CINECA and the European HPC Ecosystem

Giovanni Erbacci
Supercomputing, Applications and Innovation Department, CINECA

g.erbacci@cineca.it

Enabling Applications on Intel MIC based Parallel Architectures
PRACE Summer School, Bologna 8-11 July, 2013
CINECA non profit Consortium, made up of 57 Italian universities*, The National Institute of Oceanography and Experimental Geophysics - OGS, the CNR (National Research Council), and the Ministry of Education, University and Research (MIUR).

CINECA is the national Supercomputing Centre in Italy
CINECA is one of the 4 Hosting Members in PRACE
CINECA: the HPC Story

1969: CDC 6600  
1st system for scientific computing

1975: CDC 7600  
1st supercomputer

1985: Cray X-MP / 4 8  
1st vector supercomputer

1989: Cray Y-MP / 4 64

1993: Cray C-90 / 2 128

1994: Cray T3D 64  
1st parallel supercomputer

1995: Cray T3D 128

1998: Cray T3E 256  
1st MPP supercomputer

2002: IBM SP4 512  
1 Teraflops

2005: IBM SP5 512

2006: IBM BCX  
10 Teraflops

2009: IBM SP6  
100 Teraflops

2012: IBM BG/Q  
2 Petaflops
IBM Cluster linux

**Processor type:** 2 six-cores Intel Xeon (Esa-Core Westmere) X 5645 @ 2.4 GHz, 12MB Cache

**N. of nodes / cores:** 274 / 3288

**RAM:** 48 GB/Compute node (14 TB in total)

**Internal Network:** Infiniband with 4x QDR switches (40 Gbps)

**Accelerators:** 2 GPUs nVIDIA M2070 per node

548 GPUs in total

**Peak performance:** 32 Tflops

565 TFlops SP GPUs

293 TFlops DP GPUs

**N. 266 in Top 500 rank** (June 2013)

National and PRACE Tier-1 calls
FERMI@CINECA

- Architecture: 10 BGQ Frames
- Model: IBM-BG/Q
- Processor type: IBM PowerA2 @1.6 GHz
- Computing Cores: 163840
- Computing Nodes: 10240
- RAM: 1GByte / core (163 PByte total)
- Internal Network: 5D Torus
- Disk Space: 2PByte of scratch space
- Peak Performance: 2PFlop/s

- **N. 12 in Top 500 rank** (June 2013)
- National and PRACE Tier-0 calls
Eurora: EURopean many integrated cORE Architecture

• Hybrid cluster based on the evolution of the AURORA architecture by Eurotech
• 64 nodes Intel Sandy Bridge dual socket, (1024 cores in total)
  • 32 nodes at 2.1 GHz and
  • 32 nodes at 3.1 GHz.
• 16 GByte DDR3 RAM, 160 GByte SSD, 1 FPGA Altera Stratix V. per Node
• Interconnection networks: Infiniband QDR and 3D Torus
• Hot water cooling
• The system is equipped with:
  – 64 MIC processors (2 per node on 32 nodes)
  – 64 NVIDIA K20 accelerators (2 per node on 32 nodes)
• Peak performance (K20 accelerated) 175.7 Tflop/s
• N. 467 in Top 500 rank (June 2013)
• N. 1 in Green500 rank (June 2013)
CINECA HPC Infrastructure
Visualisation System

Virtual Theater

• 6 video-projectors BARCO SIM5
• Audio surround system
• Cylindric screen 9.4x2.7 m, angle 120°
• Ws + Nvidia cards
• RVN nodes on PLX system

Remote visualization System

• The system is based on PLX using 2 nodes with 2 NVIDIA Tesla M2070Q per node.
• Paraview visualization tool is accessible through the following technologies:
  - VirtualGL
  - TurboVNC
CINECA & Industries

CINECA provides HPC service to Industry:
- ENI (geophysics)
- BMW-Oracle (American cup, CFD structure)
- Arpa (weather forecast, Meteoclimatology)
- Dompé (pharmaceutical)

CINECA hosts the ENI HPC system:
- HP ProLiant SL390s G7 Xeon 6C X5650, Infiniband,
- Cluster Linux HP, 15360 cores
  plus some other clusters for a total of more than 30K cores
Science @ CINECA

<table>
<thead>
<tr>
<th>Scientific Area</th>
<th>National Institutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemistry</td>
<td>INFM-CNR</td>
</tr>
<tr>
<td>Physics</td>
<td>SISSA</td>
</tr>
<tr>
<td>Life Science</td>
<td>INAF</td>
</tr>
<tr>
<td>Engineering</td>
<td>INSTM</td>
</tr>
<tr>
<td>Astronomy</td>
<td>OGS</td>
</tr>
<tr>
<td>Geophysics</td>
<td>INGV</td>
</tr>
<tr>
<td>Climate</td>
<td>ICTP</td>
</tr>
<tr>
<td>Cultural Heritage</td>
<td>Academic Institutions</td>
</tr>
</tbody>
</table>

Main Activities
- Molecular Dynamics
- Material Science Simulations
- Cosmology Simulations
- Genomic Analysis
- Geophysics Simulations
- Fluid dynamics Simulations
- Engineering Applications

Application Code development/parallelization/optimization
Help desk and advanced User support
Consultancy for scientific software
Consultancy and research activities support
Scientific visualization support
CINECA HPC Training

22nd Summer School on Parallel Computing
May 20 - 31, 2013 - BOLOGNA
July 15 - 26, 2013 - ROME
September 2-13, 2013 - BOLOGNA

9th Advanced School on Parallel Computing
February 11 - 15, 2013 - BOLOGNA

12th Summer School on Scientific Visualization
June 10 - 14, 2013 - MILAN

CINECA PRACE Advanced Training Centre
The HPC Model at CINECA

From agreements with National Institutions to National HPC Agency in the European context

- Big Science – complex problems
- Support Advanced computational science projects
- HPC support for computational sciences at National and European level
- CINECA calls for advanced National Computational Projects

ISCRA Italian SuperComputing Resource Allocation

http://iscra.cineca.it

Objective: support large-scale, computationally intensive projects not possible or productive without terascale, and petascale, computing.

PRAACE CALLS
Realizing the ESFRI Vision for an HPC RI

European HPC-facilities at the top of an HPC provisioning pyramid
- Tier-0: European Centres for Petaflop/s
- Tier-1: National Centres
- Tier-2: Regional/University Centres

Creation of a European HPC ecosystem
- HPC service providers on all tiers
- Scientific and industrial user communities
- The European HPC hard- and software industry
- Other e-Infrastructures

PRACE Research Infrastructure ([www.prace-ri.eu](http://www.prace-ri.eu))
- The top level of the European HPC ecosystem
- CINECA represents Italy in PRACE
- Hosting member in PRACE
  - Tier-0 system BG/Q 2 Pflop/s (50%)
  - Tier-1 system > 5% PLX
PRACE-1IP: a Cornerstone in PRACE History

HPC part of the ESFRI Roadmap; creation of a vision involving 15 European countries

Creation of the Scientific Case

Signature of the MoU

Creation of the PRACE Research Infrastructure


HPCEUR HET PRACE Initiative PRACE RI

PRACE Preparatory Phase Project PRACE-1IP PRACE-3IP

PRACE-2IP
PRACE in Spring 2013 … has grown up

PRACE is the European High-End HPC Infrastructure

- provides access to 6 Tier-0 systems, funded by 4 countries: DE, ES, FR, IT
- is preparing a Tier-1 offer, replacing the current DECI
- has provided 5.5 Billion Core-hours to hundreds of user projects since 2010
- is providing extensive training through 6 PRACE Advanced Training Centres (PATC), seasonal schools, summer student programme, workshops, …
- currently has 25 members, representing 25 countries
- has secured national contributions of 450 M€ and 67 M€ EC funding
- is developing its strategy for 2015-2020

PRACE is user-demand driven

- Scientific Steering Committee advises on all scientific matters, in particular the access to PRACE resources through Peer Review
- User Forum provides feedback and technical requirements from end-users

Scientific Case 2012-2020: requirements and recommendations

>100 leading European Scientists contributed to this update
April, 23rd 2010 creation of the legal entity (AISBL) PRACE with seat location in Brussels, Belgium

25 PRACE Members

www.prace-ri.eu

67+ Million € from EC FP7 for preparatory and implementation phases
Grants INFSO-RI-211528, 261557, 283493, and 312763
Complemented by ~ 50 Million € from PRACE members
**Tier-0**

First production system available:
1 Petaflop/s IBM BlueGene/P (JUGENE) at GCS (Gauss Centre for Supercomputing) partner FZJ (Forschungszentrum Jülich)

**Tier-1**

Sixth production system available by January 2013: 1 Petaflop/s IBM (MareNostrum) at BSC.

**Tier-2**

Fifth production system available by August 2012: 2 Petaflop/s IBM BG/Q (FERMI) at CINECA.

Fourth production system available by mid 2012: 3 Petaflop/s IBM (SuperMUC) at GCS partner LRZ (Leibniz-Rechenzentrum).

Third production system available by the end of 2011: 1 Petaflop/s Cray (HERMIT) at GCS partner HLRS Stuttgart.

Upgrade: 5.87 Petaflop/s IBM BlueGene/Q (JUQUEEN)

Second production system available: Bull Bullx CURIE at GENCI partner CEA. Full capacity of 1.8 Petaflop/s reached by late 2011.
The HPC European e-infrastructure (ESFRI)

- 25 members, AISBL since 2010
- 530 M€ for 2010-2015 (inc 70M€ from UE)
- 6 supercomputers in 4 hosting countries, different architectures
- Research and industrial access (open R&D) for all disciplines based on excellence in science, free of charge
- Nearly 15 Pfolp/s
- 5 billion hours granted since 2010
PRACE: Tier-0 calls

• **Project Regular Access calls**
  It is open to researchers and their collaborators from recognized academic institutions and industry for projects deemed to have significant European and international impact.
  Calls for Proposals are issued twice a year and are evaluated by leading scientists and engineers in a peer-review process.

• **Preparatory Access**
  It is a simplified form of access for limited resources for the preparation of resource requests in response to Project Access Calls for Proposals.
  Type **A** (scalability tests)
  Type **B** (Enabling + Scalability tests)
  Type **C** (Enabling + Scalability tests with PRACE involvement)

www.prace-ri.eu
Top 500: some facts

1976  Cray 1 installed at Los Alamos: peak performance 160 MegaFlop/s \( (10^6 \text{ flop/s}) \)

1993  (1° Edition Top 500) No. 1 59.7 GFlop/s \( (10^{12} \text{ flop/s}) \)

1997  Teraflop/s barrier \( (10^{12} \text{ flop/s}) \)

2008  Petaflop/s \( (10^{15} \text{ flop/s}) \): Roadrunner (LANL) Rmax 1026 Gflop/s, Rpeak 1375 Gflop/s hybrid system: 6562 processors dual-core AMD Opteron accelerated with 12240 IBM Cell processors (98 TByte di RAM)

2011  11.2 Petaflop/s : K computer (SPARC64 VIIIfx 2.0GHz, Tofu interconnect) RIKEN Japan
- 62% of the systems on the top500 use processors with six or more cores
- 39 systems use GPUs as accelerators (35 NVIDIA, 2 Cell, 2 ATI Radeon)

2012  17.5 Petaflop/s (Rmax), 27.2 Petaflop/s (Rpeak) Titan DOE Oak Ridge Nat. Lab Cray XK7, Opteron 6267 2.2 GHz 560,640 cores (including 261,632 NVIDIA K20x accelerator cores.) Cray Gemini interconnect
- 84.6% of the systems on the top500 use processors with six or more cores
- 62 systems use GPUs as accelerators (56 GPU, 6 MIC Phi)

2013 (J)  33.86 PFlop/s (Rmax), 54.80 Petaflop/s (Rpeak) Tianhe-2 (MilkyWay-2) Guangzho Nat Supercomp.Centre CHINA, Cluster, Intel Xeon 2.2 GHz, Intel Xeon Phi (16,000 nodes, each with two Intel Xeon IvyBridge processors and three Xeon Phi processors 3,120,000 cores in total)
- 88 % of the systems use processors with six or more cores
- 54 systems use accelerator/co-processor technology (39 GPU, 11 MIC, 3 ATI)
Multi-core processors

- Motivation for Multi-Core
  - Exploits improved feature-size and density
  - Increases functional units per chip (spatial efficiency)
  - Limits energy consumption per operation
  - Constrains growth in processor complexity

- Challenges resulting from multi-core
  - Relies on effective exploitation of multiple-thread parallelism
    - Need for parallel computing model and parallel programming model
  - Aggravates memory wall
    - Memory bandwidth
      - Way to get data out of memory banks
      - Way to get data into multi-core processor array
    - Memory latency
    - Fragments (shared) L3 cache

Heterogeneous Multi-core Architecture

Combines different types of processors
Each optimized for a different operational modality
Synthesis favors superior performance
Accelerators: No standard (yet).

Programming is hard work.
Real HPC Crisis is with Software

A supercomputer application and software are usually much more long-lived than a hardware
  - Hardware life typically four-five years at most.
  - Fortran and C are still the main programming models

Programming is stuck
  - Arguably hasn’t changed so much since the 70’s

Software is a major cost component of modern technologies
  - The tradition in HPC system procurement is to assume that the software is free.

It’s time for a change
  - Complexity is rising dramatically
  - Challenges for the applications on Petaflop systems
  - Improvement of existing codes will become complex and partly impossible
  - The use of O(100K) cores implies dramatic optimization effort
  - New paradigm as the support of a hundred threads in one node implies new parallelization strategies
  - Implementation of new parallel programming methods in existing large applications has not always a promising perspective

There is the need for new community codes
Trends

Scalar Application

Vector

Distributed memory

MPP System, Message Passing: MPI

Multi core nodes: OpenMP

Accelerator (GPGPU, FPGA): Cuda, OpenCL and Vector

Shared Memory

Hybrid codes
Hybrid parallel programming (example)

Python: Ensemble simulations

MPI: Domain partition

OpenMP: External loop partition

VECTOR or
CUDA: assign inner loops
Iteration to GPU threads