Science & Technology Facilities Council

IBM

The IBM Blue Gene/Q: Application performance, scalability and optimisation

Mike Ashworth, Andrew Porter

Scientific Computing Department & STFC Hartree Centre

Manish Modani

IBM

STFC Daresbury Laboratory, UK

mike.ashworth@stfc.ac.uk

Overview

- Blue Gene/Q
- WRF
  - Computational Performance
  - Pure MPI vs Hybrid MPI-OpenMP
  - I/O Performance
- Conclusions

Science & Technology Facilities Council

IBM

UK Government Investment

17th Aug 2011: Prime Minister David Cameron confirmed £10M investment into STFC's Daresbury Laboratory. £7.5M for computing infrastructure

3rd Oct 2011: Chancellor George Osborne announced £145M for e-infrastructure at the Conservative Party Conference

4th Oct 2011: Science Minister David Willetts indicated £30M investment in Hartree Centre

30th Mar 2012: John Womersley CEO STFC and Simon Pendlebury IBM signed major collaboration at the Hartree Centre

Clockwise from top left

Science & Technology Facilities Council

IBM

STFC Hartree Centre

1st Feb 2013: Chancellor George Osborne officially opened the Hartree Centre

The Hartree Centre is a Research Collaboratory in association with IBM supporting innovation in science and industry

- Software development
- Applications and optimisation
- HPC on demand
- Collaboration
- Training and education

工合

Gung-Ho is a flagship project of the Hartree Centre

Blue Gene/Q

Blue Gene Evolution – 3 generations

KEY PROPERTIES	BG/L	BG/P	BG/Q
Compute Nodes			
Processor	32-bit PowerPC 440	32-bit PowerPC 450	64-bit PowerPC (A2 Core)
Processor Frequency	0.7 GHz	0.85 GHz	1.6 GHz (target)
Cores	2	4	16 + 1
FPU	Double Hummer (2x)	Double Hummer (2x)	QPU (4x)
Peak Performance	5.7 TF/rack	13.9 TF/rack	209.7 TF/full rack
Main Memory / Node	512 MB or 1 GB	2 GB or 4 GB	16 GB
Torus Network			
Dimensions	3D	3D	5D
Bandwidth	2.1 GB/s	5.1 GB/s	32 GB/s
System			
Peak Performance	360 TF (64 racks)	1 PF (72 racks)	20 PF (96 racks)
Total Power	1.5 MW (64 racks)	2.9 MW (72 racks)	~5 MW (96 racks)
Year Introduced	2004	2007	2011
Price-Performance	~18 cents/MF	~11 cents/MF	~1.4 cents/MF

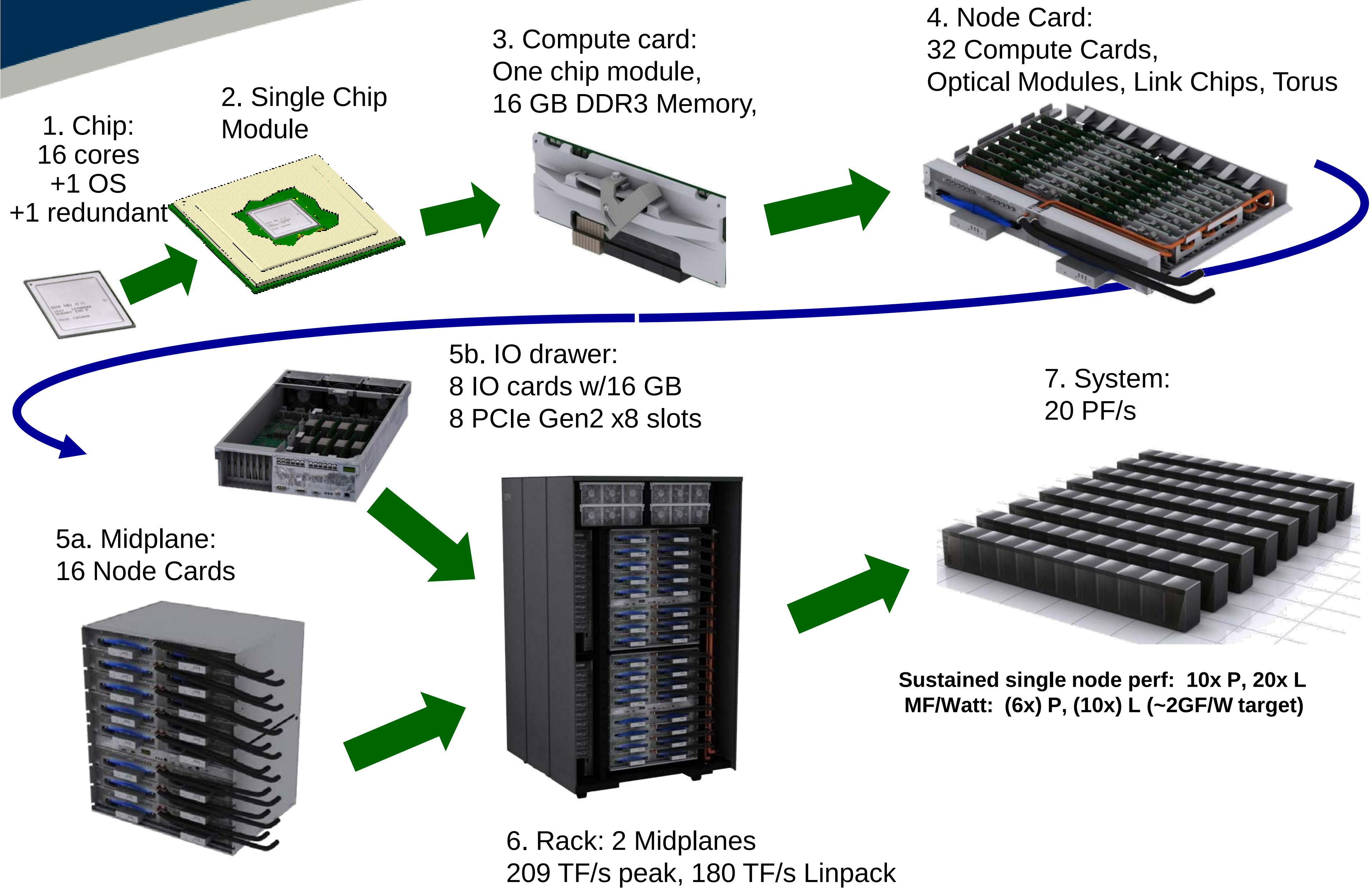
BG/Q Philosophy

- Energy efficiency through large numbers of low power cores (18 of TOP30 Green500 are BG/Q at > 2 GF/s/W
- Standard MPI applications that scale well can run on BG/Q without modification
- Hybrid mode (MPI + OpenMP) will help to exploit large numbers of cores.
- For typical systems an application scales to 4096 cores
  - 4096 MPI tasks, 1024 nodes of BG/P
- After hybridization on BG/Q could scale to 65536 cores
  - 4096 MPI tasks, 4096 nodes,16/32/64 threads/task (SMT)
- SMT can improve performance by hiding memory latency

Science & Technology Facilities Council

BG/Q Packaging

IBM



Sustained single node perf: 10x P, 20x L
MF/Watt: (6x) P, (10x) L (~2GF/W target)

Hartree Centre BG/Q

#2 system in UK (#1 2012)

#23 in the world (#13 2012)

6+1 racks

16 cores, 16 GB per node

6 racks

- 98,304 cores
- 1.26 Pflop/s peak

1 rack of BGAS (Blue Gene Advanced Storage)

- 16,384 cores

Blue Gene/Q Optimisation

- Scalability
  - Slow clock (1.6 GHz) means that it is vital to scale efficiently to larger numbers of cores
  - On the BG/P this was 3x-4x – BG/Q the gap has narrowed
- Vectorisation and FMA
  - BG/Q Linpack is 10.9 GF/s/core @ 1.6 GHz – 85% of peak
  - Relies on 8 flops per cycle, quad vector units, FMA
  - Develop your relationship with the IBM XL Fortran compiler!
- Hybrid MPI and OpenMP
  - OpenMP helps by reducing MPI costs
  - OpenMP may not scale; consider balance of tasks to threads
- SMT can be beneficial to mask memory latency
- I/O needs to be carefully considered
  - Each I/O node serves 128 compute nodes.

WRF

Science & Technology Facilities Council

IBM

Weather Research and Forecast (WRF) Model

- Regional- to global-scale model for research and operational weather-forecast systems (WRF)
- Developed through a collaboration between various US bodies (NCAR, NOAA...)
- Finite-difference scheme + physics parameterisations
- Features two dynamical cores, data assimilation system
- Software architecture for parallel computation
- F90 [+ MPI] [+ OpenMP]
- 20,000 registered users.
- Used in Academia and Industry

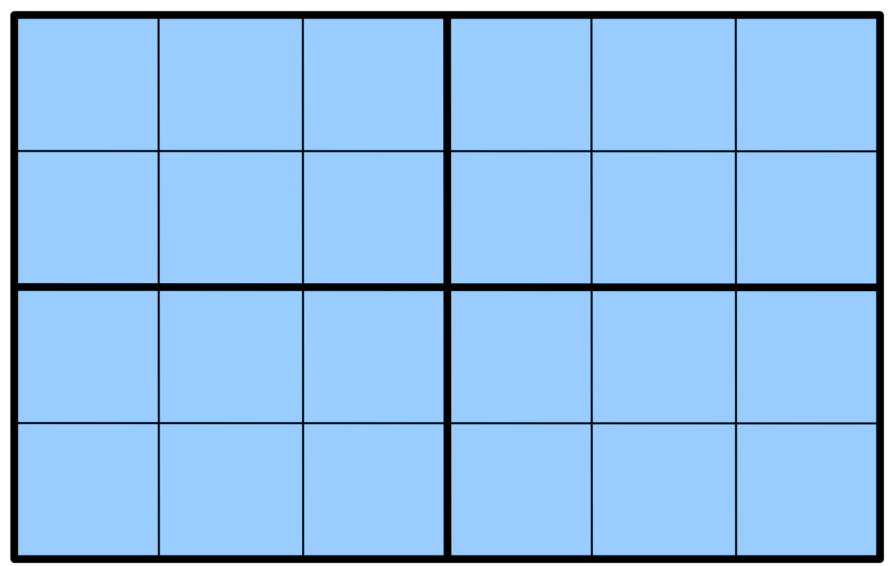
Introduction to this work

- WRF accounts for significant fraction of usage of UK national facility
- I/O is the major bottleneck in scalability
- Aim here is to investigate the WRF I/O performance at large core counts (>10000)
- Mainly through API for I/O-Layer NETCDF/PNETDF/GRIB2

WRF Parallelism

- Efficient use of large, on-chip memory cache is very important in getting high performance from chips
- Under MPI, WRF gives each process a 'patch' to work on. These patches can be further decomposed into 'tiles' (used by the OpenMP implementation)

e.g. decomposition of domain into four patches with each patch containing six tiles



WRF Domain

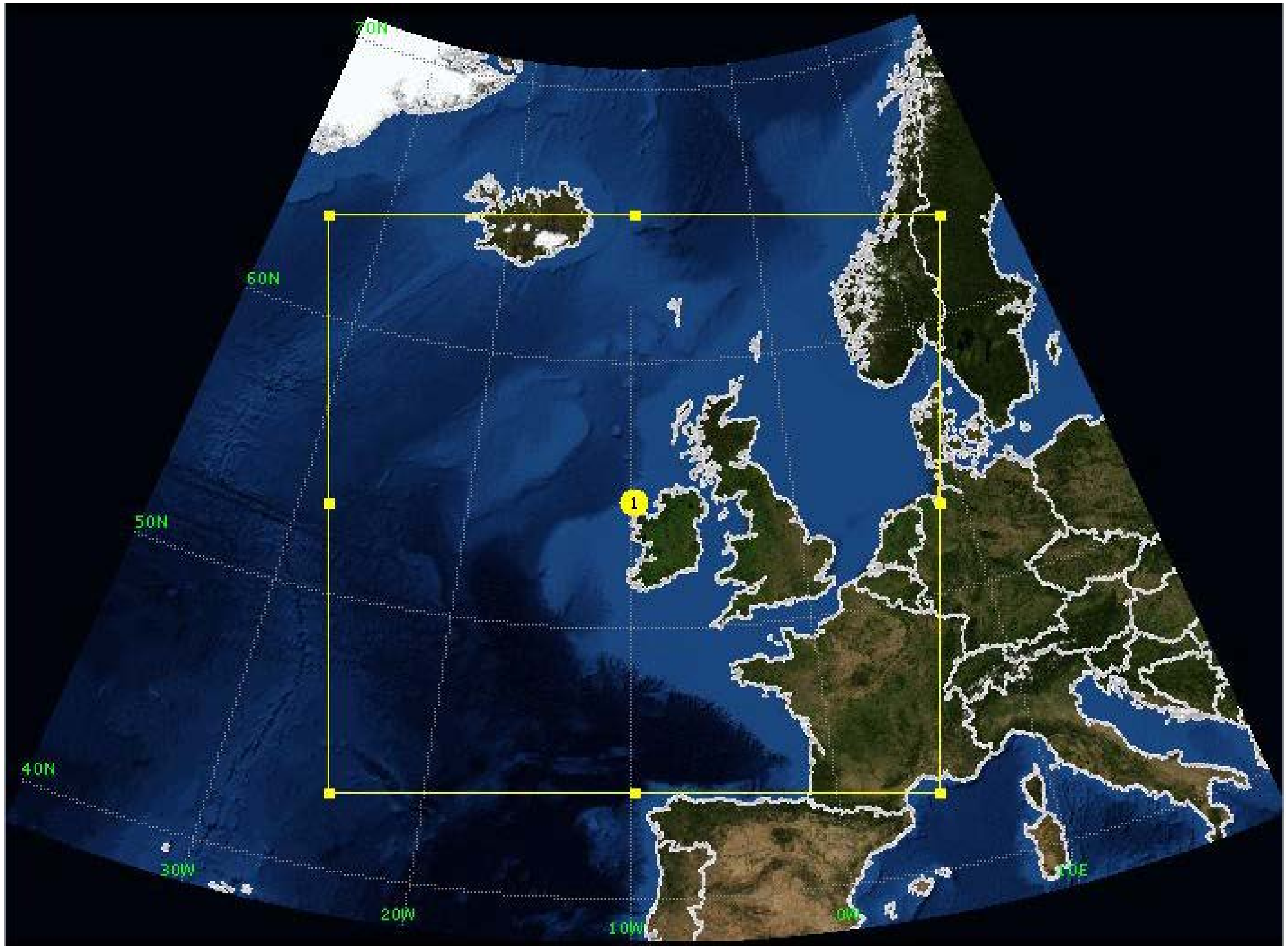
Domain Size

1200x1200x81

2km resolution

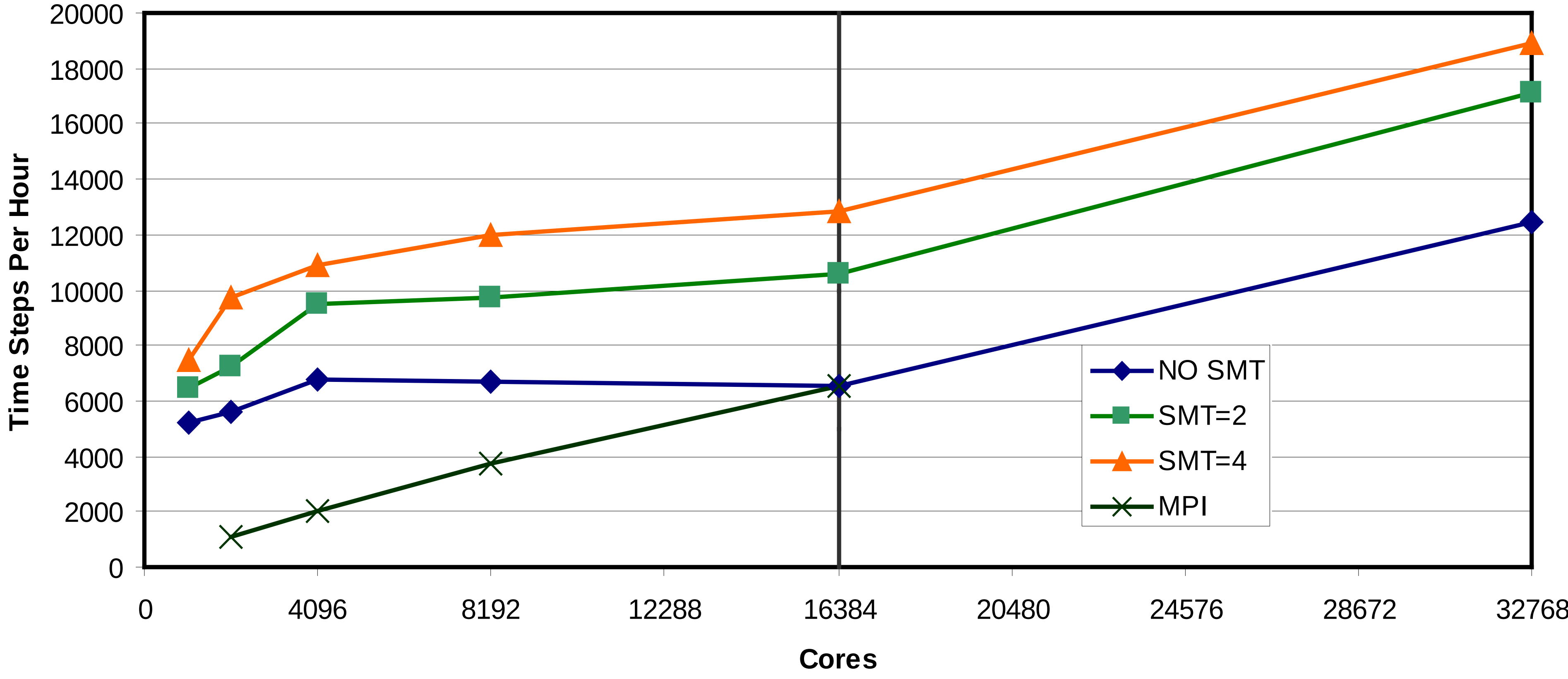
WRF 3.4.1

WRF minimum patch size of 9x9, so upper limit of 17,689 PEs for this domain



WRF Performance: Hybrid Mode

up to 32K cores



Hybrid Mode & SMT gives better performance

Approaches to I/O in WRF

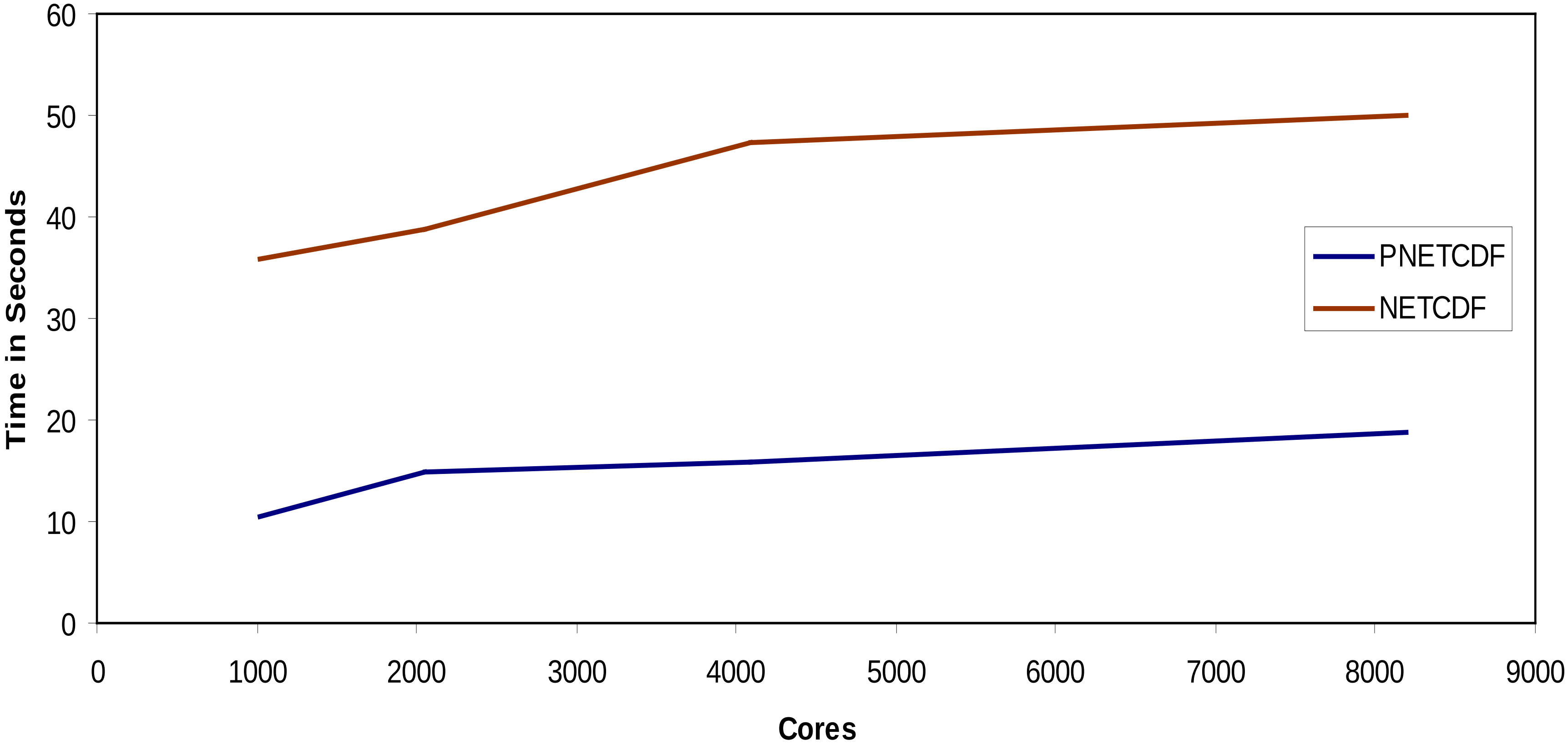
Serial I/O (default)

- Data for whole model domain gathered on ‘master’ PE which then writes to disk
- All PEs block while master is writing; does not scale; memory limitations
- Approximately 75% (22% in wrfout & 54 % in wrfrst) of wall time in I/O (on 1024 cores)

Parallel netCDF

- Every MPI process writes; also unscalable
- Approximately 25% (7% in wrfout and 16 % in wrfrst) of wall clock time in I/O

WRF pNetCDF I/O Time

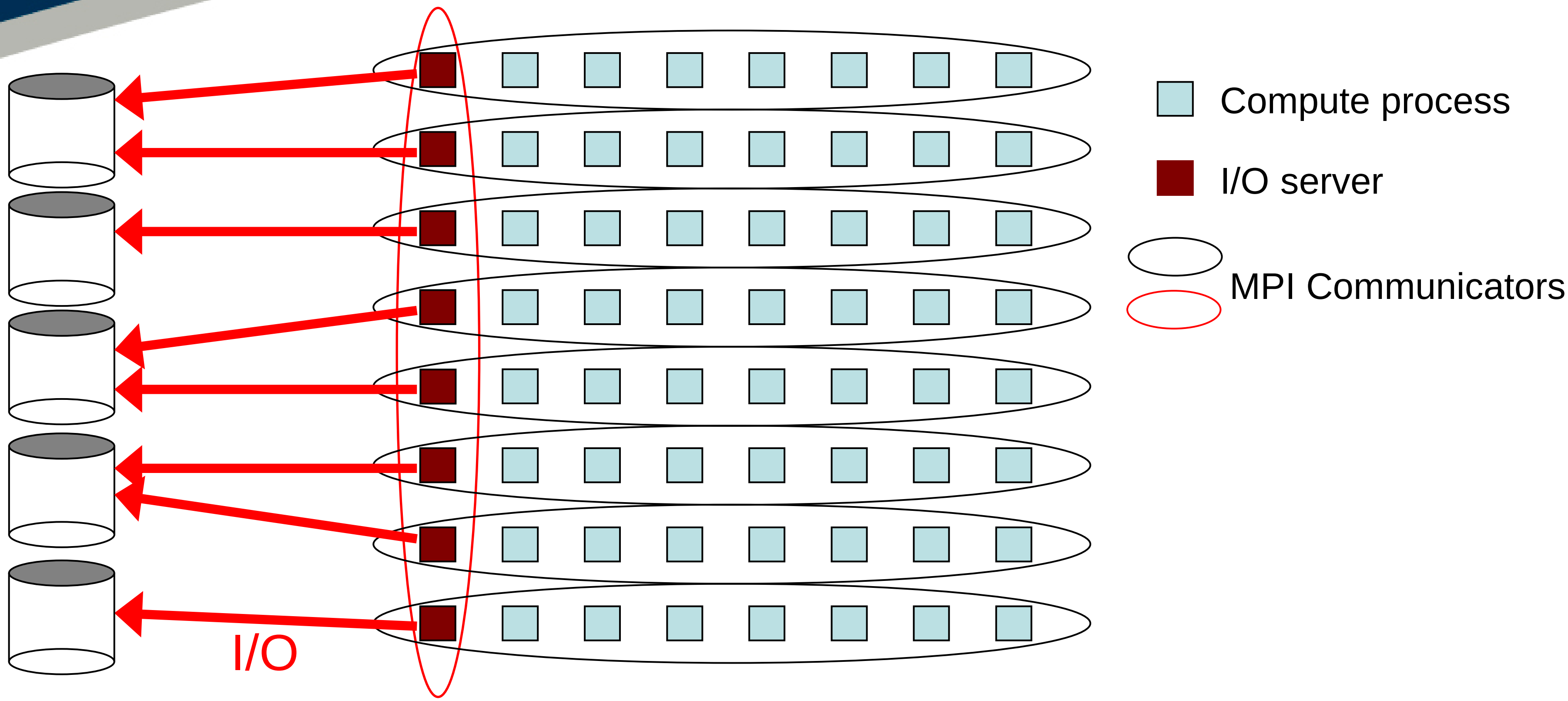


Wrfout file size: 8.3 GB, parallel Netcdf improves the I/O performance by 60%

WRF I/O Quilting

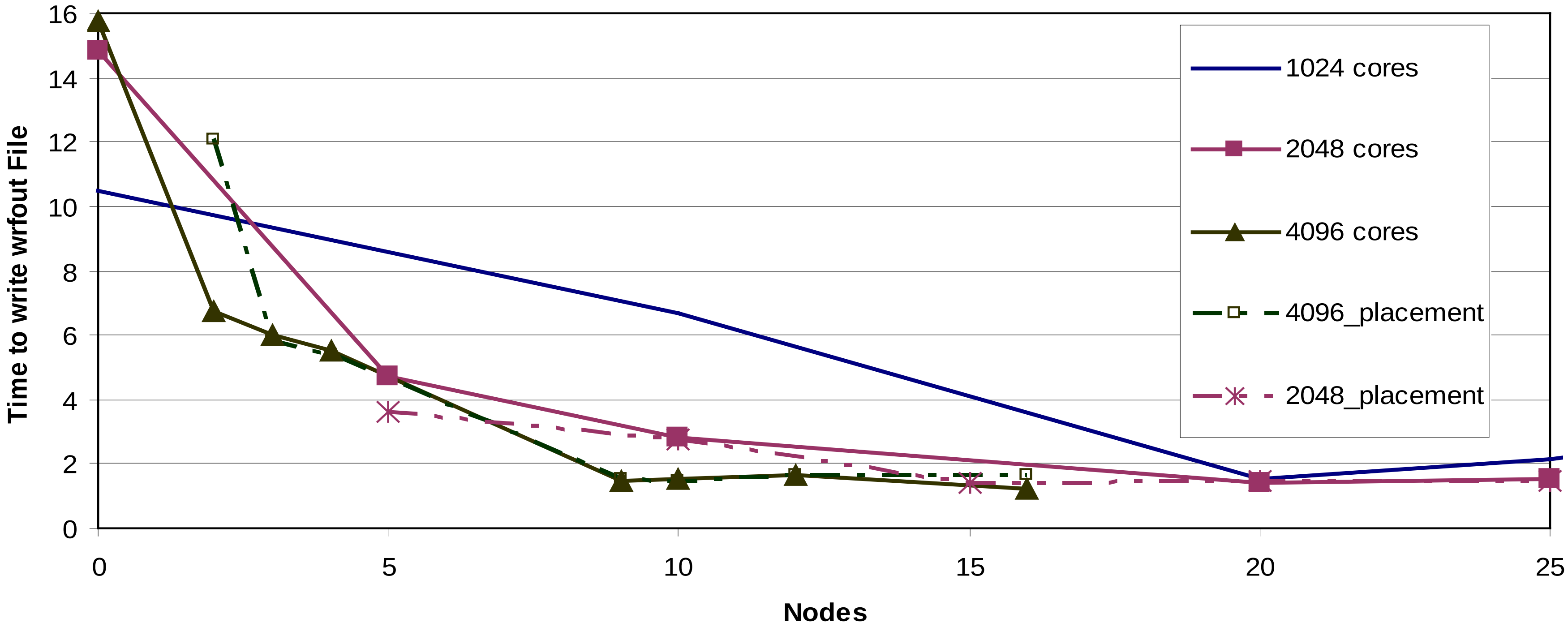
- Use dedicated I/O servers to write data
- Compute PEs are free to continue once data are sent to I/O servers
- No longer have to block while data are sent to disk
- Number of I/O servers may be tuned depending on the gather time and the parallel file system

WRF process mapping



- How best to assign compute PEs to I/O servers?
- How best to assign I/O server PEs within the pool of all PEs? (Match to hardware I/O nodes on the Blue Gene)

WRF quilting performance



- Performance investigated on 1 rack
- Best performance 20 I/O servers per rack is around 2%
- WRF cannot run > 60 quilt servers with 1 I/O group
- Task placement does not impact performance on higher number of quilt servers

Conclusions

WRF performs well

- WRF scales well on higher core counts (32k)
- Hybrid mode with SMT yields best performance
- Time spent in I/O is significant
- pNetcdf helps in reducing the I/O time significantly
- Quilting is the best option at scale
- 2% cores allocated to quilts yields the best performance

BG/Q is a highly energy efficient solution for highly scalable applications

- Allows us to develop hybrid scalable MPI-OpenMP codes to O(100,000) cores

Science & Technology Facilities Council

IBM

Publication

This work was presented at the Exascale Applications and Software Conference, 9^{th}-11^{th} April 2013, Edinburgh, UK

It will be published in a paper

"Strategies for I/O-Bound Applications at Scale",

Manish Modani and Andrew Porter,

to appear in a special edition of the journal

Advances in Engineering Software