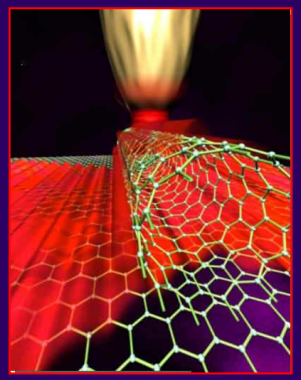
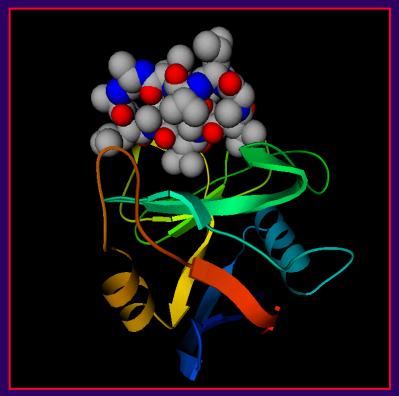
Application Performance on High-end and Commodity-class Computers



Martyn F. Guest
CCLRC Daresbury Laboratory

m.f.guest@daresbury.ac.uk



Acknowledgements:

P. Sherwood, H.J. van Dam, I.J. Bush, DL
J. Nieplocha, E. Apra, PNNL

http://www.cse.clrc.ac.uk/disco/hw-perf.shtml

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
- Application performance
  - Distributed vs. Replicated Data
  - Molecular Simulation (DLPOLY, DLMULTI and CHARMM)
  - Electronic Structure - Global Arrays (GAs) and Linear Algebra (PeIGS)
    - NWChem and GAMESS-UK
  - Materials Simulation (CPMD) and Computational Engineering (ANGUS, SBLI)
- Summary

Capability and Capacity Computing - Cost and Performance

Specification	Usage	Cost Units	CPU	Memory	I/O
MPP/ASCI 1600CPUs, IBM p690+ (1.5 TB)	HPC community	10,000	2,000	3,000	200-300
	Capability	**Commodity Systems (1.5 x N ??)**			
SMP 16-processor SGI Altix 350, Itanium 2 (10GB RAM)	Department	100	20	25	20-30
	Capacity	**Commodity Systems (1.5 x N)**			
PC Pentium -4 Xeon / 3.4GHz (512 MByte , 30 GB)	Desktop	1	1	1	1

3 year continual access to 32 CPUs:
High-end (£0.5 / CPU hour) : £420,000
in-house commodity system : £50,000

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

Commodity Systems (CSx). Evaluation Hardware

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")

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)***

Applications Performance Overview

- ***Serial (SPEC, DL) & Communication Benchmarks***
- ***Parallel Applications Performance***

1. Computational Chemistry: Molecular Simulation & Electronic Structure
2. Computational Materials Science
3. Atomic & Molecular Physics
4. Computational Engineering
5. Environmental Modelling

- Capacity-based group solution
- Issues of Cost effectiveness
- On e.g. 128-256 CPU cluster, modal job size is ~ 32 CPUs
- Increasing trend to hierarchical clusters - Gbit network with HEC core (with e.g. myrinet)

Capability and Capacity Computing
Commodity vs. Proprietary Solutions

Performance Metrics: 1999-2001

Attempted to quantify delivered performance from the Commodity-based systems against MPP (CSAR Cray T3E/1200E) and ASCI-style SMP-node platforms (e.g. SGI Origin 3800) i.e.

Performance Metric (% 32-node Cray T3E)

T (32-nodes Cray T3E/1200E) / T (32 CPUs) CSx

[$T_{32\text{-node}}$ *T3E* / $T_{32\text{-node}}$ *CS1 Pentium III/450 + FE*]

$T_{32\text{-node}}$ **T3E /** $T_{32\text{-node}}$ **CS6 Pentium III/800 + FE**

$T_{32\text{-node}}$ **T3E /** $T_{32\text{-CPU}}$ **CS2 Alpha Linux Cluster + Quadrics**

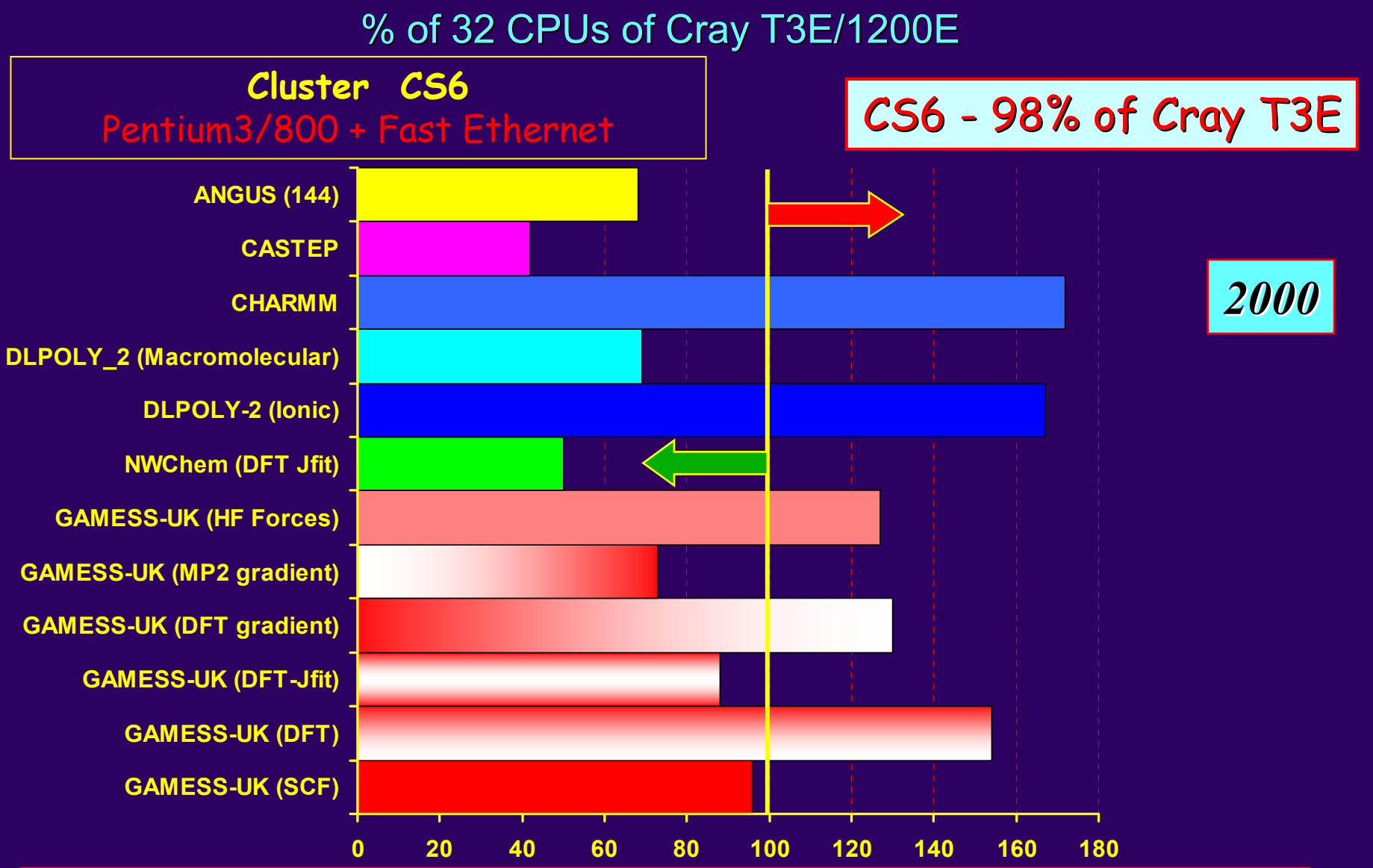
Performance Metrics: 2002

Performance Metric (% 32-node AlphaServer SC [PSC])

T (32-CPUs AlphaServer SC ES45/1000) / T (32 CPUs) CSx

$T_{32\text{-CPU}}$ **AlphaServer ES45 /** $T_{32\text{-CPU}}$ **CS9 Pentium 4 Xeon / 2000 + Myrinet 2k**

Commodity Comparisons with High-end Systems

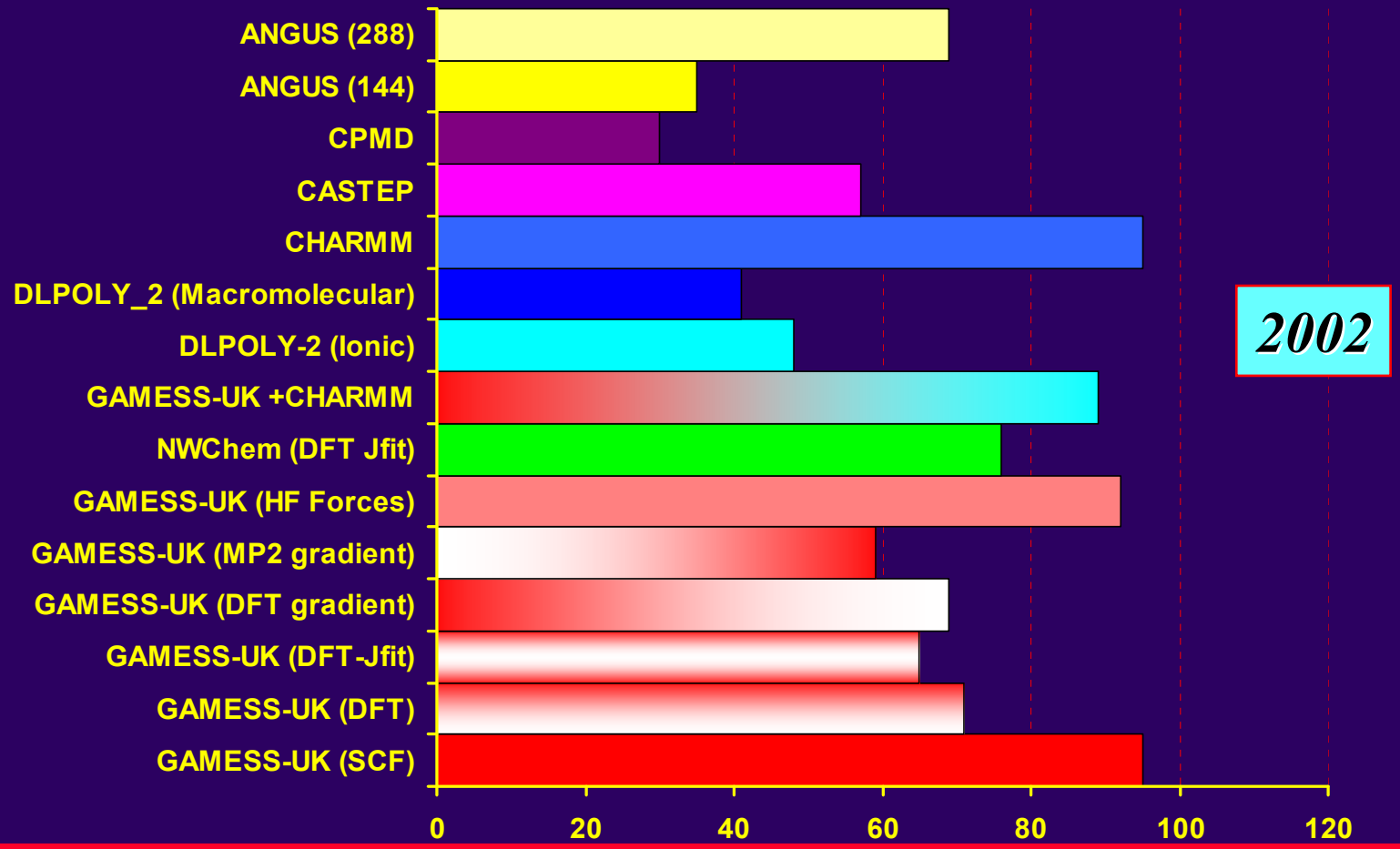


Commodity Comparisons with High-end Systems

% of 32 CPUs of Compaq AlphaServer SC ES45/1000

Cluster CS9
Pentium4/2000 Xeon + Myrinet

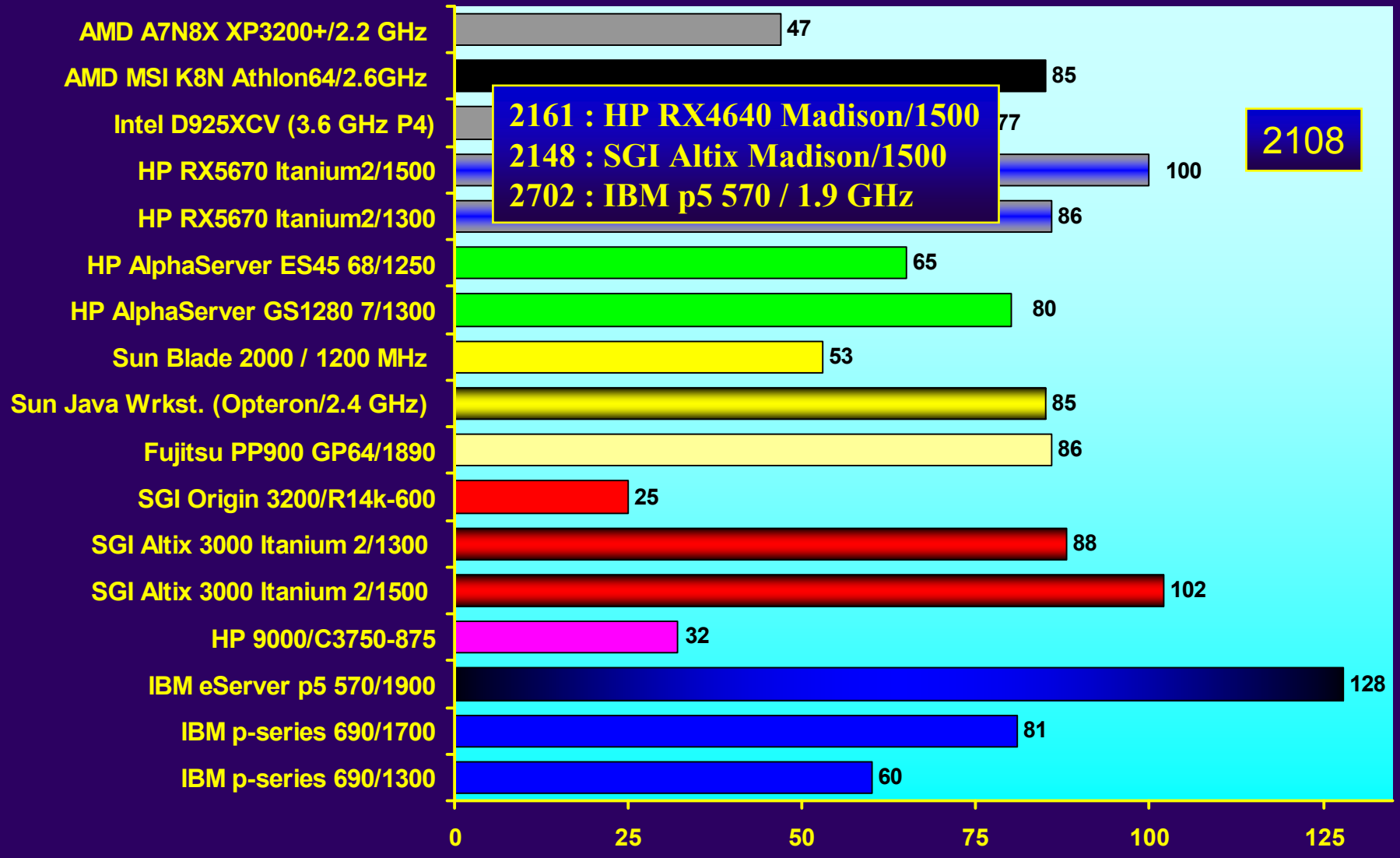
CS9 - 66% of ES45/1000



2002

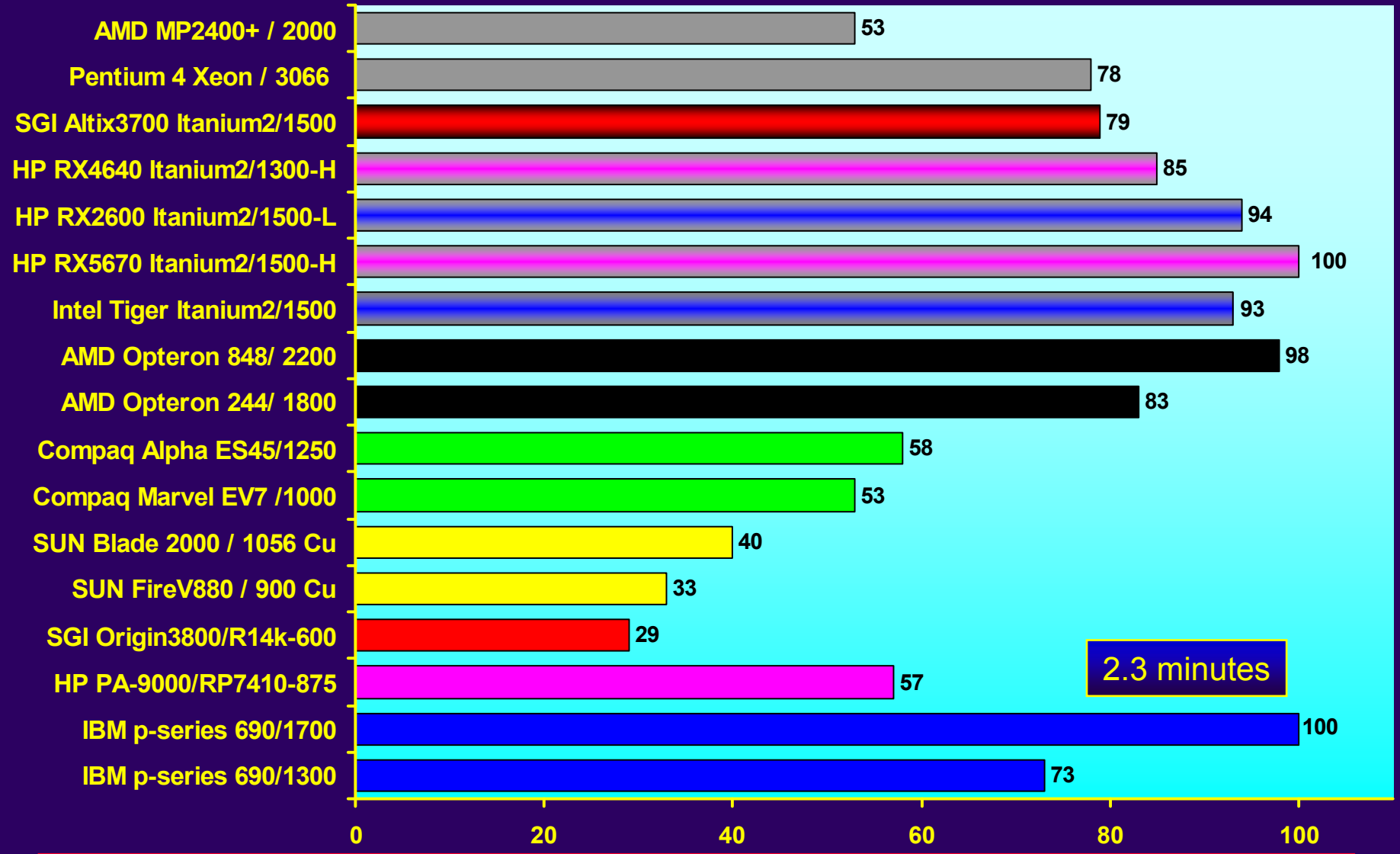
SPEC CPU 2000 - SPECfp2000

Values relative to HP RX5670 Itanium2/1.5GHz



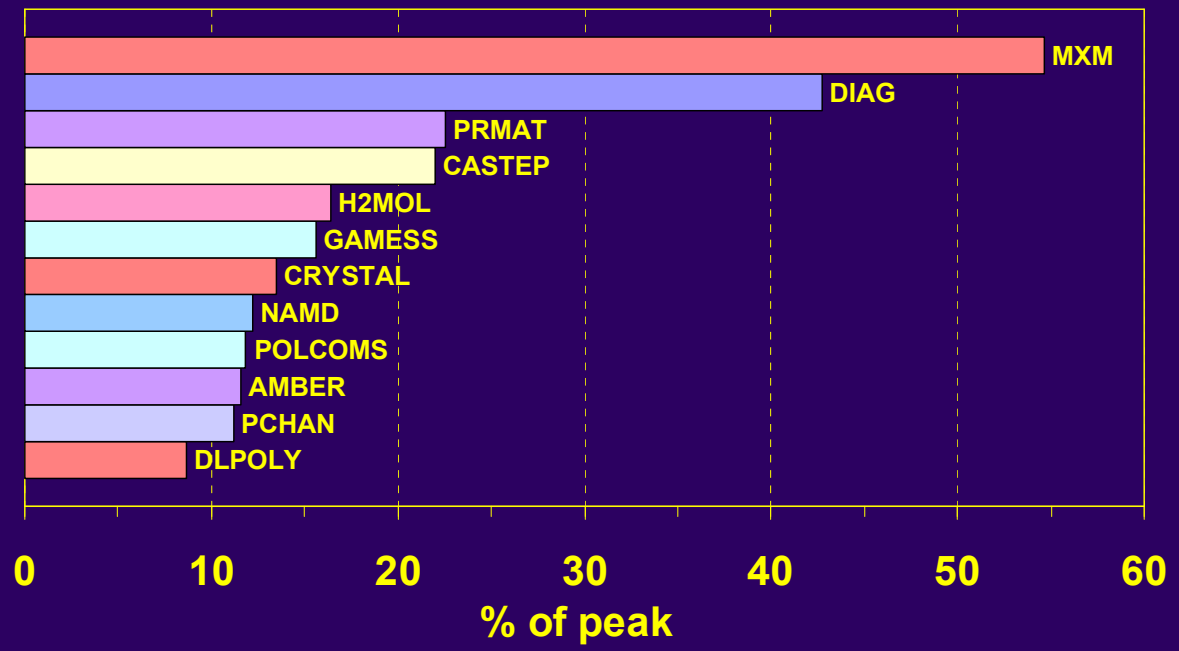
The GAMESS-UK Serial Benchmark

Performance relative to the HP RX5670 Itanium2/1.5GHz



Single CPU performance - A Case Study

- Scalability of Terascale applications is only part of the story …
- Absolute performance also depends on single cpu performance
- % of peak is seen as an important measure
- Comparison with other systems e.g. vector machines



Ran representative test cases on small numbers of IBM p690 processors for applications and some important kernels
Used IBM's *hpmlib* to measure Mflop/s
Other *hpmlib* counters can help to understand performance e.g. memory bandwidth, cache miss rates, FMA count, computational intensity etc.

Interconnects and Networking

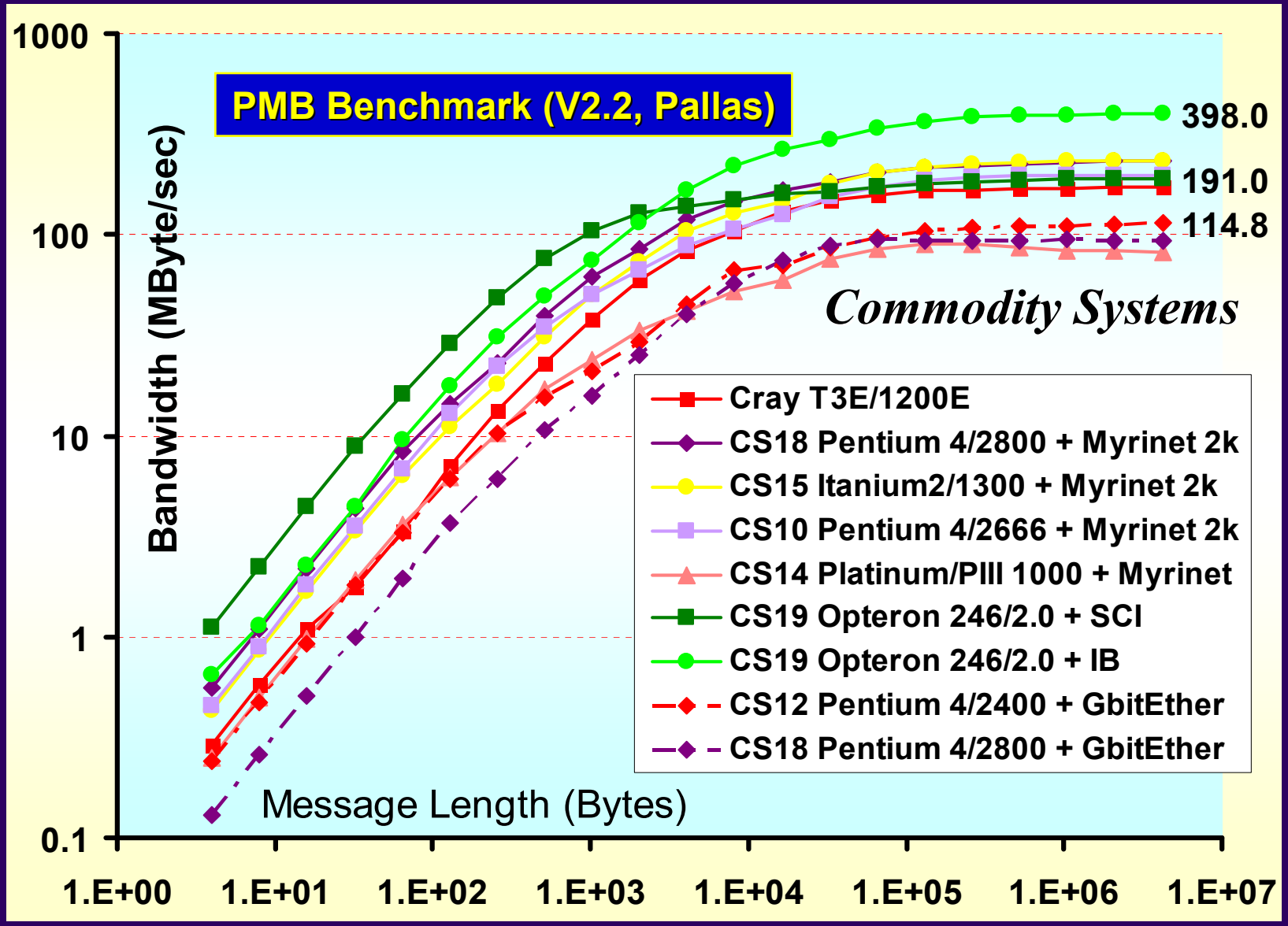
- Ethernet, Gbit etc.
- Myrinet, Quadrics and Dolphin SCI
  - 64 bit PCI implementations, with faster PCI-X options near market
- Infiniband PCI-X interconnect

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

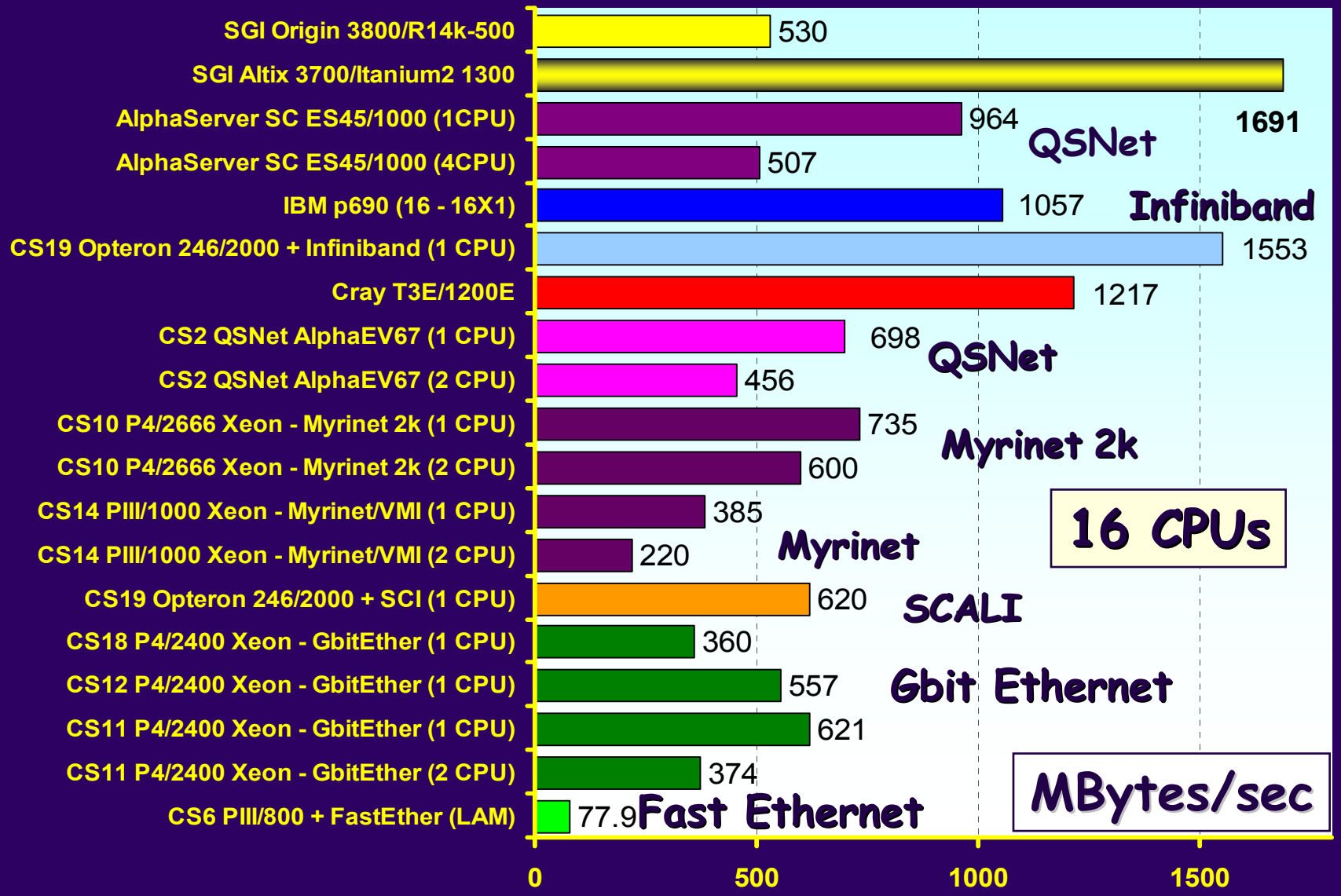
Performance of current interconnects.

	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

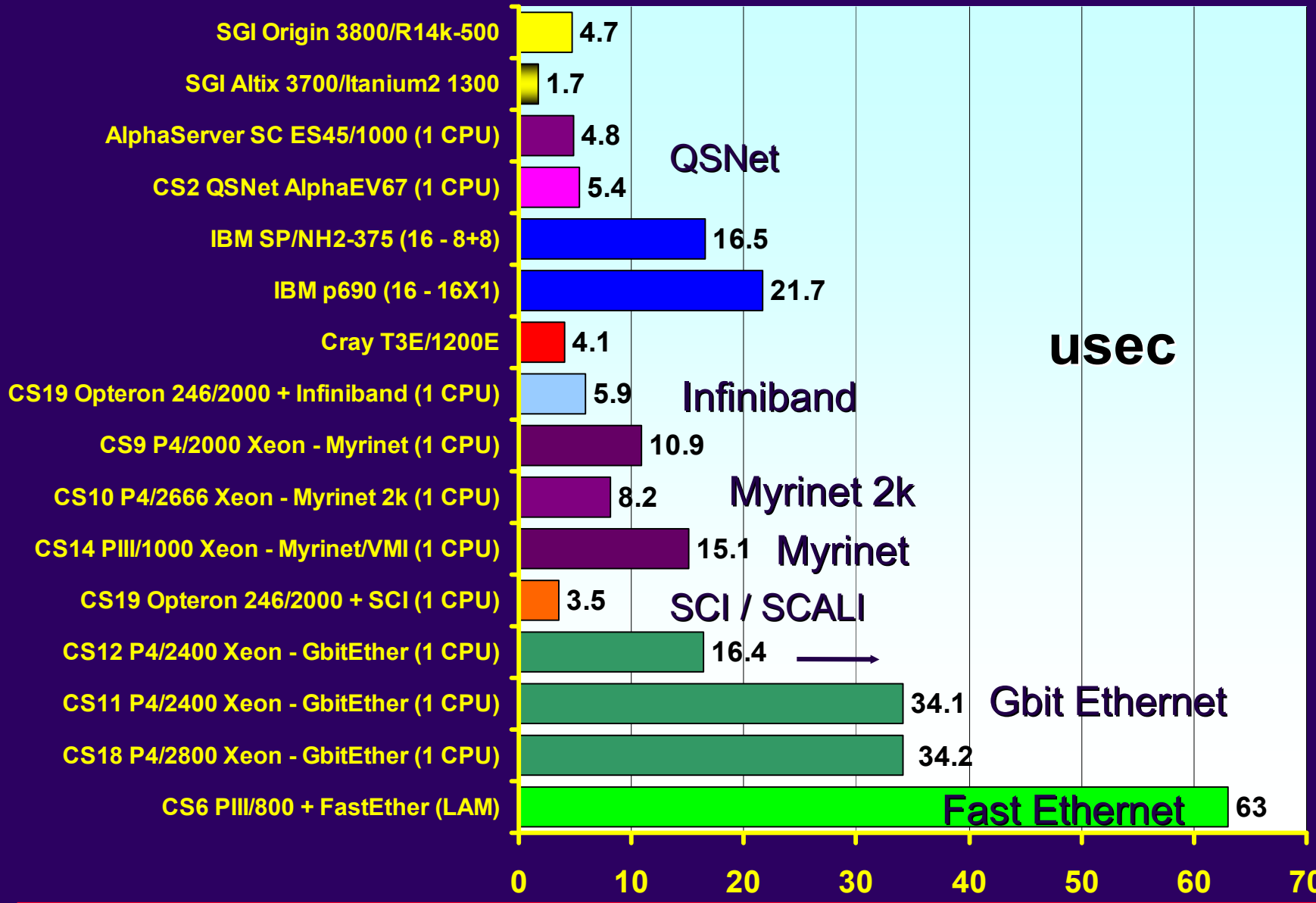
PingPong Performance



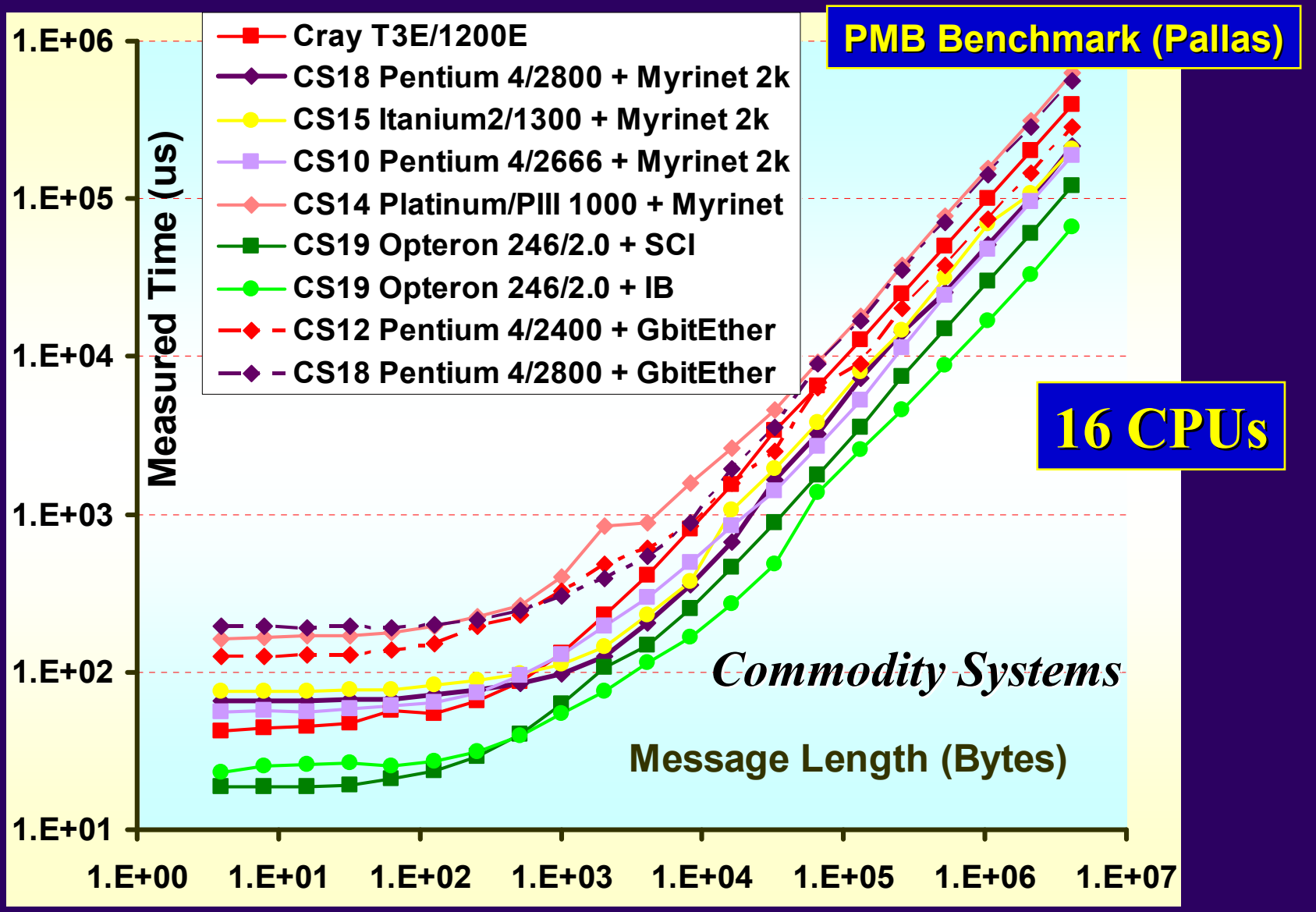
Interconnect Benchmark - EFF_BW



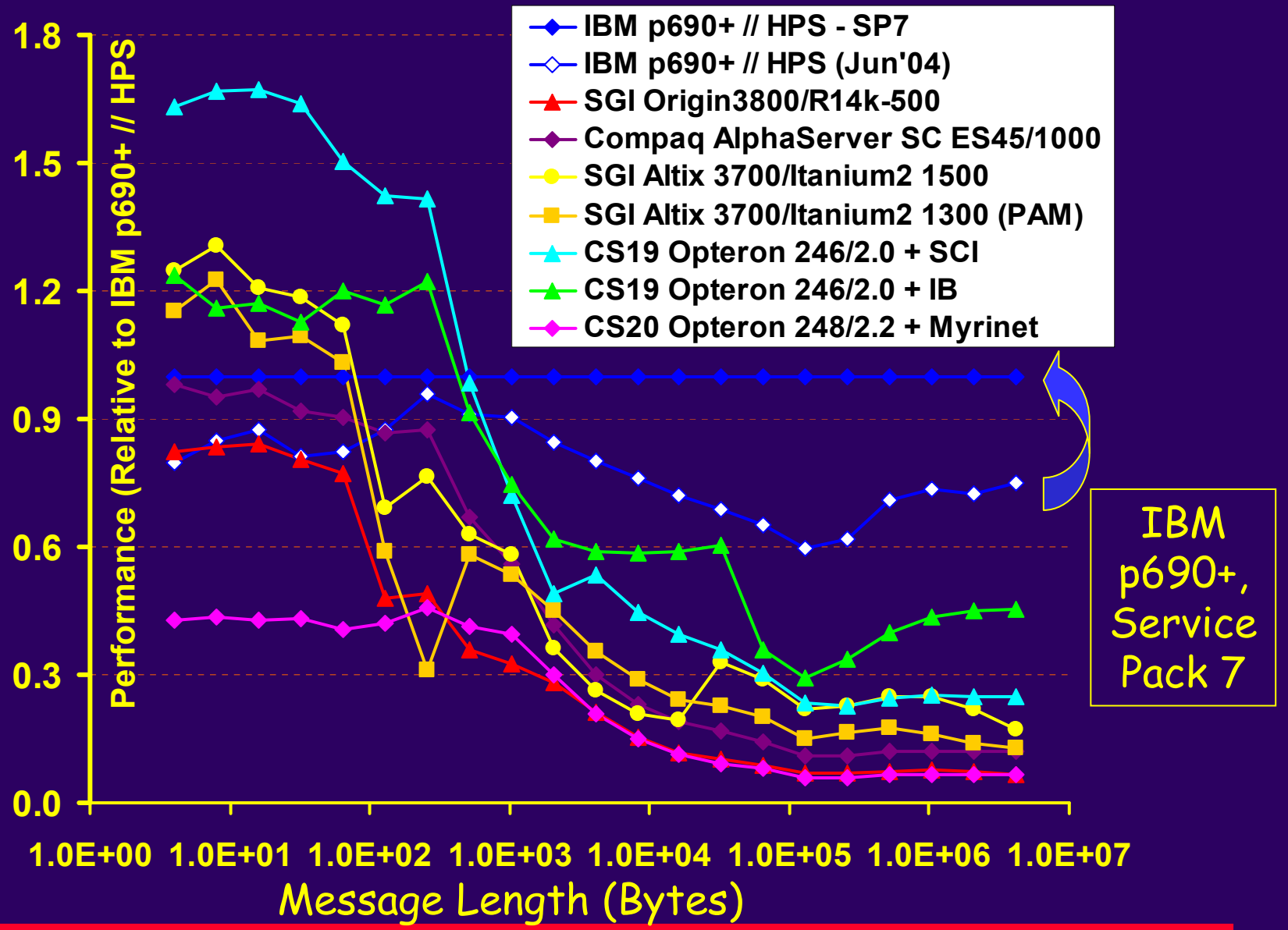
Interconnect Benchmark - Latency



MPI_allreduce Performance



64-CPU Relative Performance for Allreduce



Application Codes

Application-driven performance comparisons between Commodity-based systems and current High-end platforms :

- **Computational Chemistry**
  - **Molecular Simulation**
    - **DL_POLY, DLMULTI - parallel MD codes with many applications**
    - **CHARMM - macromolecular MD and energy minimisation**
  - **Molecular Electronic Structure**
    - **GAMESS-UK and NWChem - Ab initio Electronic structure codes**
- **Materials**
  - **CPMD, CASTEP** - Car-Parrinello plane wave codes
- **Engineering**
  - **ANGUS, SBLI** - regular-grid domain decomposition engineering codes

Performance Metric (% 32-node High-end system)

Performance Metrics: 2004

Attempt to quantify delivered performance from current Commodity-based systems against high-end platforms: SGI Altix 3700/Itanium2 1.3 GHz (CSAR service) and IBM p690+ (HPCx service) i.e.

Performance Metric (% 32-node SGI Altix 3700/Itanium 2 1.3 GHz)

T (32-CPUs SGI Altix 3700/Itanium2 1.3GHz) / T (32 CPUs) CSx

$T_{32\text{-CPU}}$ **SGI Altix 3700 /** $T_{32\text{-CPU}}$ **CS20 Opteron 248 / 2800 + Myrinet 2k**

$T_{32\text{-CPU}}$ **SGI Altix 3700 /** $T_{32\text{-CPU}}$ **CS18 Pentium 4 Xeon / 2800 + Gbit Ethernet**

Molecular Simulation

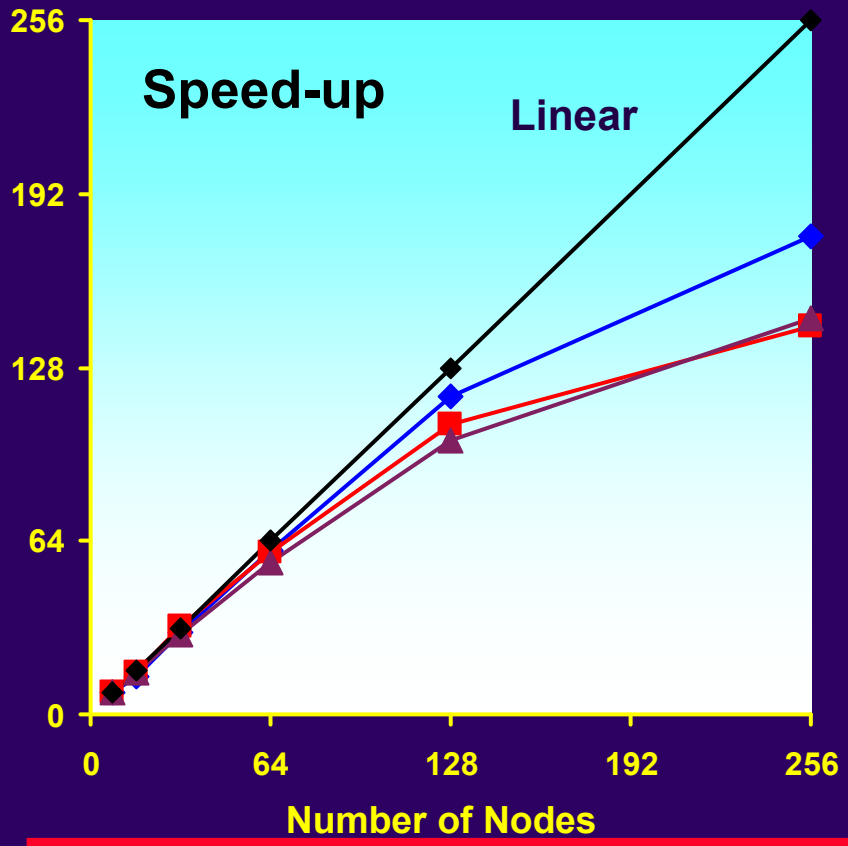
Molecular Dynamics Codes: DL_POLY, DLMULTI and CHARMM

- DL_POLY
- Developed as CCP5 parallel MD code by W. Smith and T.R. Forester
  - UK CCP5 + International user community
  - DLPOLY_2 (replicated data) and DLPOLY_3 (distributed data)
- 830 licences issued since 1994.
- Areas of application:
  - liquids, solutions, spectroscopy,ionic solids, molecular crystals,polymers,glasses, membranes, proteins, metals, solid and liquid interfaces, catalysis, clathrates,liquid crystals, biopolymers, polymer electrolytes.

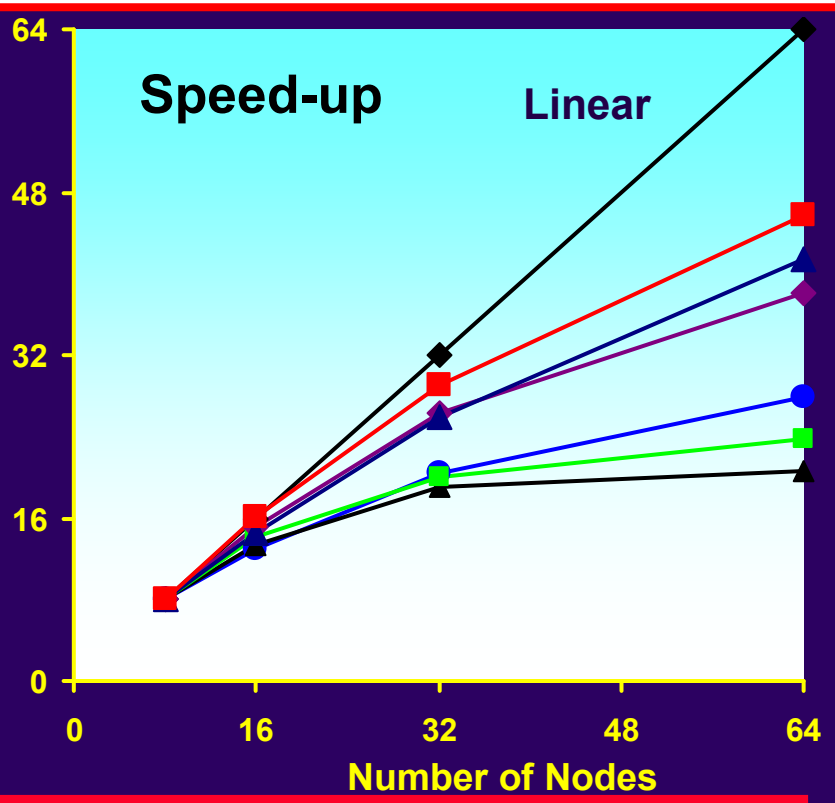
DL_POLY Parallel Benchmarks (Cray T3E/1200)

V2: Replicated Data

4. NaCl; Ewald, 27,000 ions
5. NaK-disilicate glass; 8,640 atoms,MTS+ Ewald
8. MgO microcrystal: 5,416 atoms

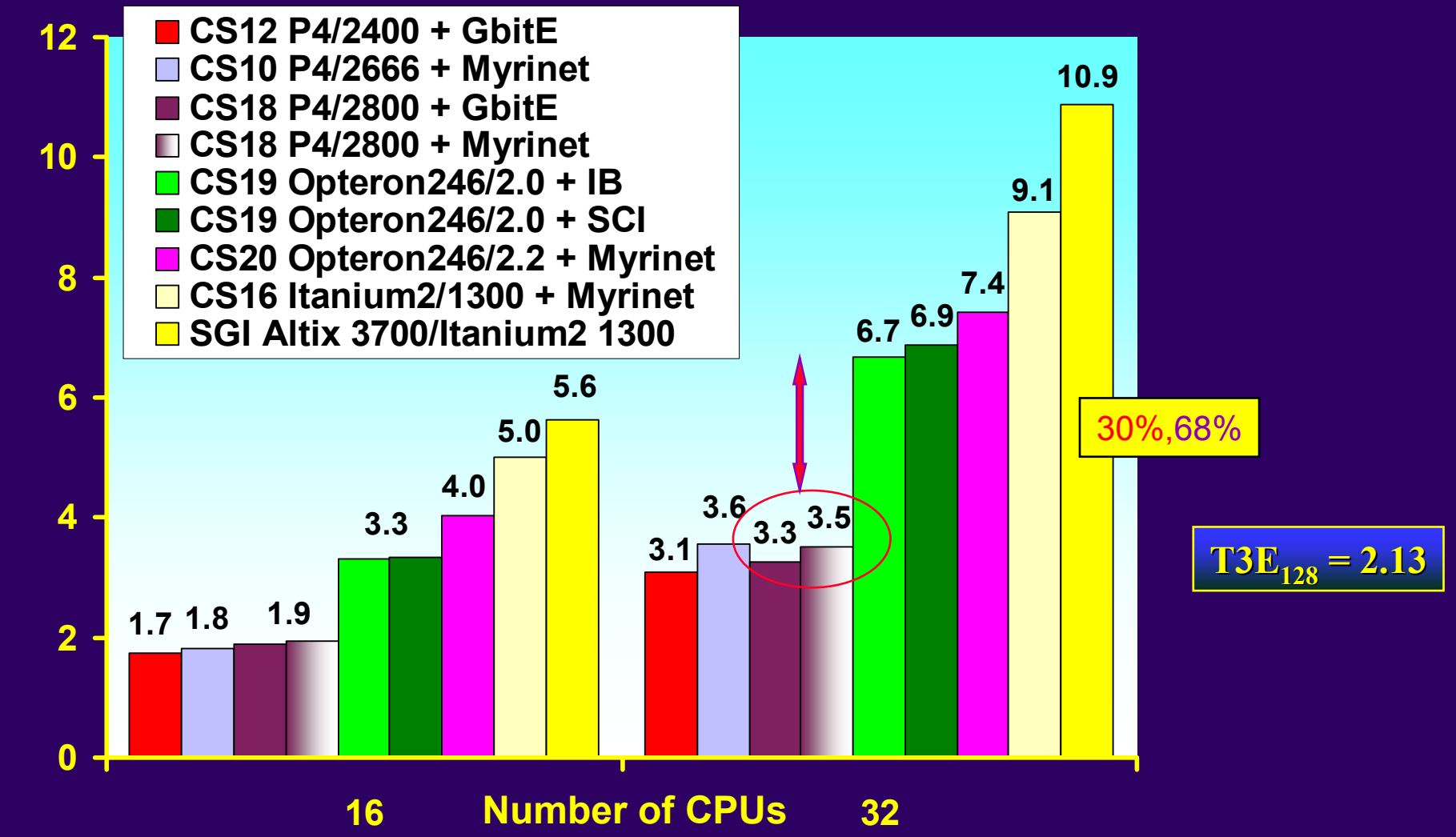


9. Model membrane/Valinomycin (MTS, 18,886)
7. Gramicidin in water (SHAKE, 12,390)
6. K/valinomycin in water (SHAKE, AMBER, 3,838)
1. Metallic Al (19,652 atoms, Sutton Chen)
3. Transferrin in Water (neutral groups + SHAKE, 27,593)
2. Peptide in water (neutral groups + SHAKE, 3993).



DL_POLY V2: Bench 4 - Commodity-based Systems

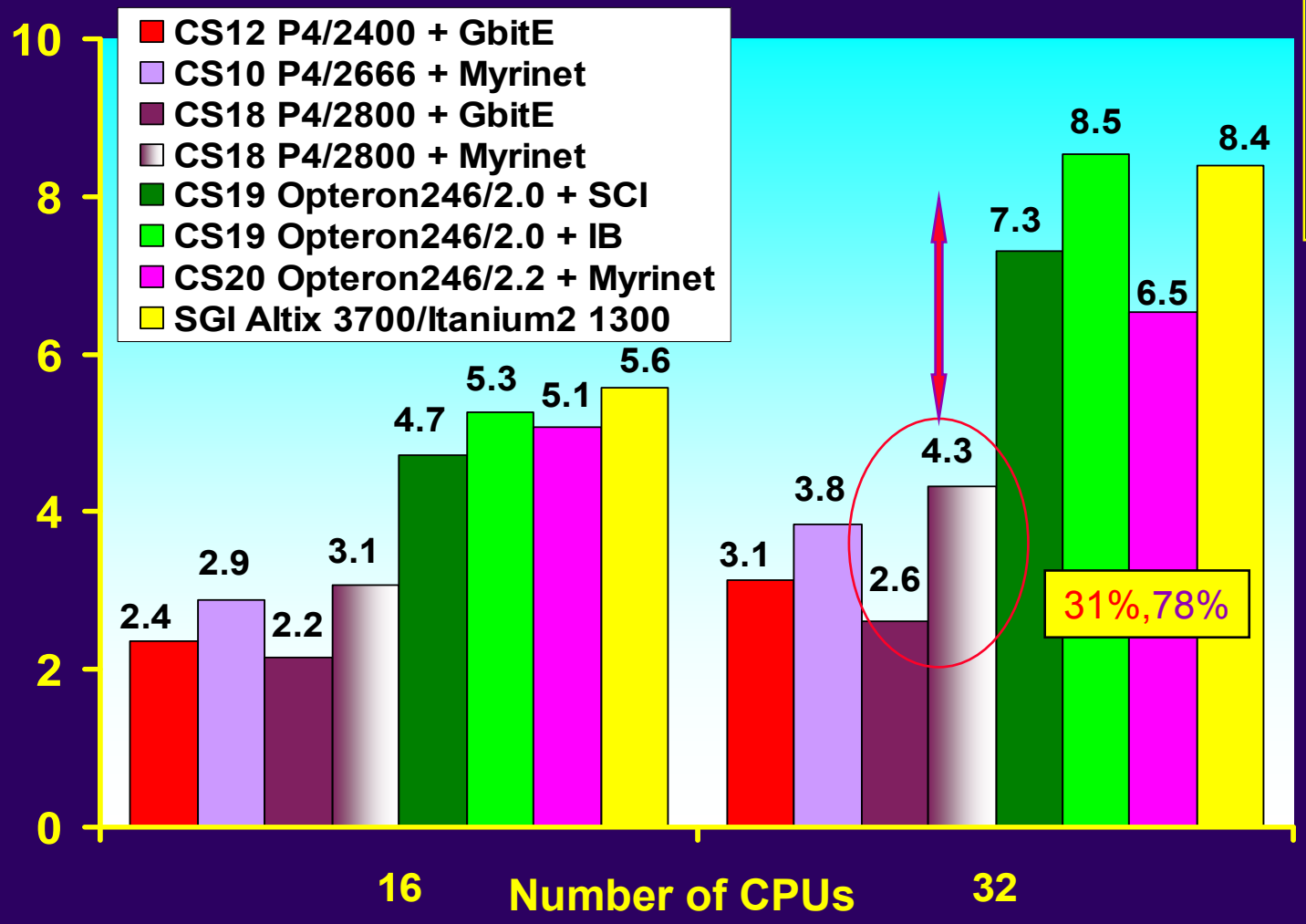
NaCl; 27,000 ions, Ewald , 75 time steps, Cutoff=24Å



$T3E_{128}$ = 2.13

DL_POLY V2: Bench 7 - Commodity-based Systems

Performance

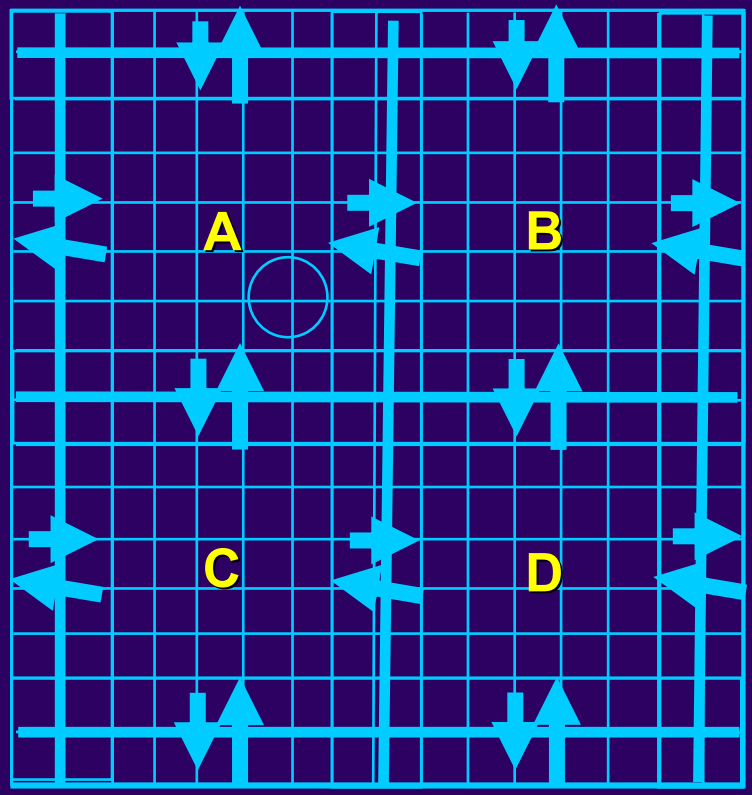


Gramicidin in water;
rigid bonds and SHAKE,
12,390 atoms,
500 time steps

$T3E_{128}$ =2.4

Migration from Replicated to Distributed data: DL_POLY-3 : Domain Decomposition

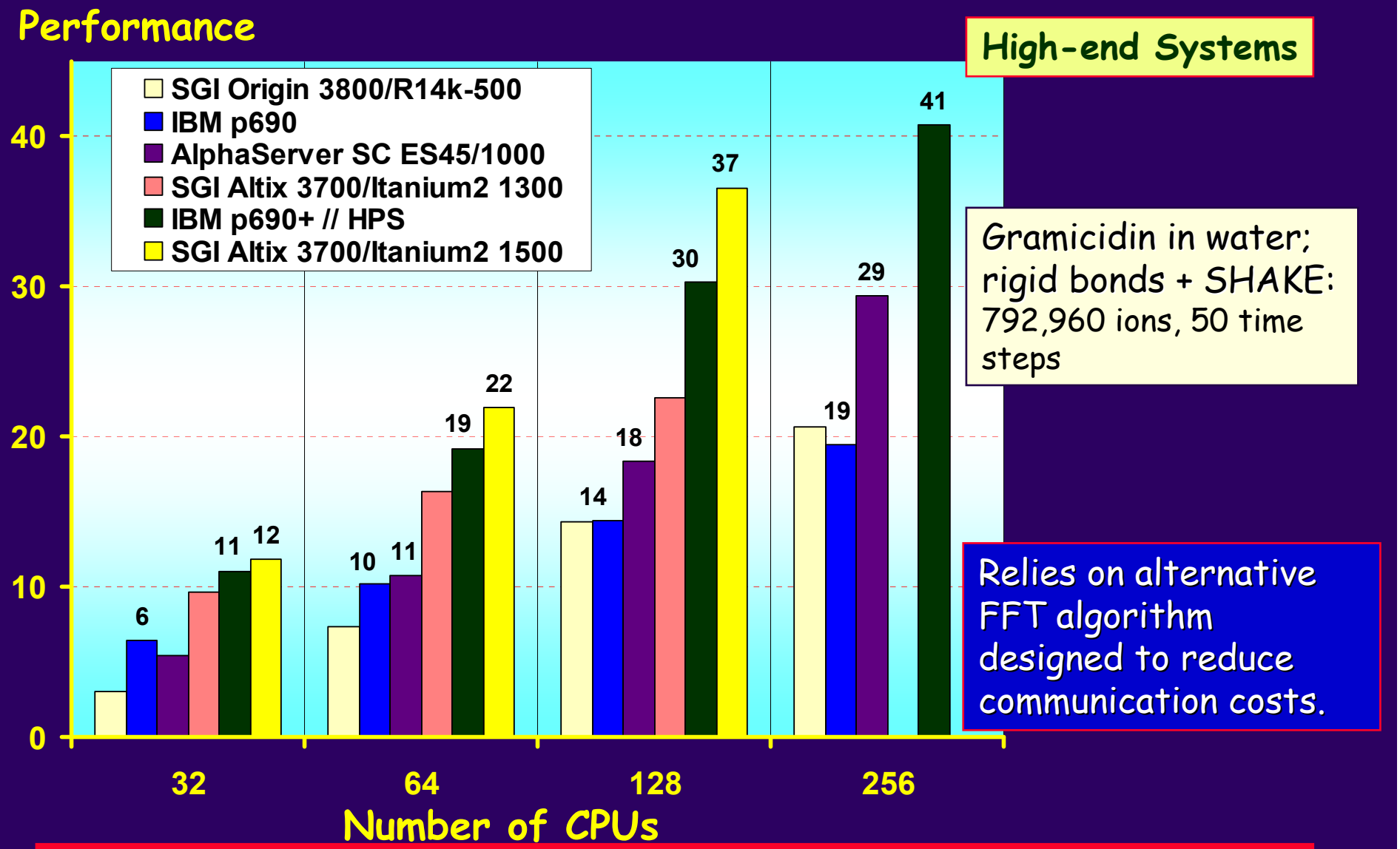
- Distribute atoms, forces across the nodes
  - More memory efficient, can address much larger cases (10^5-10^7)
- Shake and short-ranges forces require only neighbour communication
  - communications scale linearly with number of nodes
- Coulombic energy remains global
  - strategy depends on problem and machine characteristics
  - Adopt **Smooth Particle Mesh Ewald** scheme
    - includes Fourier transform smoothed charge density (reciprocal space grid typically 64x64x64 - 128x128x128)



W. Smith and T.R. Forester

http://www.cse.clrc.ac.uk/msi/software/DL_POLY

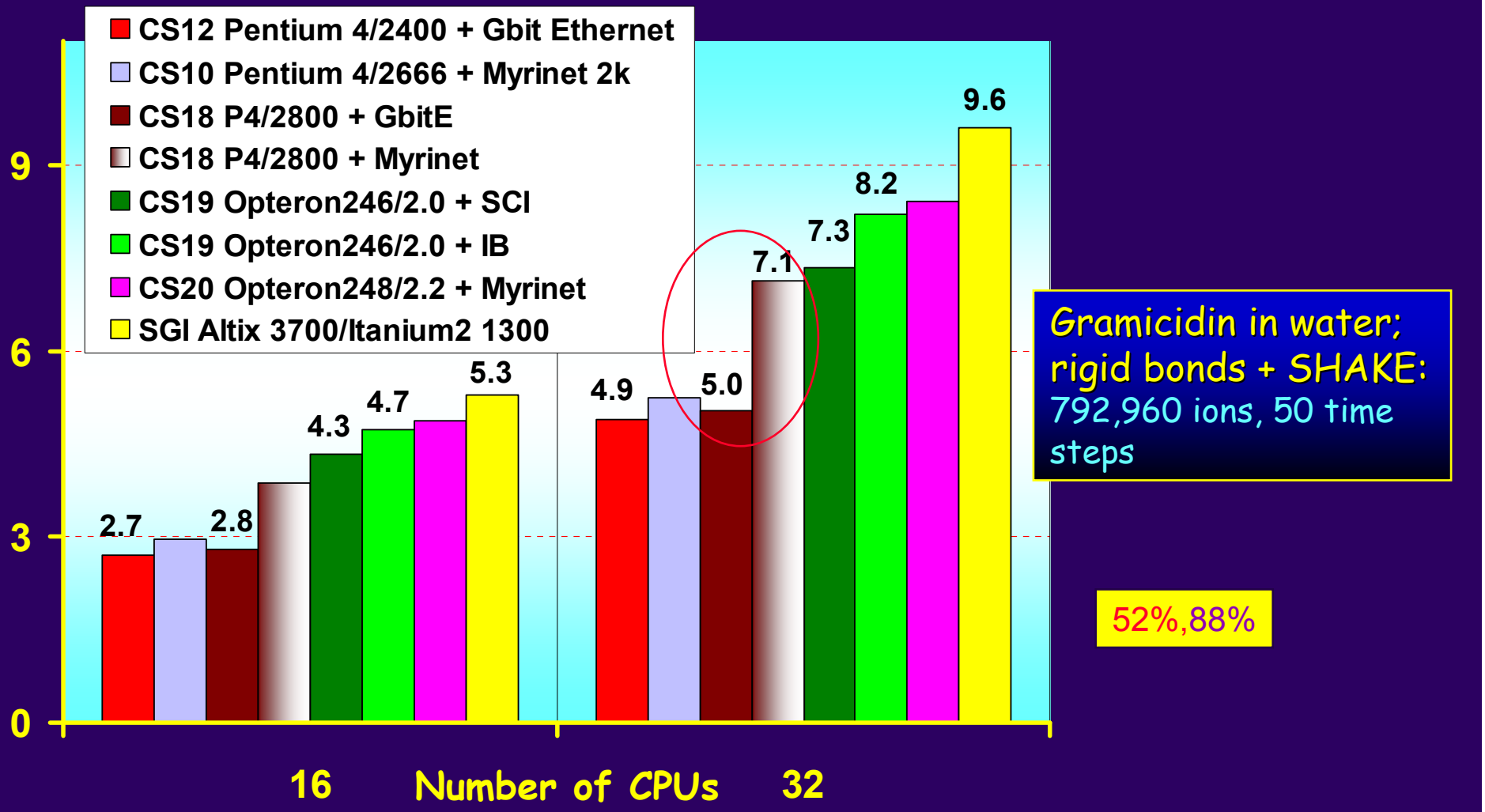
DL_POLY3 Macromolecular Simulations



DL_POLY3 Macromolecular Simulations

Performance

Commodity-based Systems



Gramicidin in water; rigid bonds + SHAKE: 792,960 ions, 50 time steps

52%,88%

Molecular Electronic Structure

Ab initio Electronic structure Codes:
NWChem and GAMESS-UK

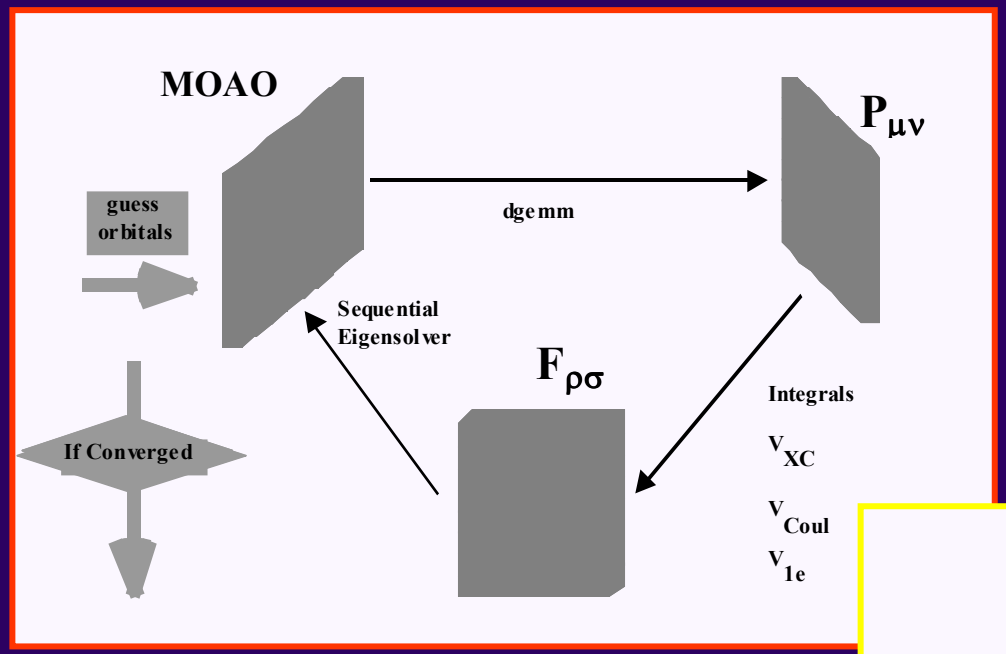
Molecular Electronic Structure

RD & DD Strategies

- Replicated Data Strategy (RD)
  - Not efficient for large parallel computers due to both system size and parallel scalability limitations.
  - Problem size driven by memory available to one CPU, not the whole machine.
  - Communications overhead – results of parallel OP must be re-replicated so that each CPU has a current version i.e. expensive global communication - time taken grows with the number of CPUs.
  - RD does not fit well with many algorithms e.g. diagonalization - increasingly important stage of SCF/DFT calculations as the system size increases.
- Distributed Data Strategy (DD) - NWChem
  - Each CPU holds only part of each of the (major) data structures, overcoming many of the drawbacks of replicated data;
    - Limit on system size is now that given by all memory on the computer,
    - global communications largely avoided as re-replication is not required,
    - many of the algorithms that can not be effectively addressed by a RD strategy fit more naturally within a distributed data one.

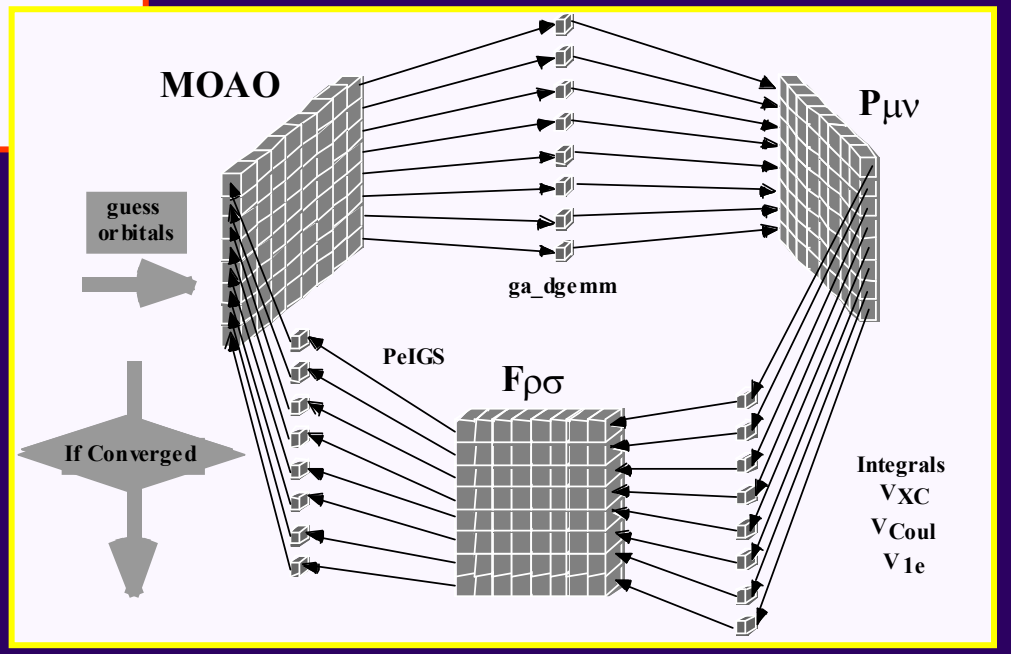
http://www.emsl.pnl.gov/docs/nwchem/nwchem.html

Distributed Data SCF



Pictorial representation of the iterative SCF process in (i) a sequential process, and (ii) a distributed data parallel process: **MOAO** represents the molecular orbitals, **P** the density matrix and **F** the Fock or Hamiltonian matrix

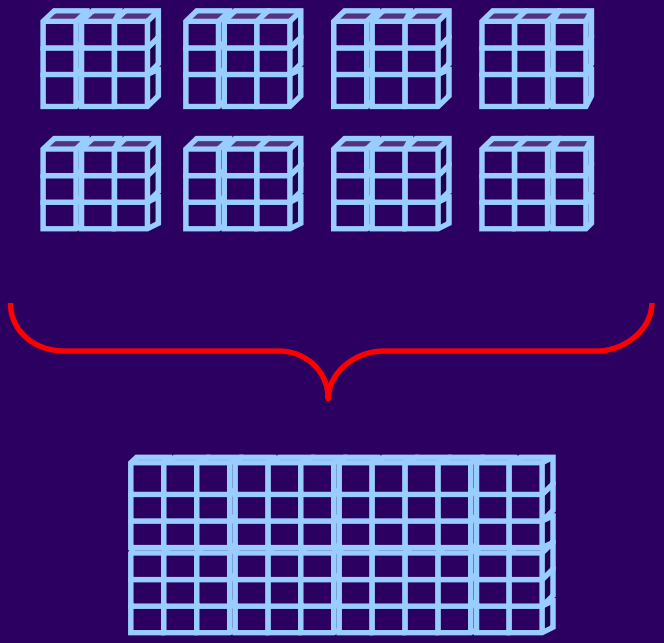
Sequential



Distributed Data

Global Arrays

Physically distributed data

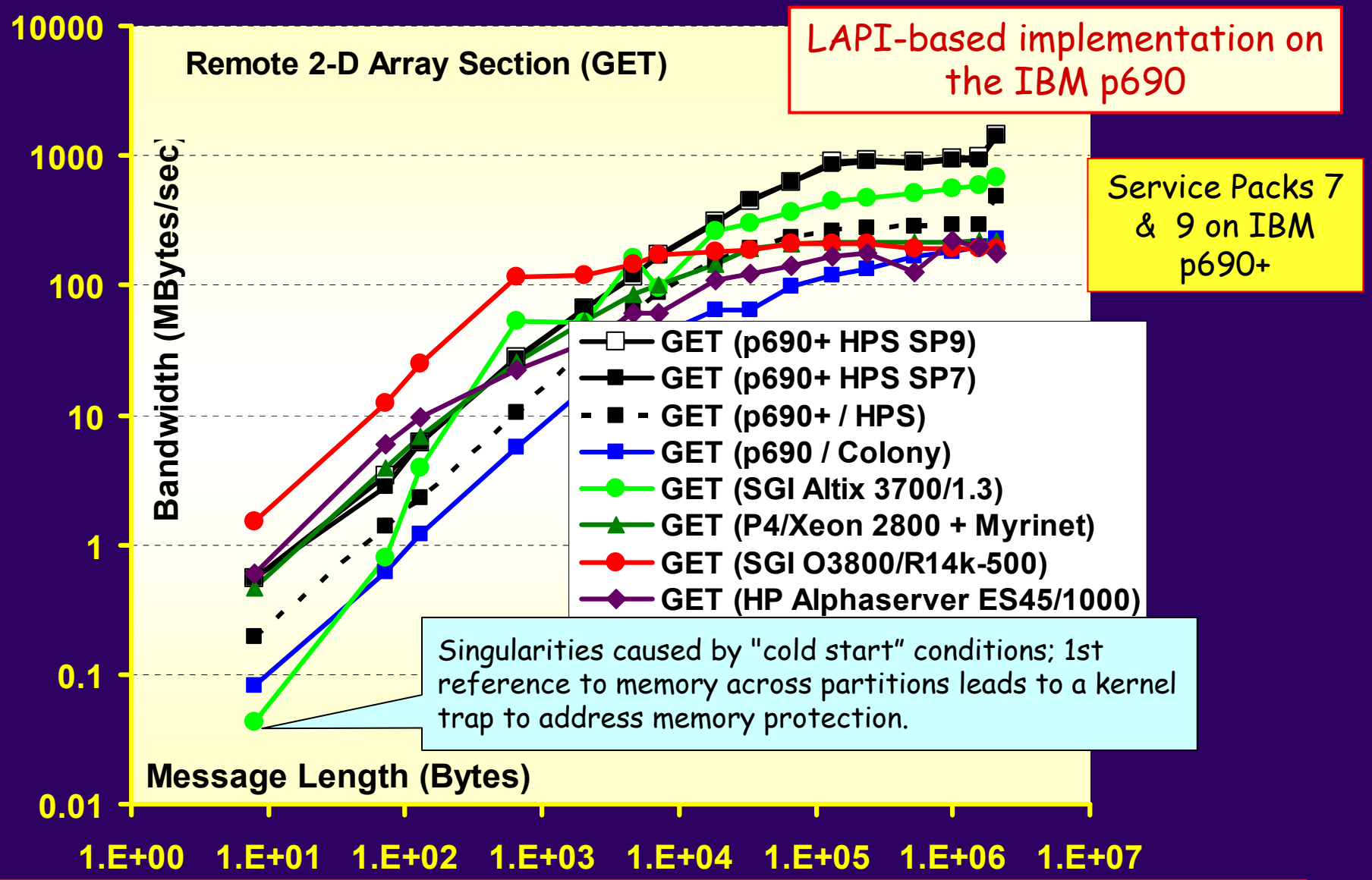


Single, shared data structure

- 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

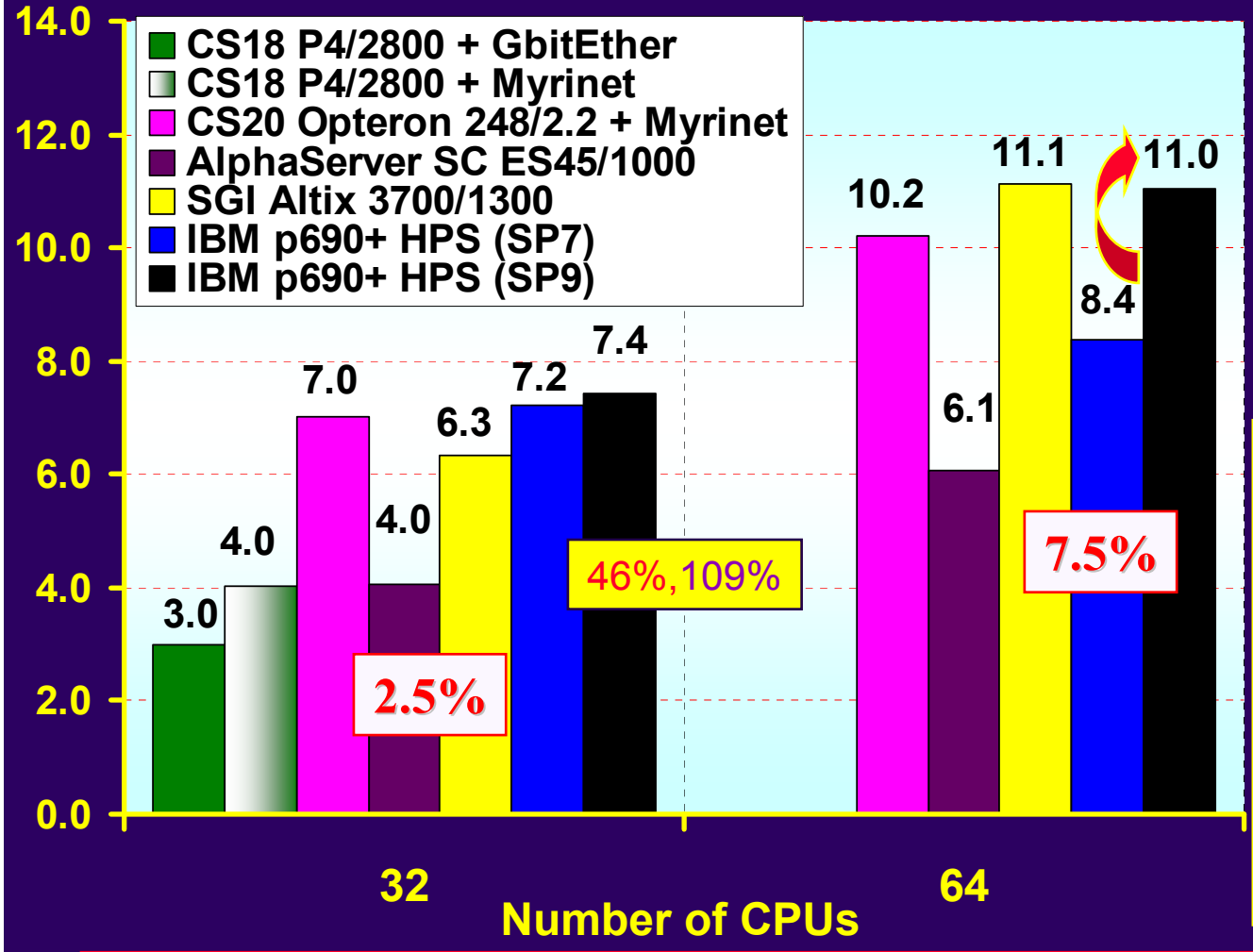
Tools developed as part of the NWChem project at PNNL; R.J. Harrison, J. Nieplocha et al.

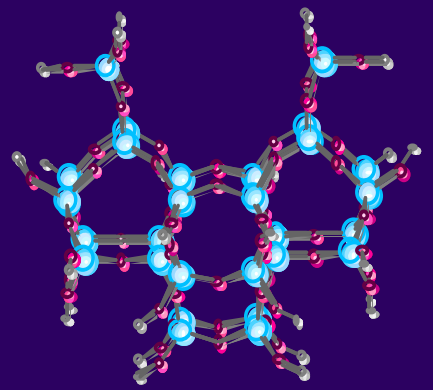
Global Array Benchmark I. GET



DFT Coulomb Fit - NWChem 4.6

Performance $Si_{28}O_{67}H_{30}$ 1687/3928





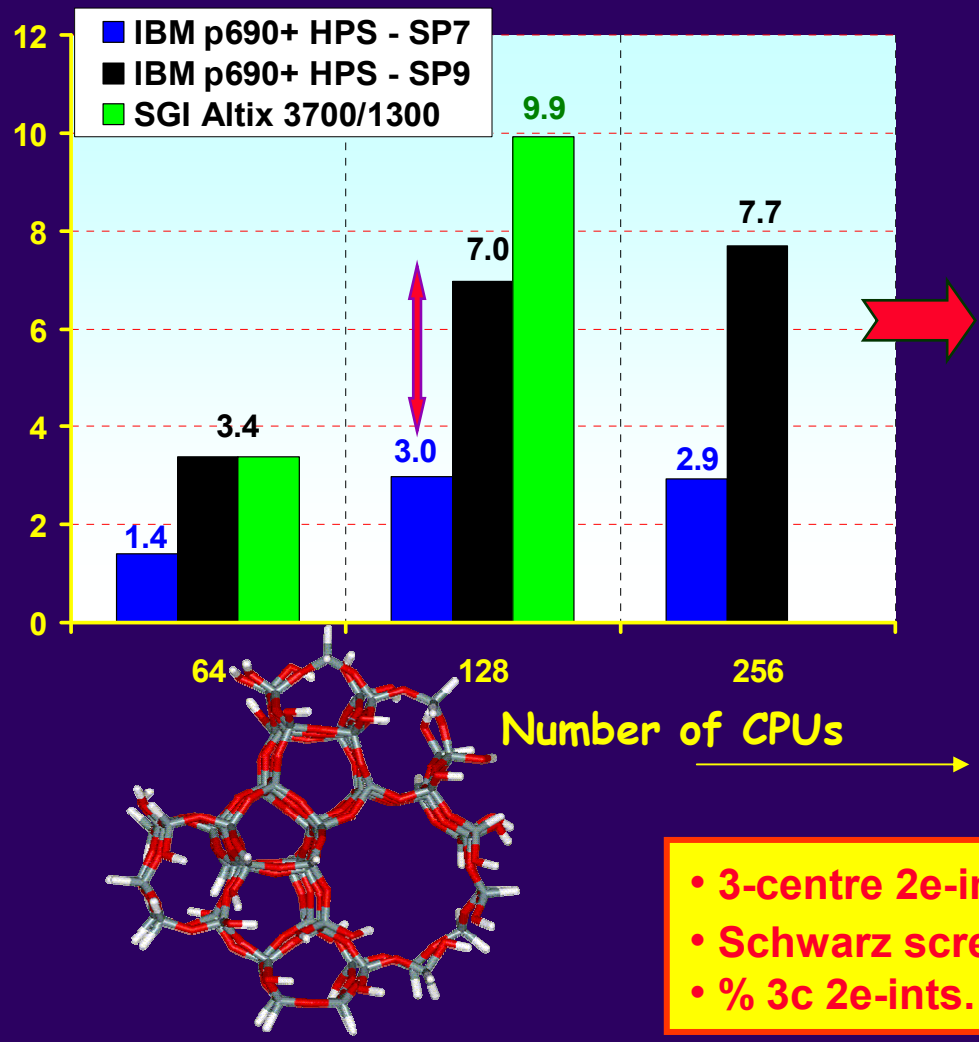
"memory 420 mb"

Basis (Godbout et al.)
DZVP - O, Si
DZVP2 - H
Fitting Basis:
DGAUSS-A1 - O, Si
DGAUSS-A2 - H
Use auxiliary fitting basis for coulomb energy, with 3 centre 2 electron integrals held in core (%).

Exploiting Global Memory: NWChem

Zeolite ZSM-5

DFT (LDA) - Fitted Coulomb Approach - use memory to avoid re-computation of the 3c 2e-integrals



- DZVP Basis (DZV_A2) and Dgauss A1_DFT Fitting basis:
  - AO basis: 3554
  - CD basis: 12713
- IBM p690+ HPS (SP9)

Wall time (13 SCF iterations):

32 CPUs = 5,499 seconds (0%)
128 CPUs= 1,434 seconds (100%)

- 3-centre 2e-integrals = 1.00×10^{12}
- Schwarz screening = 5.94×10^{9}
- % 3c 2e-ints. In core = 100%

Parallel Implementation of GAMESS-UK

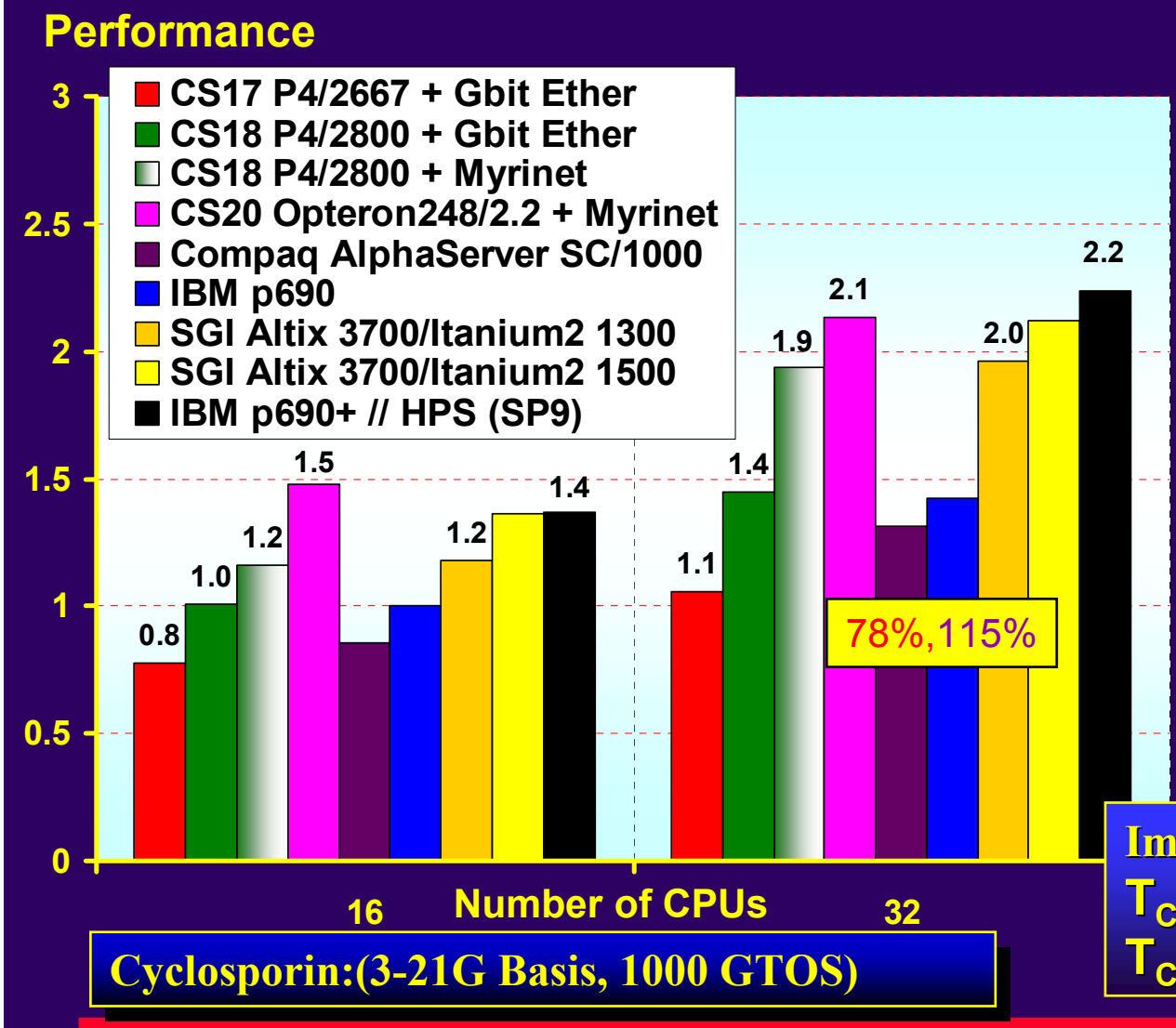
- Current approach between extremes of replicated and distributed data.
- Most data is replicated - when a parallel linear algebra OP is to be performed the data is copied into a global array, GA tools used to perform the operation, and the data copied back into a replicated object.
  - SCF and DFT
    - GA Tools for caching of I/O for restart and checkpoint files
    - Storage of 3-centre 2-e integrals in DFT Jfit
    - Linear Algebra (via Parallel Linear Algebra from NWChem Project – PeIGs) - DIIS/MMOs, Inversion of Fitting matrix)
  - MP2 gradients, SCF and DFT second derivatives
    - Distribution of <vvoo> and <vovo> integrals via GAs
  - Direct RPA Excited States
    - Replicated data with parallelisation of direct integral evaluation

http://www.cfs.dl.ac.uk

- This solves some of the problems associated with a RD strategy, but large memory requirements and the re-replication overheads remain.
- Efficiency dependent on optimised GA implementation ….

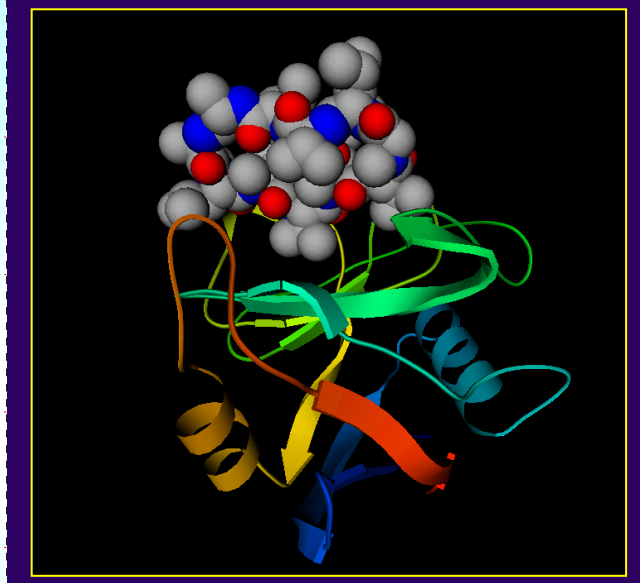
GAMESS-UK ΔSCF Performance

IBM SP/p690, High-end and Commodity-based Systems



Cyclosporin:(3-21G Basis, 1000 GTOS)

$T3E_{128} = 0.78$



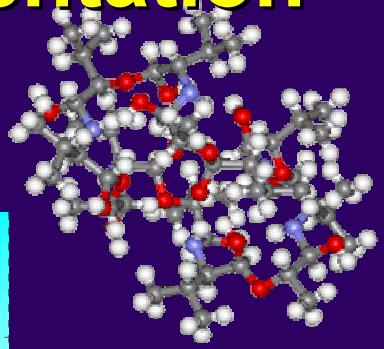
PDSYEVD (D&C Method)
(Scalapack (1.7)

Impact of Serial Linear Algebra:

$T_{CS20\text{-}Opt248}(16) = 1306$ [231]

$T_{CS20\text{-}Opt248}(32) = 1270$ [160]

GAMESS-UK - GA-based Implementation

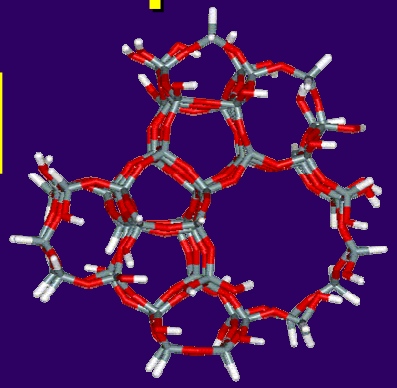


Performance

Commodity-based and High-end Systems

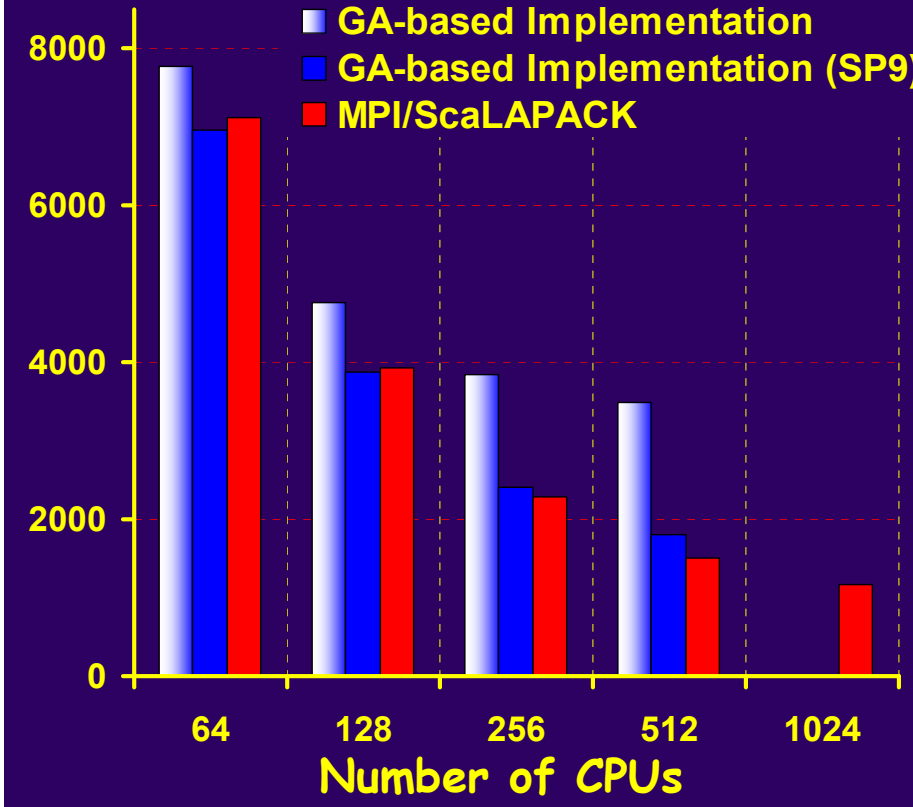
Distributed Data Implementation of GAMESS-UK

Zeolite Y cluster

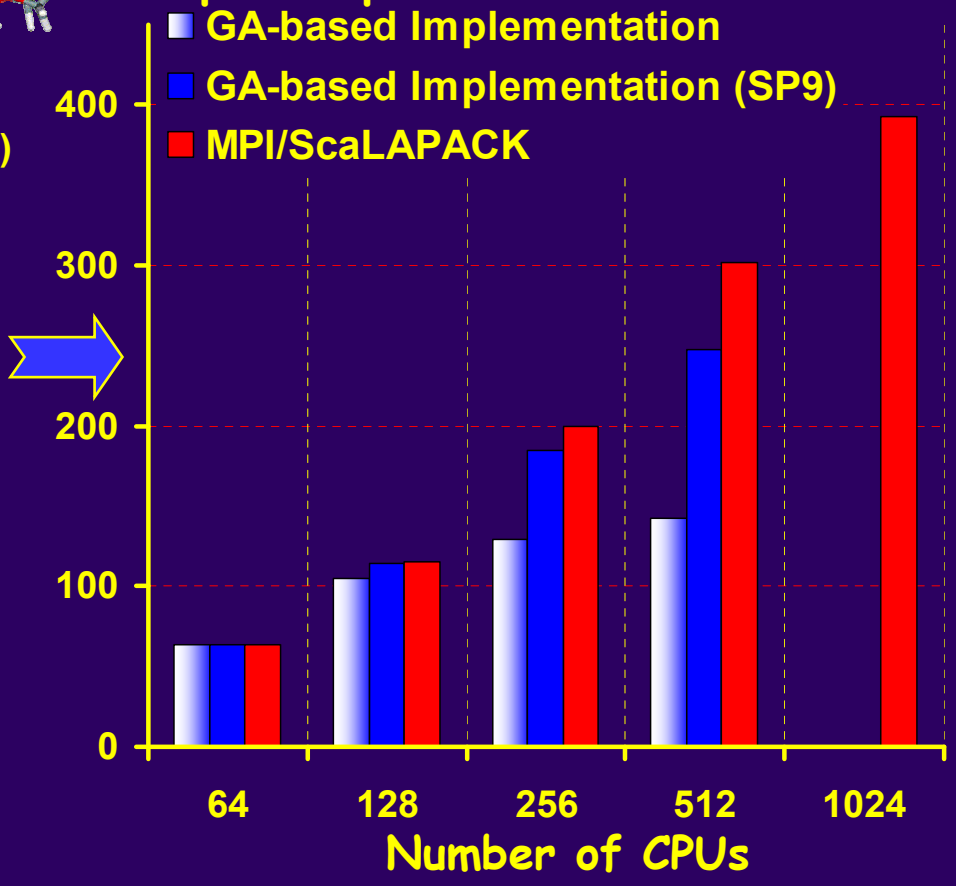


DZVP Basis (DZV_A2): 3975 GTOs
Hartree Fock (IBM p690+)

Elapsed Time (seconds)



Speed-up



Materials Simulation. Plane Wave Methods: CASTEP, CPMD

Direct minimisation of the total energy (avoiding diagonalisation)

$$\psi_j^{\vec{k}}(\vec{r}) = \sum_{\vec{G}}^{(\vec{k}+\vec{G})^2 < E_{cut}} C_{j,\vec{G}}^{\vec{k}} e^{-i(\vec{k}+\vec{G}).\vec{r}}$$

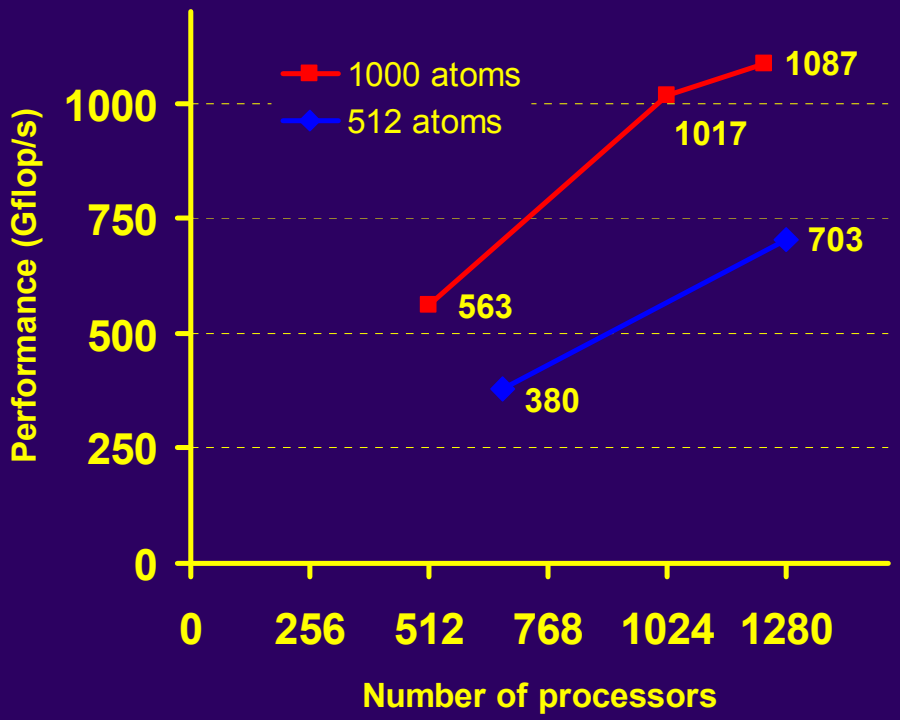
- Pseudopotentials must be used to keep the number of plane waves manageable
- Large number of basis functions N~10^6 (especially for heavy atoms).

The plane wave expansion means that the bulk of the computation comprises large 3D Fast Fourier Transforms (FFTs) between real and momentum space.

- These are distributed across the processors in various ways.
- The actual FFT routines are optimized for the cache size of the processor.

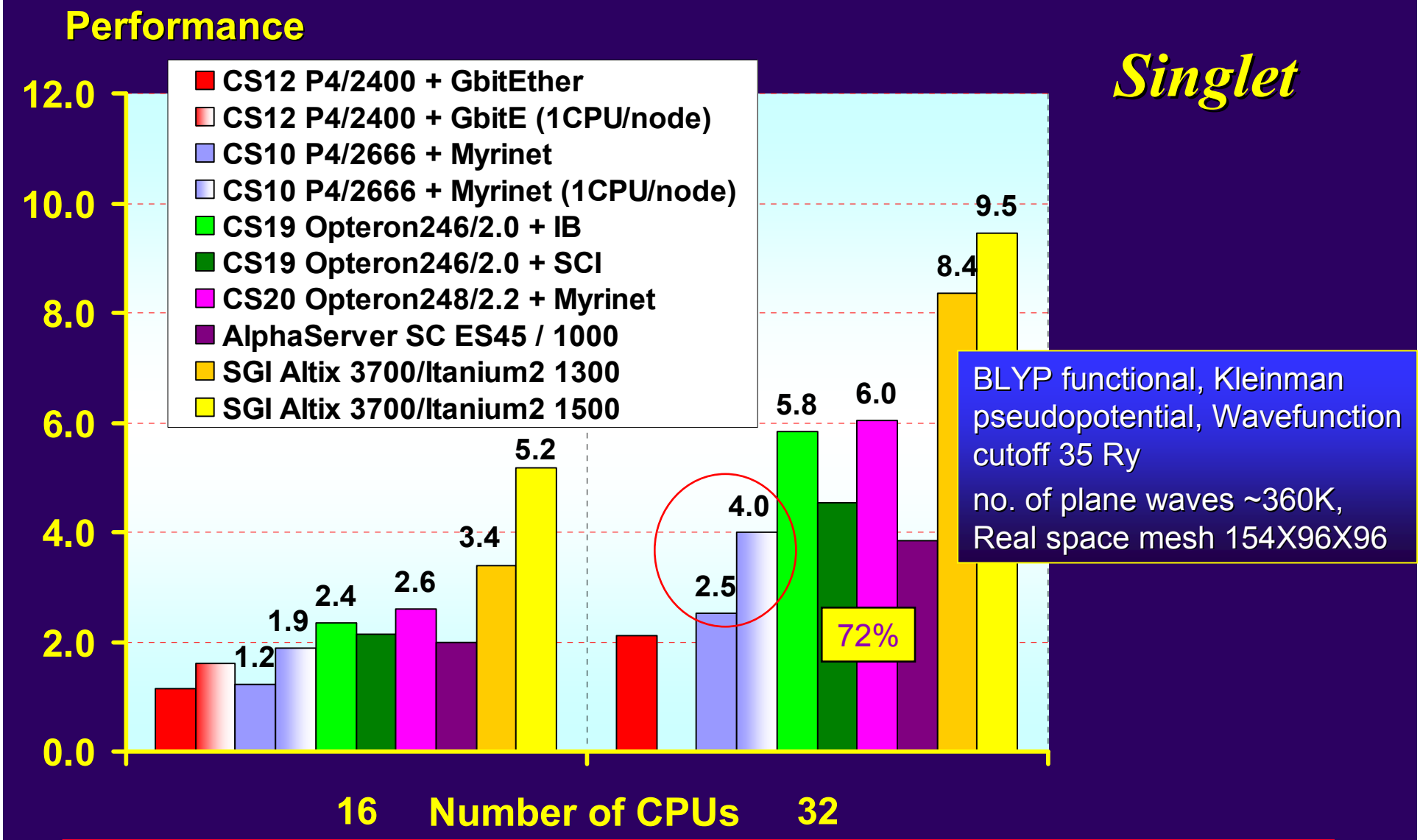
CPMD - Mixed Mode Programming

- Developed at IBM Zurich from the original Car-Parrinello Code in 1993;
- Since 2001 distributed free to academic institutions (www.cpmd.org); more than 2500 licenses in the first year in more than 50 countries.



- Performance has been tested using up to 1280 processors of HPCx (IBM p690/colony).
- The maximum performance achieved was above 1 Tflop/s for a system consisting of 1000 atoms. (~20% of peak performance).
- 1000 atoms SiC supercell, 256x256x256 Mesh, Cutoff 60 Ry (2,209,586 plane waves)
- Mixed approach instrumental to obtain these results
- Larger SMPs / better switches will help

CPMD 3.7 - C_{120} Benchmark

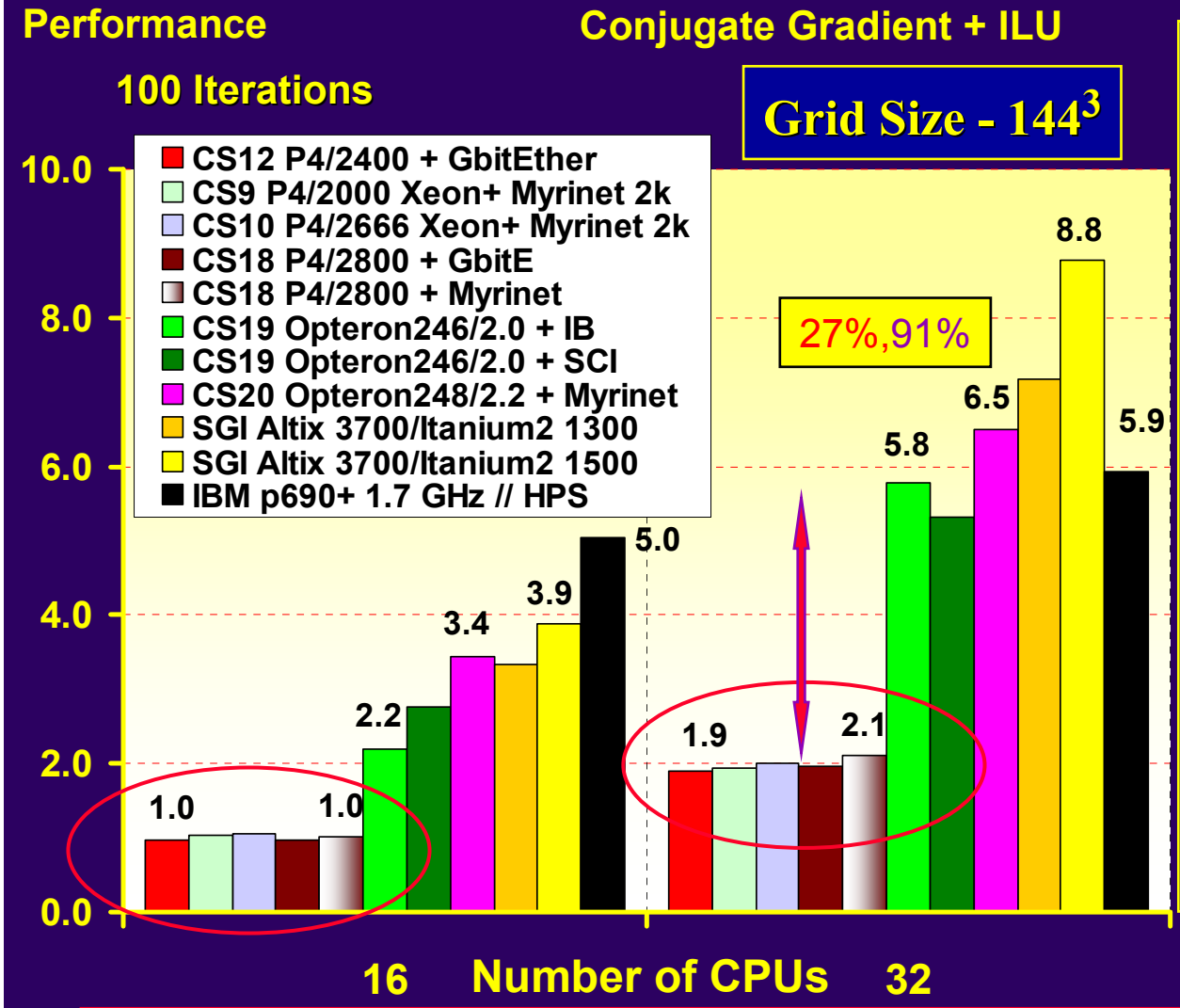


Computational Engineering

Direct numerical simulation (DNS) Codes:
ANGUS and SBLI

ANGUS: Combustion modelling (regular grid)

High-end and Commodity-based Systems



Direct numerical simulations (DNS) of turbulent pre-mixed combustion solving the augmented Navier-Stokes equations for fluid flow.

Discretisation of equations is performed using standard 2nd order central differences on a 3D-grid.

Pressure solver utilises either a conjugate gradient method with modified incomplete LU pre-conditioner or a multi-grid solver (both make extensive use of Level 1 BLAS) or fast Fourier transform.

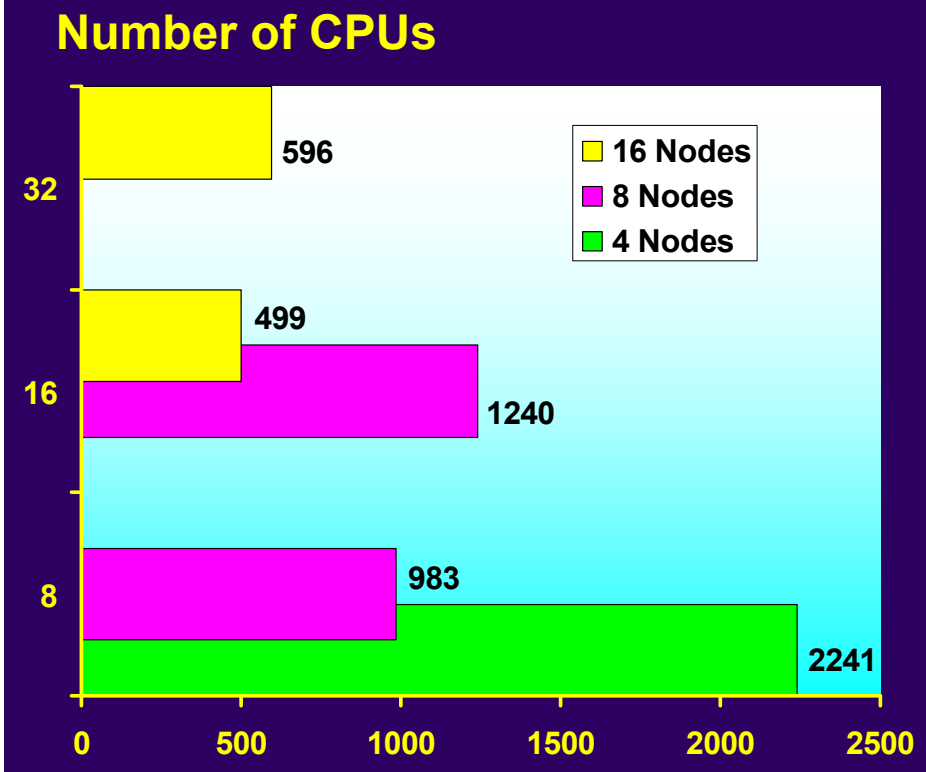
ANGUS: Combustion modelling (regular grid)

Memory Bandwidth Effects: Pentium Xeon and Opteron Systems

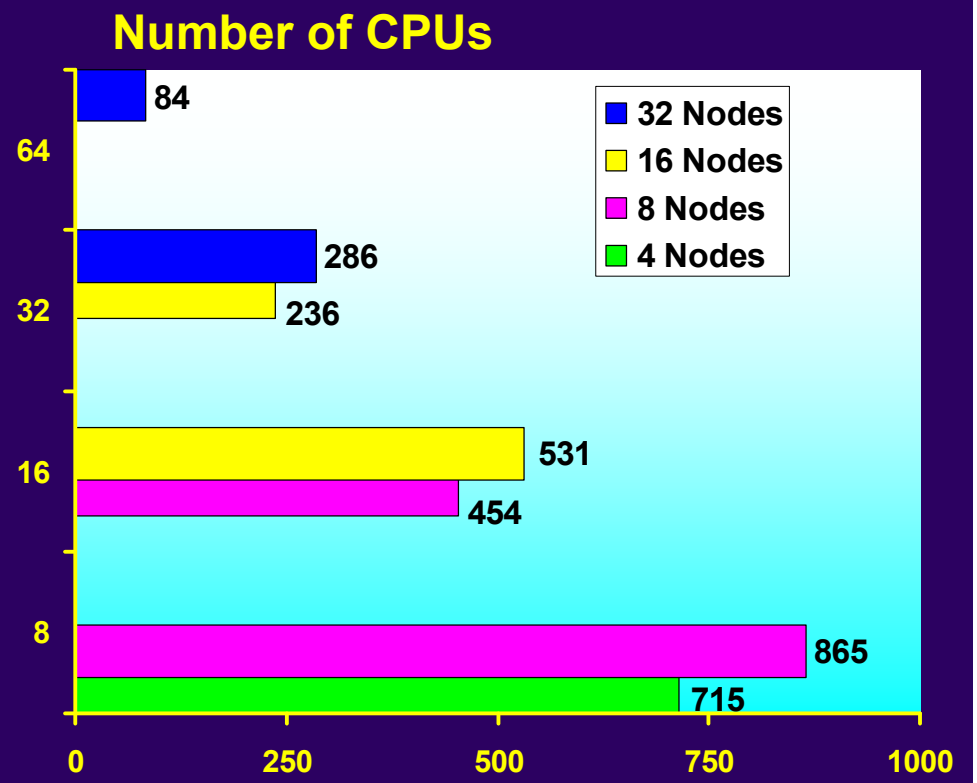
Conjugate Gradient + ILU

Grid Size - 144^3

CS18 P4 Xeon/2800 + Myrinet 2k

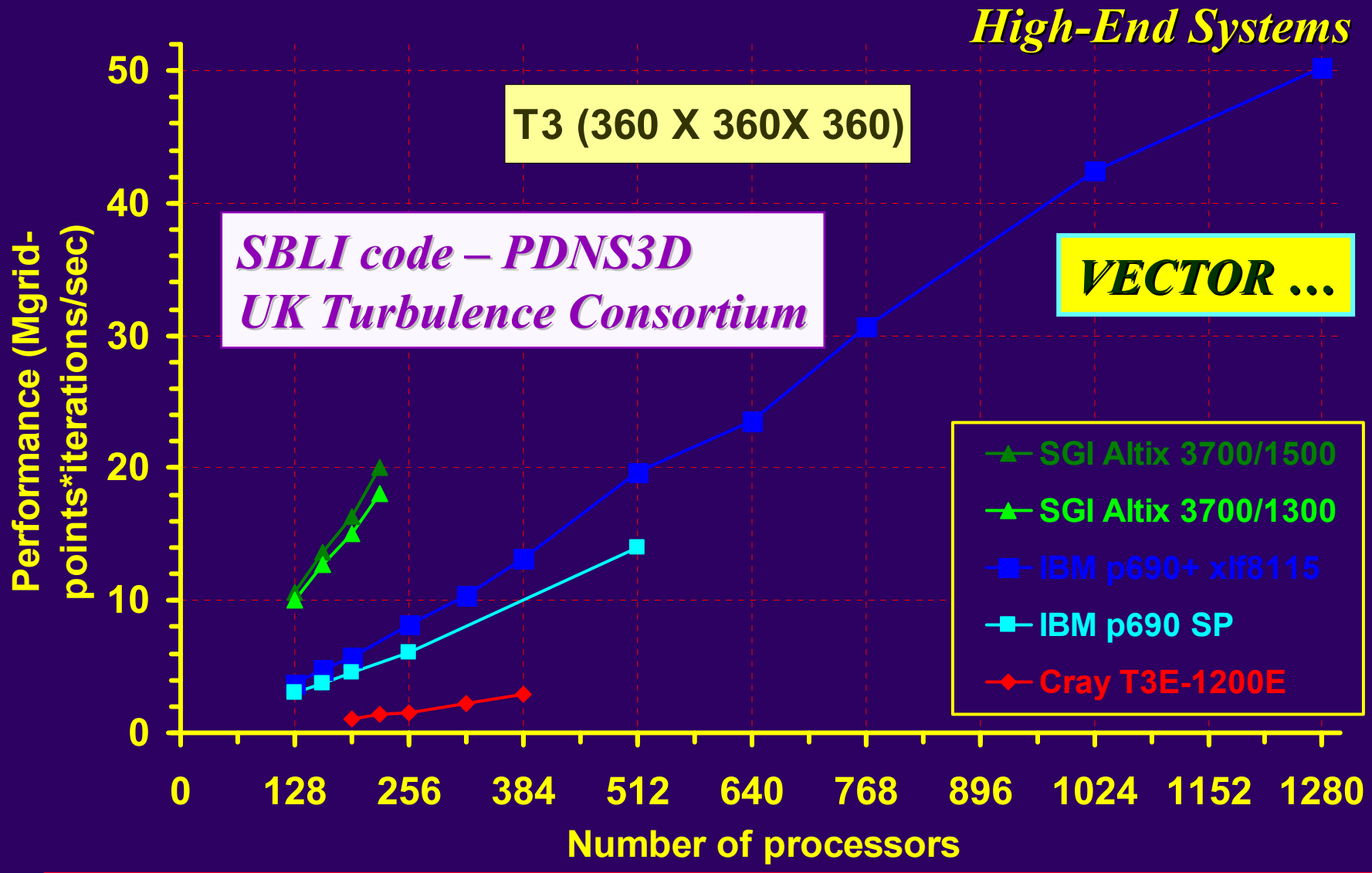


CS19 Opteron246/2.0 + SCI



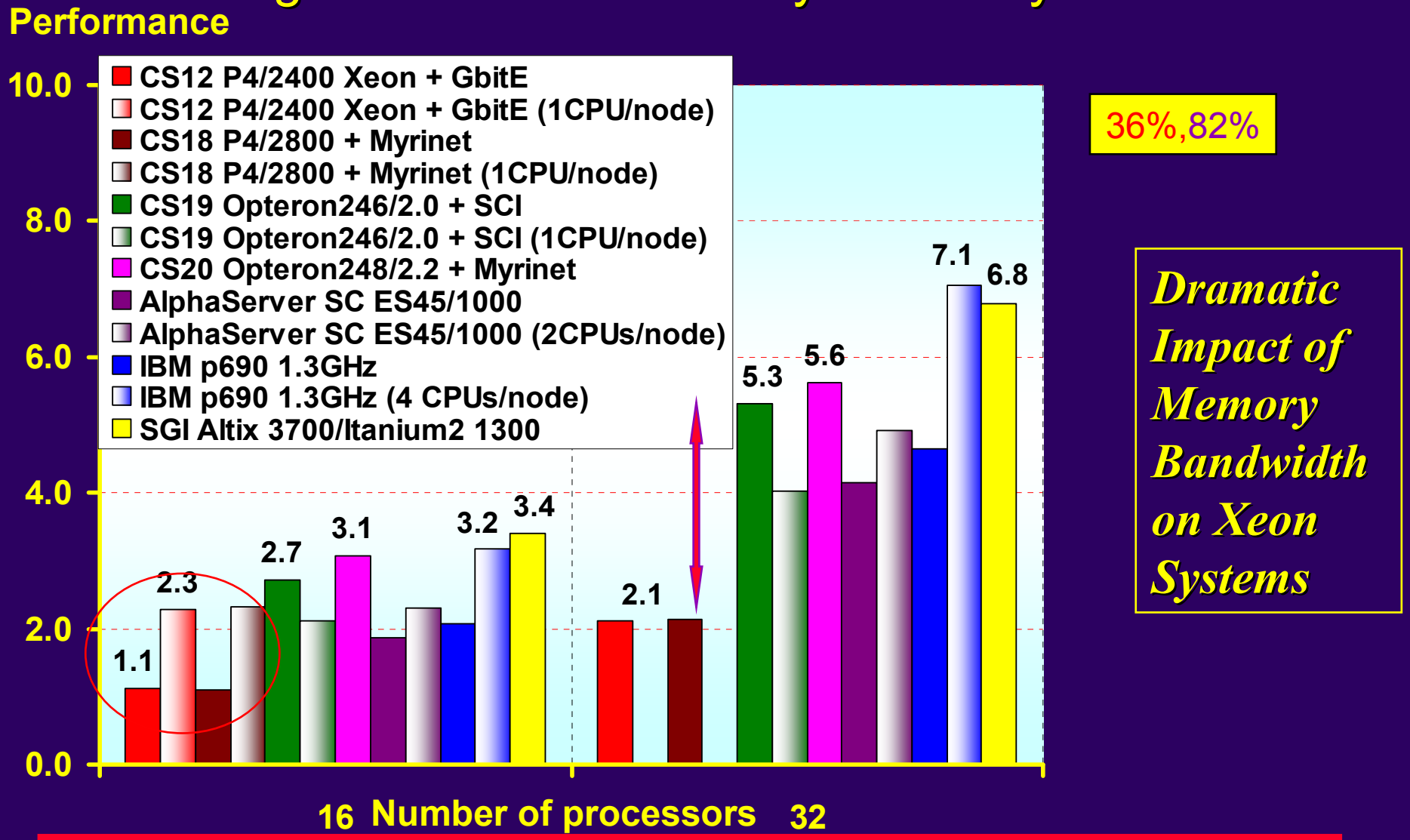
Measured Time (seconds)

Direct Numerical Simulation: 360^3 benchmark



DNS: 120^3 benchmark.

High-end and Commodity-based Systems



36%,82%

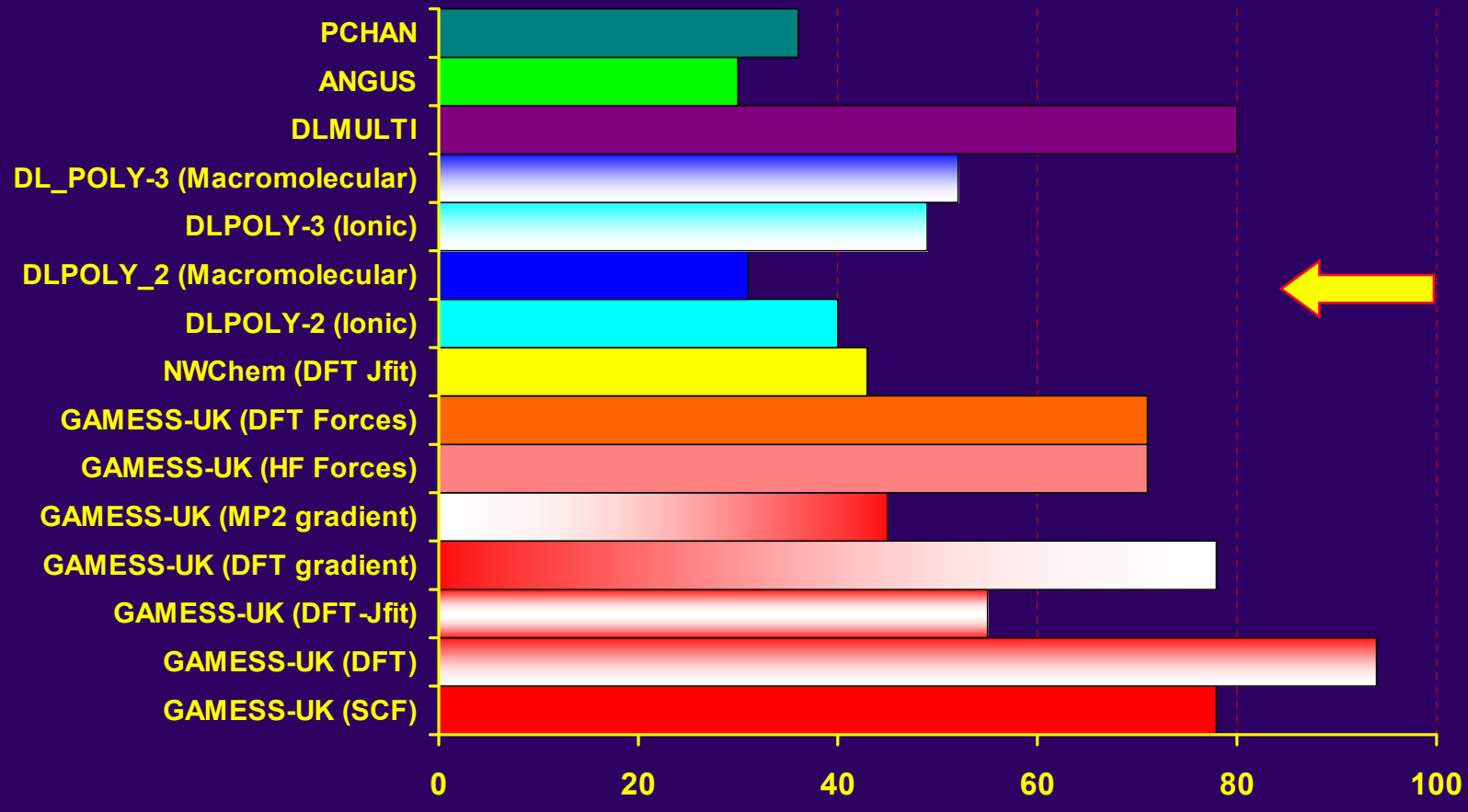
Dramatic Impact of Memory Bandwidth on Xeon Systems

Commodity Comparisons with High-end Systems

% of 32 CPUs of SGI Altix 3700 / 1300

Cluster CS18
Pentium4/2800 Xeon + Gbit Ethernet

CS18 - 49% of Altix

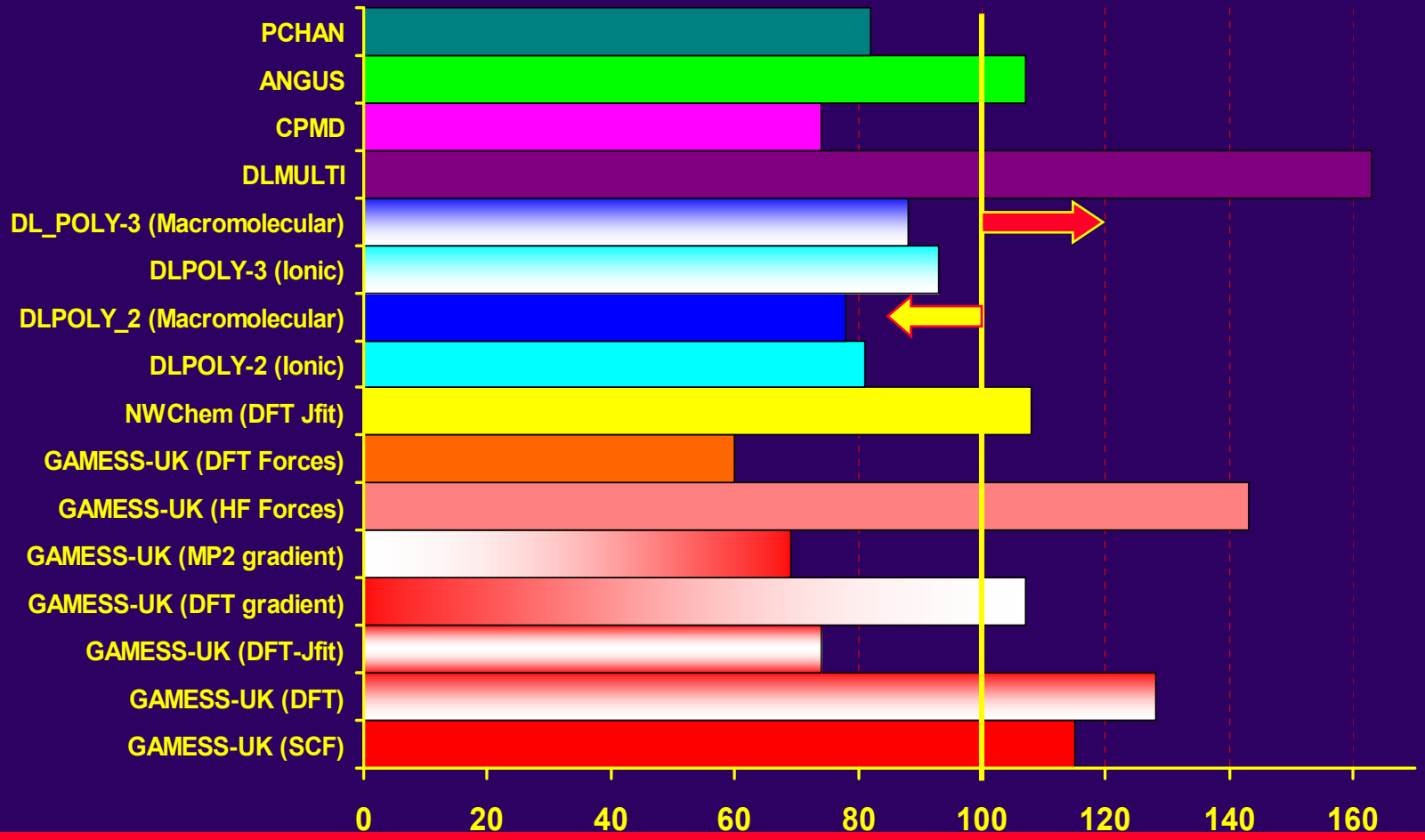


Commodity Comparisons with High-end Systems

% of 32 CPUs of SGI Altix 3700 / 1300

Cluster CS20
AMD Opteron 248/2.2 + Myrinet

CS20 - 100% of Altix



Summary

- Commodity-based and High-end Systems
  - Commodity Systems under evaluation ; CS1 – CS20
  - High-end systems from SGI, IBM and HP/Compaq
  - Serial CPU and Communications Performance; Evaluation Metrics
- Application performance
  - Molecular Simulation (replicated & distributed data)
    - DL_POLY, DLMULTI and CHARMM
  - Molecular Electronic Structure & Materials Simulation
    - NWChem, GAMESS-UK and CPMD
  - Computational Engineering
    - ANGUS, SBLI
- Notable Performance of Opteron clusters with Myrinet:
  - CS20 Opteron248/2200 cluster with Myrinet delivers on average 100% of SGI Altix 3700/1300.
  - Cost effectiveness of the Clusters (e.g., CS18) with Gbit Ethernet - delivers 49% of the SGI Altix 3700/1300 .