Optimization of HEP Analysis Activities Using a Tier2 Infrastructure

S Arezzini, G Bagliesi, T Boccali, A Ciampa, S Coscetti, E Mazzoni, S Sarkar, S Taneja
1 INFN Sezione di Pisa, Pisa, ITALY
2 Scuola Normale Superiore, Pisa, ITALY
3 Saha Institute of Nuclear Physics, Kolkata, INDIA
4 Dipartimento di Informatica dell’Universita’ di Pisa, Pisa, ITALY

Abstract. While the model for a Tier2 is well understood and implemented within the HEP Community, a refined design for Analysis specific sites has not been agreed upon as clearly. We aim to describe the solutions adopted at the INFN Pisa, the biggest Tier2 in the Italian HEP Community. A Standard Tier2 infrastructure is optimized for Grid CPU and Storage access, while a more interactive oriented use of the resources is beneficial to the final data analysis step. In this step, POSIX file storage access is easier for the average physicist, and has to be provided in a real or emulated way. Modern analysis techniques use advanced statistical tools (like RooFit and RooStat), which can make use of multi core systems. The infrastructure has to provide or create on demand computing nodes with many cores available, above the existing and less elastic Tier2 flat CPU infrastructure. At last, the users do not want to have to deal with data placement policies at the various sites, and hence a transparent WAN file access, again with a POSIX layer, must be provided, making use of the soon-to-be-installed 10 Gbit/s regional lines. Even if standalone systems with such features are possible and exist, the implementation of an Analysis site as a virtual layer over an existing Tier2 requires novel solutions; the ones used in Pisa are described here.

1. Introduction
Since their start-up phase, LHC Experiments[1] have based their computing model on the prescriptions of MONARC[2]. In this model, the centres where physicists are supposed to carry on most of their analysis activities are Tier2s and eventually Tier3s (not really present in the initial implementation used by the Experiments). Tier2 centres are specialized in hosting Monte Carlo simulations and jobs sent by users, and were tested deeply during a number of service challenges prior to the LHC collisions. For what concerns analysis models, the tests allowed to verify and optimize the user job submission via Grid infrastructures (like WLCG[3], OSG[4], …), but rarely have focussed on a later step in the data analysis chain, which involves local and interactive activities. The Grid Data Centre (GDC) in Pisa was born as a Tier2 for the CMS experiment, and has grown as a multidisciplinary centre hosting also activities from other branches of research. We faced the need to maintain a completely working Tier2 infrastructure (as requested by the CMS Experiment), while at the same time introducing features needed to optimize interactive and local analysis activities.

2. Activities at the Centre
As a CMS Tier2, the Centre must guarantee CPU cores and storage for the experiment’s needs.

1 *corresponding author: Giuseppe.Bagliesi@cern.ch
The Pisa GDC offers
- For CMS Monte Carlo production activities: ~1000 cores and 50 TB of temporary storage space;
- For CMS data analysis via Grid: ~1000 cores and a data area of ~800 TB;
- Support for local CMS analysis: 500 cores to be shared by ~50 users, and an additional 100 TB of local space;
- Support for CMS interactive activities: a set of load balanced interactive machines, with a few cores per each of the potential (50) users.
- Additional ~3000 cores and 200 TB for other users; they can be used by CMS if free.

The Centre uses LSF[5] as main tool to balance the ~5000 cores present, and guarantee the different users a correct fairshare, and different storage technologies to give access to the ~1 PB present.

3. Storage Configuration
A Center used for diverse activities must offer a solid storage infrastructure, accessible through all the protocols needed by the sum of the use cases. The possibilities offered by the Pisa GDC are described in the following.

- SRM[6] for WAN transfers (we use for that Storm[7] and dCache[8]);
- Multiple protocols for local access (dcap, Xrootd[9], POSIX);
- Xrootd for WAN direct data access ("streaming");

While all protocols have different strong and weak points, we feel that at least interactive users should be offered a simple POSIX layer, and should not be forced to learn different storage technologies. For this reason, we are decommissioning our dCache storage in favor of a Storm/GPFS[10] complete solution (complemented with a Xrootd/GPFS for streaming purposes).

Figure 1 offers a graphical description of the fluxes.

![Figure 1. Allowed LAN and WAN access protocols.](image)

4. CPU Configuration
The Center allows the use of its computing cores via a number of different technologies:
- Use the site provided User Interfaces to submit jobs using Grid infrastructures;
- Use direct local submission to LSF;
- Use interactive nodes;
- Use processes spanning many cores on a single machine (multi-threaded applications like RooFit[11] and RooStat[12]);
- Use linked processes running on different systems (like a PROOF[13] farm).
LSF serves as a common interface to direct tasks in the proper order and location. While the farming part (used for LSF direct submission and Grid submission) is standard, and consists in LSF queues governed by a fairshare system, for the interactive and multicore part a special solution in which LSF serves interactive nodes has been found. This has multiple advantages: LSF serves as a load balancer between the different interactive nodes, making sure users do not step on each other’s foot, and at the same time can reserve jobs with multiple cores, either on a single or multiple machines. LSF also takes care to kill user jobs if they exceed the resources (RAM, CPU time) declared when requesting CPU slots.

5. Site Infrastructure

The Pisa GDC exceeds largely the typical dimensions of a Department level computing centre. As such, it is equipped with enterprise level facilities for the infrastructural part.

5.1. Storage Infrastructure

At the heart of the storage infrastructure Pisa chose to use a GPFS file system, realized on two (currently) big SAN systems, DataDirect DDN S2A9900[14], hosting up to 1200 disks in total. The systems can scale up to 3.6 PB, while currently hosting slightly more than 1 PB. Connections to the GPFS are realized with 32 8 Gbit/s Fiber Channel connections, via a Fiber Channel switch. The GPFS configuration is shown in Figure 2, and is fully redundant using FC Multipath.

The GPFS Metadata are stored on a different system, optimized for high I/O and low bandwidth; we chose a DDN EF3015[15] for the purpose. Since the two current DDN S2A9900 machines are optimized for big files (they are supposed to host Data and Monte Carlo simulations), we decided recently to set up a third system optimized for smaller files, to be used to host users’ working areas. For this, we have currently opened a tender for a DDN 12000[16] new machine, half a PB in size.

5.2. Network Infrastructure

The Pisa GDC hosts currently roughly 5000 computing cores, connected to 1 PB of storage. Most of the machines are with 8+ cores per system, but we still have some older 4 core machine, and some newer 24 core systems. All in all, there are roughly 800 hosts connected as computing cores, plus

![Figure 2. GPFS configuration at the Pisa GDC.](image)
those serving as storage nodes. We decided long ago to use a flat switching matrix for the Center, using a Force10 ExaScale[16] fabric. All the computing nodes are connected via 1 Gbit/s Ethernet, while all the storage nodes (GPFS, Storm and dCache) have 10 Gbit/s fiber connections.

5.3. Rack space, Cooling, Power distribution, …

The Pisa Center is currently equipped with 34 full height racks (close to its limit), and can host machines up to a power usage of 300 KVA. All the services, plus a fraction of the computing nodes, are served by an UPS system, which guarantees site operations also in case of power transient problems. The cooling is guaranteed by 12 APC InRow[18] air sources, cooled by 3 Emerson Chillers[19] (300 kW of refrigerating capacity).

6. Monitoring

We needed to implement a complex infrastructure to guarantee an easy monitoring of the facilities. These complement the standard CMS and WLCG existing tools, focusing more on the monitoring of the infrastructure.

6.1. LSF Monitoring

We implemented a complete set of tools to provide a live and historical information of site usage, LSFMON[20]. The system is now in use in various CMS Tier2s, and also in sites not participating in HEP. A typical report can be seen in Figure 3.

![Figure 3. LSFMON historical view.](image)

6.2. Cooling and Power Monitoring

When dealing with a site as large as Pisa, it is important to have a unique view of the status of the infrastructure, for fast reaction to problems. We implemented a synoptic view of the full site status, as can be seen in Figure 4.
7. Conclusions
We described in this paper the status of the facilities available in Pisa. Starting from a standard Tier2 infrastructure, the site has optimized access to storage, CPU and network in order to allow for efficient interactive data analysis activities. The site can now be used for a diversity of use cases not envisaged at the moment of its initial planning.

The current work has been partially developed under PRIN 2008MHENNA MIUR Project.

8. References

[1] http://lhc.web.cern.ch/lhc/
[2] http://www.cern.ch/MONARC
[3] http://lcg.web.cern.ch/lcg/
[4] http://www.openscienceGrid.org
[5] http://www.platform.com/workload-management/high-performance-computing
[6] https://twiki.grid.iu.edu/bin/viewfile/Education/Syllabus/SRM.pdf
[7] http://storm.forge.cnaf.infn.it/
[8] http://www.dcach.org/
[9] http://xrootd.slac.stanford.edu/
[10] http://www-03.ibm.com/systems/software/gpfs/
[11] http://root.cern.ch/drupal/content/proof
[12] https://twiki.cern.ch/twiki/bin/view/RooStats/WebHome
[13] http://www.ddn.com/products/s2a9900
[14] http://www.ddn.com/pdfs/EF3015_Datasheet_5.pdf
[15] http://www.ddn.com/pdfs/SFA12K-20_Family_Datasheet_V1.pdf
[16] http://www.dell.com/us/enterprise/p/force10-e-series
[17] http://www.apc.com/products/family/?id=339
[19] http://www.emersonnetworkpower.com/EMEA/PRODUCTS/PRECISIONCOOLING/DATACENTERCHILLER/Pages/default.aspx
[20] http://sarkar.web.cern.ch/sarkar/doc/lsfmon.html