Performance Comparison of Cray XT4 with SGI Altix 4700, IBM POWER5+, SGI ICE 8200, and NEC SX-8 using HPCC and NPB Benchmarks

Subhash Saini and Dale Talcott

NASA Ames Research Center

Moffett Field, California, USA

and

Rolf Rabenseifner, Michael Schliephake and Katharina Benkert

High-Performance Computing-Center (HLRS)

Nobelstr. 19, D-70550 Stuttgart, Germany

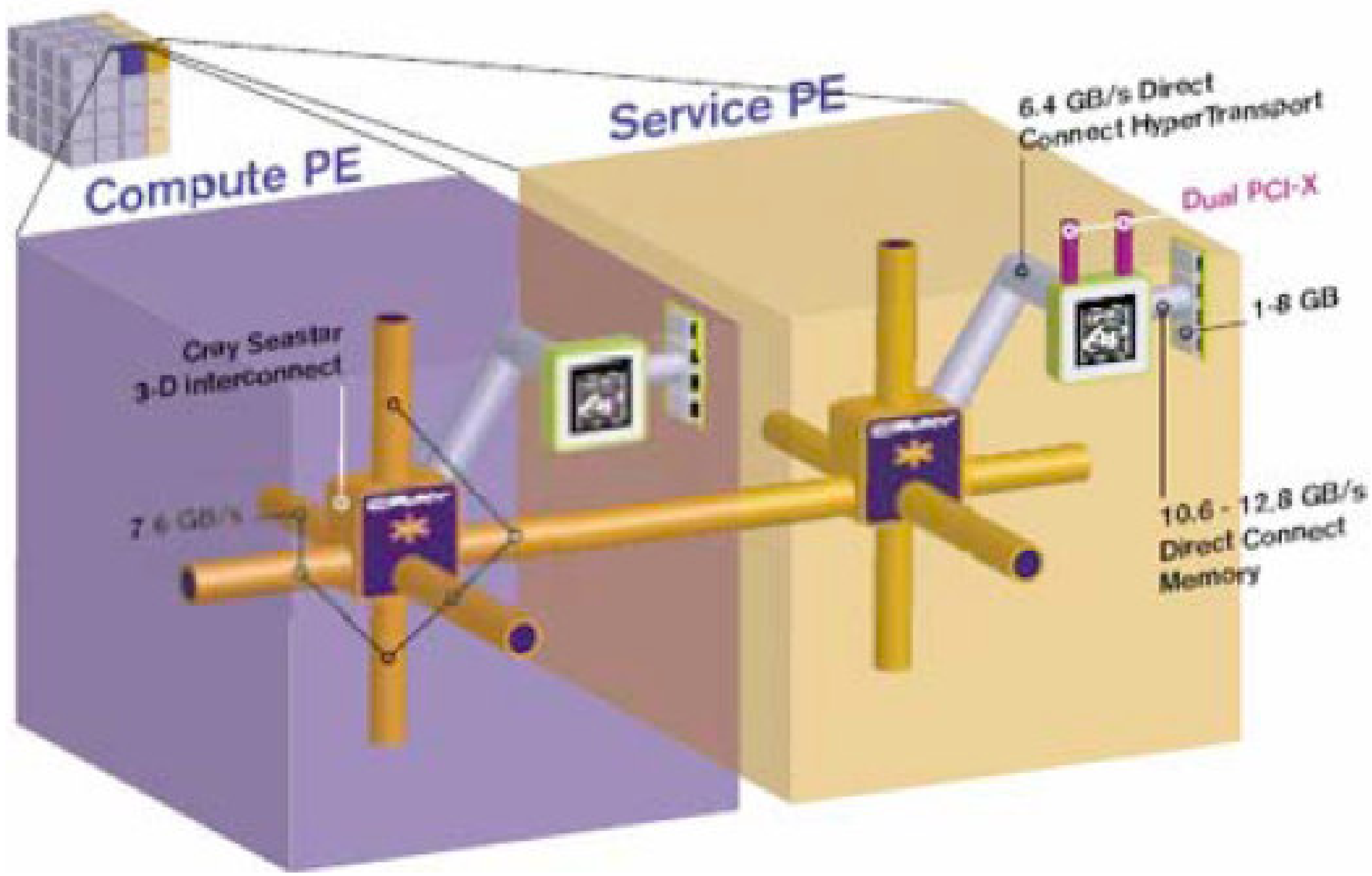
CUG 2008, May 5-8, 2008, Helsinki, Finland

Outline

- **Computing platforms**
  - Cray XT4 (NERSC-LBL, USA) - 2008
  - SGI Altix 4700 (NASA, USA) - 2007
  - IBM POWER5+ (NASA, USA) - 2007
  - SGI ICE 8200 (NASA, USA) - 2008
  - NEC SX-8 (HLRS, Germany) - 2006
- **Benchmarks**
  - HPCC 1.0 Benchmark suite
  - NPB 3.3 MPI Benchmarks
- **Summary and conclusions**

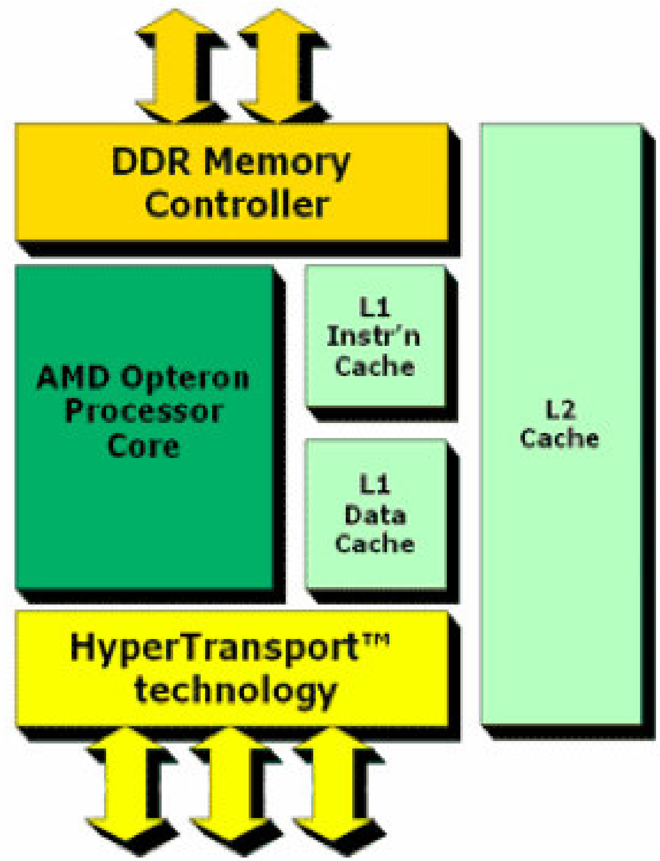
Cray XT4

Cray XT4 Scalable Architecture



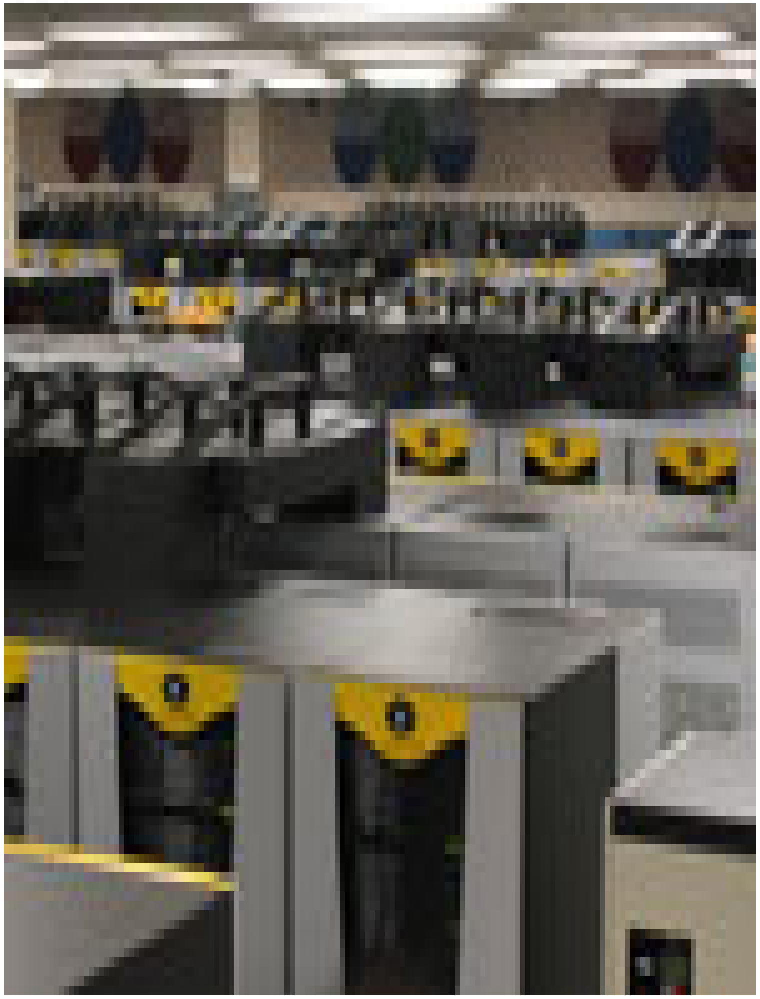
Cray XT4

- Dual-core AMD Opteron
- Core clock frequency 2.6 GHz
- Two floating operations per clock per core
- Peak performance per core is 5.2 Gflop/s
- L1 cache 64 KB (I) and 64 KB (D)
- L2 cache 1MB unified
- L3 cache is not available
- 2 cores per node
- Local memory pet node is 4 GB
- Local memory pert core is 2 GB
- Frequency of FSB is 800 MHz
- Transfer rate of FSB is 12.8 GB/s
- Interconnect is Sea Star 2
- Network topology is mesh.
- Operating system is Linux SLES 9.2
- Fortran compiler is pgi
- C compiler is Intel pgi
- MPI is Cray implementation



SGI Altix 4700 System

- Dual-core Intel Itanium 2 (Montvale)
- Core clock frequency 1.67 GHz
- Four floating operations per clock per core
- Peak performance per core is 6.67 Gflop/s
- L1 cache 32 KB (I) and 32 KB (D)
- L2 cache 256 (I+D)
- L3 cache is 9 MB on-chip
- 4 cores per node
- Local memory pet node is 8 GB
- Local memory pert core is 2 GB
- Frequency of FSB is 667 MHz
- Transfer rate of FSB is 10.6 GB/s
- Interconnect is NUMAlnk4
- Network topology is fat tree
- Operating system is Linux SLES 10
- Fortran compiler is Intel 10.0.026
- C compiler is Intel 10.0/026
- MPI is mpt-1.16.0.0



IBM POWER5+ Cluster

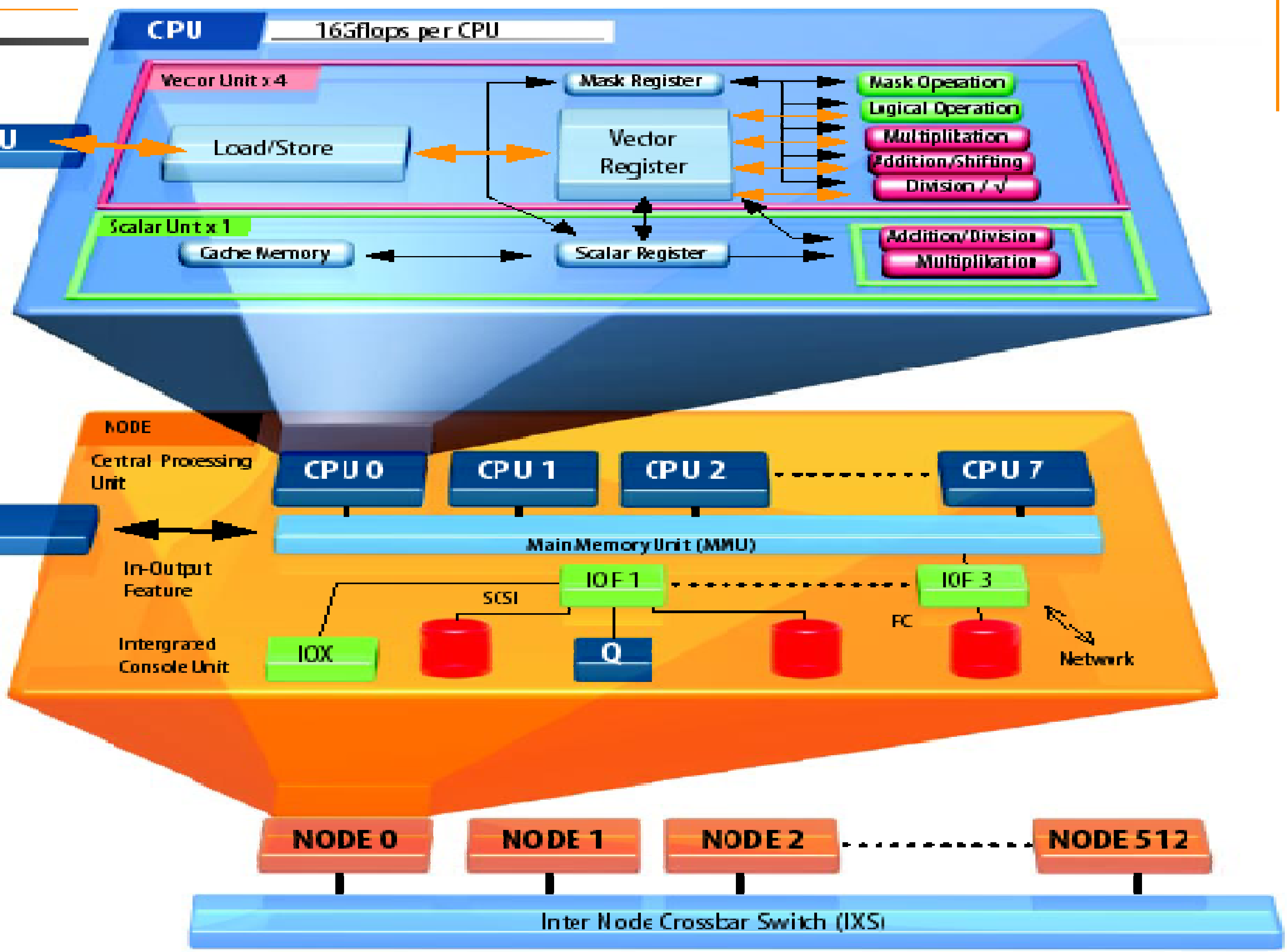
- Dual-core IBM POWER5+ processor
- Core clock frequency 1.9 GHz
- Four floating operations per clock per core
- Peak performance per core is 7.6 Gflop/s
- L1 cache 64 KB (I) and 32 KB (D)
- L2 cache 1.92 MB (I+D) shared
- L3 cache is 36 MB and is off-chip
- 16 cores per node
- Local memory pet node is 32 GB
- Local memory pert core is 2 GB
- Frequency of FSB is 533 MHz
- Transfer rate of FSB is 8.5 GB/s
- Interconnect is HPS (Federation)
- Network topology is multi-stage.
- Operating system is AIX 5.3
- Fortran compiler is xlf 10.1
- C compiler is xlc 9.0
- MPI is POE 4.3

SGI Altix ICE 8200 Cluster

- Quad-core Intel Xeon (Clovertown)
- Core clock frequency 2.66 GHz
- Four floating operations per clock per core
- Peak performance per core is 10.64 Gflop/s
- L1 cache 32 KB (I) and 32 KB (D)
- L2 cache 8 MB shared by two cores
- L3 cache is not available
- 8 cores per node
- Local memory pet node is 8 GB
- Local memory pert core is 1 GB
- Frequency of FSB is 1333 MHz
- Transfer rate of FSB is 10.7 GB/s
- Interconnect is Infiniband
- Network topology is hypercube.
- Operating system is Linux SLES 10
- Fortran compiler is Intel 10.1.008
- C compiler is Intel 10.1.008
- MPI is mpt-1.18.b30

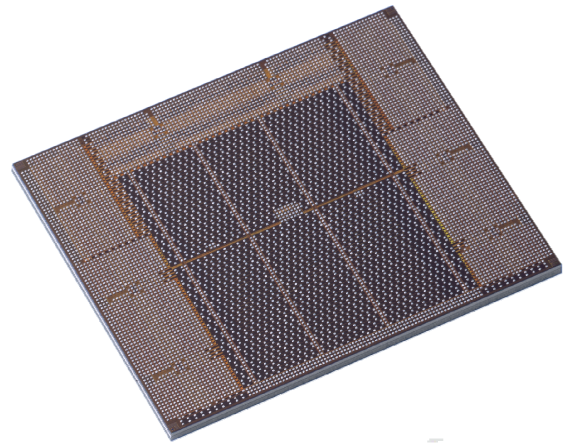
NEC SX-8 System

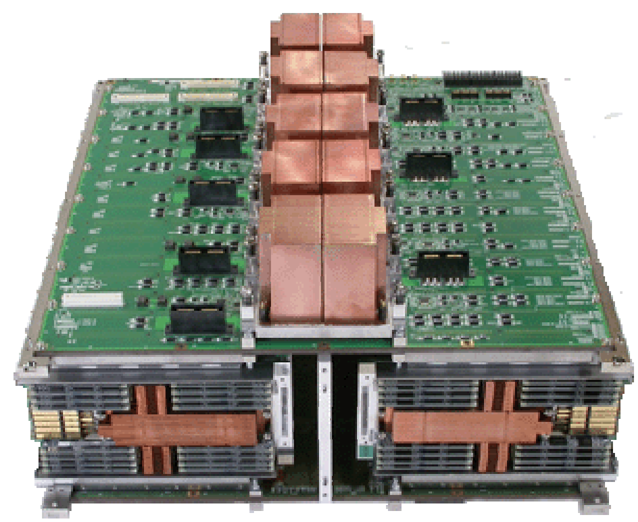
SX-8 System Architecture



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

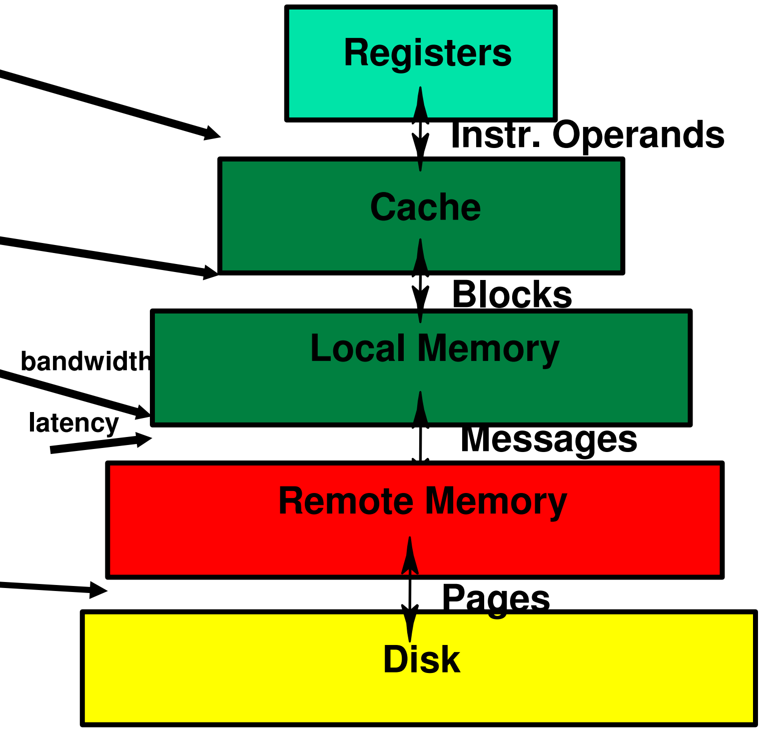
HPC Challenge Benchmarks

- 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

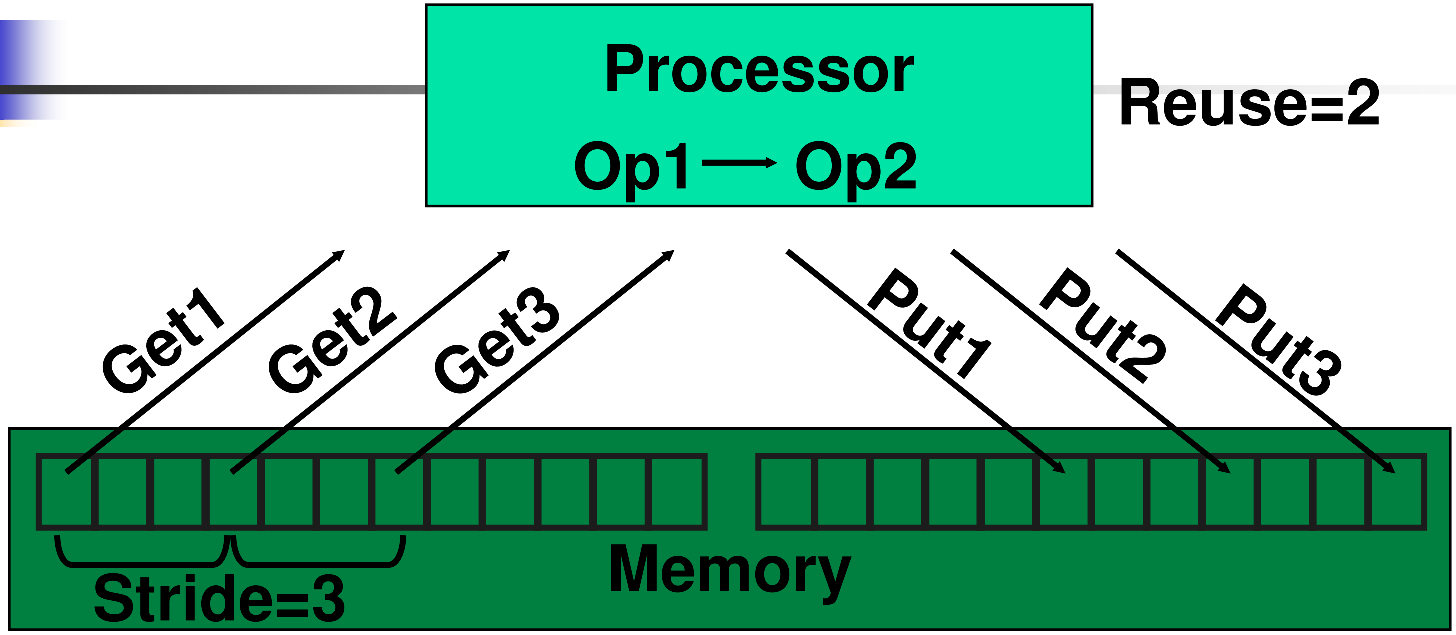
- **Top500: solves a system**
  $Ax = b$
- **STREAM: vector operations**
  $A = B + s \times C$
- **FFT: 1D Fast Fourier Transform**
  $Z = FFT(X)$
- **RandomAccess: random updates**
  $T(i) = XOR(\ T(i), r\ )$

Corresponding Memory Hierarchy



- 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

NAS Parallel Benchmarks (NPB)

- Kernel benchmarks
  - **MG:** multi-grid on a sequence of meshes, long- & short-distance communication, memory intensive
  - **FT:** discrete 3D FFTs, all-to-all communication
  - **IS:** integer sort, random memory access
  - **CG:** conjugate gradient, irregular memory access and communication
  - **EP:** embarrassingly parallel
- Application benchmarks
  - **BT:** block tri-diagonal solver
  - **SP:** scalar penta-diagonal solver
  - **LU:** lower-upper Gauss Seidel

Benchmark Classes

- Class S - small (~1 MB)
  - any quick test
- Class W - workstation (a few MB)
  - used to be, now too small
- Classes A, B, C
  - standard test problems
  - 4x size increase going from one class to the next
- Class D
  - about 16x of Class C
- Class E
  - About 16x of Class D

NPB Implementations

- The original NPB
  - paper-and-pencil specifications
  - useful for measuring efficiency of parallel computers, parallel tools for scientific applications
  - well-understood, generally accepted
  - decent reference implementations available
    - MPI (3.2.1), OpenMP (NPB3.2.1)
    - NPB 3.3
- Multi-zone versions of NPB
  - from application benchmarks: **LU-MZ**, **SP-MZ**, **BT-MZ**
  - exploit multi-level parallelism
  - test load balancing schemes
  - hybrid implementation
    - MPI+OpenMP (NPB3.2-MZ)

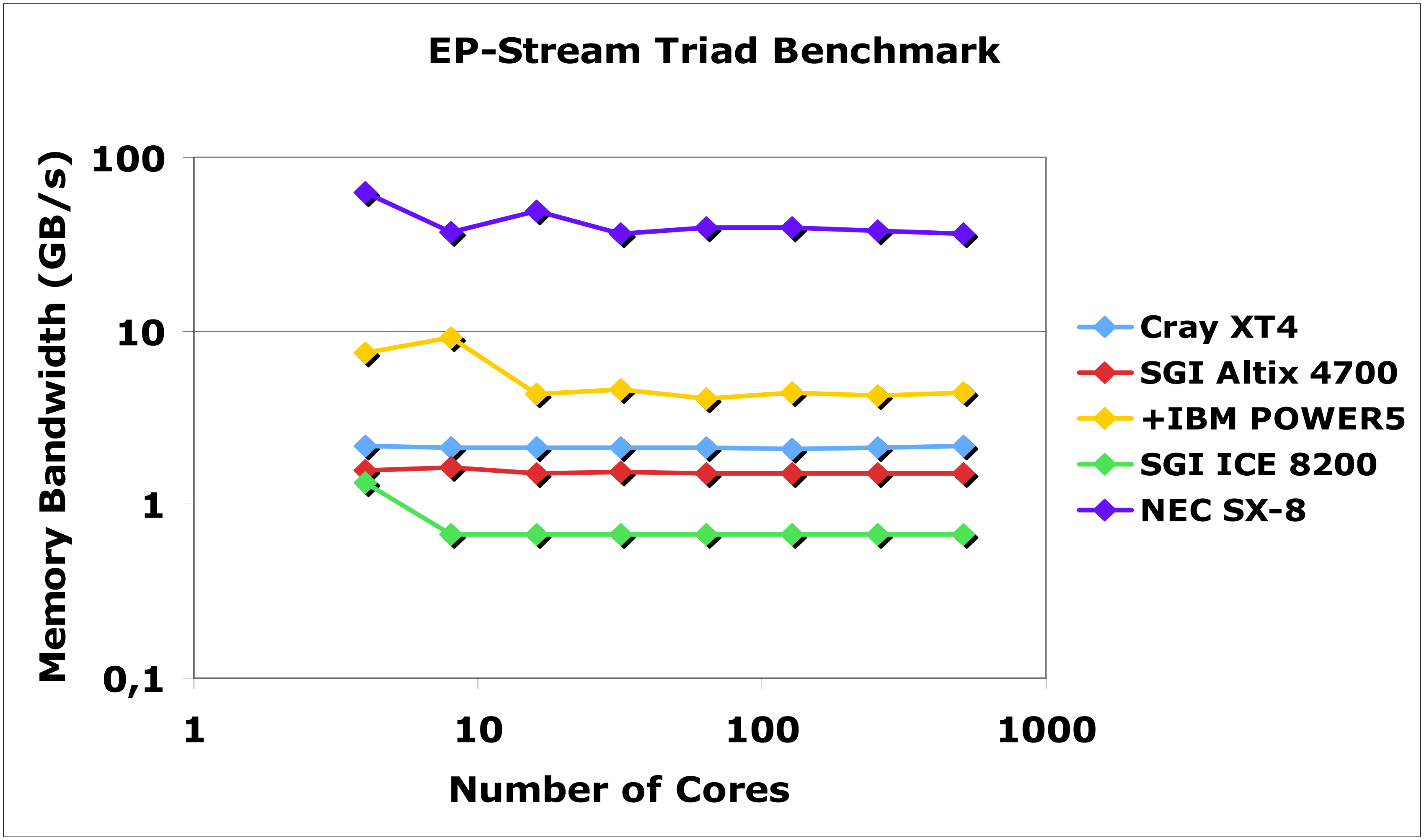
HLRS

NAS

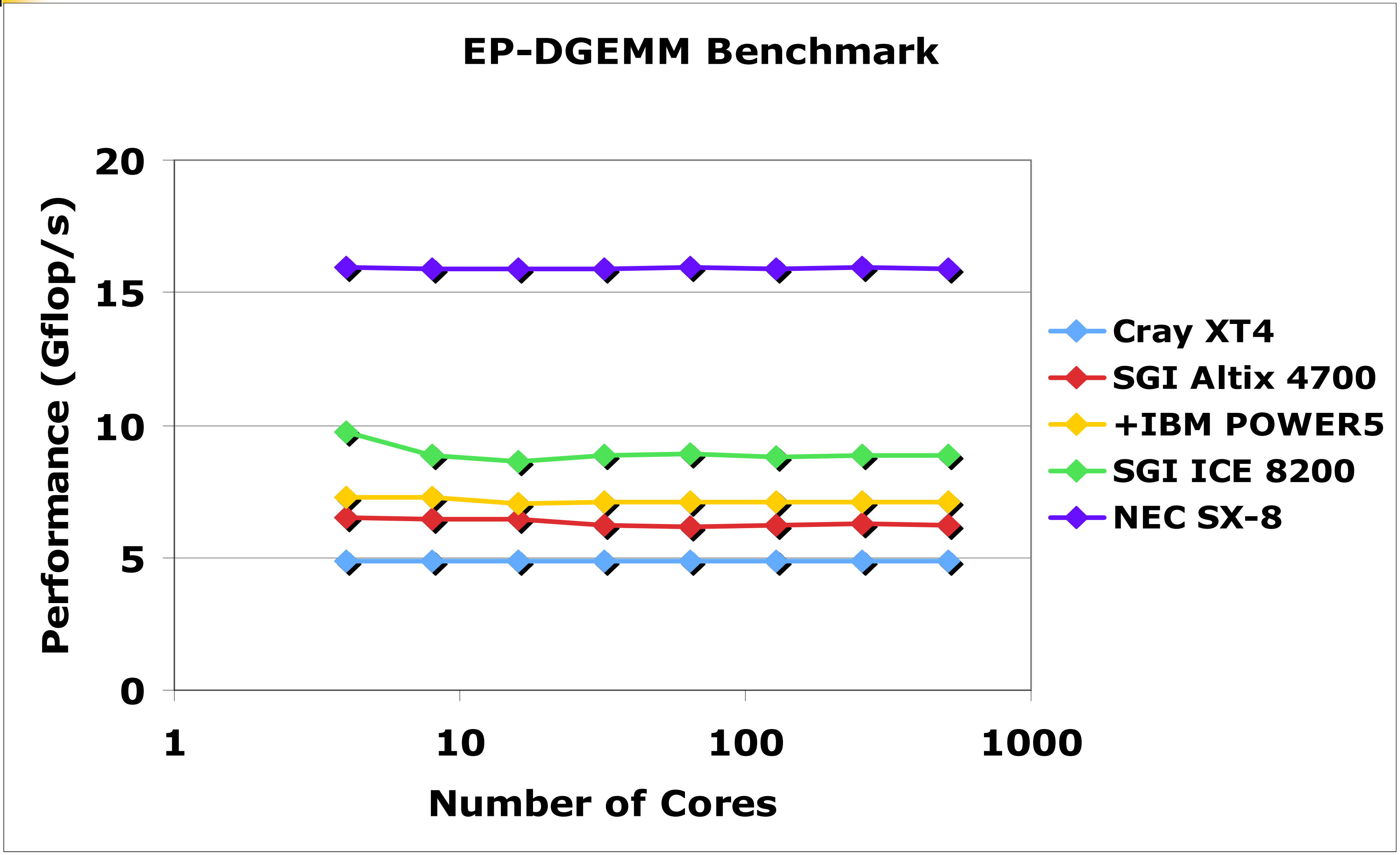
NPB and HPCC Implementations on NEC SX-8

- MPI version of NPB are written/optimized for cache based systems
  - Computational intensive benchmarks like BT, LU, FT and CG are not suitable for vector systems such as NEC SX-8 and Cray X1
  - NPB benchmarks were altered to run on NEC SX-8 making inner loops longer for appropriate vector lengths.
  - For SX-8, LU was run with SX-8 specific compiler directives for vectorization.
- HPCC 1.0 version is written/optimized for cache based systems
  - Cache based MPI FFT benchmark is not suitable for vector systems such as NEC SX-8 and Cray X1

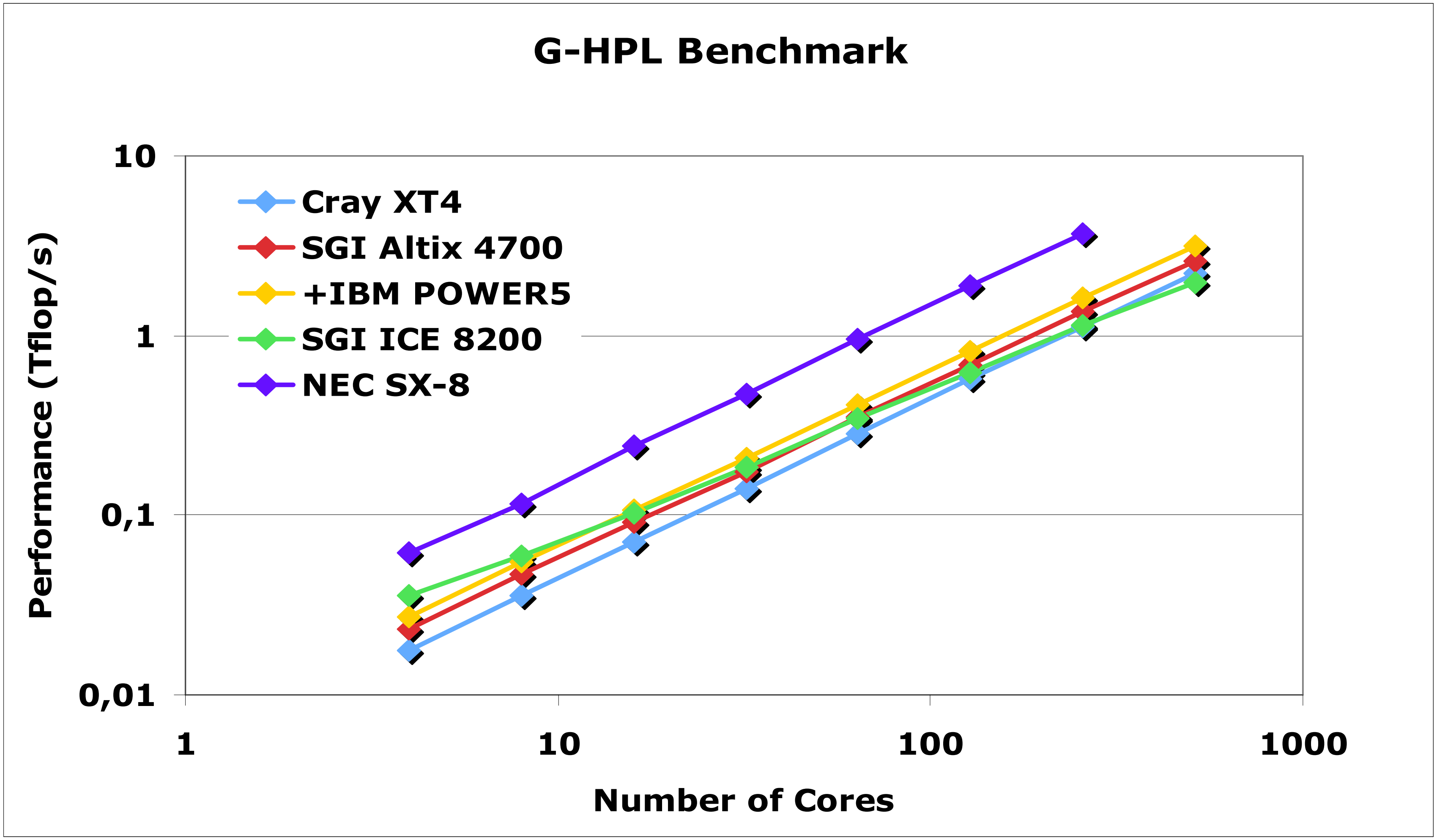
HPCC EP-Stream Benchmark



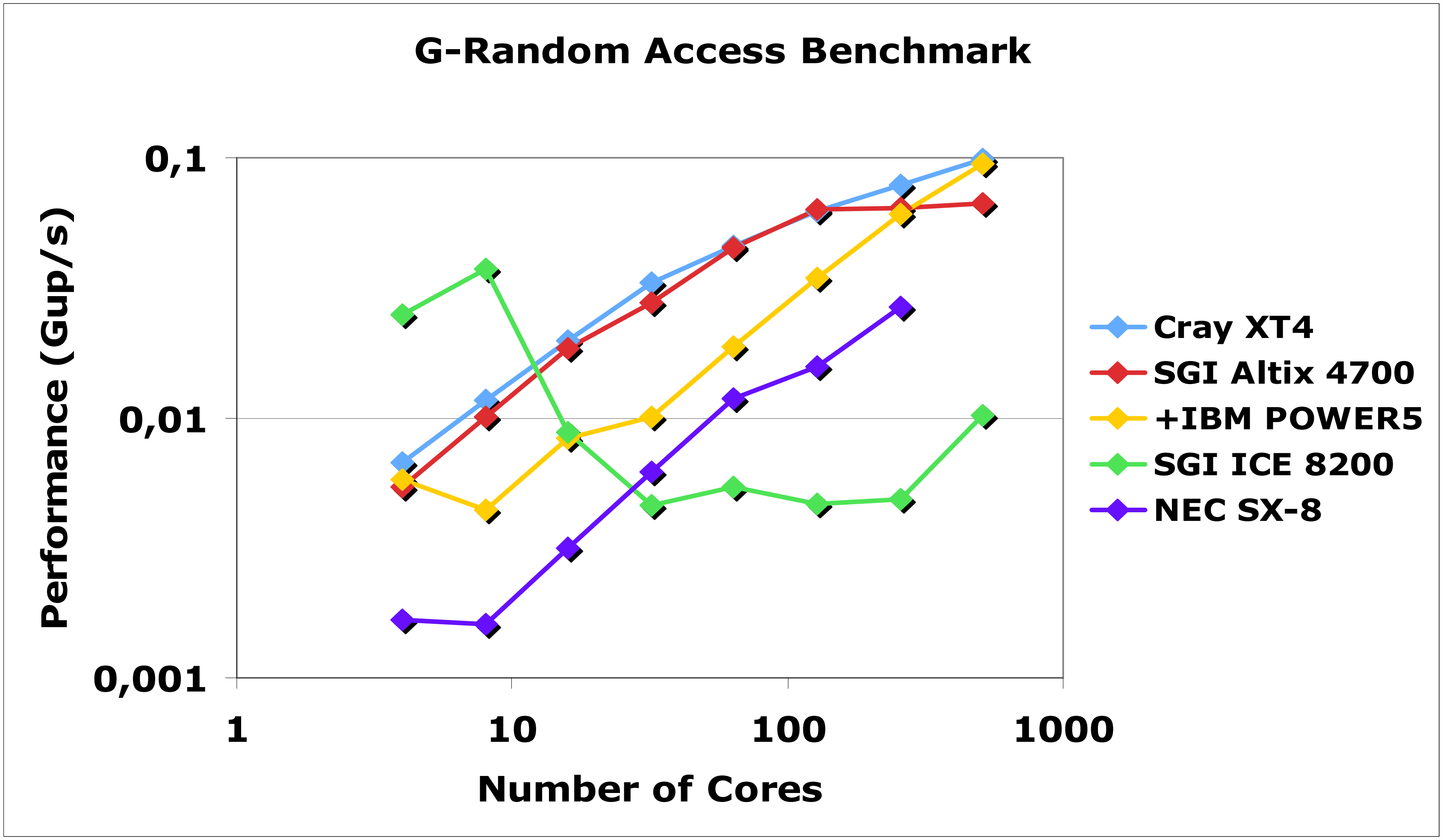
HPCC: EP-DGEMM Benchmark



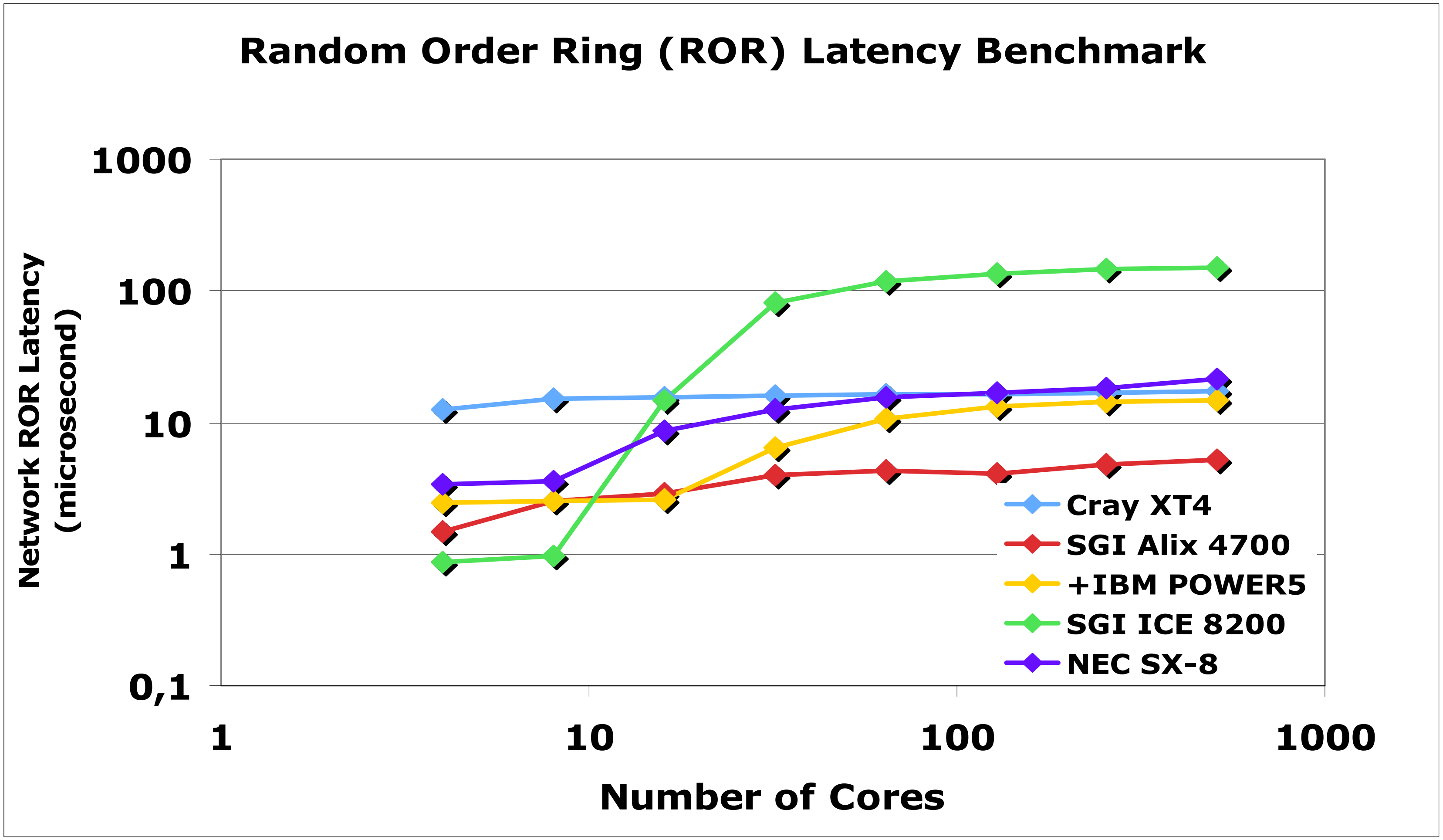
HPCC: G-HPL Benchmark



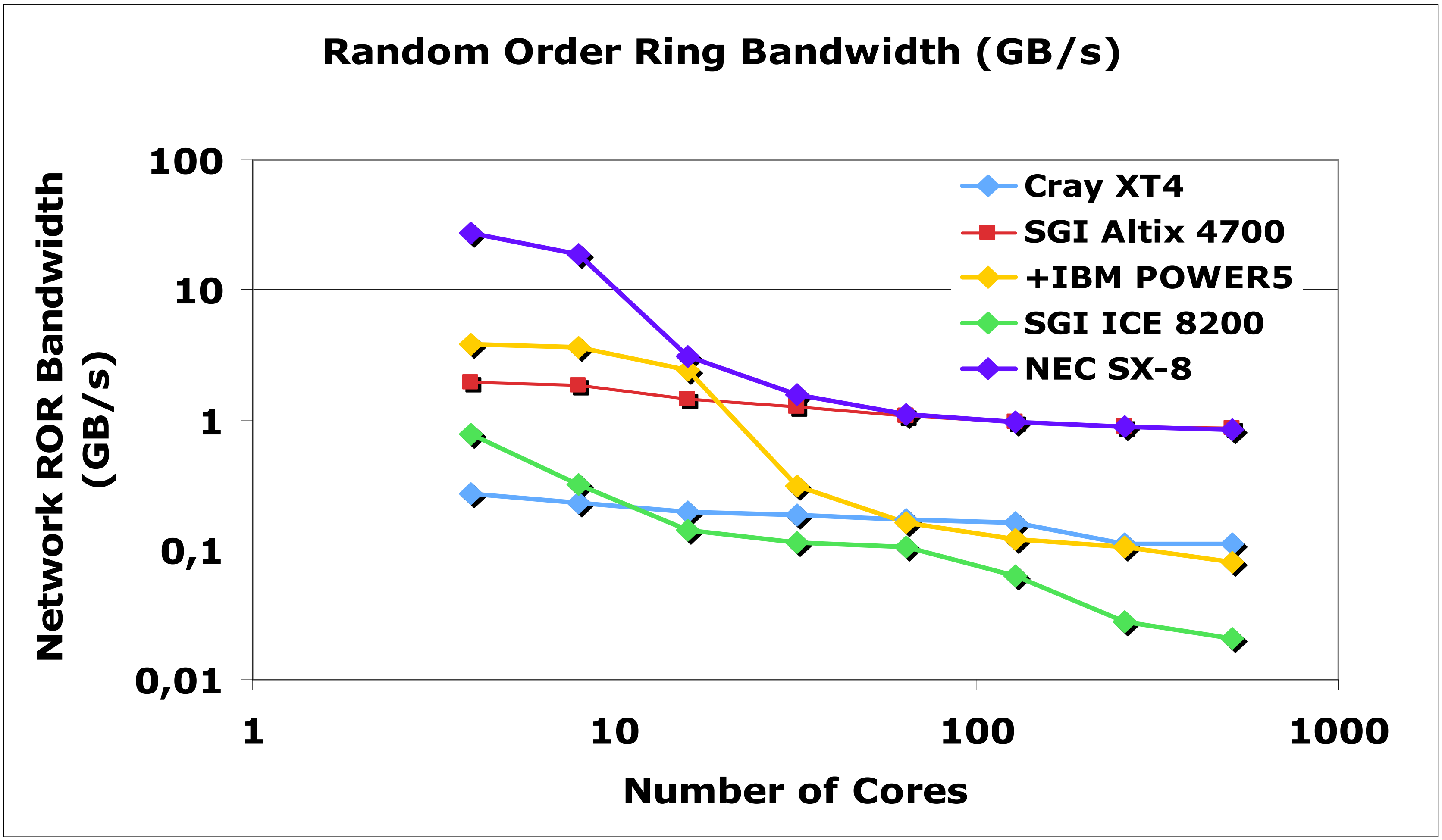
HPCC: Random Memory Access Benchmark



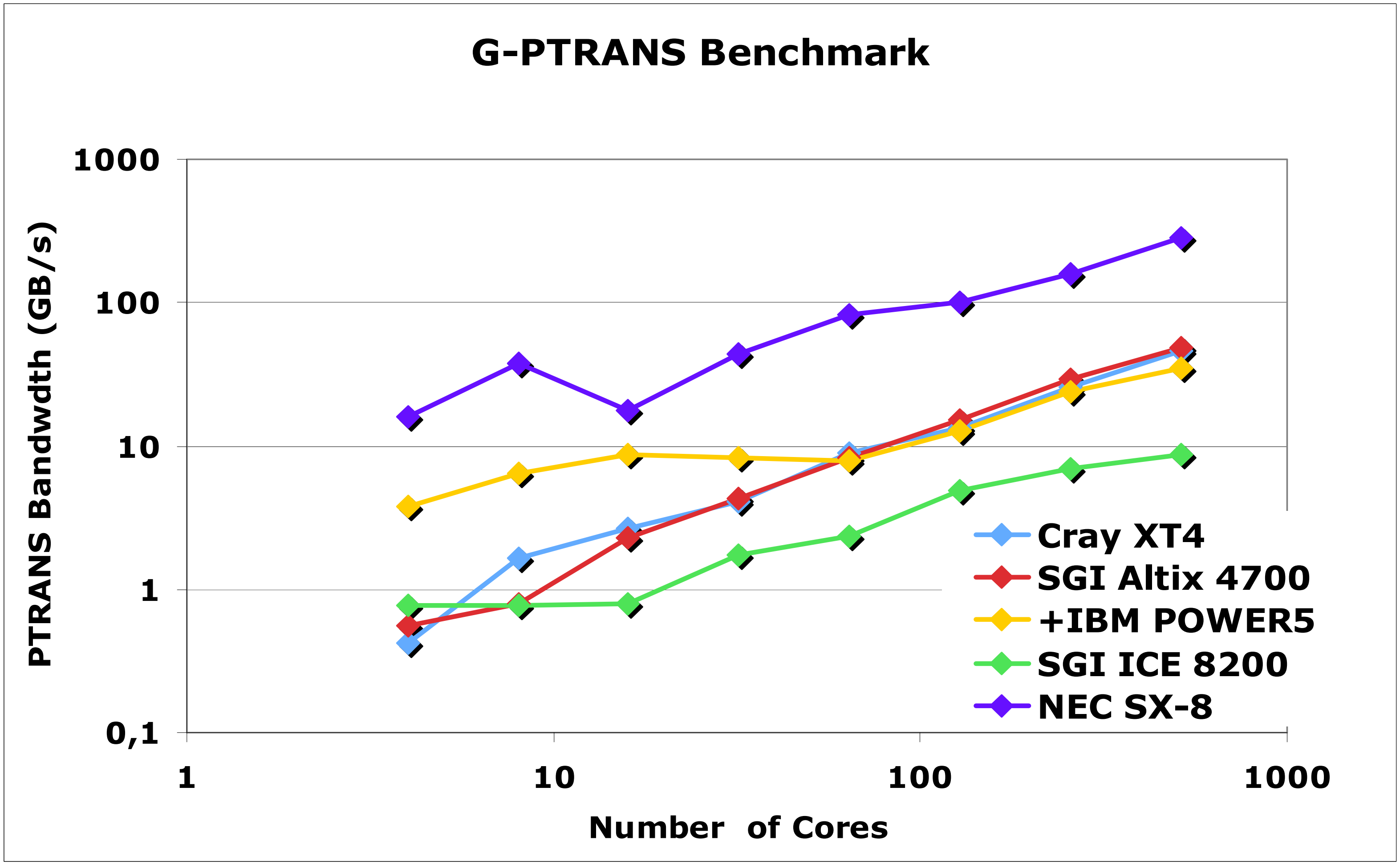
HPCC: Random Order Ring Latency Benchmark



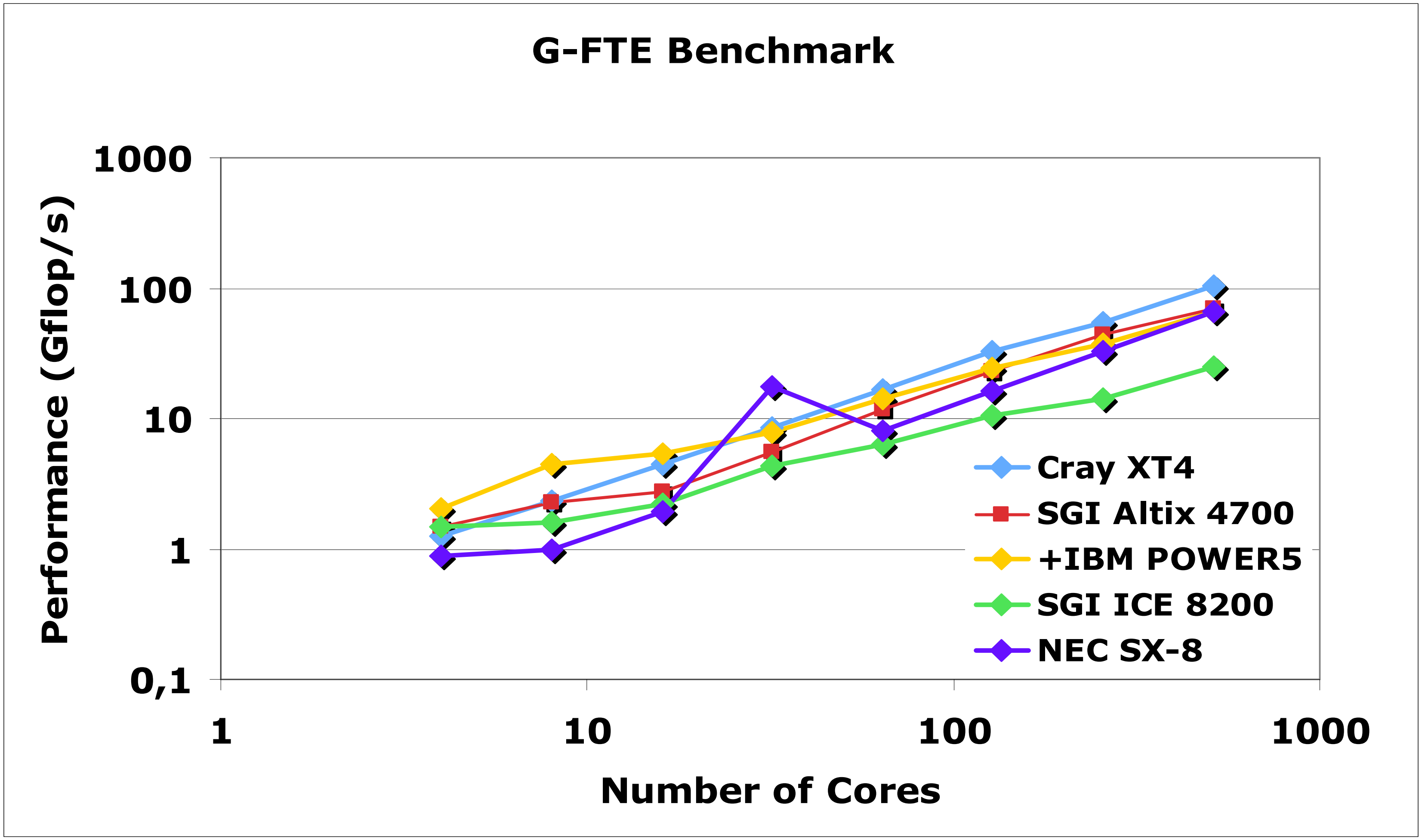
HPCC: Random Order Ring Bandwidth



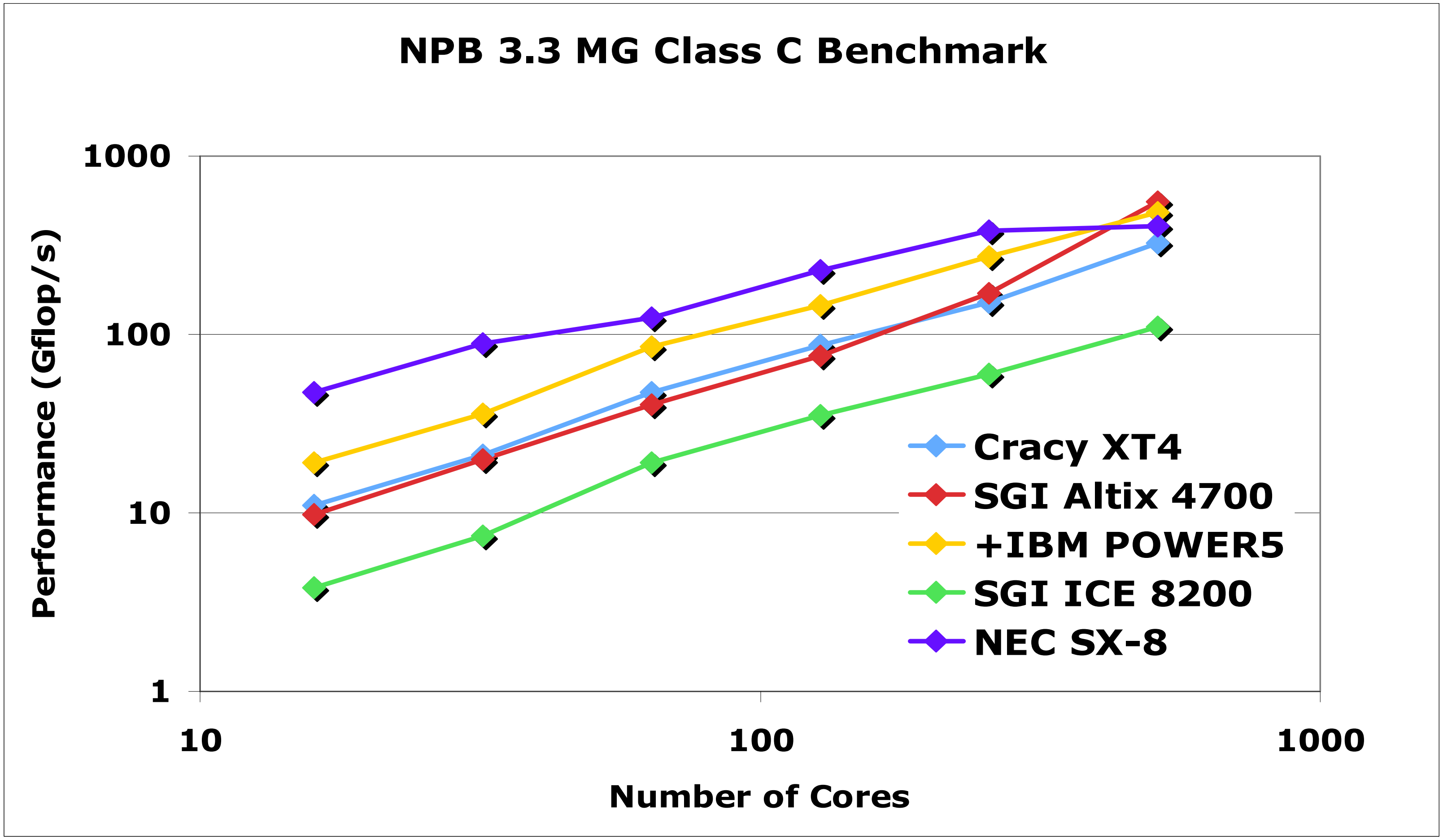
HPCC: PTRANS Benchmark



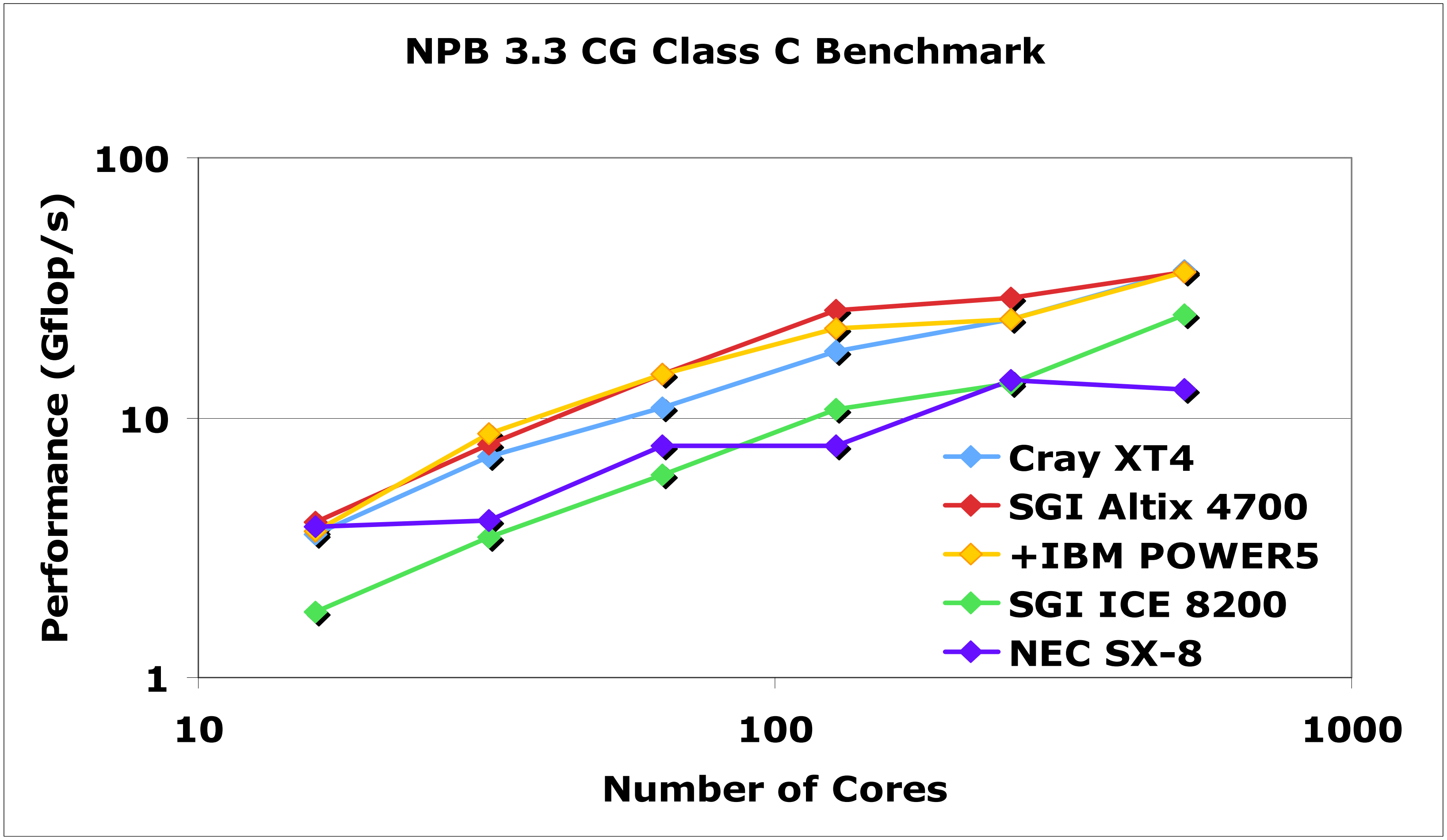
HPCC: FFTE Benchmark



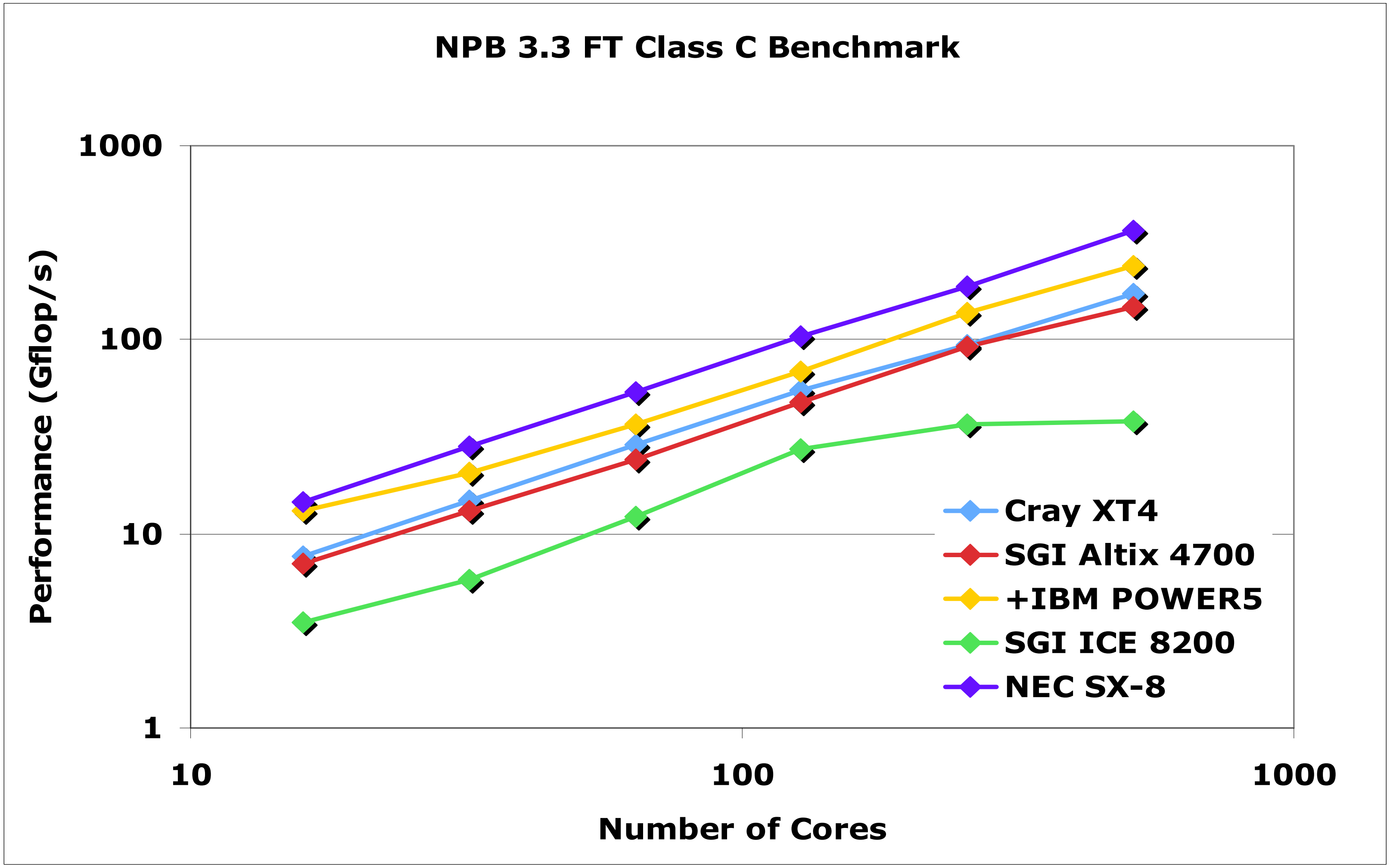
NPB MG Class C Benchmark



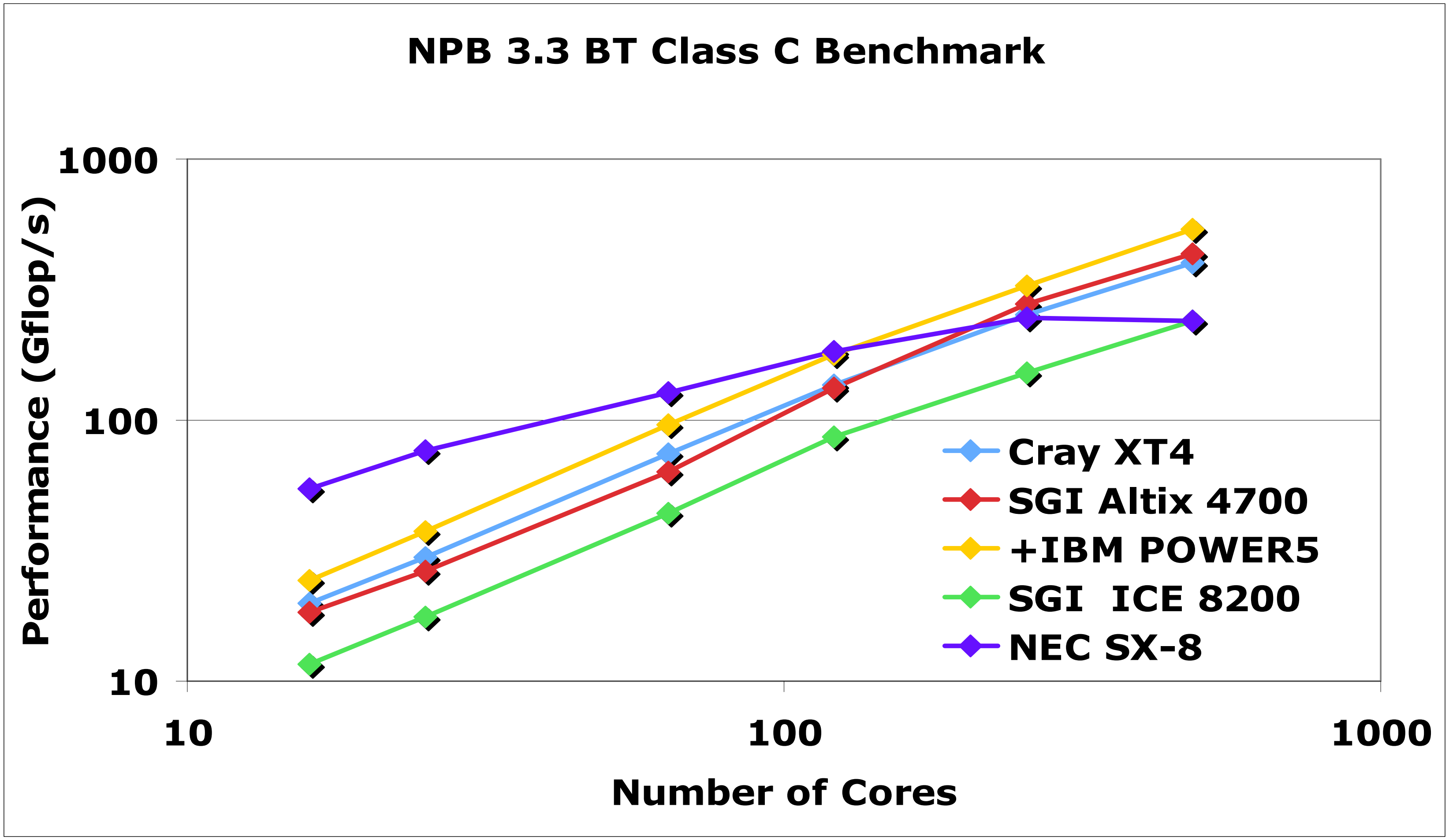
NPB CG Class C Benchmark



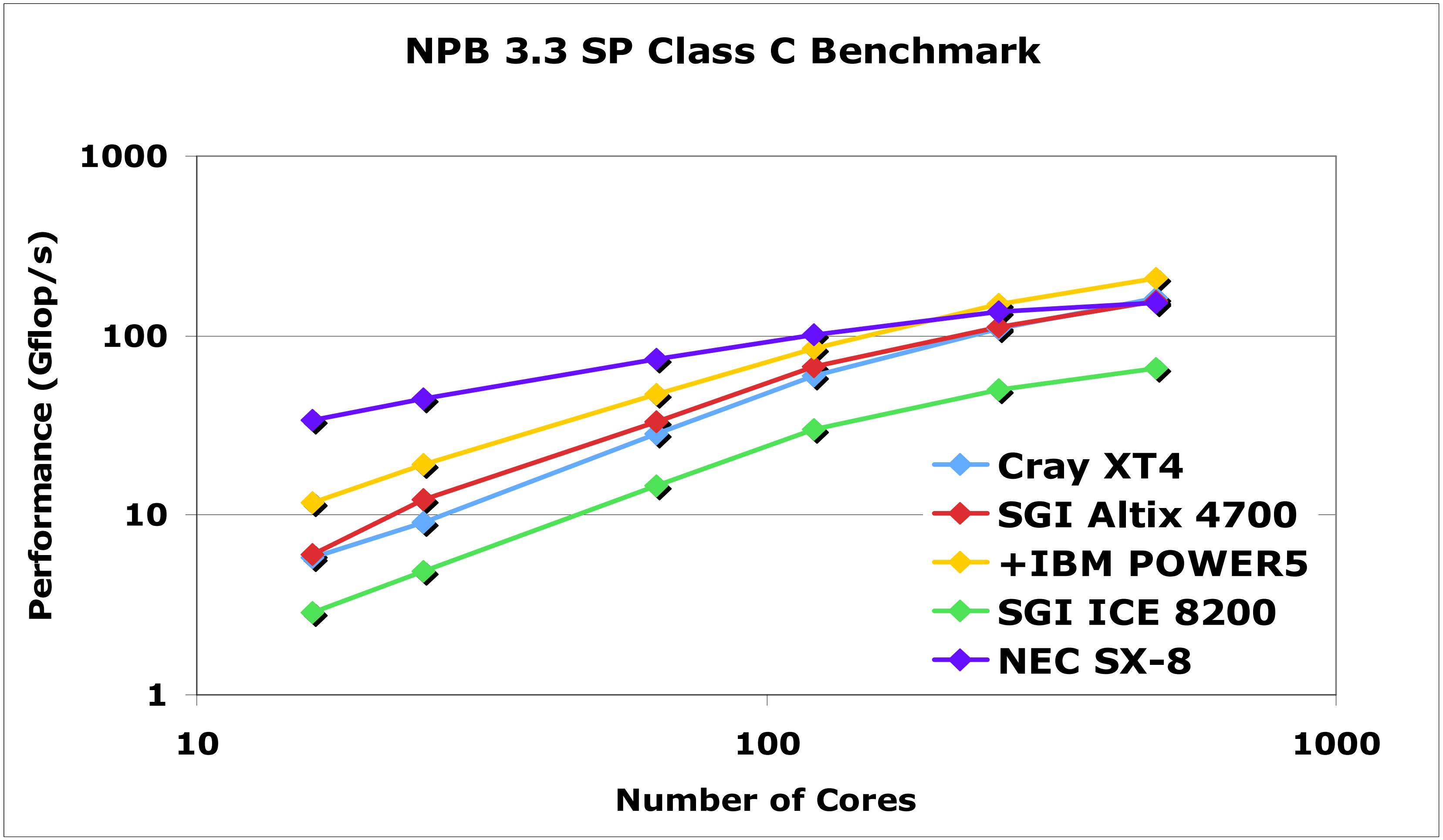
NPB FT Class C Benchmark



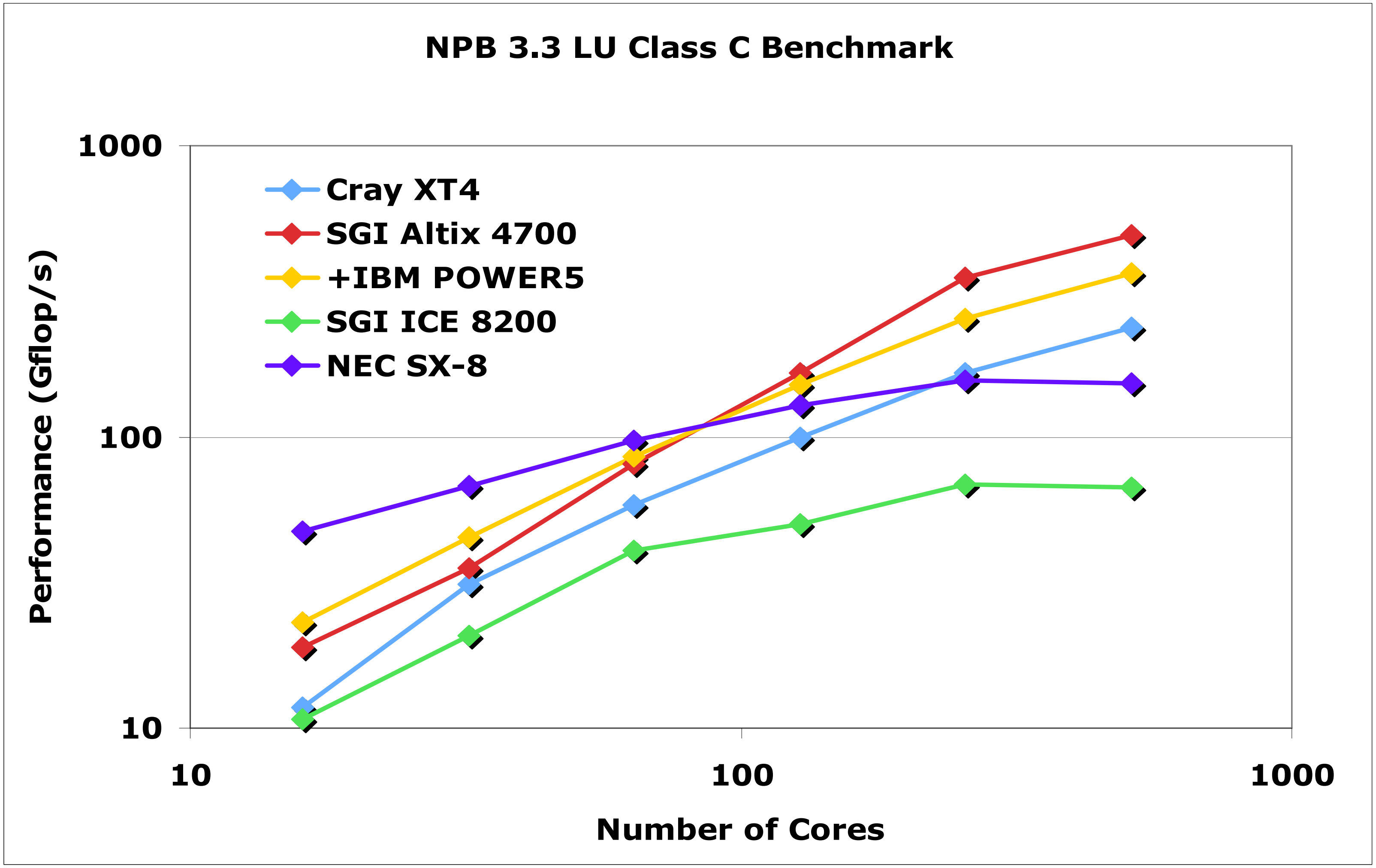
NPB BT Class C Benchmark



NPB SP Class C Benchmark



NPB LU Class C Benchmark



Summary

- Stream memory BW is highest for vector system NEC SX-8.
  Among cached based systems it is highest for IBM POWER5+ and lowest for SGI ICE 8200
- Floating point performance is highest for NEC SX-8. Among cached based systems it is highest for SGI ICE 8200 and lowest for Cray XT4
- Network random order latency is lowest for SGI Altix 4700 (NL4) and highest for ICE 8200 (IB). However, for Cray XT4 it is almost constant from 4-512 cpus
- Network random order bandwidth is highest for NEC SX-8 and lowest for SGI ICE 8200 (IB).
- Performance of PTRANS is highest for NEC SX-8 and lowest for SGI ICE 8200 (IB).

Summary

- Performance of HPCC-FFT is highest for Cray XT4 and lowest for SGI ICE 8200 (IB)
- Performance of MG is highest for NEC SX-8 and lowest for SGI ICE 8200 (IB)
- Performance of CG is
highest for IBM POWER5+ & SGI Altix 4700 and lowest for SGI ICE 8200 (IB) & NEC SX-8
- Performance of NPB FT is highest for
NEC SX-8 and lowest for SGI ICE 8200 (IB)
- Performance of NPB BT and SP is highest for NEC SX-8 and lowest for SGI ICE 8200 (IB)