# Performance Evaluation of Supercomputers using HPCC and IMB Benchmarks

**Subhash Saini**

Robert Ciotti, Brian T. N. Gunney, Alice Koniges, Don Dossa
Panagiotis Adamidis, Rolf Rabenseifner, Sunil R. Tiyyagura,
Matthias Mueller, and Rod Fatoohi

NASA Advanced Supercomputing (NAS) Division
*NASA Ames Research Center, Moffett Field, California*

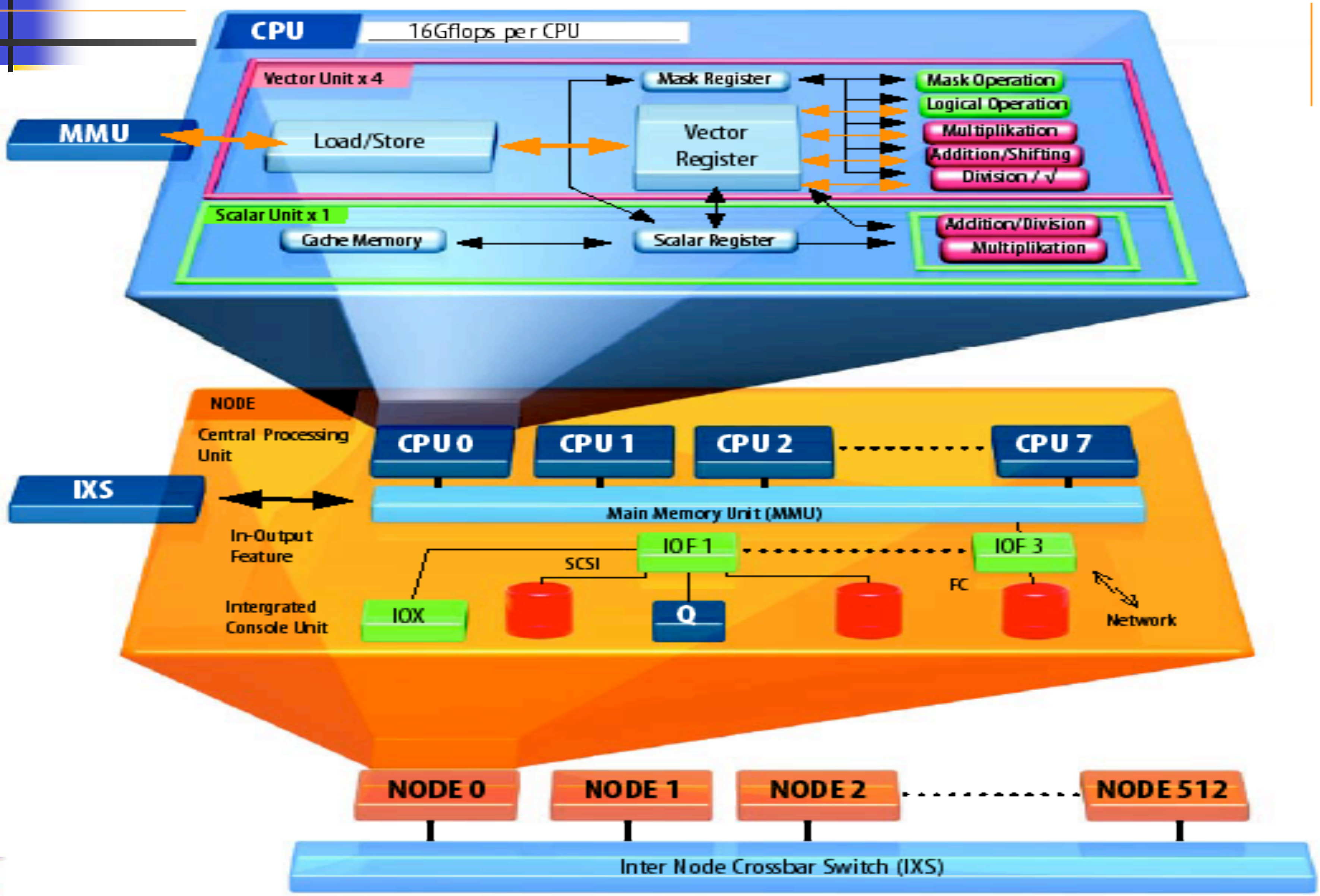
**IPDPS 2006 - PMEO, Rhodes, Greece, April 29**

# Outline

- **Computing platforms**
  - Columbia System (NASA, USA)
  - NEC SX-8 (HLRS, Germany)
  - Cray X1 (NASA, USA)
  - Cray Opteron Cluster (NASA, USA)
  - Dell POWER EDGE (NCSA, USA)
- **Benchmarks**
  - HPCC Benchmark suite
  - IMB Benchmarks
- **Summary**

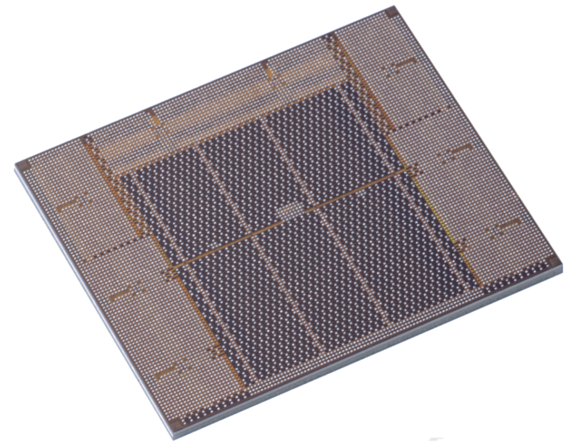
# NEC SX-8 System

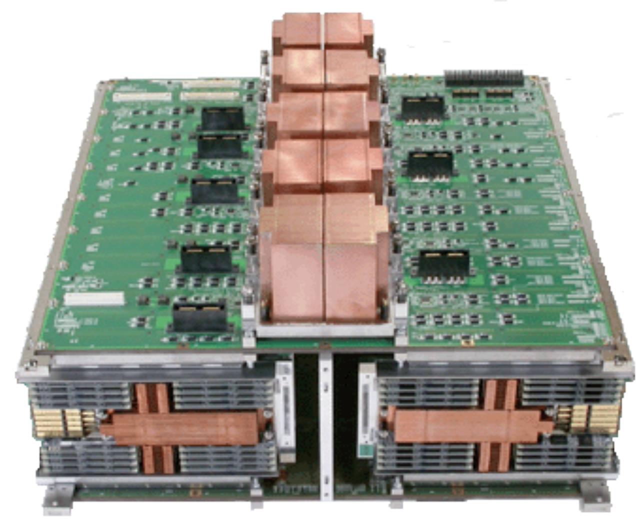
# SX-8 System Architecture



# SX-8 Technology

- Hardware dedicated to scientific and engineering applications.
- CPU: 2 GHz frequency, 90 nm-Cu technology
- 8000 I/O per CPU chip
- Hardware vector square root
- Serial signalling technology to memory, about 2000 transmitters work in parallel
- 64 GB/s memory bandwidth per CPU
- Multilayer, low-loss PCB board, replaces 20000 cables
- Optical cabling used for internode connections
- Very compact packaging.





# SX-8 specifications

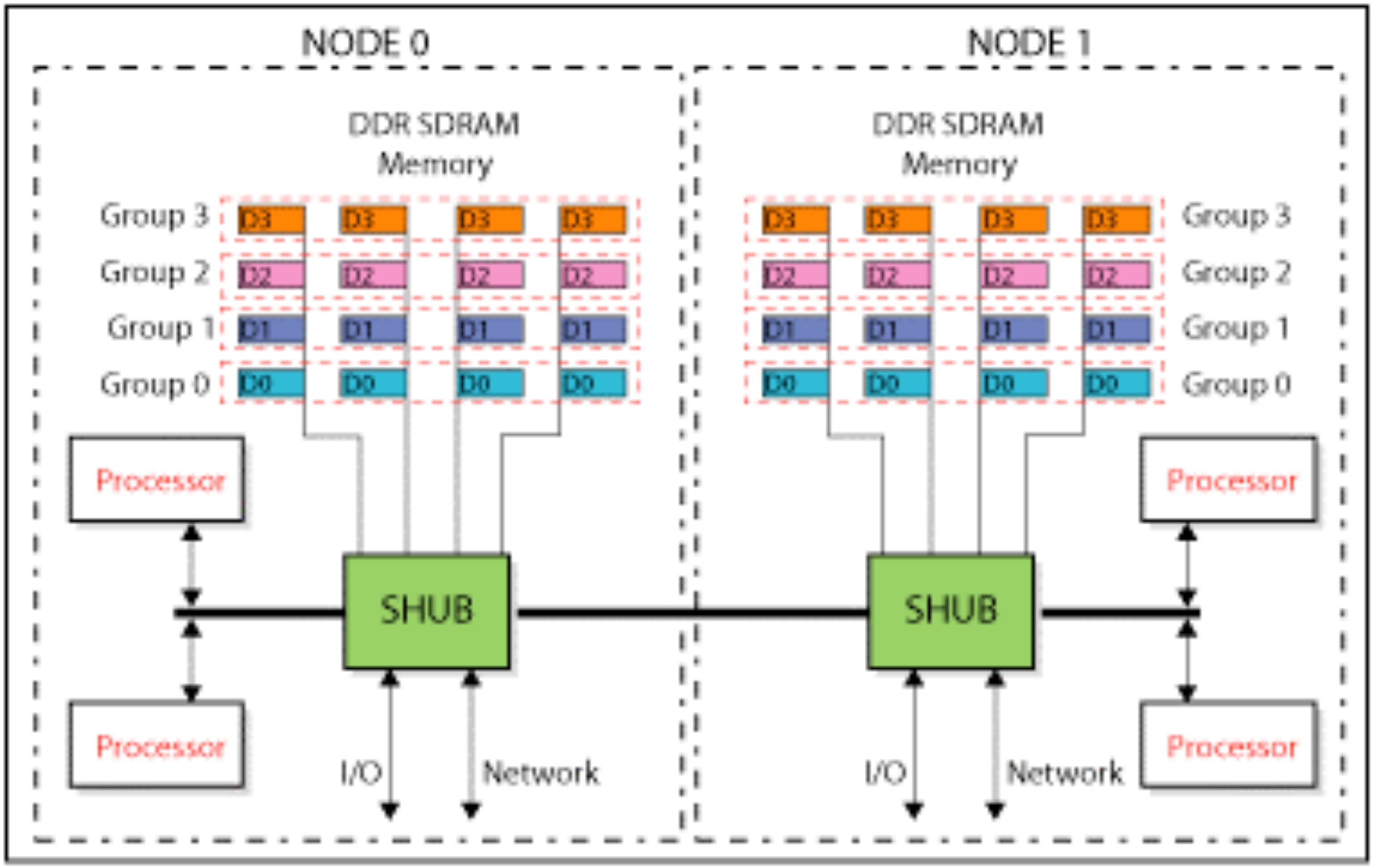
- 16 GF / CPU (vector)
- 64 GB/s memory bandwidth per CPU
- 8 CPUs / node
- 512 GB/s memory bandwidth per node
- Maximum 512 nodes
- Maximum 4096 CPUs, max 65 TFLOPS
- Internode crossbar Switch
- 16 GB/s (bi-directional) interconnect bandwidth per node
- Maximum size SX-8 is among the most powerful computers in the world

# Columbia 2048 System

- Four SGI Altix BX2 boxes with 512 processors each connected with NUMALINK4 using fat-tree topology
- Intel Itanium 2 processor with 1.6 GHz and 9 MB of L3 cache
- SGI Altix BX2 compute brick has eight Itanium 2 processors with 16 GB of local memory and four ASICs called SHUB
- In addition to NUMALINK4, InfiniBand (IB) and 10 Gbit Ethernet networks also available
- Processor peak performance is 6.4 Gflop/s; system peak of the 2048 system is 13 Tflop/s
- Measured latency and bandwidth of IB are 10.5 microseconds and 855 MB/s.

# Columbia System

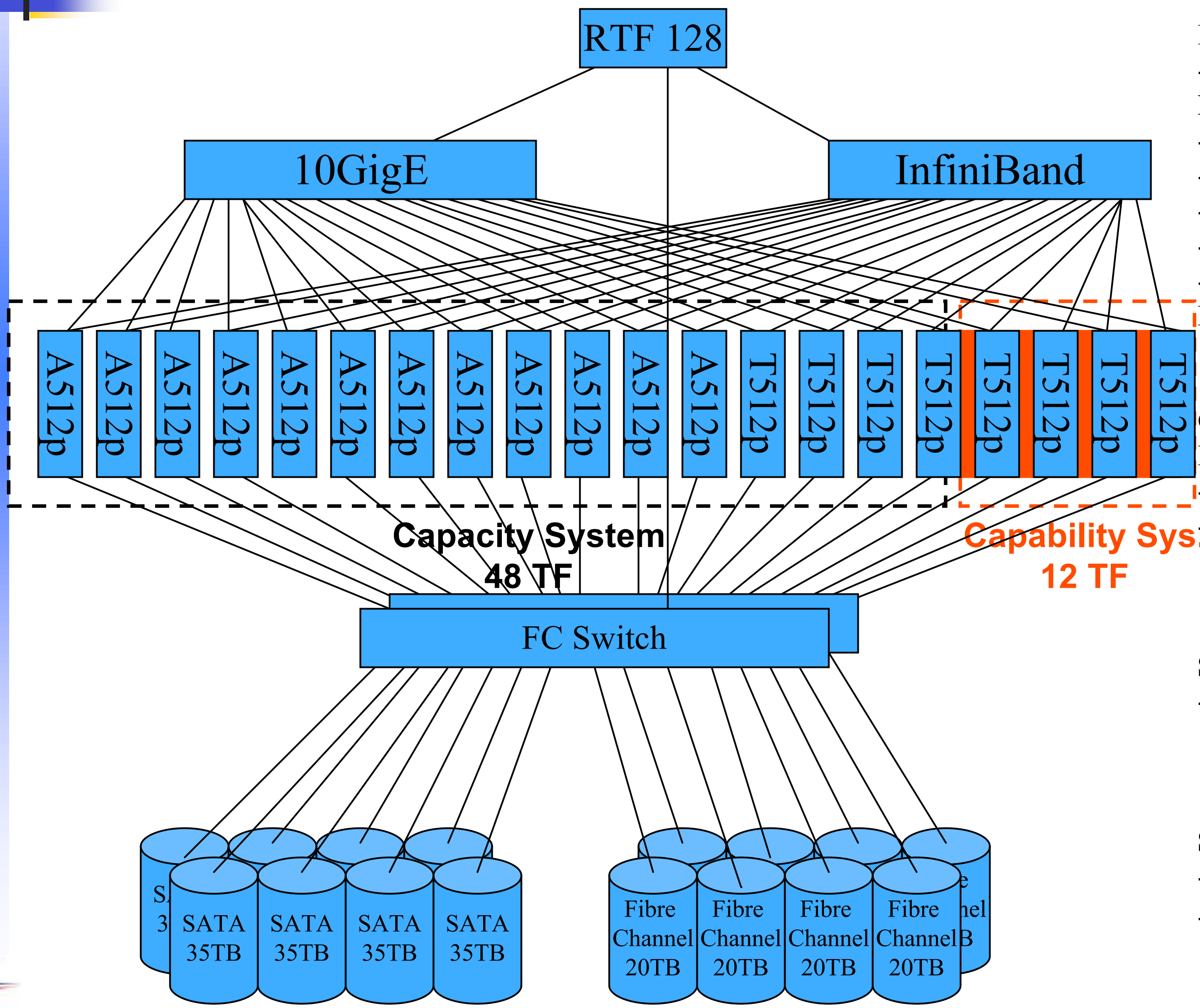
# SGI Altix 3700




- Itanium 2@ 1.5GHz (peak 6 GF/s)
- 128 FP reg, 32K L1, 256K L2, 6MB L3

- CC-NUMA in hardware
- 64-bit Linux w/ single system image -- looks like a single Linux machine but with many processors

# Columbia Configuration




**Front End**
- 128p Altix 3700 (RTF)

**Networking**
- 10GigE Switch 32-port
-10GigE Cards (1 Per 512p)
- InfiniBand Switch (288port)
- InfiniBand Cards (6 per 512p)
- Altix 3700 2BX 2048 Numalink Kits

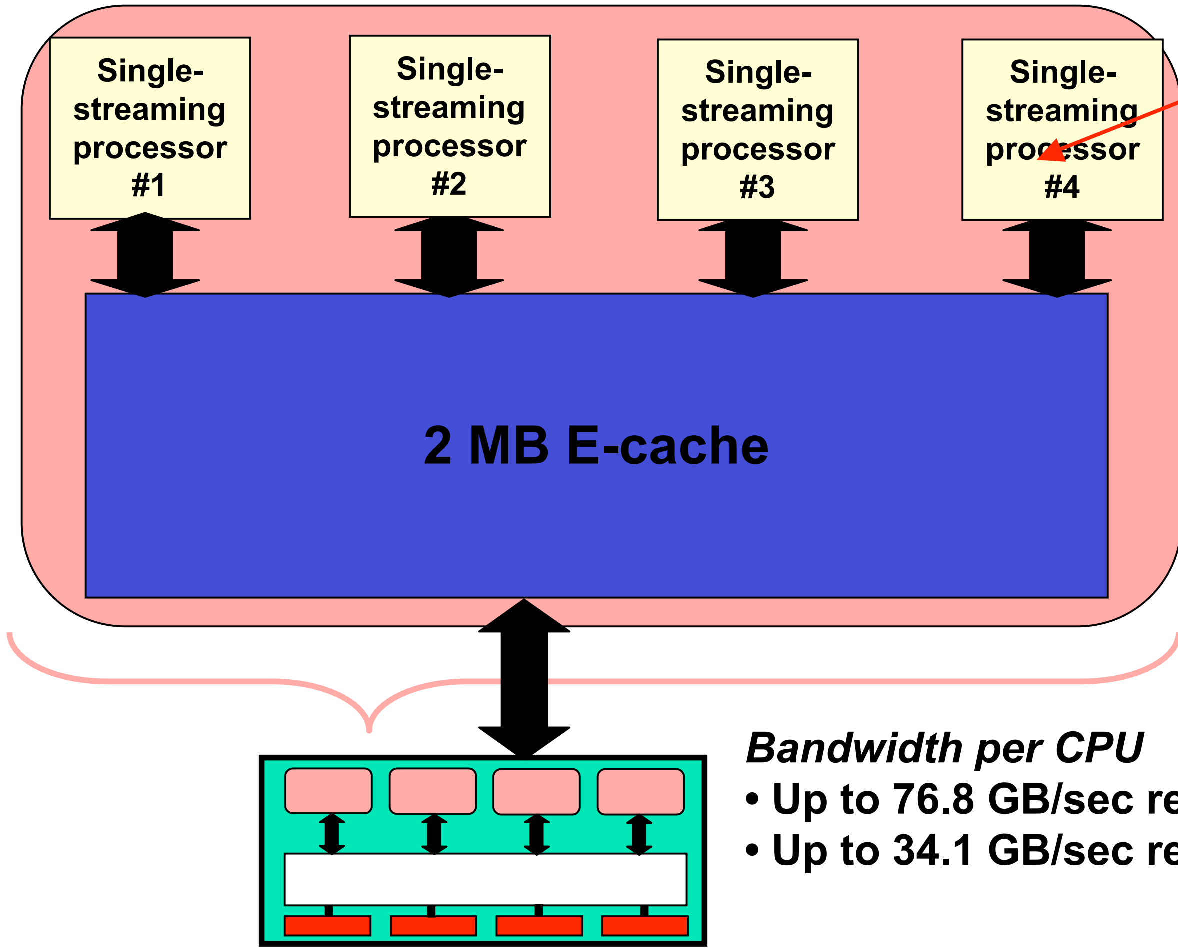
**Compute Node (Single Sys Image)**
- Altix 3700 (A) 12x512p
- Altix 3700 BX2 (T) 8x512p

**Storage Area Network**
-Brocade Switch 2x128port

**Storage (440 TB)**
-FC RAID 8x20 TB (8 Racks)
-SATARAID 8x35TB (8 Racks)

# Cray X1 CPU: Multistreaming Processor

- New Cray Vector Instruction Set Architecture (ISA)
- 64- and 32-bit operations, IEEE floating-point



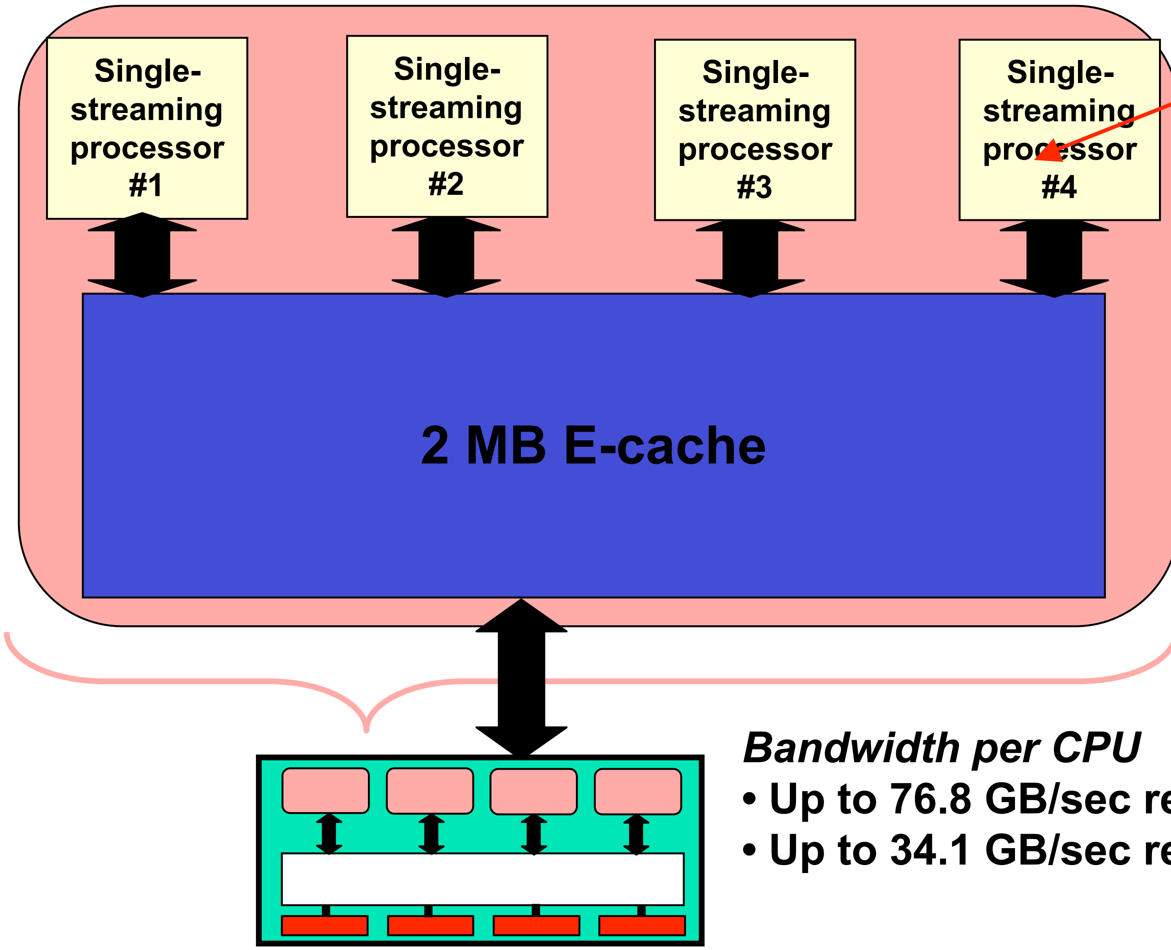

*Each Stream:*
- 2 vector pipes (32 vector regs. of 64 element ea)
- 64 A & S regs.
- Instruction & data cache

*MSP:*
- 4 x P-chips
- 4 x E-chips (cache)

*Bandwidth per CPU*
- Up to 76.8 GB/sec read/write to cache
- Up to 34.1 GB/sec read/write to memory

# Cray X1 CPU: Multistreaming Processor

- New Cray Vector Instruction Set Architecture (ISA)
- 64- and 32-bit operations, IEEE floating-point

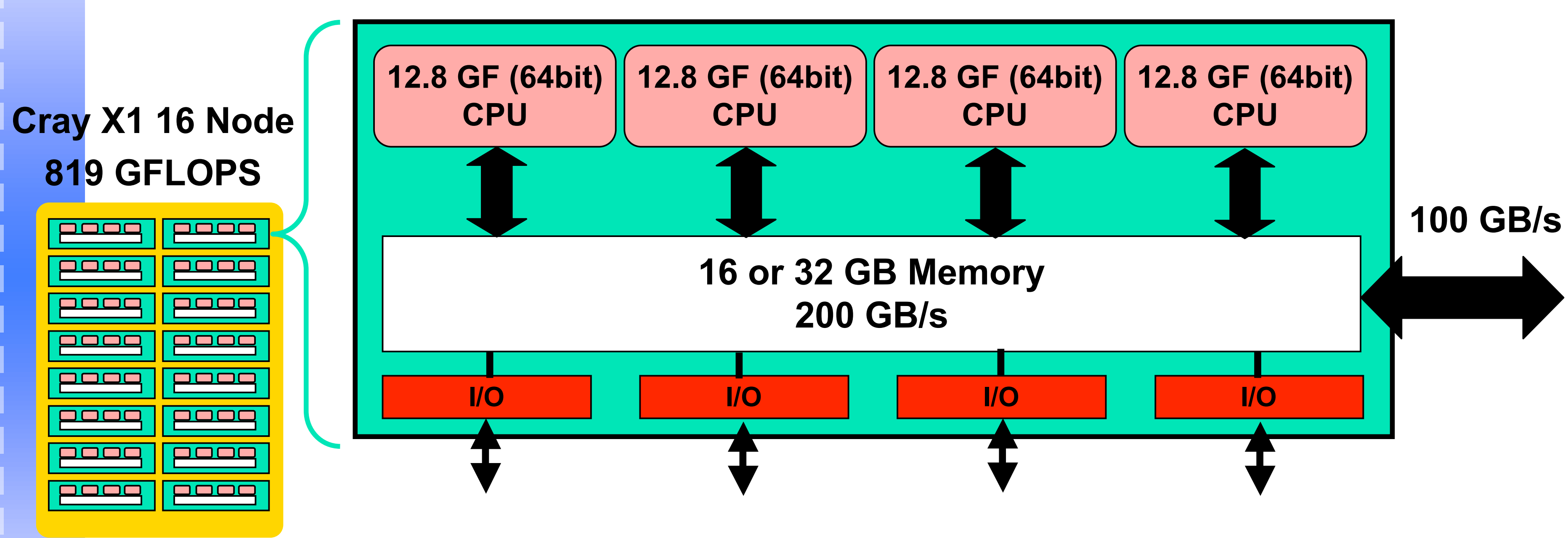



*Each Stream:*
- 2 vector pipes (32 vector regs. of 64 element ea)
- 64 A & S regs.
- Instruction & data cache

*MSP:*
- 4 x P-chips
- 4 x E-chips (cache)

*Bandwidth per CPU*
- Up to 76.8 GB/sec read/write to cache
- Up to 34.1 GB/sec read/write to memory

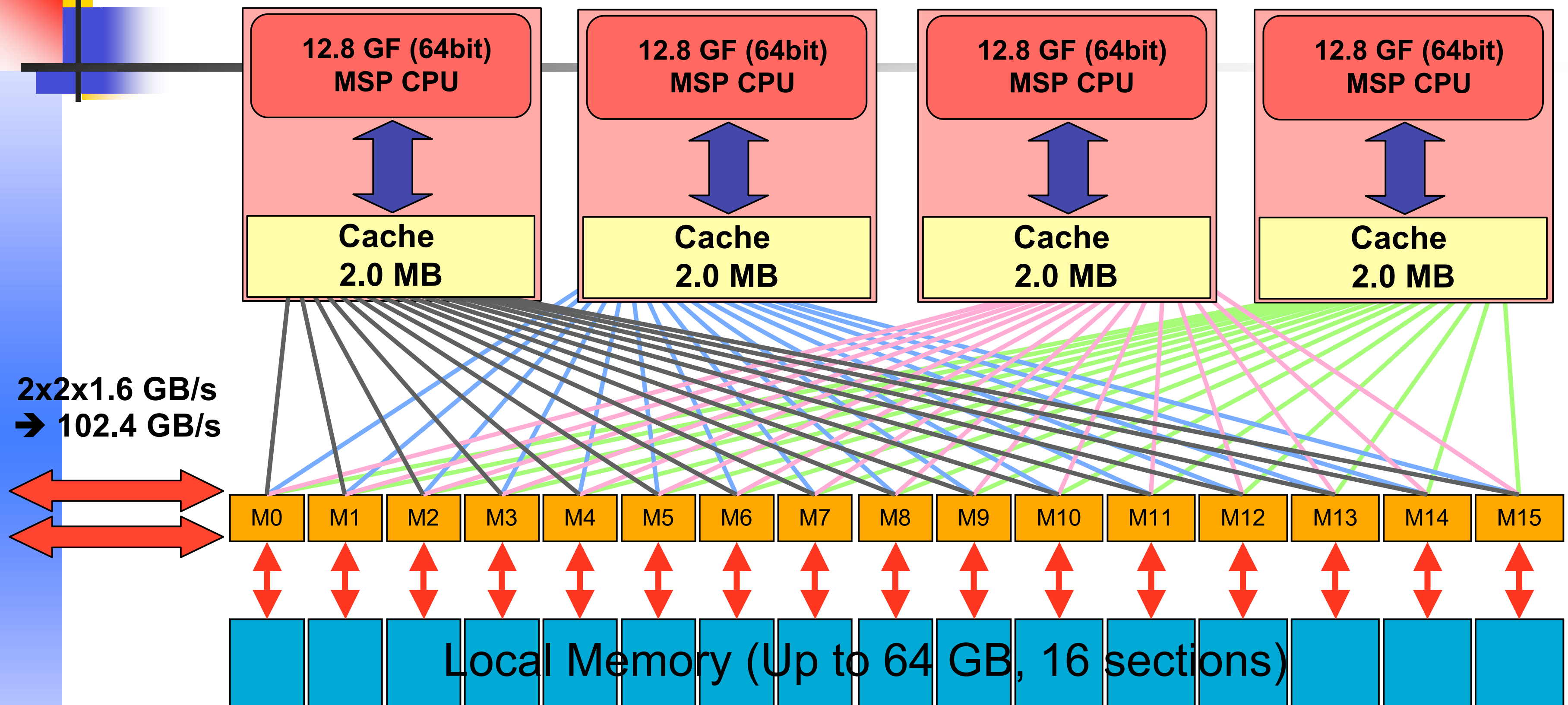
# Cray X1 Processor Node Module




X1 node board has performance roughly comparable to:

- 128 PE Cray T3E system
- 16-32 CPU Cray T90 system

# Cray X1 Node - 51.2 Gflop/s




**Interconnect network**

2 ports/M-chip
1.6 GB/s/port peak in each direction

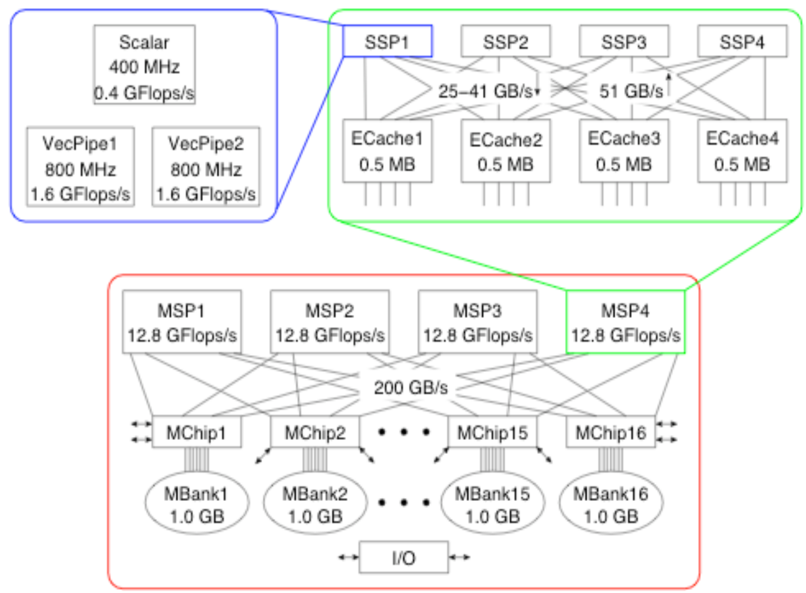
= 102.4 GB/s to the network

**Local memory**

Peak BW = 16 sections x 12.8 GB/s/section = 204.8 GB/s
Capacity = 16, 32 or 64 GB

# Cray X1 at NAS

- **Architecture**
  - 4 nodes, 16 MSPs (64 SSPs)
  - 1 node reserved for system;
    3 nodes usable for user codes
  - 1 MSP: 4 SSPs at 800 MHz, 2 MB ECache
    12.8 Gflops/s peak
  - 64 GB main memory;
    4 TB FC RAID




- **Operating Environment**
  - Unicos MP 2.4.3.4
  - Cray Fortran and C 5.2
  - PBSPro job scheduler

# Cray X1 at NAS

# Intel Xeon Cluster ("Tungsten") at NCSA

# High End Computing Platforms

**Table 2:** System characteristics of the computing platforms .

| Platform | Type | CPUs / node | Clock (GHz) | Peak/node (Gflop/s) | Network | Network Topology | Operating System | Location | Processor Vendor | System Vendor |
|---|---|---|---|---|---|---|---|---|---|---|
| SGI Altix BX2 | Scalar | 2 | 1.6 | 12.8 | NUMALINK 4 | Fat-tree | Linux (Suse) | NASA (USA) | Intel | SGI |
| Cray X1 | Vector | 4 | 0.800 | 12.8 | Proprietary | 4D-Hypercube | UNICOS | NASA (USA) | Cray | Cray |
| Cray Opteron Cluster | Scalar | 2 | 2.0 | 8.0 | Myrinet | Fat-tree | Linux (Redhat) | NASA (USA) | AMD | Cray |
| Dell Xeon Cluster | Scalar | 2 | 3.6 | 14.4 | InfiniBand | Fat-tree | Linux (Redhat) | NCSA (USA) | Intel | Dell |
| NEC SX-8 | Vector | 8 | 2.0 | 16.0 | IXS | Multi-stage Crossbar | Super-UX | HLRS (Germany) | NEC | NEC |

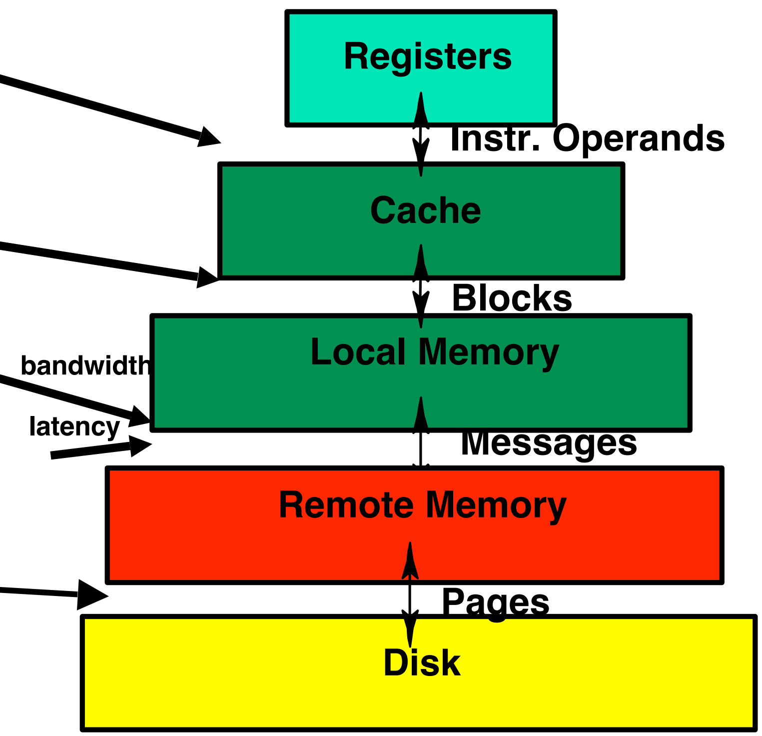
# HPC Challenge Benchmarks

- Basically consists of 7 benchmarks
  - **HPL:** floating-point execution rate for solving a linear system of equations
  - **DGEMM:** floating-point execution rate of double precision real matrix-matrix multiplication
  - **STREAM:** sustainable memory bandwidth
  - **PTRANS:** transfer rate for large data arrays from memory (total network communications capacity)
  - **RandomAccess:** rate of random memory integer updates (GUPS)
  - **FFTE:** floating-point execution rate of double-precision complex 1D discrete FFT
  - **Latency/Bandwidth:** ping-pong, random & natural ring

# HPC Challenge Benchmarks

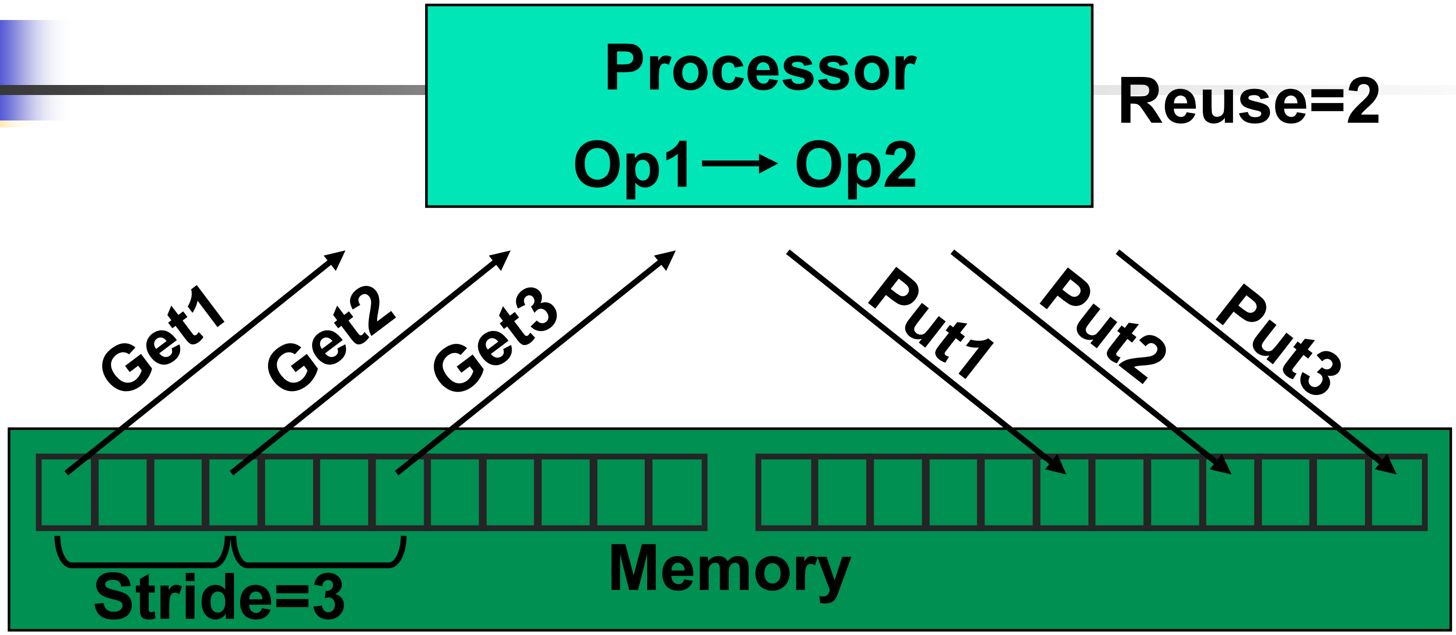
## Corresponding Memory Hierarchy

- **Top500: solves a system**
  **$Ax = b$**
- **STREAM: vector operations**
  **$A = B + s \times C$**
- **FFT: 1D Fast Fourier Transform**
  **$Z = FFT(X)$**
- **RandomAccess: random updates**
  **$T(i) = XOR( T(i), r )$**



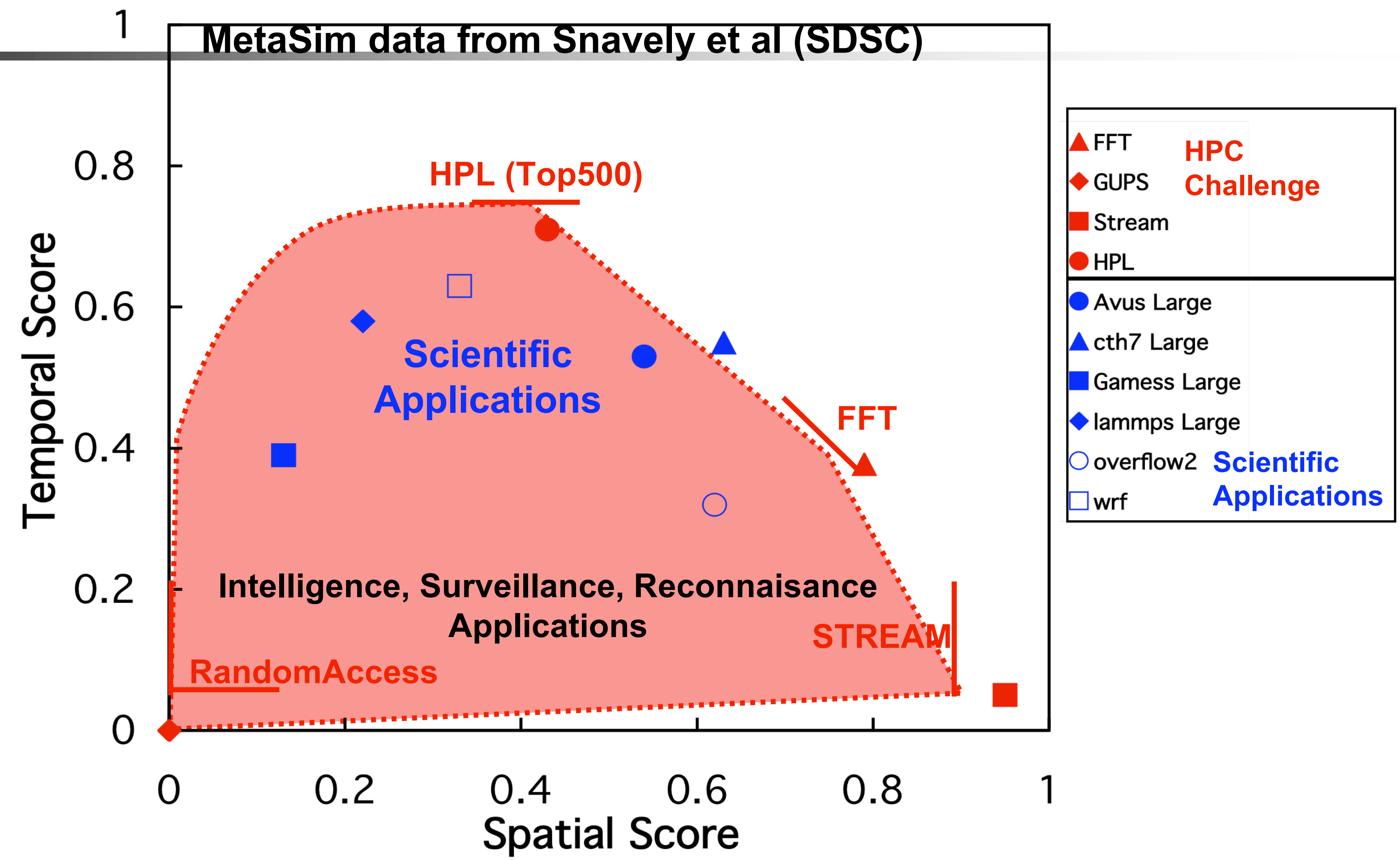

- HPCS program has developed a new suite of benchmarks (HPC Challenge)
- Each benchmark focuses on a different part of the memory hierarchy
- HPCS program performance targets will flatten the memory hierarchy, improve real application performance, and make programming easier

# Spatial and Temporal Locality




- Programs can be decomposed into memory reference patterns
- Stride is the distance between memory references
    - Programs with small strides have high "Spatial Locality"
- Reuse is the number of operations performed on each reference
    - Programs with large reuse have high "Temporal Locality"
- Can measure in real programs and correlate with HPC Challenge

# Spatial/Temporal Locality Results




- HPC Challenge bounds real applications
  - Allows us to map between applications and benchmarks

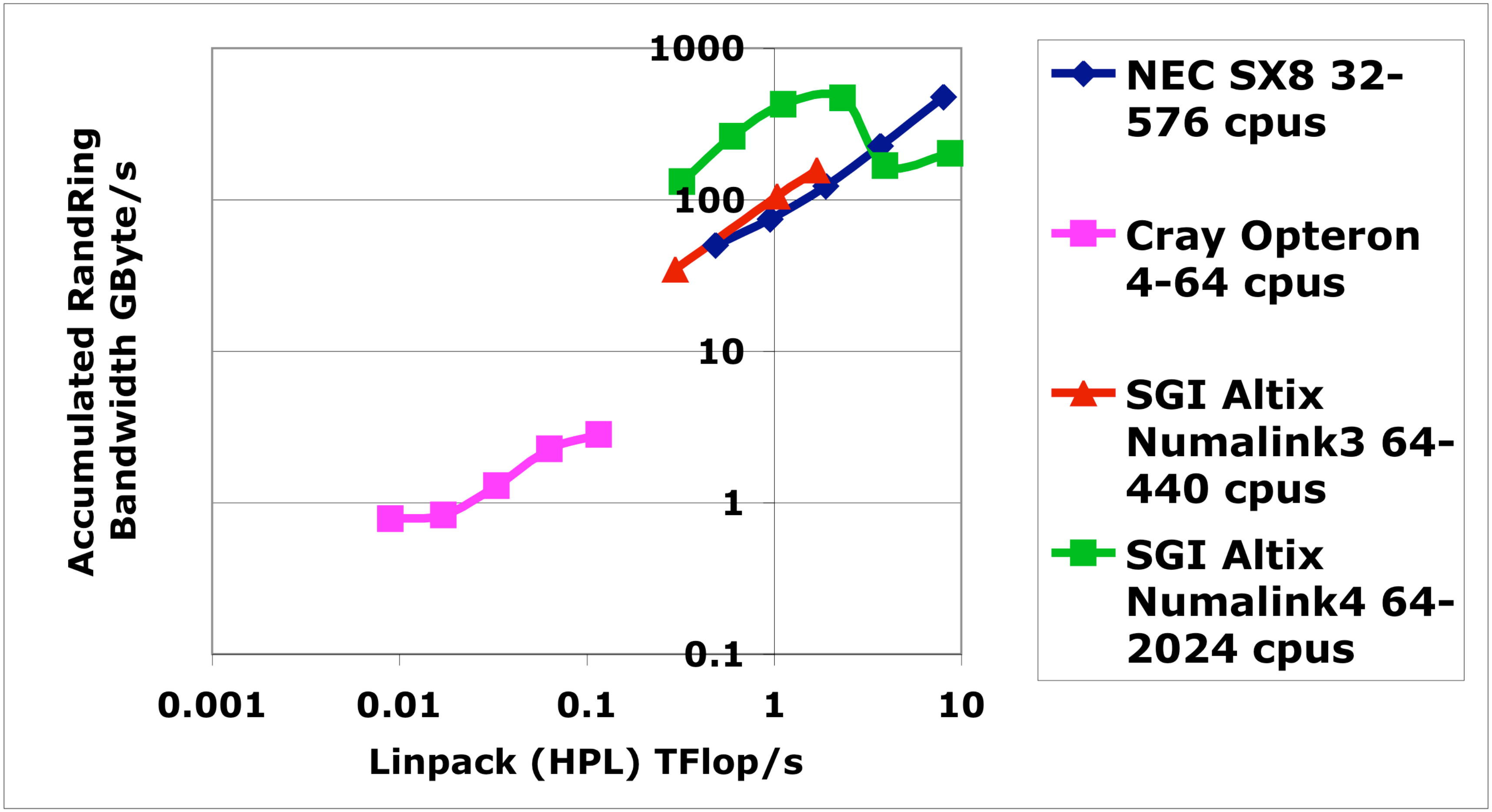
# Intel MPI Benchmarks Used

1. **Barrier:** A barrier function `MPI_Barrier` is used to synchronize all processes.
2. **Reduction:** Each processor provides *A* numbers. The global result, stored at the root processor is also *A* numbers. The number *A[i]* is the results of all the *A[i]* from the *N* processors.
3. **All_reduce:** MPI_Allreduce is similar to MPI_Reduce except that all members of the communicator group receive the reduced result.
4. **Reduce scatter:** The outcome of this operation is the same as an MPI Reduce operation followed by an MPI Scatter
5. **Allgather:** All the processes in the communicator receive the result, not only the root

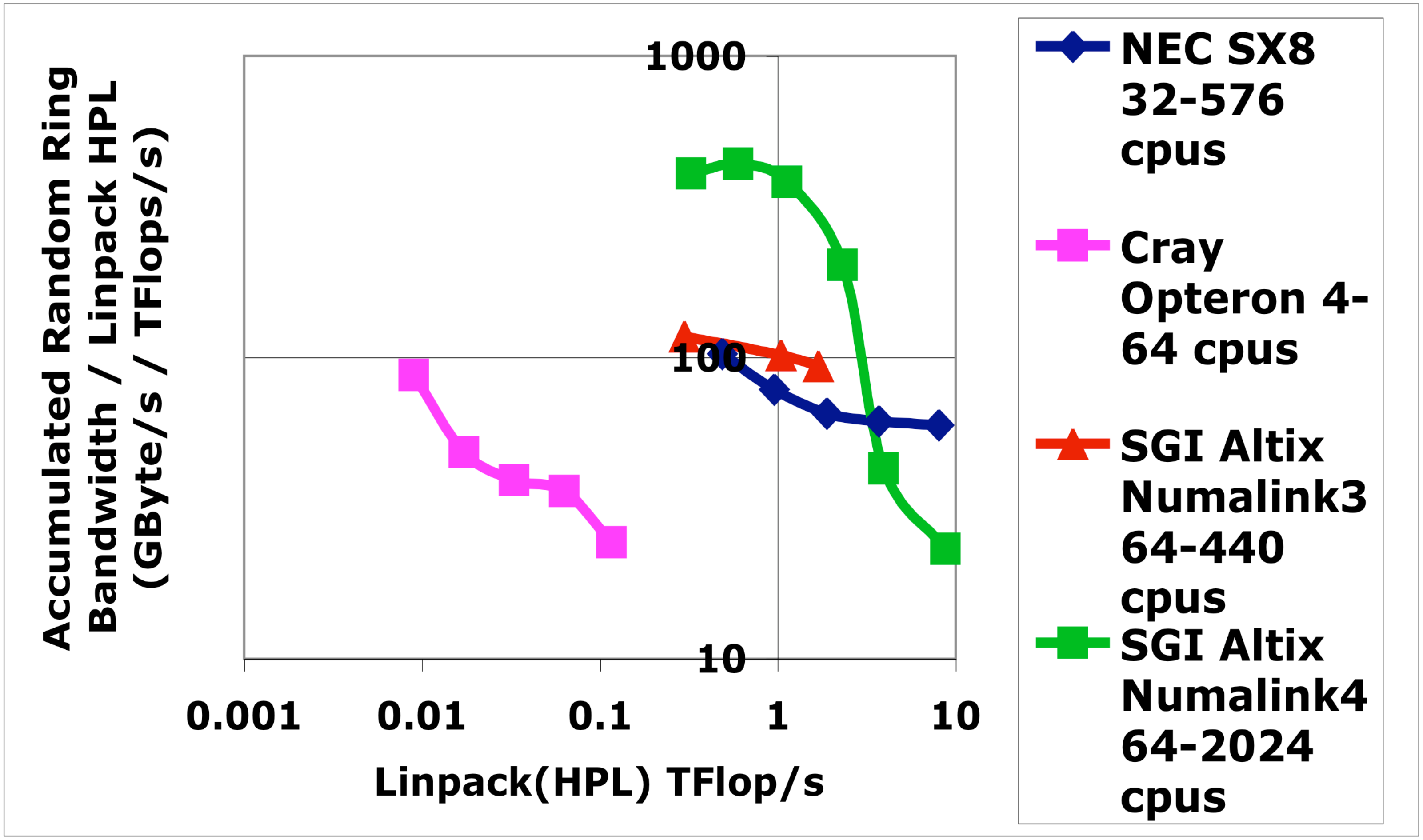
# Intel MPI Benchmarks Used

1. **Allgatherv:** it is vector variant of MPI_ALLgather.
2. **All_to_All:** Every process inputs $A*N$ bytes and receives $A*N$ bytes ($A$ bytes for each process), where $N$ is number of processes.
3. **Send_recv:** Here each process sends a message to the right and receives from the left in the chain.
4. **Exchange:** Here process exchanges data with both left and right in the chain
5. **Broadcast:** Broadcast from one processor to all members of the communicator.

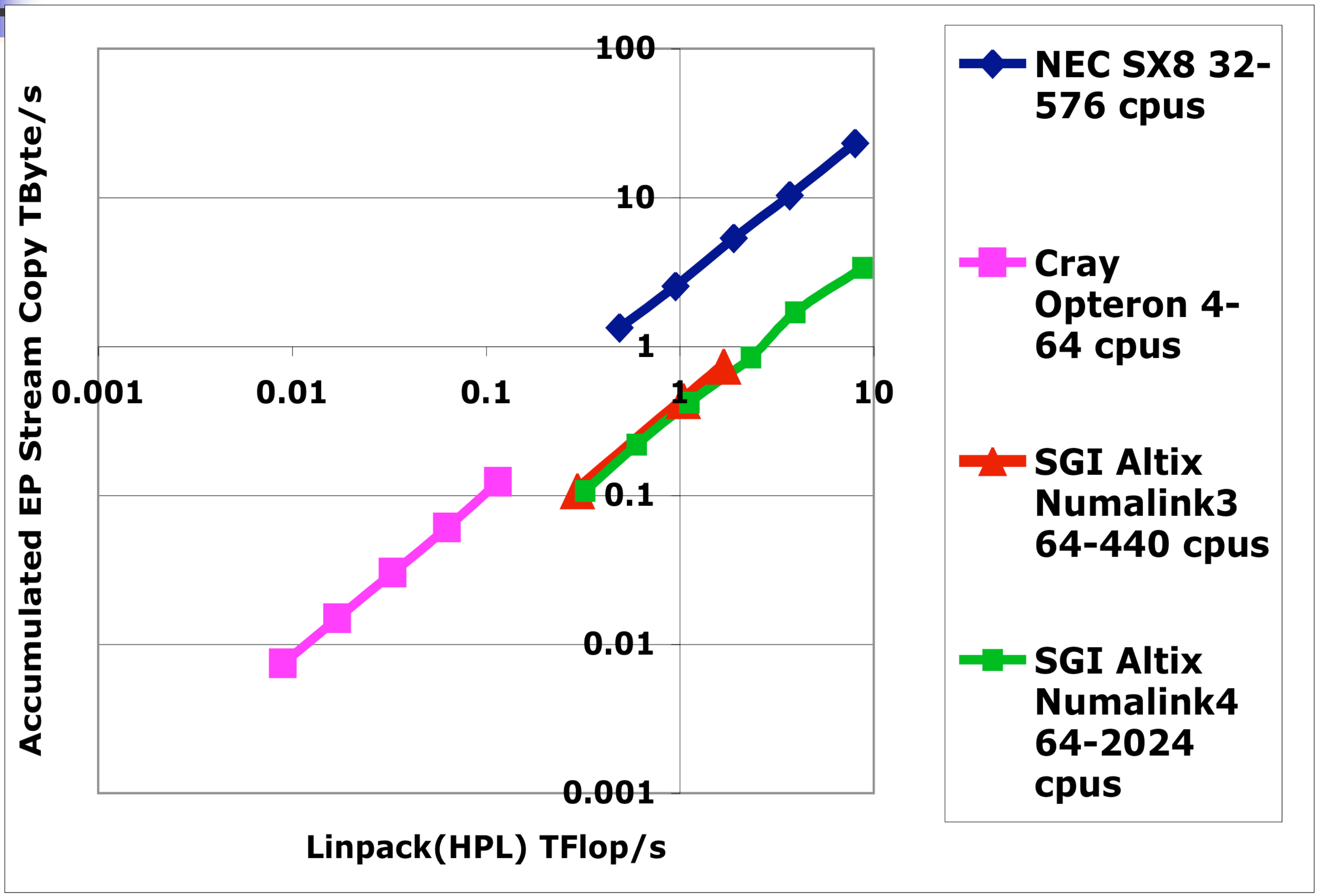
# Accumulated Random Ring BW vs HPL Performance



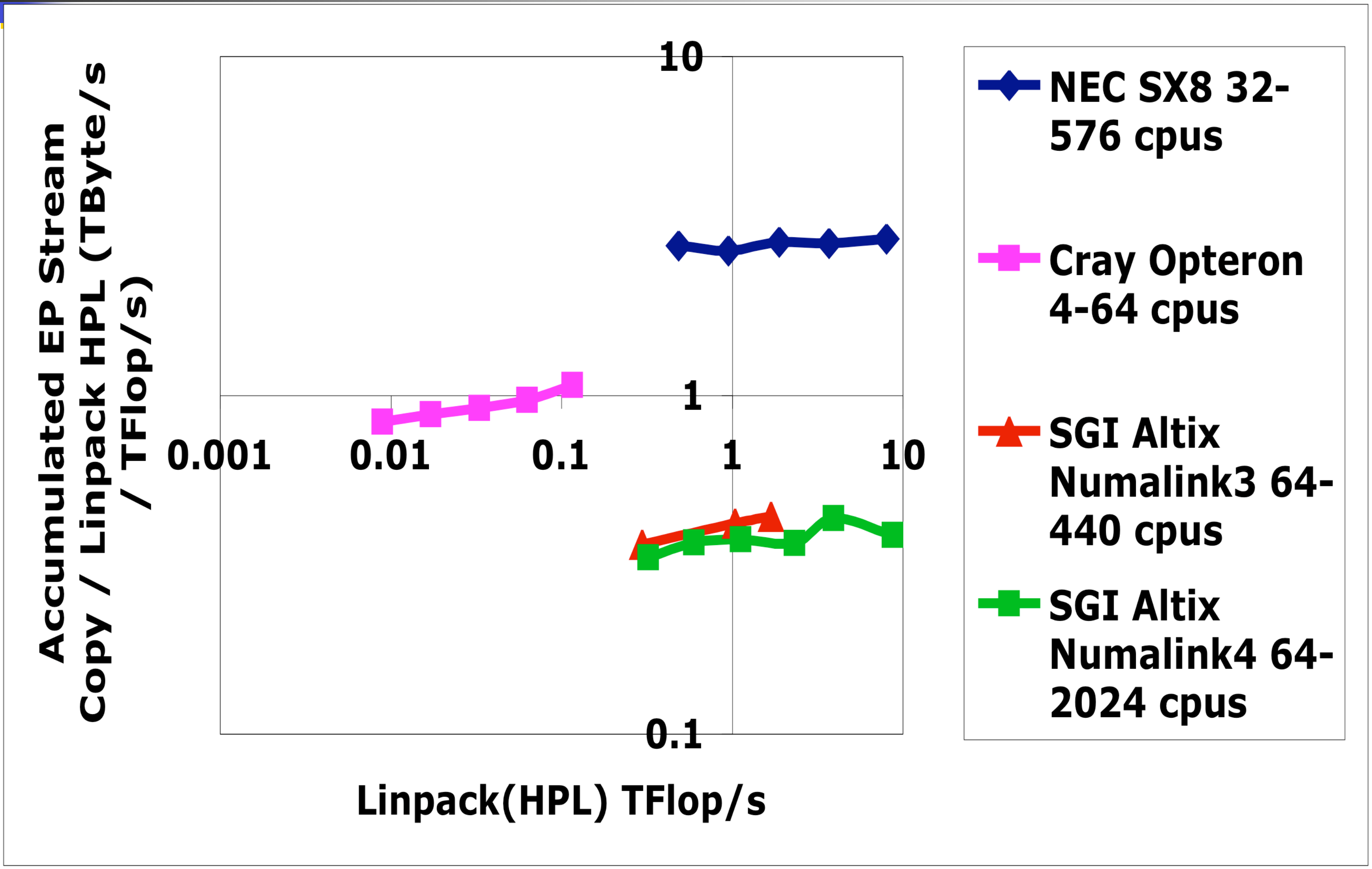
# Accumulated Random Ring BW vs HPL Performance



# Accumulated EP Stream Copy vs HPL Performance



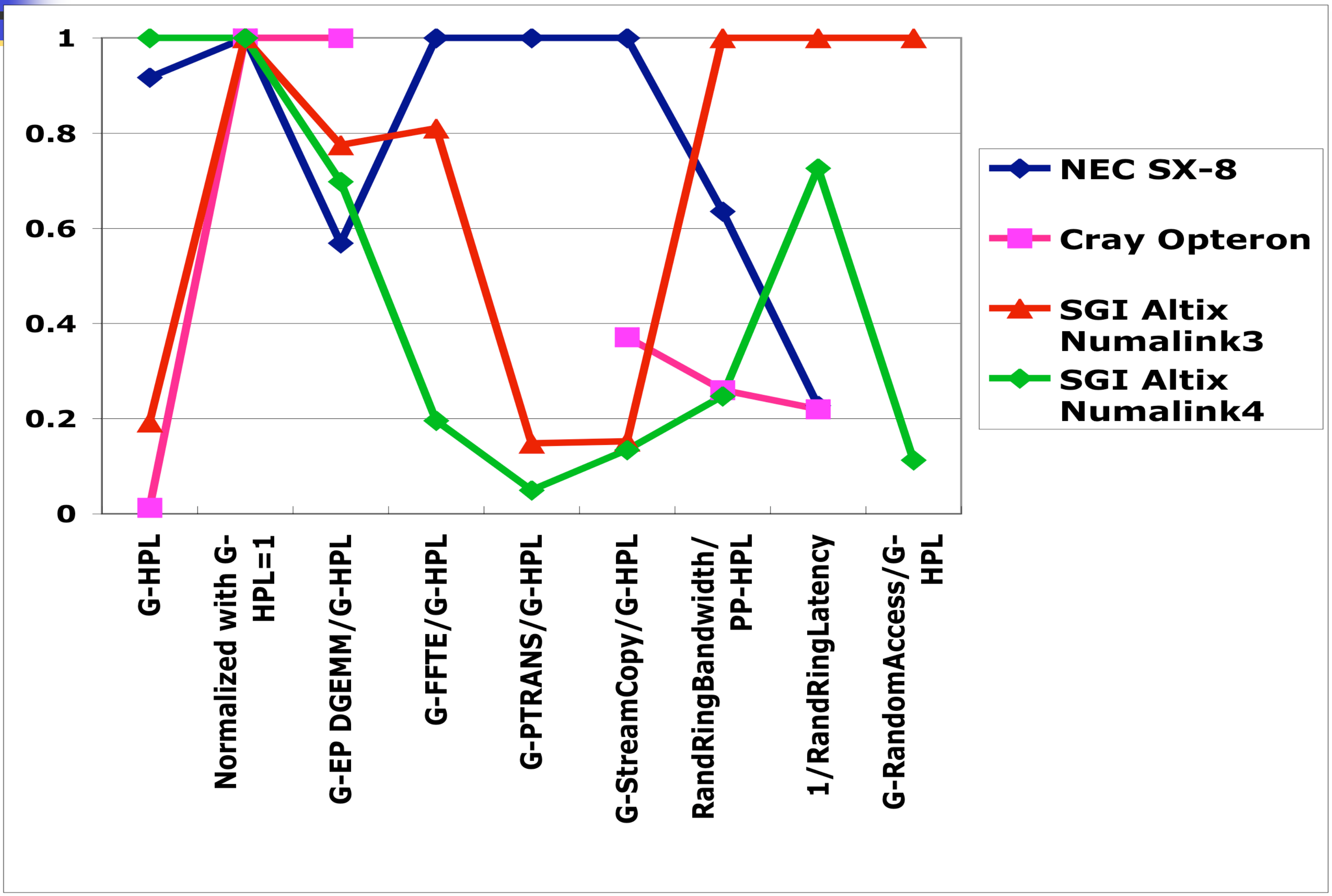
# Accumulated EP Stream Copy vs HPL Performance



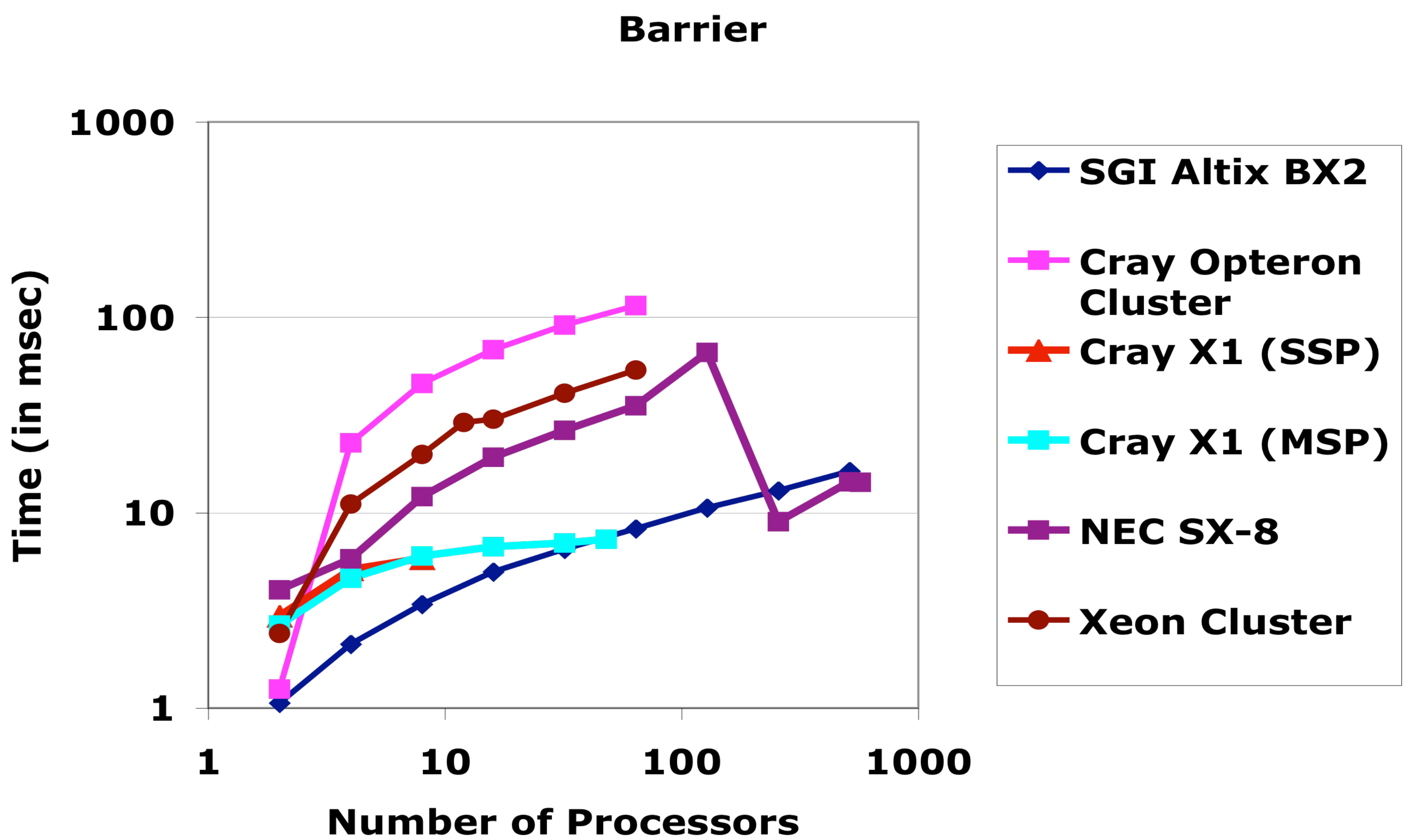
# Normalized Values of HPCC Benchmark

| Ratio | Maximum value |
|---|---|
| G-HPL | 8.729 TF/s |
| G-EP DGEMM/G-HPL | 1.925 |
| G-FFTE/G-HPL | 0.020 |
| G-Ptrans/G-HPL | 0.039 B/F |
| G-StreamCopy/G-HPL | 2.893 B/F |
| RandRingBW/PP-HPL | 0.094 B/F |
| 1/RandRingLatency | 0.197 1/μs |
| G-RandomAccess/G-HPL | 4.9e-5 Update/F |

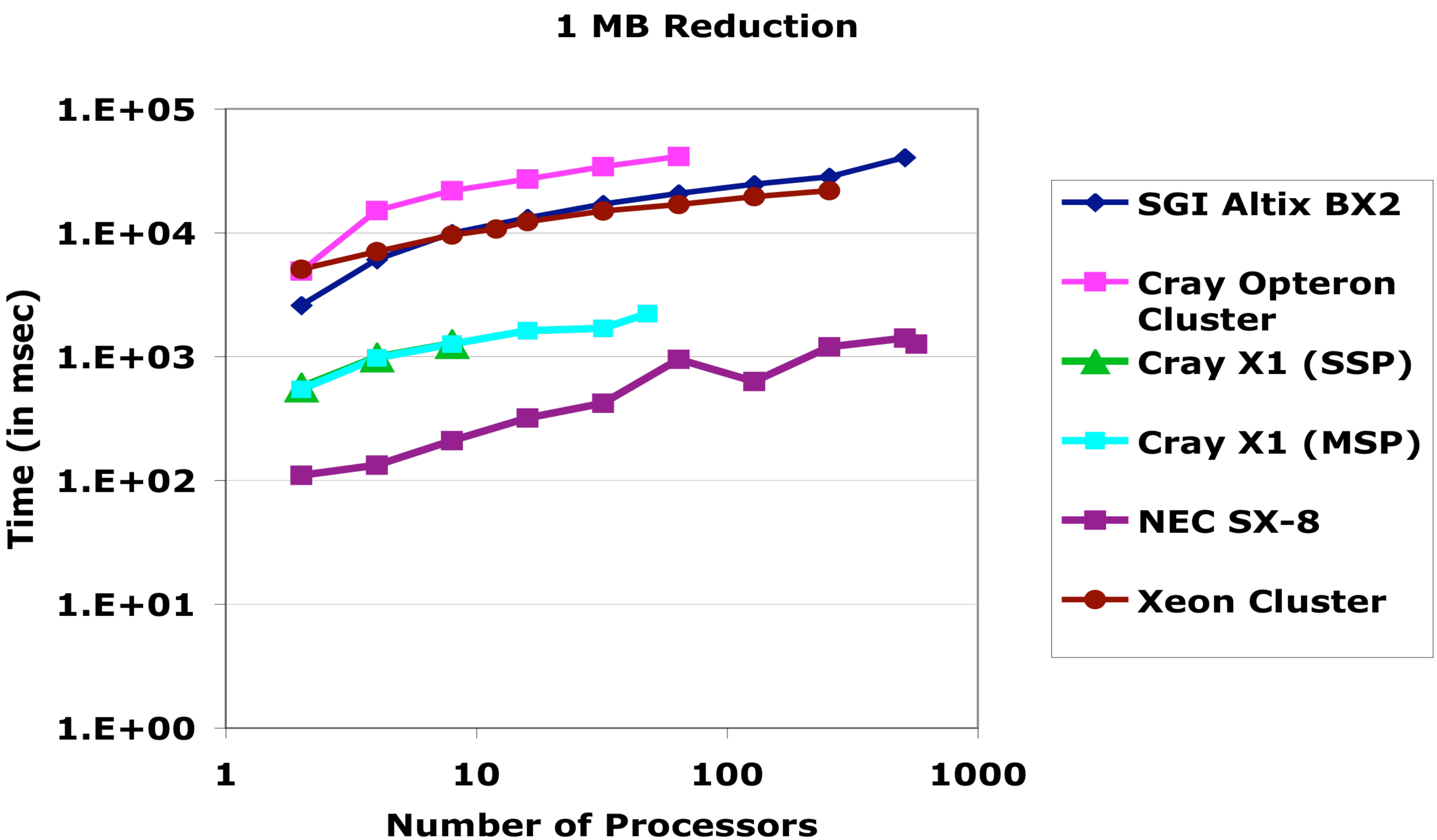
# HPCC Benchmarks Normalized with HPL Value



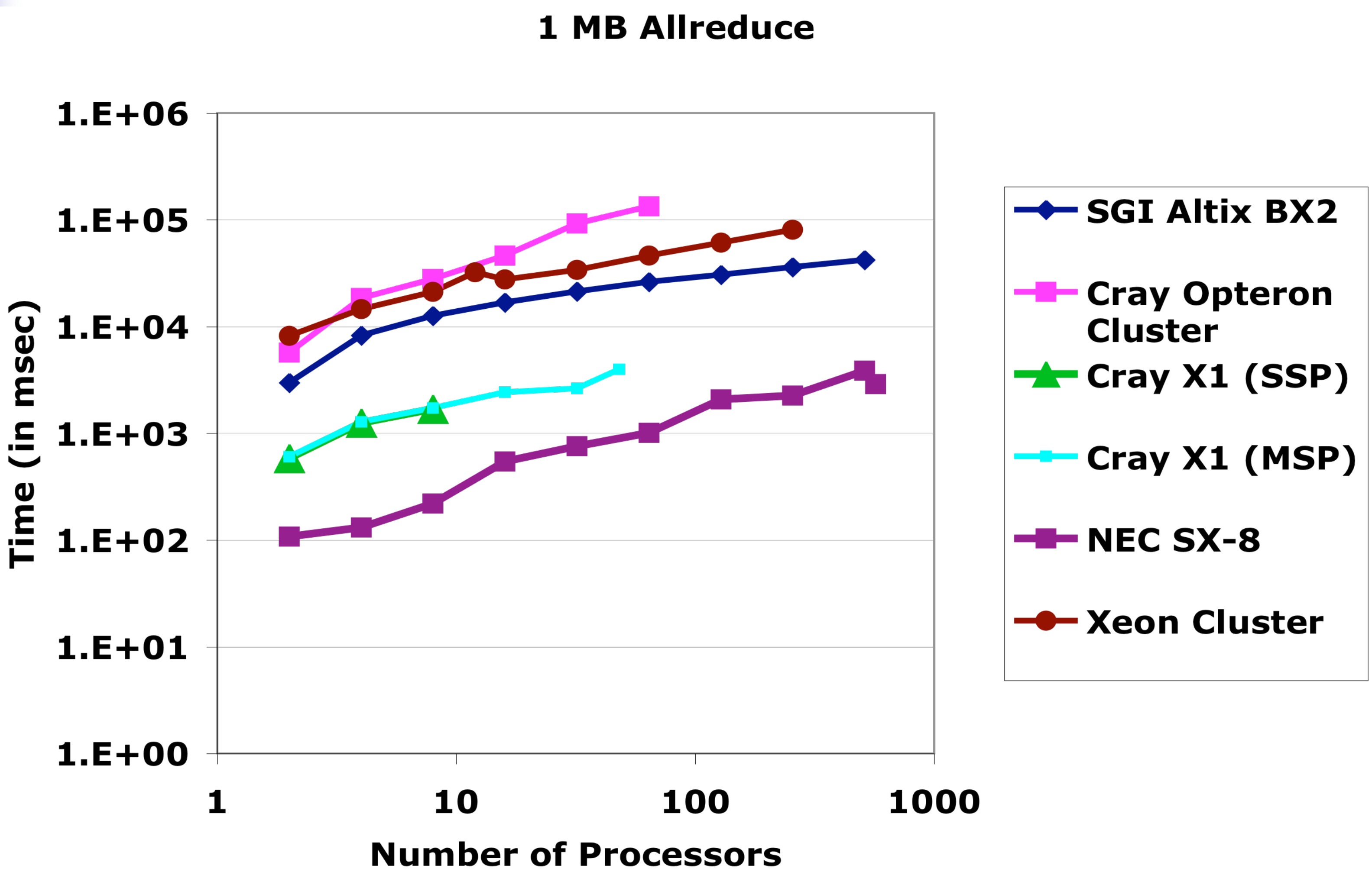
# Barrier Benchmark



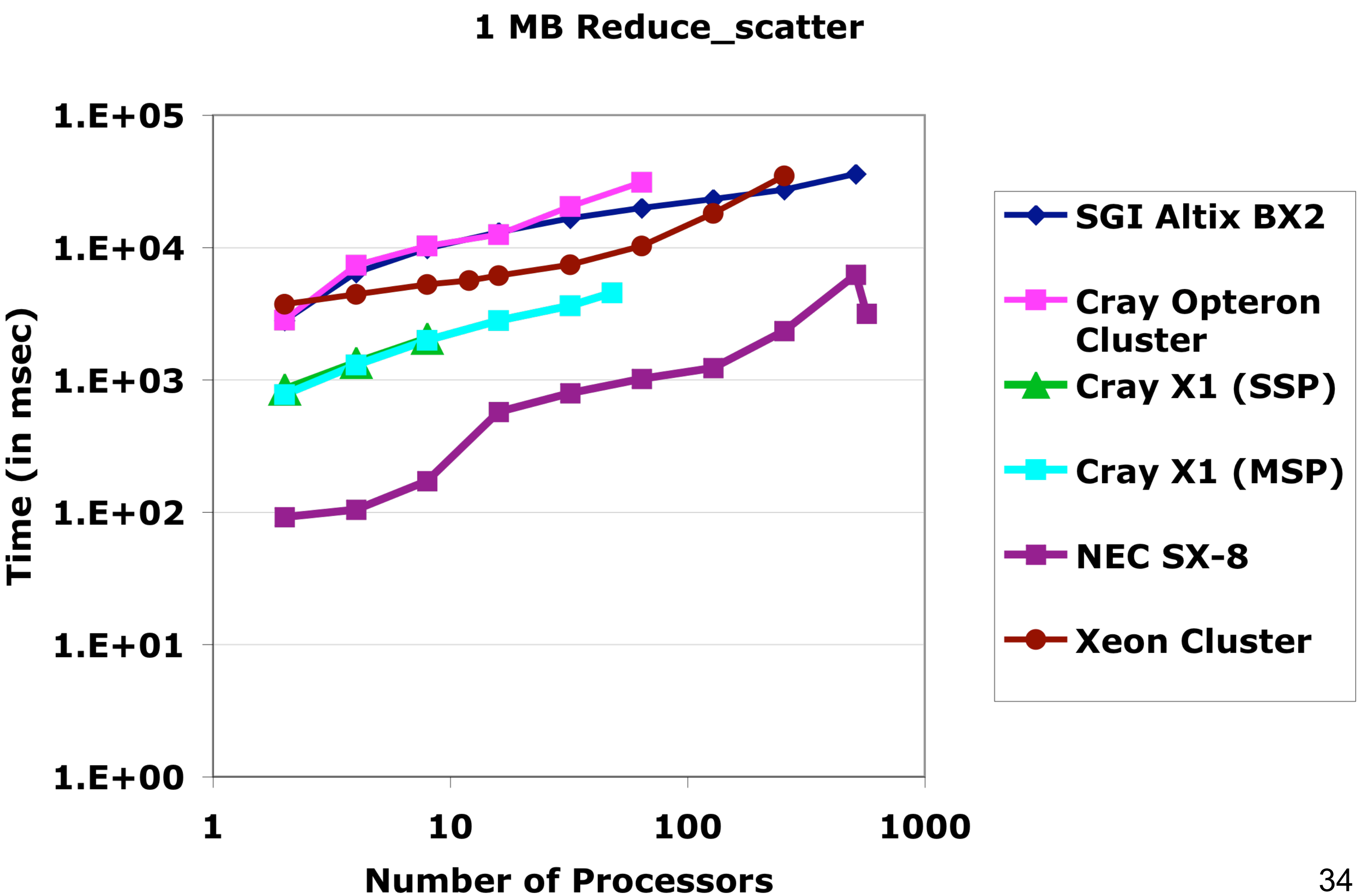
# 1 MB Reduction



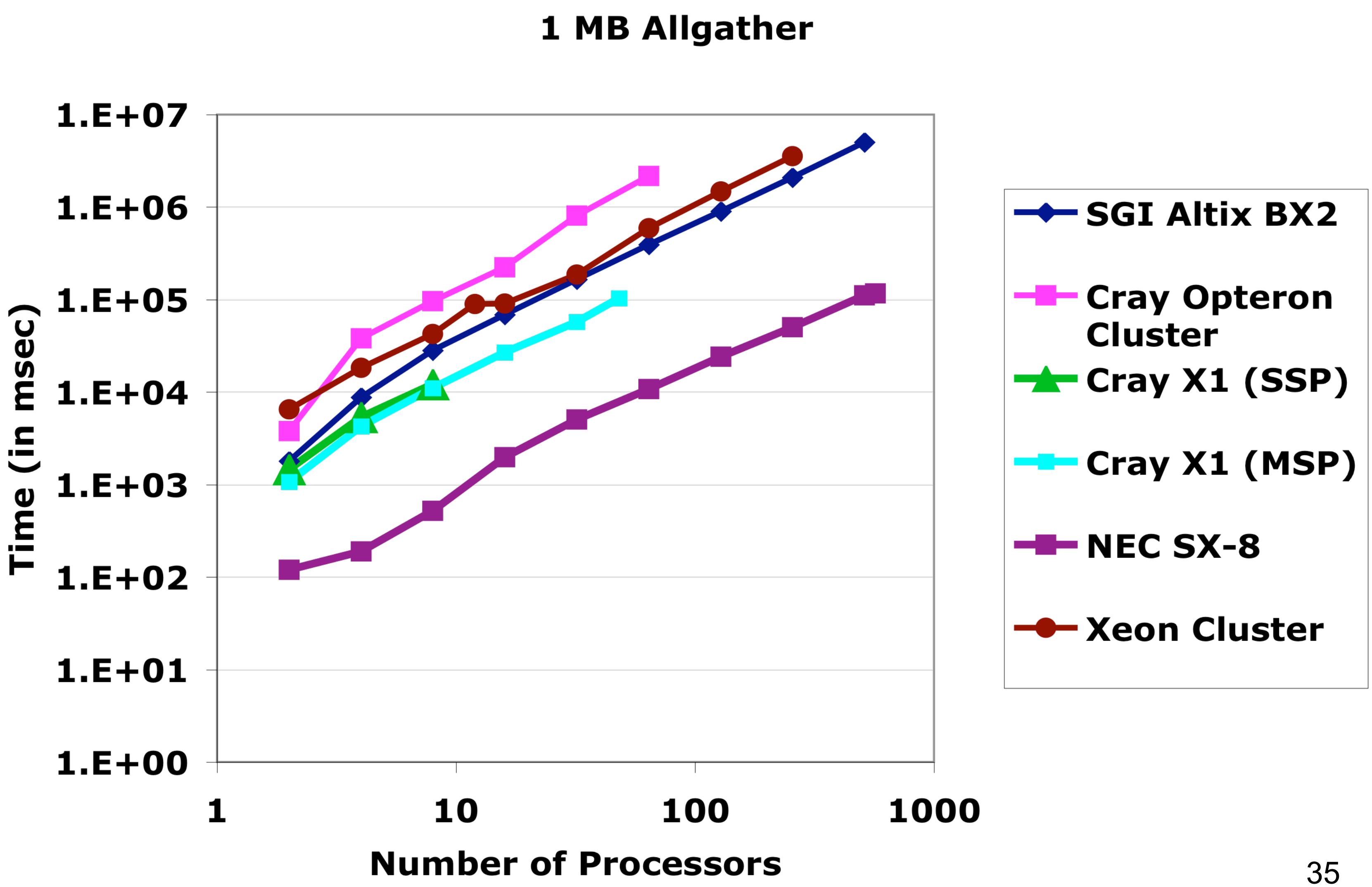
# 1 MB Allreduce



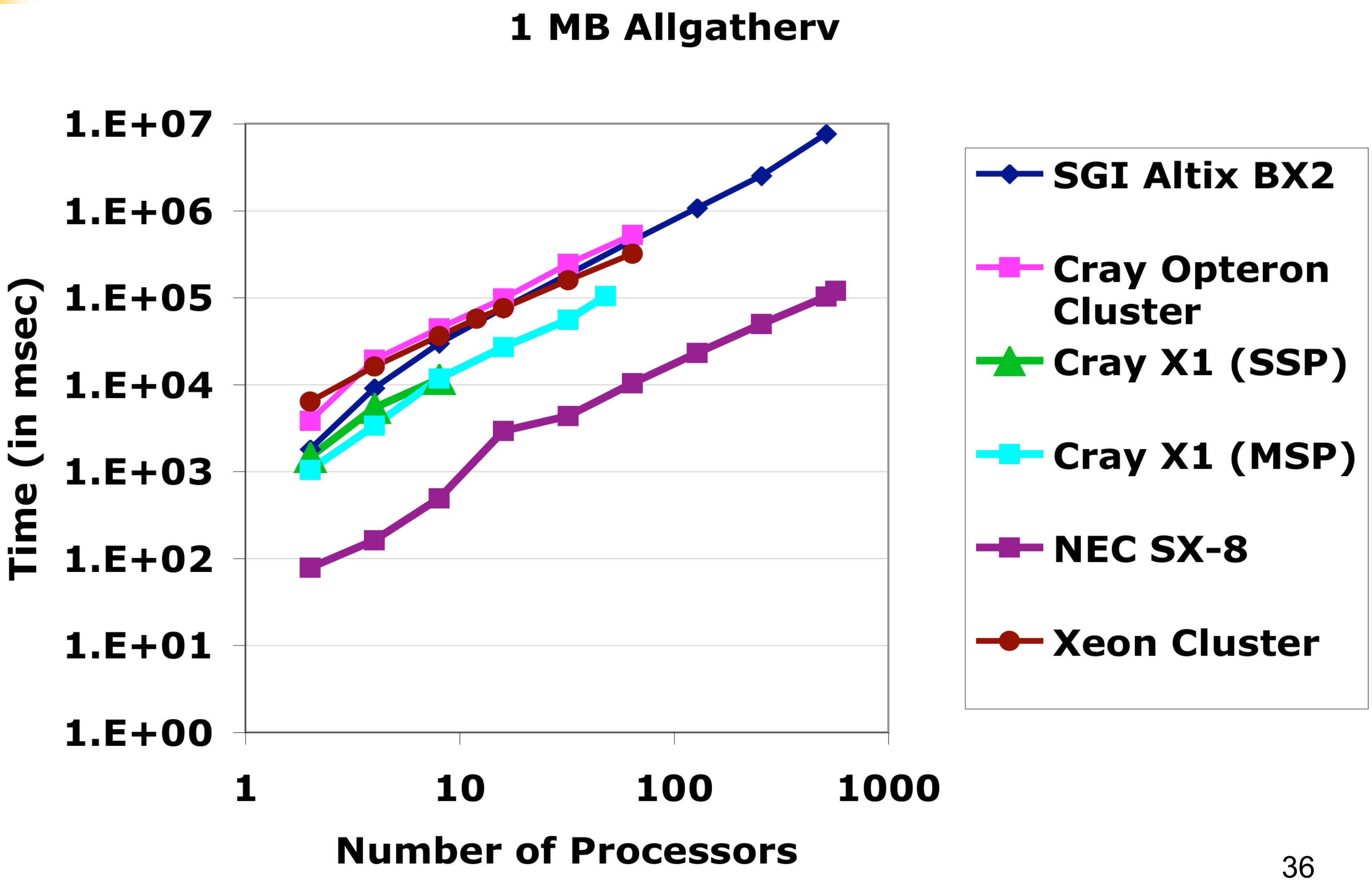
# 1 MB Reduction_scatter



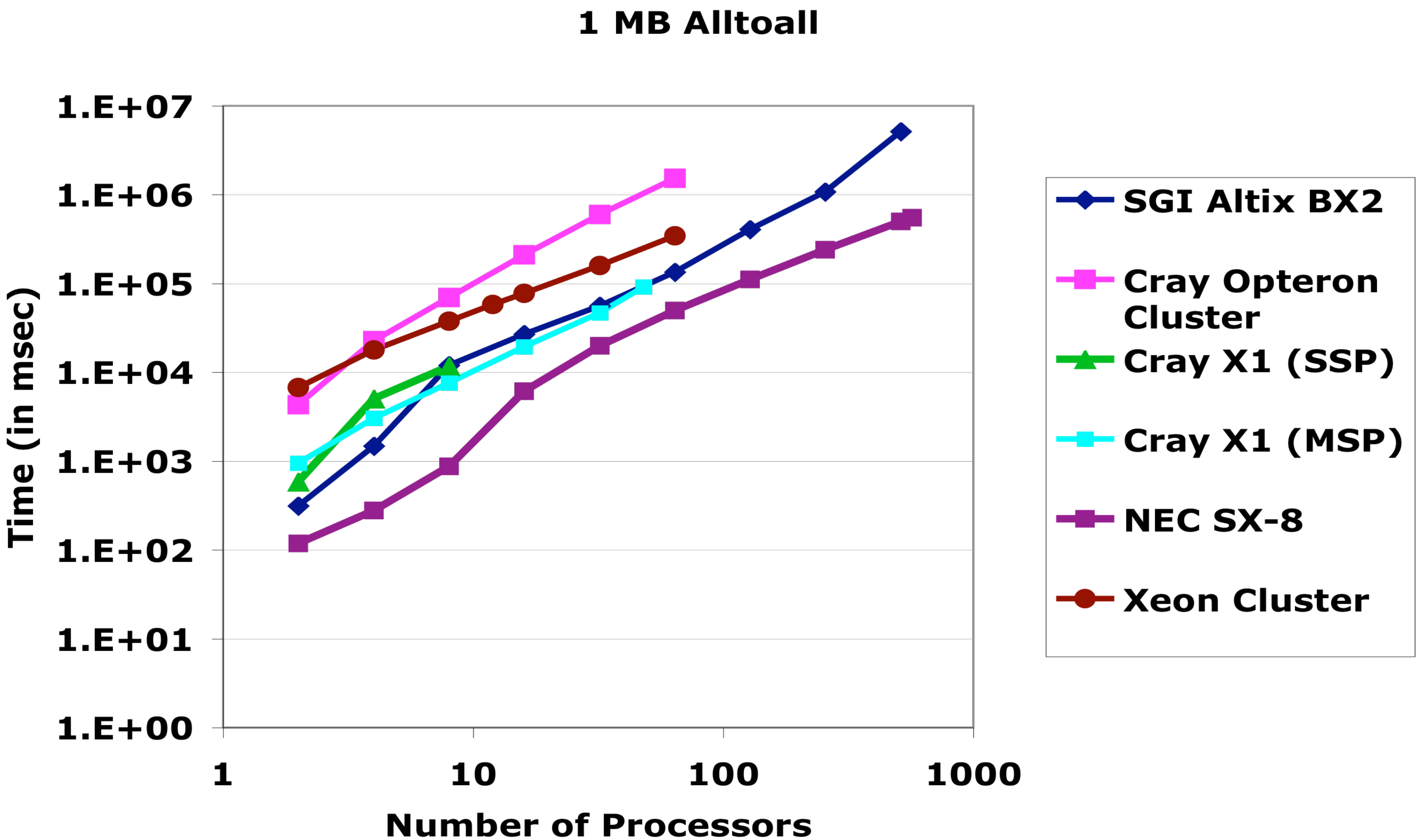
# 1 MB Allgather



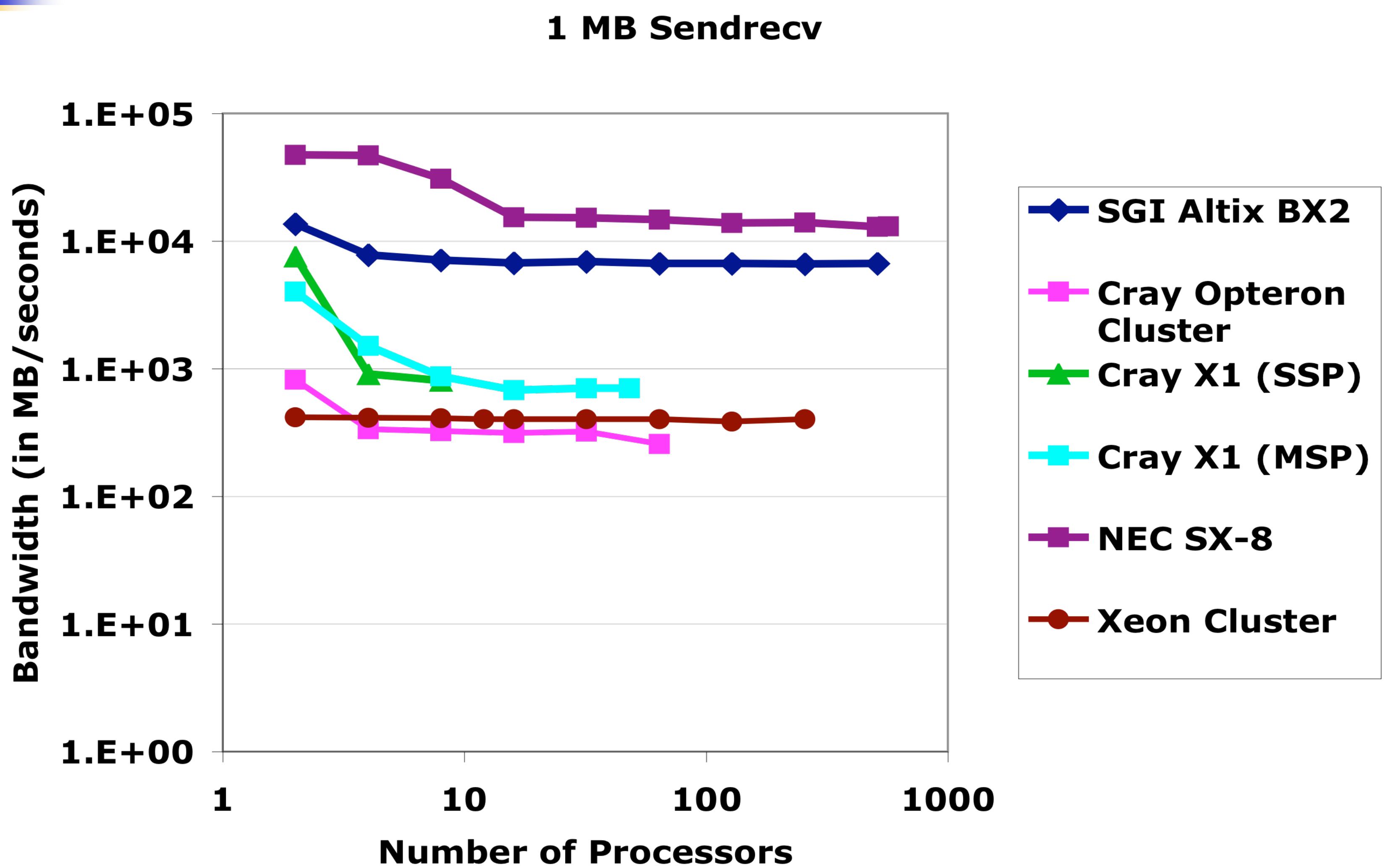
# 1 MB Allgatherv



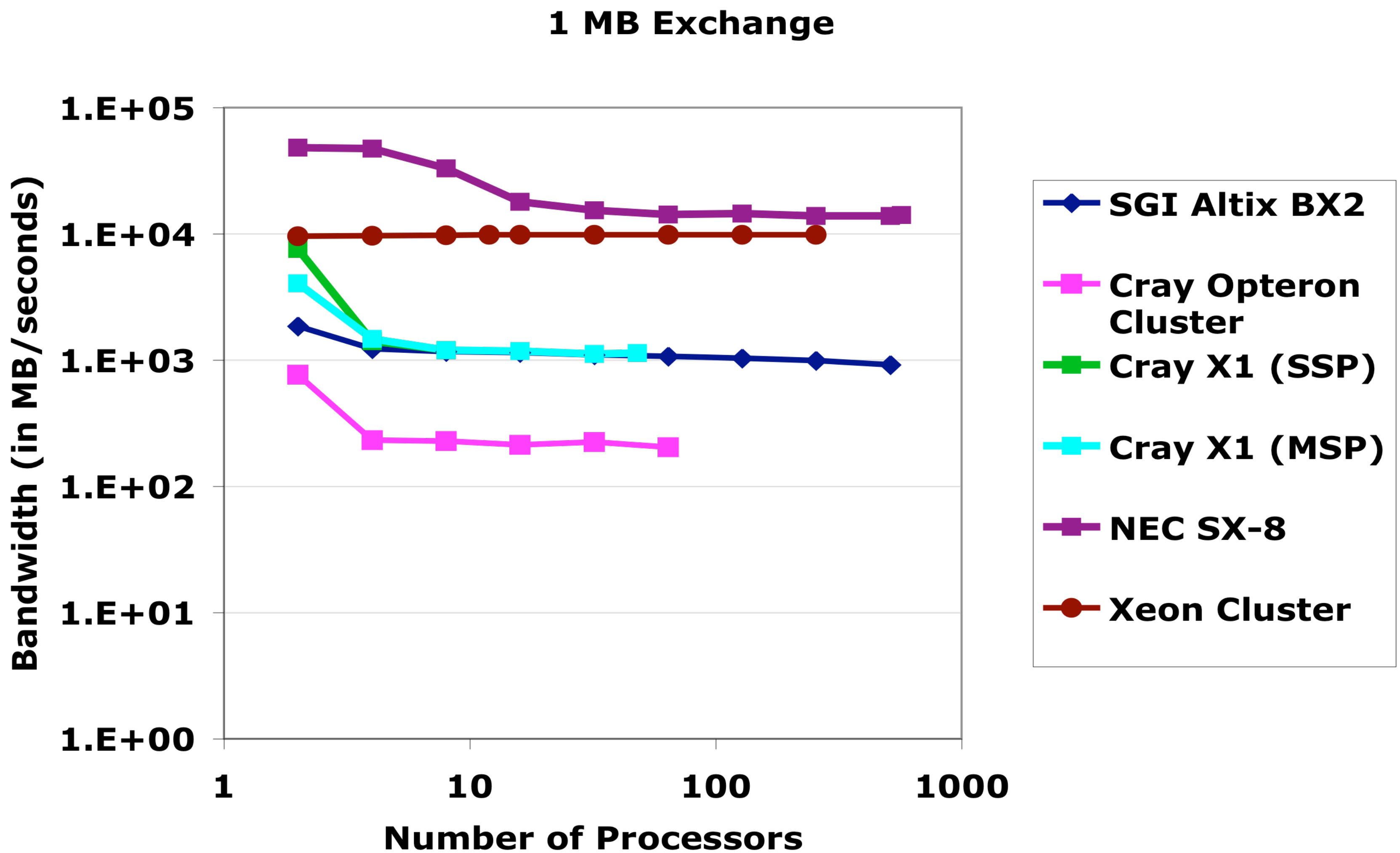
# 1 MB All_to_All



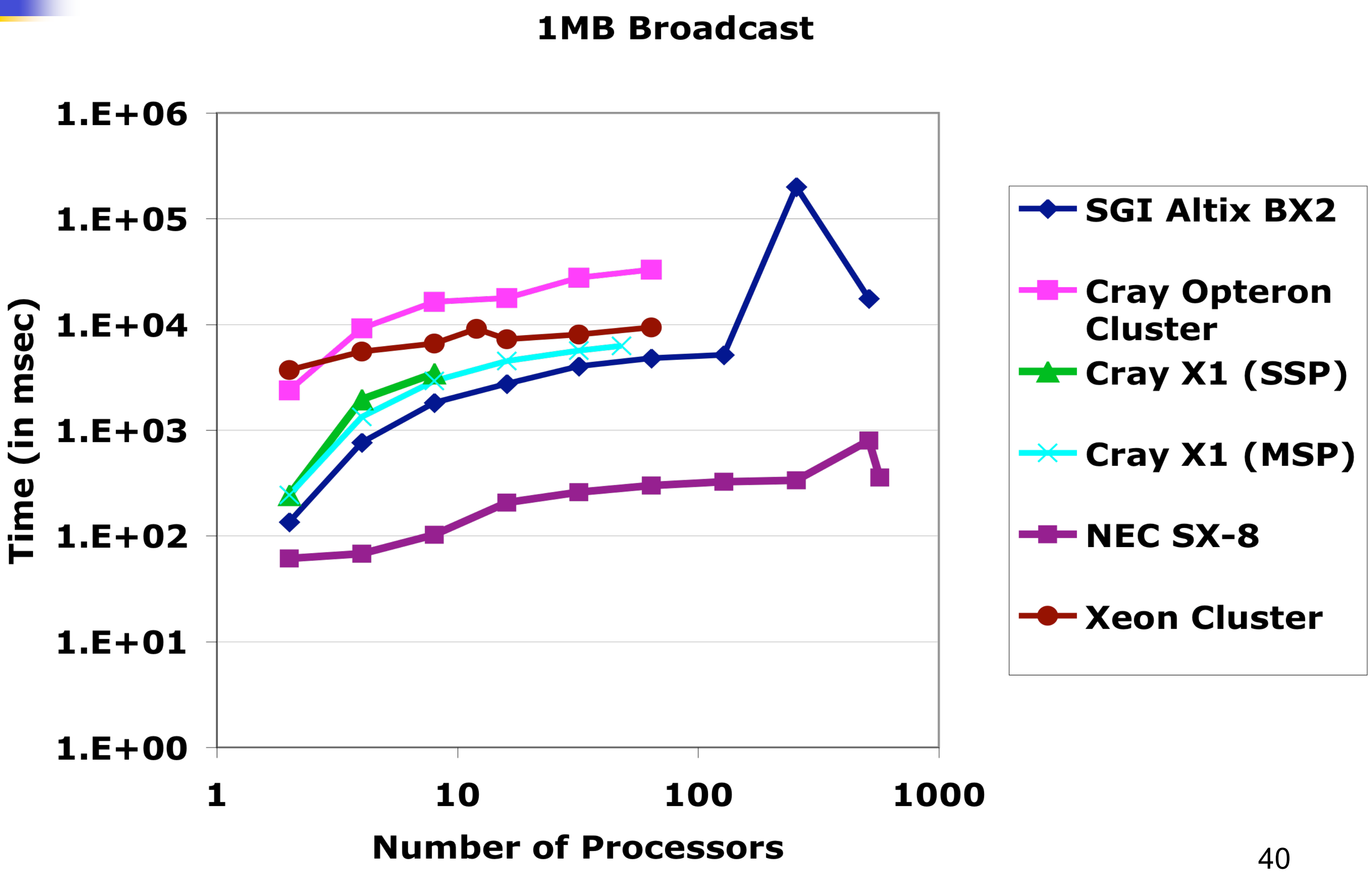
# 1 MB Send_recv



# 1 MB Exchange



# 1 MB Broadcast



# Summary

- Performance of vector systems is consistently better than all the scalar systems
- Performance of SX-8 is better than Cray X1
- Perfprmance of SGI Altix BX2 is better than Dell Xeon cluster and Cray Opteron cluster
- IXS (SX-8) > Cray X1 network > SGI Altix BX2 (NL4) > Dell Xeon cluster (IB) > Cray Opteron cluster (Myrinet).