Journal of Physics: Conference Series **396** (2012) 032051 doi:10.1088/1742-6596/396/3/032051

Examples of shared ATLAS Tier2 and Tier3 facilities

S. González de la Hoz[1], M. Villaplana[1], Y. Kemp[2], H. Wolters[3], H. Severini[4], W. Bhimji[5] on behalf of the ATLAS Collaboration

[1] Instituto de Física Corpuscular (IFIC), Valencia, Spain
[2] DESY, Hamburg, Germany
[3] LIP Coimbra, Portugal
[4] University of Oklahoma (OU), USA
[5] University of Edinburgh, UK

E-mail: santiago.gonzalez@ific.uv.es, helmut@coimbra.lip.pt, yves.kemp@desy.de,

Abstract. In this contribution, the model of shared ATLAS Tier-2 and Tier-3 facilities is explained. Data taking in ATLAS has been going on for more than two years. The Tier-2 and Tier-3 facility setup, how do we get the data, how do we enable at the same time Grid and local data access, how Tier-2 and Tier-3 activities affect the cluster differently and process of hundreds of millions of events, are described. Finally, an example of how a real physics analysis is working at these sites is shown, and this is a good occasion to see if we have developed all the Grid tools necessary for the ATLAS Distributed Computing community, and in case we do not, to try to fix it, in order to be ready for the foreseen increase in ATLAS activity in the next years.

1. Introduction to ATLAS Tier-2s and Tier-3s

ATLAS is the largest of the four experiments now taking data with proton-proton collisions at the Large Hadron Collider (LHC) at CERN, Geneva. It is a general purpose High-Energy Physics experiment, with as main aim the search for the Higgs boson and for particles predicted by theoretical models to explain the existence of dark matter. Since November of 2009 when LHC started more than 3.5 billions of events were recorded, 66 PetaBytes stored and those events are being analyzed by 3000 physicists from 174 institutes spreads around the world [1].

The ATLAS computing challenge consists of storing, processing and reprocessing, and accessing those data from any institute of the collaboration. This task has been done thanks to the ATLAS distributed computing system which consists of three classes of "regional Centres", names as Tier-0, Tier-1 and Tier-2 following the requirements and the pledged resources provided by the Worldwide LHC Computing Grid project (WLCG) [2].

[1] santiago.gonzalez@ific.uv.es, helmut@coimbra.lip.pt, yves.kemp@desy.de,

- The Tier-0: It is based at CERN and the raw data are recorded to tape and a first processing is done.
- Tier-1s: There are 10 sites, which are storing the raw data on tapes; analysis/real data on disk and a reprocessing is carried out.
- Tier-2s: There are around 80 centres taking care of analysis data on disk, providing resources for user analysis and producing the Monte Carlo simulation events.
- Tier-3s: Local computing resources for the user belonging to an institute or University.

In the original ATLAS Computing Model [3, 4] each Tier-1 centre has a group of Tier-2s centres associated to it. The data distribution over the Grid onto the Tier1 and the Tier-2 centres are managed in an organised way with this association. The data to be stored at a Tier-2 centre are delivered via its associated Tier-1 centre.

2. Evolution of the ATLAS data and Computing model

Based on the operational experiences and the monitored transfer throughput and reliability, ATLAS has decided to evolve the computing model to make flexible transfer routing, enabling more efficient job distribution to the sites in data processing.

The revised data distribution system is composed of pre-defined distribution based on the model, dynamic data placement based on the usage, and on-demand replication. The pre-defined data distribution creates "primary" replicas (to be available on disk) at Tier-1 sites for redundancy and at Tier-2 sites for end-user analysis, as well as "secondary" replicas (they are extra replicas that are created for supposedly popular data using the remaining available disk spaces) to give larger opportunity for analysis. The replicas at the Tier-1 sites are either exported from the Tier-0 or copied from the other Tier-1 sites. The replicas at Tier-2s are created from the replicas at the Tier-1 sites. The dynamic data placement is implemented in the distributed analysis system to increase "secondary" replicas based on the usage of the data, in order to reduce the waiting time of analysis jobs [5].

2.1. Network evolution: Availability and Connectivity

In order to optimize the efficiency of the whole ATLAS Computing Grid for performing analysis of data, an evaluation method has been implemented, that classifies every Tier-2 site with respect to availability and connectivity. The idea behind is to optimize the whole system in a way that the user jobs run as quickly as possible, and, in addition, that the users are able to retrieve their analysis output as quickly as possible. In order for the jobs to run, the sites that host the input data need to be available for analysis. In order for users to be able to retrieve the output, the site storage needs to be online, and it has to offer a good network connectivity to guarantee that the user gets access to the output data in an efficient way. Thus, a good network connectivity is essential as well and becomes one of the criteria for the distribution policy.

3. Tier-2 and Tier-3 activities in different institutes

Data taking in ATLAS has been going on for more than two years. This section will present the Tier-2 and Tier-3 facility setup, how to get the data, how do we enable at the same time grid and local data access using the most diffused HEP storage solution (Lustre, dCache, Xrootd, DPM, AFS, etc), and how Tier-2 and Tier-3 activities affect the cluster differently and process of hundreds of millions of events in the EGI/gLite and OSG flavour.

3.1. Germany and the National Analysis Facility (NAF)

Since 2005, DESY is a Tier-2 for the ATLAS and CMS experiments. As DESY is split over two sites, Hamburg and Zeuthen, Hamburg offers a full Tier-2 for CMS and half a Tier-2 for ATLAS while Zeuthen offers half a Tier-2 for ATLAS. At each site, the Tier-2 resources are integrated into a larger

Grid installation, which is also used by non-LHC experiments. The gLite middleware is used exclusively to offer these resources to the worldwide community.
In 2007, in the framework of a larger collaborative effort to join forces among the different physicists in Germany working in experiments, theory, detector and accelerator research at the TeV-Scale, the Helmholtz program "Physics at the Terascale", the NAF project was created. The intention was to offer additional Grid resources as well as services and resources complementary to the Grid to German physicists with a special emphasis on data analysis. The NAF is located at DESY, and benefits from its infrastructure and the proximity to physics data. Although the NAF is not explicitly called "Tier-3", it has a lot of features that characterize it as such.

The main components of the NAF are shown in the figure1:
Additional Grid CPU resources. These are fully integrated into the Grid batch resources, but are not accounted for the Tier-2 MoU numbers. A separate share is made that is accessible for ATLAS users with the VOMS proxy extension /atlas/de. These users have a higher priority compared to normal ATLAS users. Some work from ATLAS was necessary to make this extra share accessible for users of pilot jobs: A separate pilot factory is submitting using the necessary proxy extension.
Additional Grid storage resources. These are fully integrated into the existing dCache Grid Storage Element, in addition to the Tier-2 MoU pledges. To take advantage of the ATLAS data management tools, no separation is done between these resources and the Tier-2 resources, so no special space token is used.
Local Batch system plus dedicated storage. In order to complement the Grid with a more interactive environment targeted at high bandwidth analysis, a separate batch system was set up including some workgroup server for development work. Data from the dCache Grid Storage can be read with high throughput. In addition, for fast and easy turnaround, a parallel file system was attached to this cluster. Lustre was chosen and operated since 2007, but the NAF is currently deploying an IBM SONAS system as a replacement at the Hamburg site. Finally, AFS is offered to users as a global file system.
The local batch system is managed by the SGE scheduling system. Using its parallel environment, PROOF jobs can be run by users on the cluster [6].

User support on the NAF is split into different parts:

- Experiment specific support groups prepare the experiment specific parts of the NAF (like software installation in AFS) and help users with experiment specific issues (e.g. Ganga setup of NAF) [7].
- First level IT support for user problems related to the infrastructure.
- Expert users can contact the support lists of the providers directly.

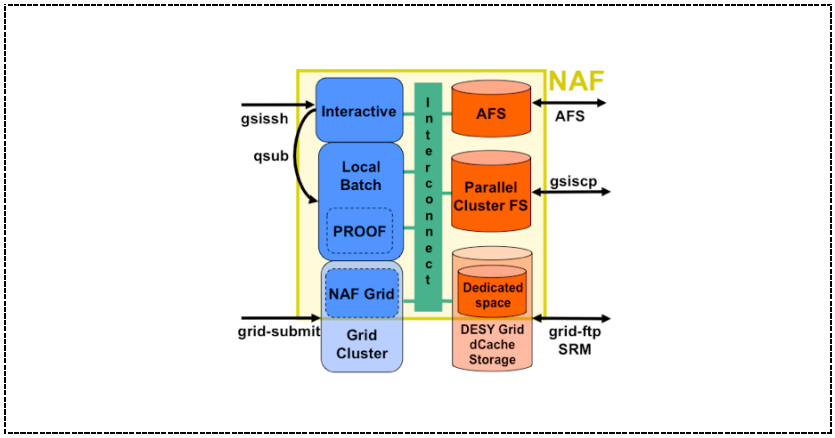


Figure 1. NAF building blocks

As the user community is split among different experiments and institutes in Germany, it is important to have formalized meetings. A monthly meeting brings together representatives from the experiments and providers. Usually once per year a general user meeting is held. The providers also take part in national experiment groups meetings.

3.2. University of Oklahoma (OUHEP, USA)

The University of Oklahoma High Energy Physics (OUHEP) group is participating in computing efforts for both the ATLAS and D0 experiments. The ATLAS computing effort is in operating the Tier-2 Center (the US Southwest Tier-2, in collaboration with UT Arlington and Langston University), running a medium sized Tier-3 desktop cluster for data analysis as well as testing, deployment, and integration of the ATLAS software and Open Science Grid (OSG) infrastructure. The D0 effort includes Monte Carlo (MC) production, data processing and reprocessing using the grid, using the SAMGrid infrastructure.

The 72 worker nodes (844 cores), 350 TB, Tier-2 cluster has been operating continuously since its deployment in early 2006 and has consistently been performing extremely well, with 100% uptime and close to 100% efficiency. The attached 350 TB storage space is seeing good utilization by ATLAS data and has shown a marked performance improvement since the last upgrade in February 2012. New hard drives for a storage solution allowed to double the size of the current Lustre storage during that upgrade, which is now divided between two DDN S2A 9900s, therefore allowing a peak local throughput of 1100 MB/s (as measured with iozone), peak transfer rates of 600 MB/s between OU and BNL, and between 250 and 350 GB/s throughput between OU and BNL can be routinely see. Lustre support from DDN is provided, which has proven to be very useful, if not necessarily inexpensive -- but worth the price when you look at the stability and support.

The smaller (170 CPU/core) OUHEP desktop computer cluster with roughly 100 TB of usable (raid6, xfs) storage functions as our Tier-3 site, with a fully operational OSG CE and Bestman2 SRM SE. It is being used by local faculty and researchers for data analysis and root/proof-lite calculations. And as back-fill D0 MC jobs are continuously running, therefore utilizing all available CPU power.

In addition, a very small testbed cluster is used for software integration and testing, both for OSG middleware and ATLAS Panda tests.

All clusters at OU run a late version of RHEL5; the Tier-2 cluster 64-bit O/S, while the Tier-3 and the testbed run 32-bit O/S. And Condor runs as batch scheduler on all clusters, which have been performing very well for all our requirements.

The number of jobs running on all OU clusters at any given time can be found at [8] as it is shown in figure 2.

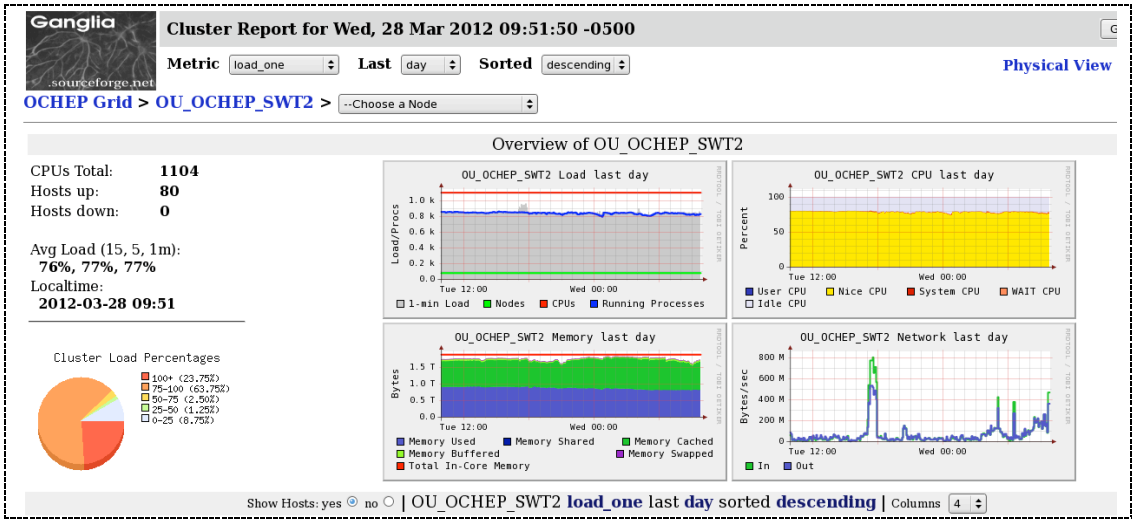


Figure 2. A Ganglia Snapshot of the OU SWT2 cluster

3.3. UK ATLAS Cloud

The UK ATLAS cloud consists of 4 federated "Tier-2" centres as is shown in figure 3 (London, NorthGrid, ScotGrid and SouthGrid) that share pledges, expertise, and in some instances resources, but that are physically located at 16 distinct sites (Birmingham, Brunel, Cambridge, Durham, ECDF, Glasgow, Imperial College, Lancaster, Liverpool, Manchester, Oxford, QMUL, RalPP, RHUL, Sheffield, UCL). Each of these is managed as separate units within ATLAS and 13 run both production and analysis activities. This is a relatively large number of Tier-2 sites for an ATLAS cloud. Conversely, while these, and other UK sites have local "Tier-3" resources, there is no formal sharing, nor Grid connectivity of those. Instead some sites make available tools for local users to easily run on their Tier-2 resources, while others have enabled access of data hosted at the nearby Tier-2 on local Tier-3 compute. The cloud also manages its local space dedicated for the users on a cloud-wide basis allowing and encouraging users from any institute to store data at other UK Tier-2s.

While some of these sites are predominately for ATLAS use, many support other VOs and for some, the ATLAS VO is not the main customer. Furthermore the sites use a variety of different hardware and storage middleware (though DPM is the most commonly used storage solution). Therefore, in order to facilitate smooth operations of this many diverse sites, there is a cloud support team, mostly comprised of local site managers with some central resource. There are also weekly phone meetings where both the core team and other interested site administrators participate. This interoperates with other UK-wide cross-VO support groups and mailing lists. This has worked well as evidenced by the recent rapid deployment of CVMFS [9] across the cloud as well as development activities in the storage and network areas covered elsewhere in these proceedings.

Figure 3. UK ATLAS cloud sites

3.4. Tier-2 and Tier-3 activities at ES Cloud (Spain and Portugal)

In Spain, there is a federated Tier-2 [10] made up of IFIC (Instituto de Física Corpuscular de Valencia), IFAE (Instituto de Física de Altas Energías de Barcelona) and UAM (Universidad Autónoma de Madrid). IFIC represents 50% of the ATLAS Spanish resources and has the responsibility to coordinate the activities of the Spanish Tier-2 federation. The members of the ATLAS user community in Portugal belong all to one research institute (LIP), with branches in Lisbon, Coimbra and Braga. The Portuguese federated Tier-2 consists of three sites, one at LIP in Coimbra, one at LIP in Lisbon and one at the National Grid Center in Lisbon. Each of the ES Cloud sites has an associated Tier-3 that shares the infrastructure of the Tier2 at the same site, using the same queuing system. Additional resources for computing and storage for the Tier3 have been added to the clusters. Whereas the storage for Tier-2 and Tier-3 are logically separated, for running the jobs a fair share queuing system was set up that guarantees that on a medium time scale the Tier2 activities get the share that corresponds to the capacity of the Tier-2 site established for the ATLAS computing Grid. On a short time scale, whenever the activities of the local user group on the Tier-3 part of the site do not occupy the complete slot of their share, additional jobs can be run for the Tier-2 activities, and vice versa. This way we can optimize the occupation of the installed infrastructure.

3.4.1. Data placement. The storage in ATLAS is organized using space tokens. These space tokens are controlled through the DDM system and they are associated to a path to a Storage Element (SE). Data distribution and size in the ES Cloud space tokens on March 2012 is shown in the figure 4. ATLASDATADISK is dedicated to real data (cosmic and collisions) as well as Monte Carlo production. In addition, other space tokens are reserved for physics groups and for users.

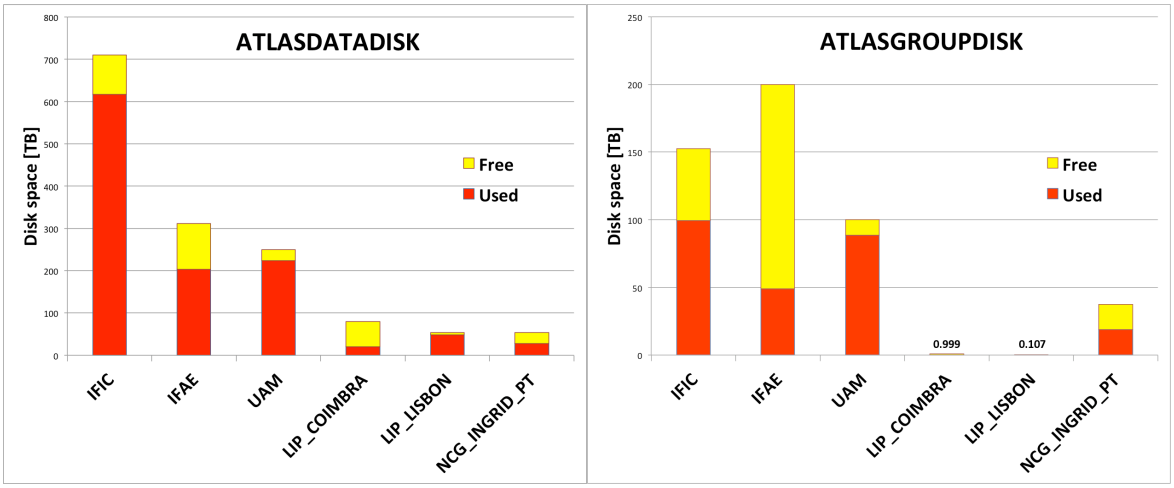
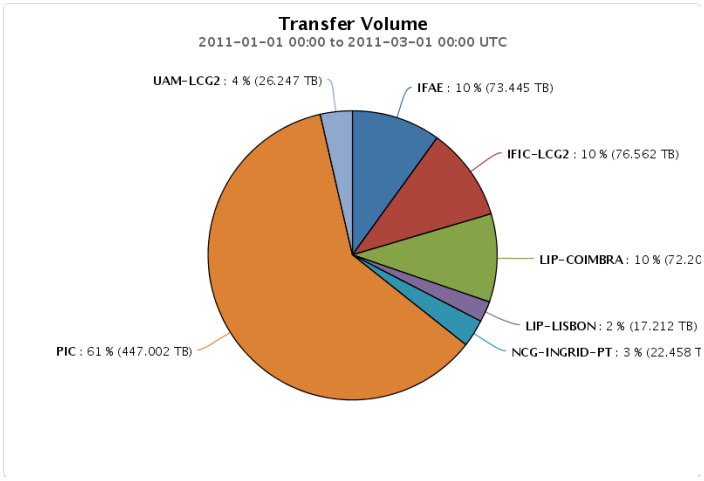


Figure 4: Status of ATLAS space tokens in the Iberian Cloud sites on March 2012.

3.4.2. Dataset replication and data access. The new data distribution strategy distinguishes between primary and secondary replicas. In order to increase analysis opportunities, secondary "extra" replicas of popular data are made using the remaining available disk space.
Primary replicas are distributed according to the computing model, at Tier-1s for redundancy and at Tier-2s for analysis. There is a dynamic placement of secondary replicas at Tier-2s based on usage as well as an on-demand replication system. On top of that, multi-cloud production and direct inter-cloud transfer make Tier-2s less dependent to Tier-1s and thus, its role has become more important. ES-Cloud Tier-2 sites are getting more datasets than Tier-1 (PIC) as shown in figure 5.



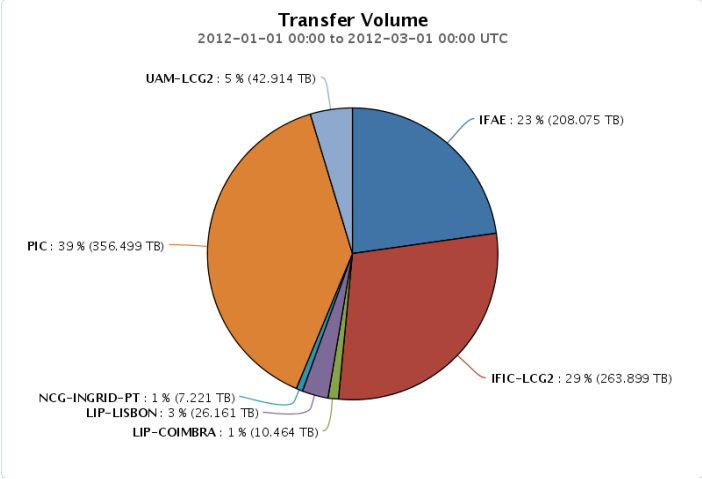
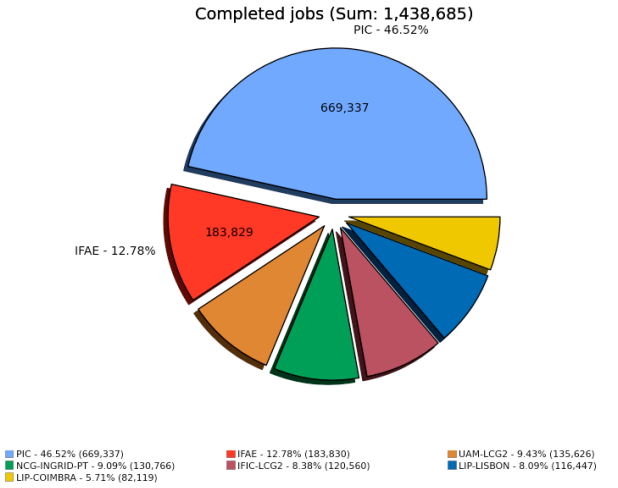


Figure 5: Data transfer volume in the ES-Cloud sites during January and February 2011(top) and 2012(bottom).

3.4.3. Job Distribution. In order to optimize our physics output and make maximal use of available CPU and disk resources production shares are fixed to limit "group production" jobs at Tier-1. Analysis share at Tier-1s has been reduced as well. Therefore, a large part of the analysis and the MC production is done at Tier-2s. Figure 6 compares the amount of jobs completed in the different sites of the ES-Cloud. It is clear that there is less weight at Tier-1 (PIC) now.



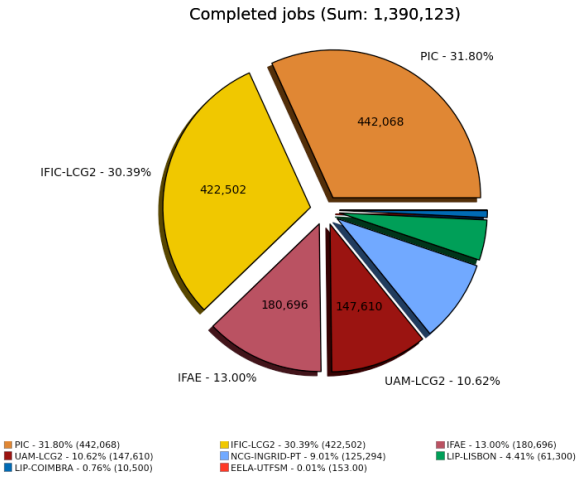


Figure 6: Completed job volume in the ES-Cloud sites during January and February 2011(top) and 2012(bottom).

3.4.4. Infrastructure and Operation. The use of automatic tools for system management tasks is essential. To install and configure the operating system, Grid middleware and the storage system, QUATTOR[11] or PUPPET[12] were used. The ATLAS software is available for the collaboration via a network file system based on HTTP, CVMFS [9], where files and file metadata are cached on demand. For the cache, a SQUID [13] server is used. The database access by ATLAS jobs is done trough SQUID plus FRONTIER [14] servers.
There are two levels of monitoring: one from the global LHC Grid (the "Service Availability Monitoring" is one of its modules), and another one internal to the site, which uses tools like *Nagios*, *Cacti* or *Ganglia*.
To avoid jeopardizing the centre availability, every site has a pre-production computing cluster where software updates are tested before they are put in place. Pre-production machines can be virtualised, via systems like *xen* or *openvz*, in order to economize resources.
The CPU usage must be shared fairly between Monte Carlo production and user analysis jobs. The percentage assigned to each role is configured in the scheduler (*Maui[15]*), which is used by the batch queue system (*Torque[16]*) to submit jobs according to their priority.

3.5. Commonalities and differences of some Tier2s and Tier3s from different clouds

It should be noted that while we reported on several different Tier-2 and Tier-3 activities in different institutes in this section, all the Tier-3s mentioned on this section are co-located with a Tier-2. Therefore, Tier-3s can profit from the technologies used by the Tier-2, without overlapping with Tier-2 activities. For instance, the Tier-3 might use the same storage system as the Tier-2. Another common point between them is that this complex system can coexist with other scientific applications (from non ATLAS users) running at the site.

All the sites reviewed in this section use the grid tools developed by ATLAS to send jobs to the Grid and to retrieve data, but they also have a non-Grid infrastructure to perform interactive analysis, for example a PROOF farm.

It is up to the different institutions to propose possible Tier-3 configurations and software setups that match the requirements of their users' analyses. Therefore there are many differences among the sites reviewed in this section. Those sites enable grid and local data access at the same time, using several different HEP storage solutions (Lustre, dCache, Xrootd, DPM, AFS, etc).

4. Example of GRID and Physics Analysis

In this section the GRID tools and the workflow used in a real example of Distributed Analysis in heavy exotic particles is explained.

4.1. GRID tools for analysis

The ATLAS collaboration has developed different GRID tools for the ATLAS scientists in order to manage data and run analysis jobs. For Data Management, ATLAS have the following specific tools:

- Don Quijote (DQ2) [17]: This tool allows user to download and register files on the GRID and get data information like name, files number, data placement, number of events, etc.
- ATLAS Metada Interface (AMI) [18]: This interface provides information for simulation related with the generation parameters and
- Data Transfer Request (DaTRi): user request tool for transferring data from one site to other site. This transfer is under restrictions and has to be approved by ATLAS Computing Management.

For running GRID analysis jobs the ATLAS community has developed two tools where all ATLAS requirements have been included, the different GRID environments (EGI-Glite, OSG, EGI-ARC) has been unified and a practical job monitoring status web page has been deployed. These tools are shown in figure 7:

- PanDA client [19]: PanDA is a specific system for production and distributed analysis for sending jobs to the GRID in an easy way for users.
- Ganga (Gaudi/Athena and Grid Alliance) [20]: A job management tool for local, batch systems and the GRID. Ganga was initially developed by LHCb, and then ATLAS started contributing to the project.

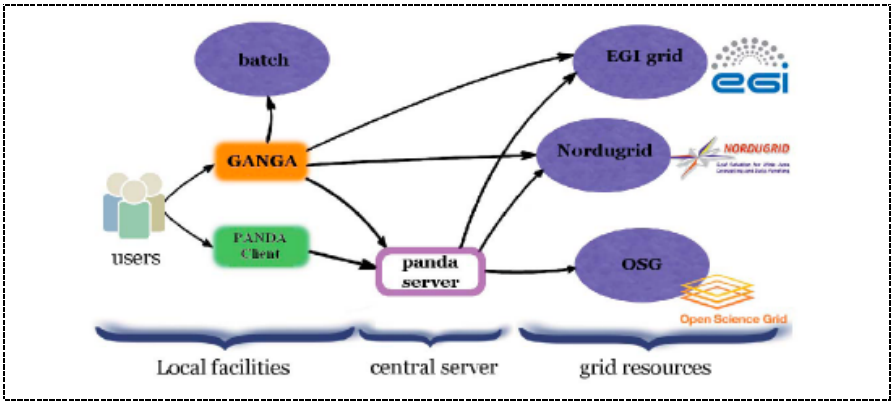


Figure 7. ATLAS Grid tools for sending jobs to the resources

4.2. Work flow

The workflow in an example of distributed analysis is the following. User has to find the input files, which can be simulated data for determining reconstruction models and final selections and/or real data for making comparisons and discoveries with the physics models. In both cases the input files size is around TeraBytes. Once the input files have been selected, the user is testing his/her analysis algorithm locally (in his/her Tier-3 infrastructure), for that, the user can download a few numbers of events using DQ2 tool. Later a first job is submitted to the GRID (to the Tier-2 centres) creating output file with only the information interesting for her/his analysis, running over millions of events and taking around 20 hours. The input of those jobs is the real or simulated data. For that submission

Ganga and/or PanDA client tool is used. Finally, a second job is submitted to the GRID (usually at Tier-3 centres) to carry out the final analysis of the events. The input files are the output of the first Grid job and the job is taking around 2 hours.

A python script is created where application address, input, output, a replica request to a Tier-3, for instance IFIC, and some splitting requirements are defined. The file is executed in Ganga or in PanDA and a GRID job will be send. When the job finished successfully, out files are transferred to the Tier-3 where the access is easy for the user. The schema is shown in figure 8.

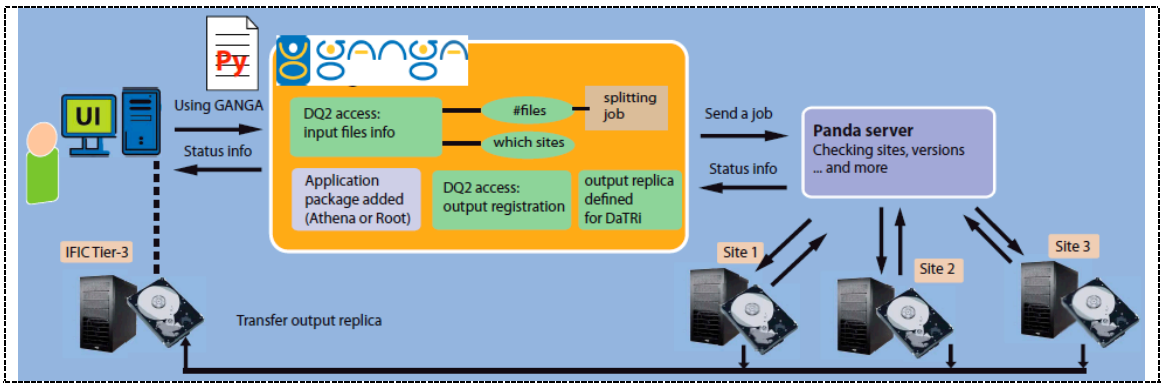


Figure 8. ATLAS Workflow for distributed analysis

For a heavy exotic particles analysis at IFIC, in two weeks 6 users sent 35728 jobs with success (jobs were executed successfully) to 64 sites (Tier-2s and Tier-1s). 1032 jobs were sent to the Spanish Tier-2 (2.89%) and 815 datasets were used as input files producing 1270 files as output. The average size of the input file was around 1.8 GB; in total we used 5.5 TB. The CPU time used to create the output files was 20 hours per job and the average size of the output file was 600 MB; having in total around 1TB.

5. Conclusions

The ATLAS distributing computing system has been evolving and improving different issues, for instance, data distribution dynamically and automatically, monitoring of production of simulated data and its distribution, analysis jobs activities, site status and network. The aim is to have a data transfer and processing beyond the original ones with a better environment for physics studies of the collaboration.

Several Tier-2 and Tier-3 facilities setups in different countries/clouds have been presented. Data placement, replication and access with the new strategy and the use of automatic tools developed by ATLAS and the sites for system management tasks have been shown. To take advantage of the ATLAS data management tools and installation software, no separation is done between Tier-3 and the Tier-2 resources, and only special space Tier-3 token is used in the sites. In the most of centres in order to complement the Grid with a more interactive environment targeted at high bandwidth analysis, a separate batch system was set up installing for instance a proof farm.

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

As far as the Grid tools go, to see if we have developed all the Grid tools necessary for the ATLAS Distributed Computing community, we think we have a mostly working set right now, but we would not go so far as to say we do not need anything else. We are sure there will always be ways to improve our current Grid tools, or replace them with new ones, which for instance scale better.

6. References

[1] ATLAS experiment, *http://www.atlas.ch*.
[2] WLCG project, *http:// lcg.web.cern.ch/LCG/Default.html*.
[3] D. Adams et al, on behalf of the ATLAS collaboration, The Atlas computing model, CERN *ATL-SOFT-2004-007, CERN-LHCC-2004-037/G-85*.
[4] R.W.L Jones and D. Barberis, The evolution of the ATLAS Computing Model, *J. Phys. Conf. Ser, 219 072037*
[5] T. Maeno et al. for the ATLAS Collaboration, Overview of ATLAS Panda Workload Management, *J. Phys.: Conf. Ser. 331 072024*
[6] Haupt A. and Kemp Y. 2010 *J. Phys. Conf. Ser.* **219** 052007
[7] Ehrenfeld W. et. al. 2011 *J. Phys. Conf. Ser.* **331** 072053
[8] *http://gratiaweb.grid.iu.edu/gratia/monbysite?facility=OU*
[9] J.Blomer et al.; "CernVM-FS: delivering scientific software to globally distributed computing resources"; *Proceedings of the first international workshop on Network-aware data management*.
[10] Iberian ATLAS Cloud response during the first LHC collisions, M Villaplana Perez et al; 2011 *J. Phys.: Conf. Ser. 331 072068 doi:10.1088/1742-6596/331/7/072068*
"ATLAS Spanish Tier2 experiences during the STEP09 period", S. Gonzlez de la Hoz et al. EGEE09 Conference. 21 September, 2009, Barcelona, Spain
[11] *http://www.quattor.org*
[12] *http://puppetlabs.com/solutions/configuration-management*
[13] *http://www.squid-cache.org*
[14] A. Dewhurst et al.; "Evolution of grid-wide access to database resident information in ATLAS using Frontier"; *Proceeding of the Computing High Energy Physics (CHEP2012) conference (contribution 400)*.
[15] *www.clusterresources.com/products/maui*
[16] *http://www.clusterresources.com/pages/products/torque-resource-manager.php*
[17] The ATLAS DDM accounting and Storage Usage Service, F. Barreiro, V. Garonne, S. Jezquel. M. Branco and M. Lassing 2010. *Journal of Physics Conference Series Volume 219 part 7*.
[18] *http://ami.in2p3.fr/opencms/opencms/AMI/www/*
[19] T. Maeno et al.: "Overview of ATLAS PanDA Workload Management"; *Proceeding of the Computing High Energy Physics conference (CHEP 2010), Taiwan, October 2010*.
[20] *http://ganga.web.cern.ch/ganga/*

Acknowledgments

We acknowledge the support of MICINN, Spain (Proj. Ref. FPA2010-21919-C03-01)