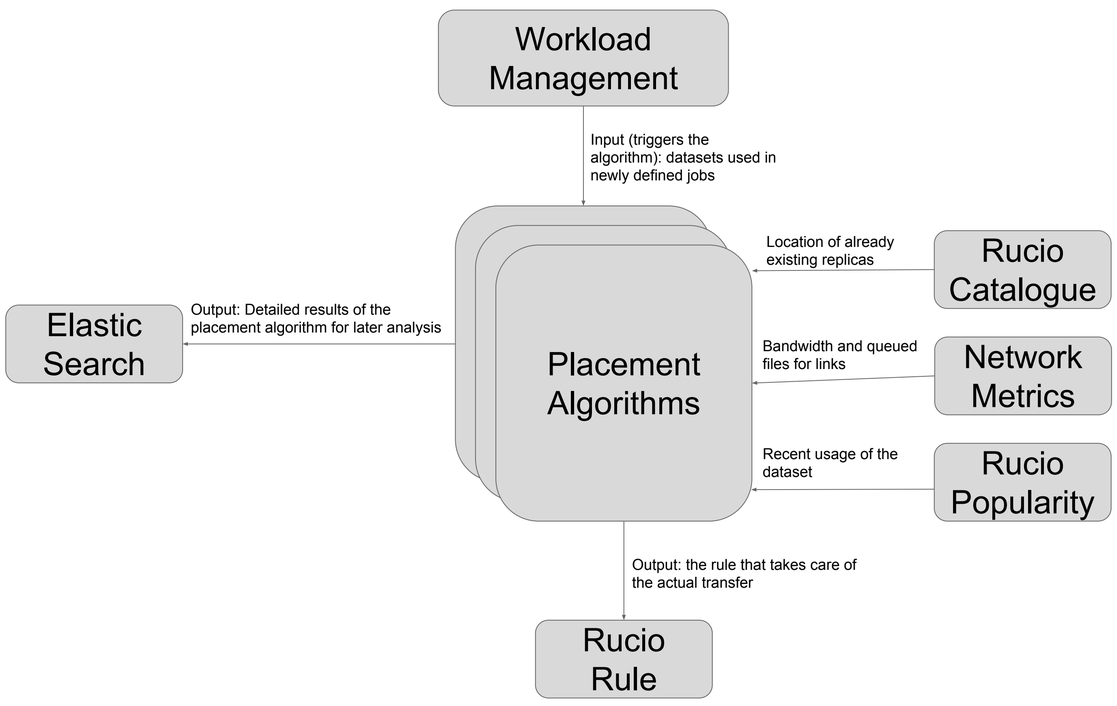
C3PO - A Dynamic Data Placement Agent for ATLAS Distributed Data Management

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

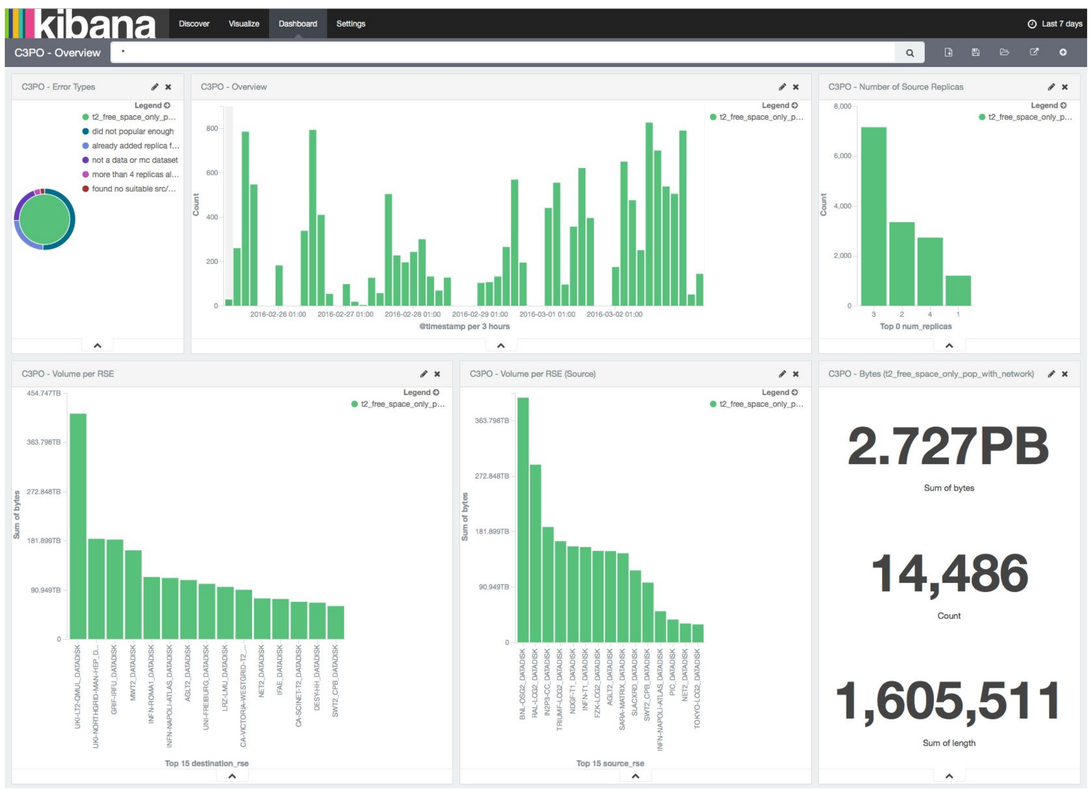
About ATLAS DDM

- The Distributed Data Management project is charged with managing all ATLAS data on the grid.
- All for the purpose of helping the collaboration store, manage and process LHC data in a heterogeneous distributed environment, like:
  - Transfer data to/from sites.
  - Delete data from sites.
- On centrally managed endpoints the oldest unused data is deleted first (Least Recent Used Mechanism).
- Together with this new tool presented in this poster popular data can quickly be replicated into space that has been freed up by deleting unpopular data.
- The tool is designed to be modular and can be easily extend and this poster shows how it has been tested with one particular algorithm.

Architecture



Monitoring

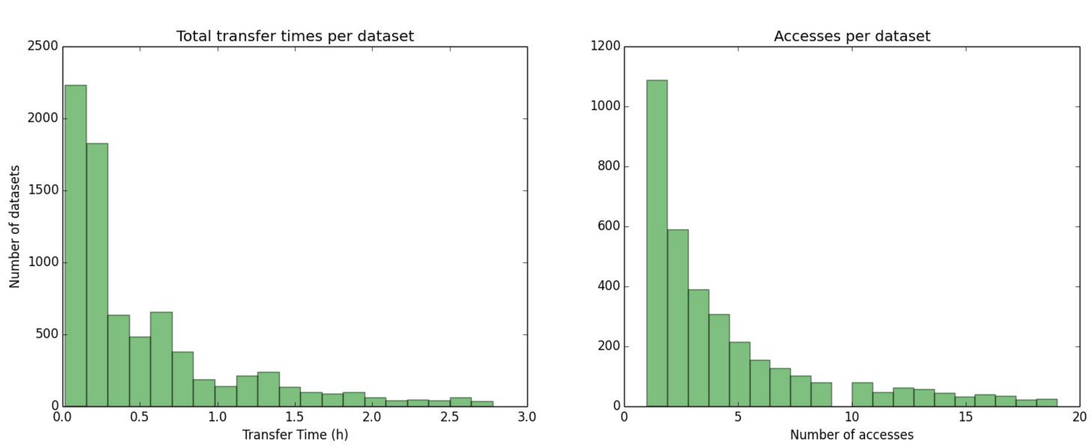


- All results of the algorithm are written to ElasticSearch.
- With this information a Kibana dashboard was built giving an overview like the volume and number of files that have been created (example above).
- In more details the top source and destination endpoints are shown helping a lot in debugging and improving the algorithm

Algorithm

1. The daemon constantly scans incoming user jobs and collects the input datasets, the placement algorithm runs for every dataset containing official Monte Carlo or detector data.
2. First the algorithm checks if there has already been a replica created in the past 24 hours, if yes it will exit.
3. It then checks how many replicas already exist and exits if a threshold is met.
4. The algorithm has a configurable threshold of files and bytes that it can create per hour and day and per destination site, if the datasets would exceed this it will exit.
5. It then checks the popularity of the dataset in the last 7 days, if it hasn't been popular enough it will again exit.
6. It will then continue to check links between sites having an existing replica and possible destination sites
7. The sites are ranked based on free space, bandwidth and queued files and they are down-ranked if a replica of a previous job has been recently created there.
8. If a suitable site has been found the algorithm submits a replication request to Rucio, which will then take care of the transfer.
9. In a last step detailed information about this decision is written to ElasticSearch for further analysis.

First results



- The results shown here are for a period in August 2016.
- During that time 7500 new replicas and 1.8 PB of data have been replicated.
- Of those newly created replicas 60% have then been used by the WMS.
- The left plot shows that most of the data have been transferred in less than an hour making it quickly available for the jobs to use it.
- The right plot shows that more than half of the accessed data have been accessed more than once.

Outlook and Future Work

- This new tool is still in an early development stage and there are multiple points that can be improved:
  a. Better destination selection by taking into account the available computing resources.
  b. Notification of the WMS that a new replica has been created so that it can trigger a rescheduling.
  c. Better selection of datasets by using machine learning techniques to find popular datasets.
  d. Better selection of source replicas by using better network metrics (predicted time-to-complete).