

Multilevel Workflow System in the ATLAS Experiment

M Borodin¹, K De², J Garcia Navarro³, D Golubkov^{4,5}, A Klimentov⁶, T Maeno⁶ and A Vaniachine⁷ on behalf of the ATLAS Collaboration

¹ Department of Elementary Particle Physics, National Research Nuclear University "MEPhI," Moscow, 117513, Russia

² Physics Department, University of Texas Arlington, Arlington, TX 76019, United States of America

³ Instituto de Fisica Corpuscular, Universidad de Valencia, E-46980 Paterna, Spain

⁴ Experimental Physics Department, Institute for High Energy Physics, Protvino, 142281, Russia

⁵ Big Data Laboratory, National Research Centre "Kurchatov Institute" Moscow, 123182, Russia

⁶ Physics Department, Brookhaven National Laboratory, Bldg. 510A, Upton, NY 11973, United States of America

⁷ High Energy Physics Division, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439, United States of America

E-mail: vaniachine@anl.gov

Abstract. The ATLAS experiment is scaling up Big Data processing for the next LHC run using a multilevel workflow system comprised of many layers. In Big Data processing ATLAS deals with datasets, not individual files. Similarly a task (comprised of many jobs) has become a unit of the ATLAS workflow in distributed computing, with about 0.8M tasks processed per year. In order to manage the diversity of LHC physics (exceeding 35K physics samples per year), the individual data processing tasks are organized into workflows. For example, the Monte Carlo workflow is composed of many steps: generate or configure hard-processes, hadronize signal and minimum-bias (pileup) events, simulate energy deposition in the ATLAS detector, digitize electronics response, simulate triggers, reconstruct data, convert the reconstructed data into ROOT ntuples for physics analysis, etc. Outputs are merged and/or filtered as necessary to optimize the chain. The bi-level workflow manager – ProdSys2 – generates actual workflow tasks and their jobs are executed across more than a hundred distributed computing sites by PanDA – the ATLAS job-level workload management system. On the outer level, the Database Engine for Tasks (DEFT) empowers production managers with templated workflow definitions. On the next level, the Job Execution and Definition Interface (JEDI) is integrated with PanDA to provide dynamic job definition tailored to the sites capabilities. We report on scaling up the production system to accommodate a growing number of requirements from main ATLAS areas: Trigger, Physics and Data Preparation.

1. Introduction

The multi-purpose nature of the ATLAS experiment [1] at the LHC resulted in continuous growth in use cases for Big Data processing, as more data and new requirements emerge. To process Big Data the ATLAS experiment adopted the dataset transformation approach, where software applications



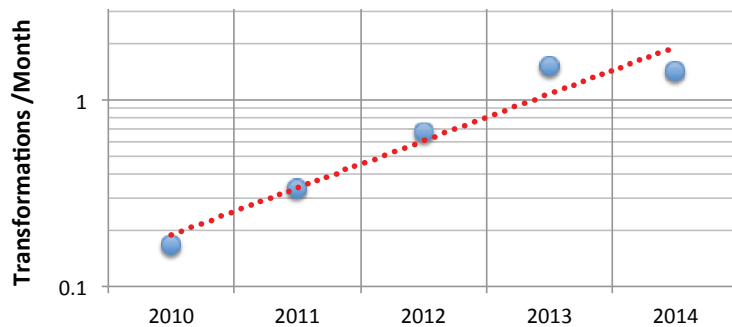


Figure 1. Continuous growth in the rate of data transformations added for Big Data processing in the ATLAS experiment.

transform the input datasets into the output datasets [2]. A success of this approach is evident from the exponential growth rate in the number of new data transformations used for Big Data processing in the ATLAS experiment (figure 1). Figure 2 shows the number of datasets produced during the major Run 1 simulation campaign dominated by datasets of SUSY physics group. The names of other physics and combined performance groups shown on the figure are: phys-exotics, phys-higgs, phys-gener (event generators), physics-sm (Standard Model), physics-top, soft-simul (software validation for the Monte Carlo simulations), perf-flavtag (flavour tagging), perf-jets, phys-beauty, perf-egamma (electron and gamma), and perf-tau.

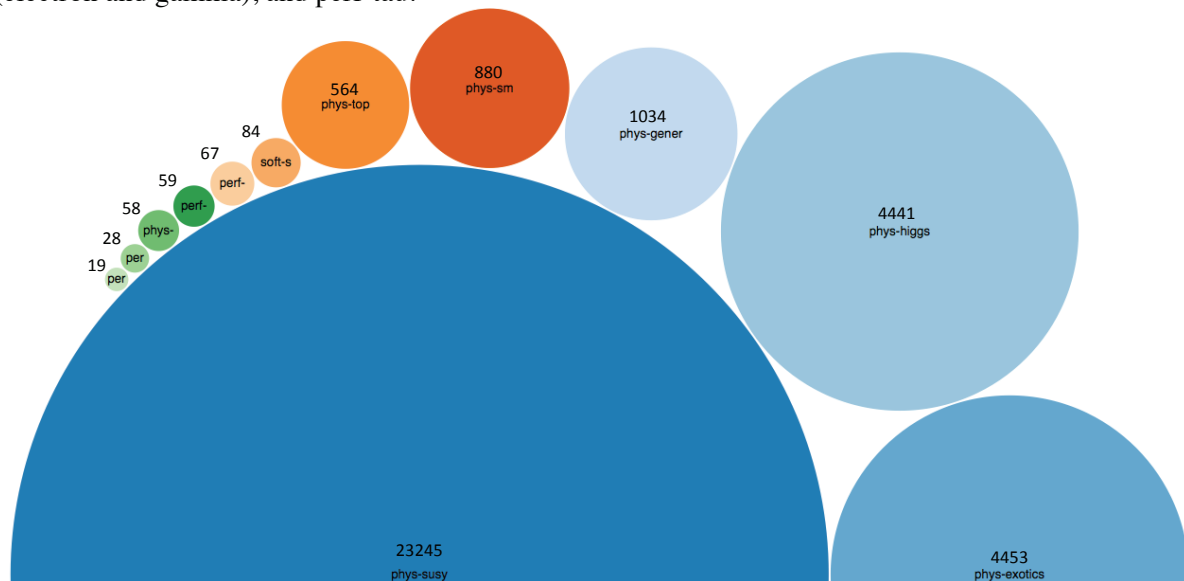


Figure 2. The number of datasets produced during the major Run-1 simulation campaign represents the scale and variety of Big Data processing requirements provided by ATLAS physics groups.

In ATLAS Big Data processing, each dataset transformation is represented by a task [3], having the individual data processing tasks organized into workflows with outputs merged and/or filtered as necessary. Figure 3 shows the Monte Carlo workflow steps: generate or configure hard-processes, hadronize signal and minimum-bias (pileup) events, simulate energy deposition in the ATLAS detector, digitize electronics response, simulate triggers, reconstruct data, convert the reconstructed data into ntuples for physics analysis, etc.



Figure 3. Monte Carlo workflow is composed of many steps.

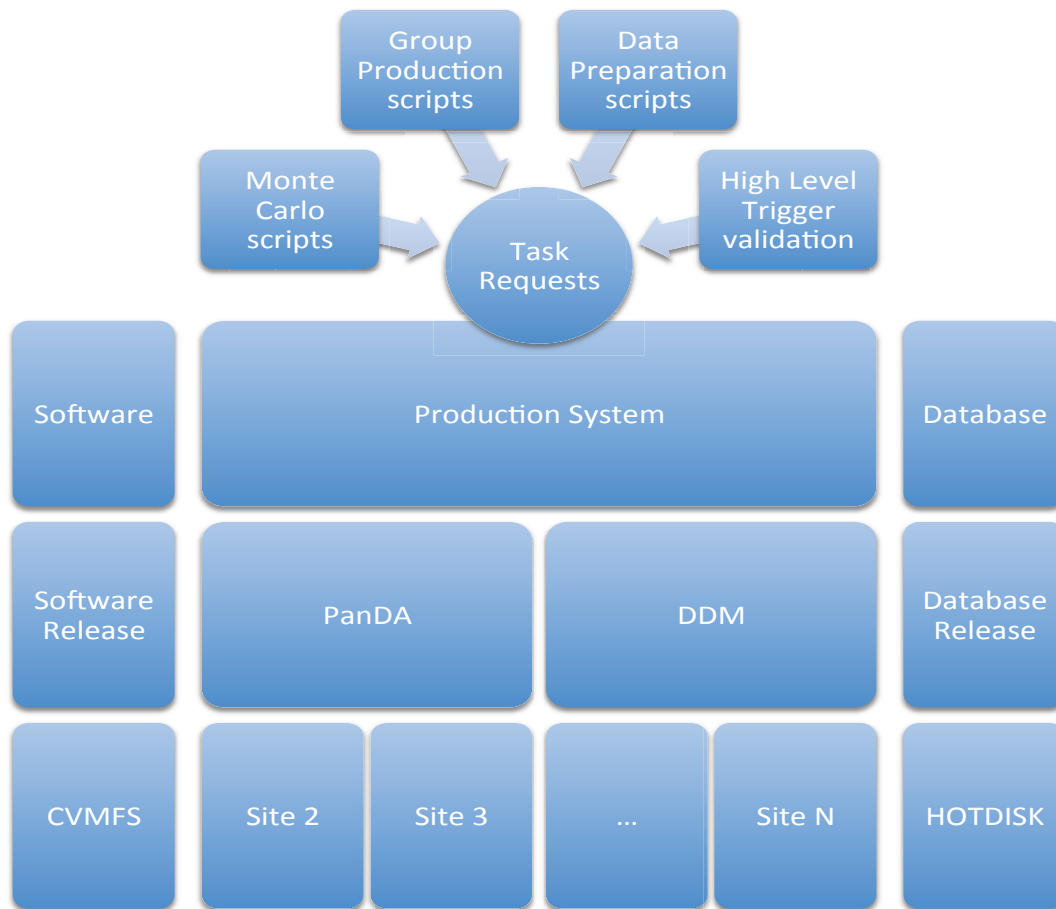


Figure 4. The ATLAS Big Data processing infrastructure during the LHC Run 1.

2. Big Data Processing

Figure 4 shows the multilevel Big Data processing infrastructure used by the ATLAS experiment during and after the LHC Run 1. The top management layer includes the Production System with the front-end Task Request interface [4]. In the middle are the workload management system PanDA [5] and the Distributed Data Management (DDM) system DQ2 [6]. The bottom layer shows Grid sites together with the software and database access technologies deployed at the sites. During Run 1, the infrastructure fully satisfied the requirements of ATLAS data reprocessing, simulations, and production by physics groups. As an example, we report below our experience with a representative use case of data reprocessing.

2.1. Run 1 Experience

A starting point for ATLAS data processing is data reconstruction. During reconstruction, the raw detector data are processed with software algorithms to identify and reconstruct physics objects such as charged particle tracks. Following the prompt reconstruction at the computing centre at CERN (the Tier-0 site), the ATLAS data are reprocessed on the Grid, which allows reconstruction of the data with updated software and calibrations improving the quality of the reconstructed data for physics analysis. The collaboration completed four major reprocessing campaigns, with up to 2 PB of data being reprocessed every year. Automatic job resubmission avoids data losses at the expense of CPU time used by the failed jobs. Table 1 shows that failures have not presented a problem, as the fraction of CPU-time used for data recovery is not significant.

Table 1. Cost of recovery from transient failures for the reconstruction jobs.

Reprocessing campaign	Input Data Volume (PB)	CPU Time Used for Reconstruction ($10^6 h$)	Fraction of CPU Time Used for Recovery (%)
2010	1	2.6	6.0
2011	1	3.1	4.2
2012	2	14.6	5.6
2013	2	4.4 ¹	3.1

Another example of the success of our task-based data transformation approach is represented by the exponential growth of the rate of production tasks submission over the years [4]. As a result, the ATLAS production tasks count exceeded 1.6 million, with each task containing hundreds or thousands of jobs submitted by PanDA for execution on the Grid.

2.2. Run 2 Preparations

The LHC shutdown presented an opportunity to reengineer the ATLAS Big Data processing infrastructure, adding extra layers to further improve the system scalability and flexibility. To prepare the production for Run 2 challenges, PanDA has been upgraded with the Job Execution and Definition Interface (JEDI) [5], the production system enhanced with the Database Engine for Tasks (DEFT) [7], and the DDM system being upgraded to Rucio [8] (figure 5). On the production system upper level, the Database Engine for Tasks (DEFT) empowers production managers with templated workflow definitions. On the lower level, the Job Execution and Definition Interface (JEDI) is integrated with PanDA to provide dynamic job definition tailored to the sites capabilities. The data transformations will be configured using the ATLAS Metadata Interface (AMI) [9] harmonizing definitions of AMI tags between Tier-0 and the production system.



Figure 5. Multi-level architecture of the ATLAS Big Data processing infrastructure for Run 2.

In the bi-level production system, the JEDI layer is coupled with PanDA, while the DEFT layer implemented as the flexible database engine for bookkeeping. These two independent layers communicate via customized JSON protocol. During task execution, the JEDI layer defines the jobs

¹ In 2013 reprocessing, 2.2 PB of input data were used for selecting about 15% of all events for reconstruction, thus reducing CPU resources vs. the 2012 reprocessing.

tailored to the actual resources: disk space, CPU-time, memory, networks, etc. In contrast, the Run 1 production system employed the static job definition.

3. Conclusions

During the LHC Run 1, the ATLAS Big Data processing infrastructure supported a diverse range of workflows handling centrally ATLAS Big Data reprocessing (reconstruction of LHC data) and Monte Carlo production (full and fast simulations, digitization and reconstruction of simulated data). The production system scalability and flexibility has been demonstrated by managing double exponential growth in the number of task requests and data transformations. The total number of tasks exceeded 1.6 million; the data transformations count exceeded 120.

The LHC shutdown provided an opportunity for upgrading the production system with new capabilities, such as automatic recovery of lost data files and tailoring jobs to the capabilities of particular sites. As the ATLAS experiment continues optimising the use of Grid computing resources in preparation for the LHC Run 2 in 2015, the next generation production system is being integrated with other layers. The commissioning is in progress, scaling up the production system for a growing number of tasks and transformations that will process data for physics analysis and other ATLAS LHC Run 2 activities.

Acknowledgments

We thank all our colleagues who contributed to the ATLAS Big Data processing infrastructure development and operations. This work was funded in part by the U. S. Department of Energy, Office of Science, High Energy Physics under Contract No. DE-AC02-06CH11357.

References

- [1] The ATLAS Collaboration 2008 The ATLAS experiment at the CERN Large Hadron Collider *J. Inst.* **3** S08003
URL <http://iopscience.iop.org/1748-0221/3/08/S08003>
- [2] Stewart G A *et al.* 2014 ATLAS Job Transforms: A Data Driven Workflow Engine *J. Phys.: Conf. Ser.* **513** 032094
URL <http://iopscience.iop.org/1742-6596/513/3/032094>
- [3] Vaniachine A V for the ATLAS Collaboration 2011 ATLAS detector data processing on the Grid *IEEE Nuclear Science Symposium and Medical Imaging Conference* pp 104-107
URL <http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6154460>
- [4] Golubkov D *et al.* 2012 ATLAS Grid Data Processing: system evolution and scalability *J. Phys.: Conf. Ser.* **396** 032049
URL <http://iopscience.iop.org/1742-6596/396/3/032049>
- [5] Maeno T *et al.* 2012 Evolution of the ATLAS PanDA production and distributed analysis system *J. Phys.: Conf. Ser.* **396** 032071
URL <http://iopscience.iop.org/1742-6596/396/3/032071>
- [6] Garonne V *et al.* 2012 The ATLAS Distributed Data Management project: Past and Future *J. Phys.: Conf. Ser.* **396** 032045
URL <http://iopscience.iop.org/1742-6596/396/3/032045>
- [7] De K *et al.* 2014 Task management in the new ATLAS production system *J. Phys.: Conf. Ser.* **513** 032078
URL <http://iopscience.iop.org/1742-6596/513/3/032078>
- [8] Garonne V *et al.* 2014 Rucio – the next generation of large scale distributed system for ATLAS data management *J. Phys.: Conf. Ser.* **513** 042021
URL <http://iopscience.iop.org/1742-6596/513/4/042021>
- [9] Fulachier J *et al.* 2014 Looking back on 10 years of the ATLAS Metadata Interface. Reflections on architecture, code design and development methods *J. Phys.: Conf. Ser.* **513** 042019
URL <http://iopscience.iop.org/1742-6596/513/4/042019>