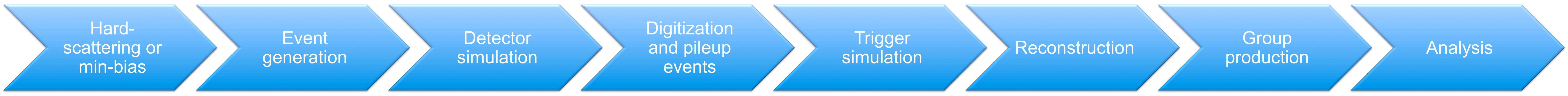
Multilevel Workflow System in the ATLAS Experiment

M. Borodin, J. Garcia, K. De, D. Golubkov, A. Klimentov, T. Maeno and A. Vaniachine for the ATLAS Collaboration

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. Each task performs the data processing as the transformation of input datasets into output datasets, 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 in workflows (Figure 1).

The LHC shutdown provided an opportunity to enhance the system architecture improving the performance and scalability [1]. The new bi-level workflow manager - ProdSys2 - generates actual workflow tasks, with their jobs executed across more than a hundred distributed computing sites by PanDA – the ATLAS job-level workload management system [2, 3] (Figure 2). The new system is being integrated with outer layers: at the top, the enhanced ATLAS Metadata Interface (AMI) [4] configures the data transformation parameters; at the bottom, the new distributed data management system Rucio [5] transfers datasets between the sites.



Hard-scattering or min-bias

Event generation

Detector simulation

Digitization and pileup events

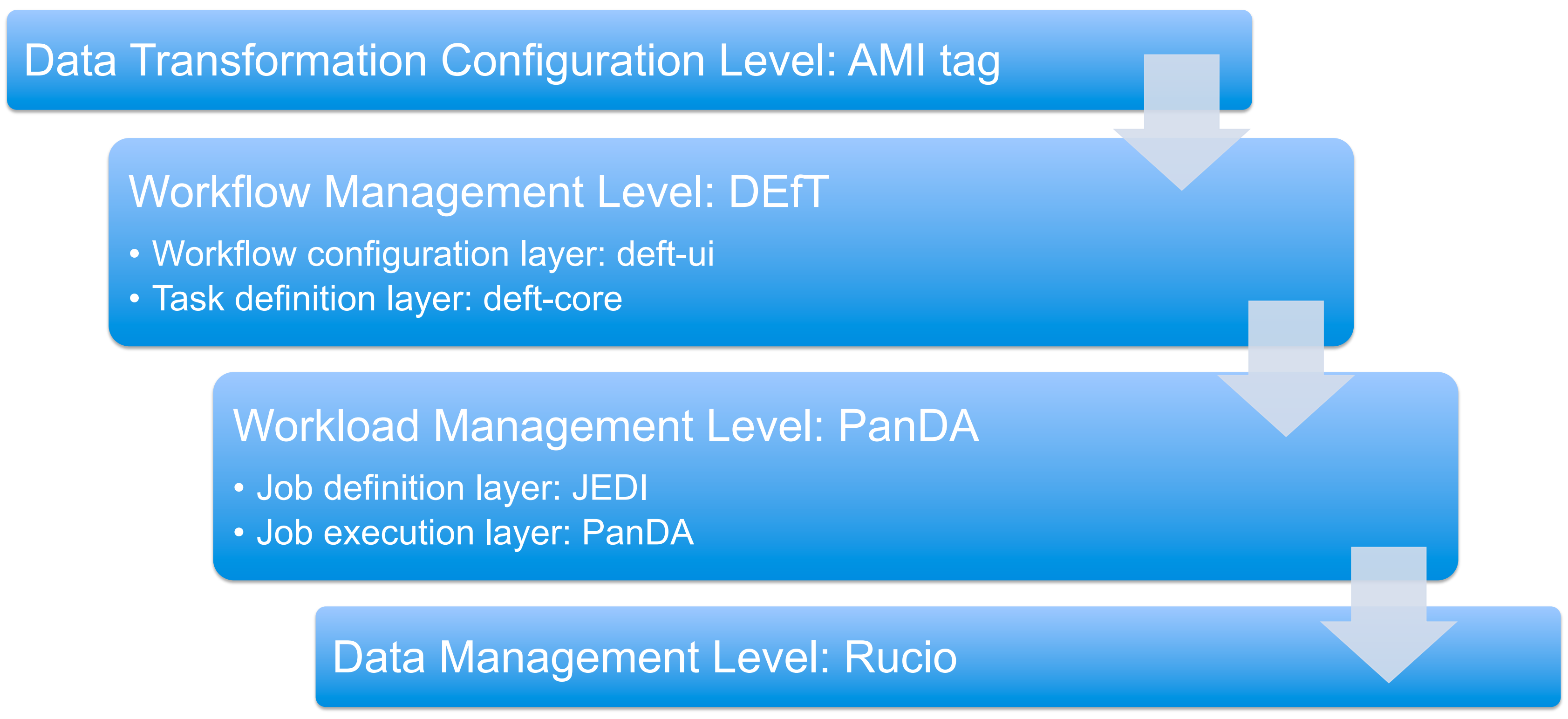
Trigger simulation

Reconstruction

Group production

Analysis

Figure 1: 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 ntuples for physics analysis, etc. Outputs are merged and/or filtered as necessary.



Data Transformation Configuration Level: AMI tag

Workflow Management Level: DEfT
- Workflow configuration layer: deft-ui
- Task definition layer: deft-core

Workload Management Level: PanDA
- Job definition layer: JEDI
- Job execution layer: PanDA

Data Management Level: Rucio

Figure 2: Multi-level architecture of the new ATLAS production system. On the upper level, the Database Engine for Tasks (DEfT) empowers production managers with templated workflow definitions [3]. 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.

Requirements: Figure 3 represents the scale and variety of requirements from physics groups, with the number of datasets dominated by datasets of SUSY grids. Figure 4 shows that the new system has to be flexible as the number of data transformations grows exponentially during LHC data taking and beyond.

Implementation: In the bi-level ProdSys2, 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.

Late binding: During task execution the dynamic job definition tailors the jobs based on the actual resources: disk space, CPU-time, memory, networks, etc. In contrast, the first production system employs a static definition of the jobs.

Analysis: The new system provides additional capabilities for ATLAS physicists.

Reprocessing: A starting point for physics analysis of LHC data is reconstruction. Following the prompt reconstruction, the ATLAS data are reprocessed on the Grid, which improves the quality of the reconstructed data for 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. The table below shows that failures are no longer a problem, as the fraction of CPU-time used for data recovery is small.

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*	3.1

* 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.

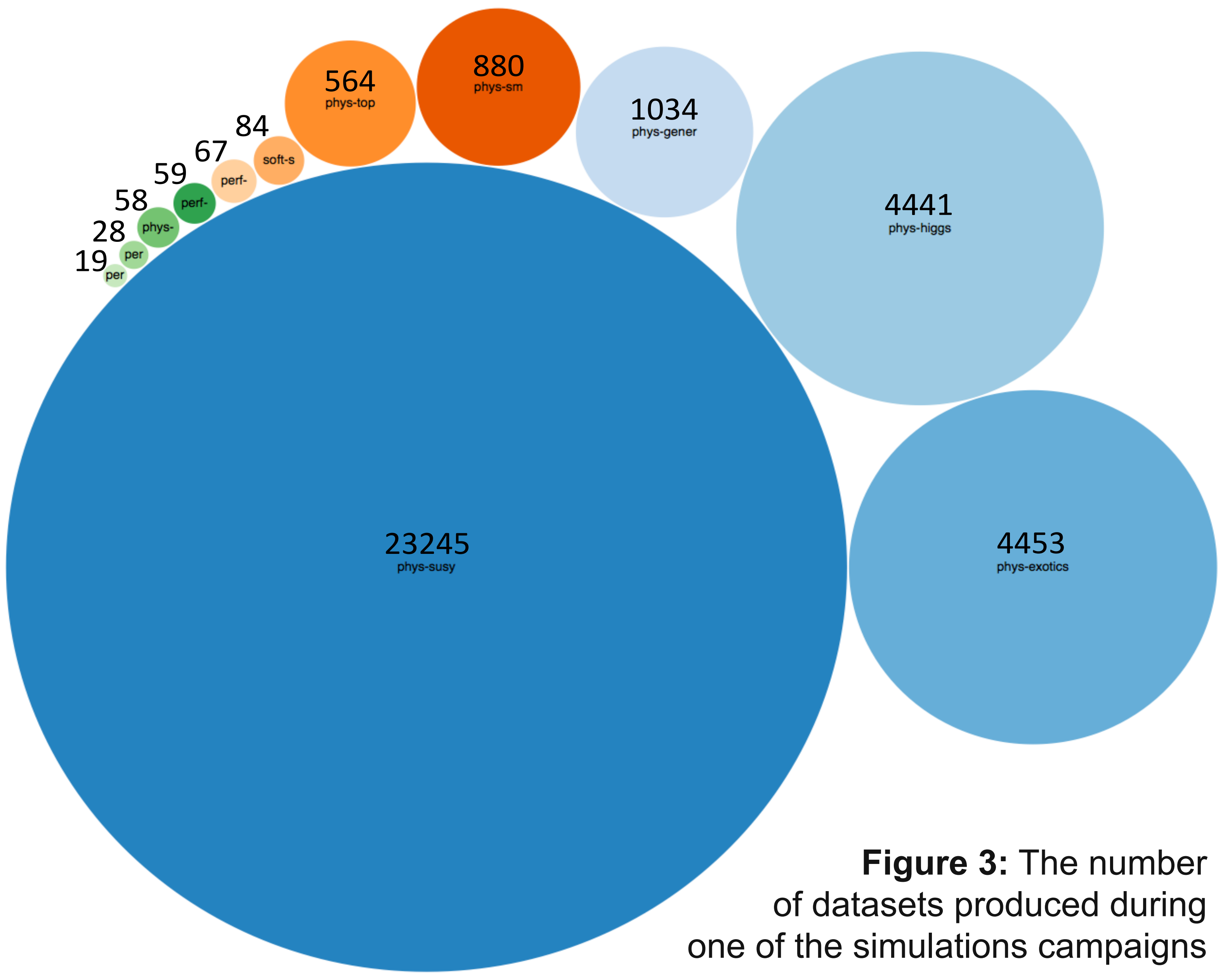


Figure 3: The number of datasets produced during one of the simulations campaigns

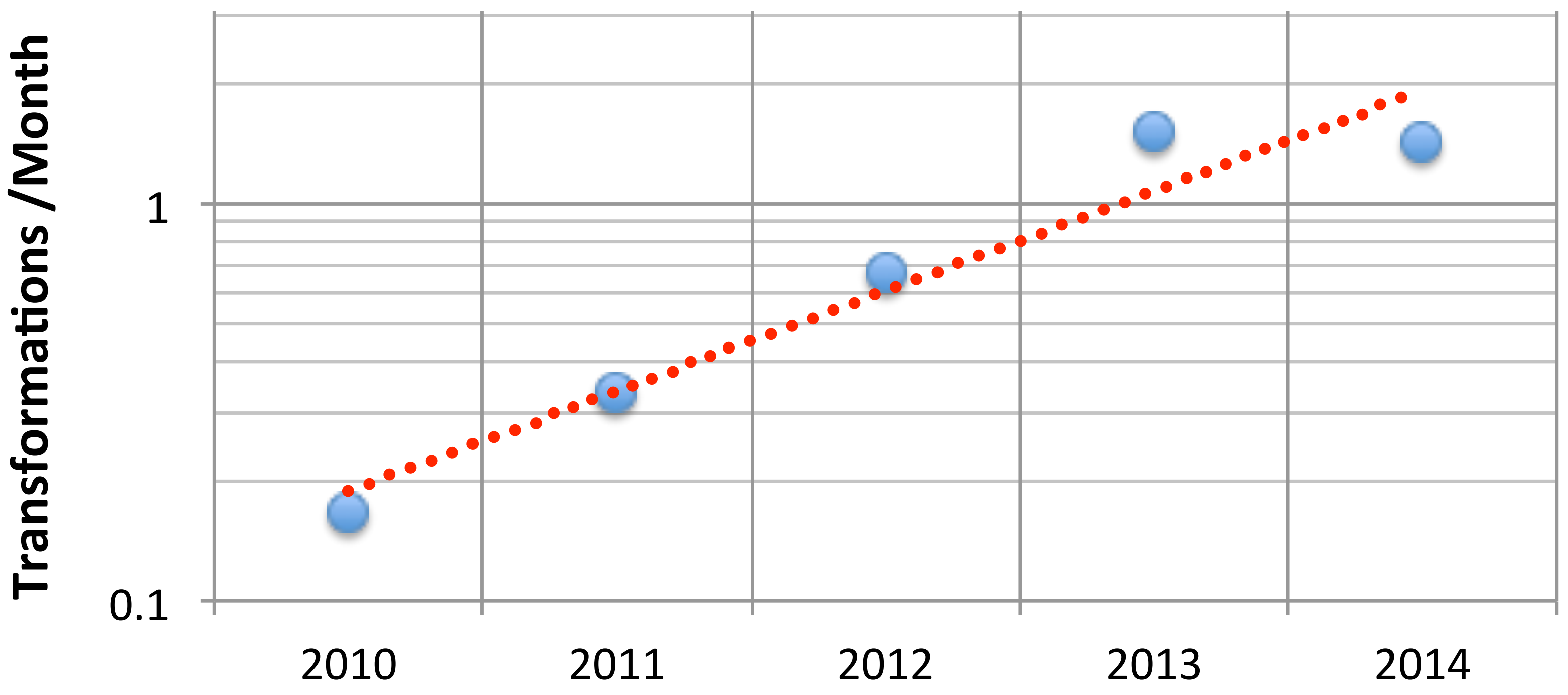


Figure 4: Continuous growth in the rate of new data transformations added to the system upon requests from the production managers.

Conclusions and next steps: The ATLAS production system fully satisfies the requirements of ATLAS data reprocessing, simulations, and production by physics groups. 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 in preparation for the LHC data taking in 2015, the next generation production system is ready for integration 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 main activity areas: Trigger, Data Preparation and Software & Computing.

References:

[1] D. Golubkov *et al.* “ATLAS Grid Data Processing: system evolution and scalability” 2012 *J. Phys.: Conf. Ser.* **396** 032049
[2] T. Maeno *et al.* “Evolution of the ATLAS PanDA workload management system for exascale computational science” 2014 *J. Phys.: Conf. Ser.* **513** 032062
[3] K. De *et al.* “Task Management in the New ATLAS Production System” 2014 *J. Phys.: Conf. Ser.* **513** 032078
[4] J. Fulachier *et al.* “Looking back on 10 years of the ATLAS Metadata Interface” 2014 *J. Phys.: Conf. Ser.* **513** 042019
[5] V. Garonne *et al.* “The ATLAS Distributed Data Management project: Past and Future” 2012 *J. Phys.: Conf. Ser.* **396** 032045

Argonne NATIONAL LABORATORY
BROOKHAVEN NATIONAL LABORATORY
IFIC INSTITUT DE FÍSICA CORPUSCULAR
ИФВЭ IHEP
UNIVERSITY OF TEXAS ARLINGTON