Iberian ATLAS Cloud response during the first LHC collisions

M Villaplana[1], G Amorós[1], G Borges[2], C Borrego[3], J Carvalho[4], M David[2], X Espinal[5], A Fernández[1], J Gomes[4], S González de la Hoz[1], M Kaci[1], A Lamas[1], J Nadal[3], M Oliveira[4], E Oliver[1], C Osuna[3], A Pacheco[3], J J Pardo[6], J del Peso[6], J Salt[1], J Sánchez[1], H Wolters[4] for the ATLAS Collaboration.

(1) Instituto de Física Corpuscular (CSIC-UVEG), Valencia, Spain
(2) Laboratório de Instrumentacao e Física Experimental de Partículas, Lisboa, Portugal
(3) Institut de Física d'Altes Energies, Universitat Autónoma de Barcelona, Spain
(4) Laboratório de Instrumentacao e Física Experimental de Partículas, Coimbra, Portugal
(5) Port d'Informació Científica, Universitat Autónoma de Barcelona, Spain
(6) Universidad Autónoma de Madrid, Madrid, Spain

Miguel.Villaplana@ific.uv.es, Jose.Salt@ific.uv.es, Santiago.Gonzalez@ific.uv.es

Abstract. The computing model of the ATLAS experiment at the LHC (Large Hadron Collider) is based on a tiered hierarchy that ranges from Tier0 (CERN) down to end-user's own resources (Tier3). According to the same computing model, the role of the Tier2s is to provide computing resources for event simulation processing and distributed data analysis. Tier3 centers, on the other hand, are the responsibility of individual institutions to define, fund, deploy and support. In this contribution we report on the operations of the ATLAS Iberian Cloud centers facing data taking and we describe some of the Tier3 facilities currently deployed at the Cloud.

1. Introduction

First beam circulated in the Large Hadron Collider on 20 November 2009. Since then, several world records have been set and a luminosity of around 10 pb^{-1} have been recorded (to September 2010). This amount of data was processed and distributed to the ATLAS Grid centres according to a predetermined plan.

The main goal of this paper is to report on the response of the Iberian ATLAS Cloud for these LHC beam events. Previously the whole ATLAS distributed system, in particular the Iberian Cloud, had been operating on Cosmic data events, which had a much lower production rate than beam data, and on Monte Carlo simulated events [1]. The approach of the report is user oriented and is organized as follows: in section 2 the operation of the Iberian Cloud during the mentioned period is given; section 3 addresses the management of data including the important issue of accessing to them, in section 4 we focus on the analysis process using the distributed Grid infrastructure also particularized to the Iberian Cloud; in section 5 a description of some of the Tier3 facilities deployed at the Cloud is given. Last section is devoted to the conclusions. An older but more extended explanation can be found in [2].

2. Operation of the Iberian ATLAS Cloud

2.1. Resources

During the Collision Event Data Analysis, the Iberian Cloud provided the hardware resources

fulfilling the ATLAS requirements of the Memorandum of Understanding (MoU), as shown in table 1.

	CPU (HS06)		DISK (TB)	
	Pledge 2010	Current	Pledge 2010	Current
IFIC	6.000	5.152	523	532
IFAE	3.000	5.000	261.5	280
UAM	3.000	3.155	261.5	307
LIP_COIMBRA	1.178	1.178	78	78
LIP_LISBON	1.261	1.261	79	79
NCG_INGRID_PT	871	871	43	43

Table 1: Hardware resources provided by the Iberian Cloud.

Disk space is managed by two distributed storage systems, namely dCache [3] at IFAE and UAM, and Lustre+StoRM [4] at IFIC, LIP_COIMBRA, LIP_LISBON and NCG_INGRID_PT. The Worker Nodes have 2 GB of RAM per CPU core to be able to run the highly demanding ATLAS production jobs.

In addition to the pure Tier2 resources, each site provides a Tier3 infrastructure for data analysis to users, which has a part based on Grid architecture and another part being a standard computing cluster. The use of the former or the latter depends on the stage of the analysis.

2.2. Operation Procedure

The use of automatic tools for system management tasks is essential. To install and configure the operating system, Grid middleware and the storage system, *QUATTOR* [5] has been the tool used. However, the ATLAS software is not installed by this tool but using a centralized procedure from the ATLAS software team, which implies to submit a Grid job.

There are two levels of monitoring: one from the global LHC Grid (the "Service Availability Monitoring" is one of its modules), and another one internal to the site, which uses tools like *Nagios* or *Cacti*.

To avoid jeopardizing the center availability, every site has a pre-production computing cluster where software updates are tested before they are put in place. Pre-production machines can be virtualised, via systems like *xen* or *openvz*, in order to economize resources.

The CPU usage must be shared fairly between Monte Carlo production and user analysis jobs. The percentage assigned to each role is configured in the scheduler (*Maui[6]*), which is used by the batch queue system (*Torque[7]*) to submit jobs according to their priority.

3. Data management of the collision events

3.1. Data model [8]

RAW data straight from the detector are reconstructed resulting in both Event Summary Data (ESD) and Analysis Object Data (AOD). The former incorporates enough information for detector performance studies, while the latter, much reduced, contains only information about physical reconstructed objects useful for physics analyses. In addition, filtered versions of ESD and AOD, derived data formats dESD and dAOD, can be created by the different physics groups to select events of their interest.

The data distribution policy states that RAW data is transferred from CERN to the 10 Tier1s keeping 3 replicas of them. Reconstructed data, ESD/AOD/dESD, is distributed to Tier1s and Tier2s. A large fraction of these data is recorded on disk in order to perform analyses for the understanding of the detector response.

Users can always submit a job to the Grid to analyze data; the job will be running at the (Tier2) site where the data are located. In addition, users can request to replicate some data to the Tier2 site they belong to, in order to access the data in a local mode. For this purpose, the Data Transfer Request web Interface (DaTRI) of the PANDA web portal was created. The DaTRI is coupled to the ATLAS Distributed Data Management (DDM) system to move the data between sites. The tool allows users to monitor the status of their requests and cloud managers to approve or reject them.

3.2. Data distribution in the space tokens (storage)

The storage in ATLAS is organized using space tokens. These space tokens are controlled through the DDM system and they are associated to a path to a Storage Element (SE). Data distribution and size in the Iberian Cloud space tokens on September 2010 is shown in the figure 1. ATLASDATADISK is dedicated to real data (cosmic and collisions). ATLASMCDISK is populated with Monte Carlo production. In addition, other space tokens are reserved for physics groups and for users.
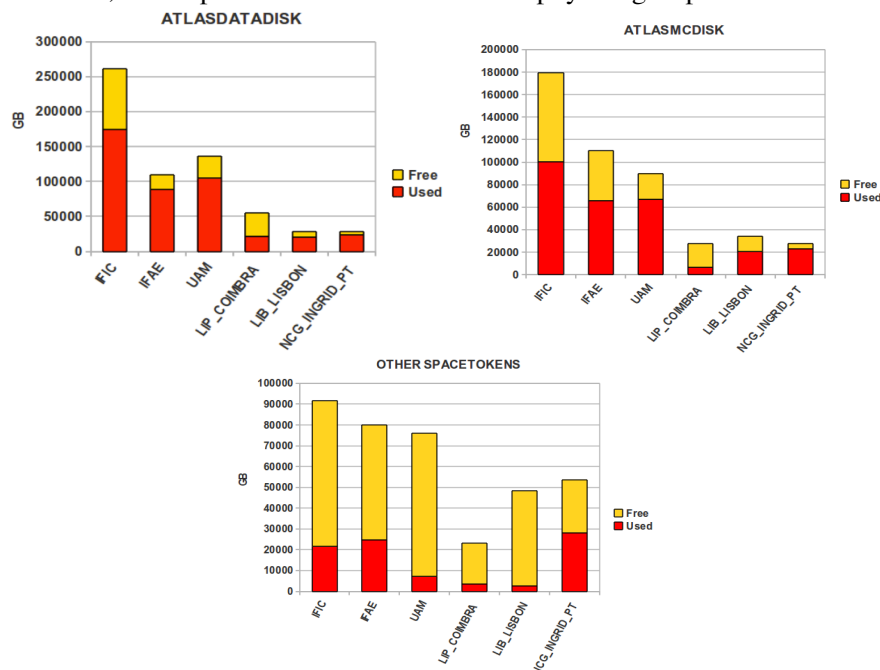

Figure 1: Status of ATLAS space tokens in the Iberian Cloud sites on September 2010.

4. Distributed Analysis

In a typical analysis the first stage is to submit a job to the Grid, which can be executed in one of the Iberian Tier2 sites if the input data are located in one of them. The output files will be stored in the corresponding Tier2 storage resources that later can be retrieved for a further analysis on the local facilities of the institute where physicists are running their analyses.

Two front-ends for ATLAS distributed analysis were used for the analysis of collision data in the Iberian Cloud: Ganga[9] and PanDA[10]. The PanDA system makes use of a central server to look for resources in the Grid and to manage the job submission. Jobs are submitted to PanDA via a python client (pathena) by whom users define job sets and the associated input datasets. The PanDA server receives the jobs defined by the user clients and places them in a global queue. The brokerage module allocates jobs to sites (prioritizing the queue based on the job type, available CPU resources on the destination site, etc.), not before it has been checked that the input data exists locally in the site to dispatch the job. Ganga is a front-end for the configuration, execution and management of computational tasks, written in python as an object oriented software for distributed analysis. Through a graphical user interface or trough scripts, the user can submit the ATLAS jobs to a large variety of

back-ends, like a local batch system, the LCG Grid, the Nordugrid or the PanDA server. Encapsulating the different technologies existing on all the possible back-ends, it allows the user to choose among them in a transparent way using a common interface as the interoperability layer.

4.1. Distributed analysis usage in Tier2 sites

Figure 2 shows the number of jobs allocated by PanDA in some of the Iberian cloud sites in a period from 2010-03-01 to 2010-10-01. Histograms contain the number of analysis jobs on collision event data. Figure 3 shows statistics of PanDA analysis jobs in the Iberian cloud sites (Tier1-PIC, Tier2-Spain and Tier2-Portugal) for the last year (October 2009 to September 2010).
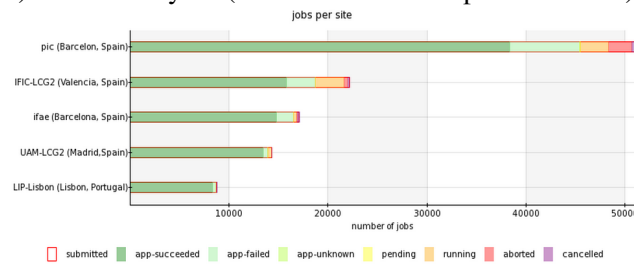

Figure 2: Number of jobs allocated by PanDA in some of the Iberian cloud sites in a period from 2010-03-01 to 2010-10-01.
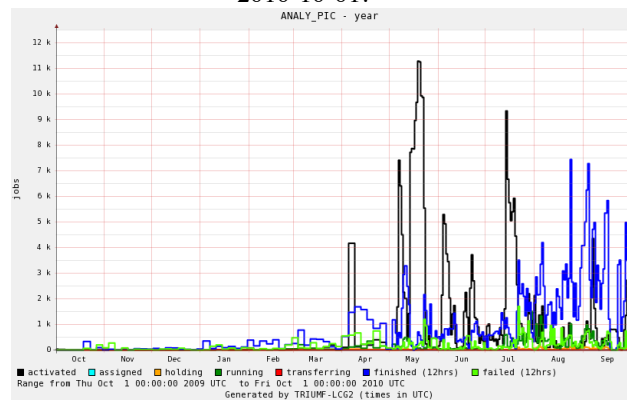

Figure 3: PanDA analysis jobs in the Iberian cloud sites (Tier1-PIC, Tier2-Spain and Tier2-Portugal) for the last year (October 2009 to September 2010).

In conformity with LHC's re-start in Spring 2010, the amount of analysis jobs arriving at the Iberian Cloud increased tremendously in April 2010, reaching peaks of more than 10000 jobs. More than 100000 jobs run at the Iberian Cloud in the period shown in figure 3.

5. Analysis at Local Facilities

To retrieve the dataset generated by the phase of the distributed analysis the user can follow two different procedures:

- Request a subscription to the ATLAS *DDM* [11] system to replicate the dataset to the LOCALGROUPDISK area. This disk space is allocated at the Tier3 facility of the institute and it is somehow connected to a Storage Element on the Grid through which the ATLAS *DDM* can perform the replication). This way the data can later be accessed locally. The replication is managed automatically by the ATLAS *DDM*.
- Use the ATLAS *DDM* client tools to download the dataset to the local disk space.

For the last phase of the analysis on n-tuples or D3PDs, every site has a local Tier3 facility separated from the Tier2 resources. These local resources are implemented using different technologies in order to create a computational facility that must be highly reliable and must have low latency. These properties are particularly important in this very last phase of the analysis due to the

high frequency of the jobs that are run on the Tier3 to obtain the final plots and results. Every site has CPU resources organized in different architectures to serve various purposes with different needs:

- Some home-built *User Interfaces* are used to perform interactive analysis on the final datasets produced in the distributed analysis phase.
- A local batch farm to provide additional computing power for analyses that need to run on local resources.
- A *Parallel ROOT Facility, PROOF* [12] farm for parallel processing of *ROOT[13]* analysis jobs.

There are 7 sites in the Iberian Cloud and there are various Tier3 setups as well, here we describe two of them. The following examples show Tier3 facilities attached to a Tier2. In both cases, Tier3's resources are split into two parts. There is a computer farm to perform interactive analysis outside the GRID framework and there are some resources that are coupled to Tier-2 in a GRID environment. To manage local users' data in the GRID environment there is a space token dedicated to Tier-3 (ATLASLOCALGROUPDISK) that has an area on a SE but points to non-pledged resources.

5.1. Tier3 at IFIC-Valencia [14, 15]. IFIC's Tier3 is attached to a Tier2 that has 50% of the Spanish Federated Tier2 as shown in figure 4. It currently has around 100 TB(60 TB under DDM control + 40 TB under IFIC control). An important feature of IFIC's Tier-3 is that it uses the same storage system as Valencia's Tier-2 (Lustre). Its central component is the Lustre file system, a shared file system for clusters. The Lustre file system is available for Linux and provides a POSIX-compilant UNIX file system interface. This interface allows users to access to the file system easily. Another important component of Lustre is the metadirectory Server (MDS), a catalog that, in IFIC's case, is the only shared resource between Tier2 and Tier3. PROOF enables interactive analysis of large sets of ROOT files in parallel on clusters of computers or many-core machines. In this case 3 disk servers are dedicated exclusively to Tier3 to avoid overlap with Tier2.
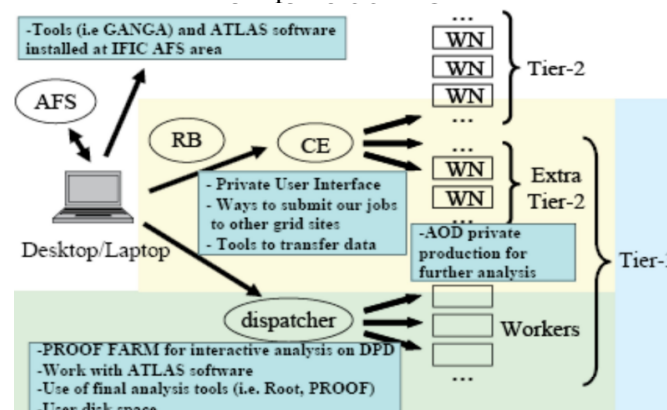

Figure 4: IFIC's Tier3 facility.

5.2. Tier3 at UAM-Madrid. Figure 5 shows UAM's Tier3 facility. It is attached to a Tier-2 as well and, in this case, it has 25% of the Spanish Federated Tier-2. It currently has 10 TB for interactive analysis, software installation, etc using NFS file system and 5 TB for LOCALGRUPDISK using dCache.
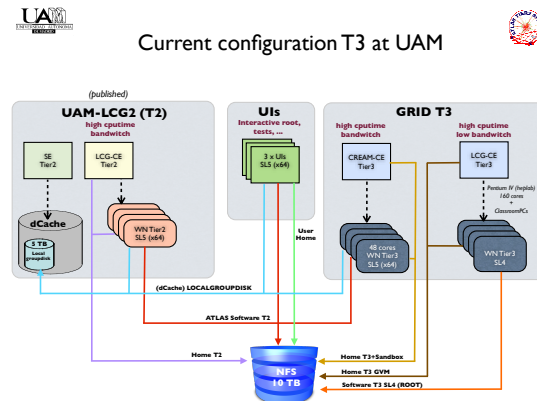
Figure 5: UAM Tier3 facility.

6. Conclusion

It has been shown in this paper that the Iberian ATLAS Cloud responded very efficiently at the different stages, from collecting data to final experimental results. The success has been made possible through:

- Receiving and storing the produced data, thanks to the high availability of its sites and the reliable services provided by the team managers.
- Providing easy and quick access to these data to users, as well as giving support to them.
- Providing the required distributed analysis tools to allow users to use the data and produce experimental results.
- Setting up infrastructures for the final analysis stages (Tier3s) managed by the Tier2 personnel.

Acknowledgements

References

[1]   "Contribution of the Iberian Grid Resources to the Production of Simulated Physics Events for the ATLAS Experiment", M.Kaci et al. ISBN 978-84-9745-549-7.
[2]   "Data analysis on the ATLAS Spanish Tier2", G. Amorós et al. ISBN 978-84-9745-549-7, pags 221-231.
[3]   http://www.dcache.org
[4]   http://wiki.lustre.org
[5]   http://www.quattor.org
[6]   www.clusterresources.com/products/maui
[7]   http://www.clusterresources.com/pages/products/torque-resource-manager.php
[8]   "The ATLAS Computing Model", D. Adams et al., ATL-SOFT-2004-007, CERN, (2004)
[9]   "Ganga: a tool for computational-task management and easy access to Grid resources", F. Brochu et al. CoRR, abs/0902.2685, 2009
[10]   "Proceedings of XII Advanced Computing and Analysis Techniques in Physics Research", P. Nilsson et al. Proceedings of Science, 2008
[11]   "Managing ATLAS data on a petabyte-scale with DQ2", M Branco et al., J. Phys.: Conf. Ser., 119 062017, 2008
[12]   "The PROOF Distributed Parallel Analysis Framework based on ROOT", M. Ballintijn et al. http://www.citebase.org/abstract?id=oai:arXiv.org: physics/0306110, 2003
[13]   http://root.cern.ch/drupal
[14]   "Analysis facility infrastructure (Tier-3) for ATLAS experiment", S. Gonzalez de la Hoz et al. Published in Eur.Phys.J.C54:691-697, 2008.
[15]   "First tests with Tier-3 facility for the ATLAS experiment at IFIC (Valencia)", M. Villaplana et al. ISBN 978-84-9745-549-7, pags 212-220.