IL NUOVO CIMENTO Vol. 34 C, N. 6 Novembre-Dicembre 2011
DOI 10.1393/ncc/i2011-11108-4

Colloquia: IFAE 2011

New data processing technologies at LHC: From Grid to Cloud Computing and beyond

A. De Salvo
INFN, Sezione di Roma - Roma, Italy

(ricevuto il 29 Luglio 2011; pubblicato online il 21 Dicembre 2011)

Summary. — Since a few years the LHC experiments at CERN are successfully using the Grid Computing Technologies for their distributed data processing activities, on a global scale. Recently, the experience gained with the current systems allowed the design of the future Computing Models, involving new technologies like Could Computing, virtualization and high performance distributed database access. In this paper we shall describe the new computational technologies of the LHC experiments at CERN, comparing them with the current models, in terms of features and performance.

PACS 89.20.Ff – Computer science and technology.
PACS 07.05.-t – Computers in experimental physics.

1. – Introduction

The LHC data handling is a big challenge for the Computing Infrastructures. An unprecedented data volume of ~ 15 PB of data is collected every year by the four experiments. The incoming data has to be analysed promptly by thousands of users, requiring more than 200 k of today's fastest CPUs, due to the event complexity, huge number of events and number of concurrent users.

Unlike in the past, the use of centralized resources is not suitable, due to the size of the data processing activites, and CERN can provide only $\sim$ 20–30% of the overall required resources. The rest, $\sim$ 70–80% of the resources, has to be provided by the World LHC Computing Grid (WLCG) partners, using the Grid paradigm, therefore allowing efficient analysis everywhere by means of a fully distributed, decentralized computing system.

The analysis models of the four LHC experiments is based on the basic concept that the data are available at remote sites, spread across the globe, and the Grid Computing tools are used to perform distributed analysis. Users can access their data and computing resources via a single "sign-on" access with an X509 personal certificate [1], thus avoiding remote logins and allowing a fine-grained allocation and management of the remote resources.

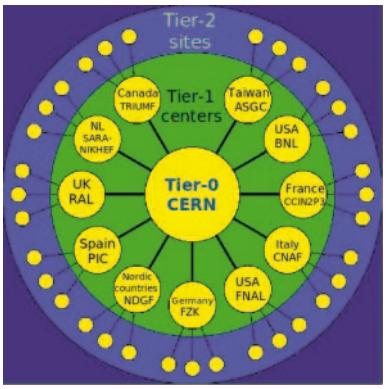

Fig. 1. – The WLCG Tier architecture.

2. – MONARC, the Grid paradigm and WLCG

The current Computing Model of the LHC experiments is based on the original MONARC [2] model, where the Computing Centers are organized hierarchically in categories, known as *Tiers* and identified by a number. CERN is the higher point of the hierarchy, and is therefore referred to as *Tier-0*. Regional centers are called *Tier-1* sites, and they are big sites, which high-performance dedicated connectivity to the Tier-0, also known as LHCOPN [3]. Attached to the Tier-1 sites there are smaller sites, called *Tier-2*s, commonly used for data analysis and specific tasks, communicating at high-speed to their corresponding Tier-1s. All the Tiers are part of the World LHC Computing Grid (WLCG [4], fig. 1), which at the moment counts more than 140 computing centers distributed in 35 countries, 12 large data centers for primary data management and ~ 40 federations of smaller Tier-2 sites.

The access to the resources of the WLCG centers is performed via the Grid Middleware, *i.e.* a software layer that makes multiple computers and data centers looking like a single system. The Grid Middleware provides several services, and in particular:

- a security layer to authorize and grant different levels of permissions;
- a dynamical Information System, to track the available resources in real time;
- job and data management services;
- monitoring and accounting tools.

The Grid Middleware, as shown in fig. 2, is an intermediate layer between the actual resources and the application and serviceware layer, allowing a transparent usage of heterogeneous resources in several sites, without any centralized entity.

3. – Distributed Dabatase access

The Grid Middleware does not cope directly with Database technologies, so the LHC experiments use traditional Relational Databases (RDBMs), based on the *Structured Query Language* (SQL) technologies, like ORACLE or MySQL. The data distribution is performed via proprietary technologies like streaming or master-slave configurations,

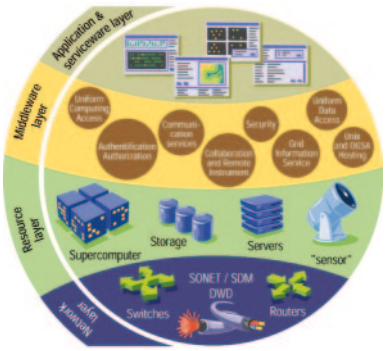

Fig. 2. – The grid middleware stack.

and usually enabled only for Tier-1 sites. However, the traditional systems showed their limitation already in the early stages, since the scalability of the direct connections from the applications was not enough to cope with the high number of concurrent jobs sent by the experiments, resulting in high-load and risk of inefficiency. To cope with the scalability limitations, a new system, called FroNTier [6] (see fig. 3), has been developed. On the server side, FroNTier is a servlet communicating with the Database backend, running under Tomcat. The clients are extensions of a standard squid proxy server, thus providing also caching and failover, taking advantage of a hierarchy of similar services. This hierarchically approach increases the scalability of the overall system and moves the bulk of the load from the Database server to a hierarchy of squid servers.

Another problem to address with the standard RDBMs is the efficient access to large amount of data, with random I/O intensive applications. In general, storing a lot of historical data on expensive transaction-oriented RDMBs is not optimal. An option to unload significant amounts of archivial-type reference data from the RDBMs is to use an high-performance, scalable system, running on commodity hardware. A few noSQL

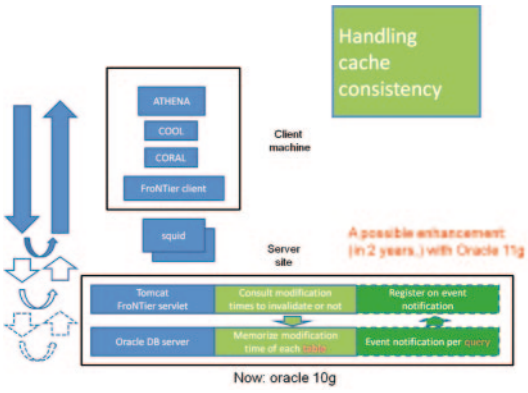

Fig. 3. – The FroNTier client and server architecture for transparent, scalable DB data access.

Databases like Cassandra, MongoDB and HBase([1]) are currently being evaluated to cope with these requirements. The noSQL solutions are mostly openSource projects, already in use by important service providers like Google, Facebook, Amazon or Yahoo! Their good scalability and application level reliability configure them as a good solution to be used for the tasks where the RDBMs are not performing in an optimal way, without being, for the moment, a full replacement for them.

4. – New data access technologies

4˙1. *Experiment software distribution*. – The experiment software is usually shipped in the clusters via Grid jobs and saved in a shared filesystem, making it accessible by all the nodes. This scenario has the advantage of being completely isolated from external problems, but it is not very efficient in terms of used disk space and scalability. To overcome this limitation the LHC experiments are gradually introducing a new method of software distribution, via an http, read-only filesystem called CernVMFS [7]. CernCVMFS has been originally created to work with CERN Virtual Machines [5], but it is a completely separated entity. It provides a Data Store with compressed chunks and file demultiplication, based on the checksum values. The file catalog has a directory structure, symlinks and SHA1 sums of regular files, granting the file integrity. CernVMFS can be mounted as a standard read-only filesystem via the *fuse* kernel module and benefits of a squid hierarchy to guarantee performance, scalability and reliability. While CernVMFS has been designed to distribute the experiment software, the ATLAS experiment is now using it to ship the condition data files too, stored as plain root files.

4˙2. *XRootd federations*. – Users performing data analysis can take a big advantage from using a unified access point for the files in the sites. This functionality can be provided by an XRootd *Global Redirector*, exposing a location-neutral, unified namespace with a single protocol. This approach, mainly introduced for sites smaller than the Tier-2 centers (*Tier-3* sites), is now also considered for Tier-2 and Tier-1 sites, since it introduces an high-performance, low-management data access paradigm. Jobs using the federated XRootd facilities can directly access read-only remote data, stored on existing Storage Technologies like dCache, GPFS or Hadoop.

5. – New Workload Management technologies

5˙1. *WMCore/WMAgent*. – The CMS experiment is currently reworking its ProdAgent architecture to address the shortfalls experienced in data processing. The new system, called WMCore/WMAgent [8], is built of 4 different components:

- *WMSpec*: the language used to describe the organizational units of a workload;
- *WorkQueue*: some chunk of work;
- *JobStateMachine*: workload manager, all state changes handled here are DB operations;
- *WMBS*: defines Job Entities, control creation rates, supervises job splitting for various tasks, handle dependencies.

([1]) `http://cassandra.apache.org/`, `http://www.mongodb.org/`, `http://hbase.apache.org/`

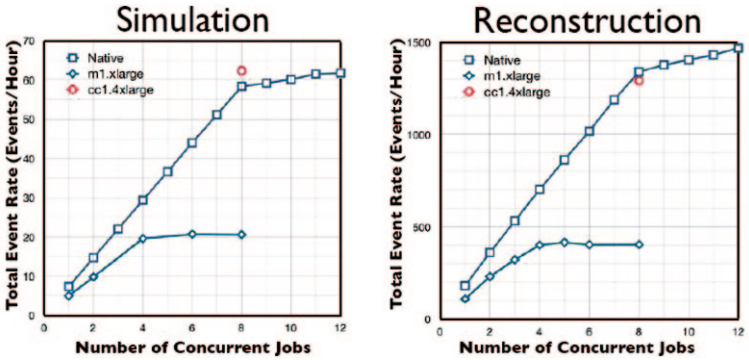

Fig. 4. – A comparison of the performance of local and cloud nodes with Simulation and Reconstruction jobs of the ATLAS experiment.

5˙2. *GlideinWMS*. – A glidein is just a properly configured execution node submitted as a Grid Job. The CMS experiment is currently working on an automated tool for submitting glideins on demand. The new system, *GlideinWMS* [9], is built up of three logical entities, only two being actual services: the glidein factories, who know about the Grid status, and the *Virtual Organization* frontend, who knows about the users and drive the factories.

5˙3. *Cloud Computing*. – Almost all the LHC experiments are interested on the Virtualization and Cloud Computing. ATLAS performed some comparisons on the performance of the Simulation and Reconstruction jobs between real machines and virtual appliances like the ones of the Amazon EC2 Cloud. The results, in fig. 4, show that properly configured Cloud nodes have a performance very close to the one of the real nodes. However, the cost is still prohibitive, therefore using clouds is very attractive, but to make it cost-effective we would currently need to be our own cloud providers.

6. – Conclusions

The WLCG Collaboration prepared, deployed and is now managing the common Computing Infrastructure of the LHC experiments, coping reasonably well so far with the large amount of data that is distributed, stored and processed every day. All the LHC experiments are successfully using the Grid Distributed Computing to perform data analysis of MonteCarlo and real data. Several new technologies are under evaluation, development or already implemented to overcome the current limitations of the system and to better adapt to the experiments' needs. The new technologies described in this paper will allow to achieve a lower maintenance level, more efficiency and optimization of the computing resources available.

* * *

Thanks to D. Barberis, Y. Yao, C. Waldman, V. Garonne, R. Santinelli, A. Sciabà and D. Spiga for the help and the material used for this work.

REFERENCES

[1] Housley R. *et al.*, *Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile*, RFC3280, April 2002.
[2] Aderholz M. *et al.*, *Models of Networked Analysis at Regional Centres for LHC Experiments*, CERN/LCB 2000-001, 24 March 2000.
[3] Bos E. *et al.*, *LHC Tier-0 to Tier-1 High-Level Network Architecture*, 30 July 2005, `http://lcg.web.cern.ch/lcg/activities/networking/LHC%20networking%20v2.dgf.doc`.
[4] Knobloch J., Robertson L. *et al.*, *LHC Computing Grid - Technical Design Report*, LCG-TDR-001, CERN-LHCC-2005-024, 20 June 2005.
[5] Buncic P. *et al.*, *A practical approach to virtualization in HEP*, *Eur. Phys. J. Plus*, **126** (2011) 13, Doi:101140/epjp/i2011-11013-1.
[6] Dykstra D., *Scaling HEP to Web Size with RESTful Protocols: The Frontier Example*, in *Proceedings of the CHEP 2010 Conference, October 2010*, CERN-CMS-CR-2010-240.
[7] Blomer J. and Fuhrmann T., *A Fully Decentralized File System Cache for the CernVM-FS*, in *International Conference on Computer Communications and Networks (ICCCN)* DOI: 10.1109/ICCCN.2010.5560054.
[8] Riahi H., Spiga D. *et al.*, *Automating CMS Calibrations using the WMAgent framework*, in *Proceeding of Science* (ACAT2010) 069.
[9] Sfiligoi I. *et al.*, *The Pilot Way to Grid Resources Using glideinWMS*, in *Proceeding of the 2009 WRI World Congress on Computer Science and Information Engineering*, Vol. **02** (IEEE Press) 2009.