Journal of Physics: Conference Series **513** (2014) 042023 doi:10.1088/1742-6596/513/4/042023

Xrootd data access for LHC experiments at the INFN-CNAF Tier-1

Daniele Gregori[1], Tommaso Boccali[2], Francesco Noferini[3], Andrea Prosperini[1], Pier Paolo Ricci[1], Vladimir Sapunenko[1], Vincenzo Vagnoni[4]

[1] INFN-CNAF, Viale Berti-Pichat 6/2, 40127, Bologna Italy
[2] INFN-Pisa, Largo B. Pontecorvo 3, 56127, Pisa Italy
[3] Centro Fermi, Piazza del Viminale 1, 00184, Roma Italy
[4] INFN-Bologna, Via Irnerio 46, 40126, Bologna Italy

E-mail: daniele.gregori@bo.infn.it

Abstract. The Mass Storage System installed at the INFN-CNAF Tier-1 is one of the biggest hierarchical storage facilities in Europe. It currently provides storage resources for about 12% of all LHC data, as well as for other experiments. The Grid Enabled Mass Storage System (GEMSS) is the current solution implemented at CNAF and it is based on a custom integration between a high performance parallel file system (General Parallel File System, GPFS) and a tape management system for long-term storage on magnetic media (Tivoli Storage Manager, TSM). Data access to Grid users is being granted since several years by the Storage Resource Manager (StoRM), an implementation of the standard SRM interface, widely adopted within the WLCG community.

The evolving requirements from the LHC experiments and other users are leading to the adoption of more flexible methods for accessing the storage. These include the implementation of the so-called storage federations, *i.e.* geographically distributed federations allowing direct file access to the federated storage between sites.

A specific integration between GEMSS and Xrootd has been developed at CNAF to match the requirements of the CMS experiment. This was already implemented for the ALICE use case, using ad-hoc Xrootd modifications. The new developments for CMS have been validated and are already available in the official Xrootd builds. This integration is currently in production and appropriate large scale tests have been made.

In this paper we present the Xrootd solutions adopted for ALICE, CMS, ATLAS and LHCb to increase the availability and optimize the overall performance.

1. Introduction

The INFN-CNAF Tier-1 [1, 2] has been the main computing centre of INFN since 2005. As a WLCG Tier-1, it provides medium and long term data storage, as well as computing resources to reconstruct and analyze data, and for running massive physics simulations needed by the LHC experiments. To allow efficient access to the storage system, we developed GEMSS [3], a system that integrates GPFS [4] for disk management, TSM [5] for tape management and StoRM [6] as a SRM interface for transparent access to the data from the applications, either locally or from remote sites.

Currently, the LHC experiments are considering other ways to access data in addition to the SRM protocol, in particular via Xrootd. The ALICE experiment has been adopting Xrootd

since the beginning of its activity, some months ago also CMS, ATLAS and LHCb have entered the scene.

2. ALICE Xrootd

The specific ALICE data management is based, as mentioned, on the Xrootd protocol to distribute and access data through an authorization system provided by the AliEn [8] framework, as for all the other services exploited by ALICE. The authorization chain is developed in the following way: a user that needs to access a given file sends a request, with his own token (proxy), to the AliEn file catalogue to know where the file is and to get an authorization signed by AliEn. Once the user gets an authorization, the request can proceed to the site which has to verify the authorization and provide the file. Therefore write/read rights are not managed at the site level, but they are listed only in the AliEn catalogue for each ALICE user.

The standard installation for ALICE is a "plain" Xrootd, and it foresees a structure with a machine dedicated to manage user requests (a redirector which redirects the user to the server which owns the file) plus a number of independent servers which contain different amounts of data. The implementation of the ALICE data access with GEMSS is slightly different with respect to the standard case. The CNAF Tier-1 storage system is based on a global filesystem, GPFS, which naturally provides a direct POSIX access, and it includes also all the Xrootd servers dedicated to the two ALICE file systems, which share all the files. This implies that CNAF has an additional layer to the standard configuration which potentially has some advantages and disadvantages. The evident disadvantage when introducing an additional layer is a degradation of performance. This potential problem has been solved by fixing an appropriate size of the Xrootd cluster, and no degradation of the performance at all has been observed. On the other hand a big advantage of such a configuration consists in an almost perfect balance of the load on the servers both in writing and reading access.

The system is also fully redundant, since every Xrootd server belongs to GPFS cluster and in the case of disappearance of one Xrootd server, the access to any file can be made through the remaining servers. An additional feature of the ALICE Xrootd configuration at CNAF is that the redirector machine is replaced by an alias DNS which points to all the Xrootd servers, so that each server basically manages itself. This choice allows the system administrator to add/remove servers in the partition on-the-fly, *e.g.* when updating the system to a new operative system or a new Xrootd version. The described configuration is replicated for the two ALICE instances at CNAF: one for the so-called T1D0 and one for the T0D1 storage class. The T1D0 corresponds to the custodial data instance and it is interfaced with TSM to provide data migration to tape and recalls from tape with a disk buffer of 230 TB and 2 dedicated Xrootd servers. The T0D1 is the disk-only instance, which is usually accessed by user analysis jobs and Monte Carlo productions. It has to serve the largest amount of requests, therefore it is characterized by 1.3 PB of disk and 6 Xrootd servers. In Fig. 1 the maximum throughput reached during 2013 by the T0D1 instance is shown, which was greater than 6 GB/s.

For the T1D0 instance, the development of an Xrootd plugin was required to manage the recalls from tape in a transparent way, while for the T0D1 instance the system was working without any extra configuration. The role of the plugin basically consists of triggering a recall from tape in the GEMSS system when an Xrootd file access was submitted. Recently, CNAF Tier-1 Xrootd server was upgraded to operating system SL6 and to Xrootd version v3.3.2 (the latest ALICE production version) in order to be aligned with the newest features of Xrootd development.

3. CMS Xrootd

The CMS computing model requires local (LAN) access to data from the jobs. It also allows a small fraction of remote direct access, with the aim of improving the reliability and exploitation

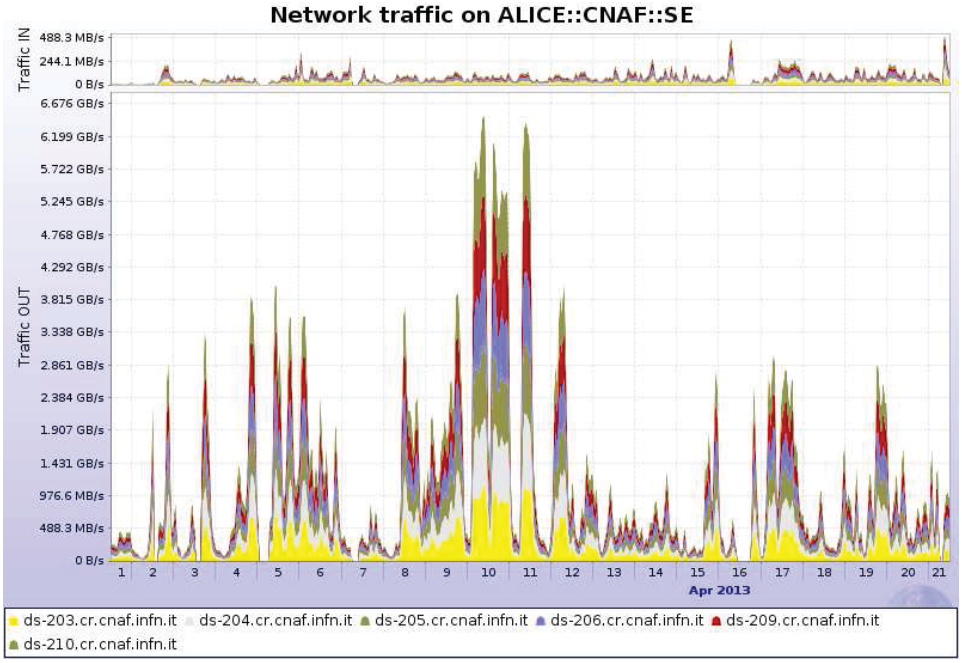

Figure 1. ALICE throughput at CNAF (T0D1) as reported by MonaLisa (ALICE monitor system) in April 2013.

of storage resources. Most jobs access data locally by design; those which do not succeed, either because of local storage problems or by explicit choice, are redirected to an Xrootd server inside the CMS federation. In Italy, the major computing centres (the CNAF Tier-1, and the Tier-2s in Legnaro, Pisa, Bari and Rome) already joined the CMS Xrootd federation, and are able to serve remote access requests, routed via a redirector hierarchy.

The GPFS-TSM (GEMSS) setup in use at CNAF by CMS does not allow for a trivial deployment of Xrootd servers. The main reason is that tape-resident files are always visible in GPFS, even if the file content is not present on disk. A standard Xrootd server would publish to the upstream redirectors all the files present in the GPFS filesystem; all the requests of tape-resident files from remote clients would then trigger tape recalls in an chaotic and inefficient way. This could put unnecessary load on the tape drives, with a potentially unlimited number of simultaneous remote requests to be handled. The adopted solution consists of hiding the tape-resident files: these files are not exported by the `cmsd`[1] process. Any request which should anyhow arrive for such files (for example via direct Xrootd server connection), is injected into the GEMSS recall system, which arbitrates the number of concurrent recalls and merges those requesting files on the same tape. After a testing phase, Xrootd developers packaged the patches in the form of a specific `stat` [9] plugin, which can be enabled at configuration level. The code is now in the Xrootd development branch and will be available to everyone shortly.

Currently, there are four Xrootd servers in production at CNAF Tier-1 for CMS. They also run the GridFTP service at the same time. Each server is equipped with a 10 Gb/s Ethernet connection and a dual channel port 8 Gb/s Fibre Channel to the storage backend over the storage area network (SAN). Servers run operating system SL 5.4 and Xrootd version 3.3.3. During the test phase, the Xrootd service was activated on a single server in order to compare if the Xrootd additional workload would remain in an acceptable range. About 30 TB of data were read via Xrootd during the normal operation of the GridFTP service. This test clearly

[1] The `cmsd` process is used to localize files in the Xrootd cluster.

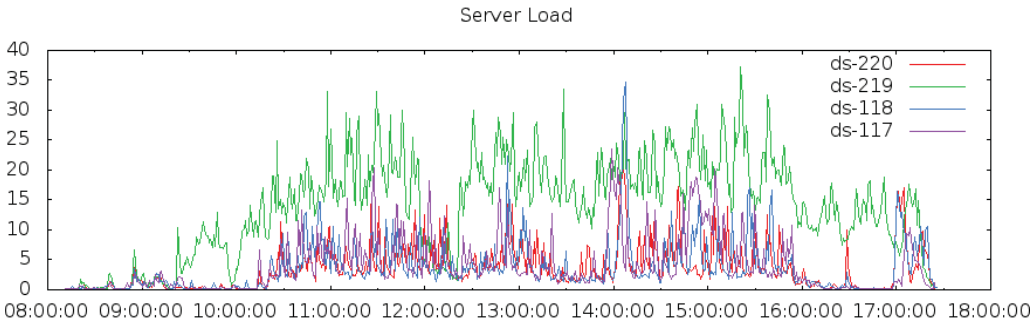

Figure 2. CPU load test of the Xrootd/GridFTP server for CMS. Each server is a GridFTP server in production, and ds-219 (in green) is also an Xrootd server, serving 30 TB of data via Xrootd. The CPU load for ds-219 is only slightly higher, but in an acceptable range.

showed that the impact of the Xrootd traffic was not degrading the overall performance of the single server. The CPU load recorded during the test is shown in Fig. 2.

4. ATLAS Xrootd

During 2013, the Tier-1 also joined the ATLAS Xrootd federation for read-only access. The configuration adopted is based on the use of four machines that work both as GridFTP and Xrootd servers. The servers are machines with 10 Gb/s network interfaces, with operating system SL5.4. The GPFS file system has a size of 4 PB and the access to disks takes place directly over the SAN. The configuration does not allow access to the tape media, hence it is not required to use the plugin developed for the CMS and ALICE user cases. Since an Italian redirector is not present the CERN one is used instead. As the four machines access the same storage, they have been grouped into a single DNS alias. This bundling allows a round-robin access to the servers, thus the CPU load and network traffic are balanced. A schematic layout of the ATLAS Xrootd system at CNAF Tier-1 is reported in Fig. 3. To ensure the proper behaviour of the service, we have also developed automatic controls that make copies of test files using the `xrdcp` [10] command from the ATLAS storage area to the local disk of a dedicated User Interface. These checks are carried out at intervals of four hours and use the hostname contained in the DNS alias. More frequent checks could verify the status of services, and in case of errors, Nagios will send an email notification.

5. LHCb Xrootd

LHCb has also recently started to use Xrootd for read-only data access. Even in this case the servers are used to run the GridFTP service as well. The peculiarity of this experiment lies in the lack of a redirector. LHCb queries its file-catalog to know the location of the files and uses StoRM to obtain the Xrootd transfer URL (tURL). All servers have direct access to the SAN, they are bound into a DNS alias and monitored by Nagios [11].

6. Conclusions

Data access via Xrootd has been adopted by all the LHC experiments, showing good performance in terms of bandwidth, comparable with GridFTP, as well as a high level of availability and reliability. Appropriate Nagios checks have been developed [12] to monitor the correct operation of daemons, the memory usage and the CPU load.

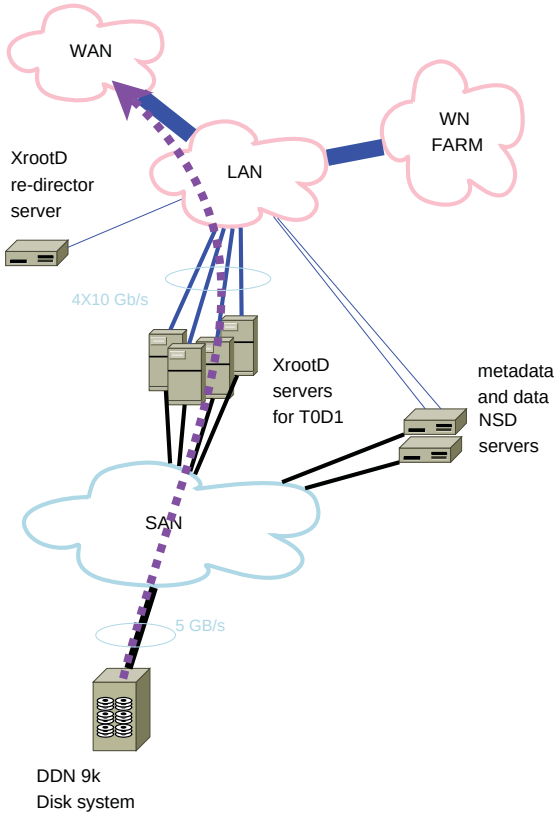

Figure 3. ATLAS storage layout: four Xrootd servers accessing files that are in the T0D1 Storage Class (disk-only).

Acknowledgements

We wish to thank Fabrizio Furano and Andrew Hanushevsky for having helped us in developing the solution adopted for the CMS Xrootd implementation, and in consolidating the initial patches into an Xrootd standard component.

Bibliography

[1] D. Gregori et Al., *INFN-CNAF activity in the TIER-1 and GRID for LHC experiments* , IEEE Parallel & Distributed Processing, 2009, E-ISBN: 978-1-4244-3750-4
[2] G. Bortolotti et Al., *The INFN Tier-1*, 2012 J. Phys.: Conf. Ser. 396 042016
[3] 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
[4] `http://www-03.ibm.com/systems/software/gpfs/`
[5] `http://www-03.ibm.com/software/products/us/en/tivostormana/`
[6] `http://italiangrid.github.io/storm/index.html`
[7] `http://www.webdav.org/`
[8] `http://alien2.cern.ch/`
[9] Manual Available Online `http://linux.die.net/man/1/stat`
[10] Manual Available Online `http://linux.die.net/man/1/xrdcp`
[11] `http://www.nagios.org/`
[12] D. Gregori et Al., *INFN-CNAF Monitor and Control System*, 2011 J. Phys.: Conf. Ser. 331 042032
[13] Manual Available Online `http://linux.die.net/man/1/curl`