Communications Benchmarks on High-End and Commodity-Class Computers

Martyn F. Guest,
CCLRC Daresbury Laboratory

m.f.guest@daresbury.ac.uk

Outline

- Background - Distributed computing at DL
- Commodity-based and High-end Systems
  - Single-node performance & the Interconnect bottleneck
  - Prototype Commodity Systems; CS1 - CS20
  - High-end Systems: IBM SP/p690+, SGI Altix 3700, plus HP/Compaq Alpha Server SC and SGI Origin 3800
  - Performance Metrics
- Communication benchmarks
  - GbitEther, Myrinet, SCI, Infiniband, Quadrics ..
  - MPI and Global Arrays (GA tools)
  - Pallas PMB MPI Benchmark - point-to-point and MPI Collectives (V 2.2 - 16, 64 and 128 CPU systems) -
  - Pallas Bandwidth Benchmark - B_EFF (EFF_BW)
  - Global Array benchmarks (V 3.4)
    - NWChem, GAMESS-UK and MOLPRO

DisCo: Technical Progress in 2003-4

Hardware and Software Evaluation:

- CPU
  - IA32, x-86 and IA64 systems -
    Intel Pentium 4 and Xeon Systems (3.06 GHz),
    AMD Opteron 246 & 248 (2.0 & 2.2 GHz)
  - *Itanium2 (Intel Tiger 1, 1.2 and 1.5 GHz;*
    HP systems, 900 MHz, 1 and 1.5 GHz;
    SGI Altix 3700 - 1.3 (3MB L3) & 1.5 GHz (6MB L3)
- Networks
  - Gigabit Ethernet options, cards, switches, channel-bonding, ….
  - SCI, Infiniband and Myrinet (P4/2400, P4/2666, P4/2800, Opteron 246 & 248 Clusters: OCF, Streamline, ClusterVision & Workstations UK), Quadrics
- System Software
  - message passing S/W (LAM MPI, LAM MPI-VIA, MPICH, VMI, SCAMPI), libraries (ATLAS, NASA, MKL, ACML, ScaLAPACK), compilers (Absoft, PGI, Intel's ifc and efc, Pathscale, GNU/g77), tools (GA tools, PNNL)
  - resource management software (PBS, TORQUE, GridEngine, **LSF** etc.)

www.cse.clrc.ac.uk/Activity/DisCo

High-End Systems Evaluated

- Cray T3E/1200E (… historical …)
  - 816 processor system at Manchester (CSAR), 600 Mz Alpha EV56 CPU, 256 MB
- IBM pseries 690 and pseries 690+ (Daresbury)
  - ***IBM p690* (8-way LPAR'd nodes, 1280 X 1.3 GHz CPUs with colony, HPCx)**
  - ***IBM p690+* (32-way nodes, 1600 X 1.7 GHz CPUs with HPS, HPCx- Phase2)**
- Compaq AlphaServer SC
  - 4-way ES40/667 A21264A (APAC) and 833 MHz SMP nodes (2 GB RAM);
  - ***TCS1 system at PSC*** (750 4-way ES45 nodes - 3,000 EV68 CPUs - 4 GB memory per node, 8MB L2 cache), Quadrics interconnect (5 usec latency, 250 MB/sec B/W)
- SGI Origin 3800
  - SARA (1000 CPUs) - Numalink with MIPS R14k/500 CPUs
- *SGI Altix 3700*
  - ***Linux Cluster - Numalink with Itanium 2 1.3 GHz CPUs, 3MB L3 cache***
    - ***CSAR ("newton" 512 CPUs) and SARA ("aster" - 416 CPUs - 7 nodes)***
  - ***ORNL ("ram" 256 CPUs with Itanium 2 1.5 GHz CPUs, 6MB L3 cache)***

Commodity Systems (CSx)

Prototype / Evaluation Hardware

Systems	Location	CPUs	Configuration
CS1	Daresbury	32	PentiumIII / 450 MHz + FE (EPSRC)
CS2	Daresbury	64	24 X dual UP2000/EV67-667, QSNet Alpha/LINUX cluster, 8 X dual CS20/EV67-833 (“loki”)
CS3	RAL	16	Athlon K7 850MHz + myrinet
CS4	*Sara*	32	Athlon K7 1.2 GHz + FE
CS6	*CLiC*	528	PentiumIII / 800 MHz; fast ethernet (Chemnitzer Cluster)
CS7	Daresbury	64	AMD K7/1000 MP + SCALI/SCI (“ukcp”)
CS8	*NCSA*	320	160 dual IBM Itanium/800 + Myrinet 2k (“titan”)
CS9	*Bristol*	96	Pentium4 Xeon/2000 + Myrinet 2k (“dirac”)
Prototype Systems			
CS0	Daresbury	10	10 CPUS, Pentium II/266
CS5	Daresbury	16	8 X dual Pentium III/933, SCALI

Commodity Systems (CSx) II.

Systems	Location	CPUs	Configuration
CS10	*Hull*	64	Pentium4 Xeon/2667 + Myrinet 2k ("eagle"), Streamline/SCORE
CS11	*Workstations*	32	Pentium4 Xeon/2667 + GbitEther, ScaMPI
CS12	*Essex*	48	Pentium4 Xeon/2400 + GbitEther ("sstream1"), Streamline/SCORE
CS13	*White Rose, Leeds*	256	Pentium4 Xeon/2200-2400 + M2k ("snowdon"), Streamline/SCORE
CS14	*NCSA*	1024	Pentium III Xeon/1000 + M2k ("platinum")
CS15	*SDSC*	128	Pentium III Xeon/ 800 + M2k ("meteor")
CS16	*SDSC*	256	dual-Itanium2/1.3 GHz + M2k ("Teragrid")
CS17	*Daresbury*	32	Pentium4 Xeon/2667 + GbitEther ("ccp1"), Streamline/SCORE
CS18	*Bradford*	78	Pentium4 Xeon/2800 + M2k/GbitE ("grendel")
CS19	*Daresbury*	64	dual-Opteron/246 2.0 GHz nodes + Infiniband, Gbit and SCI ("scaliwag")
CS20	*RAL*	256	dual-Opteron/248 2.2 GHz nodes + Myrinet ("scarf")

Interconnects and Networking

- Ethernet, Gbit etc.
- Myrinet, Quadrics and Dolphin SCI
  - 64 bit PCI implementations, with faster PCI-X options near market
- Infiniband PCI-X interconnect

- Key to any interconnect is the performance of the library implementation.
- MPICH (ANL)
- ScaMPI library from Scali
- PMB MPI Benchmarks (Pallas)

Performance of current interconnects.

	Latency (µs)	Bandwidth (MB/s)	Switch size	Message size (kB)	1MB transmit (ms)
Quadrics Elan3	5	325 (680)	128-2000	1.6	3.08
Dolphin SCI	5 (1.5 intranode)	326	N/A	1.6	3.08
Myrinet	~8 (6.3)	243 (500)	8-128	2	4.12
Mellanox Infiniband	7-10 (5.5)	800 (830)	8-96	5.6-8	1.26
Gbit. Ethernet	30-100	125	64	3.7-12.5	8.1

Communications Benchmark
PMB and EFF_BW

- PMB - Pallas MPI Benchmark Suite (V2.2)

Point-to-Point (Mbytes/sec)

PingPong; PingPing; Sendrecv; Exchange

Collective Operations (Time - usec) - as function of no.of CPUs

Allreduce; Reduce; Reduce_scatter; Allgather; Allgatherv; Alltoall; Bcast, Barrier

Message Lengths:

0 to 4194304 Bytes

- EFF_BW
- Spin off from PMB, the EFF_BW benchmark (Pallas) calculates an "effective bandwidth".
- A single integral number is calculated which includes the performance for small and large messages under participation of all available processors,
- EFF_BW uses only the PMB PingPong Benchmark for measuring startup and throughput.

Communication Benchmarks

PMB: Pallas MPI Benchmark Suite (V2.2) and B_EFF

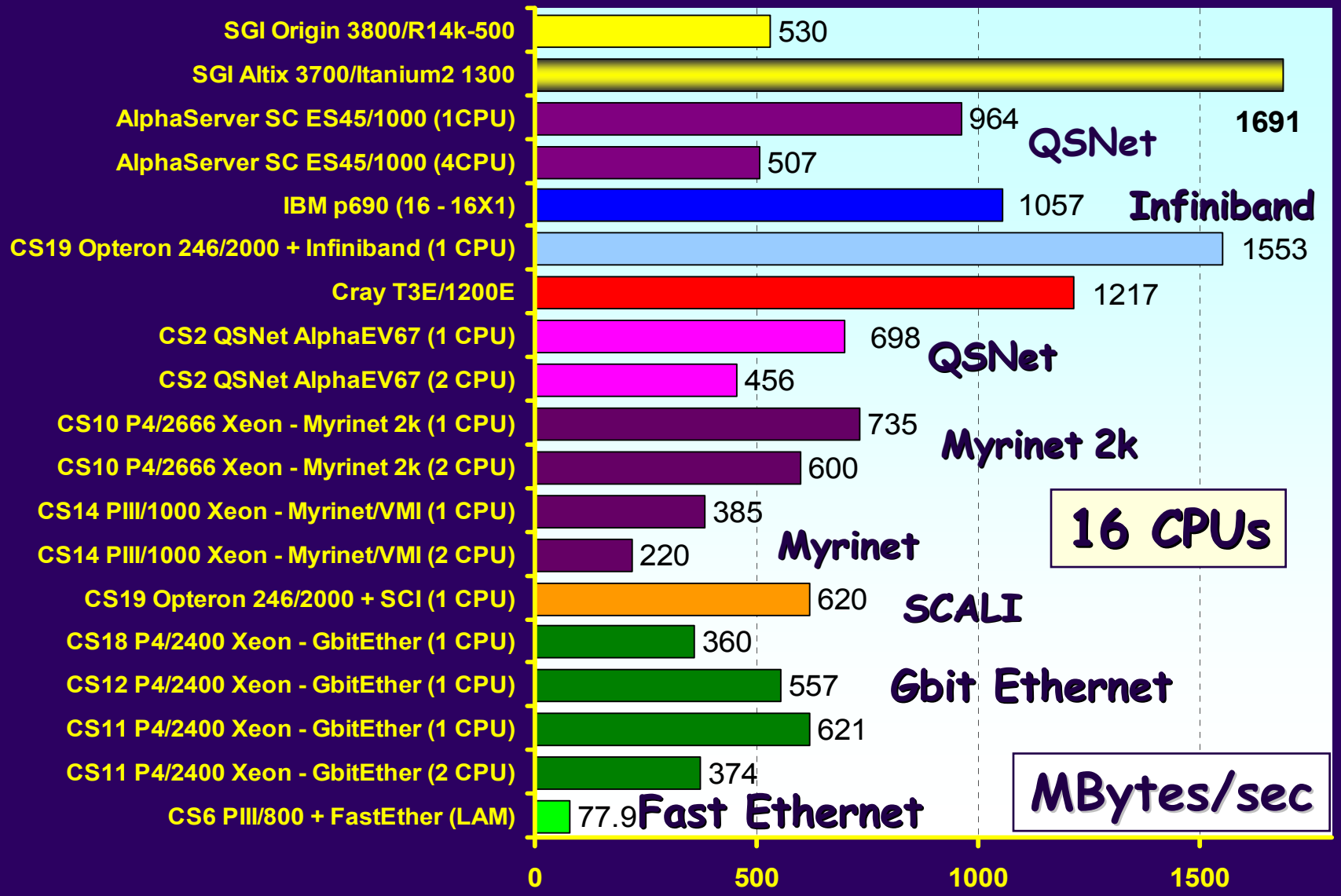
High-end Systems

- Cray T3E/1200E
- SGI Origin3800/R14k-500 (Teras)
- IBM SP/WH2-375 (“trailblazer” switch and SP/NH2-375 (single-plane colony)
- **IBM p690 (8-way LPAR, HPCx - colony “Phase 1”)**
- **IBM p690+ (32-way SMP, HPCx - HPS “”Phase2”)**
- Compaq AlphaServer SC ES45/1GHz -TCS1
- SGI Altix3700/Itanium2-1.3GHz (Newton) and 1.5 GHz (Ram)

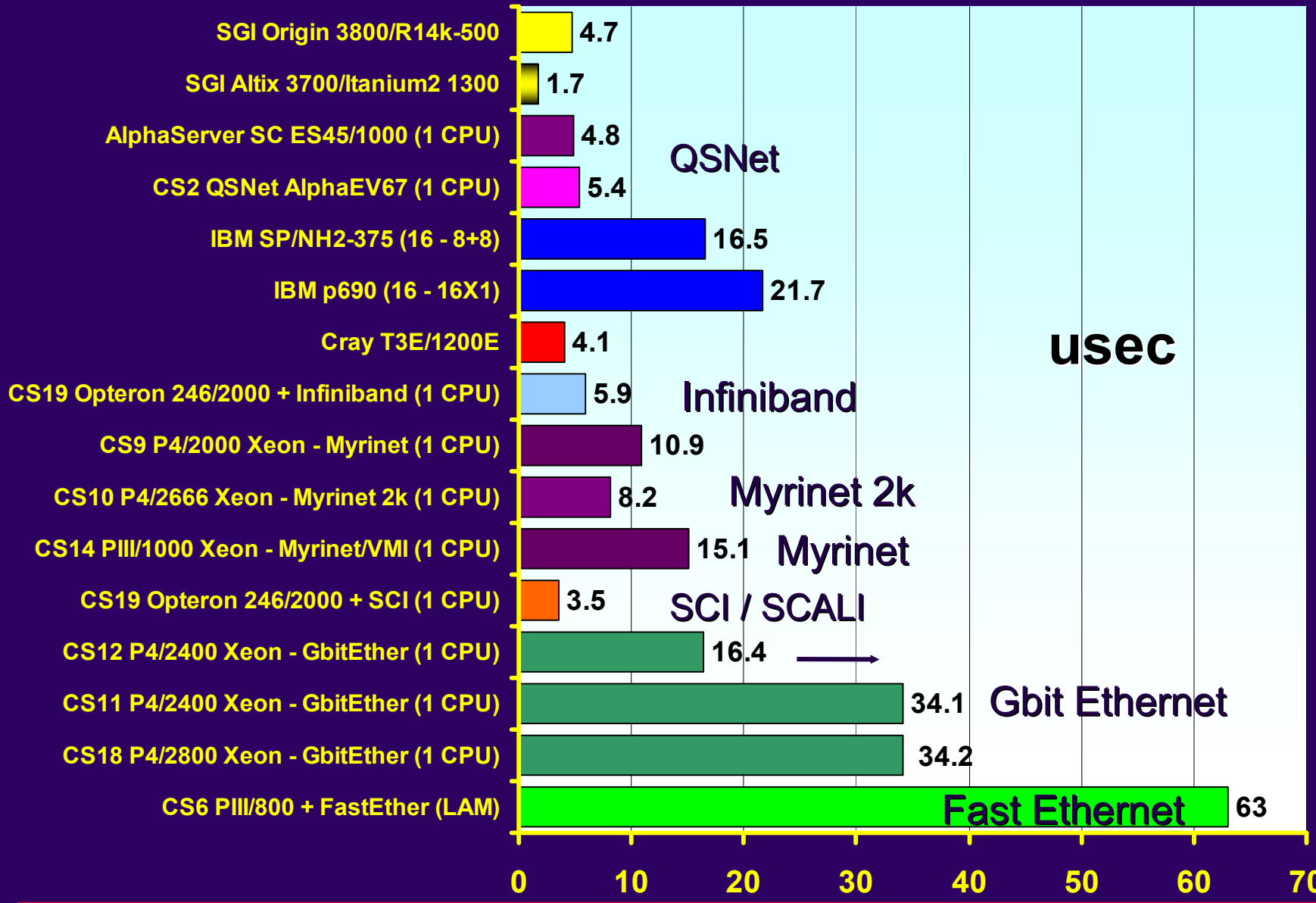
Commodity-based Systems

- CS2 Alpha Linux Cluster dual UP2000/667
- CS8 Itanium/800 + Myrinet 2k (NCSA)
- CS9 dual P4/2000 Xeon + Myrinet 2k
- CS10 P4/2666 + Myrinet 2k (Streamline, Hull U.)
- CS12 P4/2400 + GbitEther (Streamline, Essex U.)
- CS13 P4/2200-2400 + Myrinet 2k (Streamline, Leeds U.)
- CS14 PIII/1000 + Myrinet (VMI, NCSA)
- CS16 Itanium2 1300 + Myrinet2k (SDSC)
- CS18 P4/2800 + Myrinet 2k, GBitE (Clustervision, Bradford U.)
- CS19 Opteron 246/2000 + Infiniband, SCI, GbitEther (OCF, Daresbury).
- CS20 Opteron 248/2200 + Myrinet, (Streamline, RAL).

Interconnect Benchmark - EFF_BW



Interconnect Benchmark - Latency



Collective Operations

(Time - usec) - as function of no.of CPUs

Communications among groups of processors in the cluster e.g.

- Distribute data to all nodes in the cluster (scatter, all-to-all)
- Collect data from all nodes (gather)
- Collect information from all nodes to determine an overall condition e.g. minima or maxima (reduce).

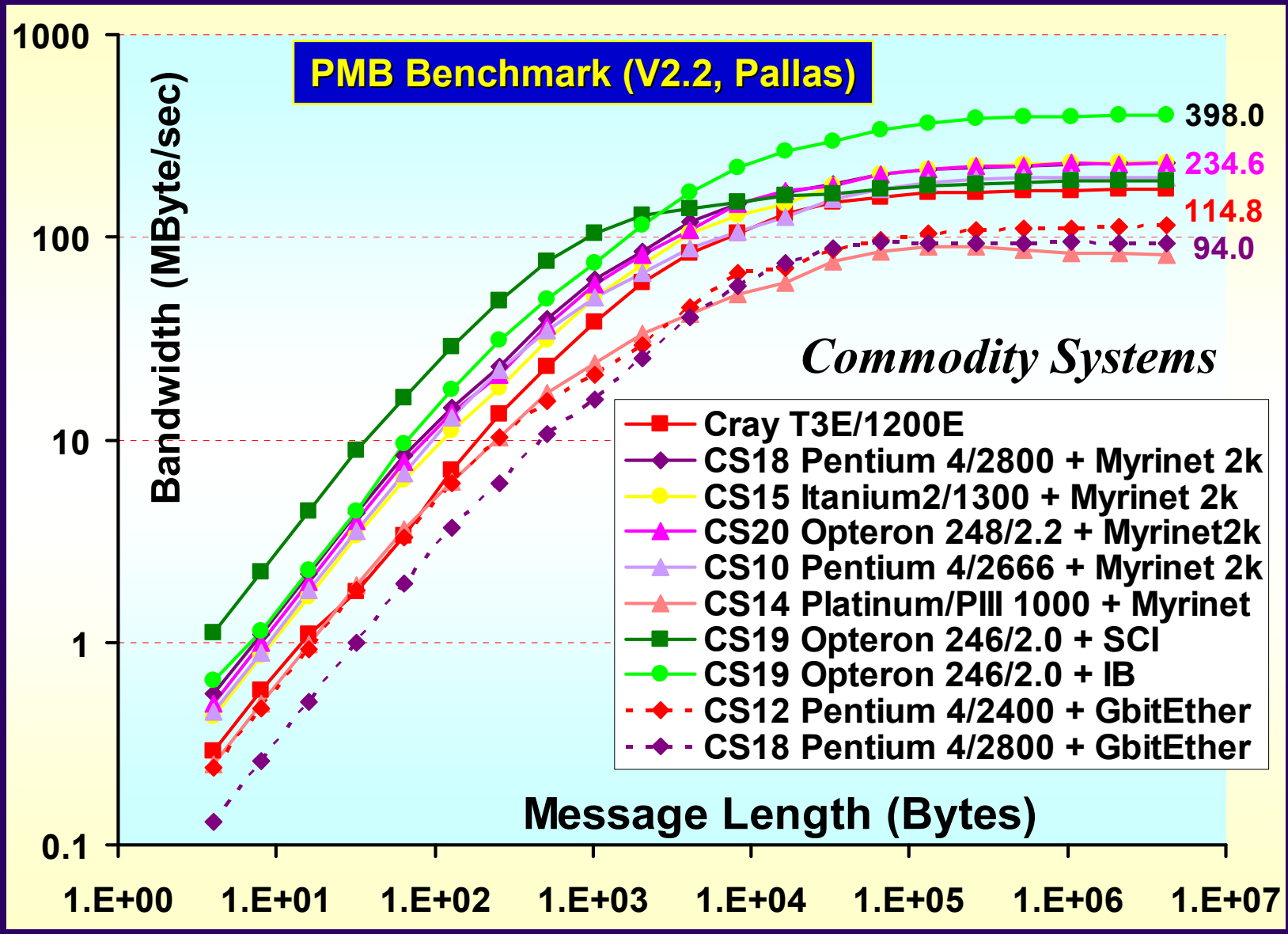
Proces 0	Proces 1	Proces 2	Proces 3	Function Used	Proces 0	Proces 1	Process 2	Proces 3
a,b,c,d	e,f,g,h	i,j,k,l	m,n,o,p	MPI_Alltoall	a,e, i,m	b,f, j,n	c,g, k,o	d,h, l,p
Send Buffer	Send Buffer	Send Buffer	Send Buffer		Receive Buffer	Receive Buffer	Receive Buffer	Receive Buffer

Alltoall

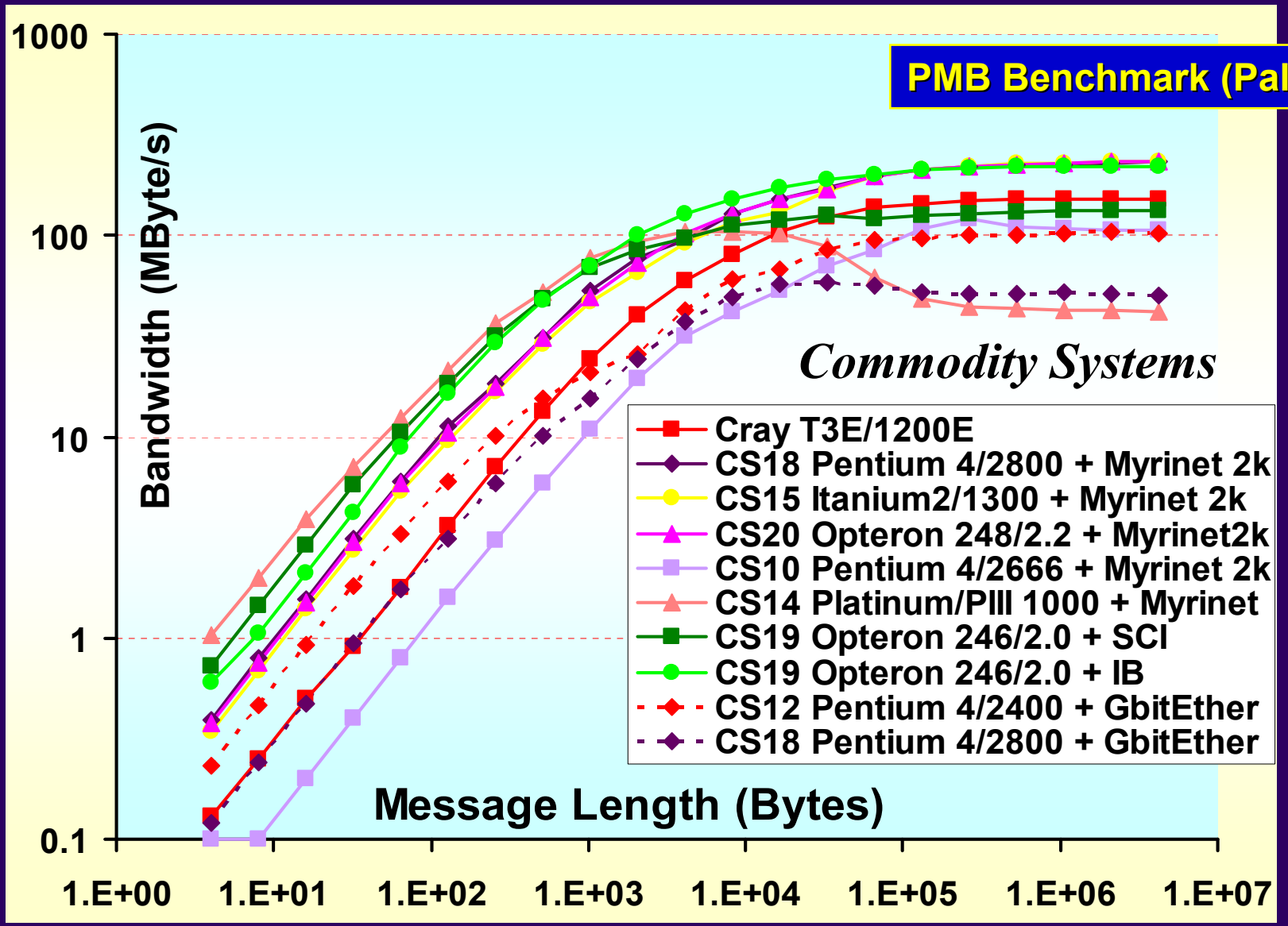
Process 1	Process 2	Process 3	Function Used	Process 1	Process 2	Process 3
a	b	c	MPI, Allgather	a,b,c	a,b,c	a,b,c
Send Buffer	Send Buffer	Send Buffer		Receive Buffer	Receive Buffer	Receive Buffer

Allgather

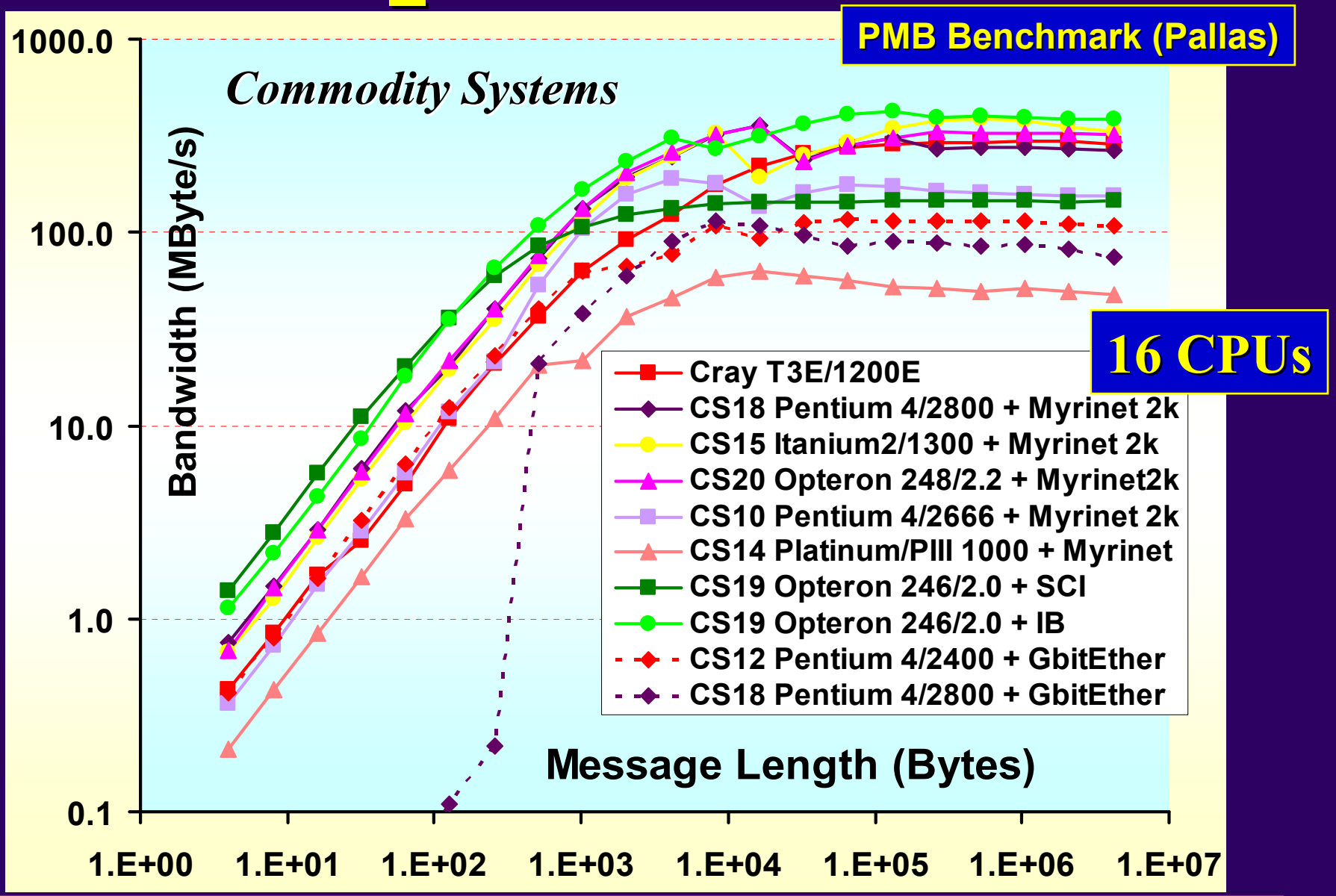
PingPong Performance



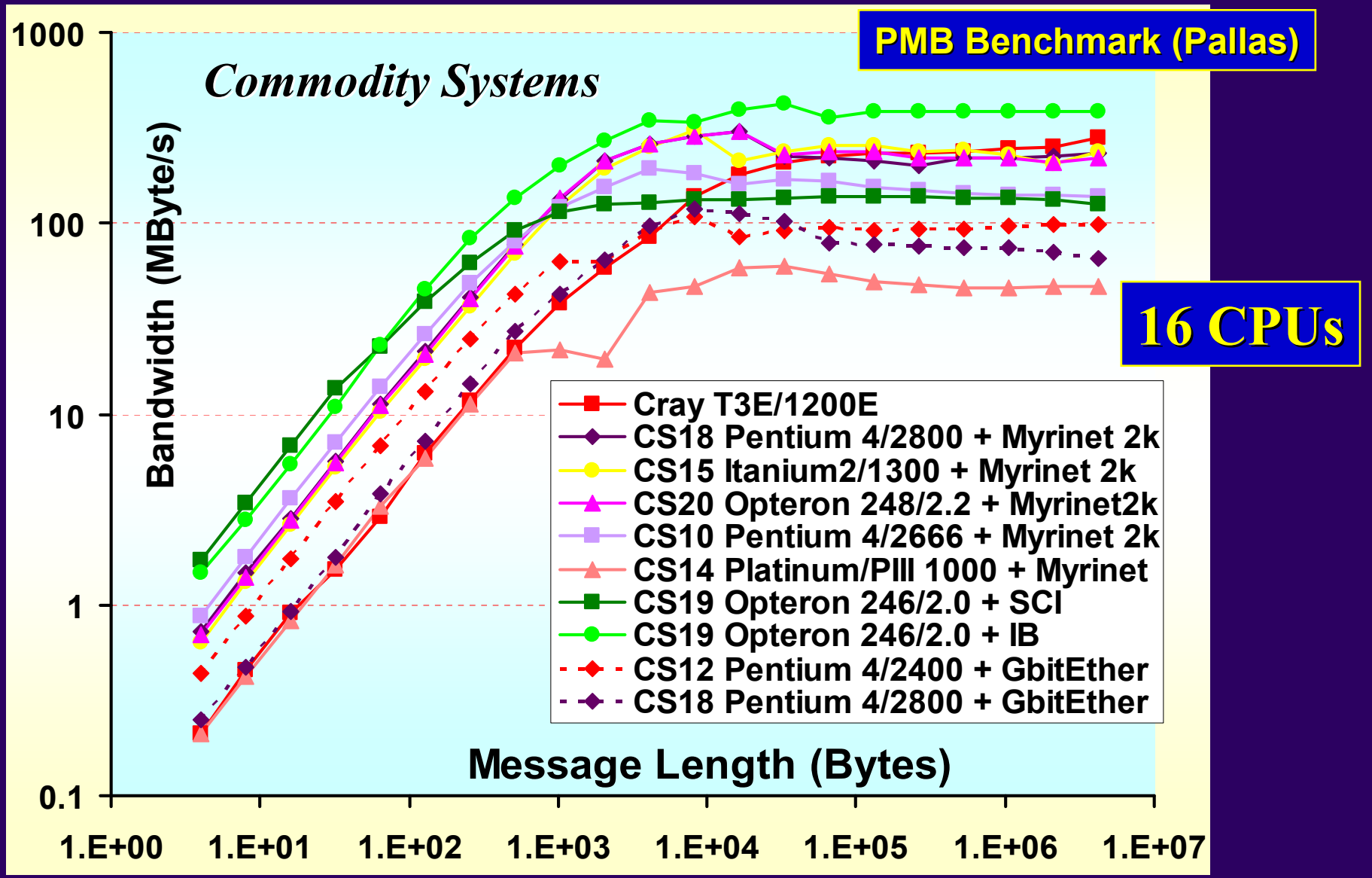
MPI_PingPing Performance



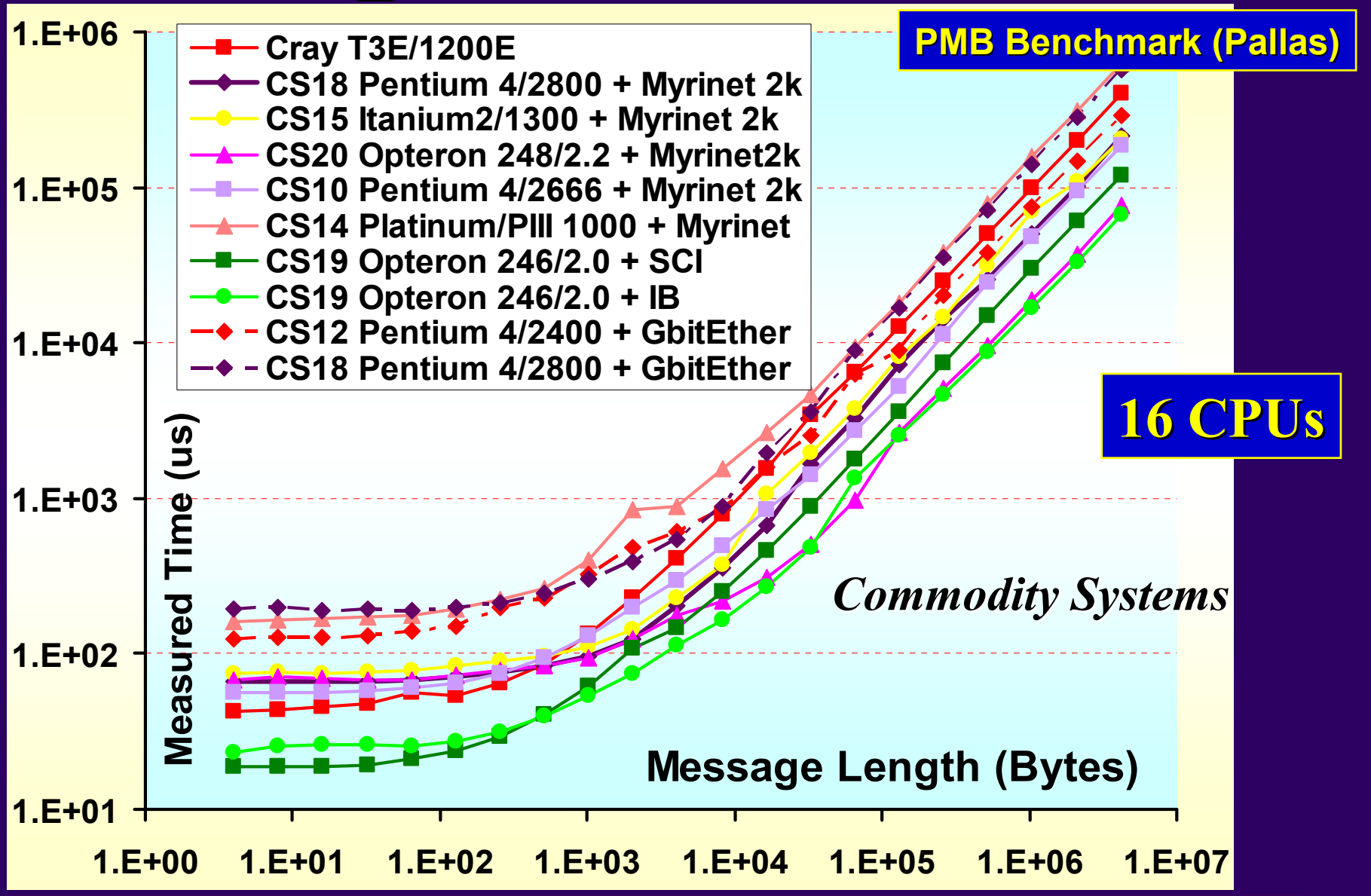
MPI_Sendrcv Performance



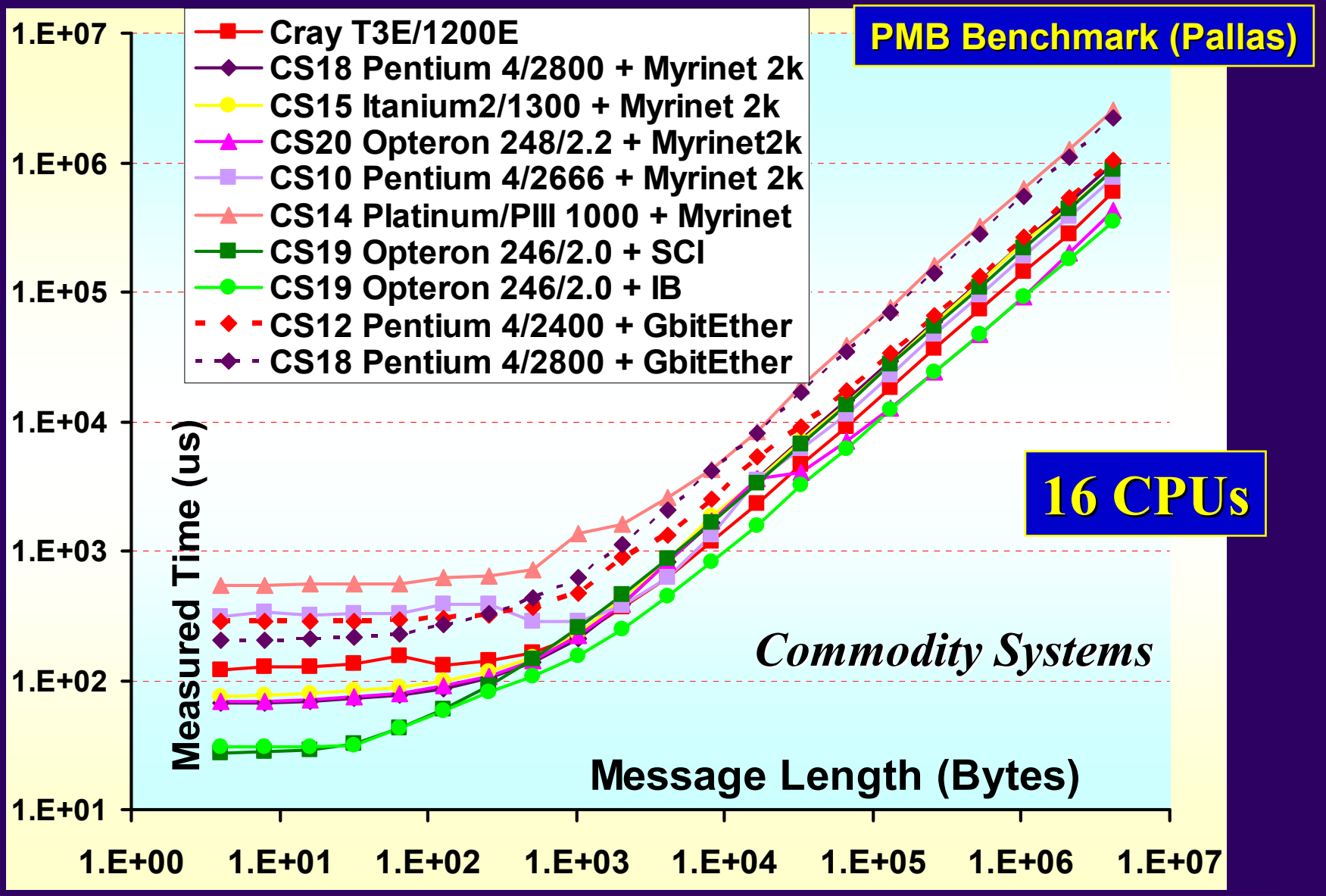
MPI_Exchange Performance



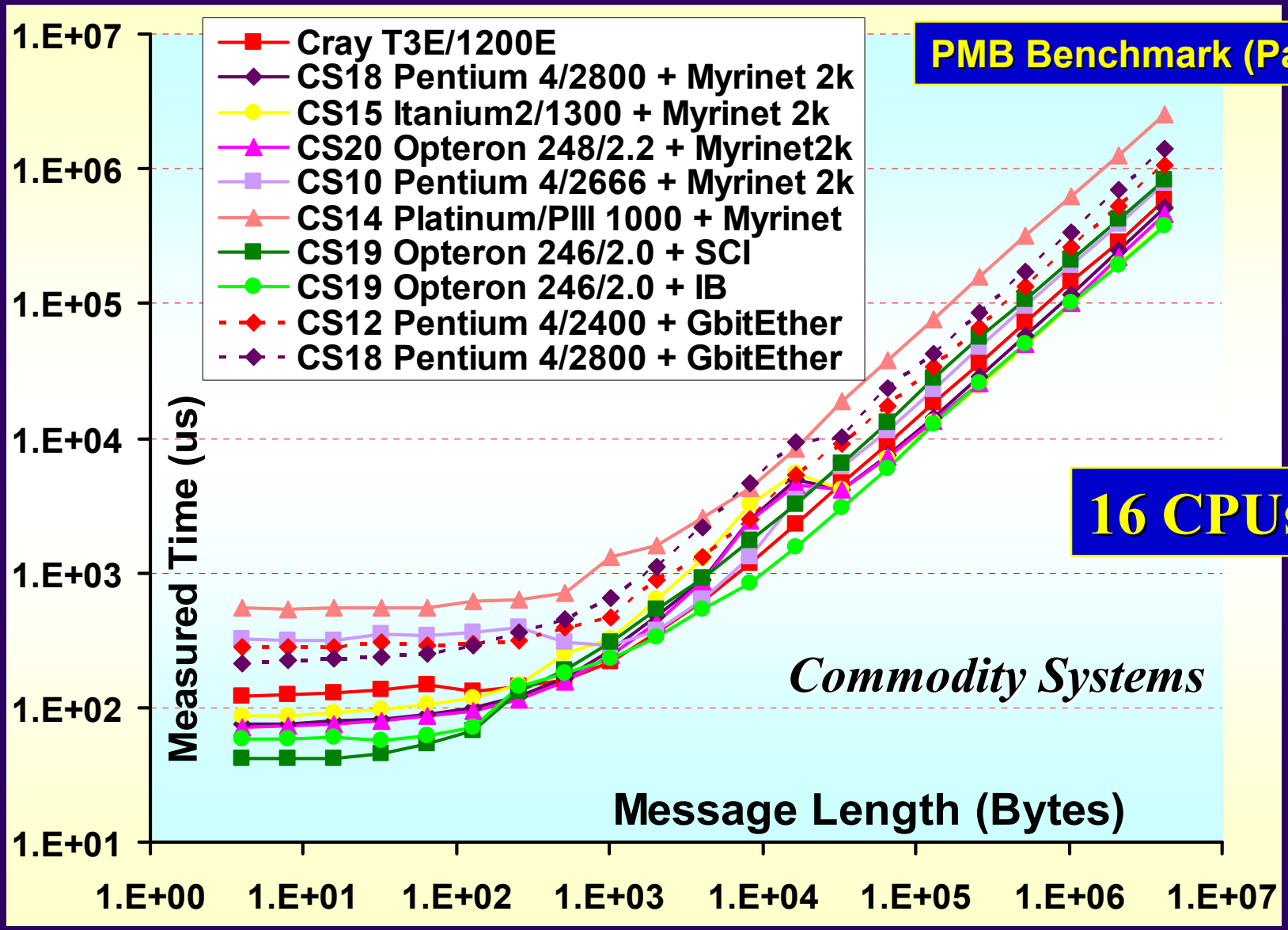
MPI_allreduce Performance



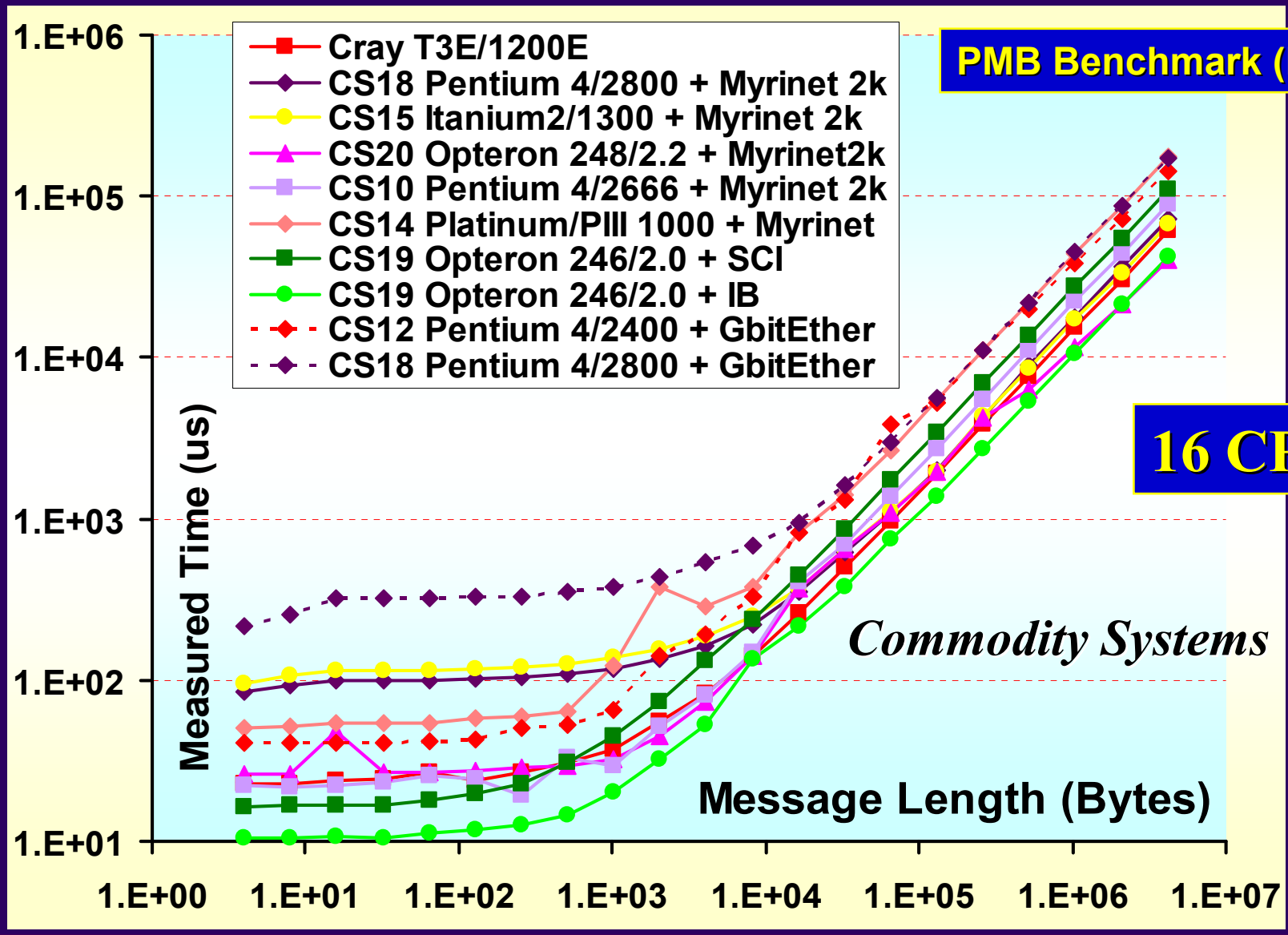
MPI_Allgather Performance



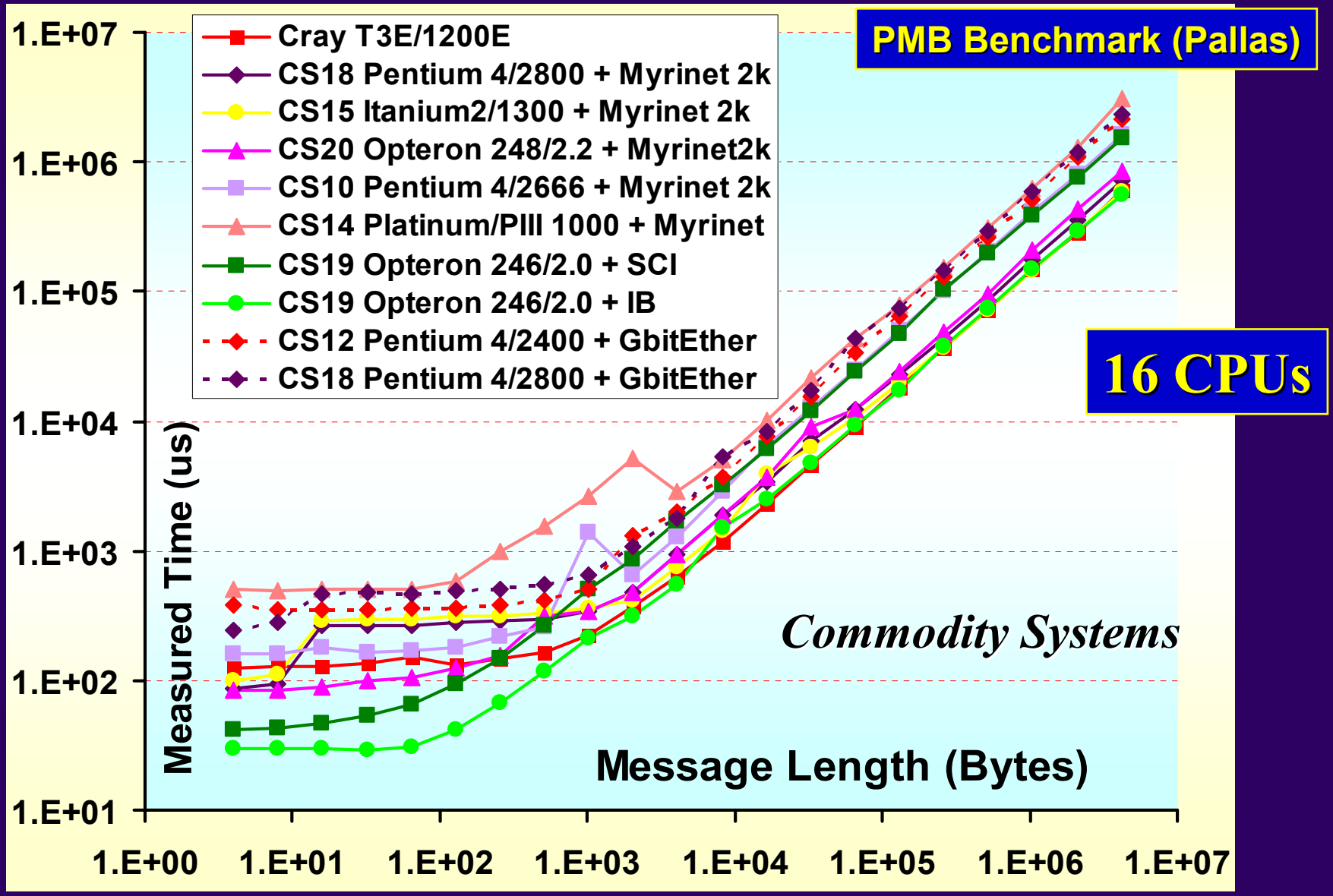
MPI_Allgatherv Performance



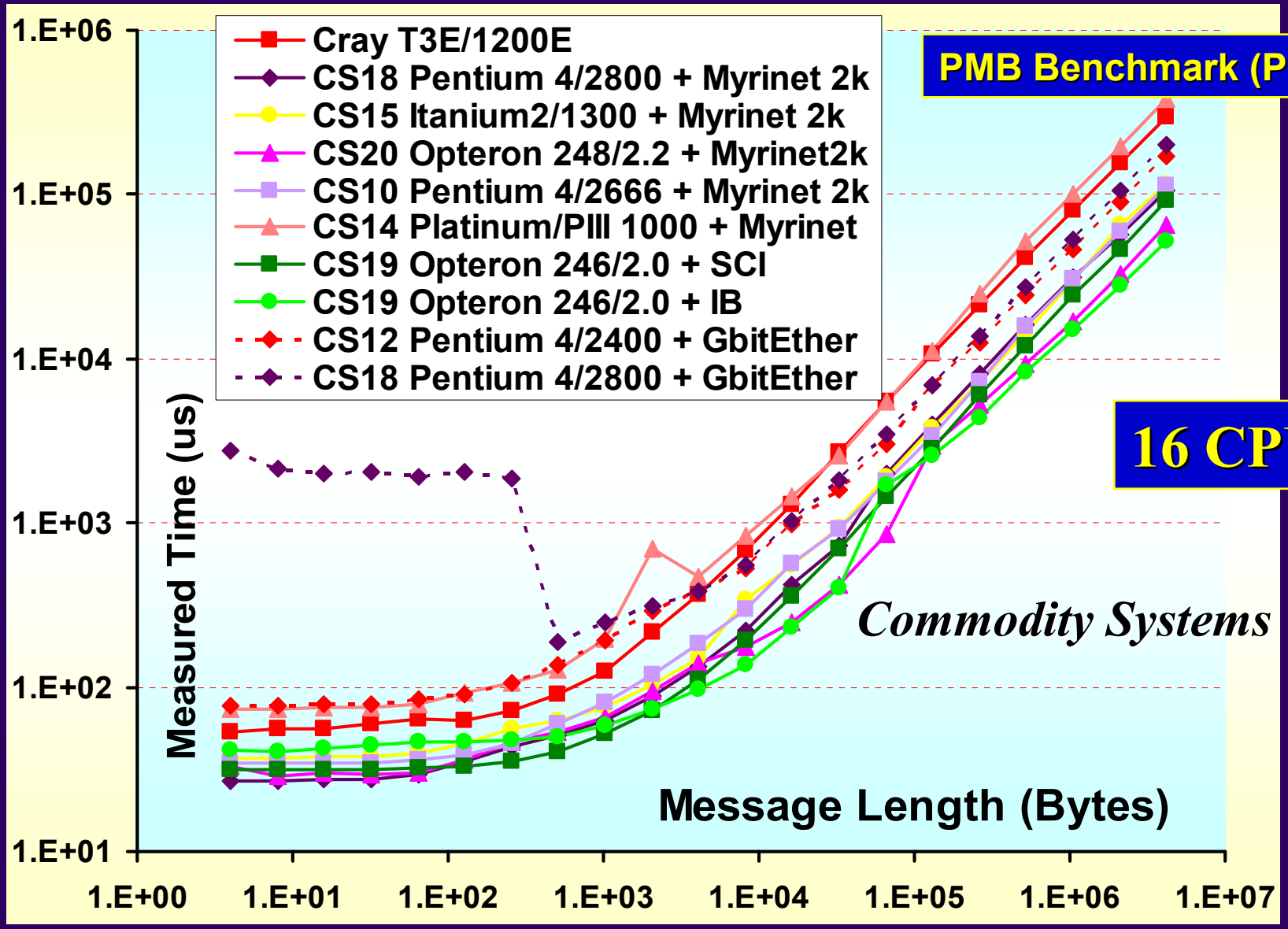
MPI_Bcast Performance



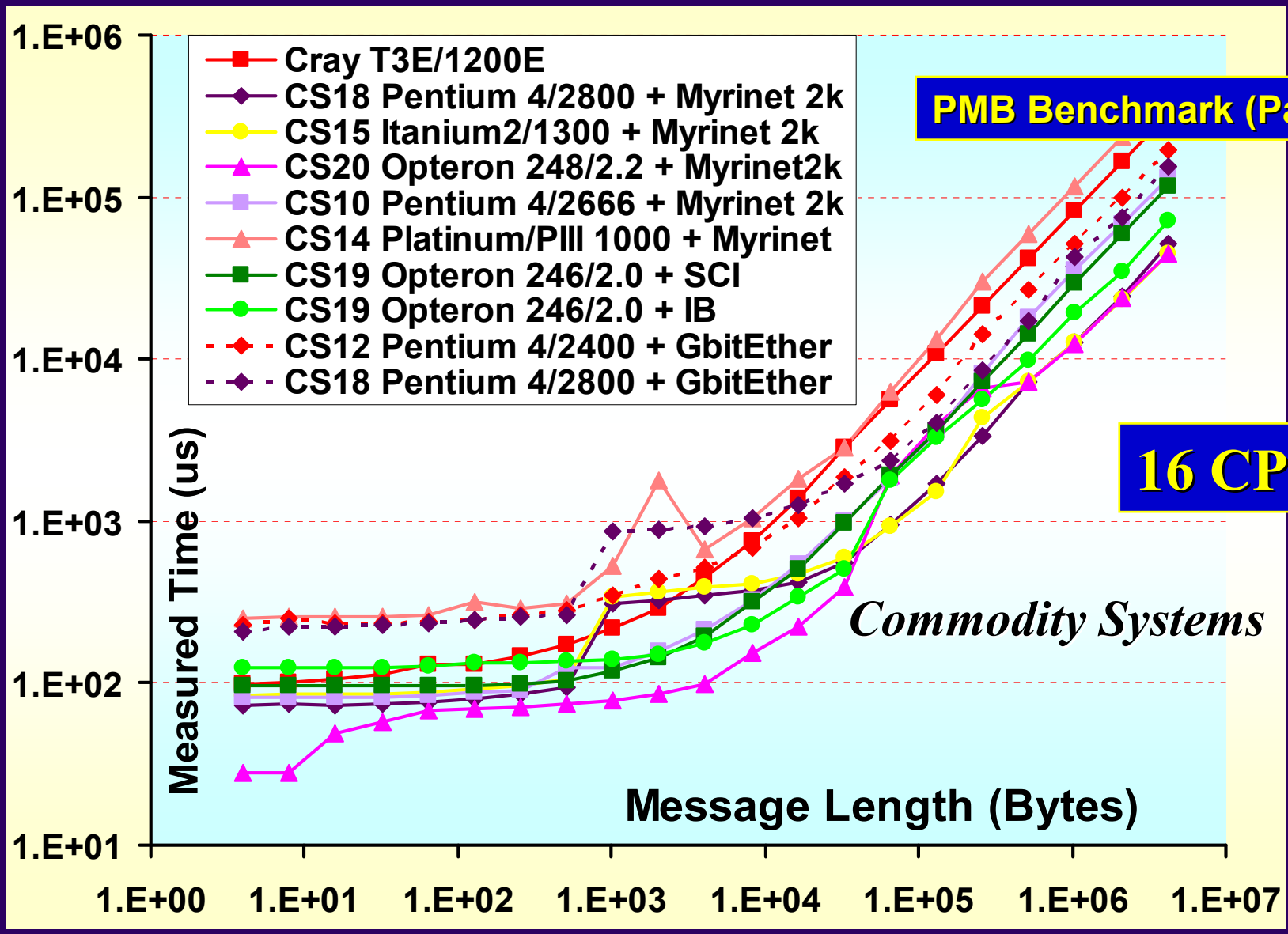
MPI_All to All Performance



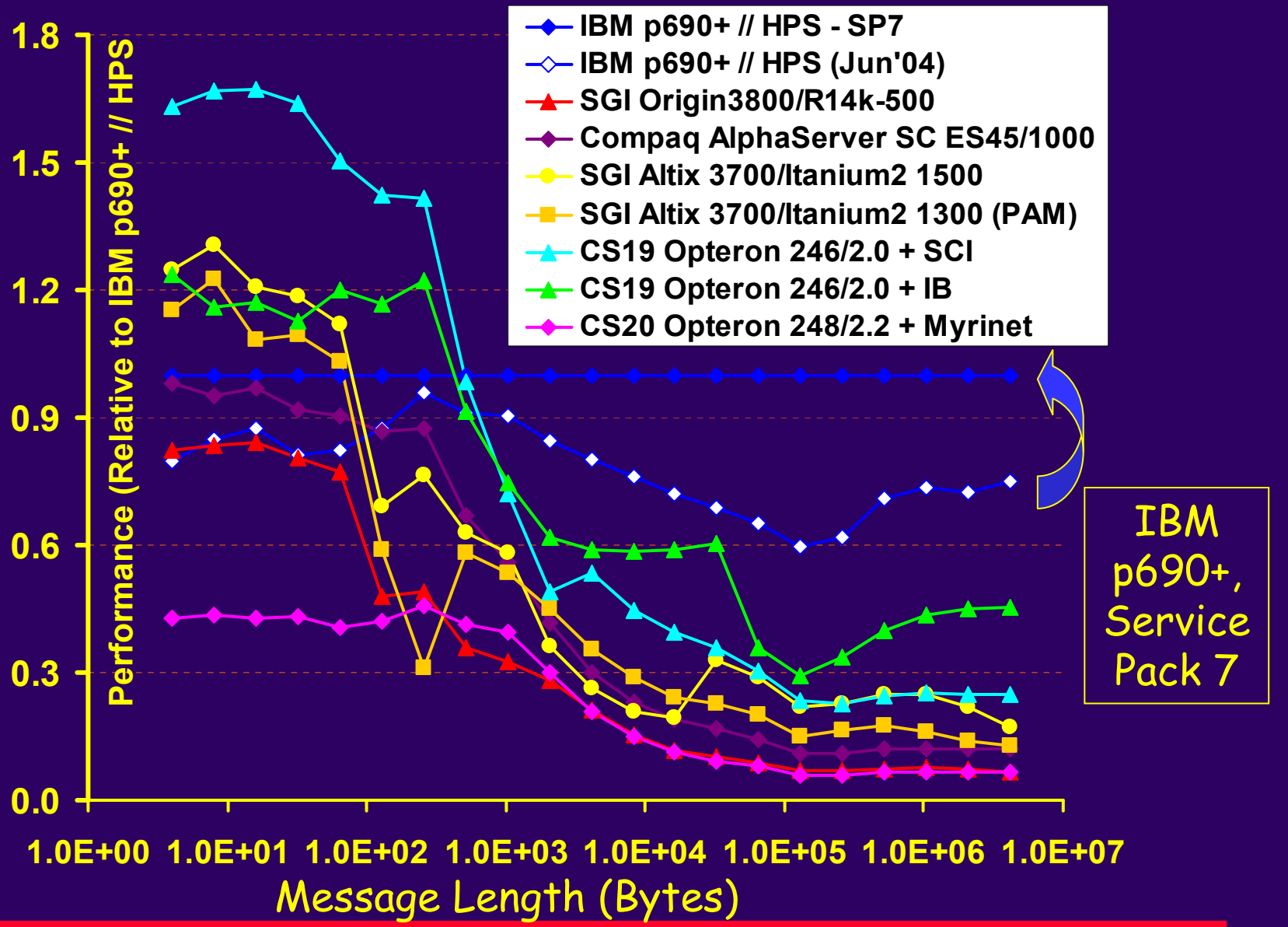
MPI_Reduce Performance



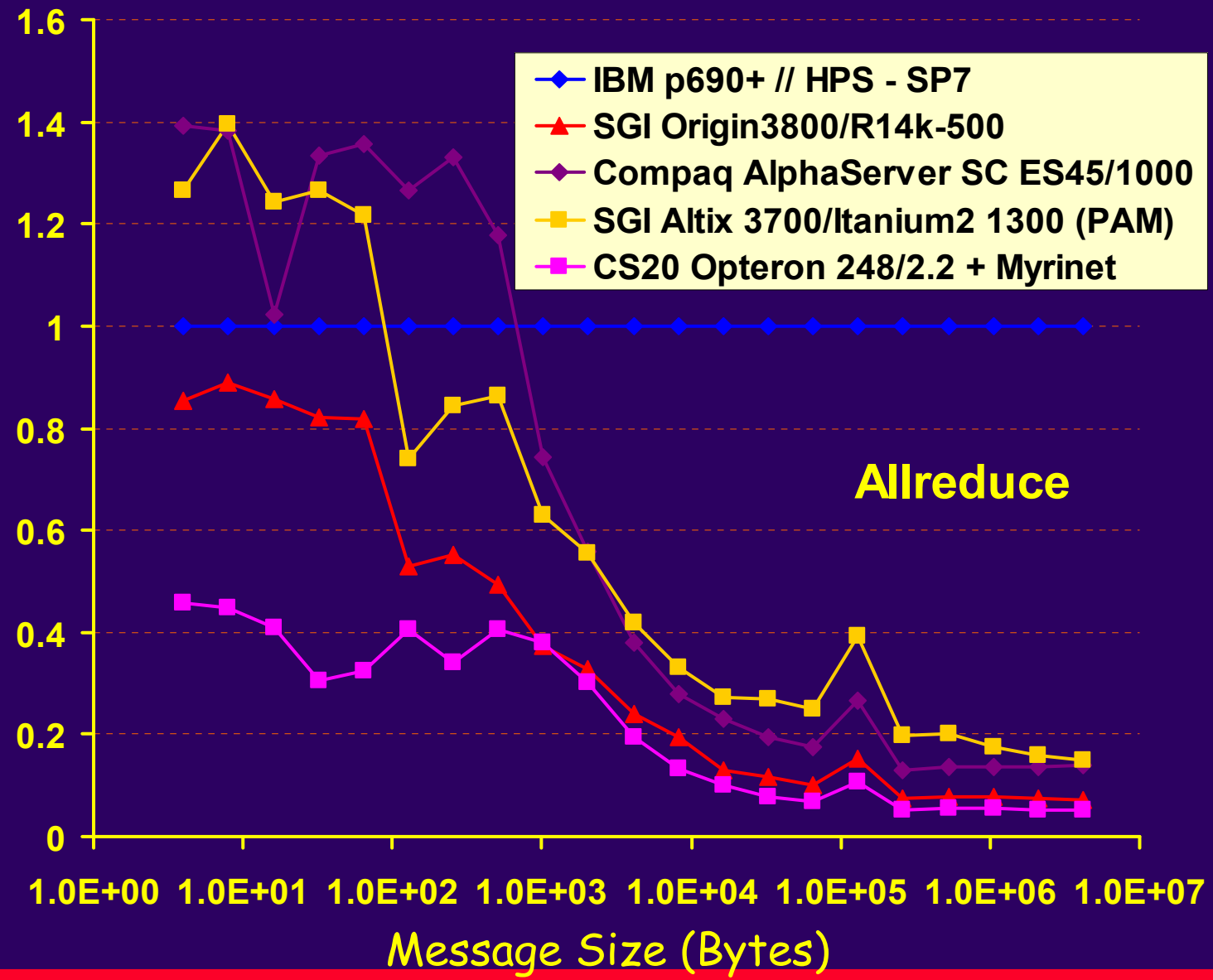
MPI_Reduce_scatter Performance



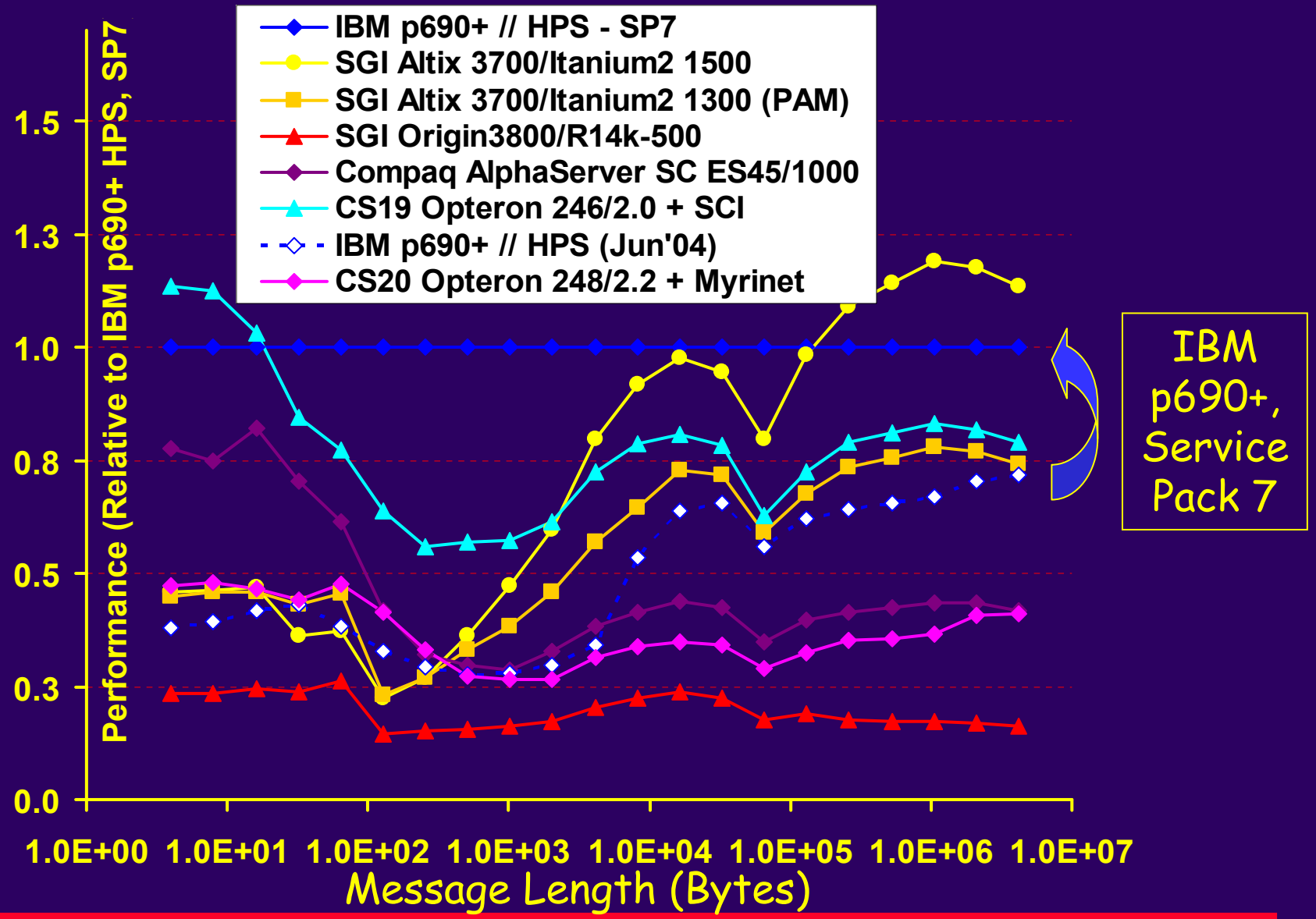
64-CPU Relative Performance for Allreduce



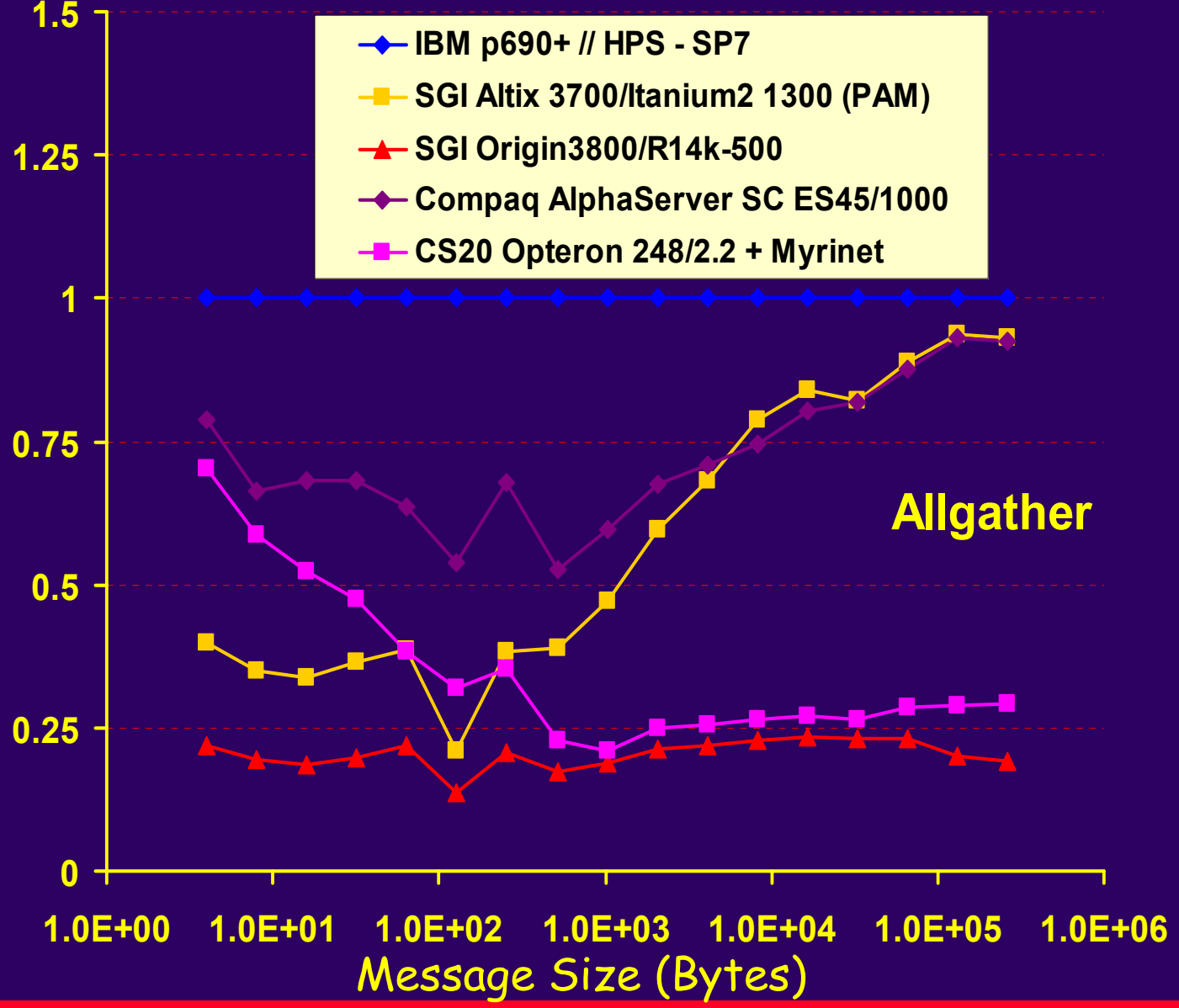
128-CPU Relative Performance for Allreduce



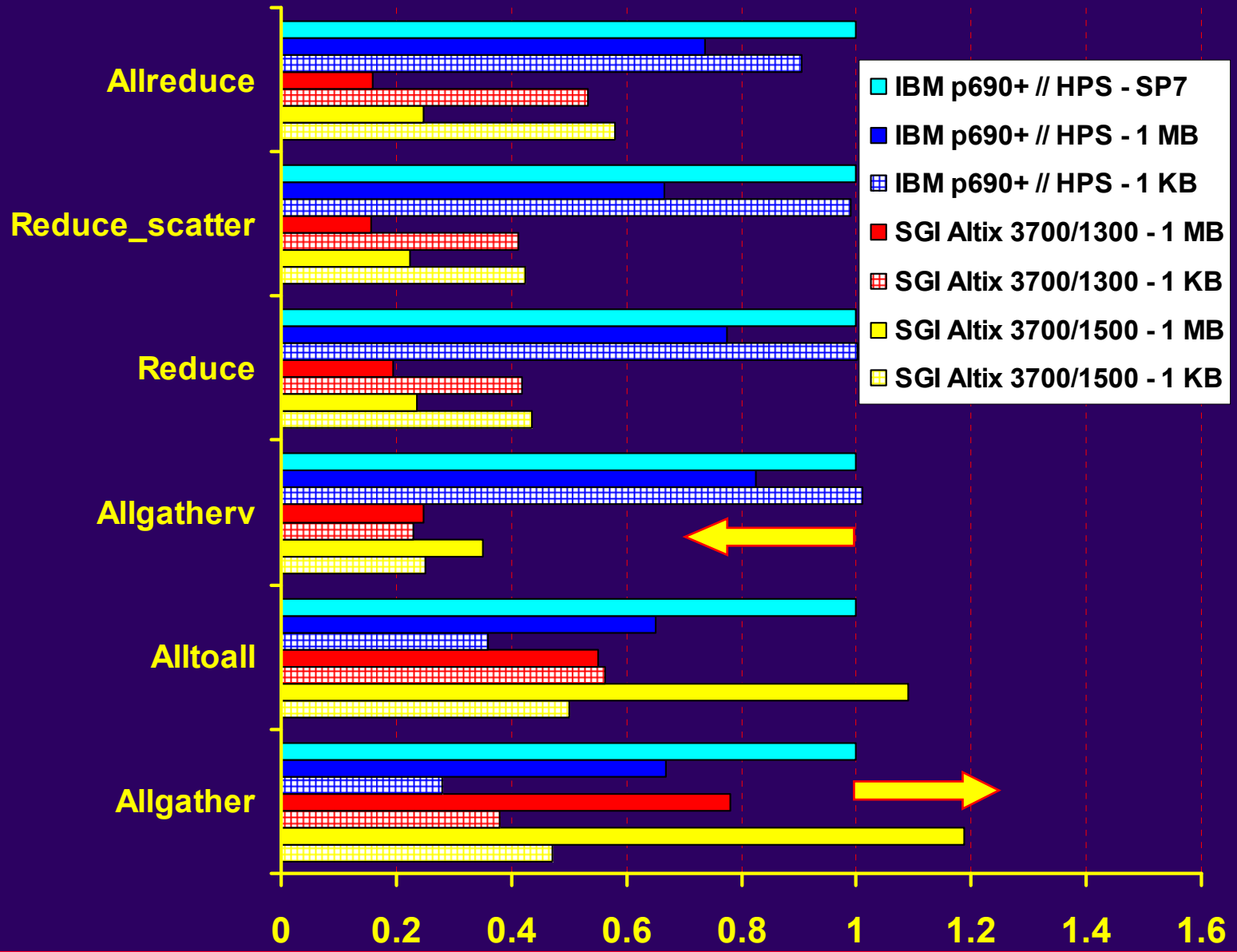
64-CPU Relative Performance for Allgather



128-CPU Relative Performance for Allgather

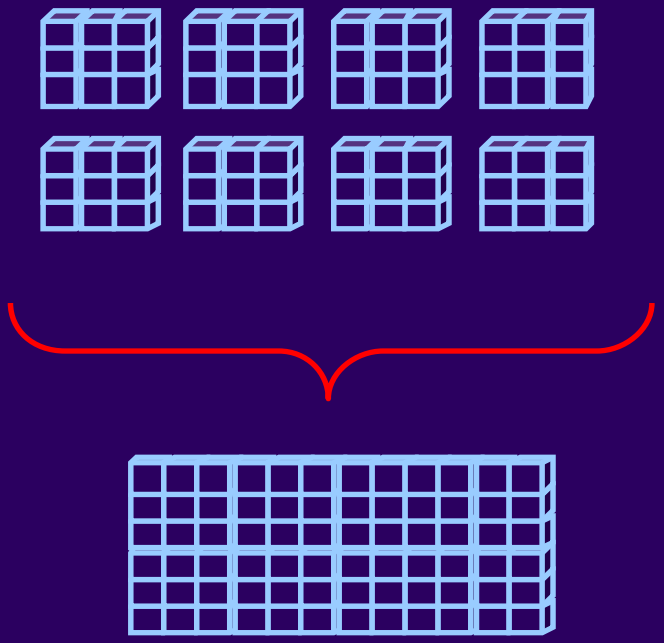


64-CPU Performance for all collective operations



Global Arrays

Physically distributed data

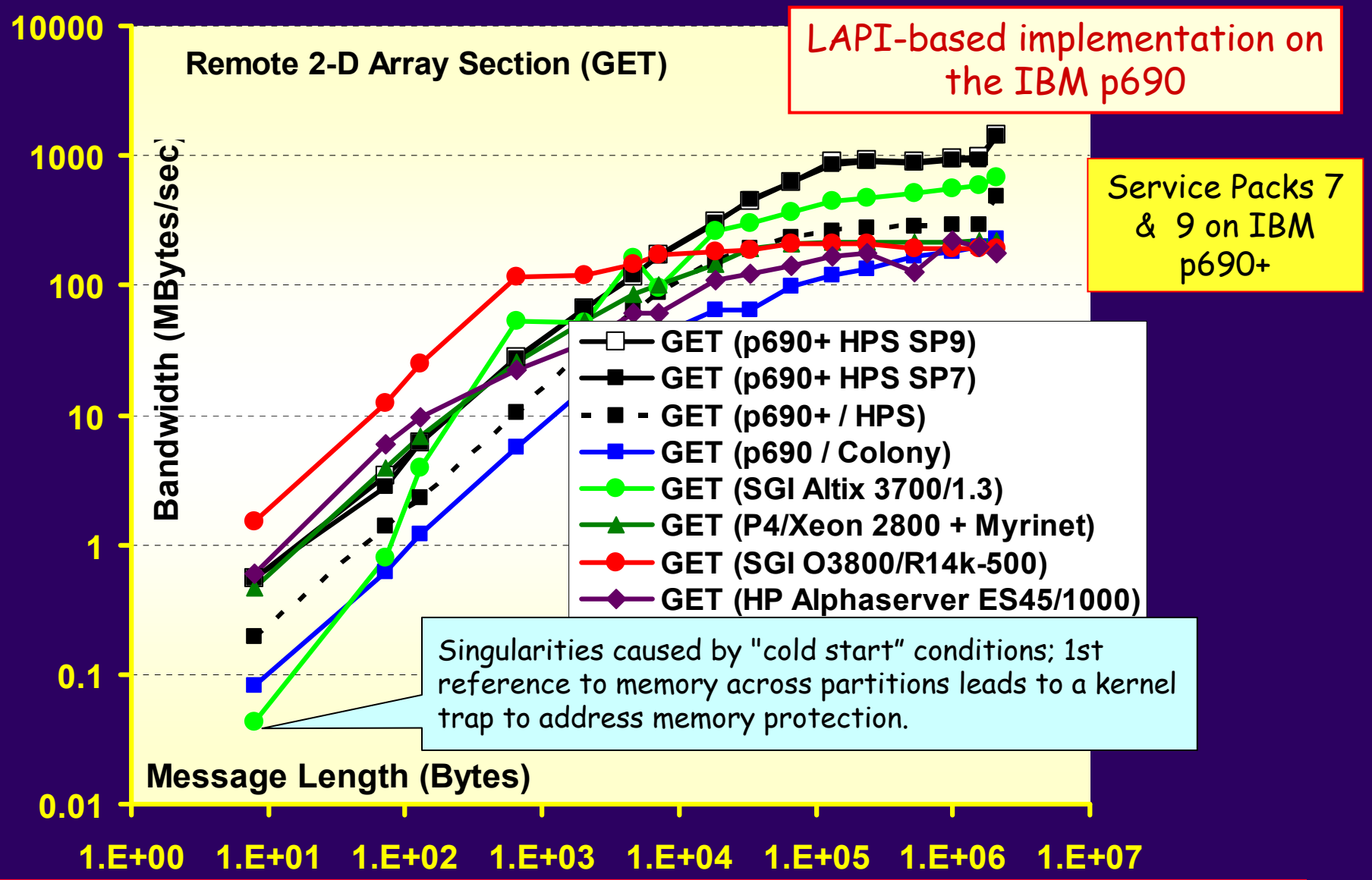


Single, shared data structure

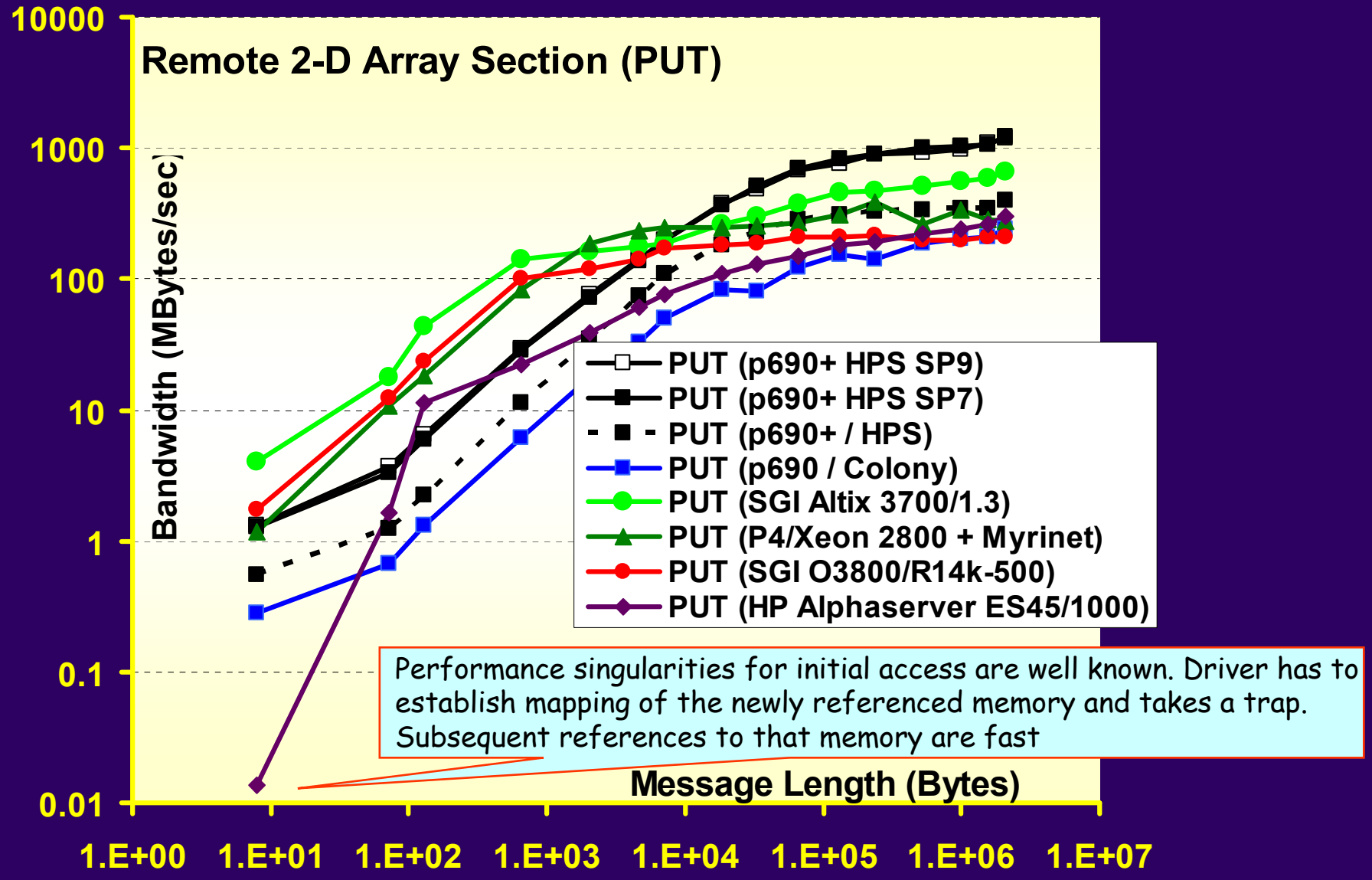
- Shared-memory-like model
  - Fast local access
  - NUMA aware and easy to use
  - MIMD and data-parallel modes
  - Inter-operates with MPI, …
- BLAS and linear algebra interface
- Ported to major parallel machines
  - IBM, Cray, SGI, dusters,...
- Originated in an HPCC project
- Used by 5 major chemistry codes, financial futures forecasting, astrophysics, computer graphics

Tools developed as part of the NWChem project at PNNL; R.J. Harrison, J. Nieplocha et al.

Global Array Benchmark I. GET



Global Array Benchmark II. PUT



Global Array Benchmark III. ACCUMULATE