Available on CMS information server

CMS CR -2009/080

The Compact Muon Solenoid Experiment

Conference Report

Mailing address: CMS CERN, CH-1211 GENEVA 23, Switzerland

07 May 2009 (v2, 14 May 2009)

The CMS experiment workflows on StoRM based storage at Tier-1 and Tier-2 centers

D Bonacorsi, I Cabrillo Bartolomé, I González Caballero, F Matorras and A Sartirana

Abstract

Approaching LHC data taking, the CMS experiment is deploying, commissioning and operating the building tools of its grid-based computing infrastructure. The commissioning program includes testing, deployment and operation of various storage solutions to support the computing workflows of the experiment. Recently, some of the Tier-1 and Tier-2 centers supporting the collaboration have started to deploy StoRM based storage systems. These are POSIX-based disk storage systems on top of which StoRM implements the Storage Resource Manager (SRM) version 2 interface allowing for a standard-based access from the Grid. In this notes we briefly describe the experience so far achieved at the CNAF Tier-1 center and at the IFCA Tier-2 center.

Presented at *CHEP - 17th International Conference on Computing in High Energy and Nuclear Physics.,21 - 27 March 2009 ,Prague,Czech Republic,15/05/2009*

The CMS experiment workflows on StoRM based storage at Tier-1 and Tier-2 centers

D Bonacorsi[1], I Cabrillo Bartolomé[2], I González Caballero[3], F Matorras[2] and A Sartirana[4]

[1]Univ. of Bologna / INFN, Italy
[2]IFCA, Univ. of Cantabria-CSIC, Spain
[3]Univ. of Oviedo, Spain
[4]LLR, École Polytechnique, France

E-mail: sartiran@llr.in2p3.it

Abstract.
Approaching LHC data taking, the CMS experiment is deploying, commissioning and operating the building tools of its grid-based computing infrastructure. The commissioning program includes testing, deployment and operation of various storage solutions to support the computing workflows of the experiment. Recently, some of the Tier-1 and Tier-2 centers supporting the collaboration have started to deploy StoRM based storage systems. These are POSIX-based disk storage systems on top of which StoRM implements the Storage Resource Manager (SRM) version 2 interface allowing for a standard-based access from the Grid. In this notes we briefly describe the experience so far achieved at the CNAF Tier-1 center and at the IFCA Tier-2 center.

1. Introduction

In preparation of LHC data taking, the CMS experiment has exercised its computing model in periodic challenges, with increasing scale and complexity, as well as in daily production level operations.

The CMS computing system relies on the tools of the Worldwide LHC Computing Grid project [2]. On top of them CMS builds its own experiment-specific applications: data placement and transfer system, distributed analysis and MonteCarlo (MC) production tools, data bookkeeping and location system, etc. The system is based on a model of "Tiers" of resources (Tier-0, Tier-1s, Tier-2s) with specific functionalities [1, 2, 3].

The Tier-0, located at CERN laboratory, provides real time reconstruction, tape archiving and distribution to Tier-1s of detector data. The Tier-1s centers, 7 national level centers, provide secondary archive and re-reconstructions of data, serve real data to Tier-2s and import MC data. The Tier-2s, about 50 smaller centers in the world, generate MC samples and provide resources and tools for Physics analysis.

A key role in the setup of the site's activities is played by the deployment and the operation of storage systems. A number of different solutions have been adopted and widely tested by the various sites supporting the collaboration: DPM, dCache, Castor, etc. Recently, an increasing number of sites have moved to adopt StoRM based storage systems [5, 4]. During the last challenges this solution has been thoroughly tested with excellent results, and is now stably deployed in everyday operations by many Tier-2s sites and one Tier-1.

In this note we will report about the experience achieved with StoRM storage at the CNAF Tier-1 center and the IFCA Tier-2 center. In section 2 we quickly review the basics of StoRM storage and of IBM General Parallel File System (GPFS), section 3 briefly describe the status of integration between StoRM and the CMS-specific applications, the experience achieved with StoRM at CNAF and IFCA is described in section 5 and 6 respectively.

2. StoRM SRM System

StoRM is a storage resource manager (SRM) for disk based storage systems implementing the SRM interface version 2.2 [5].

The SRM interface allows the Grid applications to have a standard-based access to the disk storage without any need of knowing the details of the underlying storage technology. All requests of reading/writing files, as well as executing namespace operations (listing, creating and removing directories, etc), are managed by the SRM service in a transparent way. Grid-level authentication is provided by standard GSI interface. The direct interaction with the storage system is performed by the SRM which takes care of extracting the files physical names, reserving the storage space, implementing the ownership policies, etc.

As a SRM service, StoRM works on top of any standard POSIX file system with ACL support, like GPFS[6], XFS[7] and ext3[8]. It supports GridFTP, rfio or file as access protocols (fig. 1). Access is provided by standard POSIX operations, without interacting with any external service that emulates data access, with the result of improving the performance when the underlying file system is efficient. StoRM hence takes advantage from high performance parallel and cluster file systems. ACLs provided by the underlying file systems are used by StoRM to implement the security model.

StoRM has a multilayer architecture, fig. 2, built of two main components: the front-end and the back-end. The front-end exposes the SRM web service interface, manages user authentication and stores the requests data into the database. The back-end executes all synchronous and asynchronous SRM functionalities, takes care of file and space meta-data management, enforces authorization permissions on files and interacts with file transfer services.

Thanks to a plug-in mechanism, the back-end is able to exploit the advanced functionalities provided by the file system (for example by GPFS and XFS) in order to accomplish space

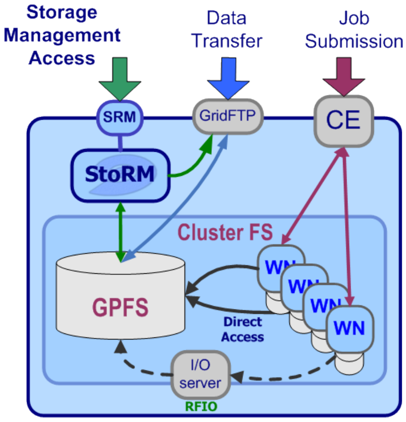

Figure 1. Access to StoRM storage by grid applications.

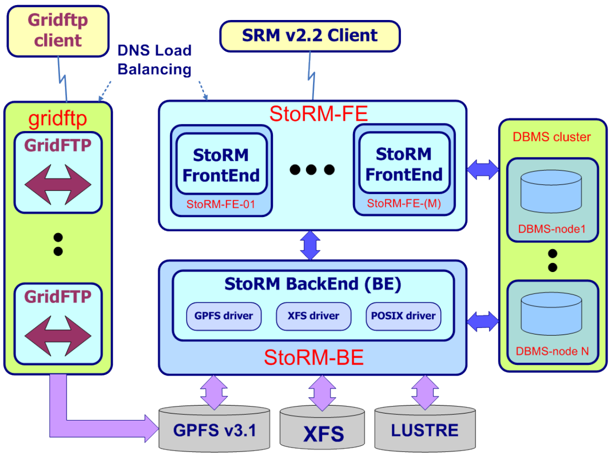

Figure 2. StoRM architecture.

reservation requests. The plug-in mechanism keeps the back-end logic decoupled from the wrapper component that provides file system support, which considerably simplify adding new support for different file systems.

2.1. GPFS

It is worth spending few words of introduction to the IBM General Parallel File System since the StoRM storage solutions which are presented in this note are based on this technology.

GPFS is a file system designed for high-performance parallel access on disk pool storage[6]. GPFS can build and serve a single file system containing up to Petabytes of data on dedicated hardware in a Storage Area Network (SAN) configuration. Moreover, GPFS can provide a software simulation of a SAN storage by the Network Shared Disk (NSD) functionality. Exploiting NSD, GPFS storage system can thus extend and serve data with high performances to wide pools of nodes (such as big farms) which cannot be included in the SAN Fiber Channel network.

3. StoRM and CMS Workflows

The deployment of StoRM based storage at the sites that support the CMS activities obviously poses some compatibility issues which have to be addressed in order to be able to perform all the CMS typical workflows. Considerable effort has been spent on this topic by the administrators of the concerned sites, by the developers of CMS applications and by the StoRM development team. In this section there is a brief description of the status of the integration between StoRM storage systems and the CMS computing.

CMS-specific applications are, in most cases, built to be compliant with the SRMv2 standards, which remarkably simplifies their integration with StoRM based storage.

For example, PhEDEx is the CMS data transfer and placement system which manages, at VO level, all the data placement requests to the CMS sites, it submits the file transfers request to the Grid services and checks that the files are correctly copied at sites. It also manages data deletion from sites. PhEDEx relies on agents located at sites for performing all local operations: files removal, file validation after transfers, etc. The site agents can be configured in order to use general SRMv2 clients while performing such local operations. In particular the SRMv2 client used by PhEDEx are the native dCache ones; dCache is, like StoRM, a storage manager system which is deployed in many Tier-1 and Tier-2 sites supporting CMS, and its clients are fully compatible with all SRMv2 standard based storage systems.

For what concerns the staging out to local storage of the output of production jobs, it is possible to set it up in such a way that it is performed by using the same dCache clients. This solution is clean and quite easy to implement but may not be optimal for this use case. dCache SRM clients use, in fact, the GridFTP protocol to write files. GridFTP doors are usually limited in number and the GridFTP server is often optimized for long distance WAN transfers: large allocated buffer, etc. A better solution is to write files directly with the POSIX 'cp' command. In order to do this in a clean way, the command should be enclosed by a PrepareToPut and a PutDone SRMv2 calls. At the moment such operation is not available in the framework of CMS applications. Despite this limitation, most of SToRM T2s are successfully using POSIX stageout exploiting some locally implemented workarounds.

For what concerns read access to local files from both production and analysis jobs, this can be performed by direct POSIX interaction with the GPFS file system. The physical path on GPFS can be provided to the jobs by the Trivial File Catalogue (TFC). TFC is a local configuration file, used by CMS jobs, which contains the rules for the logical-to-physical file name mapping. Each site has its own TFC and can modify it in such a way to best fit the local storage configuration. This turned out to be a very efficient solution leading to remarkable performances and stability.

As the use of StoRM would spread over sites, an effort to further integrate CMS applications with StoRM based storage systems would be required.

4. StoRM at a Tier-1 center: the CNAF case

INFN-CNAF is a multi-VO Tier-1 located in Bologna, Italy. CNAF supports all four LHC experiments as well as some non-LHC collaborations (BaBar, CDF, Argo, Virgo, etc). Moreover it contributes to various European development projects like LCG, EGEE, INFN Grid, etc.

4.1. Storage setup

CNAF hosts a hybrid storage system with a Castor based tape back-end storage and a StoRM/GPFS based disk-only storage. The center hosts 7 GPFS clusters, version 3.2.1-4, serving more than 1 PB of storage to approximately 600 nodes: worker nodes, user interfaces, GridFTP servers and GPFS servers. Disk storage is connected by a 2-4Gb Fiber Channel SAN system. CMS has a dedicated StoRM, version 1.3.20-03, instance with 2 front-end machines and one back-end machine, a dedicated 190TB of GPFS storage and a pool of three dedicated GridFTP servers. Presently, all GridFTP servers are Scientific Linux CERN 4 (SLC4) 32bit server accessing to the GPFS storage via LAN connection. This configuration implies a limitation in the maximum rate per servers: 1Gb/s maximum inbound/outbound aggregate; moreover, the GridFTP server on 32bit OS has shown to have memory allocation issues. A different configuration, with 64bit SLC4 GridFTP servers directly accessing the storage via SAN Fiber Channel will soon be deployed.

4.2. Experience achieved and future plans

GPFS storage has been in use by CMS at CNAF since the end of 2007 hosting the disk-only area for the unmerged MC files. These are files which are created by the MC production jobs, they are temporarily stored on disk, waiting for being merged, by other jobs, in larger files which will eventually go on tape storage. This solution showed to be remarkably stable. The StoRM endpoint was, instead, first deployed on February 2008 and tested for the first time during CCRC08 Phase I and II [10]. After some initial rump up during the challenge, the endpoint has been constantly exercised with the CMS LoadTest infrastructure (fig. 3, fig 4). Since Apr 2009 the StoRM/GPFS storage is also employed in the continuous reconstruction exercise which is performed by CMS in order to keep Tier-1s filled with processing jobs and ready in sight of data taking.

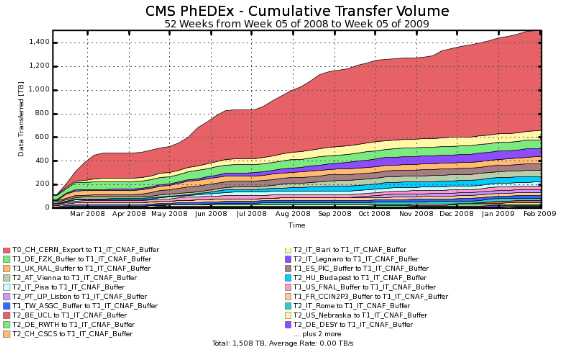

Figure 3. StoRM import transfer volume to CNAF, last year up to March 2009

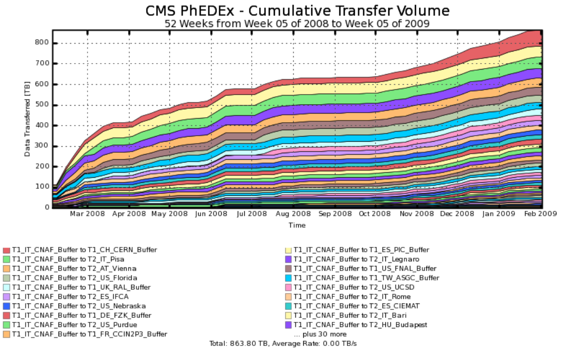

Figure 4. StoRM import transfer volume from CNAF, last year up to March 2009

CNAF is also working to deploy a full disk + tape back end solution based on StoRM, GPFS and TSM[9]. Once a, first, fully operational StoRM/TSM solution for tape backend storage will be deployed, CMS operations at CNAF will consider to start moving from Castor to StoRM storage. Detailed plans for such migration are still under evaluation.

5. StoRM at a Tier-2 centre: the IFCA case

IFCA is a multi-VO Tier-2 located in Santander, supporting CMS in LHC and other non-HEP communities: Plank in Astrophysics, statistical Physics, Biomedicine, etc. It also hosts several GRID computing projects like NGI-ES, DORII, EGEE, EGI, EUFORIA, GRID-CSIC and INTEUGRID.

5.1. Setup

IFCA is currently running StoRM version 1.3.20-03. The present configuration of the endpoint consists of a unique node hosting all StoRM services: front-end, back-end and MySQL database. High performance WAN transfers are served by a pool of 4 GridFTP servers. The underlying POSIX file system is provided by a GPFS cluster. The GPFS Storage Network is deployed on top of a private LAN. StoRM and the GridFTP servers must have access to both networks. With this configuration all the farm has access to GPFS through the usual POSIX linux commands (cp, rm, mv, etc) and the GPFS storage can be used as any other local file system. This simplifies the access to the data on the Worker Nodes and improves the performances and availability compared to other storage systems that need to implement other access protocols.

5.2. Experience with StoRM

The installation of StoRM at IFCA was proposed in order to solve the large request for data transfer and availability of some projects, in particular of CMS. The StoRM endpoint IFCA was deployed in production for the first time on March 2008. Since then the endpoint supported all the IFCA activity with good performance and availability. In fig. 5 and 6 you can see the volume of data transferred with the StoRM endpoint, inbound and outbound, in the last 52 weeks up to March 2009. From the administration point of view StoRM showed to be easy to install, the procedure is quite straightforward by simply following the INFN-GRID instructions, and easy to maintain.

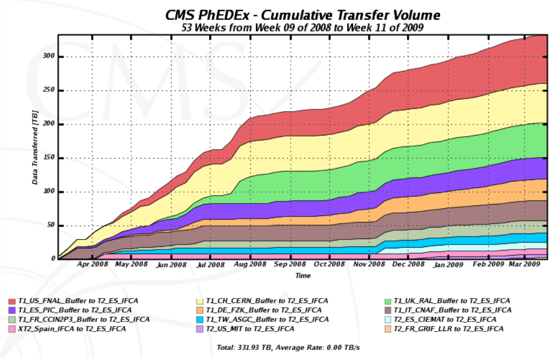

Figure 5. StoRM import transfer volume to IFCA, one year up to March 2009

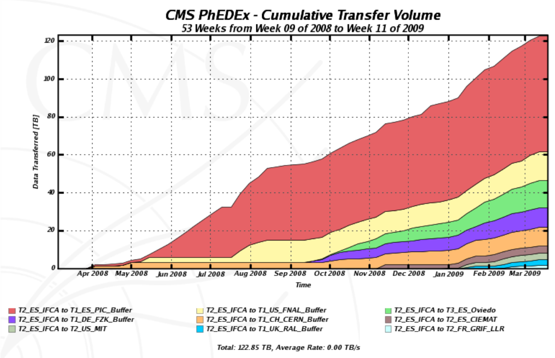

Figure 6. StoRM export transfer volume from IFCA, one year up to March 2009

6. Summary and Outlook

StoRM storage system, despite being one of the youngest which showed up in the framework of Computing in High Energy Physics, has already spread in many sites and was widely tested in the operation of LHC experiments. Many CMS sites have recently turned to operate StoRM storage systems with good results. At the moment, the system works with a good level of integration with the CMS computing framework but some further effort will be needed in order to make StoRM more and more compatible with the standard CMS workflows.

StoRM has been adopted by the IFCA Tier-2 as its main storage system and by the CNAF Tier-1 in a hybrid storage setup together with Castor. The experience so far achieved in these two sites, both from challenges and from daily operations, is very positive and promising in the sight of LHC data taking.

References

[1] C.Grandi, D.Stickland, L.Taylor et al., “The CMS Computing Model” CERN-LHCC-2004-035/G-083 (2004)

[2] I. Bird et al., “LHC computing Grid. Technical design report”, CERN-LHCC-2005-024 (2005)

[3] The CMS Collaboration “CMS Computing Technical Design Report”, CERN-LHCC-2005-023,(2005)

[4] A. Carbone *et al.*, “Performance studies of the StoRM Storage Resource Manager,” http://www.slac.stanford.edu/spires/find/hep/www?irn=8199450SPIRES entry *Proceedings of Third IEEE International Conference on e-Science and Grid Computing*

[5] R. Zappi, L. Magnoni, F. Donno and A. Ghiselli, “StoRM: Grid middleware for disk resource management,” http://www.slac.stanford.edu/spires/find/hep/www?irn=7874316SPIRES entry

[6] Schmuck, Frank; Roger Haskin (January 2002). “GPFS: A Shared-Disk File System for Large Computing Clusters” Proceedings of the FAST’02 Conference on File and Storage Technologies: 231-244, Monterey, California, USA: USENIX. Retrieved on 2008-01-18

[7] http://www.sgi.com/products/storage/software/xfs.html

[8] “Journaling the Linux ext2fs Filesystem” Stephen C. Tweedie (May 1998). Proceedings of the 4th Annual LinuxExpo, Durham, NC. http://jamesthornton.com/hotlist/linux-filesystems/ext3-journal-design.pdf. Retrieved on 2007-06-23.

[9] http://www-01.ibm.com/software/tivoli/products/storage-mgr/

[10] P. Mendez Lorenzo, A Sciaba, S. Campana et al. “The WLCG common computing readiness challenge: CCRC‘08” *3rd EGEE User Forum, Clermont-Ferrand, France, 11 - 14 Feb 2008*