資料4

エクサに向けたプロジェクト（欧州編）

東京工業大学
学術国際情報センター
松岡　聡

文部科学省
第２回HPCI検討ワーキンググループ　プレゼン資料
2012年5月14日（月）

Europe's investments

TABLE 2

GDP and Supercomputer Spending by Country (GDP: €000,000; Sales €000)

	GDP (1)	Average Supercomputer Sales Over Last Five Years (2)	Supercomputers as a Percentage of GDP	Compared to the U.S. = 100%
U.S.	10,949,000	979,126	0.0089%	100%
Europe	10,201,000	502,074	0.0049%	55%
Japan	3,874,000	212,070	0.0055%	62%
China	3,651,000	52,050	0.0014%	16%
Korea	614,070	51,569	0.0083%	93%
Hong Kong	160,200	11,886	0.0074%	83%
Singapore	140,500	12,525	0.0100%	112%

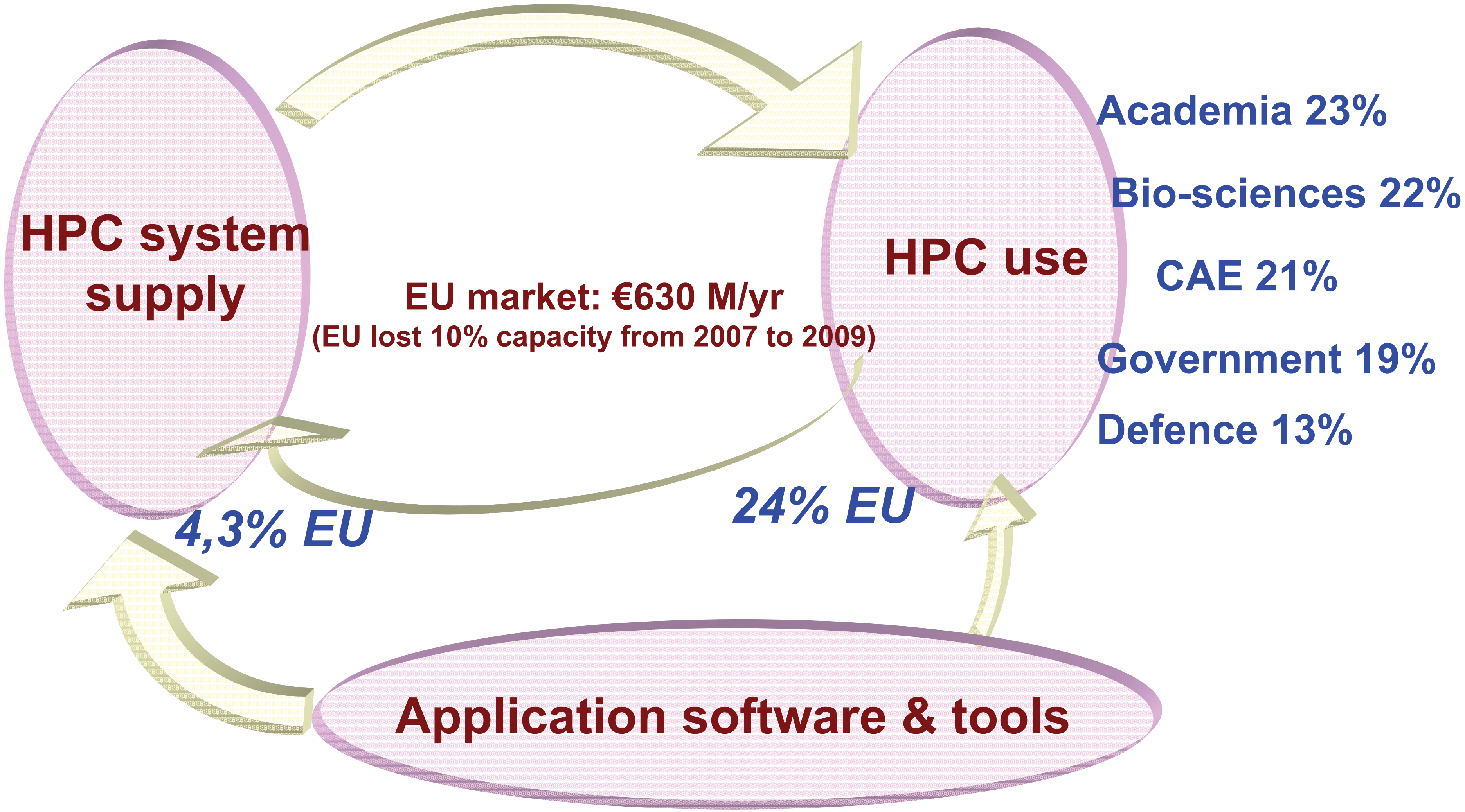
Only half of US GDP spending

Notes: (1) source: CIA World Factbook, 2009, (2) five-year average yearly spending. Supercomputing data includes server spending only

Source: IDC, 2010

HPC state of play

HPC Ecosystem



EU Projects Towards Exascale

- **FP7-INFRA-2010.2.3.1 – First Implementation Phase of the European HPC service PRACE (PRACE1)**
  - Pan-European facilitation and operation of 1-10 Petaflops supercomputers (Tier-0)
- **FP7-INFRA-CSA(Coordination and Support Action) EESI (European Exascale Software Initiative)** → **EESI2に継続**
  - Create exascale software roadmap, contribute to IESP
- **FP7-INFRA-2010.2-RI- Structuring the European Research Area (PRACE2)**
  - Evolution of DEISA2, sub- to petaflop centers, direct coordination of governance and operations with PRACE1
- **FP7-ICT-2011.9.13 Exa-scale computing**
  - R&D Towards European Exascale SC, 2011-2014
- **FP7-INFRA-2012-2.3.1 Third implementation phase of the European High-Performance Computing (HPC) service PRACE**
  - Continuation of PRACE1, join w/PRACE2 for PRACE RI

FP7-ICT-2011.9.13 Exa-scale computing 25 million Euro, 2011-2014

- Must involve major PRACE centers and tech. vendors
- 60% systems (HW/SW), 40% apps, 2014 prototype deliverable
- 6 submitted, 3 accepted
- 1. Julich-Intel-others DEEP Project
  - Hybrid Intel MIC(Booster)-Cluster architecture, Extoll network, warm water cooling(45C), scalable hybrid software stack
- 2. BSC-ARM-others Mont-Blanc
  - Multi-Chip bonding of ~75 high-performance ARM Cortex A9 processors per socket + GPUs, next-gen 40-100Gbps Ethernet switch chips, hybrid execution model
- 3. HLRS-Cray CRESTA
  - All software: scalable next generation software stack and application scaling (e.g., ECMWF-xxx, GROMACS, …)---monitoring, autotuning, PGAS compilers,
- 日本の影は<u>全く</u>ない

PRACE RI

- PRACE1: Tier-0 “PRACE” Centers
- PRACE2: Tier-1 (Regional, National) Centers
- PRACE3: Unify PRACE1&2
  - Unified Operation (like HPCI and NSF XSEDE)
  - PRACE Prototype
  - Pre-Competitive Procurement

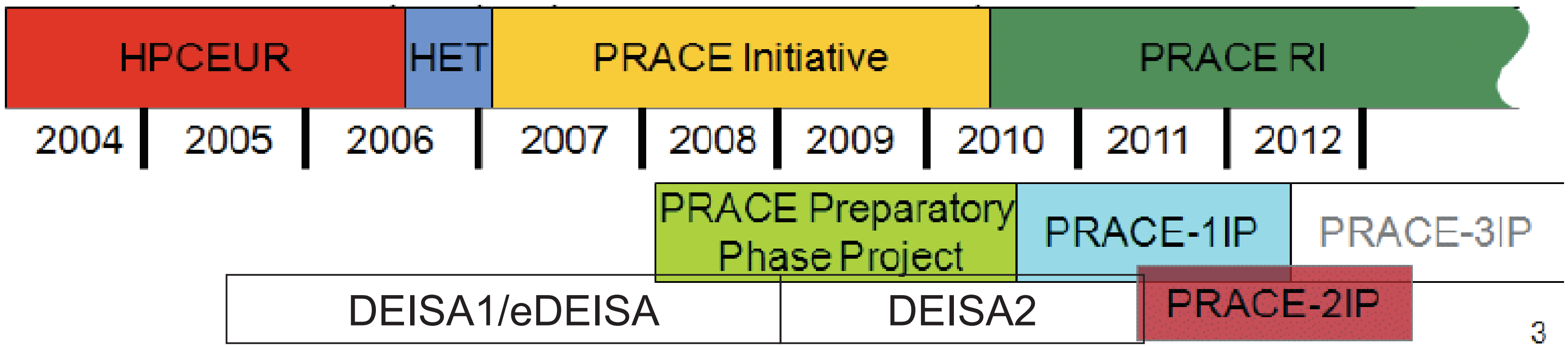


Figure 1: Evolution and history of the PRACE Research Infrastructure

From EESI to EESI2

EESI roadmaps, vision and recommendations need to be monitored, updated, on a dynamical way ... AND new issues to be addressed:

- **Extend, refine, and update** Exascale cartography (**directly in the icated WG for better analysis of each WG**) and **roadmaps** from HPC community, on software, tools, methods, R&D and industrial applications, ... *With a Gap Analysis.*
  - Including WG on ***disruptive technologies***
- **Address "Cross Cutting issues":** Data management and ploration, Uncertainties - UQ&VQ, Power & Performance, Resilience, Disruptive technologies
- **Investigation on** funding scheme and opportunities, education, -design centres, international coordination
- Operational **Software maturity** level methodology, evaluation

EESI2 Main Focus

Main focus of EESI2 are the key issues identified by EESI1 experts' panel :

Cross Cutting issues

- Power management: A power supply reduction (a factor of 50 must be achieved)
- Performance optimization, programmability, load balancing
- Fault tolerance, resilience: developing software or API (fault tolerance independently of users)
- Reproducibility, ***Uncertainties***: many phenomena studied can exhibit chaotic behaviours.
- ***Data management***: Big data, Data placement and memory access , I/O parallel, Storage ...

Software technologies issues for strong and weak scalability:

- Numerical analysis: new efficient solvers/algebra libraries, automatic massively parallel mesh-generation tool, meshless methods and particle simulation,
- Scientific software engineering: platform, standard coupling interfaces and software tools mixing legacy and new generation codes for Multi-physics, multi scale simulation,
- Coupling between stochastic and deterministic methods, UQ approach

Based on the recognized EESI1 network expertises, the **EESI2 objectives** will be to go a step forward

- in the **Exascale dynamic vision and roadmap, recommendations** for Europe
- in the proposition of **benchmarks** and **methodologies** to validate the incremental progress and **breakthrough,**
- in the **gap analysis** to reach the targeted objectives, to periodically estimate maturity, innovation

Dynamical Exascale Entry Platform (DEEP)

- Duration: 3 years
- Start: December 2011
- Overall budget: 18 Mio€
- EU-funding: 8 Mio€
- Coordinator: Jülich Supercomputing Centre
- Partners: 16 from Europe and Israel

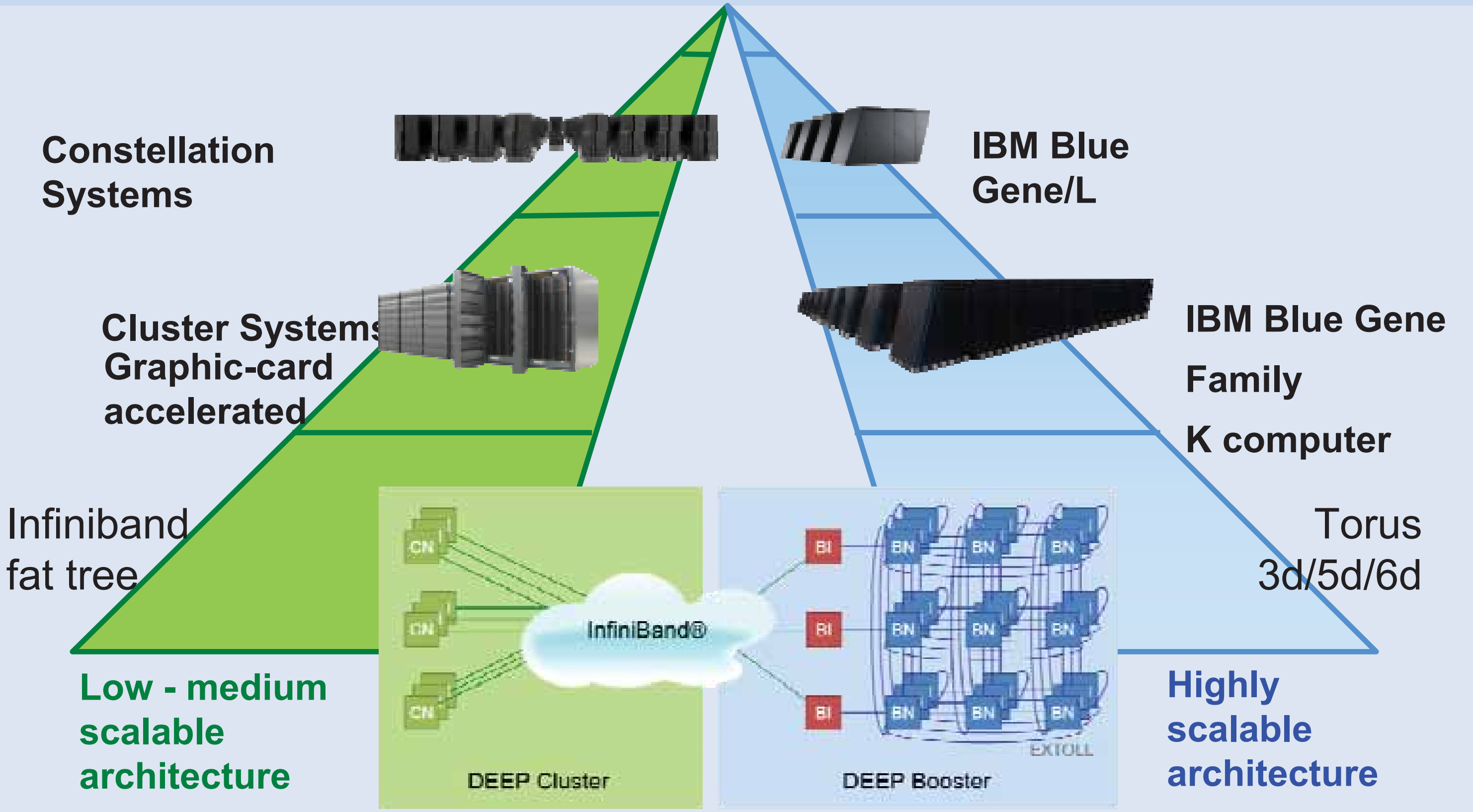
DEEP is one of three EU-founded Exascale Projects

The research leading to these results has received funding from the European Community's Seventh Framework Programme (*FP7/2007-2013*) under *Grant Agreement* n° 287530

Project Goals

- Build a prototype from an Exascale architecture:
  - With accelerators that can work and react autonomously → "Booster"
- Hardware Development:
  - Build a Booster based on Intel MIC and EXTOLL torus network
  - Energy efficient system using "hot water" cooling
- Software Development
  - Ressource-Management System
  - Programming environment, Programming models
  - Libraries and Performance analysis tools
- Porting scientific applications to demonstrate the concept

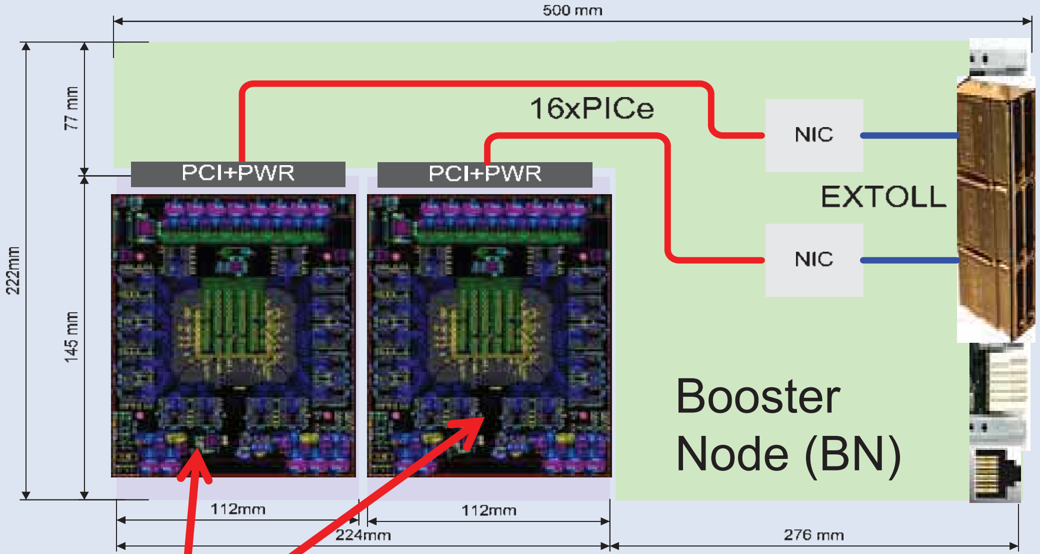
DEEP Concept



- The DEEP system is a combination of a compute cluster and a “Booster” which is a cluster of accelerators.
- A program with medium and highly scalable code parts runs on the entire system
- The highly scalable code parts will be offloaded to the Booster

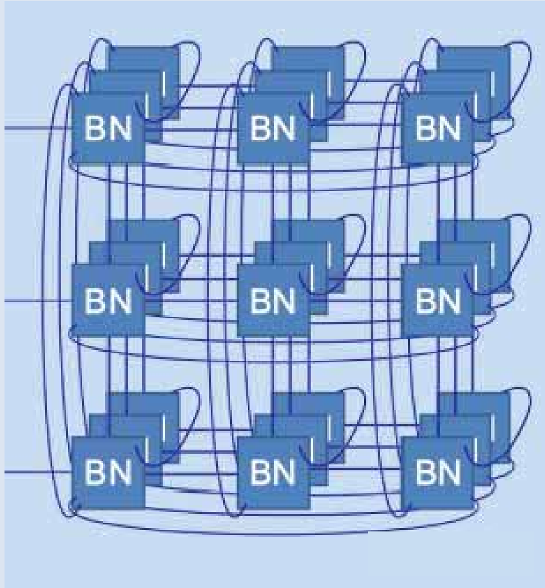
Booster Hardware

- Booster Nodes based on Intel Knights Corner processors (KNC)
- No extra CPU needed for PCIe initialization and KNC-booting



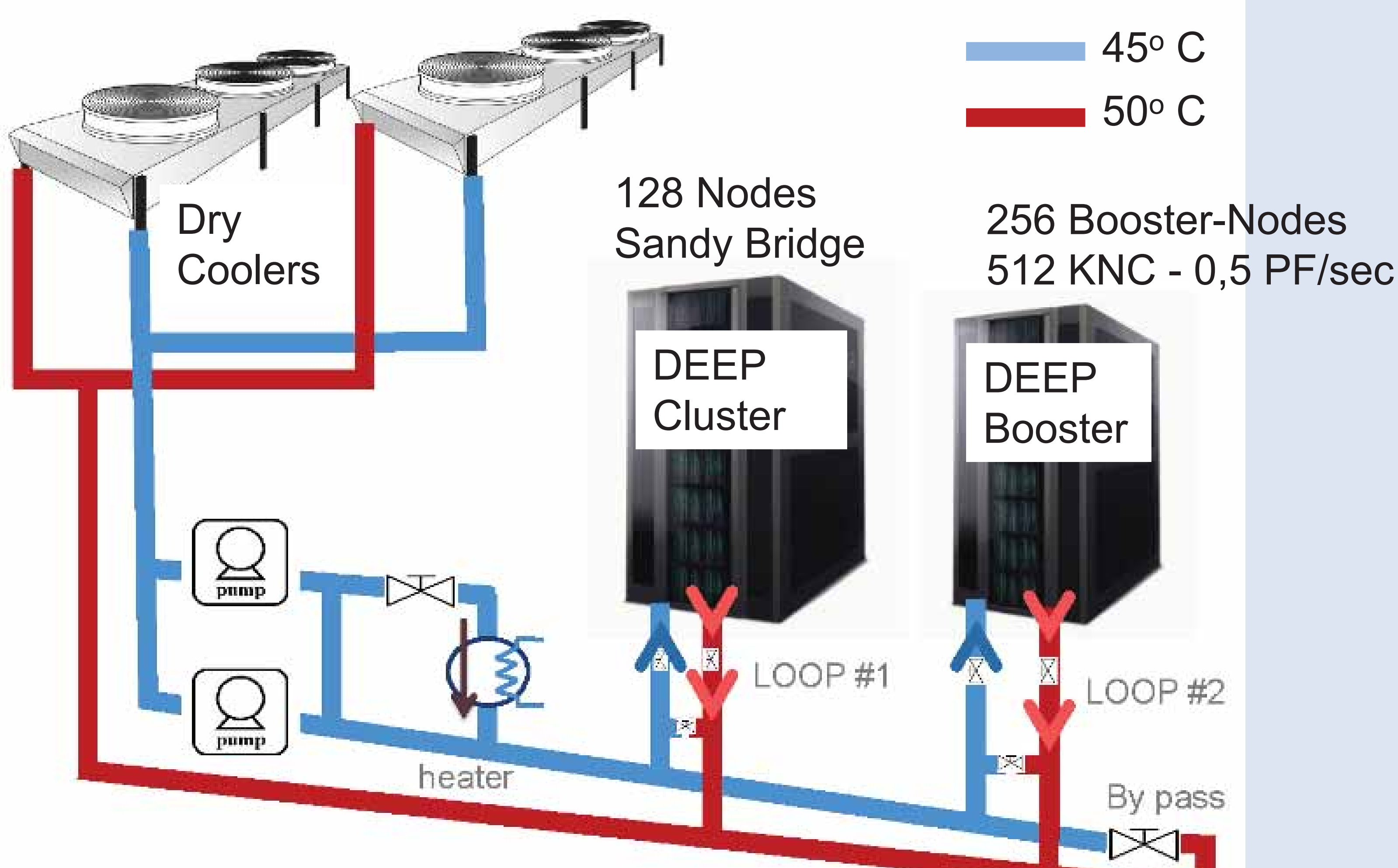
2x KNC with >50 cores each

- Booster Interconnect based on ExToll
- 3D torus topology
- NIC integrates an 8 port switch



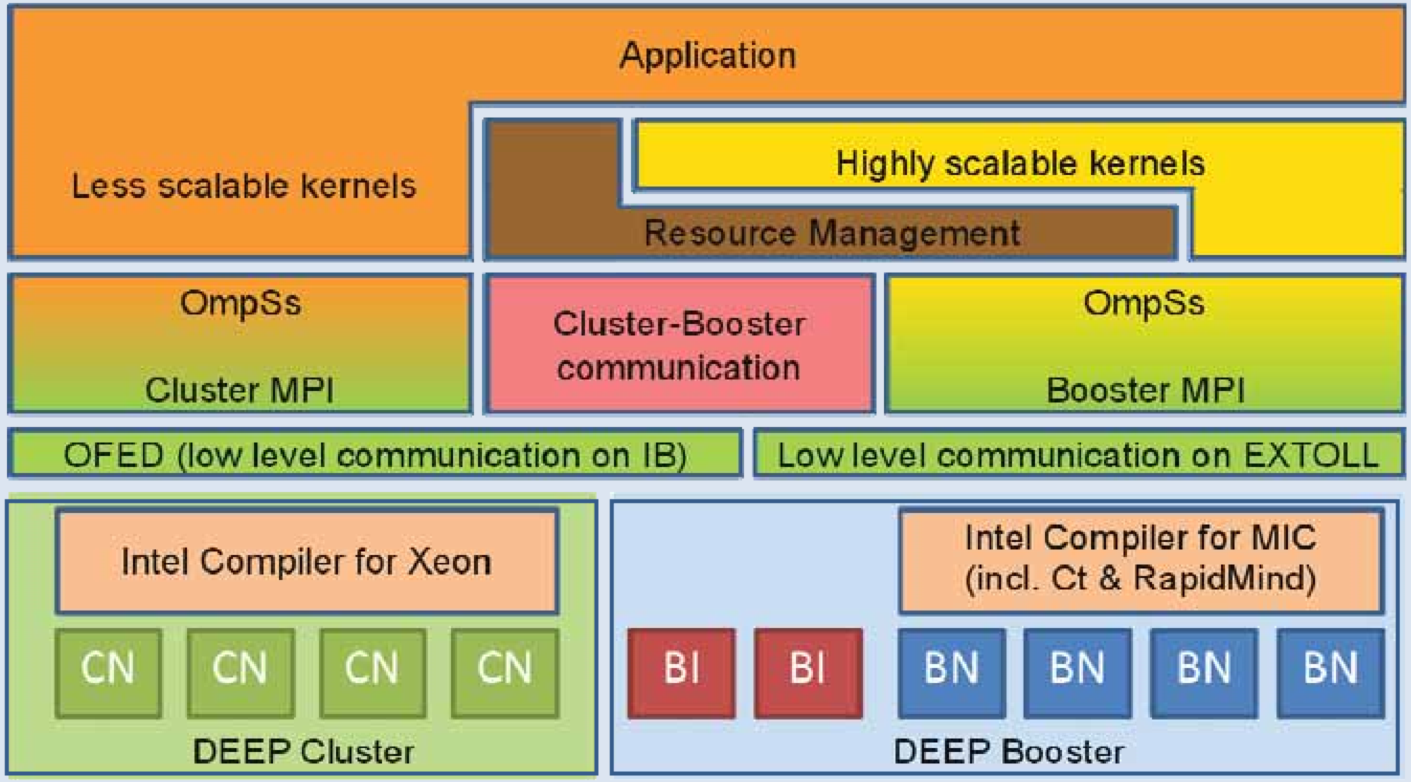
Very low latency: < 1 µsec
Large bandwidth: 32 Gbit/s

Hot-water cooling



Energy saving by „free cooling“

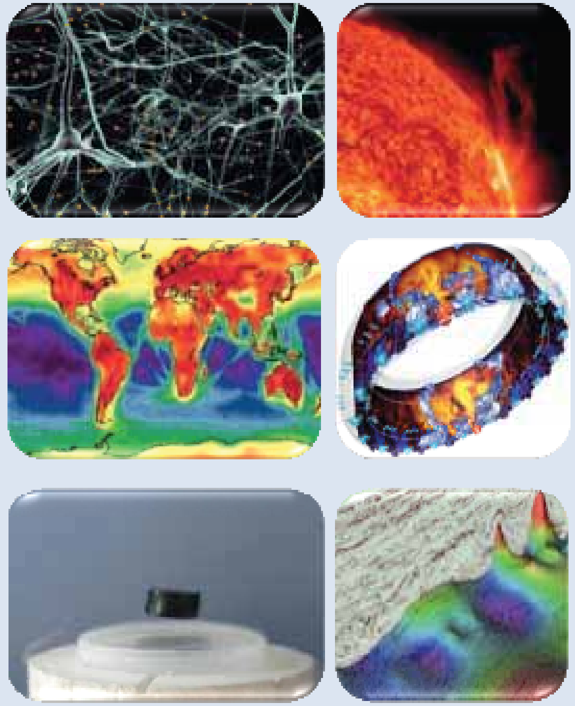
Software Architecture



Software stack provides:

- Communication between all Booster nodes and Cluster and Booster nodes
- Programming Model OmpSs with underlying MPI and OpenMP
- Dynamical allocation of sets of Booster nodes

DEEP Pilot Applications



- Pilot applications:
  - Brain simulation (EPFL)
  - Space weather simulation (KULeuven)
  - Climate simulation (CYI)
  - Computational fluid engineering (CERFACS)
  - High temperature superconductivity (CINECA)
  - Seismic imaging (CGGVS)
- Goals:
  - Evaluation of DEEP concept and its programmability
  - Performance comparison with standard architectures
  - Create of a best practice guide
  - Propose improvements to the DEEP system

Mont-BlancProject goal

- To develop an **European** exascale approach
- Based on embedded **power-efficient technology**

- Funded under FP7 Objective ICT-2011.9.13 Exa-scale computing, software and simulation
  - 3-year IP Project (October 2011 - September 2014)
  - Total budget: 14.5 M€ (8.1 M€ EC contribution),
    - 1095 Person-Month

Project objectives

- Objective 1: To deploy a **prototype HPC system** based on currently **available energy-efficient embedded technology**
  - Scalable to 50 PFLOPS on 7 MWatt
    - Competitive with Green500 leaders in 2014
  - Deploy a full HPC system software stack
- Objective 2: To design a next-generation HPC system and **new embedded technologies** targeting HPC systems that would overcome most of the limitations encountered in the prototype system
  - Scalable to 200 PFLOPS on 10 MWatt
    - Competitive with Top500 leaders in 2017
- Objective 3: To port and optimise a small number of **representative exascale applications** capable of exploiting this new generation of HPC systems
  - Up to 11 full-scale applications

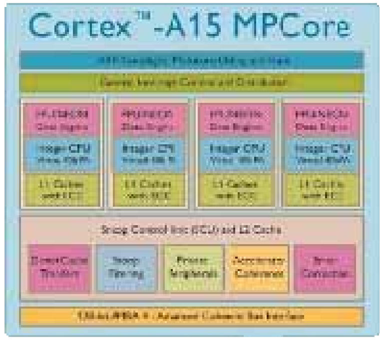
Power defines performance

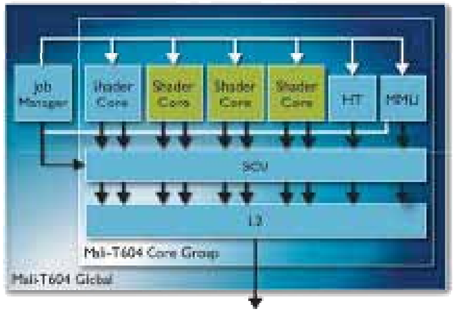
- Prototype goal: 50 PFLOPS on 7 MWatt
  - 7 GFLOPS / Watt efficiency
- Required improvement on energy efficiency
  - 3.5x over BG/Q
  - 5x over ATI GPU systems
  - 7x over Nvidia GPU systems
  - 8.5x over SPARC64 multi-core
  - 9x over Cell systems

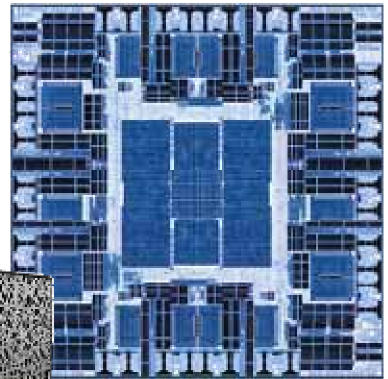
Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
1	2097.19	IBM Thomas J. Watson Research Center	NNSA/SC Blue Gene/Q Prototype 2	40.95
2	1684.20	IBM Thomas J. Watson Research Center	NNSA/SC Blue Gene/Q Prototype 1	38.80
3	1375.88	Nagasaki University	DEGIMA Cluster, Intel i5, ATI Radeon GPU, Infiniband QDR	34.24
4	958.35	GSIC Center, Tokyo Institute of Technology	HP ProLiant SL390s G7 Xeon 6C X5670, Nvidia GPU, Linux/Windows	1243.80
5	891.88	CINECA / SCS - SuperComputing Solution	iDataPlex DX360M3, Xeon 2.4, nVidia GPU, Infiniband	160.00
6	824.56	RIKEN Advanced Institute for Computational Science (AICS)	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect	9898.56
7	773.38	Forschungszentrum Juelich (FZJ)	QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus	57.54

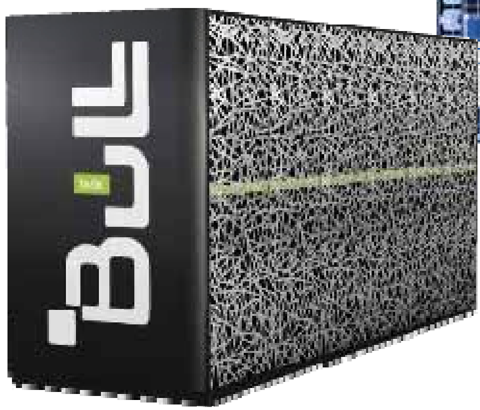
Energy-efficient building blocks

- Integrated system design built from mobile / embedded components
- ARM multicore processors
  - Nvidia Tegra / Denver, Calxeda, Marvell Armada, ST-Ericsson Nova A9600, TI OMAP 5, …
- Mobile accelerators
  - Mobile GPU
    - Nvidia GT 500M, …
  - Embedded GPU
    - Nvidia Tegra, ARM Mali T604
- Low power 10 GbE switches
  - Gnodal GS 256
- Tier-0 system integration experience
  - BullX systems in the Top10





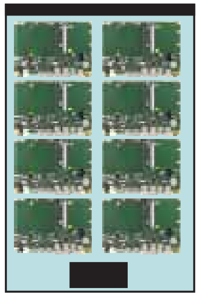




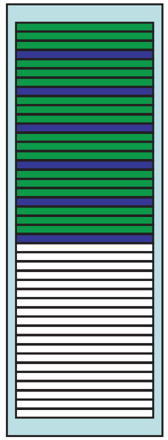
PRACE prototype @ BSC: ARM multicore

Tegra2 SoC:
2x ARM Corext-A9 Cores
2 GFLOPS
0.5 Watt

Tegra2 Q7 module:
1x Tegra2 SoC
2x ARM Corext-A9 Cores
1 GB DDR2 DRAM
2 GFLOPS
~4 Watt
1 GbE interconnect



1U Multi-board container:
1x Board container
8x Q7 carrier boards
8x Tegra2 SoC
16x ARM Corext-A9 Cores
8 GB DDR2 DRAM
16 GFLOPS
~35 Watt

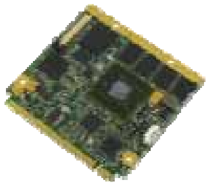


Rack:
32x Board container
10x 48-port 1GbE switches
256x Q7 carrier boards
256x Tegra2 SoC
512x ARM Corext-A9 Cores
256 GB DDR2 DRAM
512 GFLOPS
~1.7 Kwatt

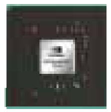
300 MFLOPS / W

- First large-scale ARM cluster prototype
- Proof-of-concept to demonstrate HPC based on low-power components
  - Built entirely from COTS components
  - Mont-Blanc integrated design could improve substantially
- Enabler for early software development and tuning
  - Open-source system software stack
  - Application development and tuning to ARM platform

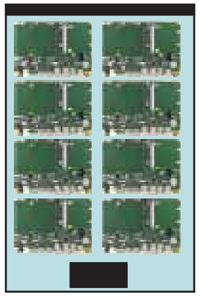
PRACE prototype @ BSC: ARM + mobile GPU



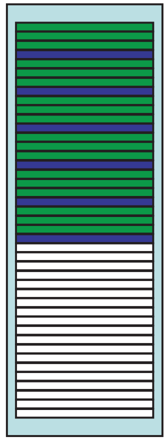
Tegra3 Q7 module:
1x Tegra3 SoC
  4x Corext-A9 @ 1.5 GHz
4 GB DDR3 DRAM
6 GFLOPS
~4 Watt
1 Gbe interconnect



Nvidia GeForce 520MX
48 CUDA cores @ 900 MHz
142 GFLOPS
12 Watts
11.8 GFLOPS / W



1U Multi-board container:
1x Board container
8x Q7 carrier boards
  32x ARM Corext-A9 Cores
  8x GT520MX GPU
32 GB DDR3 DRAM
1.2 TFLOPS
~140 Watt



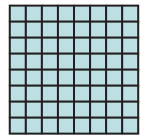
Rack:
32x Board container
10x 48-port 1GbE switches
256x Q7 carrier boards
  256x Tegra3 SoC
  1024x ARM Corext-A9 Cores
  256x GT520MX GPU
  1TB DDR3 DRAM
38 TFLOPS
~5 Kwatt

7.5 GFLOPS / W

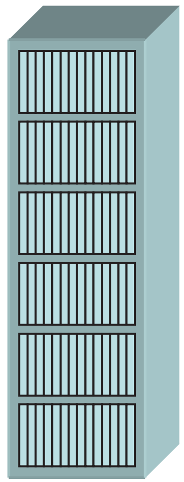
- Increasing number of Top500 systems use GPU accelerators
- Validate the use of their energy efficient counterparts
  - ARM multicore processors
  - Mobile Nvidia GPU accelerators
- Perform scalability tests to high number of compute nodes
  - Higher core count required when using low-power processors
  - Evaluate impact of limited memory and bandwidth on low-end solutions

Prototype architecture (reverse engineering)

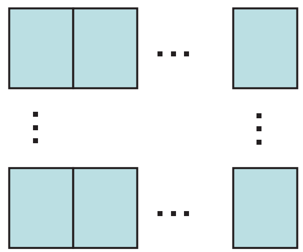
- 50 PFLOPS on 7 MWatt
- ARM Cortex-A15 CPU
  - 4 ops/cycle @ 2GHz = 8 GFLOPS
  - 65% efficiency = 5.2 GFLOPS
- Blade-based system design
  - 108 blades / rack
  - 12 sockets / blade
  - 5 Watts / socket
- 50 PFLOPS / 5.2 GFLOPS
  - 10 M cores
- 32% of 7 MWatt = 227 KWatt
  - 227 KWatt / 10 Mcores
  - 0.23 Watts / core
- 5 Watt / socket
  - 22 cores / socket
  - 4 cores + GPU / socket
  - 112 GFLOPS / socket



Multi-core chip:
5 Watts
112 GFLOPS
5.2 GFLOPS / core
4 cores + GPU / chip
0.23 Watts / core



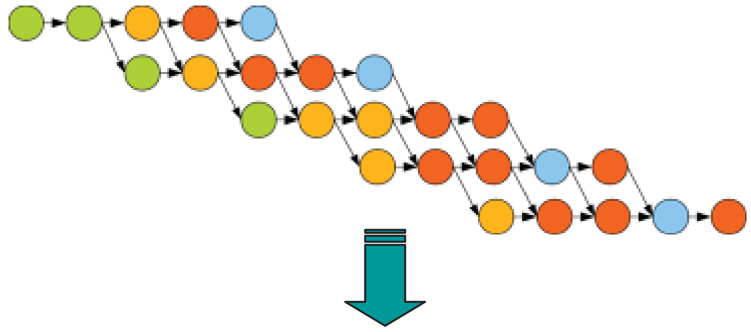
Rack:
108 compute nodes
1.296 chips
18 Kcores
140 TFLOPS
18-20 Kwatts

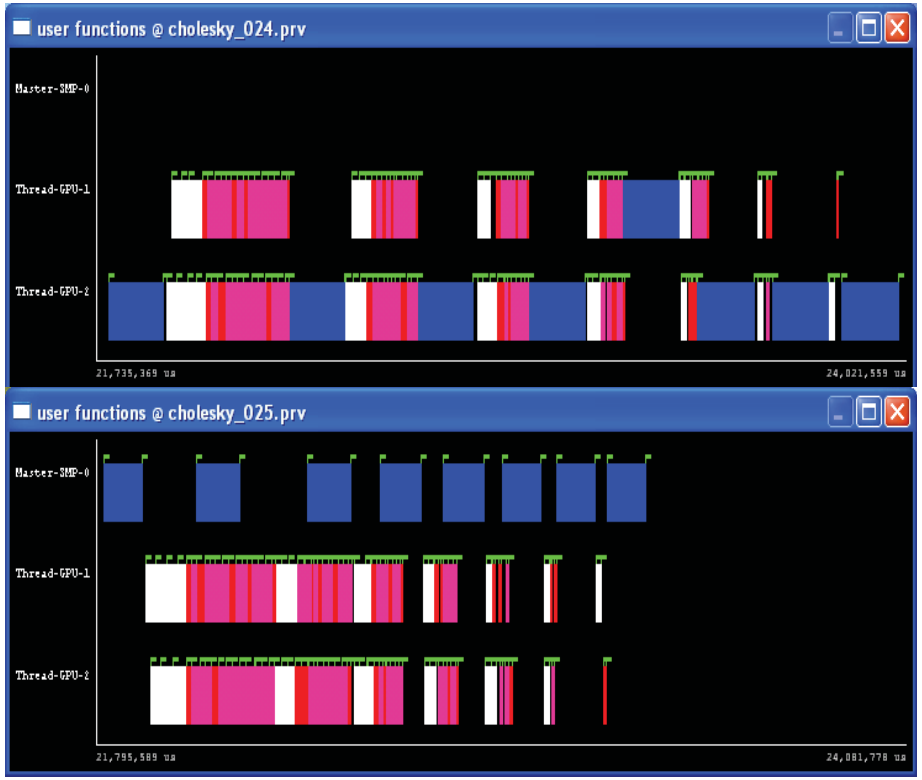


Full system:
352 racks
19 K blades
10 M cores
50 PFLOPS
7 MWatts

Hybrid MPI + OmpSs programming model

- Hide complexity from programmer
- Runtime system maps task graph to architecture
  - Many-core + accelerator exploitation
  - Asynchronous communication
    - Overlap communication + computation
  - Asynchronous data transfers
    - Overlap data transfer + computation
  - Strong scaling
    - Sustain performance with lower memory size per core
  - Locality management
    - Optimize data movement





System software porting + tuning

- Linux OS
- Filesystem
  - NFS, Lustre
- Parallel programming model + Runtime libraries
  - OmpSs, OpenMP, MPI, OpenCL
- Scientific libraries
  - ATLAS, FFTW, HDF5, LAPACK, MAGMA, ...
- Performance tools
  - Hardware performance counters
  - EXTRAE, PARAVER, SCALASCA
- Cluster management
  - Slurm, Ganglia

Target Mont-Blanc applications

- Real applications currently running in PRACE Tier-0 systems or National HPC facilities
  - YALES2 — Fluid Dynamics
  - EUTERPE — Fluid dynamics
  - SPECFEM3D — Seismic wave propagation
  - MP2C — Multi-particle collisions
  - BigDFT — Electronic structure
  - QuantumESPRESSO — Electronic structure
  - PEPC — Coulomb + gravitational forces
  - SMMP — Protein folding
  - ProFASI — Protein folding
  - COSMO — Meteorological modeling
  - BQCD — Quantum ChromoDynamics

Project results

- Prototype HPC system based on European embedded processors
  - Demonstrate potential of embedded technology for HPC
  - Target maximum power efficiency
  - Limited by currently available technology
- Design of a next-generation system
  - Full scale system paving the way towards Exascale computing
  - Proposal and definition of the required technologies to achieve it
- Open source system software stack
  - Operating system, runtime libraries, scientific libraries, performance tools
- Up to 11 full-scale scientific applications
  - Capable of exploiting the benefits of this new class of HPC architectures

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

http://www.cresta-project.eu

Partnership

- Consortium
  - Leading E … owners and
    - EPCC –
    - HLRS – … sity – Abo,
    - CSC – E … – Jyvaskyla,
    - KTH – S
  - A world lea … ondon –
    - Cray UK … UK
  - World lead … – Paris, France
    - Technisc … many (Vampir)
    - Allinea L

CREST

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)

CREST

Systemware

- Software components required for grand challenge applications to exploit future exascale platforms
- Underpinning and cross cutting technologies
  - Operating systems, fault tolerance, energy, performance optimisation
- Development environment
  - Runtime systems, compilers, programming models and languages including domain specific
- Algorithms and libraries
  - Key numerical algorithms and libraries for exascale
- Debugging and Application performance tools
  - World leaders in Allinea's DDT, TUD's Vampir and KTH's perfminer
- Pre- and post- processing of data resulting from simulations
  - Often neglected, hugely important at exascale

CRESTA

終わりに

- 欧州はPRACEで複数のペタスケールスパコンおよびその共通運用により計算科学の研究基盤(RI)の拡充に成功
  – HPCIの先駆的活動、米TeraGridよりも成功
  – PRACEの一部として欧州企業との次世代のR&D
- 更にEXAプロジェクトにおいては、欧州のIT及びHPCハード・ソフト技術を推進しエクサを目指し、米国に全面依存はしない技術開発を推進
  – 欧州独自コンペテンスの同定：ARM・低電力ネットワーク等の組み込み系、EXTOLLネットワーク、システムソフトウェア/ツールやアプリ
  – 欧州のスパコン産業の再興が最終的な目標
  – 2013も第二次のEXA開発(22mil Euro), Horizon2010へつなげる
- (米国同様) システムソフトウェアや、場合によってはハードウェアも日本との共同開発の期待は高い
  – 現実的に一部の技術は米国製を使っている
  –「どこに真のコンペテンスがあり、どこは他とコラボするか」
- HORIZON2020 (2014-)では本格的な研究開発とRIの充実が見込まれる(HPC予算が年間120億ユーロへ倍増)
  – 欧州との連携のチャンス