Journal of Physics: Conference Series **513** (2014) 032082 doi:10.1088/1742-6596/513/3/032082

Lessons learned from the ATLAS performance studies of the Iberian Cloud for the first LHC running period

V Sánchez-Martínez[1], G Borges[2], C Borrego[3], J del Peso[4], M Delfino[5,6], J Gomes[2], S González de la Hoz[1], A Pacheco Pages[3,5], J Salt[1], A Sedov[3,5], M Villaplana[1] and H Wolters[7]

[1] Instituto de Física Corpuscular (IFIC), University of Valencia and CSIC, Valencia, Spain.
[2] Laboratório de Instrumentação e Física Experimental de Partículas - LIP, Lisboa, Portugal.
[3] Institut de Física d'Altes Energies, Universitat Autònoma de Barcelona, Spain.
[4] Departamento de Fisica Teorica C-15, Universidad Autonoma de Madrid, Madrid, Spain.
[5] Port d'Informació Científica (PIC), Campus UAB, Bellaterra, Spain.
[6] Departament de Física, Universitat Autònoma de Barcelona, Barcelona, Spain.
[7] Laboratório de Instrumentação e Física Experimental de Partículas, Coimbra, Portugal.

E-mail: victoria.sanchez@ific.uv.es

Abstract. In this contribution we describe the performance of the Iberian (Spain and Portugal) ATLAS cloud during the first LHC running period (March 2010-January 2013) in the context of the GRID Computing and Data Distribution Model. The evolution of the resources for CPU, disk and tape in the Iberian Tier-1 and Tier-2s is summarized. The data distribution over all ATLAS destinations is shown, focusing on the number of files transferred and the size of the data. The status and distribution of simulation and analysis jobs within the cloud are discussed. The Distributed Analysis tools used to perform physics analysis are explained as well. Cloud performance in terms of the availability and reliability of its sites is discussed. The effect of the changes in the ATLAS Computing Model on the cloud is analyzed. Finally, the readiness of the Iberian Cloud towards the first Long Shutdown (LS1) is evaluated and an outline of the foreseen actions to take in the coming years is given. The shutdown will be a good opportunity to improve and evolve the ATLAS Distributed Computing system to prepare for the future challenges of the LHC operation.

1. Introduction

Since the Large Hadron Collider (LHC) [1] started operating (the first beam circulated on November 2009) a huge amount of data has been produced. In 2012 the LHC recorded 23 fb^{-1} of luminosity at $\sqrt{s} = 8\,TeV$. This quantity of data has been recorded, processed and distributed over all the ATLAS [2] GRID [3] centres following a given policy.

In terms of the first LHC running period (Run I, March 2010-January 2013), this paper is aimed to report on the Iberian ATLAS Cloud performance and to outline its plans for the first LHC Long Shutdown (LS1).

1.1. The ATLAS Computing and Data Distribution Model

During the first LHC running period, ATLAS was using a Computing Model [4] with a tiered hierarchy based on GRID technologies that allows high degree of decentralization and the possibility to share resources. The first level is the Tier-0 at CERN; the second level comprises 10 Tier-1 centres; the third level consists of 80 Tier-2 centres distributed world wide; the last level is the Tier-3, an end-user private analysis facility. In this first period, the ATLAS Computing Model had to change in order to better satisfy the needs of the community and to adapt to technology developements.

The network is a key component in the evolution of the ATLAS Computing Model for the Tier-2, since they must be well connected to be able to exchange data. As a consequence of the strict tiered hierarchy, when a Tier-1 was in schedule downtime the job input files and data transfer of its associated Tier-2 were affected. In the new model, Tier-2s with a good network connection are allowed to link with other Tier-1s or Tier-2s belonging to different clouds. Tier-2 well connected (T2D) can directly exchange data with other Tier-1 and even with Tier-2 of different clouds.

The Iberian ATLAS Cloud is formed by one Tier-1 (PIC), two federated Tier-2s: ES-ATLAS-T2 (50% IFIC, 25% IFAE and 25% UAM) and PT-LIP-LCG-Tier2 (50% COIMBRA and 50% INGRID) and two Tier-2s not included in this report: EELA-UTFSM (Universidad Tecnica Federico Santa Maria, Chile) and EELA-UNLP (Universidad Nacional de La Plata, Argentina).

2. Evolution of the resources

During the collision event data analysis, the Iberian ATLAS Cloud provided the hardware resources fulfilling the ATLAS requirements of the Resource Review Board of the LHCC committee, as is shown in the Table 1:

Federation	CPU (HEP-SPEC06)		DISK (TB)	
	Pledges 2013	Current	Pledges 2013	Current
ES-ATLAS-T2	18000	17800	2800	2558.3
PT-LIP-LCG-Tier2	3200	3200	220	183.0
ES-PIC	16269	16269	1785	1812.0

Table 1. Hardware resources provided by the Iberian ATLAS Cloud on September 2013.

Disk space is managed by two distributed systems, namely dCache at PIC, IFAE and UAM, and Lustre+StoRM at IFIC, LIP_COIMBRA, LIP_LISBON (at the moment, it no longer exists) 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 Tier-2 resources, each site provides a Tier-3 infrastructure for data analysis, which has a part based on GRID architecture and another part a standard computing cluster. The use of the former or the latter depends on the stage of the analysis.

3. Data Distribution over Iberian Cloud sites

The storage in ATLAS is organized using SRM[5] Space Tokens [6]. These Space Tokens are controlled through the ATLAS Distributed Data Management (DDM) system [7]. They are associated to a GRID Storage Element (SE), a container for physical files, which is the smallest unit of storage space addressable within the GRID.

DDM model distinguishes between primary and secondary replicas. Primary replicas are distributed to Tier-1s for redundancy and to Tier-2 for analysis. The remaining disk space is filled with secondary replicas of popular datasets. In the first LHC running period around

36 PB of data, which include collision events and Monte Carlo (MC) simulations jobs, have been processed in the Iberian Cloud sites.

The data are transferred over the Tiers after the reprocessing and after the ATLAS official production. Figure 1 shows the transfer throughput from ATLAS sites to Iberian Cloud during the Run I. It can be seen that there is an increase data throughput around September 2011, when the changes in the Computing Model were applied. In this period, the transfer throughput from the ATLAS sites to the Iberian Cloud sites reached $550 MB/s$ in September 2012.

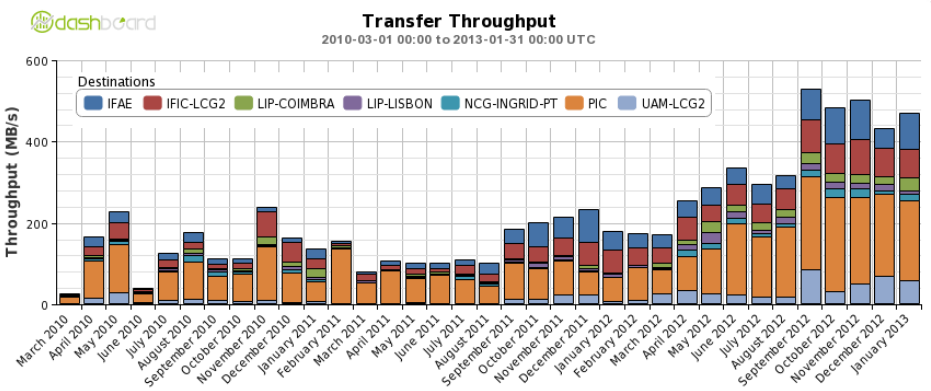

Figure 1. Throughput from the ATLAS sites to the Iberian Cloud sites.

4. Distribution of Simulation and Analysis Jobs

In order to optimize our physics output and make maximal use of available CPU and disk resources, production shares (fees) are fixed to limit "group production" jobs at Tier-1s.

The share of analysis jobs at Tier-1s has been reduced as well. The execution of simulation and analysis jobs are favoured in the Tier-2, while the reconstruction of data is favoured in the Tier-1. During the Run I, the Tier-1 has processed around 41% of the total number of jobs. The number of events processed related to these completed jobs during the Run I is around 163'000 million events.

5. Distributed Analysis Tools

The end-users are the physicists working daily on physics analyses. Typically, a physics analysis has two parts. In the first stage, physicists run an analysis program that uses a given number of collision events. These events can be stored in different datasets that are usually spread over the different sites. At this step, the Distributed Computing and Data Management Tools, based on GRID technologies, are used in an exhaustive way. The output of this first step is often a set of ROOT ntuples. In the second stage, the physicists analyze the ntuples interactively in order to get the final plots, to refine the analysis, etc.

The Distributed Analysis is using the following ATLAS tools:

(i) For Data Management:

- DQ2 (Don Quijote 2)[8]: to obtain information about data and to download and register files on GRID.
- DaTri (Data Transfer Request Interface)[9]: the end-user dataset subscription service.
- AMI (ATLAS Metadata Interface)[10]: web page for motinoring datasets, releases, number of events, etc.

(ii) For GRID Jobs:

- ganga (Gaudi/Athena and GRID Alliance)[11]: it is a job definition and submission management tool for local, batch system and the GRID.
- PanDA Client (Production and Distributed Analysis)[12]: analysis job submission tool.
  - pathena (Panda Athena): it works in the Athena runtime environment. It is a client tool to submit user-defined jobs to Distributed Analysis systems. It provides a consistent user-interface to Athena users.
  - prun (Panda Run): it is a Panda-client software which allows users to submit general jobs to Panda. It is intended to support non-Athena type analysis.
  - pbook: the next-generation of the bookkeeping application for all Panda analysis jobs.

6. Cloud performance in terms of Availability and Reliability

The site availability metrics are calculated by the Service and Availability Monitoring system (SAM)[13], which runs a range of different tests at regular intervals throughout the day. A site is considered to be available if a defined set of critical tests complete successfully.

These metrics distinguish between availability and reliability with the following definitions:

$$availability = \frac{U}{TT - TU} \qquad reliability = \frac{U}{TT - D - TU} \tag{1}$$

where U means uptime (time the site is available), TT is the total time, TU is the time status was unknow and D is the scheduled downtime.

The availability and reliability, on average, during the Run I over all Iberian Cloud sites has been always in the interval $90 - 100\%$.

Another metric to take into account is evaluated by ATLAS Computing operations. It consists of the site overall time spent online considering also the Hammer Cloud (HC) exclusions and determines the Tier-2 qualification. This is called Hammer Cloud test.The average efficiency of the HC test is greater than 90% in the last 2 years for the Iberian ATLAS Cloud sites.

7. Software and Computing for LS1

Below are shown the specific activities within the ATLAS Computing Model to be done by the Iberian ATLAS Cloud during the LS1. In addition to these tasks, the site support provided by some members of the Iberian Cloud will go on.

7.1. The deployment of the Federated ATLAS Xrootd (FAX) federation

FAX will be used to data access via a single entrance (using Xrootd's redirection tech), to read a dataset directly from WAN, to bring data to local Tier 3 Xrootd disk (storage cache) and to Users sharing non-DDM data between sites. Old untouched files will be purged when space was needed.

7.2. The contribution to the development of the EventIndex subproject

It is a complete catalogue of ATLAS events (all events, real & simulated data and all processing stages). Its contents will be, for instance: event identifiers, online trigger pattern & hit counts and references (pointers) to the events at each processing stage (RAW, ESD, AOD, NTUP) in all permanent files on storage. The main motivations for this new project were that EventTAG was designed and developed long time ago and the implementation in Oracle is an intensive labour and expensive. For that reason database technologies based on NoSQL seem well adapted to this type of application.

7.3. Operation of the data reduction framework
Some requirements are: to develop a simple mechanism to control the addition of "user data" to the new persistent format in the context of the reduction framework, to prepare the repository and operational procedures for collection and operation of reduction-framework tools, to prepare the reduction framework for inclusion of "smart slimming" and to work with ADC on the GRID integration of the data reduction framework.

8. Summary - Lessons learned
This paper has shown the Iberian ATLAS Cloud has responded very efficiently during the Run I, from collecting data to final experimental results. Some of the lessons learned about the Iberian ATLAS Cloud are:

- The change in the Computing Model has allowed to improve its performance in terms of connectivity, storage, replication, transfer, etc.
- The centres have been available around $90 - 100\%$. All its Tier-2s centres are categorized as T2D.
- All the sites have provided the resources (CPU, Disk and Tape) needed to fulfill the pledge.
- The required Distributed Analysis tools have been provided in order for the users to use/store the data and produce experimental results (i.e, the observation of a new particle in the search for the Standard Model Higgs boson).
- The LS1 is the most suitable period to improve and to evolve the tools, facilities and resources, and to make it ready for the next LHC running period.

Acknowledgement
We acknowledge to everybody who has contributed to the success of the Iberian ATLAS Cloud during the first LHC running period. Thanks for the support of Ministerio de Economía y Competitividad (Spain) and Ministério da Educação e Ciência (Portugal)

References
[1] http://public.web.cern.ch/public/en/lhc/lhc-en.html
[2] ATLAS Collaboration,"ATLAS detector and physics performance: TDR 1" 1999.
[3] http://public.web.cern.ch/public/en/lhc/Computing-en.html
[4] Adams D, Barberis D, Bee C, Hawkings R, Jarp S, Jones1 R, Malon D, Poggioli L, Poulard G, Quarrie D and Wenaus T, 2005, "The atlas computing model", CERN- LHCC-2004-037/G-085.
[5] https://sdm.lbl.gov/srm-wg/doc/SRM.v2.2.html
[6] "The atlas distributed data management project: Past and future", V. Garonne, G. A. Stewart, M. Lassnig, A. Molfetas, M. Barisits, T. Beermann, A. Nairz, L. Goossens, F. B. Megino, C. Serfon, D. Oleynik, and A. Petrosyan, CHEP 2012, vol. 396, p. 032045, 2012. 39
[7] "Managing ATLAS data on a petabyte-scale with DQ2", M Branco et al 2008 J. Phys.: Conf. Ser. 119 062017.
[8] "Managing Very-Large Distributed Datasets", Branco M, Zaluska E, de Roure D, Salgado P, Garonne V, Lassnig M and Rocha R, Proceedings of the OTM 2008 Confederated International Conferences. 775 - 92
[9] "DaTRI Notes v.0.8.5"
[10] https://twiki.cern.ch/twiki/bin/viewauth/AtlasComputing/AtlasMetadataInterface
[11] "Ganga: a tool for computational-task management and easy access to GRID resources", F. Brochu et al. CoRR, abs/0902.2685, 2009
[12] "The PanDA System in the ATLAS Experiment", Nilsson P, Caballero J, De K, Maeno T, Potekhin M, Wenaus T, XII Advanced Computing and Analysis Techniques in Physics Research 2008
[13] http://atlas.fis.utfsm.cl/atlas/adcos/AdcMonitoring.htm#Service_and_Availability_Monitor