Preparation of a Petascale Supercomputing Infrastructure for European Scientists

PRACE Status at Mid Year 2009

Dr. Jean-Philippe Nominé, CEA
Outline: PRACE Status at Mid Year 2009

- PRACE context - HPC in Europe
- PRACE objectives
- PRACE organisation
- Key achievements in 2008 and 2009
Outline

- PRACE context - HPC in Europe
- PRACE objectives
- PRACE organisation
- Key achievements in 2008 and 2009
HET: The Scientific Case

- **Weather, Climatology, Earth Science**
  - degree of warming, scenarios for our future climate.
  - understand and predict ocean properties and variations
  - weather and flood events

- **Astrophysics, Elementary particle physics, Plasma physics**
  - systems, structures which span a large range of different length and time scales
  - quantum field theories like QCD, ITER

- **Material Science, Chemistry, Nanoscience**
  - understanding complex materials, complex chemistry, nanoscience
  - the determination of electronic and transport properties

- **Life Science**
  - system biology, chromatin dynamics, large scale protein dynamics, protein association and aggregation, supramolecular systems, medicine

- **Engineering**
  - complex helicopter simulation, biomedical flows, gas turbines and internal combustion engines, forest fires, green aircraft,
  - virtual power plant
European HPC Shortcomings

• Fragmentation
  – No sustained HPC service beyond the national scale
  – No coordination of procurements

• Lack of a strong HPC industry
  – European companies mainly supply the European market (like Bull) or occupy market niches (like Numascale AS, Dolphin, ParTec)
  – Europe needs an independent access to this key technology
Outline

• PRACE context - HPC in Europe
• PRACE objectives
• PRACE organisation
• Key achievements in 2008 and 2009
Computational science infrastructure in Europe

The European Roadmap for Research Infrastructures is the first comprehensive definition at the European level.

Research Infrastructures are one of the crucial pillars of the European Research Area.

A European HPC service – impact foreseen:
- strategic competitiveness
- attractiveness for researchers
- supporting industrial development
The ESFRI Vision for a European HPC service

• European HPC-facilities at the top of an HPC provisioning pyramid
  – Tier-0: 3-5 European Centres
  – Tier-1: National Centres
  – Tier-2: Regional/University Centres

• Shape the European HPC ecosystem involving all stakeholders
  – HPC service providers on all tiers
  – Grid Infrastructures
  – Scientific and industrial user communities
  – The European HPC hardware and software industry
Expected Structuring Effects

• Renewal of Tier-0 systems every 2-3 years required
• Investments of 200 – 400 Mio € every 2-3 years will create a critical mass for strengthening the European HPC industry
  – Capability to provide independent access to this key technology
  – Access to world-class HPC systems as a competitive advantage for European science and industry

• A European HPC Service as part of the European Research Area
  – Access to world-class HPC systems for the best researchers from all European countries
First Steps and Achievements

- Production of the HPC part of the ESFRI Roadmap; Creation of a vision, involving 15 European countries
- Bringing scientists together
- Creation of the Scientific Case
- Signature of the MoU
- Submission of an FP7 project proposal
- Approval of the project
- Project start
- PRACE Initiative

Timeline:
- 2004: HPCEUR
- 2005
- 2006: HET
- 2007
- 2008: PRACE Initiative
Outline

- PRACE context - HPC in Europe
- PRACE objectives
- PRACE organisation
- Key achievements in 2008 and 2009
  - Education and training
  - Applications benchmarking and petascaling
  - Prototypes: systems for 2010
  - Prototypes: components and technologies for 2011/2012
PRACE – Project Facts

• Workplan of the PRACE Project:
  – Prepare the contracts to establish the PRACE permanent Research Infrastructure as a single Legal Entity in 2010 including governance, funding, procurement, and usage strategies.
  – Perform the technical work to prepare operation of the Tier-0 systems in 2009/2010 including deployment and benchmarking of prototypes for Petaflop/s systems and porting, optimising, peta-scaling of applications

• Project facts:
  – Partners: 16 Legal Entities from 14 countries
  – Project duration: January 2008 – December 2009
  – Project budget: 20 M €, EC funding: 10 M € - 1300 PM

PRACE is funded in part by the EC under the FP7 Capacities programme grant agreement INFSO-RI-211528
New Partners - since May 2008 - of the PRACE Initiative
Europe’s current position in HPC

91% of European HPC-power is within PRACE countries
PRACE Work Packages (classical FP7 project…)

• WP1 Management
• WP2 Organizational concept
• WP3 Dissemination, outreach and training
• WP4 Distributed computing
• WP5 Deployment of prototype systems
• WP6 Software enabling for prototype systems
• WP7 Petaflop/s systems for 2009/2010
• WP8 Future petaflop/s technologies
Outline

• PRACE context - HPC in Europe
• PRACE objectives
• PRACE organisation
• Key achievements in 2008 and 2009
  – Preparation of legal entity, governance and access
  – Education and training
  – Applications benchmarking and petascaling
  – Prototypes: systems for 2010
  – Prototypes: components and technologies for 2011/2012
Outline

• PRACE context - HPC in Europe
• PRACE objectives
• PRACE organisation
• Key achievements in 2008 and 2009
  – Preparation of legal entity, governance and access
  – Education and training
  – Applications benchmarking and petascaling
  – Prototypes: systems for 2010
  – Prototypes: components and technologies for 2011/2012
Organisational Structure

• Legal Form
  – ERI – European Research Infrastructure would have been optimal, but is not fully ready yet

• Still under discussion… as well as detailed governance model
Accessing the future PRACE RI (main user interface….)

Access Model
• Based on peer-review: “the best systems for the best science”
• Free-of-charge
• Technically: DEISA-like

Funding
• Mainly national funding through partner countries
• European contribution
• Access model has to respect national interests (ROI)
Guiding Principles & Flowchart from Initial Report on the Peer Review Process

PRINCIPLES

- Transparency
- Expert Assessment
- Confidentiality
- Managing Interests
- Right to Reply
- Ensure Fairness to the Science Proposed
- Prioritisation
- No Parallel Assessment

Flowchart:

1. Call for Proposals
2. Proposal Submission
3. Technical Assessment
   - Yes
   - Scientific Assessment
   - Applicant’s Right to Reply
   - Prioritisation by Panel
4. Reject
5. Feedback PRACE office
6. Reject
   - No
   - Funding Decision
   - Yes
   - Time Allocated

General Peer Review Management
How to get involved?
If you are a national coordinator of HPC activities and your country is not yet a member:

Join the PRACE Initiative!
(scientists: influence your national coordinators…)

Port your code to the PRACE Prototypes, prepare yourselves to petascale
– Prototypes are mainly used project-internally, but…
– … prototypes are also made available publicly for testing/porting purposes using a light-weight peer-review process
– See: http://www.prace-project.eu/prototype-access
PRACE web site

- [http://www.prace-project.eu](http://www.prace-project.eu)
- RSS feeds, news
- Various documents on line
Outline

• PRACE context - HPC in Europe
• PRACE objectives
• PRACE organisation

• Key achievements in 2008 and 2009
  – Preparation of legal entity, governance and user access
  – Education and training
  – Applications benchmarking and petascaling
  – Prototypes: systems for 2010
  – Prototypes: components and technologies for 2011/2012
Training & education

• Survey of HPC education and training needs
  – Showed a strong need for parallel programming training
  – cf PRACE web site http://www.prace-project.eu
• PRACE Petascale Summer School, August 26-29, 2008, KTH, Stockholm
  – 30+ participants
• PRACE Winter School, February 9-13, 2009, Athens, Greece
  – Almost 50 participants
• GPGPU CUDA et al. training, France, CEA+GENCI, April 2009
  – 12 participants
• Usage of PRACE prototypes for hands-on
Outline

• PRACE context - HPC in Europe
• PRACE objectives
• PRACE organisation

• Key achievements in 2008 and 2009
  – Preparation of legal entity, governance and access
  – Education and training
  – Applications benchmarking and petascaling
  – Prototypes: systems for 2010
  – Prototypes: components and technologies for 2011/2012
Representative Benchmark Suite

• Key goal of WP6 is a set of applications benchmarks
  – To be used in the procurement process for Petaflop/s systems
• Survey data helped us select highly-used applications which span the breadth of both scientific area and algorithmic ‘dwarf’
• “Living list”
  – QCD, NAMD, CPMD, Code_Saturne, GADGET, EUTERPE, NEMO, CP2K, GROMACS, NS3D, HELIUM, AVBP, TRIPOLI-4, PEPC, GPAW, ALYA, BSIT
  – WRF, Quantum_Espresso, Octopus, SPECFEM3, ELMER
• Each application will be ported to appropriate subset of prototypes
Performance Analysis Tools and Benchmarks

- **Documented:**
  - Reviews of tools
    - 9 different tool suites
  - Synthetic benchmarks
    - Computation, mixed-mode, IO, bandwidth, OS, communication
  - Application benchmarks
    - Selection and porting

- **Applications and Synthetic benchmarks integrated into JuBE**
  - Juelich Benchmark Environment
Outline

• PRACE context - HPC in Europe
• PRACE objectives
• PRACE organisation

• Key achievements in 2008 and 2009
  – Preparation of legal entity, governance and access
  – Education and training
  – Applications benchmarking and petascaling
  – Prototypes: systems for 2010
  – Prototypes: components and technologies for 2011/2012
PRACE prototypes approach

• Test & try before buying
  – Assessment of technology and architectures
• Share experience between partners
• Prepare benchmarks
• Foresee technology evolutions
• Foster collaborations between providers and users

=> A selection of systems and components, with existing, near-existing or emerging technology
PRACE prototypes TWO-FOLD approach

• (Near) existing technologies for 2009-2010 (WP7)
  – Full-featured systems, either:
    • % of large existing production systems (*scaling*)
    • or extensions of existing production systems (*techno. updates*)
    • or dedicated (smaller) systems (*new technology*)

• Emerging technologies for 2011 and beyond (WP8)
  – Mostly components/subsystems
  – As of today, strong focus on application specific, attached processors (*accelerators*)
## Selected prototypes (coordinated by WP7)

<table>
<thead>
<tr>
<th>Site</th>
<th>Architecture Vendor/Technology</th>
<th>Point of contact</th>
</tr>
</thead>
<tbody>
<tr>
<td>FZJ Germany</td>
<td><strong>MPP</strong> IBM BlueGene/P</td>
<td>Michael Stephan <a href="mailto:m.stephan@fz-juelich.de">m.stephan@fz-juelich.de</a></td>
</tr>
</tbody>
</table>
| CSC-CSCS Finland+Switzerland | **MPP** Cray XT5/XTn - AMD Opteron | Janne Ignatius [janne.ignatius@csc.fi](mailto:janne.ignatius@csc.fi)  
|                    |                                | Peter Kunszt [peter.kunszt@cscs.ch](mailto:peter.kunszt@cscs.ch) |
| CEA-FZJ France+Germany | **SMP-TN** Bull et al. Intel Xeon Nehalem | Gilles Wiber [gilles.wiber@cea.fr](mailto:gilles.wiber@cea.fr)  
|                    |                                | Norbert Eicker [n.eicker@fz-juelich.de](mailto:n.eicker@fz-juelich.de) |
| NCF/SARA Netherlands | **SMP-FN** IBM Power6          | Axel Berg [axel@sara.nl](mailto:axel@sara.nl)  
|                    |                                | Peter Michielse [michielse@nwo.nl](mailto:michielse@nwo.nl) |
| BSC Spain          | **Hybrid – fine grain** IBM Cell + Power6 | Sergi Girona [sergi.girona@bsc.es](mailto:sergi.girona@bsc.es) |
| HLRS Germany       | **Hybrid – coarse grain** NEC Vector SX/9 + x86 | Stefan Wesner [wesner@hlrs.de](mailto:wesner@hlrs.de) |
Installed prototypes

IBM BlueGene/P (FZJ) 01-2008
IBM Power6 (SARA) 07-2008
Cray XT5 (CSC) 11-2008
IBM Cell/Power (BSC) 12-2008

NEC SX9 vector part (HLRS) 02-2009
x86 part 04-2009
Bull Nehalem/INCA (CEA) 06-2009
<table>
<thead>
<tr>
<th>Feature</th>
<th>NCF</th>
<th>CSCS/CSC</th>
<th>BSC</th>
<th>FZJ</th>
<th>CEA/FZJ</th>
<th>HLRS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dedicated system - makes possible unfriendly tests</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shared large system - makes possible large runs and assessment under real production</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MPP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cluster with thin-nodes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cluster with fat nodes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Advanced (Hybrid)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specific Hardware Technologies</td>
<td>Power6</td>
<td>AMD Barcelona and Shanghai</td>
<td>IBM Cell</td>
<td>Blue Gene</td>
<td>Intel NehalemEP</td>
<td>SX9</td>
</tr>
<tr>
<td>Specific Software technologies</td>
<td>PERCS</td>
<td>MPI/OpenMP</td>
<td>CAF UPC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full featured system with storage and IO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Connected to the DEISA network</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CEA new DEISA associate partner</td>
</tr>
<tr>
<td>Collaboration with vendors</td>
<td>Software technology (IBM)</td>
<td>system reliability, performance, functionality (CRAY)</td>
<td>System design (IBM)</td>
<td>System design (BULL)</td>
<td>System design (NEC)</td>
<td></td>
</tr>
</tbody>
</table>

Note: The table above lists various features and technologies associated with different computing systems. Each cell indicates the availability or presence of a particular technology or feature in the corresponding system. The colors denote the presence, with green indicating availability and white indicating absence. The systems mentioned include NCF, CSCS/CSC, BSC, FZJ, CEA/FZJ, and HLRS.
Outline

• PRACE context - HPC in Europe
• PRACE objectives
• PRACE organisation
• Key achievements in 2008 and 2009
  – Preparation of legal entity, governance and access
  – Education and training
  – Applications benchmarking and petascaling
  – Prototypes: systems for 2010
  – Prototypes: components and technologies for 2011/2012
## Prototypes selected by WP8

<table>
<thead>
<tr>
<th>Sites</th>
<th>Hardware/Software</th>
<th>Porting effort</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEA</td>
<td>1U Tesla Server  T1070 (CUDA, CAPS, DDT) Intel Harpertown nodes</td>
<td>“Evaluate GPU accelerators and GPGPU programming models and middleware.” (e.g., <em>pollutant migration code</em> (ray tracing algorithm) to CUDA and HMPP)</td>
</tr>
<tr>
<td>CINES-LRZ</td>
<td>Hybrid SGI ICE2/UV/Nehalem-EP &amp; Nehalem-EX/ClearSpeed/ Larrabee</td>
<td>Gadget, SPECFEM3D_GLOBE, RaXml, Rinf, RandomAccess, ApexMap, Intel MPI BM</td>
</tr>
<tr>
<td>CSCS</td>
<td>Prototype PGAS language compilers (CAF + UPC for Cray XT systems)</td>
<td>“The applications chosen for this analysis will include some of those already selected as benchmark codes in WP6.”</td>
</tr>
<tr>
<td>EPCC</td>
<td>Maxwell – FPGA prototype (VHDL support &amp; consultancy + software licenses (e.g., Mitrion-C))</td>
<td>“We wish to port several of the PRACE benchmark codes to the system. The codes will be chosen based on their suitability for execution on such a system.”</td>
</tr>
</tbody>
</table>
### Prototypes selected by WP8 (cont’d)

<table>
<thead>
<tr>
<th>Sites</th>
<th>Hardware/Software</th>
<th>Porting effort</th>
</tr>
</thead>
<tbody>
<tr>
<td>FZJ (BSC)</td>
<td>eQPACE (PowerXCell cluster with special network processor)</td>
<td>Extend FPGA-based interconnect beyond QCD applications.</td>
</tr>
<tr>
<td>“Cell &amp; FPGA interconnect”</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LRZ</td>
<td>RapidMind (Streaming Processing Programming Paradym) X86, GPGPU, Cell</td>
<td>ApexMap, Multigrid, FZJ (QCD), CINECA (linear algebra kernels involved in solvers for ordinary differential equations), SNIC</td>
</tr>
<tr>
<td>“RapidMind”</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NCF</td>
<td>ClearSpeed CATS 700 units</td>
<td>Astronomical many-body simulation, Iterative sparse solvers with preconditioning, finite element code, cryomicrotome image analysis</td>
</tr>
<tr>
<td>“ClearSpeed”</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CINECA</td>
<td>I/O Subsystem (SSD, Lustre, pNFS)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>+ 1 additional prototype targeting energy efficiency</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
As a temporary conclusion…

- You do have access to PRACE prototypes
- We want to foster long-term relationships with scientific communities
  - PRACE governance should include a Users’ Forum and a Scientific Steering Committee
- Training and education are a main concern
- We will go on watching and assessing technology and architectures, then take part in R&D and influencing vendors’ roadmaps
- More to come about 2010 RI organisation
  - Keep in touch, ask us…

jean-philippe.nomine@cea.fr
Europe has already gone Petascale…

• Jugene 1 PFlop/s BlueGene at FZJ (TOP#3, June 2009)

• CEA will have 1 PFlop/s Bull TERA 100 Cluster mid 2010

• CEA and FZJ collaborate…
PRACE and CEA

- GENCI coordinates the French PRACE effort
  - Half of French manpower provided by CEA
  - Strong involvement in prototypes
  - Preparing TGCC new facility to host one of the first PRACE machines at CEA/DIF, Bruyères – near Paris

- CPU Nehalem/Bull INCA drawers
- 1024 cores, ~10 TF

- GPU Tesla/NVIDIA
- 2 S1070 servers
- Programmers models and tools
  - CUDA, HMPP

TGCC as of June 2009