Scaling up ATLAS production system for the LHC Run 2 and beyond: project ProdSys2

2015 J. Phys.: Conf. Ser. 664 062005

(http://iopscience.iop.org/1742-6596/664/6/062005)

Scaling up ATLAS production system for the LHC Run 2 and beyond: project ProdSys2

M Borodin[1,2], K De[3], J Garcia Navarro[4], D Golubkov[1,5], A Klimentov[6], T Maeno[6], and A Vaniachine[7] on behalf of the ATLAS Collaboration

[1] Big Data Laboratory, National Research Centre "Kurchatov Institute", Moscow, Russia
[2] Department of Elementary Particle Physics, National Research Nuclear University "MEPhI", Moscow, Russia
[3] Physics Department, University of Texas at Arlington, TX, United States of America
[4] Instituto de Fisica Corpuscular, Universidad de Valencia, Spain
[5] Experimental Physics Department, Institute for High Energy Physics, Protvino, 142281, 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: aak@bnl.gov

Abstract. The Big Data processing needs of the ATLAS experiment grow continuously, as more data and more use cases emerge. For Big Data processing the ATLAS experiment adopted the data transformation approach, where software applications transform the input data into outputs. In the ATLAS production system, each data transformation is represented by a task, a collection of many jobs, submitted by the ATLAS workload management system (PanDA) and executed on the Grid. Our experience shows that the rate of task submission grows exponentially over the years. To scale up the ATLAS production system for new challenges, we started the ProdSys2 project. PanDA has been upgraded with the Job Execution and Definition Interface (JEDI). Patterns in ATLAS data transformation workflows composed of many tasks provided a scalable production system framework for template definitions of the many-tasks workflows. These workflows are being implemented in the Database Engine for Tasks (DEfT) that generates individual tasks for processing by JEDI. We report on the ATLAS experience with many-task workflow patterns in preparation for the LHC Run 2.

1. Introduction

In 2015 the Large Hadron Collider will reach instantaneous luminosities exceeding $2{\cdot}10^{34}$ cm^{-2}s^{-1} and centre of mass energies of 13 TeV. The physics goals of the ATLAS experiment [1] include searches for physics beyond the Standard Model and high precision Higgs sector studies. These goals require detailed comparison of the expected physics and detector behaviour with data. A rich set of computational models is employed to provide simulated data needed for these comparisons.

To address the corresponding Big Data processing challenge, the LHC experiments employ the computational infrastructure of the Worldwide LHC Computing Grid (WLCG) – world's largest academic distributed computing environment [2]. Thanks to the outstanding LHC performance, ATLAS manages over 160 petabytes of data on more than hundred computational sites. Following Big Data processing, more than ten thousand scientists analyse LHC data in search of new phenomena. ATLAS leads the WLCG usage in the number of data processing jobs and processed data volume.

Leveraging the underlying job management system PanDA [3], the production system orchestrates ATLAS data processing applications for efficient usage of more than a hundred thousand CPU cores provided by the WLCG. In order to manage the diversity of LHC physics (exceeding 35 000 physics samples per year), the individual data processing tasks are organized into workflows. During data processing the system monitors site performance and supports dynamic sharing minimizing the workflow duration. In addition, the production system manages jobs and/or task failures enhancing the resilience.

In preparation for data taking, the ATLAS experiment is scaling up its Big Data capabilities by upgrading to a multilevel production system. In the next section we describe our experience with representative data processing use cases handled by the production system.

Table 1. Use cases representing variety of data processing requirements.

Use Case	Frequency	Workflow Length	Number of Tasks	Number of Events	Tasks Duration	Data Loss
Trigger Data	Weekly	Short	Several	Millions	Hours	no
Real Data	Yearly	Medium	Hundreds	Billions	Weeks	no
Simulated Data	Quarterly	Long	Thousands	Billions	Months	yes

2. Big data processing use cases

To process Big Data, the ATLAS experiment adopted the data transformation technique, where software applications transform the input datasets of one data type into the output datasets of another data type. In data processing ATLAS deals with datasets, not individual files. Similarly a task (comprised of many jobs) has become a unit of the workflow in ATLAS processing. The successful validation of this technique was achieved through the exponential growth rate in the number of new data transformations and data types used for data processing in the ATLAS experiment. Table 1 lists representative use cases described below. One of the differences in data processing requirements is that losses are not tolerated for the real data, while the simulated data samples tolerate losses, which reduce the statistics without physics bias.

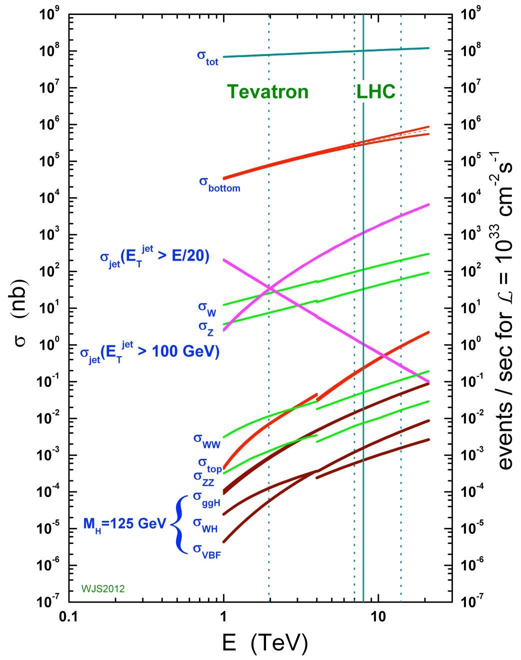


Figure 1. Comparison of cross-sections and rates for “known” and “rare” events in proton-(anti)proton collisions [4].

2.1. Trigger data processing

New physics discoveries and high precision studies of rare events require the rejection of the events arising from “known” processes by more than ten orders of magnitude (Figure 1). The multi-tier trigger system reduces volume to a manageable level. In 2015, the ATLAS experiment will have a two-tier trigger system:

- the hardware-based Level 1 trigger;
- the software-based High-Level Trigger [5];

The Trigger Data Processing happens one step before the raw data recording. Thus, any inefficiencies or mistakes may lead to unrecoverable loss of real data. To eliminate such losses, the dedicated raw-to-raw data processing technique is employed to validate trigger software and other critical trigger changes during data taking. This technique is the main tool for commissioning the trigger for data taking.

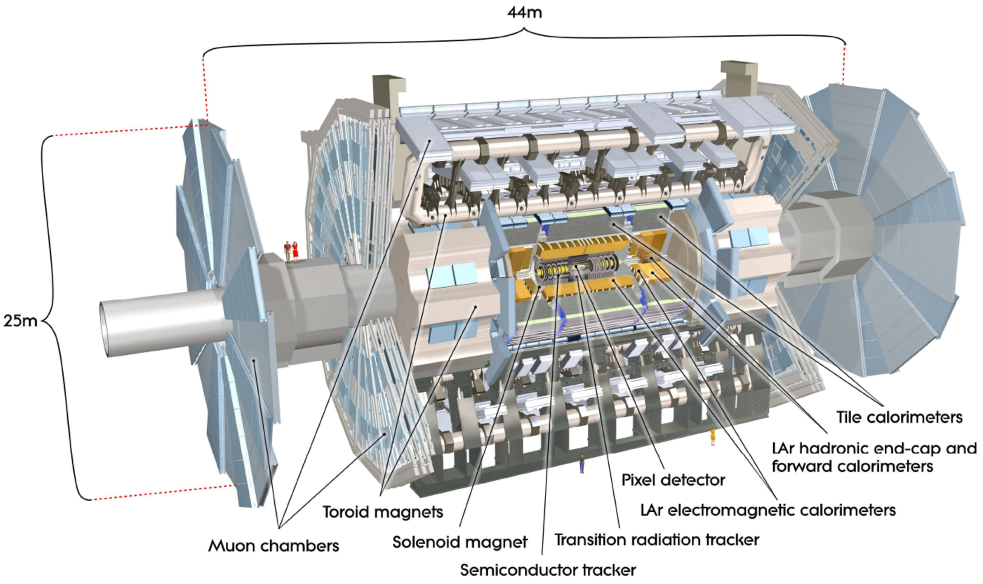

44m
25m
Tile calorimeters
LAr hadronic end-cap and forward calorimeters
Pixel detector
LAr electromagnetic calorimeters
Toroid magnets
Muon chambers
Solenoid magnet
Transition radiation tracker
Semiconductor tracker

Figure 2. The ATLAS detector.

2.2. Real data processing

The raw data from the ATLAS detector (Figure 2) are processed to produce the reconstructed data for physics analysis. During reconstruction, ATLAS applications process raw detector data to identify and reconstruct physics objects such as leptons. The distributed multi-tier data processing architecture handles the petascale data flow [6]. Since the detector data are comprised of independent events, massively parallel applications process one event at a time. Events taken during an interval of a few minutes are collected in one file. Thousands of files with events that are close in time are collected in one dataset.

The ATLAS collaboration has completed four petascale data processing campaigns on the Grid, with up to 2 PB of real data being processed every year. Table 2 lists parameters for the ATLAS yearly data processing campaigns.

Table 2. Processing campaigns for real data.

Campaign year	Input Data Volume (PB)	CPU Time Used for Reconstruction ($10^6 h$)
2010	1	2.6
2011	1	3.1
2012	2	14.6
2013[a]	2	4.4

[a] In 2013 reprocessing, 2 PB of input data were used for selecting about 15% of all events for reconstruction, thus reducing CPU resources vs. the 2012 reprocessing.

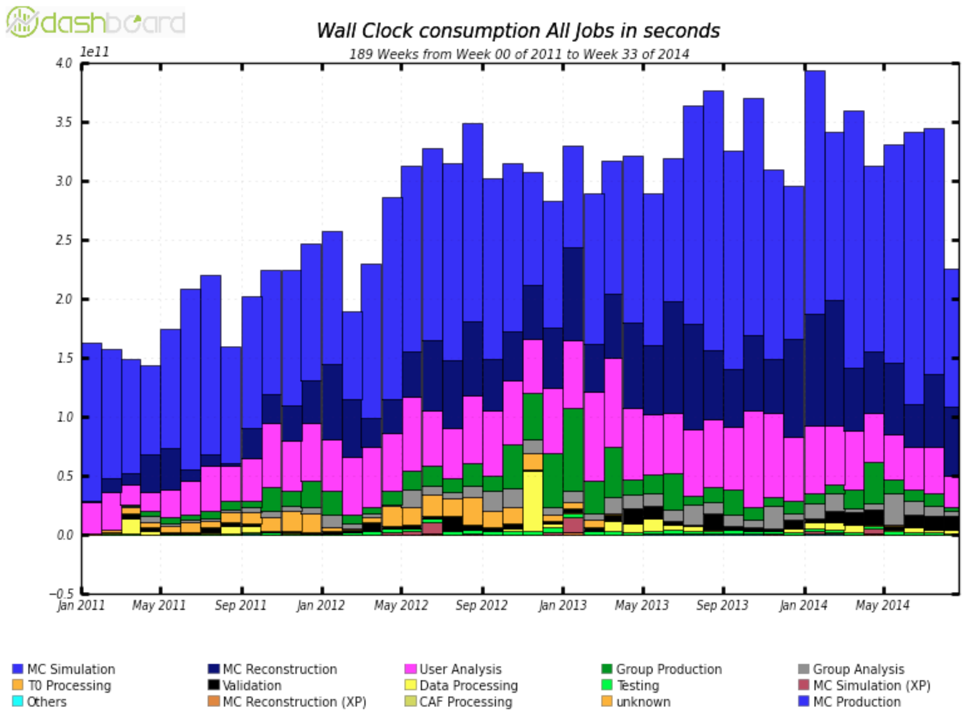


Figure 3. The consumption of computational resources in ATLAS.

2.3. Simulated data processing

The computational resources required to process the simulated data dominate the overall resource usage (Figure 3). The data processing campaigns for the simulated data correspond to the data taking periods of the real data. The LHC data taking periods of the same conditions are characterized by the same centre-of-mass energy, instantaneous luminosity, detector configuration, etc. Table 3 lists the major data processing campaigns for the simulated data, while Figure 4 shows the rate of the simulated events produced.

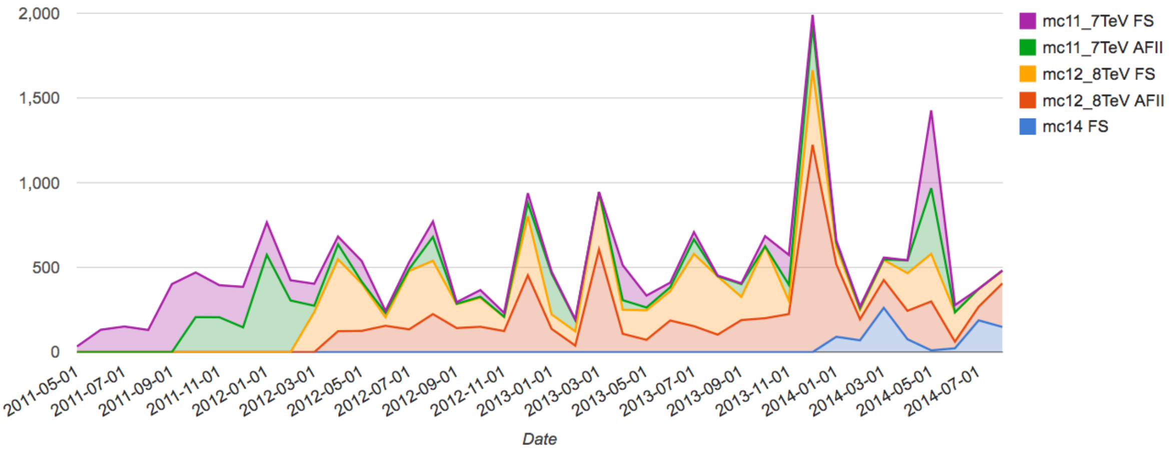


Figure 4. Monthly rate of the simulated events produced (10^6). The Full Simulations are labelled FS, the Fast Simulations are labelled AFII. Note, that the production system processes several concurrent campaigns.

Table 3. Data processing campaigns for simulated data.

Campaign Label	Data Taking Period for Real Data	Configuration	Full Simulation (10^9 events)	Fast Simulation (10^9 events)	Number of Sub-campaigns
mc11	2011	7 TeV	3.64	3.27	4
mc12	2012	8 TeV	6.37	6.43	3
mc14	2012 & 2015	8 & 13 TeV	0.85		2

The LHC instantaneous luminosities result in the presence of a large number of simultaneous collisions in the same event, overlapping the hard scattering event of interest. The presence of the minimum bias events is usually called “pileup”. To provide realistic simulation of these conditions, the data processing workflow for simulated data is composed of many steps (Figure 5): generate hard processes, hadronize signal and minimum bias (pileup) events, simulate energy deposition in the ATLAS detector, digitize electronics response, simulate triggers, reconstruct data, transform the reconstructed data into data types for physics analysis, etc. The intermediate outputs are merged and/or filtered as necessary to optimize the chain.

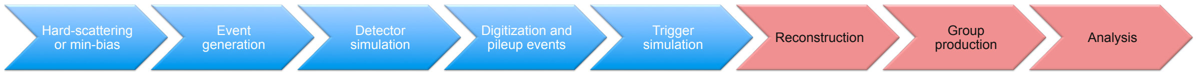


Figure 5. ATLAS simulation workflow.

An example of a more complex workflow used to simulate the ATLAS trigger using dedicated hardware for fast tracking (FTK) [7] is shown in Figure 6, where to keep the computational resources for the FTK simulation within practical limits, we split every event into 256 η-ϕ sub-regions [8]. In the three-step workflow, each event is processed by 64 jobs; each job simulates tracks in four FTK sub-regions one after another. The sub-region merging is done in two steps: producing whole regions, then whole events in the n-tuples files. The final step uses FTK tracks in trigger simulations producing the

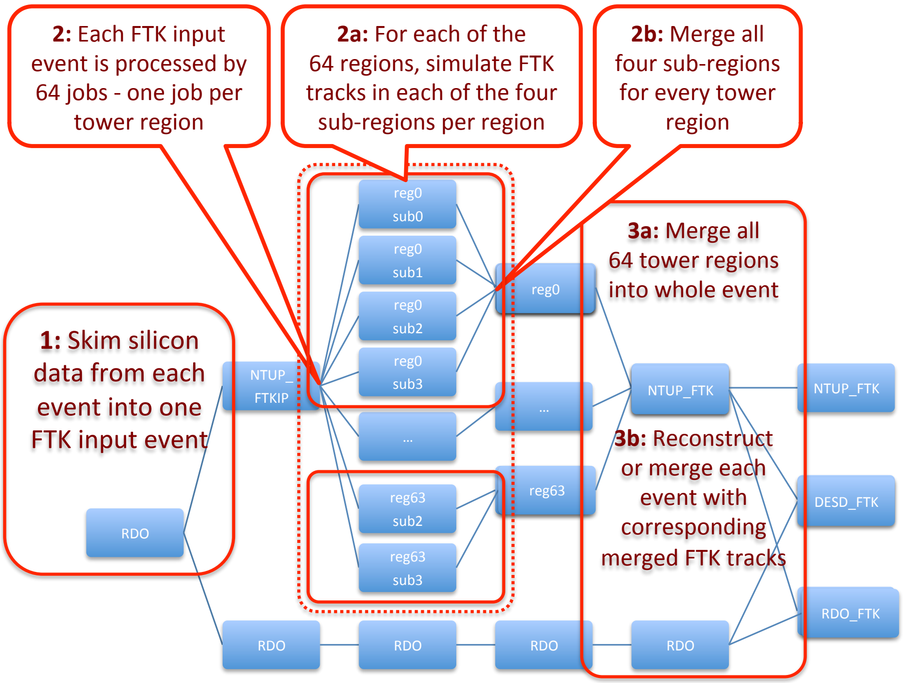

2: Each FTK input event is processed by 64 jobs - one job per tower region
2a: For each of the 64 regions, simulate FTK tracks in each of the four sub-regions per region
2b: Merge all four sub-regions for every tower region
3a: Merge all 64 tower regions into whole event
1: Skim silicon data from each event into one FTK input event
3b: Reconstruct or merge each event with corresponding merged FTK tracks
reg0 sub0
reg0 sub1
reg0 sub2
reg0 sub3
...
reg63 sub2
reg63 sub3
reg0
...
reg63
NTUP_FTKIP
NTUP_FTK
NTUP_FTK
DESD_FTK
RDO_FTK
RDO
RDO
RDO
RDO
RDO

Figure 6. The simulation of the FTK hardware.

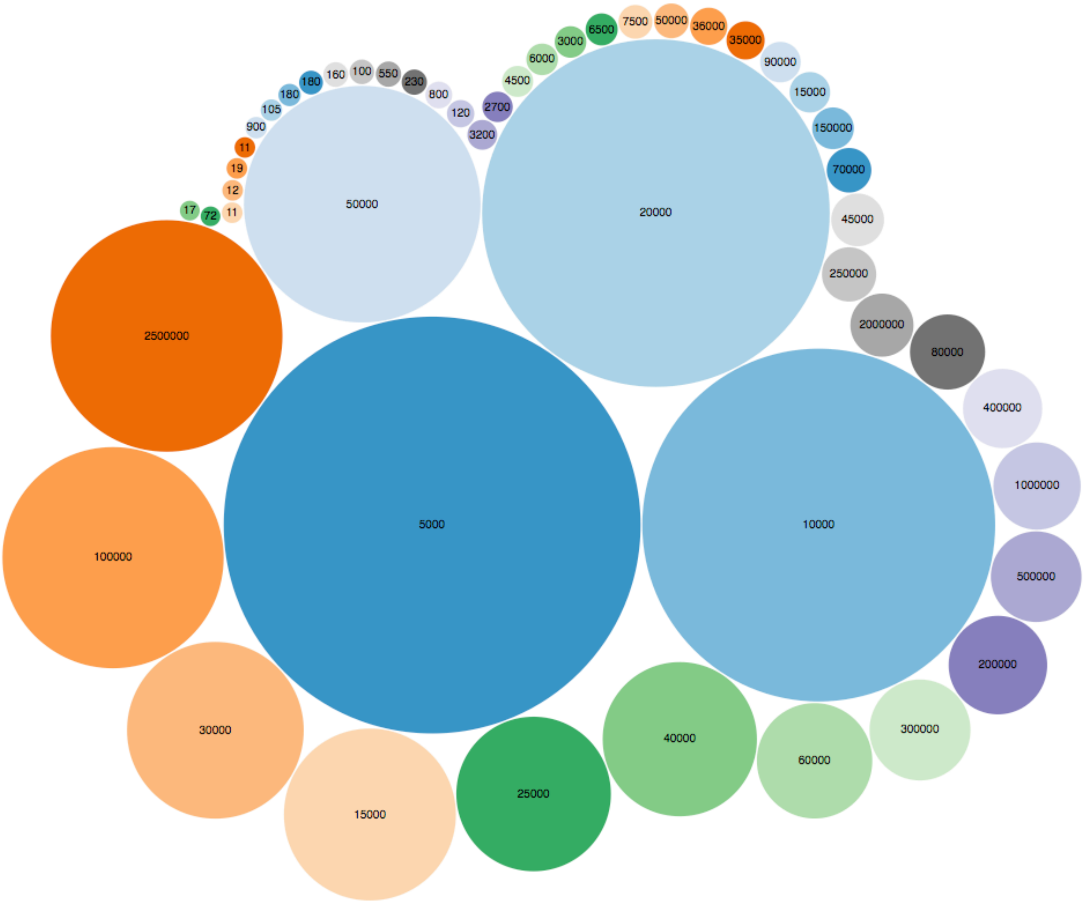
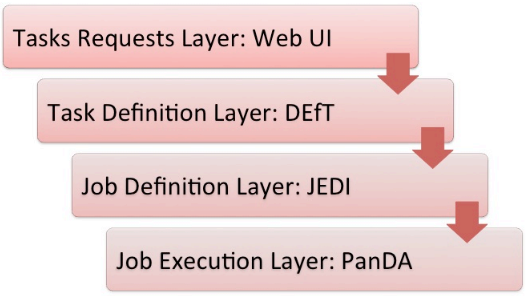


Figure 7. Big Data variety represented by event samples sizes from more than 22 000 different datasets produced during mc12 campaign. Bubble sizes are proportional to the number of the samples, labels correspond to the number of events in the sample.

reconstructed data in event summary data files or adds FTK tracks to the simulated events in the object-oriented representations of simulated detector readouts files.

Leveraging our Big Data processing techniques, four different sub-campaigns of the mc11 campaign implemented pileup conditions, detector conditions and geometries that were increasingly closer to those in real data. During the mc12 campaign, the majority of the events was simulated in the sub-campaign mc12b. Later, the mc12c sub-campaign implemented an improved detector geometry description. Figure 7 shows the variety of the simulated of event sample sizes for more than 22 000 different datasets produced during mc12 campaign. The goal of the mc14 campaign was to prepare for the 2015 data taking. The 8 TeV events were processed with improved and updated simulation, digitization and reconstruction software while using the same conditions as in the mc12 campaign. The 13 TeV campaign had the centre of mass energy expected for the 2015 data taking with estimated pileup and detector conditions. The mc14 campaign used the new ATLAS Integrated Simulation Framework [9], with multicore processing becoming the default for major simulated data processing steps: simulation, digitization and reconstruction.


Tasks Requests Layer: Web UI

Task Definition Layer: DEfT

Job Definition Layer: JEDI

Job Execution Layer: PanDA

Figure 8. Multilayer architecture of the ATLAS production system.

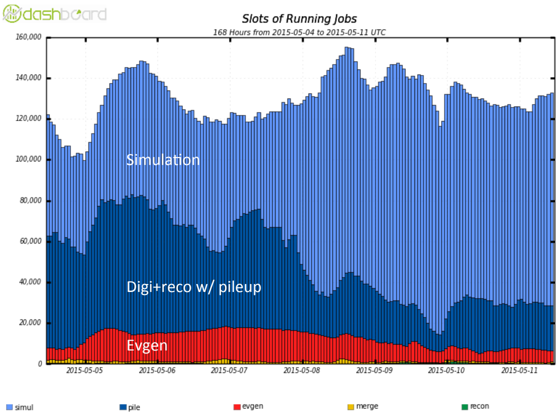


Figure 9. The number of jobs running in the ProdSys2.

3. Multilayer data processing system

The LHC shutdown provided an opportunity for upgrading the production system, making implementations more scalable, whilst retaining most valued core capabilities. To assure scalability, the production system was upgraded with extra layers. Avoiding inherent fragility of the monolithic systems, we separated the core concerns: the system logic layer is separated from the presentation layer, providing a familiar but improved interface for task requests.

The upgraded data processing system generates actual workflow tasks and their jobs are executed across more than a hundred distributed computing sites via PanDA – the ATLAS workload management system. Figure 8 shows that on top, the Task Request interface provides the presentation layer for users, while the lower Task Definition layer implements the core data processing logic that empowers production managers with template workflow definitions through the Database Engine for Tasks (DEfT) [10]. At the layer below, the Job Execution and Definition Interface (JEDI) [11] is integrated with PanDA to provide dynamic job definition tailored to the site's capabilities.

In the WLCG distributed computing environment, PanDA provides transparency of data and processing. As a result, the production system sees a unified computing facility that is used to run all data processing for the experiment, even though the sites are physically located all over the world. The production system supports a diverse range of workflows handling centrally ATLAS petascale data processing of the real and simulated data, including a mixture of both. ProdSys2 is designed to support all workflows supported by ProdSys1, and also many new workflows that would have been impossible or extremely difficult to manage in ProdSys1. The number of running jobs has approached 160k (Figure 9), with the rate of task processed by ProdSys2 exceeding that of ProdSys1 (Figure 10).

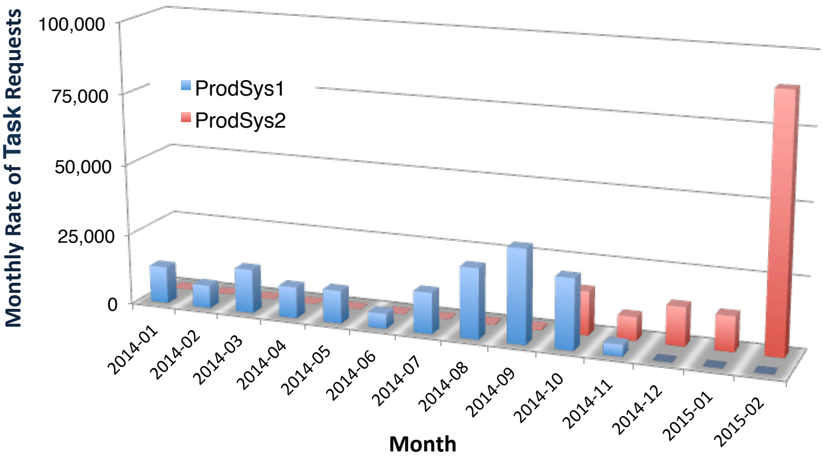


Figure 10. Monthly rate of production tasks.

4. Conclusions

During LHC data taking, the ATLAS production system unified a diverse range of workflows and special use cases including processing of both real and simulated data samples at large scales. The ATLAS production system fully satisfies the Big Data processing requirements of the ATLAS experiment through a unified approach for real data processing and simulations as well as the mixture of both. This technique has enabled the infrastructure to address a much wider range of physics analyses, with a higher level of precision, surpassing the most optimistic expectations [12]. In addition, detailed physics studies have established that the simulated data are of unprecedented quality compared to previous generations of experiments, describing the detector behaviour quite well in most analyses. The unified capabilities for real and simulated data processing have significantly enhanced ATLAS physics output, and motivated production of higher than foreseen simulated data volumes.

The LHC shutdown provided an opportunity for enhancing the production system, whilst retaining those core capabilities most valued by production managers. As the ATLAS experiment continues optimising the use of Grid computing resources for the LHC data taking, the next generation production system has been integrated with other ATLAS Distributed Computing layers – the ATLAS collaboration's dataset metadata interface and distributed data management system. Major workflows have been validated and in production for physics analysis and other ATLAS main activity areas - Trigger and Data Preparation.

Acknowledgments

We wish to thank all our colleagues who contributed to ATLAS Big Data processing activities. This work was funded in part by the U. S. Department of Energy, Office of Science, High Energy Physics and ASCR Contract No. DE-AC02-98CH10886 and Contract No. DE-AC02-06CH11357 and by the Russian Ministry of Science and Education under Contract №14.Z50.31.0024.

References

[1] The ATLAS Collaboration 2008 The ATLAS experiment at the CERN Large Hadron Collider. *J. Inst.* **3** S08003

[2] Bird I 2011 Computing for the Large Hadron Collider *Annu. Rev. Nucl. Part. S.* **61** 99

[3] Maeno T et al. 2012 Evolution of the ATLAS PanDA production and distributed analysis system *J. Phys.: Conf. Ser.* **396** 032071

[4] Stirling WJ, private communication

[5] The ATLAS Collaboration 2013 *Technical Design Report for the Phase-I Upgrade of the ATLAS TDAQ System,* LHCC Report CERN-LHCC-2013-018 (Geneva: CERN)

[6] Vaniachine AV for the ATLAS Collaboration 2011 ATLAS detector data processing on the Grid *IEEE Nuclear Science Symposium and Medical Imaging Conference* p 104

[7] The ATLAS Collaboration 2013 *ATLAS Fast TracKer (FTK) Technical Design Report,* LHCC Report CERN-LHCC-2013-007, ATLAS-TDR-021 (Geneva: CERN)

[8] Adelman J et al. 2014 *ATLAS FTK challenge: simulation of a billion-fold hardware parallelism* ATLAS Note ATL-DAQ-PROC-2014-030 (Geneva: CERN)

[9] Debenedetti C on behalf of the ATLAS Collaboration 2014 Concepts for fast large scale Monte Carlo production for the ATLAS experiment. *J. Phys.: Conf. Ser.* **513** 022006

[10] De K et al. 2014 Task management in the new ATLAS production system *J. Phys.: Conf. Ser.* **513** 032078

[11] Borodin M et al. 2014 *Multilevel Workflow System in the ATLAS Experiment* ATLAS Note ATL-SOFT-PROC-2014-005 (Geneva: CERN)

[12] Golubkov D et al. 2012 ATLAS Grid Data Processing: system evolution and scalability. *J. Phys.: Conf. Ser.* **396** 032049