ATLAS Distributed Computing Operations: Experience and improvements after 2 full years of data-taking

S. Jézéquel¹, G. Stewart² on behalf of ATLAS Collaboration

¹: LAPP, CNRS/IN2P3 and Université de Savoie, Annecy-le-Vieux, France
²: CERN, CH-1211 Geneva 23, Switzerland

E-mail: stephane.jezequel@lapp.in2p3.fr

Abstract. This paper summarizes operational experience and improvements in ATLAS computing infrastructure in 2010 and 2011. ATLAS has had 2 periods of data taking, with many more events recorded in 2011 than in 2010. It ran 3 major reprocessing campaigns. The activity in 2011 was similar to 2010, but scalability issues had to be addressed due to the increase in luminosity and trigger rate. Based on improved monitoring of ATLAS Grid computing, the evolution of computing activities (data/group production, their distribution and grid analysis) over time is presented. The main changes in the implementation of the computing model that will be shown are: the optimization of data distribution over the Grid, according to effective transfer rate and site readiness for analysis; the progressive dismantling of the cloud model, for data distribution and data processing; software installation migration to cvmfs; changing database access to a Frontier/squid infrastructure.

1. Introduction

The Large Hadron Collider (LHC) at European Organization for Nuclear Research (CERN) has been delivering stable beams colliding at 7 TeV since the first collisions at the end of March in 2010. ATLAS [1, 2], one of the general-purpose experiments at the LHC, has been taking data with a good efficiency, accumulating about 4 PB of data over the two years. To process and analyse the associated simulated events, ATLAS is heavily using the WLCG Grid. ATLAS Distributed Computing team is responsible to optimise its usage according to the available resources and the evolving requirements for the high-energy physicist community.

2. Data Management

The raw data acquired with the ATLAS detector (RAW) were recorded at a nominal rate of 200 Hz with average event size of 1.6 MB and stored into tape immediately at the Tier-0. In 2011, ATLAS decided to increase event recording rate and take as much data as possible, broadening possibilities in physics studies, that led to a higher event recording rate up to 400 Hz. Since the first collision at 7 TeV in 2010, ATLAS prompt reconstruction has produced more than 10 PB of raw and derived data (figure 1).
The data files are sent out from the Tier-0 to the Tier-1 centres and stored on tape as soon as possible after their acquisition and registration to the system. Having a tape replica for each file at a Tier-1 in addition to the original at the Tier-0 ensures a long-term protection against a possible data loss. After data calibration being performed at the CERN Analysis Facility, the first-pass processing of the RAW data with event reconstruction is carried out at the Tier-0 [3, 4], producing Event Summary Data (ESD), Analysis Object Data (AOD) and other types of derived data (dESD, dAOD, NTUP, etc.) that are also distributed over the Grid to Tier-1 and Tier-2 centres for further analysis. The average event size of ESD and AOD are of the order of 1 MB and 100 kB, respectively.

3. Site distribution

The ATLAS distributed computing system [5, 6, 7] defines three classes of “Regional Centres”, named as Tier-0, Tier-1 and Tier-2 following the WLCG MoU [8], with different roles and requirements for the pledged resources. The Tier-0 centre is located at CERN and associated with the CERN Analysis Facility. There are 10 Tier-1 and 38 Tier-2 centers over the world hosting ATLAS activities. A “Regional Centre” can be a federation of multiple sites and the number of Tier-2 sites becomes about 70 by counting site-by-site. There are other “ATLAS Grid Centres” without pledged resources, which are named as “Tier-3 centres”. Their operational model was defined in 2011 to control their massive introduction in ATLAS Grid (> 50 sites). But, since their roles, in the ATLAS distributed computing activities, are similar to those of the Tier-2 centres, they are merged to the description of the Tier-2 in this paper. The number of sites in total is about more than 130 including all the Tier-0, Tier-1, Tier-2 and other ATLAS Grid centres.

In 2011, the promised resources (also called ‘pledged’) at the Tier-0 and the CERN Analysis Facility were 75000 HEP-SPEC06 for CPU, 7 Pbytes for disk and 12.2 Pbytes for tape and at the Tier-1 centres in total were 250208 HEP-SPEC06, 27 Pbytes and 32 Pbytes respectively. At the Tier-2 centres 281 kHEP-SPEC06 for CPU and 34 Pbytes for disk were pledged in total and no requirement for tape. But the resources are not equally spread as 61% of the CPU activity is provided by the top 20 sites.

In the ATLAS Computing Model, the main roles of the Tier-1 centres are to store the replicas of RAW data permanently, to store the reconstruction outputs on disk serving as repositories with faster access, and to perform the second and the further processing of RAW data hosted at the site (i.e. reprocessing). The Tier-2 centres are the main facilities for end-user analysis because they hosts input data for the analysis jobs on disk. Each Tier-1 centre has a group of Tier-2 centres associated to it. The data distribution over the Grid onto the Tier-1 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. The Tier-2 centres are also the main resources for producing simulated data.

4. Data distribution

The data to be distributed on the Grid are registered to the ATLAS distributed data management system (DDM) [9]. At the start of 2010, the data distribution was made following the ATLAS Computing Model. However, the model was not necessarily applicable to the situation of the first years. In 2010, the main issue was related to the fact that many analysis studies had to be done on variables only available in ESDs. While those studies were supposed to use the ‘derived’ ESD (dESD) and AOD respectively, their contents were not well tuned, or the analysis codes were not well adapted to the latter formats. The data distribution model was revised accordingly to the needs including creation of extra replicas of ESD but it required to transfer and allocate more TeraBytes than expected. ATLAS also decided to put RAW data on disk for the “discovery mode”, to provide prompt access to any candidate events of new particles. In order to keep the disk usage and the data export throughput
within the available resources, ATLAS introduced (a) compression of RAW data that gives about factor 2 of reduction of data volume, and also (b) a limited lifetime of ESD with reduced number of replicas, that allows studies of detector performances during the period and making space to store the RAW data on disk. As figure 2 shows, the data export rate from the Tier-0 varied due to the changes in number of replicas and the composition of data on disk, while the volume on tape was following that of RAW data in figure 1.

![Figure 1](image1.png) ![Figure 2](image2.png)

**Figure 1.** Cumulative data volume registered at the Tier-0 over the two years by the data type. **Figure 2.** Cumulative data volume stored over the Grid and registered in DDM

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 replicas at Tier-1 sites for redundancy and at Tier-2 sites for end-user analysis with a decreasing number of replicas when the dataset ‘popularity’ decreases. The replicas at the Tier-1 sites are either exported from the Tier-0 or copied from the other Tier-1 sites, the detector data and the first pass processing output being exported from the Tier-0 and the simulated data and the output of reprocessing copied from the other Tier-1 unless they were produced at the site. The replicas at Tier-2 are created from the replicas at the Tier-1 sites. The dynamic data placement is implemented in the distributed analysis system to increase the number of replicas based on the usage of the data, in order to reduce the waiting time of analysis jobs [10]. The first implementation was made in mid-2010 and some tuning was made in beginning of 2011 after the experiences acquired with the system and the observations on the analysis job statistics. The on-demand replication covers special cases or specific requests with approval by the responsible people. Figure 2 shows the total storage occupancy over the last two years. It shows that the storage usage increased linearly with time. As figure 3 shows, the major transfer activities for the Tier-1 sites are “Tier-0 export” and “pre-defined distribution”, while the “dynamic data placement” plays a larger role for the Tier-2 sites. The peaks in May and November 2010 correspond to the data replication right after the reprocessing campaigns that produced large amounts of data in short periods. There is no such peak in 2011 because ESD were not replicated on the Grid.
5. Job activity

Data processing for ATLAS can roughly be divided into two activities, central production and end-user analysis. Among the central production activities, Monte Carlo simulation has always been running even before the start of data taking and has proven that the ATLAS data processing system [11] is robust and performing well at the scale of the ATLAS Distributed Computing. Detector data reprocessing is another important central production activity, which has been carried out several times in 2010 and 2011, providing essential data for ATLAS to produce physics results. The end-user analysis activities on the grid have also been running before 2010, using the simulated data, but they started rising significantly since the start of data taking (figure 4). In between the central production and the end-user analysis activities, there are group analysis activities, where the detector performance study groups and the physics analysis groups produce common data for end-user analysis from the central production output such as ESD and AOD. In the beginning, the group analysis activities were run in a form of end-user analysis, submitted by the individuals who were responsible for producing the group data. The activities have features very much similar to those of central production and have been formalized as “group production” once their software became standardized.
Figure 5. Data transfer activities related to data processing. Production jobs trigger transfers of input and output files, while analysis jobs are sent to the sites hosting input data and no transfers are triggered by default. However, users can request replication of the input data they are going to use, and can also specify the destination of the output. The two activities are comparable in terms of number of files.

Data processing activities also trigger data transfers. The central production jobs, including the group production, running at Tier-2 sites require transfers of input data from their associated Tier-1 centres, unless already available at the site. The output files are sent back to the associated Tier-1 and aggregated there. Group production jobs are defined with destination group spaces allocated at various sites, and the output files are automatically transferred to the destination sites. End-user analysis jobs are brokered to the sites hosting the input data, but in case they have difficulties to run the jobs at those sites, the users can request replication of the input data to some other sites where they have no problem to run the jobs. Users can also submit analysis jobs specifying a destination site for the output, which results in automatic data transfers as a set of requests in the on-demand replication system. All these transfer activities triggered by data processing can be identified in the data transfer monitoring, and as figure 5 shows, the transfers triggered by users including the output transfers to the destination sites are comparable in number of files to the transfers of input and output files for official and group production.

6. Breaking the cloud model

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

In the original model of data transfer routing, the existence of efficient network was only expected between Tier-1 sites and between Tier-1 and its associated Tier-2 sites. Thus, the transfers between two Tier-2 sites that are not associated to the same Tier-1 were routed as T2A – T1A – T1B – T2B, where T2A and T2B are the two Tier-2 sites and T1A and T1B are their associated Tier-1 sites correspondingly (figure 6a). Direct transfers between Tier-2 sites were carried out only between sites that are associated to the same Tier-1 site. However, in reality, some Tier-2 sites are efficient in data transfers from or to other Tier-1 or Tier-2 sites not in association. Based on the full mesh transfer tests from every site to every site, an auto-routing algorithm in the transfer system was implemented. It optimises the routing, either via Tier-1 sites or direct, by computing the expected transfer durations. It can differ for different file size.

To identify well connected sites, regular transfers are generated to provide measurements which are displayed in the Site Status Board framework [12]. We have identified the Tier-2 sites that are well connected to most of the Tier-1 sites. Today, half of the Tier-2 sites have this label. The others are encouraged to improve their connectivity to enter this new category.
In parallel to direct transfers, such Tier-2s can now be assigned simultaneously to many Tier-1s for the production activity. It allows assigning the computing capacity to the Tier-1 hosting the highest priority task. A shortcoming of this model is that the sum of the CPU capacity at the Tier-1 and Tier-2 sites to produce data may not be in balance against the disk capacity at the Tier-1 to host the aggregated output. As an extension, it is also possible to associate even a Tier-1 site with other Tier-1 sites so that it can contribute to the task assigned to the other Tier-1. The different models are presented in figure 6.

Figure 6. The evolution of the data processing mode. (a) In the original model, the jobs of a task assigned to a Tier-1 site, are run at the Tier-1 and its associated Tier-2 sites, with the input data sent to the Tier-2, and the output data aggregated at the Tier-1. (b) The Tier-2 sites well connected to another Tier-1 site can contribute to the tasks assigned to the Tier-1 as well as its associated Tier-1. (c) Even a Tier-1 site can contribute to the tasks assigned to another Tier-1.

7. ATLAS software and database

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 at each site, and a larger database system at each Tier-1 site. The software installation is carried out by special jobs submitted to the site that download software releases from the central repository, install them on a local shared file system at the site, and validate the installation. The detector information is put into a database at each Tier-1 site synchronized with Oracle Streams from the database at the Tier-0. To achieve scalability, full detector information such as detector geometry parameters needed for simulation jobs and detector conditions parameters for end-user analysis of simulated data are put into file-based database and distributed to every site with the DDM [13]. Thus, the initial model was to run at Tier-1 sites certain types of jobs, especially reprocessing jobs, which require the information not in the file-based database, and simulation and end-user jobs at mostly at Tier-2 sites although they can also be run at Tier-1. With this model, some bottlenecks were observed when many jobs accessed the shared file system or the Oracle database simultaneously [14].

Evolutions came with the CernVM-FS [15] and Frontier/Squid [16]. CernVM-FS is a network file system based on HTTP where files are downloaded and cached at the site 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. End 2011, it was deployed in more than 50% of the sites. This reduced the workload in software installation and the bottlenecks with the shared file systems. Frontier/Squid is a 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 this technology has removed limits with the database access, allowing the jobs running at Tier-2 sites accessing the database at Tier-1 sites. With this, any type of jobs can now run at any Grid site. The frontier deployment also enabled to reduce to 5 the number of Tier-1s hosting condition data. As a consequence, any type of processing task can be shared between Tier-1 and Tier-2 sites.

8. Grid operation tools

Running the activities over Grid with components widespread around the world at many sites (> 130), and experiencing troubles frequently, ATLAS decided to introduce regular tests to detect problems as soon as possible and to notify the responsible people, and in some cases, when necessary, to hide the problematic components while waiting for fixes. The data transfer functional tests were created to ensure a smooth data distribution. Detecting problems is important especially for the Tier-1 sites where data replication from the Tier-0 is crucial. Whenever possible, the problem should be detected and solved before the Tier-0 export resumes. The tests were introduced after the early exercises of a large-scale data distribution in 2007. Another system collects site downtime and storage information to automatically stop data transfers from and to the sites when in downtime or without sufficient free space. As a side effect, it prevents the error monitoring to be spawned by expected issues so that other issues at lower rate cannot be detected.

On the job activity side, failures in data processing affect end-user jobs more seriously than production jobs because failed production jobs are re-submitted automatically by the system, whereas end-user analysis jobs are not resubmitted automatically. The reason is that production jobs in principle are well validated and most failed jobs will finish successfully after resubmission, whereas end-user jobs may fail due to the user codes. Since problems at sites are usually reported only after a large number of failures are observed, users would suffer with many failed jobs running before actions, such as closing the queues, are taken. Therefore, it is important to detect problems as soon as possible, and close the queues at problematic sites. Resuming the queues should be done after having successful tests. In order to improve the end-user experiences in data processing on the grid, the analysis job functional tests together with automatic control of queue status have been introduced in 2010. In parallel, there is also an on-going work to resubmit end-user jobs when the failure is identified as Grid related. When production jobs are failing, it is enough to close the queue manually to avoid submitting jobs to the site, resubmitting failed jobs automatically by the system, and to send test jobs once the problem is fixed. However, the manual interventions give a load on the operations. After successful automation of queue control with the analysis job functional tests, we have introduced production job functional tests recently in 2012 in order to reduce the load of manual interventions by automatically sending test jobs and controlling the queue status based on their results. In addition to the queue control based on the functional tests, another system has also been introduced to automatically control the queues based on the downtime information. Figure 7 shows the typical instability for few sites and clearly shows the benefit to automatize the grid queue management and validation.
Figure 7: Site availability for some Tier-2s over a month period. The green/orange/yellow/red correspond to the site being labelled as running/brokeroff/testing/offline

Because of these site instabilities and since the ATLAS Grid is used permanently over the year, it is followed by people to track issues and ensure that the responsible persons cure them. ATLAS has increased the amount of automated actions over the last two years but human interventions are still necessary for:

- Identify issues not automatically identified
- In case of site issue, report and require action through the GGUS system
- In case of ATLAS tool issue, report and require action through Savannah
- Inform ATLAS Grid users about issues and help them to bypass them.

For the moment, the frequency of site issues has been kept at a constant level although the number of sites increased significantly.

These issues are followed by different shifter teams:

- Shifters in the ATLAS control room: Prompt data processing at Tier-0 and their distribution to Tier-1s and calibration Tier-2s
- Remote ATLAS Grid shifters (ADCoS): Permanent survey of all ATLAS activities on the Grid
- Remote Distributed Analysis shifter: First support line for ATLAS users on the Grid
- ATLAS Manager on Duty: Expert shifter defining the criticality of problems on the Grid and making sure that service experts are addressing problems in time.
For these duties, monitoring is a key to effective operations, and we have put a large amount of efforts. The most basic monitoring tools are made to understand the situation of ATLAS activities, such as number of successes and failures in data processing and data transfers, number of running and waiting jobs, throughput of data transfers. The monitoring system also records declared downtime information of the sites that can be taken into account for site availability calculation. During 2010 and 2011, the monitoring tools have evolved to cope with the significant increase of activity, provide more details and become more intuitive for experts or shifters. A global site monitoring called Site Status Board was built to display, in a concise way, the status of basic services as well as the efficiency of ATLAS activities. From this page, one can access more detailed information for each component.

9. Conclusion

The ATLAS distributed computing system has been running stably with the large amount of data for the activities such as data distribution from the Tier-0, production of simulated data and its distribution, group and end-user analysis jobs. The system has been evolving and improving without facing scalability issues, for example, data placement with various components to optimize the data distribution dynamically and automatically, group analysis integrated into the production system, monitoring of those activities, site status and network, constant flow of functional tests of data transfer, analysis and production to ensure smooth activities and automated actions against site instabilities, and the models for data transfers and data processing beyond the original ones. The aim is a better environment for physics studies of the collaboration, and we are looking forward to fruitful physics results.

References

[1] The ATLAS Collaboration, The ATLAS Experiment at the CERN Large Hadron Collider, JINST 3 S08003
[2] The ATLAS Collaboration, Expected Performance of the ATLAS Experiment - Detector, Trigger and Physics, hep-ex/0901.0512
[3] Elsing M., Goossens L., Nairz A. and Negri G., The ATLAS Tier-0: Overview and operational experience, J. Phys.: Conf. Ser. 219 072011
[4] I. Ueda for the ATLAS collaboration, ATLAS Operations: Experience and Evolution in the Data Taking Era, J. Phys.: Conf. Ser. 331 072034
[5] The ATLAS Collaboration, ATLAS computing : Technical Design Report, CERN ATLAS-TDR-017, CERN-LHCC-2005-022
[6] D. Adams et al. on behalf of the ATLAS Collaboration, THE ATLAS COMPUTING MODEL CERN ATL-SOFT-2004-007, CERN-LHCC-2004-037/G-085
[7] Jones R.W.L. and Barberis D., The Evolution of the ATLAS Computing Model, J. Phys.: Conf. Ser. 219 072037
[8] Memorandum of Understanding for Collaboration in the Deployment and Exploitation of the Worldwide LHC Computing Grid, http://cern.ch/LCG/mou.htm
[9] Branco M. et al., Managing ATLAS data on a petabyte-scale with DQ2, J. Phys.: Conf. Ser. 119 062017
[10] Molfetas A. et al., Popularity Framework to Process Dataset Tracers and Its Application on Dynamic Replica Reduction in the ATLAS Experiment, J. Phys.: Conf. Ser. 331 062018
[11] Maeno T. et al. for The ATLAS Collaboration, Overview of ATLAS PanDA Workload Management, J. Phys.: Conf. Ser. 331 072024
[12] Magradze E. on behalf of the ATLAS Collaboration, Automating ATLAS Computing Operations using the Site Status Board Monitoring and automatic site exclusion of the ATLAS computing activities: the ATLAS Site Status Board, presented at CHEP12

[13] Basset R. et al., Advanced technologies for scalable ATLAS conditions database access on the grid, 2010 J. Phys.: Conf. Ser. 219 042025

[14] Borodin M., Nevski P., Vaniachine A. for the ATLAS Collaboration, Scaling up ATLAS Database Release Technology for the LHC Long Run, 2011 J. Phys.: Conf. Ser. 331 042004

[15] De Salvo A. on behalf of the ATLAS Collaboration, Software installation and condition data distribution via CernVM FileSystem in ATLAS, presented at CHEP12

[16] Dewhurst A. on behalf of the ATLAS Collaboration, Evolution of grid-wide access to database resident information in ATLAS using Frontier, presented at CHEP12