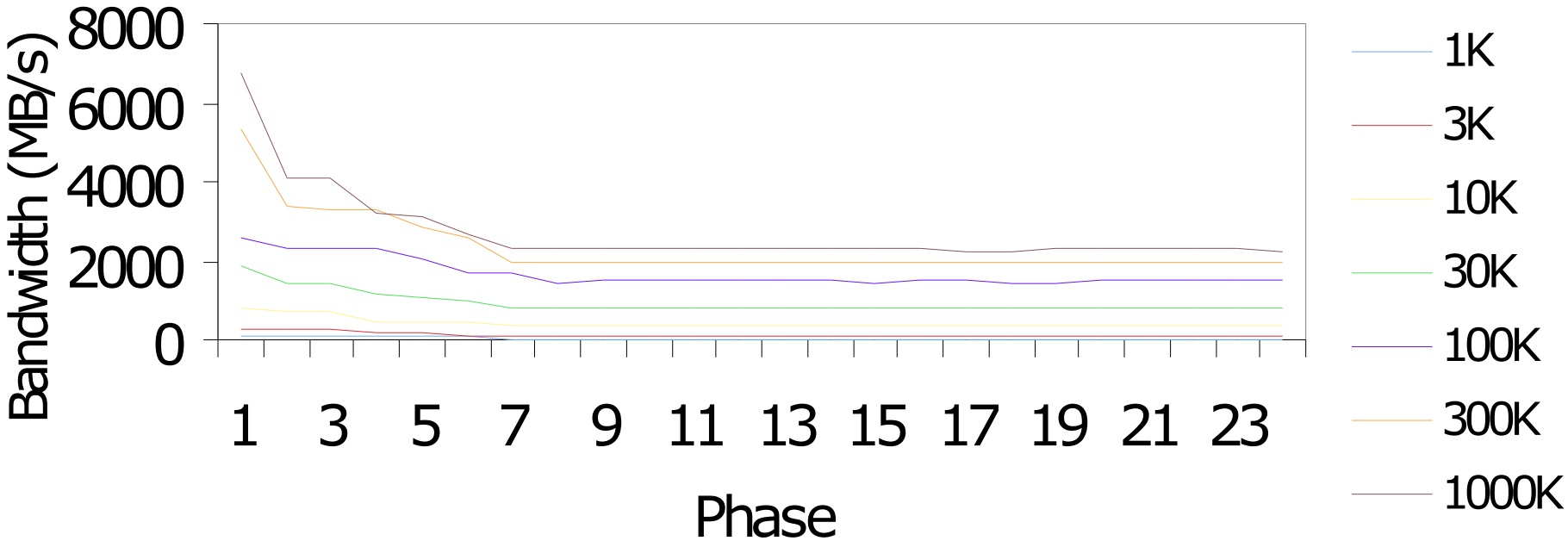
- Congestion-controlled AAPC (All-to-All Personalized Communication)
- Simple pair-wise communication
- Cumulative pair-wise communication
- Random pair-wise communication

# Results: Congestion-controlled AAPC on 512-proc Altix



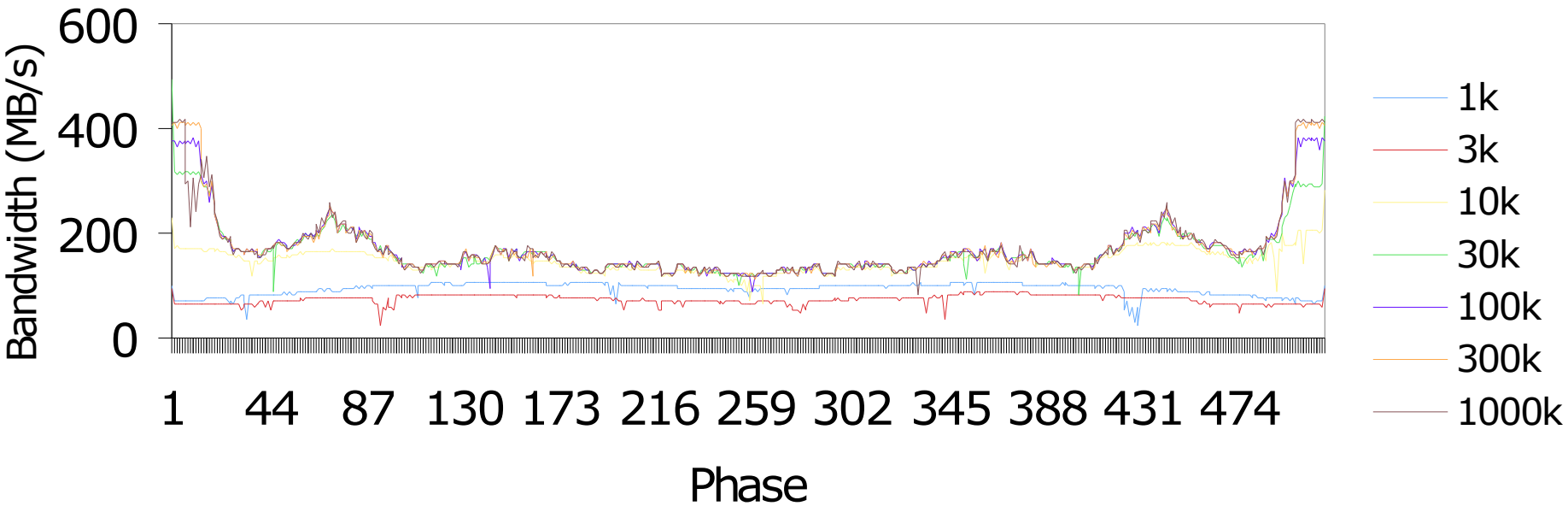

- Drop in middle phases (by factor of 5) compared to 1st & last phases

# Results: Cumulative pairwise on 48-proc Cray X1 in SSP mode



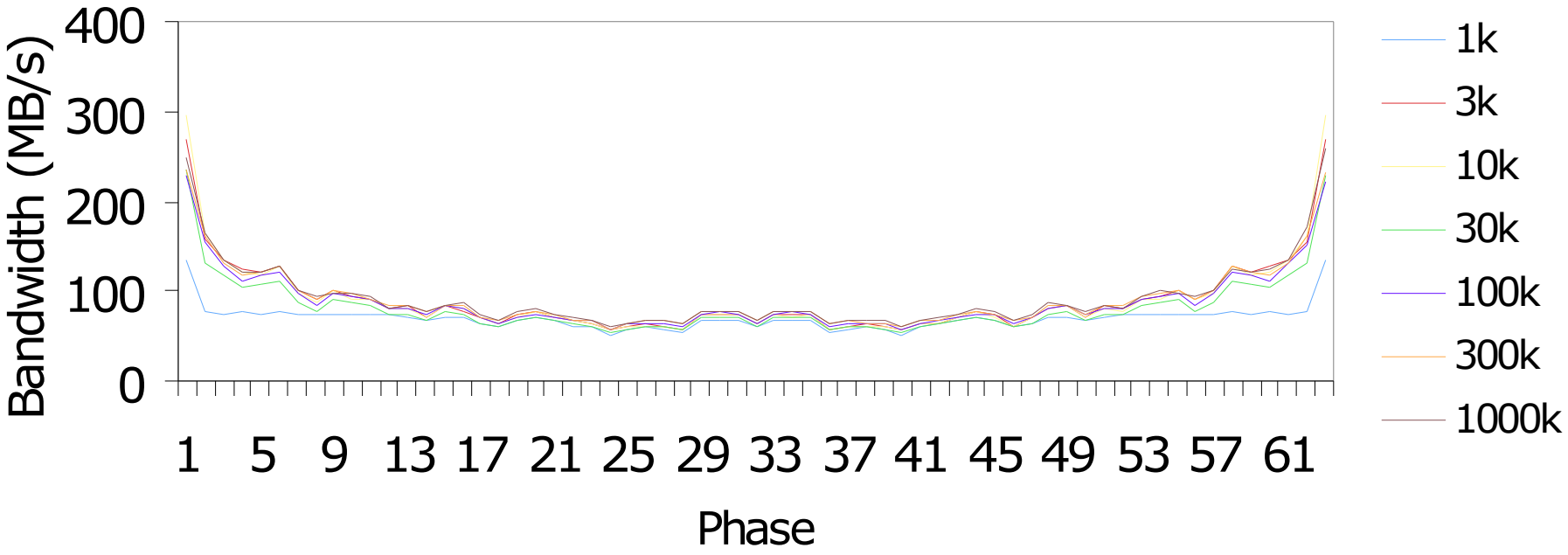

- Drop of factor of 3 as # of communicating procs increased to 8 or more pairs

# Results: Simple pairwise on 512-proc Dell PowerEdge

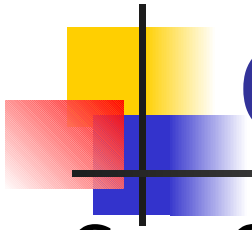



- Drop in middle phases (by factor of 3) compared to 1st & last phases

# Results: Simple pairwise on 64-proc Cray Opteron




- Drop in middle phases (by factor of 4) compared to 1st & last phases



# Conclusions

- Cray Opteron has lowest latency while Cray X1 has highest link BW
- Communication in parallel has significant impact on X1 (SSP)
- Drop of 50% in link BW due to oncoming message on all systems except on X1 (MSP)
- Significant drop in link BW as communicating processors are separated apart (from 1 to 16) on Cray Opteron
- Shared-memory systems outperformed distributed-memory systems using collective communication
- Significant drop in performance as communicating processors are far apart w/ dense communication patterns