The INFN-CNAF Tier-1 GEMSS Mass Storage System and database facility activity

Pier Paolo Ricci[1,4], **Alessandro Cavalli**[1], **Luca Dell'Agnello**[1], **Matteo Favaro**[1,] **Daniele Gregori**[1], **Andrea Prosperini**[1], **Michele Pezzi**[1], **Vladimir Sapunenko**[1], **Giovanni Zizzi**[1] **and Vincenzo Vagnoni**[2]

[1] INFN CNAF, viale Berti Pichat 6/2 40127 Bologna, Italy.
[2] INFN Sezione di Bologna, via Irnerio 46, 40126 Bologna, Italy.

E-mail: pierpaolo.ricci@cnaf.infn.it

Abstract. The consolidation of Mass Storage services at the INFN-CNAF Tier1 Storage department that has occurred during the last 5 years, resulted in a reliable, high performance and moderately easy-to-manage facility that provides data access, archive, backup and database services to several different use cases. At present, the GEMSS Mass Storage System, developed and installed at CNAF and based upon an integration between the IBM GPFS parallel filesystem and the Tivoli Storage Manager (TSM) tape management software, is one of the largest hierarchical storage sites in Europe. It provides storage resources for about 12% of LHC data, as well as for data of other non-LHC experiments. Files are accessed using standard SRM Grid services provided by the Storage Resource Manager (StoRM), also developed at CNAF. Data access is also provided by XRootD and HTTP/WebDaV endpoints. Besides these services, an Oracle database facility is in production characterized by an effective level of parallelism, redundancy and availability. This facility is running databases for storing and accessing relational data objects and for providing database services to the currently active use cases. It takes advantage of several Oracle technologies, like Real Application Cluster (RAC), Automatic Storage Manager (ASM) and Enterprise Manager centralized management tools, together with other technologies for performance optimization, ease of management and downtime reduction. The aim of the present paper is to illustrate the state-of-the-art of the INFN-CNAF Tier1 Storage department infrastructures and software services, and to give a brief outlook to forthcoming projects. A description of the administrative, monitoring and problem-tracking tools that play a primary role in managing the whole storage framework is also given.

1. The storage system of the INFN-CNAF Tier-1

In this paper we describe the main activities and the consequent results obtained during the last years concerning mass storage resources at the INFN-CNAF Tier-1. Starting from the end of 2005 the Tier-1 [1] has become the Italian national reference point for the INFN computing activities. In particular, all of the LHC experiments (ALICE, ATLAS, CMS and LHCb) use our site as a primary Tier-1 computing centre. In addition, about 20 non-LHC collaborations (e.g. CDF, Kloe, AMS2, Argo, Auger, Glast, MAGIC, Pamela, Borexino, Darkside, Virgo, etc.) currently use the computing and storage resources of the centre with a guarantee level of service of 24/7 support (24h every day non-stop service availability). The storage resources comprise 15 Petabyte (PB) of net disk space accessible in a Storage Area Network (SAN), mainly composed of RAID-6 disk arrays. Furthermore,

[4] Corresponding Author

tape storage is provided by an Oracle STK SL8500 tape library robot, characterized by a capacity of 17 PB of uncompressed tape space served by several Oracle STK T10K tape drives. The computing farm is composed of ~1000 nodes, which have direct access to 12 PB of disk space shared over 11 IBM General Parallel File System (GPFS) [2] filesystems.

In order to access these storage resources, a very reliable and high performance disk-based data access system is required, as well as an efficient Mass Storage Systems (MSS). The MSS allows the archival of several PB of data per year on tape media, that should remain available over a time scale of several years. For these reason in 2008 we started to work on the integration of our IBM GPFS disk storage infrastructure with the IBM Tivoli Storage Manager (TSM) [3], aiming at realizing a full Hierarchical Storage Management (HSM) system called GEMSS (Grid Enabled Mass Storage System). GEMSS uses StoRM as Grid Storage Resource Manager (SRM) [4], providing the standard interface used in the WLCG framework. It is distributed with standard RPM packages implementing a software layer for the interaction and optimization between GPFS, TSM and StoRM, thus providing a complete solution for tape and disk storage. The entire storage resources at the INFN-CNAF Tier-1 have been managed by GEMSS since the start of its production phase at the end of 2010 [5].

In Figure 1, a schematic representation of the main GEMSS components is reported.

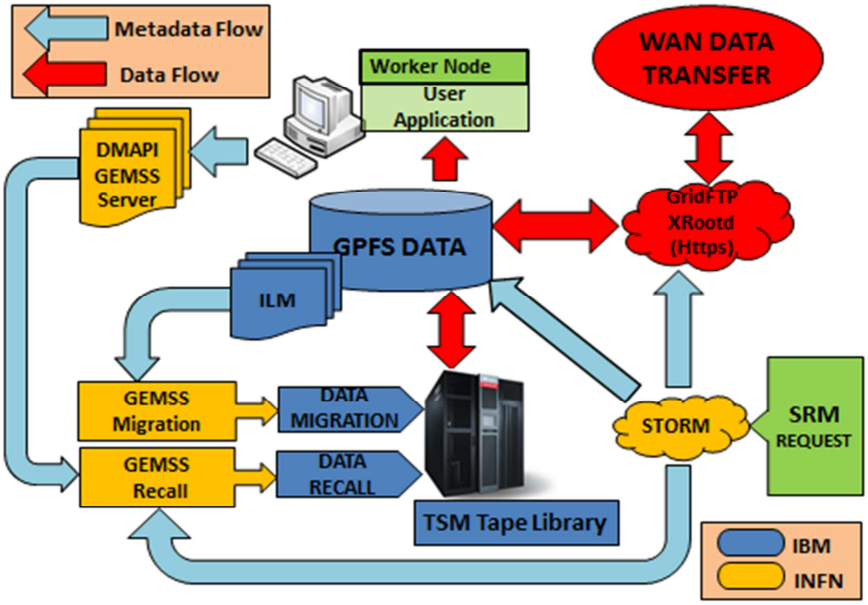

Figure 1. Logical schema of the main building blocks of the GEMSS system.

In the Figure, the five main components of the GEMSS system (some are developed by IBM and some by INFN) and the two main data and metadata flows (the pale blue and red arrows) are shown.

In particular it is possible to recognize the following main elements:

- GPFS DATA cluster and Information Lifecycle Management (ILM) engine which is the disk storage software main GPFS filesystem infrastructure with the utilization of ILM. By using ILM, the filesystem can be partitioned into a number of so-called GPFS "storage pools" implementing file placement policies and data migration rules from one pool to another according to some user-defined criteria.
- TSM (Tivoli Storage Manager) Tape Library. The tape library management system performs the data movement between disk and tape media using the "Data Migration" and "Data Recall" operations.
- StoRM, the SRM service.

- GridFTP/XRootD/(HTTPS) services. The LAN/WAN data transfer services and protocols which are used for transferring the data to and from the storage system. These protocols will be described in the next section.
- GEMSS migration and recall engine for the interaction between TSM and GPFS and for the general metadata management.

In addition, the DMAPI GEMSS server plays a central role when tape-resident data is directly accessed by User Application running on a worker node, without requesting in advance the tape recall via the SRM service. This is to allow the User Application to trigger tape recalls in an optimized and controlled way simply by accessing the files. The POSIX read requests are intercepted by the DMAPI GEMSS server and are submitted to the GEMSS system which groups and sorts the requests in an optimal way [5].

The GEMSS system is therefore the core of the INFN-CNAF Tier-1 storage infrastructure and it has been demonstrating stability, reliability and good performance over the last years of running activity. In the next section we will focus on the access to the MSS and on the specific performance monitoring tools implemented in order to improve the general service supervision.

2. MSS access and monitoring

The main entry point to data storage from the WAN at the INFN-CNAF Tier-1 is StoRM, a Grid middleware service that implements the SRM interface version 2.2 and is designed to support direct file access using native POSIX I/O calls to the storage. Actually StoRM is used by ATLAS, CMS and LHCb, and by other smaller Virtual Organizations (VOs). Data transfers are usually performed using the GridFTP service, that is also used without contacting first the SRM service by some non-LHC experiments using X509 certificates (DarkSide, Kloe, Theoretical physics groups, etc.). Besides the SRM protocol, the LHC experiments have also started considering other methods for accessing data. In particular, two interesting protocols are gaining popularity within the LHC world: XRootD [6] and HTTPS/WebDaV.

XRootD is now adopted by all LHC experiments for disk and tape data access. It is a protocol for data access where a user joined to a specific Federation can move, read and write data inside federated sites. User authentication is done using personal X509 certificates and the actual data transfer is done with checksum (adler32 algorithm) in order to grant the maximum reliability to the transfer operation. However, since tape-resident files are visible as ordinary files in a GPFS filesystem, a user could freely access such files causing massive recalls that would trigger uncontrolled operation and overload of the tape backend. For this reason, the tape recall over XRootD uses a specific ad-hoc “plugin” developed at CNAF to submit recall requests to GEMSS, when tape-resident files are directly accessed via XRootD. The specific “plug-in” is now in the XRootD development repository and it will be available to everyone in the standard XRootD release.

The HTTPS/WebDaV protocol implemented in StoRM has been used by the ATLAS and Xenon experiments. As well known in the literature, the WebDaV (Web Distributed Authoring and Versioning) protocol is an extension of the Hypertext Transfer Protocol (HTTP), and it makes the Web a readable and writable medium. It is a good framework for users to create, change and move files on a server (e.g. a web server or web share) and its main advantage lies in an intuitive and easy-to-use interface. At present, the utilization of this protocol at CNAF is somewhat limited to test activities or very specific use cases. Nevertheless, its diffusion is expected to increase quickly in the near future.

In order to monitor the performance and the correct behaviour of the whole storage system, two main monitoring tools are used in conjunction with other “minor” tool for specific monitoring of hardware storage boxes: Lemon and Nagios [7].

Lemon (LHC Era Monitoring) [8] is a CERN-developed server/client monitoring system available for Linux, where an agent running on the monitored machine gathers information using a set of standard and customized sensors (e.g. network traffic, memory and CPU usage, etc.). Information records are gathered by one central server that provides archiving over an Oracle database and a visual

interface for organizing the information in a set of graphs over a web interface. Specific sensors have been developed at CNAF, and are currently used in production in addition to the default ones. In particular, information on the GPFS performance utilization, the Fibre Channel traffic over SAN interlinks, the tape drive throughput, and the TSM operations (e.g., number of successful or failed migrations or recalls operations) have proven to be very useful for identifying problems or data traffic bottlenecks. In particular the specific Lemon Fibre Channel sensor contacts the SAN switches using SNMP commands and obtains the counters values for the total traffic throughput at a specific timestamp. After a conversion these values are displayed as standard megabyte per second values. Regarding to the TSM operations, some specific Lemon sensors have been developed in order to analyse the latest GEMSS operation logs and summarize the useful information in a more readable form. Since the GEMSS logs include all the operation of the system and are rather verbose, some specific "keywords" are printed by the GEMSS daemons in the log files for operations such as recall file or mount a tape. The Lemon sensors search for these keywords in the last part of the log files and collect the information in standard plots (e.g. the number of successful and failed recall operations with a 5 minutes granularity) and in statistical average values on daily basis.

Nagios [9] is a widespread open source monitoring tool that can be easily adapted to specific local use. At our site, Nagios is used for alerting the site administrators via email or SMS notifications, when something actually goes wrong. Basically, the Nagios monitoring daemon runs intermittent checks on hosts and services using plugins which return status information to the central server. In case of failures, depending on the logics of the configured check, Nagios can also execute corrective actions like trying to restart the service or removing a machine from a cluster, or it can simply notify the failure. Custom Nagios checks have been developed for monitoring the GPFS cluster, TSM clients and server, GridFTP, and the StoRM services. Effectively some of these services (e.g. GridFTP and StoRM front-ends) use a multi-hosts definition list in the DNS alias name for redundancy and load balancing. In case of unrecoverable error on one specific host, the Nagios central server can interact with the local DNS server and remove the failed machine from the alias name of the service, leaving only the working machines in the list.

In addition to the Lemon and Nagios tools, specific monitoring tools have been installed and customized in order to obtain performance information with the finest granularity over the storage system boxes. In particular since the majority of the storage hardware at our site is from DataDirect Network (DDN), specific plugins have been provided by the vendor for accessing the whole set of information provided by their storage systems. A dedicated Cacti [10] customized tool service has been installed and configured at our site for monitoring the DDN hardware. It collects and provides detailed information on bandwidth performance (throughput reported in megabyte per second and rate of input/output operations per second) for each single logical and physical hard disk in the DDN storage boxes. This information provides a valuable help when debugging performance problems and throughput bottlenecks.

The INFN-CNAF Tier-1 storage infrastructure has shown good stability and excellent performance figures during the last years of service, also owing to its "parallel" and "clustered" layout. Throughout several stressful computing activities the Tier-1 storage system responded well, in some cases even exceeding expectations.

In Figure 2 a Lemon plot of the LHCb GPFS filesystem performance during three weeks in fall 2013 is reported. During the period of October-November 2013 the LHCb collaboration launched a massive stripping campaign, recalling from tape to disk at the INFN-CNAF Tier-1 a total of ~700 TB (divided in ~150k files from a total of roughly 300 tapes). As shown in the Figure, the sustained throughput from the tape backend has been between 300 and 450 MB/s (twice as fast as requested by the experiment) and this has reduced the total time for the whole operation to 23 days instead of the initial estimation of two months.

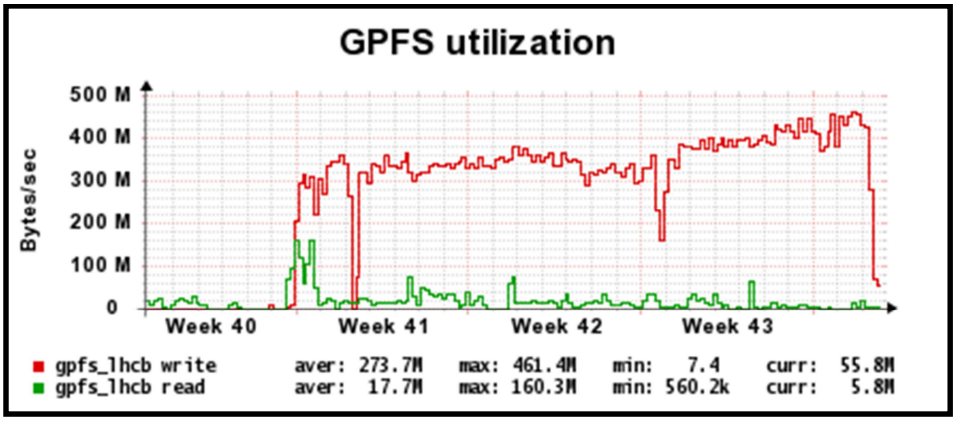

Figure 2. Lemon plot of the recall (tape to disk in red, disk to tape in green) throughput during the LHCb stripping campaign of oct-nov 2013.

3. The INFN-CNAF Tier-1 Oracle database services

The Oracle database service is one of the core services of the INFN-CNAF Tier-1. The Oracle database level of service has evolved during the last five years and now relies on different Oracle Real Application Clusters (RAC) for the maximum reliability and performance.

In Figure 3 a sketch of the infrastructure of the database service is shown. There are different Oracle instances in production and the connection with the mass storage is realized through the Tier-1 SAN. Some of the storage is currently phasing out due to hardware obsolescence and it is used only for testing purposes. A new storage box in production is now starting to replace the old disk volumes. Mixed Oracle 10g and 11g RAC clusters run over physical machines with Red Hat Enterprise Linux rel. 5 and 6 (the Oracle Linux operating system is under test). Oracle 12c is currently installed on a separate stand-alone testbed. The production storage hosts both the data files and the data of the Oracle Application and E-Business Suite related services. The services are monitored by a 10g clustered Oracle Grid Console, based on a RAC database for the repository and an active-passive console for total redundancy.

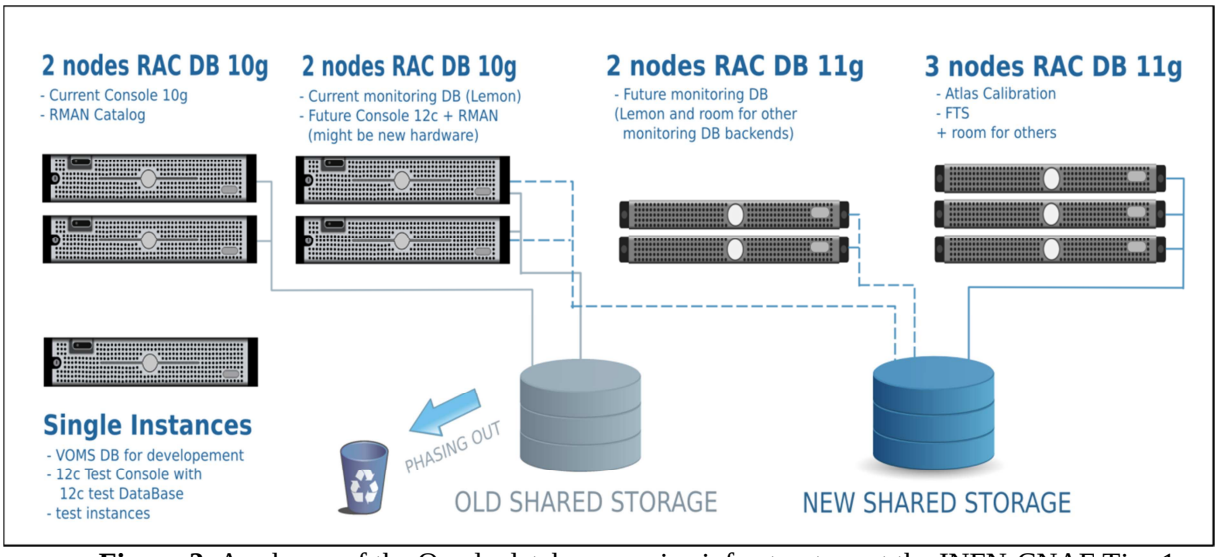

Figure 3. A schema of the Oracle database service infrastructure at the INFN-CNAF Tier-1.

At present the production services are operated on 2 RACs and a single instance:

- An Oracle 11g RAC composed of 3 nodes which runs:
    - the ATLAS Calibration database, in production within 2014;
    - the Grid File Transfer Service (FTS) database, available for WLCG and other users;
    - Oracle RMAN catalog for E-Business Suite related databases.
- An Oracle 10g RAC composed by 2 nodes which runs:
    - database for Lemon monitoring with history available for query and plot.
- A single Oracle 10g database dedicated to CNAF software developers.

Future plans and developments for the Oracle database service at CNAF include the installation of 2 nodes in a new Oracle 11g RAC cluster in order to migrate the current Lemon monitoring database and for other monitoring tool backends. Furthermore, the new Oracle 12c Cloud Control with 12c database is under test with the database service and the E-Business Suite for replacing the current 10g version. This will bring great advantages to the local monitoring and management of the RAC database, as it imples the "jump" of two major releases. In addition there are plans for a replica of the CDF FNAL database for the Long Term Data Preservation (LTDP) project [11].

All databases in production are backed-up using Oracle RMAN on separate disks for the "nearline" disk-backup, and the IBM Tivoli Data Protection (TDP) for archiving the backup over tape storage, in order to achieve the maximum data protection in case of disasters.

4. Conclusions

This paper is a site report of the INFN-CNAF Tier-1 concerning storage activities and services. Since the foundation of the Tier-1 centre, storage services have evolved in flexibility and stability with the introduction of the GEMSS storage system, that fulfilled the performance requests to the site, in addition to other clustered services like the Oracle Real Application Cluster (RAC), that helped increasing the database level of service. Furthermore, the use of monitoring tools, like Nagios, allowed tracking and automatically correcting failures of services or server hardware. The introduction of additional monitoring software packages (e.g., hardware specific ones) provided additional information that can be collected over the mass storage and that turned out to be very useful when trying to diagnose specific problems. The introduction and the wide distribution of additional methods for accessing data from the WAN (e.g. XRootD and HTTPS/WebDaV) enables the storage system to respond to the performance and flexibility requirements during the years to come. The disk and tape storage capacity in production at our Tier-1 are expected to increase quickly in the near future, and we are confident that the whole system will continue to exhibit great performance figures and reliability.

References

[1] G. Bortolotti et al. *The INFN Tier-1* 2012 J. Phys. Conf. Ser. 396 042016 Proceedings of 2012 CHEP conference.

[2] *IBM General Parallel File System Administration and Programming Reference* Version 3 Release 2 SA23-2221-01.

[3] IBM website references for Tivoli Storage Manager info and documentation: https://www.ibm.com/developerworks/wikis/display/tivolidoccentral/Tivoli+Storage+Manager and http://www-03.ibm.com/software/tivoli/products/storage-mgr/.

[4] Info about StoRM available online http://www.italiangrid.it/middleware/storm/.

[5] D. Bonacorsi et al., *The Grid Enabled Mass Storage System (GEMSS): the Storage and Data management system used at the INFN Tier1 at CNAF*, 2012 J. Phys. Conf. Ser. 396 042051 Proceedings of 2012 CHEP conference.

[6] D. Gregori et al., *Xrootd data access for LHC experiments at the INFN-CNAF Tier-1* 2013 J. Phys. Conf. Ser. 513 (2014) 042023 Proceedings of 2013 CHEP conference.

[7] S. Antonelli et al., *INFN-CNAF Monitor and Control System* 2010 J. Phys. Conf. Ser. 331 (2011) 042032 Proceedings of 2010 CHEP conference.
[8] Info about Lemon available online http://lemon.web.cern.ch/lemon/index.shtml
[9] Info about Nagios available online http://www.nagios.org/.
[10] Info about Cacti available online http://www.cacti.net/.
[11] S. Amerio et al., *Long Term Data Preservation for CDF at INFN-CNAF* 2013 J. Phys. Conf. Ser. 513 (2014) 042011 Proceedings of 2013 CHEP conference.