The ATLAS Data Carousel Project

M.Barisits, M.Borodin, A.Di Girolamo, D.Golubkov, W.Guan, J.Elmsheuser, E.Karavakis, A.Klimentov, T.Korchuganova, M.Lassnig, F-H Lin, T.Maeno, S.Padolski, D.South, and X.Zhao

25th International Conference on Computing in High-Energy and Nuclear Physics

May 17-21, 2021

BROOKHAVEN

Data Carousel

- Data Carousel: on-demand reading from tape without pre-staging
- Uses a rolling disk buffer whose size can be tuned to suit available resources and production requirements
- Key to success: rate at which data can be staged to disk at the Tier-0 and Tier-1 sites
- Technique can eventually be used for any experiment

ATLAS Data Carousel Project Phases

- Phase I : Tape Sites Evaluation (Y2018)
  - completed
  - Conduct tape staging tests, understand tape system performance at sites and define primary metrics
- Phase II : ProdSys2/Rucio/Facilities integration (Y2019-2020) *CHEP2019 talk*
  - completed
  - Address issues found in Phase I
  - Deeper integration between workflow, workload and data management systems (ProdSys2/PanDA/Rucio), plus facilities
  - Identify missing software components
- completed
- Phase III : Run production, at scale, for selected workflows (Y2020) *This talk*
- Phase IV : Use data carousel for many workflows in parallel, respect computing share per workflow. Run Data Carousel jointly for more than one experiment (Y2021)

Now we are at the middle of Phase IV (we are increasing the number of workflows running in Data Carousel mode) ***The initial goal is reached : to have data carousel in full production for LHC Run3***

ATLAS Run2 Data Reprocessing in Data Carousel Mode

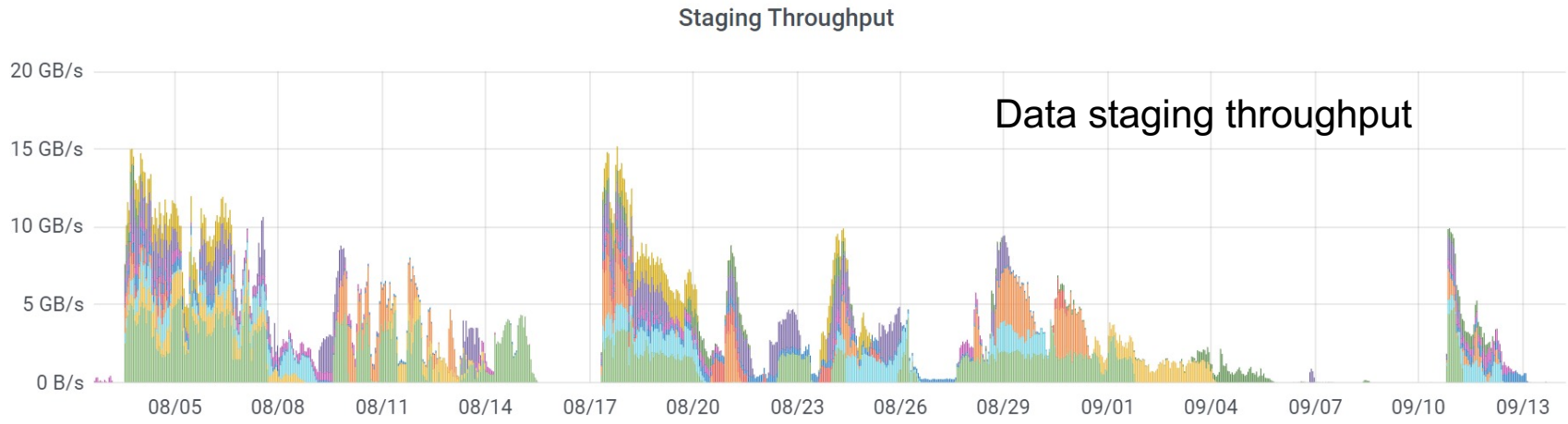
- Ultimate goal to demonstrate Data Carousel for bulk production and respect computing shares and disk buffer size
- Data have been processed in reverse order (year 2018 first). The total data volume was 18.5 PB

LHC Run2 2015-2018

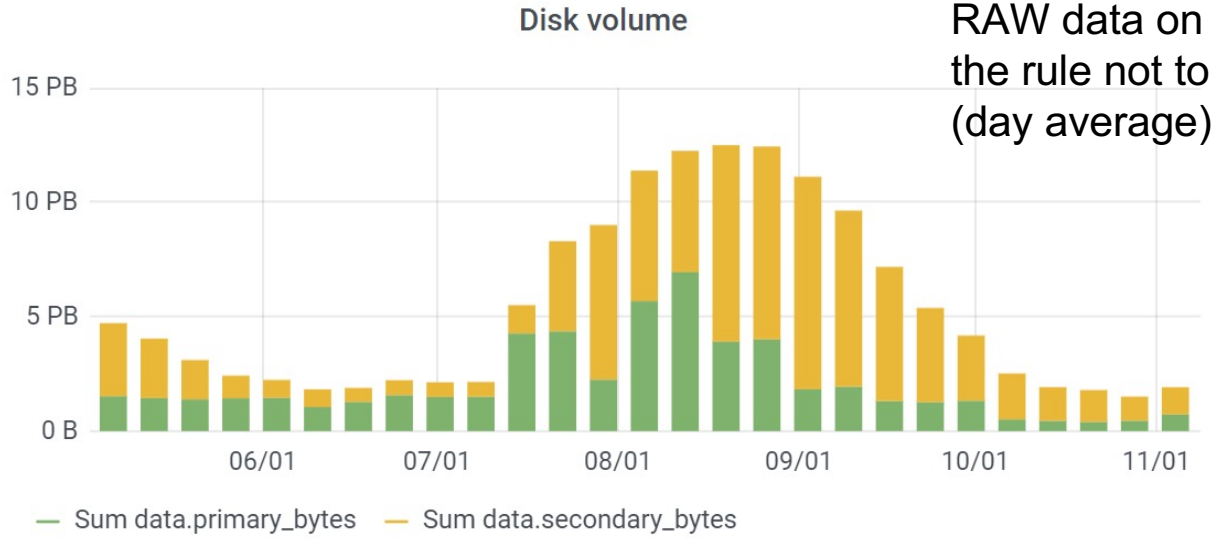
ATLAS Run2 Data Reprocessing in Data Carousel Mode

- Fine tuning before reprocessing was started
  - Tier-1, CERN, CTA, dCache teams participation in global monitoring
  - Tier-1s and CERN data staging profiles were developed and stored in the Information System (CRIC). Staging profiles were used by the ATLAS Production System
    - The Production System doesn’t send new requests to Rucio to stage a new data chunk until the previous one has reached a predefined level, usually 50+%.
  - Reprocessing shares have been defined by Physics Coordination and respected

ATLAS Run2 Data Reprocessing in Data Carousel Mode



Data staging throughput



RAW data on disks
the rule not to exceed 3 PB
(day average) was respected

BROOKHAVEN

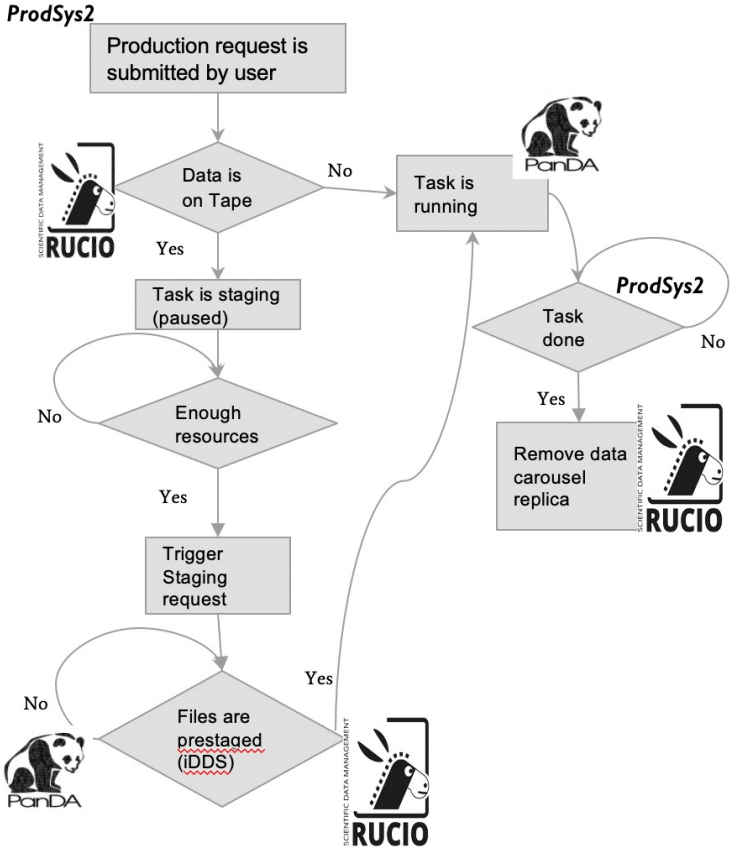
ATLAS Run2 Data Reprocessing in Data Carousel Mode

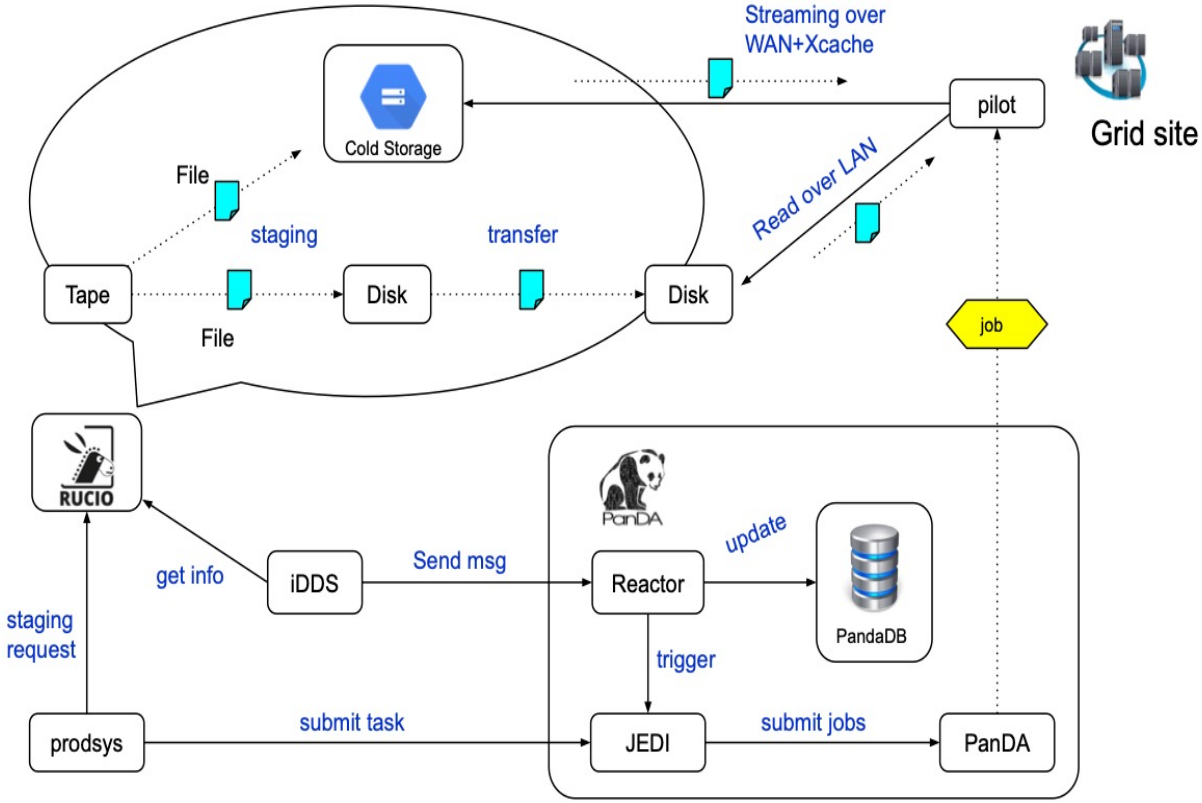
Joint sites, dCache, CTA and ATLAS effort to improve I/O performance

Sites	2018 Phase I Test (MB/s)	2020 Reprocessing (MB/s)
CERN (CTA Test)	2000	4300
BNL	866	3400
FZK	300	1600
INFN	300	1100
PIC	380	540
TRIUMF	1000	1600
CC-IN2P3	3000	3000
SARA-NIKHEF	640	1100
RAL	2000	2000
NDGF	500	600

Table 1: Stable Rucio tape throughput for the ATLAS Tier-1 sites and CERN, measured from the 2020 reprocessing campaign, with comparison to the Phase I results.

Data Carousel workflow and New distributed software component : intelligent Data Delivery Service (iDDS)





See T.Maeno et al iDDS talk

Summary and Data Carousel Today

- We successfully and quickly passed the R&D project phases involving ATLAS, FTS, dCache, CTA and the WLCG centers.
- During full Run2 data reprocessing, i.e., 18.5 PB of RAW data, ATLAS demonstrated the real Data Carousel mode in action, in a production environment with many other concurrent activities such as data writing, data rebalancing, or data consolidation between ATLAS Grid sites.
- Deep integration and communication protocols between data and workflow management systems were defined and implemented. We have evaluated the optimal file size to have more efficient tape I/O and, based on this, the file size will be increased for data produced by prompt reprocessing, i.e.,Tier-0 data processing and by the Production System.
- The first joint ATLAS-CMS test was conducted in February-March 2021 at three Tier-1s
- End-user analysis in Data Carousel mode with data staging from tape to be evaluated
- The major campaigns requesting data from tape will run in Data Carousel mode in Run3
- We continue to improve tape recall efficiency and grow tape capacity towards the needs of the HL-LHC

Acknowledgments

- *We thank our ATLAS Distributed Computing colleagues, ATLAS sites, Tier-1 ATLAS centers, CERN Tier-0 operations, CTA, and dCache teams. The work at Plekhanov University and ISP RAS is funded by the Russian Science Foundation grant (project No.19-71-30008).The work at Brookhaven National Laboratory is funded in part by the U.S. Department of Energy, Office of Science, High Energy Physics and Advanced Scientific Computing contracts. The work at University of Wisconsin-Madison is funded by the National Science Foundation under Grant No. 1836650*