Journal of Physics: Conference Series **331** (2011) 052005 doi:10.1088/1742-6596/331/5/052005

INFN-CNAF Tier-1 Storage and Data Management Systems for the LHC Experiments

D. Andreotti, S. Antonelli, A. Cavalli, C. Ciocca, S. Dal Pra, L. dell'Agnello, C. Genta, D. Gregori, B. Martelli, A. Prosperini, P.P. Ricci, L. Rinaldi, V. Sapunenko, V. Vagnoni

INFN-CNAF, Viale Berti Pichat 6/2, Bologna I-40127, Italy

E-mail: claudia.ciocca@cnaf.infn.it

Abstract.

INFN-CNAF, the italian Tier-1, houses more than 10 PB of tape space and 6 PB of disk space used by the global physics user communities including the LHC experiments. In this context, an innovative solution to provide the highest level of reliability, scalability and performances on the installed mass storage system has been developed and used on the production system. The solution, called Grid Enabled Mass Storage System (GEMSS), is based on two components: a layer between the IBM GPFS and TSM HSM, and StoRM, designed to allow direct access to the storage resources by file protocol as well as standard Grid protocols. During this year, both LHC and non-LHC experiments have agreed to use GEMSS with SRM endpoint at INFN Tier-1 after the excellent results observed during their GEMSS validation phase.

In this paper we describe the storage system setup, the data workflow and computing activities of the LHC experiments, and main results obtained at INFN-CNAF Tier-1 in the first year of LHC data taking. In particular, we show the system response to a sustained huge load on the mass storage system and the computing infrastructure.

1. Introduction

INFN-CNAF, located in Bologna, is the Information Technology Centre of INFN (National Institute of Nuclear Physics). In the framework of the Worldwide LHC Computing Grid (WLCG) INFN-CNAF is one of the eleven worldwide Tier-1 centers. The WLCG project is a global collaboration, leveraging on top of the EGI, OSG and Nordugrid communities, which allows physics communities to share the resources offered by the computing centers that are organized in a hierarchy of Tiers according to their specific mission. The data produced at the Large Hadron Collider (LHC) are distributed from CERN, identified as Tier-0, to the Tier-1 centers that make them available to about two hundreds Tier-2 centers all over the world. As Italian Tier-1 of WLCG, INFN-CNAF provides the resources of storage (disk space for short term needs and tapes for long term needs) and computing power needed for data processing and analysis to the LHC scientific community. Moreover, INFN-CNAF houses computing resources for other particle physics experiments, like BaBar at SLAC and CDF at Fermilab, as well as for astroparticle and spatial physics experiments, like the AMS, GLAST and PAMELA satellites, the Auger, ARGO and VIRGO observatories, the MAGIC telescope.

INFN-CNAF, in order to guarantee to its users a high level of reliability, scalability and performances, has developed the Grid Enabled Mass Storage System (GEMSS) [1], a new mass

storage solution. The GEMSS development started in 2006 after the exhaustive and satisfying results from a series of functional and stress tests providing a first proof of concept; these tests proved that the combination of a parallel filesystem with a storage implemented on a Storage Area Network (SAN) behaved in a much more scalable and reliable manner respect to solutions based on traditional filesystems. In more detail, a comparison work among different mass storage solutions, such as CASTOR, the IBM General Parallel File System (GPFS) [2], dCache and xrootd was performed and, therefore, IBM GPFS was chosen because of its better performances [3]. Furthermore, having to enable the access to the WLCG users, the interaction between the storage system and the Grid layers was handled by StoRM [4], an implementation of the standard Storage Resource Management (SRM) (http://sdm.lbl.gov/srm-wg) interface developed at INFN. The adoption of the StoRM layer over a suitable GPFS testbed effectively provided a clear evidence about scalability and reliability of the StoRM and GPFS coupling [5].

After this first implementation [6], two of the largest experiments hosted at the INFN Tier-1 agreed to migrate their data to the new system for their disk-only activity (a disk-only access is called D1T0 Storage Class, whereas a tape based storage is called D0T1 when tape-only or D1T1 if either disk and tape access). The first LHC experiments to adopt StoRM and GPFS as D1T0 storage were LHCb [7] and ATLAS [8]: their excellent results in official production environment were the final push for implementing a complete Mass Storage System (MSS) solution that could fulfill all the LHC requirements also for tape resident Storage Class. Firstly, to integrate the GPFS filesystem with the tape media, the "external pool" concept, introduced since GFPS 3.2, was chosen: it can be used to migrate data from and to different storage areas. Secondly a software for managing the tape library, drive and tape interactions with the data movement was needed and the selected choice was the IBM Tivoli Storage Manager (TSM) [9] software that was then interfaced with the GPFS external pool.

To get an efficient interaction between GPFS and TSM, a layer consisting of user-level programs was developed. This work has been fundamental to manage long latency operations such as recalls from tape: we achieved this by performing a clever reordering on the recall queue, thus reducing the number of tape library accesses. We also implemented facilities to handle the error conditions, providing a useful set of information to the administrators for troubleshooting purposes. Finally, the StoRM layer was extended to support tape Storage Classes such as D1T1 and D0T1. After summer 2009, the GEMSS solution that combines the StoRM, GPFS, and TSM systems was put in production with the storage at INFN-CNAF for the LHC experiments, after having migrated their data to GEMSS.

In this paper, we describe GEMSS, pointing out the way it can support physics communities.

2. GEMSS

The main components of the system are: the GPFS filesystem layer that is spread over different SAN devices to optimize the overall throughput; the TSM software that is used to manage the tape layer access; the StoRM layer that is used in conjunction with the GridFTP servers to provide remote Grid access; and different customized programs for the data migration operation (i.e., data flow from disk to tape), for the data recall process (i.e., data flow from tape to disk) and application data access.

GPFS is a clustered parallel filesystem, providing a scalable POSIX access to files. Client machines do not need to have direct access to the SAN, whereas they access the disk devices via LAN (Local Area Network) by means of the so-called GPFS Network Shared Disks (NSD) (i.e., an abstraction layer used for shipping IO requests to the disk servers). Since GPFS is a real clustered filesystem, some of the servers can be off-line without any interruption of the filesystem availability.

In our implementation, the GPFS NSD layer is decoupled from the underlying TSM services because the software services reside on different servers: hence problems affecting the TSM

layer do not influence the GPFS system. The main components of the TSM layer are the TSM central server that includes the database, and several nodes, identified as the Hierarchical Storage Manager (HSM), that physically execute the data flow between the disk and tape media. The TSM server runs the central software service: it is the reference to the different clients and provides all the needed information to the GEMSS services. The database resides on a separated high availability disk system and a cold-standby server can be easily used in case of failure of the main TSM central Server. The HSM nodes use a complete LAN-free data flow being interconnected to the SAN through Fibre Channel links; this greatly increases the performance and reduces the latency. The GEMSS services use the HSM nodes to provide the required space management in particular with the automatic migration of the unused files from the disk area. The data flow from disk to tape is controlled using periodic GPFS filesystem metadata scans: according to a set of policies specified by the administrators, a list of files to be pre-migrated to tape is created. The GEMSS services pass these lists to the TSM process running on each HSM node and the pre-migration is performed (i.e. the files are copied to tape without deleting them from disk). If the system recognizes that the disk area occupation is over a defined threshold, a migration process is run. A migration process replaces the disk copy of a file with a so-called stub-file, a pointer that can be used for retrieving the tape copy as needed. Similarly, the recalls use optimized list of files, grouped by tape volumes and ordered "as written on tape" to reduce unnecessary access to the tape volume media.

The StoRM service has a multilayer architecture composed by two main stateless components: the frontend and the backend. The frontend exposes the SRM Web service interface, manages user authentication and stores the requested data into the database. The backend is the core of StoRM service: it executes all synchronous and asynchronous SRM functionalities. The entry points for data are the GridFTP servers with WAN (Wide Area Network) access and the NSDs with LAN access directly connected to GPFS where the files are initially stored.

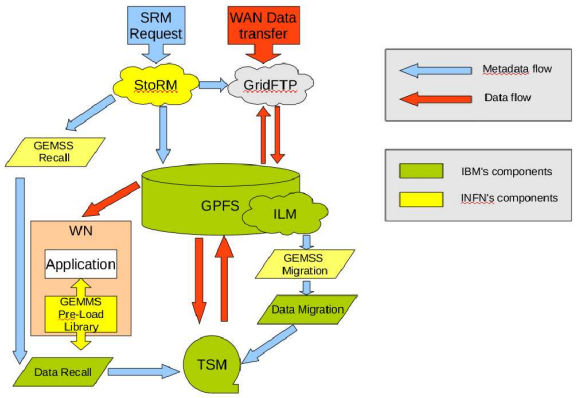

Figure 1. The data and metadata flows in GEMSS.

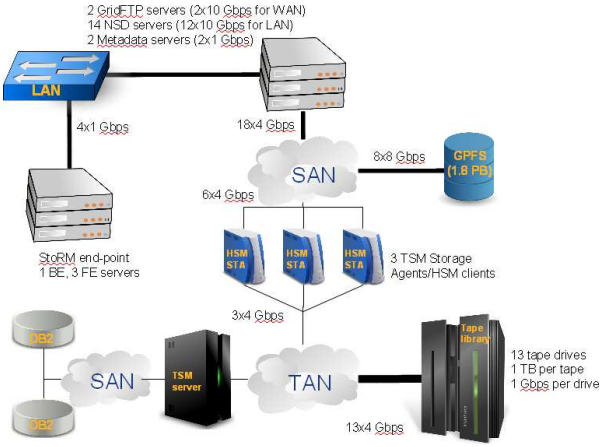

Figure 2. The CMS use case.

Figure 1 shows flows of metadata and data in GEMSS, while Figure 2 details the current implementation of GEMSS used at INFN Tier-1. Figure 2 illustrates the storage resources needed for CMS [10], a typical LHC experiment. Using 10 Gb/s LAN connections, a small number of NSD and GridFTP servers is needed to match the CMS required bandwidth. In addition, only 3 HSM nodes are required for CMS to provide the necessary disk to tape throughput and connection redundancy; each HSM is connected to the SAN dedicated to tape library, called Tape Area Network (TAN), with a 4 Gb/s connection.

3. Results

INFN-CNAF provides the resources and services necessary for all activities of the four LHC experiments, such as data storage and distribution, data processing, Monte Carlo production and data analysis. INFN Tier-1 houses a considerable amount of hardware storage resources and CPU. In 2010 the CPU power was 66 kHEP-SPEC06 [11] (to be increased to 86 kHEP-SPEC06 during 2011), while the disk and tape space were respectively 6 and 8 PB (to be increased to 8 and 10 PB during 2011). INFN-CNAF interact with CERN and others Tier-1s and Tier-2s with very-high-speed connection, 10 Gbit connectivity with Tier-0 and Tier-1s and typically 1 Gbit connectivity with Tier-2s. The overall bandwidth usage is of some Gb/s, increasing with intensive experiments data distribution and analysis up to peaks of 10 Gb/s registered during summer 2010, and well sustained by the network infrastruture. Since the start of LHC, on November 2009, more than four PB of data have been stored overall on disk and tape at INFN-CNAF. The INFN Tier-1 is dedicated mainly to data archiving of raw data from CERN, data reprocessing, distribution of data from and to other Tier-1s and Tier-2s. It also offers services for data access and analysis to researchers all over the world. Throughput from all sites to INFN Tier-1 is shown in Figures 3 for the LHC experiments (here noted as Virtual Organization) since the start of LHC and in the period of maximum activity. In the three figures respectively the average throughput per month, per day and per hour is reported. In evidence it is highlighted the peak of 120 MB/s monthly, 300 MB/s daily and 700 MB/s hourly.

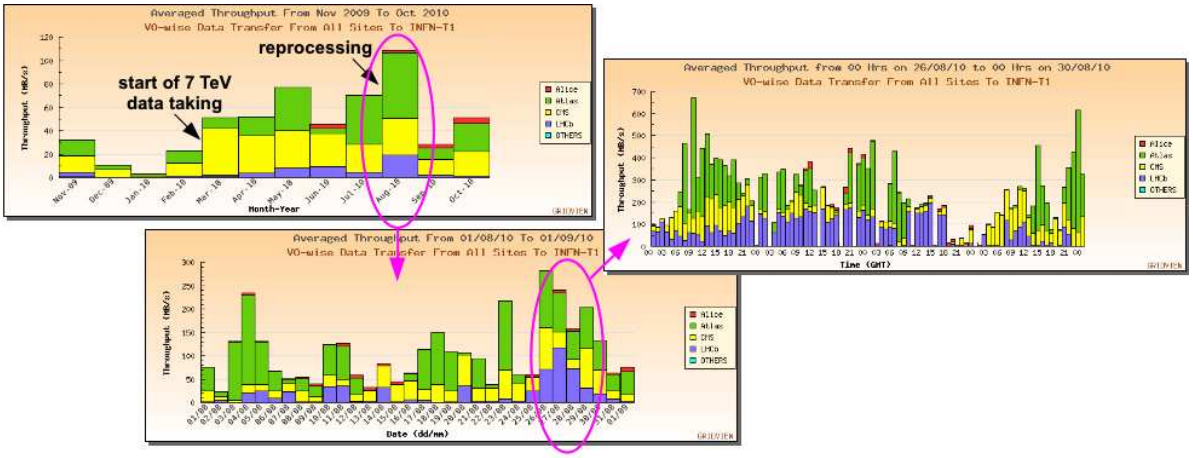

Figure 3. Monthly throughput (up left) from all WLCG sites to INFN Tier-1 per experiment in 2010. Daily (right) and hourly (bottom left) throughput per experiment respectively in August and in the most intensive four days of August.

GEMSS has been shown to very well sustain the throughput, also when more than expected and over the nominal rate.

Figures 4 show the throughput on GPFS and GridFTP servers on which, respectively, the overall storage access and the access on WAN lies. The GEMSS disk-only response is shown for the ATLAS experiment, which is the one that runs most intensive activities on disk from week 34 to week 41 (almost two months). A throughput peak of 900 MB/s on GridFTP server and of more than 3.5 GB/s on the GPFS server point out the high performances of the system.

3.1. ATLAS activities

From the ATLAS dashboard, the monitoring system for transfers between sites, an exceptional throughput of 1.2 GB/s to INFN-CNAF has been registered on August 12th 2010, as shown in

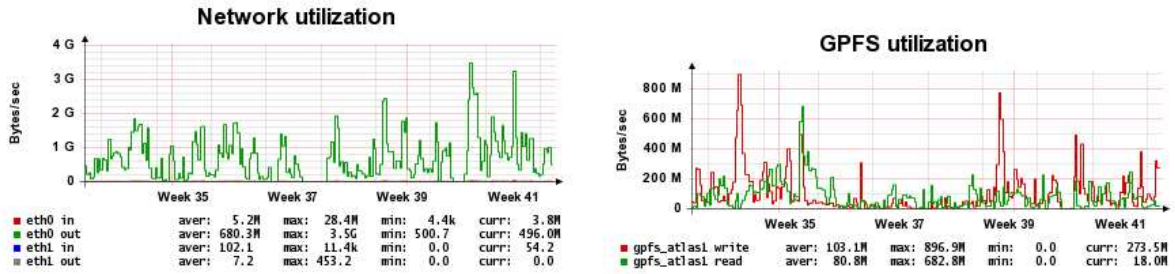

Figure 4. ATLAS jobs access (left) from computing farm and (right) ATLAS storage access on WAN.

Figure 5 on the left. The transfers activity was due to Tier-0 export, data consolidation at sites and user subscription of data to the site. ATLAS saturated the high velocity link between INFN-CNAF and CERN on July 13th and 14th 2010, and the LAN with more than four thousands jobs accessing the storage from the computing farm, an unprecedented users analysis activity, as reported in Figure 5 on the right.

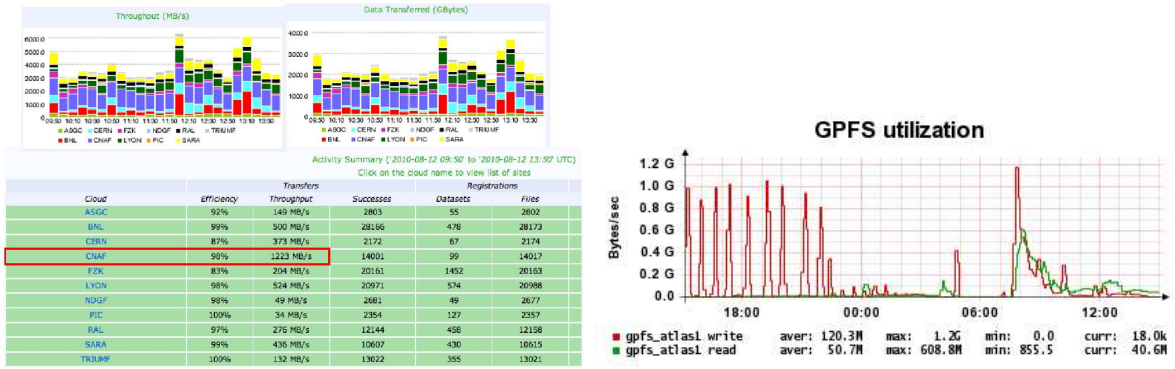

Figure 5. ATLAS throughput of 1.2 GB/s to INFN CNAF (left) registered on August 12th 2010, and (right) ATLAS saturation of the LAN with storage data access of analysis jobs from the computing farm.

3.2. CMS activities

The transfer rate for data volume moved between INFN-CNAF and other sites is reported in Figure 6 as recorded by the CMS monitoring system, Phedex. Peaks of 700 MB/s and 400 MB/s (hourly-averaged) are shown respectively for transfers to and from INFN-CNAF, during an intensive data subscription occurred in early August. Data migration from disk to tape as well as data recalls from tape, provided excellent performances. With 13 tape drives available, an average throughput of 300MB/s both for data migrations to tape and for recalls, with peaks of 700 MB/s and 600 MB/s respectively, has been reached as reported in Figure 7.

4. Conclusions

GEMSS is in production at INFN-CNAF Tier-1 for LHC experiments since the starting of LHC data taking. The figures shown during the first year of production are very promising in terms of performances, scalability and stability. With GEMSS we could fulfill all the requirements of

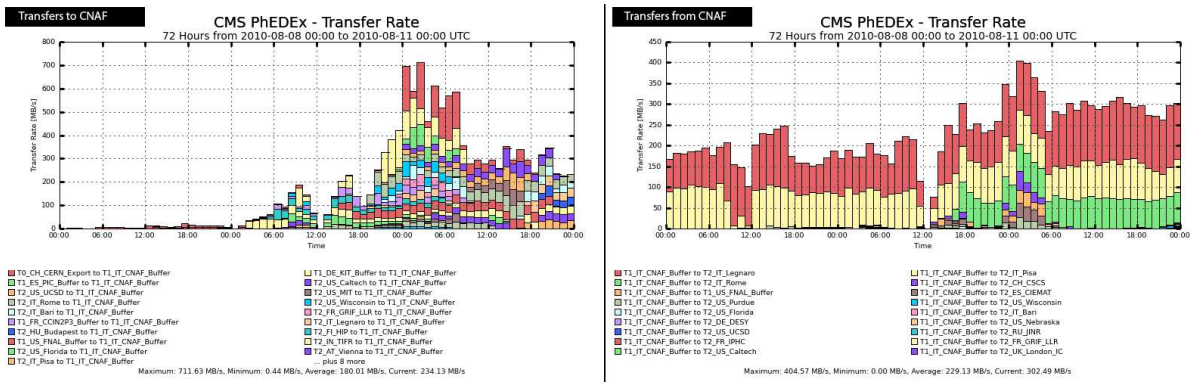

Figure 6. Transfer volume to (left) and from (right) INFN-CNAF and other CMS sites in the early August 2010.

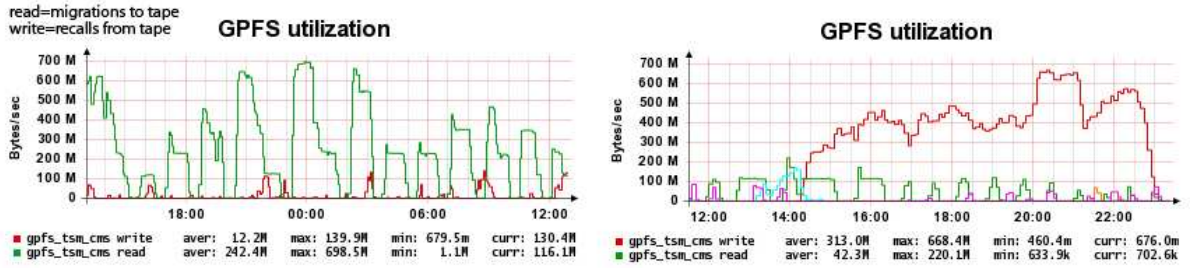

Figure 7. Migration transfer rate from disk to tape (left) and recalls transfer rate from tape (right).

the LHC experiments in terms of access to the data, both on disk and tape. Also the non-LHC experiments have started to move to the GEMSS because of its efficiency abandoning other MSS solutions.

References

[1] M. Bencivenni et al., *INFN-CNAF activity in the TIER-1 and GRID for LHC experiments*, Proceedings of IEEE International on Parallel & Distributed Processing Symposium (IPDPS), Rome, Italy, May 2009
[2] GPFS: `http://www-03.ibm.com/systems/software/gpfs/`
[3] M. Bencivenni et al., *A comparison of Data-Access Platforms for the Computing of Large Hadron Collider Experiments*, IEEE Transactions on Nuclear Science, Volume 55, Issue 3, pp. 1621–1630, ISSN: 0018-9499, 17 June 2008
[4] R. Zappi, E. Ronchieri, A. Forti, and A. Ghiselli, *An efficient Grid data access with StoRM*, S.C. Lin and E. Yen (eds.), Data Driven e-Science: Use Cases and Successful Applications of Distributed Computing Infrastructures (ISGC 2010), Springer Science + Business Media, LLC 2011
[5] A. Carbone et al., *Performance studies of the StoRM Storage Resource Manager*, Proceedings of Third IEEE International Conference on e-Science and Grid Computing (10-13 Dec. 2007), pp. 423-430
[6] A. Cavalli et al., *StoRM-GPFS-TSM: a new approach to Hierarchical Storage Management for the LHC experiments*, J. Phys.: Conf. Ser. 2010, V. 219 N. 7
[7] LHC: `http://lhcb.web.cern.ch/lhcb/`
[8] ATLAS: `http://atlas.web.cern.ch/Atlas/Collaboration/`
[9] TSM: `http://www-01.ibm.com/software/tivoli/products/storage-mgr/`
[10] CMS: `(http://cms.web.cern.ch/cms/index.html/)`
[11] http://w3.hepix.org/benchmarks/doku.php