Building a Prototype of LHC Analysis Oriented Computing Centers

G Bagliesi1, T Boccali1, G Della Ricca2, G Donvito3, M Paganoni4
1 INFN Sezione di Pisa, Pisa, ITALY
2 Università degli Studi di Trieste, Trieste, ITALY
3 Consorzio IGI and INFN Sezione di Bari, Bari, ITALY
4 Università degli Studi di Milano Bicocca, Milano, ITALY

Abstract. A Consortium between four LHC Computing Centers (Bari, Milano, Pisa and Trieste) has been formed in 2010 to prototype Analysis-oriented facilities for CMS data analysis, profiting from a grant from the Italian Ministry of Research. The Consortium aims to realize an ad-hoc infrastructure to ease the analysis activities on the huge data set collected at the LHC Collider. While "Tier2" Computing Centres, specialized in organized processing tasks like Monte Carlo simulation, are nowadays a well established concept, with years of running experience, site specialized towards end user chaotic analysis activities do not yet have a de-facto standard implementation. In our effort, we focus on all the aspects that can make the analysis tasks easier for a physics user not expert in computing. On the storage side, we are experimenting on storage techniques allowing for remote data access and on storage optimization on the typical analysis access patterns. On the networking side, we are studying the differences between flat and tiered LAN architecture, also using virtual partitioning of the same physical networking for the different use patterns. Finally, on the user side, we are developing tools and instruments to allow for an exhaustive monitoring of their processes at the site, and for an efficient support system in case of problems. We will report about the results of the test executed on different subsystem and give a description of the layout of the infrastructure in place at the site participating to the consortium.

1. Introduction

In 2009 four centers, already involved in LHC [1] activities with local support for CMS, ALICE and ATLAS, were granted funding from the MIUR (Ministero dell’Istruzione, Università e Ricerca, Italy) to develop prototypes of computing centers with a special focus on Analysis activities, in order to help the Italian LHC communities in their analysis efforts, with specific attention on CMS [2] experiment needs, as detailed in its Computing TDR [3]. The Computing Model depends for analysis activities on jobs being dispatched to remote sites using a GRID middleware, and accessing data previously pre-placed there. Interactive analysis is also happening in the same remote sites, using ad-hoc infrastructures not fully standardized in the model itself and thus left to local implementation and optimization.

We report the topics being covered, the final assessment of result being postponed to the conclusion of the project (Fall 2012).

1 a corresponding author: Giuseppe.Bagliesi@cern.ch

Published under licence by IOP Publishing Ltd
2. Activities in Bari

2.1. Testing Hardware storage solution
The INFN-Bari group have worked on testing several different hardware technologies in order to find out which characteristics could be useful to build a storage solution for a medium large farm for scientific data. We focused on the performance achievable with each of the storage solution, but also on the capabilities of resilience to failures and the overall reliability of the system. The latter is quite important in order to provide an efficient data centre that could provide a high availability to the end user.

Among the examined solutions we have:

• Traditional DAS (Direct Attached Storage)
  o They are characterized by having both disks and CPU on the same box
    ▪ This is quite a good solution in terms of overall performance and performance between CPU and disk.
    ▪ It is also easy to scale horizontally.
    ▪ The main problem with this solution is that if one server fails the data that are stored in that server become unavailable until the problem is really fixed, and this may mean also than a hardware intervention is needed.

• Server plus SAS (Serial Attached SCSI) external box of disk
  o In this case the overall available bandwidth is smaller that in the previous case
    ▪ The SAS link will indeed reduce the bandwidth between CPU and disk.
    ▪ It decouples the server exposing the storage and the storage itself, and indeed it will be possible to replace the server if it has troubles (at the price of a manual intervention).

• Complex SAN (Storage Area Network) environment with big FC (Fiber Channel) switches and O(PB) controllers
  o The storage infrastructure is based on big enterprise storage solutions
    ▪ This implies that the solution is highly fault-tolerance;
    ▪ But the cost of this kind of infrastructure could be fairly higher than the others solutions.
    ▪ The management requires advanced knowledge.

• Simple redundant storage controller with small number of server connected via Fiber Channel
  o A small redundant controller, with hundreds of TB is connected with Fiber Channel technology to two server
    ▪ This solution is able to provide easily high availability of the data to the failures of a server or of a controller.
    ▪ This is also a quite cheap solution in terms of TCO (Total Cost of Ownership), as the acquisition cost is lower than enterprise solutions, and it is easier to be managed.
    ▪ This solution could be horizontally scaled by means of adding new units
  o The performance of this solution could be very good with the help of a parallel, shared file system that could aggregate all the units together.

Our tests are showing us that the latter solution could provide a good storage infrastructure for medium large computing farm involved in scientific data analysis.

2.2. Testing software storage solution
The INFN-Bari group focused also their activities on test several software technologies for managing large storage infrastructure, both in terms of functionalities and performance. The main solutions under test are:

- dCache [4]
- Xrootd [5]
- Lustre [6]
- GPFS [7]

From the point of view of the performance all the tests performed have the aim of optimizing the CPU efficiency of the scientific application while analysing data. In particular we have executed both standard I/O test, like IOZone[8] and bonnie++[9], but specific test executing application from the end user community (in particular CMS and bioinformatics), in order to be sure to test the file-system with the real pattern of usage. In this latter case the parameters that we tend to optimize are: CPU efficiency and the overall running time.

The tests show us that the parallel/distributed file-systems like Lustre and GPFS perform quite better compared to the solution based on data access libraries. For example we extensively tested dCache vs. Lustre storage solutions by means of using a standard CMS analysis application, and we found that using comparable storage infrastructure we are able to increase the CPU efficiency from a 40% in case of dCache storage to more than 85% using Lustre. This is very important for a computing centre that is involved in physics analysis, because it will lead to a big speedup in processing performance and than a better usage of the computing resource that finally means a better TCO for the computing analysis in general.

After an extensive testing on Xrootd storage system, it is evident that this solution is performing slightly worse than Lustre or GPFS file systems, but requires far lower network utilization than the others. This becomes important when the bandwidth utilization needs to be optimized, for example when accessing files over Wide Area Network.

Another interesting feature of both Lustre and GPFS is that they are completely POSIX compliant and this could be important as the end users of the farm can easily access all the files as they are on their personal desktop. Indeed, exploiting this feature, it is possible to implement on the same file system both the scientific data storage and the user’s home directories. It gives the opportunity to consolidate all the storage services on a single solution, reducing the human effort needed to maintain the service.

After the tests we decided to migrate the production storage infrastructure on Lustre parallel file-system. Indeed, we have chosen Lustre instead of GPFS as it is open source and could be used for free, giving the possibility to keep a low TCO.

We have also executed a large number of tests in order to find out the best configuration of the read-ahead for each of the tested storage systems, to understand which value can better fit the main access pattern that typically an LHC Tier2 should serve: the HEP (High Energy Physics) data analysis. These activities led us to change the default value of 20MB of read-ahead on Lustre to 2MB, to serve efficiently data to the CPUs without wasting too much bandwidth on data that will not be used. Using this configuration it is possible to deliver a great amount of data to the available worker nodes, as it is shown in the Figure 1. Nowadays the INFN-Bari farm is composed by 3700 CPU/cores and about 1.3 Pbyte of storage, and when a large number of analysis job are submitted, it is easy to deliver more than 4 GByte/s of aggregate bandwidth among the storage and the worker nodes.
2.3. Monitoring of data access over Lustre distributed file-system

One of the main issues with the Lustre storage technology is the absence of a detailed monitoring system, to help the site admins to understand which files are used by which user(s) or node(s). This possibility is quite important in order to achieve a better monitoring of the misuse of the farm, but also for debugging purposes.

In order to solve this issue, the INFN-Bari group has developed a monitoring system that is able to look at the open file on each client node and report to a central database several information, like:

- The date and time when the file is seen as opened;
- The name of the users that is using the file;
- The name and the number of the process that is opening the file;
- The type of activity carried on the files (reading of writing);
- The name of the client host where the file is opened.

This allows the site admins to a close control of the activity of the cluster and to provide figures of accounting on the data access based on the path of the files.

2.4. Optimization of job scheduling for interactive usage

The HEP Tier2 computing centres are not only supporting grid analysis activities, but they should also support the local work of the end users, that do need an interactive facility to develop the code, analyse the data and build the final plots.

Usually the solution for this problem is to provide a number of frontend machines where the users can login and execute all the interactive activities, similarly to what happens on their own desktop. The main disadvantage of this approach is that it is quite difficult to provide a well-balanced hardware infrastructure for user login. Indeed, usually the solution is to build an ad-hoc cluster to be used for these use-cases. The load is distributed using low-level technologies, like DNS-alias (with round-robin login distribution).

Some machines could be overloaded, while others under-utilized. It could also happen that when the users are less active, the cluster is under-utilized leading to a waste of resources. On the contrary as soon as the number of the users or their number increases, the cluster could be overloaded, and it will not be easy to scale up the involved resources.
In order to find a solution to this problem we proposed and tested a new paradigm. The solution is to use “interactive jobs”. This is a feature already available on many batch systems, like Torque [10], LSF [11] and others. It allows the user to submit a job to the local batch farm and to obtain an interactive shell instead of executing a batch job.

All the CPU intensive applications can be executed on a dedicated CPU on a WN that is found among the whole Tier2 farm. In this way it is enough to have one or two frontend machine, for the simple login of the users, while all the other CPU intensive but interactive activities could be moved to the whole farm, giving the solution for dynamically scale up (or down) as soon as the load of the users increase or decrease.

Using this solution, at the moment on the INFN-Bari farm it is possible to support up to 180 local users without providing huge interactive cluster.

3. Activities in Pisa

As a CMS Tier2, the Pisa centre guarantees CPU cores and storage for the experiment’s needs.

In particular, Pisa hosts:

- For CMS Monte Carlo production activities: about 1000 cores and 50 TB of temporary storage space;
- For CMS data analysis via GRID: about 1000 cores and a data area of about 800 TB;
- Support for local CMS analysis: 500 cores to be shared by about 50 users, and an additional 100 TB of local space;
- Support for CMS interactive activities: a set of load balanced interactive machines, with a few cores per each of the potential (50) users.
- Additional about 3000 cores and 200 TB for other users; they can be used by CMS if free.

The Centre uses LSF as main tool to balance the ~5000 cores present, and guarantee the different users a correct fair share, and different storage technologies to give access to the available about 1 PB.

The storage system in Pisa is currently transitioning from dCache[4] to Storm[12]. At the heart of both systems sits a GPFS storage system, based on two DataDirect DDN9900A [13] SANs. The GPFS system serves also local INFN needs, and exceeds 1.2 PB of installed disks.

A challenge we are currently facing is about how to realize a transition without sizeable site downtimes, not acceptable in an important period of the LHC data taking. We prepared a transition plan in which only metadata are moved between the two systems, which should result in a total downtime for the site of less than one hour. The plan is going to be part of the final report the project wants to prepare, since many smaller sized sites in Italy are preparing to use Storm in the near future.

A summary of the activities going on in the Pisa institute is described in the following.

- Storage systems: a procedure on how to switch from dCache to Storm without sizeable downtime (1 hour for 1 PB data) was prepared, tested and deployed.
- Networking tests on how to mix 1Gbps, 10Gbps and InfiniBand were performed, resulting in an hybrid system were all the network traffic follows the best possible route.
- Monitoring: tools to monitor the site were written from scratch. They include
  - LSFMON [14]: which monitors the current and historical status of multiple complex clusters;
• dCacheTools [15]: used to monitor and help in debugging a dCache system. It consists of a high level Perl API (dCacheTools), which allows automated operation and recovery.

• Local user activities: a special effort has been put in optimizing the user analysis activities. This includes:
  • Load balanced on demand GRID User Interfaces served via LSF, with guaranteed services;
  • A shared disk area of PB size common to all the nodes;
  • MPI enabled and PROOF [16] enabled machines;
  • Machines with many cores made available to a single user, to allow for application level multi threading.

• National level Authentication and Authorization: since the site hosts national facilities for the CMS Experiment and for Italian theoretical physicists, we wanted to allow an easy access to the site to all the involved parties. Access via INFN-AAI is already in place for theoretical physicists, and will be available soon for the whole Italian CMS community. A first test on a real use case was used to create student accounts for the First CMS Data Analysis School[17] held in Europe, in January 2012.

• Distribution of data via WAN protocols: the CMS Pisa Tier2 is currently the biggest at Italian level, and hosts a lot of data interesting to users. We are offering our data via remote streaming, using the Xrootd protocol, via the redirector implemented in Bari.

More information and HowTos are available at the site WEB page[18].

4. Activities in Trieste

INFN-TRIESTE activities were focused on two distinct aspects: client side studies on remote data access in HEP analysis and improvement of user support.

On the aspect of the client side remote data access, real analysis tests in HEP environment (CMS 2011 7 TeV data) were performed, using as targets: local disk (for reference), remote dCache systems (like INFN-PISA), and remote Storm systems (like INFN-BARI). In the case of interactive analysis, the performance drop due to the data non-locality was of the order of 10%; we evaluated that increase in ease of analysis largely outweighs the lost performance. A modified CRAB client (the tool which allows for GRID submission of analyses in CMS) is going to be used to test also the non-interactive analysis part. This is going to happen by not forcing the jobs to land where data is, but by allowing them to be run where CPUs are available, with data accessed remotely.

On the aspect of user support improvements, the activities were focused on preparing user support infrastructures. Specifically for the use case of analysis in CMS, user support has been followed via a shared Gmail account, via CMS HyperNews system, and via the Savannah system. Just to cite a number, the HyperNews forum has seen 6000 threads opened in the last two years, for a total number of messages at least three times bigger.

5. Activities in Milano
INFN-Milano-Bicocca activities were focused on the distributed analysis aspect in HEP analysis activities, for the CMS experiment in particular. CMS uses since the start of the p-p collisions in LHC (late 2009) CRAB [19] as main tool for submitting analysis jobs to the GRID. The version currently operational is CRAB2, which consists of

- A client to be installed on the user’s User Interface;
- A server which takes care of interaction with the GRID (OSG [20] and gLite [21] flavors being supported).

The client supports also server-less operations, with direct interaction to the GRID; the benefits introduced by the server are

- Speed: the user experiences much faster commands, since the jobs are sent to the GRID in asynchronous mode;
- Automatic smart resubmission of jobs in case of failure (for example, excluding sites with excessive failure rate).

The group in Milano-Bicocca focused on one side on the maintenance of CRAB2, introducing bug fixes and new features as requested by the CMS Experiment (like special job splitting features). On the other, Milano was part of the development of the successor CRAB3, completely rewritten to use a common interface with the other tool in CMS submitting jobs to the GRID (the one used for Monte Carlo processing and data reprocessing, WMAgent).

CRAB3 is now a thin client, which does not store any information locally (stateless), but communicates with the WMAgent infrastructure. CRAB3 is supposed to enter into production phase later this year, after some months of extensive testing from expert users.

6. Conclusions

The activities carried out by the institutions involved in the project have been designed having user analysis activities in mind. All the aspects important to increase user satisfaction level in analysis have been considering, studying and deploying solutions to improve the experience.

This has included:

- Networking optimizations
- Storage optimizations
- Study on diverse hardware components
- Optimization to the facilities’ access
- Improvement in user support

The project result will be made public in the form of a report in late Summer 2012, at the project natural conclusion.

The current work has been partially developed under PRIN 2008MHENNA MIUR Project.

7. References

[1] http://lhc.web.cern.ch/lhc/
[2] The CMS experiment at the LHC, J. Instrum. 3, S08004, (doi:10.1088/1748-0221/3/08/S08004)
[3] CMS: The Computing Project Technical Design Report, CERN-LHCC-2005-023, CMS TDR 7, 20 June 2005.
[4] http://www.dcache.org/
[5] A. Dorigo, P. Elmer, F. Furano, and A. Hanushevsky. Xrootd - A highly scalable architecture for data access. WSEAS Transactions on Computers, Apr. 2005.

[6] http://www.lustre.org
[7] http://www-03.ibm.com/systems/software/gpfs/
[8] http://www.iozone.org/
[9] http://www.coker.com.au/bonnie++/
[10] http://www.adaptivecomputing.com/products/open-source/torque/
[11] http://www.platform.com/workload-management/high-performance-computing
[12] http://storm.forge.cnaf.infn.it/
[13] http://www.ddn.com/products/s2a9900
[14] http://sarkar.web.cern.ch/sarkar/doc/lsfmon.html
[15] http://sarkar.web.cern.ch/sarkar/doc/dcachetools.html

[16] Data Analysis with PROOF, XII Advanced Computing and Analysis Techniques in Physics Research, November 3-7 2008, Erice, Italy

[17] http://lpc.fnal.gov/programs/schools-workshops/cms-das-2012_entry.html
[18] http://www.pi.infn.it/tier2/

[19] CRAB: A CMS application for distributed analysis. IEEE Trans.Nucl.Sci. 56 (2009) 2850-2858

[20] The Open Science Grid. Journal of Physics: Conference Series, 78 15, Pordes R, Petravick D, Kramer B, Olson D, Livny M, Roy A, Avery P, Blackburn K, Wenaus T, Wurthwein F, Foster I, Gardner R, Wilde M, Blatecky A, McGee, J, and Quick R. 2007.

[21] http://glite.cern.ch/