The LHC Tier-1 at PIC: ten years of operations

J Flix\textsuperscript{1,2}, E Acción\textsuperscript{1,3}, V Acín\textsuperscript{1,3}, C Acosta\textsuperscript{1,3}, G Bernabeu\textsuperscript{1,2†}, J Casals\textsuperscript{1,2}, M Caubet\textsuperscript{1,2}, R Cruz\textsuperscript{1,3}, M Delfino\textsuperscript{1,4}, X Espinal\textsuperscript{1,3‡}, J M Hernandez\textsuperscript{2}, F López\textsuperscript{1,2}, G Merino\textsuperscript{1,2§}, A Pacheco Pages\textsuperscript{1,3}, A Pérez-Calero Yzquierdo\textsuperscript{1,2}, E Planas\textsuperscript{1,3}, M Porto\textsuperscript{1,2}, B Rodríguez\textsuperscript{1,3}, and A Vedaee\textsuperscript{1,3}

E-mail: jflix@pic.es

\textsuperscript{1} Port d’Informació Científica (PIC), Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Spain
\textsuperscript{2} Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), Madrid, Spain
\textsuperscript{3} Institut de Física d’Altes Energies (IFAE), Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Spain
\textsuperscript{4} Universitat Autònoma de Barcelona, Department of Physics, Bellaterra (Barcelona), Spain

This paper summarizes ten years of operational experience of the WLCG Tier-1 computer centre at Port d’Informació Científica (PIC), which serves the ATLAS, CMS and LHCb experiments. The centre, located in Barcelona (Spain), has supported all of the commissioning activities before the Large Hadron Collider (LHC) produced real collisions, it has continuously adapted to the new requirements, introducing new technologies as they became available, and it has grown significantly in resources offered to the LHC experiments, while maintaining top reliability levels as compared to other centres. Additional work has been done in the last years to reduce the power costs of the Tier-1 centre, and to prepare for the next challenges that are expected to come. Some thoughts on how the current WLCG system could evolve are also presented.
1. Introduction

The Large Hadron Collider (LHC) [1], located in the European Laboratory for Particle Physics (CERN - Geneva, Switzerland) [2], is the largest scientific instrument ever built. First collisions started by November 2009 and, since then, the LHC experiments have generated about 200 PB of raw data. Additional processed and simulated data have been generated, adding up to a total of 500 PB since the LHC started operations (Run1 [2010-2012] and Run2 [2015→]). Taking into account its expected lifetime, the LHC will be first scientific experiment ever at the Exabyte scale by 2020. Moreover, this huge amount of data is analysed by thousands of physicists around the world, which poses an additional challenge to the computing infrastructure.

The Worldwide LHC Computing Grid (WLCG) [3][4] is an innovative worldwide distributed system with more than 170 computing centres in 42 countries, which provides uniform access to resources through Grid services. The WLCG project started at CERN in 2002 with the objective of building and maintaining the computing infrastructure to store, process and analyse the data gathered by the LHC experiments. The computing centres are functionally classified in Tiers. The Tier-0, split between sites at CERN and Wigner (Budapest), receives the raw data from the experiments, records them on permanent mass storage and performs a first-pass reconstruction of the data. The raw and reconstructed data are distributed via a dedicated high-speed network (10-200 Gbps) to the Tier-1 centres, which provide definitive data reconstruction, filtering and permanent storage and management of data needed during the analysis process. Tier-1 centres also perform data-intensive analysis and re-processing, and may undertake national or regional support tasks, as well as contribute to Grid Operations Services. Tier-2 centres concentrate on simulation production, group and end-user analysis.

Twelve of the WLCG centres are identified as Tier-1 centres, which have to provide a 24x7 highly performing and reliable service, critical to the LHC operations and governed by the WLCG MoU [5], which defines the service levels to be delivered by these centres and determines the resources to be provided by each centre over a three-year period. Countries hosting a Tier-1 centres today are: Canada, France, Germany, Italy, Korea, Netherlands, the Nordic Countries, Russia, Spain, Taiwan, UK and the USA. The Spanish Tier-1 centre is deployed at Port d’Informació Científica (PIC) [6], a computing site located in the campus of the Universitat Autònoma de Barcelona, near the city of Barcelona. PIC provides services to three of the LHC experiments, ATLAS [7], CMS [8] and LHCb [9]. It accounts for 5% of the total Tier-1 resources of ATLAS and CMS, and 6.5% for LHCb, being the reference Tier-1 for Tier-2 centres in Spain and Portugal, and sites located in Valparaiso (Chile) and Marseille (France).

2. Ten years of services deployment

PIC has been and continues to be an active and successful participant in WLCG since its start [10][11]. Before the LHC data taking phase, PIC contributed to prototyping and testing of the Grid middle-ware and services that were being developed. PIC successfully participated in the Service Challenges carried out by the experiments, testing campaigns aimed to progressively ramp-up the level of load on the infrastructure under conditions as realistic as possible, achieving breakthrough performances, and showing its readiness for the LHC data taking period.
PIC has shown good performance results during Run1, Long Shutdown 1 (LS1; 2013-2014), and half of Run2, so far. During this period, it worked to be fully compliant with the requirements from the experiments, tendered and provisioned resources for data taking re-starts, and prepared all of its infrastructure for the resources increment that are expected in the next years. Figure 1 shows the PIC resources growth and usage since January 2006. PIC resources are expected to double by the end of the LHC Run2 (end of 2018), as compared to those deployed by the end of Run1. Figure 2 shows the LHC schedule since 2006, with the periods for Shutdown/Technical Stops, Data Taking, and Commissioning activities.

Figure 1: CPU (top), disk (center) and tape (bottom) resources installed (black), pledged (grey), and used (ATLAS: green; CMS: blue; LHCb: red - cumulative) at PIC, from January 2006 to December 2016.

1The Tier-1 runs in a shared computing facility. Sometimes the used CPU can be higher than the installed capacity.
By the end of 2016, the computing resources installed in PIC comprise nearly 6000 cores managed by Torque/Maui batch system and scheduler, corresponding to about 75000 HEP-SPEC06 (HS06, see Ref. [12]). The servers are typically two quad-core x86 CPUs, with at least 2GB RAM per core (recent CPU purchases are equipped with 4GB/core). Each of these nodes typically has two 10Gbps Ethernet interfaces, which are then aggregated in switches and connected to the storage infrastructure. The servers consist of Blades (HP) and Dual-Twins (Dell, Intel in oil).

The storage service at PIC is managed by dCache [13] and Enstore [14] softwares. The dCache software provides uniform and efficient access to the disk space provided by multiple file servers, and interacts with the Enstore software to interface to magnetic tape storage. By the end of 2016, 7 PB of disk space is installed, by means of around 2000 hard disks of 2, 3, 6 and 8 TB (He), distributed on around 50 servers based on x86 technology, each connected by one or two 10Gbps Ethernet, depending on the hardware. The servers brands comprises DataDirect, SGI, Intel, Dell,
The current tape infrastructure at PIC is provisioned through a Sun StorageTek 8500SL library, providing around 6650 tape slots, which are expected to cover the PIC tape needs for the coming years. By the end of 2016, Enstore manages 17 PB of tape storage, with access to a total of $\sim$8 million files. The supported technologies are LTO-4, LTO-5, T10KC, and T10KD containing 11%, 7%, 48% and 34% of the total data, respectively, in around 3800 tape cartridges. A total of 28 tape drives are installed to read/write the data (11 LTO-4, 4 LTO-5, 8 T10KC and 5 T10KD). Aggregated read/write rate has achieved hourly average rates peaking at 3.5GB/s.

2.1 2006-2016: PIC timeline of major milestones

Figure 3 summarizes the major achievements at PIC in terms of new technologies, new services, new network elements, development activities, and major infrastructure changes introduced, for the period from 2006 to 2016. The PUE\(^2\) improvement along the years is also displayed. Note that the PIC centre has only suffered two major electrical incidents (blackouts) in this ten years period, which were solved within a day. The following list highlights the most important events that occurred in each year.

- **2006**
  - **Technologies**: Introduction of 9940B tape cartridges (200 GB/tape). Disk servers with 250 GB HDDs deployed. Multi-core CPUs available (Intel Xeon 5160 - 2 cores/CPU)
  - **Services**: dCache deployment
  - **Network**: Installation of the main router Cisco Catalyst 6509. 1 Gbps WAN network
  - **LHC**: CMS and ATLAS cosmics data acquisition. Data challenges: DC06 (LHCb), CSA07 (CMS). Started 24x7 support for the experiments

- **2007**
  - **Technologies**: Introduction of LTO3 tape cartridges (400 GB/tape). Disk servers with 500 GB HDDs deployed. Installation of a IBM TS3500 tape library
  - **Services**: Deployment of Enstore (as a replacement of CASTOR)
  - **Network**: Upgrade to 10 Gbps WAN network
  - **Development**: LTO3 Enstore driver for the IBM TS3500 tape library
  - **LHC**: CMS and ATLAS cosmics data acquisition. Data challenges: DC06 (LHCb), CSA07 (CMS). Started 24x7 support for the experiments

- **2008**
  - **Technologies**: Introduction of LTO4 tape cartridges (800 GB/tape). New CPUs: Intel X5355 (4 cores/CPU) and Intel E5450 (4 cores/CPU). Installation of a StorageTek SL8500 tape library
  - **Network**: VLANs deployed. VPN established

\(^2\)Power Usage Effectiveness (PUE), introduced the Green Grid consortium, is the ratio of total amount of energy used by a computer data center facility to the energy delivered to computing equipment.
- **LHC**: Data challenges CSA08 (CMS) and global CCRC08

- **2009**
  - **Technologies**: Disk servers with 2 TB HDDs deployed. New CPUs: Intel L5420 (4 cores/CPU)
  - **Services**: Deployment of RT. Puppet adoption
  - **Network**: 2x firewalls Nokia IP1280 deployed
  - **Development**: CMS Site Readiness
  - **LHC**: Global Data challenge STEP09

- **2010**
  - **Technologies**: Introduction of LTO5 tape cartridges (1.5 TB/tape). New CPUs: Intel L5530 (4 cores/CPU)
  - **Network**: 2x ARISTA 7148SX (10 Gbps) deployed. Adoption of fibers elsewhere: Siemon @10 Gbps
  - **Development**: CVMFS for LHCb
  - **LHC**: Start of data taking (Run1)
  - **Infrastructure**: New chillers deployed in the main computing room. Compact IT room deployed (80 KVA)

- **2011**
  - **Technologies**: Introduction of T10KC tape cartridges (5 TB/tape). New CPUs: Intel X5650 (6 cores/CPU) and Intel X5645 (6 cores/CPU)
  - **Network**: LHCONET deployment
  - **Infrastructure**: Detailed study to improve the energy efficiency of the centre

- **2012**
  - **Technologies**: Disk servers with 3 TB HDDs deployed. New CPUs: Intel E5-2650 (8 cores/CPU)
  - **Services**: XrootD. LDAP. OpenNebula
  - **Network**: Installation of the main router Nexus 7009. PerfSONAR-PS deployed
  - **Development**: LHCb VMDIRAC

- **2013**
  - **Technologies**: Disk servers with 4 TB HDDs deployed
  - **Services**: NFS 4.1. Joined ATLAS FAX and CMS AAA data federations. SLURM (EGI validation). ELK. Red Hat Enterprise Virtualization. KVM
  - **Development**: HEPIX IPv6 Working Group
• 2014
  – Technologies: Disk servers with 6 TB HDDs deployed. New CPUs: Intel E5-2640v3 (12 cores/CPU) and Intel E5-2650v2 (12 cores/CPU) - 4 GB/core RAM
  – Development: MCFLOAT (efficient multi-core jobs handling under Torque/MAUI)
  – Infrastructure: New UPS (550 KVA IGBT). Free-cooling at the main computing room

• 2015
  – Technologies: Introduction of T10KD tape cartridges (8 TB/tape). Disk servers with 8 TB HDDs Helium deployed
  – Network: 2x Firewalls Fortigate F1500
  – Development: CMS dCache Middleware readiness (validation)
  – Infrastructure: Oil immersion techniques deployed (CPU - Green Cooling Revolution)

• 2016
  – Network: IPv6 enabled in production. Upgrade of WAN network to 20 Gbps

2.2 Reducing operational, maintenance and power costs at PIC

In order to pave a secured and efficient road for future site growth, PIC invested efforts in the 2013-2015 period in order to reduce maintenance, operational and power costs at the site.

In this period, the PIC computing farm power was adjusted to electricity costs, less CPU power being offered during high cost periods, and vice-versa, without affecting the annual WLCG pledges. Since the power consumption per unit of performance (Watts/HS06) has been approximately constant for the last years, the oldest CPU servers were switched on/off to modulate the computing power. The net effect was a reduction of $\sim 10\%$ in energy costs for these years.

At the end of Run1, PIC deployed a RedHat Enterprise Virtualization system (RHEV 3.4.2, KVM-based). The production system is installed over an HP Blade box with seven hypervisors, each of them equipped with 16 cores and 96GB RAM (HP Proliant BL460c) with 2x10GbE ports. The hypervisors are connected via fiber-channel to a NetApp FAS3220 (2TB, Thin Provisioning, with qcow2 image formats). Around 160 services run in the system, which reduces the number of required physical machines by a factor 10 without any impact on the reliability and performance of the services, reducing the costs substantially. In order to save licence costs, Ovirt 3.5 is being tested at scale. Four hypervisors, in a similar environment, are used to run up to 60 test services. The new setup is expected to be deployed in production along the 2017 year.
The LHC Tier-1 at PIC: ten years of operations

J Flix

Figure 3: PIC centre timeline for the period 2006-2016, showing the introduction of new technologies, services, deployments, and developments, as well as the retirement of some services and site elements. Figure is split in half to ease of readability.
Constant efforts are made to simplify and improve configuration management and automation. The majority of the services at PIC are managed by Puppet [15]. The implementation is flexible enough to rapidly evolve following changing technologies. Currently, all of the existing code is being migrated to Puppet 4. Local repositories for code development projects at PIC have been migrated from SVN to GIT, which eases the use and maintenance of code. Thanks to the large automation of services, PIC is operated with less manpower than an average Tier-1 centre [16].

Cooling is a natural place to save costs for a large computing centre. Given the growth expectations, and coinciding with the LHC shutdown period LS1, PIC improved the energy efficiency of its main computing room. This occurred during fifteen consecutive weeks of work in 2014, without any downtime, interruption or negative impact in the operations (see [17] for more details). Before this intervention, there was no separation of cold and hot air in the main computing room. Several CRAH’s (Computer Room Air Handler) managed air through a cold water battery, injecting air at 14°C to get a room temperature of 22-23°C. This system showed a PUE of 1.8, offering significant room for improvements.

Three free-cooling units, acting as indirect heat exchangers with outside air and equipped with adiabatic cooling humidifiers, replaced a number of CRAH units. Direct free-cooling was not

![Figure 4: Free-cooling units being installed in PIC, in mid June 2014.](image)

![Figure 5: PIC main computing room before (left) and after (right) completion of hot/cold air separation and confinement.](image)
considered, as the region where PIC is located has high humidity and dust concentration values. Figure 4 shows some photographs of the free-cooling units installation. Hot and cold air flows are separated, and a lower ceiling was installed in order to contain the hot air from the room, as can be seen in Figure 5. Additionally, the inlet temperature was increased to 20°C, according to ASHRAE recommendations [18].

This work was completed in September 2014. A period of one year was defined to study and adjust the system in order to reach the maximum energy efficiency. Dedicated monitors for the most important climate parameters were installed and helped to achieve the optimum performance of the installation. The yearly averaged PUE of the main computing room is 1.3. In light of this performance, electricity costs savings in 3.5 years amortize the new cooling installation.

At the end of 2014, the UPS system of the centre was at the end of its lifetime, presenting power losses up to 15%. A new UPS of 550 kVA, with Insulated-Gate Bipolar Transistor (IGBT) technology that provides efficiency in the range of 97%-99%, was installed at the beginning of 2015, impacting positively on the electricity bill of the centre. By the end of 2016, around 300 kW of IT equipment is installed in PIC.

![Figure 6: PIC members installing the first CPU units into the oil tanks from Green Cooling Revolution, at the beginning of 2016.](image)
Four Green Revolution Cooling oil immersion tanks [19] with specially designed oil to water heat exchangers were installed in the 25 m² fireproof enclosure in the basement of PIC building at the end of 2015. This new installation has a water loop that runs to the building’s roof where a redundant, modular dry cooler is installed. Each tank can hold 45U (rack units) of IT equipment and, more importantly, dissipate up to 45 kW of heat, an unprecedented 1000 Watts per 1U. The CPU servers compatible with immersion have shown prices very similar to air-cooled equipment. All of the new CPU purchases are installed in this facility, with an estimated PUE of 1.05. Figure 6 shows the installation of the first CPU units in the oil-filled tanks.

3. A reliable, high-capacity service

One of the main characteristics of Tier-1 centres, beyond a very large storage and computing capacity, is to provide these resources through services that need to be extremely reliable. Being closely connected to the detectors’ data acquisition, a maximum time for unintended interruption of the services in a Tier-1 is set to 4 hours, and a maximum degradation of Tier-0 to Tier-1 data acceptance of 6 hours. Critical services in a Tier-1 operate in 24x7 mode.

Service quality and stability are amongst the cornerstones of WLCG, therefore they are closely tracked by monitoring two metrics provided by the SAM monitoring framework: site availability and reliability. These are built from dozens of sensors, for each of the experiments, which hourly probe all of the Grid services at the site, ensuring peer pressure and guaranteeing that the reliability of WLCG services keep improving [20].

Figure 7: Tier-1 site reliabilities since 2006. PIC Tier-1 (green) monthly results can be compared to the average Tier-1 reliabilities (grey) as well as with the WLCG target for site reliability (red area), which is set to 97% since 2009, according to the WLCG MoU.

3This enclosure was previously used for an improved efficiency air-cooled room, hosting around 70% of the available CPU resources at the centre.
Figure 7 shows the monthly reliability results for PIC, the average results for the Tier-1s, and the WLCG target, which is set to 97% since 2009. The set of SAM tests was common to all of the LHC experiments, but since January 2014 the experiment dedicated tests are used instead (Figure 7 shows the average of PIC for the ATLAS, CMS and LHCb experiments, as well as the Tier-1 average for sites supporting these experiments). PIC is at the top of the reliability measures, as compared to other Tier-1s, and well above the WLCG target. These good reliability values were obtained during periods in which many new services were being deployed and interventions were made to improve the PIC site efficiency. PIC has an expert contact person per experiment on site (the liaison), communicating and coordinating priorities with the experiments and resolving operational problems, as well as a highly professional team that operates the services. This allows PIC to remain at top reliability and stability levels.

4. Current R&D projects

The PIC Tier-1 team is continuously integrating new technologies and services to improve the performance of the centre and adapting it to new requests from the LHC experiments. Some of these changes are strategic, since they will allow the center to integrate new services with less effort. There is a portfolio of R&D activities that are currently being carried out at PIC. The most important activities are:

- Testing the transition to IPv6 for middleware components, and experiment services [21], under the HEPIX IPv6 Working Group. Many services at the site have been already migrated to dual-stack.

- Coordinating the development and integration of new technologies and tools, such as multicore processing in most of the WLCG centers [22][23], and user data [24] and analysis workload [25][26] management tools, to mention a few. Additionally, high-memory queues have been deployed for HL-LHC studies which are nowadays being carried out by the experiments.

- Testing new fast-benchmarks for CPUs, which will later turn into ideas to find a replacement for HS06, under the HEPIX CPU benchmarking Working Group [27].

- Deployment of HTCondor [28] and HTCondor-CE, which will provide PIC with a flexible, robust and modern job scheduler.

- Flexible scalability approaches to provision and access commercial clouds, and use of opportunistic resources from large supercomputing centers (HPC), are being investigated.

- Participation in the HNSciCloud EU pre-commercial procurement project [29], based on a hybrid cloud schema. This is a new exploratory project for provisioning resources for the LHC and many other disciplines, which is an interesting approach towards future resources provisioning via multi-national joint procurements of Cloud resources.

- Federation of the biggest Grid sites in Spain, PIC Tier-1 and IFAE/Ciemat Tier-2s, who typically account for 60%-70% of total CPU used by WLCG in the region, providing an
unique environment to deploy an Analysis Facility for the Spanish LHC scientists, with read-access to all of the Tier-1 and Tier-2 data available for ATLAS, CMS and LHCb in the integrated site.

5. Getting prepared for the next 10 years

The Tier-1 at PIC must be prepared for many performance optimizations through software and integration and deployment of new (or improved) services that are expected to come. This requires continuous exploration of the market, profiting from new technologies as soon as they become available. The system deployed should be flexible enough to quickly adapt to the evolution of WLCG.

The LHC experiments are now exploiting all of the available resources, regardless of the WLCG tiered structure. Some experiment workflows which were exclusively run at the Tier-1 centers, are now extended to run in big and stable Tier-2 sites. The tendency is that the tiered structure will vanish, as well as some small sites that are offering resources to WLCG, which cannot cope with the rapid growth of resources that are demanded by the experiments. The countries might be suggested to provide optimized contexts to cover the whole spectrum from data processing, simulation and analysis of the LHC data, by means of federating sites or deploying resources in big centers exclusively. This Grid infrastructure would become completely integrated with cloud and opportunistic resources on demand (commercial Clouds and HPC sites), where some special workflows could be executed, accommodating for peaks of demands.

Fitting LHC computing needs within constrained costs, keeping up with the expected (high) growth of LHC computing resources, will be a real challenge, in particular during the HL-LHC era (∼2025), where the expected data delivered will be a factor x100 as compared the current LHC data volumes, per year. Technology trends will not be sufficient to cover this gap, hence the whole WLCG concept and the Computing Models of the LHC experiments will need a substantial evolution. Other scientific disciplines in the HEP and Astro-particle domains will also demand large computing resources by that time, hence evolving to a HEP-wide scientific data and computing environment is a must. A hub of knowledge and expertise will be needed to coordinate the technological development, and even integrate data and computing sites into a virtual centre, which can be more easily and transparently used by the diverse Scientific Communities.

6. Conclusions

PIC has been successfully operated as a Tier-1 in WLCG for about ten years. During this time, the center has deployed new services as they were required and it has provided its share of resources to cope with the increasing resource requirements demanded by the LHC experiments. PIC Tier-1 has been instrumental during the commissioning of WLCG and the services that have been included into the system, particularly prior data taking or in the LHC shutdown period LS1.

PIC Tier-1 has always demonstrated its readiness for the next LHC data-taking period. Benefiting from the first LHC shutdown, PIC improved its computing infrastructure, with new elements that translate in a significant reduction of the electricity costs for the next years. This work strengthens the centre for the following challenges it will be facing. In particular, the PIC team members
are very active in the computing teams of the LHC experiments, in WLCG Operations and in the HEPIX Working Groups, where many R&D activities for the future are being coordinated and developed. The paradigm and the context in ten years from now will be very different. PIC focus on staying aligned with the coordinated strategy to fulfill the goals and successfully continue supporting LHC computing activities at scale in Spain.

Acknowledgements

This work was partially supported by MINECO in Spain under grants FPA-2005-08446-C02-01, FPA2007-66152-C02-02, FPA2010-21816-C02-00, FPA2013-48082-C2-1/2-R, and FPA2016-80994-C2-1-R, which include ERDF funds from the European Union. PIC is maintained through a collaboration of IFAE and CIEMAT. The support of the ERDF fund 2007-13 is gratefully acknowledged.

References

[16] Optimising costs in WLCG operations, A. Sciabà et al, in these proceedings


[27] New CPU benchmarks: https://twiki.cern.ch/twiki/bin/view/HEPIX/CpuBenchmark

[28] HTCondor: https://research.cs.wisc.edu/htcondor

[29] HNSciCloud project: http://www.helix-nebula.eu/about-hnscicloud