PROCEEDINGS OF SCIENCE

Evolution of the ATLAS data and computing model for a Tier-2 in the EGI infrastructure

Álvaro Fernández Casaní
IFIC (CSIC/UV)
Edificio Institutos de Investigación – 22085, E-46071 Valencia, Spain
E-mail: Alvaro.Fernandez@ific.uv.es

Miguel Villaplana Pérez
IFIC (CSIC/UV)
Edificio Institutos de Investigación – 22085, E-46071 Valencia, Spain
E-mail: Miguel.Villaplana@ific.uv.es

Santiago González de la Hoz
IFIC (CSIC/UV)
Edificio Institutos de Investigación – 22085, E-46071 Valencia, Spain
E-mail: Santiago.Gonzalez@ific.uv.es

José F. Salt Cairols
IFIC (CSIC/UV)
Edificio Institutos de Investigación – 22085, E-46071 Valencia, Spain
E-mail: salt@ific.uv.es

Farida Fassi
IFIC (CSIC/UV)
Edificio Institutos de Investigación – 22085, E-46071 Valencia, Spain
E-mail: Farida.Fassi@ific.uv.es

Mohammed Kaci
IFIC (CSIC/UV)
Edificio Institutos de Investigación – 22085, E-46071 Valencia, Spain
E-mail: kaci@ific.uv.es

Alejandro Lamas
IFIC (CSIC/UV)
Edificio Institutos de Investigación – 22085, E-46071 Valencia, Spain
E-mail: Alejandro.Lamas@ific.uv.es

PoS(ISGC 2012)011

PROCEEDINGS OF SCIENCE

Elena Oliver
IFIC (CSIC/UV)
Edificio Institutos de Investigación – 22085, E-46071 Valencia, Spain
E-mail: Elena.Oliver@ific.uv.es

Javier Sánchez
IFIC (CSIC/UV)
Edificio Institutos de Investigación – 22085, E-46071 Valencia, Spain
E-mail: Javier.Sanchez@ific.uv.es

Victoria Sánchez-Martínez
IFIC (CSIC/UV)
Edificio Institutos de Investigación – 22085, E-46071 Valencia, Spain
E-mail: Victoria.Sanchez@ific.uv.es

On behalf of the ATLAS Collaboration

Since the start of the LHC pp collisions in 2010, the ATLAS computing model has moved from a strict hierarchical design, where every Tier-2 had a liaison and a network dependence on a Tier-1, to a more meshed approach with direct connections between Tiers. Evolution of ATLAS data models requires changes in ATLAS Tier-2 policy for the data replication, dynamic data caching and remote data access. It also requires rethinking the network infrastructure to enable any Tier-2 and associated Tier-3 to easily connect to any Tier-1 or Tier-2. The Tier-2 disk space is used for real, simulated, calibration and alignment, group, and user data. A cache disk space is needed for input and output data for simulations and production jobs. A number of concepts and challenges are raised in these proposals, and in this contribution we show how these changes affect an ATLAS Tier-2 and its co-located Tier-3 as a part of the EGI infrastructure. We will present the Tier-2 and Tier-3 facility setup, the data distribution and the arrangements proposed to fulfil the requirements coming from the new model, especially the requirement for any site to be the source for data replication. A given site can receive datasets from any other site "on demand", based on usage patterns, and possibly using a dynamic placement of datasets managed centrally and unused data can be centrally removed as well. The data access for users either with grid or local tools will be presented, using the EGI infrastructure and procedures, and the middleware glite flavour that is being provided by EMI releases. We use an example of a real physics analysis to show how users are working, to check the readiness of the tools and its performance within the changes being adopted coming from the evolution of the model.

The International Symposium on Grids and Clouds (ISGC) 2012
Academia Sinica, Taipei, Taiwan
February 26 – March 2, 2011

PoS(ISGC 2012)011

1. Introduction to the ATLAS data challenge

ATLAS is one of the experiments that have been working within the program of LHC (Large Hadron Collider) [1]. The proton beams provided by the LHC collide in ATLAS with an energy of 3.5TeV each, for a total collision energy of 7 TeV. The integrated luminosity recorded at the end of 2011 was 5.25 fb^{-1}. The ATLAS experiment has exported more than 25k TB of data to Tier-1s since January 2010 as shown in figure 1.

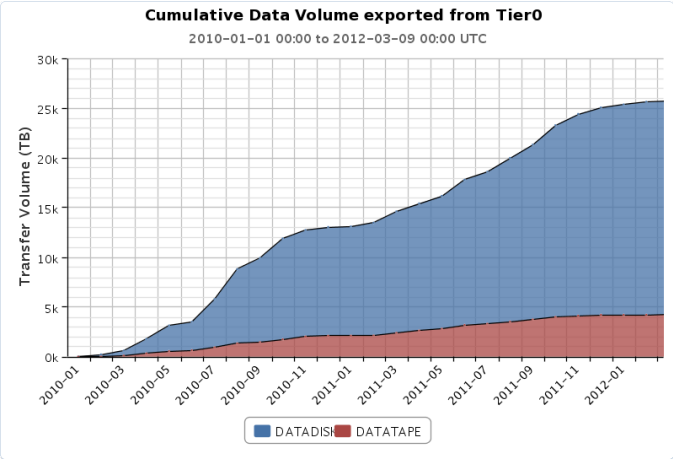

Figure 1: Cumulative data volume exported from ATLAS Tier-0 since January 2010.

Processing such volume of data has been possible thanks to the computing model based on GRID technologies [2] used by the LHC experiments. The model handles everything from the storage of raw events at CERN to physics analysis on refined data at home institutes. The ATLAS Collaboration has established the responsibilities of the various centres of the different countries for the reliable operation of the computing facilities. A proper sizing and organization of the resources, an efficient mechanism to access the data or robust algorithm development are examples of key items in the so-called solution for the steady-state period during the data taking.

1.1 IFIC Spanish Tier-2

The Worldwide LHC Computing Grid project (WLCG) groups the different types of computing centres in a tiered hierarchy that ranges from the Tier-0 at CERN to the 10 Tier-1 centres and the 40 Tier-2 centres distributed world wide. The roles of the different tiers are well established in order to produce a stable and efficient distributed computing facility for the ATLAS experiment [3].

High energy physics (HEP) groups from 3 Spanish institutions are participating in the ATLAS experiment. They are the *Institut de Física d'Altes Energies* (IFAE) at Barcelona, the *Universidad Autónoma de Madrid* (UAM) and the *Instituto de Física Corpuscular* (IFIC) of Valencia. The ATLAS Spanish Tier-2 consists of a federation of these three institutions that represents around 5% of the total ATLAS resources [3]. IFIC represents 50% of the ATLAS

Spanish resources and has the responsibility to coordinate the activities of the Spanish Tier-2 federation. In table 1 the evolution of IFIC Tier2 resources is shown.

Table1: Pledged evolution of IFIC resources..

Year	2010	**2011**	2012	2013
CPU(HS06)	6000	**6950**	6650	7223
DISK(TB)	500	**940**	1175	1325

The numbers in 2011 (bold) correspond to the current deployed resources

In 2011 IFIC has already reached 1 PB of disk storage, the total Spanish ATLAS Tier-2 reached 2 PB. In February 2012, IFIC has 6950 HS06 of CPU capacity and 940 TB of disk space.

Tier-2s with a good network connection (T2Ds) are allowed to connect to other Tier-1 sites and to Tier-2 from a different cloud. The CPU resources are more efficiently used this way and high priority tasks can be done more quickly. IFIC Tier-2 is considered to be a T2Ds and is now working as a multi-cloud site. Therefore, there are direct transfers from/to all ATLAS Tier-1.

2. Evolution of the Atlas data and computing model

2.1 Flattening the model to a Mesh

Experience during the first year of operation of the model described above has been found insufficient. Some of the tiers were not fulfilling the required level of utilization, especially Tier-2 like the one at IFIC. This situation was mainly due to two issues, being one of them operational and the other related the fact that the data flows among the Tier-2 were limited.

On the operational level, one question that limited the usability was the dependency on Tier-1 in order to run jobs. As a consequence, when our Tier-1 was in scheduled downtime the associated Tier-2 were affected.

On the other hand, job input files, as well as IFIC job outputs had to be transferred to associated Tier-1. For example, if some data produced at IFIC was needed in a Tier-2 at another cloud, dataset had to be transferred via source and destination associated Tier-1.

The solution to improve this situation was taken by ATLAS in the form of flattening the model from a tier to a mesh. Now Tier-2 can directly exchange data with Tier-1 of different clouds, and even with other Tier-2 globally.

2.2 Connectivity

Network is a key component in the evolution of the ATLAS model for the Tier-2, as they have to be well connected to be able to exchange data. In order to test connectivity, ATLAS monitors transfers among sites through analysis of the FTS (File Transfer Service) log. With this monitoring we can check for example how the transfers perform from and to a Tier2 from every other site, within the complete NxN matrix containing all the transfers. The transfers are divided in small files (<1 Mb) and big files (>1 Gb) to split transfer and initialization

components, and if they reach the minimum rates, then the site is considered Tier2 Directly connected (T2D).

During latest weeks IFIC has fulfilled transfer performance requirements, even though fluctuating values for bigger files and transfer to remote sites have been observed. For this reason, we have installed new GridFTP servers. As the transfers are now distributed among more servers IFIC is getting more stable values. Thank to this, IFIC obtained a T2D status.

2.3 Availability

Another key attribute for a Tier-2 site is high availability. ATLAS uses the HammerCloud [12] framework to test site availability, by constantly submitting typical analysis jobs to every site. If there is a minimum of 3 jobs in a row that fails, among other requirements, then the site is detected as a problematic one and it is automatically blacklisted. Then the site enters in a test mode, and it will go back online as soon as test jobs finish correctly. With this behavior, the idea is to protect user jobs from problematic sites, detecting them as soon as possible.

Sites are evaluated monthly and ranked in 4 groups (Alpha, Bravo, Charlie and Delta [7]) depending on its availability and its connectivity. A Tier2 to be qualified as an Alpha Site should be well connected (T2DS), and available more than 90% of the time. During last months IFIC has been qualifying as Alpha Site for fulfilling these requirements.

2.4 Dataset replication

Being an Alpha T2D site, IFIC is now able to replicate higher amounts of datasets. As can be seen in Figure 3, the ES-Cloud Tier-2 sites are getting now more datasets than Tier-1 (PIC) while, in the past, it used to be the Tier-1 the one getting more datasets.

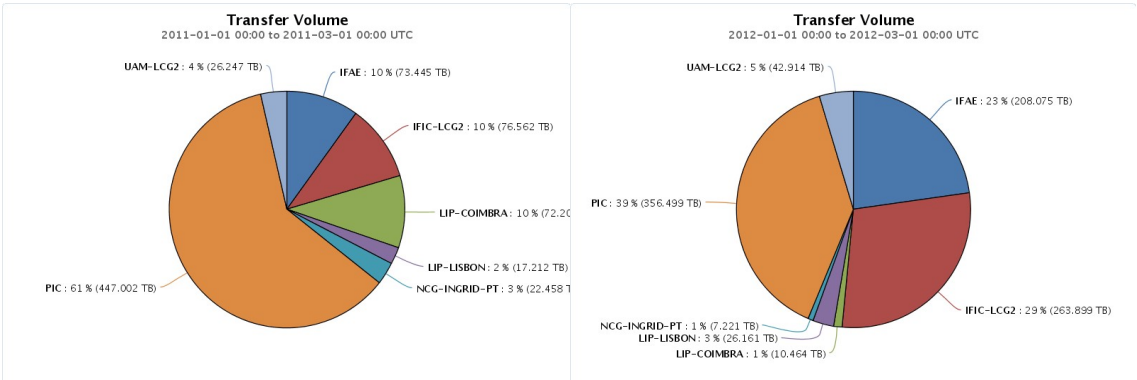

Figure 3: Data transfer volume in the ES-Cloud sites during January and February 2011(left) and 2012(right).

3. IFIC Infrastructure and operations

3.1 Computing resources

IFIC computing resources are distributed in two areas: the resources devoted exclusively to the ATLAS, and the ones provided within the Grid-CSIC project by the Spanish Research Council. The latter resources can be used for a variety of projects, and currently the usage follows the distribution of 25% for IFIC users, 25 % for CSIC projects, 25% for EGI Virtual Organizations, and 25% for Iberian Grid virtual organizations, that are also included in EGI.

Resources available to ATLAS in February 2012 are 640 cores distributed in the following CPUs:

- 48 x Xeon Quadcore E5472 3.00 GHz with 16GB RAM
- 32 x Xeon Quadcore E5520 2.77 GHz with 24GB RAM

Grid-CSIC resources comprises 1232 cores distributed as follows:

- 48 x Xeon Quadcore E5472 3.00 GHz with 16GB RAM, with Infiniband network devoted to MPI jobs.
- 106 x Xeon Quadcore E5472 3.00 GHz with 16GB RAM.

3.2 Storage resources

During last year we hit the Petabyte of storage for ATLAS at Ific institue, plus 180 TB more for the rest of the supported VOs. Initially was chosen 6 Sun Thumpers Disk Servers X4500, equipped with a combination of 500GB and 1 TB disk to provide 112 Tb of space. Adding the 13 newer X4540 with 1TB disks, sum up 442 TB more. The newest disk servers are from SuperMicro, and currently there are a total 7 of these with 2 TB disks, totaling 403 TB more.

It was chosen Lustre v1.8 [4] as a backend posix file system (see Figure 4), and with this release included pool capabilities to the installation. It allows to partition the hardware inside a given file system in order to have a better data management. In addition to this, it is possible to assign selected OSTs to an application/group of users, and we can separate heterogeneous disks in the future. There are 4 file systems devoted to different grid virtual organizations:

1. */lustre/ific.uv.es*. For the Virtual Organizations. Mounted *read only* on worker nodes (WN) and user interfaces (UI), and *read/write* on GridFTP + SRM.
2. */lustre/ific.uv.es/sw*. Devoted to software, mounted *read/write* on WNs and UIs. This is what ATLAS used before migrating to CVMFS system.
3. */lustre/ific.uv.es/grid/atlas/t3* Space for T3 users. *Read/write* on WNs and UIs.
4. *xxx.ific.uv.es@tcp:/homefs* on /rhome type lustre. Shared home for users and MPI applications. *Read/write* on WNs and UIs.

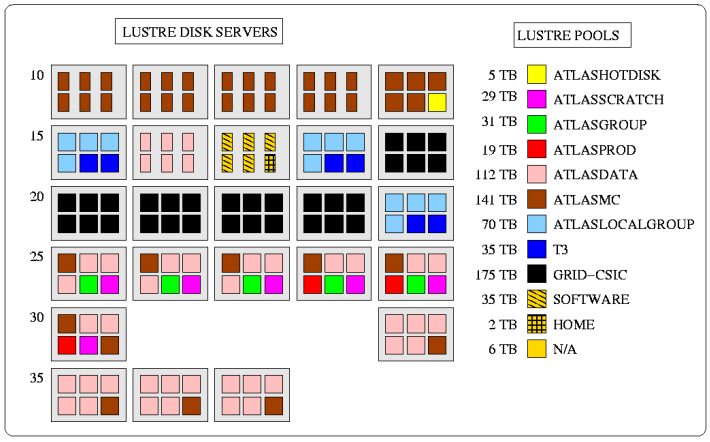

Figure 4: Lustre file system and pool distribution at IFIC.

Storm [14] is used as a SRM server for the GRID access. A plugin was developed in order for the authorization information stored at the file system for SRM access. Lately, it was seen performance degradation due to the increasing number of connections. The storm server has been upgraded to 8-core machine. However, we still see this service as a point of failure in many incidents, and we are working to find a better solution in the future.

3.3 CVMFS for Atlas Software distribution at IFIC

Last September, the infrastructure necessary for CVMFS [13] has been installed at IFIC, the same, using Squid server as Frontier, the database access point for ATLAS jobs. We are monitoring performance of the Squid Server via Cacti Tool (SNMP), and performance since then has been proved very good (Figure 5). The job setup time has been improved, which translates to a better performance for analysis jobs. On top of that, we have no more need of a central disk hosting all ATLAS software versions.

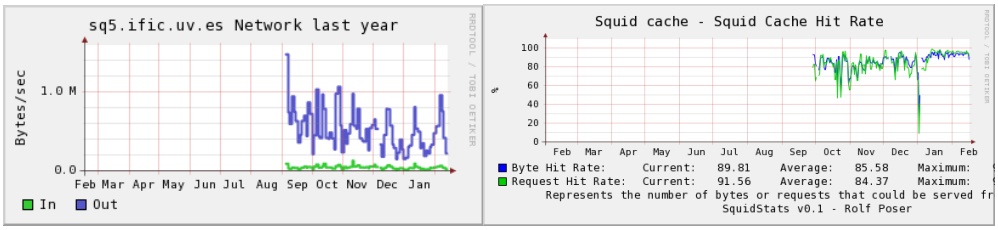

Figure 5: SQUID monitoring via CACTI tool.

3.4 Networking

A commodity 10Gb network is based on Cisco 4500 for the general infrastructure, and Cisco 6500 for the scientific infrastructure. Worker nodes, user interfaces and Sun legacy data servers have 1Gb link, and newer SuperMicro servers have 10 Gb links.

3.5 Summary of utilization and upgrade plans

In 2011, we supported 22 Virtual Organizations including ATLAS, our local institute virtual organization, European projects like vo.agata.org, and t2k.org, and all the virtual

organizations of the Ibergrid Federation. More than 3.5 million jobs have been executed, 6 million CPU hours consumed, corresponding to close to 14 million KSi2K normalized CPU time.

More than 90% of these resources are accounted to ATLAS, which runs on our resources all year long.

For next year, we already fulfill the CPU requirements for ATLAS, which is 6650 CPU HS06. For storage, 230 TB disk space will be added in April 2012, with 4 SuperMicro servers, providing 57.6TB each using 2TB disks.

4. Example of GRID and Physics analysis

From the end-user (physicists working on physics analysis) point of view, typically, a physics analysis has multiple stages. In the first stage, physicists run an analysis program that uses a given number of collision events. These events can be stored in different datasets that are usually spread at different sites. At this step, the Distributed Computing and Data Management Tools, based on GRID Technologies, are used in an exhaustive way. The output of this first step is often a set of ROOT [6] ntuples.

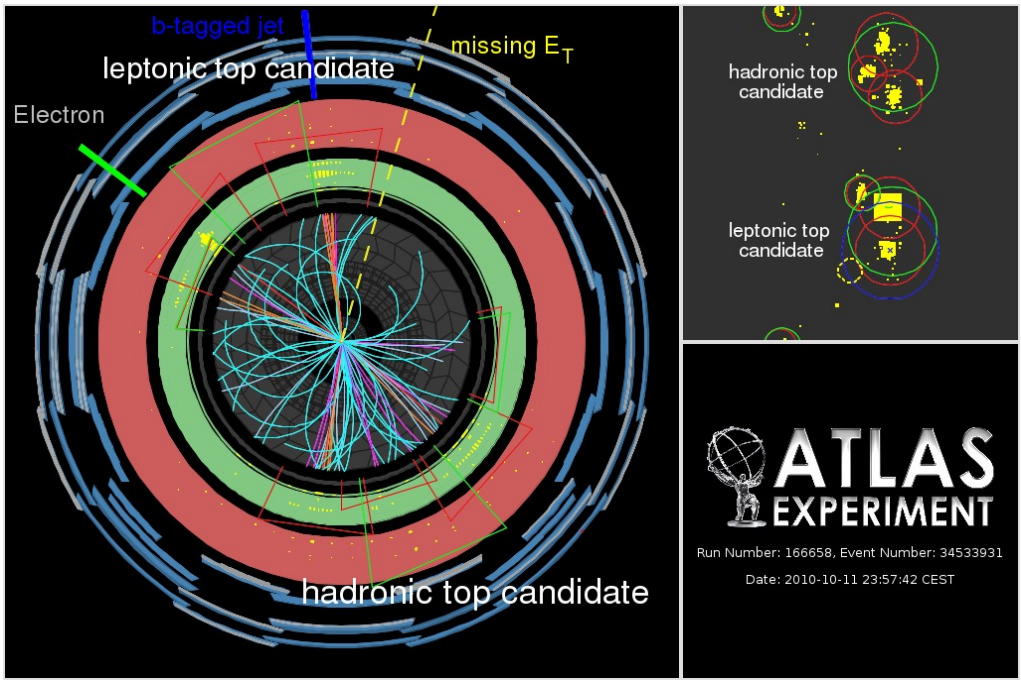

Figure 6: The three R =0.4 jets (red) that are combined to form the hadronic top candidate merge into a single jet when clustered with R =1.0 (green). These event is one the first candidates for boosted top quarks reconstructed as single jets at ATLAS.

In the latter stages, the physicists analyse the ntuples interactively in order to get the final plots, to refine the analysis, etc. A good example of the mentioned second stage is the boosted top candidate shown in Figure 6, which was presented on behalf of the ATLAS Collaboration in the Boost2011 Workshop at Princeton University. To obtain the event display shown, more than 2 fb^{-1} of ATLAS data were processed in only a few days using the ATLAS

computing infrastructure. The resulting ntuples were latter analysed in the Tier-3 of IFIC to reduce the whole sample down to a handful of events, that produced this figure which was the first of its kind ever shown.

At IFIC the Tier-3 resources are split into 2 parts (see ref. [6, 7])

- Resources coupled to IFIC Tier-2, which are included in the GRID environment and used by the IFIC-ATLAS users. Only if the resources are idle, the whole ATLAS community can use them.

- A computer farm to perform interactive analysis (PROOF [8]), which is outside the GRID Framework.

A schematic view of this Tier-3 model can be seen in Figure 7.

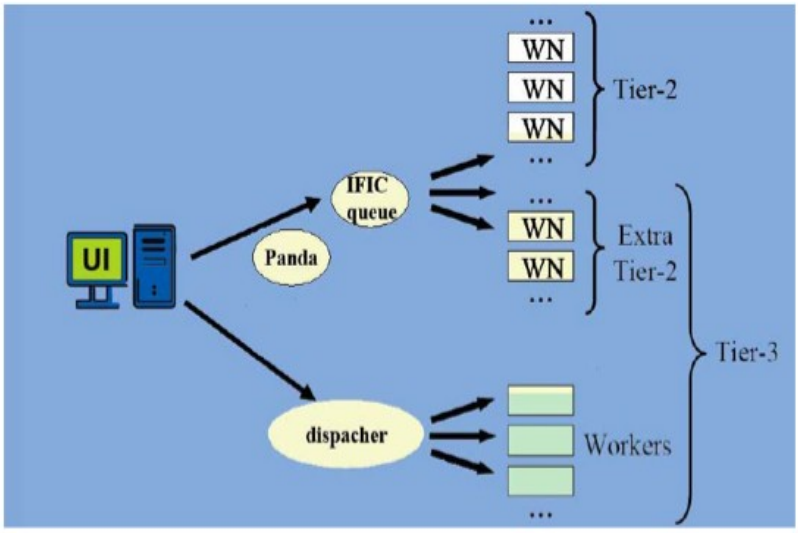

Figure 7: Schema of the IFIC Tier-3

The Distributed Analysis is using the following ATLAS tools:

1) For Data Management:

   a. Don Quijote 2 (DQ2) [9]: to obtain information about data and to download and register files on GRID.

   b. DaTri: the end-user dataset subscription service.

   c. AMI (ATLAS Metadata Interface)

2) For GRID Jobs:

   a. PanDA Client [10]: analysis job submission tool.

b. GANGA [11] (Gaudi/Athena and Grid Alliance): it is a job management tool for local, batch system and the GRID.

A deeper analysis of the daily user activity can be performed by the examination of a heavy exotic particles analysis. Simulated and real data input files represent a volume of several TB of information. The analysis activity workflow can be divided as follows:

1) Test the analysis locally, where the input files are downloaded with DQ2 tools.

2) Submit a first job to GRID which will create an output file with reduced information, the input can be either real or simulated data and the run time for a typical job is around 20 hours;

3) Submit a second job to GRID (normally at the Tier-3) with the objective of doing a refined analysis (reconstruction, application of cuts and selections, etc.). In this case the input is the output of the first job (phase 2) and the execution time is around 2 hours.

The schematic view of the activity can be seen in Figure 8. Working within the GANGA Framework, a python script is created in order to give the requirements of the job, characteristics such as application location, input-output, a replica request to IFIC or the output files splitting. The script allows sending the job to the GRID and, once the jobs finish successfully, the output files are copied to the IFIC Tier-3.

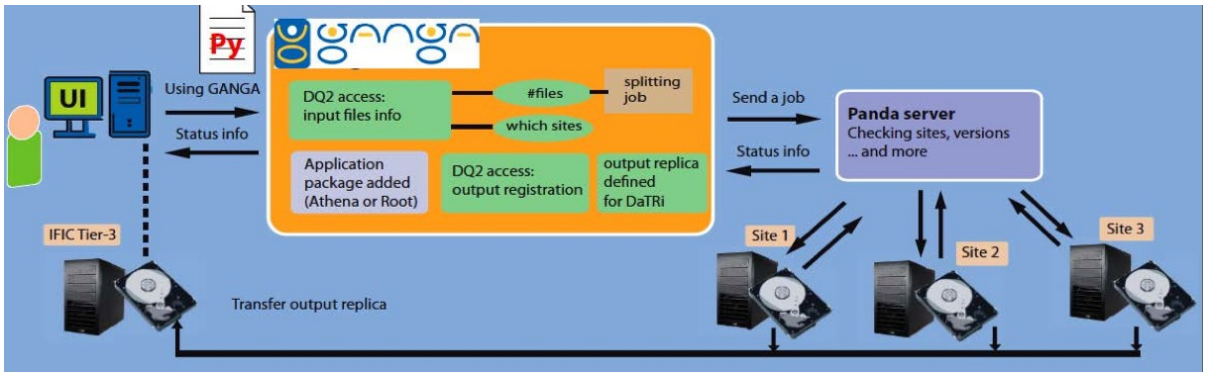

Figure 8: Schematic view of the Distributed Analysis activity (phase 2)

Moreover, the usage of the IFIC analysis resources can be obtained from the global ATLAS monitoring service. We can see the activity of the analysis jobs running at IFIC during 2011 in Figure 9.

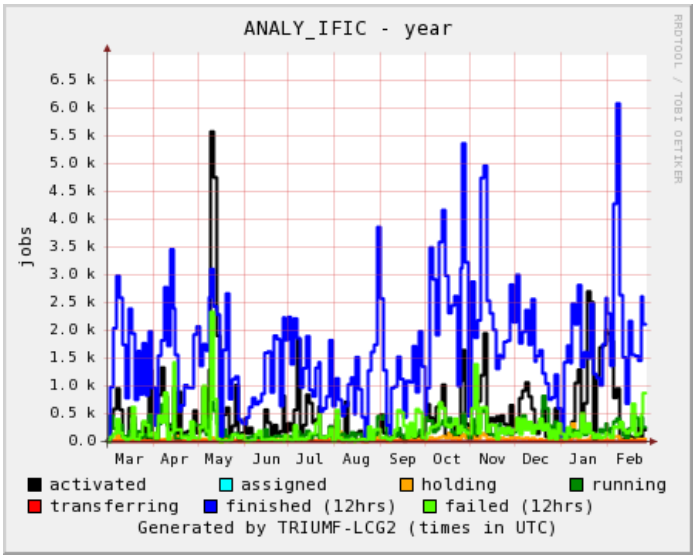

Figure 9: Analysis jobs running at IFIC during 2011

5. Conclusions

IFIC has adapted well to the changes in the computing model. Thanks to its excellent connectivity and availability, IFIC obtained the Alpha T2D status. This new status implies an increase in the number of jobs and data transfers where IFIC infrastructure performs well.

A full real analysis example has been discussed where, apart from the distributed analysis, an interactive analysis runs at the local Tier-3. This example is the best proof of the good throughput of the ATLAS Computing Model and the Tier-2 and Tier-3 at IFIC.

Acknowledgements

We acknowledge the support of MICINN, Spain (Plan Nacional de Física de Partículas FPA2010-21919-C03-01)

References

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

[2] "The ATLAS Computing Model", D. Adams et al., ATLAS Collaboration, ATL-SOFT-2004-007, CERN, (2004)

[3] "ATLAS Distributed Computing Operations in the First Two Years of Data Taking" Ueda, I for the ATLAS Collaboration, ATL-COM-SOFT-2012-007.- Geneva : CERN, 2012 - 9 p.

[4] Lustre File System, http://wiki.lustre.org

[5] ROOT, http://root.cern.ch/drupal

[6] S. Gonzalez de la Hoz et al. "Analysis facility infrastructure (Tier-3) for ATLAS experiment", Published in Eur.Phys.J.C54:691-697, 2008.

[7] M. Villaplana et al. "First tests with Tier-3 facility for the ATLAS experiment at IFIC (Valencia)", ISBN 978-84-9745-549-7, pags 212-220.

[8] M. Ballintijn et al. "The PROOF Distributed Parallel Analysis Framework based on ROOT", arXiv.org: physics/0306110, 2003

[9] "Managing ATLAS data on a petabyte-scale with DQ2", M Branco et al., J. Phys.: Conf. Ser., 119 062017, 2008

[10] "Proceedings of XII Advanced Computing and Analysis Techniques in Physics Research", P. Nilsson et al. Proceedings of Science, 2008

[11] "Ganga: a tool for computational-task management and easy access to Grid resources", F. Brochu et al. CoRR, abs/0902.2685, 2009

[12] D. van der Ster et al. HammerCloud: A Stress Testing System for Distributed Analysis. 2011 J. Phys.: Conf. Ser. 331 072036 doi:10.1088/1742-6596/331/7/072036.

[13] J Blomer et al. 2011 J. Phys.: Conf. Ser. 331 042003, "Distributing LHC application software and conditions databases using the CernVM file system".

[14] Corso E. et al. "StoRM, an SRM Implementation for LHC Analysis Farms". Proceedings of the International CHEP 2006.