Spanish ATLAS Tier-1 & Tier-2 perspective on computing over the next years

Santiago González de la Hoz[1*], *Carlos* Acosta-Silva[3,4], *Javier* Aparisi Pozo[1], *Manuel* Delfino[3,4], *Jose* del Peso[2], *Álvaro* Fernández Casani[1], *José* Flix Molina[4,5], *Esteban* Fullana Torregrosa[1], *Carlos* García Montoro[1], *Julio* Lozano Bahilo[1], *Almudena del Rocio* Montiel[2], *Andreu* Pacheco Pages[3,4], *Javier* Sánchez Martínez[1], *José* Salt[1] and *Aresh* Vedaee[3,4] on behalf of the ATLAS Collaboration

[1]Instituto de Física Corpuscular (IFIC), centro mixto CSIC – Universitat de València, Paterna, Spain
[2]Departamento de Física Teórica y CIAFF, Universidad Autónoma de Madrid, Madrid, Spain
[3]Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology, Campus UAB, 08193 Bellaterra (Barcelona), Spain
[4]Port d'Informació Científica (PIC), Campus UAB, 08913 Bellaterra (Cerdanyola del Vallès), Spain
[5]Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (Ciemat), Madrid, Spain

Abstract. Since the beginning of the WLCG Project the Spanish ATLAS computing centers have participated with reliable and stable resources as well as personnel for the ATLAS Collaboration. Our contribution to the ATLAS Tier2s and Tier1s computing resources (disk and CPUs) in the last 10 years has been around 4-5%. In 2016 an international advisory committee recommended to revise our contribution according to the participation in the ATLAS experiment. With this scenario, we are optimizing the federation of three sites located in Barcelona, Madrid and Valencia, considering that the ATLAS collaboration has developed workflows and tools to flexibly use all the resources available to the collaboration, where the tiered structure is somehow vanishing. In this contribution, we would like to show the evolution and technical updates in the ATLAS Spanish Federated Tier2 and Tier1. Some developments we are involved in, like the Event Index project, as well as the use of opportunistic resources will be useful to reach our goal. We discuss the foreseen/proposed scenario towards a sustainable computing environment for the Spanish ATLAS community in the HL-LHC period.

1 ATLAS Computing Challenges at HL-LHC period

In the High Luminosity LHC period (HL-LHC) the ATLAS computing demands far outstrip the budget dedicated to computing by the different countries. On the other hand, physics analysis and simulation from ATLAS are already limited by the current computing resources. Therefore, new techniques have to be developed to enhance a robust network, a key-

* Corresponding author: santiago.gonzalez@ific.uv.es

component in the WLCG architecture, minimize storage needs and make the most of opportunistic resources (HPCs) [1].
Assuming a flat budget for the facilities, a 20%/yr capacity growth, and a constant development effort, a ~3x shortfall of CPU resources with respect to the flat budget model is estimated for the HL-LHC (Run 4), see Figure 1. In this case, the potential use of co-processor/GPUs has not been considered.

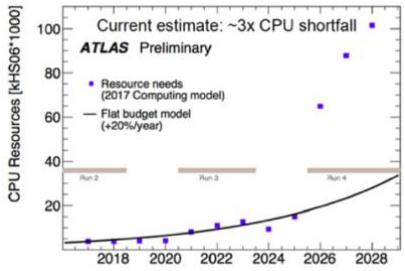

Fig. 1. Current estimate: ~3x CPU shortfall [1].

In Figure 2, we can see ~6x shortfall by today's estimate in Storage on disk. This is the ATLAS bigger problem because "opportunistic storage" basically does not exist. ATLAS collaboration is working on format size reductions, but it is hard to achieve large gains so ATLAS needs new approaches. A way to dramatically reduce the storage footprint is to increase the use of tape but tape access is slow and its use complicates the workflow.

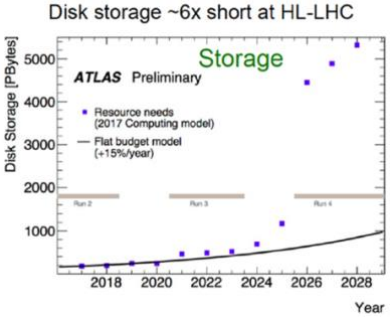

Fig. 2. Storage scaling to HL-LHC [1].

2 Spanish ATLAS Tier-1 & Tier-2 Federation

The Spanish computing community is formed by one Tier-1 (PIC), and a federated Tier2: ES-ATLAS T2 (50% IFIC, 25% IFAE and 25% UAM). So far, the community has provided the hardware resources fulfilling the ATLAS requirements of the Resource Review Board of the LHCC committee, contributing around 4% of the total resources of all ATLAS Tier-2s and 5% of all Tier-1s. These resources are integrated in the WLCG (World Wide LHC Computing Grid) project and follow the ATLAS Computing Model. Present resources provided are shown in the Table 1:

Table 1. Hardware resources provided by the Spanish community on September 2018.

Site	CPU (HEP-SPEC06)	DISK (TB)	TAPES (TB)
PIC-Tier1	46200	3265	9670
IFIC-Valencia	26751	2146	-
IFAE-Barcelona	10420	980	-
UAM-Madrid	10358	1220	-

In the recent LHC running periods around 6 PB of data have been stored in the Spanish facilities as shows the Figure 3.

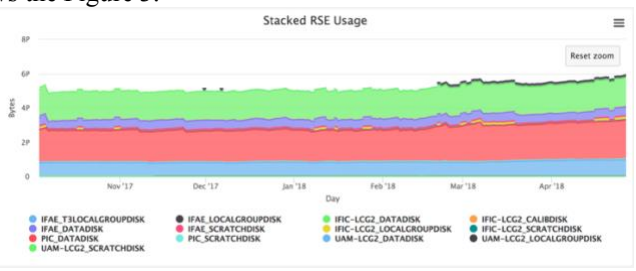

Fig. 3. Disk usage in the last 6 months in Spanish sites.

Concerning the facilities performance in terms of availability and reliability, Table 2 shows the values for the 2017 year. As can be seen, in both cases the average over all the Spanish sites is greater than 98%.

Table 2. Site Availability and Reliability performance, based on WLCG information in the year 2017.

Site	Availability (2017)	Reliability (2017)
PIC-Tier1	98.96%	99.22%
IFIC-Valencia	98.25%	98.54%
IFAE-Barcelona	98.65%	98.97%
UAM-Madrid	98.92%	99.67%

3 The Event Index (EI) Project

The ATLAS EI project indexes and catalogues all events produced by ATLAS. This allows to rapidly locate individual events and study large quantities of data. It uses industry big data tools like Hadoop for storing properties of all real and simulated data, and additionally Oracle database storage for fast dataset discovery.

IFIC-Tier2 group is involved in several areas, including the Data Production and Distributed Data Collection, and also provides tools for analysing the data. It also follows the new requirements in order to fit the analysis and production demands, and the technologies that can improve and contribute to the evolution of different areas of the project. As coordinator of the Data Production, it ensures all Tier-0 merged physics AODs and grid jobs collect info from EVNT and AOD datasets as soon as they are produced and set "ALL EVENTS AVAILABLE" in AMI. DAODs are indexed on demand. IFIC-Tier2 group is responsible for the architecture design and operation of the Distributed Data Collection, that since last year has put in production the new system based on an Object Store [2,3] with good results.

Within this new system, the EI Data Collection Supervisor is in charge of orchestrating the different components of the distributed system that indexes the datasets. Producer servers that run at CERN and the grid extract the relevant information of the datasets and temporally store them at Object Store. Producers notify the supervisor about the production. Once the production of a dataset is successfully completed, the supervisor checks the integrity and completeness of the production with other systems and consumers are informed about the validated production that has to be written into Hadoop. Consumer servers consume production taking the produced data from Object Store, according to what the supervisor has informed them. Once datasets are consumed, consumers inform the supervisor. A web application provides information of the current state of the supervisor. Through this web, the

progress of the datasets being indexed can be followed and corrective actions can be performed when needed.
With respect to the tools for analysing the data, the Trigger Counter is a web service that is able to provide meaningful information about the L1 and HLT triggers of a dataset in a human time scale by means of a very compact and optimised representation of the trigger information and some Hadoop MapReduce jobs. Given a dataset and an optional filtering expression about the presence or absence of some triggers, and/or the value of lumiblock (information on the collection of luminosity blocks) and/or bunch id of the event, it can: 1) count the occurrences of each trigger 2) compute the overlap between all the triggers, as is shown in the Figure 4. 3) provide a limited list of the events that satisfy the established expression. The service is able to provide interactive plots of the results as well as the data in JSON and CSV formats. The service is able to process millions of events and billions of triggers in a couple of minutes.

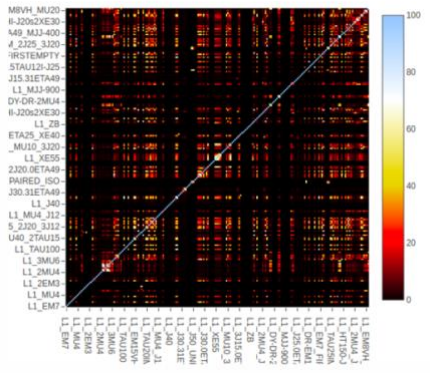

Fig. 4. A heatmap of the trigger overlaps of a dataset is shown.

4 HPCs in Spain: Use of Extra Resources

In 2017 and 2018 we got granted CPU hours in the RES (Spanish Supercomputing network). The RES allocated us 70k hours in the Lusitania supercomputer (based at CENITS) and two periods of 100k each in Mare Nostrum (based at BSC), which corresponds to use the complete PIC Tier-1 resources for about half a day. The Mare Nostrum has a peak performance of 11.15 Petaflops with 384.75 TB of main memory and 3,456 nodes (2x Intel Xeon Platinum 8160 24C at 2.1 GHz). Lusitania has 2 HP Integrity SuperDome SX2000 with 64 Itanium2 Montvale (1.6 GHz, 18 MB cache) processors each one (128 processors / 256 cores).
ATLAS production workflow does not automatically send grid jobs to remote HPC systems; indeed, it is a hard task [4] and still under work in the community. Sites need to perform some manual work in order to be integrated and efficiently use HPC resources. Common constraints from HPC systems are: no access to the outside from the compute nodes, no root privileges, or not being able to install any system wide-software, no compatible Operating System in which we can re-compile the current ATLAS software. In this scenario, it is required to send grid jobs to the HPC and to run ATLAS simulation events upon request.
The solution for installing the ATLAS SW has been to request a Singularity [5] installation in the HPCs and transfer a complete file image of the ATLAS SW in the two supercomputers because they do not have cvmfs. The input (event generation) and output (Geant4 full simulation) processed files are transferred in and from the HPCs using ssh. Both, IFIC and IFAE, use ARC Computing Element [6] as interface between the ATLAS grid infrastructure and the remote HPC installed at IFAE and IFIC.

In May 2018 the two HPCs started to simulate events fully integrated in the ATLAS grid infrastructure. Figure 5 shows the activity during the peak week from the 18th to the 25th of May. The figure shows how the HPC multiplied the computing capacity of the Spanish ATLAS computing power up to a factor of 5. The number of hours allocated to us was consumed in a few days. This exercise will be extremely valuable to assess the use of HPCs in High Energy Physics for the computing needs of the HL-LHC. The output is registered in the ATLAS catalog once it is at IFIC or at IFAE in a synchronous way.

Fig. 5. Number of jobs running in Spanish Grid resources and Spanish HPCs from 18th to 25th May.

5 Network LHCONE in Spain

LHCONE [7] is a collaborative network implemented for the LHC research community as a controlled access "walled garden" routed IP internet. It improves the use of the network to facilitate data analysis, which is mainly done in Tier-2s. Specifically, this infrastructure improves the transfers between Tier-1s and Tier-2s and from Tier-2 to Tier-2. It complements LHCOPN, which accommodates traffic of Tier-0 Tier-1 transfers.

PIC and IFAE already joined back in 2010, as soon as the infrastructure started operating. UAM has joined in 2017. LHCONE imposes a few guidelines to be followed to keep a symmetric routing, as well as some networking policies to provide a mutual trust among sites where only valid LHC traffic is routed. Such configurations are performed in the BGP (Border Gateway Protocol), in both at NSP (Network Service Provider) and site ends. Figure 6 shows the performance for LHCONE at the UAM and IFAE.

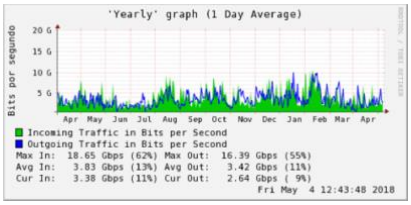

Fig. 6. Daily LHCONE traffic in RedIRIS in the last year (for PIC, IFAE & UAM).

In particular, UAM-LCG2 site joined whilst upgrading the network link. Such link was increased up to 10Gbps. In this new configuration, there is a second backup link, supplied by a different provider. The main 10Gbps line is connected to LHCONE through RedIris (Spanish academic and research network). The backup link is maintained by the University, and connected via CIEMAT to the outside. PIC Tier-1 and IFAE Tier-2 use a 20Gbps WAN network, connected to LHCONE at 10 Gbps.

At the site level, all Tier-2 services to be included in LHCONE had to be reorganized and aggregated into a single range of IPs, in order to minimize the LHCONE Prefixes, and to assure good performance of routing algorithms.

6 Federation Storage

The strong collaboration among the ATLAS Spanish sites allows ways of exploring federating access to the endpoints of storage that would be considered as a unique entry point. This idea of a regional storage federation is aligned with the WLCG Data Steering Group strategies to face the medium/long term Data Management challenges, i.e. the Run3 and HL-LHC.
The regional federation should meet the following requirements: scalability, performance, WAN access, redirection and fault tolerance.
A first attempt is being tested with DynaFed [8] at the UAM, which is a tool that meets the needs. The plan is to integrate the rest of the sites to such instance and to evaluate the tool.
On the other hand, Spanish sites are going to participate in the CERN Data Lakes project to federate storage and reduce costs (see section 8 ATLAS Core Computing). Contributing to the Eulake project, which is based on the EOS system to evaluate the potential of a new distributed storage solution to build Data Lakes [9].

7 Data Analysis (DA)

Data Analysis activity in our sites is carried out in close collaboration with the ATLAS Physicist, keeping substantial CPU resources for data analysis. In our sites, more and more frequently, Physics analysis will be performed using Machine Learning (ML) methods, that would imply the use of GPUs and much more memory than provided by ATLAS standard Worker Nodes. At the same time, user can do analysis using just their browser through tools like SWAN (Service for Web based ANalysis) [10].
The application of ML techniques has been generalized in the LHC experiments and there are already a whole set of programs, libraries, etc. that are being used for the different analysis [11]. In particular, we are interested in importing ML methods and tools developed and implemented in general environments / frameworks into the standard workflow used by an ATLAS Analysis physicist (LHC): ANACONDA, TensorFlow, etc…
In order to check these methods and tools, a physics analysis case has been chosen: the search of $t\bar{t}$ resonances in ATLAS [12]. Our work has started from the datasets placed at the UCI machine repository [13] obtaining the first results from the application of several ML methods at the GRID infrastructures (IFIC Tier-2 & Tier-3).
So far, we have made the calculations in conventional CPUs but the final goal is to streamline the whole process with the use of GPUs. The gain using GPU will be really impressive and we are in the phase of acquiring a cluster of GPUs.
These techniques together with the SWAN (web-based analysis) developments are changing the way to undertake the physics analysis making them tremendously versatile.
SWAN is a platform to perform interactive data analysis in the cloud. The original SWAN project was developed, implemented and deployed at CERN, so that its backend combines state-of-the-art software technologies with a set of existing IT services supported by CERN resources. Basically, this service works as follows: SWAN is hosted by some kind of cloud instance (CERN OpenStack [14] in the original project, and LUSTRE in the local implementation at IFIC) so the web portal and the user notebook servers run in a virtual environment. A Docker [15] container that is allocated in the virtualized infrastructure creates these environments. On the one hand, the web portal is based on the JupyterHub [16] technology. On the other hand, the web-based interface of SWAN is based on Jupyter [17] notebooks (which allow to combine code, text and plots in the same document). In general, users log with their credentials and request the creation of their notebook server by means of the JupyterHub portal. The execution of notebooks is encapsulated in personalized Docker containers, all this performed in some cloud infrastructure. Users can perform, synchronize and share their analysis with the notebooks. Finally, when a Docker container is created, it's

also linked to other resources as software repositories and data storages so that all what is needed to perform any analysis is directly accessible from the personalized environment.
IFIC's effort to implement SWAN platform was mainly focused on adapting this service to the IFIC environment. Currently the IFIC implementation is in a "proof-of-concept" stage so that the service works properly but the full platform setup is still underway. Namely, SWAN service is not available to all of the IFIC's users and some requirements have to be satisfied in terms of security and privacy, but the project is quite mature.

8 ATLAS Core Computing

PIC and IFAE are involved in the Working Groups (WG) of the WLCG project to develop tools and services for the HEP Software Foundation [18].
They are participating in Data lake WG project where the goal is to define the functionality needed to implement a federated data center concept that aims to reduce the operational cost of storage for the HL-LHC, and at the same time better manage network capacity, whilst maintaining the overall CPU efficiency. This would include the necessary qualities of service, and options for regionally distributed implementations, including the ability to flexibly respond to model changes in the balance between disk and tape. For that purpose, PIC/IFAE took part in a coordinated effort to review tape management and test tape performance across Tier-1 sites (mainly BNL, FZK, INFN, TRIUMF): namely, two tape stress tests were performed respectively around the 20th of July and August 2018, then their results and related remarks were presented to ATLAS and other site administrators. Besides reporting metrics, profiling storage capacity and revising storage settings to boost performance, the tests helped acquire a better understanding of errors, bottlenecks, and margins of improvements on the side of other intermediary grid services (like Rucio and FTS), and drawbacks affecting dCache sites (due to potential sources of miscommunication between dCache and FTS), wich was pointed to the dCache developer team.
This survey was a necessary first step within a wider strategy adopted by ATLAS to deeply revise its computing model, at least for what concerns the data processing part, soon to be re-implemented according to the so-called "ATLAS Data Carousel model (tape performance test at Tier-1s)".
On the other hand, they are contributing to the Computing Resource Information System (CRIC) task force [19]. CRIC is an evolution of AGIS (i.e. ATLAS Grid Information System) and it will readily replace BDII (and possibly OIM) in providing a unified, consistent, and detailed view of sites' computing resources to experiments. So as an information system CRIC is still experiment oriented, but no longer ATLAS specific. By design CRIC will consume information from different information sources (GocDB, Rebus, EGI, HTCondor, etc. and for its bootstrap from BDII, OIM, SiteDB, etc.) and will be accessed by different users (site administrators, experiment coordinators, robots, etc.). Therefore PIC/IFAE, among other things, are mainly contributing to the implementation of various cron plug-ins to fetch data from different sources of information; and also, to the implementation of the authentication/authorization system based on different authentication methods (SSO, VOMS, SSL-certificate, etc.) and authorization policies.

9 Conclusions

We have presented the evolution and technical updates in the ATLAS Spanish Federated Tier2 and Tier1. Spanish ATLAS Tier-1 & Tier-2 infrastructure started in 2005 and it is offering 5% and 4% of all ATLAS Tier-1 & Tier-2 resources respectively.

ATLAS will need a factor of 5 more resources at HL-LHC with respect to today and flat budget and +20%/year from technology evolution fills part of this gap. Within this scenario, storage looks like the main challenge to address. The ATLAS Spanish LHC computing community is participating to establish the main guidelines for the next years (HL-LHC) and is moving towards common implementations: Data Management, Data Lakes, HPC, Resources Federations, Grid resources, Machine/Deep learning facilities, and increasing the network bandwidth.

This work was partially supported by MICINN in Spain under grants FPA2016-75141-C2-1-R, FPA2016-75141-C2-2-R and FPA2016-80994-C2-2-R, which include FEDER funds from the European Union. PIC is maintained through a collaboration of CIEMAT and IFAE, with additional support from Universitat Autònoma de Barcelona and ERDF. The authors thankfully acknowledge the computer resources at MareNostrum and the technical support provided by BSC (FI-2017-3-0031, FI-2018-1-0019).

References

1. T. Wenaus, D. Costanzo, A. Di Girolamo, *How HL-LHC challenges inform workload management R&D:An ATLAS view*, (WLCG&HSF Workshop, Naples, 2018)
2. Á. Fernández, J. M. Orduña, and S. González, *Performance Improvements of an Event Index Distributed System*: Extended Abstract. In Proceedings of 47th International Conference on Parallel Processing, (ICPP 2018, Oregon)
3. A. Fernandez Casani, Dario Barberis, J. Sánchez, C. García Montoro, S. González de la Hoz, *Distributed Data Collection for the Next Generation ATLAS EventIndex Project.* In Proceedings of 23rd International Conference on Computing in High Energy and Nuclear Physics, (CHEP 2018, Sofia)
4. S. Campana, *Distributed Computing in LHC Run2*, J.Phys:Conf. Ser. 664 032005 (2015)
5. G. Kurtzer, V. Sochat, M. Bauer, *Singularity: Scientific containers for mobility of compute.* PLoS One 12(5) : e0177459 (2017)
6. M. Ellert et al., *Advanced Resource Connector middleware for lightweight computational Grids.* Future Generation Computer Systems 23 219-240 (2007)
7. http://lhcone.web.cern.ch
8. http://lcgdm.web.cern.ch/dynafed-dynamic-federation-project
9. X. Espinal, *Data Lake R&D,* (Joint WLCG&HSF Workshop, Naples, 2018)
10. D. Piparo, E. Tejedor, P. Mato, L. Mascetti, J. Moscicki, M. Lamamnna, *SWAN : a Service for interactive Analysis in the Cloud*, CERN-OPEN-2016-005 (2016)
11. D. Rousseau, *Machine Learning in HEP.* In Proceedings of 23rd International Conference on Computing in High Energy and Nuclear Physics, (CHEP 2018, Sofia)
12. *Searching for Exotic Particles in High Energy Physics with Deep Learning.* Nature Commun., 5.4308 (2014)
13. http://archive.ics.uci.edu/ml/datasets/HEPMASS ,2015
14. OpenStack: http://www.openstack.org
15. Docker containers http://www.docker.com
16. JupyterHub portal: http://github.com/jupyter/jupyterhub
17. Jupyter notebools: http://jupyter.org
18. J. Albrech et al., *A Roadmap for HEP Software and Computing R&D for the 2020s*, arXiv:1712.06982 [physics.comp-ph] (2017)
19. A. Anisenkov, *The ATLAS Grid Information System (AGIS) evolution for other communities.* In Proceeding of 26th International Symposium on Nuclear Electronics & Computing, (NEC 2017, Montenegro)