Performance Comparison of Capability Application Benchmarks on the IBM p5-575 and the Cray XT3

Mike Ashworth

Computational Science and Engineering Department
CCLRC Daresbury Laboratory
Warrington
UK

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

CCLRC

Abstract

The UK's flagship scientific computing service is currently provided by an IBM p5-575 cluster comprising 2560 1.5 GHz POWER5 processors operated by the HPCx Consortium. The next generation high-performance resource for UK academic computing has recently been announced. The HECToR project (High-End Computing Technology Resource) will be provided with technology from Cray Inc., the system being the follow-on to the Cray XT3, with a target performance of some 50-100 Tflop/s.

We present initial results of a benchmarking programme to compare the performance of the IBM POWER5 system and the Cray XT3 across a range of capability applications of key importance to UK science. Codes are considered from a diverse section of application domains, including environmental science, molecular simulation and computational engineering.

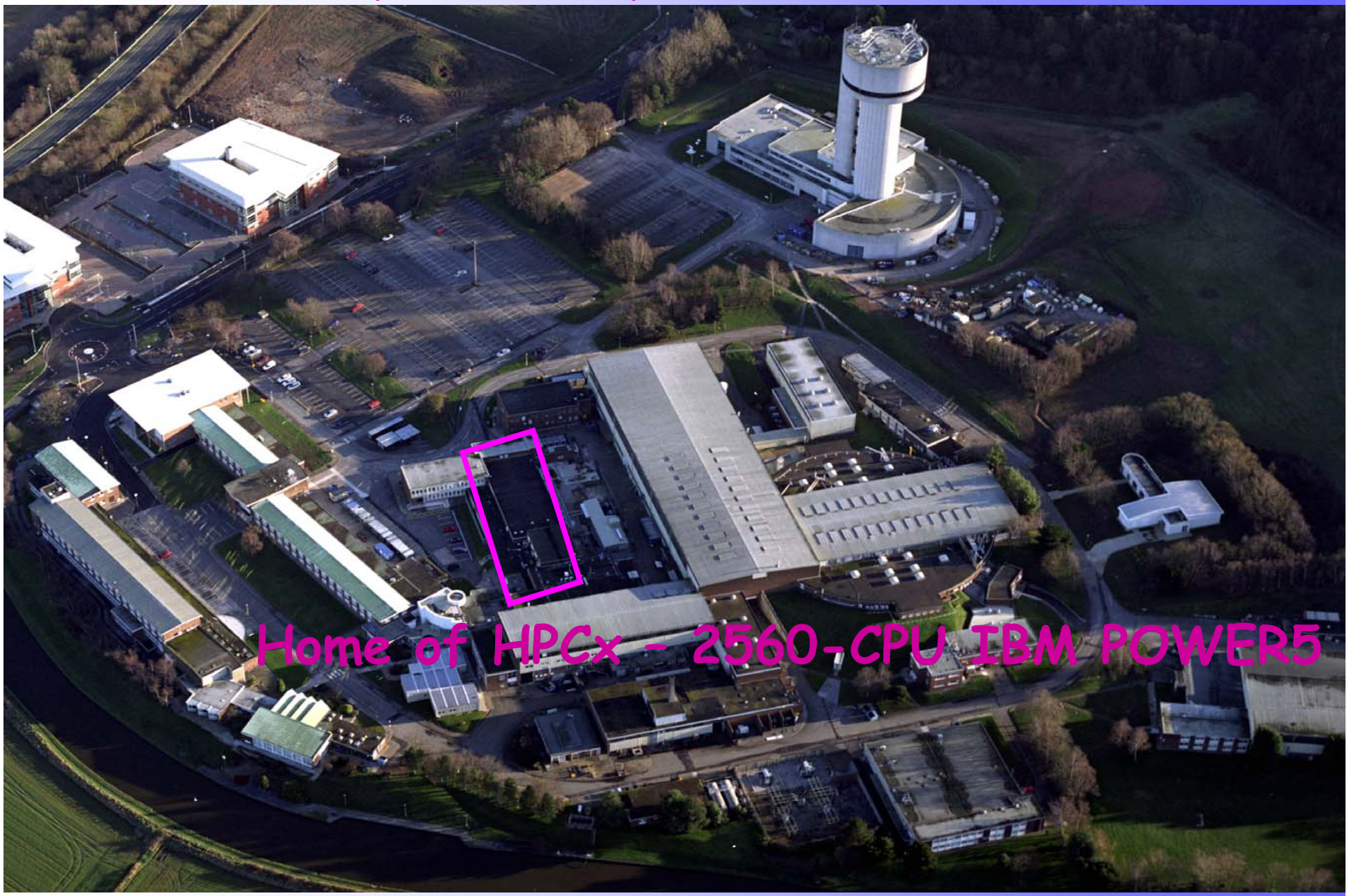
We find a range of performance with some algorithms performing better on one system and some on the other and we use low-level kernel benchmarks and performance analysis tools to examine the underlying causes of the performance.

CCLRC

Outline

- Introduction
  - CCLRC Daresbury Laboratory
  - HPCx
  - A perspective on UK academic HPC provision
  - HECToR – a new resource for UK computational science
- Performance comparison – IBM p5-575 vs. Cray XT3
  - IMB
  - POLCOMS
  - PCHAN
  - DL_POLY3
  - GAMESS-US
  - Single-core vs. dual-core
- Conclusions

Introduction


Home of HPCx – 2560-CPU IBM POWER5

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
  - '02 Phase1: 3 Tflops sustained - 1280 POWER4 cpus + SP Switch
  - '04 Phase2: 6 Tflops sustained - 1600 POWER4+ cpus + HPS
  - '05 Phase2A: Performance-neutral upgrade to 1536 POWER5 cpus
  - '06 Phase3: 12 Tflops sustained – 2560 POWER5 cpus

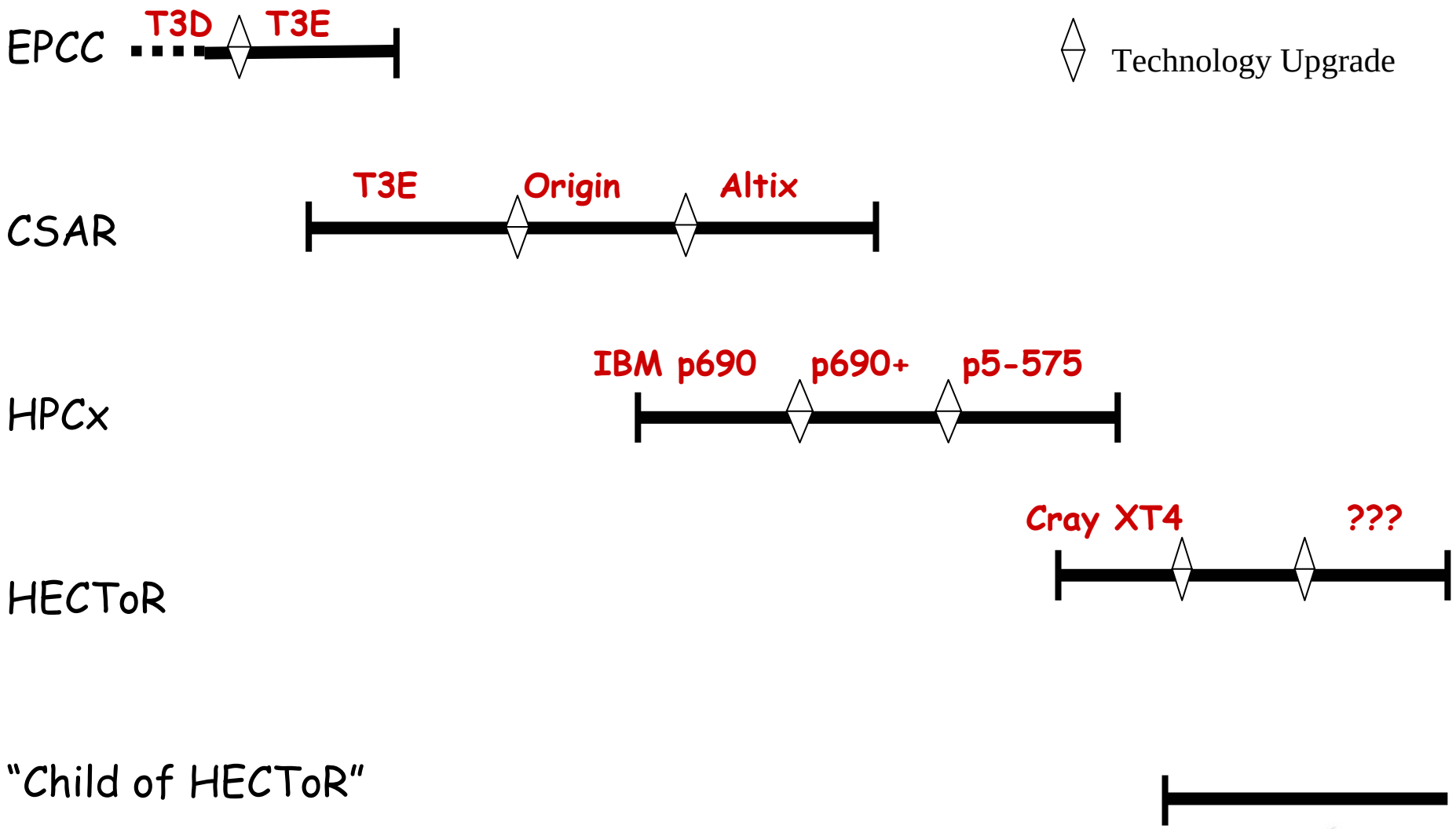
CCLRC

System doubled in size last month

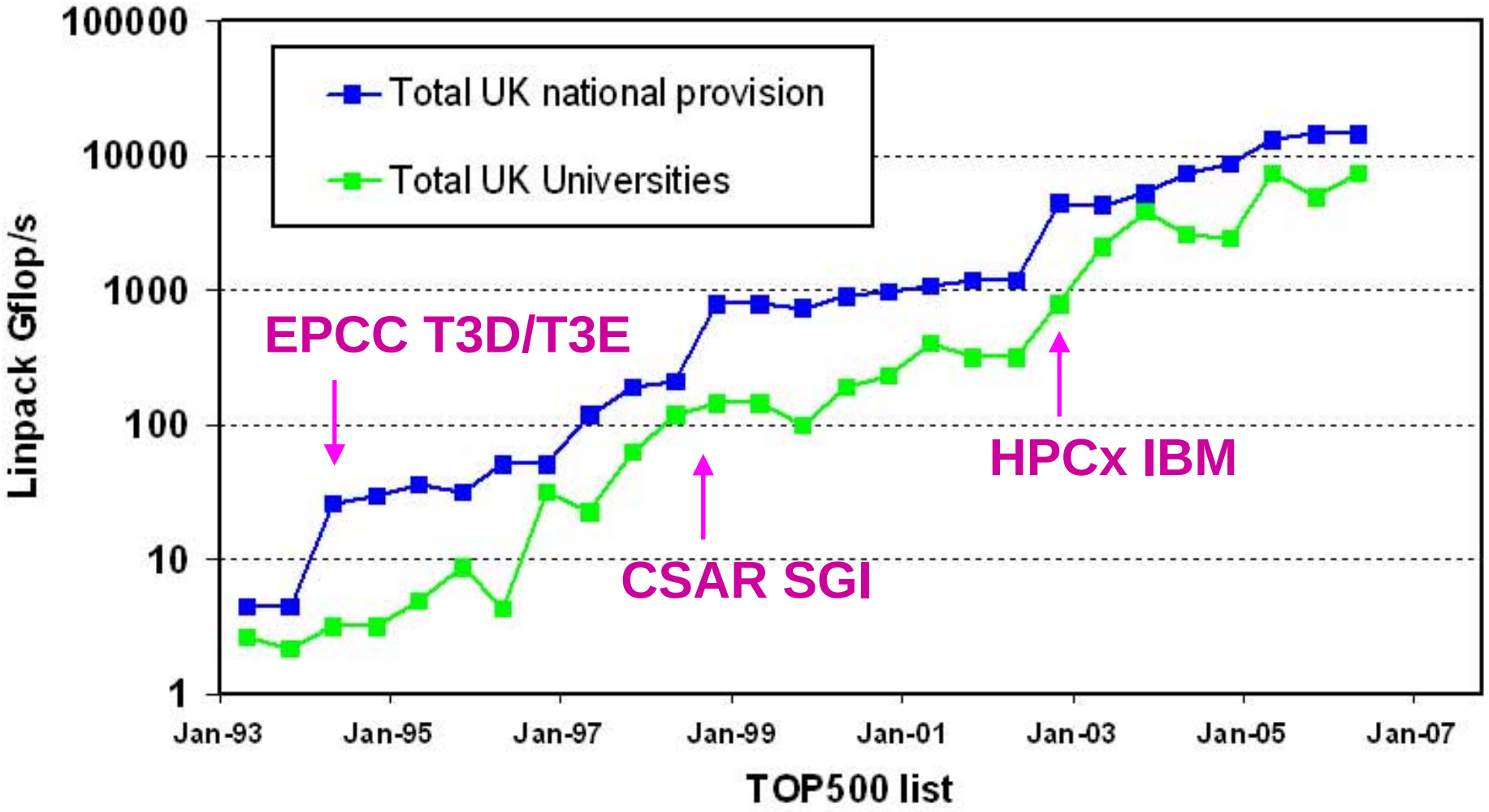
Phase3 – 2560 processors

- 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 (accommodation, management 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

UK Overlapping Services

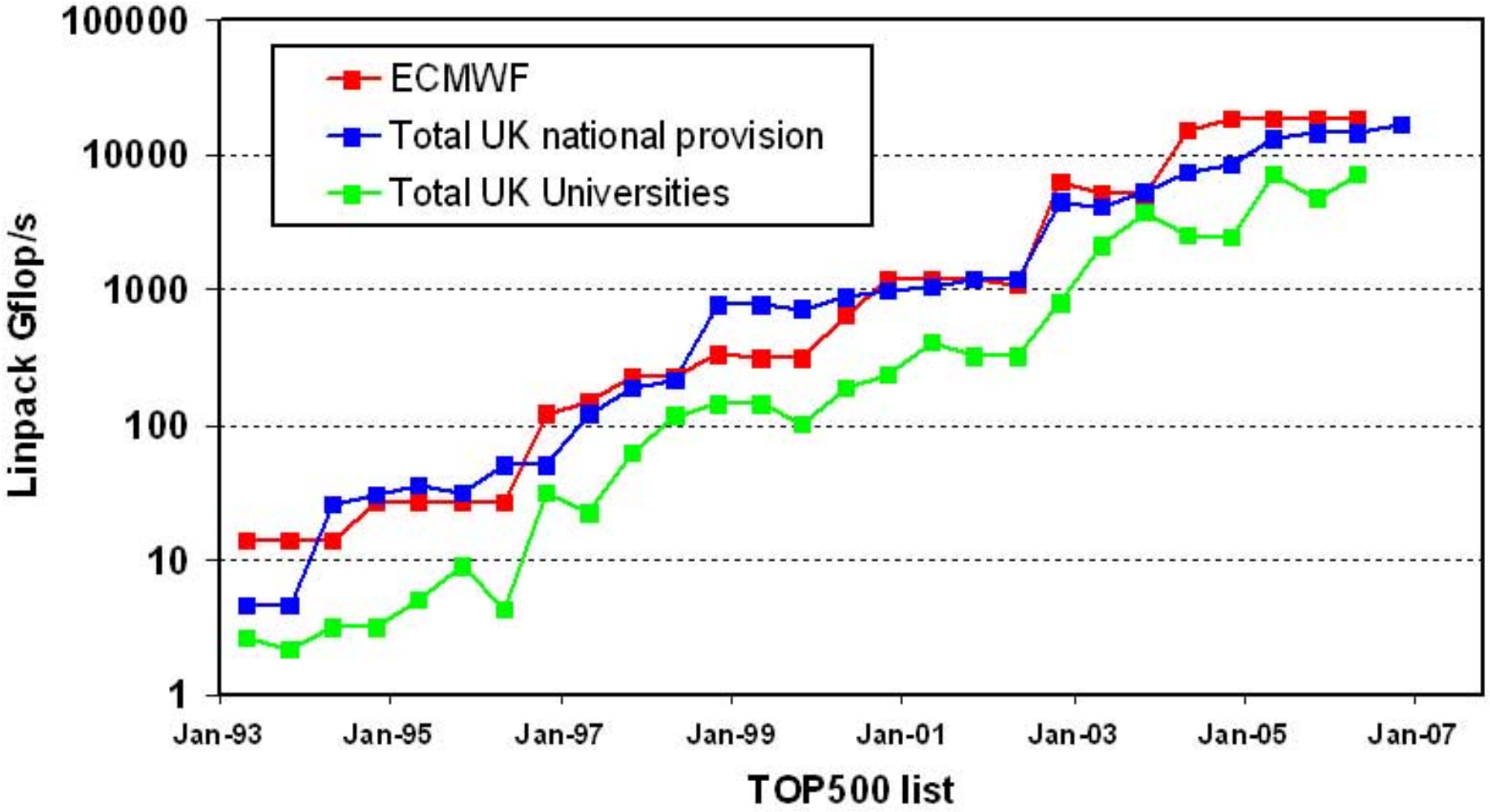


Total academic provision



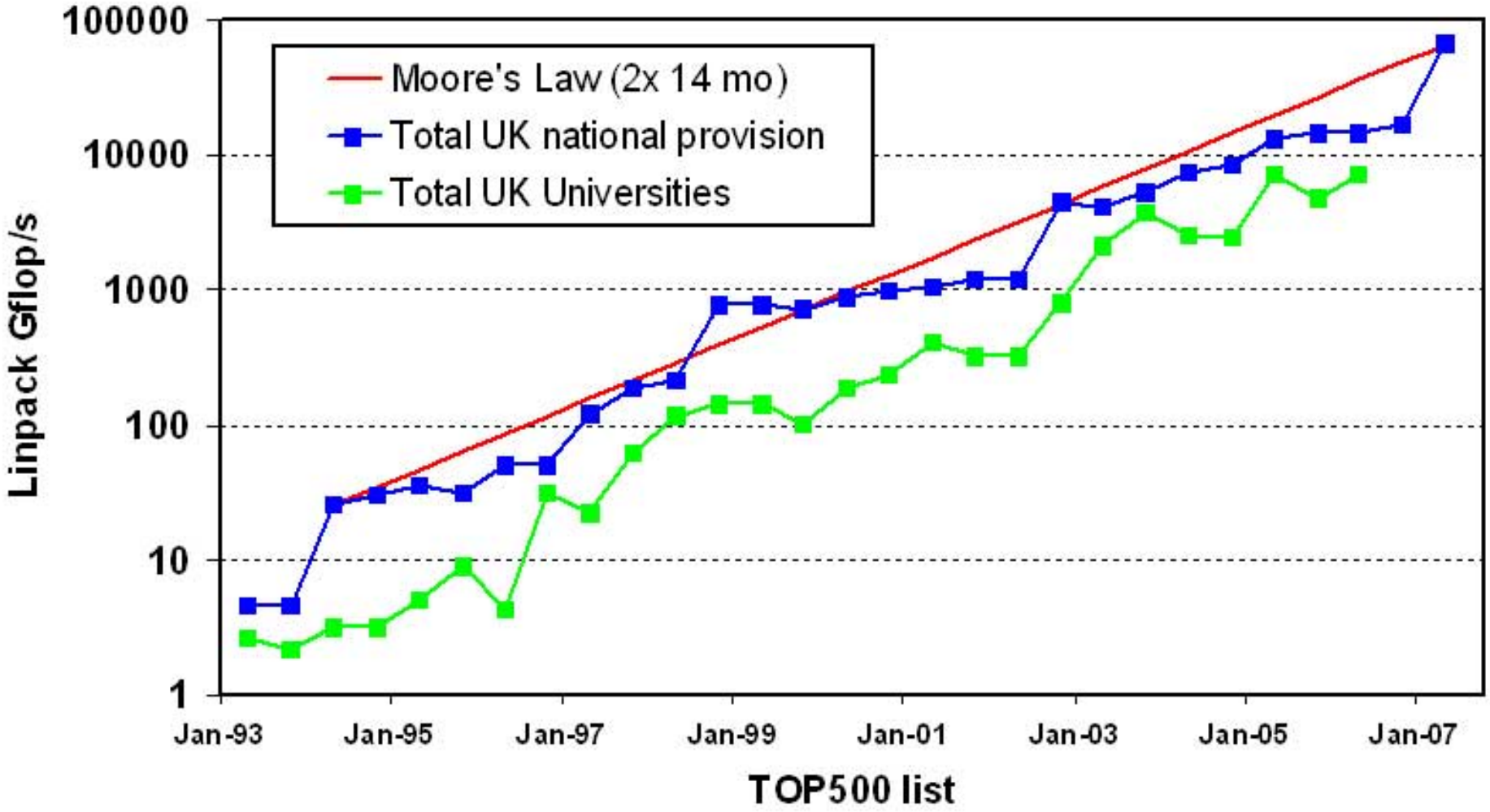
CCLRC

Total academic provision + ECMWF



CCLRC

Total academic provision + HECToR



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 (Cray are "preferred vendor" – XT4)
    - CSE support (NAG core + distributed support)
    - Accommodation & Management (2nd tender issued October 2006)

- "Child of HECToR"
  - Competition for better name!
  - Possible collaboration with UK Met Office
  - Procurement due mid-2007
  - Possible HPC Eur: european scale procurement and accommodation

CCLRC

All this provides motivation for this talk …

… a Comparison of the Cray XT3 and IBM p5-575

(current vs. "future" UK provision)

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	2560	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

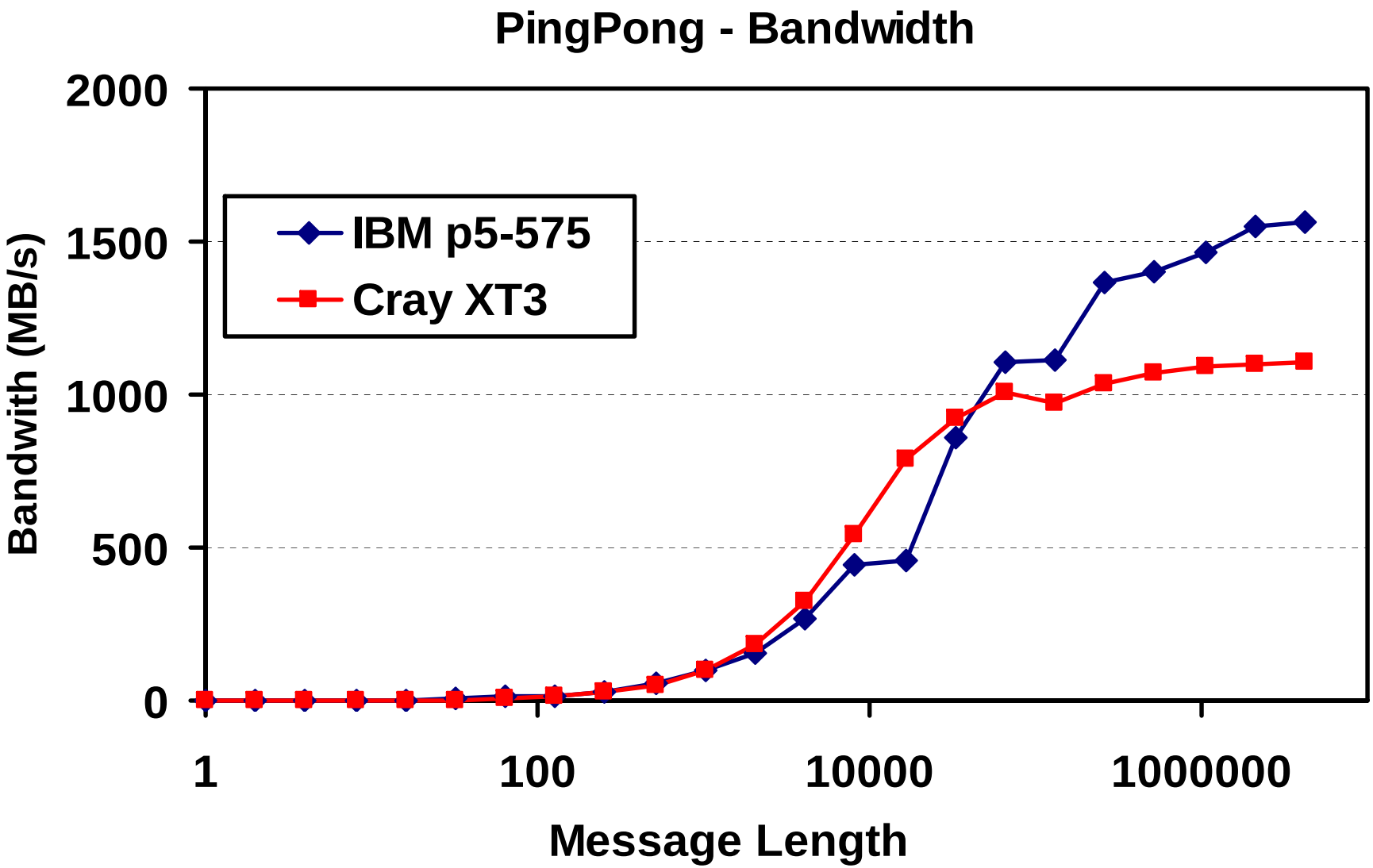
Acknowledgements

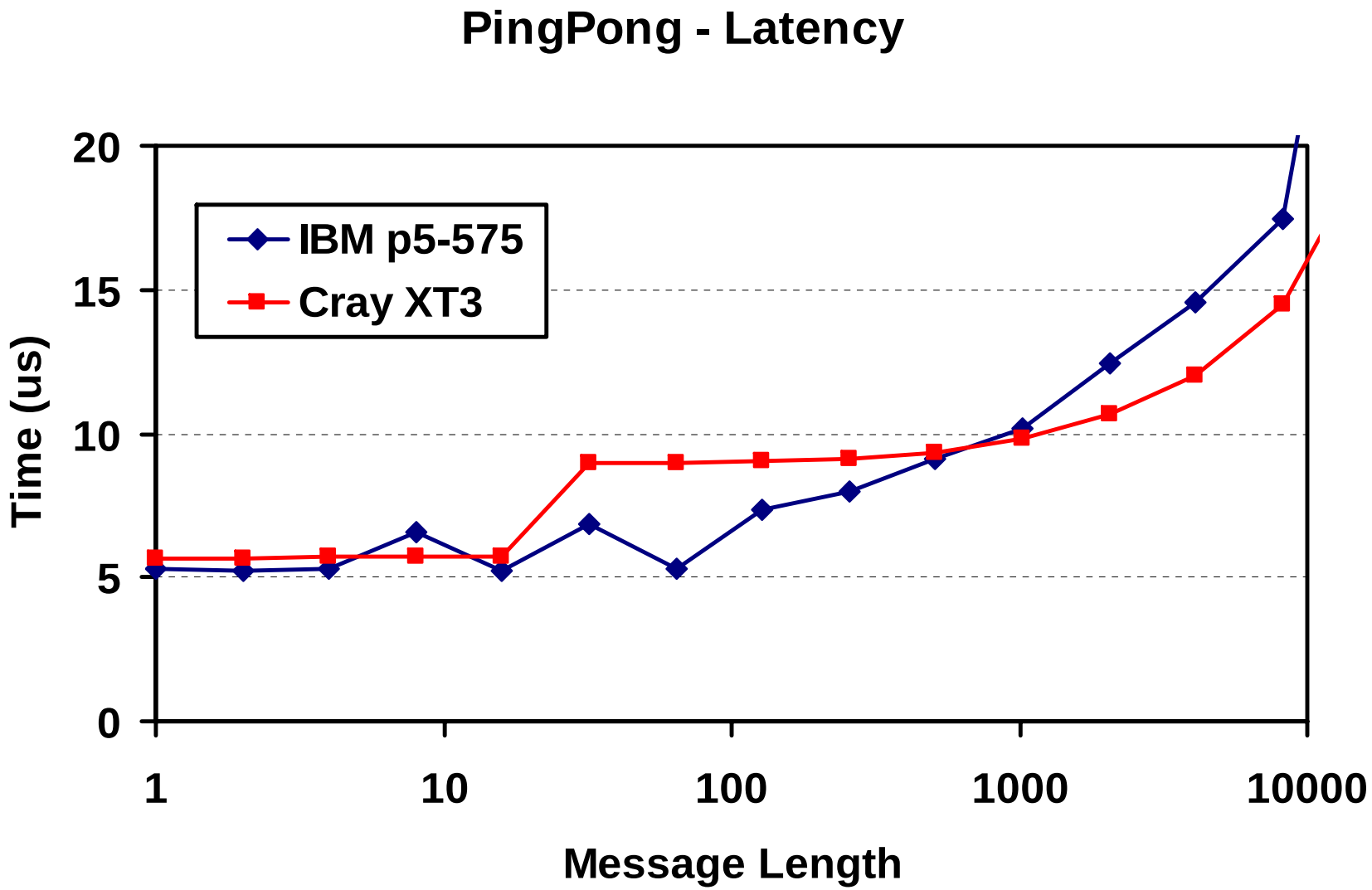
Thanks are due to the following for access to machines …

- 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
  - HPCx time
  - CSE's DisCo programme (Cray XD1)
- Council for the Central Laboratories for the Research Councils (CCLRC)
  - SCARF Streamline cluster

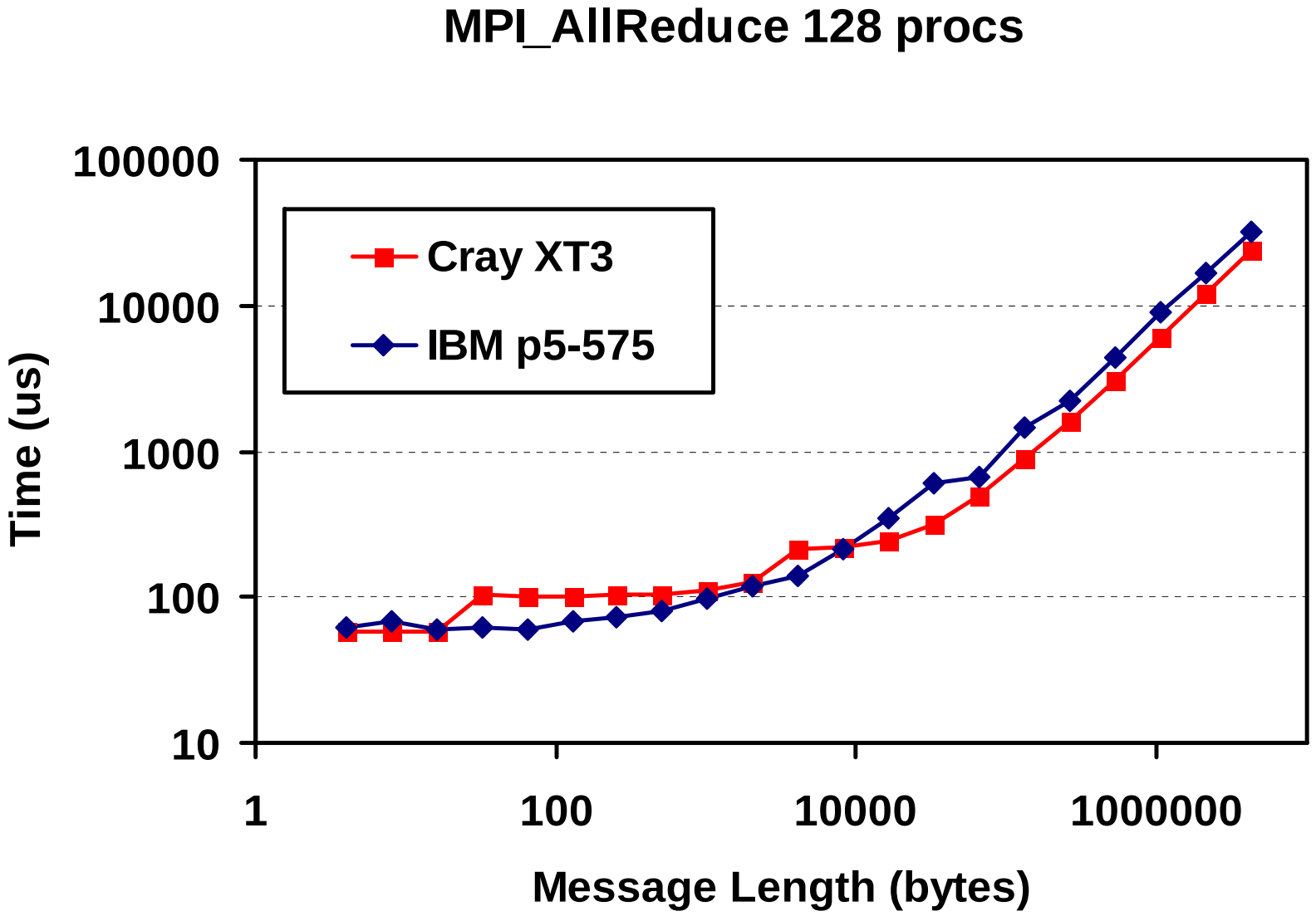
CCLRC

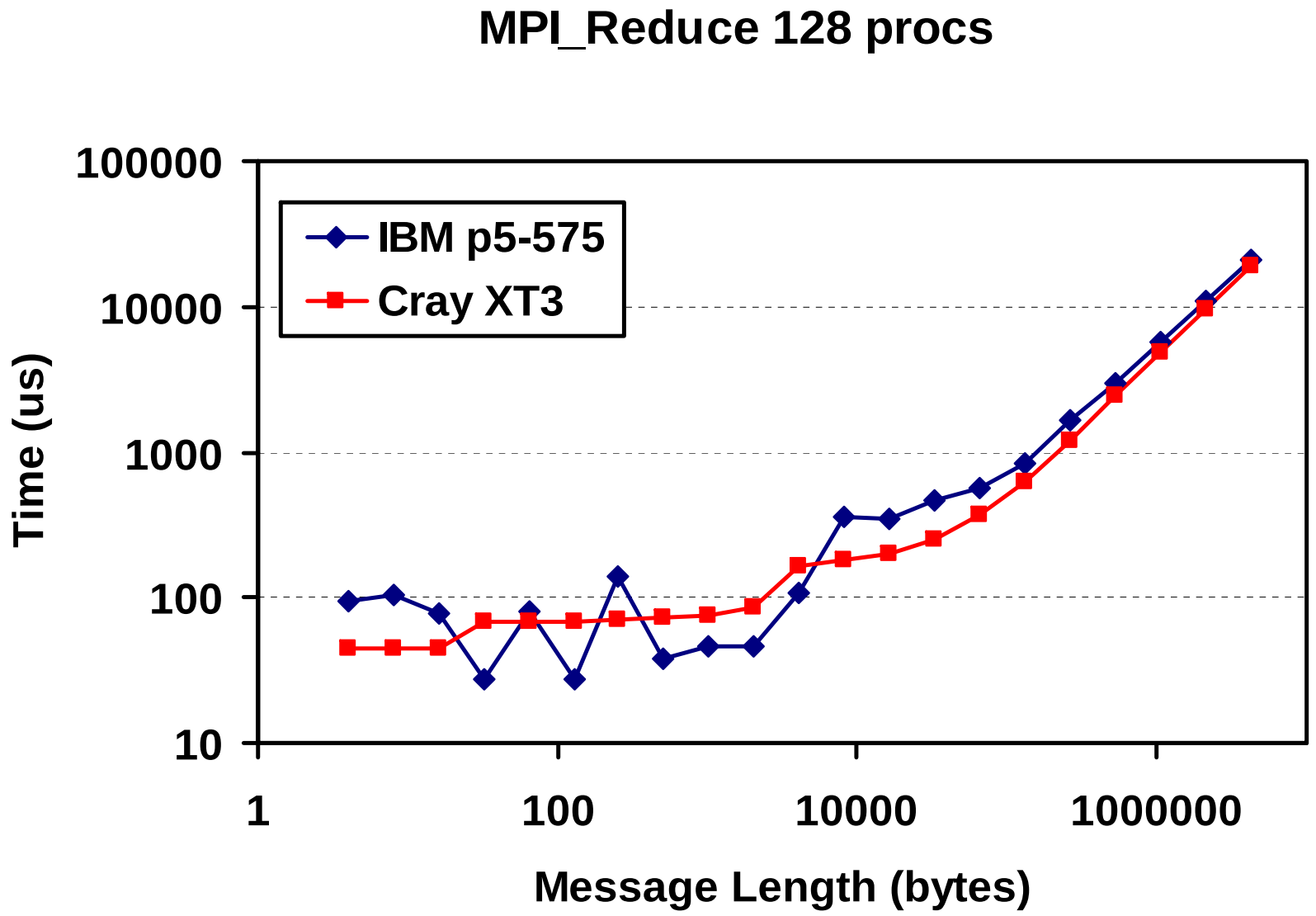
Kernel benchmarks …

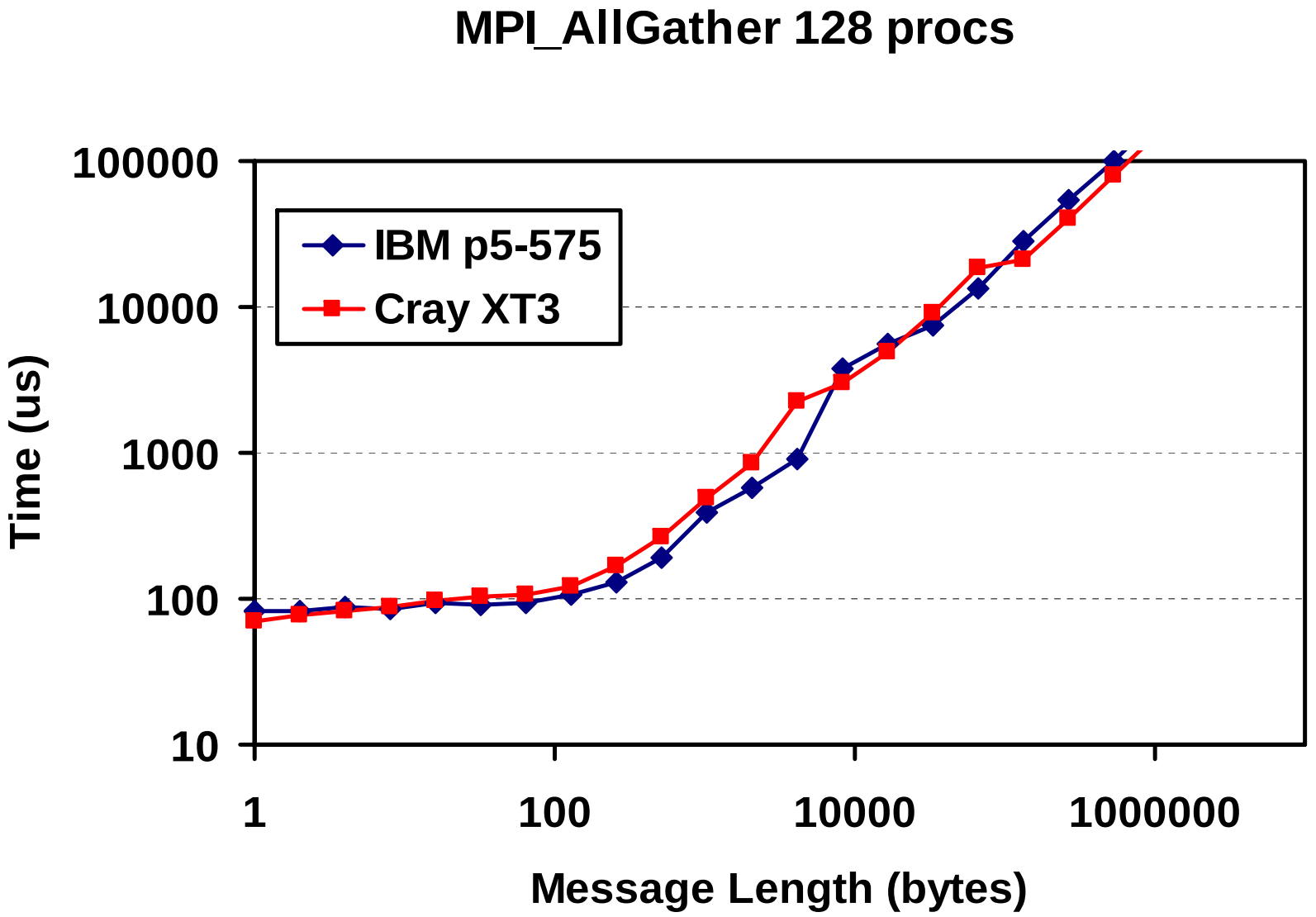

PingPong - Bandwidth
Bandwith (MB/s)
2000
1500
1000
500
0
IBM p5-575
Cray XT3
1
100
10000
1000000
Message Length

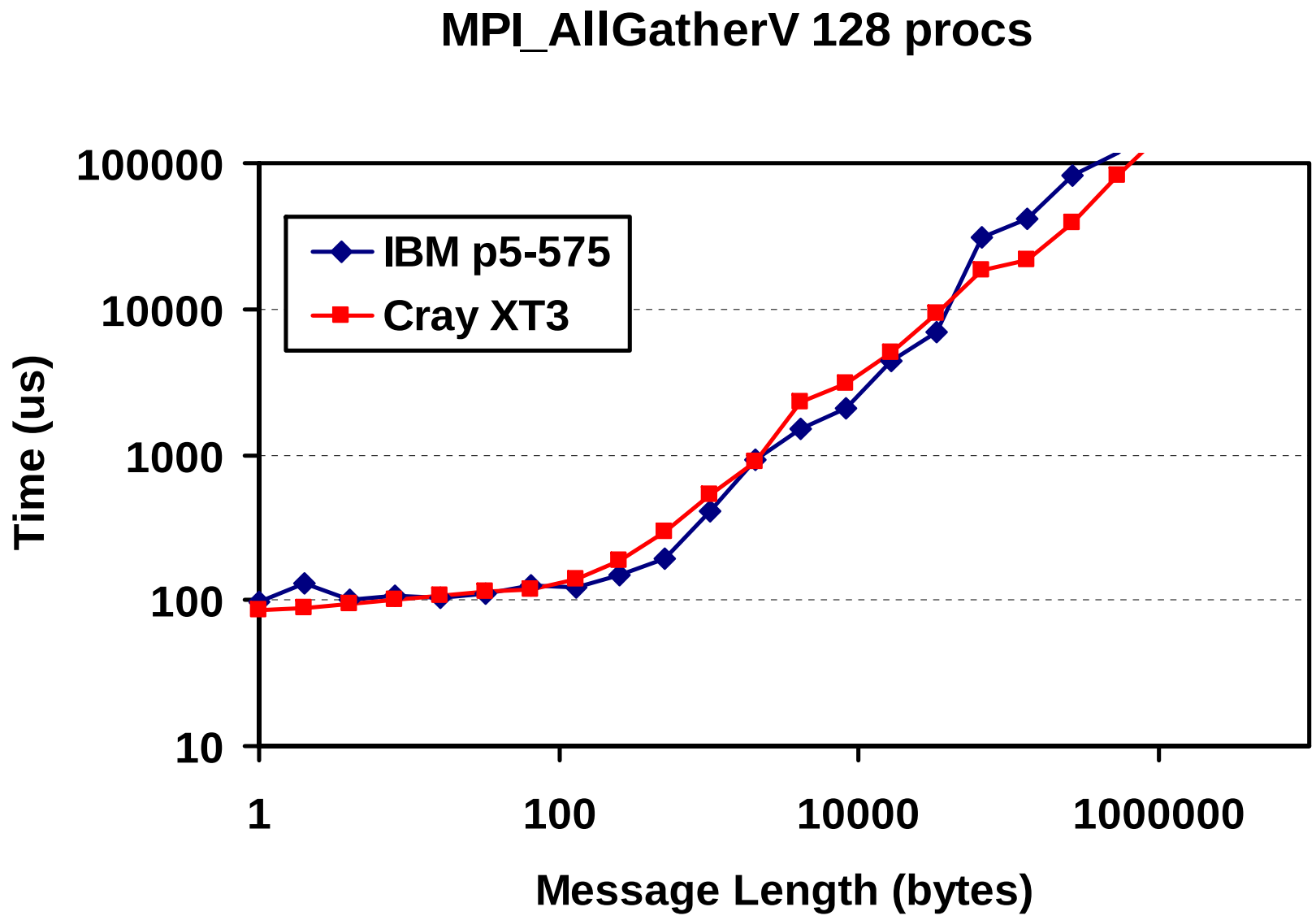

PingPong - Latency
IBM p5-575
Cray XT3
Time (us)
0
5
10
15
20
1
10
100
1000
10000
Message Length

CCLRC

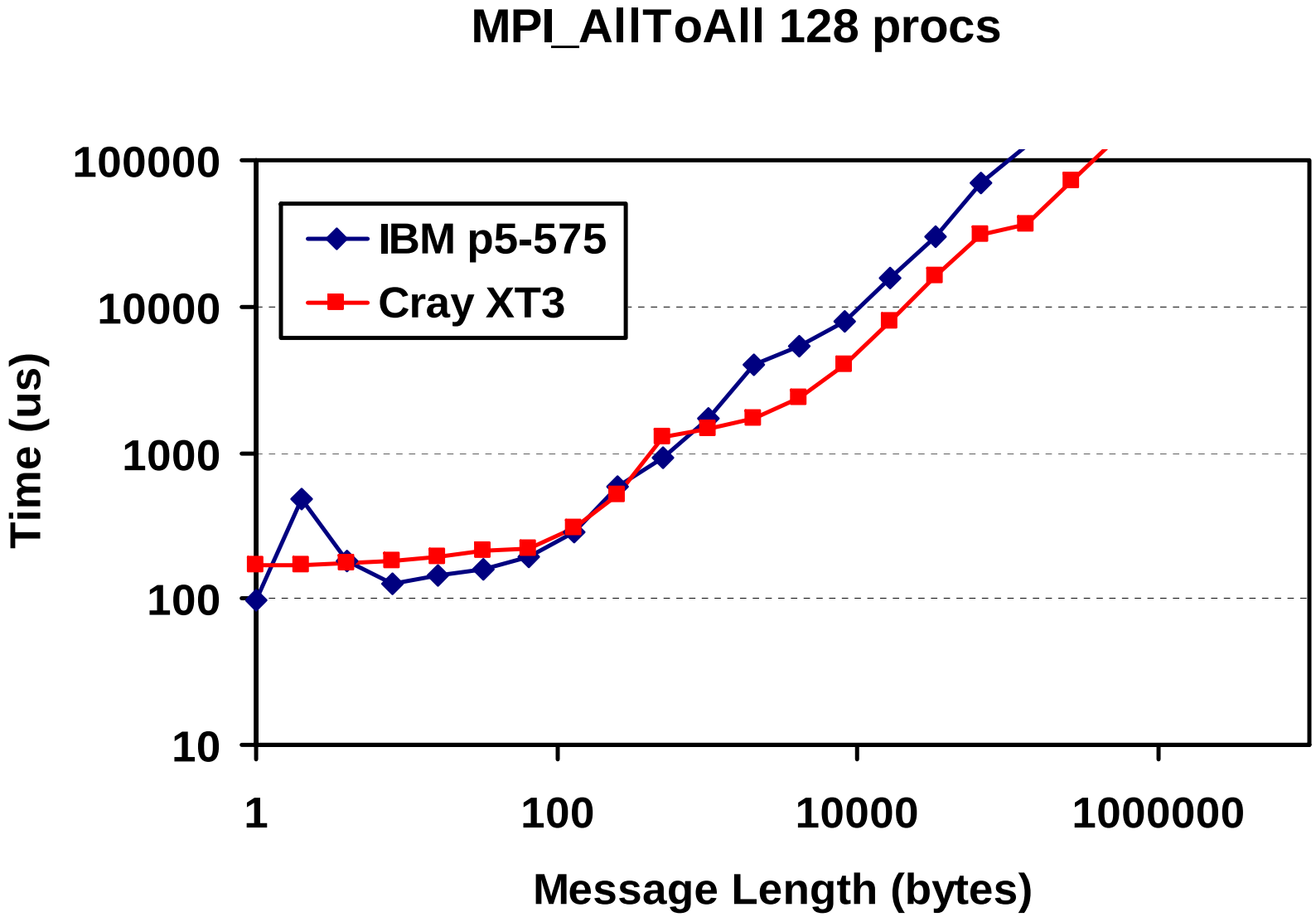

MPI_AllReduce 128 procs
100000
10000
1000
100
10
Time (us)
Cray XT3
IBM p5-575
1
100
10000
1000000
Message Length (bytes)


MPI_Reduce 128 procs
100000
10000
1000
100
10
Time (us)
IBM p5-575
Cray XT3
1
100
10000
1000000
Message Length (bytes)

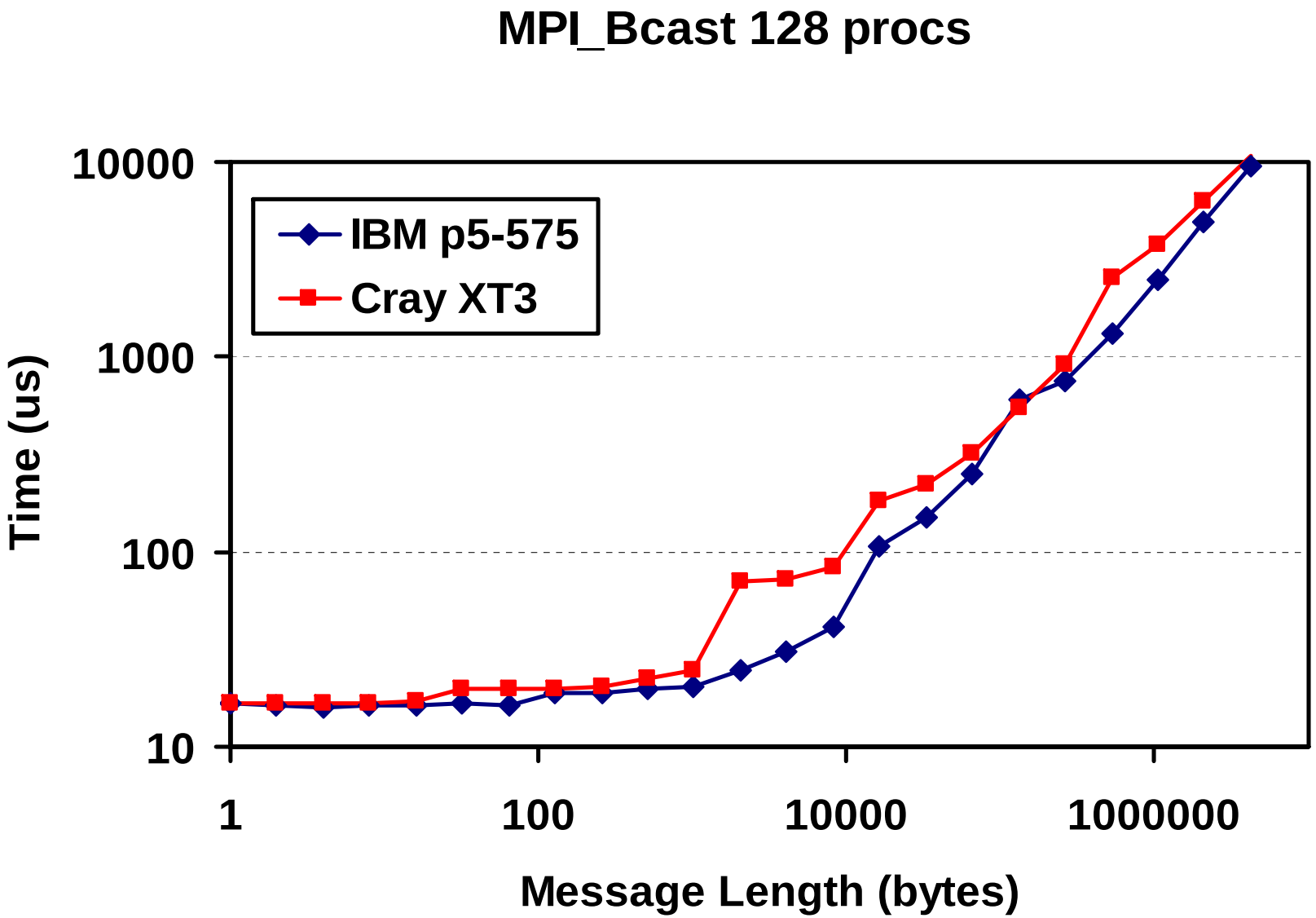

MPI_AllGather 128 procs
100000
10000
1000
100
10
Time (us)
IBM p5-575
Cray XT3
1
100
10000
1000000
Message Length (bytes)


MPI_AllGatherV 128 procs
100000
10000
1000
100
10
Time (us)
IBM p5-575
Cray XT3
1
100
10000
1000000
Message Length (bytes)

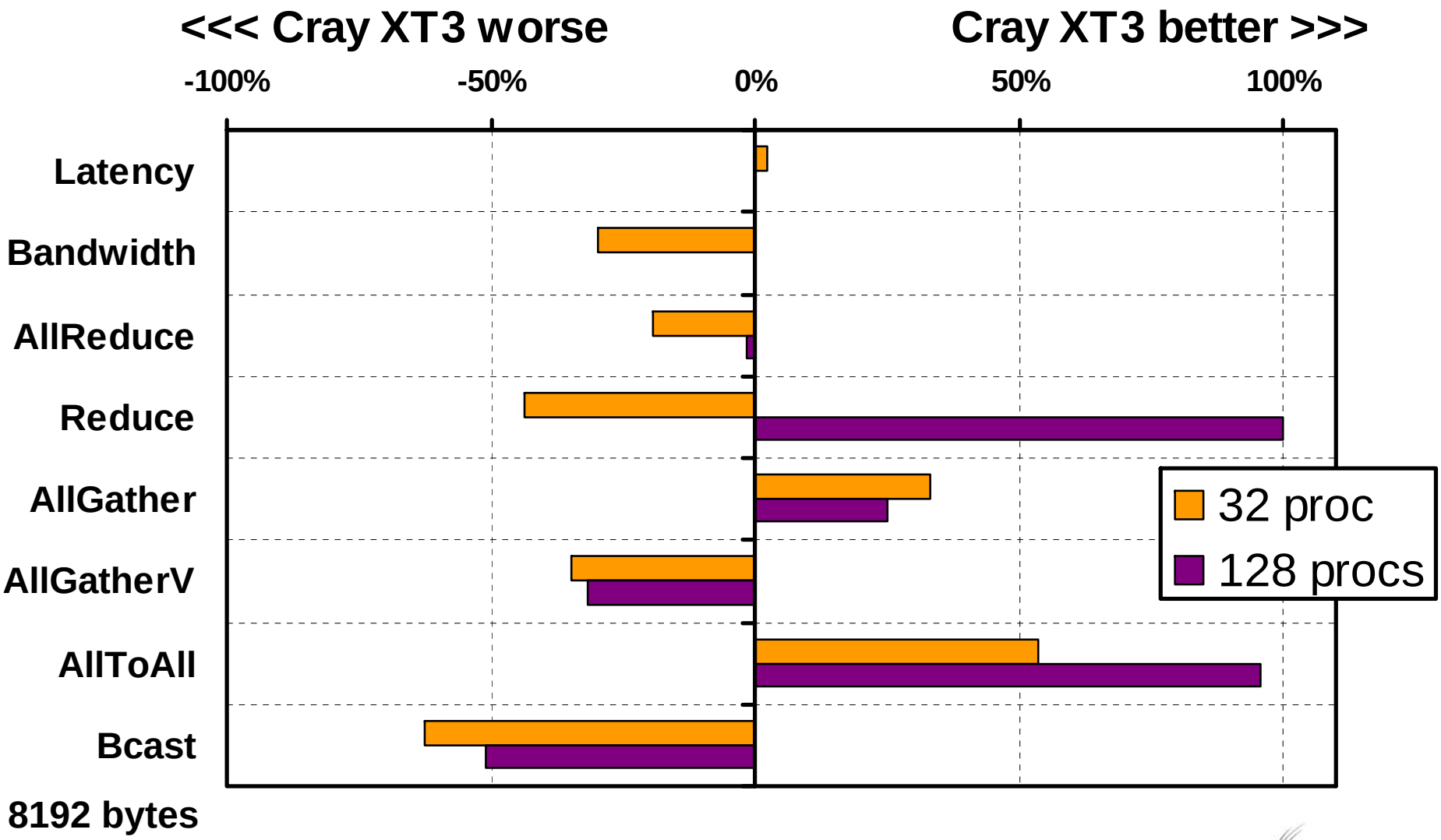
CCLRC


MPI_AllToAll 128 procs
100000
10000
1000
100
10
Time (us)
IBM p5-575
Cray XT3
1
100
10000
1000000
Message Length (bytes)

CCLRC


MPI_Bcast 128 procs
10000
1000
100
10
Time (us)
IBM p5-575
Cray XT3
1
100
10000
1000000
Message Length (bytes)

IMB Summary

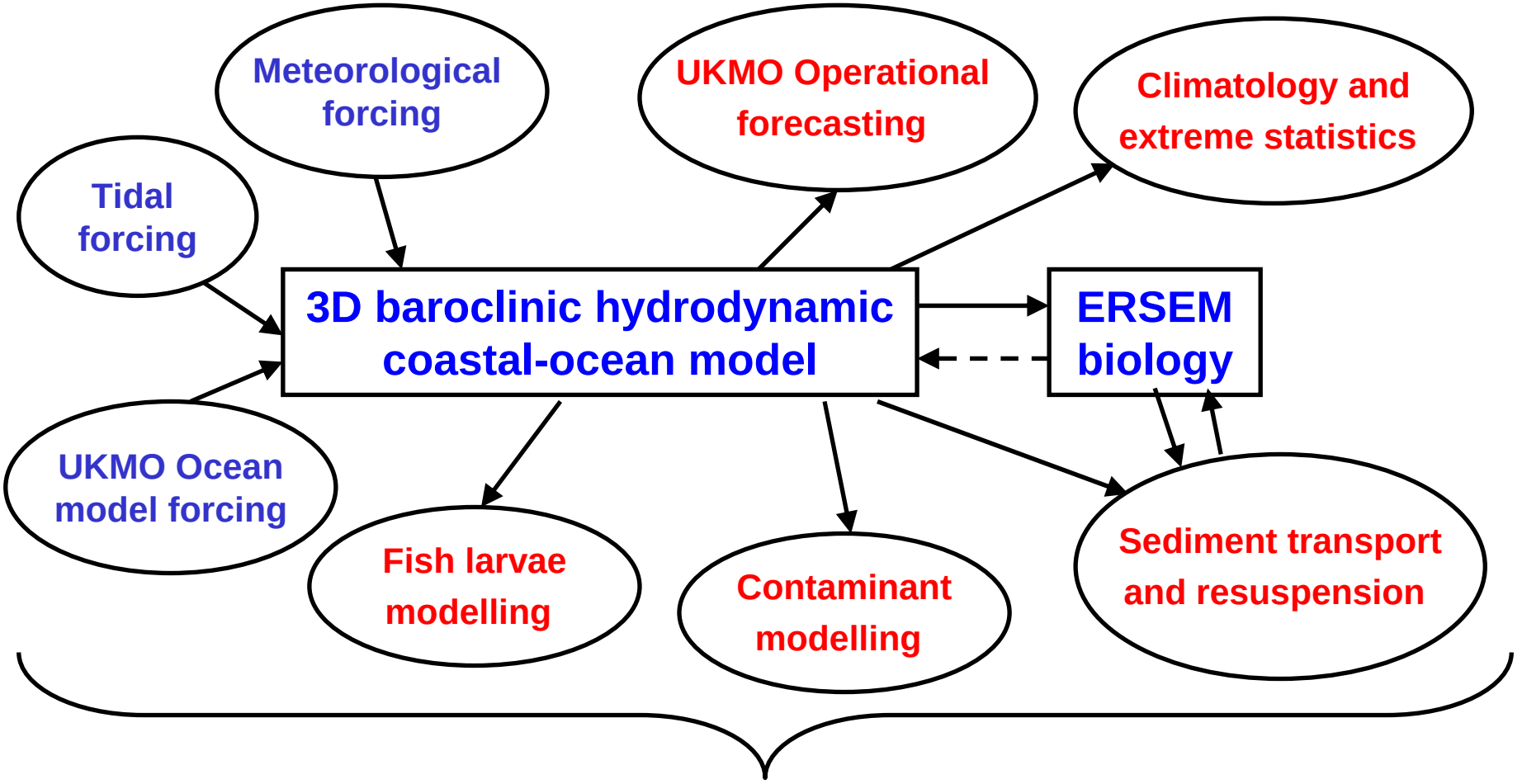

<<< Cray XT3 worse
Cray XT3 better >>>
-100%
-50%
0%
50%
100%
Latency
Bandwidth
AllReduce
Reduce
AllGather
AllGatherV
AllToAll
Bcast
8192 bytes
32 proc
128 procs

Kernel benchmarks are interesting but …

… it's the application performance that counts

- POLCOMS
  - Proudman Oceanographic Laboratory Coastal Ocean Modelling System
  - Coupled marine ecosystem modelling (ERSEM), also wave modelling (WAM), and sea-ice modelling (CICE)
  - Medium resolution continental shelf model - hydrodynamics alone.
- PDNS3D
  - UK Turbulence Consortium, led by Southampton University
  - Direct Numerical Simulation of Turbulence
  - T3 channel flow benchmark on a grid 360 x 360 x 360
- DL_POLY3
  - Molecular dynamics code developed at Daresbury Laboratory
  - Macromolecular simulation of 8 Gramicidin-A species (792,960 atoms)
- GAMESS-US
  - *ab initio* molecular quantum chemistry
  - SiC_3 benchmark
  - Si_nC_m clusters of interest in materials science and astronomy

CCLRC


Meteorological forcing
UKMO Operational forecasting
Climatology and extreme statistics
Tidal forcing
3D baroclinic hydrodynamic coastal-ocean model
ERSEM biology
UKMO Ocean model forcing
Fish larvae modelling
Contaminant modelling
Sediment transport and resuspension
Visualisation, data banking & high-performance computing

POLCOMS code characteristics

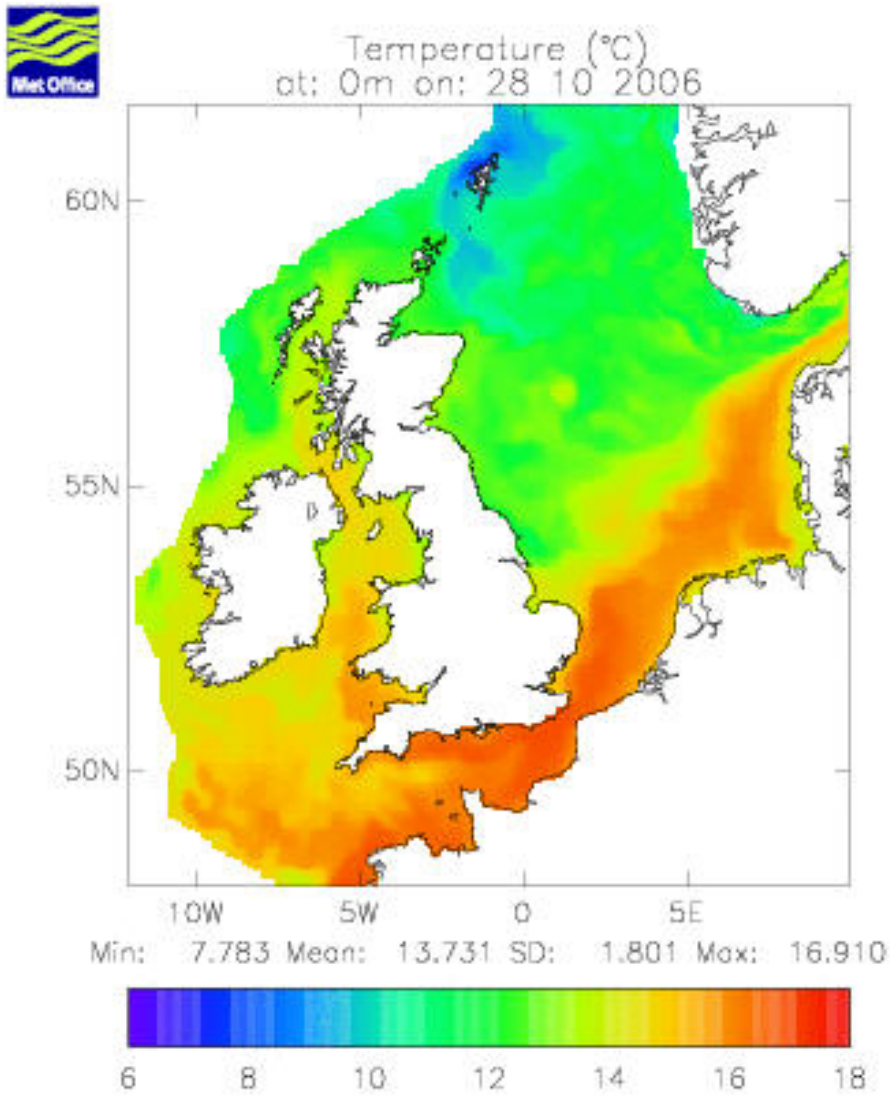
- 3D Shallow Water equations
  - Horizontal finite difference discretization on an Arakawa B-grid
  - Depth following sigma coordinate
  - Piecewise Parabolic Method (PPM) for accurate representation of sharp gradients, fronts, thermoclines etc.
  - Implicit method for vertical diffusion
  - Equations 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

CCLRC

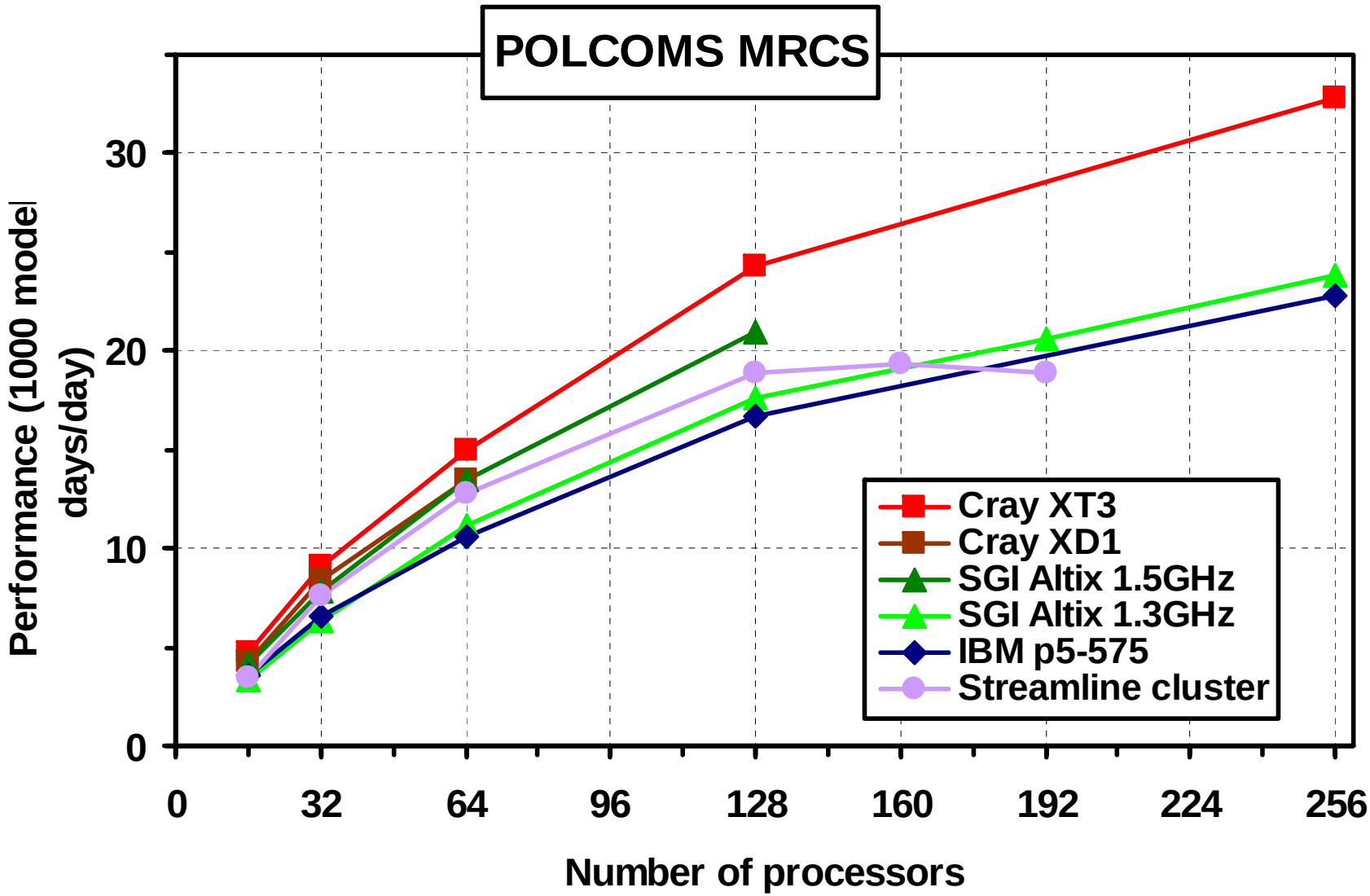
MRCS model

- 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 28th Oct 2006



http://www.metoffice.gov.uk/research/ncof/mrcs/browser.html

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

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

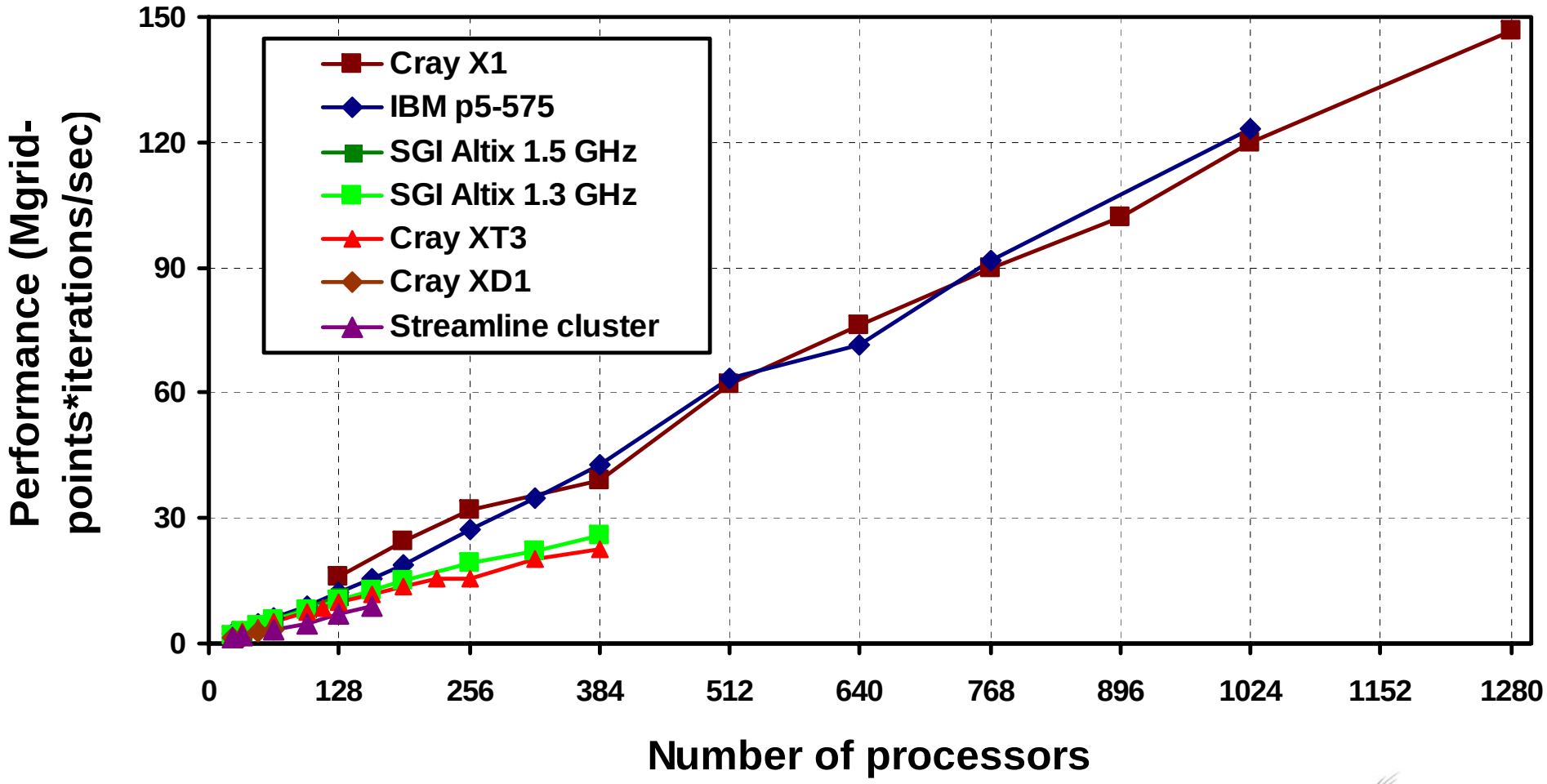
http://www.afm.ses.soton.ac.uk/

CCLRC

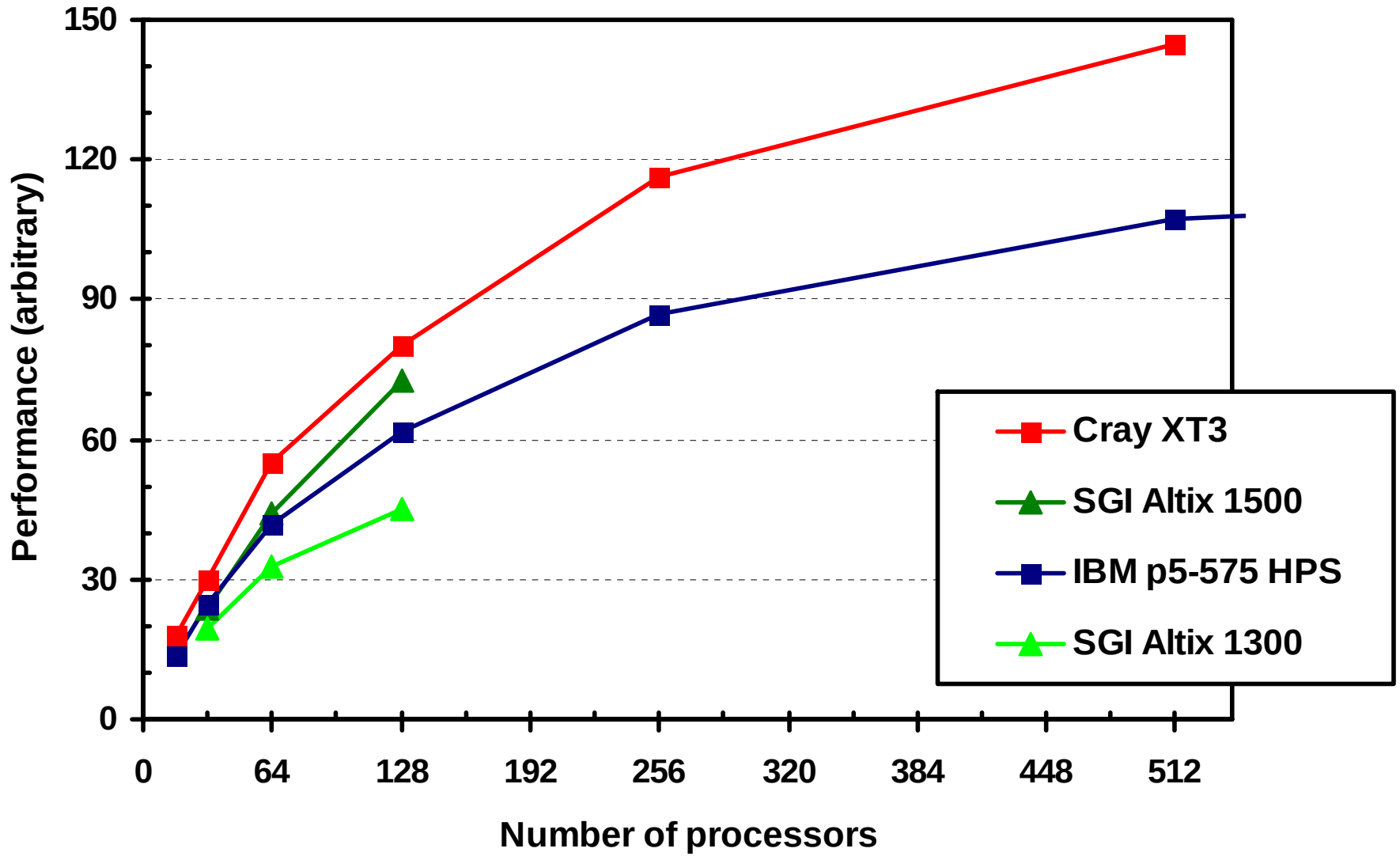
PDNS3D Code - characteristics

- 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

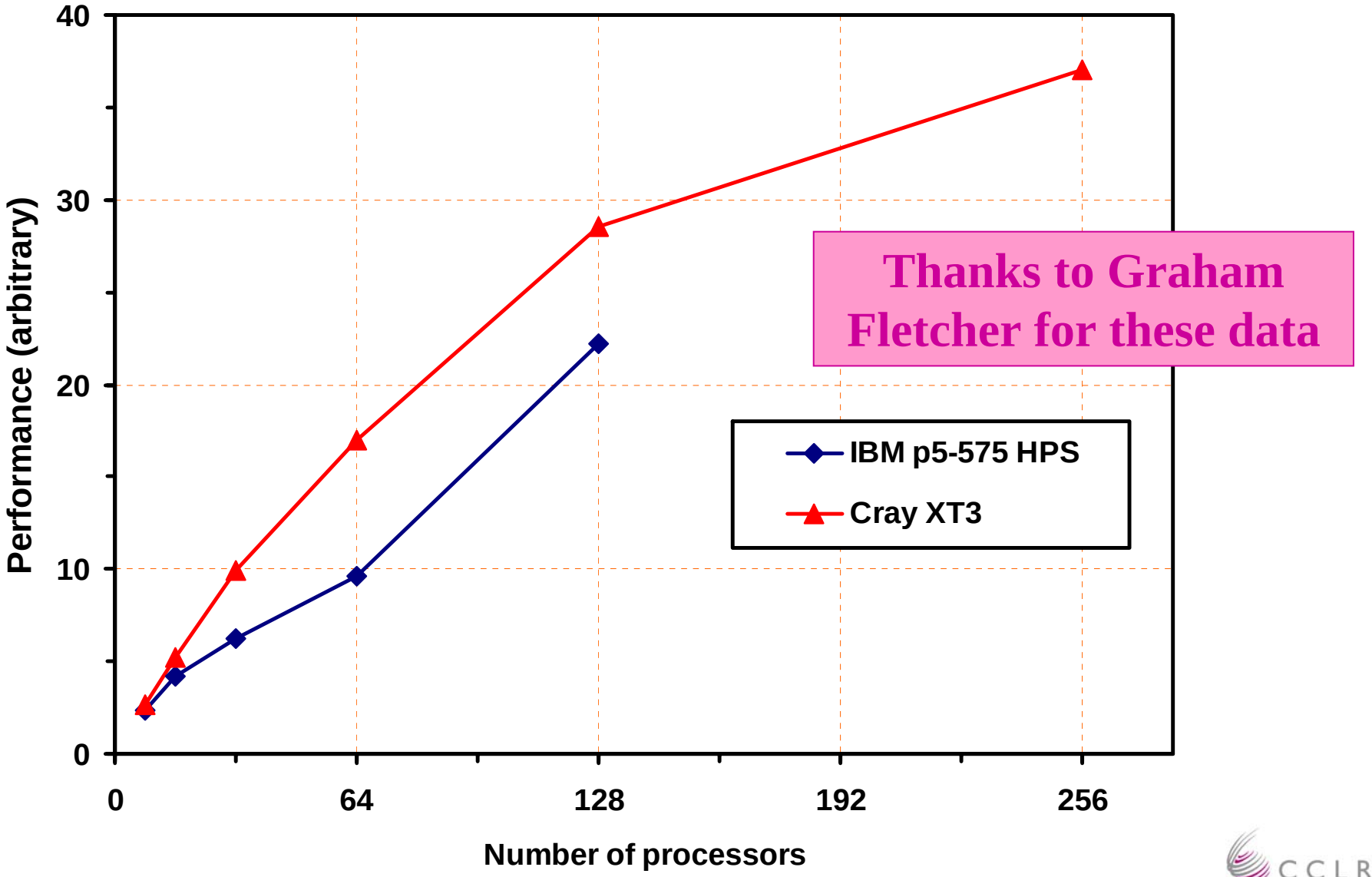
PCHAN T3 (360 x 360 x 360)



CCLRC

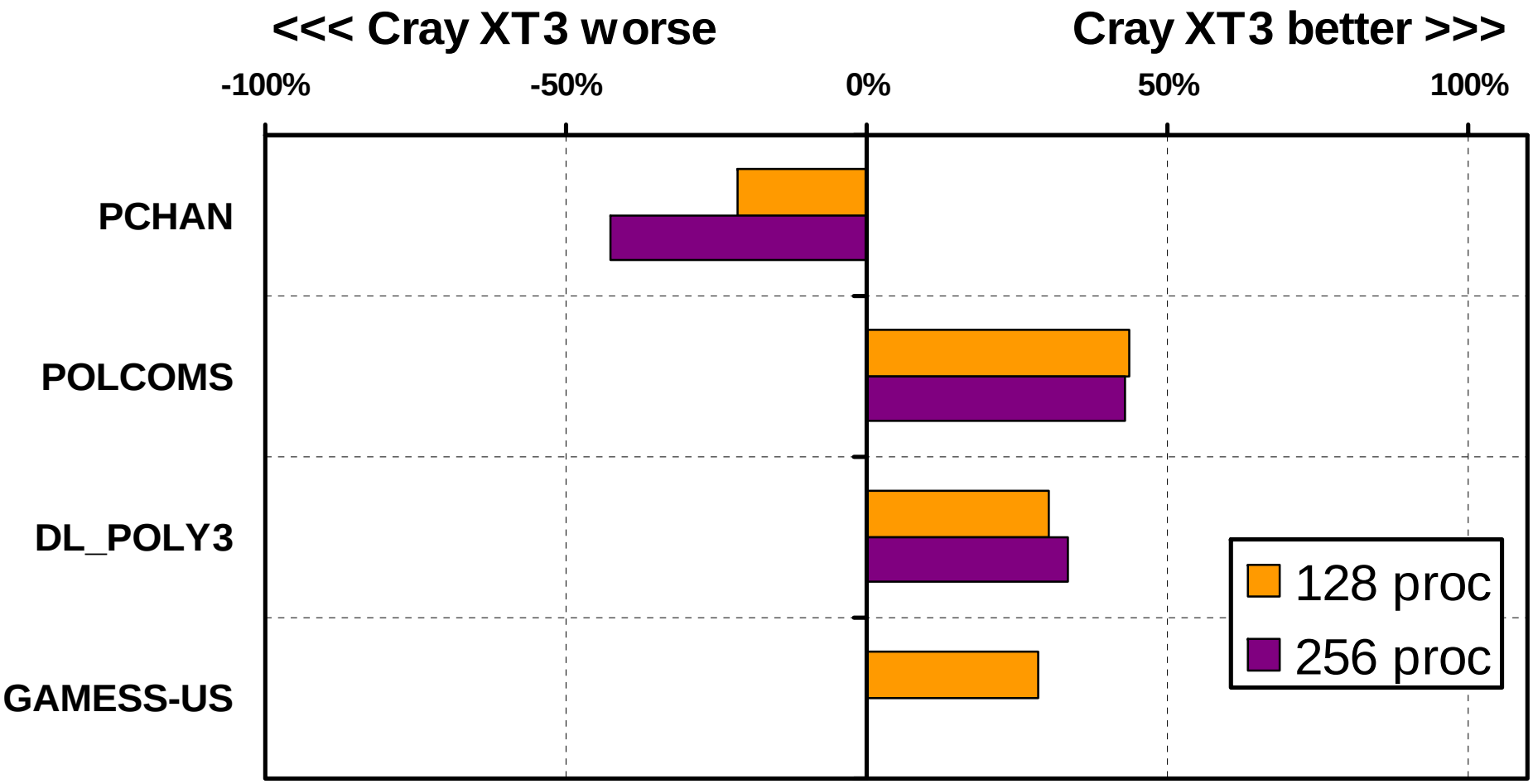

150
120
90
60
30
0
Performance (arbitrary)
0
64
128
192
256
320
384
448
512
Number of processors
Cray XT3
SGI Altix 1500
IBM p5-575 HPS
SGI Altix 1300

CCLRC


40
30
20
10
0
Performance (arbitrary)
0
64
128
192
256
Number of processors
IBM p5-575 HPS
Cray XT3
Thanks to Graham Fletcher for these data

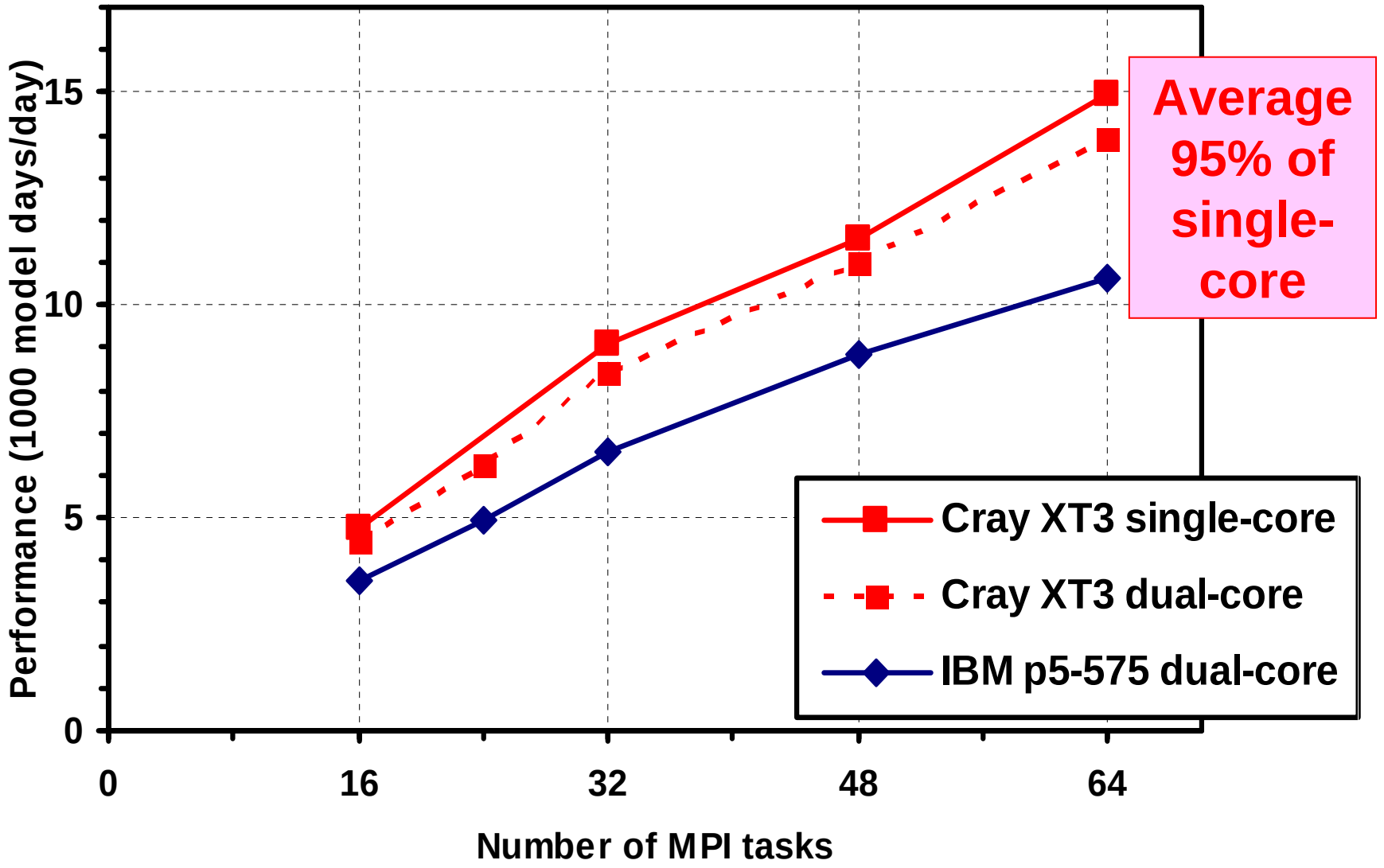
So how did the applications do?

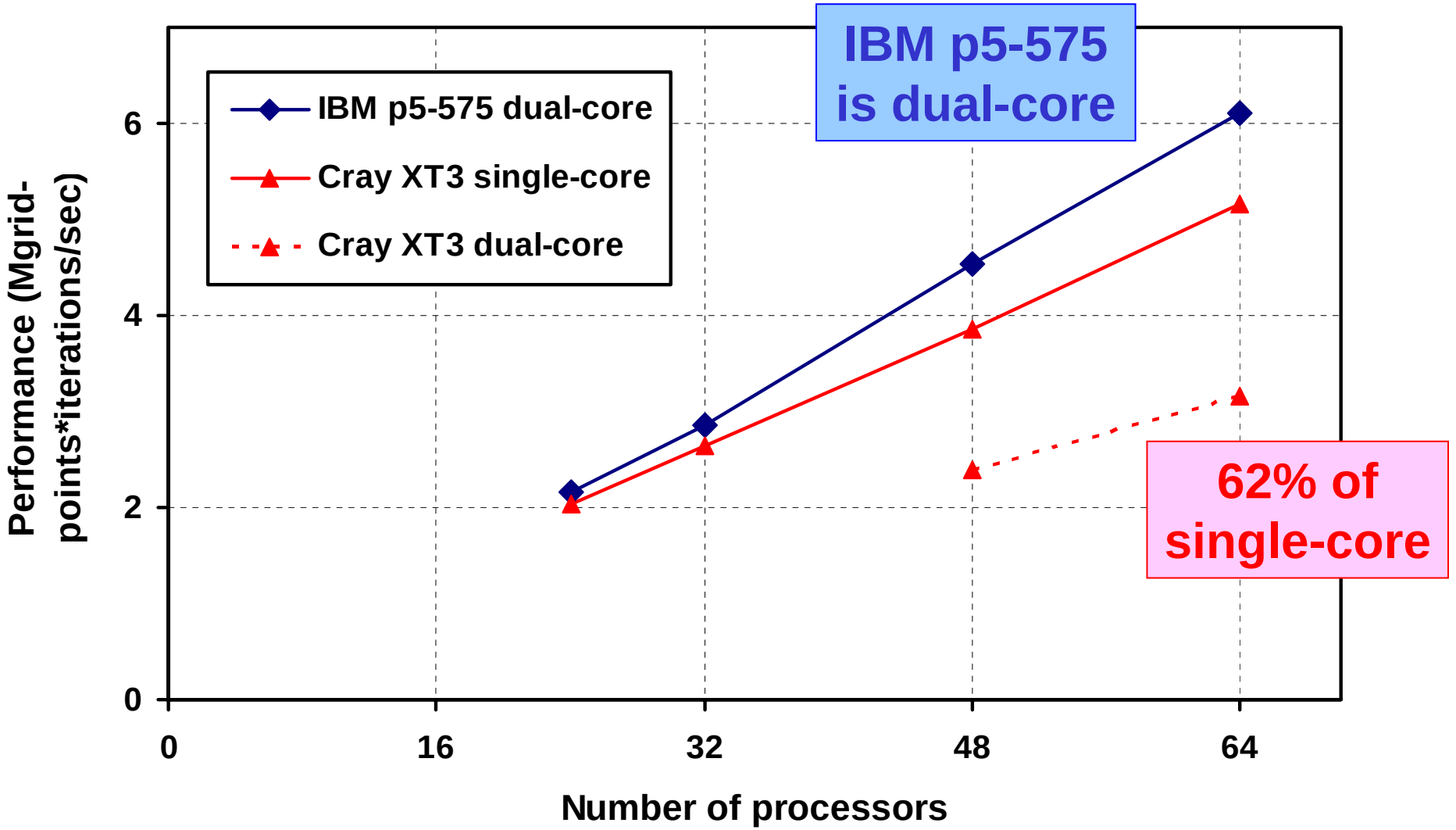
Application Summary



What about dual-core on the Cray XT3?

CCLRC


Performance (1000 model days/day)
0
5
10
15
0
16
32
48
64
Number of MPI tasks
Cray XT3 single-core
Cray XT3 dual-core
IBM p5-575 dual-core
Average 95% of single-core


IBM p5-575 is dual-core
IBM p5-575 dual-core
Cray XT3 single-core
Cray XT3 dual-core
62% of single-core
Performance (Mgrid-points*iterations/sec)
0
2
4
6
0
16
32
48
64
Number of processors

Conclusions

- We have shown initial results from an applications level benchmark comparison between IBM POWER5 and Cray XT3
- Intel MPI benchmarks are very similar
  - Latency is almost identical at 5.5-5.6 us
  - IBM HPS has better achieved bandwidth: 1.6 GB/s vs. 1.1 GB/s
- Applications performance depends on the application
  - So far … most (3:1) apps perform better on the Cray XT3
  - One memory intensive app performs much better on the IBM
  - Memory intensive apps appear to be badly affected by the move to dual-core (& multi-core?)

If you have been …

Thank you for listening

CCLRC