IL NUOVO CIMENTO Vol. 32 C, N. 2 Marzo-Aprile 2009
DOI 10.1393/ncc/i2009-10413-9

Colloquia: CSFI 2008

Storage Infrastructure at the INFN LHC Tier-1

A. Cavalli([1]), L. dell'Agnello([1]), D. Gregori([1]), L. Magnoni([1]), B. Martelli([1]), G. Peco([2]), A. Prosperini([1]), P. P. Ricci([1]), E. Ronchieri([1]), V. Sapunenko([1]), V. Vagnoni([2]) and R. Zappi([1])

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

(ricevuto il 4 Agosto 2009; pubblicato online il 6 Ottobre 2009)

Summary. — In this paper we will describe the Storage Infrastructure of the INFN-CNAF Tier-1, used to store data of High Energy Physics experiments, in particular those operating at the Large Hadron Collider.

PACS 89.20.Ff – Computer science and technology.

1. – Introduction

The growth of resource needs for the LHC experiments requires larger and larger quantities of data storage and increasing perfomance demands. In the framework of WLCG, all the Tier-1 sites will have to provide different access types to the storage systems, both disk resources for faster access and tape resources for long-term data custodial. At the INFN-CNAF Tier-1, all storage resources are organized in a Storage Area Network (SAN) and served by Linux machines used as disk and tape servers; to allow for redundancy, all the servers and the storage resources have two interconnections to the SAN. In the next sections we will describe our infrastructure in terms of hardware and software in use.

2. – StoRM

The StoRM service [1, 2] is a storage resource manager for generic disk based storage systems implementing the Storage Resource Manager (SRM) interface version 2.2 [3]. StoRM is designed to support guaranteed space reservation and direct access (native POSIX I/O call), as well as other standard Grid access protocols. It separates the data management layer from the underlying storage systems characteristics.

StoRM takes advantage from high-performance parallel file systems like GPFS (from IBM) and Lustre (from Sun Microsystems) but it is able to manage data on any other

standard POSIX file system through a driver mechanism. Cluster file systems allow several disk LUNs attached to multiple storage servers to be configured as a single file system, providing transparent parallel access to storage devices, while maintaining standard UNIX file system semantics. StoRM is able to leverage the advantages resulting from a cluster approach in a Grid environment, enabling data intensive applications to directly access data on the storage resource without interacting with any other transfer services.

A modular architecture decouples the StoRM logics from the different file systems supported, and plug-in mechanisms allow an easy integration of new file systems. With this approach, our data centre is able to choose the preferred underlying storage system maintaining the same SRM service.

Another important characteristic of StoRM is the capability to identify the physical location of a requested data without querying any database service, but simply evaluating a configuration file, an XML schema that describes the storage namespace and input parameters as the logical identifier and SRM attributes. StoRM relies on the underlying file system structure to identify the physical location of the data file.

Authorization on file and space access is another important driving feature of StoRM. To enforce permissions on files and directories, StoRM is able to use the ACL support provided by the underlying file systems. The resulting security model is highly configurable in order to satisfy the requirement coming from heterogeneous scenarios, such as the case of High Energy Physics (HEP) and Economics and Financial applications.

StoRM is an open source project and it is distributed together with the gLite middleware within the INFN Grid release [4], that collects Grid component packages with tools to automatically install and configure services on Scientific Linux based hosts. StoRM is currently adopted in the context of the World-wide LHC Computing Grid infrastructure (WLCG) in various data centres, such as the Italian Tier-1 at the INFN-CNAF institute, and different Tier-2 centres across the world.

3. – Castor

CASTOR [5] stands for CERN Advanced STORage manager and is a hierarchical storage management (HSM) system developed at CERN and used to store physics production files and user files. At our Tier-1, CASTOR has been our choice for the D0T1 (tape based storage with disk stage area as frontend) implementation for the last years with effective results. Data is copied from the user resource to the CASTOR front-end (disk servers with connected disk space) and then subsequently copied to the back-end (tape servers). Data buffering allows the system to optimize the disk-to-tape copy (migration process). Currently, we have 40 disk servers as front-end and each of them has five or six file systems. Typical size is comprised between 1.5 to 2 TB for each file system, and both XFS and EXT3 are used. Disk servers are connected to the SAN using full redundancy on FC 2 Gb/s (on latest machines 4 Gb/s) connections, with a hardware dual controller and a software Qlogic SANsurfer Path Failover (or in other cases vendor specific software). As CASTOR back-end we use tape servers (each one with dedicated tape drive) which are connected via FC to 6 LTO2 and 10 9940B tape drives which are located inside a single STK L5500 silo. The silo itself is partitioned with 2 form-factor slots. For LTO2 tapes there are 2000 slots and the other 3500 slots are for 9940B tapes (5500 slots in total). Both the technologies use 200GB cartridges, so the total capacity is roughly 1.1 PB (non-compressed). The library operations are managed with a dedicated Sun Blade v100 machine with 2 internal IDE disks with software RAID1 running ACSLS 7.0 and Solaris OS.

When accessing a file that is not already in the disk staging area, data is obtained from tape and copied to disk (recall process). Files can be stored, listed, retrieved and accessed in CASTOR using command line tools or applications built on top of different data transfer protocols like RFIO (Remote File IO), GridFTP and XROOTD. CASTOR provides a UNIX like directory hierarchy of file names. The CASTOR namespace can be viewed and manipulated only through CASTOR client commands and library calls and it is provided by a Name Server. Currently, we are running CASTOR version v2.1.6-12. The Core services run on machines with SCSI disks, hardware RAID 1 and redundant power supplies. We have distributed CASTOR Core Services basically on three different machines (Scheduler, Stager and Name Server) and on a fourth machine we run the Distributed Logging Facility (DLF) for centralizing log messages and accounting information. CASTOR is driven by a database-centric architecture. A great number of components are stateless and the code is interfaced with Oracle relational database. At CNAF CASTOR is used by the four LHC experiments (ALICE, CMS, ATLAS, LHCb) and by eight other experiments (LVD, ARGO, VIRGO, AMS, PAMELA, MAGIC, BABAR, CDF). Even if CASTOR is mainly used for the management of the tape back-end, some experiments use CASTOR as pure disk, without tape migration. CASTOR uses the LSF scheduler [6] to determine the best candidate resource (file system) for a castor job. The decision depends on the load of disk server and on the number of assigned slots per disk server. In fact, LSF slots are dedicated to each disk server (from 30 to 450) and can be modified in case of need. The SRM (v.2) interface to CASTOR, for Grid users, is provided by two end-points. One is dedicated only to the tape service class, while the other to the disk-only service class.

4. – Oracle

An Oracle database service is deployed for relational data storing/retrieving. The database service is an infrastructural layer which lies beneath several software systems ranging from experiments metadata archiving/retrieving systems, to World-wide LHC Computing Grid (WLCG) services and monitoring services. A non-exhaustive list of such database-based systems comprises: detector conditions databases (CondDB), event TAG collections, calibration databases, LFC (LCG File Catalog), FTS (File Transfer Service), VOMS (Virtual Organization Management Service), CASTOR (CERN Advanced STORage Manager), LEMON (LHC Era MONitoring).

Due to its criticality, the Oracle database services is deployed in a fully redundant manner implementing High Availability (HA) techniques at various levels:

- At the hardware storage level, RAID controllers and Storage Area Networks provide full redundancy and failover of each LUN seen by each server.
- At the software storage level, ASM (the Oracle Advanced Storage Manager) implements striping and mirroring of Oracle database blocks across a group of LUNs.
- At the database level, the Oracle Real Application Cluster (RAC) technology allows to share a database amongst multiple servers, implementing load balancing and failover.
- At the WAN level, database replication via Oracle Streams is deployed. This technique allows to ship Oracle data from CERN to CNAF, providing a disaster recovery and load balancing solution on the wide area network.

Database on-line backup is performed through Oracle Recovery Manager (RMAN) with a central catalog set-up. In the near future, we foresee to integrate the Oracle backup service with our TSM infrastructure by means of the RMAN-TSM plug-in which allows sending a backup copy to tape directly from RMAN.

Presently the Oracle database service is composed by 32 dual-core servers organized in 7 clusters, serving a total amount of 10TB of Fibre Channel storage plus 10 TB of SATA disks. Thanks to its fully HA deployment, CNAF database service has achieved an availability rate of 98.7% in 2007.

5. – GPFS-TSM

Currently, we have about 2600 TB of raw disk storage online and the space will be increased by roughly 90% during the next year. The great majority of the disks are SATA ones, due to a low-cost/good-performance compromise, but for specific applications (*e.g.*, database ones) which require a higher level of reliability and higher performances in the random I/O access, native Fibre Channel disks are employed. The choice of using Fibre Channel and SAN technology for all the disk hardware at our Tier-1 was driven by the following considerations:

- Disk-servers implement a so-called no "Single Point of Failure" (SPoF) system, where each component of the storage system is redundant.
- The SAN gives the best flexibility since we can dynamically assign new volumes or disk storage boxes to disk-servers without stopping the service.
- LAN free systems both for archiving and backup purpose to the tape facilities are possible.

The General Parallel File System (GPFS) [7] is the file system adopted to store data on disks at CNAF. GPFS is a world-wide distributed software developed by IBM, which provides a Parallel File System implementation. The idea of our implementation of GPFS is to provide a fast and reliable service, without single point of failures, parallel filesystem with direct access (POSIX file protocol) from the Farm worker nodes (*i.e.* the "clients") using block level I/O interface over standard Ethernet network. In such an implementation, the clients do not need to have direct access to the SAN, they instead contact the GPFS Network Shared Disk (NSD) disk servers using the LAN and the disk servers provide all the I/O over the SAN and the storage boxes layers.

Since GPFS works in a cluster configuration, with an opportune SAN hardware a true higly available system is possible (disk server failures just decrease the maximum theoretical bandwidth, but the file system remains always avaliable to the clients) and a single "big file system" for each Virtual Organization (VO) is possible, which is strongly preferred by users.

Furthermore, from 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 back-end such as Tivoli Storage Manager (TSM) [8], hence allowing the implementation of a complete HSM system based on GPFS and TSM.

REFERENCES

[1] Magnoni L., Zappi R. and Ghiselli A., *StoRM: a Flexible Solution for Storage Resource Manager in Grid*, in *Proceedings of the IEEE 2008 Nuclear Science Symposium (NSS-MIC 2008), 2008, Dresden, Germany.*

[2] Carbone A., Dell'Agnello L., Forti A., Ghiselli A., Lanciotti E., Magnoni L., Mazzucato M., Santinelli R., Sapunenko V., Vagnoni V. and Zappi R., *Performance studies of the StoRM Storage Resource Manager*, in *Proceedings of the 3rd IEEE International Conference on e-Science and Grid computing (eScience2007), 2007, Bangalore, India.*

[3] Sim A. and Shoshani A., *The Storage Resource Manager Interface Specification Version 2.2, Grid Final Documents*, 2008, OGSA Data Working Group, Open Grid Forum (OGF), Available [ONLINE] at `http://www.ogf.org/documents/GFD.129.pdf`.

[4] *INFNGRID Release Wiki*, Available [ONLINE] at `http://igrelease.forge.cnaf.infn.it/`.

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

[6] Load Sharing Facility (LSF) is a commercial computer software job scheduler developed by Platform Computing. More documentation is available [ONLINE] at `http://www.platform.com/Products/platform-lsf/`.

[7] IBM GPFS product, Available [ONLINE] at: `http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/topic/com.ibm.cluster.gpfs.doc/gpfsbooks.html/`.

[8] IBM Tivoli product, *IBM Tivoli Storage Management Concepts*, IBM Redbooks Series, SG24-4877.