The evolving role of Tier2s in ATLAS with the new Computing and Data Distribution model

S. González de la Hoz on behalf of Atlas Collaboration[1]
[1]Instituto de Física Corpuscular (IFIC), Valencia, Spain

E-mail: santiago.gonzalez@ific.uv.es

Abstract. Originally the ATLAS Computing and Data Distribution model assumed that the Tier-2s should keep on disk collectively at least one copy of all "active" AOD and DPD datasets. Evolution of ATLAS Computing and Data model requires changes in ATLAS Tier-2s policy for the data replication, dynamic data caching and remote data access. Tier-2 operations take place completely asynchronously with respect to data taking. Tier-2s do simulation and user analysis. Large-scale reprocessing jobs on real data are at first taking place mostly at Tier-1s but will progressively be shared with Tier-2s as well. The availability of disk space at Tier-2s is extremely important in the ATLAS Computing model as it allows more data to be readily accessible for analysis jobs to all users, independently of their geographical location. The Tier-2s disk space has been reserved for real, simulated, calibration and alignment, group, and user data. A buffer disk space is needed for input and output data for simulations jobs. Tier-2s are going to be used more efficiently. In this way Tier-1s and Tier-2s are becoming more equivalent for the network and the hierarchy of Tier-1, 2 is less strict. This paper presents the usage of Tier-2s resources in different Grid activities, caching of data at Tier-2s, and their role in the analysis in the new ATLAS Computing and Data model.

ATL-SOFT-PROC-2012-010
29/08/2012

1. Introduction

The main requirement on the Computing and Data Distribution model [1] is to provide, for all members of the ATLAS Collaboration [2], prompt access to all reconstructed data for analysis, and appropriate access to raw data for organised monitoring, calibration and alignment activities. This model relies on Grid Computing concepts to provide common Grid resources, storage and CPU, to all members of the ATLAS collaboration.

The ATLAS Computing and Data Distribution model embraces the Grid paradigm and a high degree of decentralisation and sharing of computing resources. However, as different computer facilities are better suited to different roles, a degree of hierarchy, with distinct roles at each level, remains. This should not obscure the fact that each role is necessary. The required level of computing resources means that off-site facilities will be vital to the operation of ATLAS in a way that was not the case for previous CERN-based experiments.

[1] Santiago.gonzalez@ific.uv.es

In the original model the primary event processing occurs at CERN in the Tier-0 facility. The RAW data is archived at CERN and copied (along with the primary processed data) to the 10 Tier-1 facilities around the world. These facilities:

- archive the RAW data;
- provide the reprocessing capacity;
- provide access to the various processed versions of the data;
- allow scheduled analysis of the processed data by physics analysis groups and;
- host some grid services – Logical File Catalogue (LFC), File Transfer Service (FTS) [1][3] - for its cloud (Tier-1 and its associated Tier-2s).

Derived datasets produced by the physics groups are copied to the Tier-2 facilities for further analysis. The Tier-2 sites (~80) also provide the simulation capacity for the experiment, with the simulated data housed at Tier-1s. Finally, the other ATLAS sites are labelled as Tier-3. Within ATLAS the word Tier-3 has been used for many different categories of sites/facilities, ranging from Grid sites - having the same functionality as Tier-2s - to non-Grid local facilities. Tier-3 centres participate presumably most frequently in support of the particular interests of local physicist (users at the local facility decide how these resources are used). Tier-3s must provide the software tools to access data and to perform local analysis.

The total luminosity recorded at the end of 2011 was 5.25 fb^{-1}. ATLAS has been taking and processing data with a good efficiency so far having exported to the tiered GRID hierarchy more than of 15k TB since January 2010 as can be seen in figure 1.

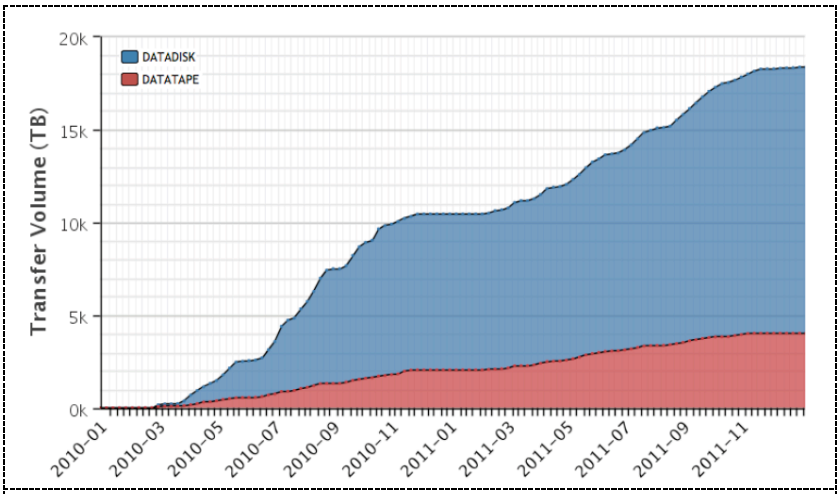

Figure 1. Cumulative data Volume exported from the Tier-0 to the Tier-1 centres over the two years by the destination storage type.

Processing such volume of data has been possible thanks to the establishment of a computing and data distribution model based on GRID technologies at the LHC experiments and, in concrete, at ATLAS.

2. ATLAS Computing and Data Distribution Model

The ATLAS Computing and Data Distribution model was initially thought as a hierarchical infrastructure, where Tier-2 centres could only receive data through their associated Tier-1 centre. Each production task was assigned to a Tier-1 where the input data was available. The task processing was shared between the Tier-1 and its Tier-2. Tier-2 could only receive data through their associated Tier-1. A Tier-1 centre, which leads its associated Tier-2s, forms with them, an ATLAS cloud. This concept is now changing [3]. Tier-2s with a good network connection are allowed to connect to other

Tier-1s and to Tier-2s from a different cloud. This way enables a more efficient usage of disk and CPU resources and high priority tasks can be done more quickly. Tier-2s considered as well connected (Tier2Ds, see section 5) can work for different Tier-1s and are called "multi-cloud sites". On the other hand, Tier-2s well connected hvae direct transfers from/to all ATLAS Tier-1s.

The original model was a working starting point but Tier-2 activity was strongly linked to the associated Tier-1 reliability. Some Tier-1s did not have associated Tier-2s and had few tasks to process while they had the storage to host a significant fraction of task outputs. In addition, Tier-2 had the computing resources to do reprocessing but was limited due to required direct access to the Tier-1 database. As a consequence, some of the sites were not used at full capacity, especially Tier-2s.

On the operational level, if the Logical File Catalog (LFC) was in downtime at Tier-1 there would not be Grid activity in its associated Tier-2s. Another example is if the Tier-1 Storage Service was in downtime, the production and data distribution to/from the Tier-2s was stopped.

The operational improvements were taken by ATLAS in the form of flattening the model from a tier to a mesh [4]. The existing network provided good connectivity to many Tier-1s or Tier-2s. This situation allows the possibility to make direct transfers from Tier-2s to Tier-1s of different clouds, and even with other Tier-2s globally as is shown in figure 2.

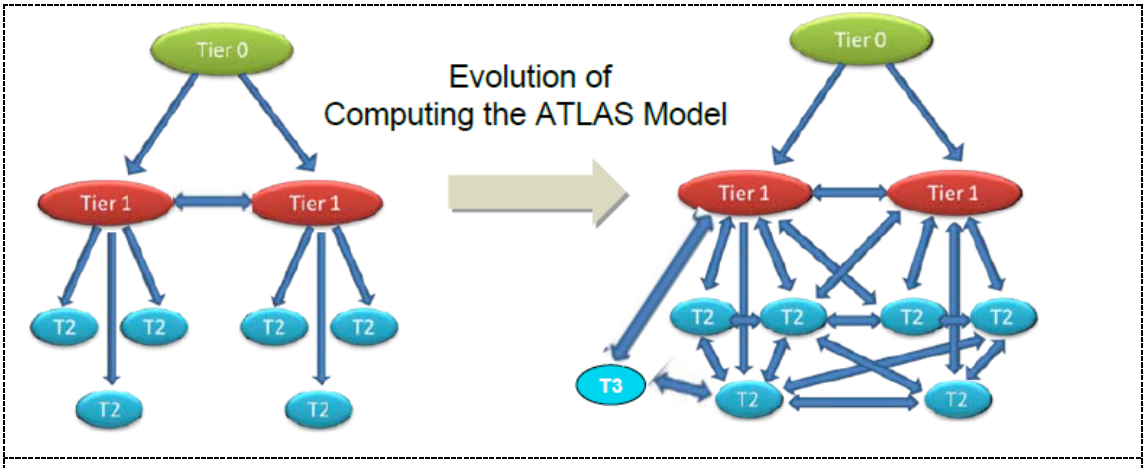

Figure 2. Evolution of the ATLAS Computing model.

Even if these changes have proved to improve the model, this is not the final solution, as Tier-2s are not always well connected to each other.

On the other hand, the ATLAS software and the detector information are necessary to run ATLAS jobs. Since it would not be very efficient to send them to the site together with each of the jobs, the initial solution was to install the software and small file-based database on a local shared file system at each site, and a larger database system at each Tier-1 site. With this model, some bottlenecks were observed when many jobs accessed the shared file system or the file-based database simultaneously.

Evolutions came with CernVM-FS [5] and Frontier/Squid [6]. CernVM-FS is a network file system based on HTTP, with which files are downloaded and cached at the sites and on the worker nodes. The ATLAS software releases and the smaller file-based database are now installed on the server at CERN, and there is no more need to install them at the site where CernVM-FS is used. This has removed the workload in software installation and the bottlenecks with the shared file systems. Frontier/Squid is an

http-based system to access database with caching, avoiding a high load on the database and latency in accessing the database from remote sites. Introduction of the system has removed limits with the database access, allowing the jobs running at Tier-2 sites accessing the database at Tier-1 sites.

The initial model was to run at Tier-1 sites certain types of jobs, for instance reprocessing jobs, which require the information not in the file-based database, and simulation and end-user jobs at mostly a Tier-2 sites although they can also be run at Tier-1. With CernVM-FS and Frontier/Squid, any type of jobs can now run at any Grid site.

3. Data Distribution and Processing activities in the first years: Dynamic data caching and disk space at Tier2s

The new data distribution policy and the new dynamic data placement algorithm were deployed in September 2011 [7], where this has significantly changed the data volume transferred to Tier-2s. ATLAS used a centralized push model for data distribution to Tier-1s and Tier-2s. At the beginning of data taking the push model was found to be effective, but not agile enough to meet the changing needs of ATLAS users. Sometimes, the early usage pattern of data turned out to be different from the pattern anticipated by the Computing Model. Sometimes the splitting of data among sites did not optimally map to user workflow. The new dynamic data placement algorithm (it not a pull model) can dynamically react to user needs much better, instead of pre-distributing data to all Tier-2s; the dynamic data placement will distribute the data by taking the following factors into account: popularity, locality, the usage pattern of the data, the distribution of CPU and storage resources, network topology between sites, site operation downtime and reliability, and so on.

Using the new Dynamic Data Placement algorithm, data replicas are distributed at Tier-2s for analysis, in order to reduce the waiting time of analysis jobs. There is a dynamic placement of data replicas at Tier-2s based on usage (popular data have many replicas) as well as an on-demand (it covers special cases or specific request with approval by the responsible people) replication system. This is shown perfectly in figure 3. Data replication explains the increase in September 2011 [8].

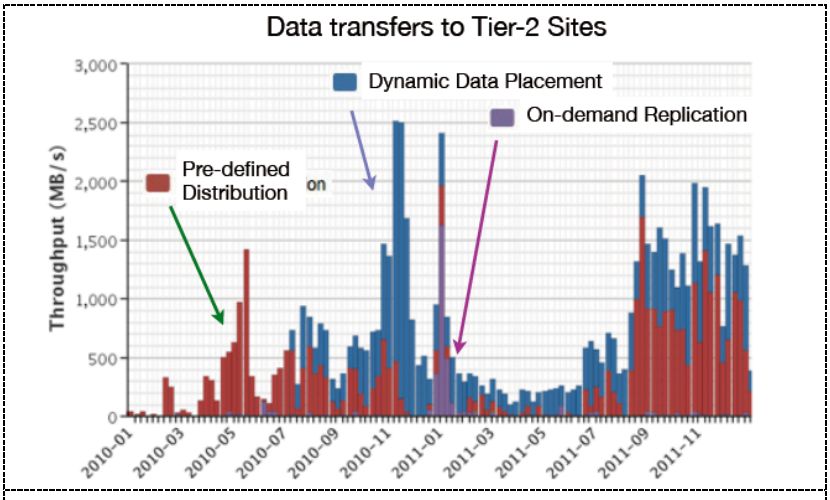

Figure 3. Data transfers to Tier-2 sites.

Tier-2s now get more datasets than Tier-1s because the disk size in Tier-2s has increased significantly while Tier-1 total size was not increasing as much. This is shown perfectly in figure 4 and 5. In figure 4 Tier-2 sites are getting now more datasets than Tier-1s (winter 2012, right) while, in the past, it used

to be the Tier-1s the one getting more datasets (winter 2011, left). In figure 5 the Tier-2s disk space usage for real and Monte Carlo data is more similar to the Tier-1 in volume. The volume of data at the Tier-2s has increased with respect to the one in the old model and there is more Monte Carlo simulation in Tier-2s than in Tier-1s.

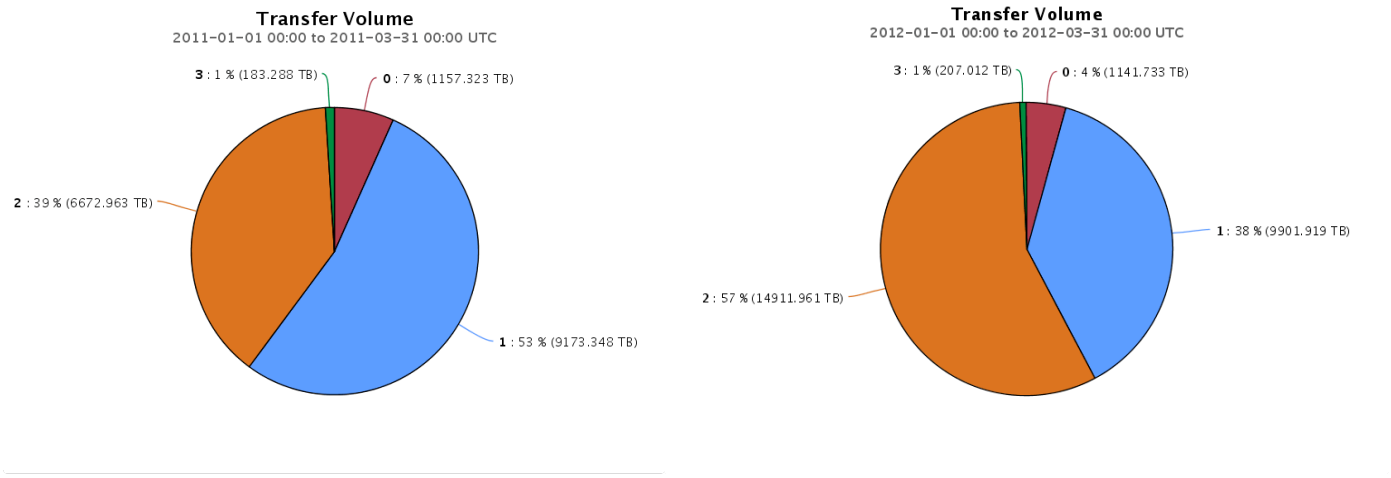

Figure 4. Data transfer volume in Tier sites from winter 2011 (left) and winter 2012 (right).

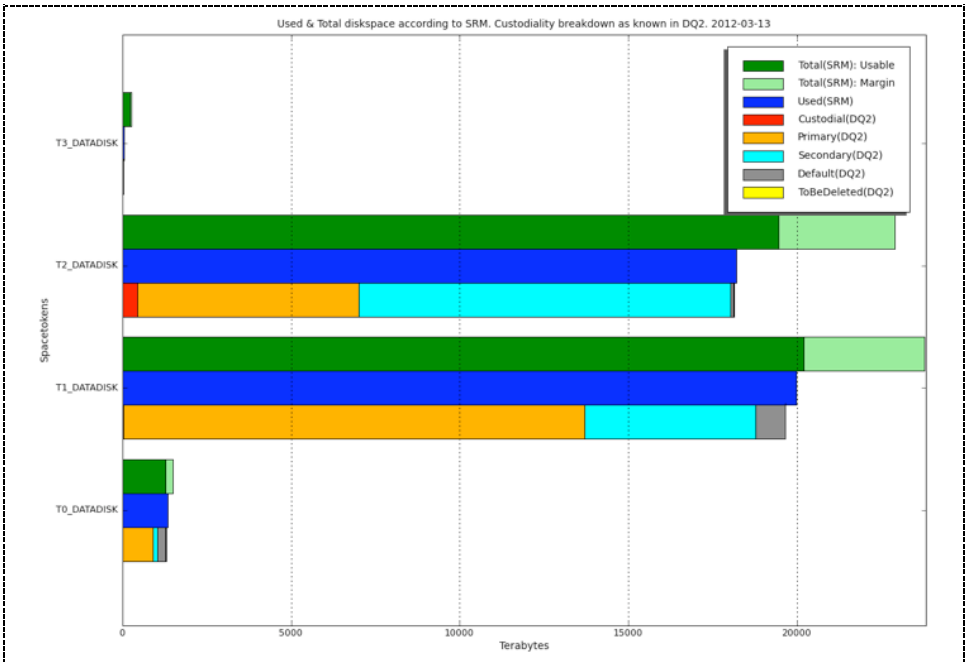

Figure 5. Used and total space for centrally managed data at different Tiers.

4. Data Processing Activities: production of simulated data

We can distinguish three main data processing activities: Official Production, End-User analysis and Group Activities.

Official Monte Carlo simulation production has been running at Tier-1s and Tier-2s sites constantly since before the start of data taking together with the reprocessing of detector data. End-user physics analysis on the Grid started rising since the start of data taking on March 2010 and finally, group activities started as "end-user analysis" of the group of physics analysis responsible of producing common data for end-user analysis. In 2011 this activities have been formalized as a "Group Production".

Figure 6 is showing the number of Monte Carlo and Analysis jobs submitted per week since February 2010.

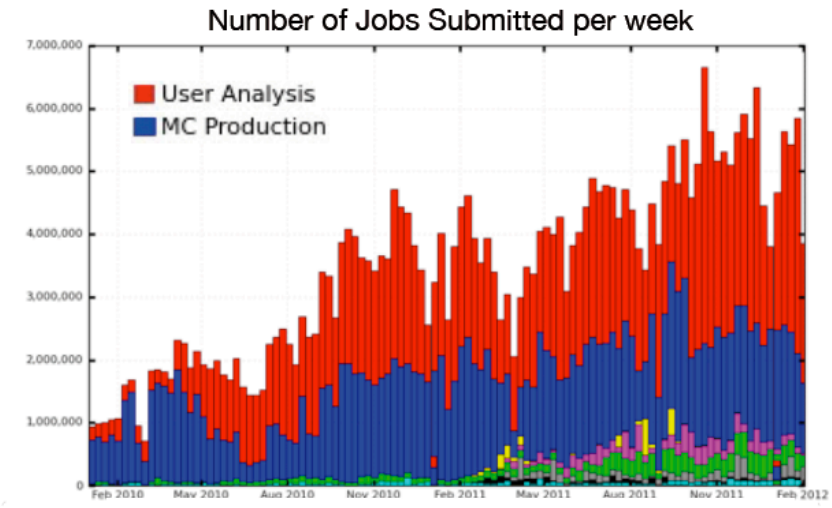

Figure 6. Number of jobs submitted per week. The growth of user analysis jobs is shown since the start of data taking.

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-1s. 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 7 shows the amount of analysis jobs splitted among different Tier types since January 2011 while figure 8 shows all activities restricted to Tier-2s from October 2010 to March 2012. The sudden increase of analysis activities in Spring 2011 is due to the preparation for summer conferences.

The task production brokering now takes into account the input dataset replicas located at Tier-2s if there is no replica in a Tier-1.

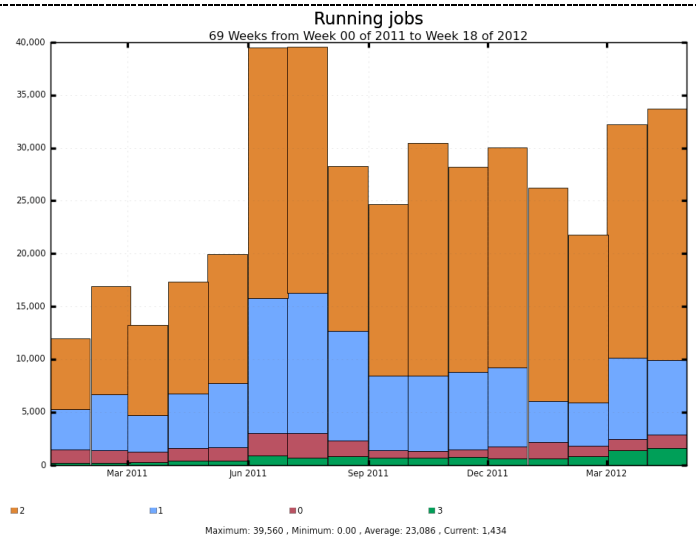

Figure 7. Number of analysis jobs running at all Tiers from January 2010 to May 2012.

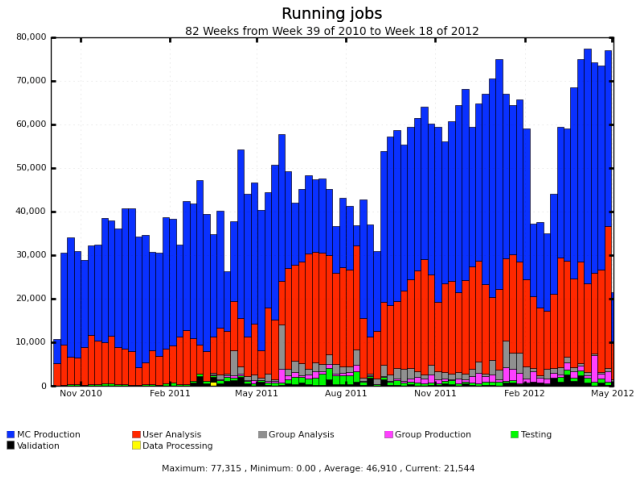

Figure 8. Number of jobs running at all Tier-2s from October 2010 to March 2012.

5. Network model: Usability for Analysis and Connectivity

The original model was based on a dedicated network among Tier-1s and a good network between a Tier-1 and its associated Tier-2s in the same cloud [9]. The initial transfer routing among Tier-2s from different clouds, was done by means of its associated Tier-1, then to the Tier-1 associated to the final destination Tier-2 site and finally to the Tier-2s. In the last year was observed that some Tier-2s did not have any problem transferring directly from/to other Tier2s, or Tier1s not associated with them.

As it has been shown, network is a key component in the evolution of the ATLAS model for the Tier-2s, as they have to be well connected to be able to exchange data among them. In order to check this

connectivity, ATLAS triggers and monitors transfers among sites; the transfer speed is estimated from the rate of multi-GB file transfer (srm overhead negligible).

In terms of connectivity, the concept of so-called Tier2Ds was introduced. Each month, the connectivity of each Tier-2 with respect to each one of the Tier-1s is reviewed. Those that fulfil a certain performance are then classified as Tier2D. Only the Tier2D will be used to run jobs that were attributed to a cloud different from the one they belong to. At the moment, the limit is set to an overall transfer performance of at least 5MB/s to at least 80% of Tier1s in ATLAS, for very large files (>1GB).

The other important point for a Tier-2 site is to be as usable for analysis as possible. Site usability for analysis is defined as the fraction of time when the site was validated to run analysis jobs. ATLAS uses the HammerCloud (HC) framework to test site usability for analysis, by constantly submitting typical analysis jobs to every site [10]. With this behavior, the idea is to protect user jobs from problematic sites, detecting them as soon as possible. Since the ATLAS point of view HC test are a much more complete and dedicated check than SAM (Service Availability Monitor) [11] tests. They run "typical" ATLAS analysis jobs, thus also verifying that ATLAS software setup is correct and usable. So if a HC job fails, a typical ATLAS job would fail too, so the site can be considered not usable for ATLAS.

In terms of usability for analysis, every Tier-2 is analysed at the end of each month, and there are four data acceptance categories:

- *alpha* – usability for analysis >90% if site is also a Tier2D (~30 sites)
- *bravo* – usability for analysis >90% but site is not a Tier2D (~20 sites)
- *charlie* – usability for analysis >80% (~5 sites)
- *delta* – usability for analysis <80% (~10 sites)

The indicated numbers of sites correspond to the status of March 2012.

Scheduled downtimes of a site and unavailability not due to the site are not considered for the classification.

At the moment, around half of the Tier2-s are classified as Tier2D, and practically all of those as *alpha* as well. Summing up the share of all Tier2Ds, they obtain at the moment (March 2012) a total share of around 75% of the data distribution.

In order to preferentially place data at reliable sites (according to the previous month), input data for analysis are preferentially distributed to "good" sites taking into account that sites in downtime are not getting data.

6. Conclusion and Prospect

The Computing and Data Distributed model continues to evolve and improve beyond the original data processing model. ATLAS is monitoring all activities, sites, network, etc and running functional tests for that purpose. The global connectivity for Tier-2s is expected to further improve with LHCONE (LHC Open Network Environment) deployment.

Tier-2 activities (monte carlo production, analysis jobs, store real data, etc.) are now less dependent to Tier-1 and participate to more critical activities in ATLAS. Tier-2s are again receiving data immediately to have a higher contribution to analysis activities. Their usability for analysis and connectivity is reported every month, requiring a good usability and connectivity for data transfer, production and to be a reliable site.

7. References

[1] D. Adams et al., ATLAS Collaboration, The ATLAS Computing Model, *ATL-SOFT-2004-007*, CERN, 2004.

[2] The ATLAS Collaboration, The ATLAS Experiment at the CERN Large Hadron Collider, *JINST* **3** S08003.

[3] R. W. L. Jones and D. Barberis, The evolution of the ATLAS computing model, *J. Phys.: Conf. Ser.* **219** 072037.

[4] S. Campana on behalf of the ATLAS collaboration.; "Evolving ATLAS computing for today's Networks" *Procceding of the Computing High Energy Physics conferences (CHEP2012) (contribution 262)*.

[5] A. De Salvo et al.; "Software installation and condition data distribution via CernVM FileSystem in ATLAS"; *Proceeding of the Computing High Energy Physics (CHEP2012) conference (contribution 349)*.

[6] 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)*.

[7] T. Maeno, D. De, S. Panitkin, for the ATLAS Collaboration, PanDA Dynamic Data Placement for ATLAS, *ATL-SOFT-PROC-2012-16*, CERN, 2012, 7p.

[8] I. Ueda for the ATLAS Collaboration, ATLAS Distributed Computing Operations in the First Two Years of Data Taking, *ATL-SOFT-PROC-2012-003,* CERN, 2012, 9p. *PoS(ISGC 2012)013*.

[9] A. Fernández, M. Villaplana, S. González de la Hoz, J. Salt on behalf of the ATLAS Collaboration, Evolution of the ATLAS data and computing model for a Tier-2 in the EGI Infrastructure, *ATL-SOFT-PROC-2012-004,* CERN, 2012, 11p.

[10] D. Van Der Ster et al.; "Improving ATLAS grid site reliability with functional test using HammerCloud"; *Proceeding of the Computing High Energy Physics (CHEP2012) conference (contribution 317)*.

[11] J. Andreeva et al.; New solutions for large scale functional tests in the WLCG infrastructure with SAM/Nagios: the experiments experience, *CERN-IT-Note-2012-020,* CERN, 2012 9p.

Acknowledgments

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