Performance of Cray Systems – Kernels, Applications & Experiences

Mike Ashworth, Miles Deegan, Martyn Guest, Christine Kitchen, Igor Kozin and Richard Wain

Computational Science and Engineering Department
CCLRC Daresbury Laboratory
Warrington
UK

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

CCLRC

Outline

- Introduction
- Application benchmarks on high-end systems
  - including Cray X1, Cray XT3 vs. systems from IBM and SGI
- The Cray XD1 as a mid-range computing resource
  - comparison with Pathscale Infinipath clusters
- Cray XD1
  - System, installation experiences at CCLRC
- FPGAs
  - early experiences of the Virtex IVs on our Cray XD1

Introduction

CCLRC Daresbury Laboratory

CSE Computational Science & Engineering Department

Collaborative Computational Projects

CSE Computational Science & Engineering Department

- CCP1 : Quantum Chemistry (Prof P Knowles)
- CCP2 : Atomic & Molecular Physics (Prof E Armour)
- CCP3 : Surface Science (Prof A Fisher)
- CCP4 : Protein Crystallography (Prof J Naismith)
- CCP5 : Molecular Simulation (Dr J Harding)
- CCP6 : Heavy Particle Dynamics (Prof J Hutson)
- CCP7 : Astronomical Spectra (Prof D Flower)
- CCP9 : Electronic Structure of Solids (Dr J Annett)
- CCP11 : Biosequences and Function (Prof D Moss)
- CCP12 : Computational Engineering (Dr S Cant)
- CCP13 : Fibre Diffraction (Prof J Squire)
- CCP14 : Powder and Single Crystal Diffraction (Dr J Cockcroft)
- CCP-B : CCP for Biomolecular Simulation (Prof C Laughton)
- CCP-N: NMR in structural biology (Dr E Laue)

Funded by:

EPSRC
BBSRC
PPARC

CCLRC

- Supports a large range of scientific and computational activities around the CCPs

- High-End Computing Programme
  - Evaluation of Novel Architecture Systems
    - Cray X1/E, Cray XT3, IBM BG/L, Liquid Computing, Lightfleet in contrast with
    - Cray XD1, IBM POWER4 and POWER5, SGI Altix plus clusters
    - Full-scale applications drawn from Computational Chemistry, Materials, Engineering and Environmental Science
  - Evaluation of Parallel Languages
    - Inter-comparison of Co-Array Fortran, UPC and Titanium
    - Forward look at Cray's Chapel, IBM's X10 and Sun's Fortress (HPCS)
  - Parallel Performance Tools
    - KOJAK, PAPI and TAU/PDT and MARMOT
  - FPGAs
  - Parallel Input/Output
    - Implementation of MPI-IO into a real application code

CCLRC

The HPCx Project

- Operated by CCLRC Daresbury Laboratory and the University of Edinburgh
- Located at Daresbury
- Six –year project from November 2002 through to 2008
- First Tera-scale academic research computing facility in the UK
- Funded by the Research Councils: EPSRC, NERC, BBSRC
- IBM is the technology partner
  - Phase1: 3 Tflops sustained - 1280 POWER4 cpus + SP Switch
  - Phase2: 6 Tflops sustained - 1600 POWER4+ cpus + HPS
  - Phase2A: Performance-neutral upgrade to 1536 POWER5 cpus
  - Phase3: 12 Tflops sustained – approx. 3000 POWER5 cpus

CCLRC

The current Phase2A HPCx

- Current UK Strategy for HPC Services
  - Funded by the UK Government Office of Science & Technology
  - Managed by EPSRC on behalf of the community
  - Initiate a competitive procurement exercise every 3 years
    - for both the system (hardware and software) and
    - service (management, accommodation and CSE support)
  - Award a 6 year contract for a service
  - Consequently, have 2 overlapping services at any one time
  - The contract to include at least one, typically two, technology upgrades
  - Resources allocated on both national systems through normal peer review process for research grants
  - Virement of resources is allowed e.g. when one of the services has a technology upgrade

CSAR National Service

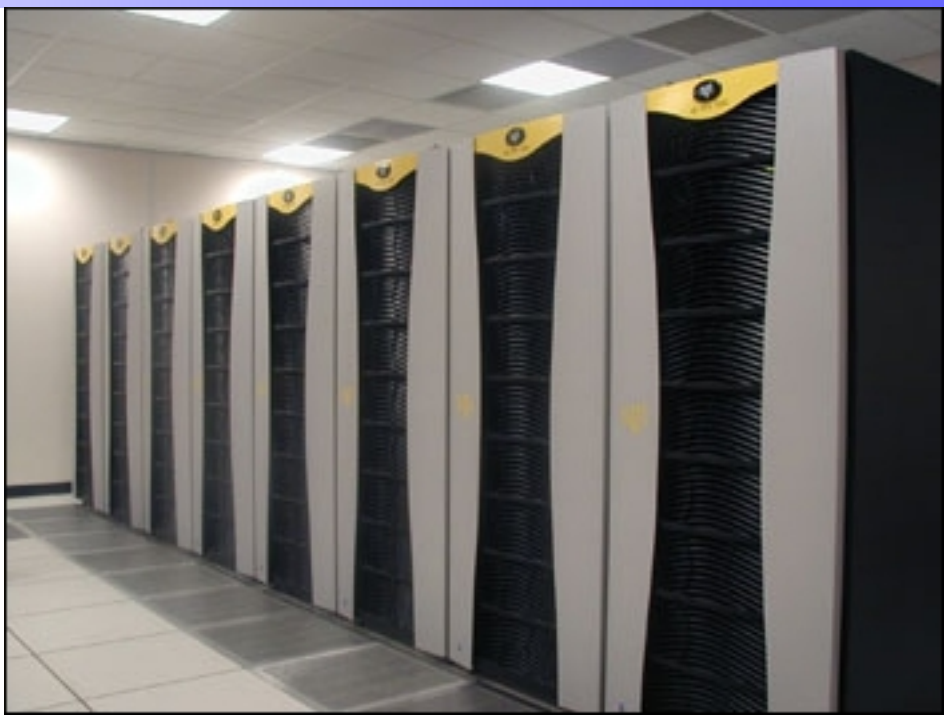
CSAR at the University of Manchester currently provides the following high performance systems:

newton: SGI Altix 3700 system - 512 Itanium2 cpus, 1 TB memory

green: SGI Origin 3800 system - 512 MIPS R12000 cpus, 512 GB GB

fermat: SGI Origin 2000 system - 128 MIPS R12000 cpus, 128 GB

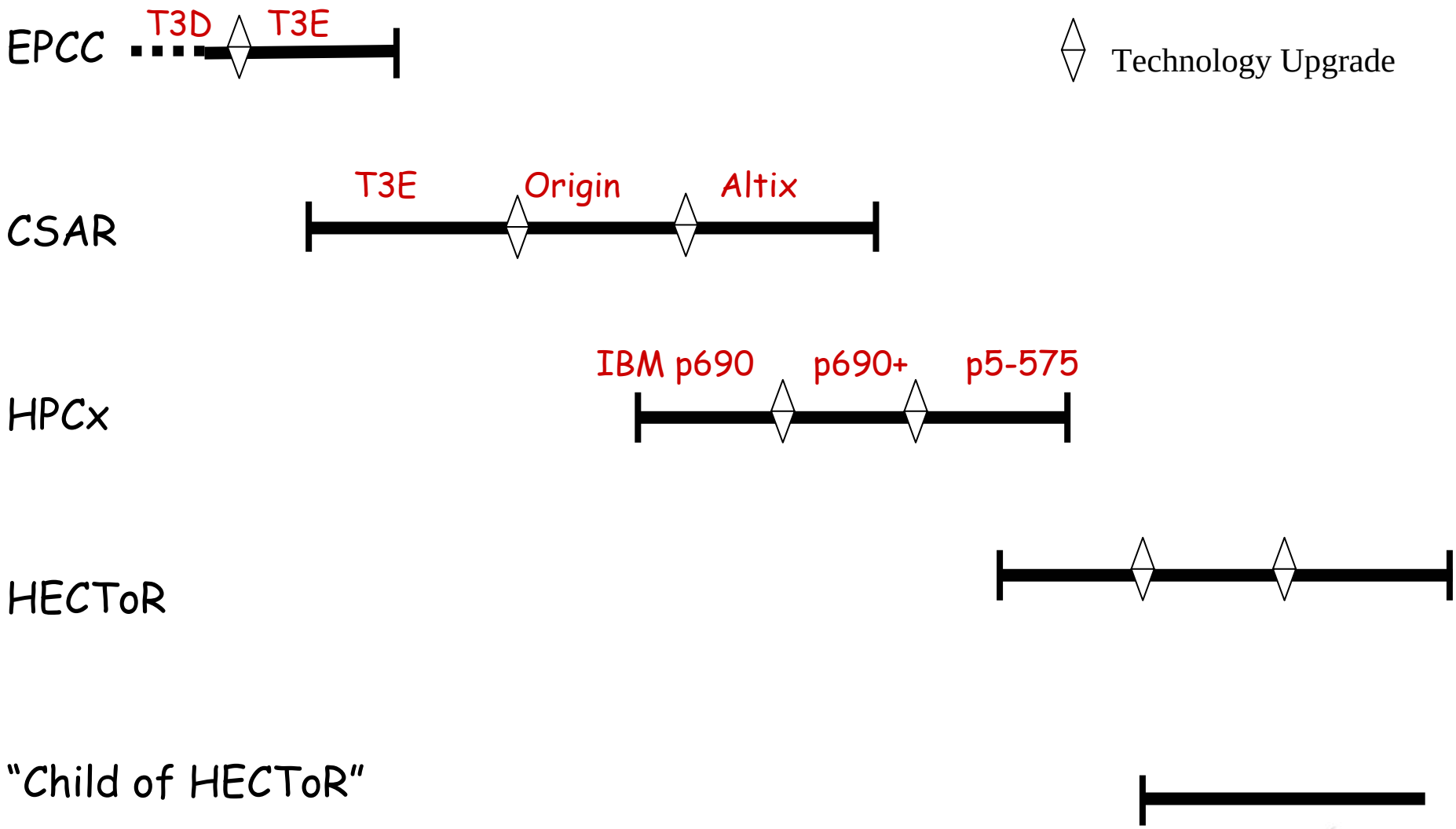
wren: SGI Origin 300 system - 16 MIPS R14000 cpus, 16 GB



CSAR services due to close at the end of June 2006

CCLRC

UK Overlapping Services



HECToR and Beyond

- High End Computing Technology Resource
  - Budget capped at £100M – including all costs over 6 years
  - Three phases with performance doubling (like HPCx)
  - 50-100 Tflop/s, 100-200 Tflop/s, 200-400 Tflop/s peak
  - Three separate procurements
    - Hardware technology (shortlist of vendors)
    - CSE support (tender issued April 2006)
    - Accommodation & Management (tender issued April 2006)

- "Child of HECToR"
  - Competition for better name!
  - Collaboration with UK Met Office
  - HPC Eur: european scale procurement and accommodation
  - Procurement due Sep 2006

- The largest ever consortium to support UK academic research using HPC:
  - University of Edinburgh
  - University of Manchester
  - CCLRC Daresbury Laboratory

"To provide UK researchers with unprecedented breadth and depth of support in the application of HPC technology to the most challenging scientific and engineering problems"

- Established through MoUs between partners
- Limited Company with Board of Directors
- Increasingly 'individual' activities will come under the HPC-UK banner
- To bid into the HECToR CSE support and A&M calls
- To position UK HPC support on a European scale

http://www.hpc-uk.ac.uk/

CCLRC

Application benchmarks on high-end systems

CCLRC

Systems under Evaluation

Make/Model	Ncpus	CPU	Interconnect	Site
Cray X1/E	4096	Cray SSP	CNS	ORNL
Cray XT3	1100	Opteron 2.6 GHz	SeaStar	CSCS
Cray XD1	72	Opteron 250 2.4 GHz	RapidArray	CCLRC
IBM	1536	POWER5 1.5 GHz	HPS	CCLRC
SGI	384	Itanium2 1.3 GHz	NUMAlink	CSAR
SGI	128	Itanium2 1.5 GHz	NUMAlink	CSAR
Streamline cluster	256	Opteron 248 2.2 GHz	Myrinet 2k	CCLRC

- all Opteron systems: used PGI compiler: -O3 –fastsse
- Cray XT3: used –small_pages (see Neil Stringfellow's talk on Thursday)
- Altix: used Intel 7.1 or 8.0 (7.1 faster for PCHAN)
- IBM: used xlf 9.1 –O3 –qarch=pwr4 –qtune=pwr4

CCLRC

Acknowledgements

Thanks are due to the following for access to machines …

- Swiss National Supercomputer Centre (CSCS)
  - Neil Stringfellow
- Oak Ridge National Laboratory (ORNL)
  - Pat Worley (Performance Evaluation Research Center)
- Engineering and Physical Sciences Research Council (EPSRC)
  - access to CSAR systems
  - "scarf" Streamline cluster
  - CSE's DisCo programme (including our Cray XD1)
  - CSE's High-End Computing programme

CCLRC

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

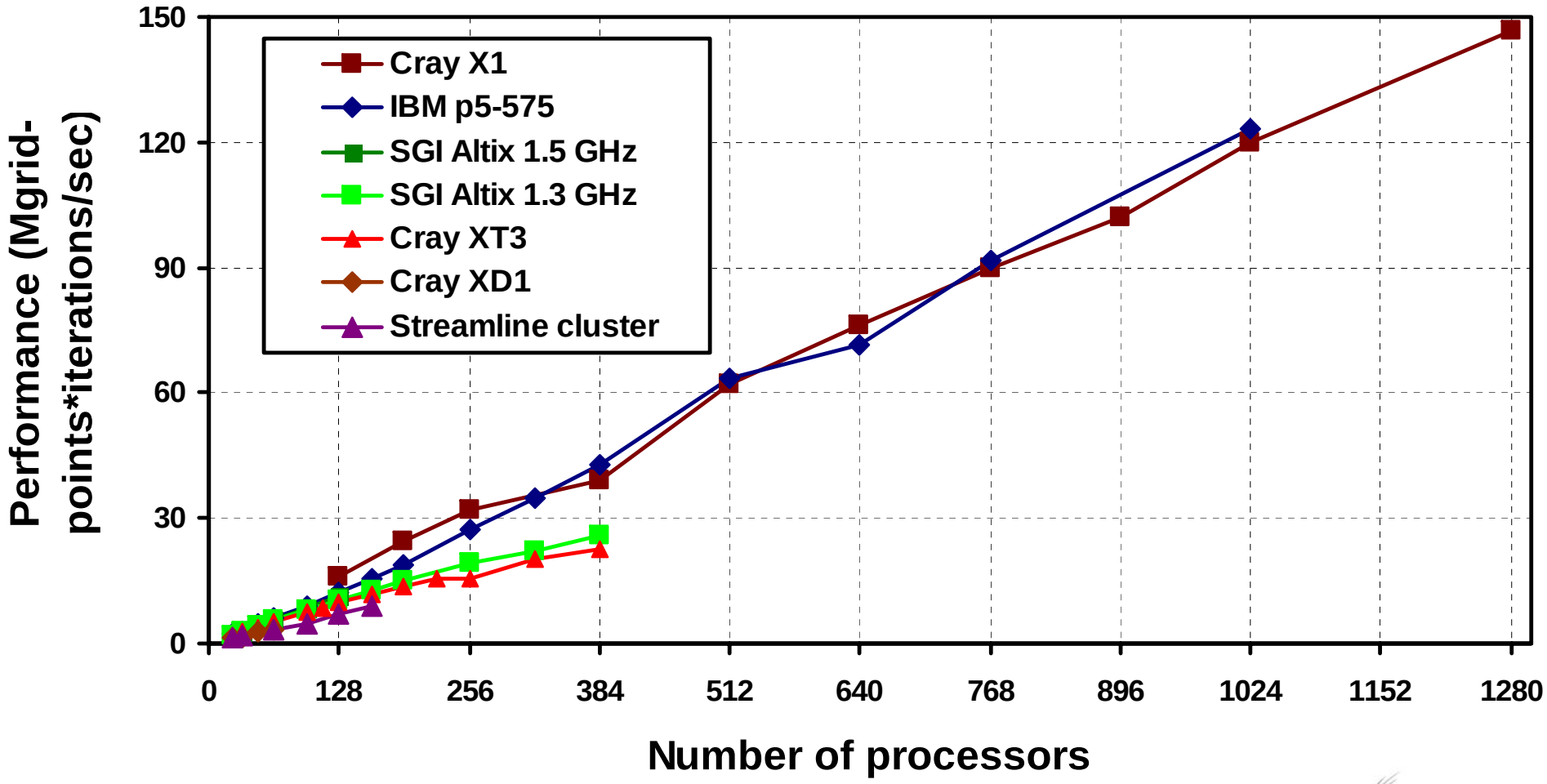
http://www.afm.ses.soton.ac.uk/

CCLRC

PDNS3D Code - characteristics

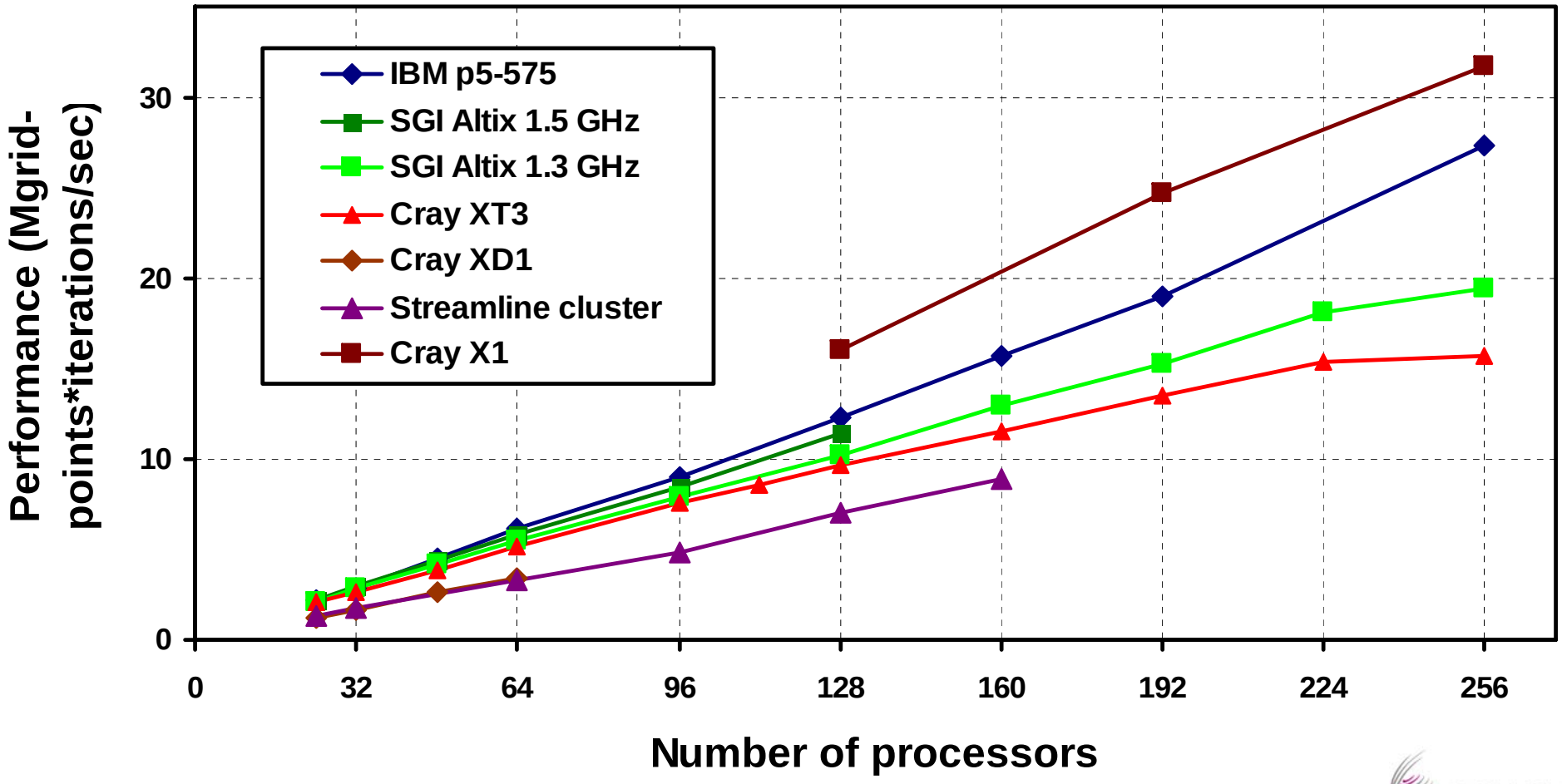
- Structured grid with finite difference formulation
- Communication limited to nearest neighbour halo exchange
- High-order methods lead to high compute/communicate ratio
- Performance profiling shows that single cpu performance is limited by memory accesses – very little cache re-use
- Vectorises well
- VERY heavy on memory accesses

PCHAN T3 (360 x 360 x 360)



CCLRC

PCHAN T3 (360 x 360 x 360)



CCLRC

Proudman Oceanographic Laboratory Coastal Ocean Modelling System (POLCOMS)

Coupled marine ecosystem modelling

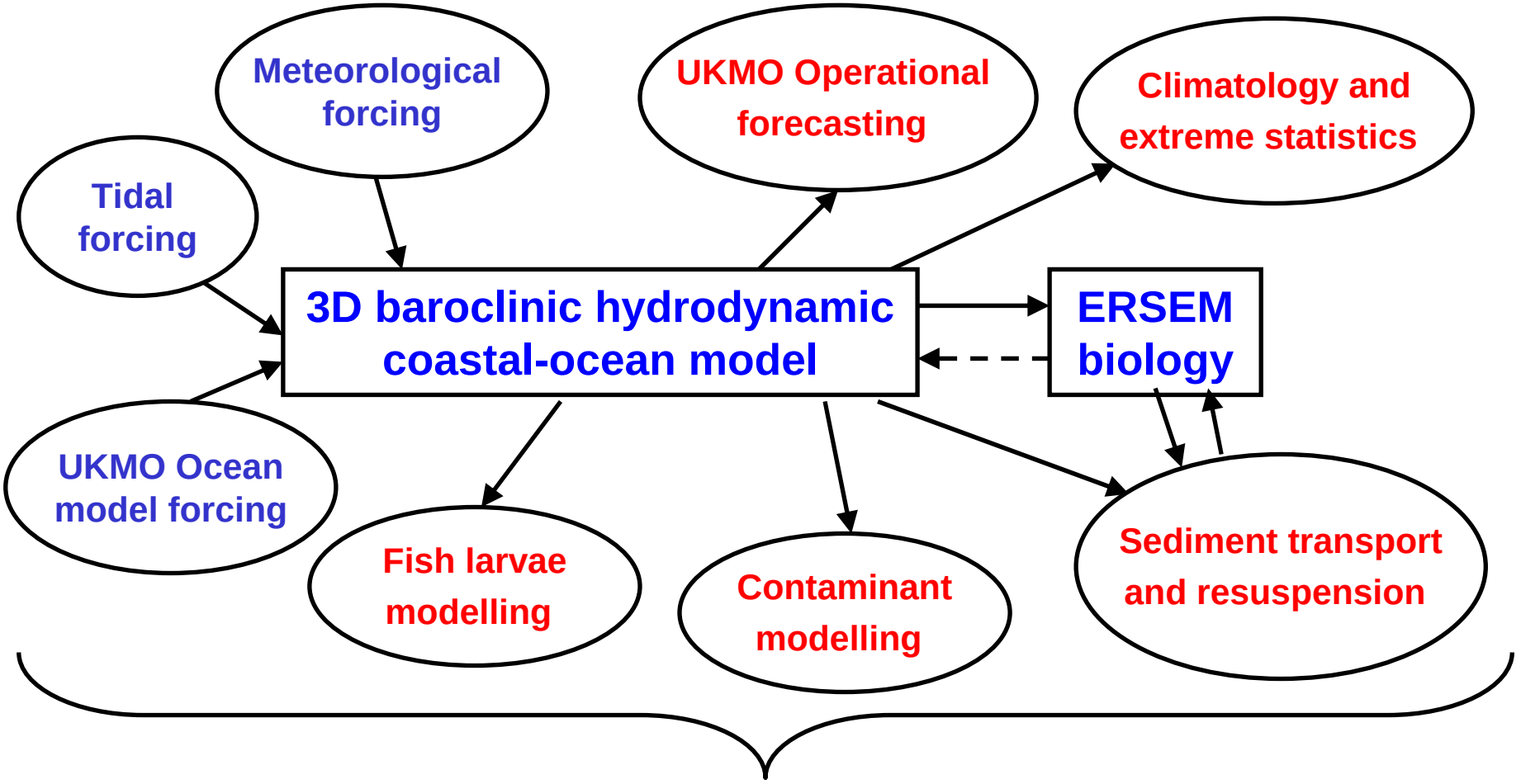
Wave modelling

Sea-ice modelling

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

CCLRC

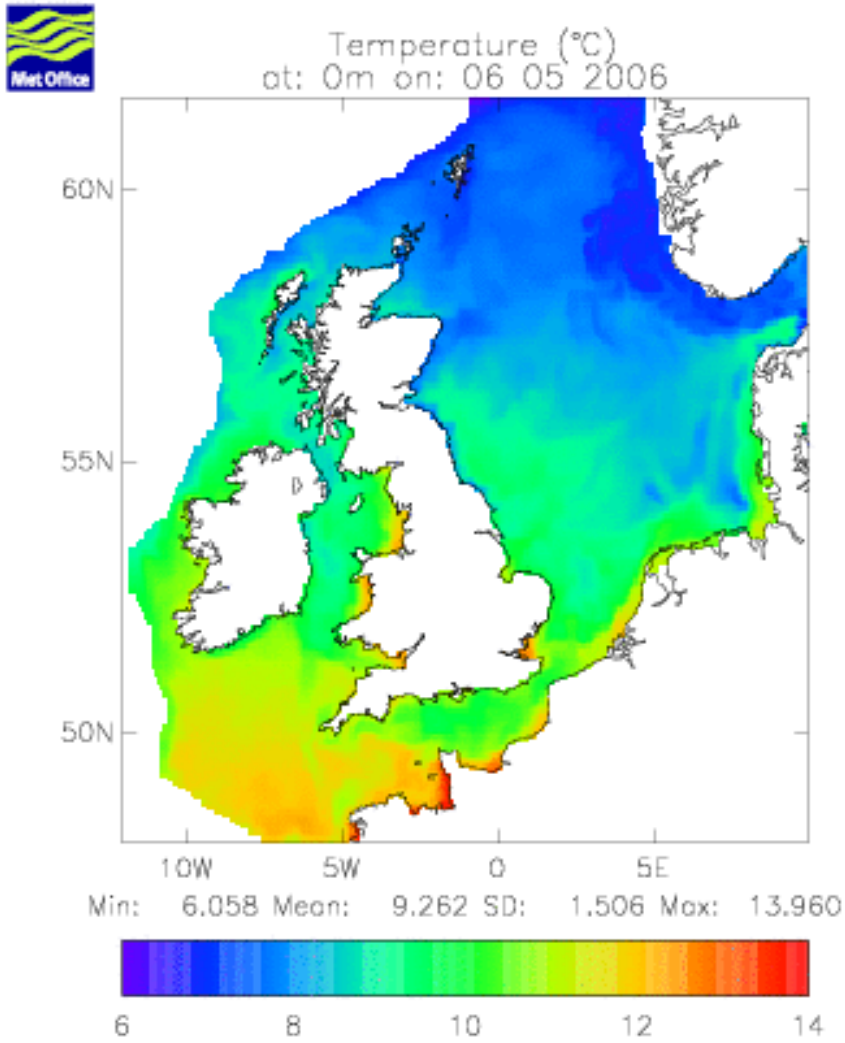
POLCOMS Structure



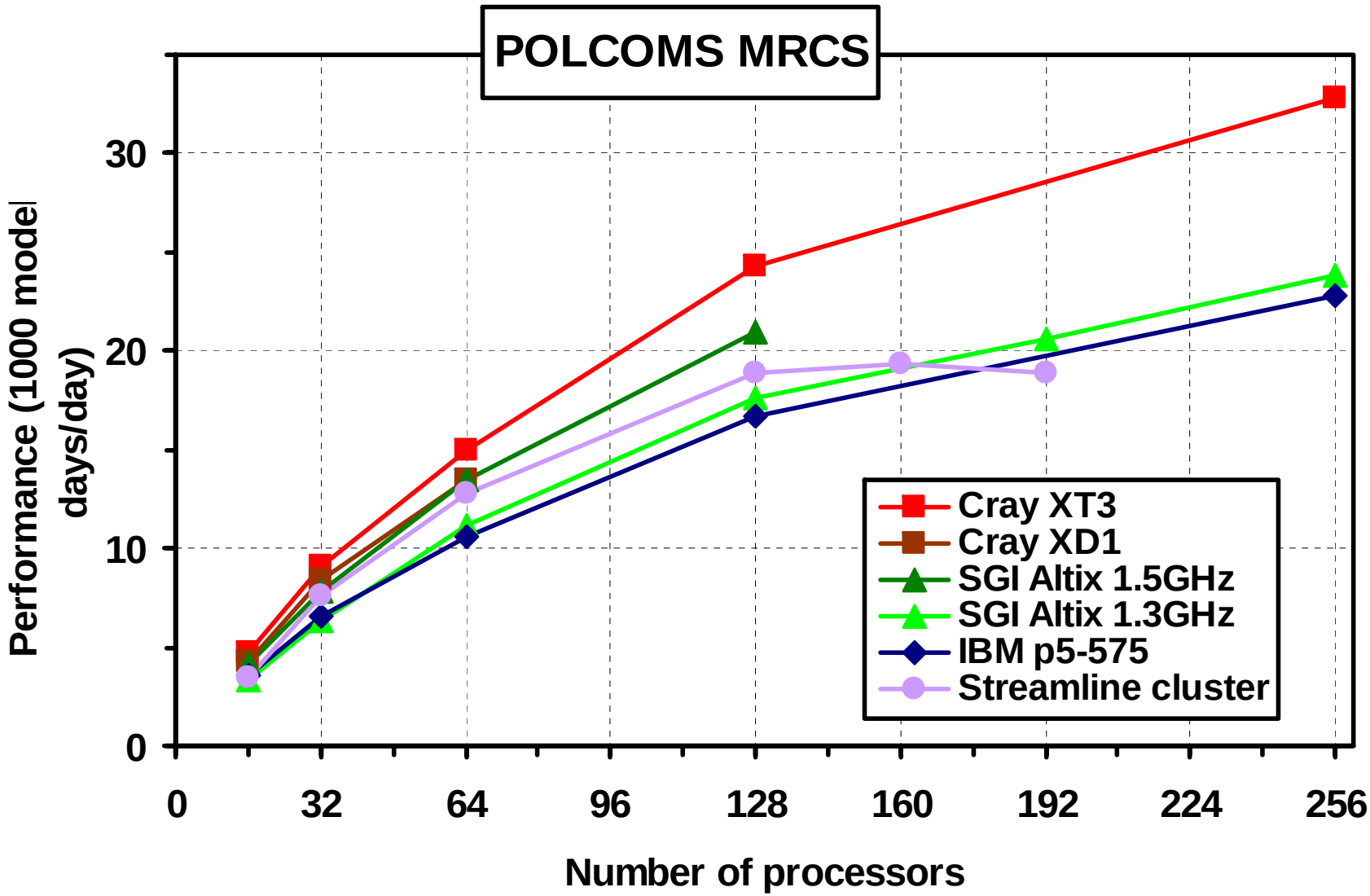
- 3D Shallow Water equations
  - Horizontal finite difference discretization on an Arakawa B-grid
  - Depth following sigma coordinate
  - Piecewise Parabolic Method (PPM) for accurate representation of sharp gradients, fronts, thermoclines etc.
  - Implicit method for vertical diffusion
  - Equations split into depth-mean and depth-fluctuating components
  - Prescribed surface elevation and density at open boundaries
  - Four-point wide relaxation zone
  - Meteorological data are used to calculate wind stress and heat flux

- Parallelisation
  - 2D horizontal decomposition with recursive bi-section
  - Nearest neighbour comms but low compute/communicate ratio
  - Compute is heavy on memory access with low cache re-use

- Medium-resolution Continental Shelf model (MRCS):
  - 1/10 degree x 1/15 degree
  - grid size is 251 x 206 x 20
  - used as an operational forecast model by the UK Met Office
  - image shows surface temperature for Saturday 6th May 2006



CCLRC


POLCOMS MRCS
Performance (1000 model days/day)
0
10
20
30
0
32
64
96
128
160
192
224
256
Number of processors
Cray XT3
Cray XD1
SGI Altix 1.5GHz
SGI Altix 1.3GHz
IBM p5-575
Streamline cluster

CCLRC

The Cray XD1 as a mid-range computing resource

CCLRC

The *Distributed Computing Group (DisCo)* provides informed technical support and strategical guidance through reports and white papers to a variety of customers;

- Science-based Groups within CSED
- The external academic community - SRIF3 assistance to Heriott Watt (Tony Newjem) – benchmarking, 6 site visits etc.
- Research Councils & Funding agencies (EPRSC, … PPARC, NERC, BBSRC)

Examples include

- An overview of *HPC integrators in the UK marketplace*
- The annual *Machine Evaluation Workshop (MEW)* which offers
  i. a two day programme of short talks
  ii. a two day technical exhibition (18 vendors in 2005)
  iii. an opportunity for delegates to access loaned systems and benchmark their own codes via the internet and on site

CCLRC

Impact on "Production" Clusters

e-science Cluster ("scarf") at the Rutherford Laboratory

Initial application benchmarking revealed poor scalability at higher processor counts compared to similar systems at DL

- Traced to excessive node polling by LSF
- 2 X performance improvement on 32 CPU parallel jobs

EM64T/Infiniband Cluster ("lisa") at SARA, Amsterdam

Application benchmarking revealed unrealistically poor performance of the Inifiniband interconnect (AlltoAll, Allgather) at higher processor counts

- Traced to inconsistent *libmpich.so* compiled for Intel compilers
- 4 X performance improvement on 32 CPU CPMD parallel jobs

Hardware and Software Evaluation:

CPUs

IA32, x86, x86-64 and IA64 systems -
AMD Opteron 850 (2.4 GHz), 852 (2.6 GHz) …
& dual-core 270 (2.0 GHz), 875 (2.2 GHz), 280 (2.4 GHz)

Itanium2 (Intel Tiger 1.5 GHz; HP systems, 1.3, 1.5 and 1.6 GHz; Bull & SGI Altix 3700 - 1.5 (6MB L3) & 1.6 GHz (9MB L3), Montecito dual-core (Bull, HP SD64000)

Networks

Gigabit Ethernet options …. ***http://www.cse.clrc.ac.uk/disco/index.shtml***

SCI, Infiniband and Myrinet (numerous clusters from: Tier-1 and integrators - OCF, Streamline, ClusterVision & Workstations UK), Quadrics, Infinipath ..

System Software

message passing S/W (LAM MPI, LAM MPI-VIA, MPICH, VMI, SCAMPI), libraries (ATLAS, NASA, MKL, ACML, ScaLAPACK), compilers (Absoft, PGI, Intel, Pathscale EKO, GNU/g77), tools (GA tools, PNNL)

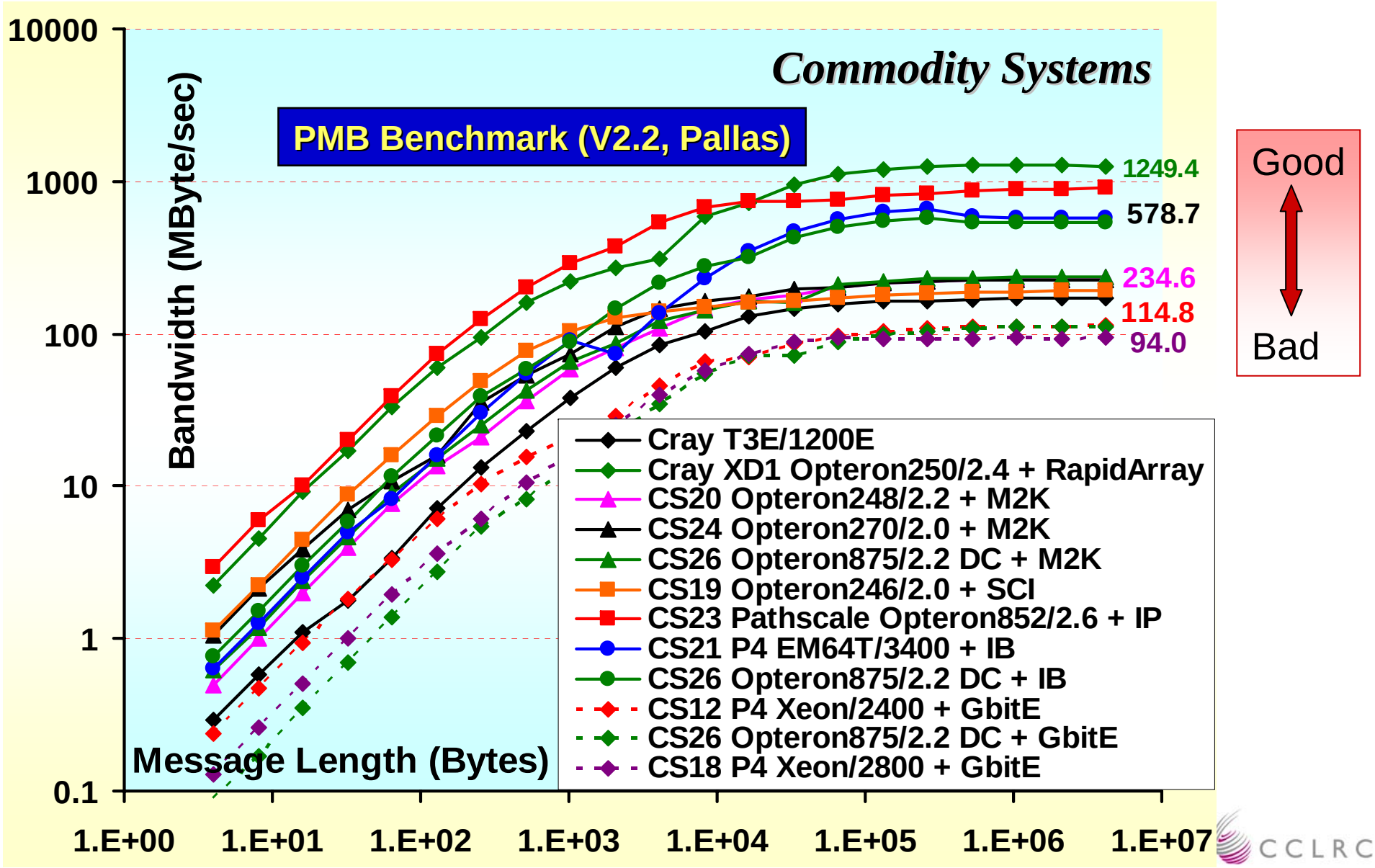
resource management software (PBS, TORQUE, GridEngine, **LSF** etc.)

Commodity Systems (CSx)

System	Location	CPUs	Configuration
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 + IB, Gbit & SCI ("scaliwag")
CS20	*RAL*	256	dual-Opteron/248 2.2 GHz nodes + M2k ("scarf")
CS21	*SARA*	256	Pentium4 Xeon/3400 EM64T + IB ("lisa")
CS22	*TACC*	256	Pentium4 Xeon/3200 EM64T +M2k ("wrangler")
CS23	Pathscale	32	dual-Opteron/852 2.6 GHz nodes + Infinipath
CS24	Leeds	256	dual-core Opteron/270 2.0 GHz nodes + GbitE ("everest")
CS25	Loughborough	64	dual-core Opteron/270 2.0 GHz nodes + M2k ("praesepe")
CS26	HP/Grenoble	64	dual-core Opteron/875 2.2 GHz nodes + M2k, IB and GBitE (HL DL585's)
CS28	AMD	512	dual-core Opteron/275 2.2 GHz nodes + Infinipath
CS29	HP/Grenoble	256	dual-core Opteron/280 2.4 GHz nodes + Infiniband (HL DL585g2's)

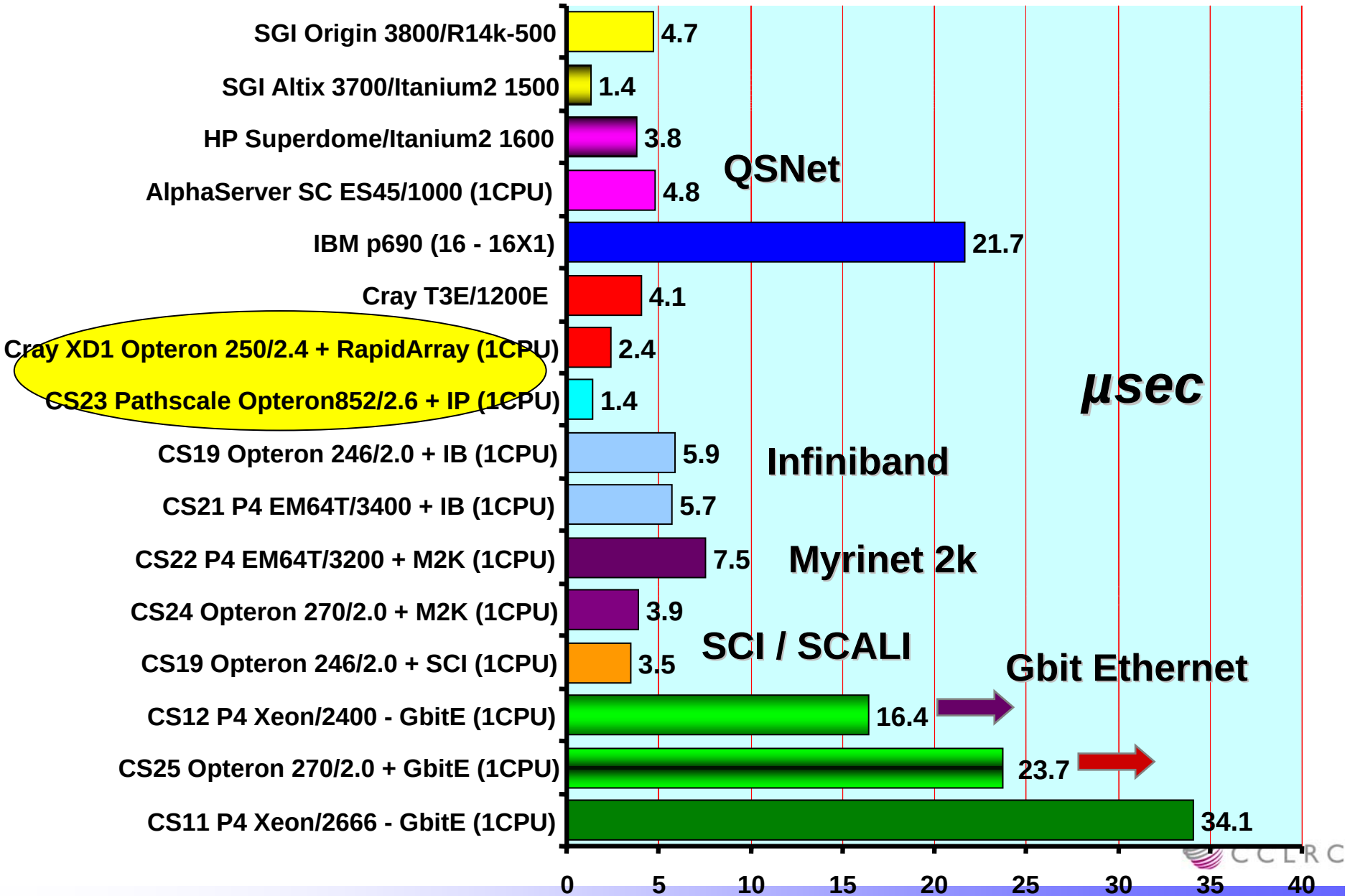
CCLRC

EFF_BW PingPong - Bandwidth

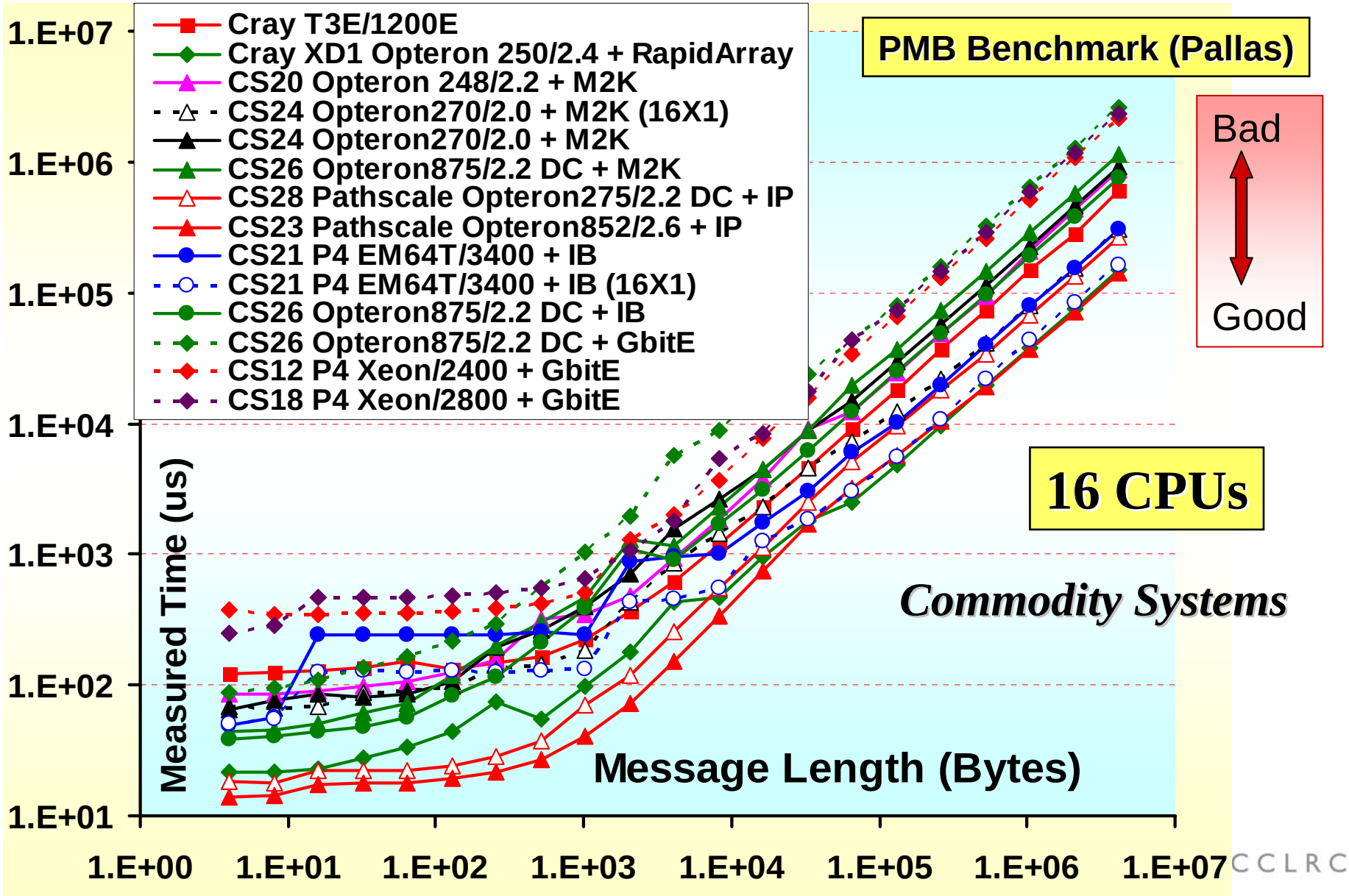


CCLRC

EFF_BW PingPong - Latency

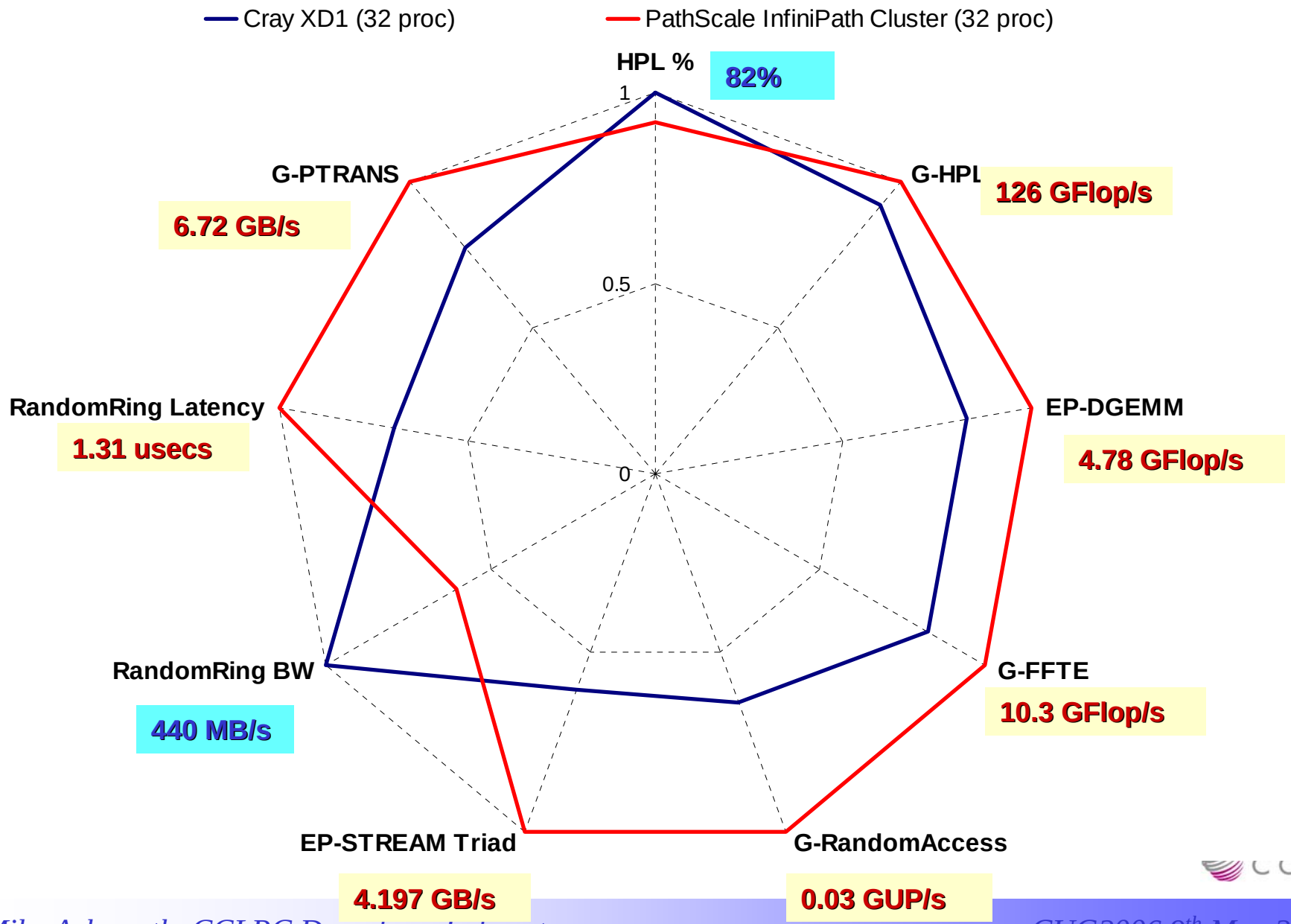


MPI Collectives - Alltoall

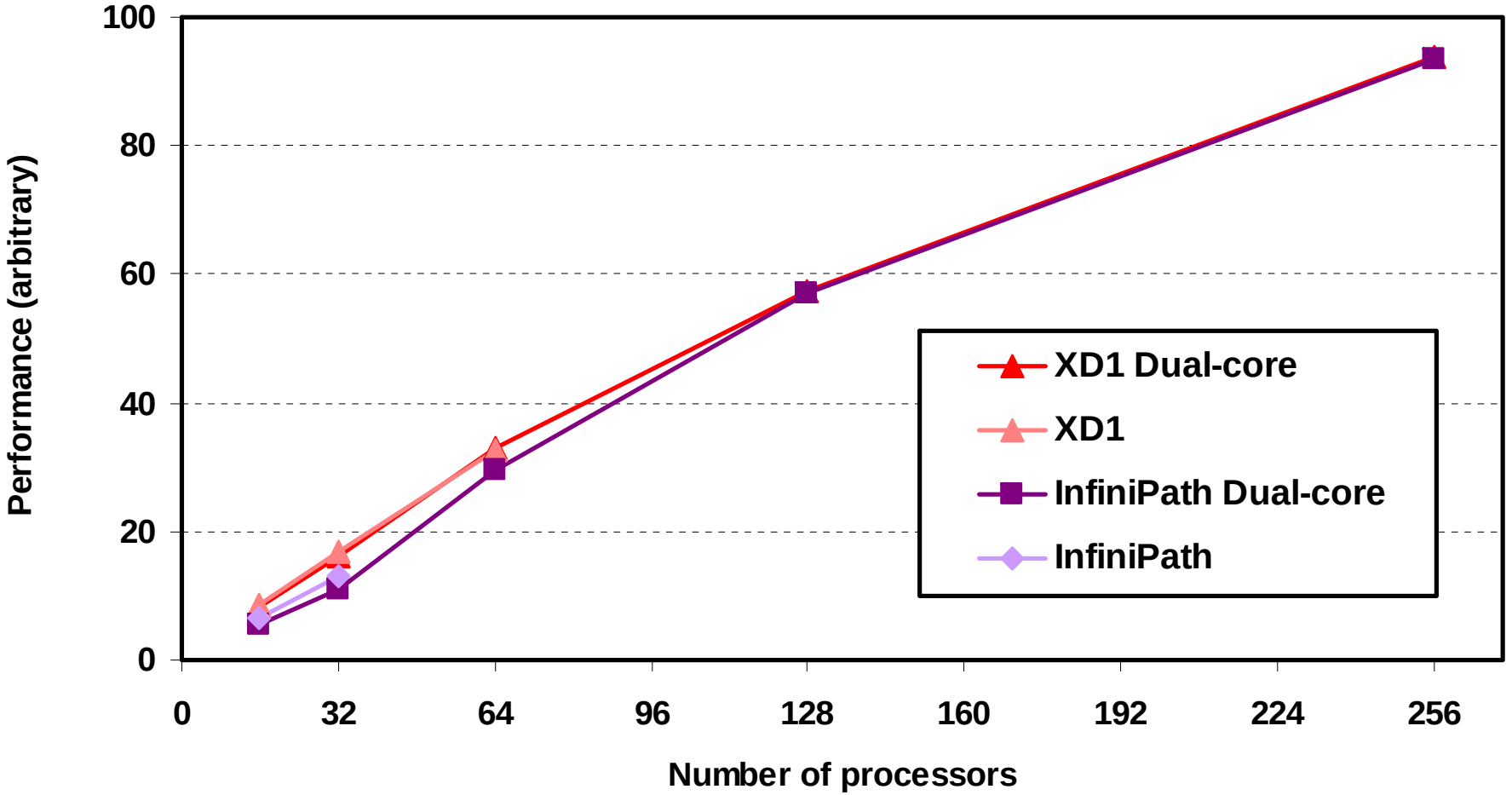


CCLRC

HPCC - XD1 vs. InfiniPath

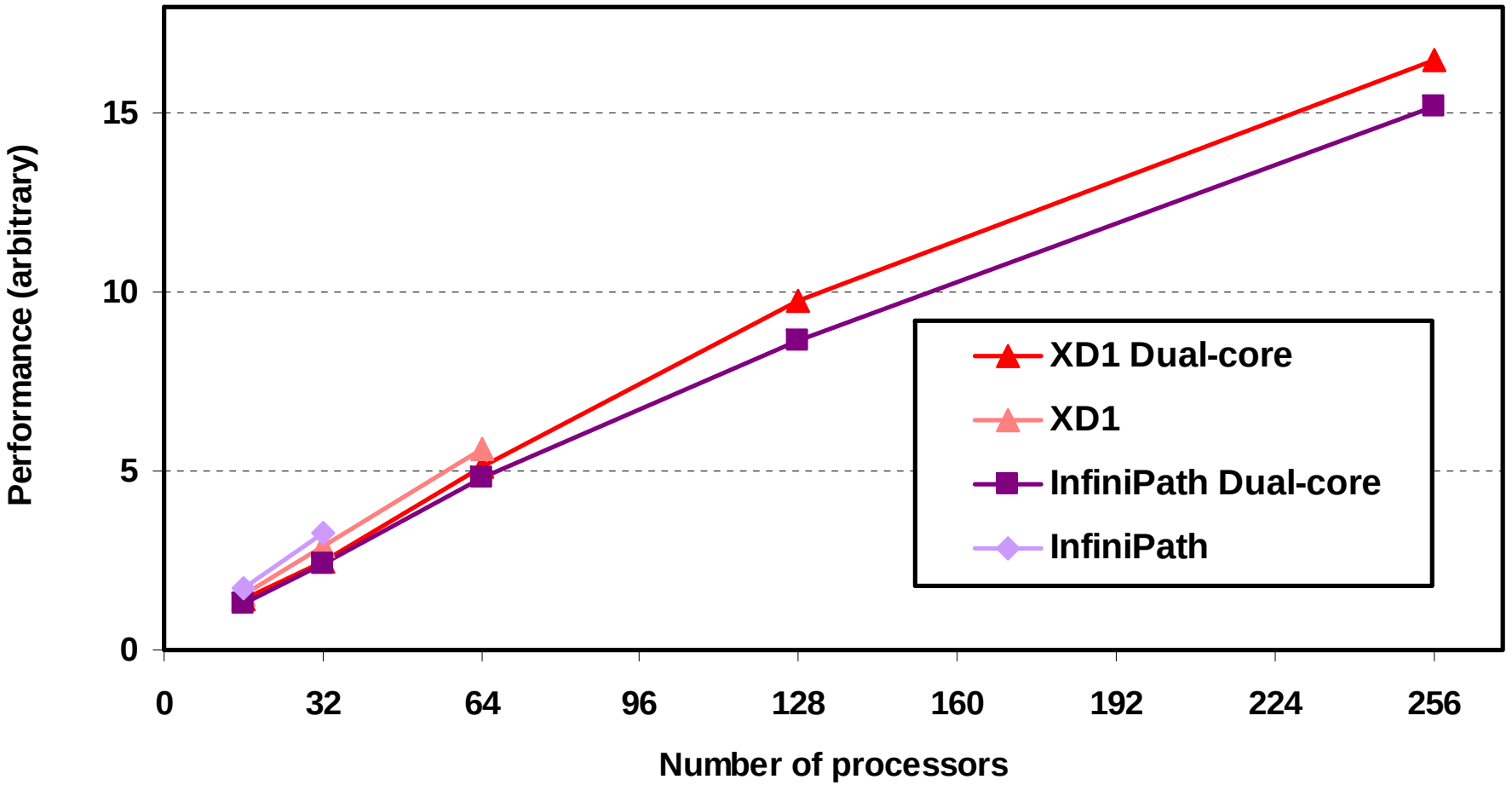


DLPOLY3 NaCl benchmark 216,000 ions



CCLRC

DLPOLY3 Gramicidin benchmark 792,960 atoms



CCLRC

Cray XD1 experiences at CCLRC

CCLRC

1. Poor choice of reselling partner – Compusys. Why could we not have the option of other resellers, or a direct sale?
2. Power supplies keep failing – four so far - and we have had to have entirely new chasses due to motherboard failures.
3. Compusys don't/won't keep spares in the UK – they have to be shipped from the US, shipments get held up at Customs, we and our users get tetchy. Up to four week turnaround for some of our power supply problems. Pathetic.
4. Compusys personnel attempting to install the system BEFORE they had been trained at Cray US facilities.
5. But once 'trained' there have still been numerous issues with their performance. We've seen quite a few engineers. Some are no longer with the company.
6. Some aspects of Sys admin difficult – numerous problems in installing software, updates, getting file systems mounted, integrating the RAID in the first place…
7. Very late delivery of FPGAs – Virtex IVs were finally delivered some months after we were led to believe they would be available.

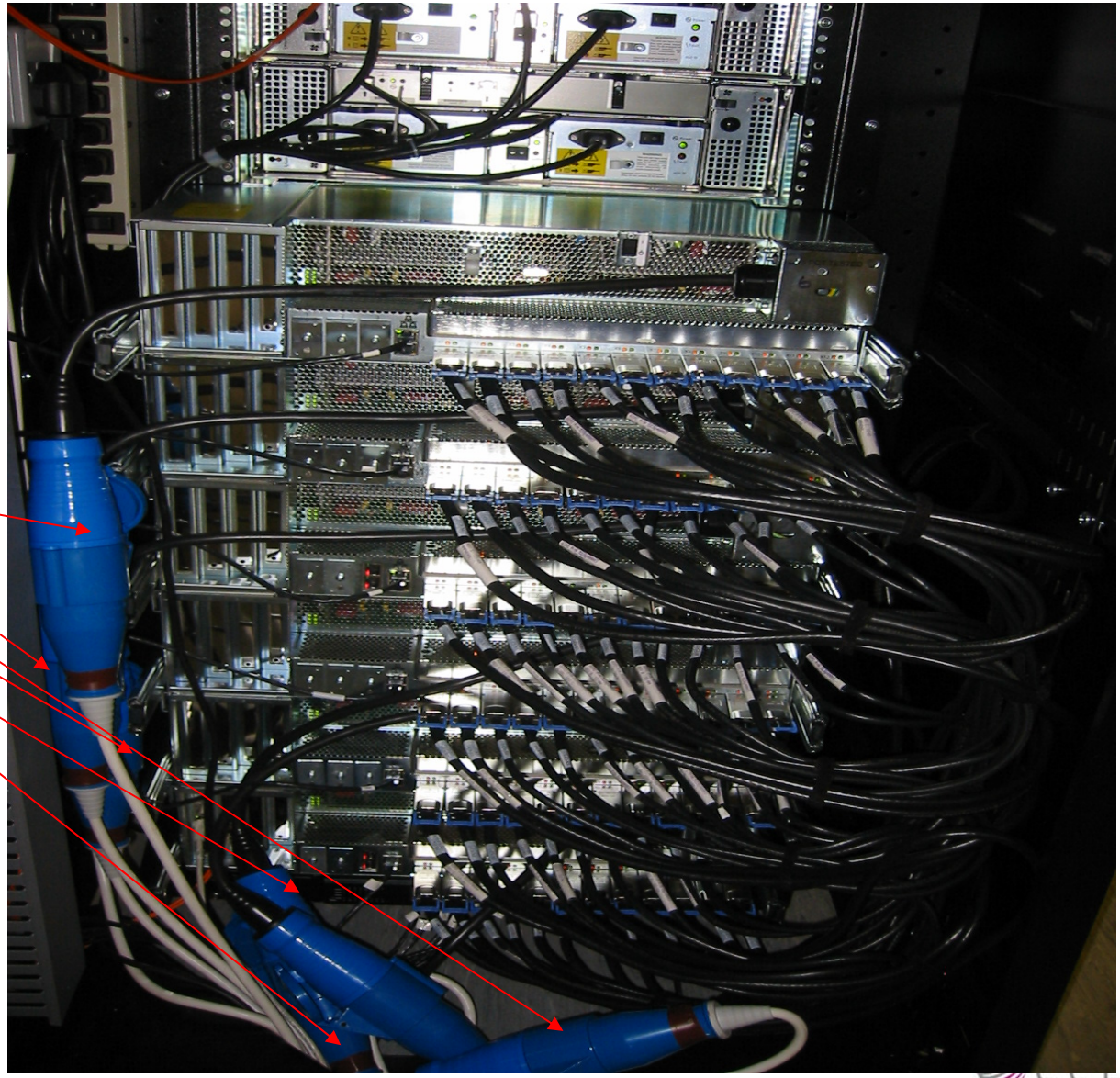
Endurance test cont.

8. The number of power feeds required is a bit excessive: one per chassis, plus power for RAID + switches = 7 X 32 A single-phase commando sockets required. We estimate that the amount of power needed requires only 3 sockets. Due to a finite supply of these things in our machine room, which is full of other kit, we have had to pay for a custom power distribution board to go inside the rack. Cray are aware of the problem but have yet to provide a solution.
9. Performance and stability problems with V 1.1 of the software; upgrading to 1.2 and 1.3 has not been straightforward (an integrated keyboard, monitor & mouse setup wouldn't be a bad idea, instead of having to plug in a laptop with a particular Linux kernel and set of patches in order to talk to the thing).
10. Errors in user/sys admin manuals have been found.
11. Second class citizenship: because we are not Cray customers officially (it's Compusys who 'sold' the system to us) For example, we are not allowed access to the CRInform web site.

All in all not the most positive of experinces, but we would like to thank ***Steve Jordan of Cray UK*** for his efforts and bailing Compusys out on numerous occasions. Without him we would probably still not have a working system.

CCLRC

socket to ‘em



A few questions

1. Are Cray committed to taking this product line forward; will there be a "XD2"?
2. What parts, if any, of the XD1 architecture will end up in future systems, e.g. the adaptive supercomputing initiative?
3. Having (once again) dipped a toe in the mid-range/departmental HPC waters are Cray now going to concentrate solely on the high-end and leave this market to the commodity resellers/integrators?
   (That's the impression we are getting in the UK.)
4. Relative expenditure profiles in the UK –
   1. HECToR (high-end) - £100M (2007-2013)
   2. SRIF (mid-range University systems - £40M (2006-2008)

   Do Cray really want to miss out on the mid-range?
5. Coprocessors: should we put our efforts into FPGAs and/or some alternative (Cell, Clearspeed, …)?

CCLRC

FPGAs

The delivery of the Virtex IVs was VERY late

However through interactions with vendors – Celoxica, Nallatech, Mitrion – and talking to other users, we have the following thoughts and initial impressions:

- VHDL/Verilog: Cray do provide some example VHDL apps and documentation with the XD1 and little else. Pretty low-level and hard for us Fortran types to get to grips with.

- Celoxica's Handel-C: as the name suggests is C-like and a support library for the XD1 is under development. We have had access to an early version and have managed to get a few simple tests run on XD1 (nothing worth reporting in detail). Perhaps not enough abstraction away from the hardware though.

Programming FPGAs cont.

- Mitrionics Mitrion-C: higher level pseudo-C language which we've found more straightforward to program with. Our main issues: (i) pricing, (ii) haven't yet quantified just how much raw performance is sacrificed cf. VHDL/Verilog in return for programmability and ease of development. Any CUG members able to comment for their applications?

XD1 vs. SGI RASC

- SGI's architecture seemingly offers greater IO bandwidth and access to global shared memory. Latest RC100 Blade provides 2 Virtex 4 LX 200 FPGAs and the architecture provides enough bandwidth to keep both FPGAs fed.
- At present the XD1 architecture which can incorporate one Virtex 4 LX 160 FPGA per node is not really able to compete with SGI's solution.
- SGI seemingly have a closer tie in with Mitrionics and seem intent on making the system as programmer friendly as possible.

How do Cray intend to respond?

CCLRC

Conclusions

- At the high-end
  - the XT3 is competitive with systems from IBM and SGI
  - the winning system is application dependent
  - the X1 is highly competitive especially for vectorisable memory-bound codes (though expensive)

- In the mid-range
  - the XD1 performs well but …
  - Infinipath clusters perform equally well at a fraction of the cost
  - a successor to the XD1 is required if Cray are going to (wish to) provide a cost-effective solution in this (highly competitive) market
  - All major processors are now dual-core:
    e.g. Power5, Opteron, Xeon, Itanium2 (Montecito)
    this is almost always a performance disadvantage due to sharing of caches, busses, memory paths etc.

CCLRC


The End