Journal of Physics: Conference Series **396** (2012) 042016 doi:10.1088/1742-6596/396/4/042016

The INFN Tier-1

G. Bortolotti, A. Cavalli, L. Chiarelli, A. Chierici, S. Dal Pra, L. dell'Agnello, D. De Girolamo, M. Donatelli, A. Ferraro, D. Gregori, A. Italiano, B. Martelli, A. Mazza, M. Onofri, A. Prosperini, P. P. Ricci, E. Ronchieri, F. Rosso, V. Sapunenko, R. Veraldi, C. Vistoli. S. Zani

INFN-CNAF V.le Berti Pichat 6/2 40127 Bologna, Italy

E-mail: luca.dellagnello@cnaf.infn.it

Abstract. INFN-CNAF is the central computing facility of INFN: it is the Italian Tier-1 for the experiments at LHC, but also one of the main Italian computing facilities for several other experiments such as BABAR, CDF, SuperB, Virgo, Argo, AMS, Pamela, MAGIC, Auger etc.. Currently there is an installed CPU capacity of 100,000 HS06, a net disk capacity of 9 PB and an equivalent amount of tape storage (these figures are going to be increased in the first half of 2012 respectively to 125,000 HS06, 12 PB and 18 PB). More than 80,000 computing jobs are executed daily on the farm, managed by LSF, accessing the storage, managed by GPFS, with an aggregate bandwidth up to several GB/s. The access to the storage system from the farm is direct through the file protocol. The interconnection of the computing resources and the data storage is based on 10 Gbps technology. The disk-servers and the storage systems are connected through a Storage Area Network allowing a complete flexibility and easiness of management; dedicated disk-servers are connected, also via the SAN, to the tape library. The INFN Tier-1 is connected to the other centers via 3x10 Gbps links (to be upgraded at the end of 2012), including the LHCOPN and to the LHCONE. In this paper we show the main results of our center after 2 full years of run of LHC.

1. Introduction

The Large Hadron Collider (LHC) has been running for more than 2 years. In the WLCG (Worldwide LHC Computing Grid) [1] structure, the Computing centres are organized according to the MONARC model [2], i.e. in a hierarchy separating the centres in Tier levels: INFN-CNAF (CNAF in the following) is a Tier-1 for all four LHC experiments (ALICE, ATLAS, CMS, LHCb). It hosts also a Tier-2 for the LHCb experiment. The Tier-1 at CNAF is the main INFN computing facility in Italy: besides the WLCG experiments, it provides computing and storage resources to several other experiments both running with accelerators data (i.e. Kloe, CDF, BABAR, SuperB) and without accelerators i.e. astro-particle physics experiments (e.g. AMS, Argo, Auger, Borexino, Fermi, Icarus, Magic, Pamela and Virgo). Even with such a large number of different experiments, with different requirements and computing models, our strategy, since the beginning of the Tier-1 activity, has been to offer the same set of services to all of them. Therefore we have only one farm shared among all the experiments, and the same Mass Storage System, GEMSS is used by all experiments at CNAF. The EMI (European Middleware Initiative) Grid middleware is used to operate both computing and storage resources. Moreover, all the services are redundant both at hardware and software levels in order to guarantee a true

24x7 support (e.g. storage services can support the failure of up to 50% of the disk-servers without service interruption): this is also possible due to a proactive monitoring system (see section 6). Both the uniformity of offered services and their internal redundancy allow to manage all the resources, namely more than 1000 servers, 12 PB of disk space and 1 tape library with a technical staff of 20 FTEs. During these years of LHC data taking, with the consolidation of the services offered to the experiments, a non negligible effort has been devoted to strengthen the internal structure of the CNAF Tier-1: the four internal sub-groups (Infrastructure, Farming, Network and Storage) will be briefly described in the following sections. The last section will describe our monitoring system.

2. Infrastructure

The CNAF Tier-1 has $\sim$ 120 equipped racks distributed in an area of $\sim$ 1000 square meters (only part of which are currently in production).

In order to guarantee the best continuous uptime, requirement for a Tier-1, the infrastructure has been built to meet key requirements in terms of reliability, redundancy and maintainability.

The key features of the electrical system are two 2500 kVA transformers (with an extra 2500 kVA transformer to guarantee a N+1 redundancy) and two 1700 kVA rotary diesel UPS systems (i.e. each of them is composed by a high-mass spinning flywheel and a diesel engine) providing redundant emergency power for computational, storage and network units up to 3400 kVA. An additional diesel engine provides emergency power for chilling units up to 1200 kVA. The energy, through two completely independent electric lines (one from each rotary UPS) is distributed to the racks through busbars, providing 2N redundancy at rack level: therefore servers with two power supplies can avoid the vast majority of site infrastructure electrical failures occurring between the uninterruptible power supply and the computer equipment, thus guaranteeing a dramatic uptime improvement [3]. In order to lower the Power Usage Effectiveness (PUE) of our system from the current value (1.6) to 1.5, we are studying the optimization of the two rotary UPS systems which are apparently very efficient under an high load while for the time being our IT absorbed power is of the order of 600 KW.

The cooling system is based on 6 chillers (able to provide free-cooling) which guarantee 2 MW cooling power with an outside air temperature of 40 ^{o}C. The circulation of the chilled water is guaranteed by 2+2 pumps (2N redundancy). The racks and the air conditioning units (50 kW of heat dissipation each) are structured inside high density islands, in which hot and cold aisle are physically separated in order to avoid hot and cold air mixing, thus increasing the cooling efficiency.

The Tier-1 computing centre is also equipped with a gaseous fire suppression system.

All the system critical points are monitored by an ad-hoc supervisory and control system enabling us to spot problems and failures.

3. Farming

The Tier-1 common cluster (also including the co-located Tier-2 resources) delivers 125 kHS06 distributed on 10,000 CPU cores in use by about 20 experiments: more than 80,000 computing jobs are executed every day on the computing farm. Part of the jobs are executed on Virtual Machines which are transparently and dynamically provisioned as Virtual computing nodes using the INFN Worker Nodes on Demand Service (WNoDeS) [4]. For each experiment, usually only one queue is defined on the farm. The job scheduling is performed by LSF [5] scheduler and the computing resource allocation is based on the 'Hierarchical Fairshare' scheduling. No resources are dedicated to any experiment and the slot allocation to a job is done according to the following principles: it is directly proportional to the assigned share (namely the resource quota on the farm) of the experiment (or subgroup) the submitter of the job belongs to, and

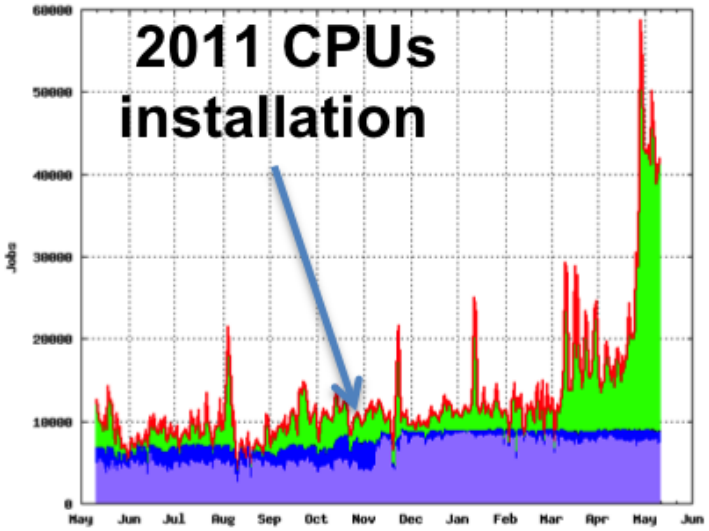

Figure 1. Tier-1 farm usage (May 2011 - May 2012)

inversely proportional to the historical resource usage. This strategy allows to maximize resource utilization: the Tier-1 farm is nearly always 100% full of jobs (see fig. 1).

We are currently testing some extension to allow the "job packing", i.e. the clustering of a set of jobs on a subset of computing nodes, to allow some special use cases such as parallel processing and the "whole node" i.e. the allocation of a whole computing node to a process).

Each computing node delivers one job slot per physical core: we are considering to enable the hyper-trading mechanism (and hence the number of job slots per motherboard) taking into account the upper limit given by the network (for each job a bandwidth of 5 MB/s is provided while, currently, each computing node has 1 Gbps link).

The computing resources can be accessed both via grid and directly: the direct access is provided for "legacy" experiments or for those not fully exploiting the potential of the grid.

Currently we install and manage the farm, such as most of the other Tier-1, with Quattor [6] but we are considering also Puppet.

4. Network

The Tier-1 internal network is based on four core switches linked together by 6x10Gbit interconnections. We have a total of 248 10 Gbps ports, number which is currently increasing. All the racks containing computing nodes are equipped each with two gigabit switches with up to 20 Gbps uplinks to the core in order to guarantee the above mentioned 5 MB/s for each job to the storage resources. On the other hand, the storage servers are connected with Fiber Channel Interfaces to the Storage Area Network and via a 10 Gbit interface to the network $\sim$ 15% of storage resources still use 2xGbps connections).

CNAF is also one of the main nodes of GARR, the Italian academic and research network [7]. The Tier-1 is connected to CERN and the other Tier-1s with two dedicated 10Gbit connections (LHCOPN) with a 10 Gbps backup line through KIT, the German Tier-1. CNAF is also interconnected to the LHCONE. A further 10Gbit link is used for Tier-1 $\Leftrightarrow$ Tier-2 and general purpose traffic (see fig. 2).

It is foreseen an upgrade for the LHCOPN connection to 100 Gbps at the beginning of 2013.

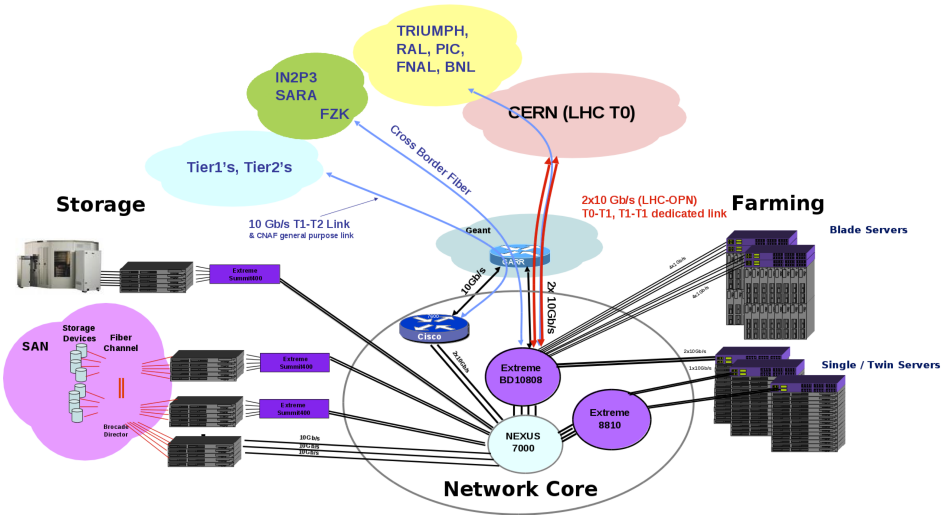

Figure 2. CNAF network infrastructure

5. Storage

CNAF is currently using a model where all our hardware storage systems (disk and tape drives) are accessed through Storage Area Network (SAN) switches and servers running Scientific Linux as operating system and equipped with redundant HBA (Fibre Channel Host Bus Adapter). This has demonstrated so far to be a robust, stable and very flexible approach. The disk storage for a total of $\sim$ 9 PetaByte (PB) of net disk space (to be increased to $\sim$ 12 PB by the end of Q2 2012) is composed by several storage systems (7 Data Direct DDN systems S2A 9950 for a total of 7 PB, 7 EMC2 CX3-80 + 1 EMC2 CX4-960 for a total of 1.9 PB and 3 Fujitsu Eternus DX400 S2 for a total of 2.8 PB currently under installation) and is served to the farm through 40 disk servers with 10 Gb/s Ethernet connection and 90 disk servers with 2x1 Gb/s Ethernet connection (the gigabit server being for the oldest systems, namely the EMC CX).

The tape storage (14 PB of uncompressed space) is provided by an Oracle SUN SL8500 tape library with 20 Oracle T10KB drives (100 MB/s of bandwidth and 9000 1TB tape cartridges) and 10 Oracle T10KC drives (200 MB/s of bandwidth and 1000 5TB tape cartridges). The connections to the tape drives use a subset of the Fibre Channel SAN that is referred to as Tape Area Network (TAN).

Both the disk space and the tape space, structured as a Hierarchical Mass Storage system, is managed by GEMSS [8], the Grid Enable Mass Storage System, an home made integration of IBM General Parallel File System (GPFS)[9] with the IBM Tivoli Storage Manager (TSM) [10].

GPFS is a clustered file-system which provides a scalable POSIX file access: the farm computing nodes do not need a direct connection to the storage backend thought the SAN, but they access the data using the GPFS disk-servers (which have redundant Fibre Channel connections to the SAN).

The disk space is subdivided into different GPFS clusters: a total of seven clusters is used, one cluster for each of LHC experiment, one dedicated to SuperB/BaBar and two shared between no-LHC users. A further GPFS cluster comprising only the farm computing nodes is provided. In general, for each experiment, a disk only file-system and file-system with HSM features are provided (see fig. 3). The farm computing nodes and the User Interfaces statically mount these file-systems (which are POSIX compliant) and the access to the data is performed using the file protocol as they were local to the nodes. Therefore roughly 12000 CPU cores corresponding to

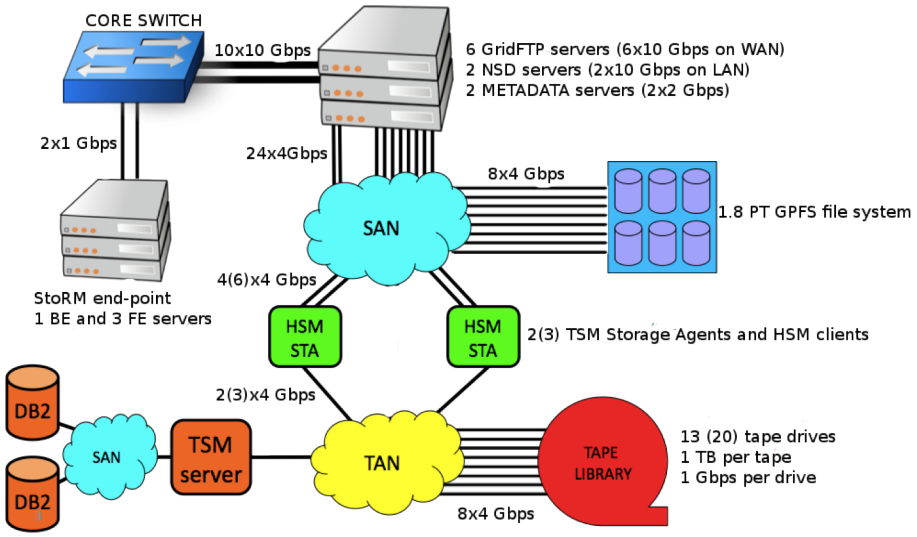

Figure 3. GEMSS layout for a typical experiment served at CNAF

a computing power of about 125 kHS06 are currently directly accessing via POSIX the GEMSS filesystems.

The storage resources are also accessible from the WAN by means of GridFTP servers through the storage resource manager (SRM) protocol.

StoRM [11] is a storage resource manager developed by INFN, which allows Grid applications to interact with storage resources through standard POSIX calls. Its modular design allows for high availability and scalability. StoRM is able to take advantage of high performance file systems like GPFS to improve performance.

Finally, an Oracle clustered database infrastructure is deployed for relational data storing/retrieving. Oracle technologies like Automatic Storage Manager, Real Application Clusters and Streams are exploited. Presently the Oracle database service is composed by 32 dual-core servers organized in 7 clusters serving a total amount of 5TB of data.

6. Monitoring

The Tier-1 monitoring and control system is composed by two independent levels.

The first layer is composed by a set of Nagios servers (at least one for logical service, such as farming, storage and network) that collect information from all the relevant services and hosts. When an error occurs on one of the monitored services or hosts, the controlling Nagios server is able to perform predefined actions such as trying to restart the interested service, remount a file-system or, in case of a cluster service based on a DNS alias, to delete the faulty host from the DNS alias (and to reregister when the hosts is back into operation). In any case, the Nagios server signals the issue via email or, in case of a blocking problem, via SMS to the on-duty experts.

The second layer is the dashboard: it collects all relevant information from the Nagios servers about servers, services, worker nodes and switch availability status, presenting it on a private web page accessible to authorized users. This page allows the on duty people to have a synthetic vision of the status of the center (see fig. 4).

Other monitor informations, for statistical and historical purposes, are gathered through Lemon (for the servers CPU and RAM usage, GPFS throughput etc.. are collected) and MRTG (it collects all the network bandwidth statistics).

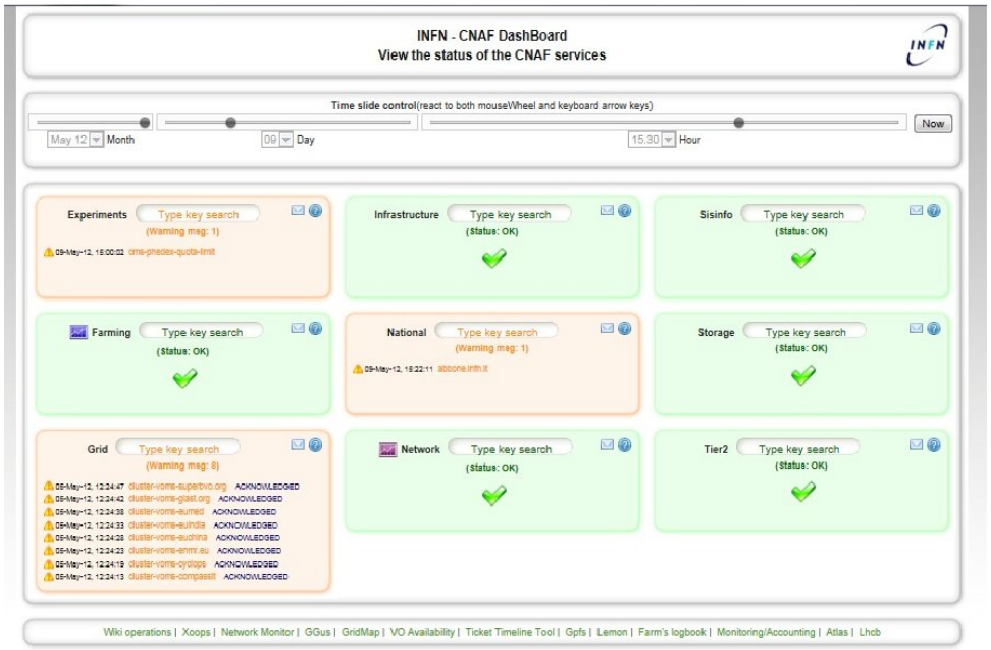

Figure 4. The CNAF Tier-1 dashboard

References

[1] The Worldwide LHC Computing Grid web site: `http://lcg.web.cern.ch/LCG`
[2] The Models of Networked Analysis at Regional Centres for LHC Experiments web site: `http://monarc.web.cern.ch/MONARC/`
[3] Pitt Turner IV W, Seader J.H. and Brill K.G., `Tier Classifications Define Site Infrastructure Performance`, **10** white paper of The Uptime Institute
[4] The Worker Nodes on Demands Service web site: `http://web.infn.it/wnodes/index.php/wnodes`
[5] The IBM Platform LSF web site: `http://www-03.ibm.com/systems/technicalcomputing/platformcomputing/products/lsf/index.html`
[6] The Quattor installation system web site: `http://www.quattor.org/`
[7] The GARR Research Network. web site: `http://www.garr.it/`
[8] D. Andreotti, et al, `INFN-CNAF Tier-1 Storage and Data Management Systems for the LHC Experiments` International Conference on Computing in High Energy and Nuclear Physics (CHEP 2010) IOP Publishing Journal of Physics: Conference Series 331 (2011) 052005 doi:10.1088/1742-6596/331/5/052005
[9] The IBM General Parallel File System web site: `http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/topic/com.ibm.cluster.gpfs.doc/gpfsbooks.html/`
[10] The IBM Tivoli Storage Manager website: `http://www-01.ibm.com/software/tivoli/products/storage-mgr/`
[11] Forti A., Magnoni L., Vagnoni V., Zappi R. et al. `Storm: a SRM Solution on Disk Based Storage System` in Proceedings of the Cracow Grid Workshop 2006 (CGW2006), Cracow, Poland;