THE HECToR SUPERCOMPUTER

AT THE PETASCALE AND BEYOND

Dr Mark Parsons

EPCC Executive Director
Associate Dean for e-Research

The University of Edinburgh

Talk outline

- Introduction to supercomputing
- HECToR
  - From Phase 1 to Phase 3
  - Introduction to AMD Bulldozer Architecture
  - Some examples of usage
- From the Petascale to the Exascale
  - What are the next challenges in supercomputing
  - Why we're at a key point in the evolution of supercomputing
- Thanks due to
  - Alan Simpson and Jeremy Nowell (EPCC)
  - George Mozdzynski (ECMWF)

EPCC

- EPCC is the supercomputing centre at The University of Edinburgh
- Founded in 1990 as focus for work in parallel computing
- Hosting national HPC services since 1994 for academia
- 70 staff highly skilled staff
- Wide variety of projects and stakeholders
  - UK Research Councils
  - Scottish Enterprise
  - European Commission
  - Scottish and UK industry
- Working with industry and commerce since 1990
  - Software development and consultancy
  - Provision of on-demand HPC to industry

What are FLOPS?

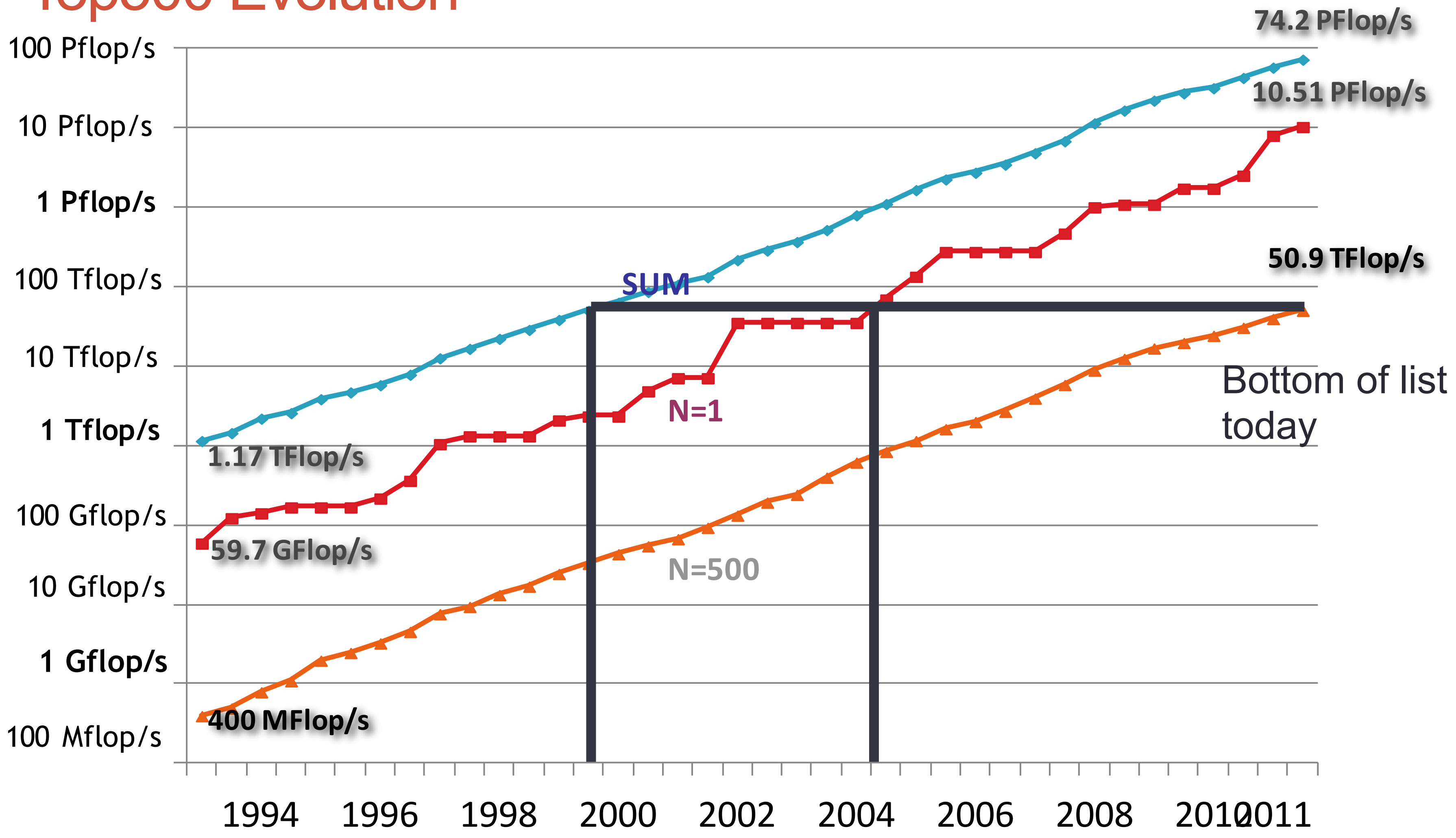
FLOPS = Floating point Operations Per Second

10 Petaflops = 10^{16} FLOPS

= 10,000,000,000,000,000 FLOPS

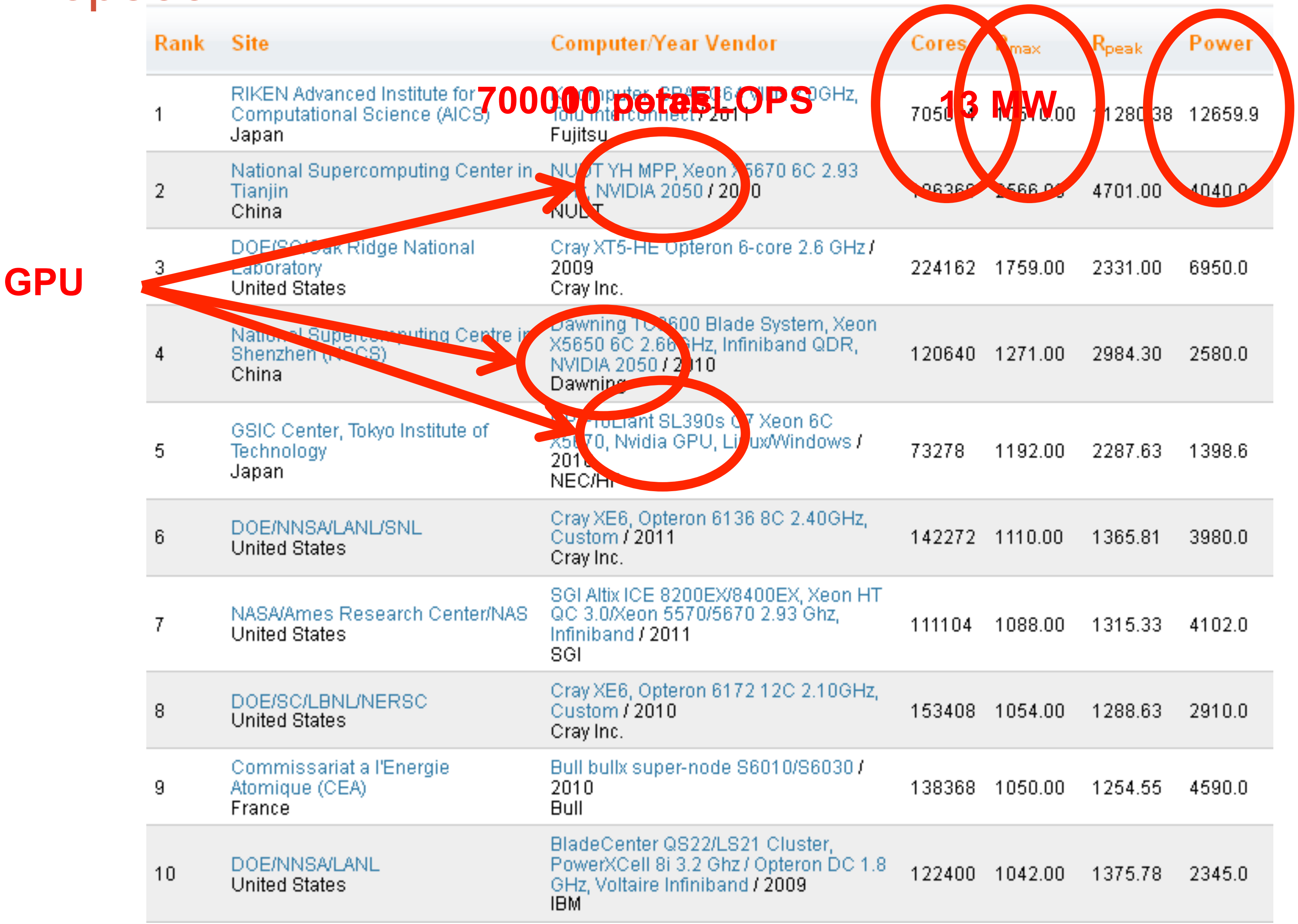
= 1,000,000 FLOPS for every person on the planet

Top500 Evolution



source: www.top500.org

Top500



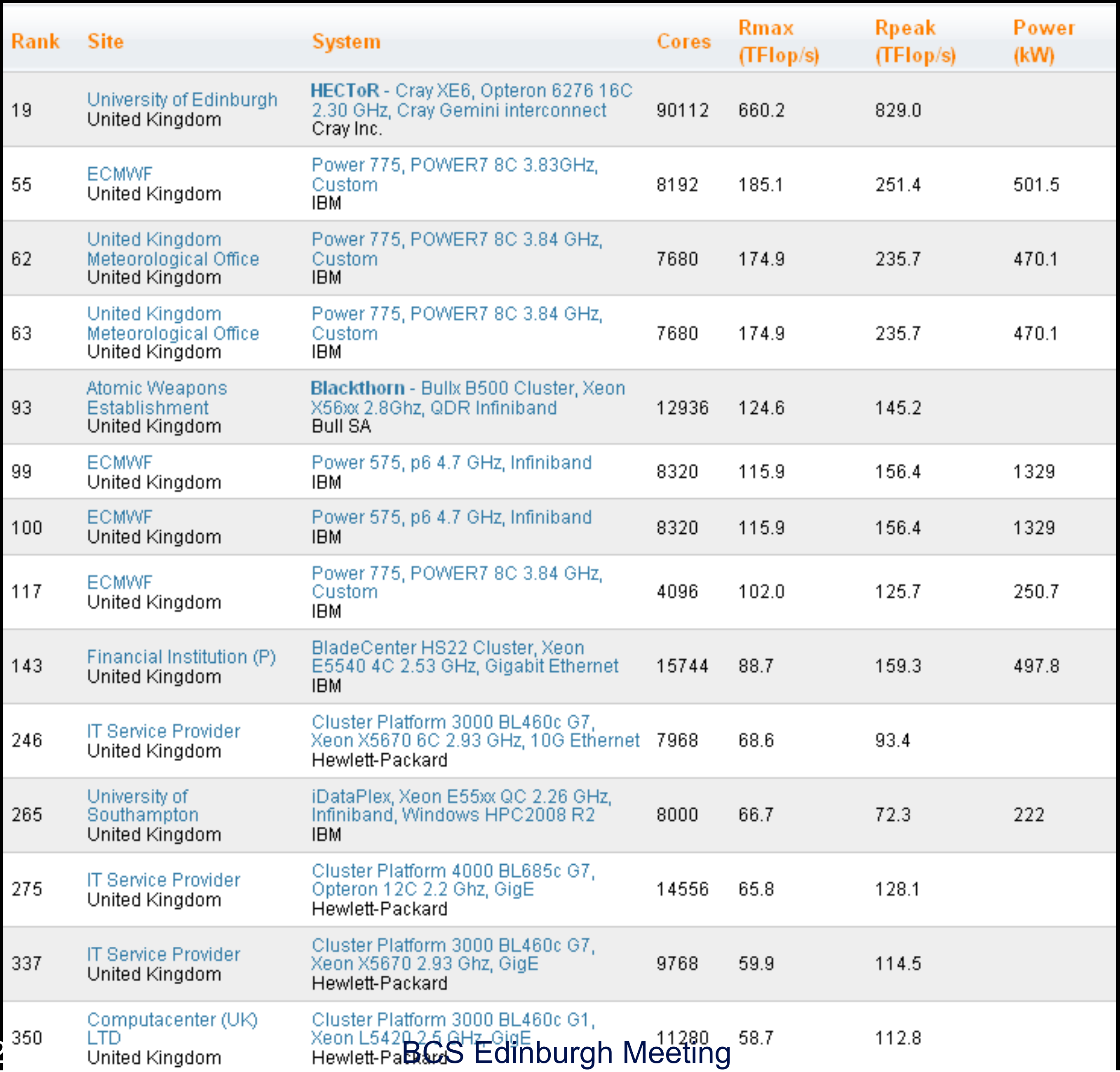
Rank	Site	Computer/Year Vendor	Cores	R_{max}	R_{peak}	Power
1	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect / 2011 Fujitsu	705024	10510.00	11280.38	12659.9
2	National Supercomputing Center in Tianjin China	NUDT YH MPP, Xeon X5670 6C 2.93 GHz, NVIDIA 2050 / 2010 NUDT	186368	2566.00	4701.00	4040.0
3	DOE/SC/Oak Ridge National Laboratory United States	Cray XT5-HE Opteron 6-core 2.6 GHz / 2009 Cray Inc.	224162	1759.00	2331.00	6950.0
4	National Supercomputing Centre in Shenzhen (NSCS) China	Dawning TC3600 Blade System, Xeon X5650 6C 2.66GHz, Infiniband QDR, NVIDIA 2050 / 2010 Dawning	120640	1271.00	2984.30	2580.0
5	GSIC Center, Tokyo Institute of Technology Japan	HP ProLiant SL390s G7 Xeon 6C X5670, Nvidia GPU, Linux/Windows / 2010 NEC/HP	73278	1192.00	2287.63	1398.6
6	DOE/NNSA/LANL/SNL United States	Cray XE6, Opteron 6136 8C 2.40GHz, Custom / 2011 Cray Inc.	142272	1110.00	1365.81	3980.0
7	NASA/Ames Research Center/NAS United States	SGI Altix ICE 8200EX/8400EX, Xeon HT QC 3.0/Xeon 5570/5670 2.93 Ghz, Infiniband / 2011 SGI	111104	1088.00	1315.33	4102.0
8	DOE/SC/LBNL/NERSC United States	Cray XE6, Opteron 6172 12C 2.10GHz, Custom / 2010 Cray Inc.	153408	1054.00	1288.63	2910.0
9	Commissariat a l'Energie Atomique (CEA) France	Bull bullx super-node S6010/S6030 / 2010 Bull	138368	1050.00	1254.55	4590.0
10	DOE/NNSA/LANL United States	BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz, Voltaire Infiniband / 2009 IBM	122400	1042.00	1375.78	2345.0

source: www.top500.org

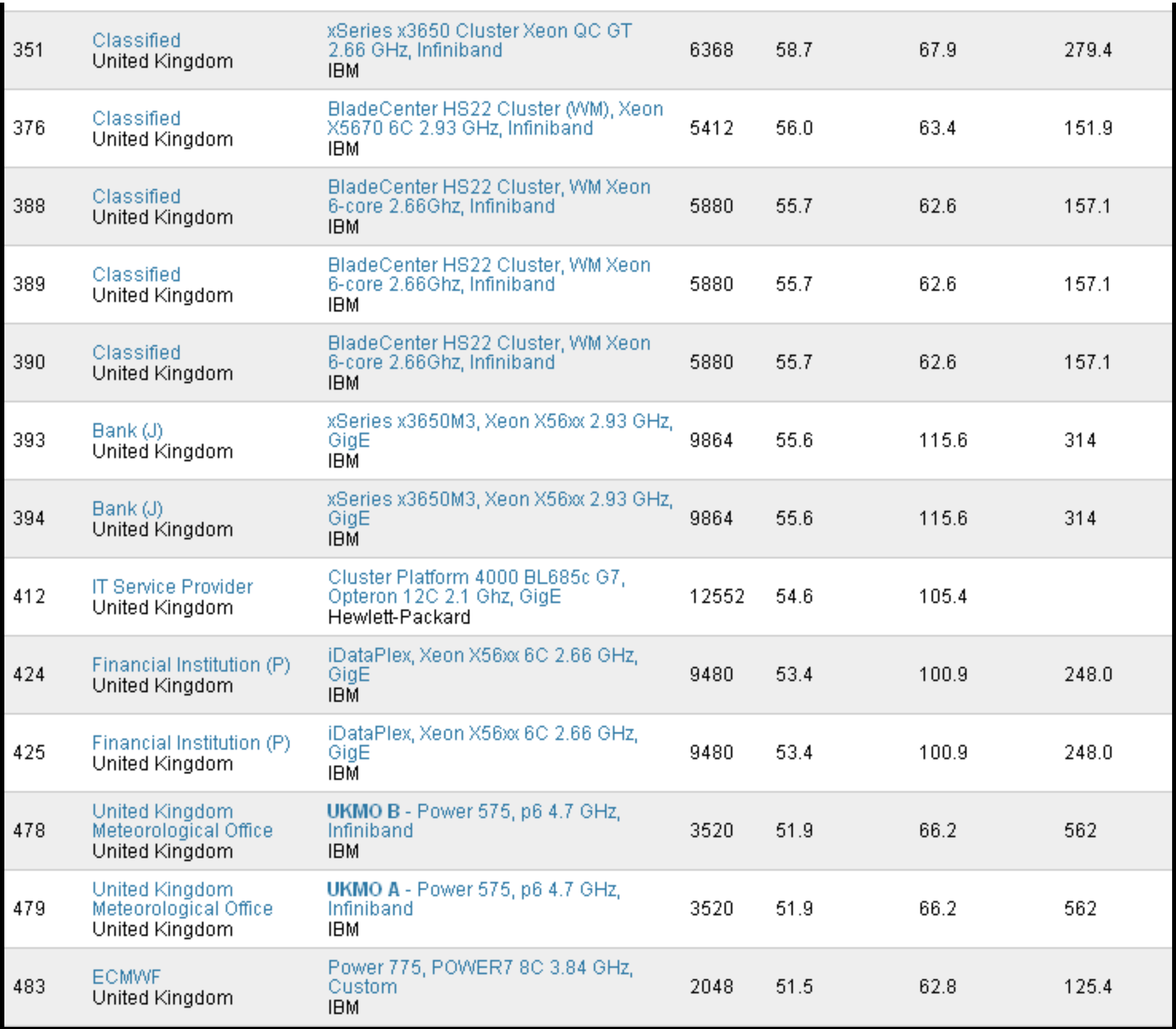
Last update: November 2011

The Japanese K-Computer

Top 500 – UK



Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
19	University of Edinburgh United Kingdom	**HECToR** - Cray XE6, Opteron 6276 16C 2.30 GHz, Cray Gemini interconnect Cray Inc.	90112	660.2	829.0	
55	ECMWF United Kingdom	Power 775, POWER7 8C 3.83GHz, Custom IBM	8192	185.1	251.4	501.5
62	United Kingdom Meteorological Office United Kingdom	Power 775, POWER7 8C 3.84 GHz, Custom IBM	7680	174.9	235.7	470.1
63	United Kingdom Meteorological Office United Kingdom	Power 775, POWER7 8C 3.84 GHz, Custom IBM	7680	174.9	235.7	470.1
93	Atomic Weapons Establishment United Kingdom	**Blackthorn** - Bullx B500 Cluster, Xeon X56xx 2.8Ghz, QDR Infiniband Bull SA	12936	124.6	145.2	
99	ECMWF United Kingdom	Power 575, p6 4.7 GHz, Infiniband IBM	8320	115.9	156.4	1329
100	ECMWF United Kingdom	Power 575, p6 4.7 GHz, Infiniband IBM	8320	115.9	156.4	1329
117	ECMWF United Kingdom	Power 775, POWER7 8C 3.84 GHz, Custom IBM	4096	102.0	125.7	250.7
143	Financial Institution (P) United Kingdom	BladeCenter HS22 Cluster, Xeon E5540 4C 2.53 GHz, Gigabit Ethernet IBM	15744	88.7	159.3	497.8
246	IT Service Provider United Kingdom	Cluster Platform 3000 BL460c G7, Xeon X5670 6C 2.93 GHz, 10G Ethernet Hewlett-Packard	7968	68.6	93.4	
265	University of Southampton United Kingdom	iDataPlex, Xeon E55xx QC 2.26 GHz, Infiniband, Windows HPC2008 R2 IBM	8000	66.7	72.3	222
275	IT Service Provider United Kingdom	Cluster Platform 4000 BL685c G7, Opteron 12C 2.2 Ghz, GigE Hewlett-Packard	14556	65.8	128.1	
337	IT Service Provider United Kingdom	Cluster Platform 3000 BL460c G7, Xeon X5670 2.93 Ghz, GigE Hewlett-Packard	9768	59.9	114.5	
350	Computacenter (UK) LTD United Kingdom	Cluster Platform 3000 BL460c G1, Xeon L5420 2.5 GHz, GigE Hewlett-Packard	11280	58.7	112.8	



351	Classified United Kingdom	xSeries x3650 Cluster Xeon QC GT 2.66 GHz, Infiniband IBM	6368	58.7	67.9	279.4
376	Classified United Kingdom	BladeCenter HS22 Cluster (WM), Xeon X5670 6C 2.93 GHz, Infiniband IBM	5412	56.0	63.4	151.9
388	Classified United Kingdom	BladeCenter HS22 Cluster, WM Xeon 6-core 2.66Ghz, Infiniband IBM	5880	55.7	62.6	157.1
389	Classified United Kingdom	BladeCenter HS22 Cluster, WM Xeon 6-core 2.66Ghz, Infiniband IBM	5880	55.7	62.6	157.1
390	Classified United Kingdom	BladeCenter HS22 Cluster, WM Xeon 6-core 2.66Ghz, Infiniband IBM	5880	55.7	62.6	157.1
393	Bank (J) United Kingdom	xSeries x3650M3, Xeon X56xx 2.93 GHz, GigE IBM	9864	55.6	115.6	314
394	Bank (J) United Kingdom	xSeries x3650M3, Xeon X56xx 2.93 GHz, GigE IBM	9864	55.6	115.6	314
412	IT Service Provider United Kingdom	Cluster Platform 4000 BL685c G7, Opteron 12C 2.1 Ghz, GigE Hewlett-Packard	12552	54.6	105.4	
424	Financial Institution (P) United Kingdom	iDataPlex, Xeon X56xx 6C 2.66 GHz, GigE IBM	9480	53.4	100.9	248.0
425	Financial Institution (P) United Kingdom	iDataPlex, Xeon X56xx 6C 2.66 GHz, GigE IBM	9480	53.4	100.9	248.0
478	United Kingdom Meteorological Office United Kingdom	**UKMO B** - Power 575, p6 4.7 GHz, Infiniband IBM	3520	51.9	66.2	562
479	United Kingdom Meteorological Office United Kingdom	**UKMO A** - Power 575, p6 4.7 GHz, Infiniband IBM	3520	51.9	66.2	562
483	ECMWF United Kingdom	Power 775, POWER7 8C 3.84 GHz, Custom IBM	2048	51.5	62.8	125.4

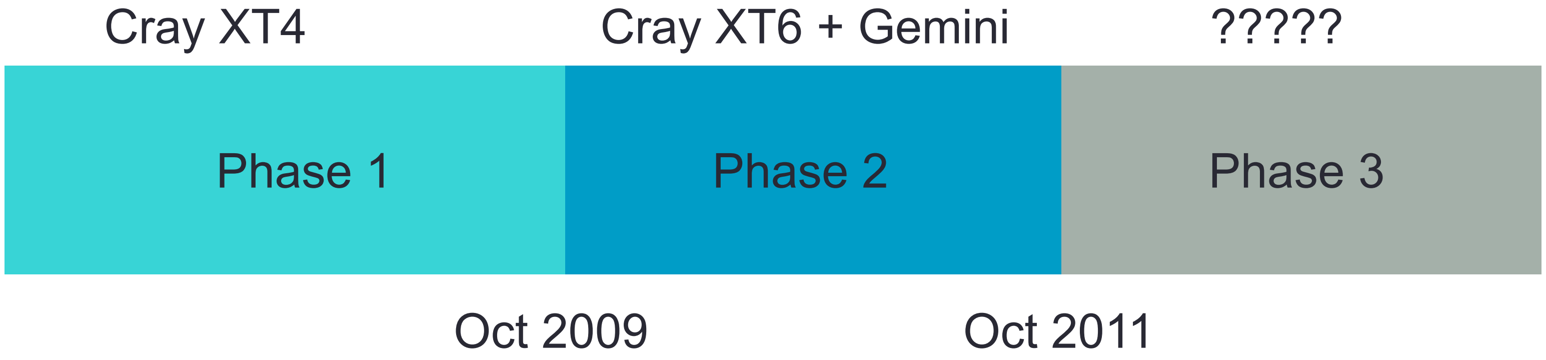
source: www.top500.org

The National HPC Service: HECToR

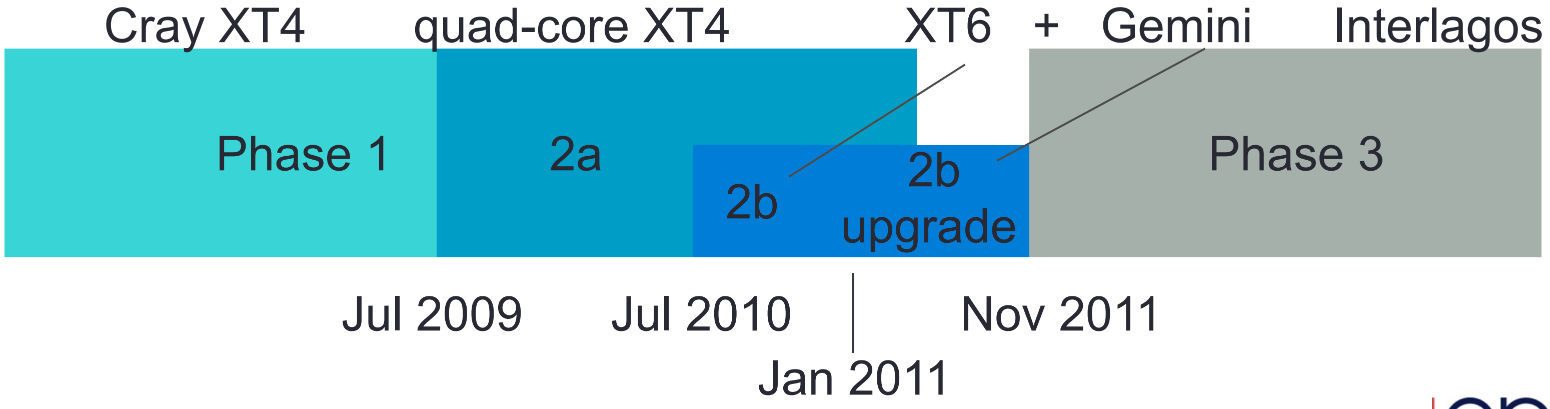
- HECToR: UK's national HPC service
    - £115M project from 2007-2013
    - Hosted by EPCC at our Advanced Computing Facility
    - Cray XE6 system
    - Recently upgraded to 90,112 AMD Interlagos cores (>800TF)
- HECToR partners
    - RCUK – UK Research funding councils
        - Led by EPSRC
        - EPSRC, BBSRC and NERC are the "Partner Research Councils"
        - But all Research Councils can gain access the system
            - Including STFC Daresbury Laboratory who provide some of the systems support
    - EPCC via UOE HPCX Ltd - host and operate the system
    - Cray Inc – HPC hardware
    - NAG Ltd – Computational science and engineering support

The HECToR Roadmap

- In the beginning



- …but then the new processor was early and Gemini was late



HECToR Phase 1 installation in 2007

April
2007

Edinburgh: new building in progress

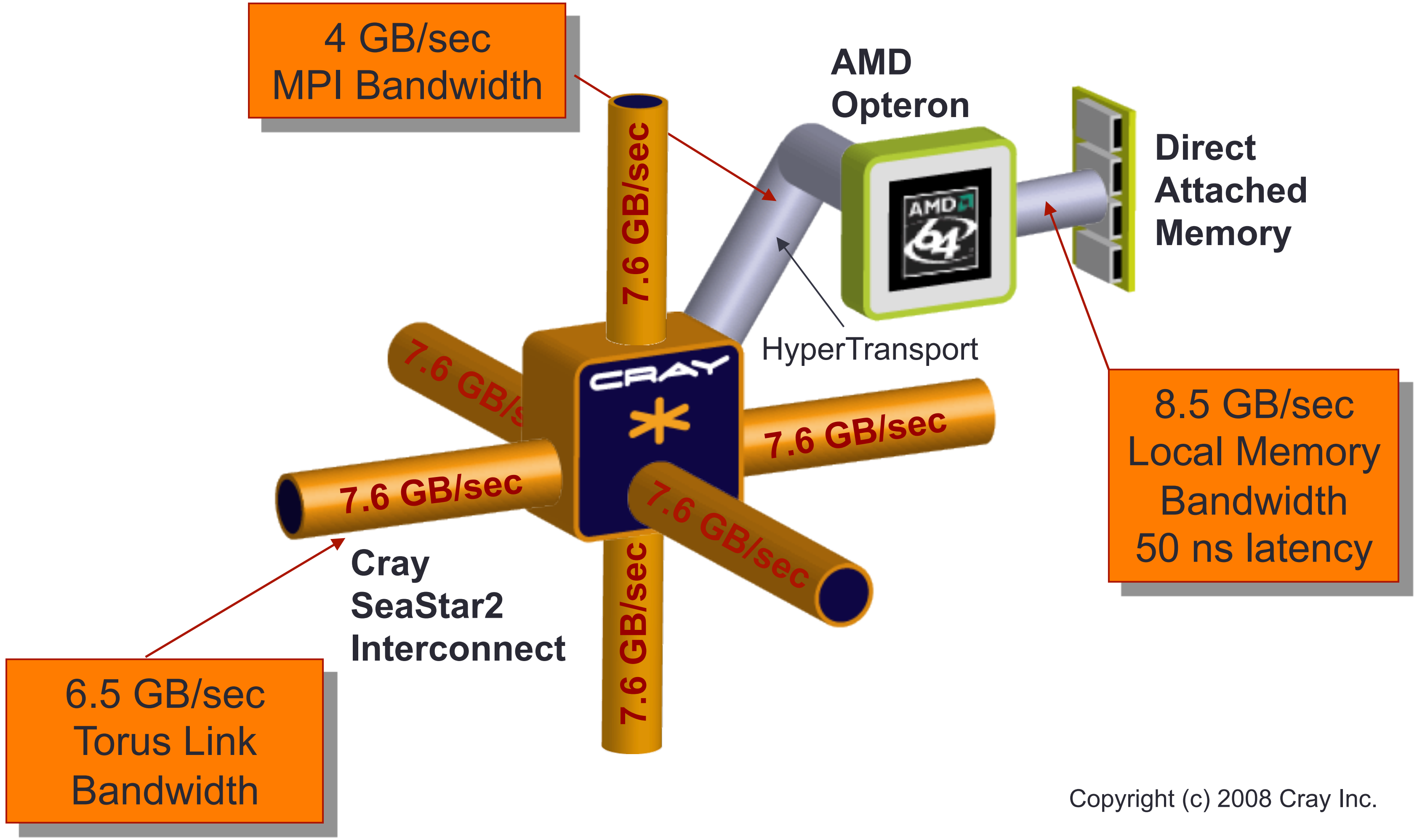
Edinburgh: Test and Development System (one XT4 cabinet) installed

August
2007

Edinburgh: Full 60 Cabinet System installed

HECToR Phase 1 at the ACF

HECToR Phase 1 Cray XT4 Processing Element



4 GB/sec
MPI Bandwidth

AMD
Opteron

Direct
Attached
Memory

HyperTransport

8.5 GB/sec
Local Memory
Bandwidth
50 ns latency

Cray
SeaStar2
Interconnect

6.5 GB/sec
Torus Link
Bandwidth

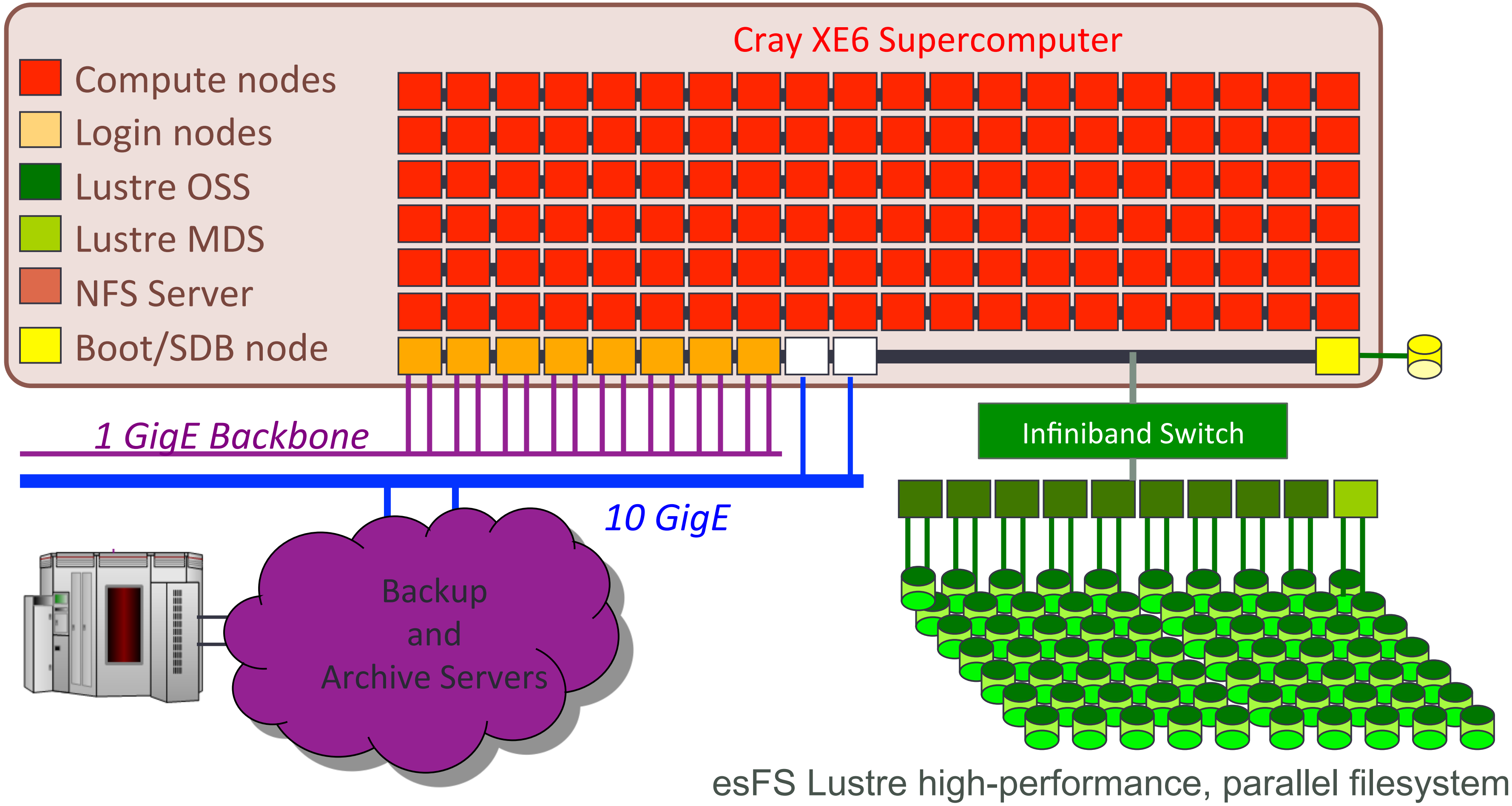
HECToR at the ACF until Nov 2011

- Phase 2b
    - 20 cabinet Cray XE6
    - 44,544 cores
    - 59.4Tb memory
    - Gemini interconnect
    - 360 Tflops
- Phase 2a (additional until May 2011)
    - 33 cabinet Cray XT4
    - 12,288 cores
    - 24Tb memory
    - 1 cabinet Cray X2 with 112 vector processors

HECToR Phase 3

- 30-cabinet Cray XE6 system
- 2816 nodes, 90,112 cores
- Each node has
  - 2×16-core AMD Opterons (2.3GHz Interlagos)
  - 32 GB memory
- Peak of over 830 TF
- 90 TB of memory

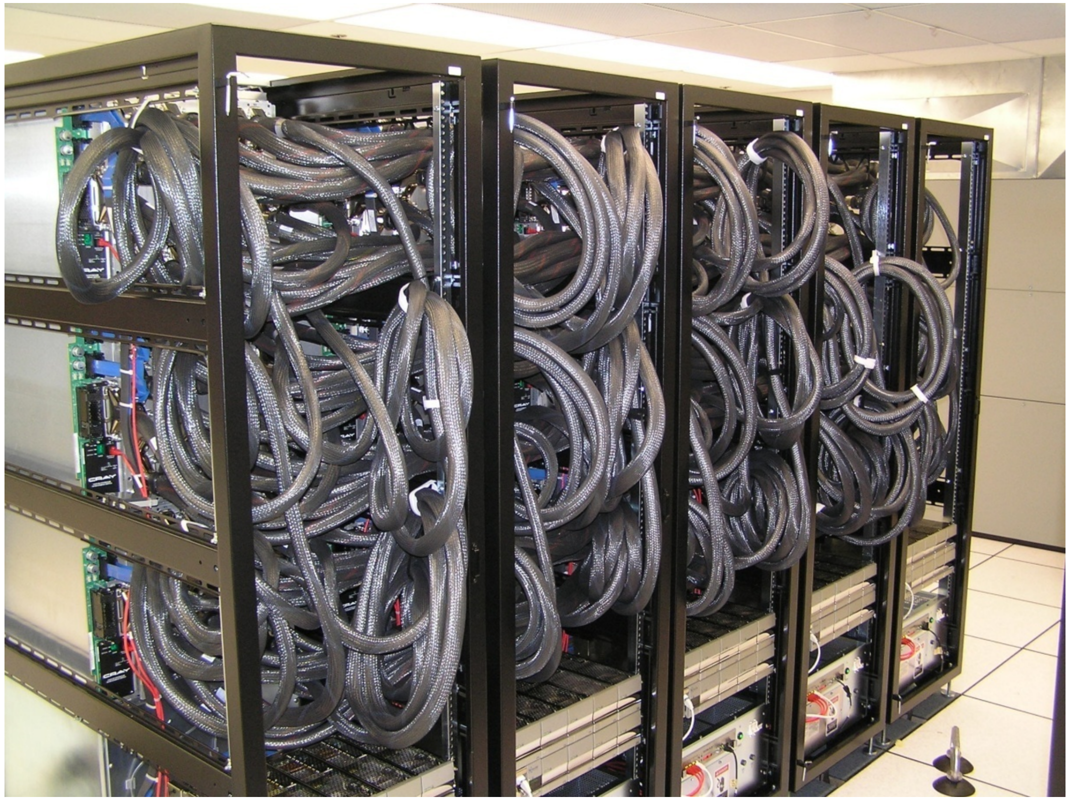
HECToR Service



esFS Lustre high-performance, parallel filesystem

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 100,000+ cores
- Major Phase 2b upgrade was Gemini interconnect



AMD Bulldozer Architecture

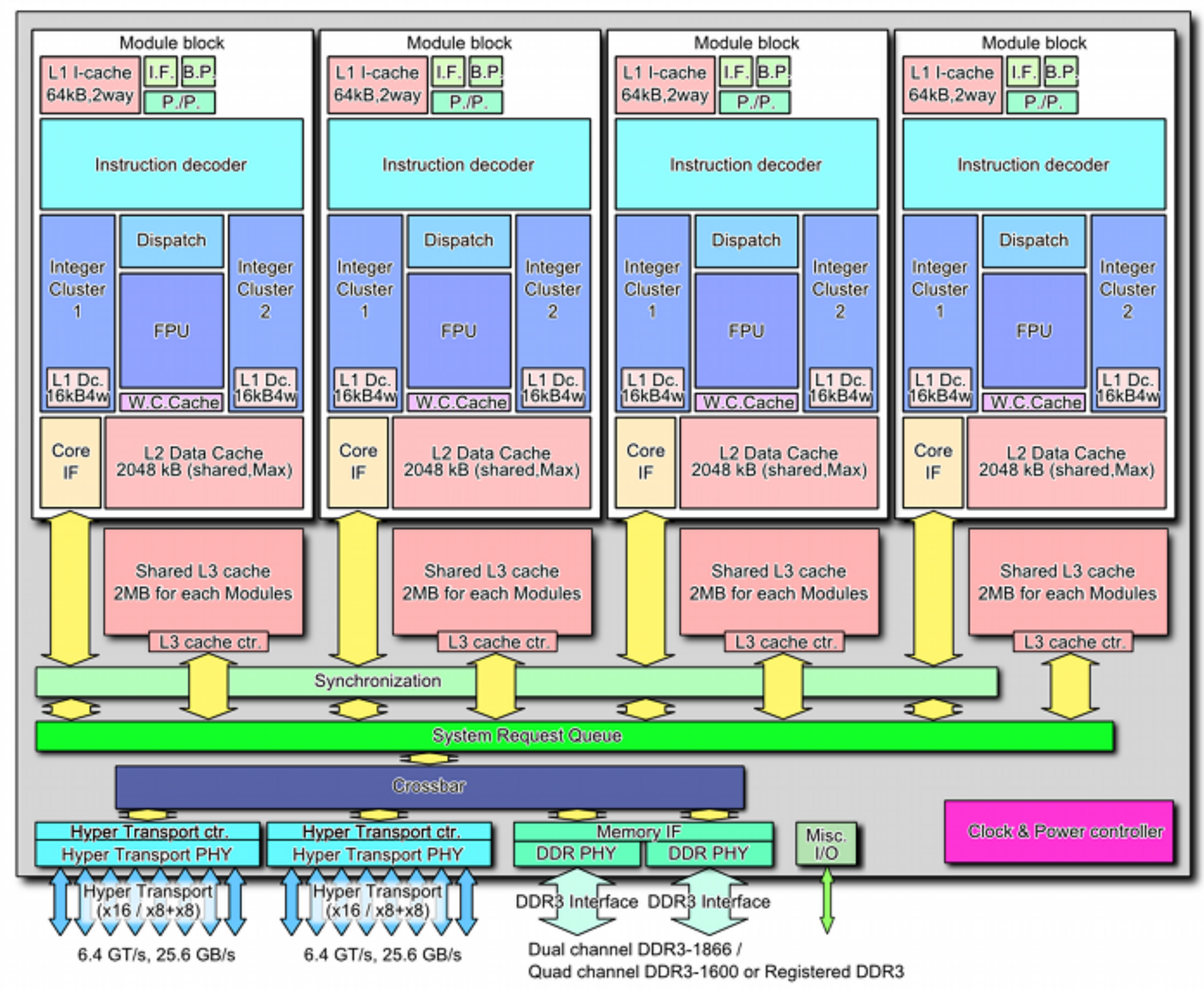
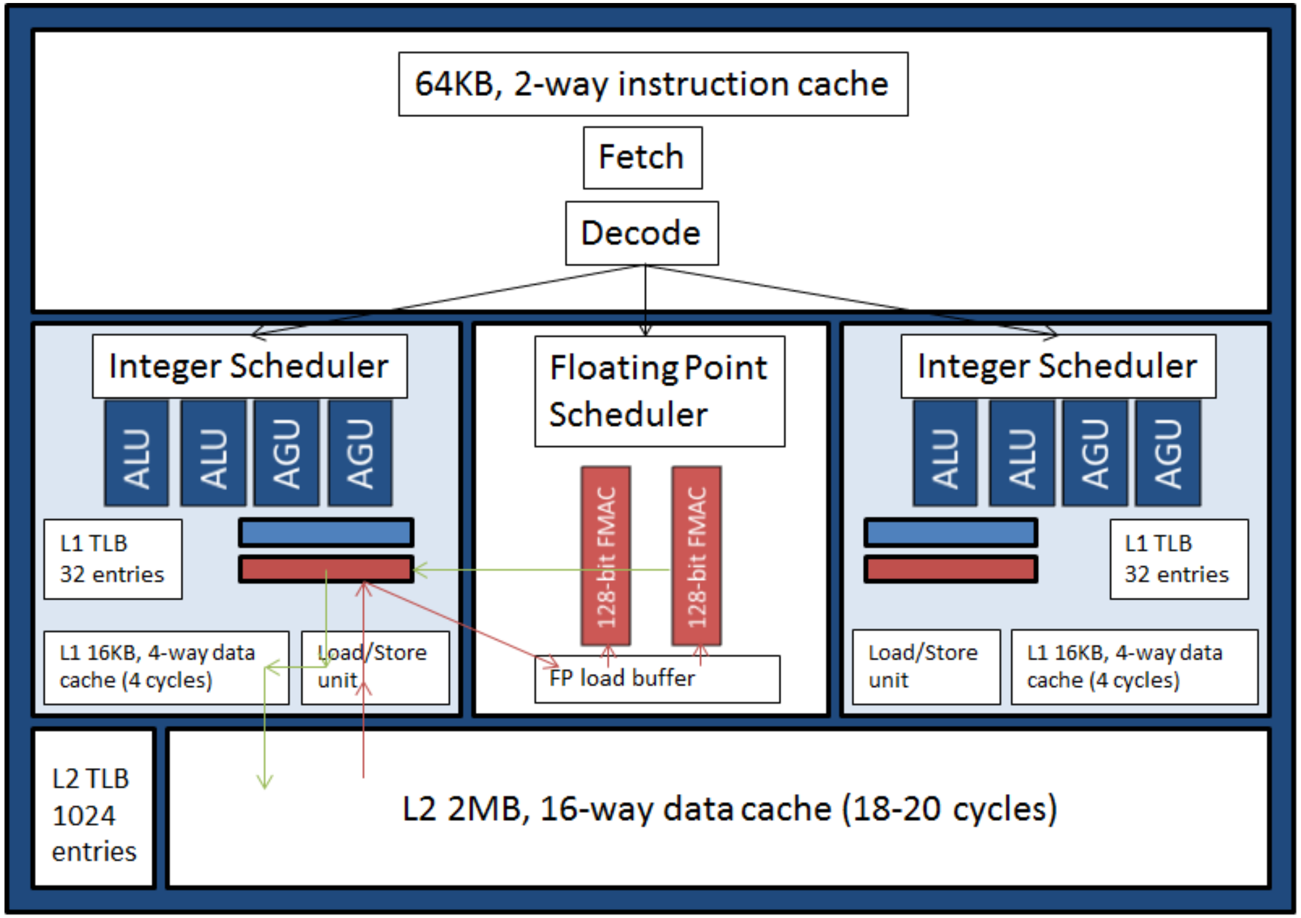


Image courtesy of Wikipedia

Interlagos dual bulldozer-core module

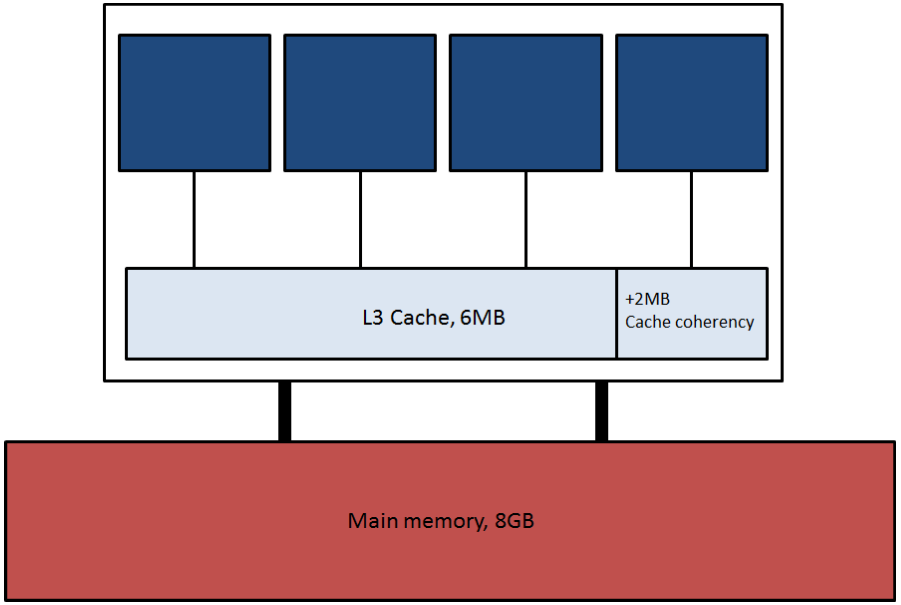


🟦 16x x86_64 general purpose integer registers

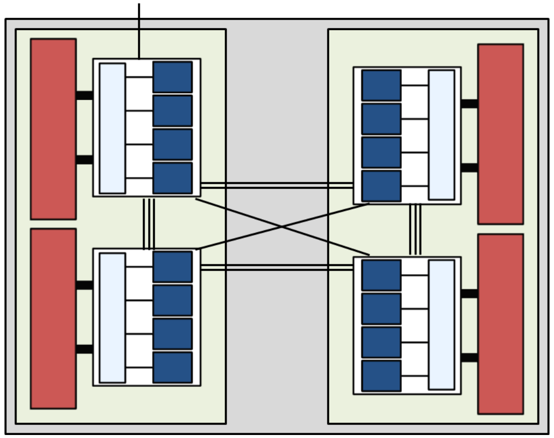
🟥 16x 256 bit AVX registers (ymm0-ymm15) / 16x 128 bit SSE registers (xmm0-xmm15)

Interlagos processor

- Each blue square represents a module containing two cores
- The four modules share a 6MB L3 cache
- A processor socket consists of two dies like this



- A HECToR node consists of two processors
- NUMA topology between dies and sockets
- Hypertransport throughout plus link to Gemini interconnect

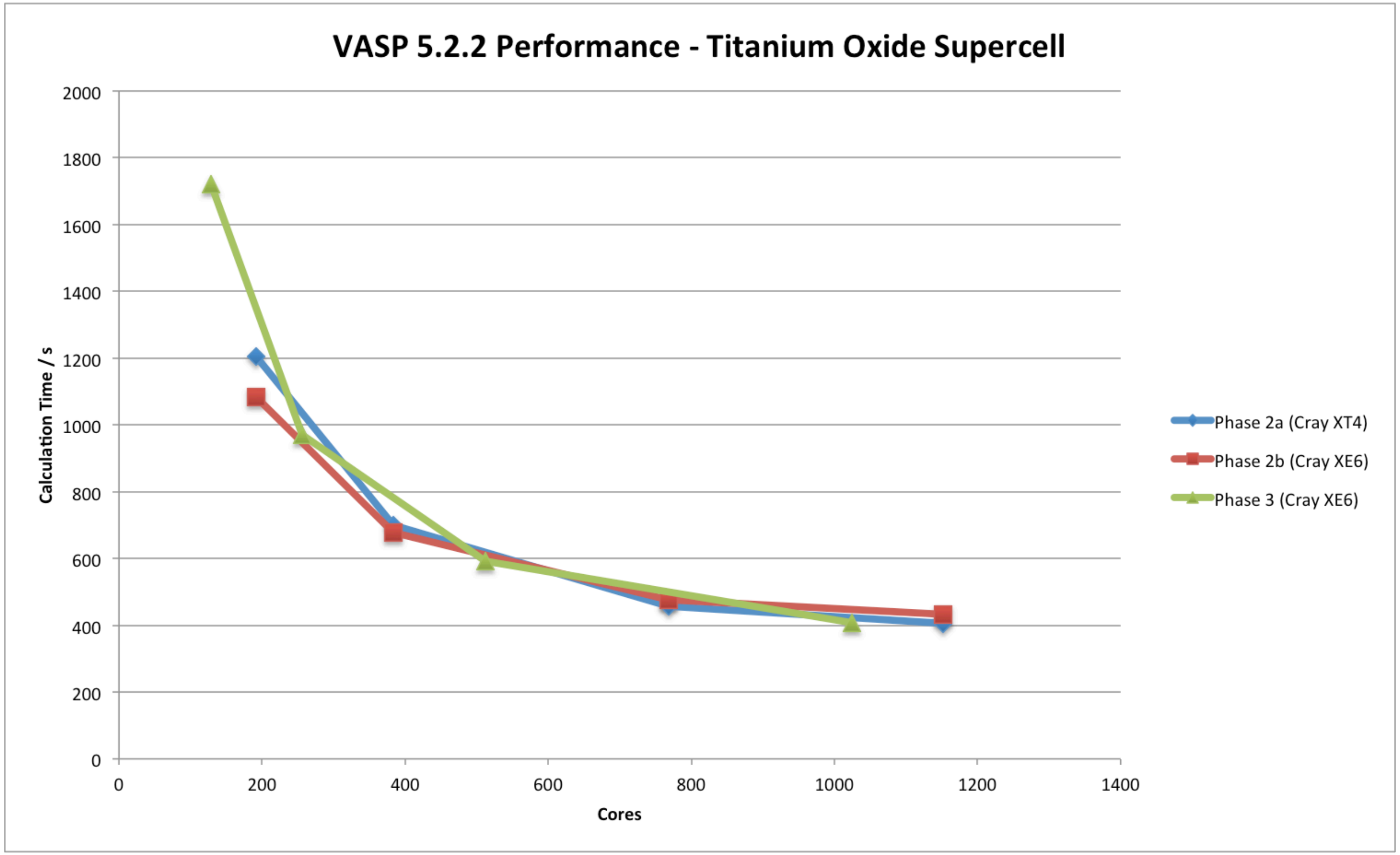


Comparison with Phase 2b nodes

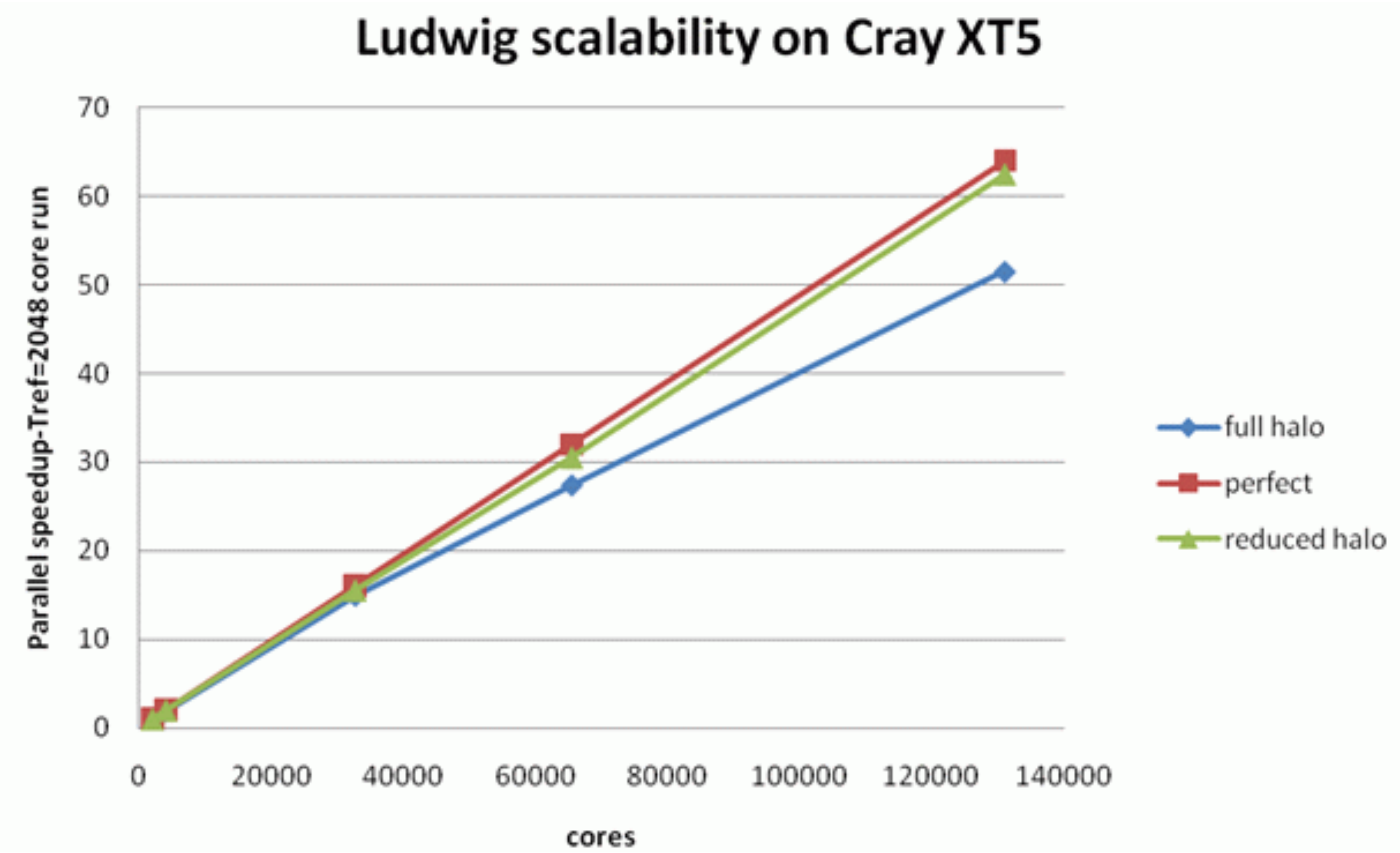
	Phase 2b (Opteron 61xx)	Phase 3 (Opteron 62xx)
Cores	24	32
Clock Speed	2.1 GHz	2.3 GHz
Memory	32 GB (1.3 GB/core)	32 GB (1 GB/core)
Memory Bandwidth	42.6 GB/s (3.55 GB/s per core)	51.2 GB/s (3.2 GB/s per core, 6.4 GB/s per module)
Vector Instructions	MMX, SSE, SSE2, SSE3, SSE4a	+ SSE4.1, SSE4.2, AVX, XOP, FMA4

- What does all this mean for code performance?

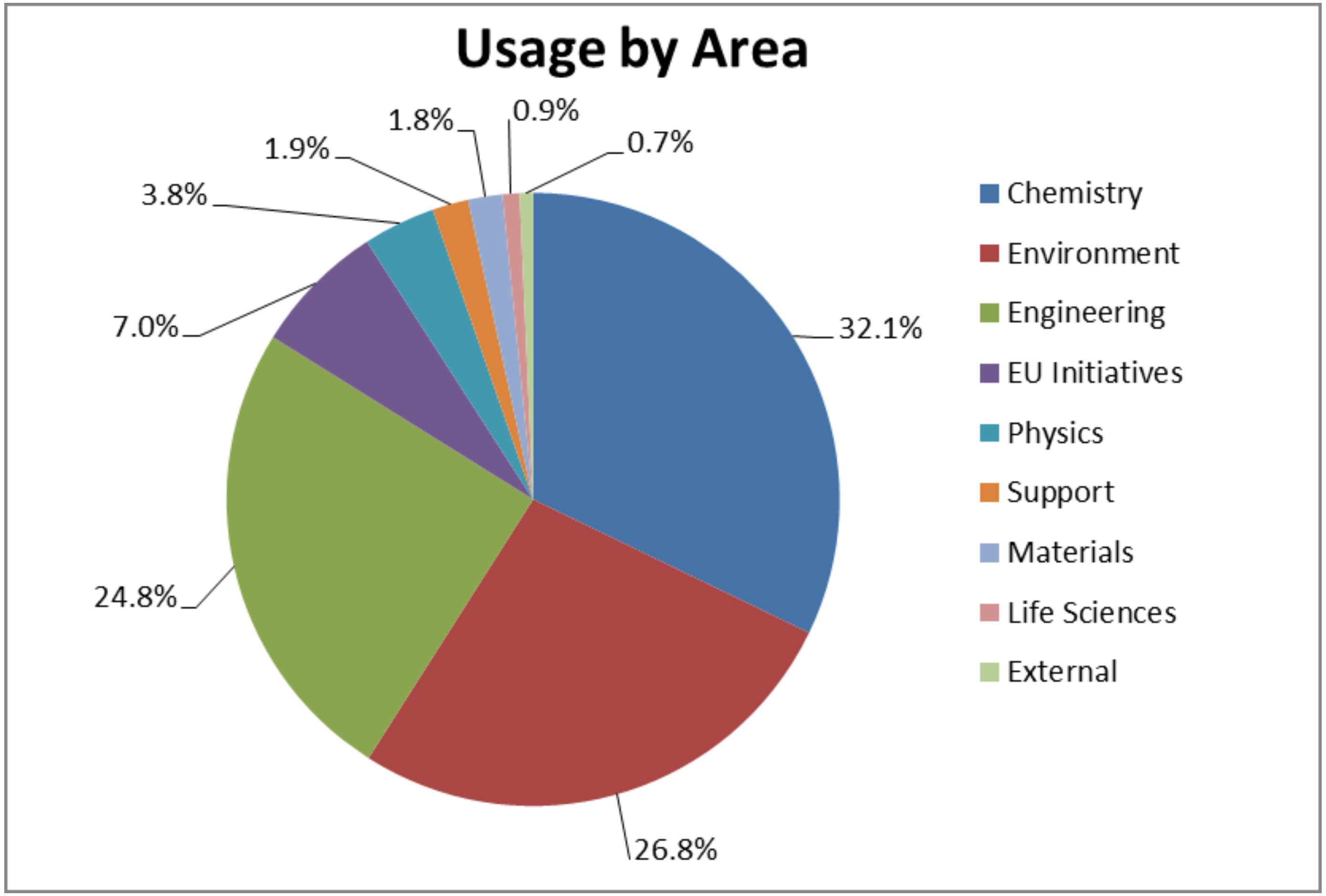
Phase 3 Performance Comparison



Scaling to very large core counts

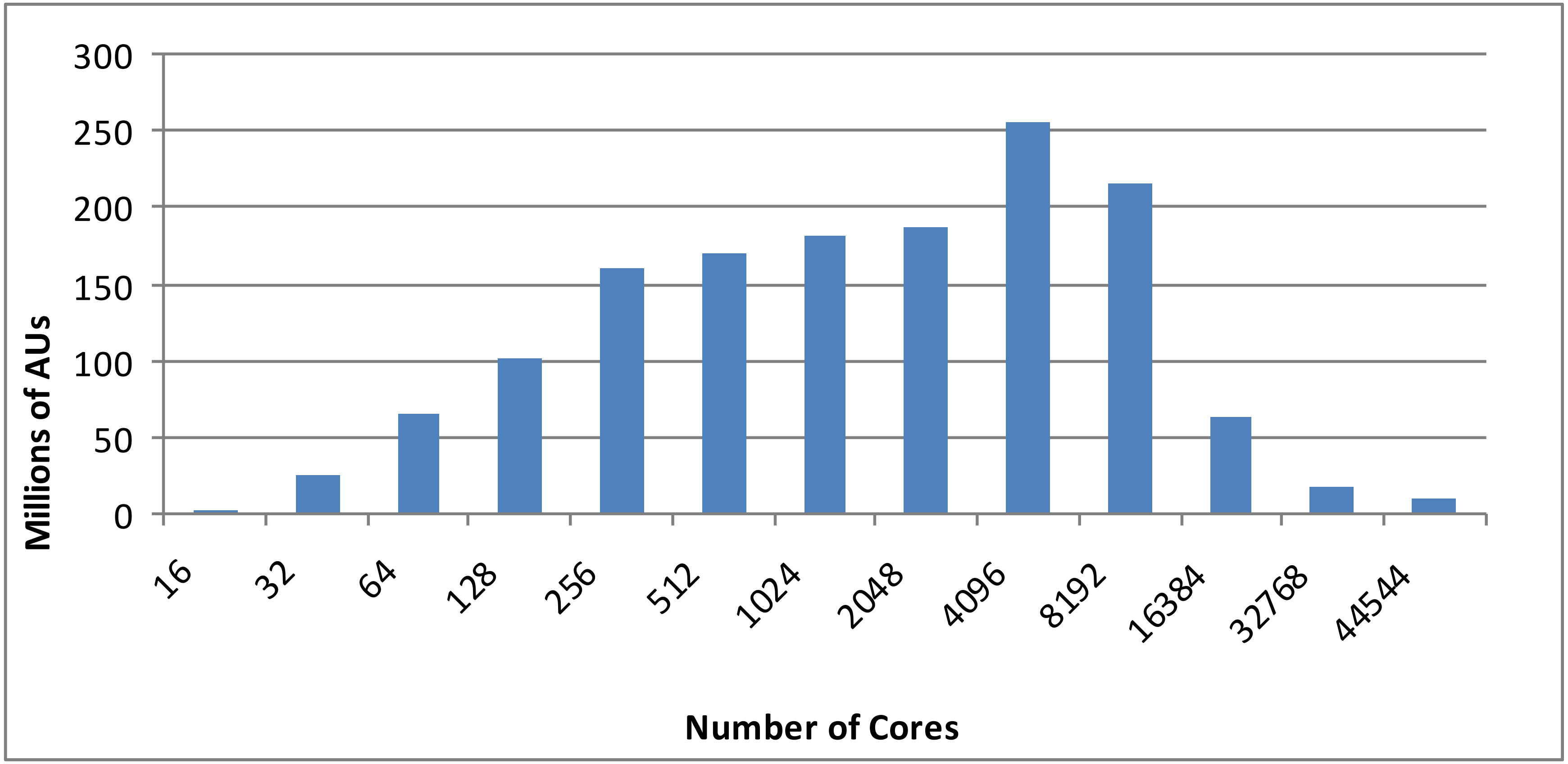


Who uses HECToR?



Currently we have around 1,800 active users

Job Size



Most heavily-used job size is 4096 cores

The Exascale Challenge

Supercomputing today

- Programming model is one of a set of distinct memories distributed over homogeneous microprocessors
  - Each microprocessor generally runs a Unix-like OS
- Data transfers between the microprocessors are managed explicitly by the application
  - With the exception of PGAS languages and some shared-memory technologies
- Almost all programs are written in sequential Fortran, C or C++
- The majority use MPI (Message Passing Interface) for data transfers between microprocessors
- Some applications which exploit parallel threads on each microprocessor use a hybrid model
  - Shared memory on the microprocessor, distributed memory outwith
  - This holds promise for many applications, but is still rare
- There is some use of accelerators – predominately GPGPU – but this is not yet mainstream

A looming problem …

- We are at a complex juncture in the history of supercomputing
- For the past 20 years supercomputing has “hitched a lift” on the microprocessor revolution driven by the PC and gaming
- Hardware has been surprisingly stable
- EPCC in 1994 had the 512 processor Cray T3D system
    - 0.0768 TFlops peak
- EPCC in 2010 retired the 2,560 processor IBM HPCx system
    - 15.36 TFlops peak – 200 x faster but only 5 x more processors ...
- The programming models for these systems were very similar
- Today we have the systems with more than 100,000 cores
    - … and yet the programming model hasn’t changed

Hardware is leaving software behind

- Hardware is leaving many HPC users and codes behind
- Many codes scale to less than 512 cores
  - These will soon be desktop systems
- Less than 10 codes in EU today will scale on capability systems with 100,000+ cores
  - HECToR (90,112) and HERMIT (113,664) are already at this level
  - Germany’s Jugene system has almost 300,000
- Many industrial codes scale very poorly – some codes will soon find a laptop processor a challenge!
- Much hope is pinned on accelerator technology
  - But this has its own set of parallelism and programming challenges
  - Many porting projects to GPGPU have taken *much* longer than expected
- Homogeneity ➔ Heterogeneity

Software is leaving algorithms behind

- (Like the OS) few mathematical algorithms have been designed with parallelism in mind
  - … the parallelism is then "just a matter of implementation"
- This approach generates much duplication of effort as components are custom-built for each application
  - … but the years of development and debugging inhibits change and users are reluctant to risk a reduction in scientific output while rewriting takes place
- Exascale brings us to a "tipping point"
  - Without fundamental algorithmic changes progress in many areas will be limited
- This doesn't just apply to exascale
  - It is apparent in the vast majority of parallel codes today

Hardware for exascale

- A number of studies have looked at hardware designs for exascale
- These have identified key hardware challenges
  - Power – using today's technology we would need > 1 GWatts
  - Memory – both power and performance
  - Processor – scalability, massive parallelism and power
  - Resiliency – component failures will be continuous

- What can we draw from these studies?
  - Hardware will have to be designed against a power budget
  - Massive heterogeneous parallelism
    - Non-homogeneous computing is here
    - For GPGPUs or MIC – the challenge is the scale of the parallelism
  - Heterogeneous, highly complex memory and network architectures
- Not clear how much exascale systems will be able to influence hardware developments

System characteristics – Aggressive Strawman (2007)

Characteristic	
Flops – peak (PF)	997
- microprocessors	223,872
- cores/microprocessor	742
Cache (TB)	37.2
DRAM (PB)	3.58
Total power (MW)	67.7
Memory bandwidth (B/s per flops)	0.0025
Network bandwidth (B/s per flops)	0.0008

220 million cores !!!

CRESTA

- **C**ollaborative **R**esearch into **E**xascale **S**ystemware, **T**ools and **A**pplications
- Developing techniques and solutions which address the most difficult challenges that computing at the exascale can provide
- Focus is predominately on software not hardware
- Funded via FP7 by DG-INFSO
- Project started 1st October 2011
- Three year duration
- 13 partners, EPCC project coordinator
- €12 million costs, €8.57 million funding

Sverige
Sweden
Suomi
Finland
Gulf of Bothnia
Norge
Norway
Bergen
Oslo
Stockholm
Helsinki
Tallinn
Eesti
Estonia
Baltic Sea
Göteborg
Latvija
Latvia
North Sea
Glasgow
Edinburgh
United Kingdom
Danmark
Denmark
Malmö
Lietuva
Lithuania
Belfast
Gdańsk
Minsk
Vilnius
Ireland
Manchester
Liverpool
Hamburg
Kiel
Szczecin
Berlin
Беларусь
Belarus
Birmingham
London
Nederland
Hannover
Bydgoszcz
Polska
Poland
Dresden
België
Belgique
Belgium
Deutschland
Germany
Kraków
Česká Rep
Czech Rep
Ostrava
Celtic Sea
München
Wien
Slovensko
Slovakia
Freiburg im Breisgau
Österreich
Austria
Zürich
Vaduz
Magyarország
Hungary
France
Moldova
Zagreb
Bay of Biscay
Lyon
Milano
Beograd
România
Romania
Torino
Bologna
Hrvatska
Croatia
Genova
Србија
Serbia
Marseille
Italia
Sarajevo
București
Toulouse

CRESTA and hardware co-design

- All vendors have the same hardware challenges
- CRESTA has Cray as a hardware (and software) partner
  - We are collaborating with Cray in a hardware context
  - But our results are valid for all efforts to build exascale systems
  - … and will be publicly available
- It would be possible to build an exascale system today … there's no hardware reason why not
  - China announced it will build 2 x 100Pflop systems in next 3 years at IESP
- But the system will be unusable from a software application point of view … and almost certainly the systemware (OS, compilers, debuggers, etc.) will struggle too
- CRESTA is therefore working from a broad understanding of what exascale systems will be like and focussing its efforts on applications

Key principles behind CRESTA

- Two strand project
  - Building and exploring appropriate *systemware* for exascale platforms
  - Enabling a set of key *co-design* applications for exascale
- Co-design is at the heart of the project. Co-design applications:
  - provide guidance and feedback to the systemware development process
  - integrate and benefit from this development in a cyclical process
- Employing both incremental and disruptive solutions
  - Exascale requires both approaches
  - Particularly true for applications at the limit of scaling today
  - Solutions will also help codes scale at the peta- and tera-scales
- Committed to open source for interfaces, standards and new software

Co-design Applications

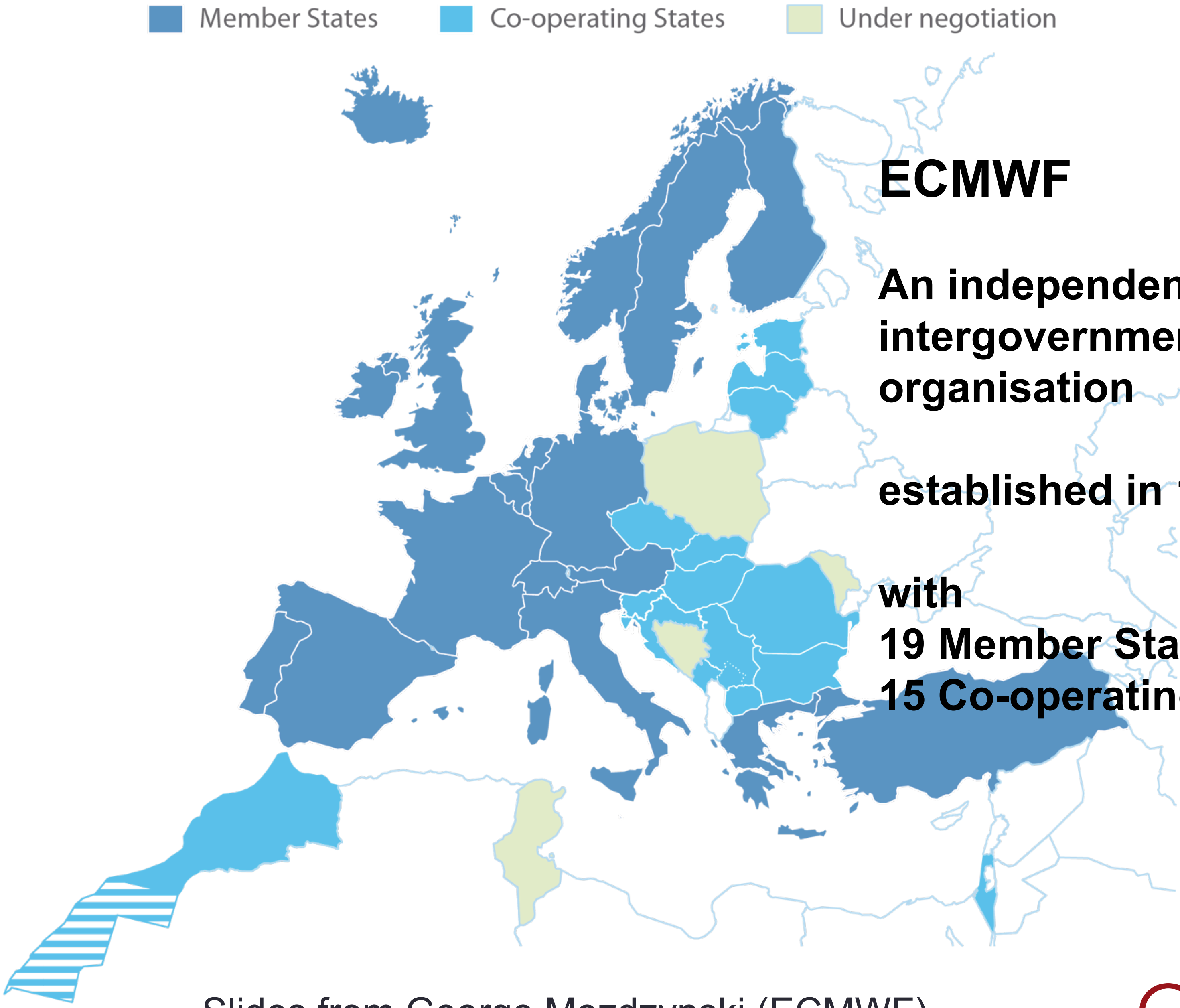
- Exceptional group of six applications used by academia and industry to solve critical grand challenge issues
- Applications are either developed in Europe or have a large European user base
- Enabling Europe to be at the forefront of solving world-class science challenges

Application	Grand challenge	Partner responsible
GROMACS	Biomolecular systems	KTH (Sweden)
ELMFIRE	Fusion energy	ABO (Finland)
HemeLB	Virtual Physiological Human	UCL (UK) / JYU (Finland)
IFS	Numerical weather prediction	ECMWF (European)
OpenFOAM	Engineering	EPCC / HLRS / ECP
Nek5000	Engineering	KTH (Sweden)

CRESTA uses incremental and disruptive approaches

- Example: FFTs are a challenge at exascale because
  - Very large number of HPC applications use them
  - Distributed memory parallel FFT is already a major performance issue today – we accept some FFTs will not scale further
- Two approaches:

Incremental approach	Disruptive approach
• Through optimisations, performance modelling and co-design application feedback • Look to achieve maximum performance at exascale and understand limitations e.g. through sub-domains, overlap of compute and communications	• Work with co-design applications to consider alternative algorithms • Crucial we understand maximum performance before very major application redesigns undertaken



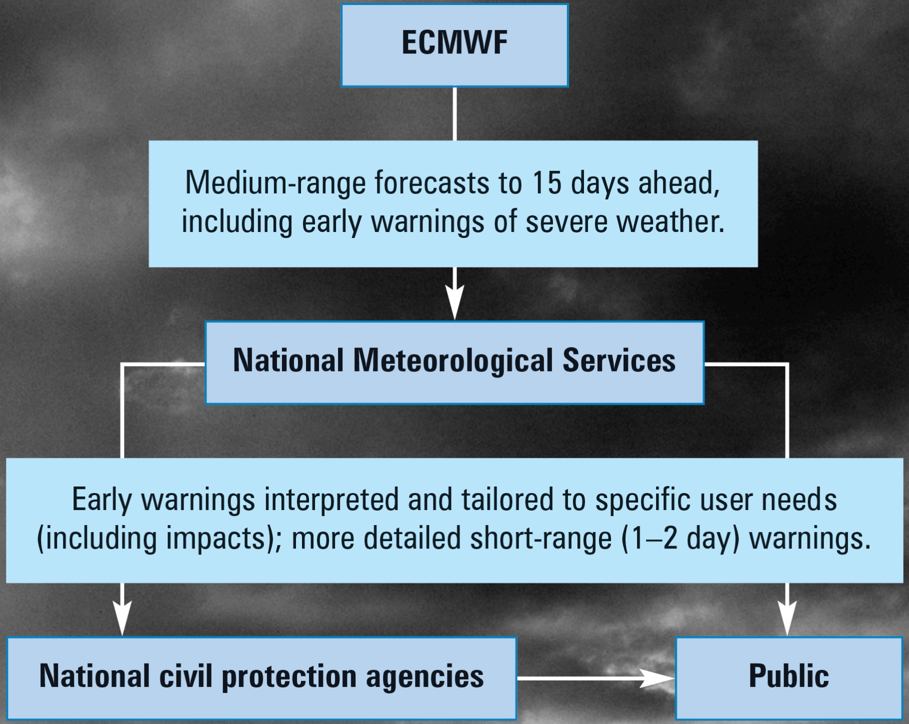
ECMWF

An independent intergovernmental organisation

established in 1975

with
19 Member States
15 Co-operating States

Slides from George Mozdzynski (ECMWF)



ECMWF

Medium-range forecasts to 15 days ahead, including early warnings of severe weather.

National Meteorological Services

Early warnings interpreted and tailored to specific user needs (including impacts); more detailed short-range (1–2 day) warnings.

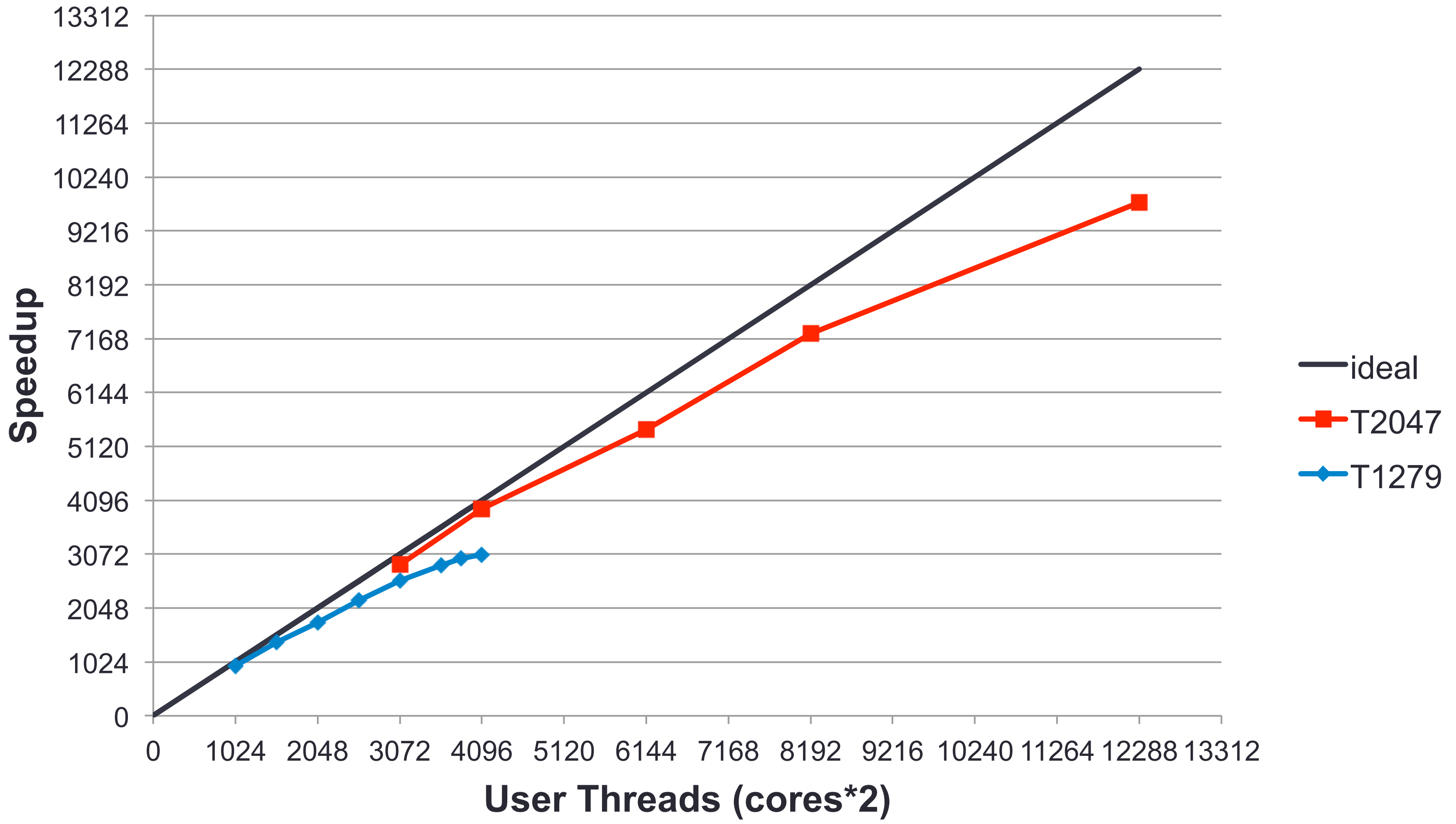
National civil protection agencies

Public

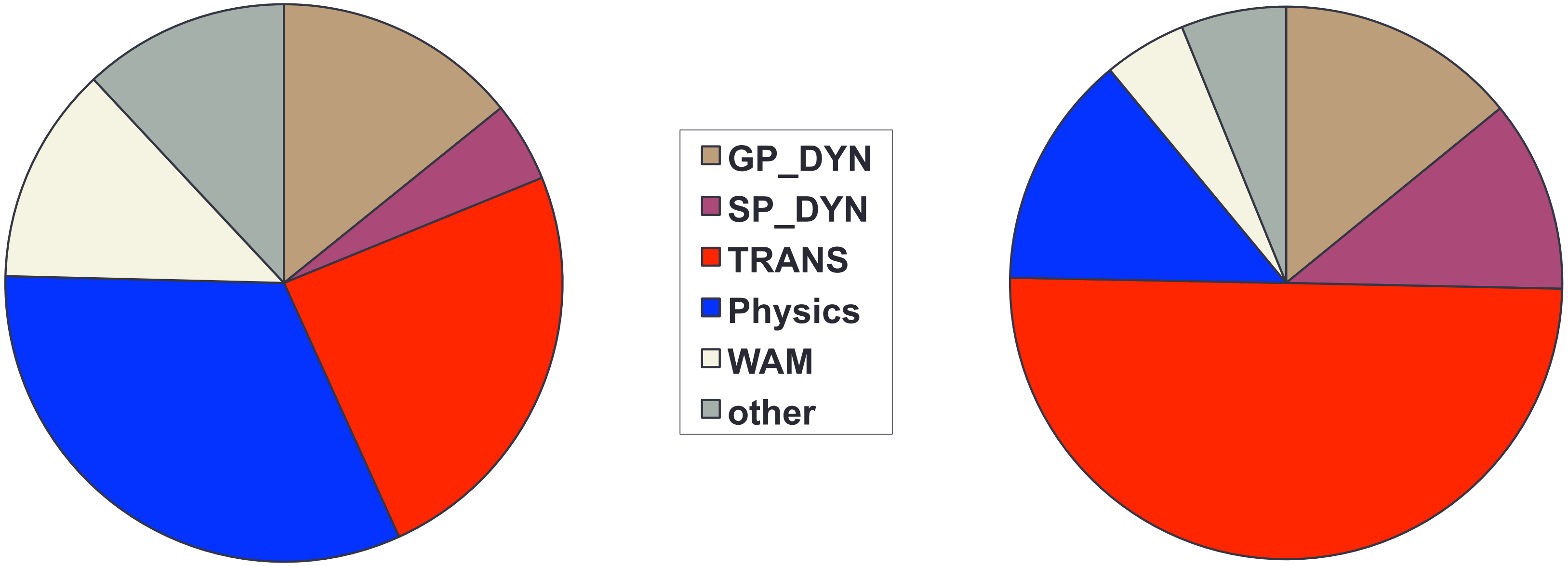
IFS model: current and planned model resolutions

IFS model resolution	Envisaged Operational Implementation	Grid point spacing (km)	Time-step (seconds)
T1279	2010	16	600
T2047	2014-2015	10	450
T3999	2020-2021	5	240
T7999	2025-2026	2.5	120

IFS model speedup on IBM Power6 (~2010)



Computational Cost at T2047 and T3999



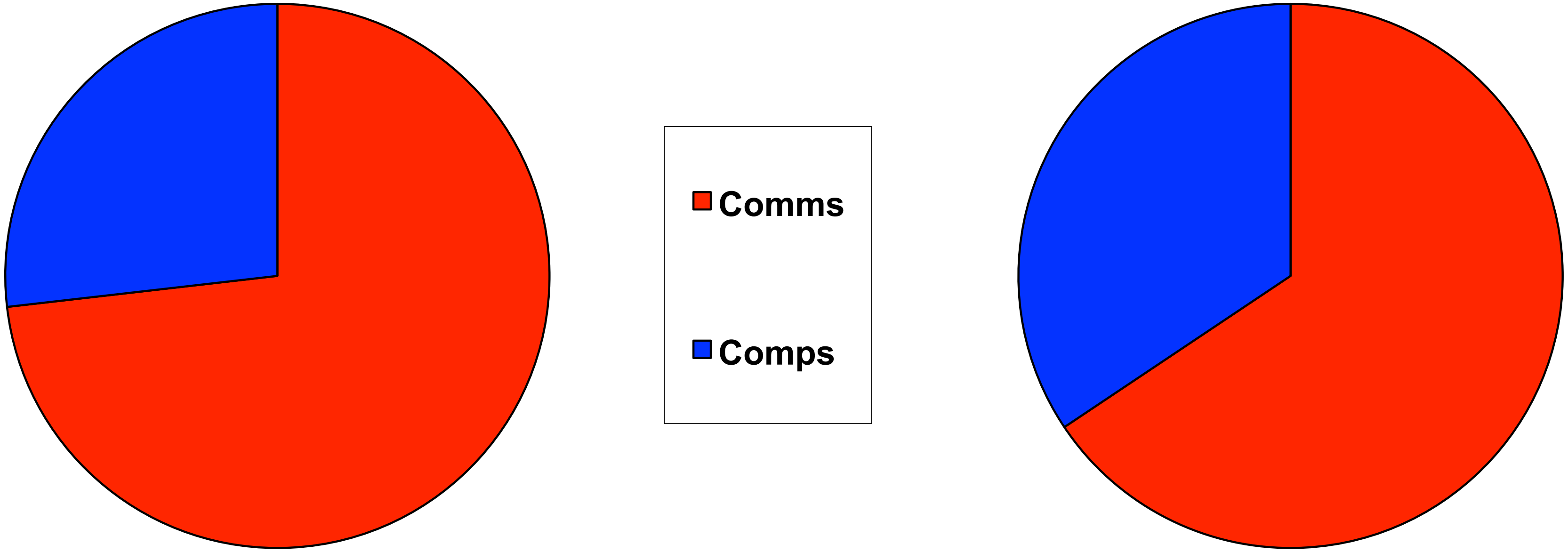
Hydrostatic T_L2047

Tstep=450s, 5.8s/Tstep
With 256x16 ibm_power6

Non-Hydrostatic T_L3999

Tstep=240s, 13.6s/Tstep
With 512x16 ibm_power6

Breakdown of TRANS cost: Computations vs. Communications



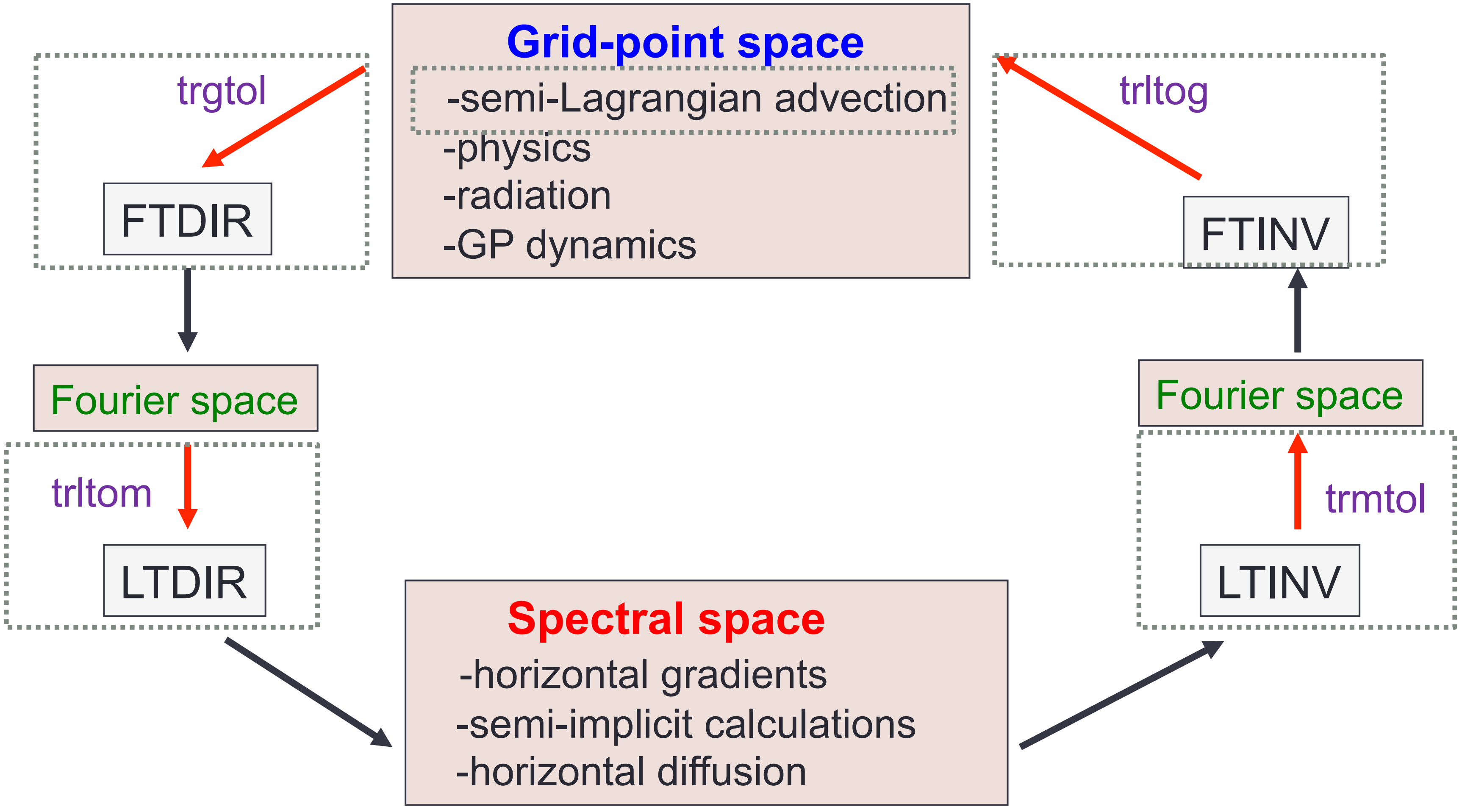
H T_L2047 ~2015

NH T_L3999 ~2020

Data sent/received: 117.8GB/s

Data sent/received: 289.6GB/s

Planned IFS optimisations for [Tera,Peta,Exa]scale



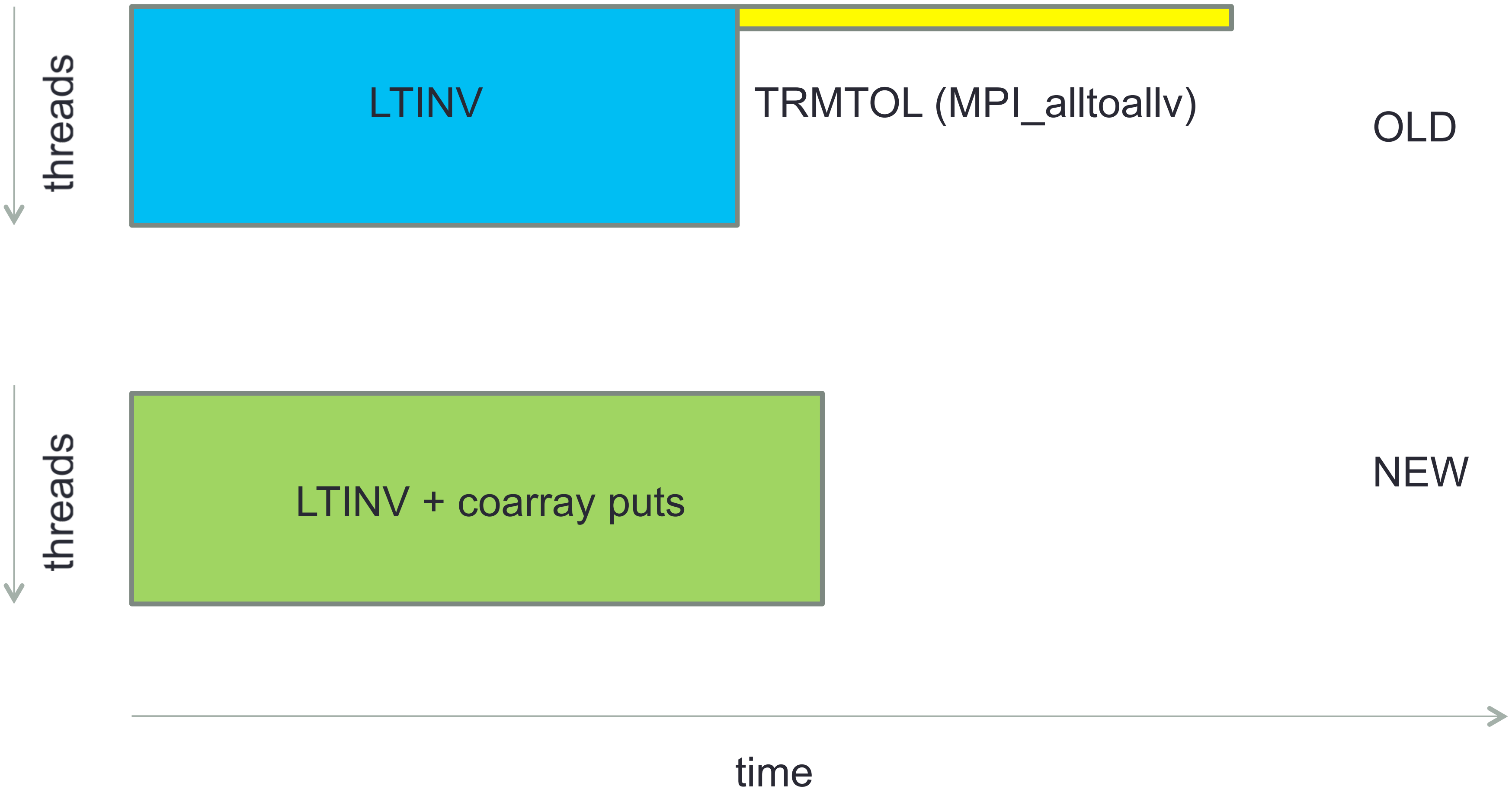
Grid-point space

- -semi-Lagrangian advection
- -physics
- -radiation
- -GP dynamics

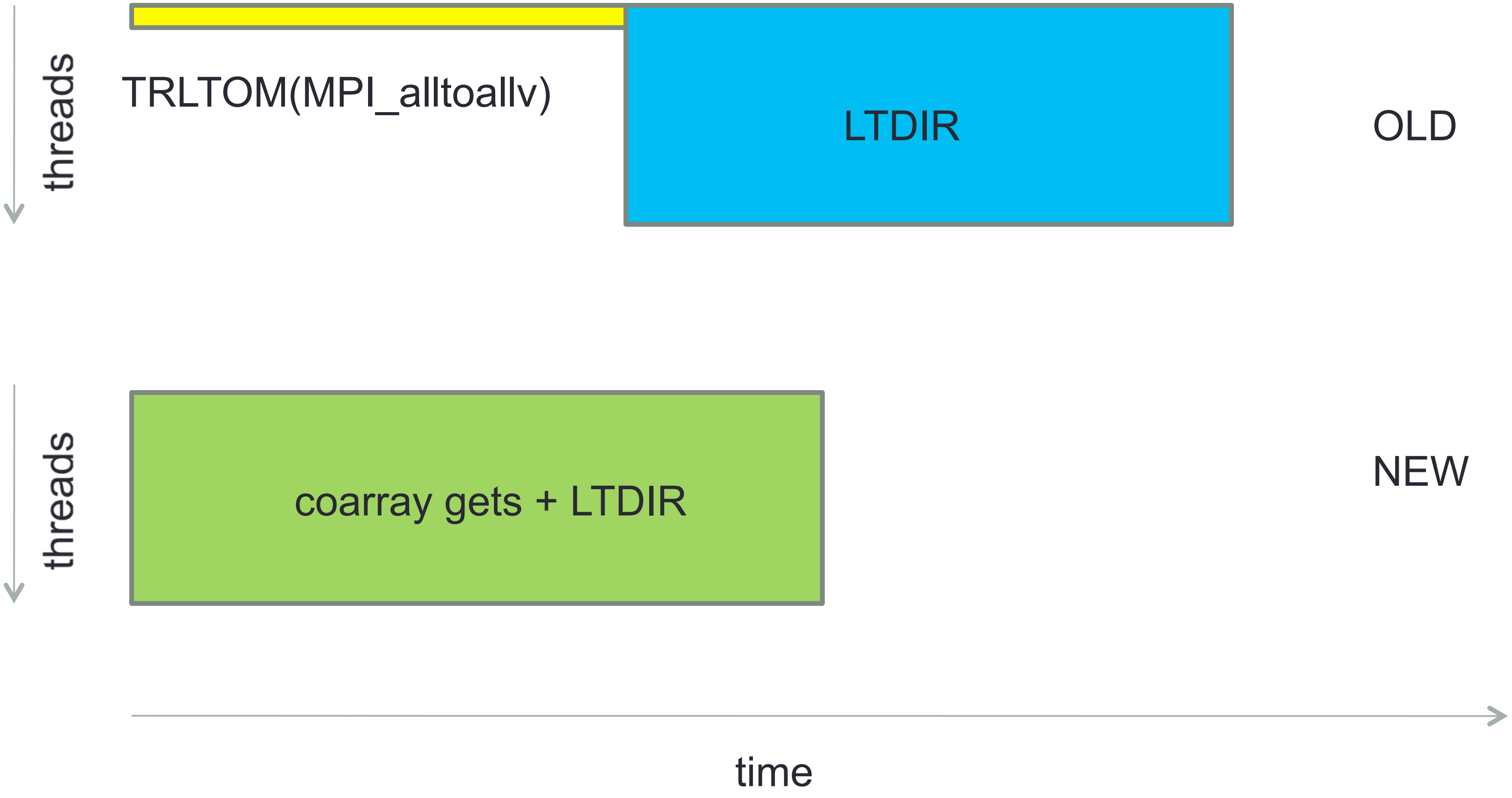
Spectral space

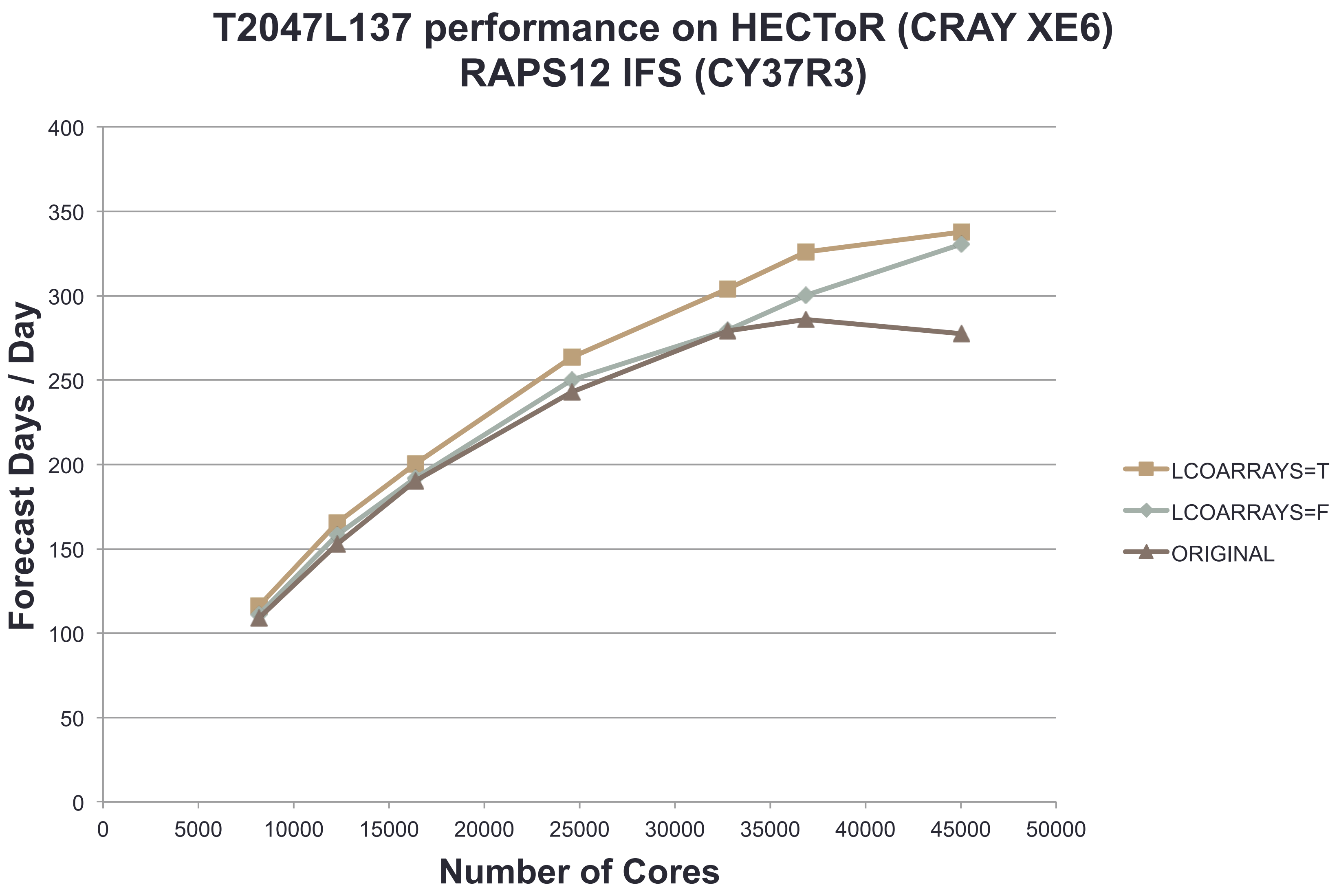
- -horizontal gradients
- -semi-implicit calculations
- -horizontal diffusion

Overlap Legendre transforms with associated transpositions



Overlap Legendre transforms with associated transpositions/2




T2047L137 performance on HECToR (CRAY XE6)
RAPS12 IFS (CY37R3)
Forecast Days / Day
0
50
100
150
200
250
300
350
400
Number of Cores
0
5000
10000
15000
20000
25000
30000
35000
40000
45000
50000
LCOARRAYS=T
LCOARRAYS=F
ORIGINAL

Final words

- HECToR has been a challenging, exciting service to deliver
- It’s grown from 12,000 to 90,000 cores
- Huge variety of science is performed on HECToR every day

- But parallel supercomputing in the next decade faces many challenges
- We’ve reached the Petascale incrementally – we can’t take the same route to Exascale
- Supercomputing faces its biggest challenge since the 1980s

… when will Edinburgh host an Exascale computer?

Thank you!