CRAY® THE SUPERCOMPUTER COMPANY

Performance Lessons from the Cray XT3

Jeff Larkin

Cray, Inc.

larkin@cray.com

2/13/2006

The 7th LCI International Conference on Clusters

Overview

- Designing a HPC system
- Cray XT3: About the Parts
  - AMD Opteron Processor
  - Cray SeaStar Network
  - Unicos/lc Kernel
- Benchmark Results
- Application Results

Designing a System

- Architectural decisions must be made very early
  - Processor
  - Network
  - OS
- Early decisions are often difficult/expensive to change
- Months or years may pass between initial design decisions and final results

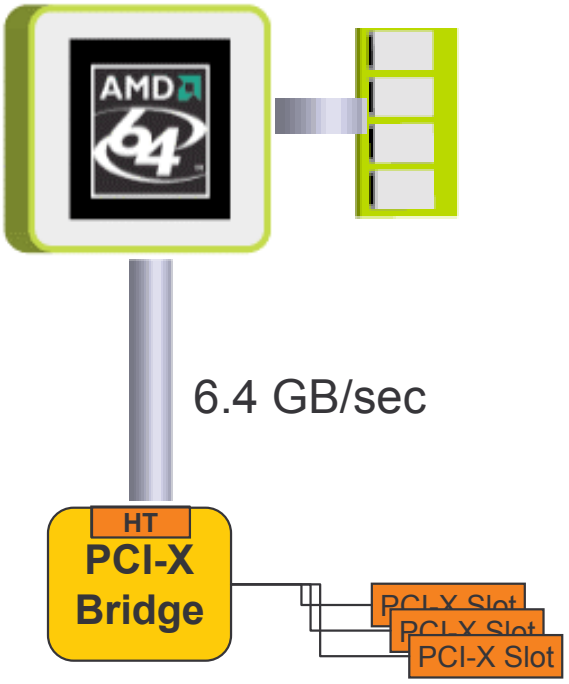
AMD Opteron: The Basics

- 64-bit x86 Architecture
- 128k Fully-associative L1 Cache
- 1MB 4-way Associative L2 Cache
- Integrated Memory Controller
  - No Northbridge
  - Low Memory Latency
- HyperTransport
  - High-bandwidth from the processor
  - Open Standard

AMD Opteron: The Basics

- 64-bit x86 Architecture
- 128k Fully-associative L1 Cache
- 1MB 4-way Associative L2 Cache
- Integrated Memory Controller
  - No Northbridge
  - Low Memory Latency
- HyperTransport
  - High-bandwidth from the processor
  - Open Standard

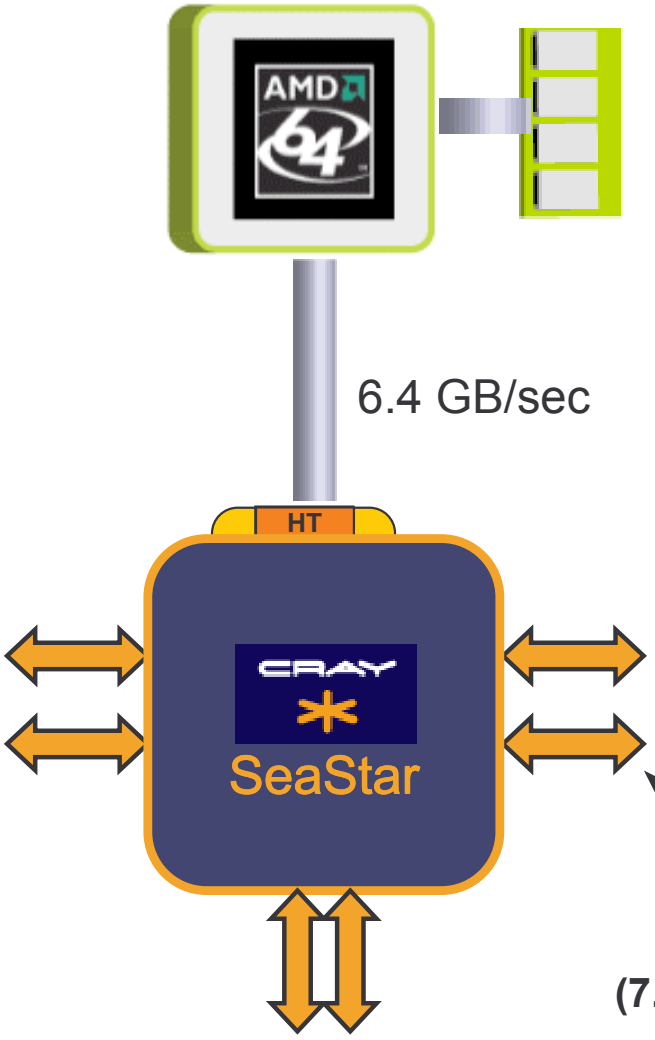
AMD Opteron: Generic System



- **DDR memory controller and function of Northbridge is pulled onto the Opteron die. Memory latency reduced to 60-90 ns**
- **No Northbridge chip results in savings in heat, power, complexity and an increase in performance**
- **Interface off the chip is an open standard (HyperTransport)**

AMD Opteron: Generic System

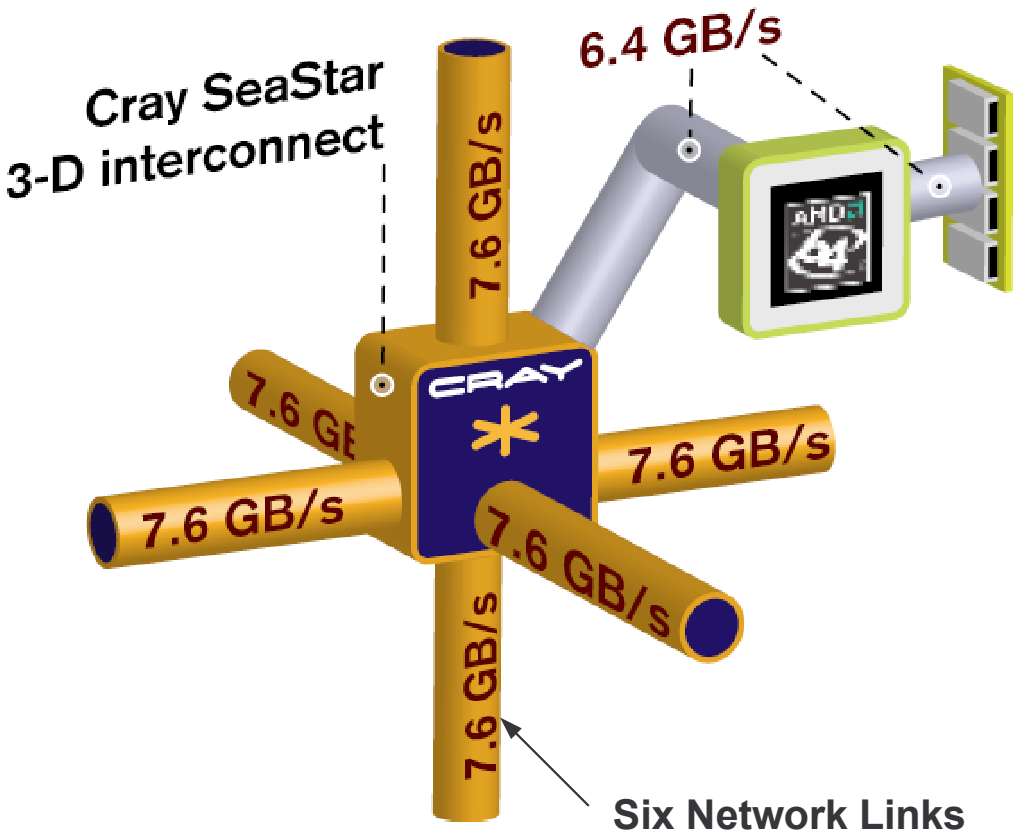
CRAY XT3 PE



- DDR memory controller and function of Northbridge is pulled onto the Opteron die. Memory latency reduced to 60-90 ns
- No Northbridge chip results in savings in heat, power, complexity and an increase in performance
- Interface off the chip is an open standard (HyperTransport)

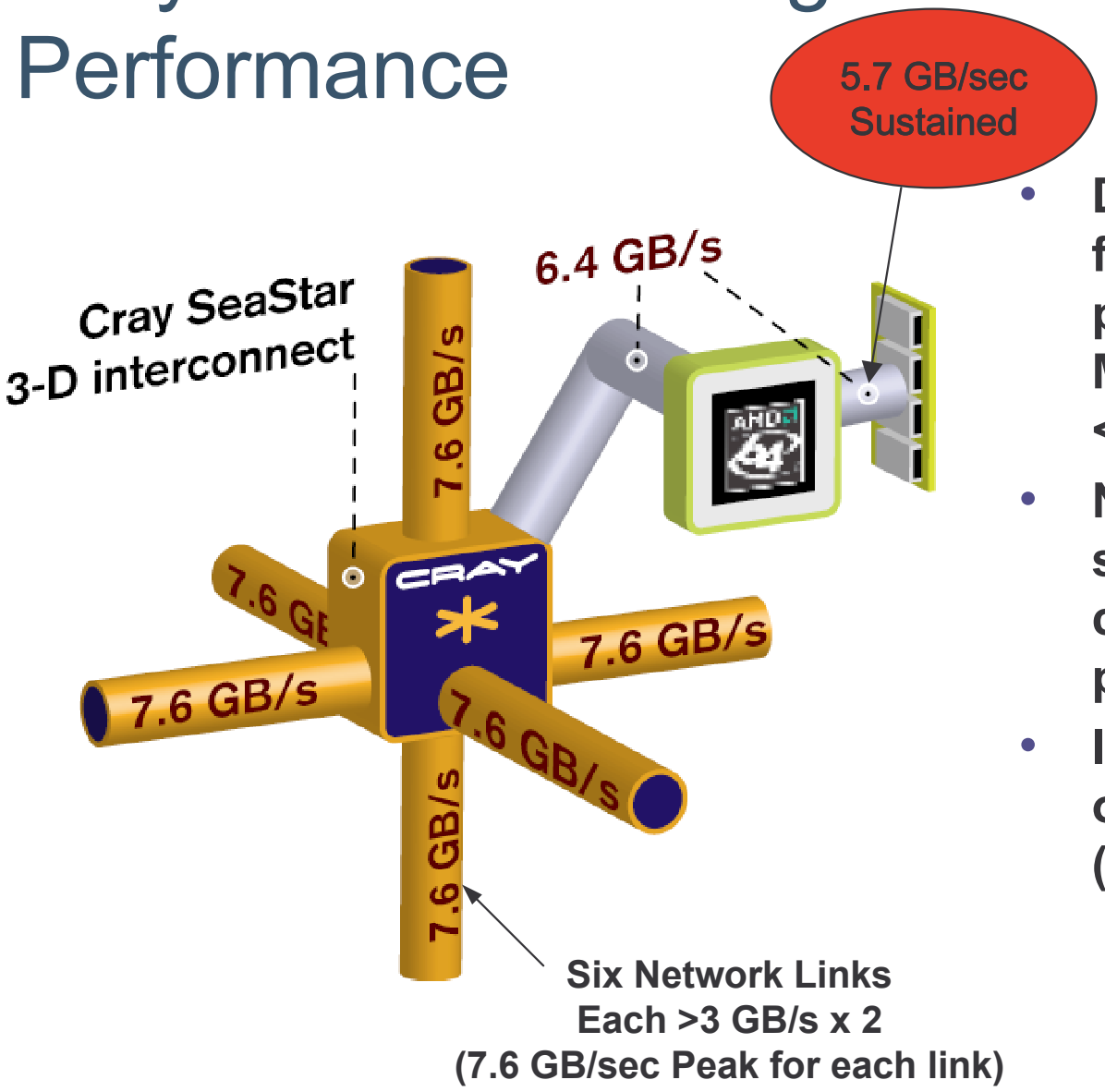
Six Network Links
Each >3 GB/s x 2
(7.6 GB/sec Peak for each link)

Cray XT3 Processing Element: Measured Performance



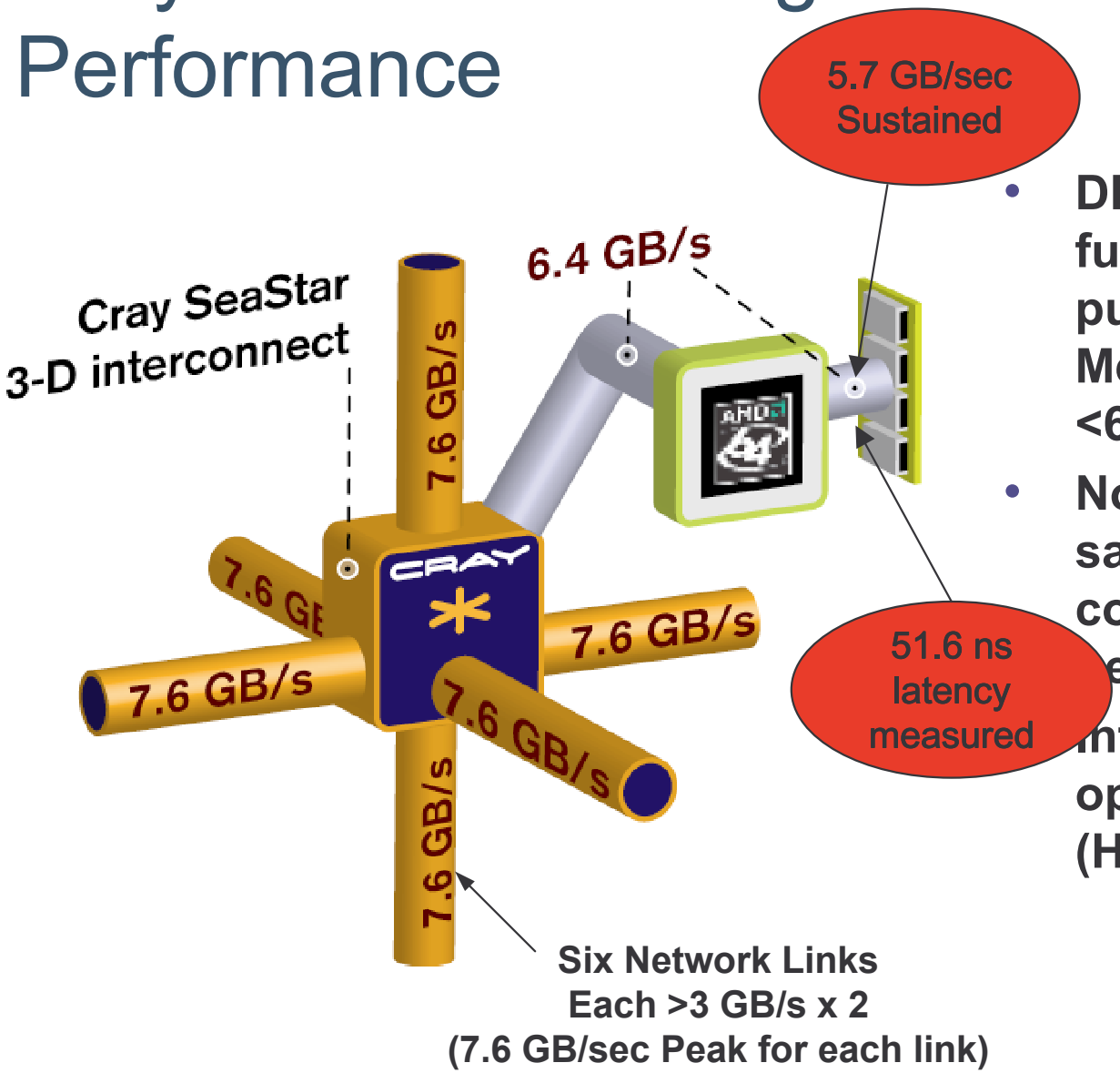
- DDR memory controller and function of Northbridge is pulled onto the Opteron die. Memory latency reduced to <60 ns
- No Northbridge chip results in savings in heat, power, complexity and an increase in performance
- Interface off the chip is an open standard (HyperTransport)

Cray XT3 Processing Element: Measured Performance



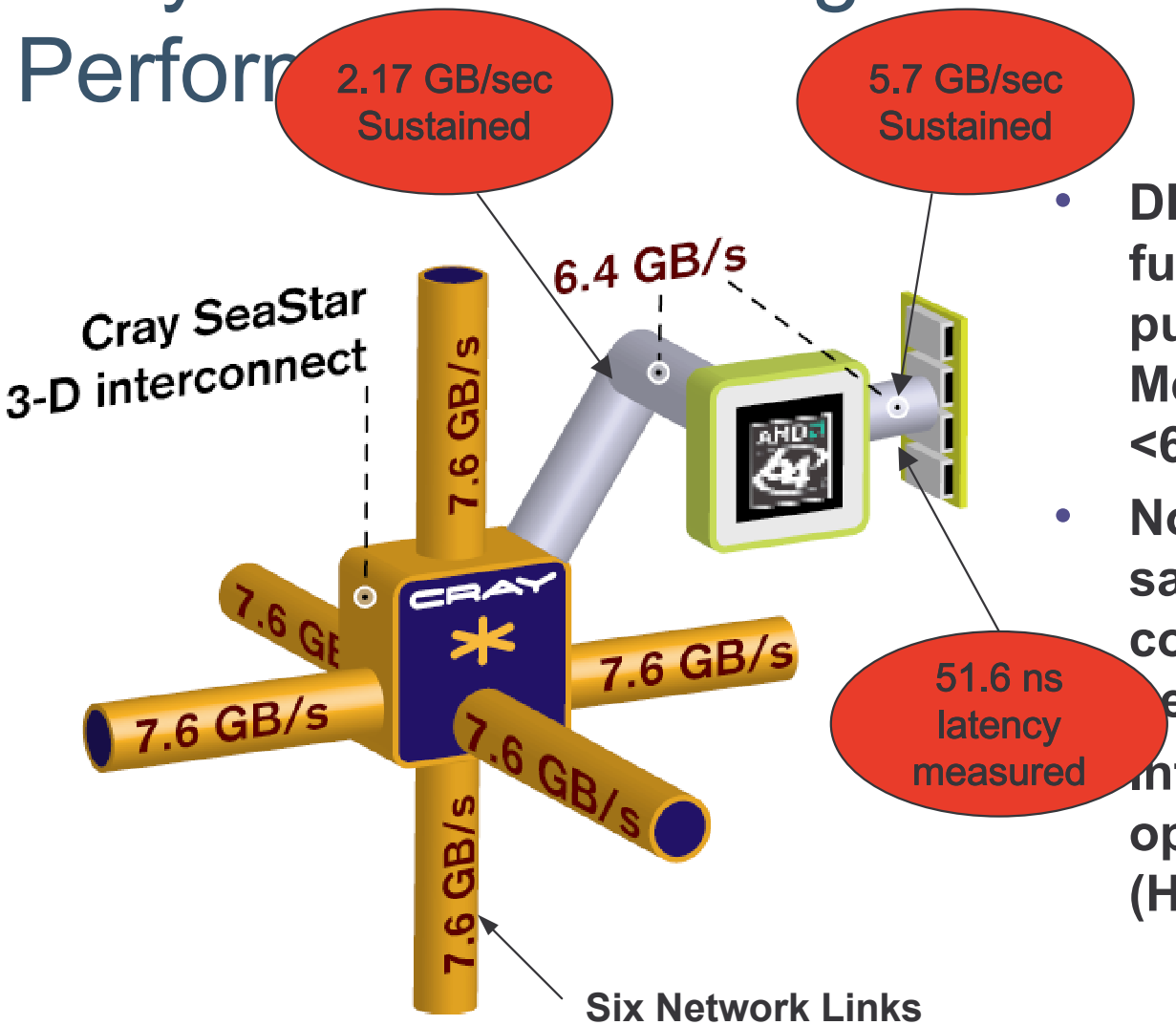
- DDR memory controller and function of Northbridge is pulled onto the Opteron die. Memory latency reduced to <60 ns
- No Northbridge chip results in savings in heat, power, complexity and an increase in performance
- Interface off the chip is an open standard (HyperTransport)

Cray XT3 Processing Element: Measured Performance



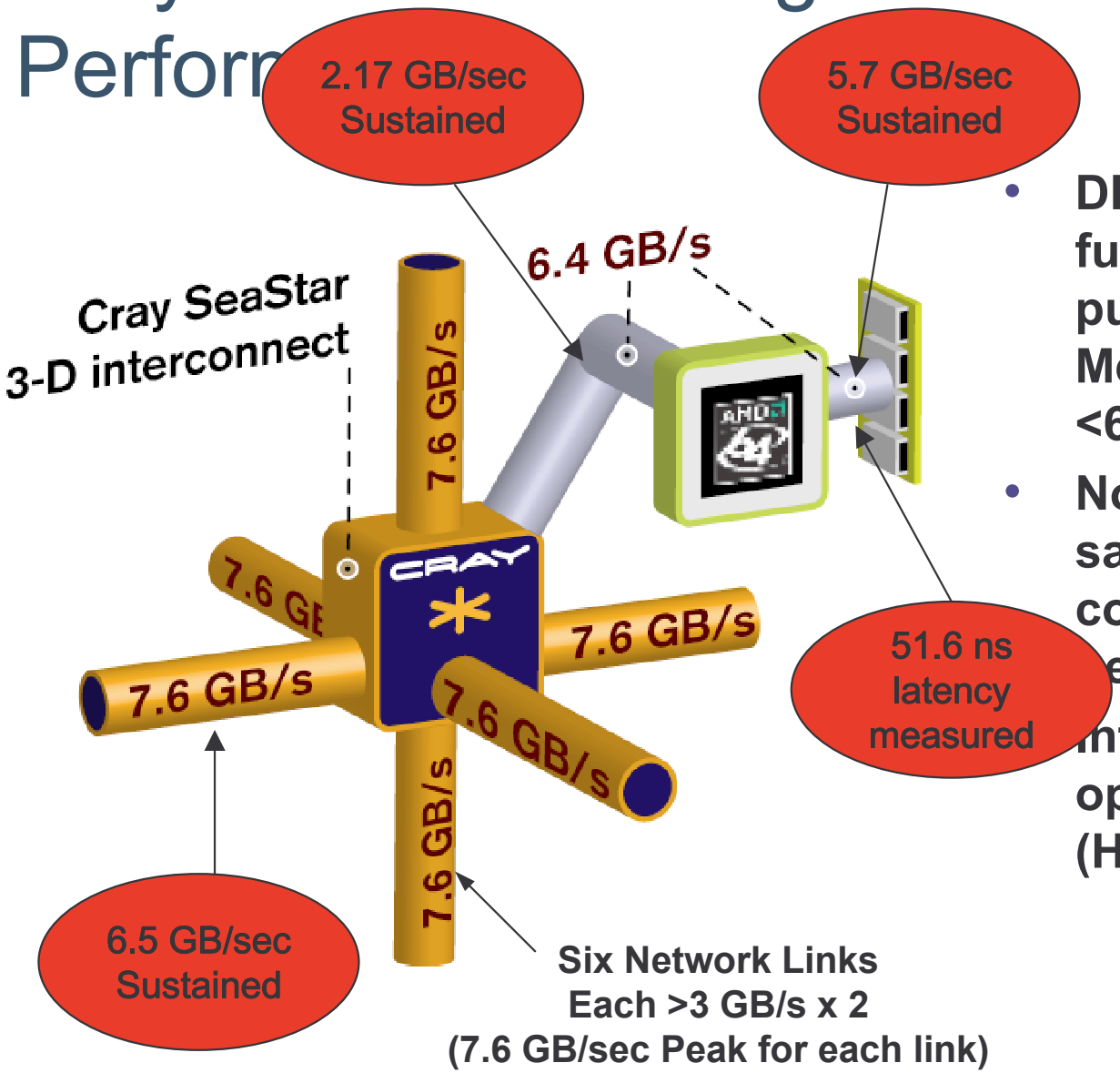
- DDR memory controller and function of Northbridge is pulled onto the Opteron die. Memory latency reduced to <60 ns
- No Northbridge chip results in savings in heat, power, complexity and an increase in performance
- Interface off the chip is an open standard (HyperTransport)

Cray XT3 Processing Element: Measured Perform



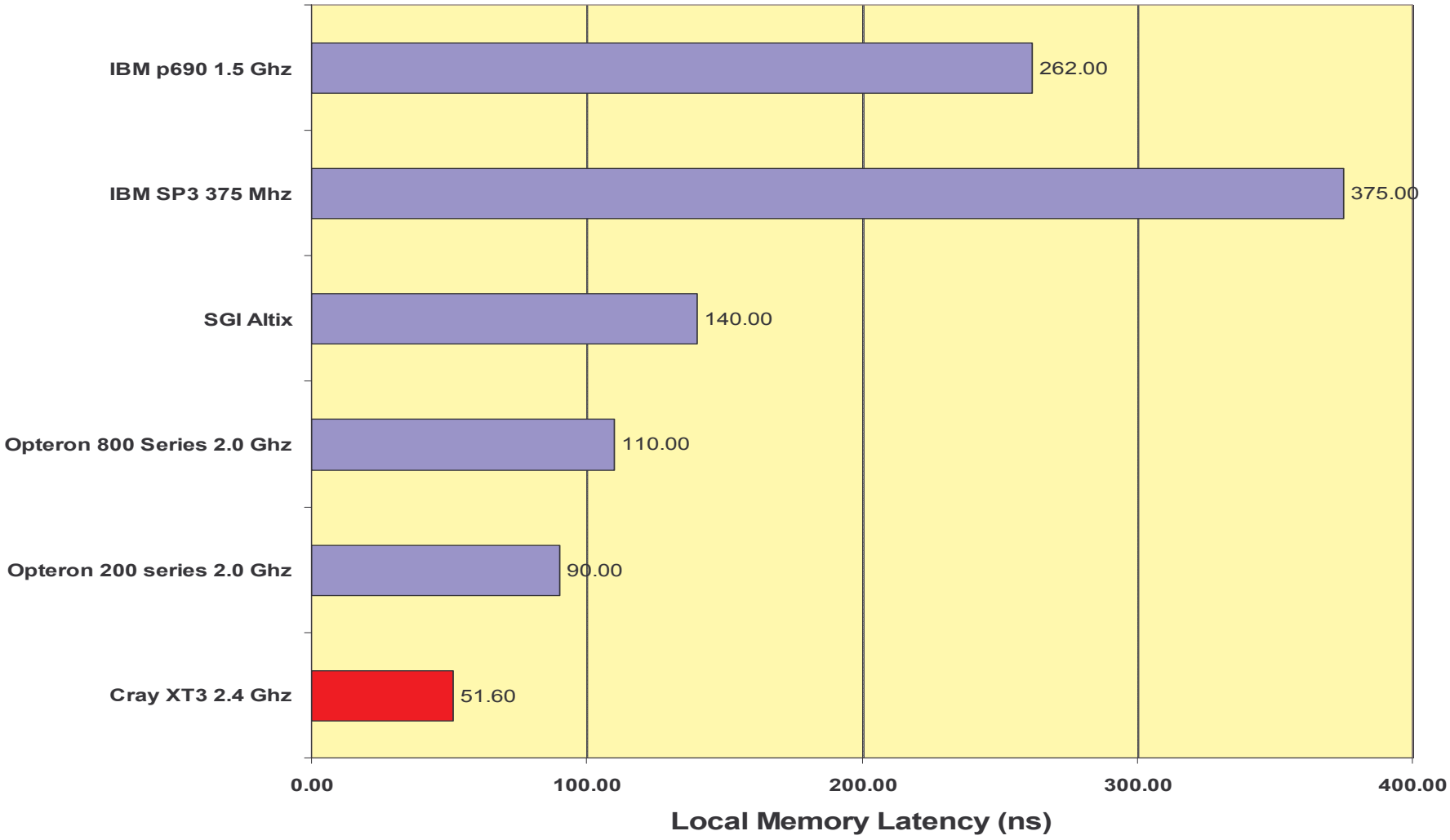
- DDR memory controller and function of Northbridge is pulled onto the Opteron die. Memory latency reduced to <60 ns
- No Northbridge chip results in savings in heat, power, complexity and an increase in performance
- Interface off the chip is an open standard (HyperTransport)

Cray XT3 Processing Element: Measured Perform



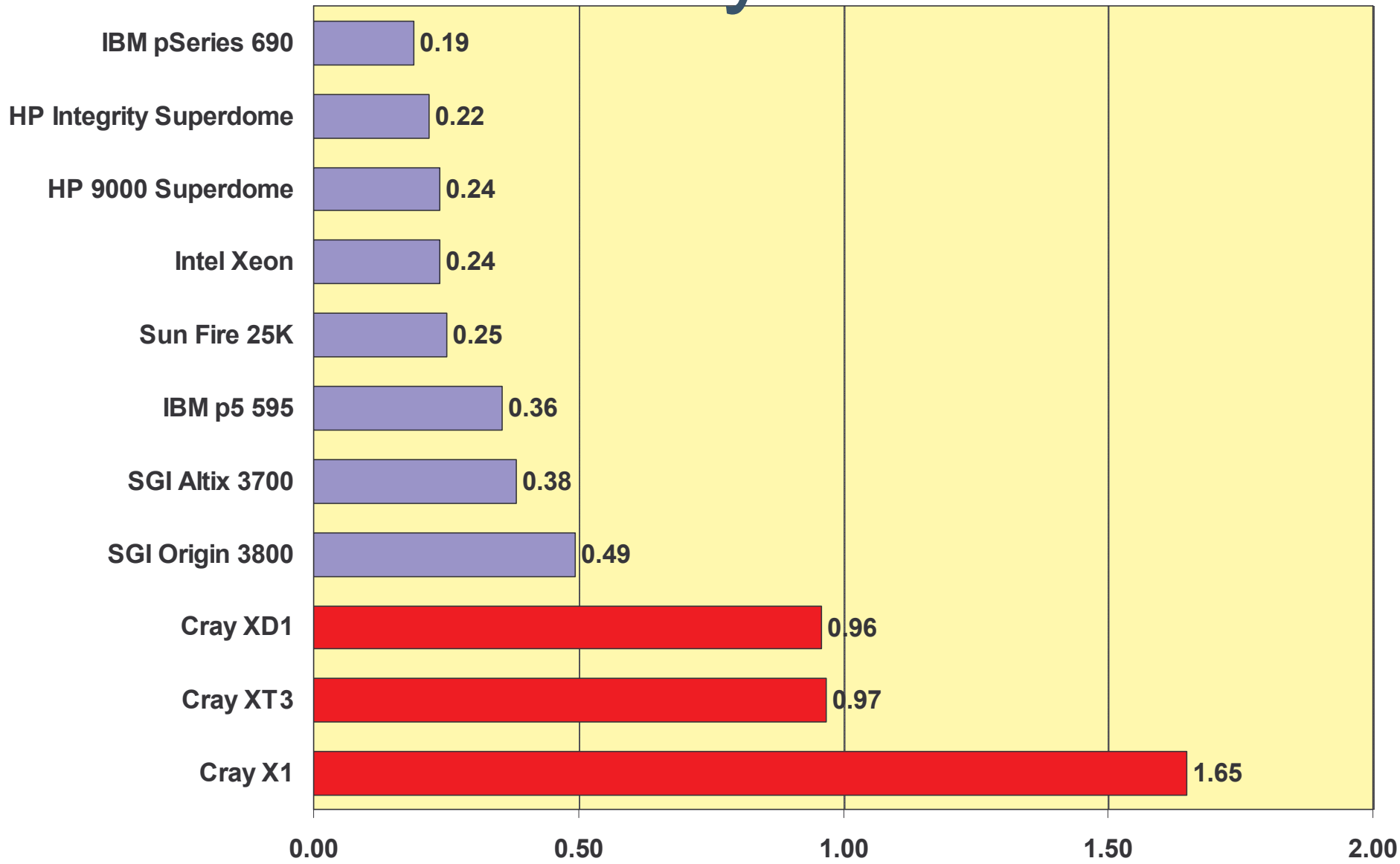
- DDR memory controller and function of Northbridge is pulled onto the Opteron die. Memory latency reduced to <60 ns
- No Northbridge chip results in savings in heat, power, complexity and an increase in performance
- Interface off the chip is an open standard (HyperTransport)

Memory Latency



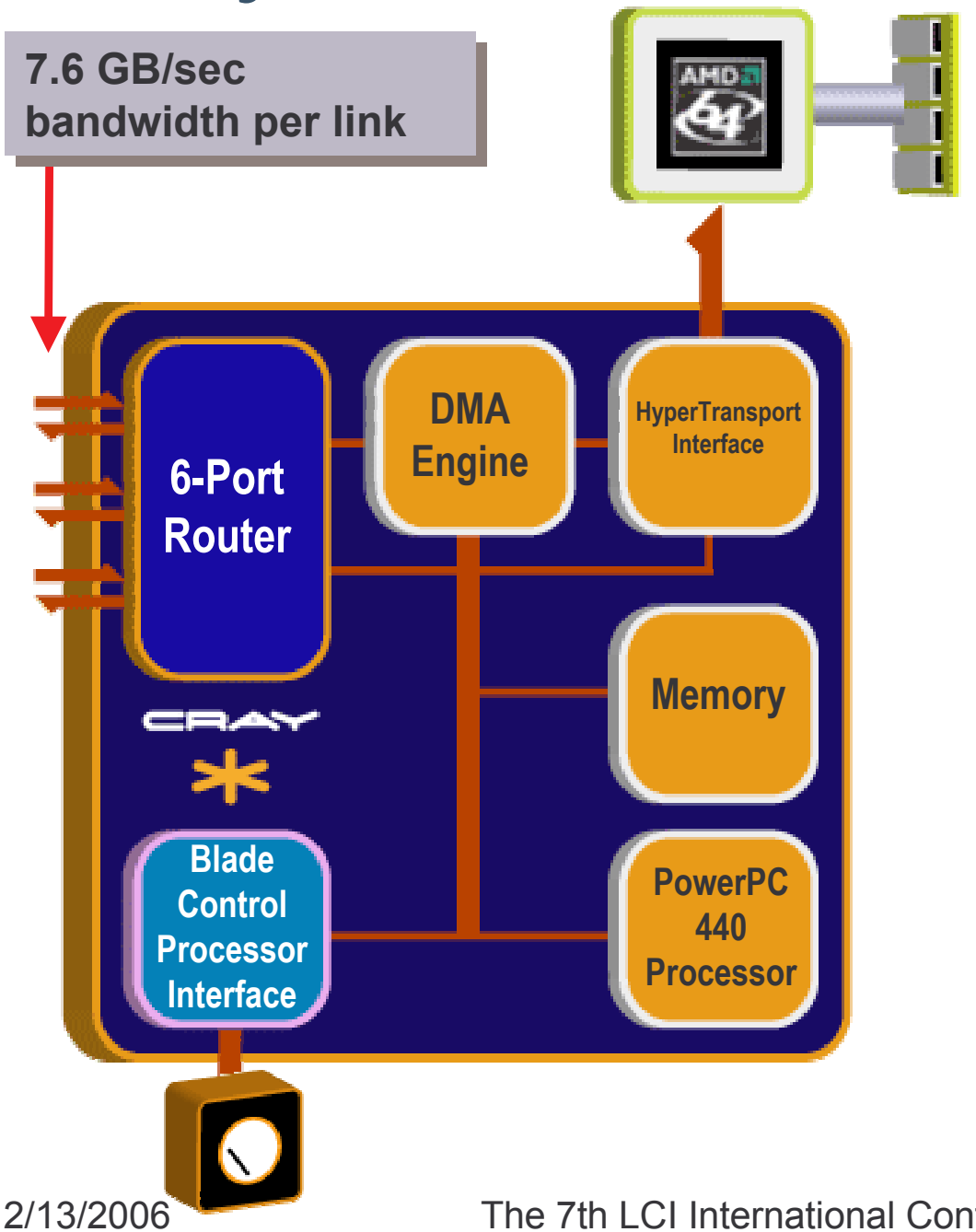
Single Processor architecture yields lowest memory latency

Measured Memory Balance



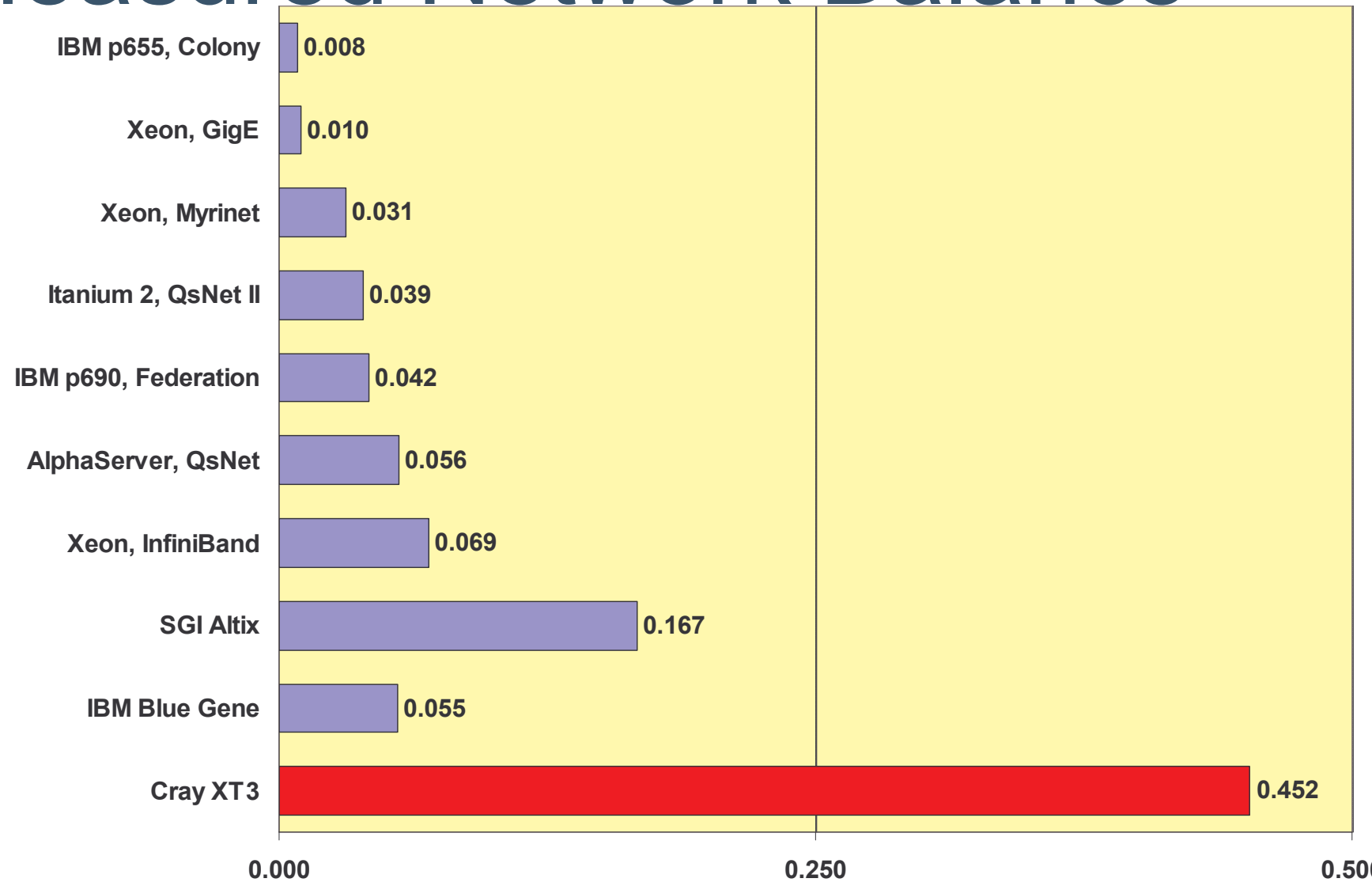
B/F calculated from memory bandwidth measured via STREAM Triad benchmark

Cray SeaStar Internals



- Each Processor is directly connected to a dedicated SeaStar
- Each SeaStar contains a 6-Port router *and* communications engine
- Provides serial connection to the Cray RAS and Management System

Measured Network Balance

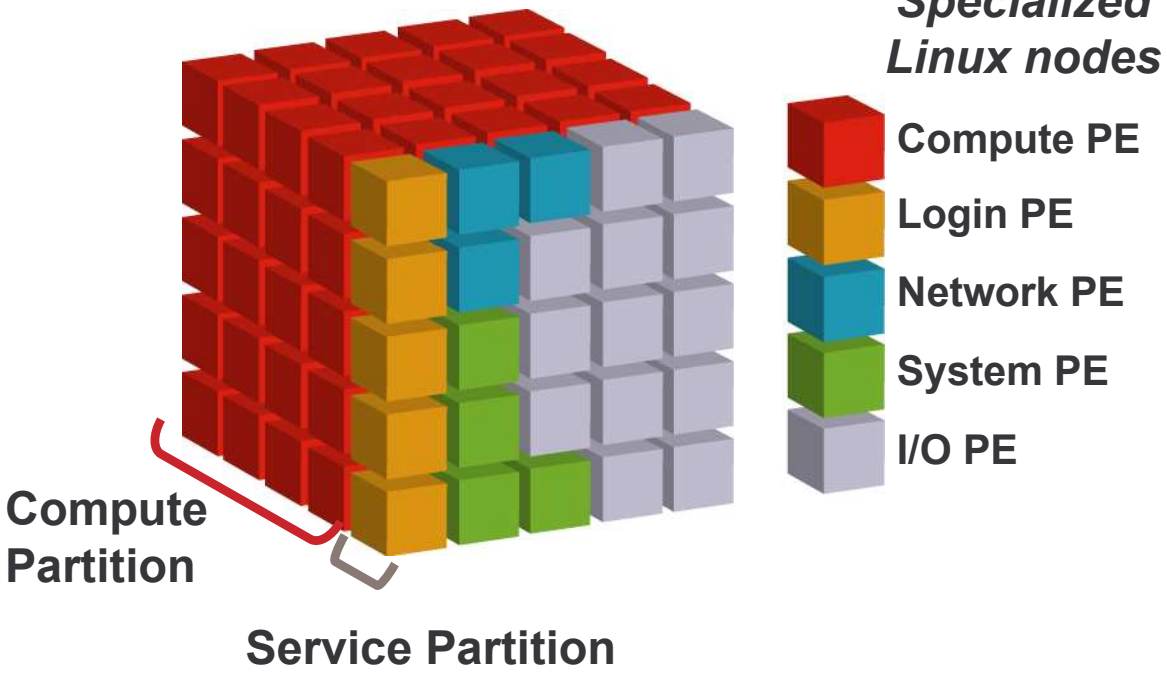


Network bandwidth is the maximum bidirectional data exchange rate between two nodes using MPI

What is OS Jitter?

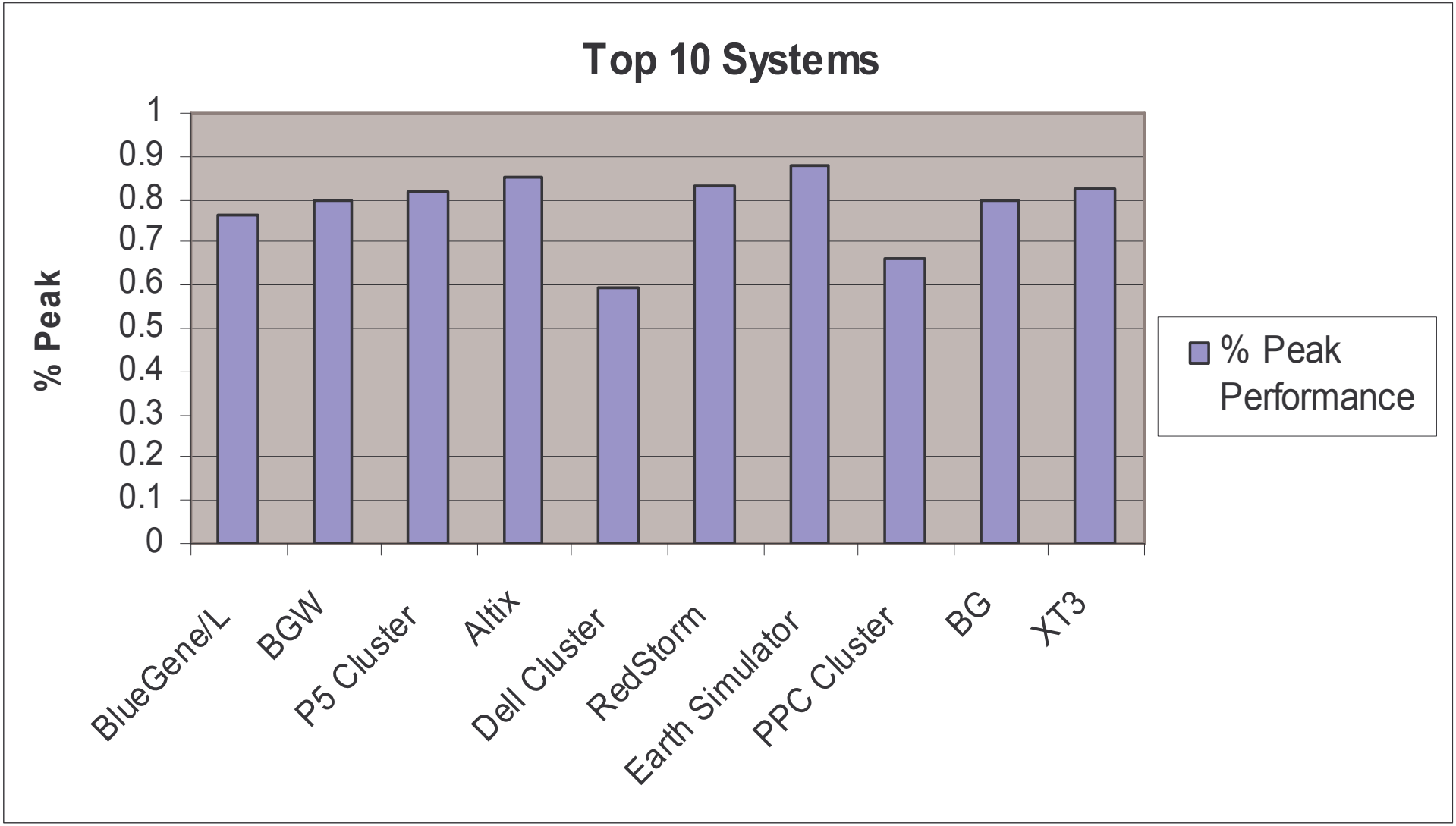
- In order to provide certain services, the OS must wake-up periodically
  - Network and I/O Requests
  - Daemon Processes
  - System Calls and Threads
- These wake up calls interrupt user time
- As more processors are added, more interruptions occur across the machine
- This white noise on the system is known as “OS Jitter” and is often the limiting factor for system scaling.

Scalable Software Architecture: UNICOS/lc



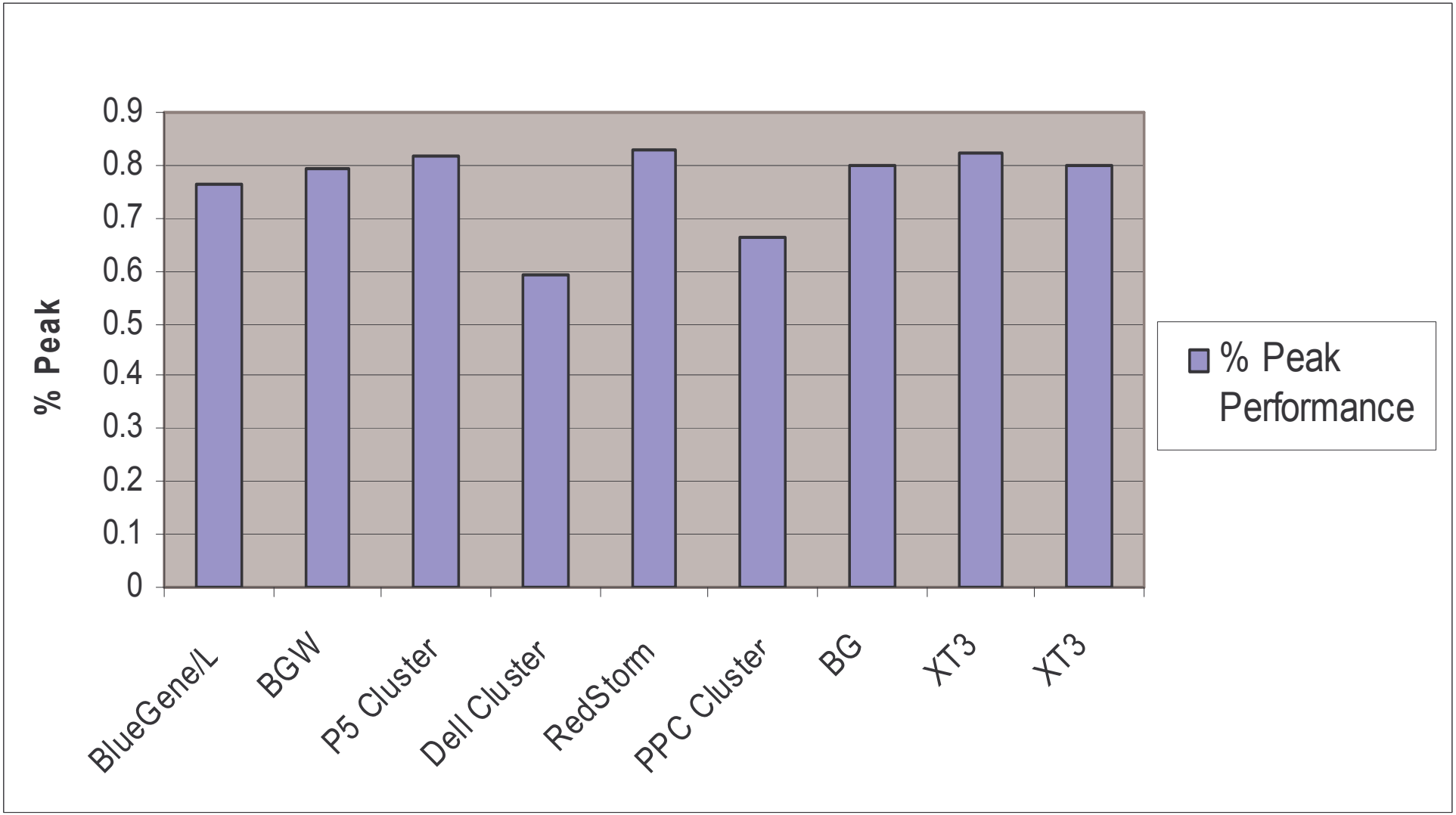
- Full featured Linux on Service PEs, Microkernel on Compute PEs.
- Service PEs specialize by function
- Contiguous memory layout used on compute processors to streamline communications
- Software Architecture eliminates OS "Jitter"
  - 100 ms interrupt times
  - Will be synchronized if required
  - OS heartbeat checked once per second.
- Software Architecture enables reproducible run times

Top500 Linpack Scaling



Data from 11/2005 Top500 list (top500.org)

Top500 Linpack Scaling



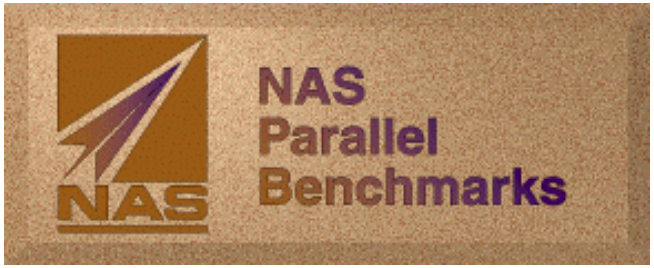
Data from 11/2005 Top500 list (top500.org)

XT3 Benchmark Results

2/13/2006

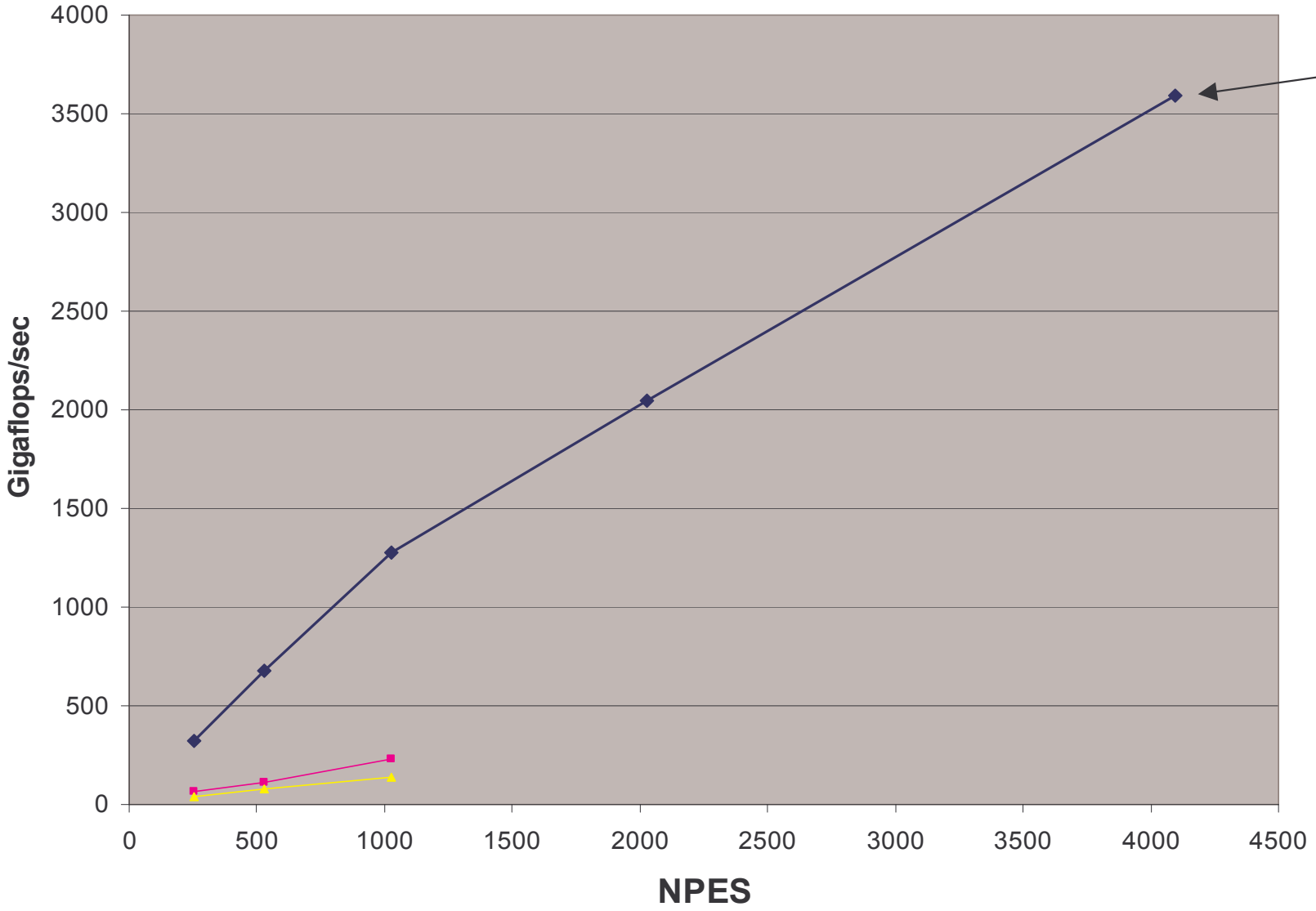
The 7th LCI International Conference on Clusters

NAS Parallel Benchmarks

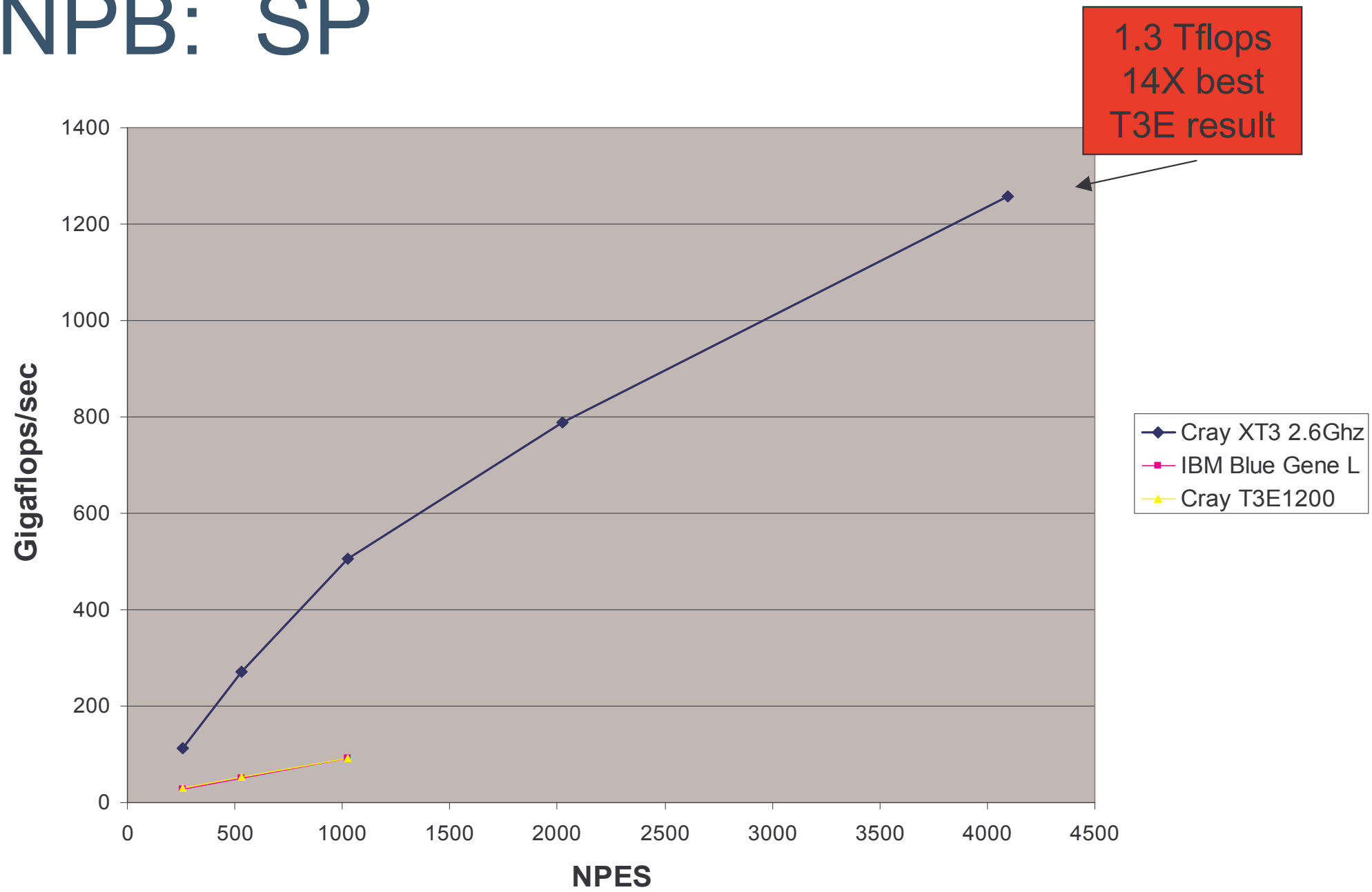


- The first time we looked at these was in 1991 on a Cray Y-MP
- The last time we looked at these was in 1999 on a 1024 Processor Cray T3E 1200
  - We tuned codes for E-registers, shmem, etc.
- We recently ran these on a 4096 PE Cray XT3
  - Codes were the ASIS MPI version (3.2) (one exception)
  - -O3 -fastsse
- IBM Blue Gene Results from Argonne
  - http://www-unix.mcs.anl.gov/~kaushik/bgl/npb_results.htm
  - Run in co-processor mode
  - -O3 -qarch=440d -qtune=440 -qbgl -qmaxmem=64000

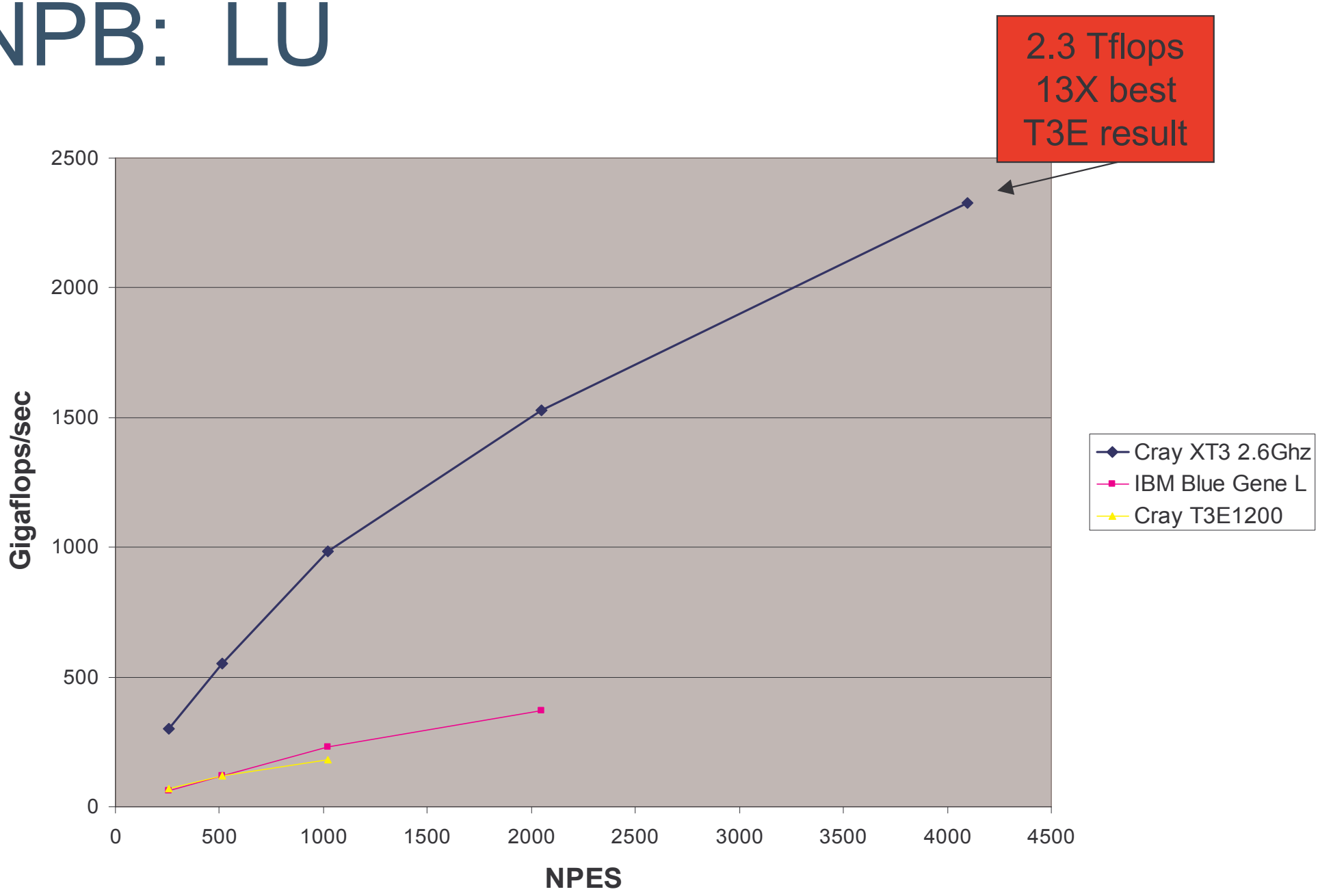
NPB: BT



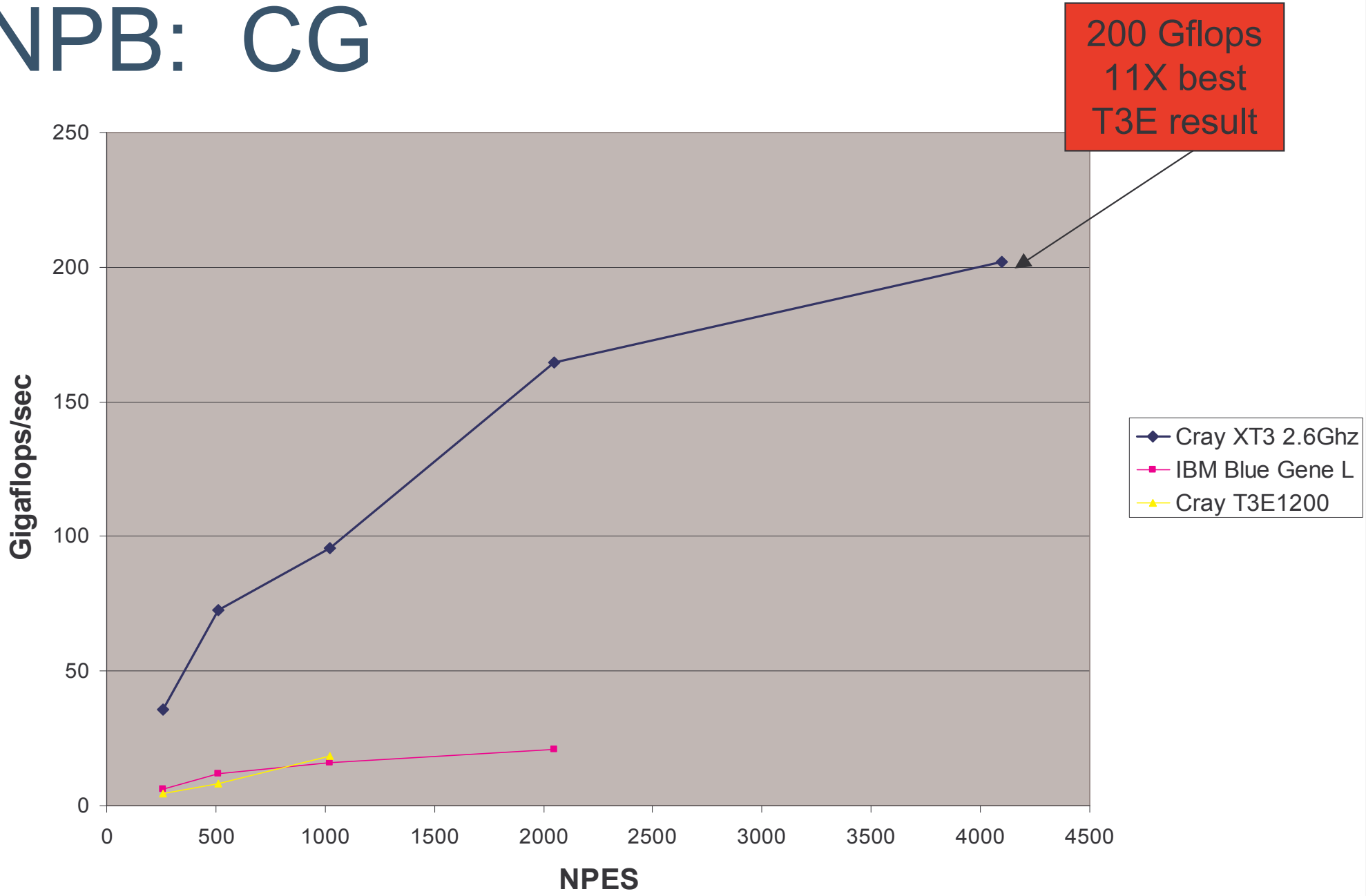
NPB: SP



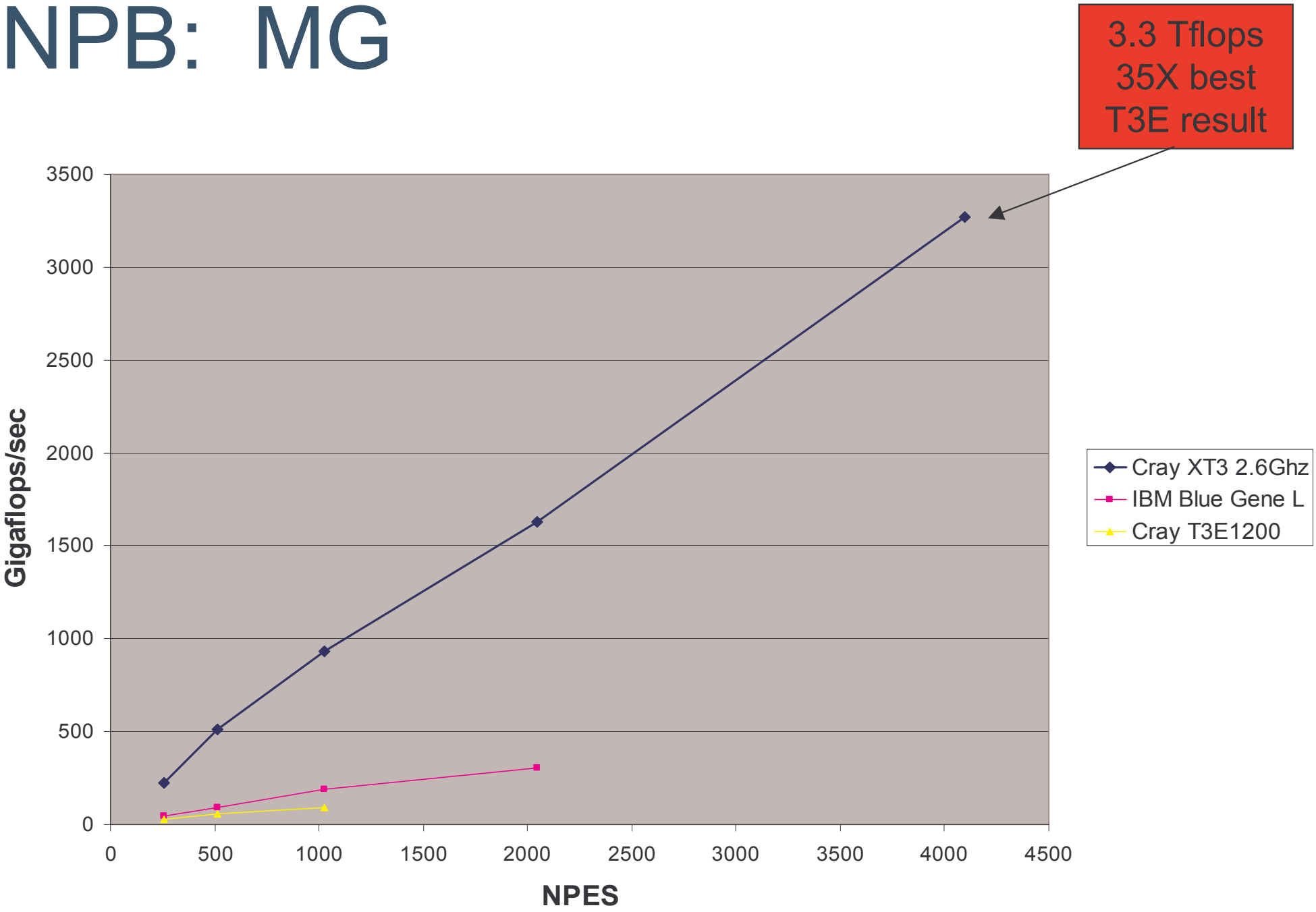
NPB: LU



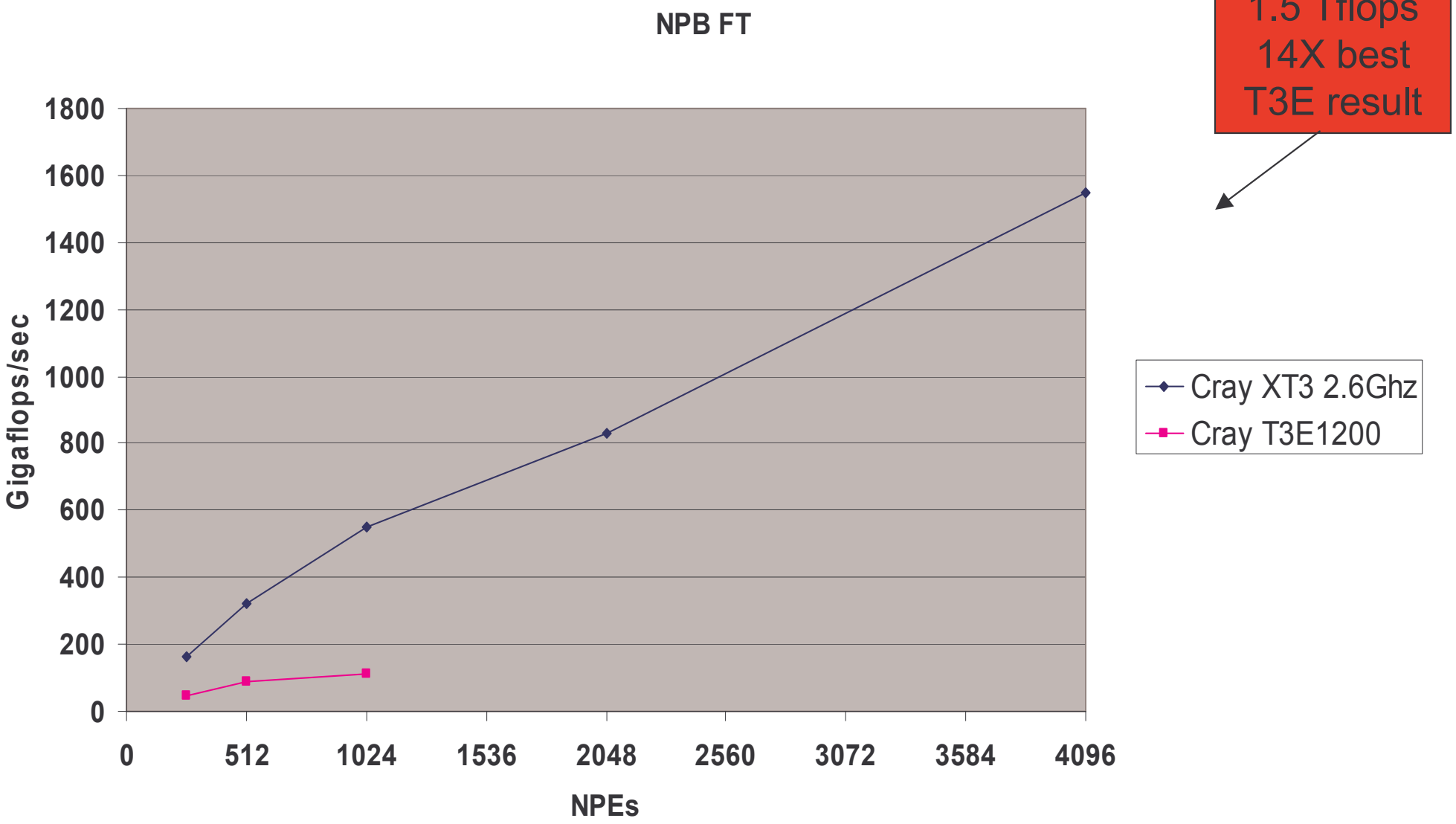
NPB: CG



NPB: MG



NPB: FT

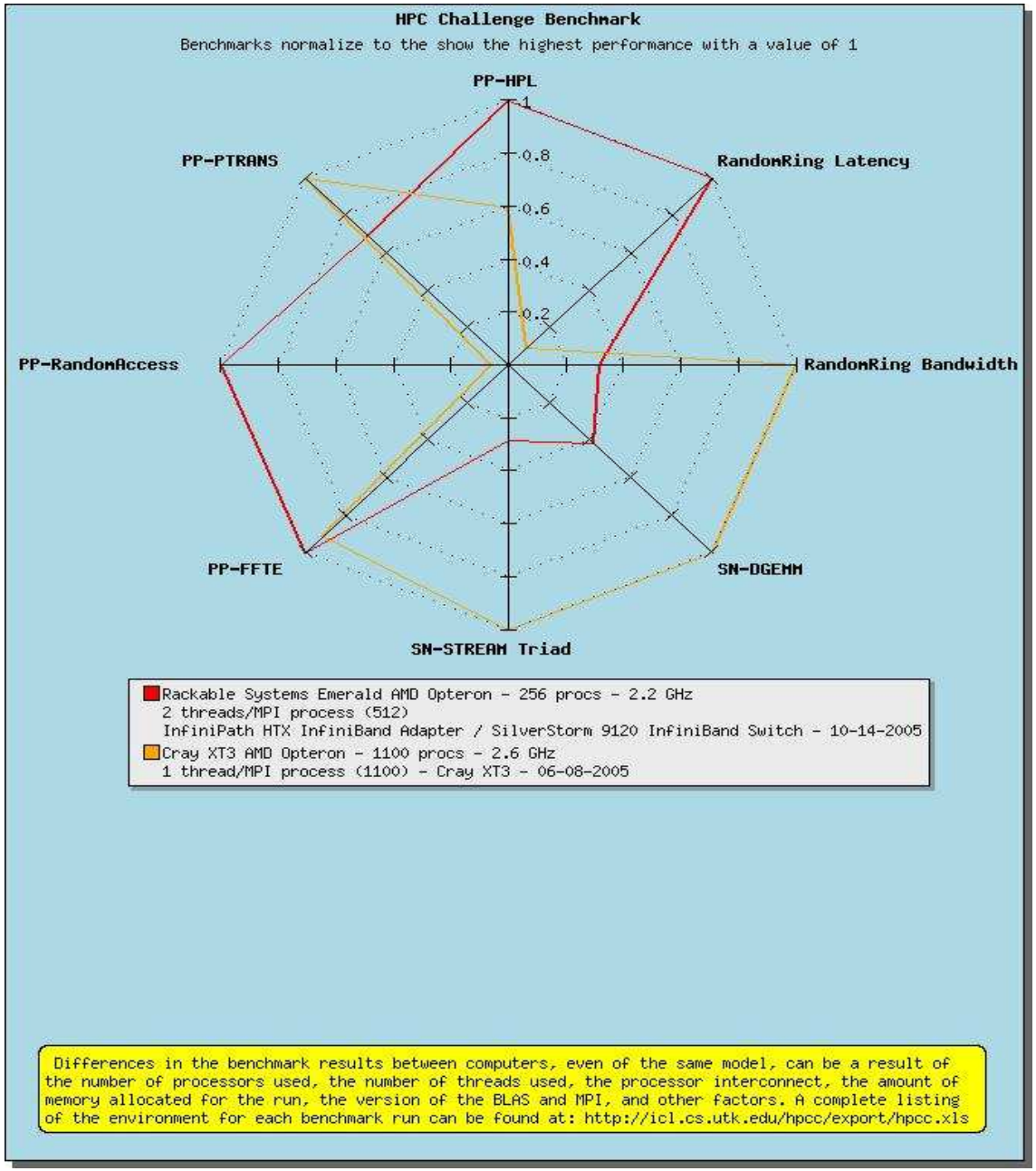


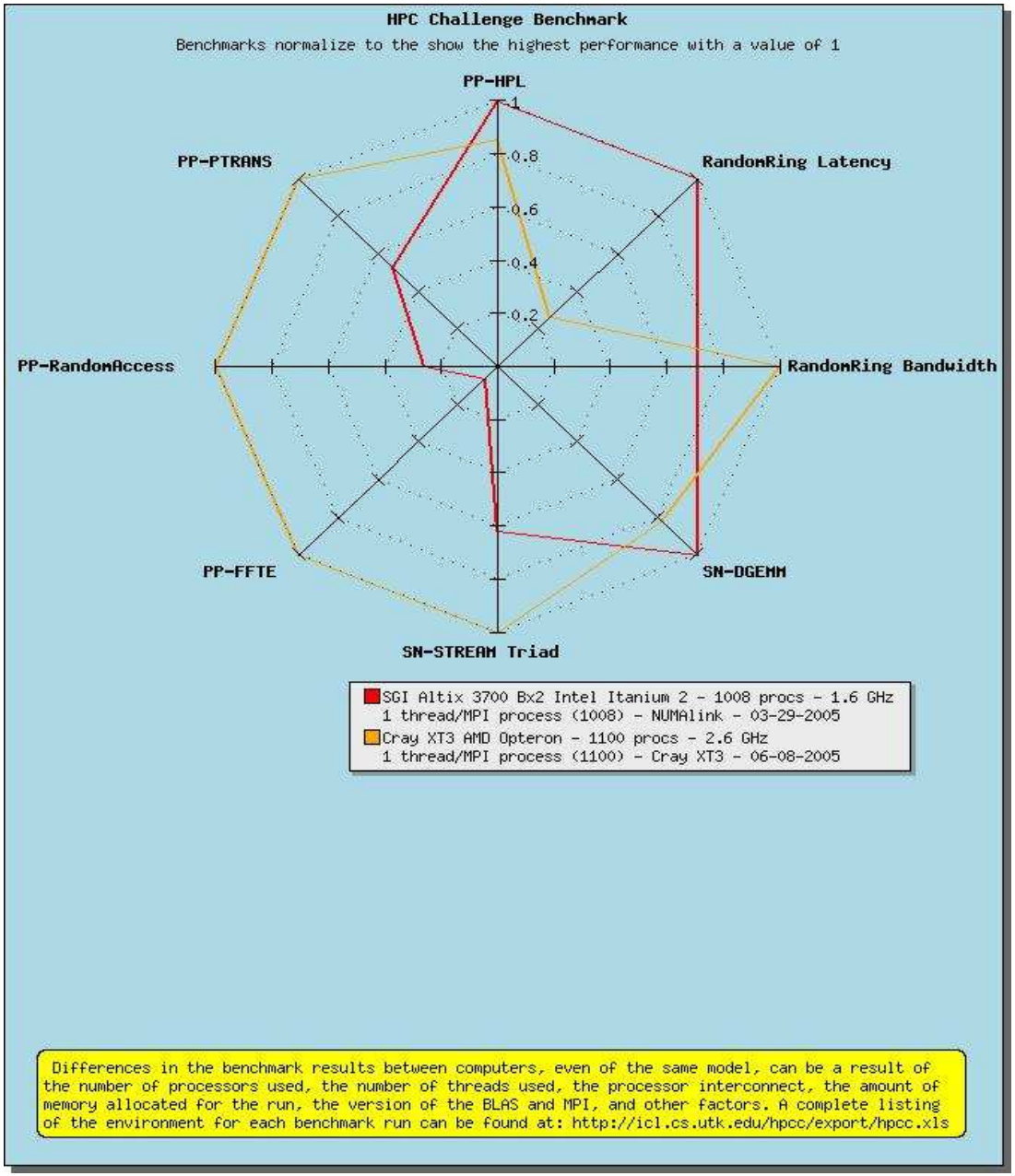
CRAY® THE SUPERCOMPUTER COMPANY

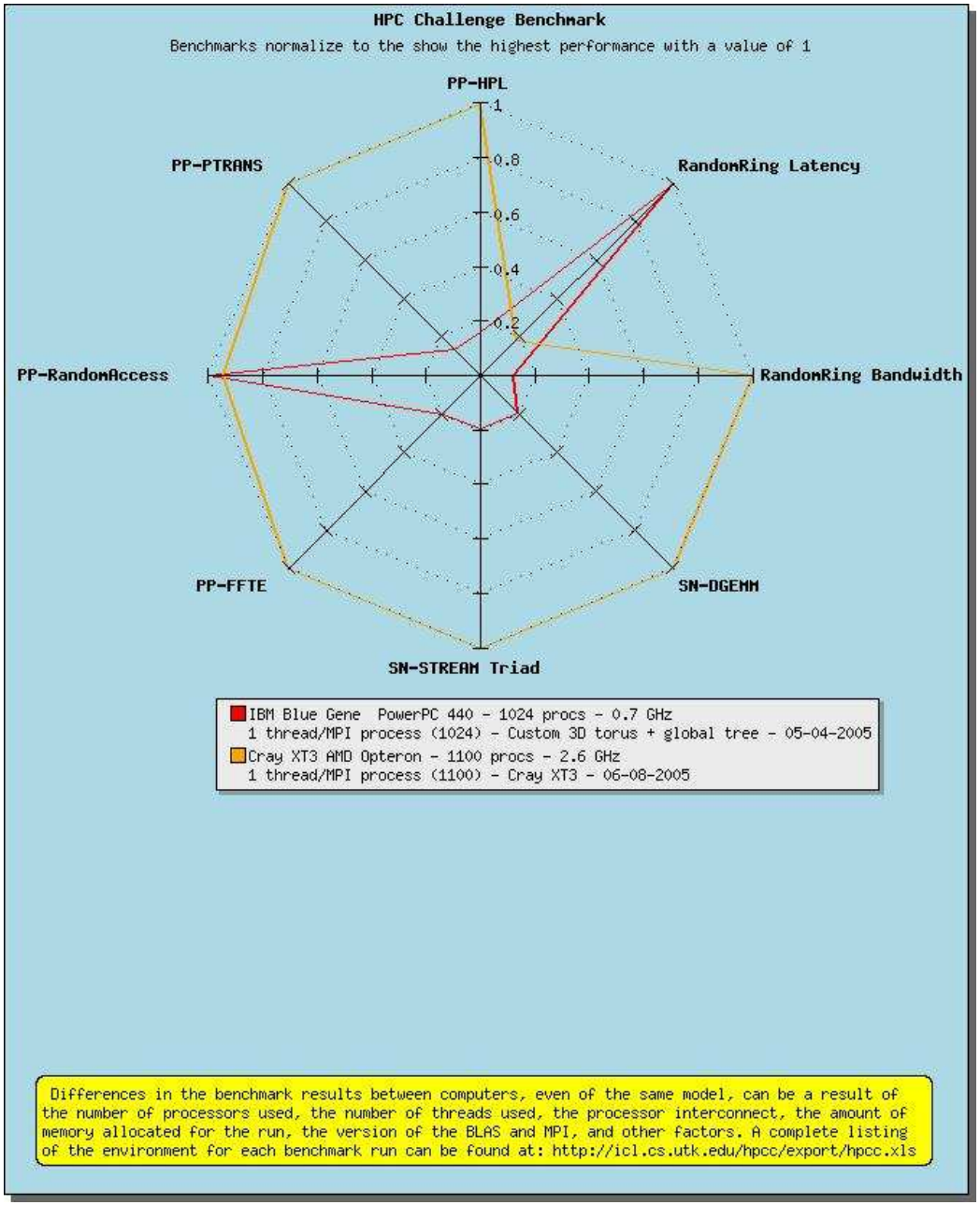
HPC Challenge

2/13/2006

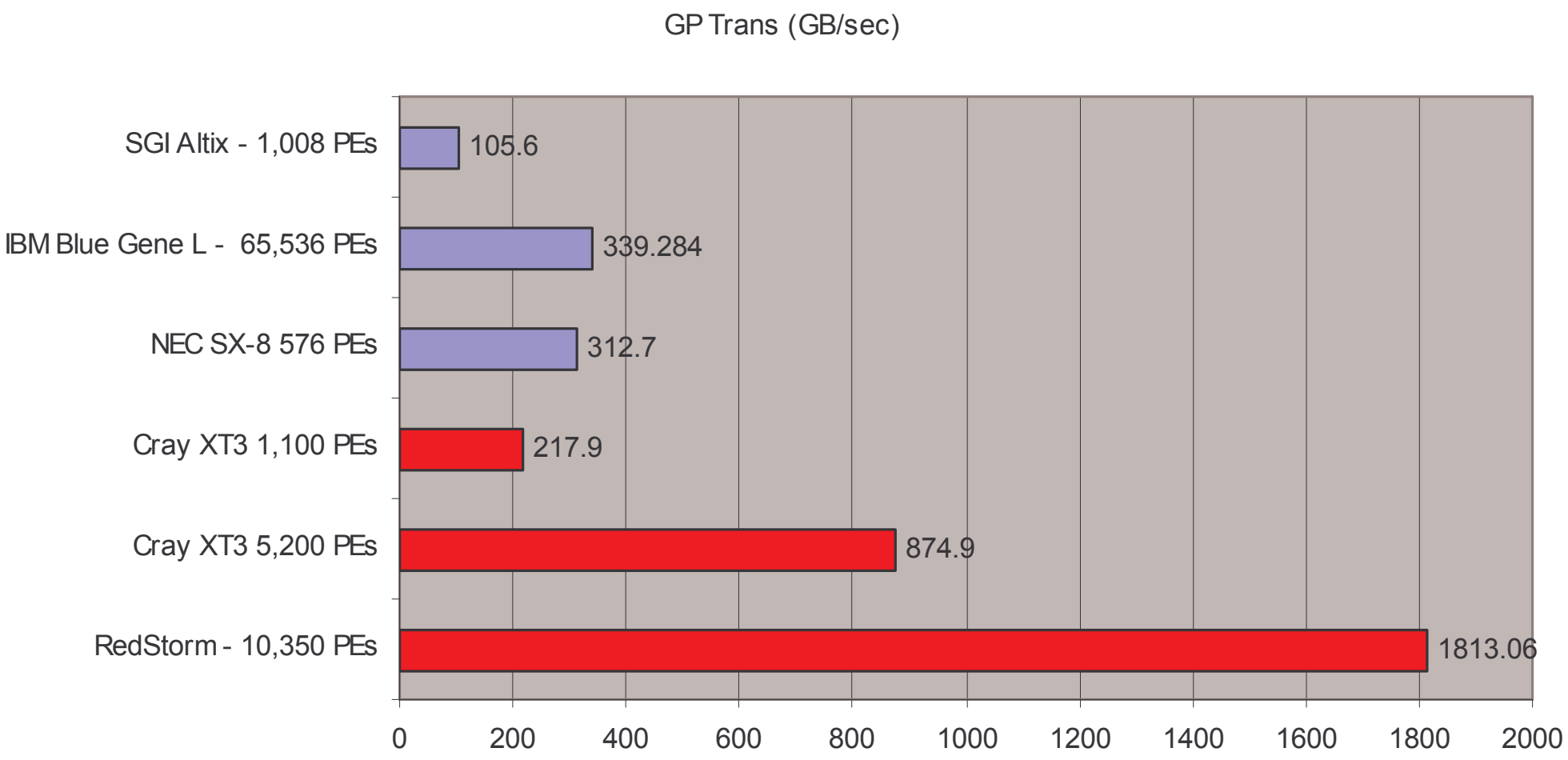
The 7th LCI International Conference on Clusters



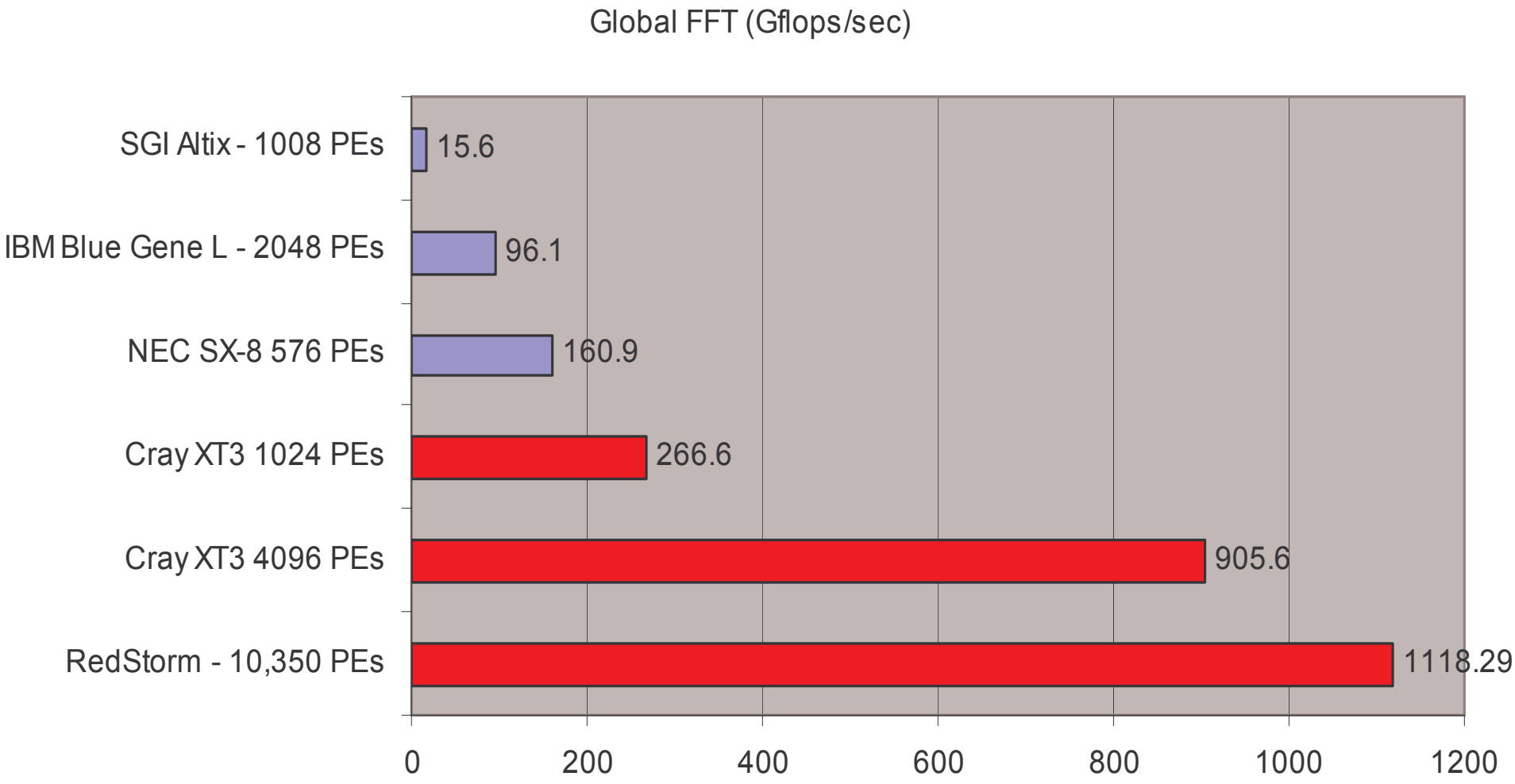




Ptrans (Global Bandwidth)



Global FFT (Gflops/sec)

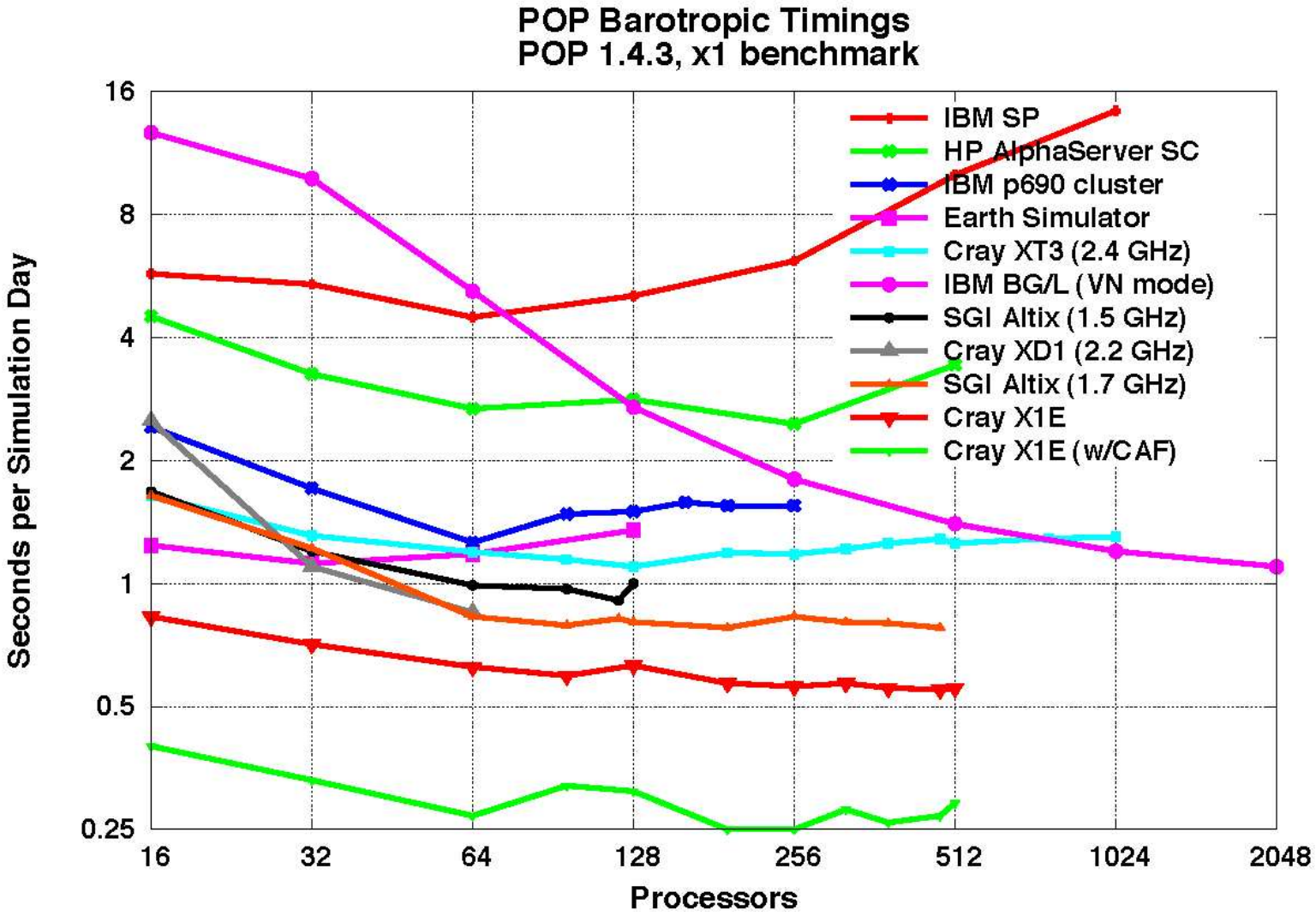


XT3 Application Results

2/13/2006

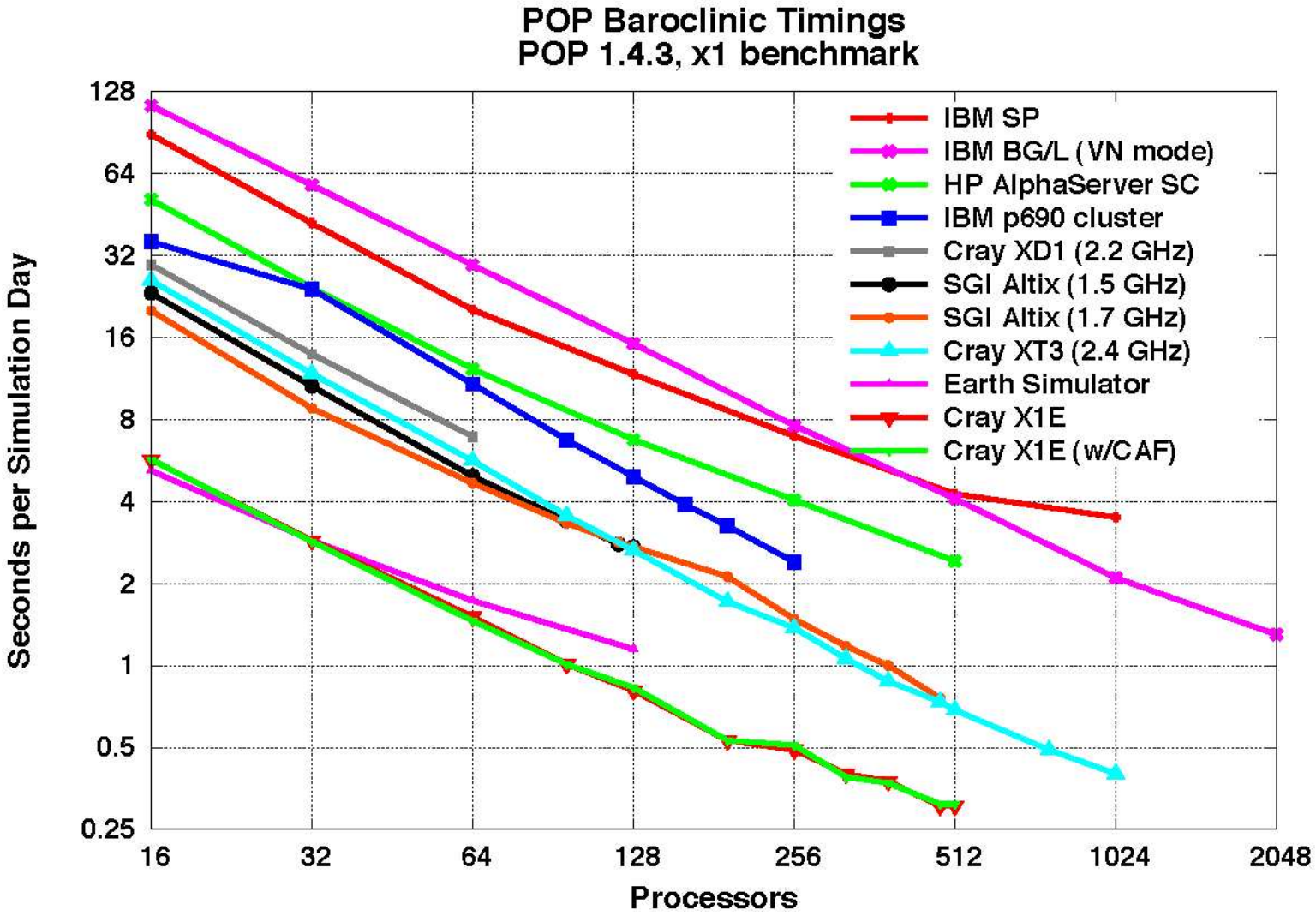
The 7th LCI International Conference on Clusters

POP Barotropic



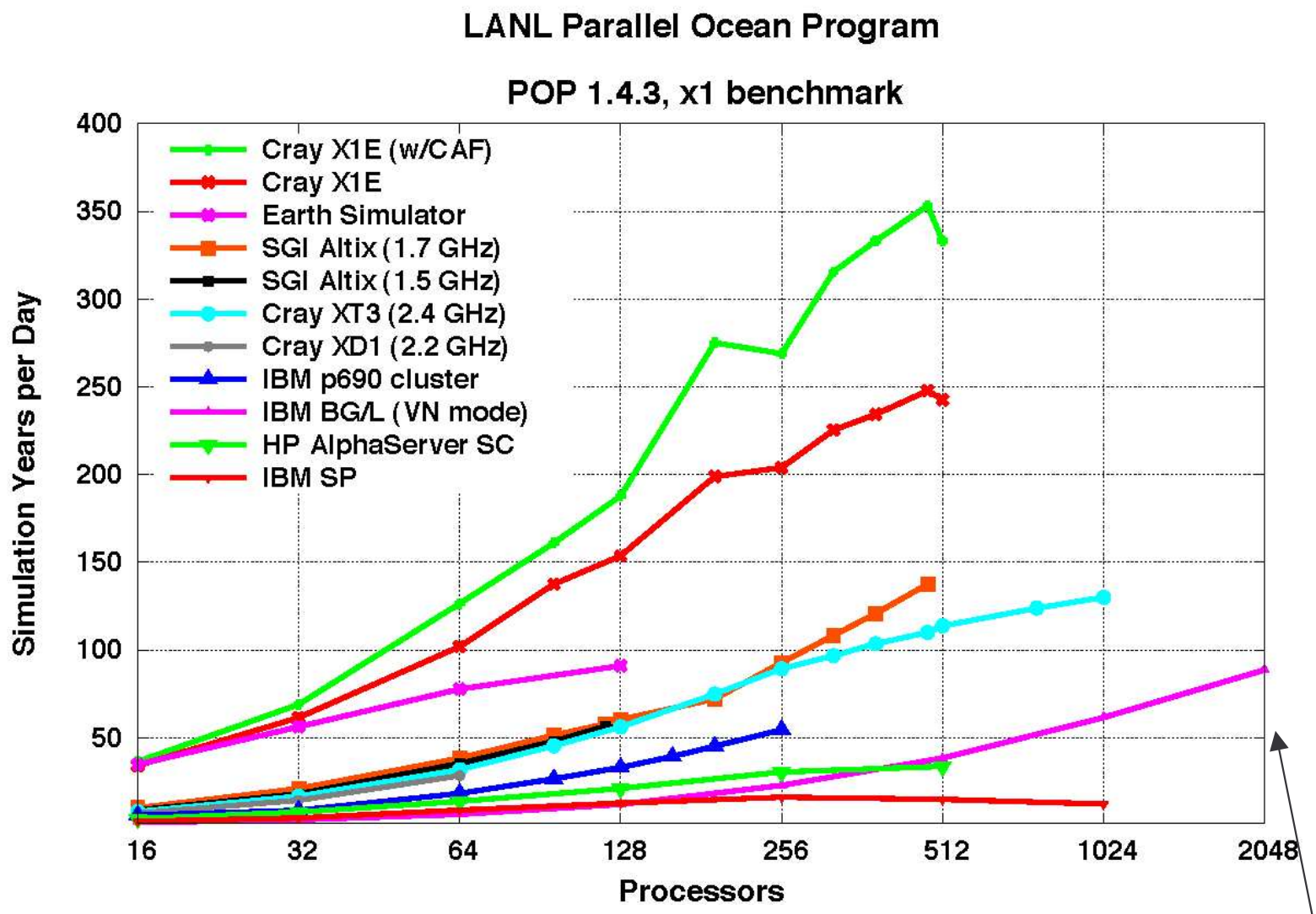
Lower is better

POP Baroclinic



Lower is better

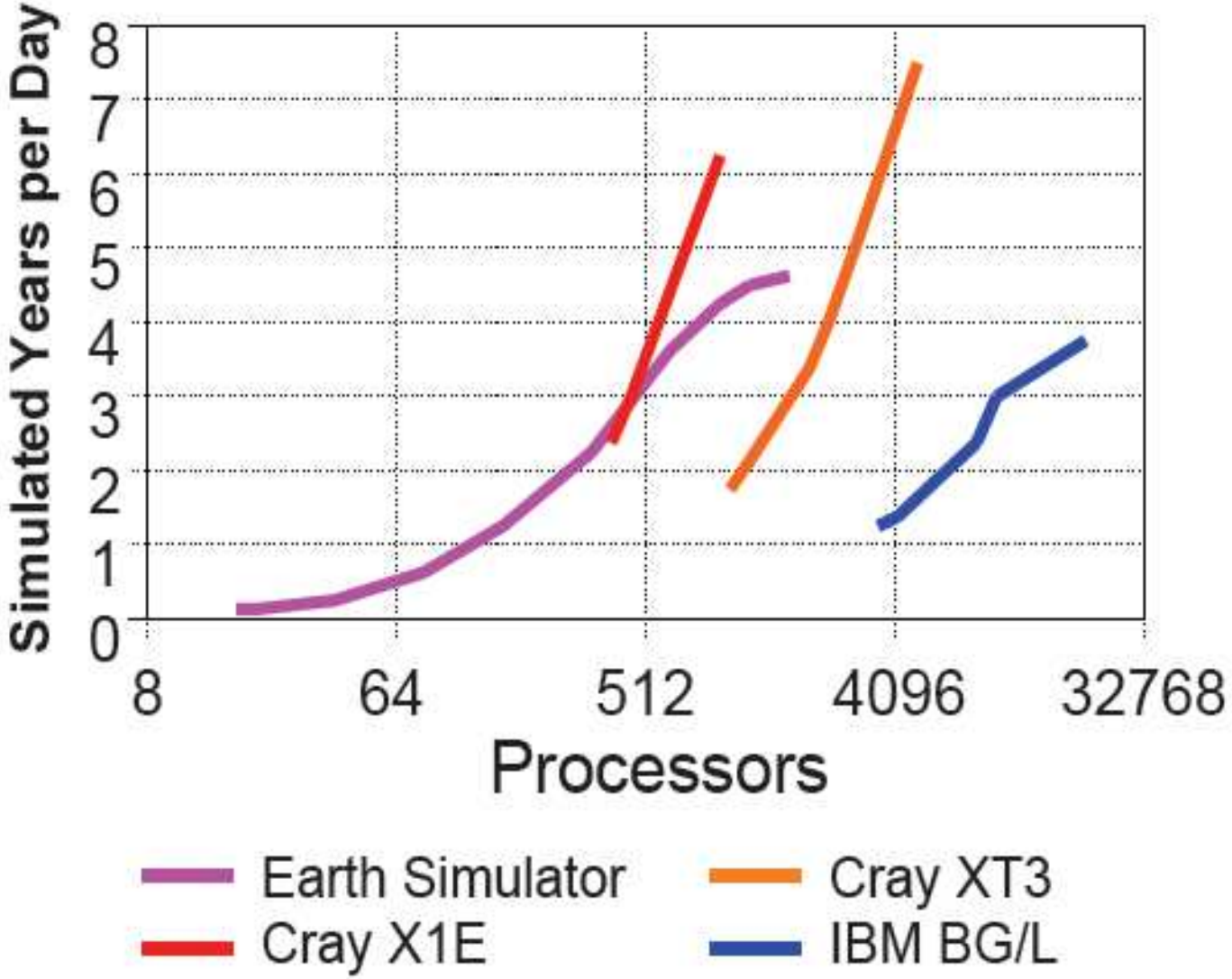
POP Overall Performance



Higher is better

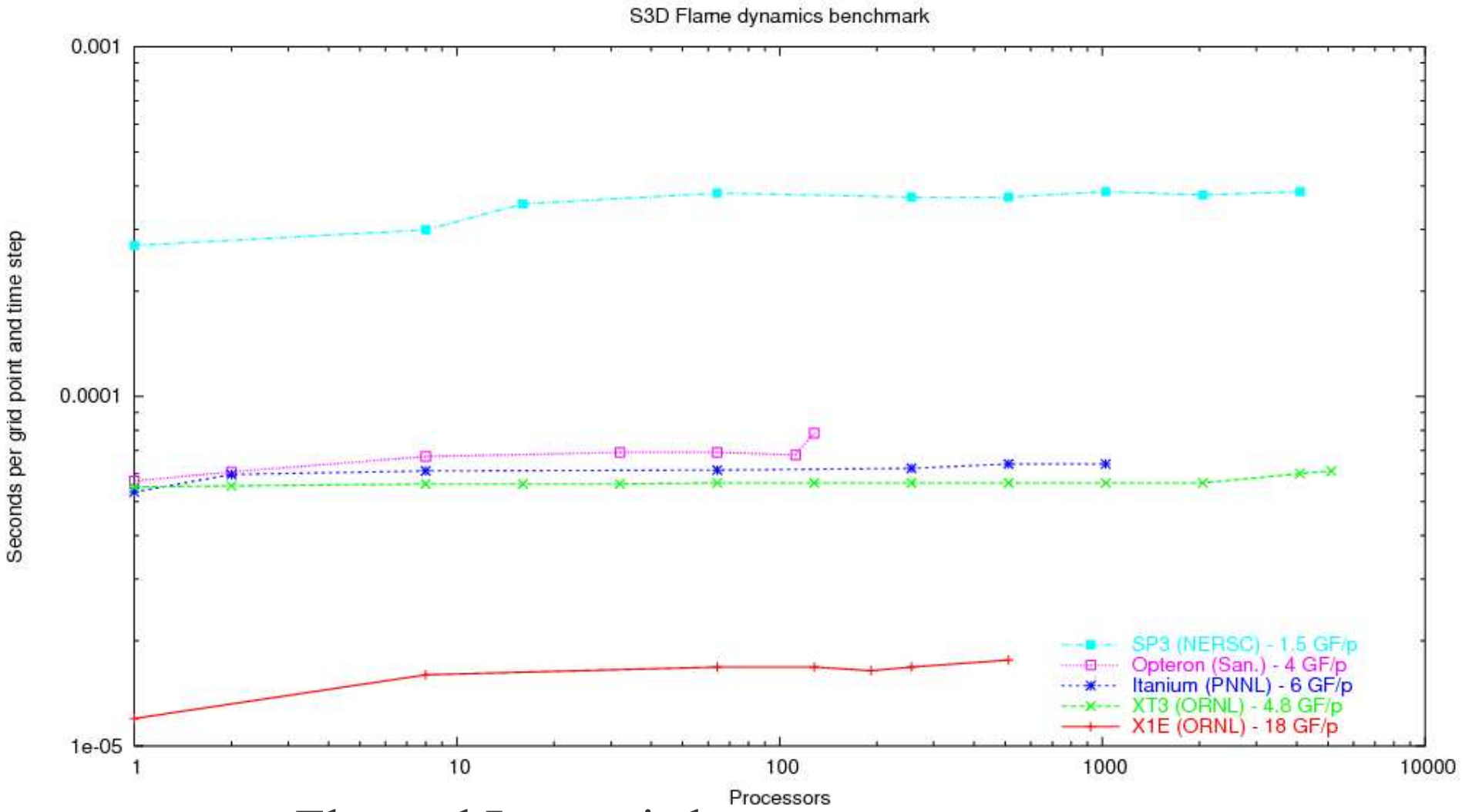
60 surface cells/processor

POP 1.4.3 tenth-degree benchmark



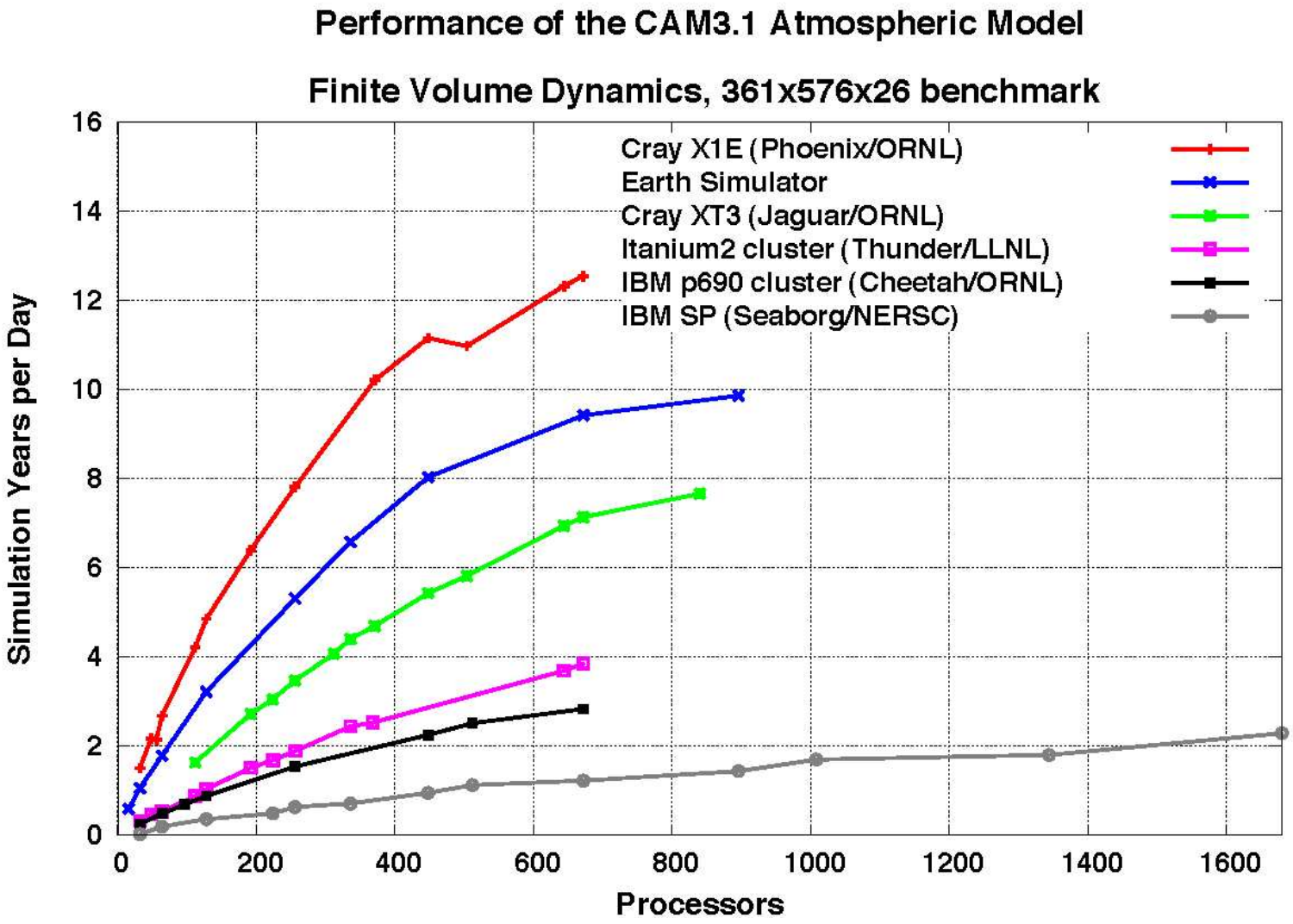
Higher is better

S3D Flame Dynamics



Flat and Lower is better

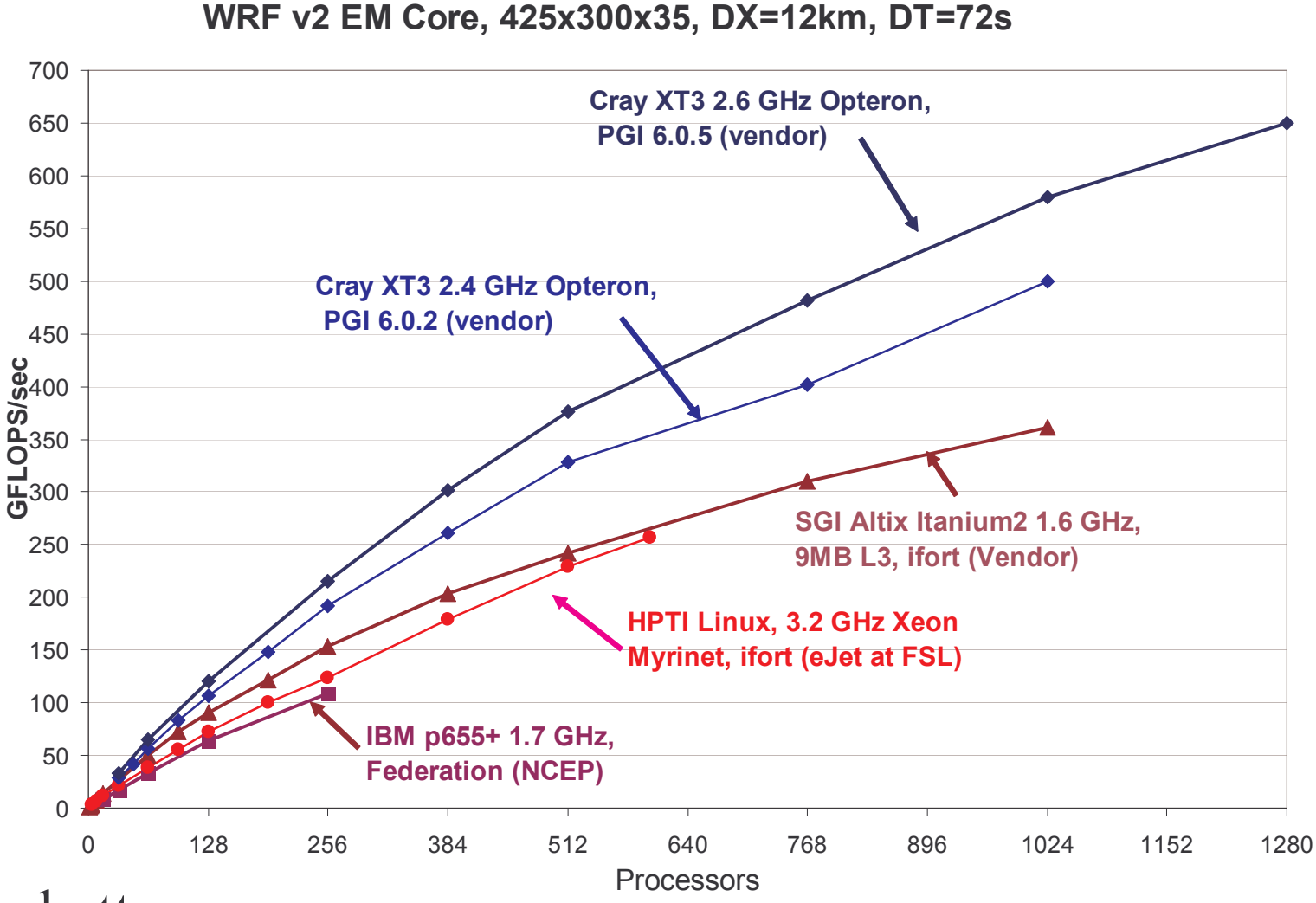
CAM Atmospheric Model



Higher is better

WRF Performance on XT3

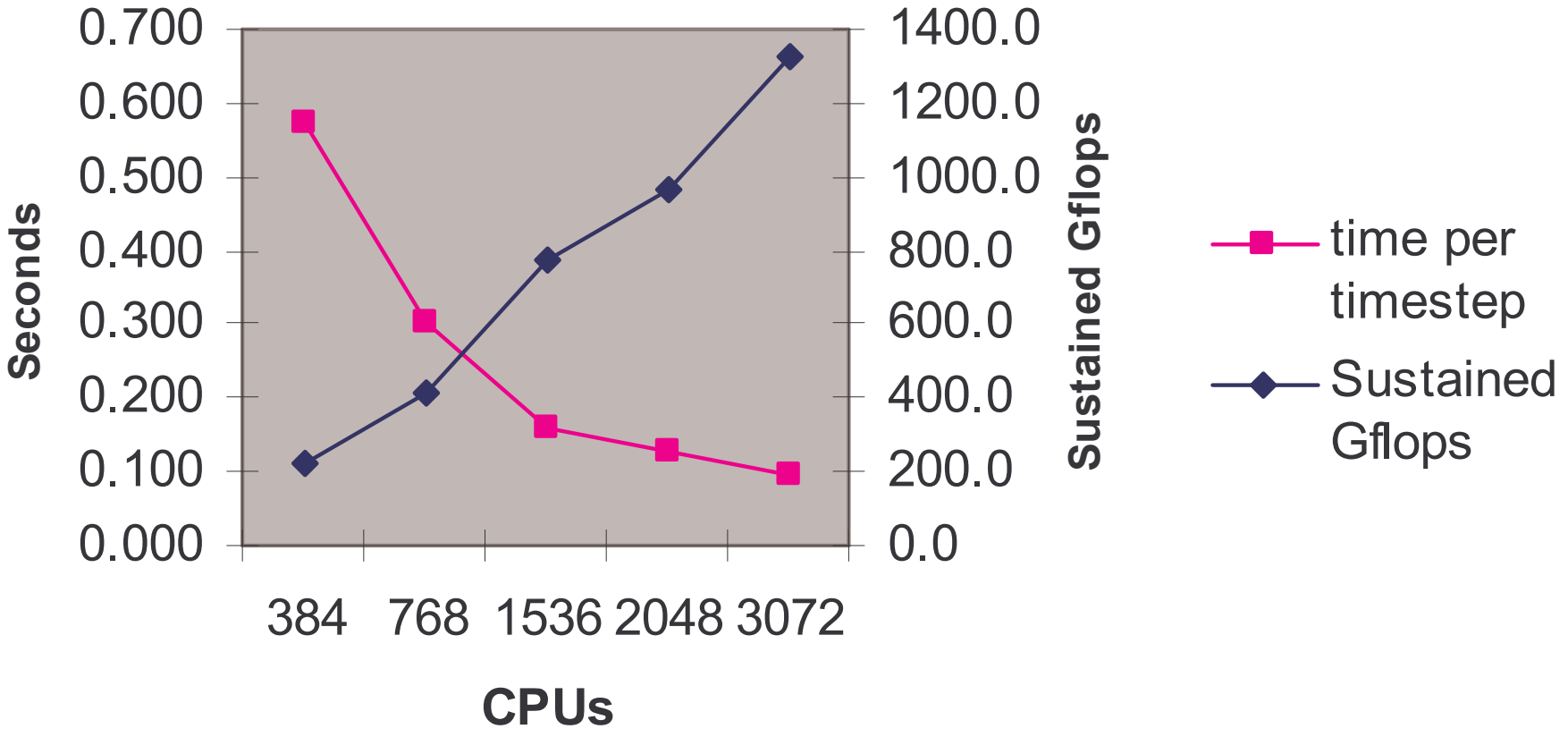
WRF THE WEATHER RESEARCH & FORECASTING MODEL



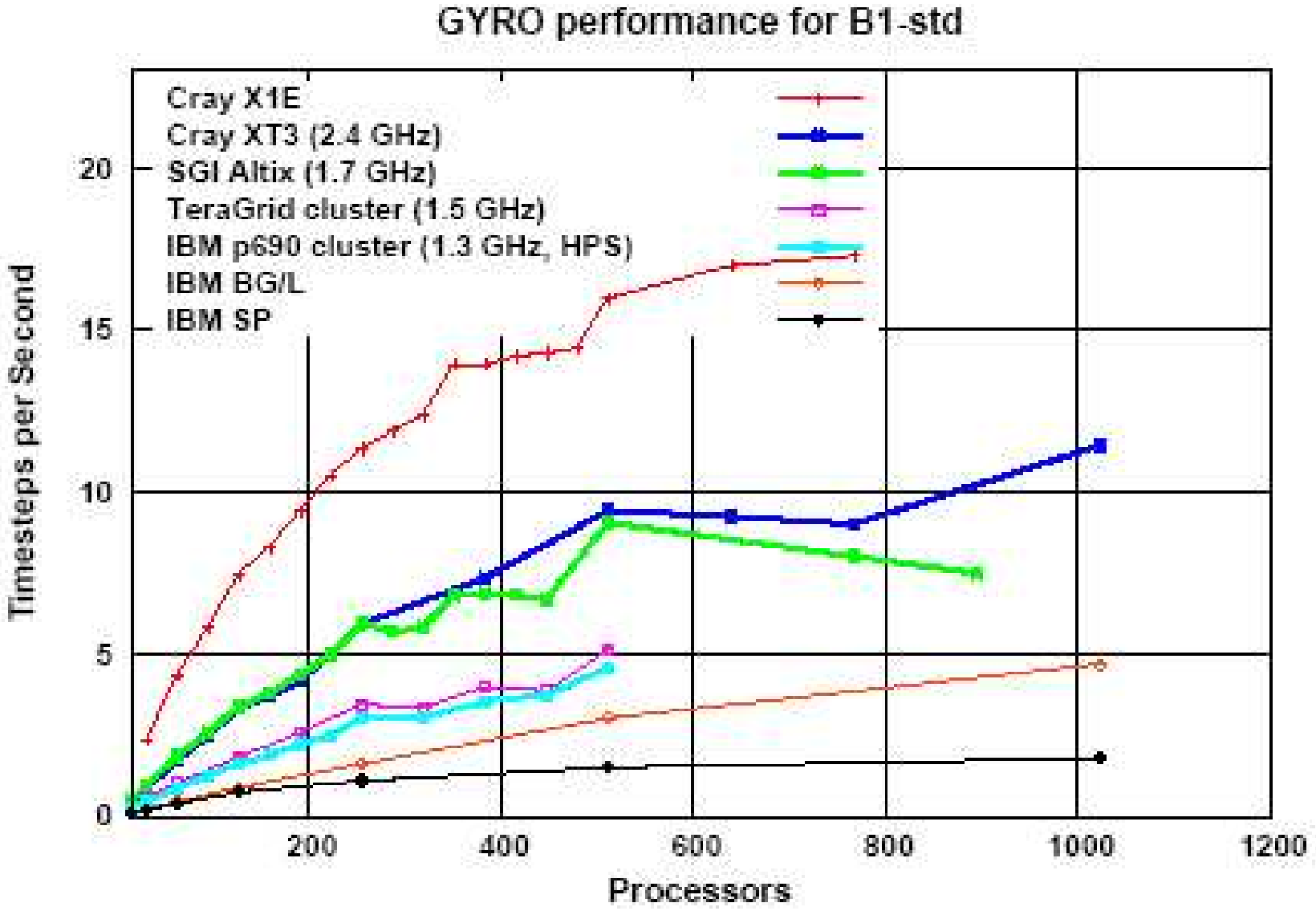
Higher is better

ECHAM5 T255L60 Performance on XT3

Max-Planck-Institut für Meteorologie
MaxPlanck Institute for Meteorology

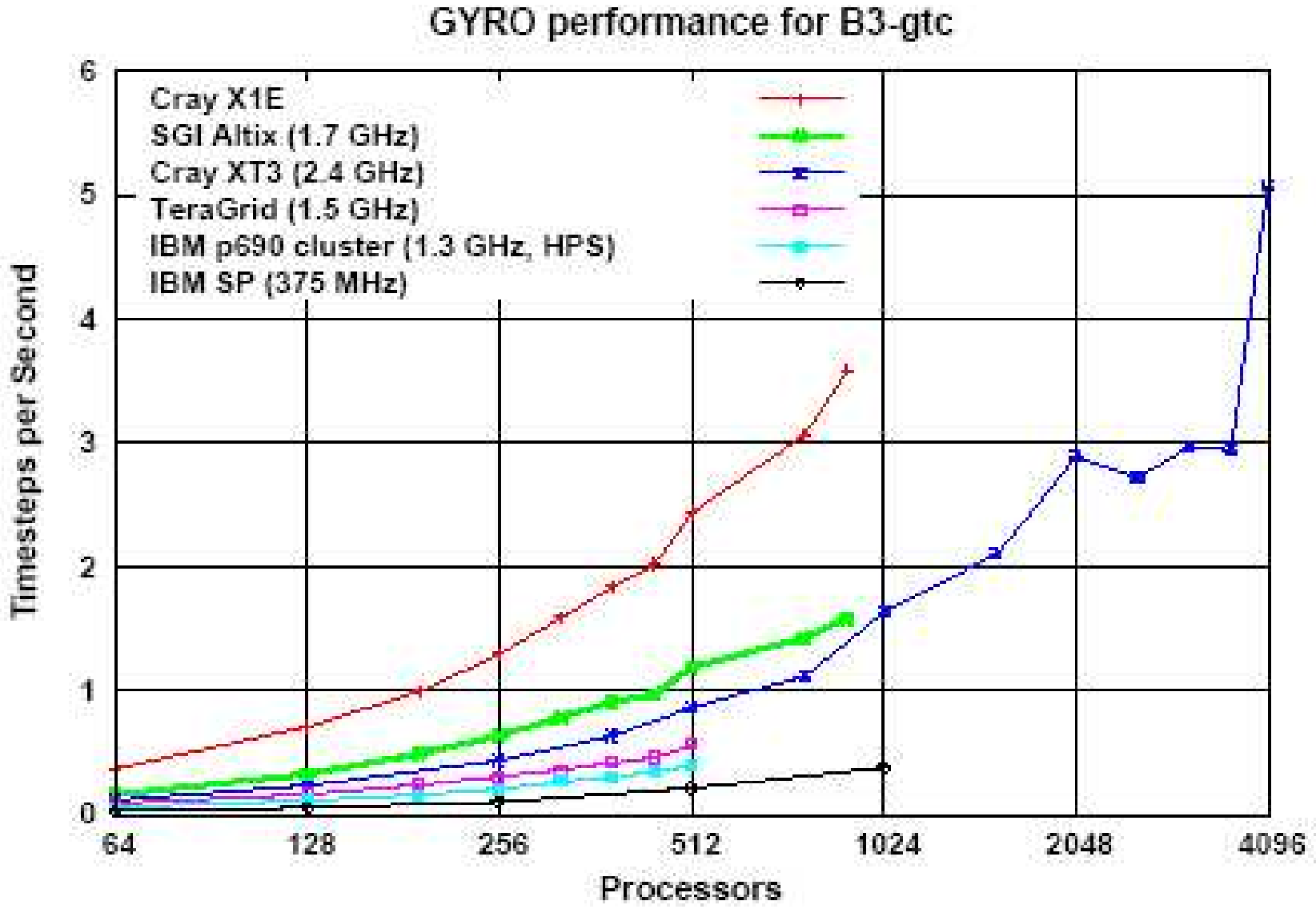


GYRO B1



Higher is better

GYRO B3



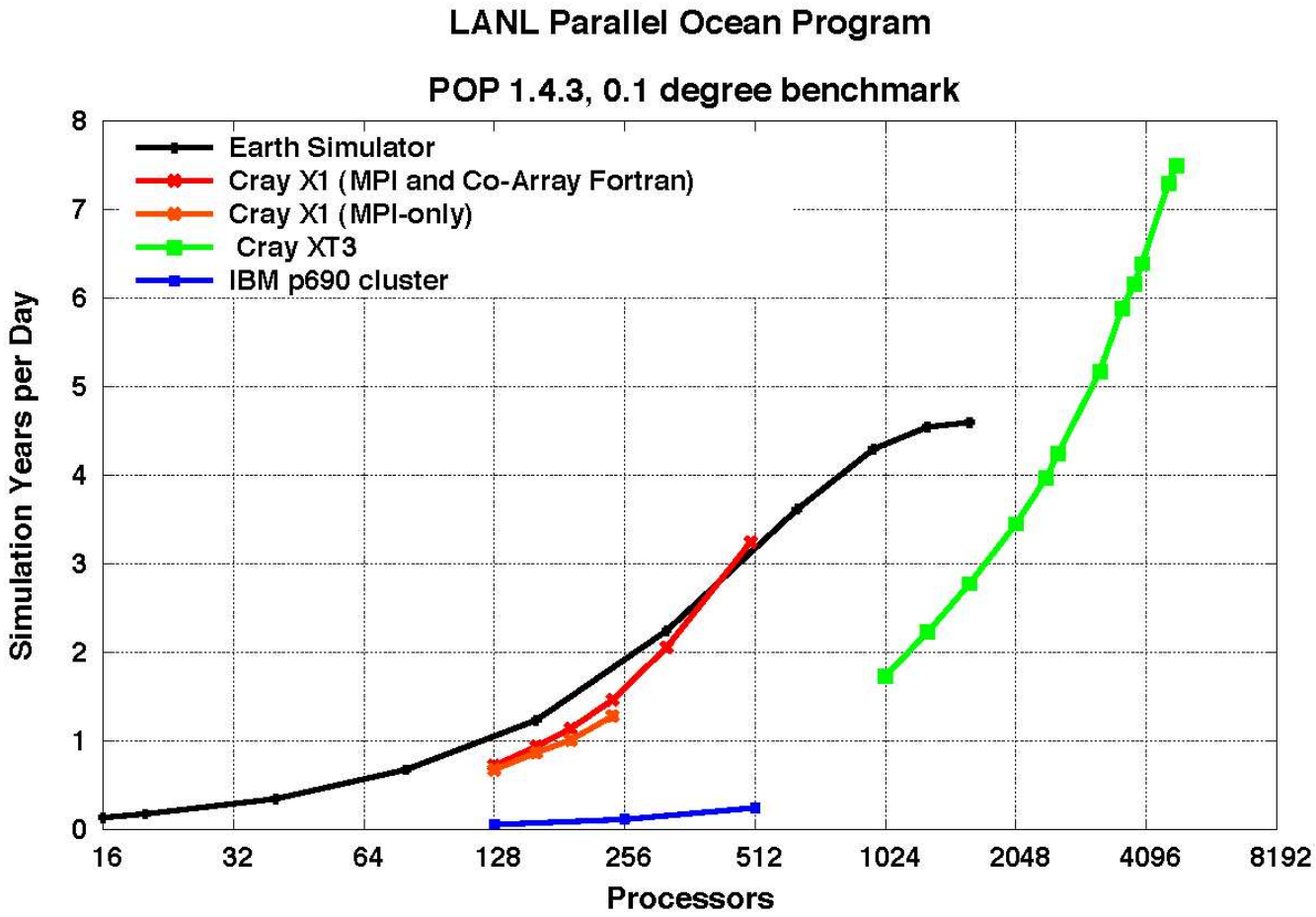
Higher is better

Conclusions

- We believe that we made the right choices
  - AMD Opteron provides excellent memory latency and injection bandwidth
  - SeaStar network gives extremely good network balance
  - Microkernel gives excellent scalability by reducing OS jitter.

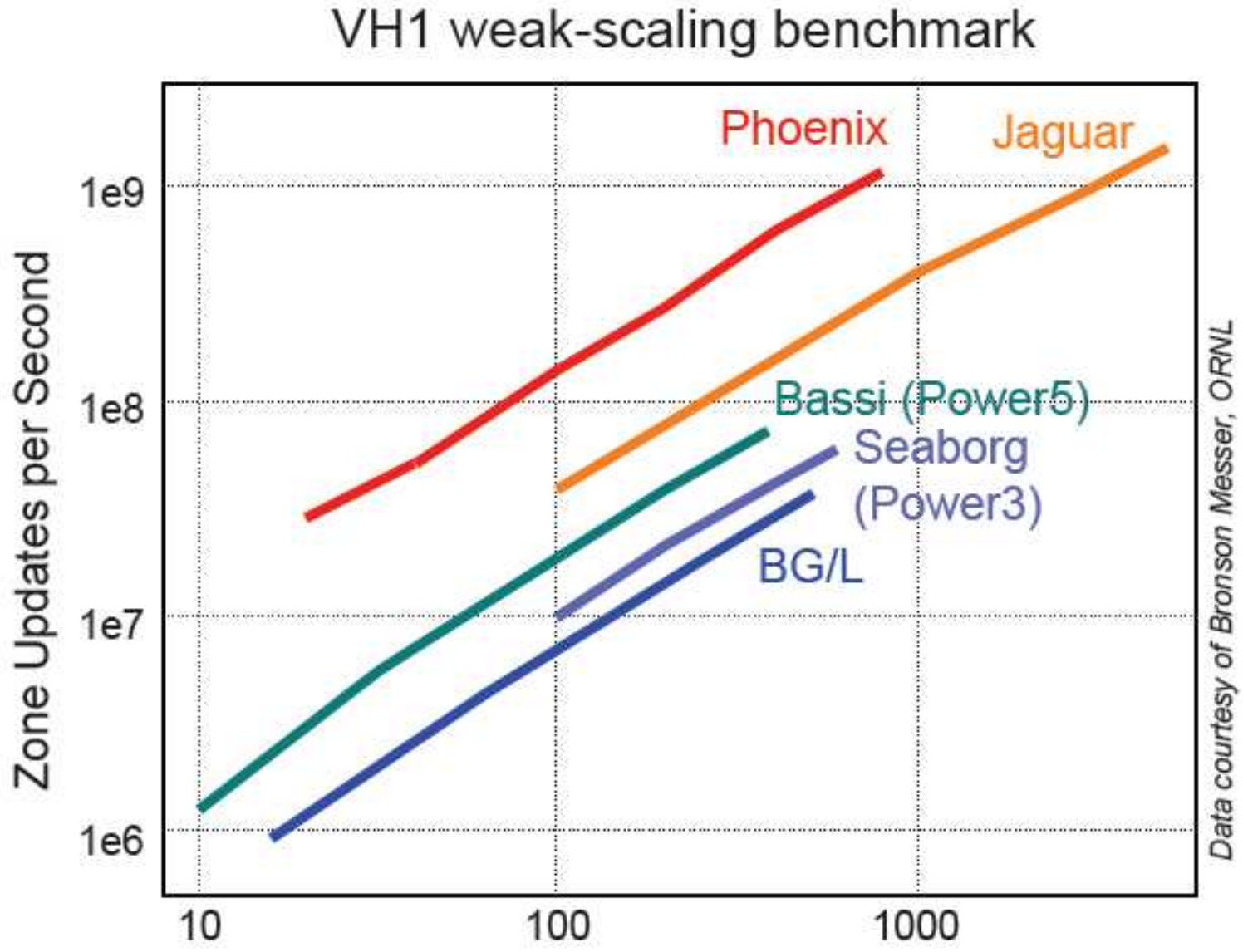
EXPLORE . SIMULATE . CREATE

Extra 1

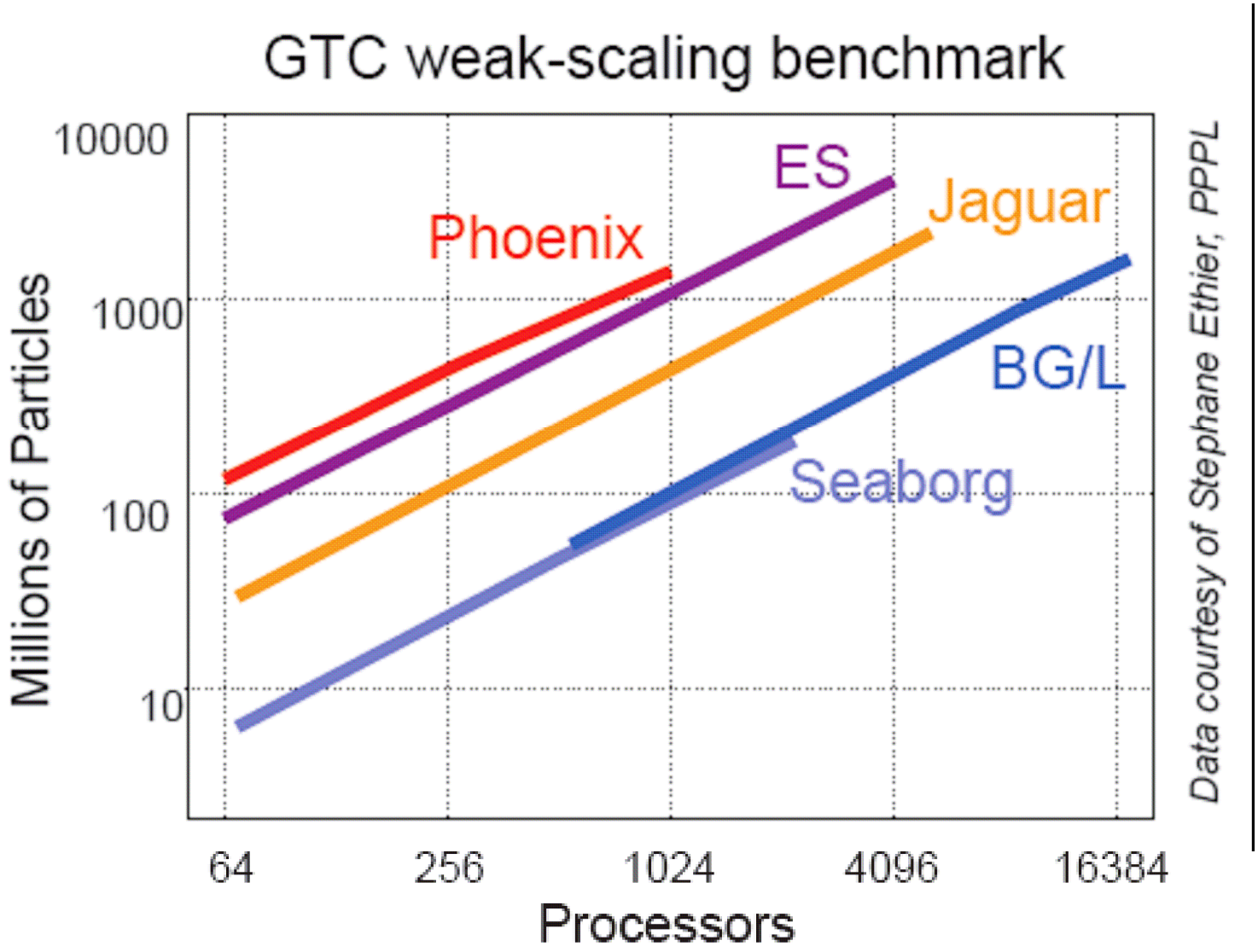


Higher is better

Extra 2



Extra 3


GTC weak-scaling benchmark
Millions of Particles
10000
1000
100
10
64
256
1024
4096
16384
Processors
ES
Phoenix
Jaguar
BG/L
Seaborg
Data courtesy of Stephane Ethier, PPPL