The ATLAS Data Management System Rucio

Supporting LHC Run-2 and beyond

Martin Barisits, Thomas Beermann, Vincent Garonne, Tomas Javurek, Mario Lassnig, Cedric Serfon
on behalf of the ATLAS Collaboration

Contact: *rucio-dev@cern.ch*

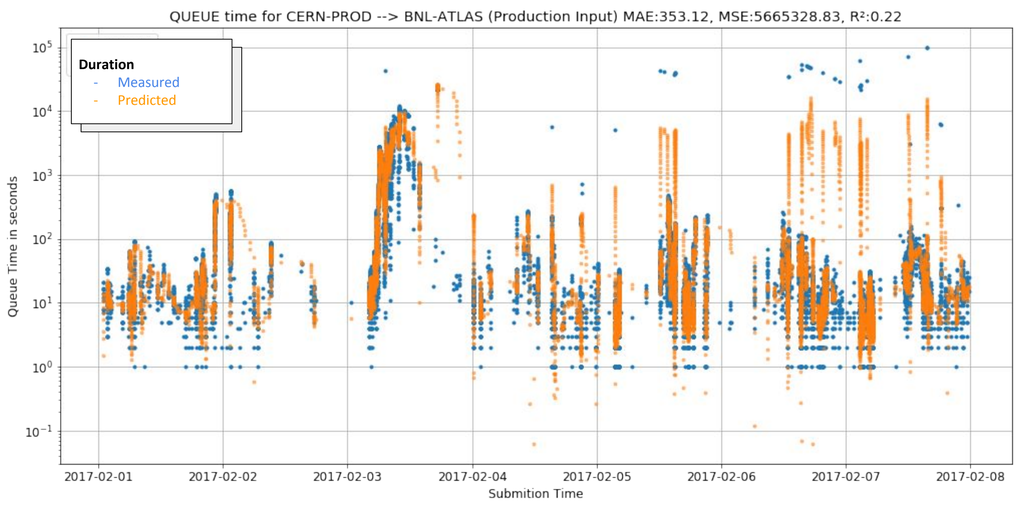
About Rucio

- Rucio is the Distributed Data Management system in charge of managing all ATLAS data on the grid.
- The main purpose of the system is to help the collaboration to store, manage and process LHC data in a heterogeneous distributed environment. Typical tasks are:
  - Transfer data to/from sites.
  - Delete data from sites.
  - Enforce the experiment computing model

The system manages **290 Petabytes** of physics data across more than **130 data centres** globally, with more than **830 million files**.

Machine learning for data management

- Data distribution can be seen as distance quantification problem. Given a use case, dataset, and location returns a sorted list of storage systems, based on distance.
- *Use Case 1: Estimate data transfers durations*
  - Use transfer duration model as distance quantifier
  - Full queueing + network time model
  - Holt-Winters Exponentially Weighted Moving Average
  - Currently evaluating LSTM recurrent networks

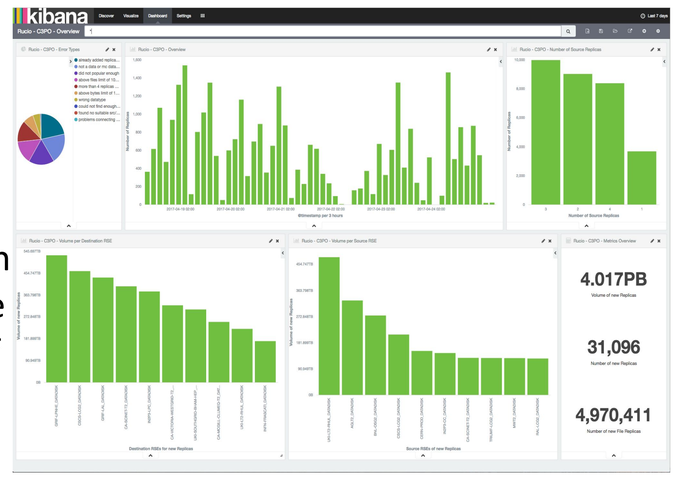


- *Use Case 2: Detecting transfer event anomalies*
  - Two approaches are currently ongoing
    - Estimation of transfer failure intervals on links
    - Throughput degradation on links
      - Google Summer of Code 2017 project
  - Compare predicted transfer duration against live system metrics to trigger alarms

System automations

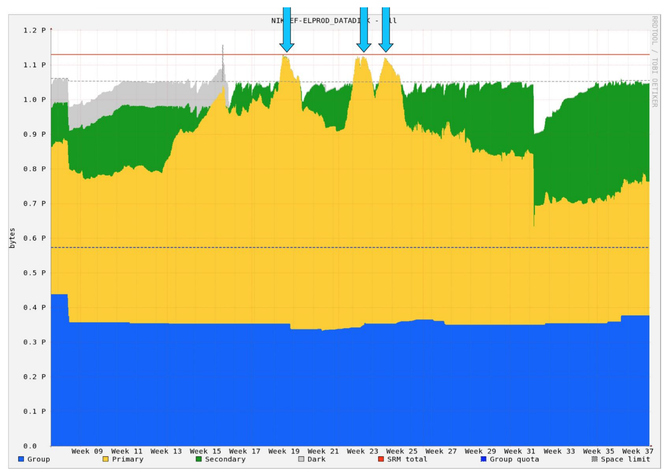
Dynamic creation of extra replicas for popular data:

- Daemon which continuously monitors incoming user analysis jobs and then decides if and where to make an extra copy of the input dataset
- Decision is based on
  - Past popularity of the input dataset
  - Already existing replicas
  - Free space on potential destination sites
  - Network speed between sites hosting an existing replica and potential destination site
- Running in pre-production mode for nearly a year now and already created replicas are evaluated to optimise the decision algorithm.
- Detailed monitoring available giving real-time information about the decision process, e.g., chosen source and destination sites, already available replicas and reasons if the algorithm decides not to make a new replica.



Rebalancing of data:

- Daemon which automatically and dynamically rebalances data to protect the system from overloads due to imbalances in computing slots and available capacity
- The plot shows the behaviour of the daemon for one RSE. If an imbalance is detected (blue arrows) primary data is moved to another site and thereby freeing space on the problematic site



Object stores

The vast majority of cloud storage in the market leverages an object-based storage architecture using the S3 protocol.

- Object-based S3 storage available for ATLAS:
  BNL (Amazon, Ceph), Lancaster (Amazon),
  RAL (Ceph), CERN (Ceph), MTW2 (Ceph)
- Use cases:
  - Log files management
  - The ATLAS Event Service
- Rucio allows the integration of non-posix namespaces with a flexible mapping of protocol+endpoint. Other authentication mechanisms than X509 usually based on access and secret keys for object-based storage.
- Two protocols have been implemented: s3:// and s3+rucio://
- s3:// requires to have the Access and Secret keys available, prefered way on central machines where we have access to them in a secure way, e.g., central deletion.
- s3+rucio relies on pre-signed URLs which gives temporary read/write access to the object.

Beyond ATLAS

- Originally developed for ATLAS, Rucio is now in use by other experiments like **Alpha Magnetic Spectrometer (AMS)** and **Xenon1t**
- Multi-experiment features were developed, for example:
  - Support for other database backends (e.g., MariaDB, PostgreSQL)
  - Integration with open-source monitoring frameworks (e.g., Elasticsearch)
  - New deployment options (e.g., docker images)
  - Other potential collaborations are under investigation, e.g., COMPASS