Cray Centre of Excellence for HECToR: Activities and Future Projects

Jason Beech-Brandt, Tom Edwards and Kevin Roy

HECToR User Group Meeting 2010

12 Oct 2010

Manchester

Activities

- Phase 2b – Cray XT6
- Current and Past Projects
  - SBLI
  - LUDWIG
  - HiGEM
  - GADGET
  - HELIUM
- Future Projects
  - Cray XE6
  - GEMINI results
- New Cray Centre of Excellence for HECToR website!
- Cray Exascale Research Initiative Europe
- Training and Conferences

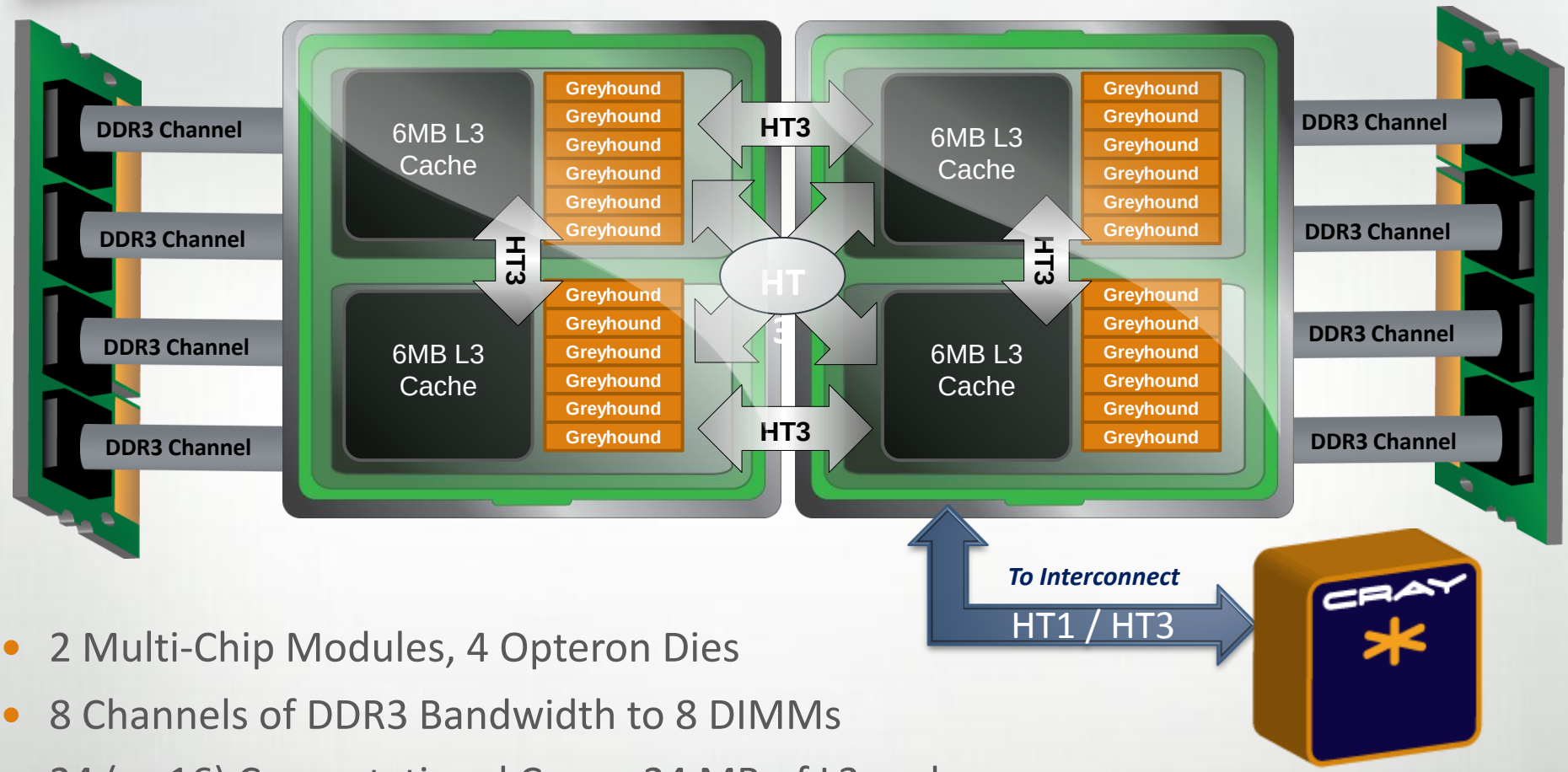
Cray Systems at HECToR

- Cray XT4 Compute Nodes
- 3072 x 2.3 GHz Quad Core Opteron Nodes
- 12,238 Cores
- 8GB per node – 2GB per Core

- Cray X2 Compute Nodes
- 28 x 4 Cray X2 Compute nodes
- 32 GB per node 8GB per Core

- Cray XT6 Compute Nodes
- 1856 x 2.1 GHz Dual 12 Core Opteron Nodes
- 44,544 Cores
- 32GB per node – 1.33 GB per Core

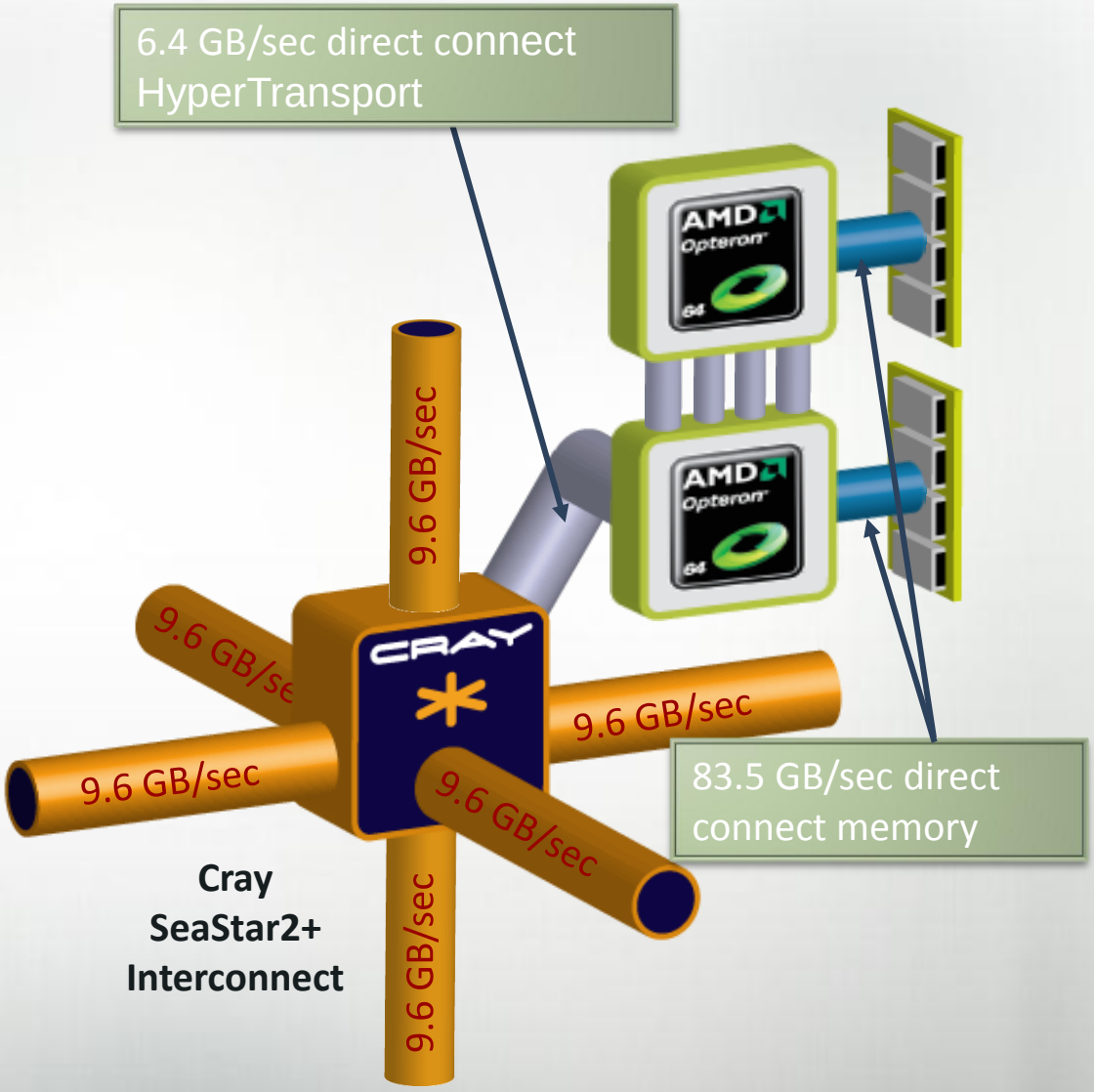
Cray XT6 Node Details: 24-core Magny Cours



- 2 Multi-Chip Modules, 4 Opteron Dies
- 8 Channels of DDR3 Bandwidth to 8 DIMMs
- 24 (or 16) Computational Cores, 24 MB of L3 cache
- Dies are fully connected with HT3
- Snoop Filter Feature Allows 4 Die SMP to scale well

Cray XT6 Node

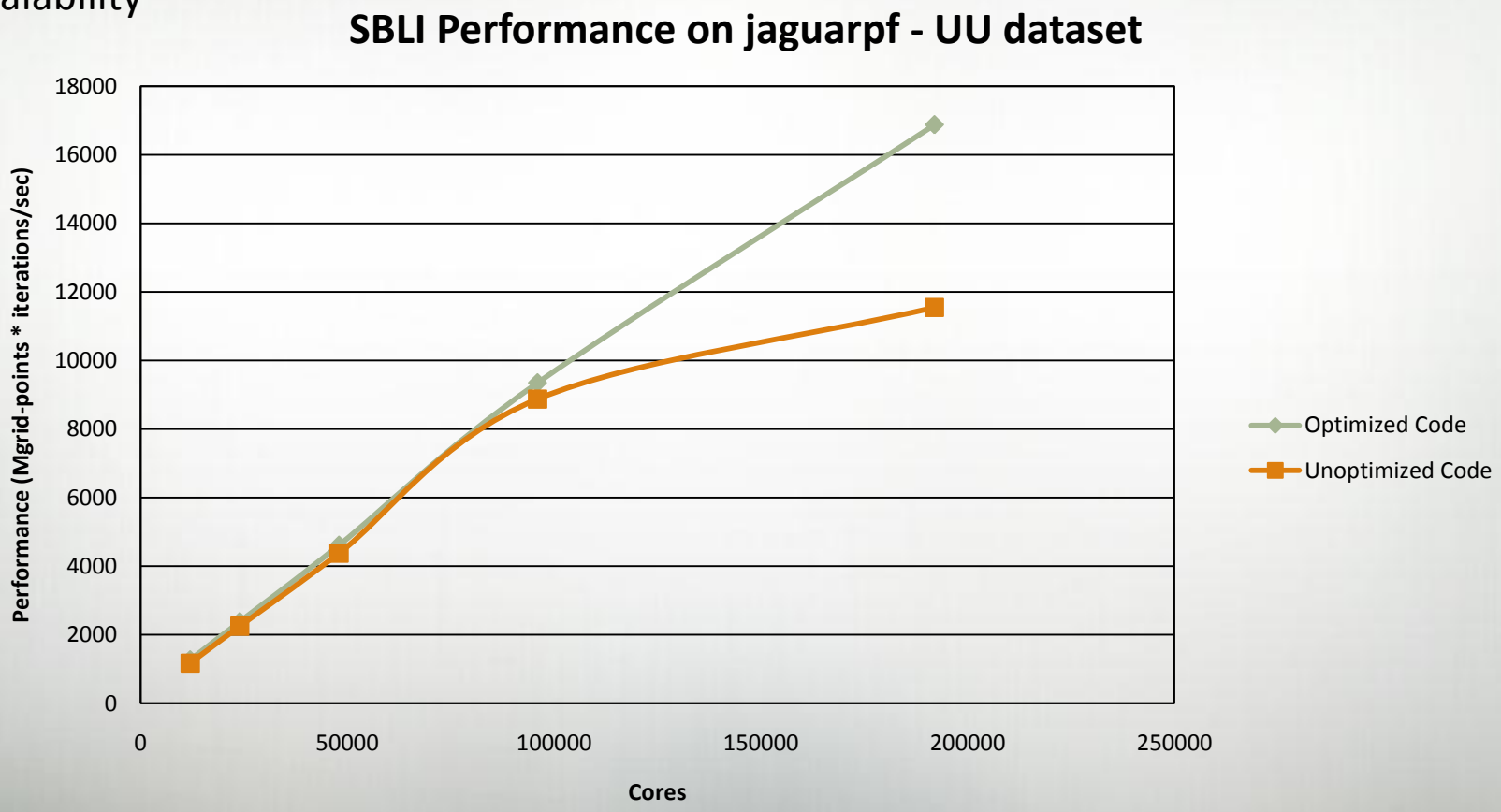
Characteristics	
Number of Cores	**24**
Peak Performance MC-12 (2.1)	**202Gflops/sec**
Memory Size	**32GB per node**
Memory Bandwidth	**83.5 GB/sec**



Projects Update

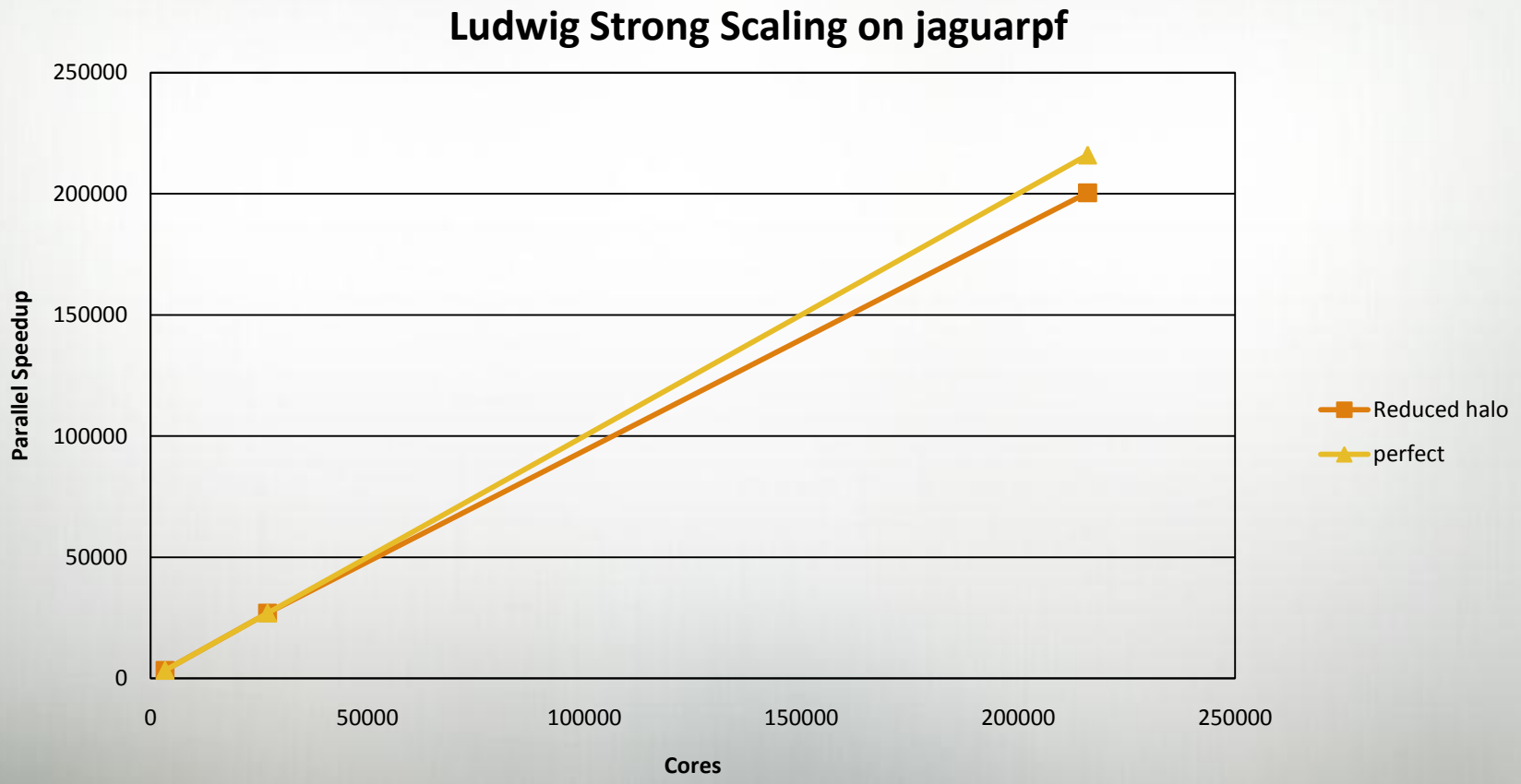
SBLI on jaguarpf – the importance of preposting receives

- User reported large runs on jaguarpf were dying with MPI buffer overflows
- Preposting of MPI receives, and appropriate MPI environment variable settings, alleviates this issue, and provides much better scalability



LUDWIG

- Lattice-Boltzmann code by Dr Kevin Stratford (EPCC) and Prof. Mike Cates (University of Edinburgh – Dept of Physics)
- Scaling very well up to 216000 cores
  - Smaller global lattice than ran previously on barcelona jaguarpf – really pushing the strong scaling



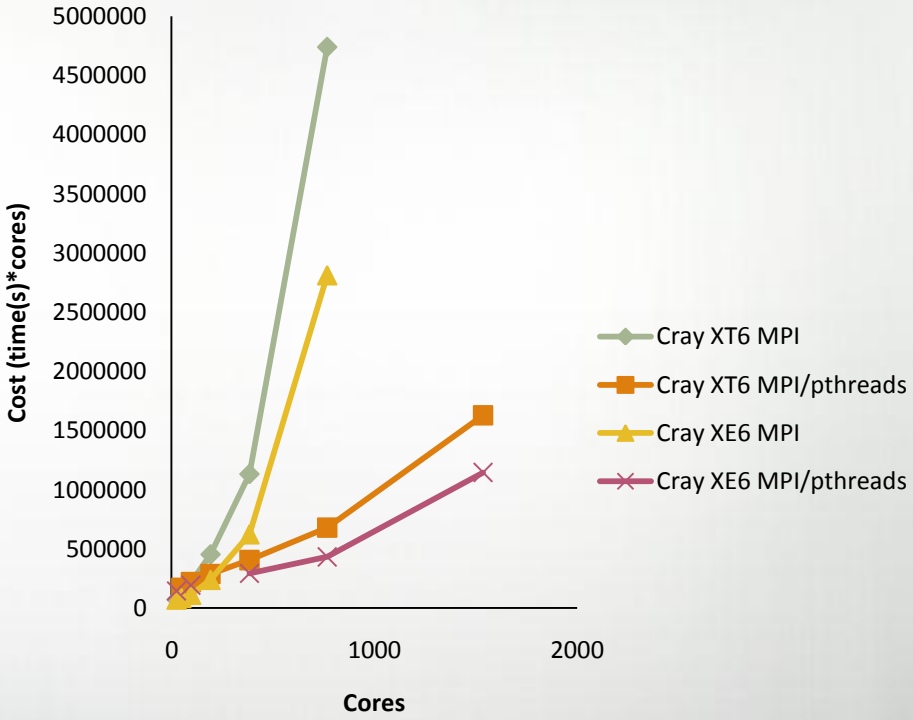
HiGEM – Climate Modelling

- Dr Len Shaffrey, University of Reading
- High resolution coupled climate model
- The CoE is collaborating with NCAS researchers to refine UM I/O performance
  - I/O servers now overlap communication and computation.
  - First stages complete
- Currently applied to a single version of the Unified Model
  - Requires a moderate effort to back port code to earlier versions of the UM
    - Large numbers of different versions of the UM being used on HECToR
- I/O server technique has been employed in other applications as well – HELIUM, CARP

GADGET3 – Computational Cosmology

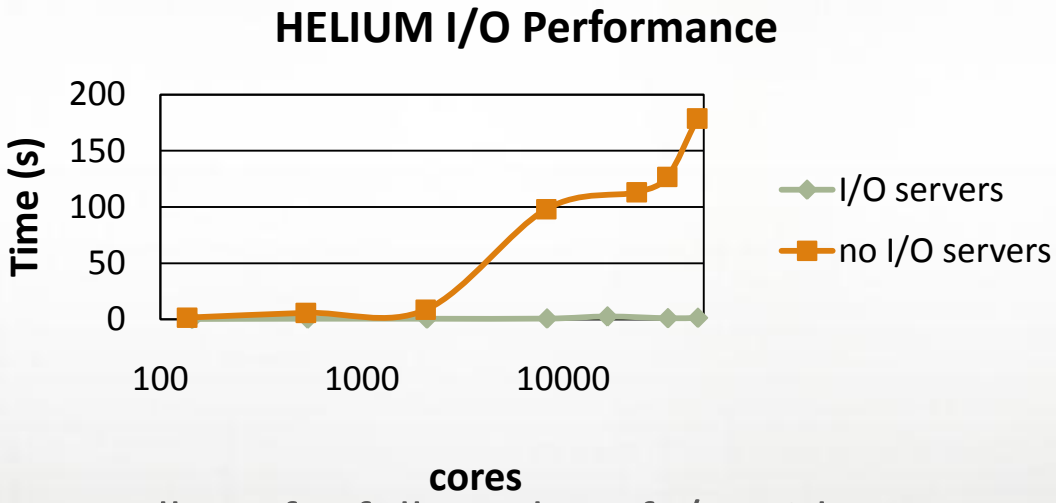
- Dr Adrian Jenkins, Prof. Richard Bower, Dr Tom Theuns - Durham University
- Dr Robert Crain – Swinburne University
- The CoE is collaborating with Institute for Computational Cosmology (ICC) researchers to better understand and improve the scaling of GADGET3.
- Code exhibits significant load imbalance with current datasets at small core counts
  - 50% of the time is spent waiting at MPI collectives on 64 cores
- CoE is looking to improve the load balancing in the code
  - Performance characterization of new hybrid MPI/pthreads code

GADGET3 scaling Cray XT6 and Cray XE6



HELIUM

- Collaboration with Prof. Ken Taylor, Queen's University Belfast
- CoE work involved implementation of an I/O server
- Work reported at the Cray User Group in Edinburgh – May 2010
  - "Optimising Application Performance using I/O Servers on Cray XT Technology"



- Use of I/O server allows for full overlap of I/O with computation and communication
  - I/O time hidden as I/O requests are offloaded to dedicated I/O servers

HELIUM

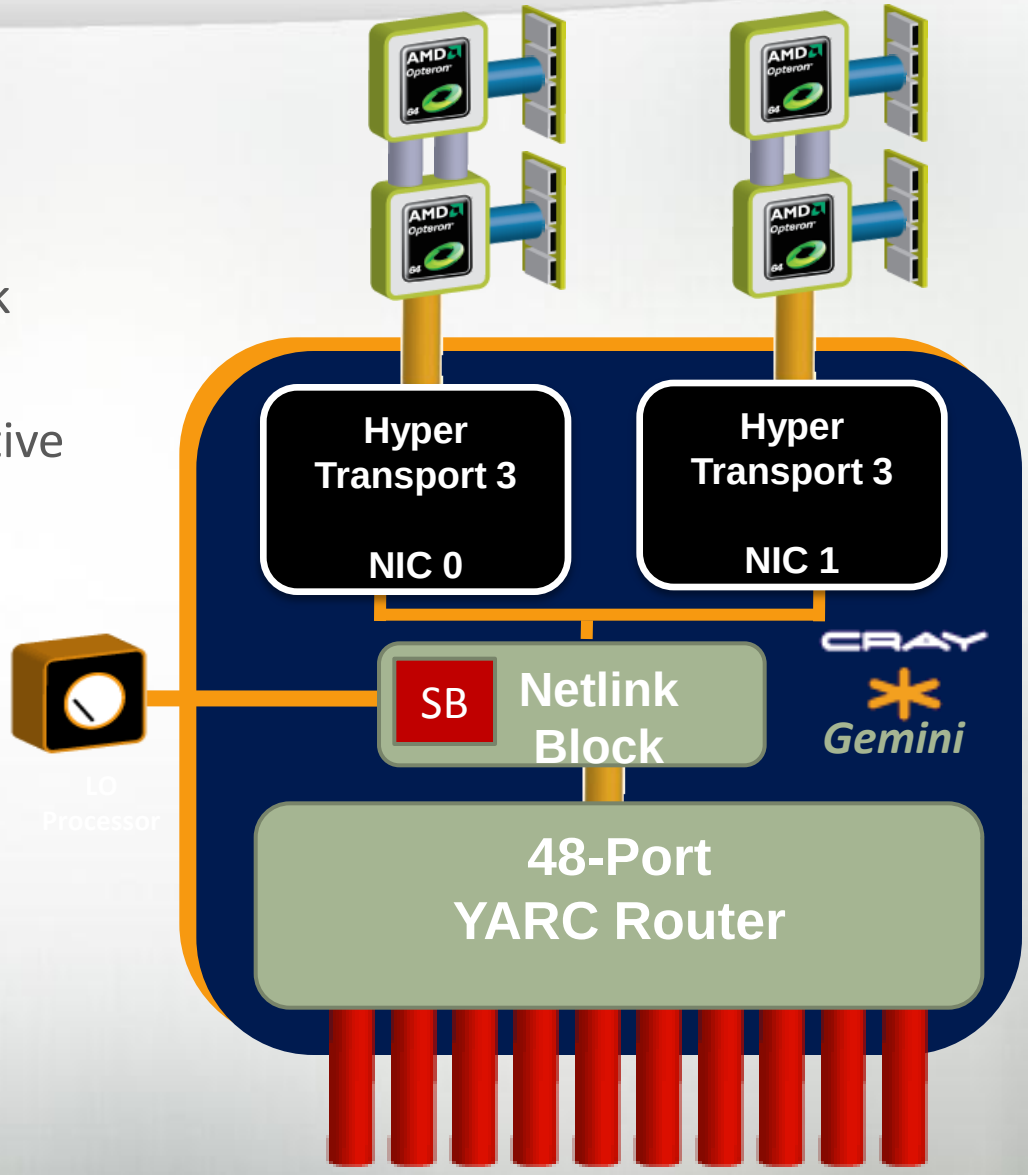
- Both of the 800 nm runs were successful.
  - Both required 16,110 cores - near the maximum available on HECToR.
  - The results confirmed present understanding of the high-energy cutoff law.
  - The results were consistent with results at 390nm, 248nm, and 195nm.

- Calculations at higher intensities are not really feasible on 16,110 cores of HECToR, and remain very much a grand challenge.

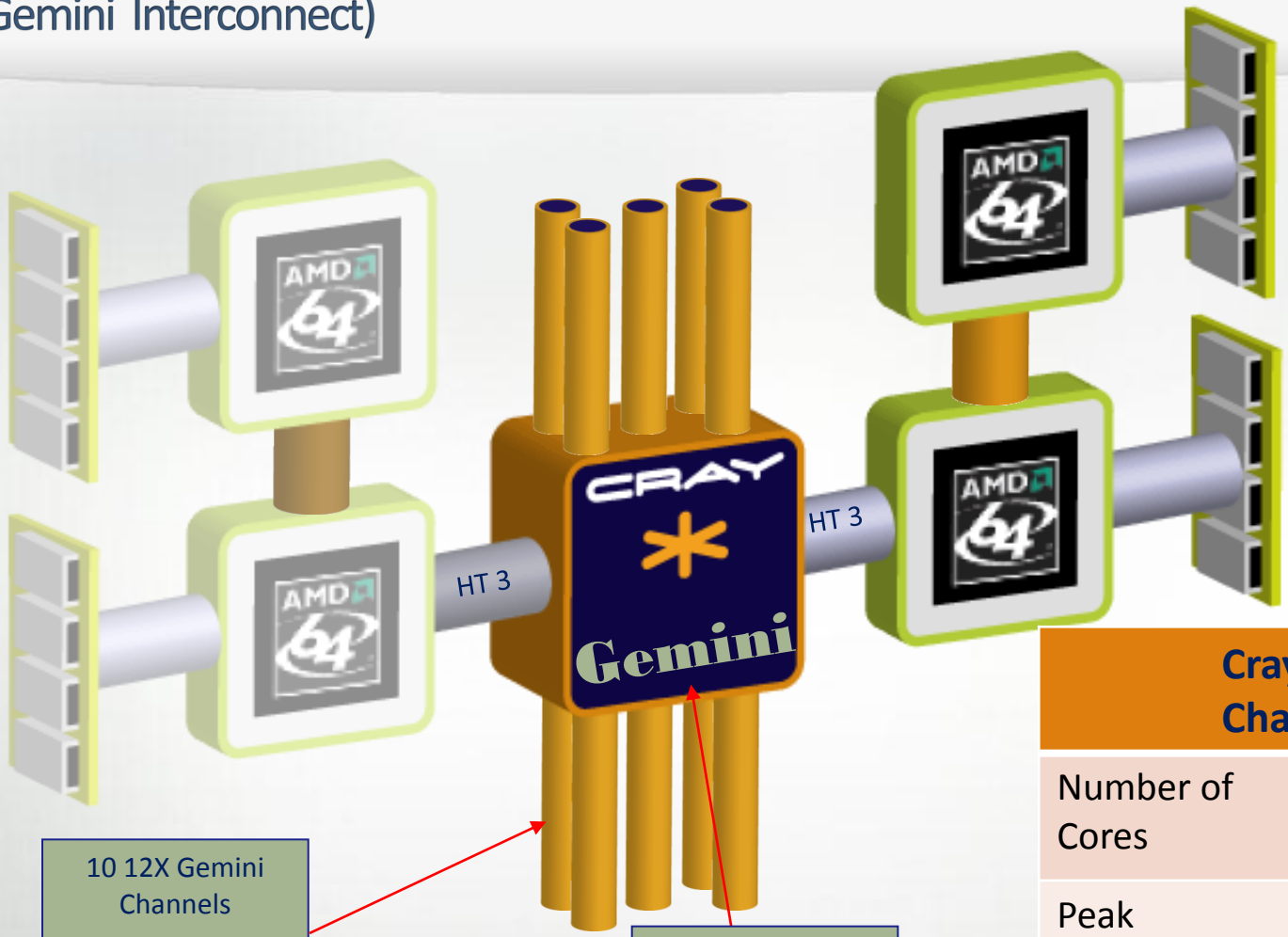
Cray XE6

Cray Gemini ASIC

- Supports 2 Nodes per ASIC
- 168 GB/sec routing capacity
- Scales to over 100,000 network endpoints
- Link Level Reliability and Adaptive Routing
- Advanced Resiliency Features
- Provides global address space
- Advanced NIC designed to efficiently support
  - MPI
  - One-sided MPI
  - Shmem
  - UPC, Coarray FORTRAN



Two Magny Cours Cray XE6 Nodes (Gemini Interconnect)



Cray XE6 Node Characteristics	
Number of Cores	16 or 24 MC
Peak Performance	210 Gflops
Memory Size	32 or 64 GB
Memory Bandwidth	85 GB/s

Gemini vs SeaStar – Topology

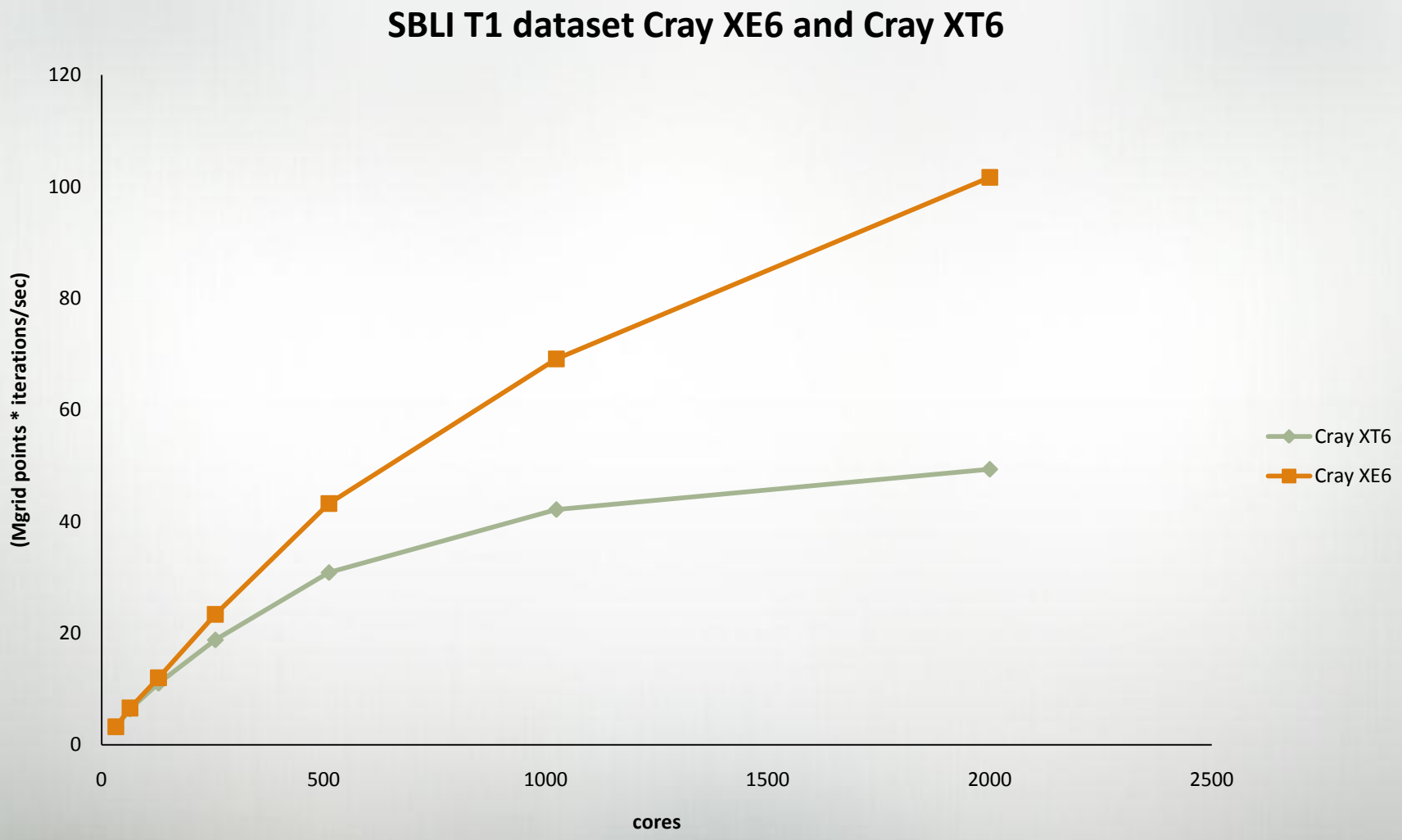
Gemini MPI Features

- Like SeaStar, Gemini has a DMA offload engine allowing large transfers to proceed asynchronously
- Gemini provides low-overhead OS-bypass features for short transfers
  - MPI latency around 1.4us in current release
  - NIC provides for many *millions* of MPI messages per second – 20 times better than Seastar
- Much improved one-sided communication mechanism with hardware support
- AMOs provide a faster synchronization method for barriers
- Gemini supports adaptive routing, which
  - Reduces problems with network hot spots
  - Allows MPI to survive link failures
- Much improved injection bandwidth – 6 GB/s user data injection with HT3 link

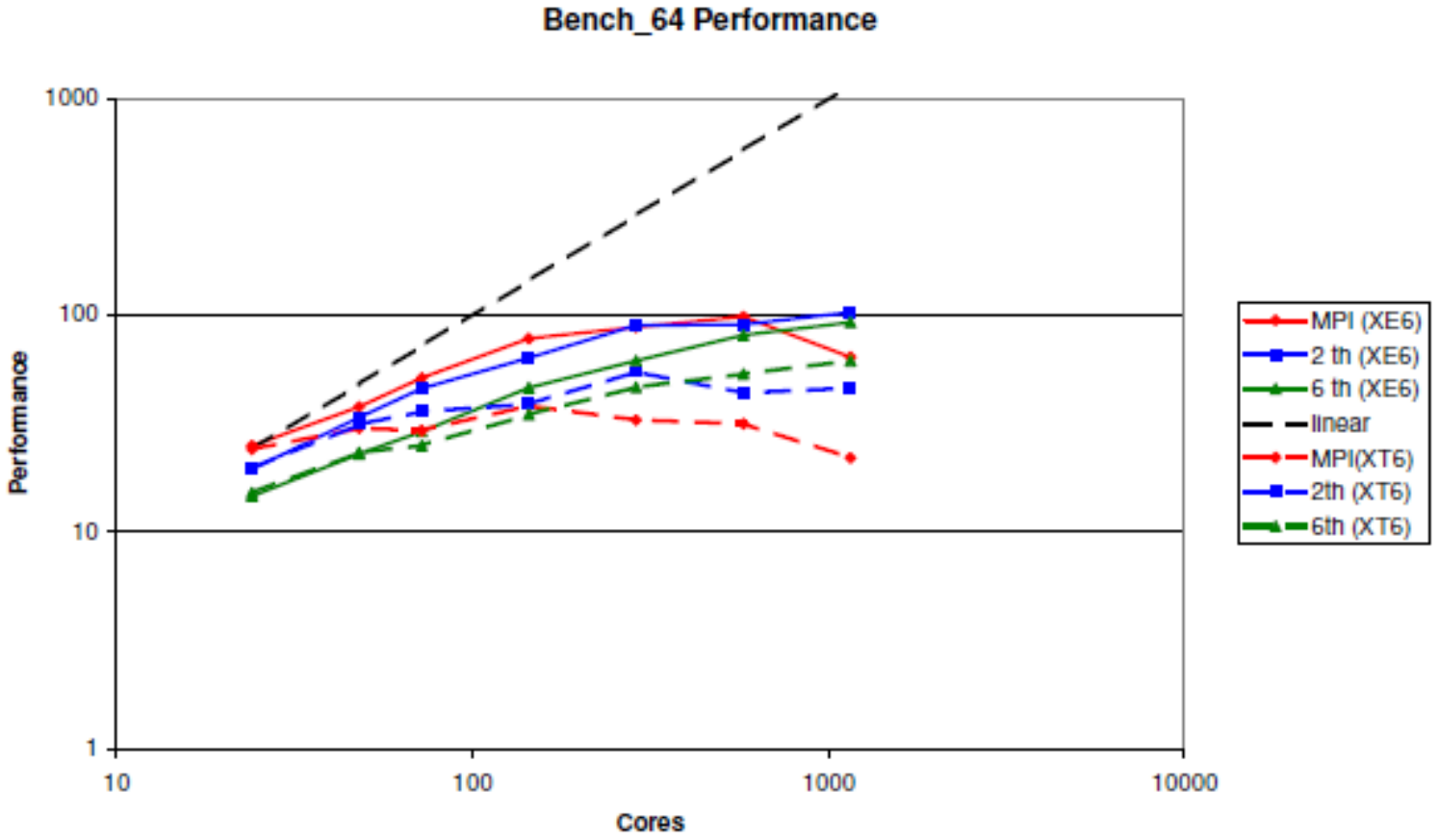
Gemini Advanced Features

- Globally addressable memory provides efficient support for UPC, Co-array FORTRAN, Shmem and Global Arrays
  - Cray compiler targets this capability directly
- Pipelined global loads and stores
  - Allows for fast irregular communication patterns
- Atomic memory operations
  - Provides fast synchronization needed for one-sided communication models

SBLI Cray XT6 vs Cray XE6

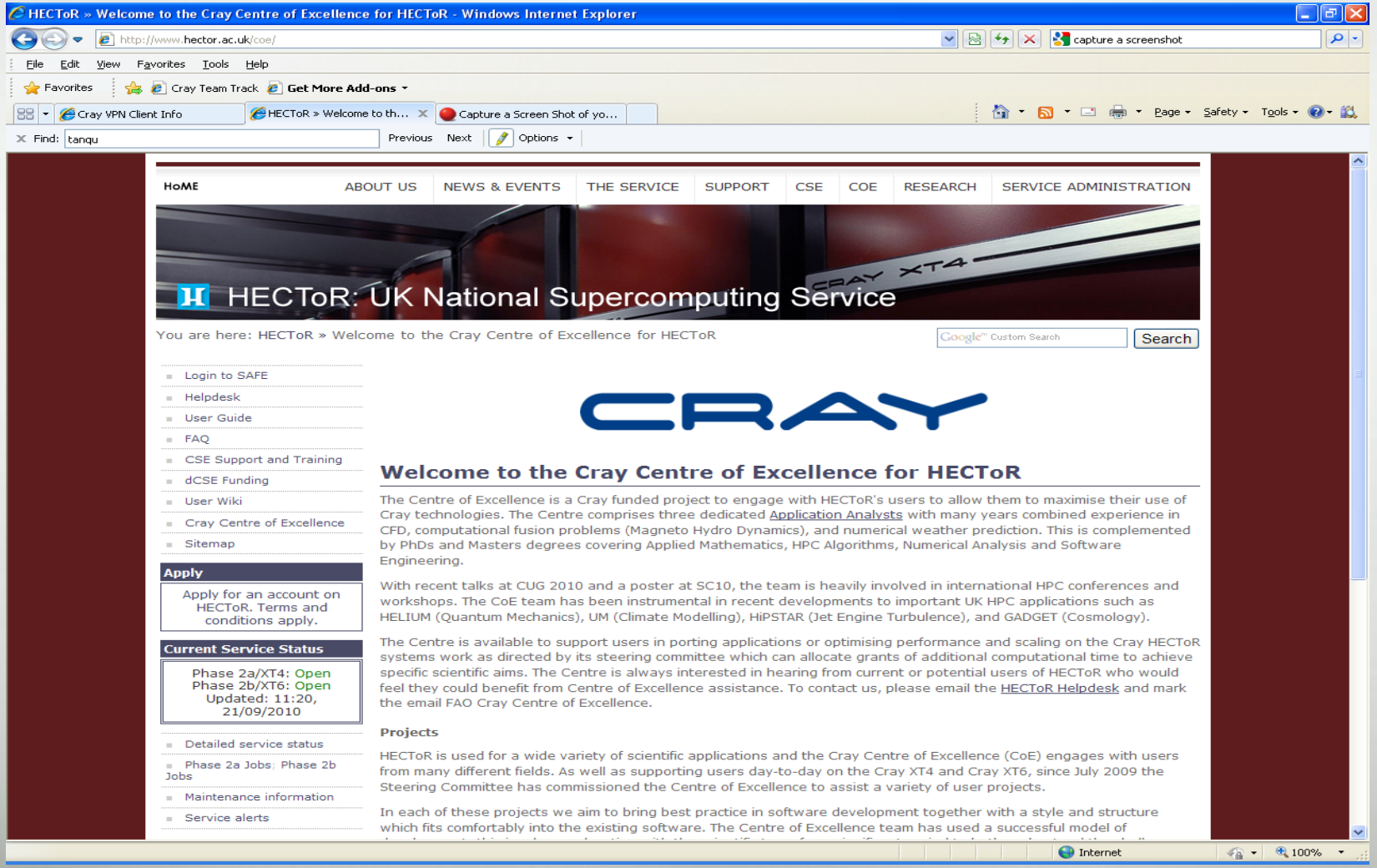


CP2K – courtesy of Iain Bethune, EPCC



Cray Centre of Excellence for HECToR Website

- http://www.hector.ac.uk/coe/



Cray Exascale Research Initiative Europe

- Launched December 2009
  - Initial research partners are EPCC and CSCS
  - Cray increased Edinburgh based team with Alistair Hart and Harvey Richardson
- Exploring how real applications can exploit future Exascale architectures, specifically:
  - Programming models for PGAS languages
  - Programming GPU accelerators
  - Improved algorithms for FFTs
  - Network and I/O profiling
- Strong interactions with Cray R&D

CoE/Exascale Training and Conferences

- CUG 2010 24-27 May
  - Using I/O servers to improve performance on Cray XT technology– Tom Edwards and Kevin Roy
- NAIS HPC Workshop 23-24 June
  - PGAS Programming with UPC and Fortran coarrays – Jason Beech-Brandt and Harvey Richardson
- DEISA Training Course 14-16 September
  - Performance Optimization on Cray XT Systems
  - Coarray Fortran Tutorial
- Hybrid Manycore Programming and Accelerators in High Performance Computing 22 October
  - Challenges in Programming Multicore and Accelerated Nodes – John Levesque and Alistair Hart
- SC10
  - Poster - Performance Tuning of Scientific Applications on HPC Systems
  - Poster – GPU Acceleration of Scientific Applications
  - Tutorial – Introduction to Coarray Fortran
- EPCC MSc project supervision
  - Global Arrays and NWChem Tuning on the Cray XT – Jason Beech-Brandt
  - Co-arrays for sparse matrix methods – Tom Edwards
- Visits to NAG and Daresbury and various user sites (Reading, Southampton, Durham, Cambridge amongst others)

- Questions?

- Jason Beech-Brandt: jason@cray.com
- Tom Edwards: tedwards@cray.com
- Kevin Roy: kroy@cray.com