HECToR

The new UK National High Performance Computing Service

Dr Mark Parsons
Commercial Director, EPCC
m.parsons@epcc.ed.ac.uk
+44 131 650 5022

Summary

- Why we need supercomputers
- The HECToR Service
- Technologies behind HECToR
- Who uses HECToR
- The challenges facing supercomputing
- A sneak preview
- Concluding remarks

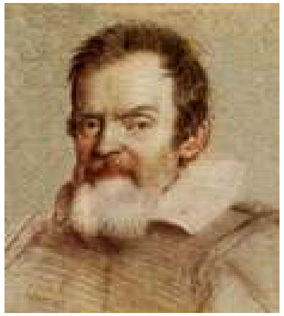
Many thanks: Mike Brown, Alan Gray, Fiona Reid and Alan Simpson – EPCC

Jason Beech-Brandt – Cray Inc.

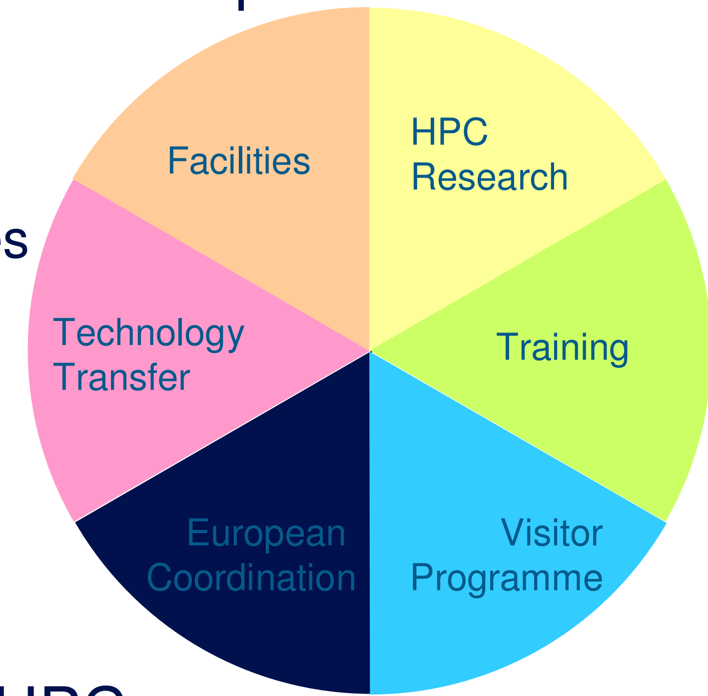
Unless otherwise noted this talk is Copyright (c) The University of Edinburgh 2008

A brief history of science

- Science has evolved for 2,500 years
- First there was THEORY
  – Led by the Greeks in 500BC
- Then there was EXPERIMENT
  – Developed in Europe from 1600AD
- Since 1980s we have also had SIMULATION
  – Edinburgh can rightfully claim to be world leading
- We use simulation for problems that are too big, too small, too distant, too quick, too slow to experiment with
- Computational science has driven high performance computing for the past 30 years

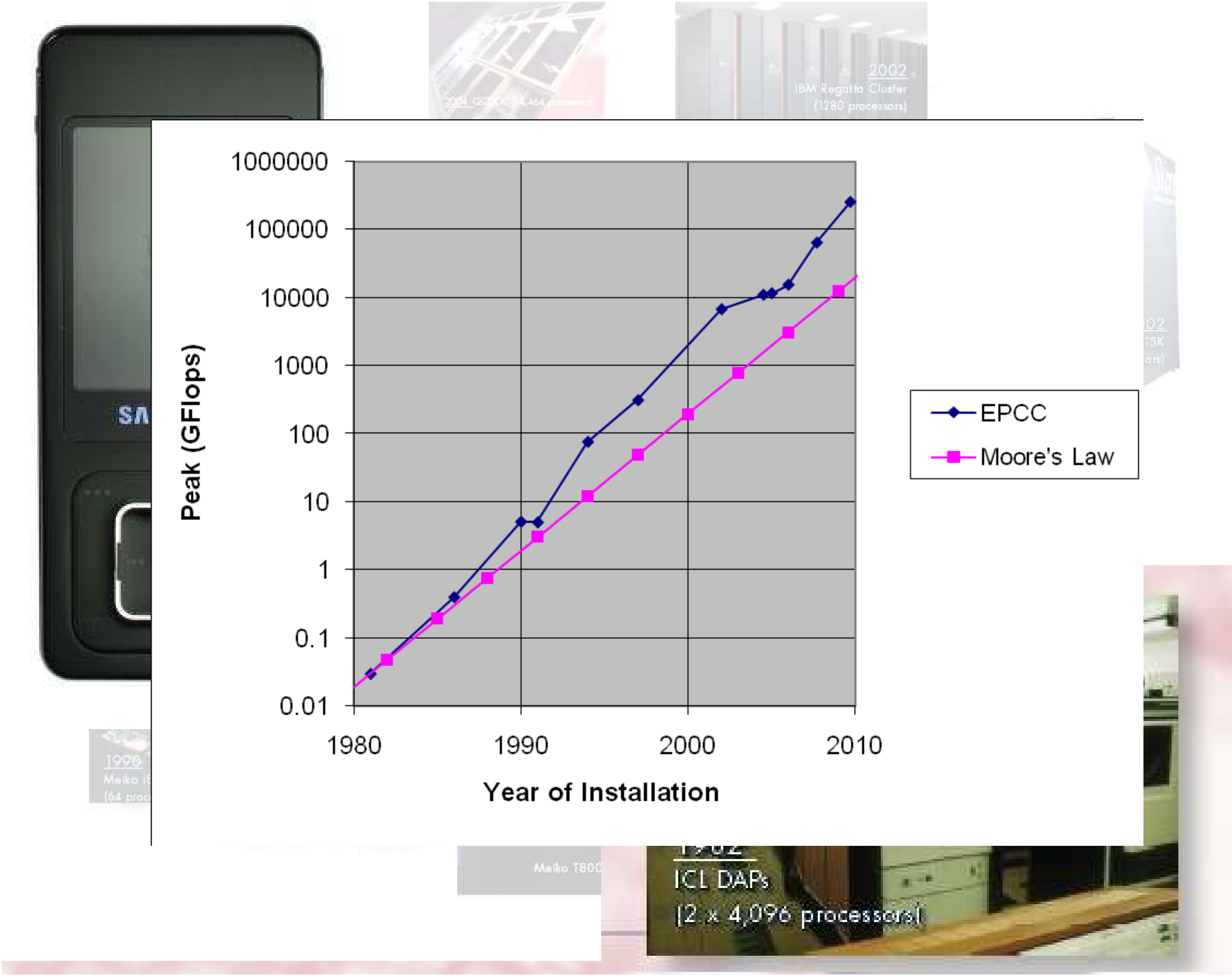


- The University of Edinburgh founded EPCC in 1990 to act as the focus for its interests in simulation
- Today, EPCC is the leading centre for computational science in Europe
  - 80 permanent staff
  - Managing all UK national HPC facilities
  - Work 50:50 academia and industry
- Aim is to rival the big US centres
  - eg. NCSA at the University of Illinois
- In 2007 we won the contract to host the new UK National Service HPC service - HECToR



20 years of hardware

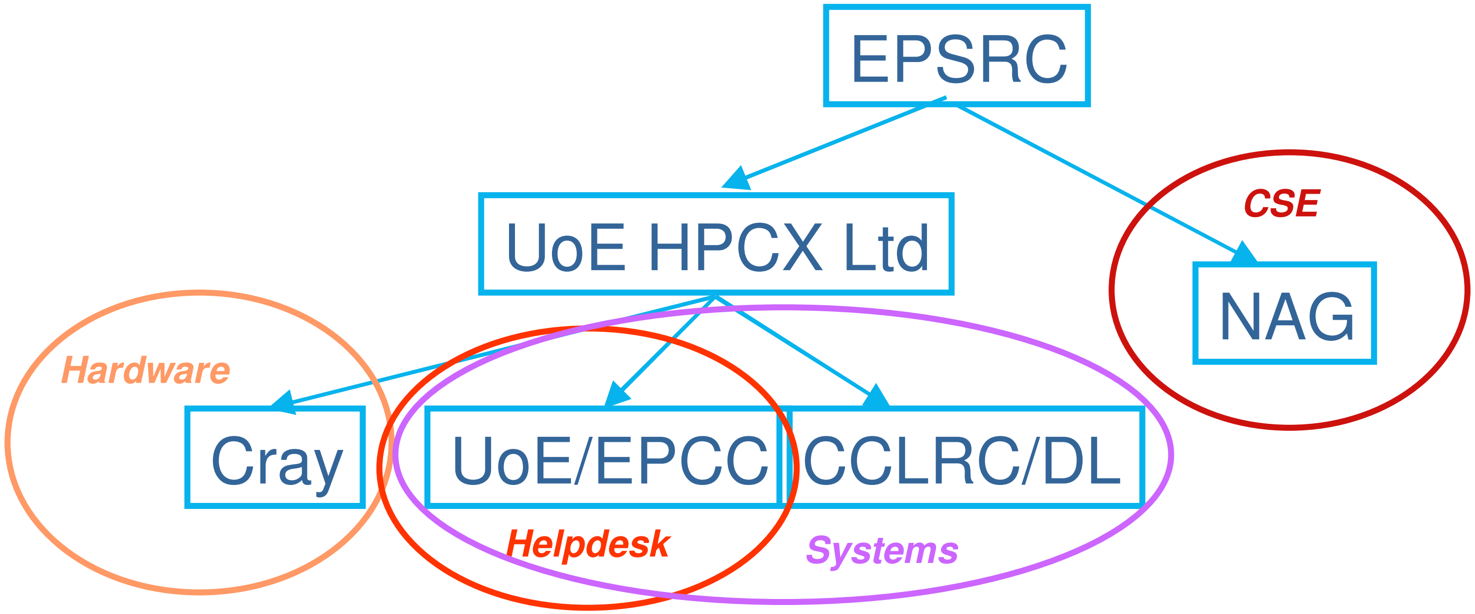
20 years of hardware



The HECToR Service

- HECToR: **H**igh **E**nd **C**omputing **T**erascale **R**esource
- Procured for UK scientists by Engineering and Physical Sciences Research Council – EPSRC
- Competitive process involving three procurements
  - Hardware – *CRAY*
  - Accommodation and Management – *UOE HPCX LTD*
  - Computational Science and Engineering Support – *NAG*
- EPCC won the A&M procurement through its company – UoE HPCx Ltd
- HECToR is located at The University of Edinburgh

Contractual Structure and Roles



- UoE HPCx Ltd already holds contract for HPCx service
  - Wholly-owned subsidiary of University of Edinburgh
- UoE HPCx Ltd awarded main contract for HECToR Service Provision
  - Runs from 2007 to 2013
  - Subcontracts: Hardware (Cray), Helpdesk (EPCC), Systems (EPCC+DL)
- CSE support from NAG is separate
- Total contract value is around £115 million

HECToR Installation Timeline

February 2007

Signing of HECToR Contracts

March 2007

Edinburgh: laying foundations for new plant room

Chippewa Falls, WI: XT4 Cabinets being assembled

HECToR Installation Timeline

April 2007

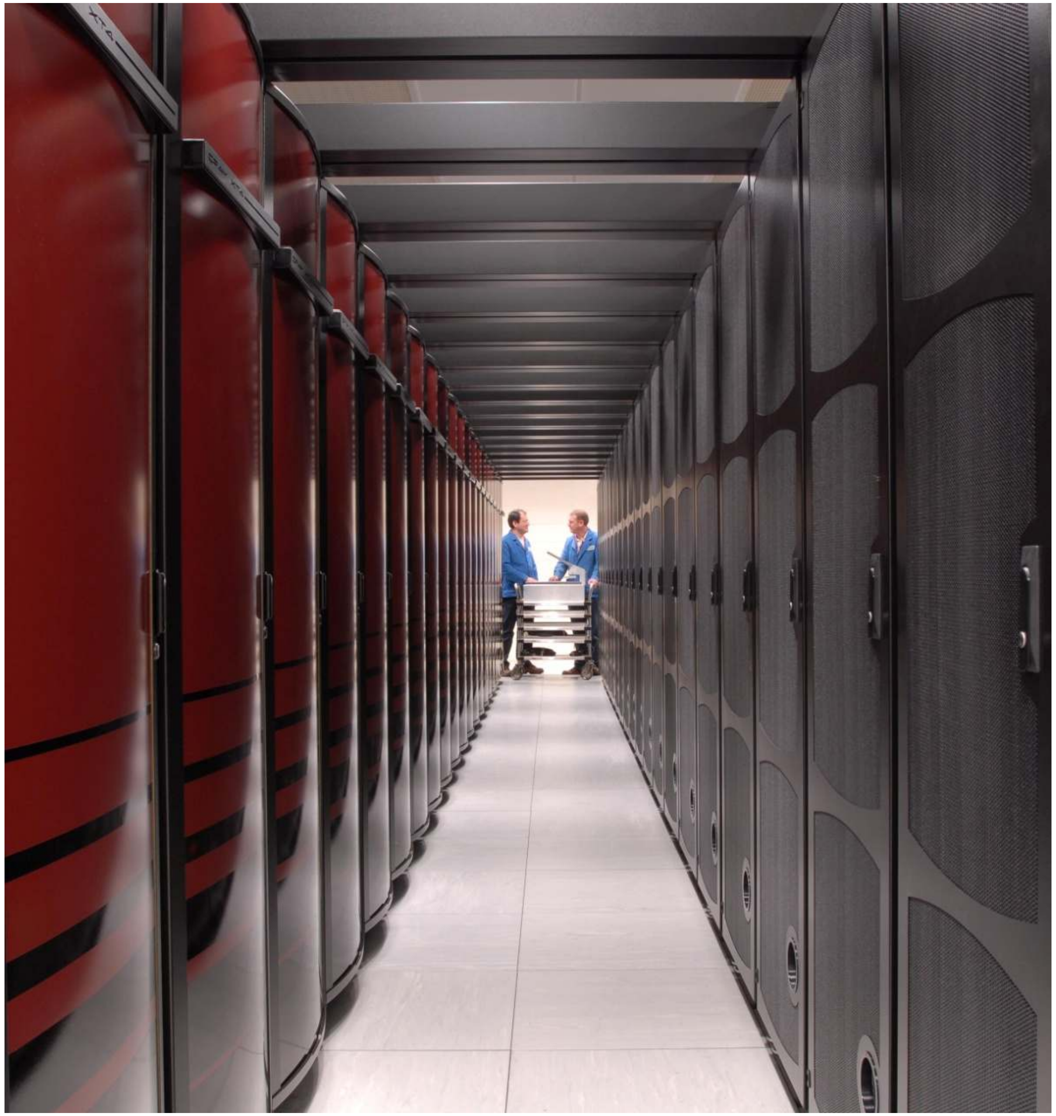
Edinburgh: new building in progress

Edinburgh: Test and Development System (one XT4 cabinet) installed

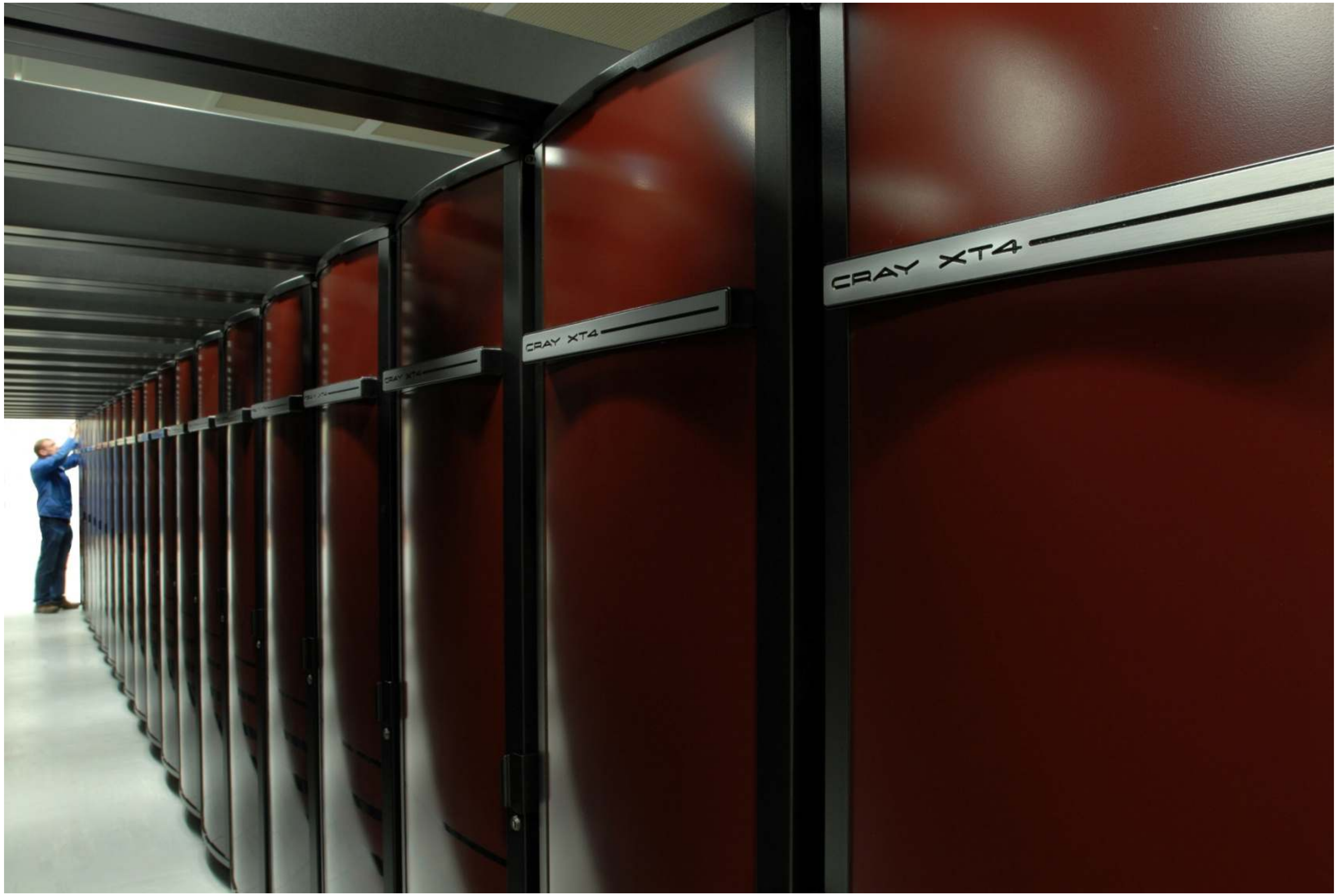
August 2007

Edinburgh: Full 60 Cabinet System installed

HECToR at the ACF



HECToR at the ACF



Advanced Computing Facility

- Constructed 1976 for the University of Edinburgh
  - 1 x 600 m² Computer Room
  - 24-stage DX-based cooling servicing the room through 4 vast walk-in air-handling units
  - "conventional" downflow system
- Refurbished 2004 as the Advanced Computing Facility
  - 2 x 300 m² Computer Rooms (one active, one empty concrete shell)
  - all new chilled-water based plant services, with capacity of 1.2MW
- Major expansion 2007 for HECToR
  - 2nd Computer Room brought into operation
  - new-build external plant room to support massive uplift in required capacity
  - new HV electrical provision (up to 7MW)

Power and Cooling

Power and Cooling

Two national services

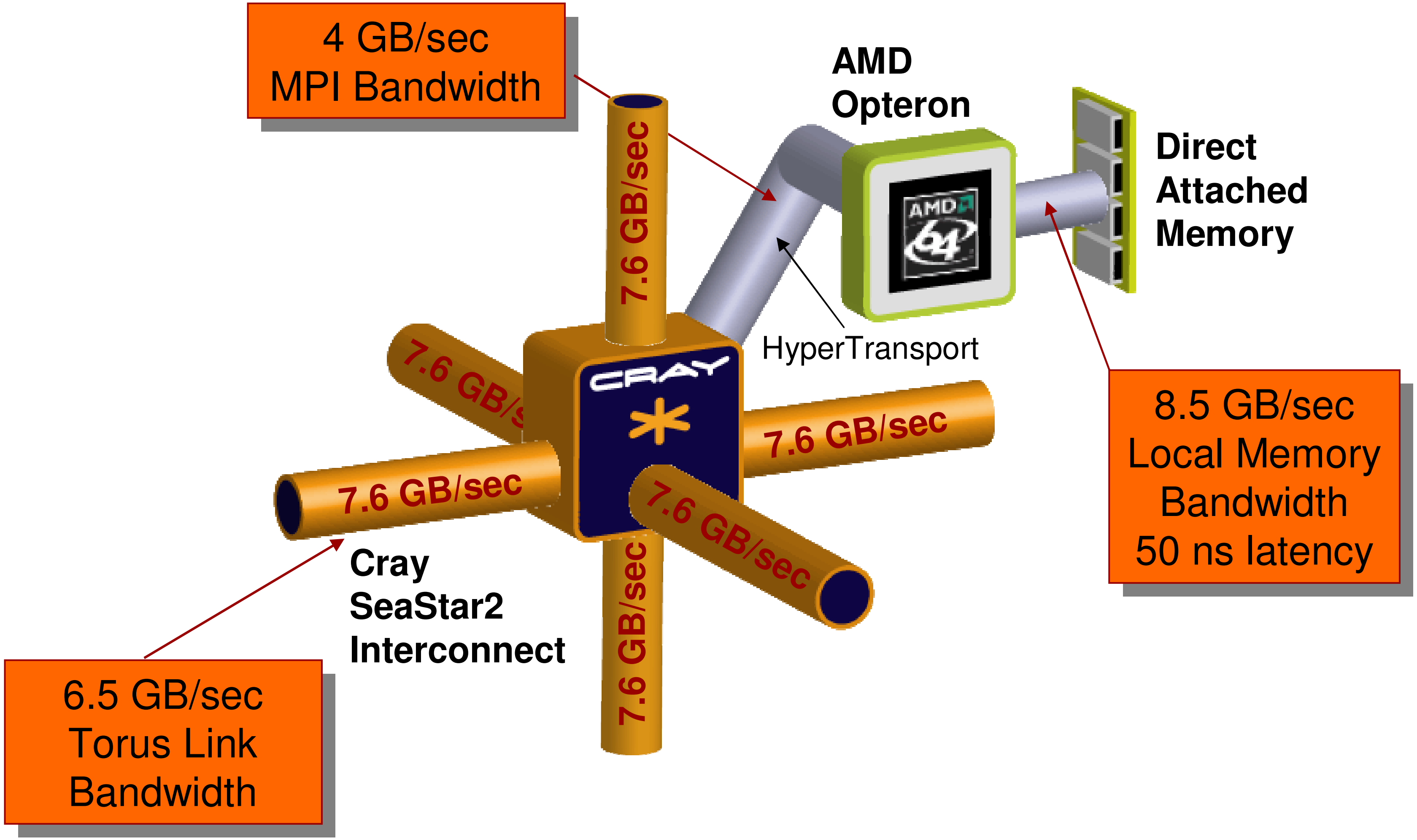
- HPCx (Phase 3): 160 IBM e-Server p575 nodes
  - SMP cluster, 16 Power5 1.5 GHz cores per node
  - 32 GB of RAM per node (2 GB per core)
  - 5TB total RAM
  - IBM HPS interconnect (aka Federation)
  - 12.9 TFLOP/s Linpack, No 101 on top500

- HECToR (Phase 1): Cray XT4
  - MPP, 5664 nodes, 2 Opteron 2.8 GHz cores per node
  - 6 GB of RAM per node (3 GB per core)
  - 33TB total RAM
  - Cray Seastar2 torus network
  - 54.6 TFLOP/s Linpack, No 17 on top500

The old and the new (cont)

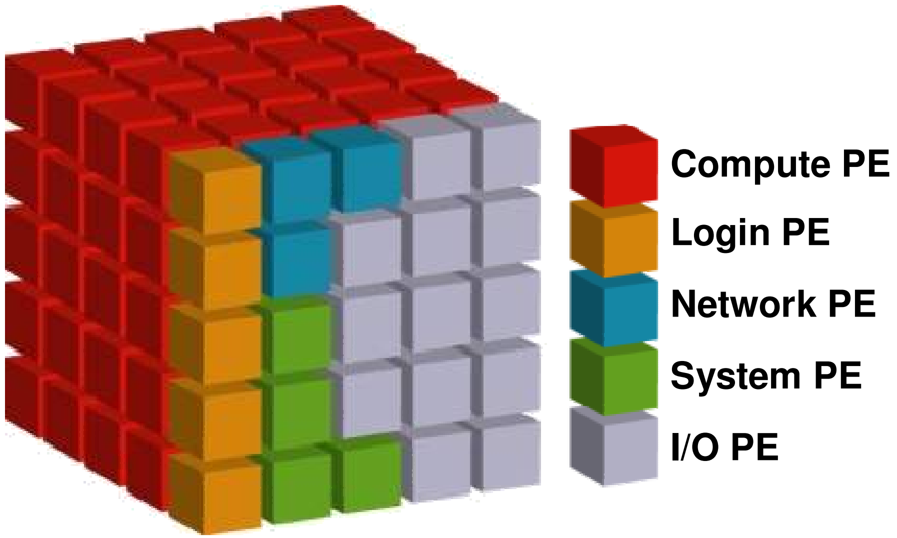
	HPCx	**HECToR**
Chip	IBM Power5 (dual core)	AMD Opteron (dual core)
Clock	1.5 GHz	2.8 GHz
FPUs	2 FMA	1 M, 1 A
Peak Perf/core	6.0 GFlop/s	5.6 GFlop/s
cores	2560	11328
Peak Perf	15.4 TFLOP/s	63.4 TFLOP/s
Linpack	12.9 TFLOP/s	54.6 TFLOP/s

The Cray XT4 Processing Element



Copyright (c) 2008 Cray Inc.

Scalable Software Architecture: UNICOS/lc



Service Partition

Specialized Linux nodes

- Microkernel on Compute PEs, full featured Linux on Service PEs.
- Service PEs specialize by function
- Software Architecture eliminates OS “Jitter”
- Software Architecture enables reproducible run times
- Large machines boot in under 30 minutes, including filesystem

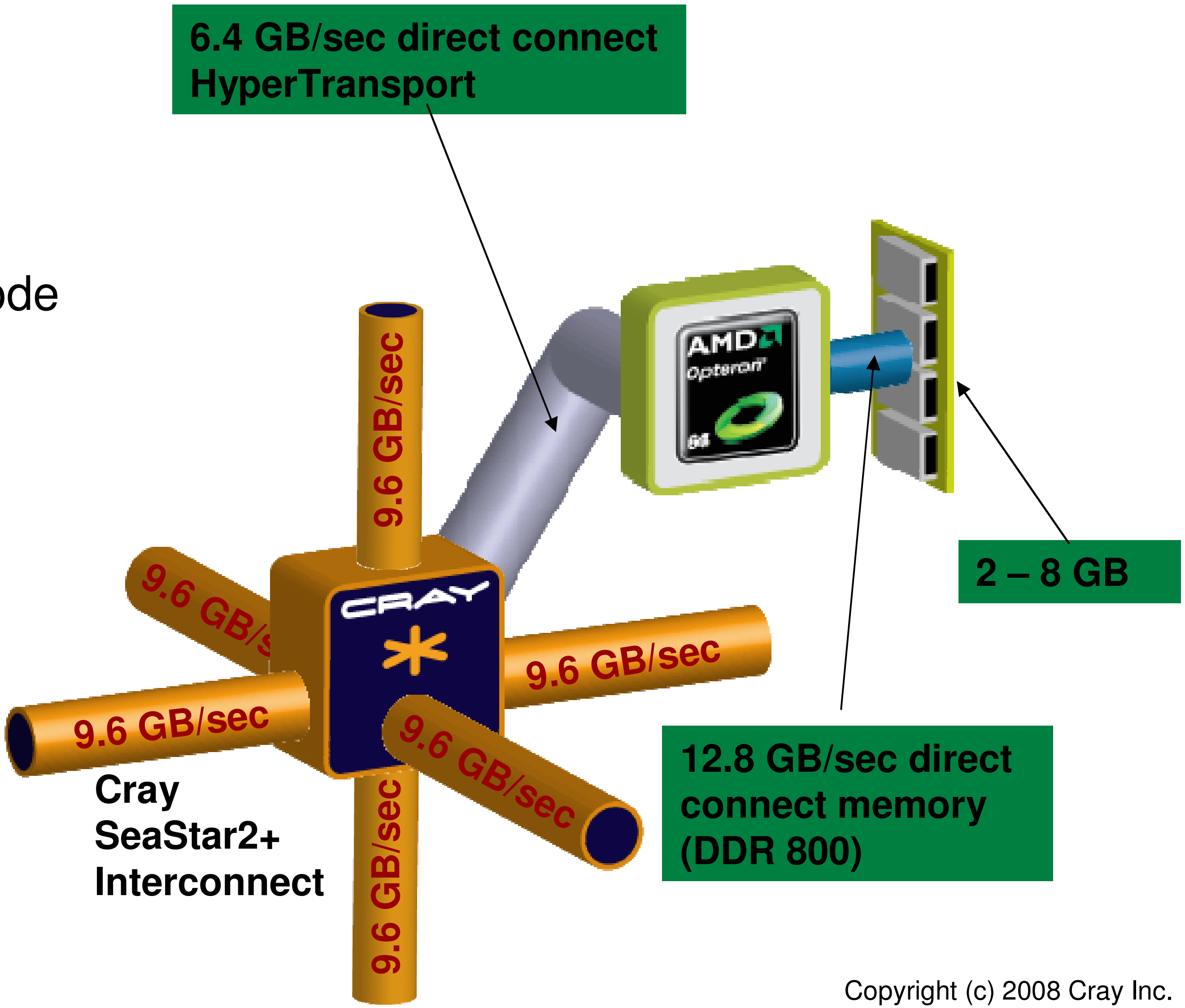
Copyright (c) 2008 Cray Inc.

Technology refreshes

- Cray have 4-year contract for hardware provision
  - Plus possible extension for years 5 and 6
- Phase 1 (accepted: September 2007):
  - 60TFlop Cray XT4
- Vector system (installed last week)
  - 2TFlop Cray X2 vector system (a “BlackWidow”)
- Phase 2 (installation: Summer 2009):
  - ~60Tflop Cray XT4 (quadcore upgrade)
  - ~200TFlop Cray (tba)
- Phase 3 (installation: Summer 2011):
  - technology supplier subject to future tender
  - anticipate infrastructure requirements approx as per Phase 2

Cray XT4 Quadcore Node

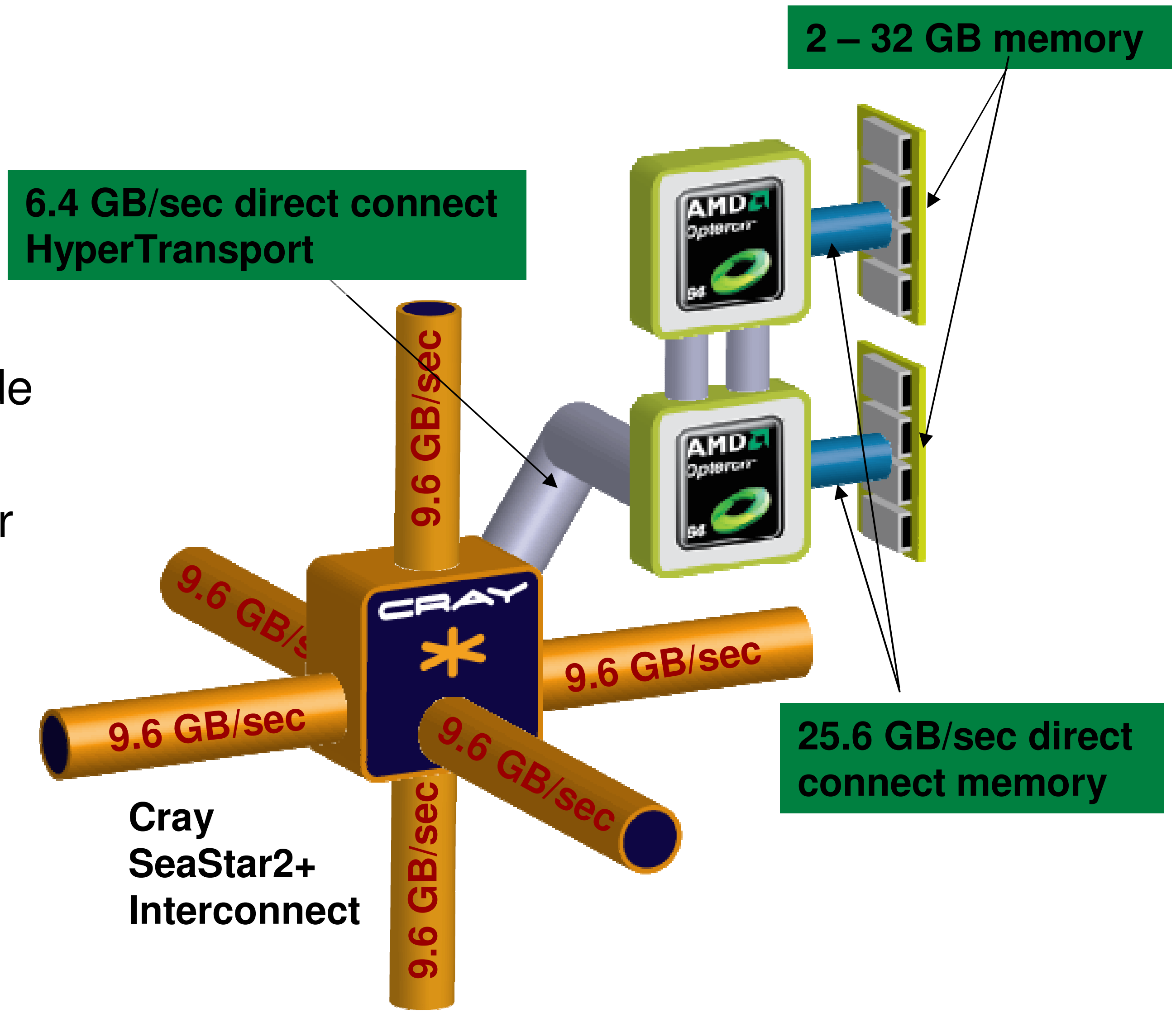
- 4-way SMP
- >35 Gflops per node
- Up to 8 GB per node
- OpenMP Support within socket



Copyright (c) 2008 Cray Inc.

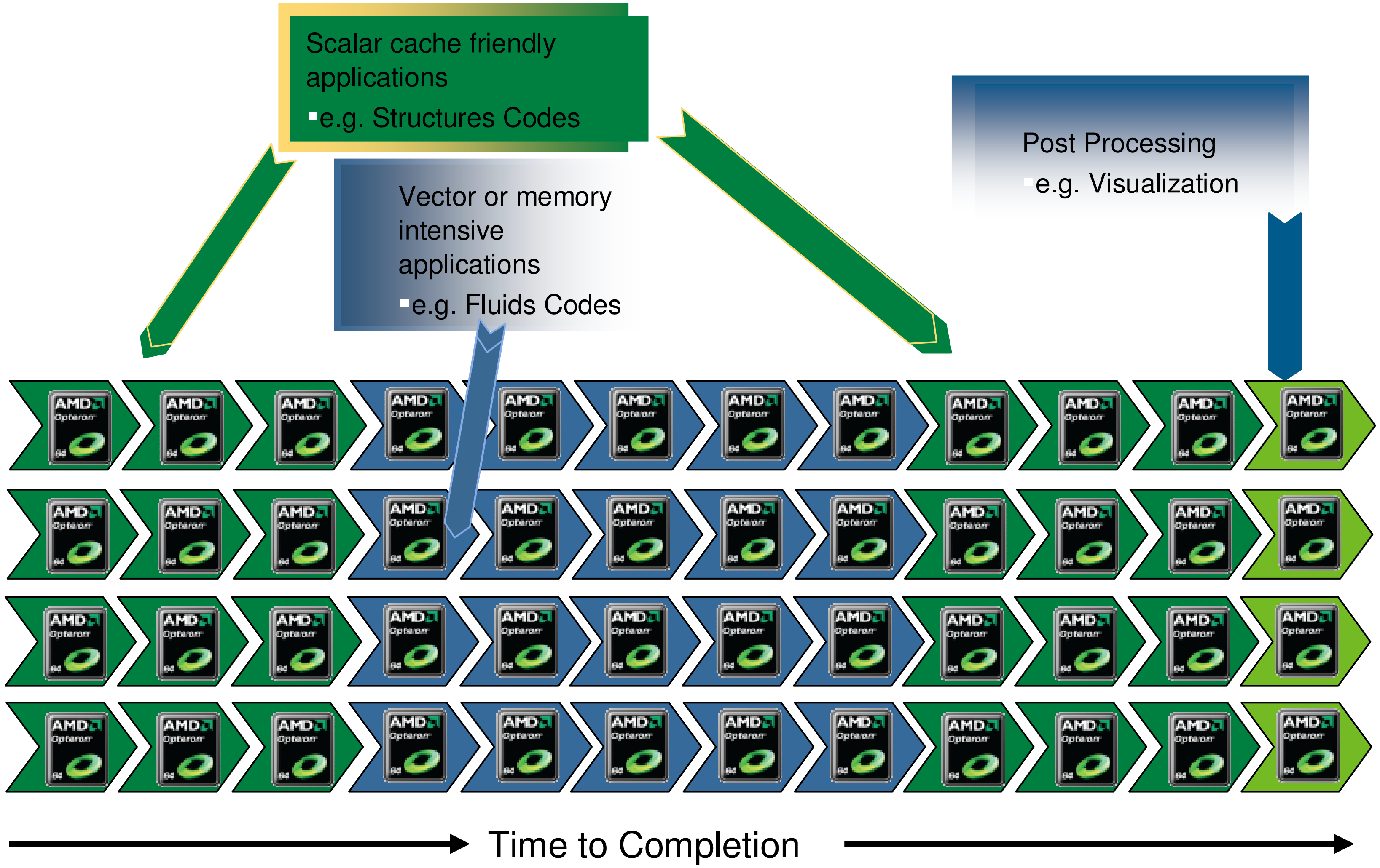
Cray XT5 Quadcore Node

- 8-way SMP
- >70 Gflops per node
- Up to 32 GB of shared memory per node
- OpenMP Support



Copyright (c) 2008 Cray Inc.

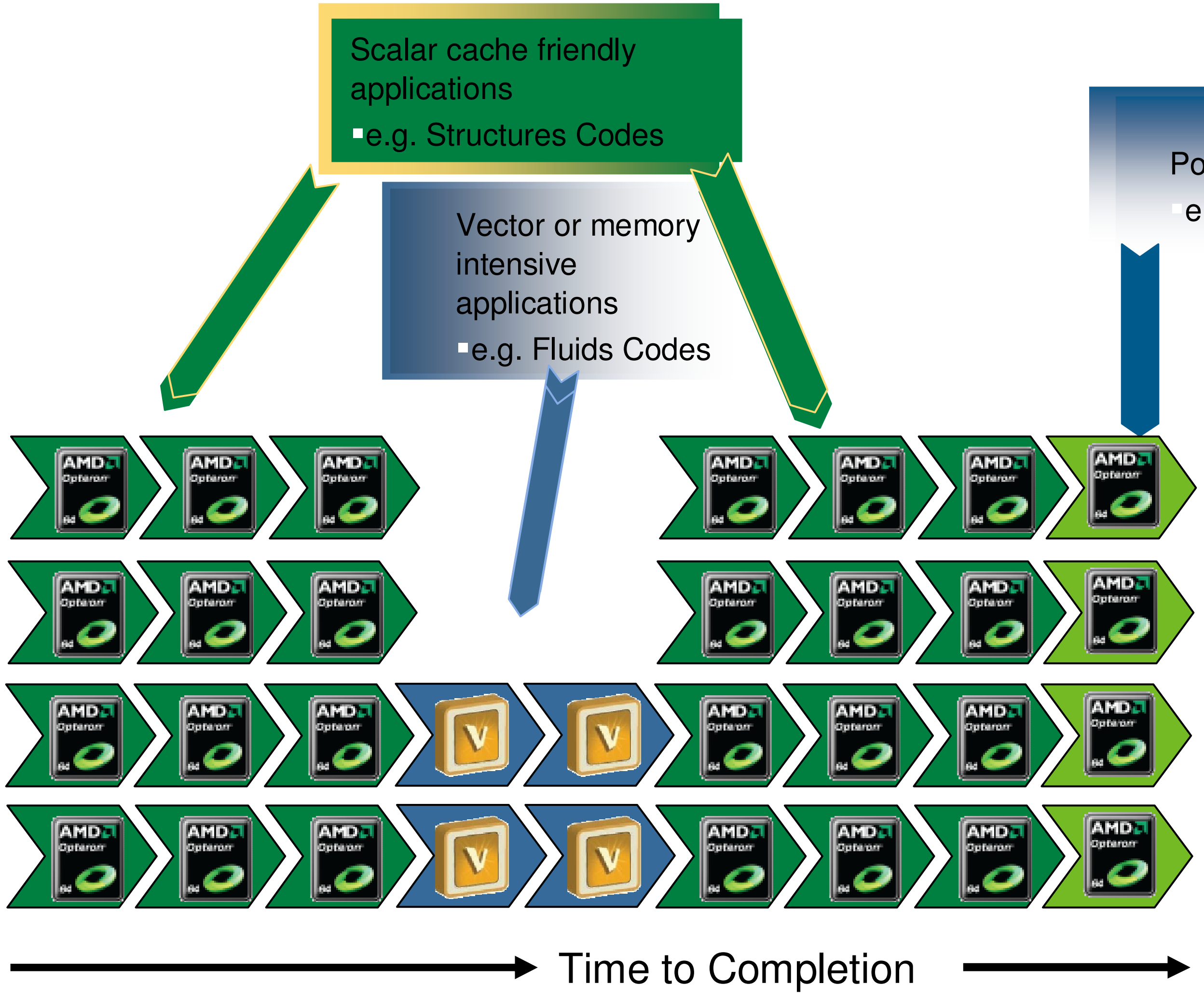
First, a workflow within a homogeneous environment



Copyright (c) 2008 Cray Inc.

Hybrid Systems

Now, the same workflow within a heterogeneous environment



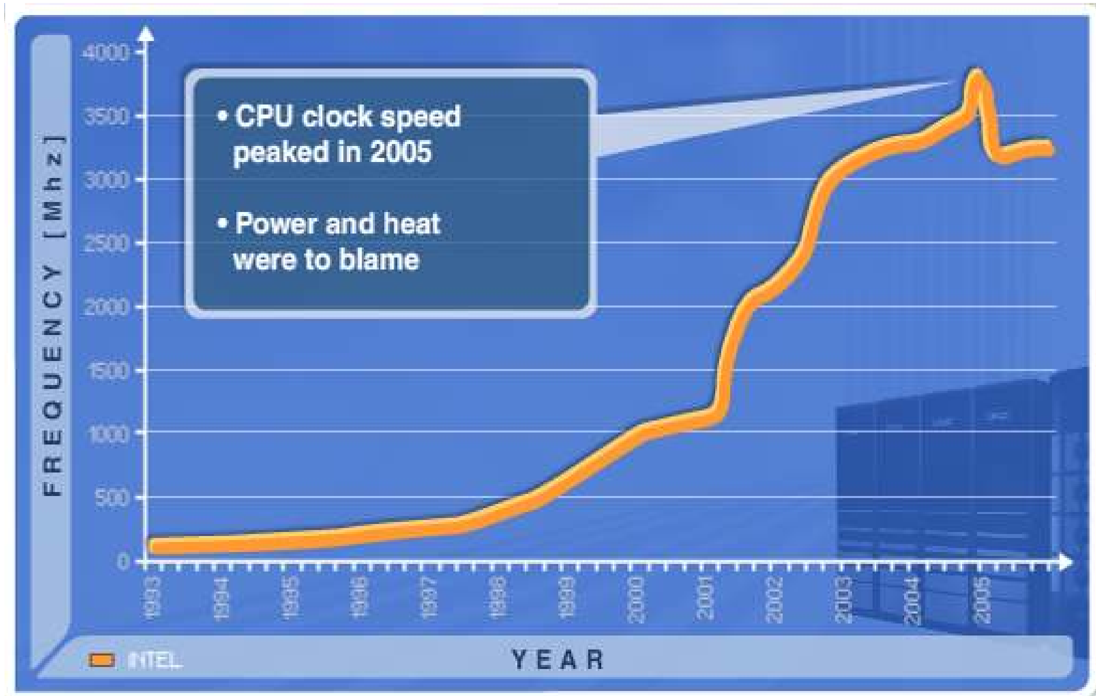
Improved time to solution

Copyright (c) 2008 Cray Inc.

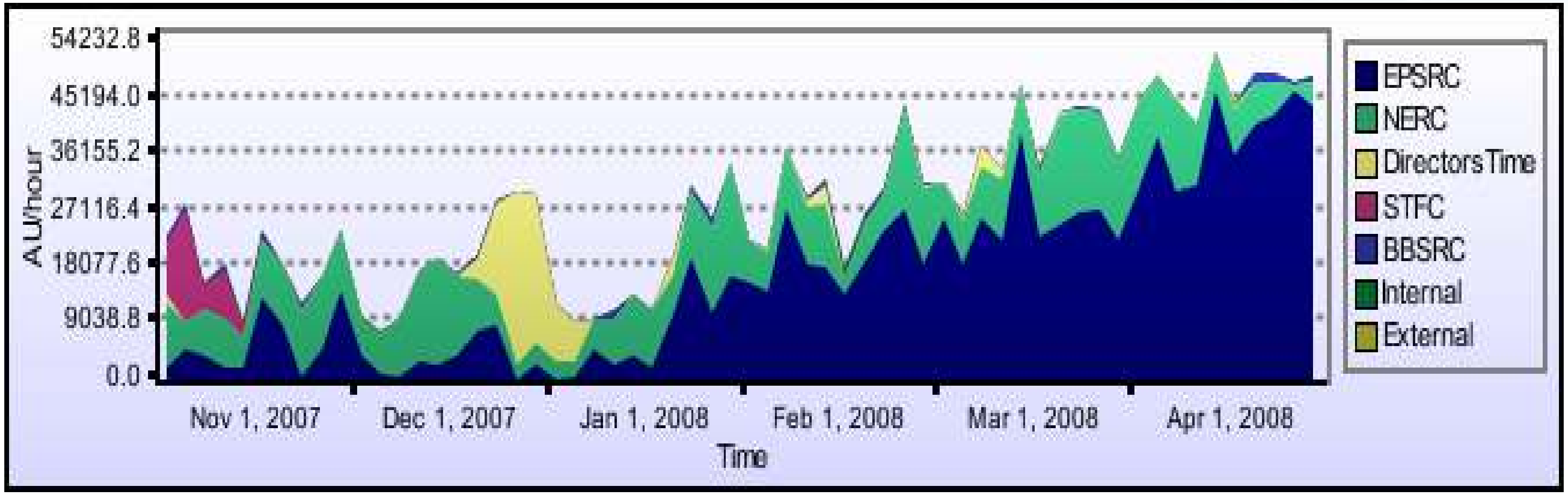
Slide 24

HECToR the Hybrid

- With the addition of the X2 last week - HECToR is Cray's first commercial hybrid system worldwide
- Clock speed, memory bandwidth, heat and power issues are driving people to look at new HPC solutions

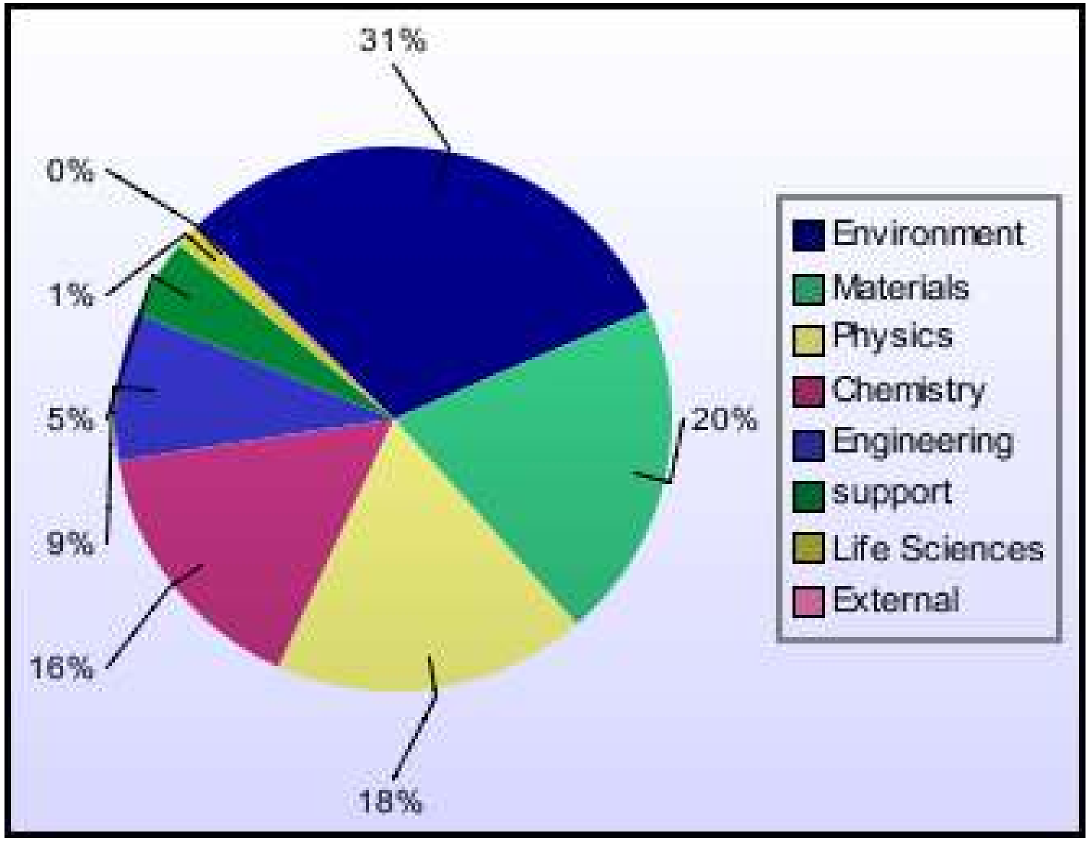


Who uses HECToR?

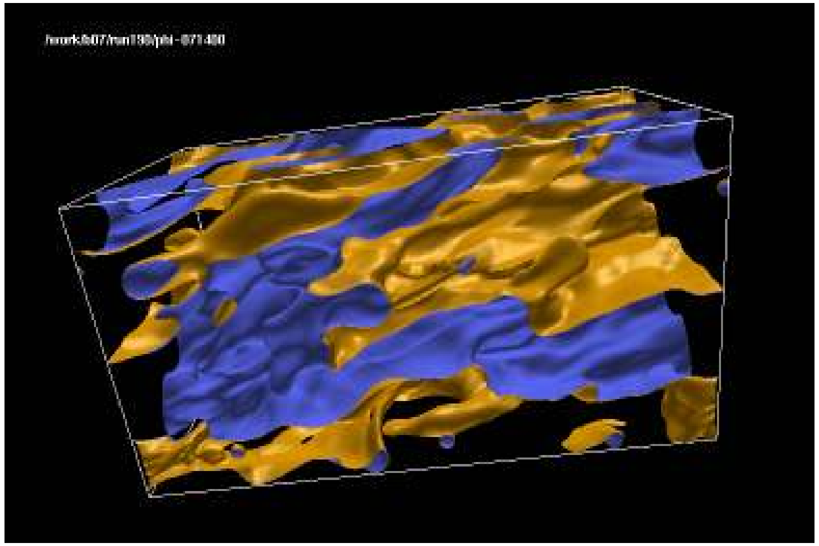


- Early user service opened in September 2007
- Full service opened on 15th October 2007
- Now have over 400 users with around 84% utilisation
- A wide variety of scientific consortia use the system
- Industry use now beginning

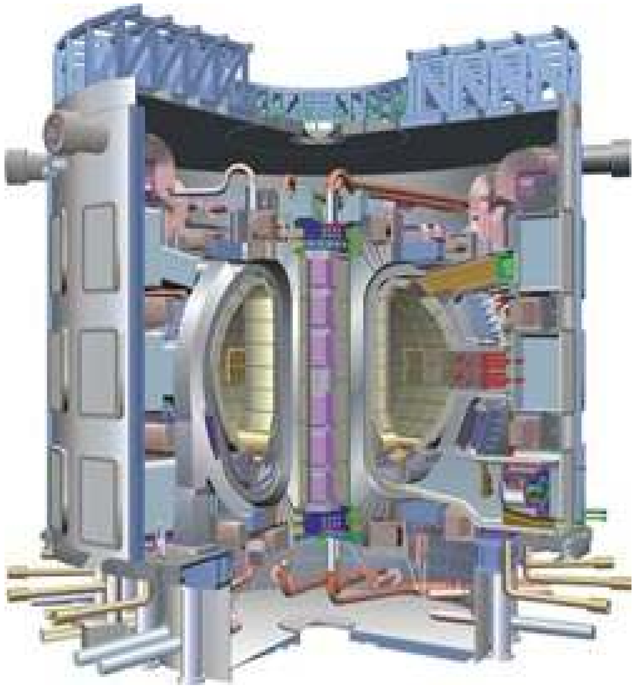
Who uses HECToR?



Fluid dynamics - Ludwig



- Ludwig
  - Lattice Boltzmann code for solving the incompressible Navier-Stokes equations
  - Used to study complex fluids
  - Code uses a regular domain decomposition with local boundary exchanges between the subdomains
  - Two problems considered, one with a binary fluid mixture, the other with shear flow

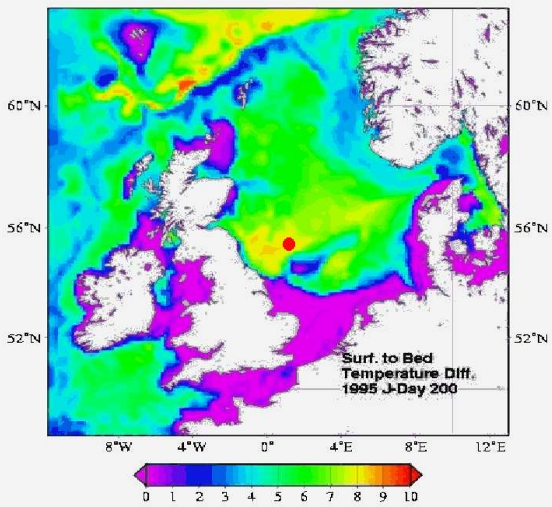


ITER tokamak reactor

(www.iter.org)

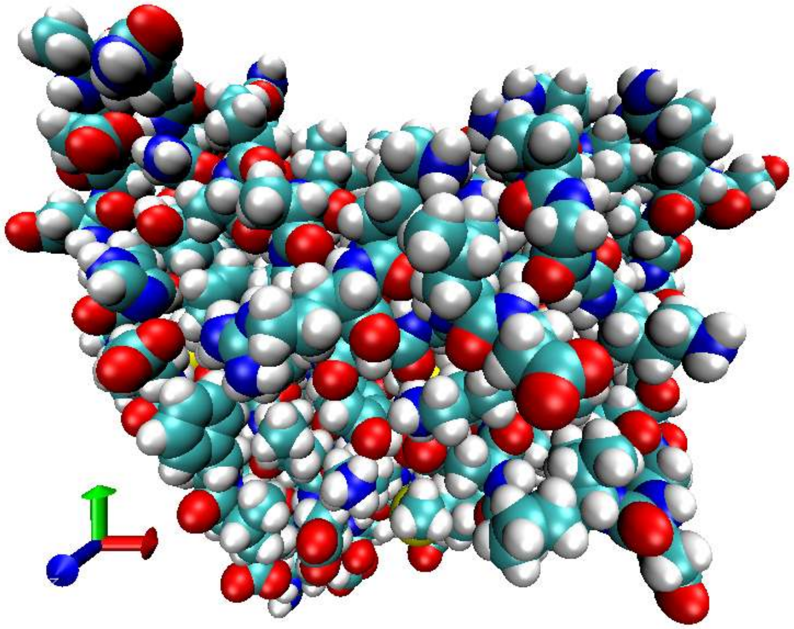
- Centori
  - simulates the fluid flow inside a tokamak reactor developed by UKAEA Fusion in collaboration with EPCC
- GS2
  - Gyrokinetic simulations of low-frequency turbulence in tokamak developed by Bill Dorland et al.

Ocean Modelling: POLCOMS



- Proudman Oceanographic Laboratory Coastal Ocean Modelling System (POLCOMS)
  - Simulation of the marine environment
  - Applications include coastal engineering, offshore industries, fisheries management, marine pollution monitoring, weather forecasting and climate research
  - Uses 3-dimensional hydrodynamic model

Molecular dynamics



Protein Dihydrofolate Reductase

- DL_POLY
  - general purpose molecular dynamics package which can be used to simulate systems with very large numbers of atoms
- LAMMPS
  - Classical Molecular Dynamics - can simulate wide range of materials
- NAMD
  - classical molecular dynamics code designed for high-performance simulation of large biomolecular systems
- AMBER
  - General purpose biomolecular simulation package
- GROMACS
  - General purpose MD package - specialises in biochemical systems, e.g. protiens, lipids etc

A parallel future?

- There are many challenges facing HPC today
- As processors have grown faster they've got hotter
- Manufacturers have responded with multicore processors
- We've entered a second golden age of parallelism
- But
  - Multicore processors are generally clocked slower than single core
  - Memory bandwidth is not increasing commensurately
  - It takes considerable effort to parallelise a code
  - Many codes do not scale

Dual Core v. Quad Core

Dual Core

- Core
  - 2.6Ghz clock frequency
  - SSE SIMD FPU (2flops/cycle = 5.2GF peak)
- Cache Hierarchy
  - L1 Dcache/Icache: 64k/core
  - L2 D/I cache: 1M/core
  - SW Prefetch and loads to L1
  - Evictions and HW prefetch to L2
- Memory
  - Dual Channel DDR2
  - 10GB/s peak @ 667MHz
  - 8GB/s nominal STREAMs
- Power
  - 103W

Quad Core

- Core
  - 2.1Ghz clock frequency
  - SSE SIMD FPU (4flops/cycle = 8.4GF peak)
- Cache Hierarchy
  - L1 Dcache/Icache: 64k/core
  - L2 D/I cache: 512 KB/core
  - L3 Shared cache 2MB/Socket
  - SW Prefetch and loads to L1,L2,L3
  - Evictions and HW prefetch to L1,L2,L3
- Memory
  - Dual Channel DDR2
  - 12GB/s peak @ 800MHz
  - 10GB/s nominal STREAMs
- Power
  - 75W

Power and cooling

- New 470m² plant room for HECToR – 1.5x the area of the room it services
- UPS provides 10-20 mins autonomy – must keep cooling running when powering HECToR – diesel engines
- Currently HECToR uses around 1.2MW
- We have provision at the ACF up to 7MW
- Rack power continues to increase:
  - 2002 – IBMp690 10kW per rack
  - 2007 – HECToR Phase 1 18kW per rack
  - 2009 – HECToR Phase 2 38kW per rack (estimate)
- Now at limits of direct air cooling – next generation must use water cooling – much more efficient

Power and cooling (cont)

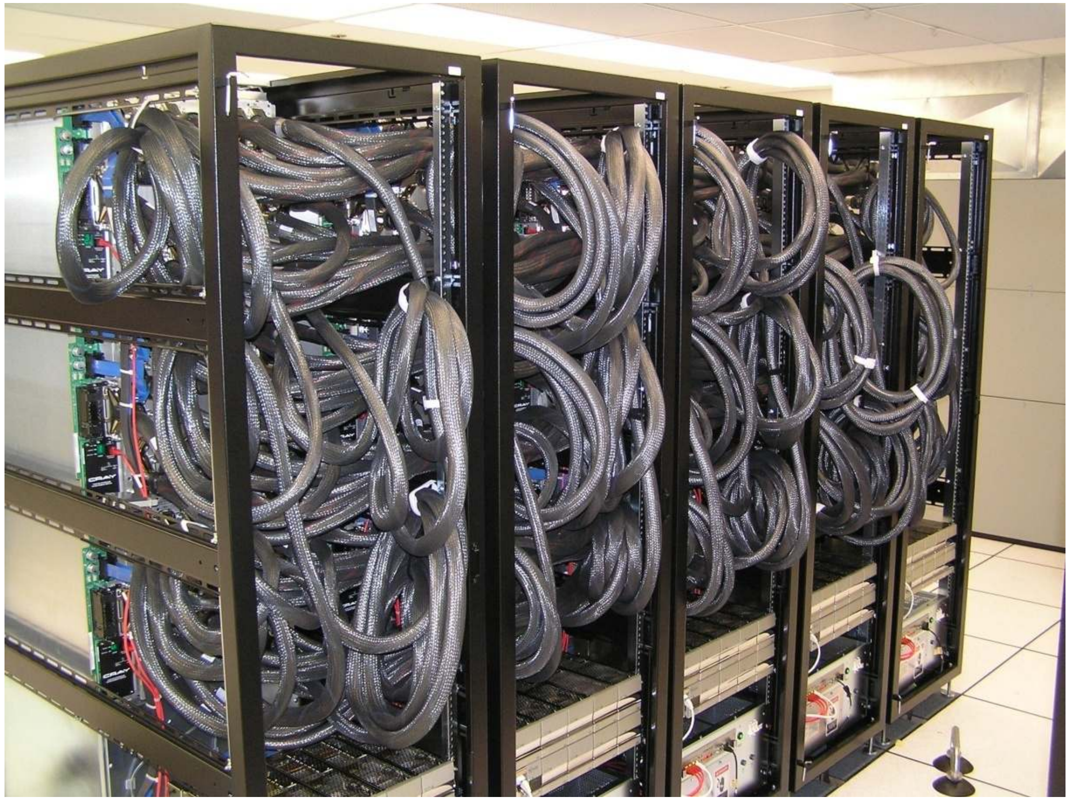
- The average off-coil air temperature is maintained with ease in the range: 12.7° - 13.3° (in excess of design spec)
- The average chilled-water flow temperature is maintained in the range: 7.7° - 8.3° (load independent)
- The average chilled-water return temperature is maintained in the range: 13.7° - 14.3°
- 60 m^3 per sec of air at mean 13° is supplied into the sub-floor
- Chilled-water flow rate is maintained at 40 litres per second
  - 144,000 litres per hour
- Because we use “free cooling” when possible the cooling overhead can be brought well below 20% over the year

Parallel scaling

- To make use of highly parallel systems the performance of a code must scale linearly with the number of processors it is executed on
- Many do not - due to
  - Memory bandwidth issues in an SMP environment
    - While Taiwanese memory producers are producing bigger and bigger devices they're not getting faster
  - Communication latency and bandwidth issues
- A key problem facing many commercial simulation codes (known as ISV codes) is scalability
  - Many ISV codes only scale to 16 – 32 processors

A room full of PCs is not a supercomputer

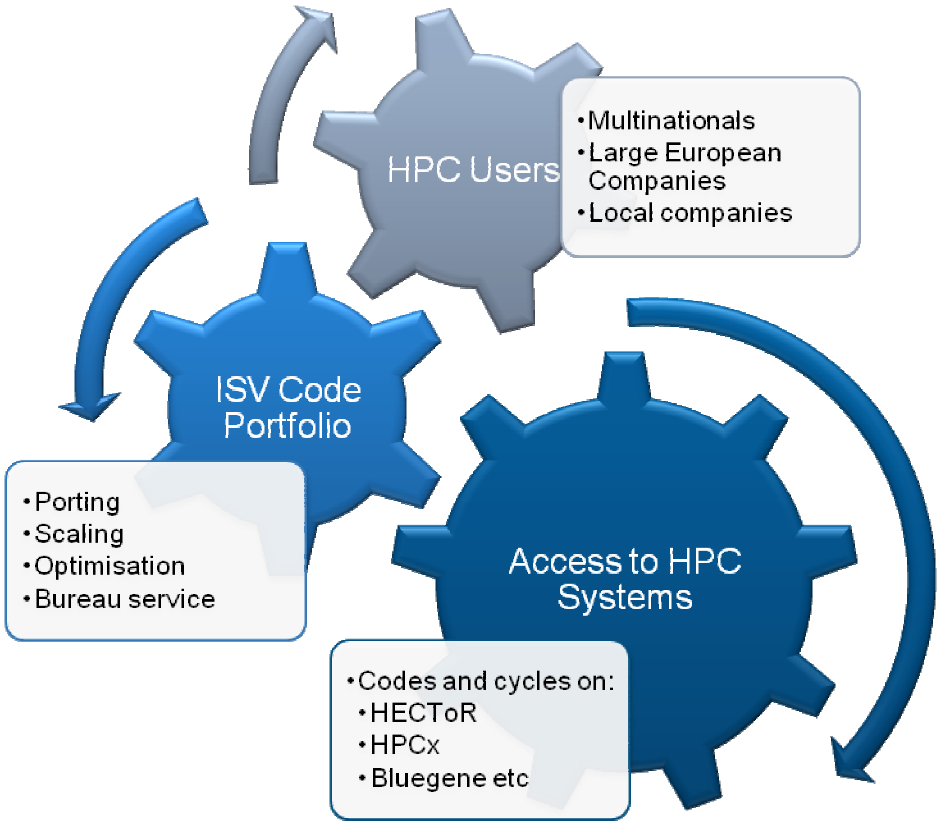
- HECToR is expensive because of its communications network
- Designed for
  - High bandwidth
  - Low latency
- Mandatory requirement to scale to 10,000+ cores



A sneak preview

- EPCC has a unique opportunity to work with ISVs and industry users to improve their use of highly parallel systems
- Over the next 6 months we're creating the *EPCC Industry Simulation Centre*
- Drivers
  - Our existing work with companies over past 18 years – 50% of our £4.7million turnover comes from working with industry
  - Pay-per-use access machines – HECToR, HPCx, Bluegene/L etc
  - Our expertise in optimising and scaling codes for our scientific users
  - Much greater use of simulation by large companies
  - Too little use by smaller Scottish companies
  - Our relationships with hardware vendors – Cray, IBM etc
  - Our desire to prepare for a Petascale system in 2010

The ISC Ecosystem



- SE funding to engage Scottish business
- Builds on existing infrastructure
- Once established – income from cycle sales will feed back into ISV code work
- Strong sales and marketing activity
- Need to partner with hardware vendors to use their ISV contacts
- ISV codes and bespoke codes

Conclusion

- At 60 TFlops, HECToR is one of the most powerful computers in the world today
- It’s a fantastic asset for Scotland
- It serves the UK scientific community *and* the business community
- We’re at a very interesting moment in computing
- The days of easy programmability are over
- We’re entering a new golden age of parallel computing!

Thanks and questions