Large storage infrastructures for high-throughput computing and clouds

A. Cavalli[1], L. dell'Agnello[1], M. Dibenedetto[3], M. Favaro[1], D. Gregori[1], M. Pezzi[1], A. Prosperini[1], P.P. Ricci[1], E. Ronchieri[1], V. Sapunenko[1], V. Vagnoni[2], V. Venturi[3], G. Zizzi[1]

[1] INFN-CNAF, Bologna
[2] INFN Bologna
[3] INFN-CNAF/IGI, Bologna

Abstract

In recent years, large scientific collaborations in various research fields have been accumulating unprecedented amounts of data, reaching in some cases the scale of several tens of PetaBytes (PB) per year. This is e.g. the case of the Large Hadron Collider (LHC) experiments at CERN. The experience gained in this sector can be of great importance for other communities which are now starting to face similar needs, especially in the view of exploiting existing data centres to serve storage resources (either by means of Cloud or Grid abstractions) to several different research groups and experiments. A very large installation of a Mass Storage System (MSS), comprising online (disk) and nearline (tape) media amounting to 10-100 PB of available space, is a complex system composed of many layers, out of which the higher level (e.g. Cloud) interface is just the tip of the iceberg. The components of such systems are e.g.:

- Hardware: disk media and controllers, fibre-channel network for Storage Area Network (SAN) and Tape Area Network (TAN), 10 GigaEthernet network, 10 GigaEthernet disk-servers and WAN data-movers, Tape-servers, etc.;
- Software: parallel filesystems, tape storage managers, file transfer services, storage management services, monitoring and accounting services, high level interfaces and abstractions, etc..

The INFN Tier-1 at CNAF is the main INFN computing facility and one of largest European computing sites. Operational since 2005, the Tier-1 is part of the WLCG (World-wide Computing Grid) collaboration, which provides the computing and storage infrastructures for the experiments at the LHC. Similarly to the other ten WLCG Tier-1 centres around the world, CNAF provides computing and storage resources, both as disk and tape. The fraction of CNAF resources is about 13% of the total available at all the WLCG Tier-1's. Currently CNAF houses more than 11 Petabytes of disk space and 14 Petabytes of tape space, used by more than 20 world-wide physics communities (not only LHC ones). The access to the storage is granted through standard protocols according to the Storage Resource Manager (SRM) specification adopted by WLCG

community (and more generally on the Grid). SRM is an abstraction layer that allows users to access the storage through a common interface. The web service interface described in the SRM specifications provides a way to transparently move files to and from the Grid, with a free choice of transfer protocol and with a well defined level of service. It provides support for the most common file system-like operations enriched with fine grained commands to control storage space management. Behind such an interface, data centres can made their own choices about MSS hardware and software solutions. There are several SRM implementations supporting a variety of storage configurations.

The CNAF Tier-1 has developed a general solution for a highly scalable and robust MSS. It is a modular system composed from industrial standards: the General Parallel File System (GPFS) and the Tivoli Storage Manager (TSM) both from IBM with an ad hoc interface layer developed by INFN, and StoRM, an implementation of the SRM specification, also developed by INFN. Amongst other things, the system implements a clever model for file recalls from tape, that minimize mechanical operations of tape robotics such as mounting, un-mounting, and seeking.
Experience, gained over several years of production has demonstrated the efficiency and the completeness of this solution.

StoRM has been designed around the driving principle of leveraging the advantages of cluster filesystems like GPFS and Lustre. Furthermore, in the last development branch, StoRM provides a brand new WebDAV interface. This interface hides the details of the SRM protocol and allows users to mount remote storage as a partition of their own desktops, or to simply browse data in a storage element via a web browser with or without x509 authentication.

In this contribution we aim at giving a complete overview of how a multi-PB scale storage infrastructure works and is operated in production, from the lower level hardware layers to the higher level software interfaces. We will also present the main achievements and performance figures obtained during these last years of activity by the main scientific collaborations working at the INFN-CNAF Tier-1. The design of an efficient, robust and reliable (large) Cloud storage installation relies on the correct choice of the many components involved in the system and on how they work in cooperation. We wish to share our experience with other communities.