ATLAS Distributed Computing Experience and Performance During the LHC Run-2

A Filipčič\textsuperscript{1} for the ATLAS Collaboration

\textsuperscript{1} Jozef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
E-mail: andrej.filipcic@ijs.si

Abstract.

ATLAS Distributed Computing during LHC Run-1 was challenged by steadily increasing computing, storage and network requirements. In addition, the complexity of processing task workflows and their associated data management requirements led to a new paradigm in the ATLAS computing model for Run-2, accompanied by extensive evolution and redesign of the workflow and data management systems. The new systems were put into production at the end of 2014, and gained robustness and maturity during 2015 data taking. ProdSys2, the new request and task interface; JEDI, the dynamic job execution engine developed as an extension to PanDA; and Rucio, the new data management system, form the core of Run-2 ATLAS distributed computing engine.

One of the big changes for Run-2 was the adoption of the Derivation Framework, which moves the chaotic CPU and data intensive part of the user analysis into the centrally organized train production, delivering derived AOD datasets to user groups for final analysis. The effectiveness of the new model was demonstrated through the delivery of analysis datasets to users just one week after data taking, by completing the calibration loop, Tier-0 processing and train production steps promptly. The great flexibility of the new system also makes it possible to execute part of the Tier-0 processing on the grid when Tier-0 resources experience a backlog during high data-taking periods.

The introduction of the data lifetime model, where each dataset is assigned a finite lifetime (with extensions possible for frequently accessed data), was made possible by Rucio. Thanks to this the storage crises experienced in Run-1 have not reappeared during Run-2. In addition, the distinction between Tier-1 and Tier-2 disk storage, now largely artificial given the quality of Tier-2 resources and their networking, has been removed through the introduction of dynamic ATLAS clouds that group the storage endpoint nucleus and its close-by execution satellite sites. All stable ATLAS sites are now able to store unique or primary copies of the datasets.

ATLAS Distributed Computing is further evolving to speed up request processing by introducing network awareness, using machine learning and optimisation of the latencies during the execution of the full chain of tasks. The Event Service, a new workflow and job execution engine, is designed around check-pointing at the level of event processing to use opportunistic resources more efficiently.

ATLAS has been extensively exploring possibilities of using computing resources extending beyond conventional grid sites in the WLCG fabric to deliver as many computing cycles as possible and thereby enhance the significance of the Monte-Carlo samples to deliver better physics results.

The exploitation of opportunistic resources was at an early stage throughout 2015, at the level of 10\% of the total ATLAS computing power, but in the next few years it is expected to deliver much more. In addition, demonstrating the ability to use an opportunistic resource can lead to securing ATLAS allocations on the facility, hence the importance of this work goes beyond merely the initial CPU cycles gained.

In this paper, we give an overview and compare the performance, development effort, flexibility and robustness of the various approaches.
1. Run-2 Experience
During Run-2, the Large Hadron Collider (LHC) delivered an unprecedented number of collision to the experiments, especially in the year 2016, when the delivered integrated luminosity exceeded the planned one by 50% (Figure 1).

![Figure 1. ATLAS Integrated luminosity during LHC Run-1 and Run-2.](image1)

The computing resources allocated to ATLAS [1] were not sufficient to cope with the increased data taking rate in 2016 and subsequently higher Monte-Carlo samples demand, however many ATLAS computing sites provided much more CPU power than officially pledged ensuring the smooth processing and on-time delivery of results to the physicists. In addition, the new production framework developed during the Long Shutdown 1 performed extremely well and efficiently exploited all the resources available to ATLAS, including the sites that opportunistically provided CPU power.

2. Production and Data Management System Evolution
The new demands for complex production workflows during Run-1 required a complete overhaul of the production and data management systems (Figure 2).

![Figure 2. New ATLAS production and data management system.](image2)

Rucio [2] is an optimised and scalable data management system with many improvements and features, such as short transfer latency, efficient and fast data placement engine, extended dataset and file metadata support. It is also demonstrated to be highly scalable and could satisfy ATLAS needs at least till the end of Run-3.

The workflow management system consists of two major components. ProdSys2 [3] is the request and task management system with a workflow engine relying on task transformation of input datasets to the output datasets and automatically organising the chains of linked tasks. Any kind of a new transformation can be quickly implemented due to its efficient architecture and modularity. JEDI [4] is the new payload management system, providing the job workflow interaction with the sites and automatically managing the task requirements and resource allocations. Analysis and production are both using the same workflow infrastructure.

Numerous improvements have been made in ATLAS computing to improve the computing performance for Run-2. Due to the rigidity of the MONARC model [5] which did not suite well the actual distribution of ATLAS resources and underestimated the role of the network connectivity between the ATLAS computing sites, a more flexible model of dynamic storage and execution sites has been established. Every site with a sufficiently stable storage and good network connectivity can store primary data (Nucleus site), while the close sites in terms of network throughput and latency, including the Nucleus site, execute the payloads (Satellite sites) [6]. All the associations are fully dynamic at the task and job brokering level.
In the past, most of the jobs were allocated predefined and static amount of resources. While this worked sufficiently well for the startup of Run-1, already towards its end it become clear that the old model was not good enough. The job memory requirements and execution walltimes can change drastically for various payloads, typically due to more complex data or Monte-Carlo event structure. Each task is now probed at the startup, where the first 10 jobs measure the memory and cpu consumption. JEDI automatically tunes the subsequent jobs of the task to better fit the expected resource consumption. If jobs are failing for exceeded resource limits, the resources for job retries are also automatically increased.

At the request level, the tasks close to completion are proactively pushed to finish the pending payloads, especially in cases where the requests need to complete for close deadline, typically due to approaching physics conference.

3. Flexible and Dynamic Resource Allocation
Increased memory usage of simulation and reconstruction jobs led to development of AthenaMP [7] where memory is efficiently shared by processes of a multi-core payload, thus reducing the overall memory consumption.

![Figure 3. Slots of running jobs by core count.](image)

![Figure 4. CPU usage by activity in 2016.](image)

This inevitably required a restructuring of the job allocation strategies, to support a dynamic reservation and scheduling of the multi-core batch jobs. Figure 3 shows a distribution of cpu slots per core count during the year 2016 demonstrating a fully adaptable infrastructure and allocation for multi-core payloads.

Some activities, typically the upgrade studies or heavy-ion data reconstruction, require much more memory than typically available on computing sites. The custom high-memory queues and sites with fully dynamic resource allocation provide a way to execute these extreme payloads on expense of the unused compute cores on the worker nodes.

4. CPU Usage During 2016
ATLAS has been using much more CPU resources than pledged (Figure 4). The major contribution comes from leveraging the so-called over-pledge resources, provided by WLCG sites [8] which have larger CPU capacity installed, typically due to shared nature of their activities.

The over-pledged contribution amounts to 50% in 2016. In addition, ATLAS is efficiently using the non-WLCG sites which opportunistically provide computing power for production. Such sites are High-Performance Computing centers in US, Europe and China, academic and
commercial cloud infrastructures, shared national computing facilities and volunteers providing their home personal computers through ATLAS@Home [9].

While the contribution of opportunistic sites was relatively low in 2016 (15%), a high increase is expected in the following few years, when production on such sites will stabilise and will be transparently included in ATLAS production infrastructure. Also, a significant increase in allocations on those sites is expected in the near future.

5. Data Management

ATLAS data volume exceeded 200PB in 2016 on disk and tape storage endpoints. There were more than 100,000 datasets with close to one billion files stored in the ATLAS computing centers. The primary data is partially replicated, but due to high demand for storage, the replication factor is only at 0.3, as shown in Figure 5.

The data transfers between the sites were reaching an aggregated traffic of 20GB/s weekly average, with daily average up to 40GB/s in some periods of high activity (Figure 6). Rucio was able to transfer more than 50 files per second, where the largest contribution to transfers was the delivery of input files to the payload execution sites. In particular, the most input/output demanding activity was production of the derived datasets (DAOD) which served as the main source for physics analysis.

Figure 5. ATLAS data volume distribution.  
Figure 6. Data transfers between ATLAS sites aggregated by transfer activity.

Figure 7. Lifetime model effectively cleans old unused data.  
Figure 8. CPU consumption of ATLAS processing activities.
To deal with a larger demand for storage space, a new data persistency model has been put in production. The so-called lifetime model ensures that each data type is set a finite lifetime. The precious RAW data has infinite lifetime, while all the other datasets are limited, from 6 months for DAOD which have typically fast turnaround, to up to 3 years for cpu-expensive Monte-Carlo production. Figure 7 shows the effect of the lifetime model after being put in production, where the volume of data older than 1 year has been drastically reduced. Lifetime of frequently used data is automatically extended. The monthly cleanup procedure involves an announcement of the data to be removed, while the physics groups can ask for extensions of datasets in interest.

6. Production Activities in 2016
The major production activity in 2016 was the Monte-Carlo simulation and reconstruction campaign (MC15c), that used 94% of CPU resources and produced more than five billion of fully simulated events and reconstructed them with full 2016 data-taking conditions (Figure 8).

The second major campaign was processing of 2015 heavy-ion data, which was delayed due to unexpectedly large cpu and memory consumption and was not promptly processed on Tier-0 as initially planned. Together with a partial proton-proton data reprocessing, it consumed about 4% of CPU resources.

The third major activity, at least in terms of the processed data volume, was the derivation production with several campaigns, typically taking two to three weeks to fully process Monte-Carlo and real data. As shown in Figure 9, the derivation production is based on simultaneous production of various output streams of analysis datasets (DAODs), as requested by ATLAS physics groups. The benefit to this new mode is the organised production which optimizes the data processing and transfers by packing many analysis filtering end data enrichment steps in a single payload (train). An important achievement of derivation production was a prompt delivery of final physics samples to physicists as fast as on week after data was recorder by the ATLAS detector.

7. Upcoming Features
The ATLAS production system and infrastructure is still evolving. Some major new features are being developed.

![Figure 9. Derivation Production workflow.](image-url)
One of the major difficulties in the past was a proper allocation of CPU resources to specific activities and finding the correct balance between Monte-Carlo production, data reprocessing, derivation production and physics analysis. The global fair-share mechanism with nested activity shares will enable to allocate a desired number of CPU slots to a particular activity and its sub-activities, to guarantee that a particular campaign will be completed by the targeted date, or boost a selected activity on a short time scale to complete the urgent delivery results. This will in the future reduce the dependency on the fixed and site specific batch system shares which are typically used to balance the CPU usage between the production and analysis.

Machine learning techniques and data analytics are becoming widely used in ATLAS computing. The logs of all the computing activities are pushed to the central analytics platform enabling in-detail monitoring and studies of complex interactions between job execution, data transfers and placements, network throughput, task progress and infrastructure reliability or failures. Network Weather Service is measuring the overall network and transfers performance between all the sites and is used to optimize the payload and data distribution. Job error and analysis algorithms are being developed to automatically handle site or central services failures.

To better exploit opportunistic or short lived resources, ATLAS is developing the ATLAS Event Service [10] where a long running batch job executes many short payloads with the granularity as high as a single event processing. The ATLAS Event Service is under commissioning for the ATLAS Monte-Carlo simulation, while other workflows will be considered as well. An additional benefit of the ATLAS Event Service for grid sites will be the reduction of the site downtime impact on production. The jobs can run till the worker nodes are shutdown for maintenance, so the long draining periods can be greatly reduced.

8. Job Execution on Opportunistic Resources

The difficulties of using opportunistic resources come from architectural differences such as unavailability of grid services, the absence of network connectivity on worker nodes or inability to use standard authorization protocols. Nevertheless, ATLAS has been extremely successful in running production payloads on a variety of sites, thanks largely to the job execution workflow design in which the job assignment, input data provisioning and execution steps are clearly separated and can be offloaded to custom services (see Figures 10 and 11). To transparently include the opportunistic sites in the ATLAS central production system, several models with supporting services have been developed to mimic the functionality of a full WLCG site. Some are extending Computing Element services to manage job submission to non-standard local resource management systems, some are incorporating pilot functionality on edge services managing the batch systems, while the others emulate a grid site inside a fully virtualised cloud environment. To overcome those issues, the execution workflow had to be separated in steps that need external connectivity and steps than can be included on the worker node. The basic payload workflow steps are: payload distribution, input data staging, software distribution, database prefetching, payload execution, output data transfers, and communication with central services. All the steps but the payload execution need to be offloaded to an external service, either local to a site, or remote.

The ATLAS software is usually distributed through cvmfs [11], while for HPCs, a partial copy of cvmfs needs to be provided on the shared filesystem. The other options are using parrot [12] or docker [13] images pre-filled with a targeted software release. ATLAS@Home uses pre-filled cvmfs cache to reduce the network traffic. The database is typically provided through a DBRelease file including an sqlite subset of conditions for a targeted Monte-Carlo simulation campaign. Some implementations of custom step execution separation are shown in Figure 12, each targeting a specific subset of computing sites. The ATLAS production system was designed from ground up to provide a seamless distribution of job steps, while the new workflow and data management system enabled even tighter integration.
9. Conclusions
Big efforts to evolve and partially redesign the ATLAS production and data management systems have payed off. The system is robust, mature, performant and much lighter in operational effort than the system used during Run-1. It can easily cope with higher LHC performance in Run-2 than initially expected. At present, there are no scaling issues. Each subsystem was tested to sustain 5-10 times larger load than currently used. Although this might not be sufficient in the longer term for Run-4, the architecture is horizontally scalable and modular to enable component improvement and further scalability optimisations.

The processing power, storage capacities and network were used to their limits in 2016. Therefore, a careful planning of CPU and storage usage for further campaigns will be necessary, cleaning the unused data more aggressively and better optimizing the network usage to avoid the saturation. Usage of opportunistic sites is crucial for ATLAS in the upcoming few years and can bring a significant boost of CPU resources.
References

[1] ATLAS Collaboration 2008 JINST 3 S08005
[2] Garonne V et al, on behalf of the ATLAS Collaboration 2016 Experiences with the new ATLAS distributed data management system. Proceedings of the 22nd International Conference on Computing in High Energy and Nuclear Physics, J. Phys.: Conf. Ser.
[3] Borodin M et al, on behalf of the ATLAS Collaboration 2016 The ATLAS production system evolution. new data processing and analysis paradigm for the LHC run2 and high-luminosity. Proceedings of the 22nd International Conference on Computing in High Energy and Nuclear Physics, J. Phys.: Conf. Ser.
[4] Maeno T et al, on behalf of the ATLAS Collaboration 2011 J. Phys.: Conf. Ser. 331 072024
[5] The MONARC project web site URL http://monarc.web.cern.ch/MONARC/
[6] Barreiro Megino F H et al, on behalf of the ATLAS Collaboration 2016 ATLAS world-cloud and networking in panda. Proceedings of the 22nd International Conference on Computing in High Energy and Nuclear Physics, J. Phys.: Conf. Ser.
[7] Calafiura P et al, on behalf of the ATLAS Collaboration 2015 J. Phys.: Conf. Ser. 664 072050
[8] Worldwide LHC Computing Grid web site URL http://cern.ch/lcg
[9] Cameron D et al, on behalf of the ATLAS Collaboration 2016 Volunteer computing experience with ATLAS@Home. Proceedings of the 22nd International Conference on Computing in High Energy and Nuclear Physics, J. Phys.: Conf. Ser.
[10] Tsulaia V et al, on behalf of the ATLAS Collaboration 2016 Production experience with the ATLAS event service Proceedings of the 22nd International Conference on Computing in High Energy and Nuclear Physics, J. Phys.: Conf. Ser.
[11] CernVM File System web site URL http://cernvm.cern.ch/portal/filesystem
[12] Parrot web site URL https://cernvm.cern.ch/portal/filesystem/parrot
[13] Docker web site URL https://www.docker.com/