PROCEEDINGS OF SCIENCE

Mass Storage System for Disk and Tape resources at the Tier1.

Pier Paolo Ricci [1]
INFN CNAF
Viale Berti Pichat 6/2, 40127 Bologna ITALY
E-mail: pierpaolo.ricci@cnaf.infn.it

Luca dell'Agnello, Alessandro Cavalli, Daniele Gregori, Barbara Martelli, Andrea Prosperini, Elisabetta Ronchieri, Vladimir Sapunenko, Vincenzo Vagnoni[2], Dejan Vitlacil
INFN CNAF
Viale Berti Pichat 6/2,40127 Bologna ITALY
E-mail:; luca.dellagnello@cnaf.infn.it; alessandro.cavalli@cnaf.infn.it; daniele.gregori@cnaf.infn.it; barbara.martelli@cnaf.infn.it; andrea.prosperini@cnaf.infn.it; elisabetta.ronchieri@cnaf.infn.it vladimir.sapunenko@cnaf.infn.it; vincenzo.vagnoni@bo.infn.it; dejan.vitlacil@cnaf.infn.it

Abstract: The activities in the last 5 years for the storage access at the INFN CNAF Tier1 can be enlisted under two different solutions efficiently used in production: the CASTOR software, developed by CERN, for Hierarchical Storage Manager (HSM), and the General Parallel File System (GPFS), by IBM, for the disk resource management. In addition, since last year, a promising alternative solution for the HSM, using Tivoli Storage Manager (TSM) and GPFS, has been under intensive test. This paper reports the description of the current hardware and software installation with an outlook on the last GPFS and TSM tests results.

XII Advanced Computing and Analysis Techniques in Physics Research
Erice, Italy
3-7 November, 2008

[1] Speaker
[2] INFN Sezione di Bologna, v. Irnerio 46, 40126 Bologna, ITALY.

PoS(ACAT08)030

1.Introduction

The growing resources requests from the LHC experiments require bigger quantities of data storage and increasing perfomance demands [1]. Moreover a higher level of resources and services stability is required. In the framework of LCG, all the Tier1 sites, will have to provide three different access types to the storage systems [2], the so-called Storage Classes (SC):

Disk0-Tape1 (D0T1). In this SC the data are saved on tape and the disk copy is considered only as a temporary buffer (usually referred as disk cache or staging area) automatically managed by the system. Usually data are automatically deleted from the staging area when the occupancy is higher than a configurable threshold and the system is somewhat sure that data have already been copied to tape.

Disk1-Tape1 (D1T1). The data are permanently saved both on tape and on disk. For this SC, the sizes of disk space and tape space are, by definition, identical; the management of the disk space is delegated to the Virtual Organization (VO) itself.

Disk1-Tape0 (D1T0). In this case there is no guaranteed copy on tape and the management of the disk space, as in the case of D1T1, is responsibility of the VO owning the data.

In the next chapter we will briefly describe the hardware resources we currently use in production as disk and tape storage in our INFN CNAF Tier1. In successive chapter a description of our Castor installation is provided with a small introduction to the Oracle databases services we run. Castor is used at our Tier1 as the primary D0T1 storage class service implementation. Afterwards we report the status of our GPFS pure disk storage pool management as our D1T0 service class and an introduction of the Tivoli Storage Manager installation. A particular attention is turned to the possibility of using this software as a tape extension of the GPFS storage disk pool. In the last chapter the preliminary and promising results from an activity of tests in collaboration with the LHCb group of Bologna are reported which show the feasibility of the implementation of the D1T1 service class.

2. Disk and Tape Storage resources

In our model all our storage resources are accessed using Storage Area Network (SAN) [3] switches and Linux Machines used as diskservers with redundant HBA (host bus adapter that is fibre channel card) for the data access operations. We have had in production this model for many years and it proved a stable and very flexible implementation.

At the moment we have 1250 TB (TeraByte) of RAW disk storage online composed by the following hardware storage controllers or “storage boxes” [4]:

N. 9 Infortrends A16F-R1211-M2	**TOTAL: 56 TB**
N. 2 SUN STK Bladestore	**TOTAL: 80 TB**
N. 4 IBM FastT900 (DS 4500)	**TOTAL: 160 TB**
N. 5 SUN STK Flexline FLX680	**TOTAL: 290 TB**
N. 3 DELL EMC CX3-80	**TOTAL: 680 TB**

The TeraByte reported should be considered as Raw space: the theoretical number calculated by multiplication of the single Hard Disk capacity and the number of these disks in the storage boxes. Using our standard linux filesystems (ext3, xfs or GPFS as described afterwards) over RAID-5 show that the net space (data space available to the users) is about a 15-25% less, strongly depending on the hardware controller of the storage box. An installation of 8 additional EMC CX3-80 as the result of our 2007 tender is currently under way and this will increase our disk resources to a total of nearly 2.5 PB.

All the storage boxes have their particular monitoring system for managing the storage RAID status and providing alerts in case of failures. In case of some vendors an automatic call-home and remote supervision system has been implemented which helps in problem resolution and drastically lower the response time from the support technicians. About the technology of the disks most of them are SATA [5] due to the good low-cost/good-performance compromise but for specific applications (like the Oracle database cluster described in the next chapter) which requires a higher level of reliability and higher performance in the random I/O access, native Fibre Channel disks has been provided.

The choice of using Fibre Channel and SAN technology for all the disk hardware at our Tier1 gives some good advantages:

• diskservers could implement a No Single Point of Failure (NSPF[1]) system where each component of the storage system is redundant (storage array controllers, SAN switches, and server HBA). If software supports it, a cluster approach is possible (as in the case of GPFS clusters).

• The SAN gives the best flexibility: we can dynamically assign new volumes or disk storage boxes to diskservers without the need of stopping the production. This is possible since the SAN is a “dynamic network” where targets (disk storage boxes) and initiator hosts (diskservers) could be added without interrupting data flow [6].

• It is possible to use opportune tools (most fibre channel switches includes vendor-specific tool) for monitoring the SAN. This could really help to monitor I/O bandwidth on devices and recognize bottleneck.

• LAN free systems for archiving and backup purpose to the tape facilities is possible (clearly the tape drive must be connected to the SAN too).

Our SAN infrastructure is based on Brocade switches supplied by the tenders for the disk storage. Two Fabric Director Switches (one SilkWorm 24000 with 128 2 Gbit/s ports and one 48000 with 224 4 Gbit/s ports) represent the core of the SAN, while two SilkWorm 3900 (total of 64 ports) are connected as peripheral switches. The access to the storage is provided by dedicated server machines (diskservers) with Scientific Linux as Operating System. Currently we have in production a total of ~90 diskservers with redundant Qlogic HBA connections to the SAN and Gigabit connection to the LAN which provide the front-ends to the Farm worker node (currently a thousand nodes with a computing capability of 9000 KSPI2k[2]).

[1] A "single point of failure" is any single piece of equipment that, if it fails, can bring your entire operation to a halt. A "No Single Point of Failure" system is a hardware layout where the operations don't stop if a single piece of equipment (i.e. hardware) fails.

[2] see www.spec.org for the related benchmark tool and KSPI2k definition.

About the tape resources we run in production 2 tape libraries:

- A SUN L5500 silo partitioned with 2000 tapes cartridges for the 6 LTO-2 drives and 3500 tape for the 10 9940B drives. The total capacity of the L5500 is about 1 PB of uncompressed data.
- A SUN SL8500 with 8 redundant robot changer with 8 T1000A drives in production (500GB/tape capacity and 110 MB/s bandwidth) and an actual capacity of 2 PB. An upgrade to 10000 slots and 20 T1000B Drives (with 1TB/tape capacity) is currently under installation and test phase and it will provide the library with a total 10 PB capacity.

The total tape space capacity reported is usually the raw space since the data archived in our tape facility from the LHC and HEP experiments are usually already compressed in archives or they are essentially binaries with a low possibility to be substantially reduced.
In Fig. 1 the whole hardware connections outline is reported.

3. Database and Castor services

3.1 Database services

The main goal of the database service is to provide high availability, scalability and reliability inside the Oracle [7] service. This could be fulfilled through a modular architecture based on the following blocks:

- Oracle Automatic Storage Management (ASM) [8] volume manager, for the storage management implementation of redundancy and striping in an Oracle oriented way.
- Oracle Real Application Cluster (RAC) [9] where the database is shared across several nodes with failover and load balancing capabilities.
- Oracle Streams for geographical data redundancy.

In addition to these three software layers the hardware has been particularly selected using dual power redundant server with RAID-1 (mirroring) system disks. These database servers use a dual path fibre channel layer to the main storage which resides on an EMC3-80 with 20TB RAW of fibre channel disks (due to the better performance and Mean Time Between Failure compared to the other technologies).
The RAC clusters are designed in order to grant service in case of hardware/software failure or patches upgrade on single machines where the primary instance (database) run. At present 5 instances are used for the Castor services and additional instances are used for specific service backend such as the LCG File Catalog (LFC) [10] for ATLAS and LHCb, the Lemon Monitoring system database [11], and the SRM (Storage Resource Manager) [12] catalog.
The Oracle Streams are implemented for geographical redundancy of the LHCb condition database which resides at CERN Tier0 and is replicated to the Tier1s.
The total number of Oracle servers is 32, 24 of them are configured in 12 clusters under the Oracle Real Application Cluster and the remaining 8 are configures as single-server instances. We currently run a number of 30 database instances online. Measuring the number of hours when one instance was offline (scheduled down were not counted) during the last year is possible to compute the reliability rate of 98,7% for the 2007, which is a very good result. This

is due the intrinsic reliability which is archived combining the native Oracle failover tool (RAC) with redundant software layers like ASM and likewise redundant dedicated hardware.

3.2 Castor services

CASTOR [13] has been our choice for the D0T1 Service Class implementation for the last years and provides a good solution with somewhat high requirements in terms of man power for optimizing the system and for standard administration. In CASTOR, the data is copied from user resources to the CASTOR front-end (diskservers) and then subsequently copied to back-end (tapeservers). Staging the data on the disk space assigned through the diskservers is one of the most important phases in CASTOR, as in every Hierarchical Storage Manager Software. Actually the operation of accessing data that is already in the staging area does not require to trigger recalls from the tape facilities, with obvious performance benefits. Currently at CNAF we run 40 CASTOR diskservers connected to a SAN at full redundancy fibre channel connections and each of them has five or six small XFS filesystems of 2 TB size each. A big number of diskservers with no more than 10 TB disk space each is required in our CASTOR implementation to avoid congestion on single machines in case of heavy accessed data and to provide a better load balancing. For accessing the tape drive described in Chapter 1 we use dedicated tapeservers which are connected using direct fibre channel connections. Currently we run CASTOR version 2.1.7-17 with all core services on machines with SCSI disks, hardware RAID1, redundant power supplies and Scientific Linux 4 as Operating System while tapeservers and diskservers have lower level hardware. Since CASTOR core services implement stateless components the code is interfaced with Oracle and the service needs a number of dedicated Oracle instances on Real Application Cluster as described in the previous paragraph.

Our CASTOR installation is used from four LHC experiments (ALICE, CMS, ATLAS, LHCb) and from eight others (LVD, ARGO, VIRGO, AMS, PAMELA, MAGIC, BABAR, CDF), with a total capacity of roughly 300 TB of disk staging area and 1 PB of tape space. CASTOR uses LSF scheduler [14] to determine the best candidate resource (diskserver and the relative filesystems) for a CASTOR job which is substantially an I/O operation; using the “LSF slots” parameter is possible to define the maximum number of concurrent accesses for one specific diskserver thus avoiding the bottleneck problem. Anyway some diskservers are used both for file transfers and for the reconstruction and analysis activity in the farm. In these specific cases a limit in the maximum “LSF slots” is not very useful since the farm job submission system could be limited by the number of corresponding slots on the designated diskserver. Therefore it is clear that tuning of the diskserver distribution and of the CASTOR LSF system is crucial for the overall efficiency of the system. The frontends to the grid users are provided using DNS-balanced SRM v.2 endpoints (currently SRM v.1 is still provided but phasing out) whereas monitoring over the whole system is carried out using Lemon software. The Lemon software stores data on one Oracle dedicated instance and it is the CERN suggested monitoring tool with a strong integration with CASTOR. It could provide data about aggregation throughput between the different parts of the system as well as status of the core services daemons in CASTOR.

4. GPFS and TSM services

GPFS [15] is a worldwide distributed software development by IBM which provides a General Parallel File System implementation. The General Parallel Filesystem gives the possibility to have a single file hierarchy (or directory) seen by a set of clients, by aggregating the resources. These resources are primarily disks where the data files are striped, but also computing power and network bandwidth are included. So far, the clients have multiple paths to the data providing a redundant and load-balanced system and in the same way this implementation reduces every potential bottleneck as well as increasing the total bandwidth. The clients can also simultaneously access the same files in a concurrent way since the global coherence is somewhat granted by the parallel filesystem cluster itself.

The idea of our implementation of GPFS is to provide a fast and reliable (with No Single Point of Failure) parallel filesystem with direct access (posix file protocol) from the Farm worker nodes (the so called "clients") using block level I/O interface over standard Ethernet network. In such an implementation the clients don't need to have direct access to the SAN, they instead contacts the GPFS Network Shared Disk (NSD) diskservers using the LAN and the diskservers provides all the I/O over the SAN and the "storage boxes" layers.

Since GPFS is a cluster, with an opportune SAN hardware a true full NSPF is possible (diskservers failures just decrease the theorical bandwidth but the filesystem is still avaliable to the clients) and a single "big filesystem" for each VO could be possible, which is strongly preferred by users. Previous tests [16] also showed us that the usage of parallel I/O drastically increase and optimize the disk performances compared to other system (like CASTOR disk pools). We had run for more that 3 years GPFS at our Tier1 and in the current implementation all the Farm worker nodes (roughly 1000 nodes which act as clients) access the GPFS filesystem using LAN and NSD configuration. A minimum of 8 diskservers are assigned to a single storage box so the filesystem relative to that box will be online as long as at least 1 out of 8 servers is up. We currently run 27 GPFS file systems in production at CNAF (~ 720 net TB) mounted on all farm worker nodes. The client have also WAN grid access to the Storage Class over GPFS using the SRM v.2 compliant tool INFN StoRM [17], which implements the SRM interface over parallel file system [18].

Furthermore in GPFS v.3.2 the concept of "external storage pool" was introduced. The "external storage pool" extends the use of a policy driven migration/recall system to a tape storage backend such as TSM or other software. This is accomplished by means of the so-called Information Lifecycle Management (ILM) engine, which is able to interpret SQL-like policy scripts written by the system administrators that are executed by GPFS to perform specific data management operations on the filesystems. In such a way specific scripts has been written for migrating data from the GPFS filesystem to the tape backend in a sort of D1T1 Storage Class prototype. The "natural" choice for managing tape storage extension for GPFS is IBM Tivoli Storage Manager [19]. In such a prototype GPFS policy engine automatically builds candidate

lists and passes them to the scripts while Tivoli Storage Manager (TSM) actually moves the data to the tape library.

Our experience with TSM resides on an agreement with IBM to use the software until ready for full production in integration with GPFS, with a strong collaboration with the development team for the migration/recall optimization features which are the most crucial and open matters. In fact we actually run server Version 5.5 while a beta version 6.1.0 client is installed for testing the improved migration algorithms. Also with the usage of the TSM "storage agent" modules a future LAN-free migration/recall approach will be possible. In this approach a drive should be connected to a dedicated SAN portion (Tape Area Network or TAN) and all data transfer from disk to tape and vice versa will occur on fibre channel. Other important features of TSM are that the software could also be easily used as a standard backup system for backup operation and the fact that TSM uses an internal database for storing filesytem metadata that could be easily duplicated for a redundant system.

5. GPFS/TSM prototype test layout and results.

The schema of the hardware involved in the prototype of the TSM tape extension of the GPFS disk pool is reported in Fig. 2. The test layout therefore consists of the following elements:

- A 40TB GPFS File system (v.3.2.0-3) served by 4 I/O GPFS NSD diskservers. A redundant 4 Gbit/s Fibre Channel interconnection between servers and disks array is provided using the local SAN. Filesystem backend storage is provided by an EMC CX3-80 storage array. The diskservers are Dell 1950 equipped with Xeon 1.66 GHz dual-core bi-processors with 4 MB L2 Cache, 4 GB of RAM, and 1 Gigabit Ethernet link. The Operating System (OS) installed was the 64 bits Scientific Linux CERN (SLC) version 4.4.
- One dedicated machine runs the TSM server software v.5.5
- 2 servers (same hardware as the diskservers) are used as TSM front-ends, each one acting as GPFS client (reads and writes the filesystem using the LAN) and acting also as TSM client (reads and writes from/to tapes using Fibre Channel connection)
- 3 LTO-2 tape drives from the SUN L5500 tape library dedicated to the testbed

The preliminary tests consist in a data transfer of LHCb files from CERN to the CNAF GPFS/TSM prototype using the File Transfer Service (FTS) [20] and SRM. An automatic migration of the data files from the GPFS disk pool to TSM while the data was being transferred by FTS was implemented using the scripts described in Chapter 4. The data files consist of 8 TB of LHCb stripped DST (realistic production data) in which the file size is mostly of 4 and 2 GB with a bit of other sizes in addition. The number of file used in this data transfer was roughly 2.500. Fig. 3 shows the results of the transfer phase. The black curve shows the net data throughput from CERN to CNAF whereas the red curve shows the net data throughput from GPFS to the TSM tape backend (the migration operation). The drop at the time of about

30.000 s was a temporary interruption on the FTS. The graph shows that the 8 TB data sets were transferred from CERN to tape in 150.000 s (almost 2 days) and the sustained migration rate was about 50 MB/s with two LTO-2 drives and 65 MB/s with three LTO-2 drives. At the end of the transfer test no tape migration error was detected which means that every file written to the GPFS disk pool was copied to the TSM tape backend without any problem or "retry" operation. In Fig. 4 the distribution of file entries versus retention time is plotted. The retention time is defined as the time since a specific file is written until it is migrated to tape (in other words it's the time since file only resides on the GPFS disk pool without a corresponding copy on the TSM tape backend). The distribution clearly shows that most of the files were migrated within less than 3 hours with a tail up to 8 hours. The tail distribution is due to the fact that for a specific period the CERN-to-CNAF throughput raised to 80 MB/s, overcoming the max performance of tape migration at that time. It is very important to show that in this case of "oversize" FTS transfer throughput, GPFS/TSM is able to accumulate a queue of files and migrate them later.

After these preliminary results a successive short phase of pre-production was run. Nearly 40 TB of D1T1 LHCb production data were successfully stored with an average 70 MB/s sustained throughput. A really good result of zero tape migration failures was achieved proving the high reliability of the TSM backend and of the related GPFS migration system. This was indeed a really promising start.

In addition a test of a complete deletion of directories of the GPFS filesystem and successive full recovery from the TSM tape system has been made (using the TSM metadata database); it was possible to rebuild the original filesystem.

6. Conclusion and future activity

This paper is a site report from the INFN CNAF Tier1 Storage Group activities focusing primary on the hardware system and on the Database, Castor, and GPFS activities at our site. It also briefly summarizes the promising implementation and the first results of the new GPFS/TSM prototype. The GPFS/TSM prototype with the SRM StoRM interface proves itself as a good and reliable D1T1 system and LHCb is still using this system for production data. The next step will be the implementation of the D0T1 storage class in the GPFS/TSM system in close collaboration with the IBM development team. Since in D0T1 systems there is no guarantee that the file resides also on disk the GPFS/TSM has a high probability to trigger a tape access when a user requests a file. Therefore recall operations become crucial and a good optimization in accessing data stored on tapes becomes of primary importance. Also the LAN-Free access to the tape facilities should be carefully tested. In a LAN-Free system the data transfer between the GPFS and TSM layers use the SAN/TAN infrastructure instead of the LAN Ethernet network. This could translate in an improvement on the overall performance and a substantial decrease in the LAN data traffic.

References

[1] I. Bird et al. "*LHC computing Grid Technical design report"*, CERNLHCC-2005-024.

[2] A Carbone et. al. *"A Novel Approach for Mass Storage Data Custodial",* proceeding of the 2008 Nuclear Science Symposium, Medical Imaging Conference and 16th Room Temperature Semiconductor Detector Workshop 19 - 25 October 2008 Dresden, Germany.

[3] W. Curtis Preston *"Using SANS and NAS",* O'reilly press

[4] For further details about vendors and related hardware consult the following websites:
http://www.infortrend.com/
http://www.sun.com/
http://www-03.ibm.com/systems/storage/disk/ds4000/ds4500/index.html
http://www.emc.com/products/detail/hardware/clariion-cx3-model-80.htm

[5] The Serial ATA (SATA) computer bus is a storage-interface for connecting host bus adapters (most commonly integrated into laptop computers and desktop motherboards) to mass storage devices (such as hard disk drives and optical drives). See also http://www.ata-atapi.com/sata.html

[6] Tom Clark *"Designing Storage Area Networks 2nd edition",* Addison Wesley Press

[7] *"Oracle Database Administrator's Guide10g Release 2 (10.2)"*
http://download.oracle.com/docs/cd/B19306_01/server.102/b14231/toc.htm

[8] *"Oracle Database 10g Automatic Storage Management Overview and Technical Best Practise" http://www.oracle.com/technology/products/database/asm/pdf/asm_10gr2_bestpractices%2009-07.pdf*

[9] *"Oracle Database Oracle Clusterware and Oracle Real Application Clusters Installation Guide 10g Release 2 (10.2) for Linux", Oracle Press*

[10] *"LCG File Catalog (LFC) administrators' guide"*
https://twiki.cern.ch/twiki/bin/view/LCG/LfcAdminGuide

[11] Lemon is a server/client based monitoring system, more details could be found in the following website:
http://lemon.web.cern.ch/lemon/docs.shtml

[12] F. Donno et al. *"Storage Element Model for SRm 2.2 and GLUE schema description*", CERN

[13] For more documentation about the CASTOR software and RFIO protocol:
http://castor.web.cern.ch/castor/

[14] Load Sharing Facility (LSF) is a commercial computer software job scheduler developed by Platform Computing.
http://www.platform.com/Products/platform-lsf/

[15] General Parallel File System documentation. Available online:
http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/topic/com.ibm.cluster.gpfs.doc/gpfsbooks.html.

[16] Ricci P.P. et. al. "*Experience with Fabric Storage Area Network and HSM Software at the Tier1 INFN CNAF",* proceeding of XI International Workshop on Advanced Computing and Analysis Techniques in Physics Research ACAT 2007 http://pos.sissa.it

[17] http://storm.forge.cnaf.infn.it/

[18] 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.

[19] *"IBM Tivoli Storage Management Concepts",* IBM Redbooks Series, SG24-4877.

[20] FTS, File Transfer Service, developed as part of the gLite middleware stack *"gLite 3.1 User Guide" par. 7.6,* https://edms.cern.ch/file/722398//gLite-3-UserGuide.html

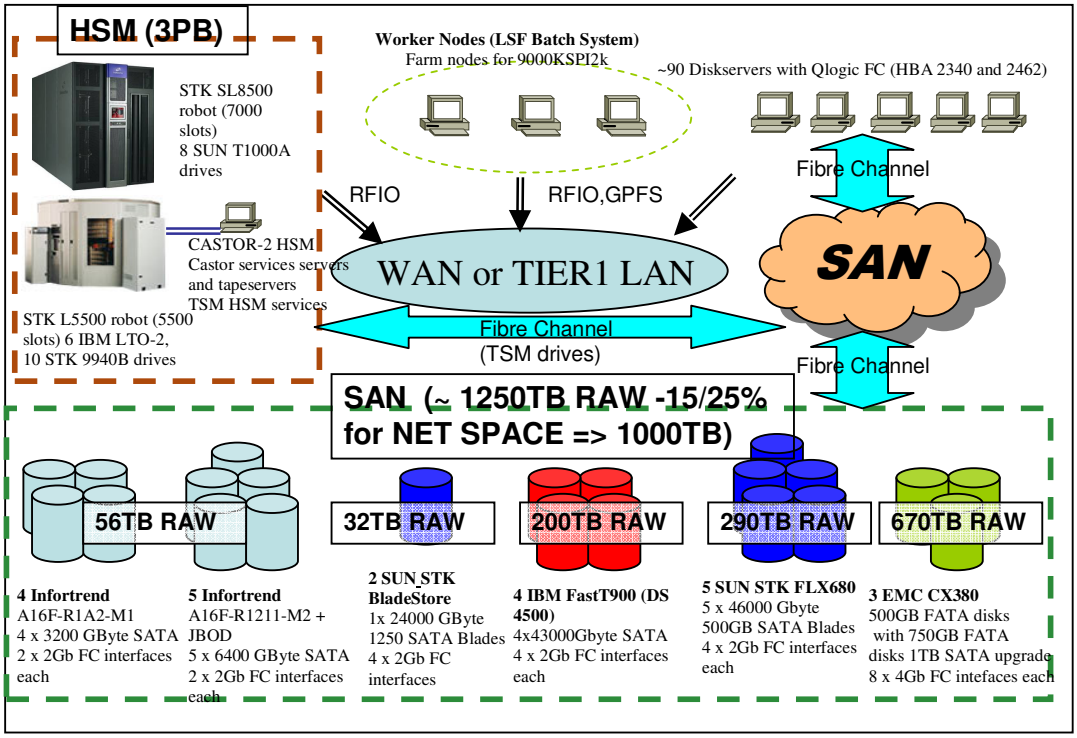

Fig.1: Schema of the INFN CNAF Tier1 Storage connections

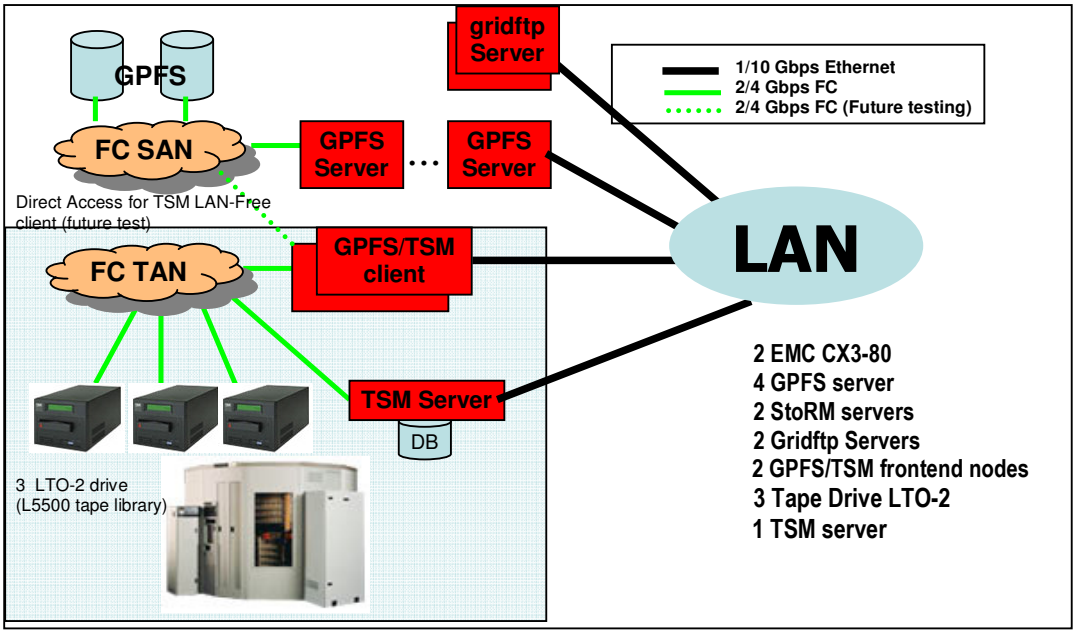

Fig.2: Schema of the TSM prototype connections.

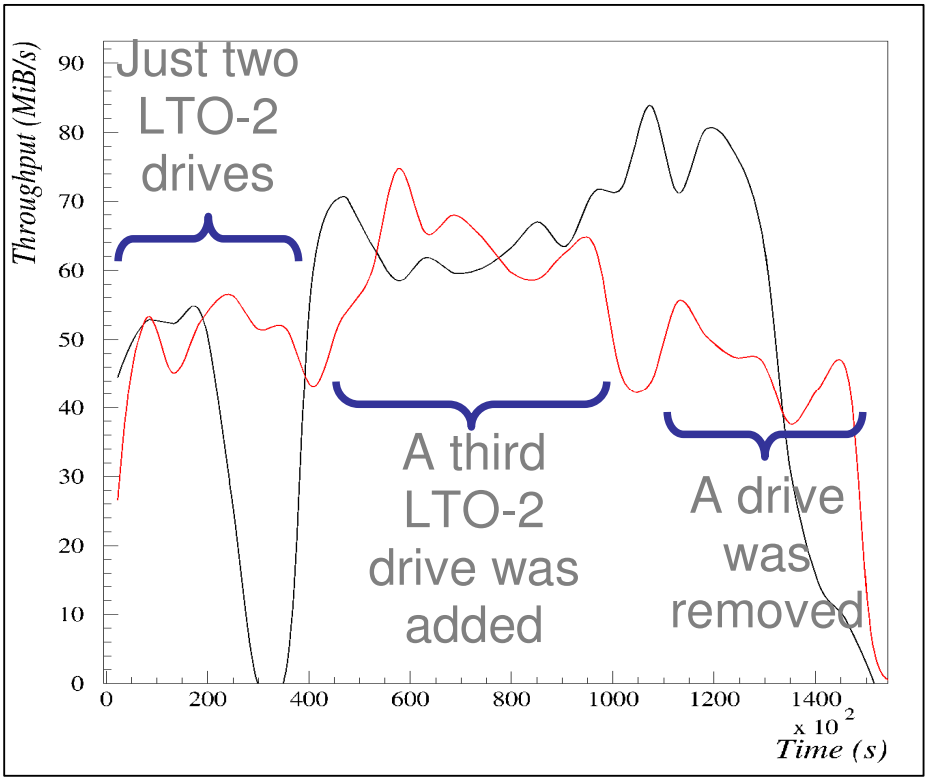

Fig.3: Result of the transfer phase from CERN to the local GPFS/TSM system

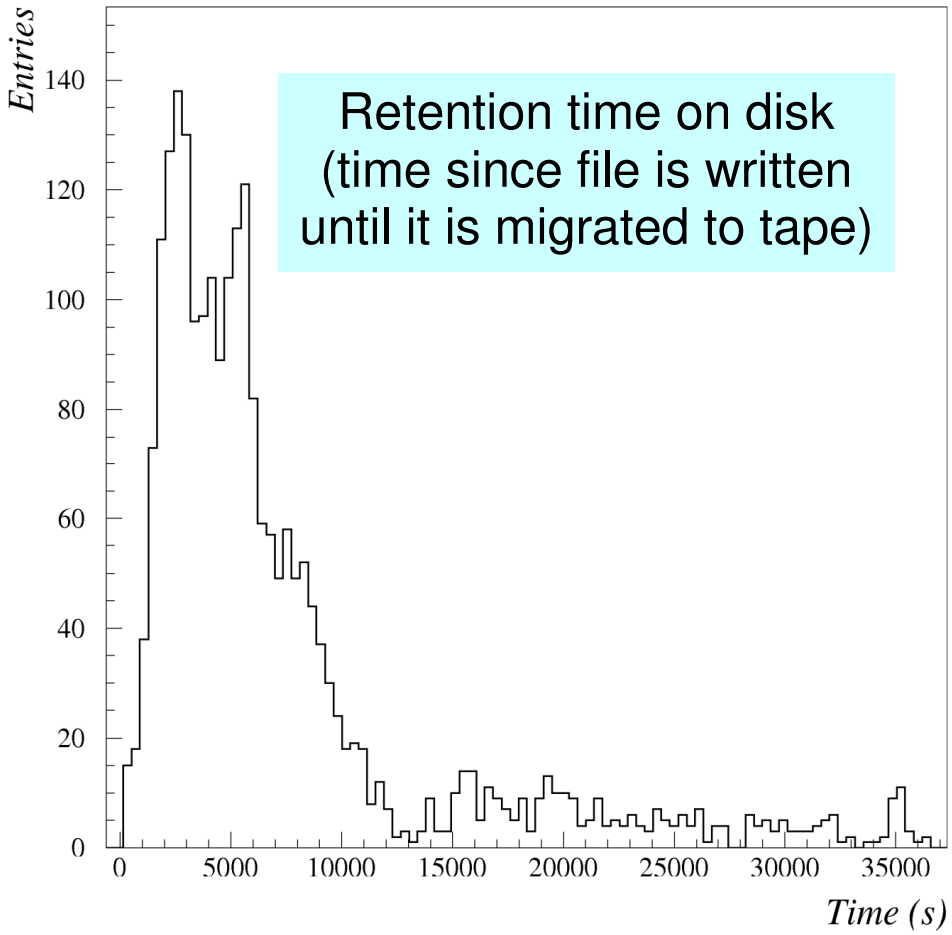

Fig.4: Distribution of the retention time on disk of the files in the GPFS/TSM system