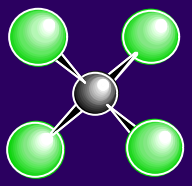
Benchmarking Linux Clusters: Application Performance on High-End and Commodity-class Computers



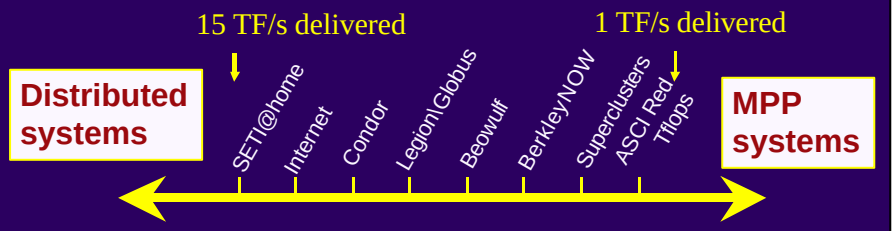
1. SINGLE PROCESSOR BENCHMARKS: Performance of Various Computers in Computational Chemistry
2. APPLICATION PERFORMANCE on HIGH-END and COMMODITY-TYPE COMPUTERS

Martyn F. Guest
Computational Science and Engineering Department
m.f.guest@dl.ac.uk

Outline

- Cost-effective high-end, departmental and desktop computation
  - Cluster Computing - where do clusters fit?
  - Background to Cluster Computing at Daresbury Laboratory
- Commodity-based Systems
  - **Single-node performance** & the **Interconnect bottleneck**
  - Prototype Systems; CS1 - CS7
- High-end Systems
  - Cray T3E/1200E, IBM SP/WH2-375, SGI Origin 3800, Compaq Alpha Server SC and Cray SuperCluster
- Application performance on High-end and Commodity Clusters
  - Electronic Structure and Molecular Simulation (next lecture)
  - Materials Simulation
    - **CRYSTAL, CPMD & CASTEP**
  - Engineering (ANGUS & FLITE3D)
- Application performance analysis
  - VAMPIR and instrumenting the GA Tools

Cluster Computing -Where Do Clusters Fit?



- Gather (unused) resources
- System SW manages resources
- System SW adds value
- 10% - 20% overhead is OK
- Resources drive applications
- Time to completion is not critical
- Time-shared
- Commercial. PopularPower, United Devices, Centrata, ProcessTree, Applied Meta, etc.

- Bounded set of resources
- Apps grow to consume all cycles
- Application manages resources
- System SW gets in the way
- 5% overhead is maximum
- Apps drive purchase of equipment
- Real-time constraints
- Space-shared

Capability Computing - Cost Effectiveness & Dependencies

Specification	Usage	Cost Units	CPU	Memory	I/O
MPP/ASCI 1280 CPUs, IBM SP (1 TB)	HPC community	10,000	2,000	2,000	200-300
	Capability	Commodity Systems (1.5 x N ??)			
SMP 16-processor SGI Origin 3800, R14k (8GB RAM)	Department	100	15	20	20-30
	Capacity	Commodity Systems (1.5 x N)			
PC Pentium-4 / 2GHz (512 MByte, 30 GB)	Desktop	1	1	1	1

3 year continual access to 32 CPUs:
High-end (£0.5 / CPU hour) : £419,358
in-house Beowulf : £50,000

CLRC's DisCo Programme

The Distributed Computing Support Programme (DisCo) at Daresbury covers a variety of activities:

- Courses
  - Graphics and Visualisation
  - System Management (Compaq/DEC, SGI, IBM, SUN, HP and Linux)
  - Fortran 90
- Newsletters (HPCGrid)
- Machine Evaluation Workshop (MEW)
- Information Services
  - Anonymous FTP/ WWW software archive
- Telephone and e-mail support
- Surveys and User Database
- Technical Developments (Beowulf systems)

http://www.cse.clrc.ac.uk/Activity/Disco

Daresbury Beowulf Project

High End

Departmental

Objective I

- **Inexpensive and robust way to "build" a 32-node commodity based ~20 Gflop server (IA32 / IPF / Alpha) for < £100K. Scalability is desired to 128 node systems;**
- **Add commodity based 10 TByte archive for < £100K.**
  - **Costs: expected to decrease by a factor of two every 2 years**
- **Departmental resource (JREI/JIF - EPSRC funded)**
- **Clustered Application solutions**
- **GRID Demonstrator**

Objective II

- **Extend developments of I to demonstrate scalability to a 512-node system. Prototype a 128-node commodity based ~100 Gflop server (Alpha/IA32/IPF)**
- **Demonstrate both metacomputing and visualisation potential through collaborative partnership.**
- **Support a broad range of applications, with a focus on computational chemistry.**
- **High-end resource that will be competitive with current high-end systems and ASCI SMP clusters.**

High-end Commodity-based Systems

"Need for real workloads on large clusters. To date mainly experimental systems, so failures are viewed as inevitable", but ...

Cluster	Site	Nodes/Processors	Peak (Gflop)	CPU	Interconnect
1. Locus	Locus, USA	708/1416	1416	PIII/1000	FastEth.
2. Biopentium	Inpharmatica	800/1220	1061	PIII/700+	FastEth.
3. Genesis	Shell, NL	1030/1038	1037	PIII/1000	GigabitE
4. Platinum	NCSA	516/1032	1032	PIII/1000	Myrinet
5. RHIC Brookhaven		638/1276	991	PIII/PII.	FastEth.
6. CBRC Magi	AIST,Japan	520/1040	967	PIII/933	Myrinet
7. Score III	RWCP, Japan	512/1024	955	PIII/933	Myrinet
8. ICE Box	Utah, USA	303/388815		PIII + AMD	FastEth.
9. Incyte	Incyte, USA	767/1511	754	PIII/450+	GigabitE
10. CPLANT	Sandia	628	628	EV6/500	Myrinet

Technical Progress in 2000/2001

Hardware and Software Evaluation:

- CPU
  - PC systems - Intel 1 GHz Pentium III, Pentium 4 (2.0 GHz), AMD K7 1.4 GHz
  - *Itanium (SGI Troon, HP RX4610)*
  - Alpha systems - DS20E/667, ES40/833, CS20/833, UP2000 (667 & 833 MHz)
- Networks
  - Fast Ethernet options, cards, switches, channel-bonding, ….
  - 100Mbit switch,
  - SCI, QSNet and Myrinet interconnect (Cray Supercluster, CS20)
- System Software
  - message passing S/W (LAM MPI, LAM MPI-VIA (100 us to 60 us), MPICH),libraries (ATLAS, NASA, MKL), compilers (Absoft, PGI, Intel's ifc, GNU/g77), GA tools (PNNL)
  - resource management software (LobosQ, PBS, Beowulf, **LSF** etc.)

Joint Research Equipment Initiative (JREI)

- Analysis of JREI'2000 (no limit) and JREI'2001 (£200K limit)
- Previous JREI competitions dominated by SMP servers
  e.g. SGI - Origin 2000, Onyx2, VR.
- Emerging role of Beowulf Systems ...

Equipment Category	Number of Proposals JREI'2000	JREI'2001
Proprietary systems	16	18
Proprietary Clusters	1	2
IA32/AMD Clusters	2	13
VR, visualisation	5	3
TOTAL	24	36

Benchmarking Linux Clusters:
SINGLE PROCESSOR PERFORMANCE:

Outline

- SPEC (Standard Performance Evaluation Corporation)
  - SPEC 95 and SPEC CPU 2000
- Single Processor Benchmarks (i.e. serial) - SPECfp ?
  - Matrix and application "kernels"
  - Application packages (GAMESS-UK, DL_POLY)
  - Comparison involves over 150 computers (supercomputers, workstations, PCs and MPP nodes)
  - URLs:
    - Powerpoint Presentation
      - http://www.dl.ac.uk/TCSC/disco/Benchmarks/ppoint/index.htm
    - Paper
      - http://www.dl.ac.uk/TCSC/disco/Benchmarks/paper/compchem.html

SPARSE MMO FORTRAN BENCHMARK, 1988-2001

(R = A X B; Single CPU Performance, seconds)

1988

Sparsity in B matrix	0%	50%
VAX 8200 (VMS V4.5)	1152	593.1
GOULD NP-1 (scalar)	148	76.0
FPS-164 (OPT=3)	60.0	32.8
HP/Apollo DN10020	44.2	22.8
CONVEX C-120	33.0	20.6
ALIANT FX/8 (3CES)	32.0	18.9
FPS-264 (OPT=3)	15.9	8.6
Cyber-205(OPT=DPRS)	19.2	10.2
IBM 3090-150 VF	11.3	6.0
Cray 1S (COS 1.14)	4.8	2.9
Cray XMP/4 (COS 1.15)	2.8	1.8

2001

Sparsity in B matrix	0%	50%
DEC Alpha PWS 433	1.4	0.7
Pentium III/550	1.1	0.6
DEC Alpha 8400/5-625	1.0	0.5
IBM RS/6000-397	0.8	0.4
SGI Origin2000/250	0.8	0.4
Cray T3E/1200	0.6	0.3
SGI Origin2000/R12k-400	0.4	0.3
SGI Origin3800/R14k-500	0.4	0.2
Compaq Alpha ES40/833	0.3	0.2
HP PA/9000 J6000/552	0.3	0.2
IBM RS/6000 44P/270	0.2	0.1
Compaq Alpha ES45/1000	0.2	0.1
Pentium 4/2000	0.1	0.1
IBM p-Series 690	0.1	0.0
FUJITSU VPP-300	1.2	0.8
CRAY Y-MP C98/4256	0.9	0.6
NEC SX-5	0.2	0.2

MACHINES UNDER EVALUATION

Machine	Processor
AMD Athlon	KZ - 1.0, 1.2, 1.4 GHz
Pentium - PC	Pentium II - 400, 450, 500, 550
	Pentium III - 850, 866, 1 GHz
	Pentium 4 - 1.4, 1.5, 2.0 GHz
SUN Ultra80/450	UltraSPARC-2 / 450 MHz
SUN 1000 Model 1750	UltraSPARC-3 / 750 MHz
SUN Fire 6800 / 900 Cu UltraSPARC-3 / 900 MHz	
HP PA-9000 / J5000	PA8500 / 440 MHz
HP PA-9000 / J6000	PA8600 / 552 MHz
HP PA-9000 / J6700	PA8700 / 750 MHz
HP RX4610	Itanium / 733 MHz (2 MB L3)
Compaq AlphaServer DS20	AXP A21264 / 500 MHz
Compaq AlphaServer DS20E	AXP A21264 / 667 MHz
Compaq AlphaServer ES40	AXP A21264A / 667 MHz
API UP2000 6/833	AXP A21264A/833 MHz (4Mb L2)
Compaq AlphaServer ES40	AXP A21264A / 833 MHz
Compaq AlphaServer ES45	AXP A21264C / 1000 MHz

Machine	Processor
SGI Origin3800/R14k	R14000/R14010 500 MHz
SGI Origin3800/R12k	R12000/R12010 400 MHz, SGI Origin2000/R12k R12000/R12010 400 MHz, SGI Origin2000/R12k R12000/R12010 300 MHz
SGI Octane2/ R12k	R12000/R12010 400 MHz
SGI O2/R12k-SC	R12000/R12010 270 MHz
IBM RS/6000-44P/270	RS/6000 Power3 375 MHz
IBM p-Series 690	RS/6000 /POWER4 1.3GHz

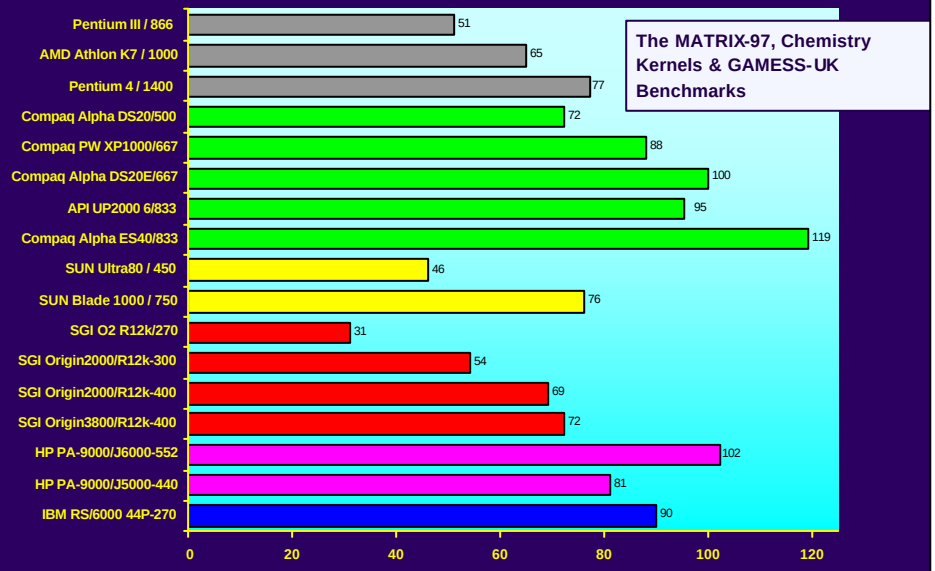
Supercomputers

NEC SX-5, SX-4	Cray Y-MP/J90-10,
FUJITSU VPP/300,	Cray YMP C98/4256

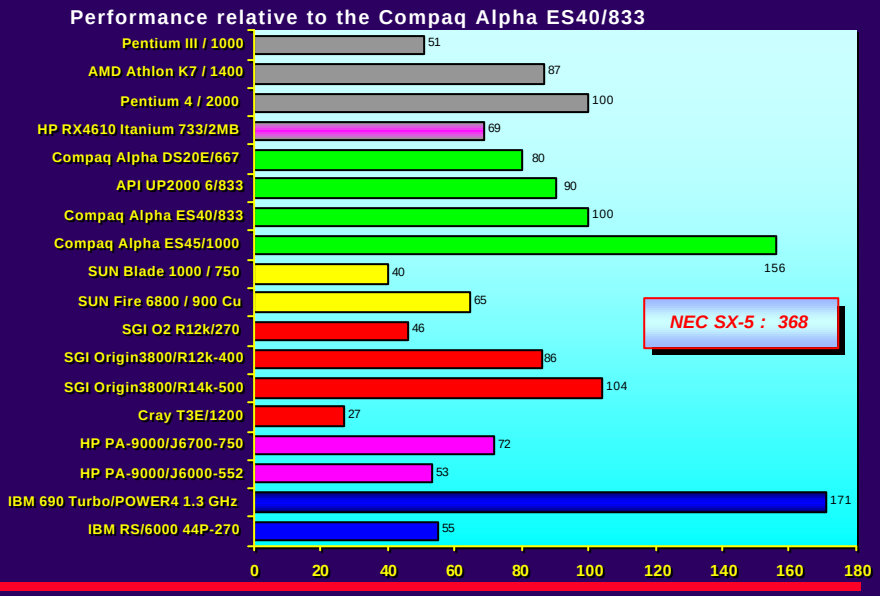
MPP Nodes

Cray T3E/1200	AXP EV56 600 MHz
IBM SP2 / P2SC	P2SC 120, 160 MHz
IBM SP / Power3	RS/6000 WH2 - 375 MHz
	RS/6000 NH - 222 MHz

MEW11 - Summary PI relative to the Compaq Alpha DS20E/667



The Whetstone-97 Benchmark.



SPEC Benchmarks

NOT:
- Graphics
- Network
- I/O

- Compute intensive categories
  - integer versus floating point
  - conservative versus aggressive compilation
  - speed versus throughput
- Composite Metrics - SPEC95

http://www.specbench.org

	SPEED	THROUGHPUT
Aggressive		
	SPECint95	SPECint_rate95
	SPECfp95	SPECfp_rate95
Conservative		
	SPECint_base95	SPECint_rate_base95
	SPECfp_base95	SPECfp_rate_base95

- SPECratio - T(measured system) / T (reference)
- Reference = Sun SPARCstation 10/40
- SPECfp95 - geometric mean of 10 ratios, one for each benchmark
- SPECint95 - geometric mean of 8 ratios, one for each benchmark

SPEC95 Products - Floating point Benchmark Suite (SPECfp95)

Benchmark	Reference † Time (secs)	Application Area
101.tomcatv	3600	Fluid Dynamics / Geometric translation
102.swim	8600	Weather Prediction
103.su2cor	1400	Quantum Physics
104.hydro2d	2400	Astrophysics
107.mgrid	2500	Electromagnetism
110.applu	2200	Fluid Dynamics
125.turb3d	4100	Turbulence simulation
141.apsi	2100	Weather prediction
145.fpppp	9600	Computational Chemistry
146.wave	3000	Electromagnetics

† SUN SPARCstation 10/40

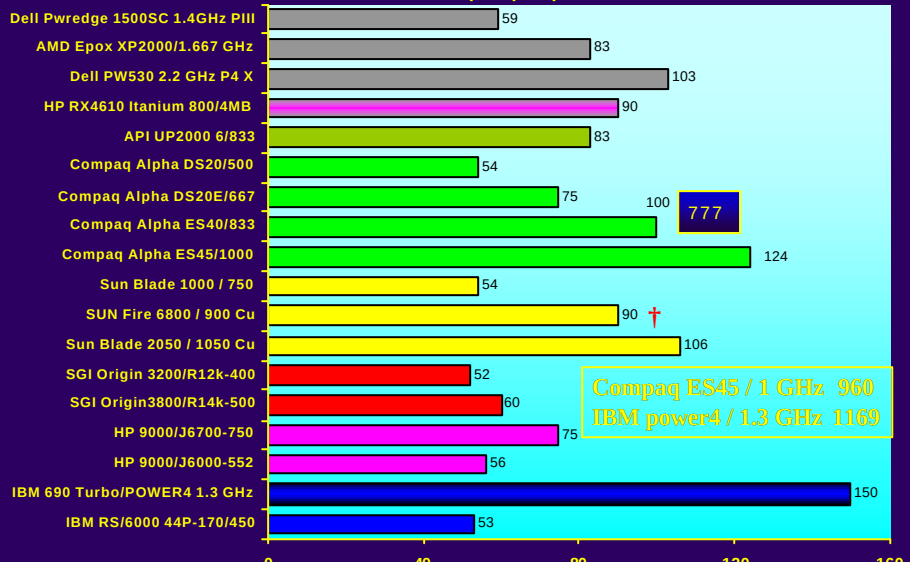
SPEC CPU 2000 - Floating point Benchmark Suite (SPECfp2000)

Benchmark	Language	Description
168.wupwise	F77	Physics: Quantum chromodynamics
171.swim	F77	Shallow water modelling
172.mgrid	F77	Multigrid solver: 3D potential field
173.applu	F77	Partial differential equations
177.mesa	C	3D graphics library
178.galgel	F90	Computational fluid dynamics
179.art	C	Image recognition / neural networks
183.equake	C	Seismic wave propagation simulation
187.facerec	F90	Image processing: Face recognition
188.ammp	C	Computational chemistry
189.lucas	F90	Number theory / primality testing
191.fma3d	F90	Finite-element crash simulation
200.sixtrack	F77	Nuclear physics accelerator design
301.apsi	F77	Metereology: Pollutant distribution

Reference: 300 MHz Ultra 5/10 = 100

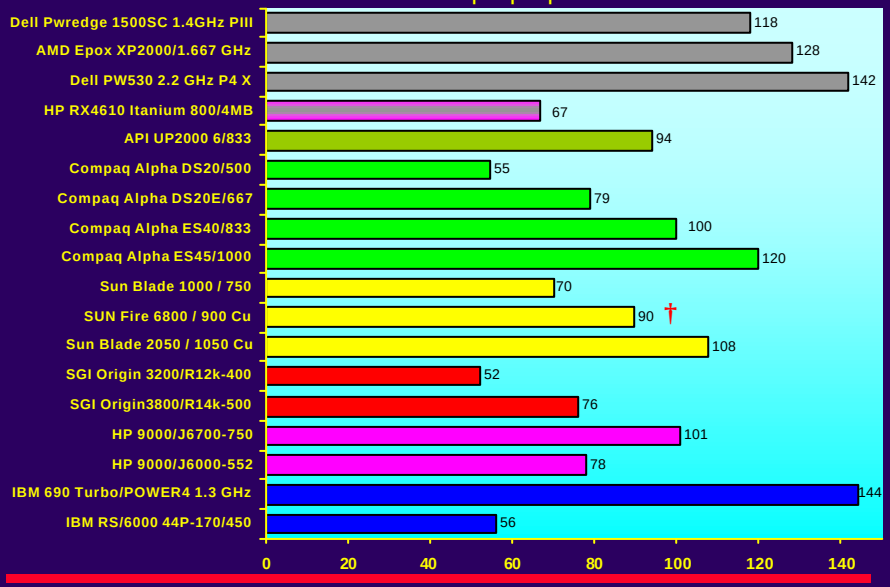
SPEC CPU 2000 - SPECfp2000

Values relative to Compaq Alpha ES40/833



SPEC CPU 2000 - SPECint2000

Values relative to Compaq Alpha ES40/833



Single Processor Benchmark Suite

- Benchmark suite developed to incorporate:
  - Matrix “kernels” (MATRIX-89 and MATRIX-97)
  - Application “kernels”
  - Application packages (e.g. GAMESS-UK, DL_POLY)
- Implemented on Supercomputers, servers (superminis), workstations, PCs and parallel machines
  - Matrix Operations / Matrix multiplication and matrix diagonalisation
  - Computational Chemistry Kernels - four typical application kernels (direct-SCF, MD, QMC and Jacobi eigen solver)
  - STREAM (memory bandwidth)
  - Quantum Chemistry Calculations - twelve typical applications, including SCF, direct-SCF, CASSCF, MCSCF, direct-CI and MRD-CI, MP2, 2nd derivatives (GAMESS-UK-89 and GAMESS-UK-99)
  - Molecular Dynamics Calculations - six typical simulations

SINGLE PROCESSOR BENCHMARKS I. Matrix Operations

MATRIX OPERATIONS (MATRIX-89 and MATRIX-97)

- SPARSE Matrix Multiply BenchMark
  - MMO operation is central to the efficient operation of modern QC codes. In this benchmark a series of MMOs (R = A X B) are performed involving matrices of increasing order:
    - MATRIX-89: 10, 20, 30, … , 100 (B is sparse)
    - MATRIX-97: 50, 100, 150, … , 500 (B is sparse)

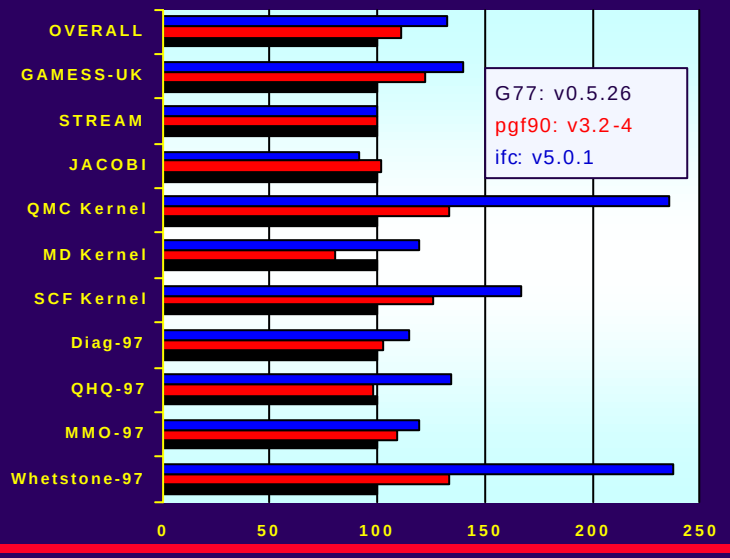
Memory bandwidth

- Diagonalisation Benchmark
  - Based on diagonalising a series of real symmetric matrices. Measures the performance of 8 routines from mathematical libraries and QC codes:
    - MATRIX-89: 10, 20, 30, … , 100
    - MATRIX-97: 50, 100, 150, 200, 250, 300
- Q†HQ Benchmark
  - Designed to extend MMO benchmark by allowing for the use of library routines e.g. BLAS. Uses both a scalar and vector algorithm:
    - MATRIX-89: 10, 20, 30, … , 150
    - MATRIX-97: 20, 40, 60, … , 300

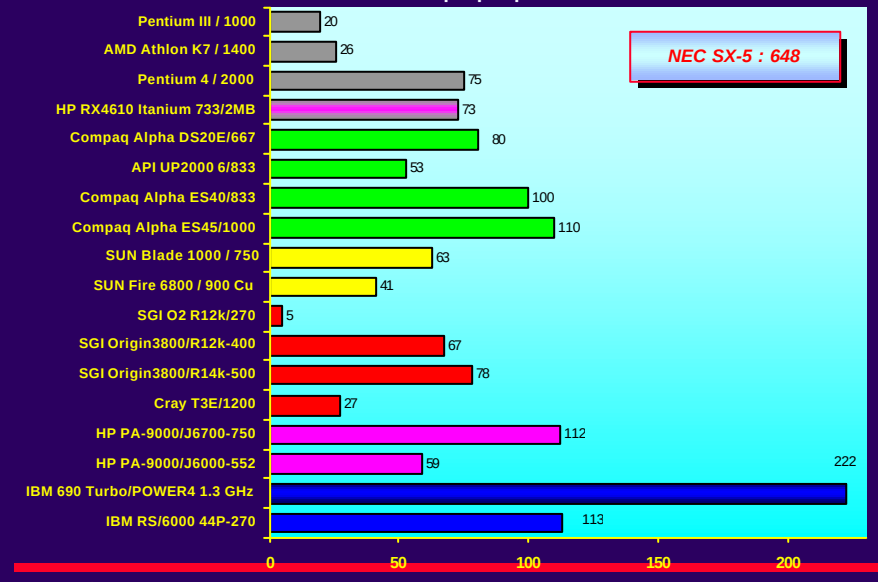
-lblas

Fortran Compilers - Performance

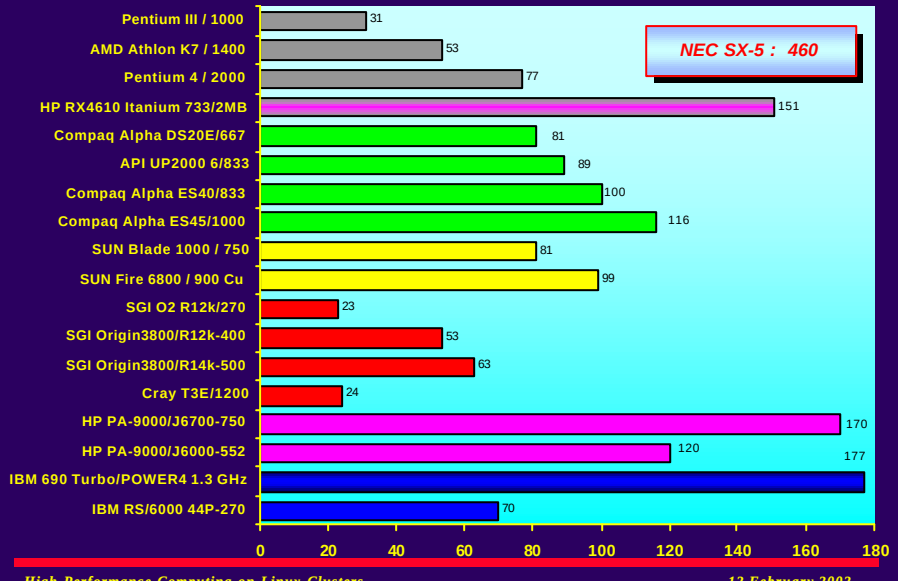
Performance relative to GNU g77 on Pentium 4/2000



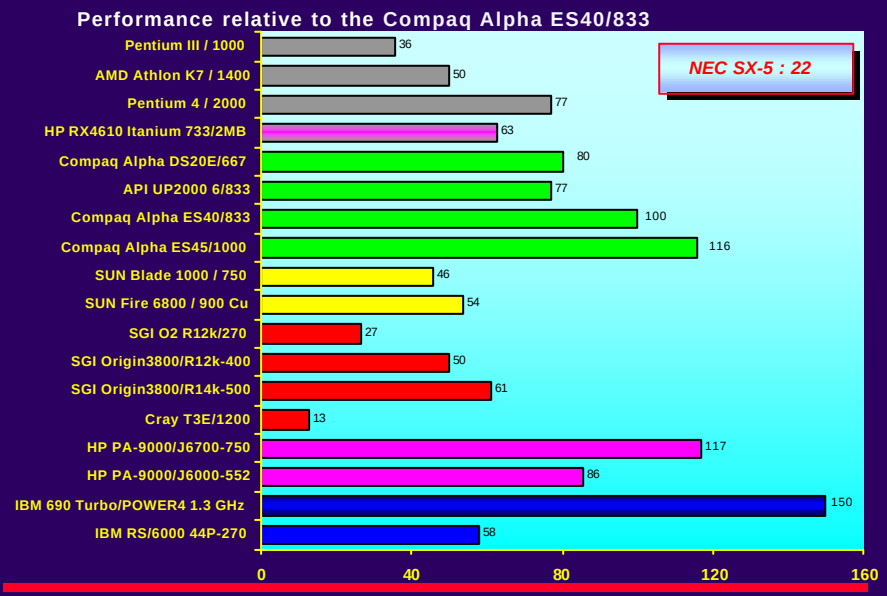
Matrix-97: SPARSE MMO Benchmark.



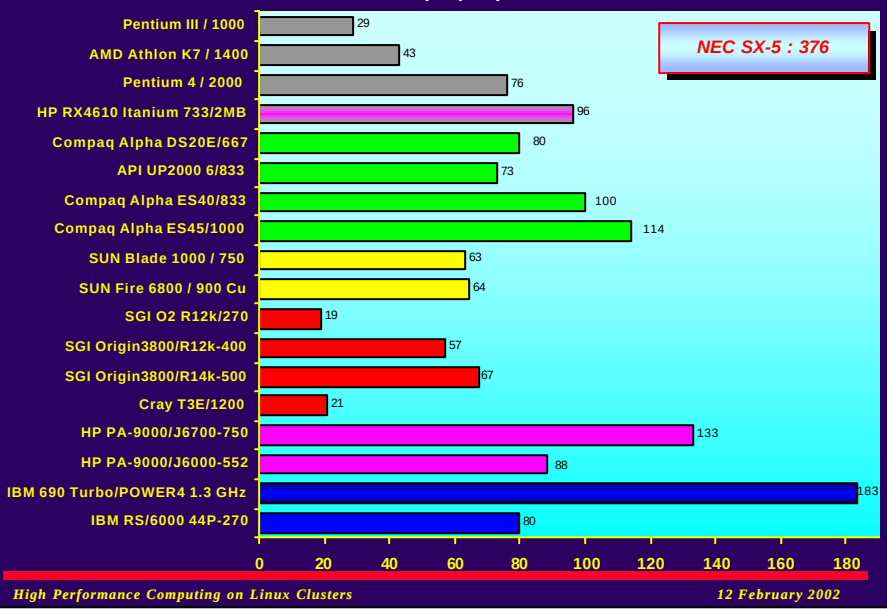
Matrix-97: Q†HQ MMO Benchmark.



Matrix-97: Diagonalisation Benchmark



The Matrix-97 Benchmarks.



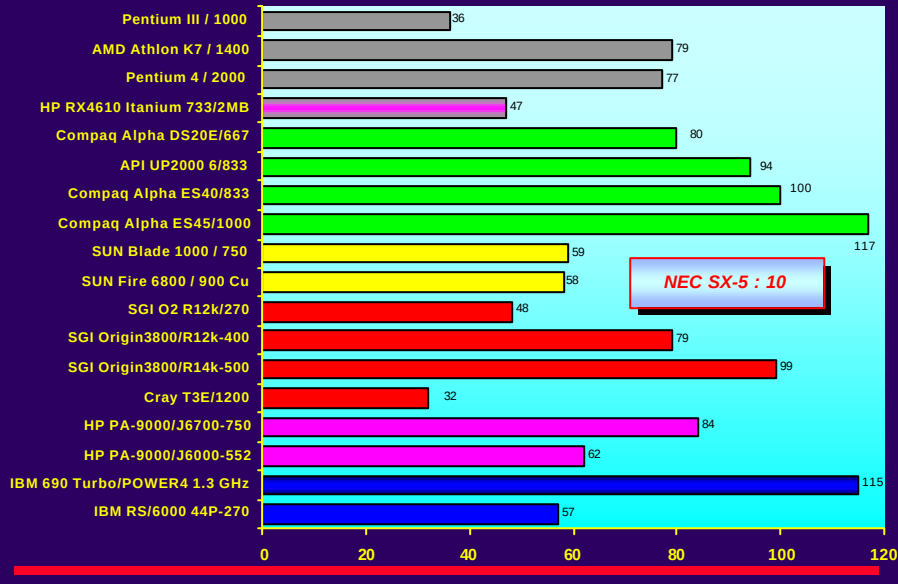
COMPUTATIONAL CHEMISTRY KERNELS

Programs that are realistic models of actual chemical applications or algorithms

- Self consistent Field (SCF)
  - SCF code using distributed primitive 1s gaussian functions as a basis (thus emulating the use of s, p, functions); performs direct-SCF calculation on Be_4 (60 functions).
- Molecular Dynamics (MD)
  - This code bounces a few thousand argon atoms around in a box with periodic boundary conditions. LJ pair-wise interactions are used with integration of the Newtonian equations of motion.
- Quantum Monte Carlo (QMC)
  - This code evaluates the energy of the simplest explicitly correlated electronic wavefunction for the He atom using a variational monte-carlo method without importance sampling.
- Jacobi iterative linear equation solver (JACOBI)
  - JACOBI uses a naive jacobi iterative algorithm to solve a linear equation. All the time is spent in a large matrix-vector multiplication.

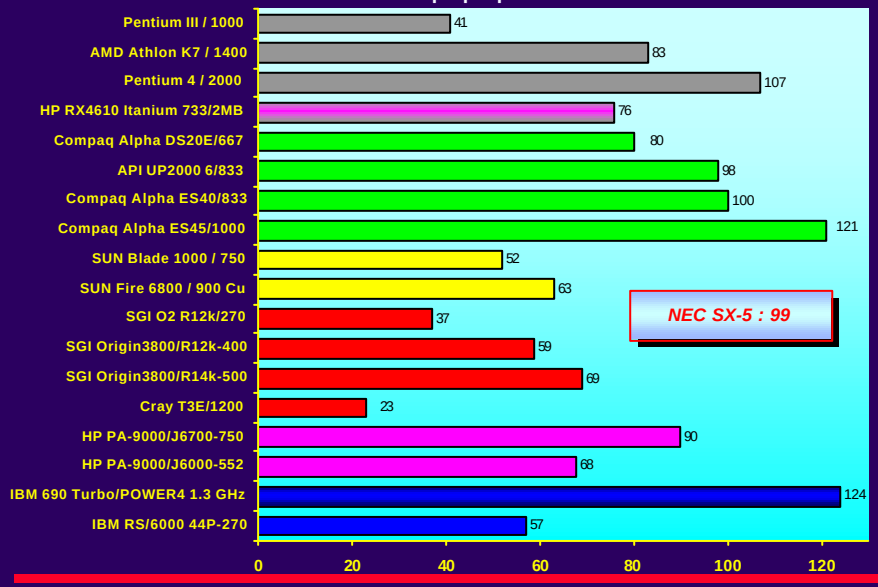
Computational Chemistry Kernels - Direct-SCF.

Performance relative to the Compaq Alpha ES40/833



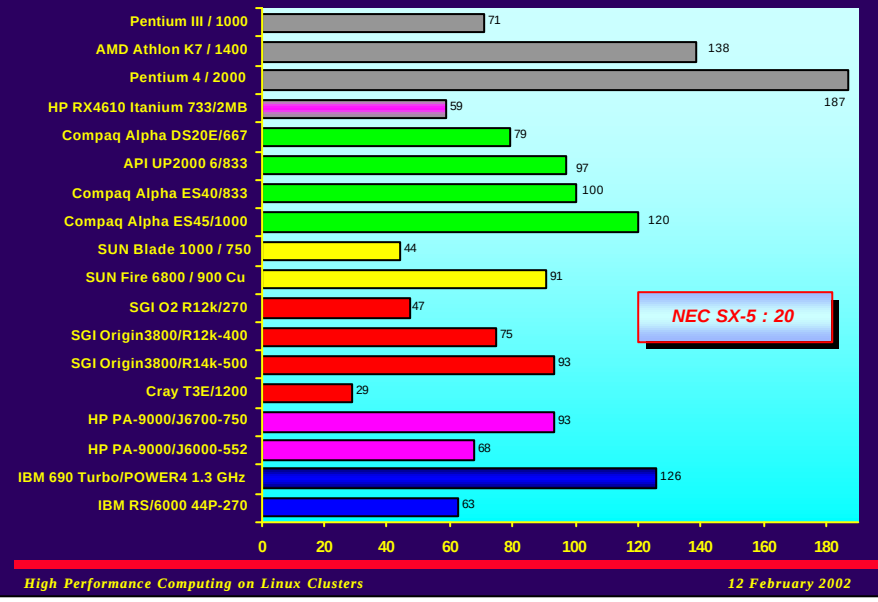
Chemistry Kernels - Molecular Dynamics.

Performance relative to the Compaq Alpha ES40/833



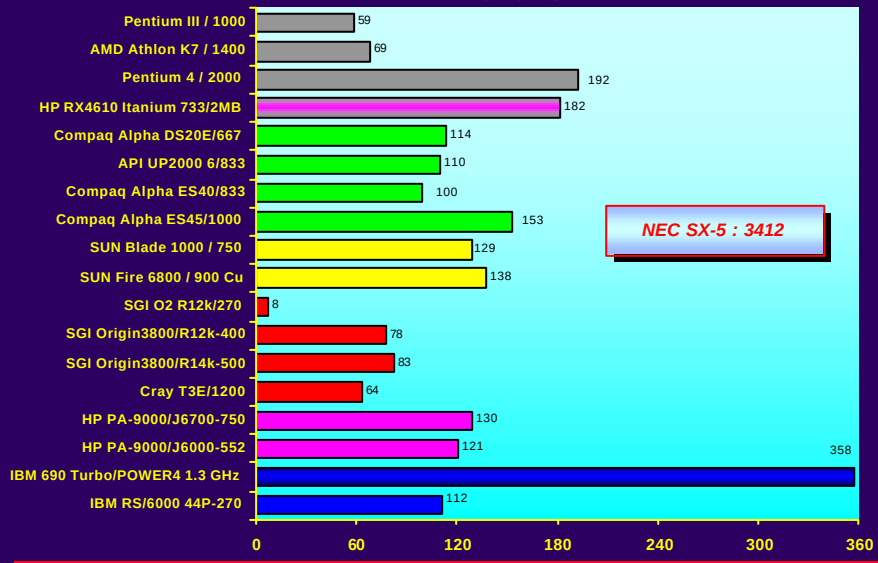
Chemistry Kernels - Quantum Monte Carlo.

Performance relative to the Compaq Alpha ES40/833

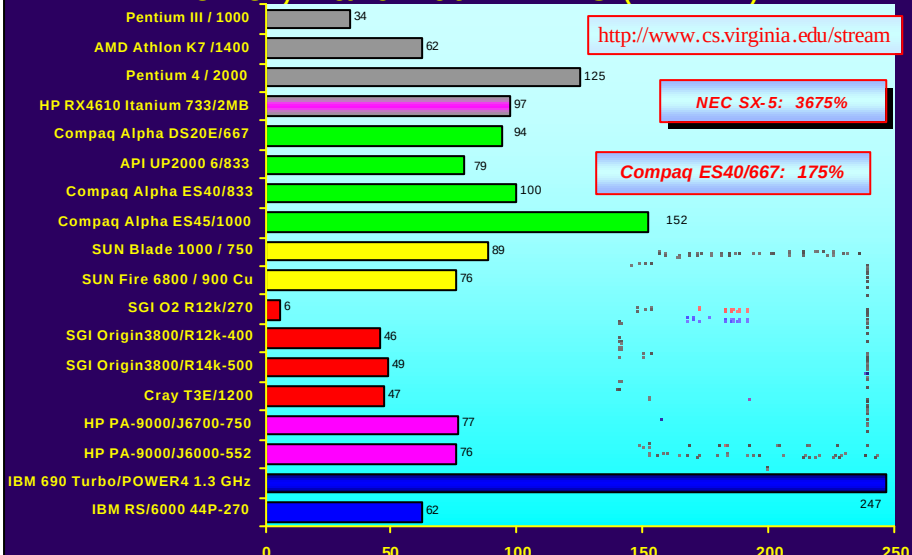


Chemistry Kernels - Jacobi Solver.

Performance relative to the Compaq Alpha ES40/833

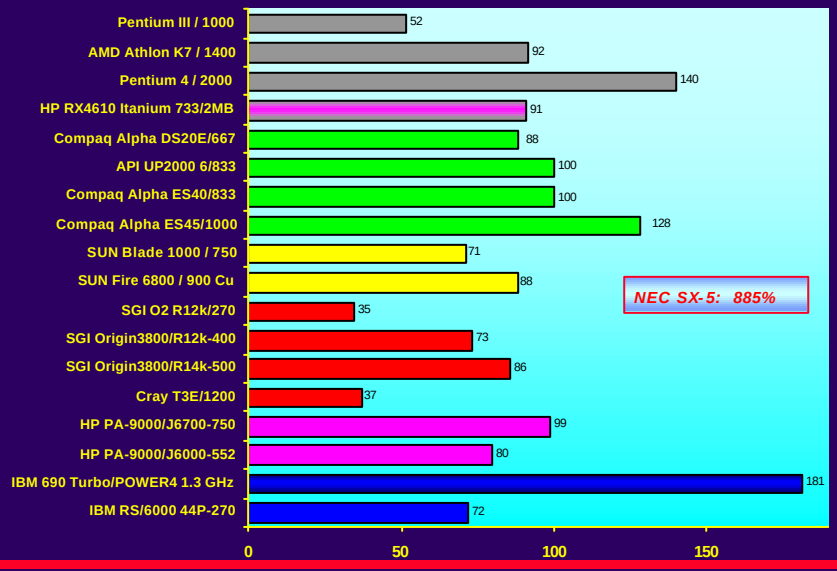


STREAM: Measured Sustainable Memory Bandwidth in HPC (TRIAD)



Computational Chemistry Kernels.

Performance relative to the Compaq Alpha ES40/833



Ab Initio MOLECULAR ELECTRONIC STRUCTURE.

Computational Activities

- Investigation of chemical reaction pathways and mechanisms;
- Computation of molecular equilibrium geometries or transition state structures;
- Harmonic vibrational analysis, prediction of vibrational frequencies and Infrared and Raman Intensities;
- Evaluation of molecular properties (multipole moments, electrostatic properties);
- Evaluation of ionization potentials and electronic excitation energies;
- Treatment of “Large” molecules;
  - direct-SCF techniques, ECPs
  - QM/MM, free energy perturbation calculations

MOLECULAR ELECTRONIC STRUCTURE SOFTWARE

- GAUSSIAN-98
  - Gaussian Inc. (Pittsburgh)
    - M.J. Frisch and co-workers
- CADPAC
  - R.D. Amos
    - Cambridge University
- MOLPRO
  - P.J. Knowles and H.J. Werner
    - Birmingham University
- GAMESS and GAMESS-UK
  - M.W. Schmidt, N. Dakota State
  - M.F. Guest at al., Daresbury
- SPARTAN
  - W. Hehre
    - University of California
- TURBOMOLE (MSI)
  - R. Ahlrichs
    - University of Karlsruhe
- QCHEM
  - M. Head Gordon et al
    - Berkeley
- NWChem
  - J. Nichols, R.J. Harrison ….
    - EMSL / PNNL

GAMESS-UK

GAMESS-UK is the general purpose ab initio molecular electronic structure program for performing SCF-, MCSCF- and DFT-gradient calculations, together with a variety of techniques for post Hartree Fock calculations.

- The program is derived from the original GAMESS code, obtained from Michel Dupuis in 1981 (then at the NRCC), and has been extensively modified and enhanced over the past decade.
- This work has included contributions from numerous authors†, and has been conducted largely at the CCLRC Daresbury Laboratory, under the auspices of the UK's Collaborative Computational Project No. 1 (CCP1). Other major sources that have assisted in the on-going development and support of the program include various academic funding agencies in the Netherlands, and ICI plc.

Additional information on the code may be found from links at:

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

† *M.F. Guest, J.H. van Lenthe, J. Kendrick, K. Schoffel & P. Sherwood, with contributions from R.D. Amos, R.J. Buenker, H.H. van Dam, M. Dupuis, N.C. Handy, I.H. Hillier, P.J. Knowles, V. Bonacic-Koutecky, W. von Niessen, R.J. Harrison, A.P. Rendell, V.R. Saunders, A.J. Stone and D.Tozer.*

DL_POLY: A Parallel Molecular Dynamics Simulation Package

- Developed as CCP5 parallel MD code by W. Smith and T.R. Forester
- UK + International user community
- Adopted by Materials Consortium 1995

Boundary Conditions

- None (e.g. isolated macromolecules)
- Cubic periodic boundaries
- Orthorhombic periodic boundaries
- Parallelpiped periodic boundaries
- Truncated octahedral periodic boundaries
- Rhombic dodecahedral periodic boundaries
- Slabs (i.e. x,y periodic, z nonperiodic)

Target Systems

- Atomic systems & mixtures (Ne, Ar, etc.)
- Ionic melts & crystals (NaCl, KCl etc.)
- Polarisable ionics (ZSM-5, MgO etc.)
- Molecular liquids & solids (CCl_4, Bz etc.)
- Molecular ionics (KNO_3, NH_4Cl, H_2O etc.)
- Synthetic polymers ($[PhCHCH_2]_n$ etc.)
- Biopolymers and macromolecules
- Polymer electrolytes, Membranes,
- Aqueous solutions, Metals

MD Algorithms/Ensembles

- Verlet leapfrog, Verlet leapfrog + RD-SHAKE
- Rigid units with FIQA and RD-SHAKE
- Linked rigid units with QSHAKE
- Constant T (Berendsen) with Verlet leapfrog and with RD-SHAKE
- Constant T (Evans) with Verlet leapfrog and with RD-SHAKE
- Constant T (Hoover) with Verlet leapfrog

GAMESS-UK and DL_POLY Benchmarks

GAMESS-UK

12 Typical QC Calculations

Module	Basis (GTOs)	Species
1. SCF	STO-3G (124)	Morphine
2. SCF	6-31G (154)	$C_6H_3(NO_2)_3$
3. ECP Geometry	ECPDZ (70)	Na_7Mg^+
4. Direct-SCF	6-31G (82)	Cytosine
5. CAS-geometry	TZVP (52)	H_2CO
6. MCSCF	EXT1 (74)	H_2CO
7. Direct-CI	EXT2 (64)	H_2CO/H_2+CO
8. MRD-CI (26M)	ECP (59)	$TiCl_4$
9. MP2-geometry	6-31G* (70)	H_3SiNCO
10. SCF 2nd derivs.	6-31G (64)	C_5H_5N
11. MP2 2nd derivs.	6-31G* (60)	C_4
12. Direct-MP2	DZP (76)	C_5H_5N

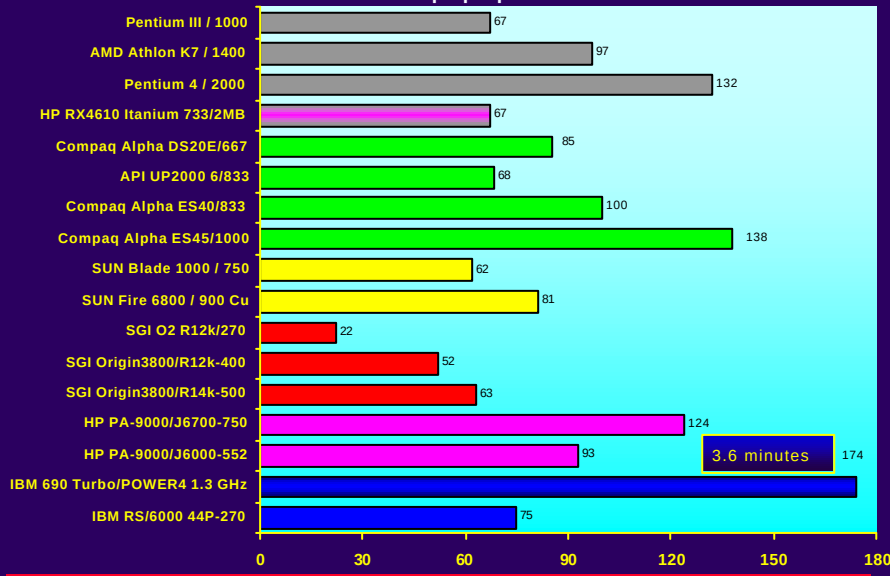
DL_POLY

Six Typical Simulations

Simulation	Atoms	Time steps
1. Na-K disilicate glass	1080	300
2. Metallic Al with Sutton-Chen potential	256	8000
3. Valinomycin in 1223 water molecules	3837	100
4. Dynamic shell model water with 1024 sites	768	1000
5. Dynamic shell model MgCl2 with 1280 sites	768	1000
6. Model membrane, 2 membrane chains, 202 solute and 2746 solvent molecules	3148	1000

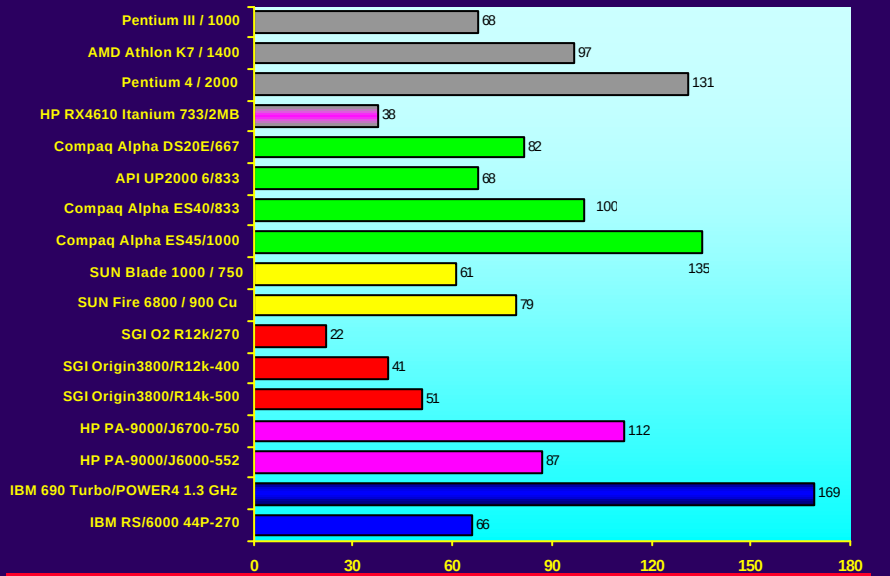
The GAMESS-UK Benchmark I. CPU

Performance relative to the Compaq Alpha ES40/833

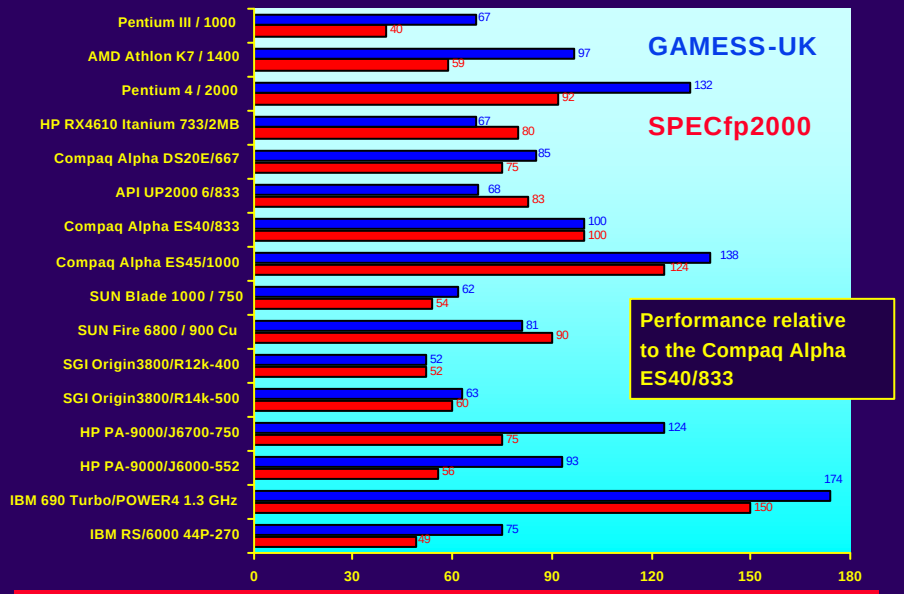


The GAMESS-UK Benchmark II. Wall

Performance relative to the Compaq Alpha ES40/833



SPECfp2000 and the GAMESS-UK Benchmark



The GAMESS-UK-99 Benchmark

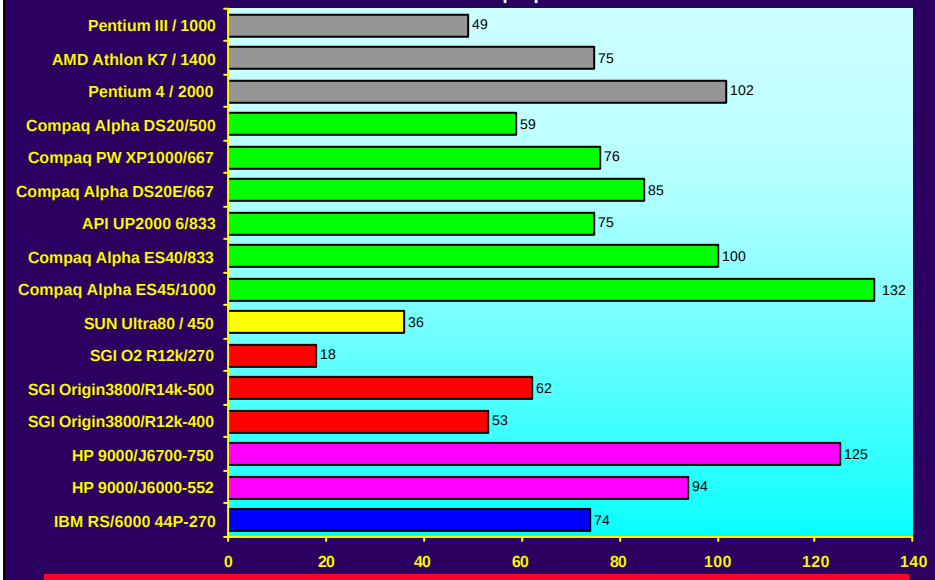
10 Typical QC Calculations

Module	Basis (GTOs)	Species
1. Direct- SCF	6-31G (227)	Morphine
2. SCF	6-31G** (265)	$C_6H_3(NO_2)_3$
3. DFT B3LYP	6-311G* (167)	Cytosine
4. MCSCF	CC-PVTZ (100)	H_2CO
5. Direct-CI	CC-PVTZ (100)	H_2CO/H_2+CO
6. CCSD(T)	TZV+2d+1f (144)	C_4
7. MP2-geometry	TZVP (105)	H_3SiNCO
8. SCF 2nd derivs.	6-311G** (144)	C_5H_5N
9. MP2 2nd derivs.	TZVP(C2d) (104)	C_4
10. Direct-MP2	6-31G* (130)	Cytosine

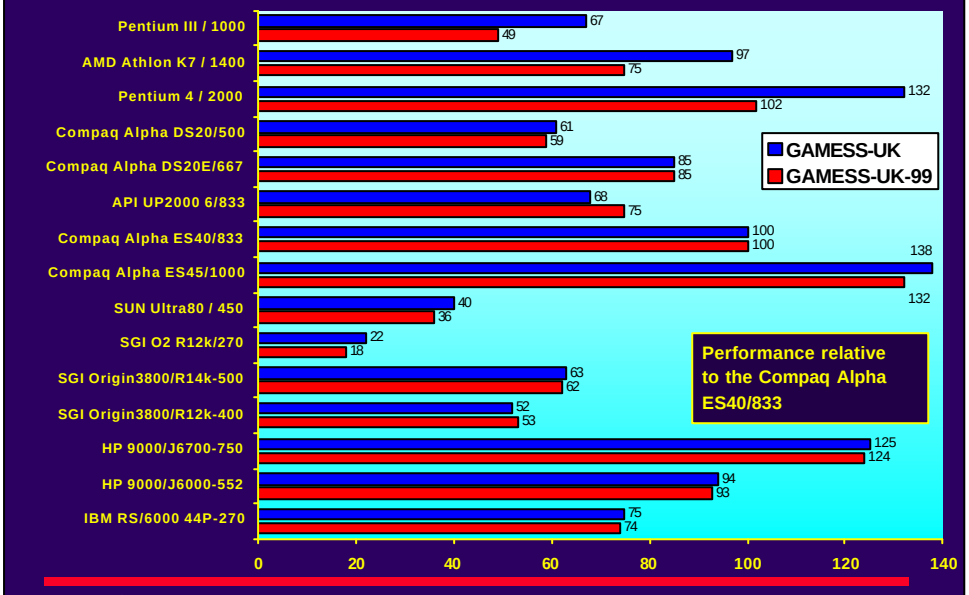
Benchmark Time: 49.9 minutes on Compaq AlphaServer ES40/833

The GAMESS-UK-99 Benchmark

Performance relative to the Compaq ES40/833

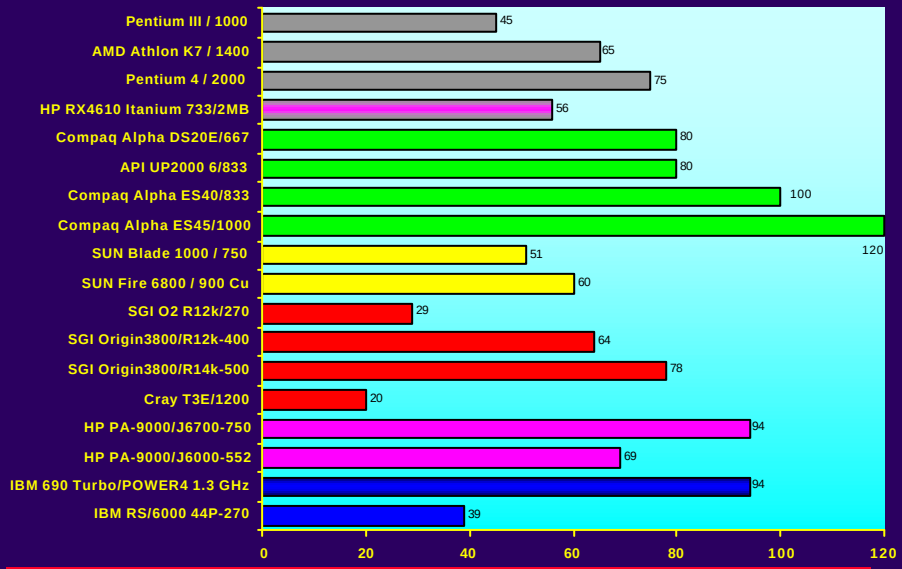


The GAMESS-UK and GAMESS-UK-99 Benchmark

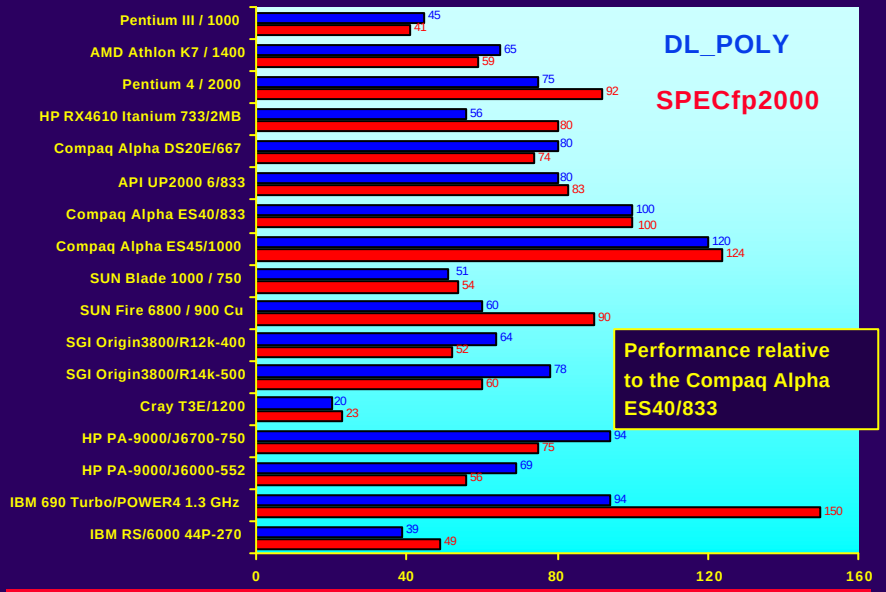


The DL_POLY Benchmark.

Performance relative to the Compaq Alpha ES40/833

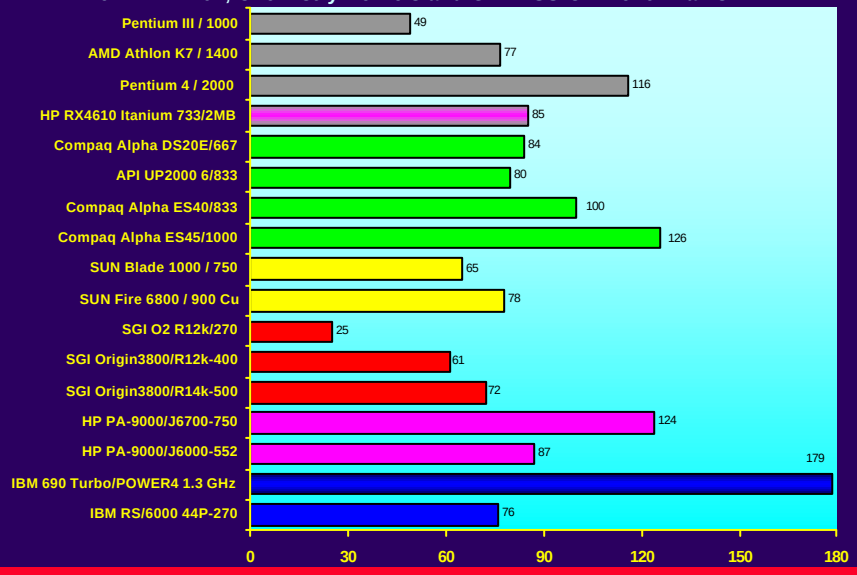


SPECfp2000 and the DL_POLY Benchmark

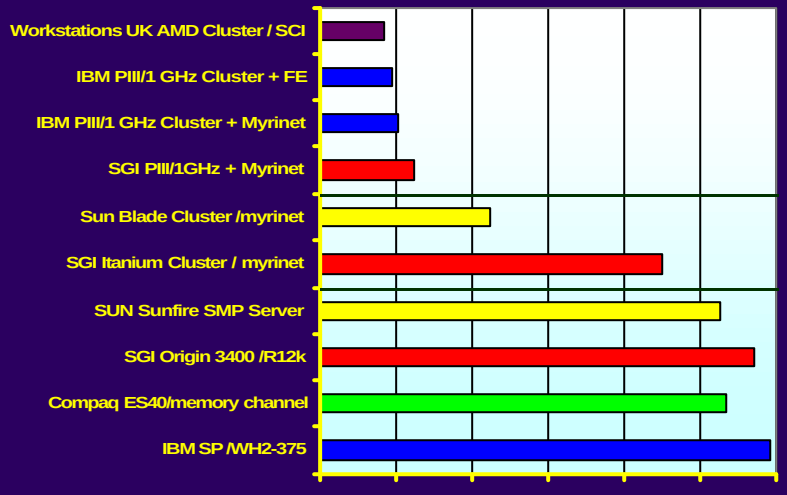


Summary Index relative to the Compaq Alpha ES40/833

The MATRIX-97, Chemistry Kernels and GAMESS-UK Benchmarks



Machine Costs: Proprietary & Commodity Solutions

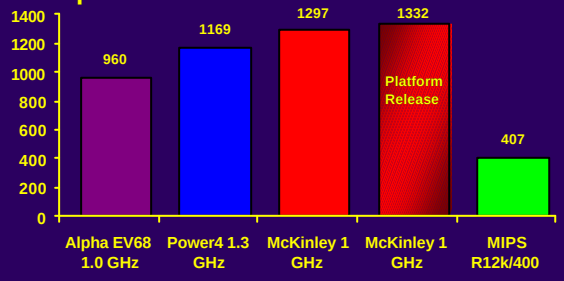


JREI 2001

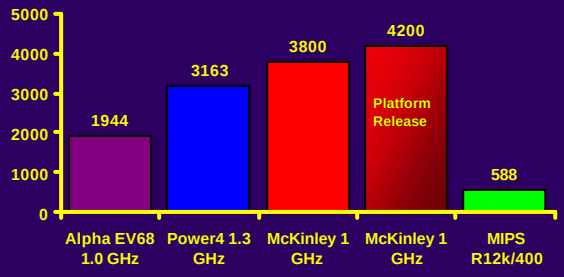
Cost / CPU

SPECfp2000 and STREAM

SPECfp2000



STREAM



SINGLE PROCESSOR BENCHMARKS: SUMMARY

- Distributed Computing Support (DisCo)
  - Summary of Activities
  - http://www.cse.clrc.ac.uk/Activity/Disco
- SPECfp and Computational Chemistry Benchmark (serial)
  - Comparison involves 150 computers (supercomputers, workstations, PCs and MPP nodes)
  - Matrix'89 and Matrix'97 kernels (MMO, diagonalisation)
  - Application "kernels" (SCF, MD, QMC and JACOBI + STREAM)
  - Application packages (GAMESS-UK, DL_POLY)
- Machine COSTS vs. Performance: URLs:
  - Powerpoint presentation
    - http://www.dl.ac.uk/TCSC/disco/Benchmarks/ppoint/index.htm
  - Paper
    - http://www.dl.ac.uk/TCSC/disco/Benchmarks/paper/compchem.html

Benchmarking Linux Clusters: Application Performance on High-End and Commodity-class Computers

http://www.dl.ac.uk/CFS/parallel/benchmarks

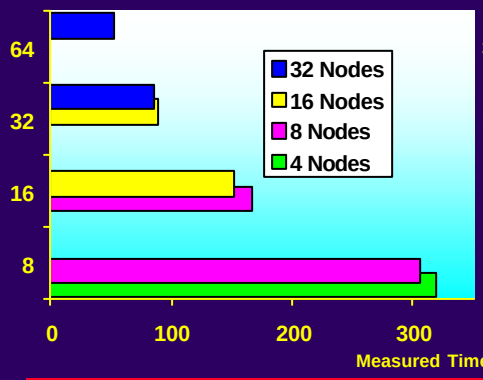
Single or Dual-Processors

- The decision to buy single or dual processor PCs is driven by the nature of the target application(s). Consider number of scenarios:
  - The application working set fits in main memory.
  - The application is characterised by demanding memory bandwidth requirements.
  - The application requires a large amount of disk I/O.
- Remember that both CPUs on a dual processor system must share the same network communications, potentially halving the communications bandwidth relative to a single CPU system.

Memory Bandwidth Effects

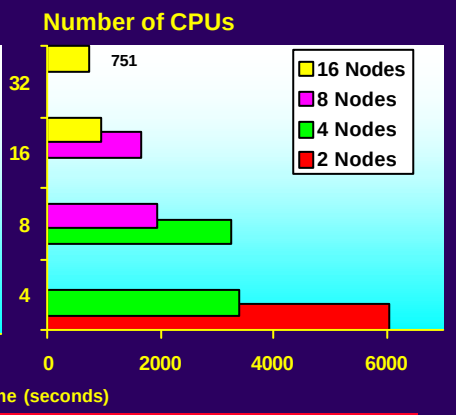
DL_POLY: Ewald Benchmark

CS7 AMD K7/1000 MP + SCI



ANGUS: Conjugate Gradient + ILU

CS2 QSNet Alpha Cluster



I/O Benchmark

Application has large I/O requirements - e.g. In periodic HF code, CRYSTAL, the conventional SCF calculation processes 2e-integrals file from disk.

Single and dual Intel Pentium II-266, 128MB SDRAM / CPU

	CPU time (s)	Elapsed T (min)	Efficiency (%)
Single CPU	390.22	11:26.43	56.8%
Dual CPU	407.45 + 409.04	30:40.49 + 30:53.02	22.1 + 22.0

Stripe the integrals file across 2 disks:

	CPU time (s)	Elapsed T (min)	Efficiency (%)
Dual CPU	425.65 + 423.48	12:10.48 + 10:30.59	58.2 + 67.1

Commodity Systems (CSx)
Prototype / Evaluation Hardware

Systems	Location	CPUs	Configuration
CS1	Daresbury	32	Pentium III / 450 MHz; fast ethernet (EPSRC)
CS2	Daresbury	64	24 X dual UP2000/EV67-667, QSNet Alpha/LINUX cluster, 8 X dual CS20/EV67-833
CS3	RAL	16	Athlon AMD K7 850MHz; myrinet interconnect
CS4	Sara	32	Athlon AMD K7 1.2 GHz; fast ethernet
CS6	CLiC	528	Pentium III / 800 MHz; fast ethernet (Chemnitzer Cluster)
CS7	Daresbury	64	AMD K7/1000 MP; SCALI SCI interconnect ("ukcp")
Protoype Systems			
CS0	Daresbury	10	10 CPUS, Pentium II/266
CS5	Daresbury	16	8 X dual Pentium III/933, SCALI

Commodity Systems (CS1)

- Master Node
  - 440LX based motherboard with on-board Adaptec Ultra Wide SCSI
  - Intel Pentium II-266 CPU
  - 256MB SDRAM
  - 4MB AGP Graphics card
  - 2 x PCI Fast Ethernet NIC
  - 4GB (system) and 9GB (user) Ultra Wide SCSI disk

- Compute Nodes (x32)
  - 440BX based motherboard
  - Intel Pentium III-450 CPU
  - 256MB ECC PC100 SDRAM
  - Minimal AGP Graphics card
  - 2 x PCI Fast Ethernet NIC
  - 10GB U-DMA (ATA-33) disk
- Networking
  - 2 x Extreme Summit 48 Fast Ethernet switches

Commodity Systems (CS2, Loki)

- Master node
  - UP2000 with 2 X 667 MHz Alpha 21264A (4MB L2), 2GB SDRAM; 36GB UW-SCSI user disk
- 64 compute processors:
  - 48 x 667 MHz + 16 x 833 MHz.
  - EV67 21264A with 512MB ECC SDRAM (1GB per node), 9GB SCSI disk
  - QSNet PCI Adaptor + 100 Mb/s Ethernet NIC (for NFS and console management).
- 2nd Fast Ethernet switch dedicated to the parallel communications traffic.
- High Performance Interconnect:
  - 128-way QsNet Elite switch chassis
  - Extreme Summit48 Ethernet switch
- System Software:
  - RH 6.2 LINUX, Compaq C/Fortran compilers, libraries etc., GCC, RMS for LINUX, MPICH, IP and SHMEM libs

Commodity Systems (CS3, Wulfgar)

- Master Node:
  - ASUS K7M slot A Motherboard
  - AMD 650 MHz Athlon
  - 256 MB PC100 ECC SDRAM
  - 4 MB AGP Graphics Card
  - Intel 10/100 Fast Ethernet NIC
  - 15 GB IBM IDE ATA-66 system disk
  - 36 GB IBM UW SCSI disk
  - Adaptec 2940 U2W SCSI controller

- Compute Nodes (x16):
  - ASUS K7M slot A Motherboard
  - AMD 850 MHz Athlon
  - 256 MB PC100 ECC SDRAM
  - 4 MB AGP Graphics Card
  - Intel 10/100 Fast Ethernet NIC
  - 15 GB IBM IDE ATA-66 system disk
  - Myrinet PCI64A LAN card
- Networking:
  - 24 port 3COM SuperStore II 3300 10/100 switch
  - 16 port Myrinet LAN switch

Commodity Systems (CS4, SARA Yellow)

- Yellow cluster, processors:
  - AMD Athlon 1200 MHz (initially 700 MHz)
  - 8Gbyte/node disk
- Network:
  - 100Mbit switched ethernet between nodes
  - 1Gbit ethernet to the fileservers
- Fileservers (one for each cluster):
  - Processor: Pentium III 650 MHz
  - Memory: 1024 Mbyte
  - Disk: 200 Gbyte Raid diskarray (Mylex)
- Software:
  - Operating system: Debian kernel 2.2.17
  - Bootprocessing system: BpBatch
- C compilers:
  - Gcc 2.95.2, Portland Group C 3.1-3
- Fortran compilers:
  - f2c, g77 2.95.2, Portland Group Fortran90 3.1-3
- Libraries:
  - MPICH, PVM3, NAG mark 19, Scalapack
- Batch system: PBS

Commodity Systems (CS6, CLiC)

- Compute nodes (x528)
  - Intel PentiumIII, 800 MHz
  - ASUS-Mainboard CUBX
  - 512 MB SDRAM PC100
  - Seagate Barracuda II 20,4 GB
  - gfxcard ATI 3D Carger 4MB
  - 2 x network adapter Level One FNC 0108TX
- Server nodes (x2)
  - Intel PentiumIII, 800 MHz
  - ASUS-Mainboard CUBX
  - 1 GB SDRAM PC100
  - Seagate Barracuda II 20,4 GB
  - floppy, gfxcard ATI 3D 4MB
  - 2 x Gigabit Ethernet Server Adapter SK 9843-SK-Net GE SX
- 1 server for managing power-on
- Power supply: USV 125 kVA

- 1st network (internal)
  - Extreme Black Diamond 6x 96-Port 10/100BASE-TX Module
- 2nd network (with uplink to campusnet)
  - Cisco Catalyst:
    - 3x 4000 Gigabit Ethernet Module
    - 13x 3548 XL Enterprise Edition
    - 28x 1000BASE-SX GBIC converter
- Software:
  - Linux Redhat 6.2, kernel 2.2.16-3
  - PVM, MPI, BLAS-3.0-4, lapack-3.0-4
  - PBS

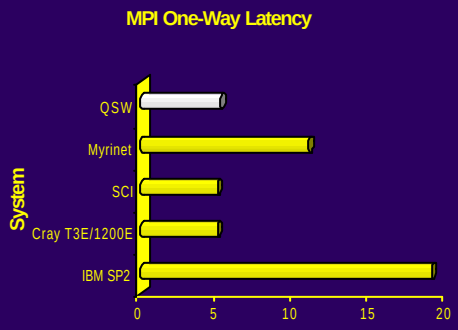
Commodity Systems (CS7, ukcp)

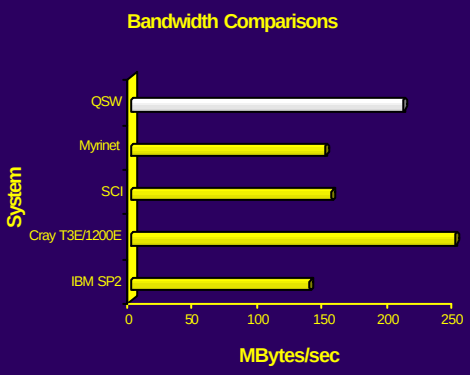
- Master node
  - Dual Pentium III 933MHz, 1GB SDRAM; 60GB IDE user disk, 20GB IDE system disk
- 64 compute processors:
  - 32 dual 1GHz Athlon MP
  - DDR ECC SDRAM (1GB per node), 20GB IDE disk
  - 64-bit Dolphin SCI 2D PCI card + 100 Mb/s Ethernet NIC (for NFS and console management).
- System Software:
  - RH 7.1 LINUX, PGI C/Fortran compilers, libraries etc., GCC, ScaliWorld (MPI, PBS, desktop tools).
  - *CS5* (8 dual Pentium III/930 nodes) provided initial access to SCALI/SCI.

High-End Systems

- Cray T3E/1200E
  - 816 processor system at Manchester (CSAR service)
  - 600 Mz EV56 Alpha processor with 256 MB memory
- IBM SP/WH2-375 (32 CPU system at DL)
  - 4-way Winterhawk2 SMP “thin nodes” with 2 GB memory
  - 375 MHz Power3-II processors with 8 MB L2 cache
- Compaq AlphaServer SC - 667 (APAC) and 833 MHz CPUs
  - 4-way ES40/667 and /833 SMP nodes with 2 GB memory
  - Alpha 21264a (EV67) CPUs with 8 MB L2 cache
  - Quadrics “fat tree” interconnect (5 usec latency, 150 MB/sec B/W)
- SGI Origin 3800
  - SARA (1000 CPUs) - Numalink with R14k/500 & R12k/400 CPUs
- Cray Supercluster at Eagen
  - Linux Alpha Cluster (96 X API CS20s - dual 833 MHz EV67 CPUs)
  - Myrinet interconnect, Red Hat 6.2

Interconnect Comparison - MPI





References
SP Switch Performance (IBM Document) 1998
Fujitsu April 1997
HLRS Germany 1997
NASA Ames Laboratory 1997
QSW 1998

Communications Benchmark
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:

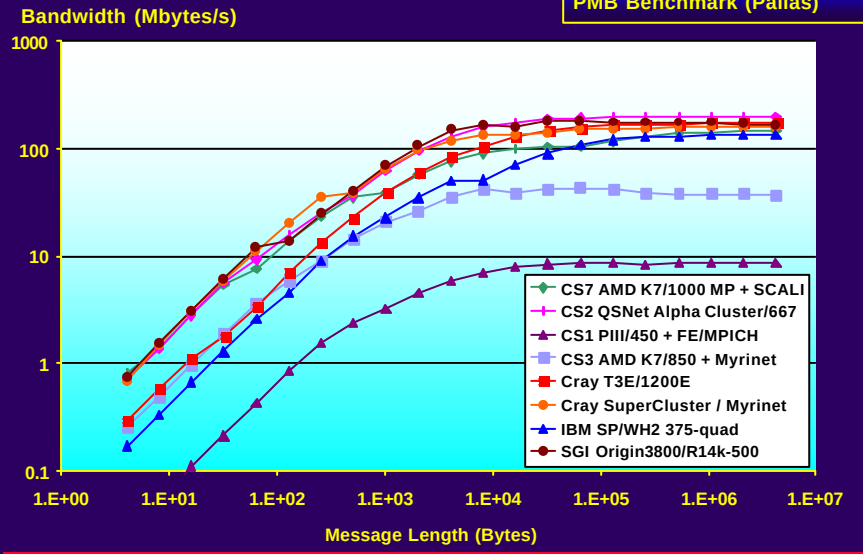
0 to 4194304 Bytes

Systems Investigated

- Cray T3E/1200E
- SGI Origin3800/R14k -500 (Teras)
- IBM SP/WH2-375 - using both 2- & 4- CPUs per quad-SMP node
- Cray SuperCluster (833MHz EV68 dual CS20 nodes with myrinet)
- **IBM Regatta H (intra node)**
- CS1 PIII/450 + Fast ethernet, MPICH
- CS2 Alpha Linux Cluster dual UP2000/667
- CS3 AMD Athlon/850 + Myrinet, MPICH
- CS5 dual PIII/930 + SCALI interconnect
- CS6 PIII/800 + Fast ethernet, MPICH
- CS7 AMD K7/1000 MP + SCALI

PingPong Performance

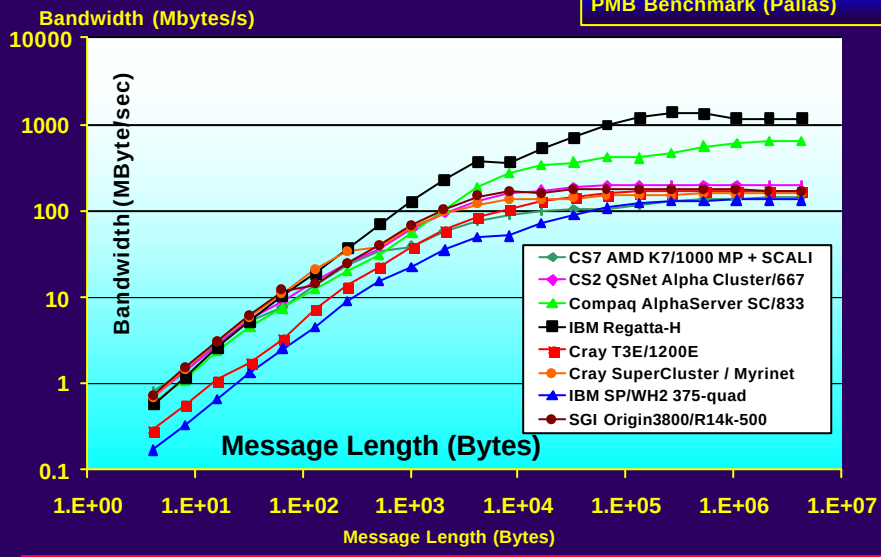
PMB Benchmark (Pallas)



Computational Science and Engineering Department Daresbury Laboratory

PingPong Performance

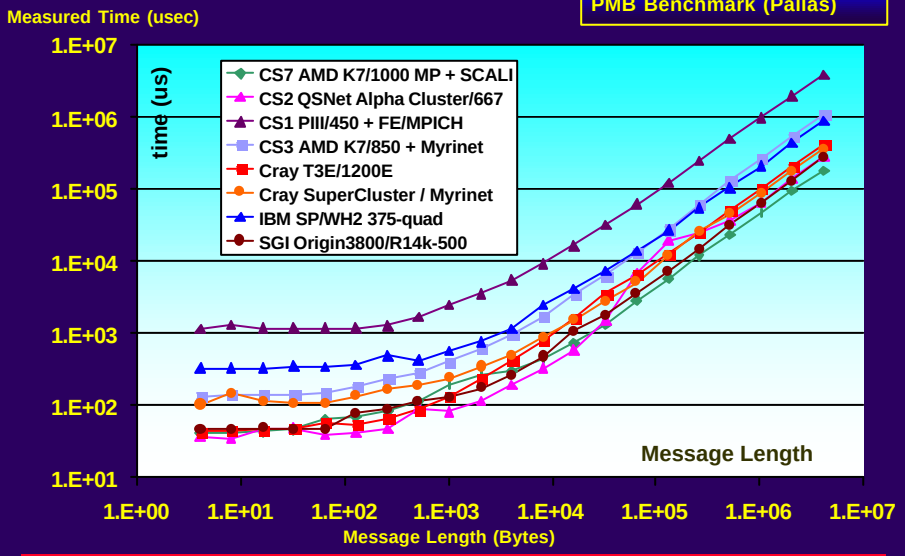
PMB Benchmark (Pallas)



High Performance Computing on Linux Clusters 12 February 2002

16 CPUs MPI_allreduce Performance

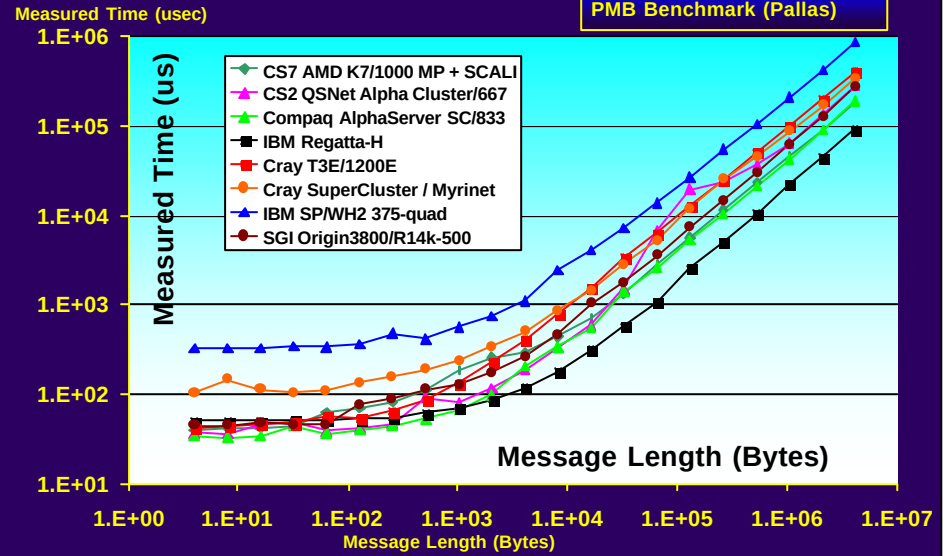
PMB Benchmark (Pallas)



Computational Science and Engineering Department Daresbury Laboratory

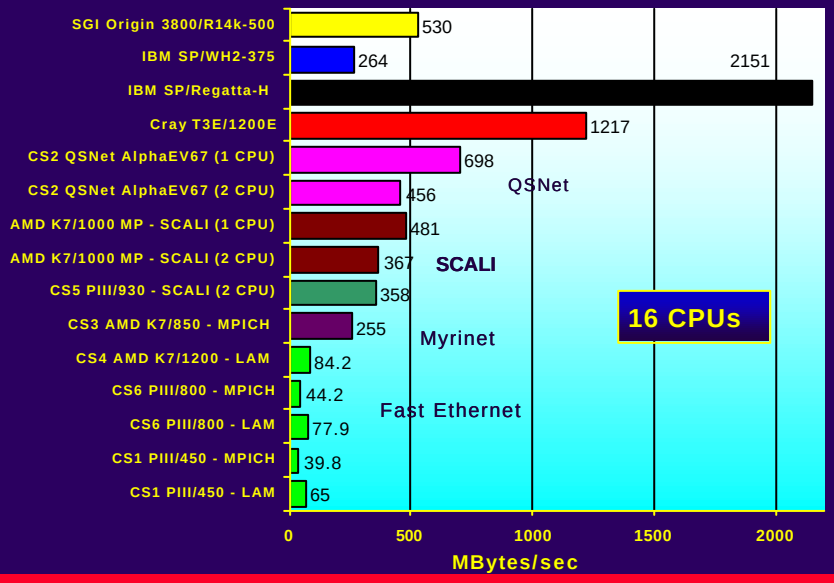
16 CPUs MPI_allreduce Performance

PMB Benchmark (Pallas)

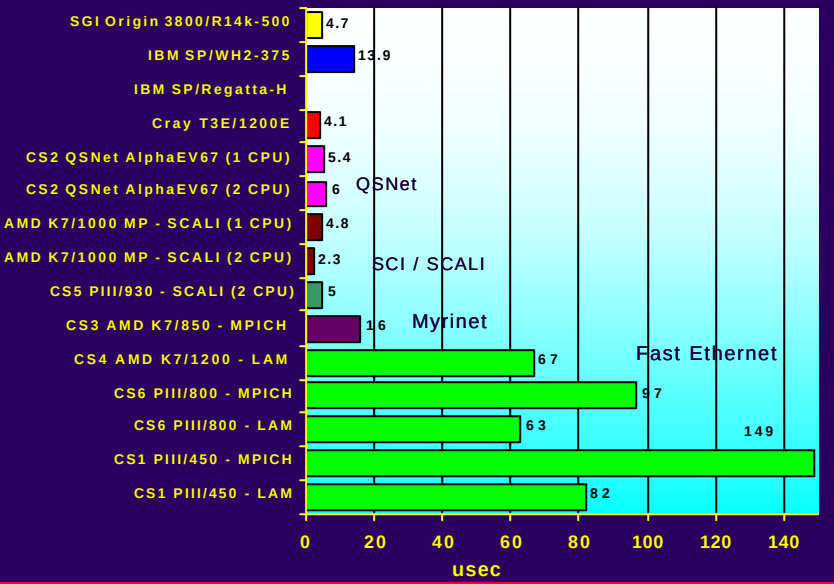


High Performance Computing on Linux Clusters 12 February 2002

Interconnect Benchmark - EFF_BW



Interconnect Benchmark - Latency



QSNet Alpha/LINUX Cluster

- Performance of Collective MPI operations
- Shared Memory Performance:
  - Cost effectiveness of UP2000 stems from use of commodity SDRAMS (c.f. STREAM figures) and 1/2 number of paths to memory as Compaq DS20 platform (effect on inter-node MPI performance);
  - Expected to impact on applications with heavy memory usage
- Effect of variations in messaging schemes:
  - Programming model for QLC-alpha uses flat MPI/SHMEM model implemented across clustered SMP solutions.
  - Intra node comms. Supported as either direct memory copies or via Elan NIC. Default for intra-node in MPI is shared memory while in SHMEM interface default is to use Elan. Complex - messages currently poll ...
- LINUX Issues and Performance:
  - Page colouring: support for multi-way associative caches (L2) to provide optimum strategy for cache reuse.

Application Codes

Performance comparisons between Commodity-based systems and both current MPP (CSAR Cray T3E/1200E) and ASCI-style SMP-node platforms (IBM SP / WH2-375, Compaq AlphaServer SC (ES40/6-667, 6-833), SGI Origin 3800 and Prototype Cray Supercluster:

- Chemistry and Materials
  - **DL_POLY - parallel MD code with many applications**
  - **GAMESS-UK, NWChem & Turbomole - Ab initio Electronic structure codes**
  - **CRYSTAL, VASP, CASTEP** - UKCP Car-Parrinello total energy materials code
- Molecular Biology
  - **CHARMM (NIH and BASF), DL_POLY, ...**
- Engineering
  - **ANGUS** - regular-grid domain decomposition engineering code with conjugate-gradient and multi-grid solvers
  - **FLITE3D** - finite-element irregular-grid engineering code
- Climate Modelling
  - **Unified Model** (Atmospheric Climate Modelling)

Performance Metric (% 32-node Cray T3E)

Performance Metrics

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

Performance Metric (% 32-node Cray T3E)

1. 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 + Quadrix**

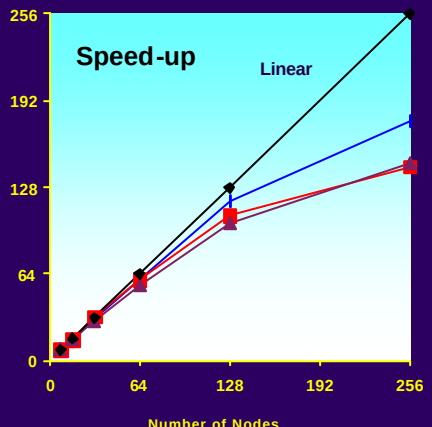
2. T (32-CPUs SGI Origin 3800) / T (32 CPUs) CS2 Alpha Linux Cluster

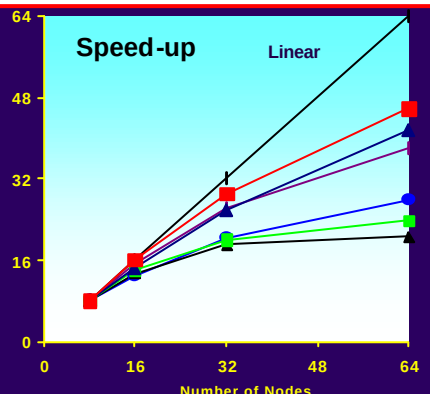
$T_{32\text{-CPU}}$ **SGI Origin 3800 /** $T_{32\text{-CPU}}$ **CS2 Alpha Linux Cluster**

DL_POLY Parallel Benchmarks (Cray T3E/1200)

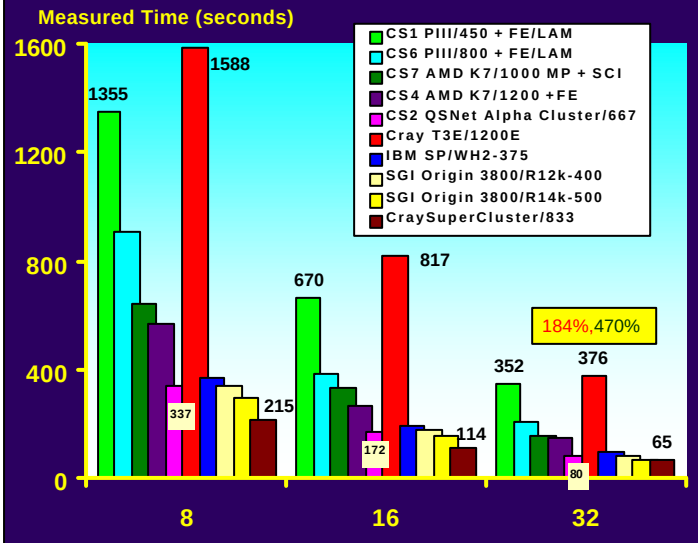
4. NaCl; MTS Ewald, 27,000 ions
5. NaK-disilicate glass; 8,640 atoms, Ewald
8. MgO microcrystal: 5,416 atoms

9. Model membrane/valinomycin (MTS, 18,886)
7. Gramicidin in water (SHAKE, 13,390)
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: Cray/T3E, High-end and Commodity-based Systems I.



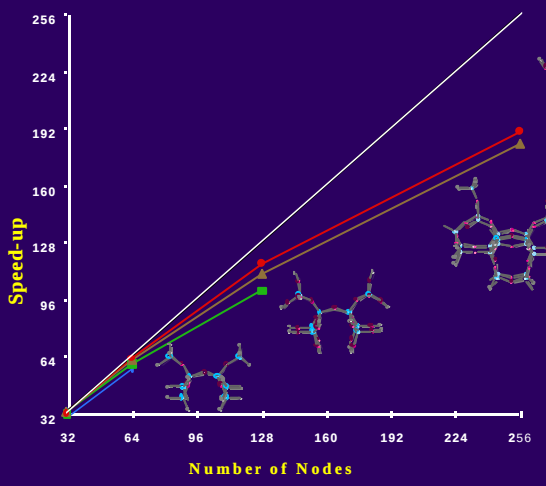
Bench 4.
NaCl; 27,000 ions,
Ewald + MTS,
75 time steps,
Cutoff=24Å

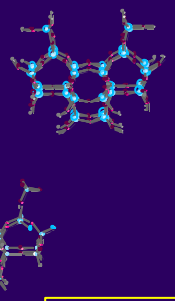
$T3E_{128} = 94$

High-End Computational Chemistry
The NWChem Software

- Developed as part of the construction of the Environmental Molecular Sciences Laboratory (EMSL) at PNNL.
- Funded to be used as an integrated component in solving DOE's grand challenge environmental restoration problems
- Designed and developed to be a highly efficient and portable MPP computational chemistry package, providing computational chemistry solutions which are scalable with respect to chemical system size as well as MPP hardware size
- Extensible framework supporting development of new methods in computational chemistry; NWChem Architecture
  - Object-oriented design
    - abstraction, data hiding, handles, APIs
  - Parallel programming model
    - non-uniform memory access, global arrays (GAs)
  - Infrastructure
    - **Global Arrays (GA)** , Parallel I/O, RTDB, MA, **Linear algebra (PeiGS)** ...

Measured Parallel Efficiency for NWChem - DFT on IBM-SP; Wall Times to Solution for SCF Convergence



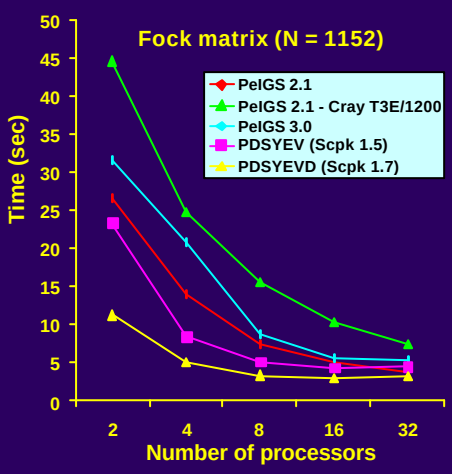


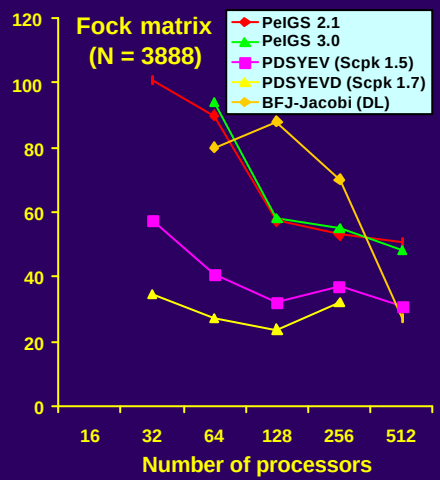
Zeolite Fragment	Basis AO/CD	Number of Nodes	Wall Time to Solution
$Si_8O_7H_{18}$	347/832	64	238s
$Si_8O_{25}H_{18}$	617/1444	128	364s
$Si_{26}O_{37}H_{36}$	1199/2818	256	1137s
$Si_{28}O_{67}H_{30}$	1687/3928	256	2766s

Parallel Eigensolvers

Real symmetric eigenvalue problems

SGI Origin 3800/R12k-400 ("green")



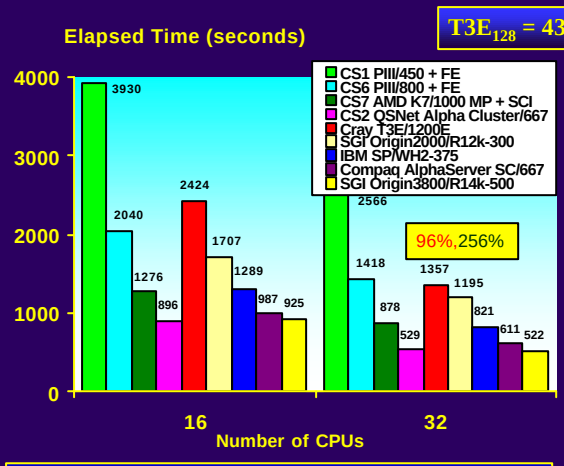


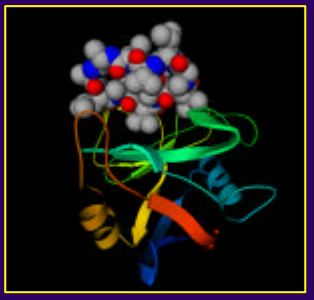
Parallel Implementations of GAMESS-UK

- Extensive use of Global Array (GA) Tools and Parallel Linear Algebra from NWChem Project (EMSL)
- SCF and DFT energies and gradients
  - Replicated data, but …
  - 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 PeIGs, DIIS/MMOs, Inversion of 2c-2e matrix)
- SCF second derivatives
  - Distribution of <vvoo> and <vovo> integrals via GAs
- MP2 gradients
  - Distribution of <vvoo> and <vovo> integrals via GAs

GAMESS-UK ΔSCF Performance

Cray T3E/1200E, High-end and Commodity-based Systems





Impact of Serial Linear Algebra:

$T_{IBM\text{-}SP}(16)$ = 2656 [1289]

$T_{IBM\text{-}SP}(32)$ = 2184 [821]

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

Materials Simulation Codes

Plane Wave DFT Codes:

- CASTEP
- VASP
- CPMD

These codes have similar functionality, power and problems. CASTEP is the flagship code of UKCP and hence subsequent discussions will focus on this.

Local Gaussian Basis Set Codes:

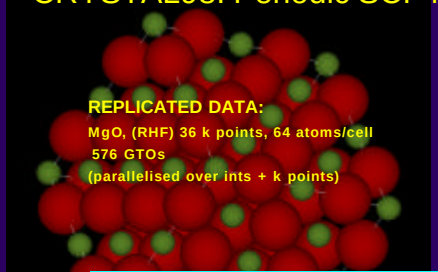
- CRYSTAL

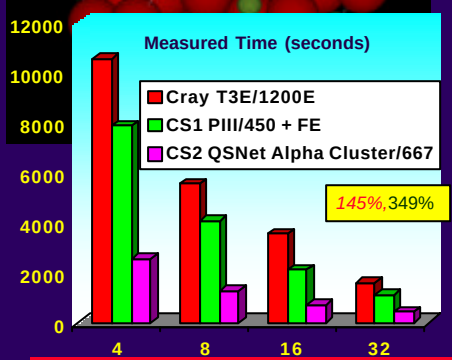
This code presents a different set of problems when considering performance on HPC(x).

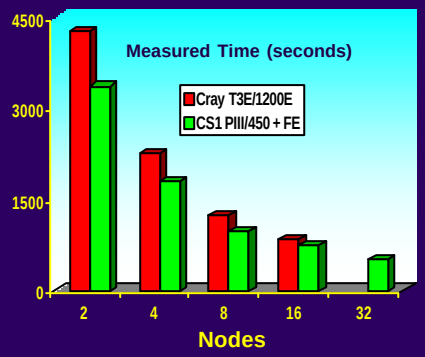
SIESTA and CONQUEST:

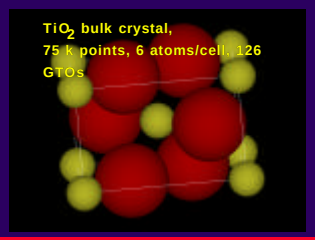
- O(n) scaling codes which will be extremely attractive to users.
- Both are currently development rather than production codes.

CRYSTAL98: Periodic SCF for MgO and TiO_2



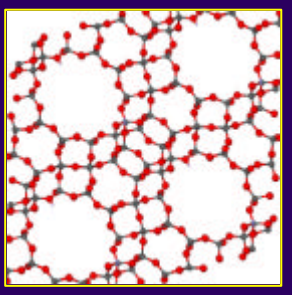


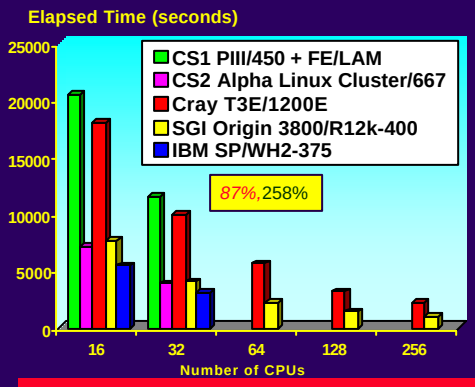


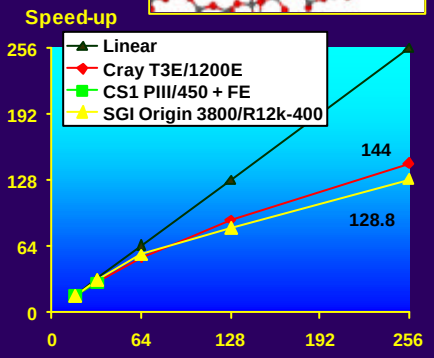


CRYSTAL - 2000

- Distributed Data implementation
- Benchmark:
  - An Acid Centre in Zeolite-Y (Faujasite)
  - Single point energy
  - 145 atoms / cell, No symmetry / 8k-points
  - 2208 basis functions, (6-21G*)







Plane Wave Methods: CASTEP

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

- Direct minimisation of the total energy (avoiding diagonalisation)
- 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.

CASTEP - The UK Car-Parrinello Consortium

UK Car-Parrinello Consortium

- The Cambridge Serial Total Energy Package CASTEP (M. Payne et al.) calculates the total energy, forces and stresses in a 3D-periodic system.
- Rev. Mod.Phys. 64 (1992) 1045
- DFT, plane-waves, pseudo-potentials & FFT's

CASTEP 4.2 β Key Features:

- Ultrasoft pseudo-potentials with non-linear core corrections
- Range of minimisation methods: Density Mixing, RM-DIIS, Conjugate Gradients band-by-band & all-bands. Full structural relaxation and MD
- LD and GGAs, spin-polarisation

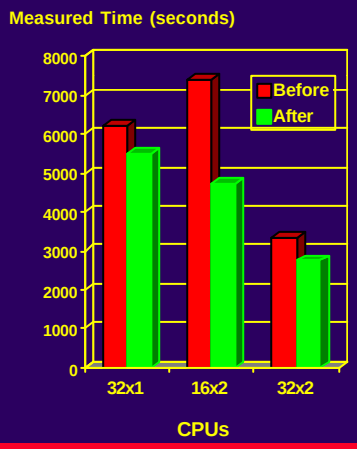
Parallelization of CASTEP

- A Number of parallelization methods are implemented:
  - **k-point:** a processor holds all the wavefunction for a k-point (MPI_ALLTOALLV is NOT required) BUT for large unit cells $N_k \Rightarrow 1$ i.e small CPU count.
  - **G-vector:** a processor holds part of the wavefunction for all k-points (MPI_ALLTOALLV is over ALL CPUs) i.e. biggest systems with 1 K point
  - **mixed kG:** k-points are allocated amongst processors, the wavefunctions sub-allocated amongst processors associated with their particular k-points i.e. MPI_ALLTOALLV is over N_{CPUs} / N_k - intermediate cases.
- On HPC hardware the desired method is either k or kG as this minimizes inter-processor communication, specifically MPI_ALLTOALLV.
- However, on large numbers of processors such distributions will still be problematic. New algorithms will therefore need to be developed to overcome latency problems.

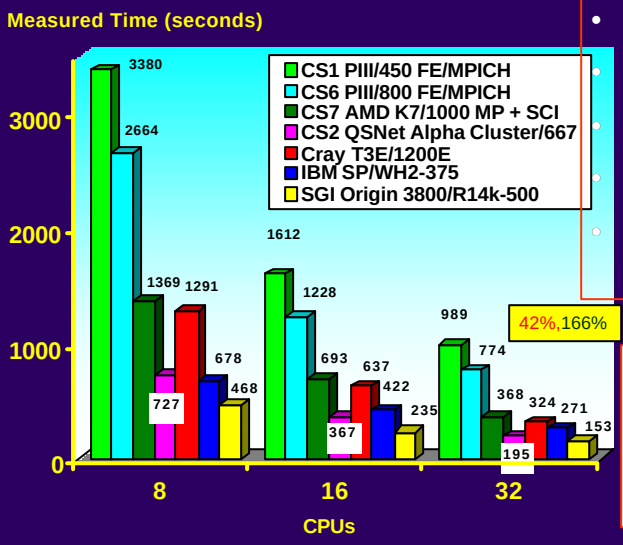
CPU Optimizations:
Efficiency on Commodity-based Systems

- CASTEP on the Cray T3E and SGI Origin 3800 systems use FFT code fully optimized for the processor L1-cache.
- Extended to other cache-based processors. FFT operations are performed on chunks of data that fit in L1-cache and run at maximum speed. i.e. FFT exploited efficiently on cache-based CPUs.
- Example:
  64 processor (32 dual 1GHz AMD K7) system, SCALI interconnect

TiN: A TiN 32 atom slab, 8 k-points, single point energy calculation with Mulliken analysis,



CASTEP 4.2 - Parallel Benchmark



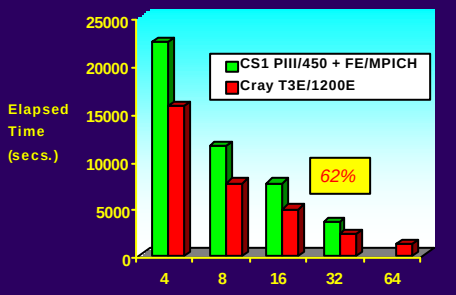
Chabazite

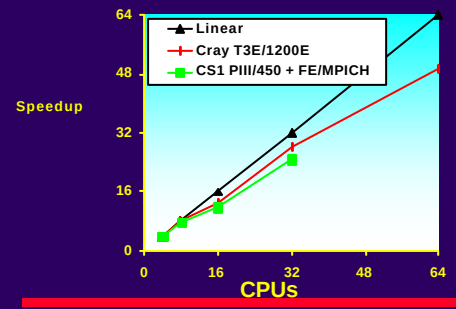
- Acid sites in a zeolite. (Si_{11} O_{24} Al H)
- Vanderbilt ulatrasoft pseudo-potential
- Pulay density mixing minimiser scheme
- single k point total energy, 96 bands
- 15045 plane waves on 3D FFT grid size = 54x54x54; convergence in 17 SCF cycles

42%,166%

	Time (comms)
IBM SP/WH2-375	157
Cray T3E/1200E	90
CS1 PIII/450+FE	660
CS6 PIII/800+FE	600
CS7 AMD K7/1000 + SCI	242
CS2 QSNet Alpha	111
SGI Origin 3800/R14k	71

CPMD - Car-Parrinello Molecular Dynamics





CPMD

- Version 3.3: Hutter, Alavi, Deutsh, Bernasconi, St. Goedecker, Marx, Tuckerman and Parrinello (1995-1999)
- DFT, plane-waves, pseudo-potentials and FFT's

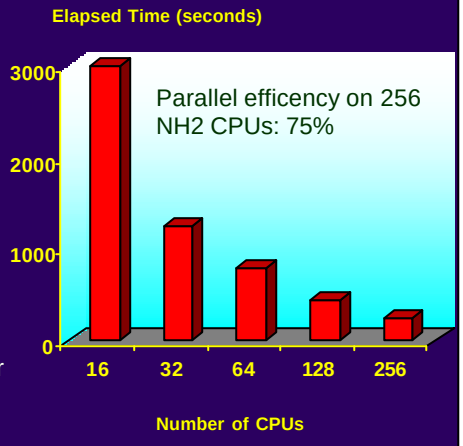
Benchmark Example: Liquid Water

- Physical Specifications:
  32 molecules, Simple cubic periodic box of length 9.86 A, Temperature 300K
- MD parameters;
  Time step 7 au = 0.169 fs; Length test run 200 steps = 34 fs
- Electronic Structure;
  BLYP functional, Trouillier Martins pseudopotential, Recriprocal space cutoff 70 Ry = 952 eV

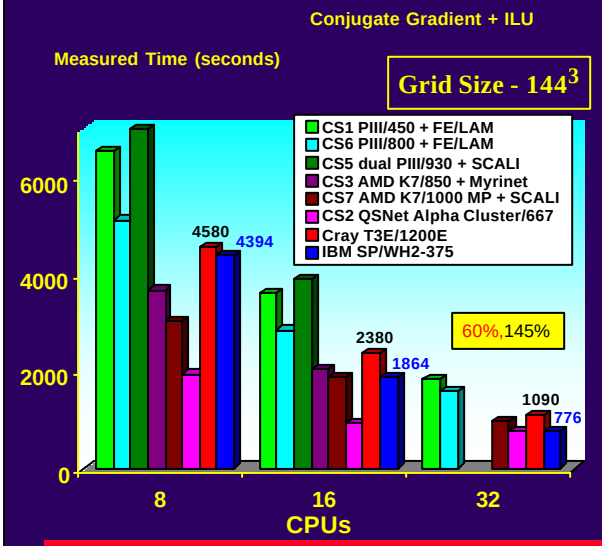
Sprik and Vuilleumier (Cambridge)

CPMD on High-end Computers

- **Performances on NH2 nodes**
  - 256 CPUs and 8 MPI tasks/node including IO (RD30WFN / WR30WFN)
  - 390 Mflops per CPU, 99.8 Gflops
  - 252*252*252 - 8 MPI tpn
  - Use of ESSL routines instead of Lapack; Some routines were "OpenMPed"
- **Mixed mode** MPI and OMP (IBM)
- CPMD is dominated by ESSL SMP routines (ROTATE and OVLAP).
- 4 MPI tpn and 2 SMP tpn is 1.36 faster than flat MPI on a given MESH
- **Power4 estimates:** The key is FFTCOM (MPall2all). Average speedup of 2.0 on a R-H LPARd system



ANGUS: Combustion modelling (regular grid)
The Cray T3E/1200, IBM SP/WH2 and Beowulf Systems

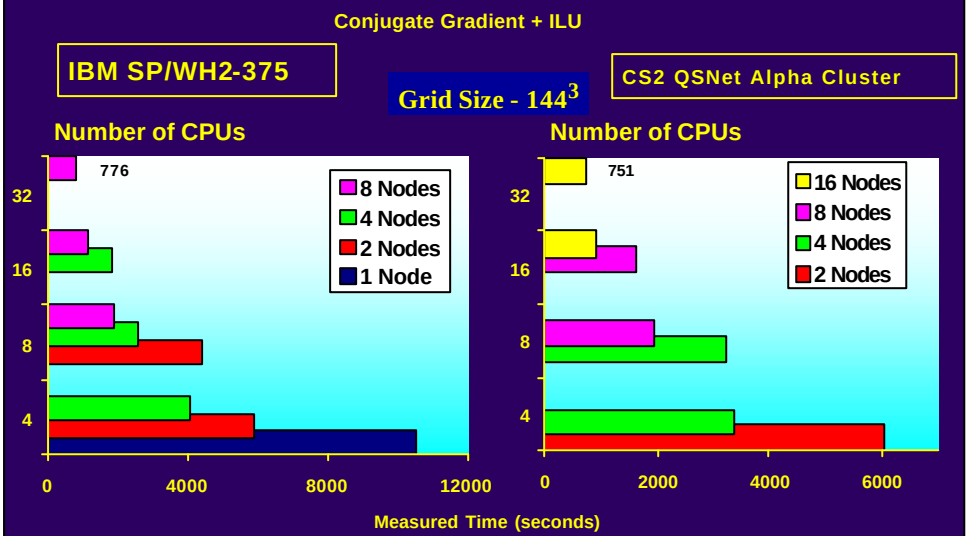


Direct numerical simulations (DNS) of turbulent pre-mixed combustion solving the augmented Navier-Stokes equations for fluid flow.

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.

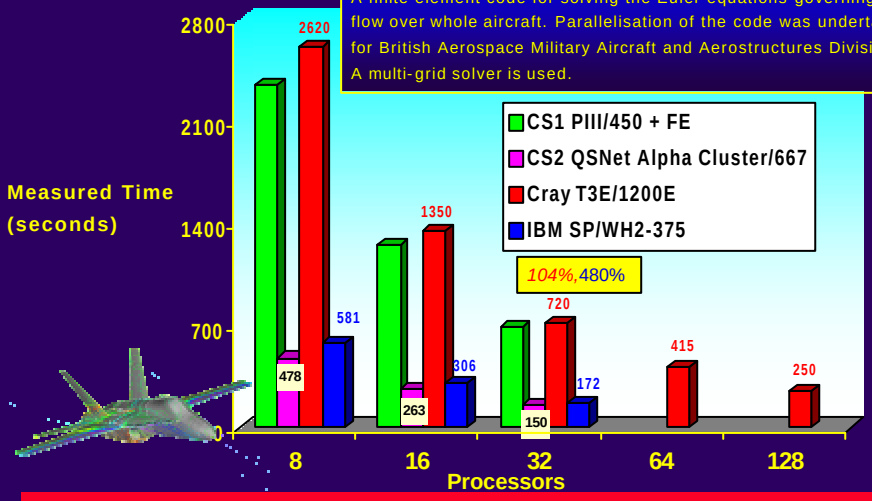
ANGUS: Combustion modelling (regular grid)
Memory Bandwidth Effects: The IBM SP and Alpha Linux System



FLITE3D: An Industrial Aerospace Code

F18 test, 3444350 elements

A finite-element code for solving the Euler equations governing air flow over whole aircraft. Parallelisation of the code was undertaken for British Aerospace Military Aircraft and Aerostructures Division. A multi-grid solver is used.

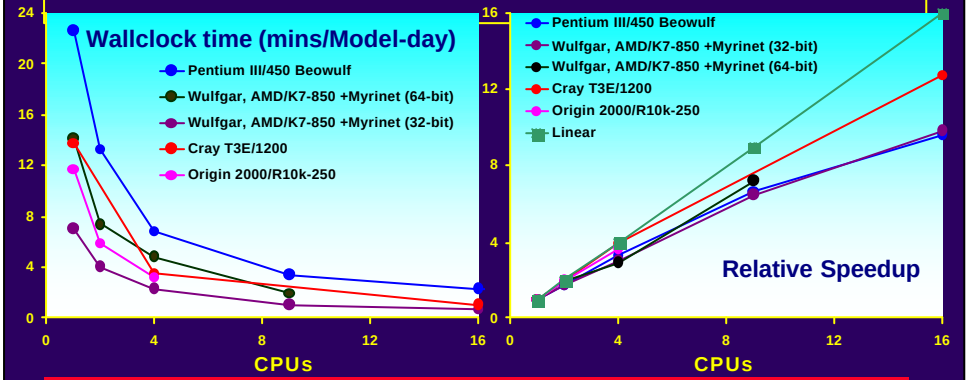


Atmospheric Climate Modelling

University of Reading and UK Meteorological Office

The Unified Model system consists of the full UKMO operational forecasting suite & the coupled Ocean-Atmosphere model used for climate prediction.

For a 5-day model period in atmosphere-only configuration, the T3E & Origin 2000 produce good speedups and the Beowulf systems scale well to 8 processors. This version of the model has many short messages and barriers & so is not particularly well coded.



Beowulf Comparisons with the T3E & O3800/R14k-500

CSx - PentiumIII + FE
% of 32-node Cray T3E/1200E

GAMESS-UK	CS1	CS6
SCF	53-69%	96%
DFT	65-85%	130-178%
DFT (Jfit)	44-77%	65-131%
DFT Gradient	90%	130%
MP2 Gradient	44%	73%
SCF Forces	80%	127%
NWChem (DFT Jfit)		50-60%
REALC	67%	
CRYSTAL	145%	
DLPOLY		
Ewald-based	95-107%	151-184%
bond constraints	34-56%	69%
CHARMM	96%	172%
CASTEP	33%	42%
CPMD	62%	
ANGUS	60%	68%
FLITE3D	104%	

CS2 - QSNet Alpha Linux Cluster
% of 32-node Cray T3E and O3800/R14k-500

GAMESS-UK		
SCF	256%	99%
DFT †	301-361%	99%
DFT (Jfit)	219-379%	89-100%
DFT Gradient †	289%	89%
MP2 Gradient	228%	87%
SCF Forces	154%	86%
NWChem (DFT Jfit) †	150-288%	74-135%
CRYSTAL †	349%	
DLPOLY		
Ewald-based †	363-470%	95%
bond constraints	143-260%	82%
CHARMM †	404%	78%
CASTEP	166%	78%
ANGUS	145%	
FLITE3D †	480%	

Collaborations: Performance Modelling

Collaboration between Pallas GmbH and CLRC to investigate performance measurement and prediction on both PC & Alpha-based Beowulf systems, and on HPC servers, using Vampir and Dimemas.

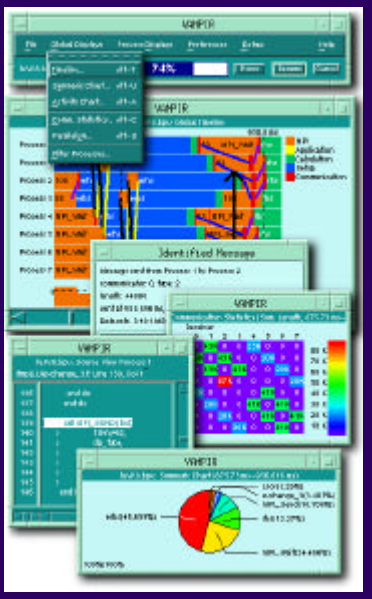
Objectives

- Determine relative merits of clusters and low/medium servers for a number of applications (including those from industry).
- Separation of algorithmic & architectural components of performance.
- Performance prediction on various machine and network architectures.

pallas

Vampir 2.5

Visualization and Analysis of MPI Programs



Performance Analysis of GA-based Applications using Vampir

GAMESS-UK on High-end and Commodity class machines

- **extensions to handle GA applications**

Summary

- Cluster Computing - where do clusters fit?
- Commodity-based Systems
  - **Single-node performance and Interconnects**
  - Prototype Systems; CS1 - CS7
- Application performance on Commodity Clusters
  - Application performance analysis - role of VAMPIR
  - Comparison with High-end Systems
    - Cray, IBM, SGI, and Compaq
  - Electronic Structure and Molecular Simulation
    - GAs and Linear Algebra (PeIGS)
    - GAMESS-UK, NWChem ,DL_POLY and CHARMM
  - Materials
    - **CRYSTAL, CPMD & CASTEP**
  - **Computational Engineering**
    - **ANGUS & FLITE3D**
  - Linux Alpha Cluster delivers between 150-400% of T3E/1200E, 78-100% of SGI Origin 3800/R14k-500

CS2 - QSNet Alpha Linux Cluster
% of 32 CPU O3800/R14k-500

GAMESS-UK	
SCF	99%
DFT	99%
DFT (Jfit)	89-100%
DFT Gradient	89%
MP2 Gradient	87%
SCF Forces	86%
DLPOLY	
Ewald-based	95%
bond constraints	82%
CASTEP	78%