Direct exploitation of a top 500 Supercomputer for Analysis of CMS Data

I Cabrillo\textsuperscript{1}, L Cabellos\textsuperscript{1}, J Marco\textsuperscript{1}, J Fernandez\textsuperscript{2} and I Gonzalez\textsuperscript{2}

\textsuperscript{1} Instituto de Fisica de Cantabria (IFCA), Avda. de los castros s/n, Santander, Spain
\textsuperscript{2} Dep. Physics, University of Oviedo (UNIOVI), Calvo Sotelo s/n, Oviedo, Spain
E-mail: cabrillo@ifca.unican.es

Abstract. The Altamira Supercomputer hosted at the Instituto de Fisica de Cantabria (IFCA) entered in operation in summer 2012. Its last generation FDR Infiniband network used (for message passing) in parallel jobs, supports the connection to General Parallel File System (GPFS) servers, enabling an efficient simultaneous processing of multiple data demanding jobs. Sharing a common GPFS system and a single LDAP-based identification with the existing Grid clusters at IFCA allows CMS researchers to exploit the large instantaneous capacity of this supercomputer to execute analysis jobs. The detailed experience describing this opportunistic use for skimming and final analysis of CMS 2012 data for a specific physics channel, resulting in an order of magnitude reduction of the waiting time, is presented.

1. Introduction

By 2012 CMS\textsuperscript{[1]} researchers at IFCA and Universidad de Oviedo were using Grid-based Tier-2 resources for skimming jobs and final data analysis in leptonic channels, including WW, Higgs and top studies. Despite the significant volume of resources available at the local Tier-2 center, execution of multiple skimming jobs requiring simultaneous access to data was found to encounter problems, mainly due to the high demands on I/O posed to the computing cluster and the GPFS\textsuperscript{[2]} file system.

Researchers however were constrained in time due to the need to deliver results to the CMS collaboration in time for submission to summer/winter conferences, and could not afford a scaled submission of jobs that could delay one month or more the production of n-tuples used in the final stage of physics analysis.

At the same time, the University of Cantabria acquired a new and powerful supercomputer, named Altamira, oriented to basic and applied research, and designed to efficiently support large data processing.

This paper describes the integration of this large High Performance Computing (HPC) resource in a framework for CMS researchers at IFCA in an efficient way: minimizing the effort for system managers and with a negligible learning curve for these final users. This integration framework is described below.

2. Computing Framework at the Institute of Physics of Cantabria (IFCA)

IFCA is a joint research center of the University of Cantabria (UC) and the National Research Council in Spain (CSIC). It focuses on basic research areas: High Energy Physics, Astrophysics,
Statistical Physics and Distributed Computing.

IFCA operates a Tier-2 CMS center in the Worldwide LHC Computing Grid (WLCG)[3], several clusters integrated in the National Grid Initiative (NGI)[4] and in the Federated Cloud within EGI, with a CPU capacity of more than 3,600 cores.

Data Storage is managed through a high performance system using the General Parallel File System (GPFS) with a total raw capacity of more than 2 Petabytes.

All these systems share an internal 10 Gb/s Ethernet network backbone, and are connected using dark fiber to the RedIris (Spanish National Research network) regional access point, integrated in the Spanish dark fiber backbone, RedIris Nova, directly linked to the GEANT[5] network.

Researchers at IFCA use local accounts managed by LDAP[6], and get access to individual and group areas in the global GPFS area. Batch systems are based on SGE[7], and Grid users submit jobs to these systems and to any other Grid site using scripts and individual certificates. Data access is provided by the Storage to Resource Manager (StoRM)[8] system, an interface between GPFS and SRM[9].

3. Altamira Supercomputer

Altamira is a general purpose supercomputer, integrating a cluster based on Intel servers using the latest generation Infiniband network. Altamira was designed at IFCA in collaboration with IBM and BSC (Barcelona Supercomputing Center). It includes 240 iDataplex dx360m4 nodes, each having two SandyBridge E5-2670 2.6 Hz/1600 20 MB cache Intel processors, 64 GB RAM, 500 GB SATA II HDD. It also integrates seven iDataplex dx360m3 nodes, each one with two nVidia Tesla M2090 GPUs, and eleven ps702 IBM Power7 servers, to maintain compatibility with the previous supercomputing node that was based on JS20 servers. A picture showing the complete system installed at IFCA is shown in figure 1.

![Figure 1. Altamira Supercomputer installed at IFCA.](image)

The connection between nodes uses the latest Infiniband generation FDR, provided by Mellanox, which allows very low latency between nodes (less than 1 microsecond), and very high throughput (40 Gb/s). It is configured in a FAT tree topology, providing a non-blocking network architecture.

Thanks to the excellent performance of this network and the use of the latest Intel processors, Altamira reached over 80 TFlops and entered the Top 500 list[10] in June 2012 in position 358.

Later in 2012 the Altamira cluster was reconfigured and the cluster currently hosted at IFCA includes 158 dx360m4 and 5 dx360m3 nodes, with a total power exceeding 50 TFlops.
Altamira users include researchers in national and international projects of the University of Cantabria, and researchers applying to the Red Española de Supercomputación (RES), accessing through ssh to a login node using username/password, and submitting parallel jobs to the SLURM[11] resource manager.

4. Integration in the Data Center

As stated, the IFCA Data Center hosts both Grid systems and the Altamira Supercomputer, but manages them separately. On one side, the information required by CMS researchers as input to the analysis is stored under GPFS in CMS Tier-2. On the other, Altamira supercomputing nodes can handle in parallel a large number of jobs (up to 2,500) with demanding data needs, thanks to the availability of Infiniband at each node.

The solution proposed was to integrate the authentication system using the same LDAP service, and employ it also for authentication in GPFS, providing a common home area to users and the possibility to easily manage data under a common GPFS path from both systems.

This integration required the use of the Ethernet network in Altamira nodes to access GPFS data in the Tier-2, while Infiniband high performance network could be used in Altamira nodes when required by the execution of a large number of jobs at the same time.

The details of this integration are given in the next sections.

5. Integration Details at Hardware and Software Level

GPFS requires an IP network to maintain cluster integrity, so the first and mandatory step was to deploy an Ethernet network in all Altamira nodes. A general description of this hardware integration is described in the scheme shown in figure 2.

![Figure 2. Hardware layout at IFCA Data Center.](image-url)
All Altamira nodes were connected to the IFCA Ethernet layer using a 1 Gb/s network interface to a top of the rack Ethernet switch, plus a network trunk of 10 Gb/s + 10 Gb/s to the IFCA network backbone. This network connection is mainly used to maintain the integrity on GPFS cluster, but it can also be used for data transfer between GPFS areas as will be shown. The scheme is depicted in figure 3.

![Figure 3. Altamira connection at IFCA Ethernet backbone.](image)

Four GPFS servers manage Altamira storage: they are directly connected by 8 Gb/s Fiber Channel to an IBM DCS3700 enclosure with a raw capacity of 1 Petabyte, and by Infiniband FDR to all Altamira nodes.

The Infiniband network uses the Remote Direct Memory Access (RDMA)[12] mechanism, which supports zero-copy networking, enabling the network adapter to transfer data directly from or to application memory, eliminating in this way the need to copy data between application memory and data buffers in the operating system.

GPFS is able to leverage the improvements that involve the use of RDMA, but this is only applicable for versions newer than 3.4. We have used the 3.5.0 update 2 for this configuration.

Once the Infiniband network is up (this can be checked using ibstat and ibstatus commands) the open device has to be captured so GPFS knows the interface to be used for data transfer. In order to use native IB support with GPFS instead of IP over IB, the verbsRdma[13] module must be enabled and the rdma device must have the verbsPort setting.

```
# mmchconfig verbsrdma=enable,verbsrdma="dev" -N "node1,node2,...,nodeN"
```

A restart of GPFS on all the Infiniband nodes (node1,..., nodeN) is also required:

```
Loading modules from /lib/modules/2.6.32-358.14.1.el6.x86_64(extra
Module Size Used by
mmfs26 1762439 0
mmfslinux 310536 1 mmfs26
tracevol 29456 2 mmfs26,mmfslinux
```
Additionally, the GPFS servers in Altamira use 1 Gb/s Ethernet interfaces for data exchange communication with the Tier-2 GPFS space. Using the RDMA improvements for GPFS over Infiniband results in a higher performance data access compared with access over Ethernet, as can be seen from the results obtained with the gpfsperf command tool quoted in table 1.

Table 1. Data access results of gpfsperf tests using different clients and servers. Measurement unit is MBytes/s

|       | CREATE | READ   | WRITE  | READ 8TH | READ 16TH | READ 32TH |
|-------|--------|--------|--------|----------|-----------|-----------|
| Eth Server | 1200   | 920    | 1200   | 925      | 930       | 92        |
| Eth Client  | 460    | 300    | 455    | 308      | 315       | 305       |
| IB Server | 1700   | 150    | 1850   | 1170     | 1420      | 1370      |
| IB Client  | 1600   | 2290   | 1600   | 1132     | 1135      | 1135      |

Finally, to enable the Altamira cluster to execute CMS jobs, it was necessary that these nodes had access to the CMS software area. This was made possible using the CVMFS[14] file system and CMS Squid cache servers.

The Altamira batch queue is limited to 72 h of execution for a job, and these jobs typically use between 32 and 512 cores. Short jobs (below 6 h) are prioritized in order to optimize node filling. The best way to exploit this configuration for CMS skimming jobs was to submit jobs from a single script requesting more than 150 cores in parallel mode, taking advantage of Altamira’s data access through Infiniband.

In summary, the integration, as shown in figure 4, provided:

- Access to the Tier-2 file system;
- Access to the CMS software area;
- Access to the Tier-2 squid cache servers;
- Multiple job batch submission (wrapping 150 jobs).

6. User Success Story

A real application test taking advantage of the Altamira facility was performed on April 2013 for a CMS SUSY dileptonic search analysis developed by the IFCA-Oviedo physics group. It comprised skimming tasks over the CMS 2012 full data set and MC samples, and ROOT tree production taking as input the output of the mentioned skimming.

The production required a total of 250,000 CPU hours, and more than 2,000 million events were processed. In total more than 23 TB of input data samples were read: 18 TB through the Ethernet network, and about 5.6 TB with the Infiniband network. In terms of output, over 8 TB of data were written to disk using the Infiniband network: 5.6 TB in the skimming step (CMSSW data format) and 2.5 TB in ROOT tree format for final user analysis.

While the estimated processing time for this analysis using the Tier-2 resources was around two months, the production using Altamira was made in one week. In addition, the saturation in the Tier-2 SRM system, that could be expected from thousands of Grid jobs staging the 8 TB of output files, was avoided.
7. Conclusions
In summary Altamira has proved to be an ideal system for large data processing enabling an efficient extra power for analysis at peak periods.

The integration between IFCA Tier-2 resources and the Altamira Supercomputer has been possible thanks to services like LDAP and GPFS that connect the access to both systems.

More than 500,000 hours have been used already during 2013 by researchers working on CMS analysis related to both Higgs and SUSY searches and standard model cross section measurements in dilepton channels.

The performance measured for GPFS over Infiniband shows a very good data transfer rate to any node, avoiding CPU overheads, increasing the job efficiency and minimizing the failure rate from over 20% in the GRID model down to 2% or less running in Altamira.

Specific research community software (for the CMS collaboration in this case) was installed with low managerial effort, using CVMFS, and this experience can be extrapolated to other areas.

The researchers had only to introduce minor modifications in their job submission scripts and they were able to reduce their execution time by an order of magnitude, from months to weeks, or from weeks to days, resulting in delivery of physics results in time for presentation at summer/winter conferences.

References
[1] CMS collaboration at LHC, see http://cms.cern.ch/
[2] IBM 2012 General Parallel File System. Version 3 Release 5
[3] Worldwide LHC Computing Grid — WLCG, see http://wlcg.web.cern.ch/
[4] The Spanish National Grid Initiative, http://www.es-ngi.es/
[5] GEANT is the high bandwidth pan-European research and education backbone that interconnects National Research and Education Networks (NRENs) across Europe: http://www.geant.net/Pages/default.aspx
[6] The Lightweight Directory Access Protocol, LDAP, is an application protocol for accessing and maintaining distributed directory information services over an Internet Protocol network, see Koutsonikola V., Vakali A.. LDAP: framework, practices, and trends. IEEE Internet Computing, 2004, 8: 66-72

[7] Son of Grid Engine project, see https://arc.liv.ac.uk/trac/SGE

[8] StoRM, and SRM implementation for LHC, http://storm.forge.cnaf.infn.it/

[9] A. Sim and A. Shoshani, The Storage Resource Manager Interface Specification Version 2.2, http://www.ogf.org/documents/GFD.154.pdf

[10] Top 500 web site, June 2013, available at http://www.top500.org/

[11] Simple Linux Utility for Resource Management (SLURM), a free and open-source job scheduler for the Linux kernel used by many of the world’s supercomputers and computer clusters: https://computing.llnl.gov/linux/slurm/

[12] MELLANOX 2012 RDMA Aware Networks Programming. Version 1.4

[13] T Bedeir 2010 RDMA Read and Write with IB Verbs

[14] CernVM File System (CernVM-FS) is a network file system based on HTTP and optimized to deliver experiment software in a fast, scalable, and reliable way, http://cernvm.cern.ch/portal/filesystem