








Exploitation of the MareNostrum 4 HPC using ARC-CE

Carles Acosta-Silva^{1,2}, José Del Peso³, Esteban Fullana Torregrosa⁴, Santiago González de la Hoz⁴, Andrés Pacheco Pages^{1,2}*, José Salt⁴, and Javier Sánchez Martínez⁴

¹Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology, Campus UAB, 08193 Bellaterra (Barcelona), Spain

²Port d'Informació Científica (PIC), Campus UAB, 08913 Bellaterra (Cerdanyola del Vallès), Spain

³Departamento de Física Teórica y CIAFF, Universidad Autónoma de Madrid, Madrid, Spain

⁴Institut de Física Corpuscular (IFIC), Centro Mixto CSIC - Universitat de València, Paterna, Spain

Abstract.

The resources of the HPC centers are a potential aid to meet the future challenges of HL-LHC [1] in terms of computational requirements. Spanish HPC centers have recently been used to implement all necessary edge services to integrate resources into the LHC experiment workflow management system. In this article, we describe the integration of ATLAS with the extension plan to other LHC experiments. We chose to configure a dedicated ARC-CE [2] and interact with the HPC login and transfer nodes using ssh commands.

The repository that includes a partial copy of the ATLAS experiment software on CVMFS is packaged in a singularity image to overcome network isolation for HPC nodes and reduce software requirements. ATLAS provided the initial container, and the authors adapted it to the specific HPC environment. This article shows the Spanish contribution to the simulation of experiments after the Spanish Ministry of Science agreement and the Barcelona Supercomputing Center (BSC), the center that operates MareNostrum 4. Finally, we discuss some challenges to take advantage of the next generation of HPC machines with heterogeneous architecture combining CPU and GPU.

1 Introduction

HPC resources help meet the future challenges of the High Luminosity LHC [1] (HL-LHC) period in terms of CPU requirements, which the budget for high energy physics programs cannot fully fund. The Spanish WLCG centers [3] are making an effort to integrate local HPC resources into the workflow management systems of the LHC experiments. In the case of this article, we show the results after BSC's Marenostrum 4 HPC was integrated as a shared resource by the three WLCG centers that provide computing resources to ATLAS. These centers are located in Madrid (UAM), Valencia (IFIC), and Barcelona (PIC) [4].

In section 2, we explain the organization of HPCs in Spain and describe the MareNostrum 4 HPCs. In section 3, we present the details of the implementation. Section 4 shows the results in 2020 regarding resources consumed in Spain for the ATLAS experiment [5]. Finally, in section 5, we discuss possible new developments.

*e-mail: pacheco@ifae.es

2 The HPC situation in Spain

In Spain, there is no single HPC center. A Supercomputing Network created in March 2007 by the Ministry of Education and Science with centers spread throughout the Spanish geography.

2.1 The RES Spanish Supercomputing Network

The Spanish Supercomputing Network (RES) [6] is a distributed infrastructure consisting of 14 interconnected supercomputers throughout Spain. They offer high-performance computing resources to the scientific community. The Barcelona Supercomputing Center (BSC) coordinates this network, which aims to manage high-performance computing technologies to promote science and innovation of excellence in Spain.

The first RES center integrated into ATLAS was the Lusitania Cenits cluster in Extremadura with 25 Teraflops. This step as a starting point was essential because the computing nodes in that cluster had Internet connectivity. In 2018, we integrated MareNostrum 4 (BSC, Barcelona) with 13.7 Petaflops in the ATLAS production system to significantly increase computing resources. In the case of MareNostrum 4, the computing nodes did not have any connectivity to the Internet.

Research projects that need to use HPC resources periodically submit a request for calculation hours to an external committee of the Spanish Supercomputing Network to be evaluated. In 2020 the BSC included High Energy Physical Computing as a strategic project following the directives of the Ministry of Science. The resources of this agreement began in mid-2020, and the contribution to the production of simulated ATLAS events has increased considerably since then. The status of the strategic project is reviewed annually and has a direct impact on the hours granted at MareNostrum 4.

In the future, the authors plan to continue requesting these HPC resources also at the European level (PRACE [7] and EuroHPC [8]) to increase the fraction of CPU that Spanish HPCs contribute to ATLAS together with other LHC experiments if it is possible. It is worth mentioning that the calls to request computer hours at HPC are competitive among all scientific fields; therefore, justification at the level of the physics research lines is essential to the success of the grants.

2.2 The infrastructure of the MareNostrum 4

The MareNostrum 4 machine is the most powerful supercomputer in Spain, with its 13.7 Petaflops of computing power. The general-purpose block has 48 racks with 3,456 Lenovo SD530 nodes. Each node has two Intel Xeon Platinum (Skylake) chips, each with 24 processors with 48 cores per physical node, which is equivalent to 165,888 processors and main memory of 390 Terabytes. MareNostrum 4's operating system is Suse Linux, the batch system is Slurm [9], and distributed storage uses GPFS [10].

The BSC has a disk storage capacity of 14 Petabytes connected to a National Big Data infrastructure. The center connects with European research centers and universities through the RedIris and Geant networks [11] and is the Spanish center in PRACE (Partnership for Advanced Computing in Europe). A new supercomputer called MareNostrum 5 [12] of 200 Petaflops is planned for 2022, co-financed with funds from the European EuroHPC program.

3 Integration of the Spanish HPC resources in LHC experiments

The LHC experiments have a centralized computer task submission system. Although it is not the same for each of them, they are pilot frames in which a pilot job extracts the

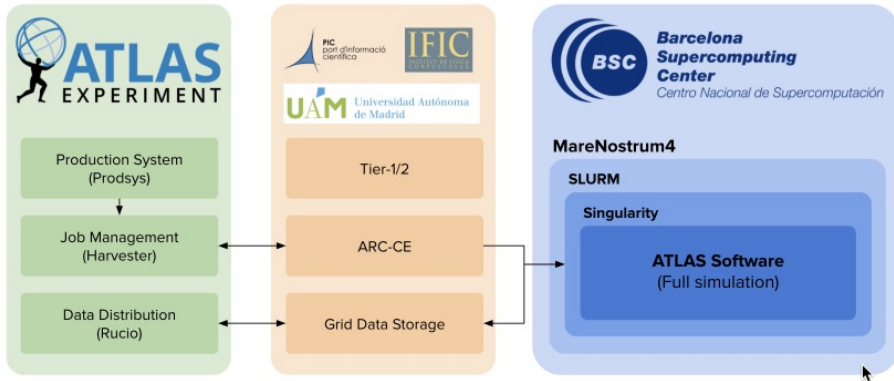


Figure 1. Flow diagram between the ATLAS production system and the execution of jobs on the MareNostrum 4 supercomputer.

payload from a specific experiment queue. CMS uses glideinWMS for CMS, ATLAS uses Panda and LHCb uses Dirac. There are several possibilities to integrate HPC centers in experiments, depending on their architecture and characteristics. Figure 1 illustrates the flow between the ATLAS production system (Prodsys) and the job submission in the specific case of MareNostrum 4. When the ARC-CE receives a job, the specific job description of the flow management system is obtained. Then the ARC-CE sends a job to the batch system of the HPC with the payload, interacting with the HPC MareNostrum 4 login and transfer nodes using ssh commands. The experiment software is included in a partial snapshot of the CVMFS Atlas-specific repositories [14]. It is packaged in a container singularity image [15] to overcome network isolation for HPC worker nodes and reduce software requirements.

3.1 Implementation using ARC-CE

The ARC-CE [2] was developed at Nordugrid as an interconnection system for the LHC experiment network with local batch systems in computing centers. A large number of Grid sites have adopted it. Jobs submitted to the ARC-CE from the involved WLCG resource centers are transformed into local jobs within MareNostrum 4 and forwarded to the center's Slurm queuing system. Once the work is finished, the ARC-CE is in charge of copying the results to the ATLAS computer system, recording them as if they had been executed on the Grid. It is important to note that jobs are submitted using a singularity image with CentOS and that while they are running, they cannot connect to the outside world. Unlike other HPC centers, MareNostrum 4 compute nodes do not have connectivity. In the implementation used for the results shown, the jobs were configured to use an entire 48-core physical machine to run.

Jobs on the ARC Compute Element generally follow these steps:

1. The ATLAS production system connects to the ARC-CE job submission interface. The client tools create a proxy certificate using the user's credentials and connect to a *Virtual Organization Membership Service (VOMS)*.
2. Using the X.509 public key infrastructure processes, the client and server authenticate each other based on trusted CA credentials previously installed on both ends.

3. The ARC-CE authorizes the user based on configurable rules and maps the identity of the user proxy to a local account.
4. The ATLAS production system grants the user's credentials to the ARC-CE to act on behalf of the user when transferring files.
5. A job description is sent from the ATLAS production system to the server.
6. The job is accepted, and a directory is created that will be the home of the session. The metadata about the job is written to the ARC-CE server control directory.
7. The client receives the location of the session directory and, if there are local input files, they will be uploaded through the file access interface through the GridFTP server.
8. If the job description specifies input files at remote locations, the ARC-CE server obtains the required files and places them in the job's working directory.
9. When all the files written in the job description are present, a suitable job *script* is created and sent to the queuing system configured in the MareNostrum 4 (Slurm).
10. During this time, the ATLAS production system can continuously access the job files so that any intermediate results can be verified.
11. Information provider *scripts* periodically monitor the status of the job, updating the information in the control directory.
12. When the job on MareNostrum 4 ends, the ARC-CE server uploads, maintains, or removes the resulting output files according to the job description. The job status and files in the ATLAS production system are updated.

We have implemented an instance of the ARC-CE in each of the centers located in Madrid, Valencia, and Barcelona. If one of the centers fails, jobs can continue to be sent to MareNostrum 4 through another center.

3.2 Other implementations

Other successful implementations are under the test phase in MareNostrum 4; one is based on Harvester[16], used by ATLAS, which allows using thousands of cores per job thanks to using MPI. One of the exciting features of the harvester implementation is that it can better benefit from the MareNostrum 4 HPC's massively parallel components. The other is an extension of HTCondor for HPCs without connectivity at the computing node level presented in vCHEP2021[17]. All these possibilities allow various functionalities that profit specific characteristics of the HPCs.

4 Results from MareNostrum 4 integration in Spain

In 2020, the MareNostrum 4 was fully integrated into the ATLAS computing framework. During that year, the HPC contributed 30% of the total contribution to the ATLAS computing by the Spanish cloud (see Fig. 2 left). Among all the different types of computing, the MareNostrum 4 HPC contributed only to the simulation effort. Figure 2 right shows that the MareNostrum 4 HPC contributed to 64% of the cumulative simulation provided by WLCG Spain to ATLAS.

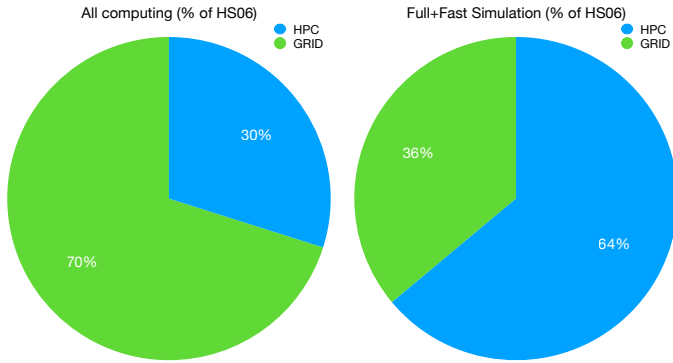


Figure 2. Left: Proportion of HS06 (s) provided by GRID resources (green) and the MareNostrum 4 HPC (blue) in total contribution to the ATLAS computing by the Spanish cloud in 2020. Right: Proportion of HS06 (s) provided by GRID resources (green) and the MareNostrum 4 HPC (blue) only in simulation jobs during 2020 in Spain.

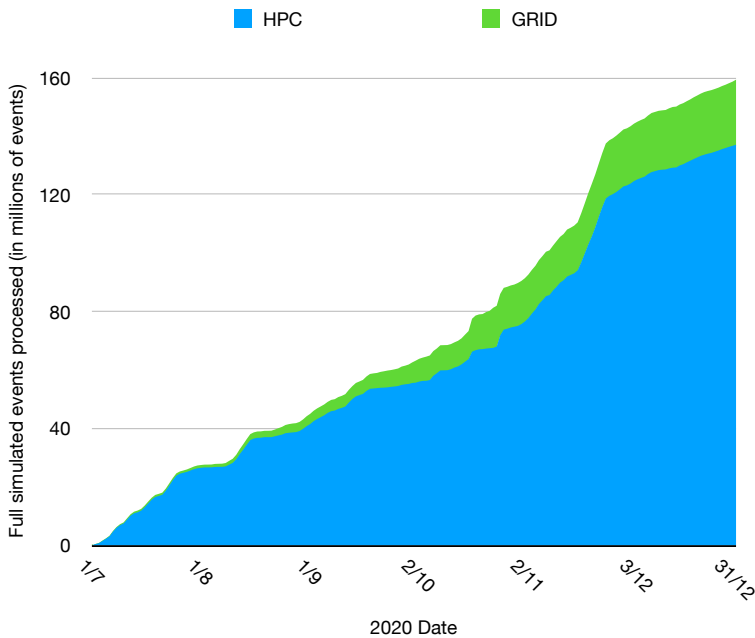


Figure 3. Time evolution of the accumulative number of full simulated events processed by GRID resources (green) and MareNostrum 4 HPC (blue) since the start of the agreement between Spanish Ministry of Science and the BSC in July 2020.

In 2020, the Spanish Ministry of Science and the BSC’s agreement started to take effect in the second half of 2020. Figure 3 right shows the impact of the new framework in terms of processed full simulated events. The figure indicates that since the effective start of the agreement, July and forward, most of those events were processed in MareNostrum 4.

The Spanish WLCG centers involved have not observed bottlenecks in network transfers between them and the BSC and the Spanish. The BSC has a high-speed connection with the backbone of the Spanish academic network.

5 New developments

The transition to increasing use of HPC requires the improvement in software distribution based on containers solution [13]. In addition, the next generation of supercomputer centers is evolving from pure computational facilities to resources for extreme data processing, e.g., Big Data, High-Performance Data Analysis, and Artificial Intelligence (Machine and Deep Learning). Most of the computing power will be provided through heterogeneous architecture mixing CPUs and GPUs.

5.1 Software distribution using containers

The ATLAS software installation in HPCs is a non-trivial task that has to be performed every time a new ATLAS software release comes out. The installed environment is often different from those in CVMFS [14], a tool used to distributed LHC experiments' software environment to their Grid sites, due to various constraints such as hardwired path starting with /cvmfs in software and configuration files. Since most HPC centers now support container technologies one way or the other, it is possible to put the ATLAS software environment in a container. This option allows distributing the ATLAS software environment consistently, without high labor cost.

In our case, ATLAS software experts build containers with singularity and used them to successfully address the issues of software distribution and metadata I/O reduction on HPCs where CVMFS is not widely available. ATLAS experts efficiently extract ATLAS software from CVMFS, update containers quickly, and make their sizes small enough to be transportable, but this is a static image. As the ATLAS collaboration puts more software in CVMFS, the method used to update this container needs to be improved. The option is to separate different versions of the software releases in separate containers is the privileged solution that we need to implement in our integration.

5.2 Analysis jobs

Matching LHC analysis workflows with an HPC system is not a trivial effort. HPC facilities usually have strict site policies for security reasons and may be based on various specialized hardware setups. On the other hand, LHC workloads are designed to run on "standard WLCG sites." Consequently, integrating LHC analysis workflows on HPC centers poses two rather distinct types of challenges in the software area (e.g., efficiently exploiting low-memory many-core CPUs, or porting applications to GPUs) and in the computing area (e.g., managing job scheduling and software/data distribution).

The plan for integrating analysis jobs is to identify those jobs in which physicists have generated analysis containers and run them on the supercomputer transparently, similar to how simulation jobs are run. The difference is that the container is created by the person analyzing the data and not by the official ATLAS software group.

5.3 GPU jobs

The LHC C/C++ code was initially designed and developed for single x86 CPU cores, but it has been modified for multicore CPUs. The code can be recompiled for other CPU platforms

as well, such as ARM and Power9. However, to use a large computing capacity available on GPUs at the HPC centers, some kernel code must be re-engineered, and new algorithmic code must be developed.

One of the risks today is that the next generation of HPC centers will rely heavily on accelerators (GPUs). The use of GPUs for the LHC experiments is currently a challenge. In the simulation case, which is a substantial part of the experiments' computational resources, the LHC experiments are highly dependent on the development of versions of GEANT [18] for use by GPUs. However, more and more data analysis uses ML techniques and accelerators as fundamental techniques.

In MareNostrum 4, the machines do not have GPUs. The next upgrade, MareNostrum 5, expected in 2022, will have GPUs. We plan using a cluster called Minotauro [19] from the same center to send jobs that use GPUs until it is available.

6 Conclusions

We have shown that the Spanish WLCG centers providing ATLAS resources can integrate the national HPC centers into LHC computing workflows. This paper focuses on simulation tasks submitted using jobs from the ATLAS experiment in 2020.

The integration of HPCs into the LHC workflow was possible due to two key elements: the ARC-CE interface and the use of singularity. This integration, developed during 2018 and 2019 and using Lusitania and MareNostrum 4 HPC machines, reached full speed in 2020. During that year, 30% of all the Spanish cloud ATLAS computing contribution was done by MareNostrum 4. In addition, regarding only simulation jobs, MareNostrum 4 brought 64% of the total. That contribution intensified after the agreement between the Spanish Ministry and the BSC. Both the use of MareNostrum 4 and the agreement Ministry-BSC have proven to be very promising for the future Spanish contribution to the HL-LHC computing.

Acknowledgements

We want to acknowledge support from the MICINN in Spain under grants PID2019-110942RB-C22, PID2019-104301RB-C21, PID2019-104301RB-C22 and FPA2016-80994-C2-2-R, FPA2016-75141-C2-1-R, FPA2016-75141-C2-2-R, including FEDER funds from the European Union. In addition, we gratefully acknowledge the computing centres and personnel of the Worldwide LHC Computing Grid and other centres for delivering so effectively the computing infrastructure essential to our analyses. The authors thankfully acknowledge the computer resources and the technical support provided by BSC and Cenits. (FI-2020-1-0001,FI-2020-1-0027, FI-2020-2-0001,FI-2020-2-0010, FI-2020-2-0004, FI-2020-3-0001, FI-2020-3-0006, FI-2021-1-002, FI-2021-1-0006, FI-2021-1-0003).

References

- [1] L. Evans and P. Bryant (editors), JINST **3** S08001 (2008)
- [2] Ellert, Mattias; et al. "Advanced Resource Connector middleware for lightweight computational Grids". Future Generation Computer Systems. Volume 23, Issue 2, February 2007, Pages 219-240. doi:10.1016/j.future.2006.05.008
- [3] J. Shiers, Comput. Phys. Commun. **177** 219–223 (2007)
- [4] S. González de la Hoz et al., EPJ Web of Conf. **245** 07027 (2020)
- [5] ATLAS Collaboration, JINST **3** S08003 (2008)
- [6] *Spanish Supercomputing Network*, <https://www.res.es/en>

- [7] *Partnership for Advanced Computing in Europe*, <https://prace-ri.eu/>
- [8] *European High Performance Computing Joint Undertaking*, <https://eurohpc-ju.europa.eu/>
- [9] Simple Linux Utility for Resource Management, A. Yoo, M. Jette, and M. Grondona, Job Scheduling Strategies for Parallel Processing, volume 2862 of Lecture Notes in Computer Science, pages 44-60, Springer-Verlag, 2003.
- [10] Schmuck, Frank; Roger Haskin (January 2002). "GPFS: A Shared-Disk File System for Large Computing Clusters" (PDF). Proceedings of the FAST'02 Conference on File and Storage Technologies. Monterey, California, US: USENIX. pp. 231–244. ISBN 1-880446-03-0. Retrieved 2008-01-18.
- [11] *GÉANT pan-European network*, https://www.geant.org/Networks/Pan-European_network/Pages/Home.aspx
- [12] Meet Europe's new supercomputer: MareNostrum 5 takes on global rivals for power <http://cern.ch/go/mh6l>
- [13] N. Ozturk et al. "Containerization in ATLAS Software Development and Data Production", these proceedings
- [14] J. Blomer et al, "CernVM-FS: delivering scientific software to globally distributed computing resources", NDM '11: Proceedings of the first international workshop on Network-aware data management. November 2011 Pages 49–56 <https://doi.org/10.1145/2110217.2110225>
- [15] G. M. Kurtzer, V. Sochat and M. W. Bauer, PLoS ONE 12(5) e0177459 (2017)
- [16] Fernando Barreiro et al., "Managing the ATLAS Grid through Harvester", EPJ Web of Conferences 245, 03010 (2020) <https://doi.org/10.1051/epjconf/202024503010>
- [17] C. Acosta-Silva et al., "Exploitation of network-segregated CPU resources in CMS", these proceedings
- [18] 'Recent developments in GEant4', J. Allison et al . NIM A 835 (2016) 186-225
- [19] *MinoTauro* <https://www.bsc.es/es/marenostrum/minotauro>