IL NUOVO CIMENTO Vol. 32 C, N. 2 Marzo-Aprile 2009
DOI 10.1393/ncc/i2009-10403-y

Colloquia: CSFI 2008

Data management technologies in LHC Computing Grid

R. Zappi
INFN CNAF - viale Berti-Pichat 6/2, I-40127 Bologna, Italy

(ricevuto il 10 Settembre 2009; pubblicato online il 18 Settembre 2009)

Summary. — Performance, reliability and scalability in data management are cornerstones in the context of the computing Grid, where the volumes of data to be moved are huge, and the data analysis must be supported by high-performance and scalable storage resources. Nowadays, the data management issues are particularly important, considering the large data size and I/O load that the Large Hadron Collider (LHC) at CERN is going to produce. The Enabling Grids for Enabling e-Science (EGEE) EU funded project, where the Italian National Institute for Nuclear Physics (INFN) is a key member, is responsible to release and maintain the currently world's largest production grid, a sophisticated hierarchy of data management and storage tools developed to help physicist, but also other scientific communities, in the face of distributed data management problems. This paper reviews the main technologies employed for storage and data management in EGEE, and in the associated Worldwide LHC Computing Grid (WLCG) project.

PACS 89.20.Ff – Computer science and technology.

1. – Introduction

The Worldwide LHC Computing Grid (WLCG) [1] is the largest Grid infrastructure in operation today, comprising more than 250 sites spread over 45 countries on 5 continents. It is based on a hierarchical, multi-tier architecture composed by geographically distributed computing centers. The computer model is designed to have the power of a real-time processing and storing of the huge sustained data flow produced by the High Energy Physics detectors located at CERN, the European Laboratory for Particle Physics. There are four major experiments at LHC: ALICE, ATLAS, CMS and LHCb, and each computing model used by experiments rely on the WLCG for the necessary storage and computing resources. The main contributors to the WLCG currently are the Enabling Grids for E-sciencE (EGEE) project [2], the Open Science Grid (OSG) [3], and the Nordic Data Grid Facility (NDGF) [4], but there are also important contributions coming from minor initiatives [5]. In the data grids delivered by WLCG, a keystone is represented by middleware responsible for the data management. The data management middleware must offer a high degree of control, to ensure that data is placed correctly for processing and safe keeping; but it also must offer a sufficient abstraction of the large variety of storage resources, in order to present them uniformly and then enabling interoperability. This short paper reviews the main middleware components employed for storage and data management in EGEE, and the associated Worldwide LHC Computing Grid (WLCG) project.

2. – Analysis of data from the LHC experiments

When the Large Hadron Collider (LHC) at CERN, the largest experiment of High Energy Physics (HEP) ever built, will be operational at its full capacity, a sustained flow of several GBs per second of rough data will be produced by four attached experiments. The huge data flow generated by the LHC experiments will be processed in real time, reducing the data flow to about 300 MB/s, and then stored at CERN. Beside storing this data at CERN, which represents the Tier-0 in the WLCG architecture, the data will be further distributed among the 11 Tier-1 centers around the world (one is located in Bologna, Italy, at INFN-CNAF [6]) in real time for further processing, with a transfer rate around 10 Gb/s. Subsequently, parts of this data will be sent to hundreds of associated Tier 2s, and finally data will be available to research department and to physicists. Tier 0 and 1, in contrast to Tier 2s, are responsible for storing data safely on tape media and as such build the global backup of this precious information. Vice versa, the Tier 2s have as main responsibility to reprocess the data coming from the Tier-1s. In order to process the data, massive CPU farms at the various sites must support reliable and large storage resources and, at the same time, must allow a fast data-access from any computer node within the CPU farm. Large volumes of data must be moved from Tier 0 to the various Tier-1s to be stored in a safe place, and redistributed to many Tier-2s to be reprocessed and analyzed. In this scenario of distributed data model, where the volumes of data to be moved are huge, a paramount is the reliability of data storage and data movement.

3. – Grid technologies

Data storage resources are one of the basic building blocks of a data Grid. There exists a very wide variety of storage solutions to consider, ranging from a memory stick to a multi-petabyte tape silo. Ultimately, all the solutions store data offering different levels of Quality of Service (QoS) and have different semantics for data access, both for reading and writing. In the context of data grid, the storage resource represents the combination of the storage hardware and the controlling software (*e.g.*, file system), and the data and its availability is managed by the Grid middleware. The storage resources in the WLCG middleware stack are the so-called Storage Elements (SE).

3˙1. *Storage resource.* – Each experiment produces a few (5 to 10) Petabytes of data each year that need to be stored permanently. Due to the large storage requirement for data produced by HEP detector and by simulation, it is not economically feasible to store all data on disks. Therefore, hierarchical storage systems with both disk and tape storage media are used to archive the main copy of data, while the replicas are distributed to Tier 2s and other minor data centers and stored in disk solutions, supported by cluster file system or a pool of commodity storage systems. Example of Hierarchical Storage Systems and MSS are HPSS (LBNL, ORNL, BNL), dCache/Enstore (Fermi, DESY), JasMINE (Jlab), Castor(CERN), MSS (NCAR). While storage resources based on disk systems are, for example, DPM (CERN), dCache(DESY, Fermi) and other commercial solution, like GPFS (IBM) [7] and Lustre (SUN Microsystem) [8].

3˙2. *Storage Element.* – The Storage Element (SE) is a uniform interface to the above different type of storage resources in a data Grid. The SE provides a set of capabilities to grid application and user. The SE provides the storage space to store files, an interface can be used to access the files directly through a POSIX-like protocol, a service to manage the available space and the lifetime of files, and a capability to queue and manage file transfers between the local SE and remote SEs.

4. – Data management

Current data computing model in WLCG adopts a write-once-read-many access model for files. A file once created, written, and closed need not be changed. Nonetheless, the files can be replaced with a newer version: simply removing a version, with all its replicas, and created an updated version. The read-only access model assumption simplifies data coherency issues and enables high throughput data access. Moreover, creating remote, read-only copies (replicas) of files, allow a reduction of access latency and improve data locality, as well, of course, increase the reliability of key files. The key middleware components in WLCG are focused to provide tools for managing files in a distributed context. The files are organized in a global namespace, managed by a File Catalog Service. The EGEE middleware provides high-level data management clients and services, like the File Transfer Service, to distribute data files in a reliable and efficient way. In this way, end users are protected from the complexities of the storage services and catalog implementations as well as transport and access protocols.

4˙1. *File naming.* – In a Grid environment, files can have replicas at many different sites. Ideally, the users do not need to know where a file is located, as they use logical names for the files that the Data Management services will use to locate and access them. In order to guarantee a unique namespace, several different file name conventions are known: Grid Unique IDentifier (GUID), Logical File Name (LFN), Storage URL (SURL) and Transport URL (TURL). While the GUIDs and LFNs identify a file irrespective of its location, the SURLs and TURLs contain information about where a physical replica is located, and how it can be accessed.

4˙2. *LCG File Catalog (LFC).* – The number of files stored in the data can be enormous. These should be accessible from any grid site, without requiring the user to know their physical location. A file can have as many replicas as needed and can be stored in any Storage Element. To provide a full flexibility to the users, a file catalog has been designed to allow a user belonging to a certain Virtual Organization (VO) to locate and retrieve files using a Logical File Name (LFN) from any WLCG grid site. The Data Management team at CERN has been developed the LCG File Catalog (LFC) [9]. Currently, the LFC is widely used by many VOs as File Catalog solution.

4˙3. *Storage Resource Manager (SRM).* – Storage Resource Manager (SRM) services are Grid services providing interfaces to storage resources, as well as advanced functionality such as dynamic space allocation and file management on shared storage resources on the Grid. Moreover, SRM service guarantees secure file availability with lifetime support, and it performs an automatic garbage collection to prevent clogging of storage systems. SRMs are based on a common specification that emerged over time and evolved into an international collaboration. Currently, the SRM standard interface is a Proposed Recommendation (P-REC, GFD.129) [10], available under the organisation of the Open Grid Forum (OGF) (specifically, by the Grid Storage Management WG (GSM-WG) working group [11]).

4˙4. *GridFTP Service*. – GridFTP [12] is an extension of the standard File Transfer Protocol (FTP) to incorporate X.509 based security (RFC2228), allowing seamless interaction with Grid Security Infrastructure (GSI). Among the features that extend the functionalities of a classical FTP protocol, GridFTP also includes extensions such as multi-streaming, which can help to utilise the full bandwidth available on network links. GridFTP is defined as part of Globus toolkit, under the organisation of the Open Grid Forum (OGF) (specifically, by the GridFTP working group). It represents the essential basic mechanism by which data is imported to and exported from the SE. Normally, the GridFTP transfer will be invoked indirectly via the File Transfer Service.

4˙5. *File Transfer Service (FTS)*. – The File Transfer Service (FTS) [13] is a Grid fabric infrastructure service designed to transfer data files between sites. The FTS uses low-level services and tools to perform file movement with a reliable and manageable way. FTS use the concept of channel as logical unit of management of the service. Each channel corresponds to a specific point-to-point link between two sites or between groups of sites. The users' transfer requests are assigned to different channels upon submission, they are managed as a single transfer job. FTS manages transfer jobs on behalf of users: it manages the network and the storage resource at both ends. FTS instructs the Storage Resource Manager (SRM) endpoint of the source and destination sites in order to prepare the two storage elements to transmit and receive data, and finally it manages the transfer through the commands to the GridFTP server hosted remotely. FTS also ensures the management of any failures of the transfer job making a number of automatic retries.

REFERENCES

[1] *Worldwide LHC Computing Grid*, Web site. URL `http://lcg.web.cern.ch/LCG`.

[2] *Enabling Grids for E-sciencE (EGEE)*, Web site. URL `http://www.eu-egee.org/`.

[3] *Open Science Grid*, Web site. URL `http://www.opensciencegrid.org/`.

[4] *NDGF - Nordic DataGrid Facility*, Web site. URL `http://www.ndgf.org/`.

[5] *INFN Grid project*, Web site. URL `http://grid.infn.it/`.

[6] *INFN - CNAF*, Web site. URL `http://www.cnaf.infn.it`.

[7] *IBM General Parallel File System (GPFS)*, Web site. URL `http://www-03.ibm.com/systems/clusters/software/gpfs/index.html`.

[8] *Lustre*, Web site. URL `http://www.lustre.org`.

[9] *LCG File Catalog (LFC) administrators' guide*, Web site. URL `https://uimon.cern.ch/twiki/bin/view/LCG/LfcAdminGuide`.

[10] Shoshani A. *et al.*, *The Storage Resource Manager Interface Specification Version 2.2*, `http://www.ogf.org/documents/GFD.129.pdf` (April 2008).

[11] *Grid Storage Management WG (GSM-WG)*, Web site. URL `https://forge.gridforum.org/projects/gsm-wg/`.

[12] *GridFTP, Globus toolkit*, Web site. URL `http://globus.org/toolkit/docs/4.2/4.2.1/data/gridftp/`.

[13] Baud J. P. and Casey J., *Evolution of lcg-2 data management*, presented at *Proceedings of CHEP04, Interlaken, Switzerland, September* 2004.