Computing activities at the Spanish Tier-1 & Tier-2s for the ATLAS Experiment in the LHC Run-3 period and towards High Luminosity

Pablo Collado Soto[4,*], *Esther* Acción García[1,2], *Vanesa* Acin[1,2], *Carles* Acosta-Silva[1,2], *Helena* Burriel Navarro[3], *Josu* Cantero[3], *Jose* Flix Molina[5], *Jose Enrique* García Navarro[3], *Santiago* González de la Hoz[3], *Andreu* Pacheco Pages[2], *Jose* del Peso[4], *Elena* Planas Teruel[1,2], *Jose* Salt[3], *Javier* Sánchez Martínez[3], *Marc* Santamaría Riba[6], *Emma* Torro Pastor[3], *Roberto* Uzum[3], and *Miguel* Villaplana Pérez[3] on behalf of the ATLAS Computing Activity

[1]Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology, Campus UAB, Bellatera (Barcelona), Spain
[2]Port d'Informació Científica (PIC), Campus UAB, Bellaterra (Barcelona), Spain
[3]Institut de Física Corpuscular (IFIC), University of Valencia and CSIC, Valencia, Spain
[4]Departamento de Física Teórica y CIAFF, Universidad Autónoma de Madrid, Spain
[5]Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), Madrid, Spain
[6]Universidad Autónoma de Barcelona (UAB), Barcelona, Spain

Abstract. This contribution showcases the Spanish Tier-1 and Tier-2s' contribution to the computing of the ATLAS experiment at the LHC during the Run-3 period. The Tier-1 and Tier-2 Grid infrastructures, encompassing data storage, processing, and involvement in software development and computing tasks for the experiment, will undergo updates to enhance efficiency and visibility within the experiment. Central to our efforts is to engage actively with the various challenges inherent in research and development, in preparation for the upcoming, more intricate phase represented by the High-Luminosity LHC (HL-LHC). In tackling these issues, we capitalize on National High Performance Computers like MareNostrum, part of the Supercomputing Spanish Network. A new activity in this work is the development and implementation of what we call the "Facility for Interactive Distributed Analysis". This initiative aims to facilitate data analysis work for physicists at Spanish centres (IFIC, UAM, and IFAE) by orchestrating the distributed nature of initial analysis phases with subsequent interactive phases involving reduced data files. The ultimate goal is to reduce the time to produce publishable physics results or contributions tailored for workshops and conferences. The ATLAS Tier-1 and Tier-2 sites in Spain have contributed and will continue to contribute significantly to research and development in computing.

ATL-SOFT-PROC-2025-005
09 January 2025

1 Introduction

In the original computing model of the ATLAS experiment [1] at the LHC at CERN, Tier-1 centres had associated Tier-2 centres, which were close-by in terms of network connectivity,

*e-mail: pablo.collado@uam.es

and they formed a cloud, which in ATLAS means a regional setup of one Tier-1 and its associated Tier-2s in a certain geographical area. All data flowed to and from Tier-2s via their associated Tier-1s.

However, an increasing number of Tier-2 sites gained exceptional global network connectivity and the ability to directly exchange data among themselves. As a result of these advancements, the ATLAS computing model underwent a transformation to a mesh model, which includes Nucleus sites and Satellite sites [2]. Tier-2s with a significant storage and excellent network connections are designated as "Nucleus" sites, passing job production on to smaller Tier-2s (Satellites) in any cloud, exchanging data directly. Currently, all the Spanish ATLAS Tier-2 sites are of the Nucleus type.

2 Spanish ATLAS Tier-1 and Tier-2s

From ATLAS' point of view, the Spanish Cloud (ES) [3] contributes to the ATLAS experiment with a Tier-1 and a Tier-2 site. The Tier-1 site is located at PIC in Barcelona and is co-located with IFAE. Aside from its contribution to ATLAS, PIC also participates in the computing activities of two other LHC experiments, CMS and LHCb. The Tier-2 is federated across three different locations: IFIC in Valencia, IFAE in Barcelona and UAM in Madrid. Both the Tier-1 and Tier-2 represent 4% of the entirety of ATLAS's computing in their respective Tier classes, and all their resources are tightly integrated in the World Wide LHC Computing Grid (WLCG) project [4] whilst strictly following the ATLAS computing model.

In Table 1 the Spanish resources in September 2024, together with their availability and reliability, are shown. The ES cloud is at the top of availability and reliability ranks.

Table 1. Resources and performance of the Spanish cloud facilities for the ATLAS experiment in September 2024.

Site	CPU (kHS23)	Disk (PB)	TAPE (PB)	Availability (09/2024)	Reliability (09/2024)
PIC	60.6	6.52	18.08	100%	100%
IFIC	44	3.41	-	99.8%	99%
IFAE	17.5	1.60	-	100%	100%
UAM	16.6	1.08	-	98%	97%

In terms of software stacks, the different physical sites deploy and operate similar tools and services. Storage Elements (SE) are based both on dCache [5] and Lustre [6] plus StoRM [7]. Computing Elements (CE) are also varied with both HTCondor [8] and ARC [9] plus SLURM [10] currently in operation. Leveraging different combinations of services allows the staff operating the sites to become familiar with different solutions, thus opening up a wider range of options for tackling the challenges set forth by the HL-LHC.

3 Grid and HPC ATLAS Tier-1 and Tier-2 jobs

The Tier-1 and the Tier-2 have been built and are maintained to operate as ATLAS Grid computing sites. They can execute both data analysis and Monte Carlo event generation for the ATLAS experiment. As "Nucleus" sites we are stable enough to guarantee data processing and analysis. The jobs completed between January and September 2024 at the Spanish Cloud are shown in Figure 1. More than 5 million jobs were completed with Grid resources alone and an additional 0.8 million jobs were completed by HPC resources. Figure 1 also shows

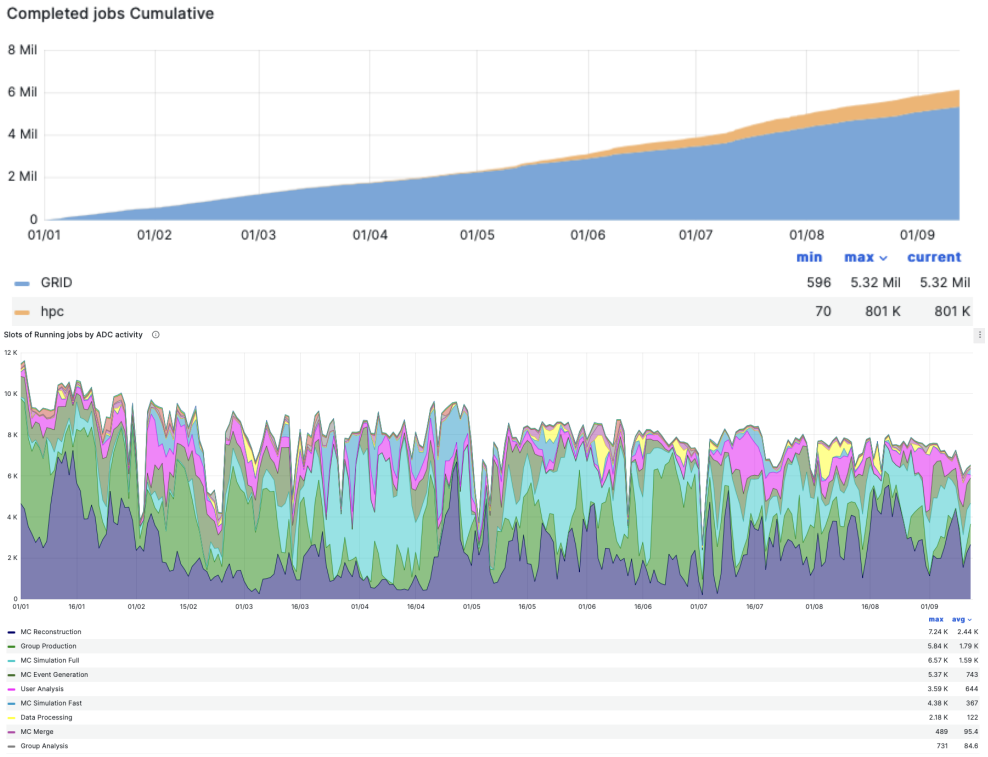

Figure 1. Top: More than 5.5 million jobs (analysis and production) completed at the ATLAS Spanish sites since the beginning of 2024 until September 2024. Bottom: Slots running on Grid resources alone (i.e. with no HPC contribution).

the amount of concurrently running jobs in Spanish Grid Resources: the HPC contribution is excluded given only simulation jobs are run in these high performance system.

The main HPC centre in Spain, the Barcelona Supercomputing Center (BSC), which is integrated into the Spanish supercomputing network (RES) [11], included LHC computing in its strategic projects list in 2021. The BSC updated their flagship HPC cluster, MareNostrum 4, to MareNostrum 5 in April 2024. Given the Spanish Cloud's long-standing expertise [12], we were capable of exploiting this improved cluster with minimal delay from the three WLCG sites that provide computing resources to ATLAS, located in Madrid (UAM), Valencia (IFIC), and Barcelona (IFAE). An instance of the ARC-CE [9] has been deployed in each of the three centres. MareNostrum 5, like its predecessor, only accepts job submissions and data transfers over SSH given its lack of outbound connectivity. HPC resources are only being leveraged for running simulation jobs where the workflow has been validated using Singularity containers whose images are pre-placed on MareNostrum 5's shared GPFS filesystem. Jobs sent from LHC resource explotation centres to the ARC CEs deployed on the different sites are transformed jobs suitable for running on MareNostrum's air-gapped infrastructure. Resource management in the context of MareNostrum leverages the SLURM job scheduling system (see Figure 2). Once jobs finish, the ARC CEs are in charge of copying the results back into the ATLAS computing system, recording them exactly as if the job had run on regular Grid resources. It is paramount to note how jobs running within MareNostrum 5 cannot make any outbound connections given aforementioned lack of connectivity.

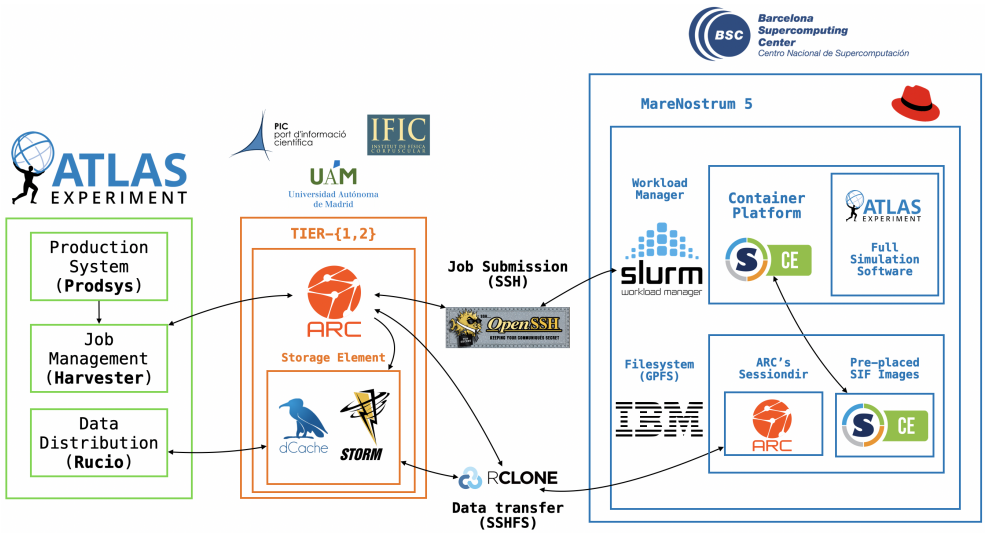

Figure 2. Flow diagram between the ATLAS production system and the execution of jobs on the MareNostrum 5 supercomputer.

Since 2020 we have an agreement between the Spanish Ministry of Science and the BSC to have around 30 million hours approved at MareNostrum 4 (and now 5) every year by ATLAS through Spanish gateways, which corresponds to 50% of the simulation jobs assigned to Spain. Note how among all the different types of computing jobs, the MareNostrum 4 and 5 HPCs contribute only to the simulation effort. This explains how the total contribution by HPCs as part of the ES cloud hovers around 30%: they only handle 50% of the simulation effort which, in turn, is a fraction of the overall computing effort carried out by the ES cloud. As stated before, during the beginning of the Run 3 and until September 2024, the MareNostrum 4 and 5 HPCs provided around 30% of the total contribution to ATLAS computing by the Spanish cloud (see Figure 3). In 2024 we have observed a slight increase (from 30% to 38%) given the inauguration of the more performant MareNostrum 5 HPC.

4 ATLAS Event Index

The Event Index project [13, 14], aims to provide a catalogue of real and simulated data of all events in all processing stages, which is needed to meet multiple use cases and search criteria. Billions of events (the total size of all events is on the order of PetaBytes) have been indexed so far since 2015. Some of the use cases are: event picking, duplicated event checking, overlap (construct the overlap matrix identifying common events across different files), trigger check (count or give an event list based on trigger selection), event skimming, and trigger counter.

The latest developments are aimed at optimizing storage and operational resources in order to accommodate the larger amount of data produced by ATLAS in Run 3 (35 billion new real event records each year). This is expected to increase in the future with a prediction of 100 billion events per year at the HL-LHC. At IFIC we have improved the data collection system [15], and the new storage schemas using HBase/Apache Phoenix for the final data backend. Both the chosen software stack and the extract-validate-load data lifecycle are clear examples of strategies applied in the industry when dealing with data-heavy applications

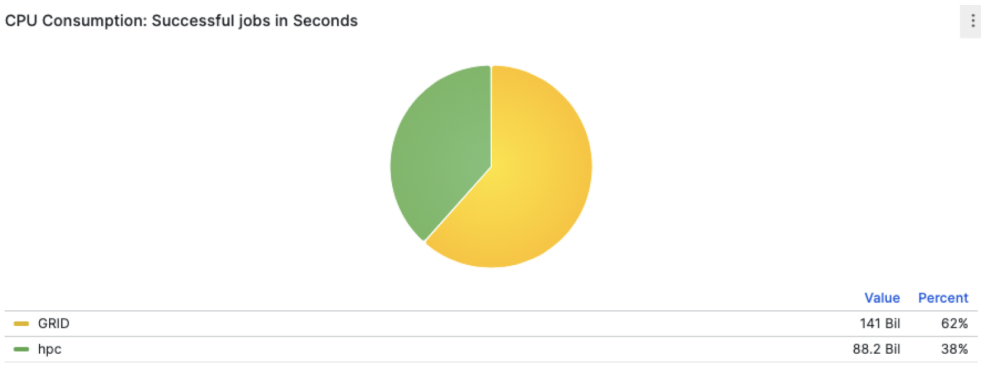

Figure 3. Proportion of HS23(s) provided by Grid resources (yellow) and the MareNostrum 4 and 5 HPCs (green) of the total contribution to ATLAS computing by the Spanish cloud in 2024. The migration from MN4 to MN5 explains the slight increase of the HPC share with respect to previous years.

such as this one. This new system has been in operation since Spring 2022 and performing excellently with promising results.

5 Data Management

In the recent LHC running periods (Run 2 and Run 3) more than 10 PB of data have been stored in the Spanish facilities as shown in Figure 4. This figure only shows data stored on disk: an additional 18.6 PB are stored on tape at PIC.

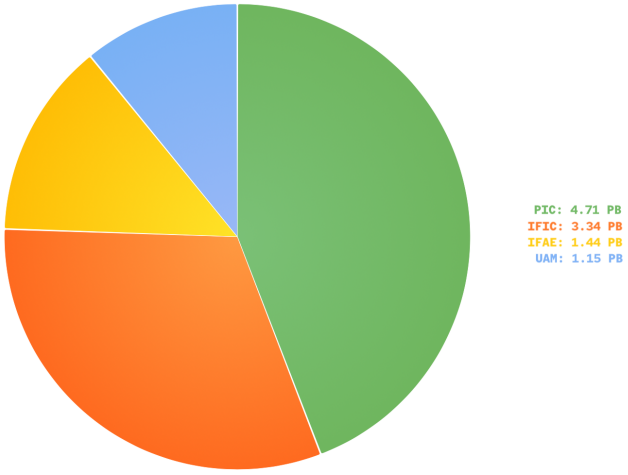

Figure 4. Current disk usage in Spanish sites.

To enhance Data Organization, Management, and Access (DOMA) in the Spanish cloud, a transition is underway towards a network-centric model known as the Datalake model [16]. This model aims to streamline operations by reducing the number of facilities responsible for storage services, minimizing data replication, and eliminating the requirement for CPUs

and storage to be physically co-located. Achieving this objective entails delivering content over a high-speed (minimum 100 Gbps) and reliable WAN, as well as implementing caching mechanisms. By ensuring a robust and efficient network infrastructure, the feasibility of establishing a Datalake is significantly enhanced. Achieving these conditions hinges on the support provided by Spain's National Research and Education Network (NREN), RedIRIS. Currently, PIC, IFAE and IFIC all have 100 Gbps links dedicated to LHC traffic whilst UAM's update to that same rate is imminent. The outlook for these data rates is nothing short of bright: PIC expects to attain a 500 Gbps rate by 2026 and RedIRIS is rolling out 400 Gbps trunks in their backbone that will be ready for the HL-LHC era.

The R&D efforts carried out at the Spanish sites for this transition are focused on minimizing the cost of storage infrastructure, eliminating data duplication, adopting smaller data formats (kilobytes per event) for analysis, and enabling access to CPU resources from external WLCG resources such as High-Performance Computing (HPC), Graphics Processing Units (GPUs), and Cloud platforms.

6 Analysis Facilities

In the original Grid computing model, Analysis Facilities, known as Tier-3s, played a crucial role in data analysis. These sites were often located alongside Tier-1s and Tier-2s but dedicated to local groups' private use. Unlike the standardized resources of the WLCG Grid deployment, Tier-3s were not pledged and did not require standard resources.

The primary objective of Analysis Facilities is to assist physicists in reducing the time-to-insight, enabling iterative data exploration. By leveraging these facilities, physicists can significantly accelerate the data analysis process, allowing for the efficient handling of larger volumes of data without proportional increases in time requirements. This significant reduction in waiting time will greatly enhance the productivity and competitiveness of our physics communities. These high-level objectives have always been and continue to be true. However, in the face of the upcoming HL-LHC era it is even more crucial to remark how the evolution and enhancement of these Analysis Facilities can aid in curbing the delays stemming from an increase of a factor of ten in the data generated by the different LHC experiments.

In order to fulfil their intended purpose, Analysis Facilities should portray a series of features:

1. The facilities should enable local access to reduced data samples (e.g., DAOD_PHYSLITE) with minimal latency from the compute infrastructure. This calls for hundreds of TBs of dedicated storage.

2. Adequate processing resources, including CPUs, GPUs, and memory, should be available, prioritized, and capable of handling the required workload.

3. The infrastructure should be designed for efficiency, employing techniques like scheduling based on systems such as HTCondor [17]. It should also be elastic, allowing for dynamic expansion to HPC/Cloud resources during peak loads.

4. The facilities should provide enhanced support for commonly used software tools such as the ROOT and the Python ecosystem, ensuring seamless integration and optimal performance.

5. They should actively encourage and facilitate the use of shared repositories for analysis routines and workflows to foster collaboration and code reuse.

6. A knowledgeable data/code manager should play a crucial role as a liaison, facilitating access to the technology and assisting users in effectively utilizing the facilities.

7. Modern software stacks and User Interfaces should be made available to the end users. These include Dask [18], coffea [19] and Jupyter [20] among others.

The Analysis Facilities being deployed at the Spanish sites typically comprise dedicated storage resources with capacities on the order of several hundred terabytes. For processing, CPU resources are utilized interactively or through a batch system, predominantly HTCondor. Software delivery is primarily facilitated via CVMFS [21], with a growing presence of container-based solutions. Although GPU resources are available, they are not always exclusively dedicated to the ATLAS teams. The network infrastructure consists of a LAN with multiple 10 Gbps capacity, supporting high-speed data throughput between the facility's processing and storage nodes. Additionally, there is a wide-area network (WAN) connectivity of 100 Gbps, ensuring seamless access to the WLCG dedicated network for data lake operations.

At PIC in Barcelona, a deployment of Jupyter Notebook instances [22] have been implemented. At IFIC, the KORE [23] system has also been deployed and is actively being tested by users; it also leverages Jupyter Hub as the user-facing interface. These instances can be created through a dedicated portal, providing the flexibility to select different resource profiles (CPU, memory, GPU) for resource allocation at the PIC farm [24]. On the other hand, at IFIC in Valencia, a GPU infrastructure named ARTEMISA [25] is actively utilized for computational tasks and to take profit of these advantages for analysis.

7 Conclusions and Perspectives

The ATLAS Spanish Tier-1 and Tier-2s have more than 19 years of experience in the deployment of LHC computing components and their successful operation. The sites are actively participating in the evolution of the computing models through the integration and update of ingredients/tools provided by ATLAS. The HL-LHC provides unprecedented opportunities for particle physics, yet its implementation poses technical and logistical challenges. Therefore an essential plan for upgrading the computing infrastructure and optimizing data analysis methods is necessary for fully realizing the HL-LHC's potential and overcoming its hurdles. The work of all the Spanish sites together with the NREN are concrete steps taken toward achieving that goal.

This work was partially supported by MICINN in Spain under grants PID2022-142604OB-C21, PID2022-142604OB-C22, PID2022-136323OB-C21, PID2022-142604OB-C22, PID2022-136323OB-C22 and by the GenT program of Generalitat Valenciana. The authors thankfully acknowledge the computer resources at MareNostrum, and the technical support provided by BSC.

References

[1] ATLAS Collaboration, *ATLAS Experiment at the CERN Large Hadron Collider, JINST* **3**, S08003 (2008).

[2] J. Elmsheus er et al., *Overview of the ATLAS distributed computing system, EPJ Web Conf.* **214**, (2019).

[3] S. Gonzalez de la Hoz et al., *Computing activities at the Spanish Tier-1 and Tier-2s for the ATLAS experiment towards the LHC Run 3 period, EPJ Web Conf.* **245**, (2020).

[4] Worldwide LHC Computing Grid project: http://wlcg.web.cern.ch

[5] dCache: https://www.dcache.org

[6] Lustre: https://www.lustre.org
[7] StoRM: http://italiangrid.github.io/storm/
[8] HTCondor: https://htcondor.org
[9] M. Ellert et al., *Advanced Resource Connector middleware for lightweight computational Grids, Future Generation Computer System* **23**, 219-240. doi:10.1016/j.future.2006.05.008 (2007).
[10] SLURM: https://slurm.schedmd.com/overview.html
[11] Spanish Supercomputing Network: http://www.res.es/en
[12] C. Acosta-Silva et al., *Exploitation of the MareNostrum 4 HPC using ARC-CE, EPJ Web Conf.* **251**, (2021).
[13] E. Gallas et al., *Deployment and Operation of the ATLAS Event Index for LHC Run 3,* these proceedings (2023).
[14] D. Barberis et al., *The ATLAS EventIndex: A BigData Catalogue for ALL ATLAS Experiment Events, Comput.Softw.Big Sci.* **7**, (2023).
[15] C. Garcia Montoro et al., *HBase/Phoenix-based Data Collection and Storage for the ATLAS EventIndex, EPJ Web of Conferences*, (2024).
[16] I. Bird et al., *Architecture and prototype of a WLCG data lake for HL-LHC, EPJ Web Conf.* **214**, (2019).
[17] D. Thain, et al., *Distributed Computing in Practice: The Condor Experience Concurrency and computation: practice and experience*, **17(2-4)** 323-356. DOI: https://doi.org/10.1002/cpe.938 (2004).
[18] Dask: https://www.dask.org
[19] N. Smith, et al., *Coffea Columnar Object Framework For Effective Analysis, EPJ Web Conf.* **245**. DOI: 10.1051/epjconf/202024506012 (2020).
[20] B. Granger, et al., *Jupyter: Thinking and Storytelling With Code and Data, Computing in Science & Engineering* **23**. DOI: 10.1109/MCSE.2021.3059263 (2021).
[21] J. Blomer et al., *Delivering LHC Software to HPC Compute Elements with CernVM-FS, In Proceedings of the first international workshop on Network-aware data management* **49-56**. DOI: 10.1145/2110217.2110225 (2011).
[22] Jupyter: https://jupyter.org
[23] KORE: https://kore.ific.uv.es
[24] A. Delgado Peris et al., *Spanish Analysis Facility at CIEMAT, EPJ Web of Conf.* **295**. DOI: 10.1051/epjconf/202429507045 (2024).
[25] ARTEMISA: https://artemisa.ific.uv.es/web/