Architecture and performances of the CRESCO HPC system

S. Migliori, G. Bracco, P. D'Angelo, D. Giammattei, M. De Rosa, S. Pierattini, S. Podda, A. Quintiliani, S. Raia, A. Funel, G. Richelli (*)

ENEA-FIM, L.Tevere Thaon di Revel (Roma) Italy (*) IBM Italia

Summary

This poster describes the architecture and the performances of the high performance computing (HPC) system that has been installed to provide the required computing power to the CRESCO project applications and the dedicated activity required to integrate CRESCO HPC system into ENEA-GRID infrastructure.

CRESCO (Computational Research Center for Complex Systems) is an ENEA Project, co-funded by the Italian Ministry of University and Research (MUR). ENEA is the Italian National Agency for New Technologies, Energy and Environment.
The project is functionally built around a HPC platform and 3 scientific thematic laboratories:

- the Computing Science Laboratory, hosting activities on HW and SW design, GRID technology and HPC platform management
- the Computational Systems Biology Laboratory, with activities in the Life Science domain, ranging from the "post-omic" sciences (genomics, interactomics, metabolomics) to Systems Biology;
- the Complex Networks Systems Laboratory, hosting activities on complex technological infrastructures, for the analysis of Large National Critical Infrastructures

CRESCO HPC system consists of more then 2700 computing cores, divided into three main sections. A section is dedicated to applications with high memory requirements, a section is dedicated to high scalable applications and a third experimental section provides systems with Cell processors, FPGA and high performance video adapters dedicated to computational applications. High bandwidth and low latency connections are provided by an InfiniBand 4xDDR network. The main storage consists of an IBM/DDN 9550 system with 160 TB raw data, organized in a GPFS file system.

CRESCO HPC system has been integrated into ENEA-GRID infrastructure which has been developed to provide a unified environment for all the main ENEA HPC resources.

The main software components of ENEA-GRID are the multi-site resource manager LSF Multicluster, the OpenAFS distributed file system, the integrated Kerberos 5 authentication and a Java and Web based Graphical User interface making use of CITRIX technologies.

The choice of mature, reliable and multi-platform software components has permitted along the years to integrate in a GRID oriented infrastructure HPC resources at the state of the art performances, with minimal changes in the user environment.

The architecture of the CRESCO HPC system

CRESCO HPC system has been designed with the aim of offering a general purpose facility based on the leading multi-core x86_64 technology.

In order to provide the best environment for different types of applications the system consists of two main sections respectively oriented (1) for **high memory request** and moderate parallel scalability and (2) for limited memory and **high scalability** cases. Both sections are interconnected by a common Infiniband 4X DDR network (IB) and can operate as a single large integrated system.

The first main section is composed by 42 fat nodes IBM x3850-M2/x3950-M2 with 4 Xeon Quad-Core Tigerton E7330 processors (2.4GHz/1066MHz/6MB L2), 32 GB RAM (12 extra-fat nodes with 64 GB RAM). The total number of cores in the first section is so equal to 672. The x3950-M2 nodes (8) are coupled into 3x smp systems (48 cores/192 GB, 2 units of 3 nodes each) and a 2x smp (32 cores/128 GB, 1 unit of 2 nodes).

The second main section is composed by 256 blades IBM HS21, each supporting dual Xeon Quad-Core Clovertown E5345 processors (2.33GHz/1333MHz/8MB L2), 16 GB RAM for total of 2048 cores. The blades are hosted by the 14 slots blades chassis for a total of 19 chassis and each blade has a dedicated IB connection.

The larger system created by joining the two main sections has 2720 cores and 5.824 TB RAM.

A third **experimental section** consists of 3 subsections dedicated to special processor architectures:

- 4 blades IBM QS21 with 2 Cell BE Processors 3.2 Ghz each.
- 6 nodes IBM x3755, 4 sockets AMD Dualcore 8222 equipped with a FPGA VIRTEX5 LX330 card.
- 4 node IBM x 3755, 4 sockets AMD Dualcore 8222 with a NVIDIA Quadro FX 4500 X2 video card.

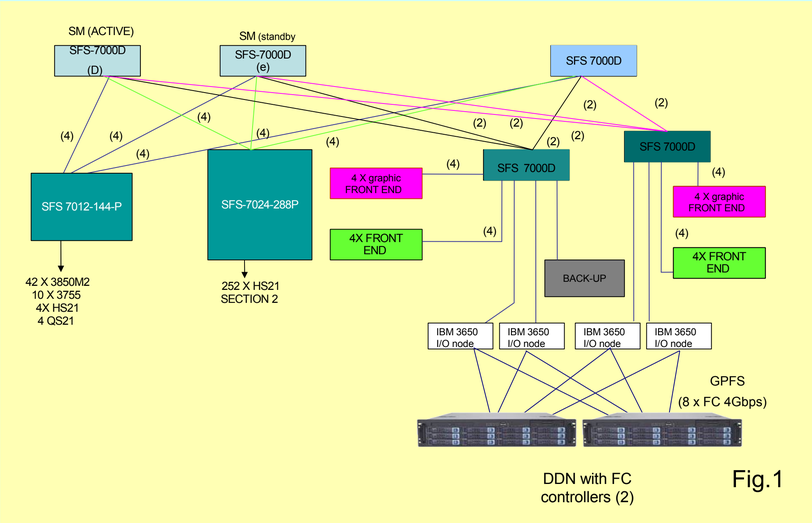


Fig.1

Fig. 1 Architecture of the InfiniBand network including the IBM/DDN 9550 storage system.

The 4 I/O Nodes, directly FC attached to the storage, are the GPFS NSD servers.

The IB network is based on a CISCO SFS 7024 (288 ports), SFS 7012 (144 ports) and 5 SFS 7000 (120 ports). The HCA type for section1 and section 2 hosts are respectively, MT25204 and MT25418.

The Ethernet network consists of one CISCO 4506 (240 ports), 3 CISCO 4948 (144 ports) and 3 CISCO 3750G (144 ports).

The storage of CRESCO HPC system is provided by an IBM DCS9550 system, 160 TB raw space based on 500 GB SATA Hard Disk. An IBM Tape Library IBM TS3500 provides the backup facility.

The power required to run the system reaches 150 kw under heavy load and proper air cooling systems have been provided.

The operating system is RedHat EL 5.1 and the usual set of Portland and Intel Compilers are available. The machine installation has been performed using systemimager [http://wiki.systemimager.org] and 256 nodes are automatically installed in 25' from a configured "golden" node..

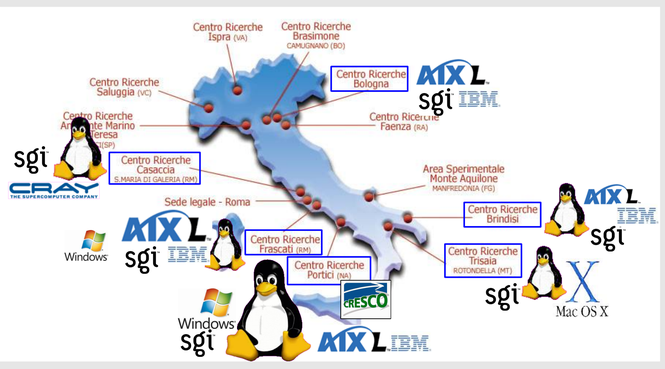
A GPFS parallel file system is shared via Infiniband between all the computing nodes of all the main section of the system. User homes are located in an OpenAFS file system

The three sections together with other 35 service machines (front-end, controls, file-servers, installation servers) and storage and network components make use of a total of 18 standard racks (19", 42 U).

CRESCO Project Computer Room 05/2008

CRESCO location:
Portici (Naples)

ENEA: the Italian National Agency for New Technologies, Energy and Environment



12 Research sites and 6 computer centres managing multi-platform resources for serial & parallel computation and graphical post processing.

ENEA GRID mission [started 1998]:

- ENEA-GRID provides an unified user environment and an homogeneous access method for all ENEA researchers and their collaborators, irrespective of their location.
- ENEA-GRID optimizes the resource utilization

ENEA GRID architecture

GRID functionalities (unique authentication, authorization, resource access and resource discovery) are provided using "mature", multi-platform components:

Distributed File System: OpenAFS + Kerberos V

Resource Manager: LSF Multicluster [www.platform.com]

Unified user interface: Java & Citrix Technologies

These components constitute the ENEA-GRID Middleware.

ENEA-GRID computational resources

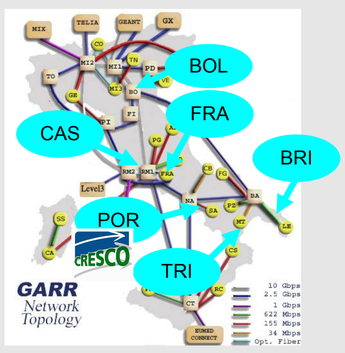
- Hardware (before CRESCO HPC system!!):
  - ~100 hosts and ~650 cpu : IBM SP; SGI Altix & Onyx; Linux clusters 32/ia64/x86_64; Apple cluster; Windows servers. Most relevant resources: IBM SP5 258 cpu; 3 frames of IBM SP4 96 cpu
- Software:
  - Commercial codes (fluent, ansys, abaqus..)
  - Research codes. (mcpn/x, eranos, fluka...
  - Elaboration environments (Matlab, IDL, Scilab...)

ENEA Network connection

ENEA computational resources are distributed over WAN, connected by GARR, the Italian Academic & Research Network

ENEA-GARR
9 PoP, 18-400 Mbps

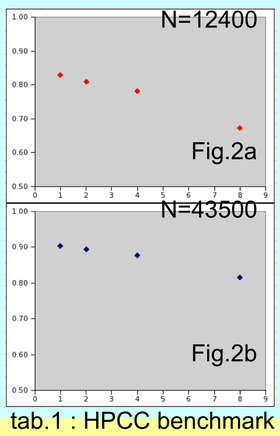
Brindisi 150 Mb/s
Bologna 30 Mb/s
Casaccia 100 Mb/s
Frascati 155 Mb/s
Portici 1000 Mb/s
Trisaia 18 Mb/s
Palermo
Pisa
Roma Sede



Performances

CRESCO HPC system has been installed in May 2008 and is currently under the certification tests.

The HPL benchmark (compiled with gcc 4.1.2, g77 3.4.6 and GotoBLAS 1.24, openmpi 1.2.5) has been run with a result of **17.1 Tflops** using 2500 cores, with 1250 processes each with 2 threads, using 15 GB of the available 16 GB for a case of size N=720000. In the run both sections of the system have been used.



The figures 2 a,b show the cpu efficiency in the HPL benchmark run on a single blade, dual socket Xeon E5345, for two cases with different problem sizes: N=12400 and N=43500, versus the number of threads, up to 8 (1 thread/core)

In the case of the larger size the efficiency of the benchmark with 8 threads is larger than 80%

tab.1 : HPCC benchmark

G-HPL (Tflop/s)	11688
EP-DGEMM Triad (GB/s) (*)	8.36
G-PTRANS (GB/s)	53
EP-Stream Triad (GB/s) (*)	0.37
Stream Triad (GB/s)	2.94
G-Random Access (Gups)	9
FFTE (Gflop/s) (*)	40
Random Ring Lat (usec)	5.72
Random Ring BW (GB/s)	0.48

The HPCC benchmark has been run on a subset of the section2 , 244 nodes and 1952 cores, with 1 process/node and 8 threads/node with a reduced size case N=400000.

The results are shown in tab.1, where (*) marks the cases where the multithreads role has to be investigated.

http://www.afs.enea.it/project/eneagrid http://www.cresco.enea.it