PROCEEDINGS OF SCIENCE

The Present and Future Challenges of Distributed Computing in the ATLAS experiment

Ikuo Ueda[1]
On behalf of the ATLAS Collaboration
The University of Tokyo, International Center for Elementary Particle Physics
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
E-mail: i.ueda@cern.ch

The ATLAS experiment has collected more than 5 fb^{-1} of data in 2011 at the energy of 7 TeV. Several billions of events had been promptly reconstructed and stored in the ATLAS remote data centres spanning tens of petabytes of disk and tape storage. In addition, a similar amount of data has been simulated on the Grid to study the detector performance and efficiencies. The data processing and distribution on the Grid sites with more than 100.000 computing cores is centrally controlled by the system developed by ATLAS, managing a coherent data processing and analysis of about one million jobs daily. An increased collision energy of 8 TeV in 2012 and much larger expected data collection rate due to improved LHC operation impose new requirements on the system and suggest a further evolution of the computing model to be able the meet the new challenges in the future. The experience of large-scale data processing and analysis on the Grid is presented through the evolving model and organisation of the ATLAS Distributed Computing system.

36th International Conference on High Energy Physics
July 4-11, 2012
Melbourne, Australia

[1] Speaker

1. Introduction

The ATLAS experiment [1] at the LHC has collected a large amount of data, more than 5 fb^{-1} and 6 fb^{-1} of proton-proton collisions at 7 TeV and 8 TeV respectively, by the end of June 2012, adding up to 4.4 billion events, and nearly 170 μb^{-1} of heavy-ion collisions with 0.4 billion events.

The ATLAS Distributed Computing System [2, 3] manages data processing and analysis jobs running on over 100,000 computing cores, as well as data transfers to and accesses from about 50 PB of disk and about 30 PB of tape storage over about 130 sites world-wide. The sites are organised in tiered "Regional Centres", namely Tier-0, Tier-1 and Tier-2, with different roles and requirements for the resources. Tier-0 is located at CERN and records raw data from the ATLAS detector onto tape and performs first-pass data processing with prompt calibration. There are 10 Tier-1 centres, whose principal roles are to store the replicas of raw data in tape for a long-term protection against a possible data loss and to serve as repositories of the processed data on disk, and to perform further processing of raw data (reprocessing). The Tier-2 centres serve as the main facility for group and end-user analyses and host the data for analysis jobs on disk. A "Regional Centre" can be a federation of multiple sites and 38 Tier-2 centres add up to about 80 sites. In addition, there are other sites functioning as Tier-2, but without pledged resources. Production of Monte Carlo simulation data is carried out wherever possible and capable. Off-Grid facilities utilised by end-users, downloading data and running analyses, are not in the scope of this paper.

The ATLAS Computing Model defines the number of replicas and further distribution policies of the processed data for analysis by working groups and end-users. The model defines the following data types:

- RAW – raw data from the detector, need processing before analysis
- ESD (Event Summary Data) – output of event reconstruction
- AOD (Analysis Object Data) – data for physics analysis in ATLAS-wide format
- TAG – event-level metadata, short event summaries primarily for event selection
- DPD (Derived Physics Data) – data for analysis in group- or user-specific format for faster iteration. One of the sub-types is NTUP for end-user analysis.

2. Evolution and operations of the ATLAS distributed computing in the first years

2.1 Adjusting the ATLAS Computing Model

During the first years of data-taking, the model was adjusted to varying real conditions [4]. In 2010, the data distribution plans were revised and dynamic data placement [5] was introduced following observations of usage pattern on data types. In 2011, ATLAS decided to decrease event filtering rate and take as much data as possible, broadening possibilities in physics studies, increasing the event recording rate from the initial nominal of 200 Hz to maximum 400 Hz. In addition, it was also decided to put RAW data on disk at Tier-1 centres. In order to keep the disk usage and the data export throughput within the available resources, RAW data were compressed at Tier-0 for about a factor 2 in size (table 1), and ESD datasets

were removed from the disks at Tier-1 centres after a lifetime of several weeks expecting prompt studies of detector performances using ESD are done before that.

	2010	2011	2012
Trigger Rate	< 200 Hz	< 400 Hz	< 400 Hz
RAW	1.7MB	1.1MB	(1.2MB)
Compressed		0.66 MB	0.73 MB
ESD	1.05 MB	1.21 MB	1.86 MB
AOD	0.09 MB	0.16 MB	0.19 MB

Table 1. Evolution of the trigger rate and the event size since 2010.

In the ATLAS Computing Model, each Tier-1 centre has a group of Tier-2 centres associated to it and the data distribution over the Grid is managed in an organised way with this association. The distributed data management system (DDM/DQ2) [6] was first implemented in a way that data transfers between Tier-2 centres associated to different Tier-1 centres were made via the Tier-1 centres. Based on operational experiences and the measured transfer rates, DDM has evolved, moving from a "tree" topology to a "mesh" of sites of any Tiers, to enable direct transfers when estimated transfer times are shorter based on the measured statistics in the recent past, for efficiency, less load to the system, and end-user convenience [7].

Among the other changes made in the model, introducing the Frontier/Squid has enabled remote access to the databases at Tier-0 and Tier-1 centres from any site [8]. It is now possible, for example, to run reprocessing jobs that require detector conditions data at Tier-2 centres. The model for installing the ATLAS software releases at sites has been simplified using CernVM-FS [9], a network file system, with which files are downloaded from the central repositories and cached locally. The ATLAS software releases are now installed in the repository at CERN, and files are downloaded and cached at some sites and on the individual worker nodes. There is no more need to submit special jobs and pre-install the software to the sites where CernVM-FS is used. It also removes problems due to a load on the shared file systems at the sites.

In all these evolution processes, improvements and new implementation of monitoring services and tools were essential in understanding the activities and needs [10].

2.2 Operations of the ATLAS distributed computing

The ATLAS computing systems has been running stably on the large scale. The Tier-0 system with a comprehensive monitoring suite [4] has been running first-pass data processing from prompt trigger streams keeping up with the ATLAS event recording rate with the improving LHC performance. The system has dedicated resources with about 3000 job slots to ensure prompt processing of RAW data, as well as a flexibility to utilise non-dedicated resources reaching up to a total of 7500 jobs. The data quality from the first-pass processing with the prompt calibration loop in 2012 has been high enough so that the data were used in physics analysis for the ATLAS results so far without reprocessing. The Tier-0 system has registered more than 15 PB of data in DDM since the first collision at 7 TeV in 2010 (figure 1).

The ATLAS production and distributed analysis system (PanDA) [11] manages about a million of jobs daily, running more than 100,000 jobs concurrently, which has doubled since

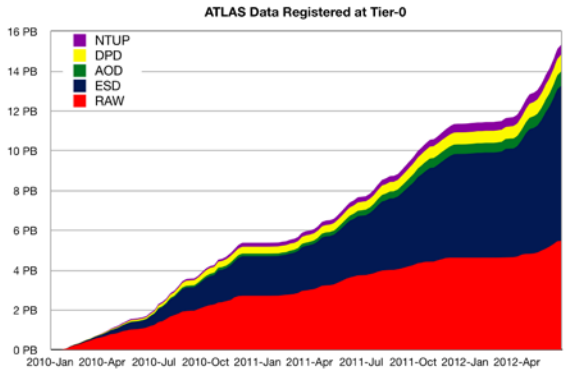

Figure 1. Cumulative data volume registered at the Tier-0 since 2010.

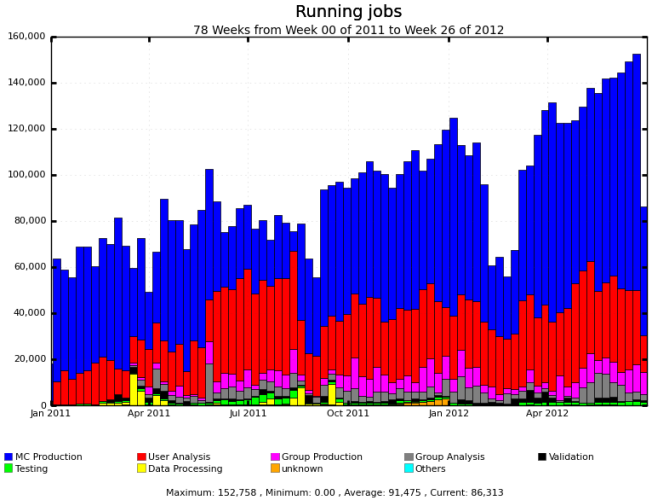

Figure 2. Data processing activities since January 2011.
(ATLAS DDM Dashboard)

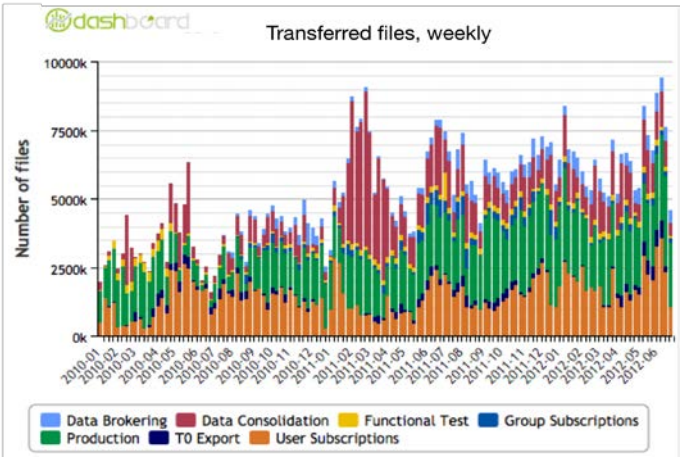

Figure 3. Data transfer activities since January 2010.
(ATLAS Job Historical Views Dashboard)

January of 2011 (figure 2). The majority of the jobs are production of Monte Carlo simulation data and end-user analysis. Occasional reprocessing of RAW data has also been carried out to improve the data quality. ATLAS distributed data management system (DDM) transfers about a million files daily between the Grid sites (figure 3). The major components are planned

distribution of the official data and transfers of the output of the end-user analysis. End-user analysis jobs are submitted with powerful commandline tools and their outputs are transferred to users' "home" site on the Grid, either on request or automatically when the destination is specified at the submission time. The stable operation has been ensured with the shifts organisation [12] as well as an essential help from the automatic control tools to disable and enable the computing and storage resources within the ATLAS systems [13, 14], minimising manual interventions.

3.Further evolution

The ATLAS distributed computing system has been built with the sites equipped with the Grid middleware on top of batch systems, where a job slot corresponds to a physical or logical computing core. This is changing with the new developments. AthenaMP [15], a multi-core implementation of the ATLAS software framework, has been developed for an optimised utilisation of resources. Virtualisation and Cloud computing have been tested with some adjustments in the system, so that we can utilise academic and commercial cloud resources in the ATLAS computing system [16]. It is now an option for the future recource deployment.

The design of the current system is based on the "Data Grid" concept, i.e. "jobs go to data", where the jobs are submitted to the sites hosting the input data to be accessed via LAN. Then, jobs need to be re-assigned to another site when the data at the site is not available for some reason, and popular datasets are replicated to more sites for a higher accessibility. Storage federation provides new access modes and redundancy enabling data access via WAN, where jobs can access data on shared storage resources, with files hosted at remote sites. Such remote access can be more efficient than replicating the whole data locally, since analysis jobs may not need all the information in a file. A system of Xrootd redirectors is a possible working solution today. Work was carried out within US-ATLAS computing facility to develop the concept and to study the performance. Testing is now being extended to a global system.

There are new developments in the ATLAS distributed computing systems, for more features, flexibility and less manual interventions. Rucio is the next ATLAS distributed data management system [6], which enables global management of the space, rather than per-site. It is being developed utilising not only the relational database (Oracle) used in the current system, but also a non-relational structured storage (Hadoop). JEDI is the next ATLAS system for task and job definition integrated into PanDA, which allows dynamic job definition with automatic adjustments of job specifications. The new systems are to be ready for the LHC restart in 2014.

4. Conclusions

The ATLAS distributed computing system has been running extremely well for the large-scale data processing, distribution and analysis. The model and the system has evolved for more flexibility and efficiency, adjusting to varying real conditions during the first years of data-taking. The implication is a more dynamic system. The model and the system continue to evolve with trends of technologies in the computing industry followed closely and used whenever possible. Some new developments are ongoing for more features and flexibility that are

considered necessary based on the experiences, and also to sustain the load in the coming years with new technologies and concepts. These are to be ready for the LHC restart in 2014.

References

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

[2] The ATLAS Collaboration, *ATLAS computing : Technical Design Report*, CERN ATLAS-TDR-017, CERN-LHCC-2005-022

[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-085

R.W.L. Jones et al., *The Evolution of the ATLAS Computing Model*, J. Phys.: Conf. Ser. **219** 072037

[4] I. Ueda for the ATLAS collaboration, *ATLAS Operations: Experience and Evolution in the Data Taking Era*, J. Phys.: Conf. Ser. **331** 072034

I. Ueda for the ATLAS collaboration, *ATLAS Distributed Computing Operations in the First Two Years of Data Taking*, PoS(ISGC 2012)013.

S. Jézéquel et al. on behalf of ATLAS Collaboration, *ATLAS Distributed Computing Operations: Experience and improvements after 2 full years of data-taking*, ATL-SOFT-PROC-2012-026

[5] T. Maeno et al. for The ATLAS Collaboration, *PD2P : PanDA Dynamic Data Placement for ATLAS*, CERN ATL-SOFT-PROC-2012-016

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

V. Garonne et al. on behalf of the ATLAS Collaboration, *The ATLAS Distributed Data Management project: Past and Future*, CERN ATL-SOFT-PROC-2012-049

[7] S. Campana et al. for ATLAS Collaboration, *Evolving ATLAS Computing For Today's Networks*, CERN ATL-SOFT-PROC-2012-027

[8] D. Barberis et al. on behalf of the ATLAS collaboration, *Evolution of grid-wide access to database resident information in ATLAS using Frontier*, CERN ATL-SOFT-PROC-2012-058

[9] A. De Salvo, et al. on behalf of the ATLAS collaboration, *Software installation and condition data distribution via CernVM FileSystem in ATLAS*, CERN ATL-SOFT-PROC-2012-030

[10] J. Schovancová for the ATLAS Collaboration, *ATLAS Distributed Computing Monitoring tools after full 2 years of LHC data taking*, CERN ATL-SOFT-PROC-2012-028

[11] T. Maeno et al. for The ATLAS Collaboration, *Overview of ATLAS PanDA Workload Management*, J. Phys.: Conf. Ser. **331** 072024

T. Maeno et al. for The ATLAS Collaboration, *Evolution of the ATLAS PanDA Production and Distributed Analysis System*, CERN ATL-SOFT-PROC-2012-011

[12] J. Schovancová et al. for the ATLAS Collaboration, *ATLAS Distributed Computing Shift Operation in the first 2 full years of LHC data taking*, CERN ATL-SOFT-PROC-2012-015

[13] F. Legger et al, *Improving ATLAS grid site reliability with functional tests using HammerCloud*, CERN ATL-SOFT-PROC-2012-007

[14] J. Andreeva et al. for the ATLAS Collaboration, *Automating ATLAS Computing Operations using the Site Status Board*, CERN ATL-SOFT-PROC-2012-048

[15] P. Crooks et al. on behalf of the ATLAS collaboration, *Multi-core job submission and grid resource scheduling for ATLAS AthenaMP*, CERN ATL-SOFT-PROC-2012-029

[16] F. H. Barreiro Megino et al. on behalf of the ATLAS Collaboration, *Exploiting Virtualization and Cloud Computing in ATLAS*, CERN ATL-SOFT-PROC-2012-044