INFN-Pisa scientific computation environment
(GRID, HPC and Interactive Analysis)

S. Arezzini, A. Carboni, G. Caruso, A. Ciampa, S. Coscetti, E. Mazzoni, S. Piras
I.N.F.N. Sezione di Pisa, Largo B. Pontecorvo 6, 56127 Pisa (Italy)
E-mail: enrico.mazzoni@pi.infn.it

Abstract. The INFN-Pisa Tier2 infrastructure is described, optimized not only for GRID CPU and Storage access, but also for a more interactive use of the resources in order to provide good solutions for the final data analysis step. The Data Center, equipped with about 6700 production cores, permits the use of modern analysis techniques realized via advanced statistical tools (like RooFit and RooStat) implemented in multicore systems. In particular a POSIX file storage access integrated with standard SRM access is provided. Therefore the unified storage infrastructure is described, based on GPFS and Xrootd, used both for SRM data repository and interactive POSIX access. Such a common infrastructure allows a transparent access to the Tier2 data to the users for their interactive analysis.

The organization of a specialized many cores CPU facility devoted to interactive analysis is also described along with the login mechanism integrated with the INFN-AAI (National INFN Infrastructure) to extend the site access and use to a geographical distributed community. Such infrastructure is used also for a national computing facility in use to the INFN theoretical community, it enables a synergic use of computing and storage resources.

Our Center initially developed for the HEP community is now growing and includes also HPC resources fully integrated. In recent years has been installed and managed a cluster facility (1000 cores, parallel use via InfiniBand connection) and we are now updating this facility that will provide resources for all the intermediate level HPC computing needs of the INFN theoretical national community.

1. Introduction
The INFN-Pisa Scientific Computational Center was born as a Tier2 for the CMS experiments in the early 2000s. Since then it has been evolved the infrastructure to cope the later step in the data analysis chain, which involves local and interactive activities. In the last few years it has grown as a multidisciplinary center hosting many others facilities for activities of many INFN community above all the theoretical one.

2. Computation resources
INFN Pisa computational environment counts a total number of about 6700 core, 1.3 PB of disk space and a total WAN bandwidth of 2x10Gb/s.

All the computing nodes are grouped in 15 farms and 4 clusters, each one housing identical or similar hardware. Clusters are equipped with high speed interconnections (both InfiniBand or Myrinet) for parallel computing, farms have no high speed low latency network connection and are used for serial jobs.
2.1. Grid
About 4500 cores are dedicated to Grid Computing: all INFN VO are supported, with CMS and THEOPHYS as stockholders. The center is a Tier2 for the CMS Experiment. The core count includes nodes from the Tramontana cluster, actually an HPC resource, initially intended to be used for both serial and parallel jobs via Grid submission for the national theoretical community. Nodes in the Tramontana cluster are connected via InfiniBand DDR 20Gb/s, each with 8 cpu cores and 8 GB of memory. All the grid resources can be used also by direct local submission.

2.2. HPC
The HPC infrastructure is composed by 4 type of resources:

(i) 1600 cores are part of the newest cluster, Theocluster, with InfiniBand QDR 40Gb/s. Each node has 4x16 cpu cores and 512 GB of memory in order to fulfill both cpu/memory bound computations.

(ii) 240 cores are part of the Theonuc farm, here high memory is a key feature, total available RAM is more than 800GB.

(iii) 240 cores are part of the Diablo cluster, with Myrinet interfaces. This cluster is born in collaboration with the Engineering Department of Pisa University and is mainly used for CFD (Computational Fluid Dynamics) using specialized software (both commercial and open) such as Fluent[1], Starcd[2] and OpenFoam[3].

(iv) 130 cores are part of the FAI (Facility For Interactive Analysis), these machines can be accessed (via LSF[4] scheduler) in interactive mode and provide users scientific software.

2.3. Advanced coprocessors
Another field of interest is advanced microprocessor, both GPUs and Intel Mic are available. Seven servers hosts many types of Nvidia GPUs:
- 24 Tesla T10
- 4 Tesla M2090
- 8 Kepler K20

The K20 GPUs are currently tested in cluster configuration using InfiniBand QDR 40Gb/s interconnections.

A single server hosts 2 Intel PHI 5110P which are mainly used for R&D.

3. Data storage and filesystem
After the initial development phase of Tier2 in which the focus was the CPU power, the site has developed a project for a storage infrastructure that would meet the following requirements:
- Architecture made of enterprise-class components;
- Able to cope as much as possible in an automatic manner to the failure of one of the hardware components;
- Would ensure scalable performance and self-balancing in case of failures;
- Able to meet the various needs of disk access: POSIX access (both batch and interactive), SRM;
- That could adapt to the needs of the entire site and not just the specific ones of Tier2;
- That would guarantee a high degree of abstraction between the physical purchase of resources and the availability of those resources for the end users;
- Ease of management and maintenance and minimize the down time due to system maintenance.
The structure that has allowed us to meet these design requirements is based on enterprise-class storage systems Data Direct Network, two S2A9900 and one SFA12k, ensuring high standards of redundancy and excellent performance for the access mode of a typical Tier2 (sequential access of large files). Above such hardware structure we use GPFS[5] as a software platform, creating a cluster of NSD server deputies only export the file system to the nodes that need to access. Using separate systems (both storage and NSD servers) for the metadata has improved performance for interactive activities mitigating the natural inclination of our storage infrastructure for sequential access, due to hardware characteristics of storage systems and filesystem parameters (block size, etc...).

In the initial phase of the life of this structure, the SRM system used by the Tier2 was dCache[6] that, although not required a parallel filesystem such as GPFS, could benefit from its use in terms of redundancy and availability of data allowing decoupling the data from the physical server which granted access. The infrastructure has also demonstrated its versatility by allowing the change of SRM moving from dCache to StoRM[7] without having to move a single bit of data, but simply renaming the existing files in the filesystem.

Last great success of the GPFS structure is the possibility, now implemented by more than a year, to access the data of the Tier2 via Xrootd[8] protocol, also in this case is a great advantage to be able to rely on a unified filesystem infrastructure.

Simultaneously to the typical Tier2 data access should be noted the increasingly use of the storage infrastructure in other ways. Typically as interactive activities or others POSIX access to storage areas, thus allowing the use of local Tier2 data in a Tier3 way.

The same structure ensures access to the data, in all the ways described so far, to all other communities of scientific calculation that lean to the site INFN-Pisa, in particular for the HPC use of the center.

Finally, it is worth noting the high degree of management and redundancy inherent in GPFS that has allowed us to manage systems updates, even major ones, with continuous availability of the filesystem (to the limit with performance degradation) or overcome severe crisis situations, such as those generated after sudden electrical problems, thus ensuring the data reliability and availability.

4. Network architecture

4.1. Geographical connection

The site geographical connection is made via two aggregate Ethernet interfaces of 10Gb/s. The physical link is made through two dark fiber of the University Pisa metro network that are running on independent paths both inside buildings and in the city. These two optical paths are connected on the border router Juniper MX80 that provides both the geographic and inner routing for the Scientific Computational Center and general purpose activities. In this way there is a geographical connection 20Gb/s capable of ensuring redundancy for both kinds of traffic (scientific and general purpose).

Our site is connected to the network LHCOne this is accomplished by routing the traffic of Tier2 on a dedicated default route different from the one used for the general purpose traffic. Due to the redundancy of the physical connection is not required to establish a redundancy mechanism at the routing level.

4.2. Internal LAN connection

The LAN is based on a single concentration switch Force10 E1200 exascale, this switch provides 8x90 1Gb/s connections and 20 of 10Gb/s, all full wire speed. The 10GE connections are used for storage servers (GPFS) and gridftp server. The links of computational nodes are 1Gb/s as well as those of the remaining grid services (CE, LSF, BDII etc ...). The 1GE ports are all copper while those 10GE are all in multimode fiber.
From the addressing point of view all the computational nodes use private IP addresses, public IP are used just for central services. To optimize the performance the servers that need to ensure access to the disk (GPFS and gridftp) are equipped with both types of address, in this way the computational nodes have Layer2 access to storage having guaranteed maximum bandwidth. Inside the site traffic between the two worlds is routed by the border router while geographical outgoing connectivity of computational nodes is guaranteed by a suitable set of the NAT servers.

5. Access and job submission
Access to all the computing resources of the center is granted by a set of user interfaces to all the INFN Theoretical Physics Community and CMS staff and is based upon INFN-AAI credentials. INFN-AAI is a project that aims to create a unique authentication and authorization infrastructure for INFN, with these unique credentials users can access the facility avoiding complex registration procedures.

User interfaces are intended to be used just for job submission, for this reason just a small set of software is available on those machines. Job submission is implemented via LSF Scheduler, both versions 7 and 9 are currently used, and is based over batch queues. Both grid and local submissions are possible. Each queue addresses a set of identical (or similar) servers and is used for specific computing needs like high memory jobs, parallel/serial jobs, interactive jobs, GPU specific jobs and so on. The correct access priority among user is granted by LSF fairshare metrics, we are currently testing an implementation which mixes fairshare with runtime accounting. The monitoring of the computing nodes is based on Ganglia, we have a in house developed web interface for job monitoring and we are currently investigating usability of IBM RTM[9] (Real Time Monitoring) and PAC[10] (Portal Access Control) to address respectively a full featured monitoring of the farms and a web interface for job submission.

6. Collaborations
The INFN-Pisa Scientific Computing Center is open to external collaboration from many years. These collaboration are in two groups:

- GRID: we are open to all the VO recognized by INFN. So we are open to accept GRID jobs from many scientific institutions in all the world. There are mainly two VO that use the INFN-Pisa Grid Center heavily:
  - Compchem: 389251 jobs submitted for 17625824683 sec of CPU-time
  - Biomed: 1231953 jobs submitted for 6036744969 sec of CPU-time
- Industrial collaboration. In this case the work is mainly based on CFD applications. As already mentioned, at INFN-Pisa we have a strong collaboration with the Aerospace Engineering Department of the Pisa University: we have three clusters dedicated to the parallel CFD (and structure) simulations. This collaboration led to external simulations for many industrial firms, performed together with the engineers. The industrial collaboration, during the last years, included: Ferrari Auto, Ducati, Oto Melara, Piaggio, Finmeccanica and many yacht local manufacturers.

7. Conclusions
During the past years our Computing Center has made some important choices. In particular we realized a strong integration between very different resources in order to optimize them. We were asked to produce a greater quantity of services with the same amount of resources, so we had to organise different use cases running on the same hardware and with similar software. This required and continues to require great organizational effort but we believe this is the right way to combine good quality with significant savings.
References
[1] ANSYS Fluent
http://www.ansys.com/Products/Simulation+Technology/Fluid+Dynamics/Fluid+Dynamics+Products/ANSYS+Fluent

[2] Adapco STAR-CD
http://www.cd-adapco.com/products/star-cd

[3] OpenFOAM
http://www.openfoam.com

[4] LSF scheduler
http://en.wikipedia.org/wiki/Platform_LSF

[5] General Parallel File System
http://www.ibm.com/systems/software/gpfs

[6] dCache SRM
http://www.dcache.org

[7] StoRM SRM
http://stormforge.cnaf.infn.it

[8] XrootD
http://xrootd.org/

[9] IBM Platform RTM
http://www.ibm.com/systems/technicalcomputing/platformcomputing/products/lsf/rtm.html

[10] IBM Platform Application Center
http://www.ibm.com/systems/technicalcomputing/platformcomputing/products/lsf/applicationcenter.html