HPC*x* : A New Resource for UK Computational Science

Mike Ashworth, Ian J. Bush, Martyn F. Guest, Martin Plummer and Andrew G. Sunderland

CLRC Daresbury Laboratory, UK

m.f.guest@dl.ac.uk

and

Stephen Booth, David S. Henty, Lorna Smith and Kevin Stratford

EPCC, University of Edinburgh, UK

http://www.hpcx.ac.uk/

Outline

- HPCx Overview
  - HPCx Consortium
  - HPCx Technology - Phases 1, 2 and 3 (2002-2007)
- Performance Overview of Strategic Applications:
  - Computational Materials
  - Molecular Simulation
  - Molecular Electronic Structure
  - Atomic and Molecular Physics
  - Computational Engineering
  - Environmental Science
- Evaluation across a range of Current High-End Systems:
  - IBM SP/p690, SGI Origin 3800/R14k-500, HP/Compaq AlphaServer SC ES45/1000 and Cray T3E/1200E
- Summary

Applications, and not H/W driven

HPCx Project overview

HPCx CAPABILITY COMPUTING

- A joint venture between the Edinburgh Parallel Computing Centre (EPCC) at the University of Edinburgh and the Daresbury Laboratory of the Central Laboratory for the Research Councils (CLRC)
- Project funded to £53M (~$120M) by UK Government
- Established to operate and support the principal academic and research computing service for the UK
- Principal objective being to provide a Capability Computing service to run scientific applications that could not be run on any other available computing platform
- Six-year project with defined performance requirements at year 0, year 2 and year 4 so as to match Moore's Law
- IBM chosen as the technology partner with Power4 based p690 platform, and the "*best available interconnect*"

Consortium partners

HPCx CAPABILITY COMPUTING

- EPCC (University of Edinburgh)
  - established in 1991 as the University's interdisciplinary focus for high-performance computing and commercial exploitation arm
  - has hosted specialised HPC services for the UK's QCD community since 1989. 5Tflop QCDOC system due 2003 in project with Columbia, IBM and Brookhaven National Laboratory
  - operated and supported UK national services on CRAY T3D and T3E systems from 1994 until 2002
- CLRC (Daresbury Laboratory)
  - HPC service provider to the UK academic community for > 25 yrs
  - research, development & support centre for leading edge academic engineering and physical science simulation codes
  - distributed computing support centre for COTS processor & network technologies, evaluating scalability and performance
  - UK grid support centre

HPCx Technology Phase 1

HPCx CAPABILITY COMPUTING

Phase 1 (Dec. 2002): 3 TFlop/s Rmax Linpack

- 40 Regatta-H SMP compute systems (1.28 TB memory)
  - 32 x 1.3GHz processors, 32 GB memory; 4 x 8-way LPARs
- 2 Regatta-H I/O systems
  - 16 x 1.3GHz processors (Regatta-HPC), 4 GPFS LPARS
  - 2 HSM/backup LPARS, 18TB EXP500 fibre-channel global filesystem
- Switch Interconnect
  - Existing SP Switch2 with "Colony" PCI adapters in all LPARs (20 us latency, 350 MB/s bandwidth)
  - Each compute node has two connections into switch fabric (dual plane)
  - 160 x 8-way compute nodes in total
- Ranked #9 in the TOP500 list (November 2002)

HPCx Technology Phases 2 & 3

HPCx CAPABILITY COMPUTING

Phase 2 (2004): 6 TFlop/s Rmax Linpack

- >40 Regatta-H+ compute systems
  - 32 x 1.8GHz processors, 32 GB memory, full SMP mode (no LPAR)
- 3 Regatta-H I/O systems (Double the capabilities of Phase 1)
- "Federation" switch fabric
  - bandwidth quadrupled, ~5-10 microsecond latency, Connect to GX bus directly

Phase 3 (2006): 12 TFlop/s Rmax Linpack

- >40 Regatta-H+ compute systems
  - unchanged from Phase 2
- >40 additional Regatta-H+ compute systems
  - double the existing configuration
- 4 Regatta I/O systems (Double the capabilities of Phase 2)

Open to Alternative Technology Solutions (IPF, BlueGene/L ..)

HPCx - Phase 1 Technology at Daresbury

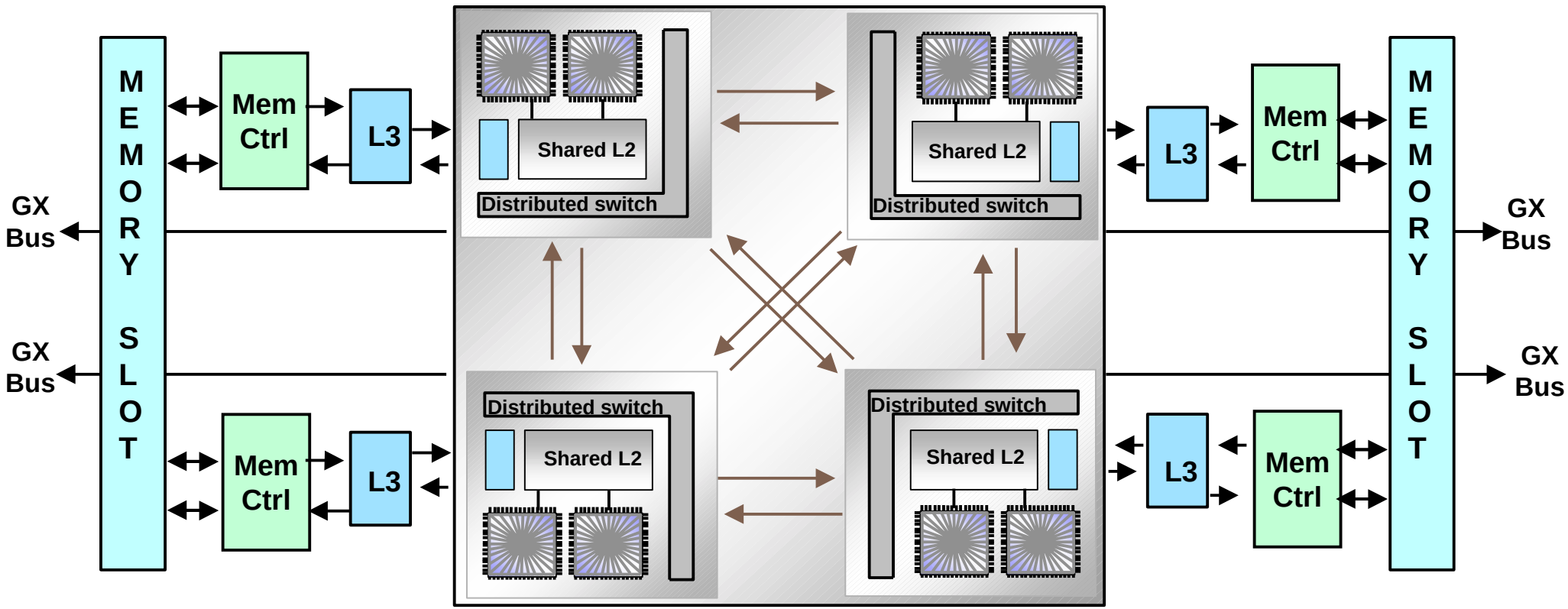
HPCx
CAPABILITY COMPUTING

July 2002

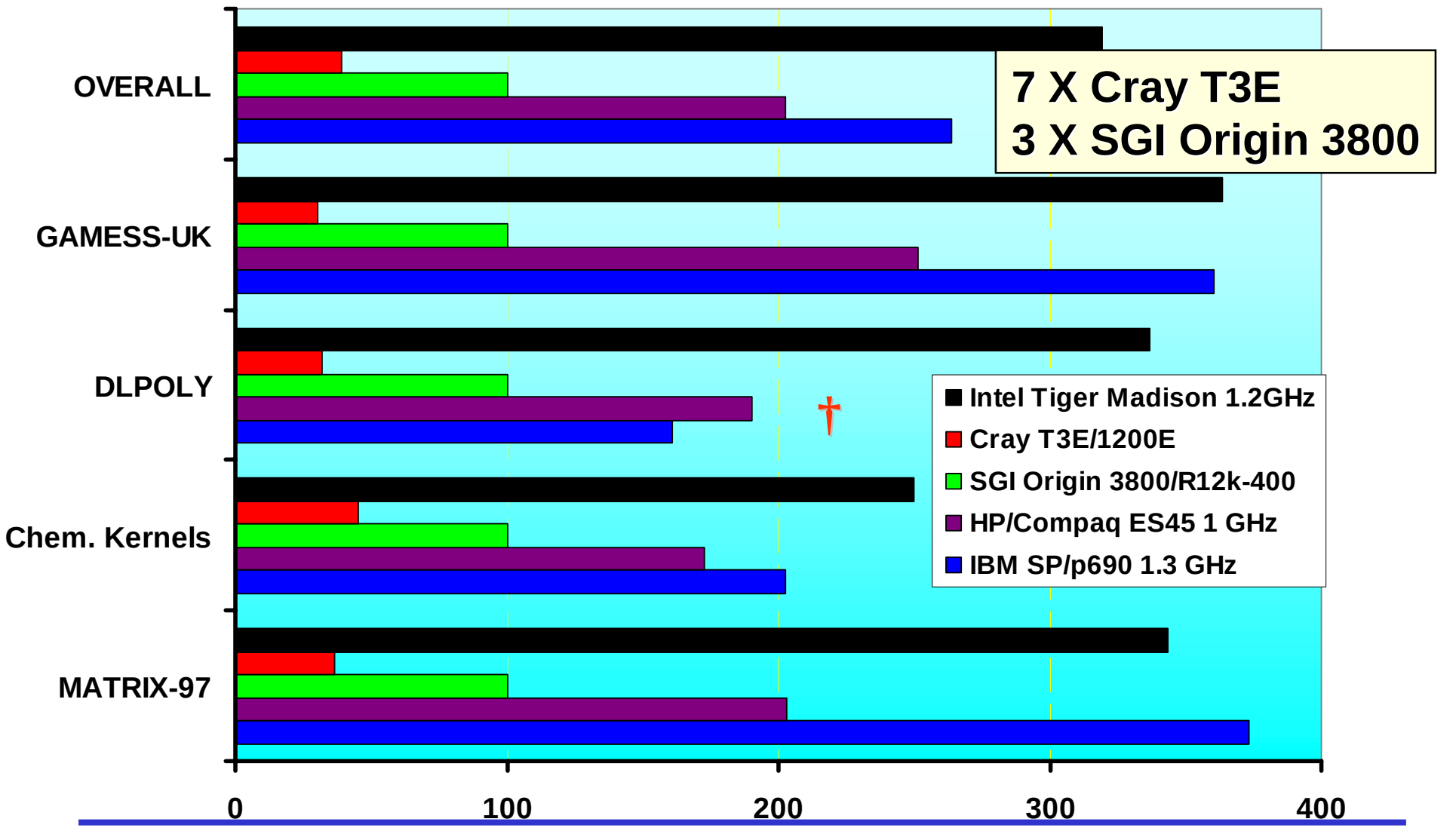
November 2002

IBM p-series 690Turbo:Multi-chip Module (MCM)

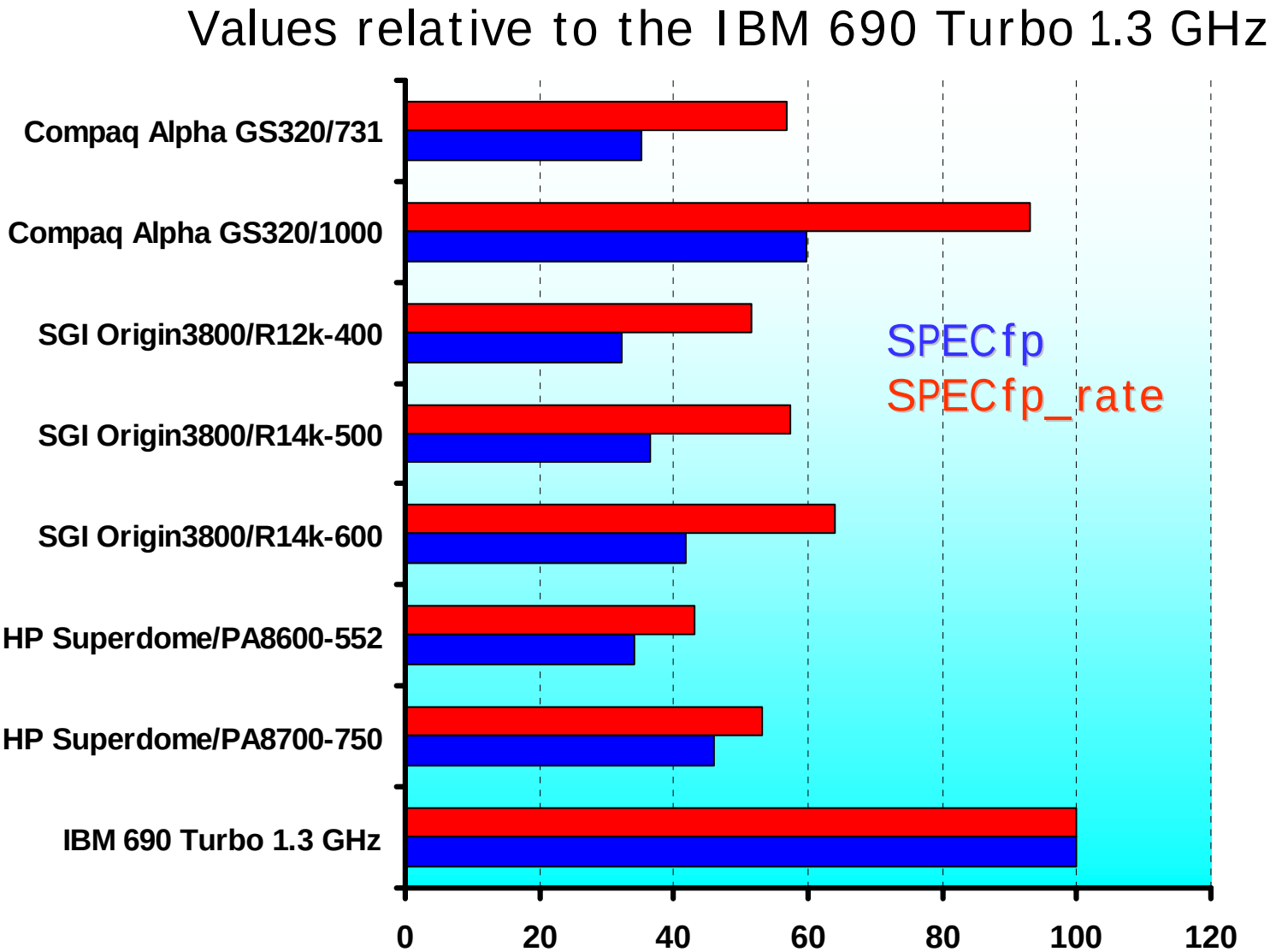
Four POWER4 chips (8 processors) on an MCM, with two associated memory slots



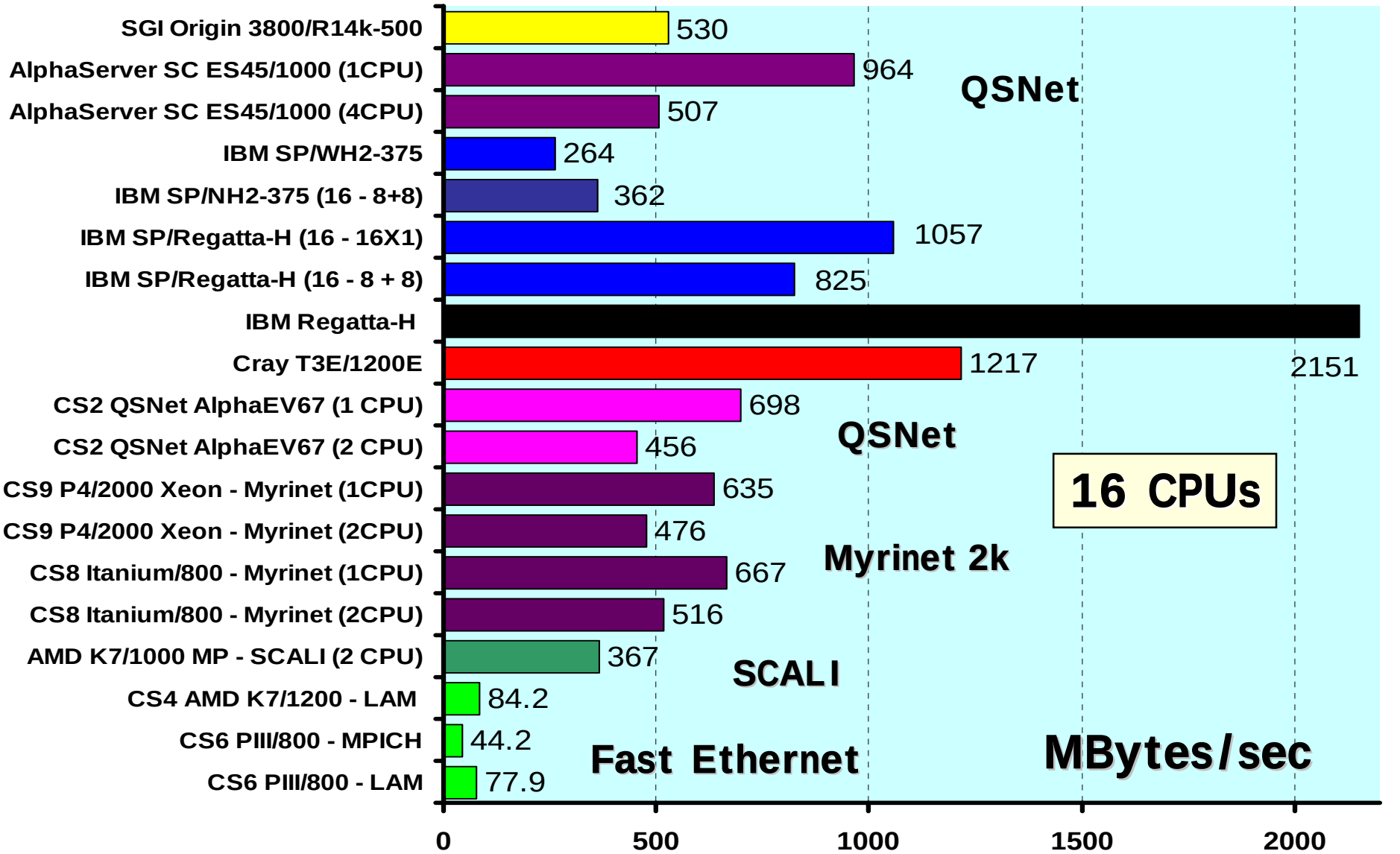
Serial Benchmark Summary



SPEC CPU2000: SPECfp vs SPECfp_rate (32 CPUs)



Interconnect Benchmark - EFF_BW



Capability Benchmarking and Application Tuning

HPCx CAPABILITY COMPUTING

- Materials Science
  - CASTEP, AIMPRO & CRYSTAL
- Molecular Simulation
  - DL-POLY & NAMD
- Atomic & Molecular Physics
  - PFARM and H2MOL
- Molecular Electronic Structure
  - GAMESS-UK & NWChem
- Computational Engineering
  - PDNS3D
- Environmental Science
  - POLCOMS

HPCx Terascale Applications Team

Systems Used In Performance Analysis

HPCx CAPABILITY COMPUTING

- IBM Systems
  - IBM SP/Regatta-H (1024 procs, 8-way LPARs) HPCx system at DL
  - Regatta-H (32-way) and Regatta HPC (16-way) (Montpelier)
  - SP/Regatta-H (8-way LPARs, 1.3 GHz) at ORNL
- HP/Compaq AlphaServer SC
  - 4-way ES40/667 (APAC) and 833 MHz SMP nodes ;
  - TCS1 system at PSC: 750 4-way ES45 nodes - 3,000 EV68 1 GHz CPUs, with 4 GB memory per node
  - Quadrics “fat tree” interconnect (5 usec latency, 250+ MB/sec B/W)
- SGI Origin 3800
  - SARA (1000 CPUs) - NumaLink - with R14k/500 and R12k/400 CPUs
  - CSAR (512 CPUs) - NumaLink - R12k/400
- Cray T3E/1200E
  - CSAR (788 CPUs)

Materials Science

HPCx CAPABILITY COMPUTING

AIMPRO

(Ab Initio Modelling PROgram)

Patrick Briddon et al, Newcastle University

http://aimpro.ncl.ac.uk/

CRYSTAL

Properties of crystalline systems
periodic HF or DFT Kohn-Sham Hamiltonian
various hybrid approximations

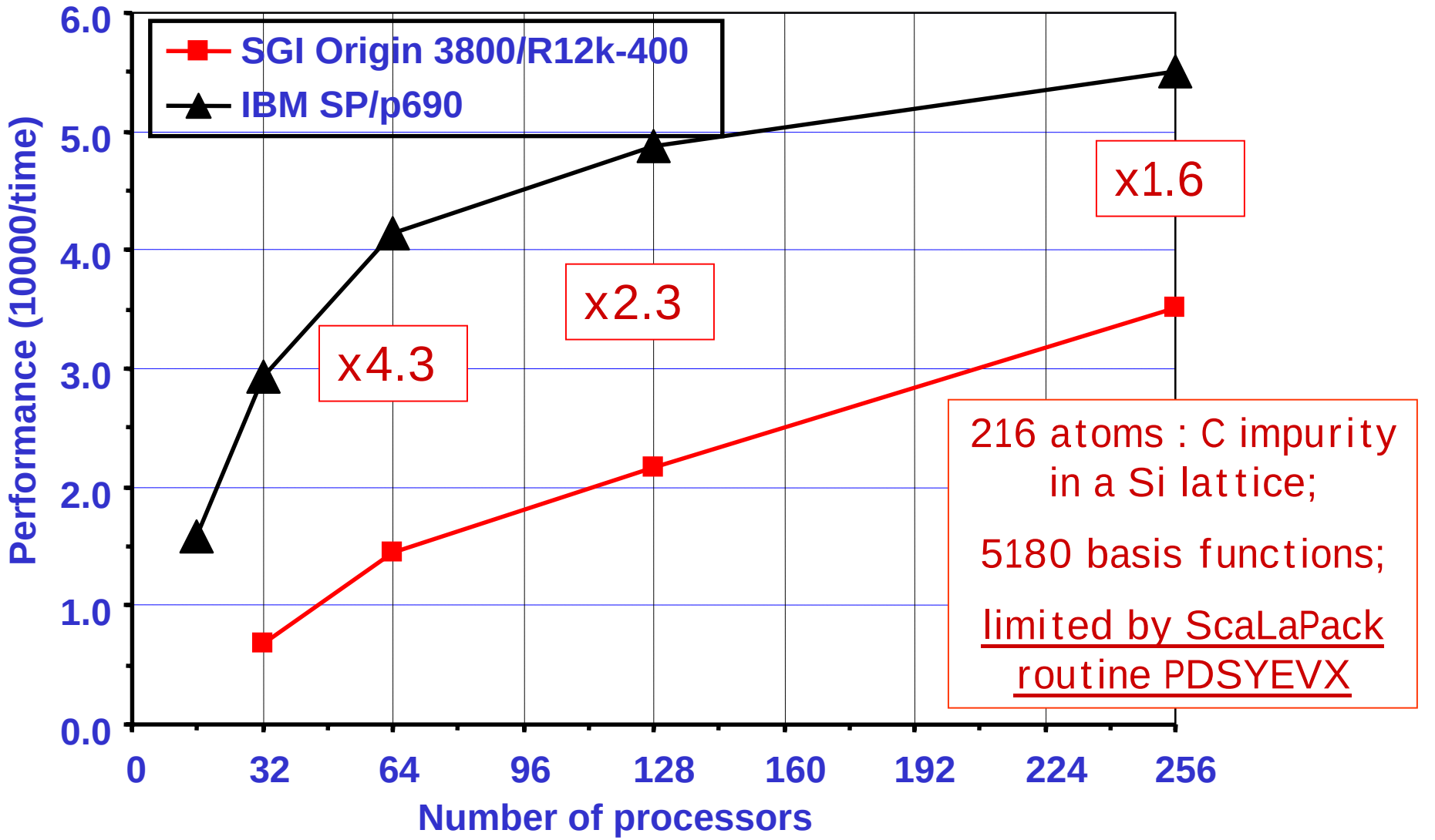
http://www.cse.clrc.ac.uk/cmg/CRYSTAL/

CASTEP

CAmbridge Serial Total Energy Package

http://www.cse.clrc.ac.uk/cmg/NETWORKS/UKCP/

The AIMPRO benchmark

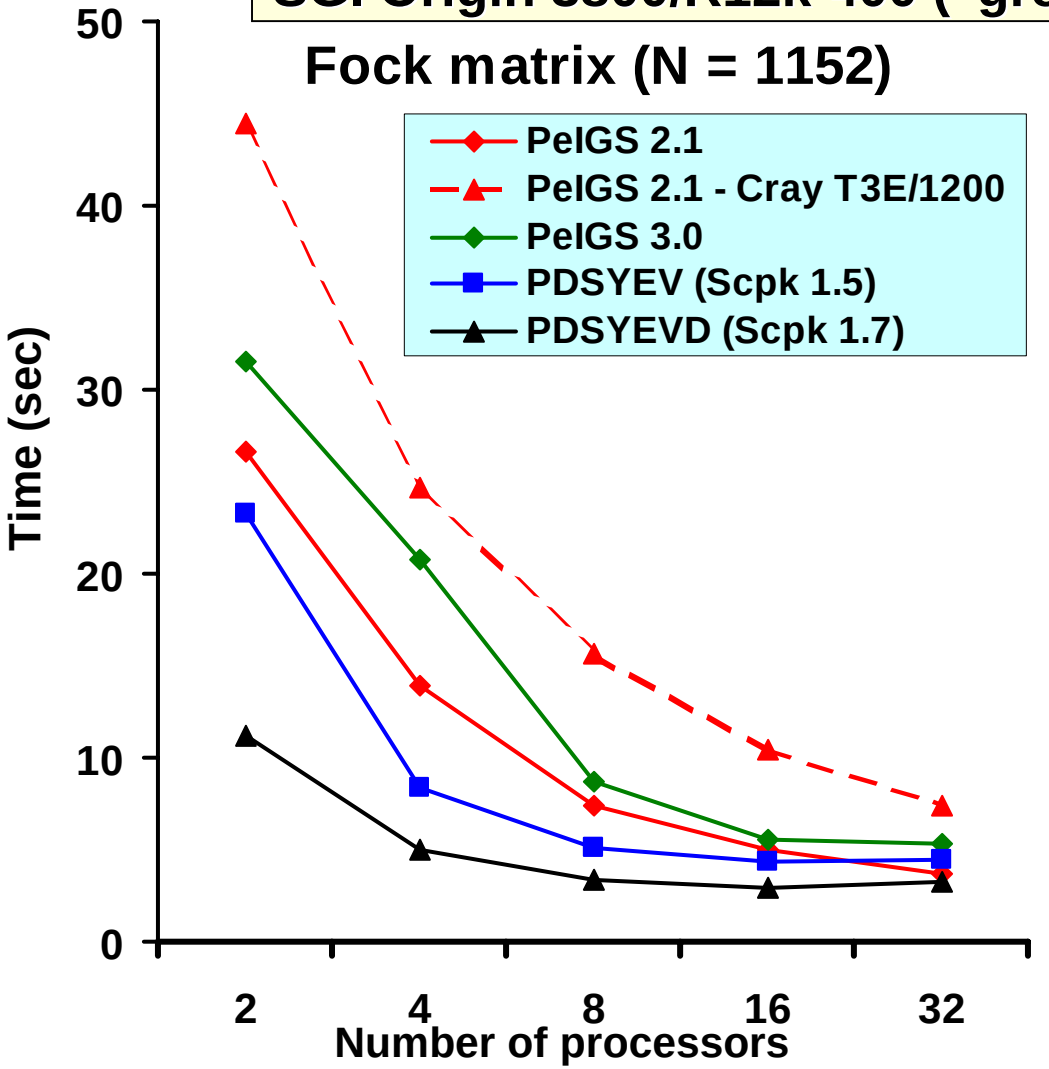


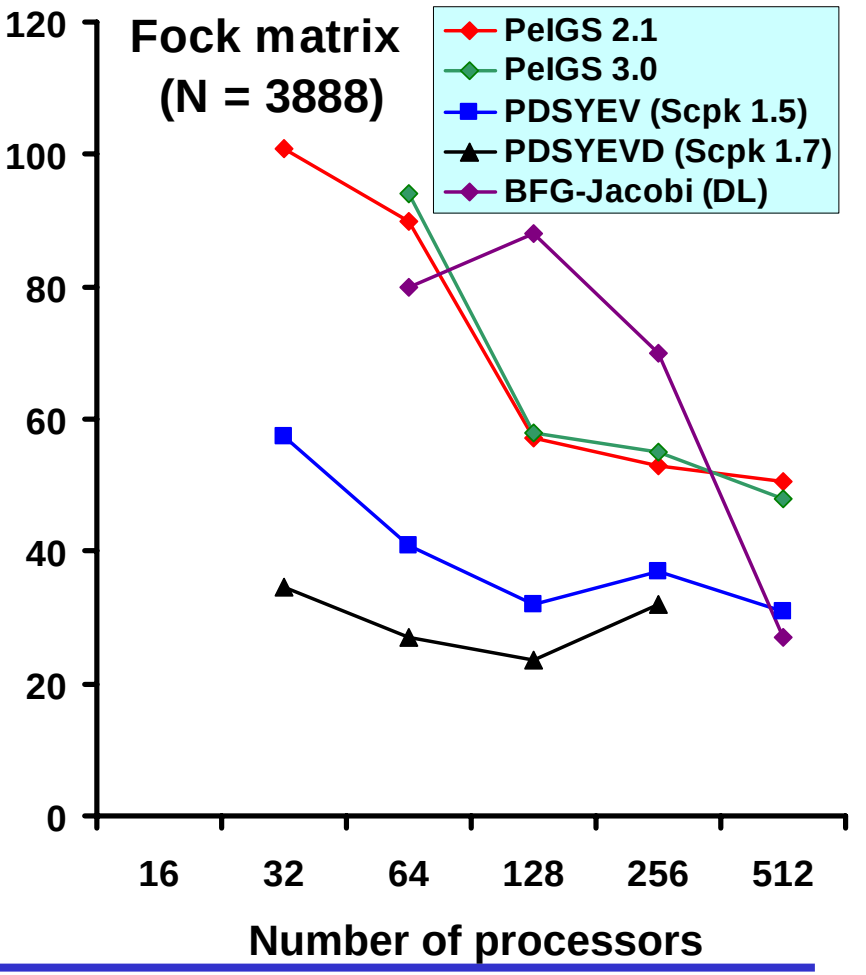
Scalability of Numerical Algorithms I.

HPCx CAPABILITY COMPUTING

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

Real symmetric eigenvalue problems

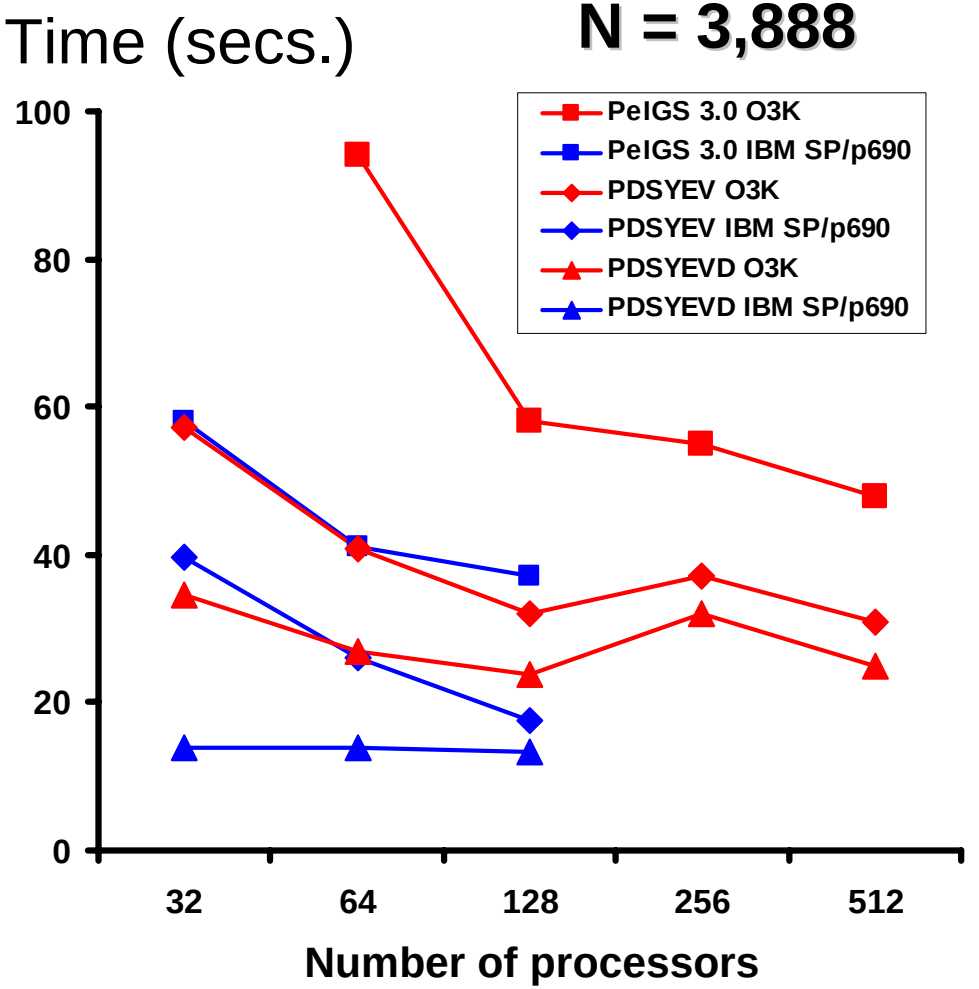




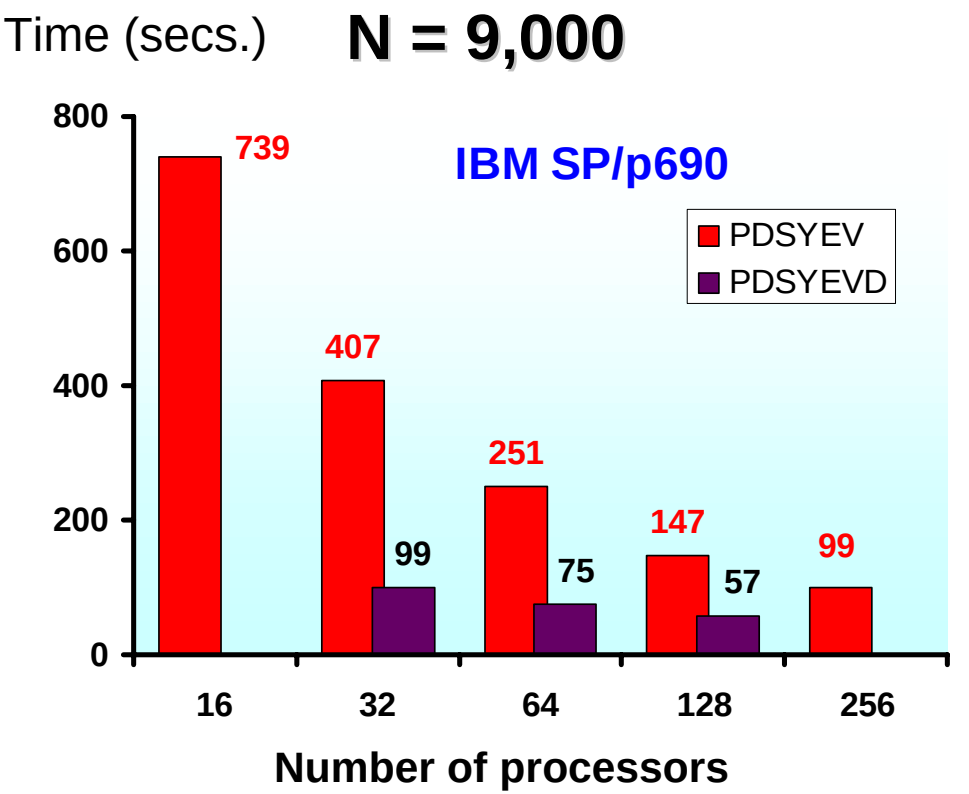
Scalability of Numerical Algorithms II.

HPCx CAPABILITY COMPUTING

IBM SP/p690 and SGI Origin O3800/R12k



Real symmetric eigenvalue problems



Materials Simulation. Plane Wave Methods: CASTEP

HPCx CAPABILITY COMPUTING

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).

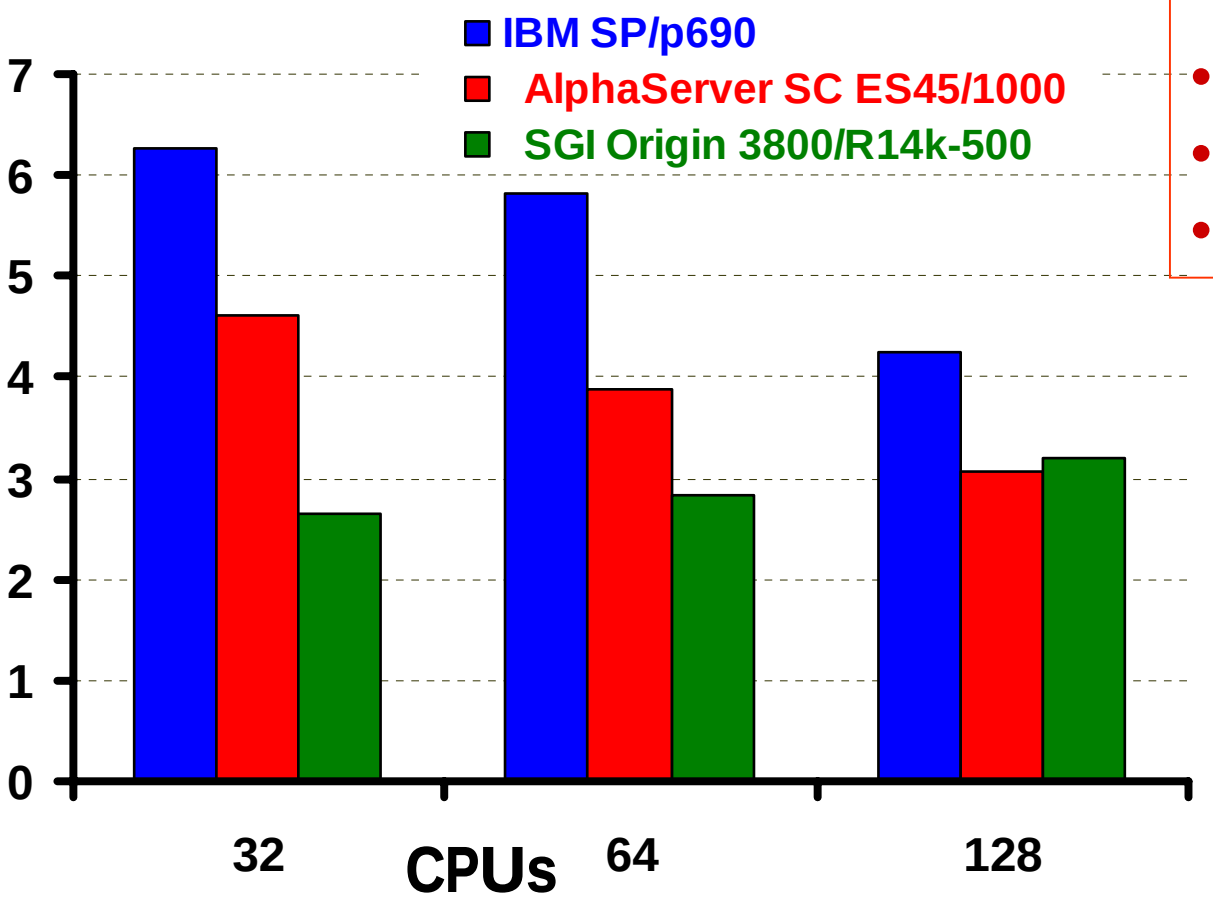
The plane wave expansion means that the bulk of the computation comprises large 3D Fast Fourier Transforms (FFTs) between real and momentum space.

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

CASTEP 4.2 - kG Parallel Benchmark

HPCx CAPABILITY COMPUTING

Performance Relative to the Cray T3E/1200E



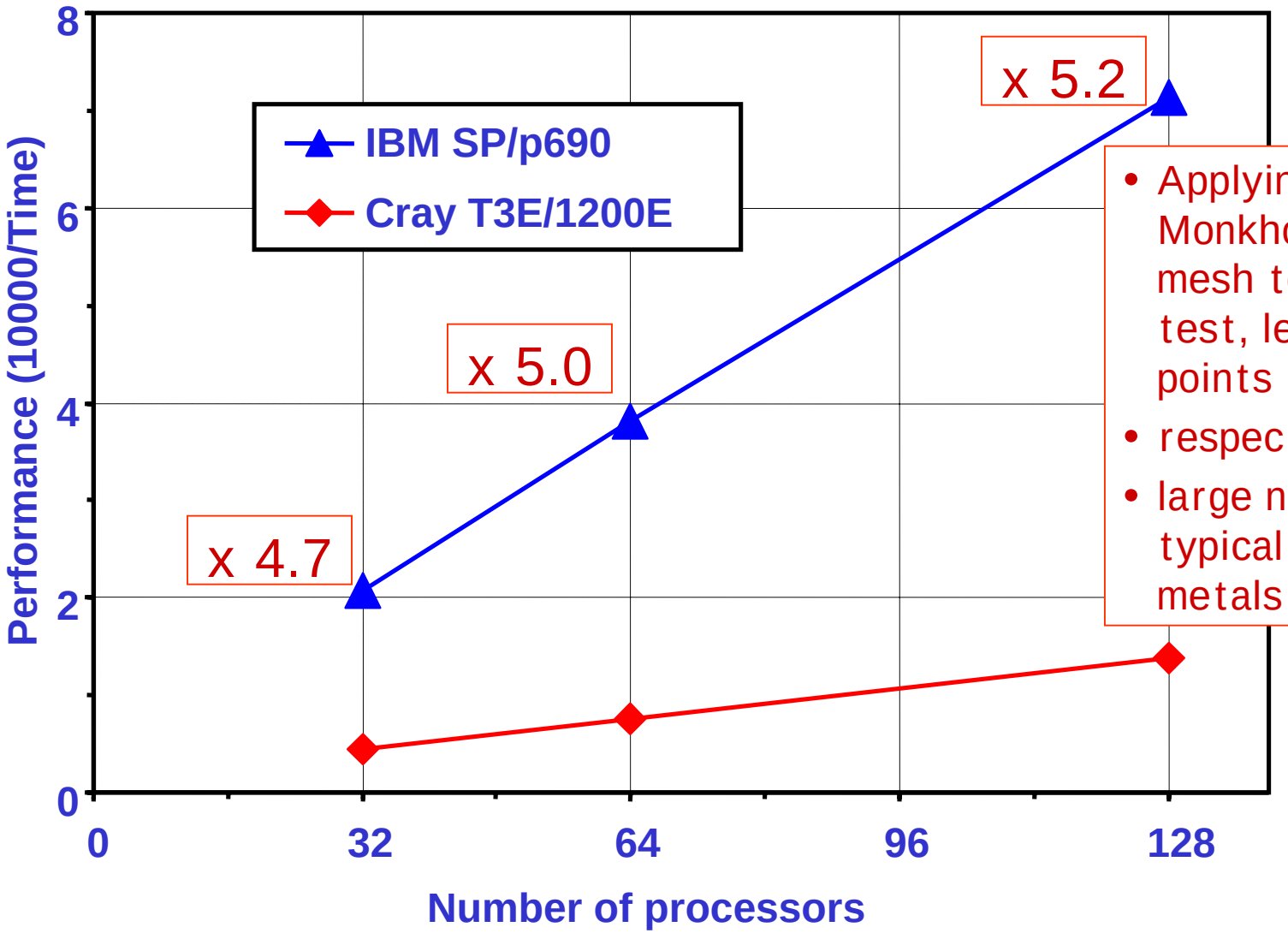
TiN: A 33 atom slab of TiN, 8 k points, single energy calculation

- 88,000 plane waves
- 3D FFT: 108X36X36
- Vanderbilt pseudopotential

Bottleneck:

Data Transformation associated with 3D FFT& MPI_AlltoAllV

CASTEP TiN Benchmark, 32 k-points

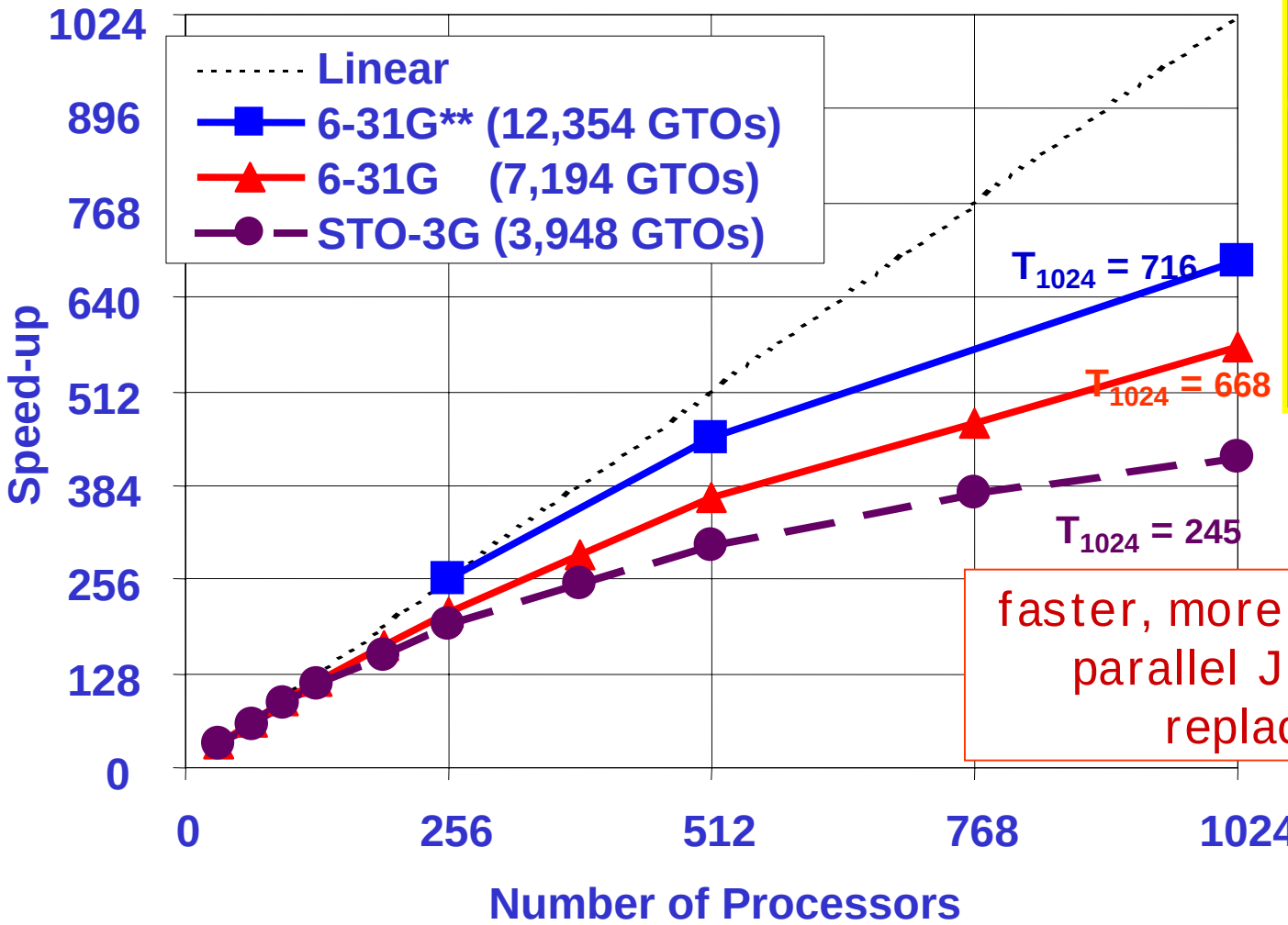


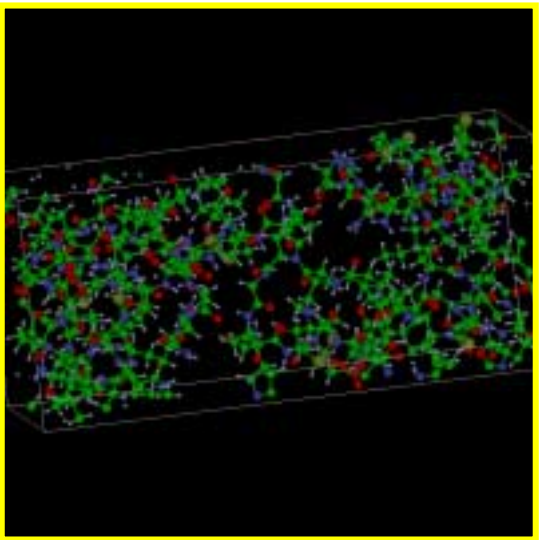
- Applying a dense Monkhorst Pack (MP) mesh to the 8 k-point test, leading to 32 k-points
- respectable performance
- large numbers of k-points typical for calculations of metals

Scalability of CRYSTAL for Crystalline Crambin

HPCx CAPABILITY COMPUTING

IBM SP/p690 HPCx





Structure of Crambin is derived from XRD data at 0.52 Å (1284 atoms).

faster, more stable version of the parallel Jacobi diagonalizer replaces ScaLaPack

Molecular Simulation

HPCx CAPABILITY COMPUTING

DL_POLY

W. Smith and T.R. Forester, CLRC Daresbury Laboratory

General purpose molecular dynamics simulation package

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

NAMD

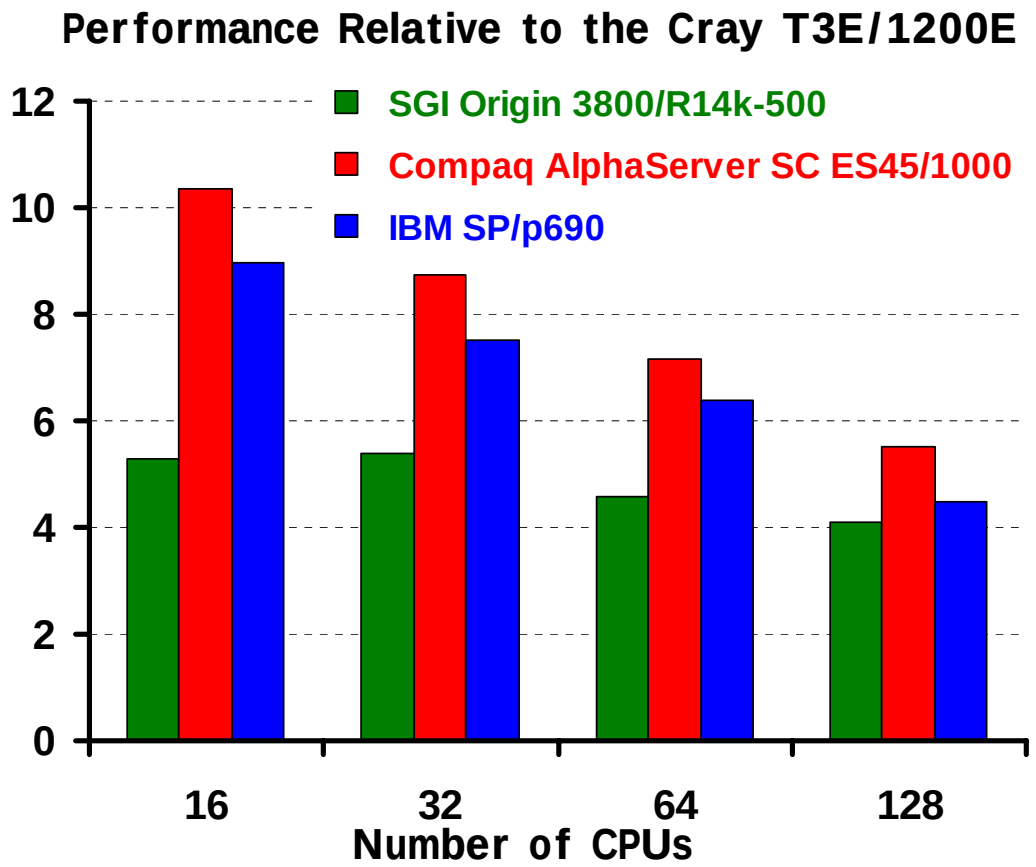
Theoretical and Computational Biophysics Group, NIH

- Parallel, object-oriented molecular dynamics code
- High-performance simulation of large biomolecular systems
- Scales to hundreds of processors on high-end parallel platforms

http://www.ks.uiuc.edu/Research/namd/

DL_POLY V2: Replicated Data

HPCx CAPABILITY COMPUTING

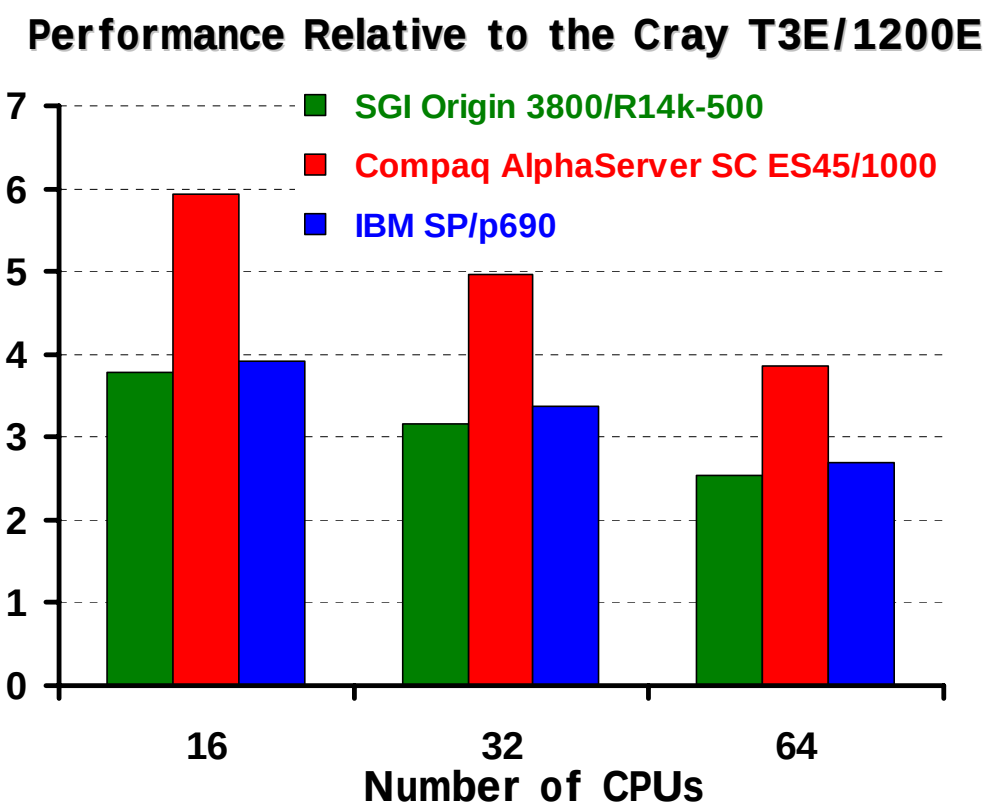


Bench 4. NaCl; 27,000 ions, Ewald, 75 time steps, Cutoff=24Å

Ionic Simulations

Macromolecular Simulations

Bench 7: Gramicidin in water; rigid bonds and SHAKE, 12,390 atoms, 500 time steps



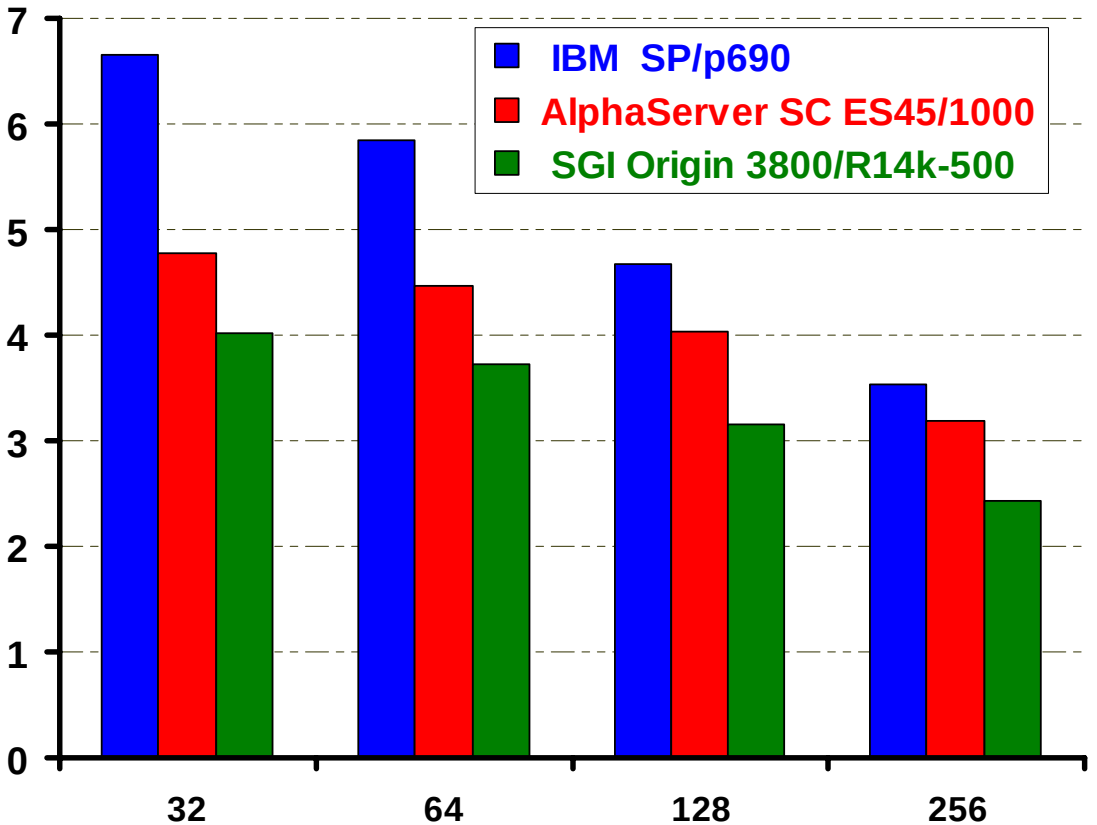
DL_POLY3 Coulomb Energy Performance

HPCx CAPABILITY COMPUTING

- Distributed Data
- SPME, with revised FFT Scheme

DL_POLY-3
216,000 ions, 200 time steps,
Cutoff=12Å

Performance Relative to the Cray T3E/1200E



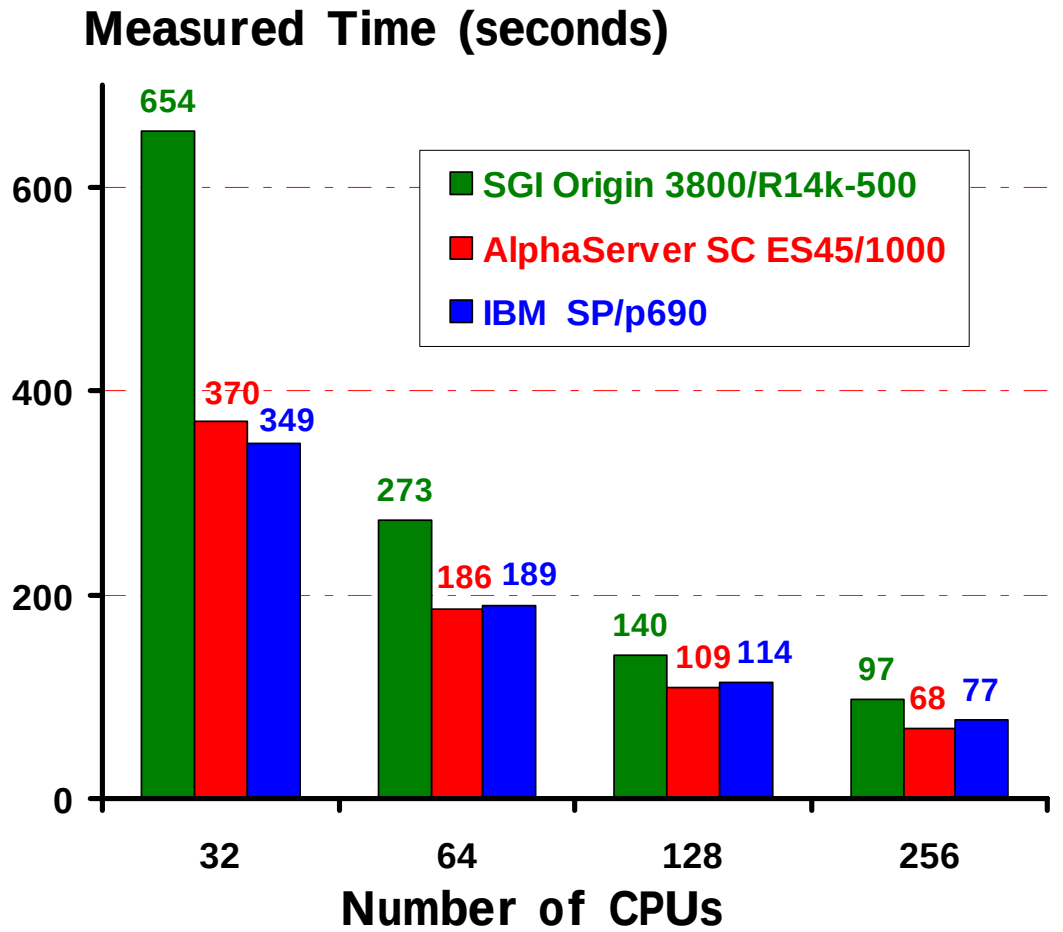
Alternative FFT algorithm to reduce communication costs:

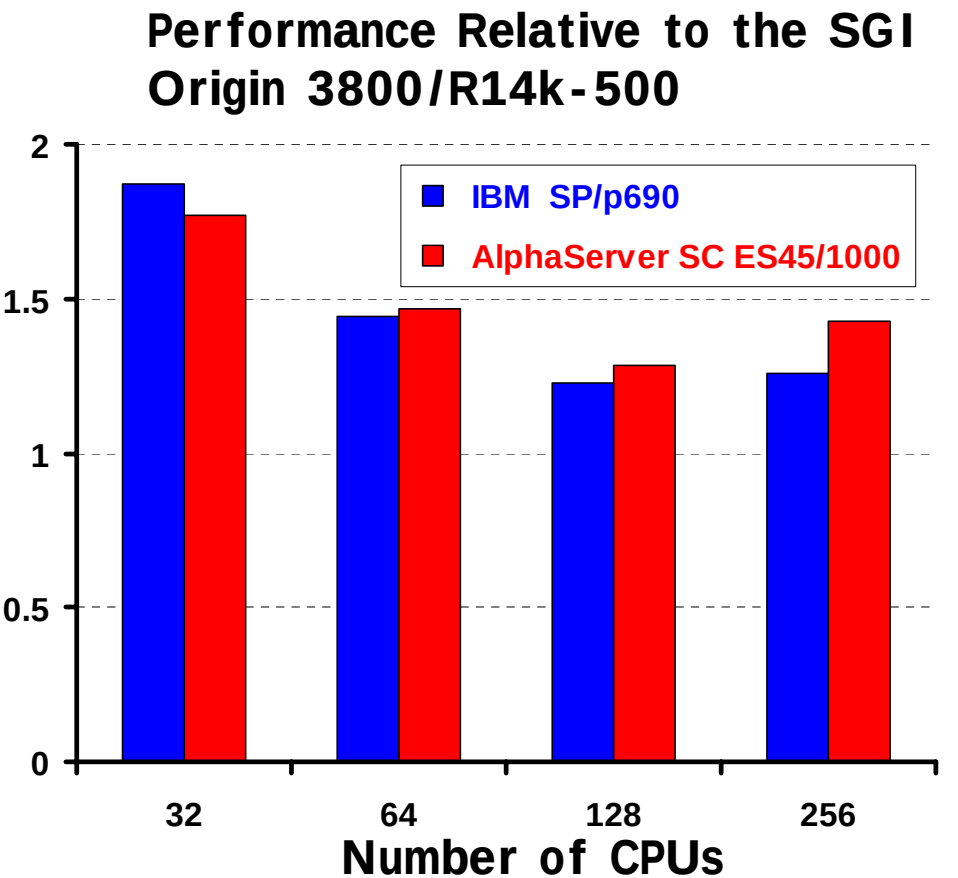
- 3D FFT performed as a series of 1D FFTs, each involving communications only between blocks in a given column
- More data is transferred, but in far fewer messages
- Rather than all-to-all, the communications are column-wise only

DL_POLY3 Macromolecular Simulations

HPCx CAPABILITY COMPUTING

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



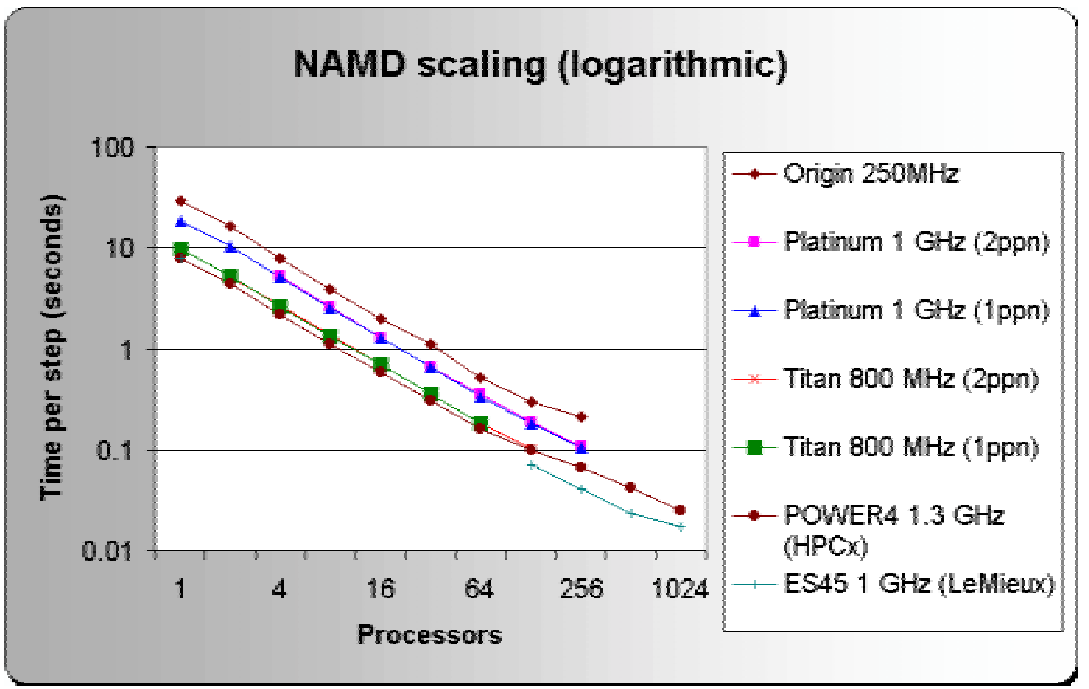


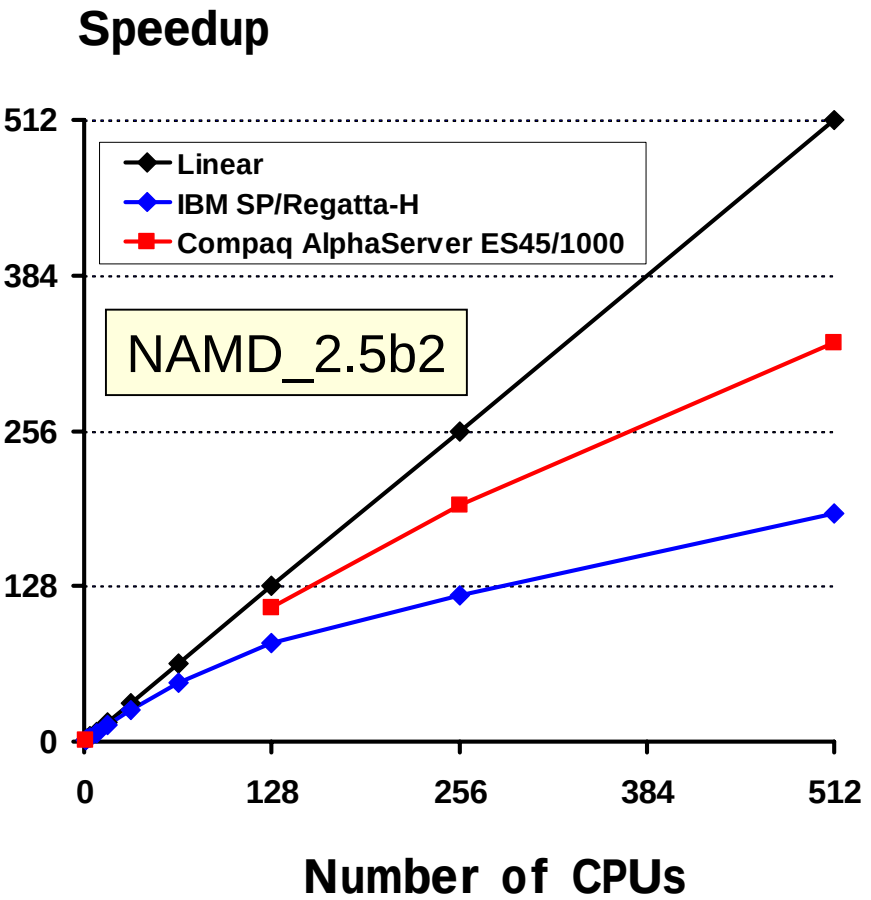
Molecular Simulation - NAMD Scaling

HPCx CAPABILITY COMPUTING

- standard NAMD ApoA-I benchmark, a system comprising 92,442 atoms, with 12Å cutoff and PME every 4 time steps.
- scalability improves with larger simulations - speedup of 778 on 1024 CPUs of TCS-1 in a 327K particle simulation of F_1-ATPase.

http://www.ks.uiuc.edu/Research/namd/



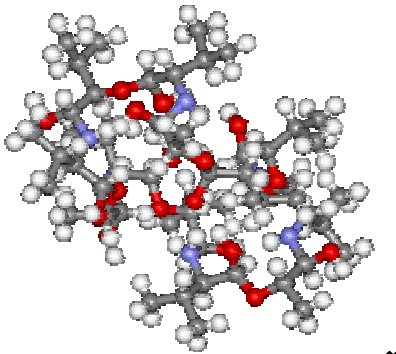


Ab Initio Molecular Electronic Structure

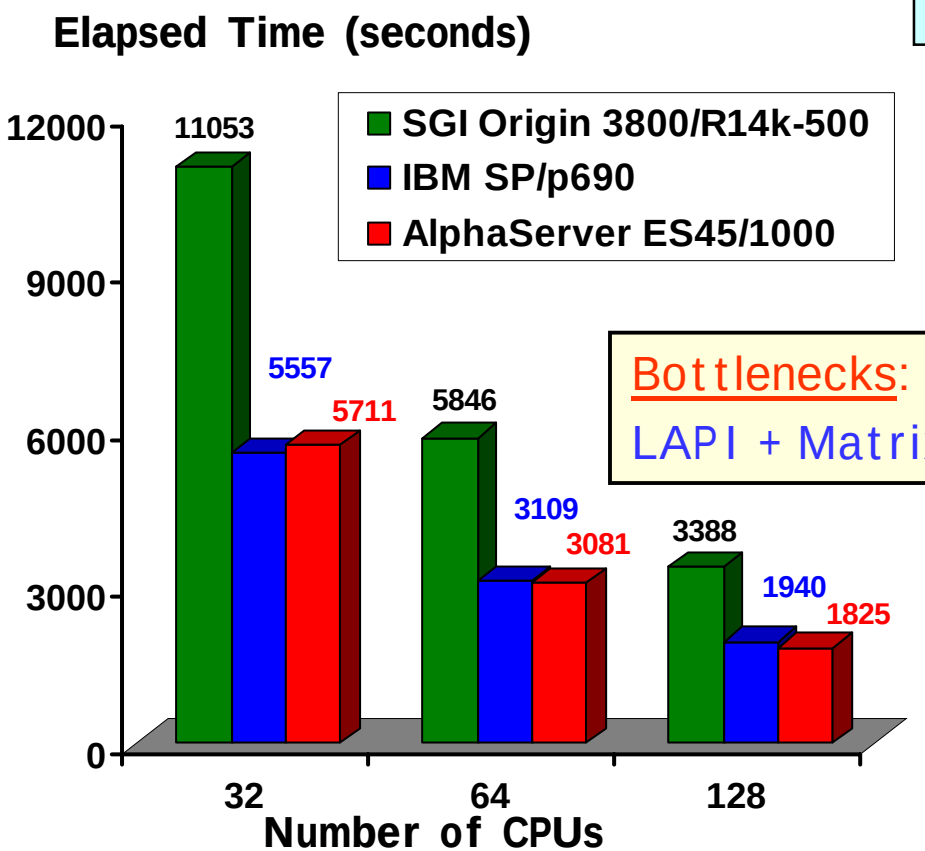
HPCx CAPABILITY COMPUTING

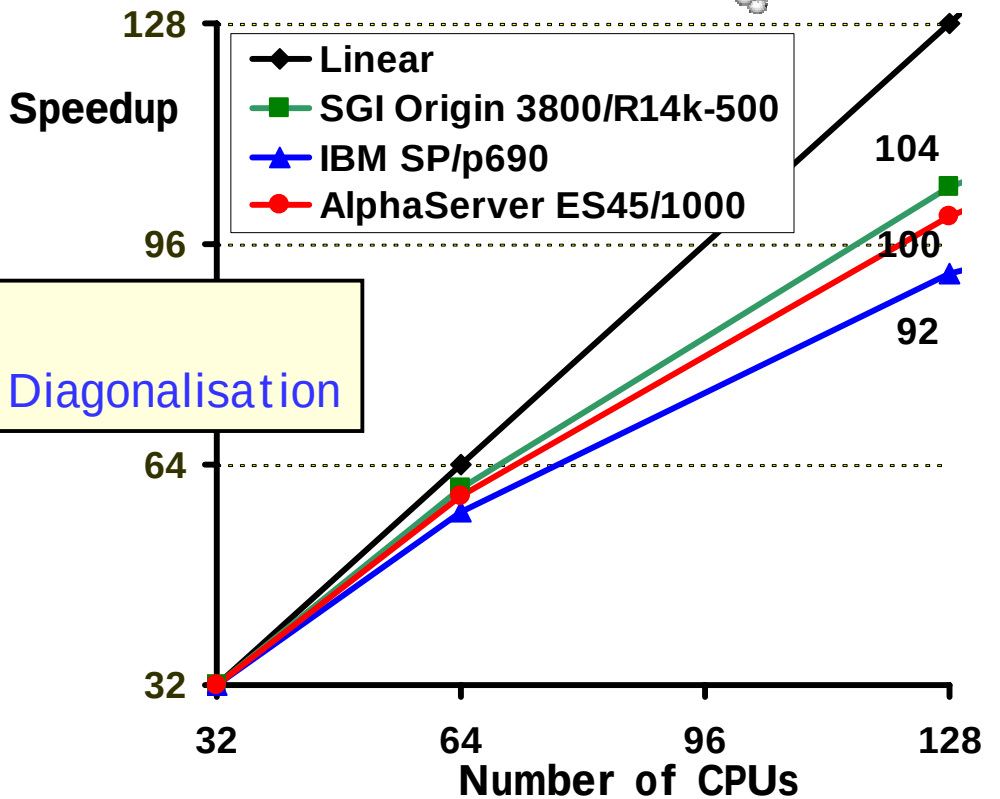
GAMESS-UK - DFT Calculations:

- Global Array (GA) Tools from PNNL
- Parallel Eigen Solvers (PeIGS)



Valinomycin (DFT HCTH):
Basis: DZVP2_A2 (Dgauss)
(1620 GTOs)





Bottlenecks:
LAPI + Matrix Diagonalisation

Atomic and Molecular Physics

PFARM

Queen's University Belfast,
CLRC Daresbury Laboratory

R-matrix formalism to treat applications such as the description of the edge region in Tokamak plasmas (fusion power research) and for the interpretation of astrophysical spectra

H2MOL

Queen's University Belfast

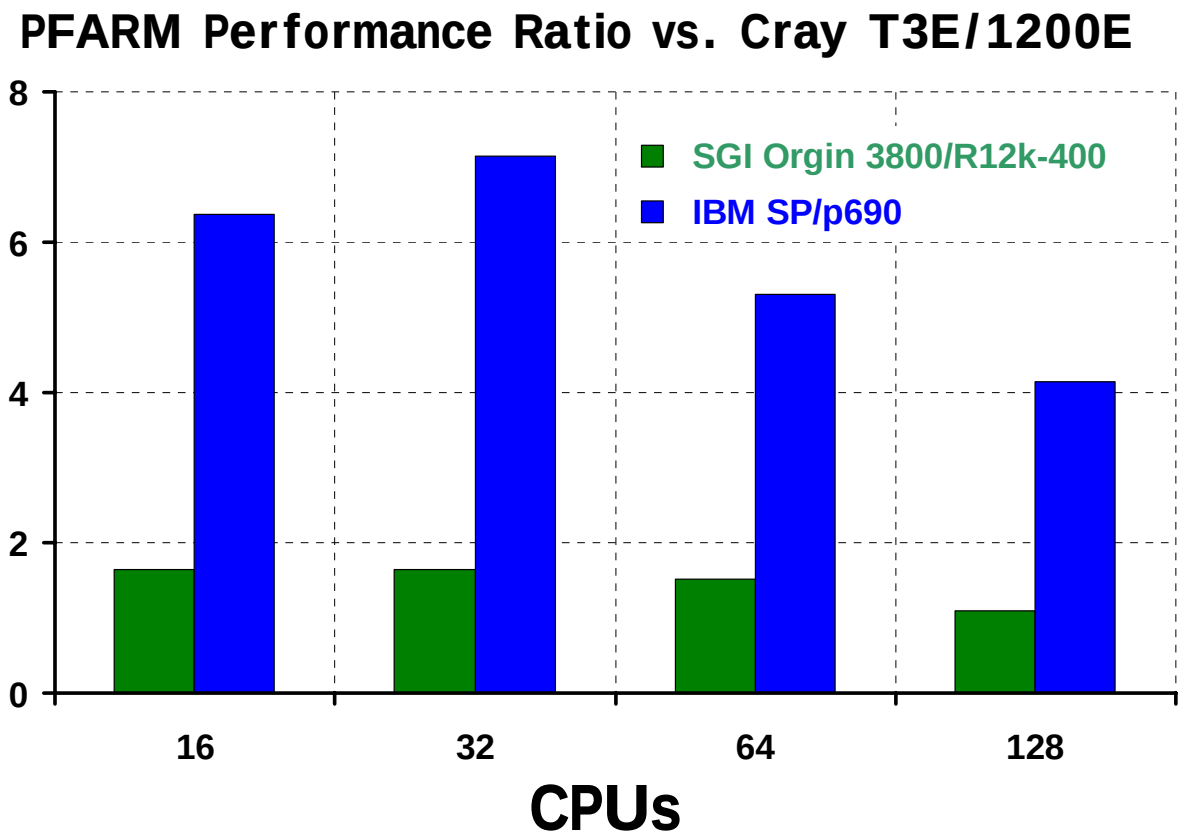
Solves the time-dependent Schrodinger equation to calculate energy distributions for laser-driven dissociative ionization of H_2 molecule.

A&M Physics: Electron-Atom Collisions

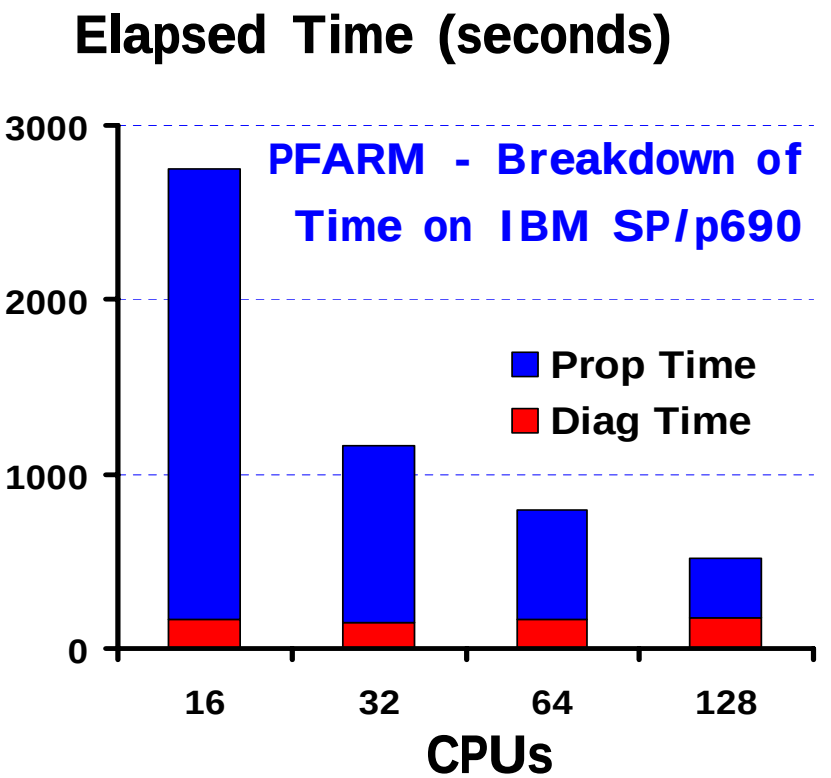
HPCx CAPABILITY COMPUTING

- R-matrix theory - efficient methods for investigating electron-atom and electron-molecule collisions.
- Calculation involves integration of up to 10^3 coupled channels i.e. 2^{nd} order linear differential equations.
- External Region Calculation Timings:
  - Data from internal region calculations (from disk)
  - 2 stage approach - Diagonalisation [PeIGS (20%, 4K)] and functional task parallelisation (80%) - BLAS3 dominated
  - Systolic processor pipeline approach.
  - Coarse-grained parallelism ensures scalable performance.
  - Asynchronous communications minimises communication costs.
- Benchmark Example for PFARM application

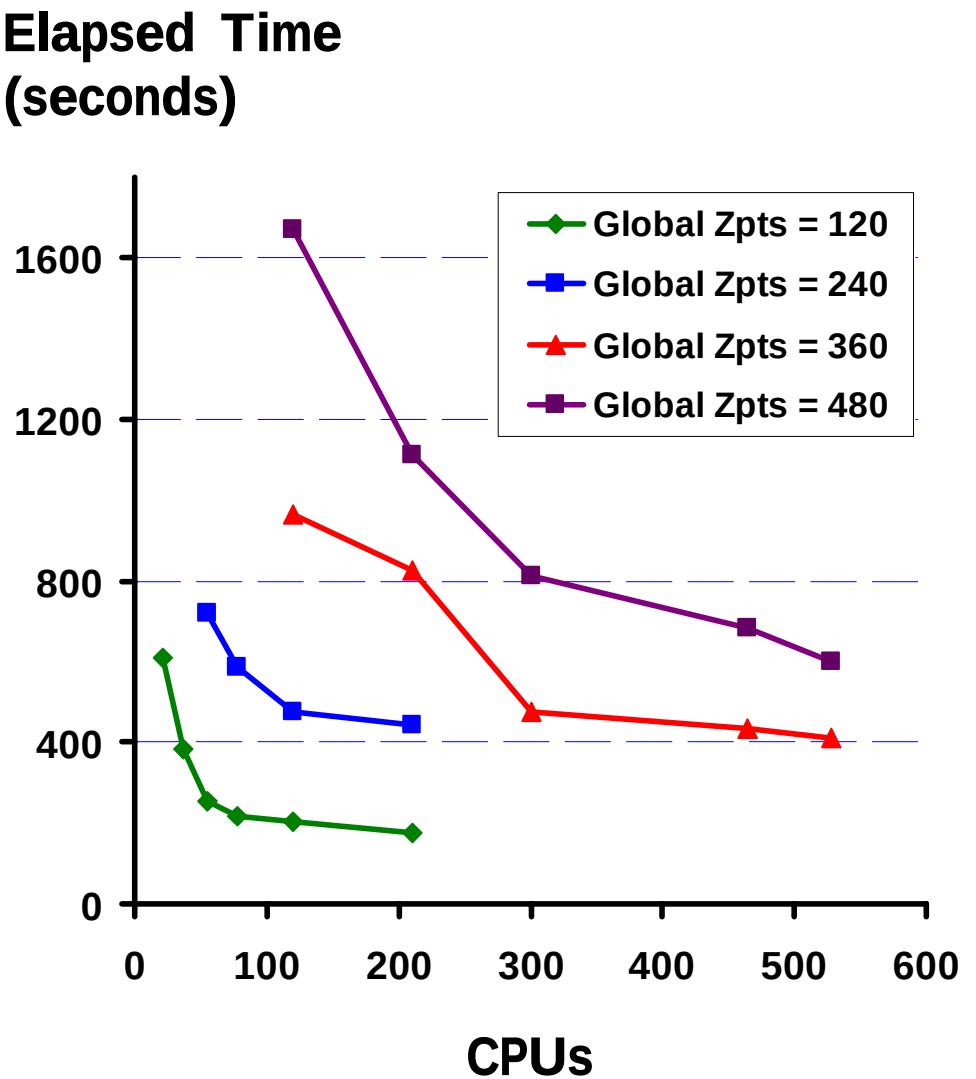
External Region Calculation Timings



Bottleneck:
Matrix Diagonalisation



H2MOL - Performance on the IBM SP/p690



- Solves the time-dependent Schrodinger equation to calculate energy distributions for laser-driven dissociative ionization of H_2 molecule.
- Cylindrical computational grid of ϕ, ρ and Z co-ordinates. Z points are distributed over processors arranged logically in a triangular grid.
- Most time spent calculating 5-point finite difference schemes and in ZGEMM. MPI collectives relatively expensive for extremely large processor grids.
- Main optimisations: improving ZGEMM performance for small matrix sizes and using asynchronous message passing.
- Improved scalability for larger grid sizes.

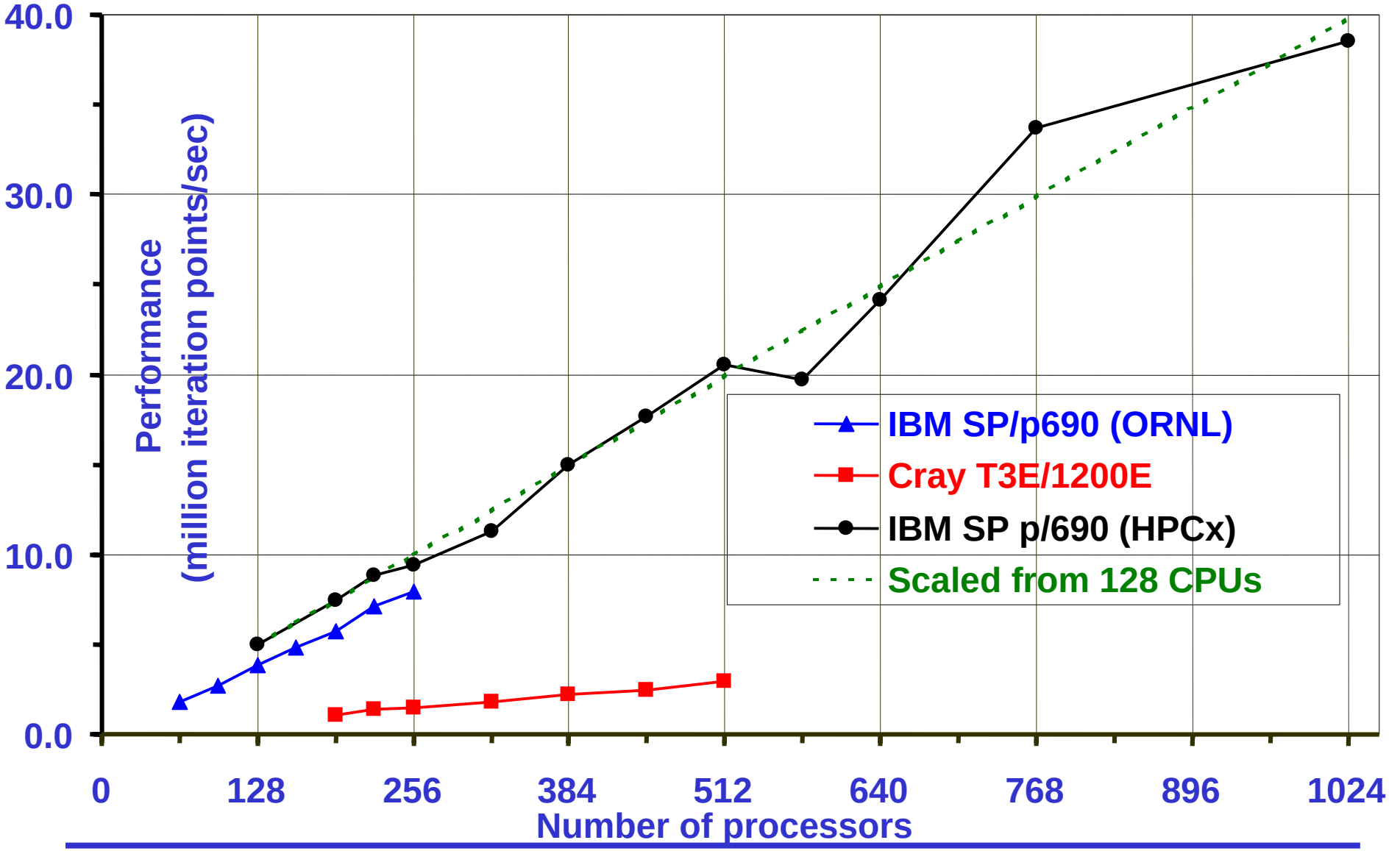
UK Turbulence Consortium

Led by Prof. Neil Sandham, University of Southampton

- Focus on compute-intensive methods (Direct Numerical Simulation, Large Eddy Simulation, etc) for the simulation of turbulent flows
- Shock boundary layer interaction modelling - critical for accurate aerodynamic design but still poorly understood

http://www.afm.ses.soton.ac.uk/

Direct Numerical Simulation: 360^3 benchmark

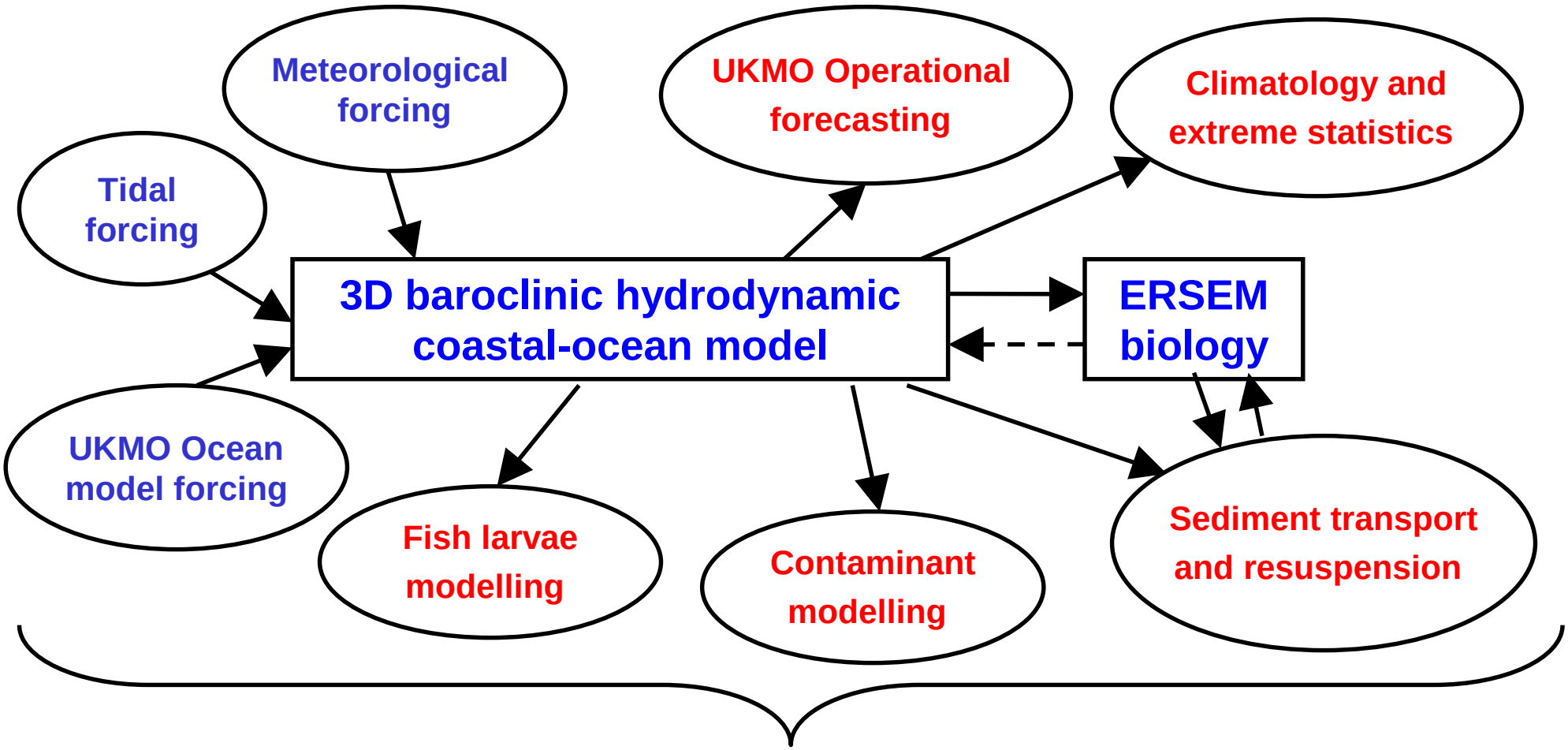


Proudman Oceanographic Laboratory Coastal Ocean Modelling System (POLCOMS)

Multidisciplinary Studies in coastal/shelf environments

http://www.pol.ac.uk/home/research/polcoms/

POLCOMS Structure

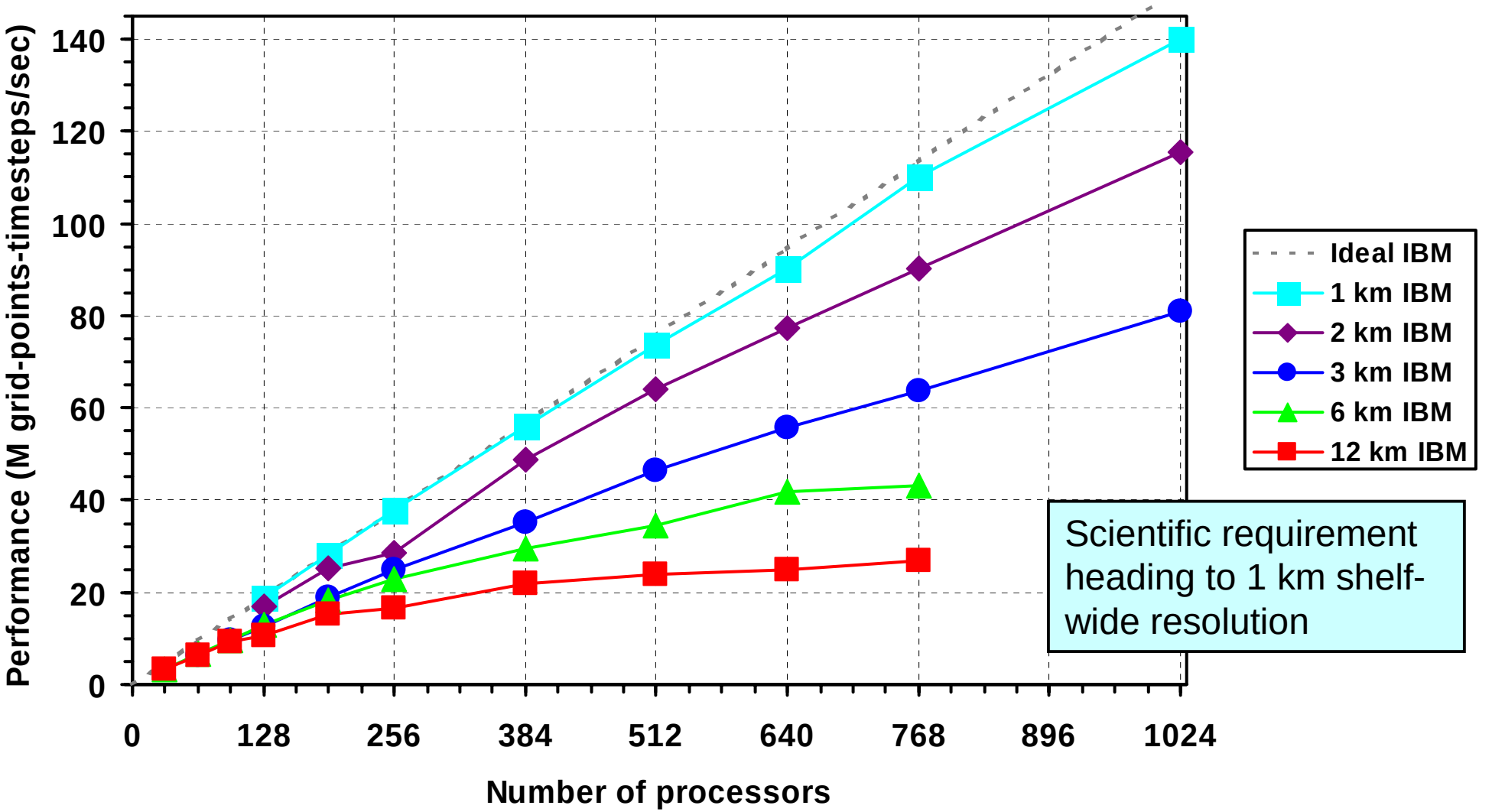


Realistic physical forcing to interact with, and transport, environmental parameters

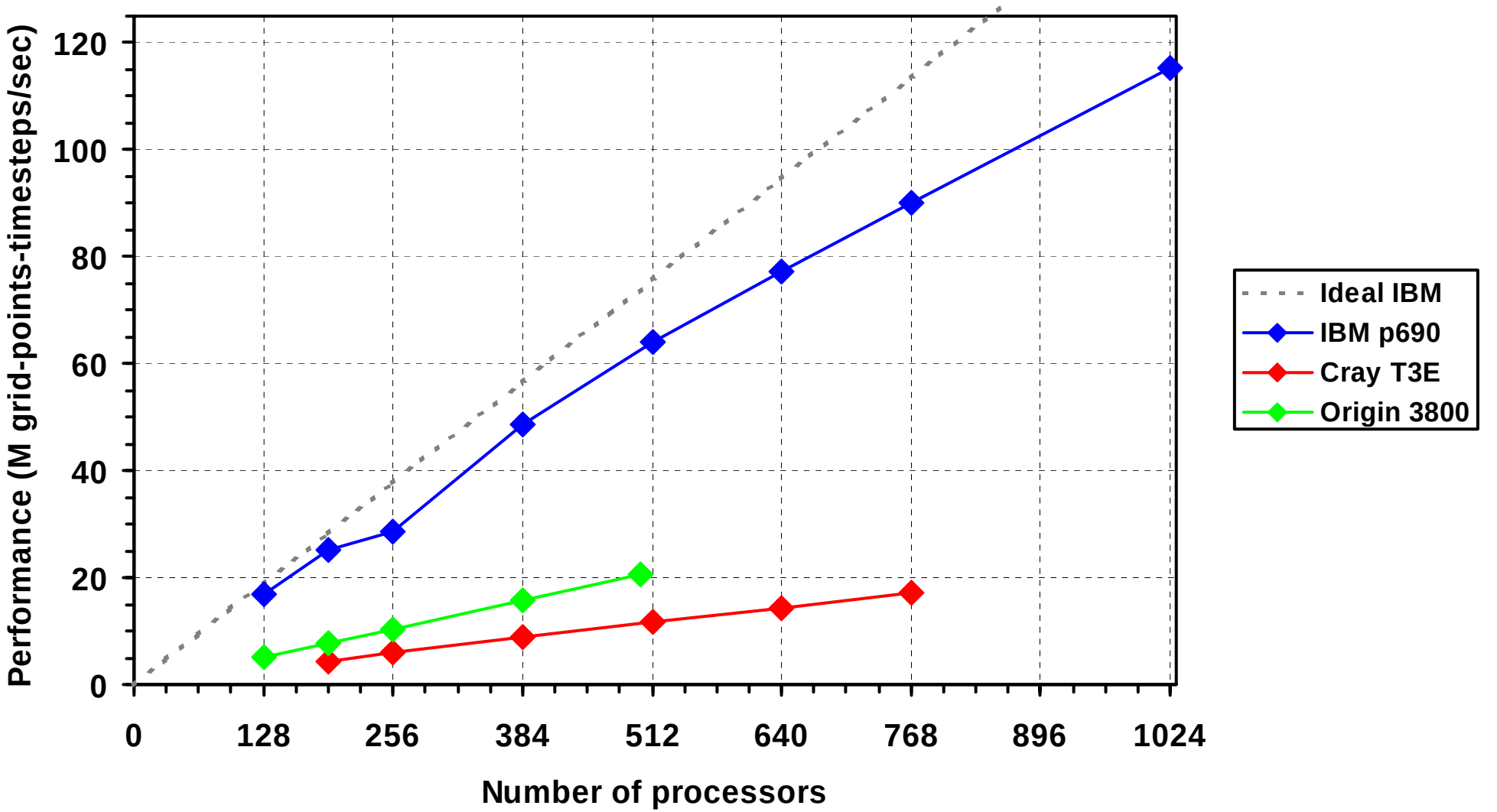
POLCOMS resolution benchmark

- Standard workhorse model is a 12 km resolution grid covering the whole of the north-west European shelf (198 x 224 x 34)
- Want to maintain accuracy in the presence of eddies, fronts, steep topography, thermoclines etc.
- Scientific requirement heading to 1 km shelf-wide resolution
- Design set of benchmarks
  - 12 km (200 x 200 x 34)
  - 6 km (400 x 400 x 34)
  - 3 km (800 x 800 x 34)
  - 2 km (1200 x 1200 x 34)
  - 1 km (2400 x 2400 x 34)
- Fixed number of timesteps => decreasing run length
- Short run, subtract start-up and shut-down times
- Performance metric is gridpoints * timesteps / time

POLCOMS resolution b/m : IBM SP/p690



POLCOMS 2 km b/m : All systems



Strategy for Capability Computing

HPCx CAPABILITY COMPUTING

1. Performance Attributes of Key Applications
   - Trouble-shooting with Vampir & Paraver
2. Scalability of Numerical Algorithms
   - Parallel eigensolvers
3. Optimisation of Communication Collectives
   - MPI_ALLTOALLV and CASTEP
4. Memory-driven Approaches
   - in-core SCF & DFT, direct minimisation & CRYSTAL
5. Terascaling Applications
   - NWChem, NAMD ...
6. Migration from replicated to distributed data
   - DL_POLY-3
7. Scientific drivers amenable to Capability Computing
   - Enhanced Sampling Methods, Replica Methods

HPCx Terascale Applications Team

Efficient Serial Execution

Summary

- UK has a New Facility for Capability Computing: HPCx
  - 66% Technology, 33% Support
- Key Strategic Applications Areas
  - Materials Science, Molecular Simulation, Molecular Electronic Structure, A&M Physics, Computational Engineering, Environmental Science
- HPCx Terascale ApplicationsTeam
  - Strategy for Capability Computing
- Range of Performance Results
  - size matters !
  - limited scalability for applications:
    - with global communications (CASTEP)
    - featuring linear algebra routines with extensive communication requirements (AIMPRO)
  - Linear scaling to 1024 processors for nearest neighbour CFD codes (PDNS3D, POLCOMS)

Acknowledgements

- HPCx Terascaling Team
  - Mike Ashworth
  - Ian Bush
  - Martyn Guest
  - David Henty
  - Martin Plummer
  - Lorna Smith
  - Kevin Stratford
  - Andrew Sunderland
- IBM Technical Support
  - Luigi Brochard et al.
- NCSA Rick Kufrin (NAMD)
- CSAR Computing Service — Cray T3E ‘turing’, Origin 3800 R12k-400 ‘green’
- ORNL — IBM Regatta ‘cheetah’
- SARA — Origin 3800 R14k-500 ‘teras’
- PSC — AlphaServer SC ES45-1000