HPC applications

✓ Classes of HPC applications
  ▪ The 7/16 dwarfs
✓ HPC in Europe – PRACE, DEISA, GÉANT2
  ▪ PRACE survey
  ▪ PRACE prototypes
  ▪ PRACE training
✓ GÉANT3 and FEDERICA?

Major HPC Application areas

- Astronomy and cosmology
- Computational chemistry
- Computational engineering
- Computational fluid dynamics
- Condensed matter physics
- Earth and climate science
- Life science
- Particle physics
- Plasma physics
- Other
The “Dwarfs” classes
(the Berkeley scheme)

The dwarfs are those algorithm types which constitute classes where:

- Membership in a class is defined by similarity in computation and data movement
- The dwarfs are specified at a high level of abstraction to allow reasoning about their behaviour across a broad range of applications.
- Programs that are members of a particular class can be implemented differently and the underlying numerical methods may change over time, but the claim is that the underlying patterns have persisted through generations of changes and will remain important into the future.

A dwarf is therefore a grouping of kernels that share both computational and data structure (and they use similar numerical libraries)

http://www.eecs.berkeley.edu/Pubs/TechRpts/2006/EECS-2006-183.html

The initial “seven Dwarfs” classes
(the Berkeley scheme)

1. Dense linear algebra (e.g., BLAS) – data is stored in dense matrices or vectors and access is often via unit-level strides. Typical algorithm would be Cholesky decomposition for symmetric systems or Gaussian elimination for non-symmetric systems.
2. Sparse linear algebra (e.g., SpMV) – data is stored in compressed format as it largely consists of zeros and is therefore accessed via an index-based load. Typical algorithm would be Conjugate Gradient or any of the Krylov methods.
3. Spectral methods (e.g., FFT) – data is in frequency domain and requires a transform to convert to spatial/temporal domain. They are typified by, but not restricted to, FFT.
4. N-body methods (e.g., Barnes-Hut) – data consists of discrete particle bodies that interact with each other and/or the “environment”.
5. Structured grids (Cactus) – Represented by a regular grid. Points on grid are conceptually updated together via equations linking them to other grids. There is high spatial locality. Updates may be in place or between 2 versions of the grid.
6. Unstructured grid (e.g., ABAQUS) – data is stored in terms of the locality and connectivity to other data. Points on grid are conceptually updated together, but updates require multiple levels of redirection.
7. Map reduce methods (e.g., Monte Carlo) – embarrassingly parallel problems, such as Monte Carlo methods, where calculations are independent of each other.
Extensions to the original Seven Dwarfs (the Berkeley scheme)

8. **Combinational Logic** (e.g., encryption) - Functions that are implemented with logical functions and stored state.

9. **Graph traversal** (e.g., Quicksort) - Visits many nodes in a graph by following successive edges. These applications typically involve many levels of indirection, and a relatively small amount of computation.


11. **Backtrack and Branch+Bound** - Finds an optimal solution by recursively dividing the feasible region into subdomains, and then pruning subproblems that are suboptimal.

12. **Construct Graphical Models** - Constructs graphs that represent random variables as nodes and conditional dependencies as edges. Examples include Bayesian networks and Hidden Markov Models.

13. **Finite State Machine** - A system whose behavior is defined by states, transitions defined by inputs and the current state, and events associated with transitions or states.
HPC in Europe

PRACE, DEISA, GEANT2

- HPC systems, applications, prototypes, training
- PRACE objective: Petaflop supercomputers  
  http://www.prace-project.eu
- DEISA objective: European Virtual Supercomputer  
  http://www.deisa.eu
- GEANT2 objective: high-bandwidth DEISA interconnect  
  http://www.geant2.net/

PRACE partners

- Principal Partners
  - France: GENCI – Grand Equipement national pour le Calcul Intensif
  - Germany: GCS – GAUSS Centre for Supercomputing
  - The Netherlands: NCF – Netherlands Computing Facilities Foundation
  - Spain: BSC – Barcelona Supercomputing Center - Centro Nacional de Supercomputación
  - UK: EPSRC – Engineering and Physical Sciences Research Council

- General Partners
  - Austria: GUP – Institut für Graphische und Parallele Datenverarbeitung der Johannes Kepler Universität
  - Finland: CSC – The Finnish IT Center for Science
  - Greece: GREEK – Greek Research and Technology Network
  - Italy: CINECA – Consorzio Interuniversitario
  - Norway: UNINETT Sigma AS
  - Poland: PSNC – Poznan Supercomputing and Networking Center
  - Portugal: UT-LCA – Universidade de Coimbra – Laboratório de Computação Avançada
  - Sweden: SNIC – Swedish National Infrastructure for Computing
  - Switzerland: ETH – Swiss Federal Institute of Technology Zurich, CSCS – Swiss National Supercomputing Centre

- Additional General Partners of the PRACE Initiative
  - Countries that have signed the PRACE Memorandum of Understanding:
    - Cyprus: The Computation-based Science and Research Center (CSTRC)
    - Ireland: Irish Centre for High-End Computing
    - Serbia: The Institute of Physics, Belgrade
    - Turkey: National Center for High Performance Computing
DEISA partners

GÉANT2

A few Facts & Figures...

- 29 POPs
- serving 30 NRENs
- ~12000 km of fibre
- >120 (own) 10G lambdas
- 22 (leased) 10G lambdas
- + some lower speed links
- NREN accesses at up to 10Gbps (+ backup) + P2P
- 4 x 10G to North America
- POP in NY
- connections to other R&E networks...
- Abilene (Internet2), ESnet, CA*net4, SINET, TENET, EUMEDCONNECT, RedCLARA, TEIN2, India

- GEANT3 Next
3-Tier architecture

Steps Towards an Integrated European HPC Infrastructure

- PRACE: Petaflop supercomputer
- DEISA: Virtual Supercomputer

Tier0 / Tier1 Top Layer of the HPC Ecosystem

- PRACE: Designing an infrastructure that will enable the operation of shared petascale European systems
- Enhancing performance in selected sites and providing wide access to shared systems

In the DEISA-2 environment, scientists will naturally have access to several distributed platforms, and shared systems will have to access remote data repositories. We will be in a situation similar to TeraGrid.

DEISA-2: strong integration of Tier0 and Tier1 systems (automatically provides wide, seamless and efficient access to shared systems and data repositories)

The DEISA-1 services have been tailored for this mode of operation. There is a positive feedback between the two orthogonal lines of action:

- DEISA is paving the way to the efficient operation of Tier systems.
- Tier systems will drive the massive adoption of the DEISA services.
Europe’s current position in HPC

Aggregated LINPACK Performance in PetaFlops in November Top 500 Lists

75% of European HPC-power is within PRACE Principal Partner countries

The PRACE survey utilisation matrix
(distribution of applications into the 7 dwarfs)

The utilisation matrix is made up of 70 categories: 10 scientific areas and 7 algorithmic “dwarfs”. The figure in each cell is an estimate of the number of Tflop/s burned in each category. White boxes are those with no usage. Orange boxes are those with usage greater than zero, but less than 5 Tflop/s usage. Red boxes signify usage greater than 5 Tflop/s.

A dwarf is a grouping of kernels that share both computational and data structure.
PRACE HPC systems

<table>
<thead>
<tr>
<th>Name</th>
<th>centre</th>
<th>manufacturer</th>
<th>model</th>
<th>factory</th>
<th>architecture</th>
<th>cores</th>
<th>Rpeak</th>
<th>Rmax</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jaguar</td>
<td>FZJ</td>
<td>IBM</td>
<td>Blue Gene/P</td>
<td>IBM</td>
<td>MPP</td>
<td>65536</td>
<td>180</td>
<td>222.82</td>
</tr>
<tr>
<td>MareNostrum</td>
<td>HSC</td>
<td>IBM</td>
<td>JS21 cluster</td>
<td>TNC</td>
<td>94208</td>
<td>63800</td>
<td>1024</td>
<td></td>
</tr>
<tr>
<td>Helios AbbVie</td>
<td>DG</td>
<td>IBM</td>
<td>JS21 cluster</td>
<td>TNC</td>
<td>62259</td>
<td>56520</td>
<td>9728</td>
<td></td>
</tr>
<tr>
<td>HECToR</td>
<td>EPSCC</td>
<td>Cray</td>
<td>XT4</td>
<td>MPP</td>
<td>63431</td>
<td>60480</td>
<td>11123</td>
<td></td>
</tr>
<tr>
<td>Noisco</td>
<td>SNIC</td>
<td>HP</td>
<td>Cray</td>
<td>MPP</td>
<td>48066</td>
<td>44408</td>
<td>6440</td>
<td></td>
</tr>
<tr>
<td>Pallas</td>
<td>GEMINI</td>
<td>IBM</td>
<td>Blue Gene/P</td>
<td>MPP</td>
<td>49152</td>
<td>42130</td>
<td>7680</td>
<td></td>
</tr>
<tr>
<td>Hexagon</td>
<td>SIMA</td>
<td>Cray</td>
<td>XT4</td>
<td>MPP</td>
<td>51700</td>
<td>42000</td>
<td>5552</td>
<td></td>
</tr>
<tr>
<td>Galera</td>
<td>MDRS</td>
<td>IBM</td>
<td>Blue Gene/P</td>
<td>MPP</td>
<td>49104</td>
<td>48190</td>
<td>6176</td>
<td></td>
</tr>
<tr>
<td>Jubi</td>
<td>FZJ</td>
<td>IBM</td>
<td>Blue Gene/P</td>
<td>TNC</td>
<td>45875</td>
<td>37330</td>
<td>16384</td>
<td></td>
</tr>
<tr>
<td>BCX</td>
<td>CINECA</td>
<td>IBM</td>
<td>LS21</td>
<td>TNC</td>
<td>53248</td>
<td>19910</td>
<td>5120</td>
<td></td>
</tr>
<tr>
<td>Mallo</td>
<td>SIMA</td>
<td>HP</td>
<td>EL640</td>
<td>MPP</td>
<td>59090</td>
<td>15090</td>
<td>5532</td>
<td></td>
</tr>
<tr>
<td>Peters</td>
<td>OTTER</td>
<td>IBM</td>
<td>XT4</td>
<td>MPP</td>
<td>15700</td>
<td>14220</td>
<td>3528</td>
<td></td>
</tr>
<tr>
<td>HPC’s</td>
<td>EPSRC</td>
<td>IBM</td>
<td>F575 cluster</td>
<td>MPP</td>
<td>15560</td>
<td>12940</td>
<td>2560</td>
<td></td>
</tr>
<tr>
<td>Hyippys</td>
<td>NEC</td>
<td>IBM</td>
<td>p575 cluster</td>
<td>MPP</td>
<td>14502</td>
<td>11400</td>
<td>1920</td>
<td></td>
</tr>
<tr>
<td>Legion</td>
<td>EPSRC</td>
<td>IBM</td>
<td>Blue Gene/P</td>
<td>MPP</td>
<td>15290</td>
<td>11110</td>
<td>4996</td>
<td></td>
</tr>
<tr>
<td>Texta-UX</td>
<td>HLRG</td>
<td>NEC</td>
<td>SX8</td>
<td>MPP</td>
<td>9216</td>
<td>8923</td>
<td>576</td>
<td></td>
</tr>
<tr>
<td>Lods</td>
<td>CSC</td>
<td>IBM</td>
<td>CP400 EL ProLiant</td>
<td>TNC</td>
<td>10835</td>
<td>9843</td>
<td>5652</td>
<td></td>
</tr>
<tr>
<td>kriselkico</td>
<td>CIC</td>
<td>HP</td>
<td>Supercomputer Centre</td>
<td>MPP</td>
<td>10094</td>
<td>9020</td>
<td>2176</td>
<td></td>
</tr>
<tr>
<td>Jumpp</td>
<td>FZJ</td>
<td>IBM</td>
<td>Blue Gene/P</td>
<td>TNC</td>
<td>8921</td>
<td>5558</td>
<td>1312</td>
<td></td>
</tr>
<tr>
<td>ZAHIR</td>
<td>GEMINI</td>
<td>IBM</td>
<td>cluster</td>
<td>MPP</td>
<td>6350</td>
<td>3900</td>
<td>1024</td>
<td></td>
</tr>
<tr>
<td>HERA</td>
<td>GEMINI</td>
<td>IBM</td>
<td>IBM</td>
<td>MPP</td>
<td>3000</td>
<td>3700</td>
<td>384</td>
<td></td>
</tr>
<tr>
<td>XCS</td>
<td>SIMA</td>
<td>IBM</td>
<td>IBM</td>
<td>MPP</td>
<td>2200</td>
<td>1600</td>
<td>520</td>
<td></td>
</tr>
<tr>
<td>Mriista-UC-DEN</td>
<td>MDRS</td>
<td>IBM</td>
<td>e325/san</td>
<td>MPP</td>
<td>1277</td>
<td>1182</td>
<td>336</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: PRACE partner systems included in survey. MPP – Massively Parallel Processing, TNC – Thin Node Cluster, PRC – Fat Node Cluster, VEC – Vector.
Performance per system (in Tflops)

Total Compute power by architecture type

Fat-node Cluster 14%
Vector 1%
MPP 50%
Thin-node Cluster 35%
System availability and utilisation

Job size distribution by system

Job size in cores
Mean job size as a % of system size

Aggregate distribution of LEFs by job size

(LEF = LINPACK Equivalent Flop/s)
Scientific area distribution by system

Aggregate distribution of LEFs by scientific area
Number of users and $R_{\text{max}}$ per user

On average 113Gflops/user = 20-25 cores

Application name

---

<table>
<thead>
<tr>
<th>Application Name</th>
<th>LID: User Setup</th>
<th>Number of cores using LID: core</th>
</tr>
</thead>
<tbody>
<tr>
<td>fft</td>
<td>254</td>
<td>1</td>
</tr>
<tr>
<td>Programs</td>
<td>234</td>
<td>1</td>
</tr>
<tr>
<td>AGR</td>
<td>228</td>
<td>1</td>
</tr>
<tr>
<td>MSES</td>
<td>220</td>
<td>1</td>
</tr>
<tr>
<td>ECO</td>
<td>210</td>
<td>1</td>
</tr>
<tr>
<td>CAF</td>
<td>200</td>
<td>1</td>
</tr>
<tr>
<td>CHA</td>
<td>190</td>
<td>1</td>
</tr>
<tr>
<td>OCT</td>
<td>180</td>
<td>1</td>
</tr>
<tr>
<td>OCT</td>
<td>170</td>
<td>1</td>
</tr>
<tr>
<td>OCT</td>
<td>160</td>
<td>1</td>
</tr>
<tr>
<td>OCT</td>
<td>150</td>
<td>1</td>
</tr>
<tr>
<td>OCT</td>
<td>140</td>
<td>1</td>
</tr>
<tr>
<td>OCT</td>
<td>130</td>
<td>1</td>
</tr>
<tr>
<td>OCT</td>
<td>120</td>
<td>1</td>
</tr>
<tr>
<td>OCT</td>
<td>110</td>
<td>1</td>
</tr>
<tr>
<td>OCT</td>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>OCT</td>
<td>90</td>
<td>1</td>
</tr>
<tr>
<td>OCT</td>
<td>80</td>
<td>1</td>
</tr>
<tr>
<td>OCT</td>
<td>70</td>
<td>1</td>
</tr>
<tr>
<td>OCT</td>
<td>60</td>
<td>1</td>
</tr>
<tr>
<td>OCT</td>
<td>50</td>
<td>1</td>
</tr>
<tr>
<td>OCT</td>
<td>40</td>
<td>1</td>
</tr>
<tr>
<td>OCT</td>
<td>30</td>
<td>1</td>
</tr>
<tr>
<td>OCT</td>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td>OCT</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>OCT</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Constantin Halatsis
Base language usage by applications

<table>
<thead>
<tr>
<th>Language</th>
<th>No. of applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fortran90</td>
<td>50</td>
</tr>
<tr>
<td>C90</td>
<td>22</td>
</tr>
<tr>
<td>Fortran77</td>
<td>15</td>
</tr>
<tr>
<td>C++</td>
<td>10</td>
</tr>
<tr>
<td>C99</td>
<td>7</td>
</tr>
<tr>
<td>Python</td>
<td>3</td>
</tr>
<tr>
<td>Perl</td>
<td>2</td>
</tr>
<tr>
<td>Mathematica</td>
<td>1</td>
</tr>
</tbody>
</table>

Parallelisation

- Of the 69 applications, all but two use MPI for parallelisation.
- The exceptions are Gaussian (OpenMP) and BLAST (sequential).
- Of the 67 MPI applications, six also have standalone OpenMP versions and three have stand alone SHMEM versions.
- Ten applications have hybrid MPI/OpenMP implementations, two have hybrid MPI/SHMEM versions and one has a hybrid MPI/Posix threads version.
- Only one application was reported as using MPI2 single sided communication.
Distribution of applications usage by scientific area

- Particle Physics: 41.4%
- Computational Chemistry: 28.1%
- Condensed Matter Physics: 12.4%
- CFD: 4.3%
- Astronomy & Cosmology: 6.5%
- Earth & Climate: 3.2%
- Life Sciences: 4.1%
- Plasma Physics: 1.9%
- Computational Engineering: 1.7%
- Other: 0.0%

---

Distribution of applications usage by job size

- > 2048: 24.7%
- 33 - 128: 24.6%
- 512 - 2048: 18.8%
- 129 - 512: 16.5%
- < 32: 15.2%
Distribution of applications usage by algorithmic dwarves

Usage matrix from the surveys

<table>
<thead>
<tr>
<th>Area/Dwarf</th>
<th>Dense linear algebra</th>
<th>Spectral methods</th>
<th>Structured grids</th>
<th>Sparse linear algebra</th>
<th>Particle methods</th>
<th>Unstructured grids</th>
<th>Map reduce methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Astronomy and Cosmology</td>
<td>0.00</td>
<td>0.62</td>
<td>4.15</td>
<td>2.26</td>
<td>0.43</td>
<td>2.99</td>
<td>0.08</td>
</tr>
<tr>
<td>Computational Chemistry</td>
<td>18.05</td>
<td>14.69</td>
<td>1.14</td>
<td>2.79</td>
<td>7.40</td>
<td>0.99</td>
<td>13.99</td>
</tr>
<tr>
<td>Computational Engineering</td>
<td>0.08</td>
<td>0.94</td>
<td>0.93</td>
<td>0.58</td>
<td>0.96</td>
<td>0.82</td>
<td>1.99</td>
</tr>
<tr>
<td>Computational Fluid Dynamics</td>
<td>0.00</td>
<td>1.70</td>
<td>4.98</td>
<td>0.06</td>
<td>0.00</td>
<td>0.84</td>
<td>0.08</td>
</tr>
<tr>
<td>Condensed Matter Physics</td>
<td>0.01</td>
<td>14.33</td>
<td>0.96</td>
<td>0.06</td>
<td>1.76</td>
<td>0.78</td>
<td>2.74</td>
</tr>
<tr>
<td>Earth and Climate Science</td>
<td>0.00</td>
<td>0.70</td>
<td>3.31</td>
<td>0.00</td>
<td>0.00</td>
<td>0.22</td>
<td>0.00</td>
</tr>
<tr>
<td>Life Science</td>
<td>0.00</td>
<td>4.22</td>
<td>0.92</td>
<td>0.00</td>
<td>0.04</td>
<td>0.78</td>
<td>0.40</td>
</tr>
<tr>
<td>Particle Physics</td>
<td>15.54</td>
<td>0.00</td>
<td>4.32</td>
<td>0.83</td>
<td>0.10</td>
<td>0.00</td>
<td>33.22</td>
</tr>
<tr>
<td>Plasma Physics</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.42</td>
<td>0.61</td>
</tr>
<tr>
<td>Other</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

The figures in boxes is the total number of LEFs (Tflop/s) used in that particular scientific area and dwarf. White boxes are those with no usage. Orange boxes are those with usage greater than zero, but less than 5 Tflop/s usage. Red boxes signify usage greater than 5 Tflop/s. Bold figures denote those boxes that have an application representing it in the proposed list (see next section). Italicised figures are those that have an application representing that box in the extended list.
### List of applications in each of the scientific area/dwarf category

<table>
<thead>
<tr>
<th>Area/Dwarf</th>
<th>Home Base</th>
<th>Specialized</th>
<th>Unspecialized</th>
<th>Scientific Area</th>
<th>DOE</th>
<th>Other</th>
<th>Unpublished</th>
<th>Map reduce</th>
</tr>
</thead>
<tbody>
<tr>
<td>Astronomy and Cosmology</td>
<td>NO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Computational Condensed Matter</td>
<td>YES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Computational Fluid Dynamics</td>
<td>YES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Condensed Matter Physics</td>
<td>YES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Earth and Climate Science</td>
<td>YES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Life Sciences</td>
<td>YES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nuclear Physics</td>
<td>YES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>YES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Core list of applications to be included in the PRACE benchmarks for the 0-Tier prototypes

<table>
<thead>
<tr>
<th>Application</th>
<th>Scientific Area/Description/URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>QCD benchmark</td>
<td>Particle physics This is a synthetic benchmark application designed to include all key QCD algorithms.</td>
</tr>
</tbody>
</table>

http://www.mpi.org.de/fileadmin/00的朋友好友喜欢.html
It seems that a future 0 Tier Petaflop/s system might run a large number of map-reduce type jobs, particularly in the area of particle physics.
PRACE prototypes (0 Tier)

<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>MPP IBM BlueGene/P</td>
<td>Michael Stephan <a href="mailto:m.stephan@fzjulich.de">m.stephan@fzjulich.de</a></td>
</tr>
<tr>
<td>CSC-CSCS Finland+Switzerland</td>
<td>MPP Cray XT5/XTn - AMD Opteron</td>
<td>Janne Ignatius <a href="mailto:janne.ignatius@csc.fi">janne.ignatius@csc.fi</a>, Peter Kursat <a href="mailto:peter.kursat@csc.de">peter.kursat@csc.de</a></td>
</tr>
<tr>
<td>CEA-FJZ France+Germany</td>
<td>SMP-TN Bull et al. Intel Xeon Nehalem</td>
<td>Gilles Wibar <a href="mailto:gilles.wibar@cea.fr">gilles.wibar@cea.fr</a>, Horbert Eicher eicher@<a href="mailto:fj-fJP@julich.de">fj-fJP@julich.de</a></td>
</tr>
<tr>
<td>NCF Netherlands</td>
<td>SMP-FN IBM Power 6</td>
<td>Axel Berg <a href="mailto:ace@bnn.nl">ace@bnn.nl</a>, Peter Michalese <a href="mailto:michalse@bwo.nl">michalse@bwo.nl</a></td>
</tr>
<tr>
<td>BSC Spain</td>
<td>Hybrid - fine grain, IBM Cell + Power6</td>
<td>Sergi Girona <a href="mailto:girona@besc.es">girona@besc.es</a></td>
</tr>
<tr>
<td>HLRS Germany</td>
<td>Hybrid - coarse grain, NEC Vector SX/9 + x86</td>
<td>Stefan Weener <a href="mailto:weener@hlrs.de">weener@hlrs.de</a></td>
</tr>
</tbody>
</table>

Call for access to prototypes

http://www.prace-project.eu/prototype-access

The general conditions for prototype access are:

- Prototypes are intended for testing and not for standard production work.
- As a general rule, application support (e.g., code development, optimisation, parallelisation, etc.) during prototype testing is limited.
- Prototype access is granted after technical review by the PRACE centre hosting the prototype.
- The maximum project duration is 3 months.
- Applicants must agree that applicants' names and affiliations, as well as a summary of the project purpose and the results achieved during prototype testing, may be made publicly available in the PRACE website and may be used in PRACE deliverables and other PRACE documents.
- Applicants must agree to summarise the results of the project in an end report to be sent to the PRACE centre hosting the prototype not later than 1 month after project conclusion.
- The project leader is responsible for signing and fulfilling the usage agreements issued by the PRACE centre hosting the prototype.
- Proposals are accessed by the PRACE Prototype Access Committee with representatives from the prototype host centre and representatives from PRACE WP5 and WP6.
- Applicants have the right to reply to the granting decisions. Application replies will be handled by the PRACE Technical Board.
FZJ prototype (Germany)

FZJ

- Specific features:
  - Access to a large existing MPP system, already 1/4 PF with an architecture expandable to 1 PF
  - Contribution to the PRACE project:
    - Application scaling, optimization and benchmarking including:
      • Communications
      • I/O
      - Large scale operations on selected applications
      - Assessment of electrical power usage
  - Availability July 2008

CSC/CSCS prototype (Finland, Switzerland)

CSC / CSCS

- Specific features:
  - Prototype installed in CSC, joint effort with CSCS
  - Funding goes fully for the dedicated system
  - As additional in-kind contribution access to a larger existing system – similar architecture
  - Contribution to the PRACE project:
    - Access to a prototype with MPP architecture and fast processors:
      • AMD Oiteon, SeaStar2+ 36-tonus network
      - Early access to AMD new generation of processors: all processors (Barcelona) replaced by Shanghai (XTn)
      - Additional focus on hybrid MPI/OpenMP parallel programming by CSCS
      - Capability testing on the CSC XT system
  - Availability December 2008
CEA / FZJ

- Specific features:
  - Combination of dedicated test system at CEA and a large production system of same architecture at FZJ
  - Contribution to the PRACE project
    - Early access to a prototype of a new product designed by BULL (pre GA machine)
      - High density blade based system with new Intel Nehalem processors
      - Optimized for HPC
      - Scalable to Petatop/s
      - Water cooling
    - Scalability testing on the FZJ part of the prototype
- Availability: March 2009

NCF prototype (Netherlands)

- Specific features:
  - Large shared memory (4-8 GB/core) and fast I/O configuration
  - Contribution to the PRACE project:
    - Access to the new IBM Power6 processors and IBM Power Cluster fat node architecture
    - Focus on HPC software from US DARPA/PERCS research addressing specific petascale issues—early access
    - Specific test nodes for aggressive experimentations
    - Capability testing and assessments on large production system
    - Very high density (>50 kW/rack), water cooled nodes
- Availability: October 2008
Barcelona SC prototype (Spain)

BSC

- Specific features:
  - Dedicated ‘fine grain’ hybrid system

- Contribution to the PRACE project:
  - New Power 6 + Cell processor integration
    - Comparable to US PF RoadRunner but different CPUs
  - Programming techniques and tools for
    - CPU+accelerators
  - Operation of an hybrid system: queuing system, file
    - system, accounting, system administration, ...
  - Assessment of electrical power usage

- Availability December 2008

HLRS prototype (Germany)

HLRS

- Specific features
  - Unique ‘System of Systems’ concept
    - Multi-physics / multi-scale apps on optimized hardware
    - Hybrid configuration Vector (SX8) + Scalar (Nehalem)
    - Highly innovative configuration
    - Scalable (e.g. with Cell, GPU, FPGA, ...)
    - Shared file system and heterogeneous network
  - Concept enables industry-related applications

- Contribution to the PRACE project:
  - New Programming models and methods
  - Close collaboration with vendor (joint Linux OS porting)
  - Necessary intermediate step towards new
    - hybrid systems (more tightly coupled)
  - Specific I/O and network challenges can be investigated

- Availability March 2009
Second set of PRACE prototypes - 1

- CINECA (HPC consortium of 36 universities, Italy) will address metadata performance of I/O subsystem solutions for petascale machines on an HP XC4 prototype and study NFS and pNFS over RDMA. The usage of SSD based technology to accelerate metadata performance will be tested with Lustre, NFS and pNFS and compared with traditional disks technology based solutions.

- EPSRC-EPCC (Edinburgh Parallel Computing Centre, UK) within the FPGA High Performance Computing Alliance (FHPCA) will evaluate porting efforts and the ratio of performance versus power consumption of PRACE benchmarks on their “Maxwell” FPGA prototype supercomputer.

Second set of PRACE prototypes - 2

- ETHZ-CSCS (Swiss National Supercomputing Centre, Switzerland) will study new parallel programming paradigms like the Partitioned Global Address Space (PGAS) programming approaches (Co-Array Fortran, UPC) and the upcoming DARPA High Productivity Computer System (HPCS) language (like Cray’s Chapel) on a 3328 cores Cray XT3 system.

- FZJ (Forschungszentrum Jülich, Germany) will provide a power efficient special-purpose architecture called eQPACE for lattice Quantum Chromodynamics (QCD). This 25.6 TFlop/s peak performance prototype is based on IBM PowerXCell 8i processors and a custom 3d-torus interconnect implemented within FPGAs supporting presently only nearest-neighbor communication. One of the main goals will be to extend the concept to general all-to-all communication.
Second set of PRACE prototypes - 3

- **GENCI-CEA** (Commissariat à l’Energie Atomique, France) offers a hybrid system composed by *nVIDIA Tesla S1070* coupled with BULL Novascale R425 systems. The purpose of this prototype will be to evaluate different programming environments like CUDA, HMPP (from CAPS Entreprise), OpenCL and the GPU aware version of Allinea’s DDT debugger.

- **BAdW-LRZ** (Leibniz Supercomputing Centre, Germany) will assess the productivity of the new *data stream parallel programming language RapidMind* on x86 multicore systems and multiple accelerators (*nVIDIA and AMD/ATI GPUs, IBM Cell and Intel Larrabee*).

Second set of PRACE prototypes - 4

- **BAdW-LRZ** and **GENCI-CINES** (Centre Informatique National de l’Enseignement Supérieur, France) will install a joint prototype in Garching and Montpellier based on SGI thin nodes (ICE system) and fat nodes (UltraViolet) coupled with *Clearspeed/PetaPath (e710 boards)* and *Intel Larrabee GPUs*. The two partners are planning to evaluate novel hardware (Intel Nehalem-EP/EX processors, NUMAlinkS and 4X QDR Infiniband networks, Clearspeed/PetaPath accelerators and Intel Larrabee manycore GPUs) as well as software components (*Lustre filesystem*) on synthetic benchmarks and real applications.

- **NCF-SARA** (Stichting Academisch Rekencentrum Amsterdam, Netherlands) will install a compact hybrid system composed by 12 standard Intel compute nodes coupled with 12 *ClearSpeed/PetaPath CATS 700* units primarily dedicated to the evaluation of large-scale applications (astrophysics, iterative solvers, geophysics and image analysis). Comparisons with GPU based versions of some applications are planned as well as collaboration with the joint BAdW-LRZ / GENCI-CINES prototype.
PRACE training
immediate requirements

• Training in mixed-mode (hybrid) programming
  – Two-thirds of respondents indicated that they had no competency in hybrid programming techniques
• Increased awareness of Partitioned Global Address Space (PGAS) languages
  – 90% of respondents are unfamiliar with this approach to parallel programming
• Information on latest developments in HPC and parallel programming
  – 80% of respondents consider themselves inadequately informed
• High quality training material on visualization
  – The vast majority of respondents rated the training they had received in visualization tools and techniques as ‘poor’
• Promotion of the use of numerical libraries
  – Proficiency and awareness of numerical libraries as rated as ‘low’ by most users

PRACE training
short-term requirements-1

• Formal training courses in modern Fortran programming
  – Fortran was considered the most important traditional language
  – Courses should cover the more sophisticated constructs and promote modern software engineering principles
• Training material on code optimization, debugging and code testing
• Training material on parallel I/O
  – Two-thirds of respondents are using parallel I/O but the vast majority have little or no knowledge in this field
PRACE training
short-term requirements-2

• Training in multi-core programming
  – Just one-fifth of respondents had received training in programming for multi-core architectures
  – The vast majority of respondents consider they have no proficiency in topics such as multi-core cache optimization, and multi-core memory and bandwidth management

• Training in standard Unix/Linux and HPC system skills
  – 70% of respondents would benefit from training material in areas such as Unix environments, batch systems, and development tools including version control systems

PRACE training
long-term requirements

• Increased awareness of developments in next generation programming languages
  – 90% of respondents are unfamiliar with next-generation HPC programming languages such as Chapel, X10 and Fortress

• Training in scripting languages
  – There is an increased uptake of scripting languages (particularly for providing the glue within workflows)
  – 60% of respondents considered the need for training in scripting languages as either somewhat or very important

• Training in distributed computing and Grid tools
  – The emergence of Grid technology as a fundamental component of a distributed computing infrastructure requires that users be familiar with Grid tools to enable exploitation of the PRACE HPC Research Infrastructure
PRACE training
genral requirements

- **Face-to-face training was considered the most useful channel for delivering training**
  - Training must be given by experts in the HPC field, a factor that would influence users’ willingness to attend

- **Existing training material must be improved**
  - 90% of respondents consider that training material is inadequate
  - No training course was designated as “excellent” by any user
  - 50% of users aren’t adequately served by their local HPC centre
  - PRACE should work with local centres to ensure training needs and expectations are met for all users
  - 95% of respondents consider that they would benefit from a pan-European centralized training repository
  - Material must be world-class quality and regularly updated