The ATLAS Data Carousel Project Status and Plans

M.Barisits, M.Borodin, A.DiGirolamo, J.Elmsheuser, A.Klimentov, T.Korchuganova, M.Lassnig, T.Maeno, S.Padolski, R.Walker and X.Zhao

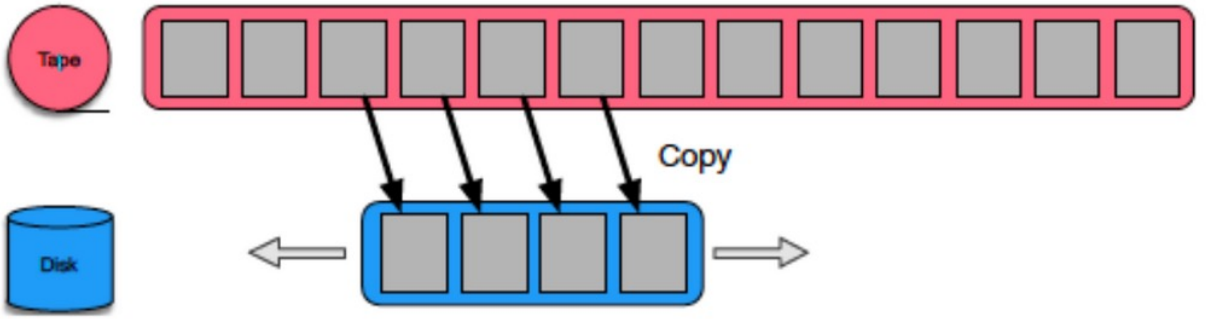
20th International Workshop on Advanced Computing and Analysis Techniques in Physics Research

December 1, 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 Grid sites
- Technique can eventually be used for any experiment
  - Two tape challenges during 2021 to address I/O tape performance at WLCG Tier-1s (4 LHC VOs)
- In ATLAS production today for data reprocessing, derivation and Monte-Carlo production



BROOKHAVEN

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) — 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
- Phase III : Run production, at scale, for selected workflows (Y2020) — completed
- Phase IV :
  - Use data carousel for many workflows in parallel
  - Respect computing share per workflow.
  - Run Data Carousel jointly for more than one experiment
  - Address ”smart writing” challenge

Now we are at the middle of Phase IV

(we increased the number of workflows running in Data Carousel mode during 2021)

The initial goal is reached : to have data carousel in full production for LHC Run3

ATLAS Run2 Data Reprocessing in Data Carousel Mode

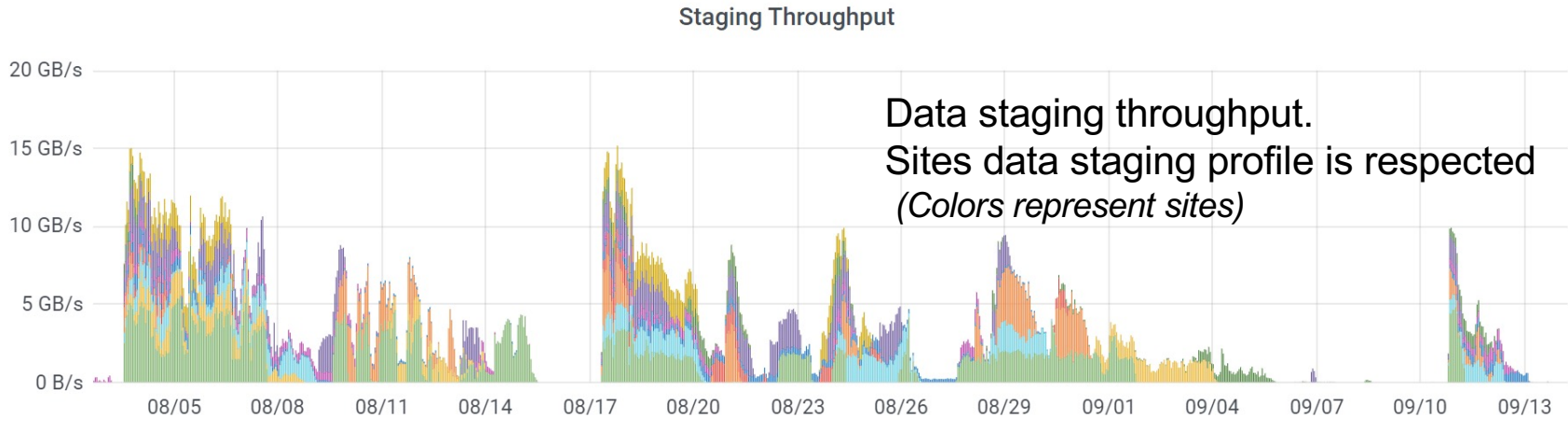
- Ultimate goal to demonstrate Data Carousel for bulk production and respect computing shares and disk buffer size
- Data have been (re)processed in reverse order (year 2018 first). The total data volume was 18.5 PB

LHC Run2 2015-2018

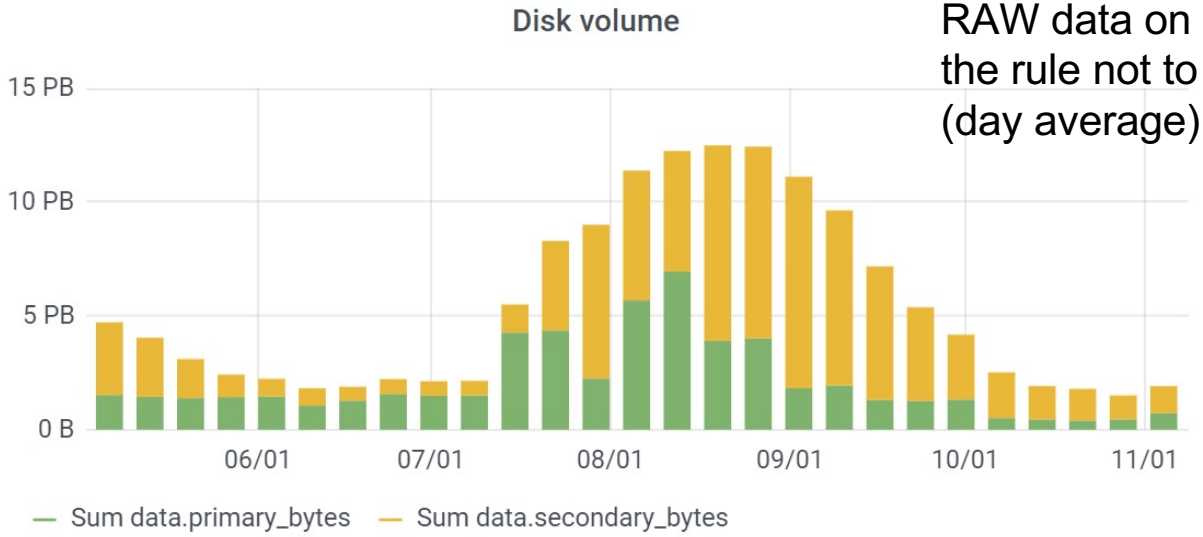
ATLAS Run2 Data Reprocessing in Data Carousel Mode

- 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 for all Tier-1s
    - *Site staging profile : The Production System doesn't send new requests to Data Management System (Rucio) to stage a new data chunk until the previous one has reached a predefined level, usually 50-75%.*
  - Reprocessing shares have been defined by Physics Coordination and respected

ATLAS Run2 Data Reprocessing in Data Carousel Mode



Data staging throughput.
Sites data staging profile is respected
(Colors represent sites)



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

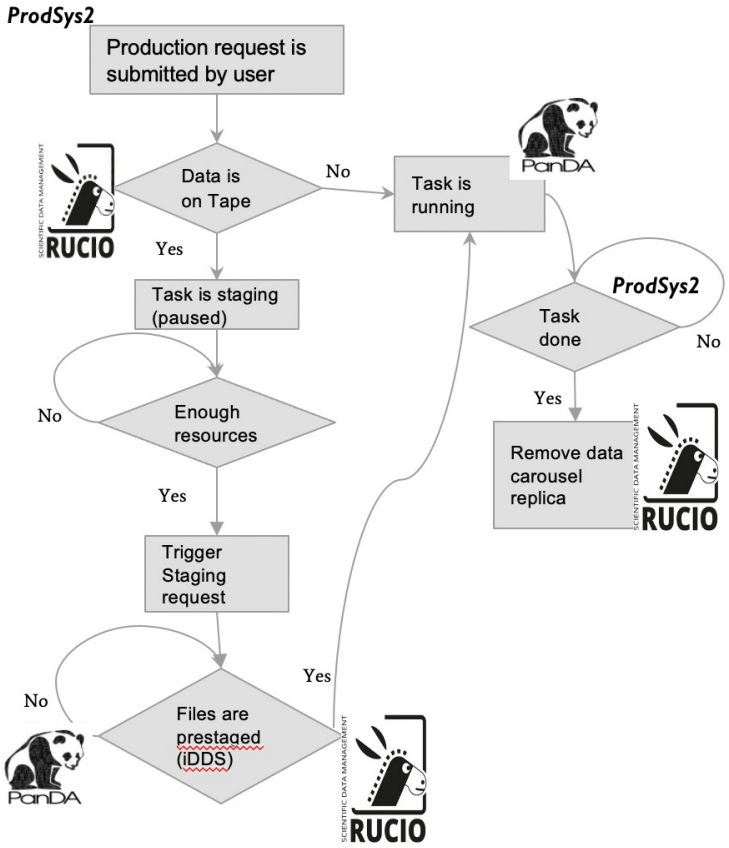
Primary – persistent data on disks
Secondary – transient (cached) data

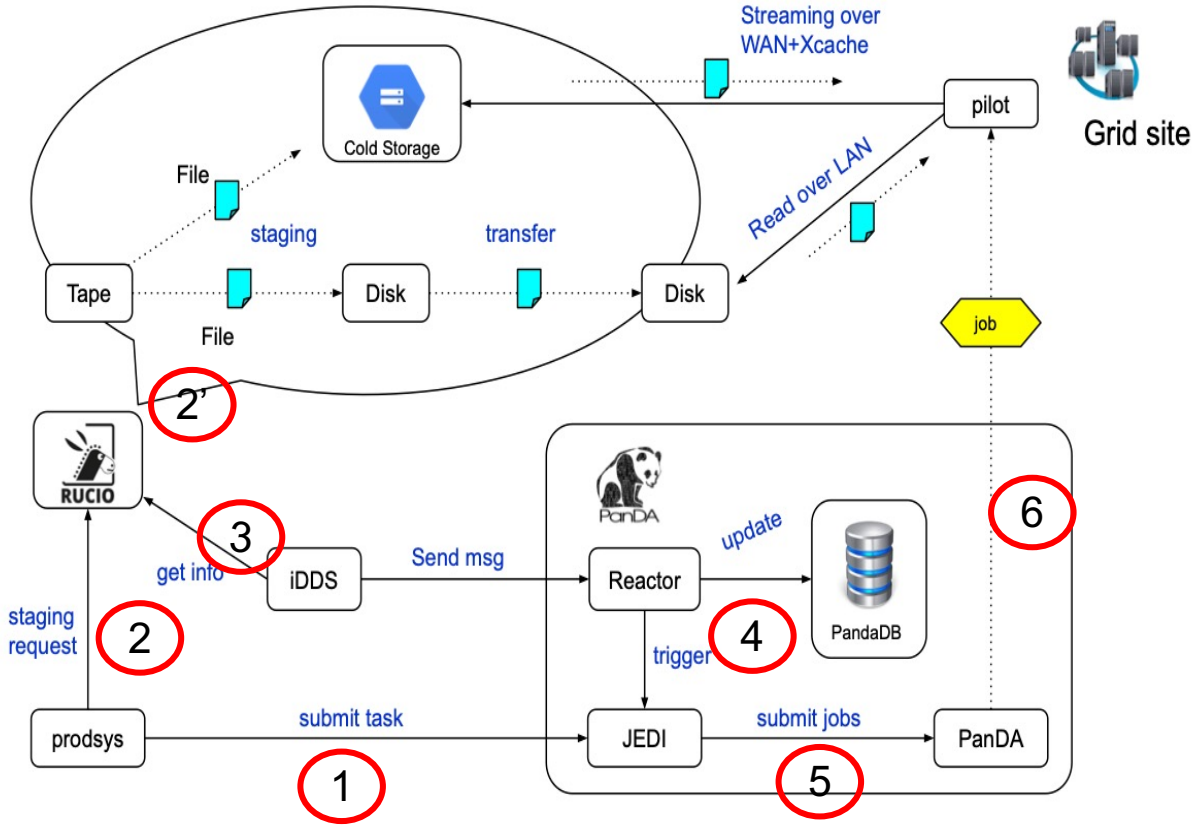
BROOKHAVEN

Data Carousel and ATLAS Production Workflows

- ATLAS runs Monte-Carlo and derivation production, data reprocessing in Data Carousel mode
  - overall staging throughput is improved at CERN and Tier-1s
  - several software (algorithmic) improvements to address staging tails and optimize data placement and processing tasks brokerage
    - iDDS – intelligent data delivery service helps to improve an overall data staging / processing performance

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



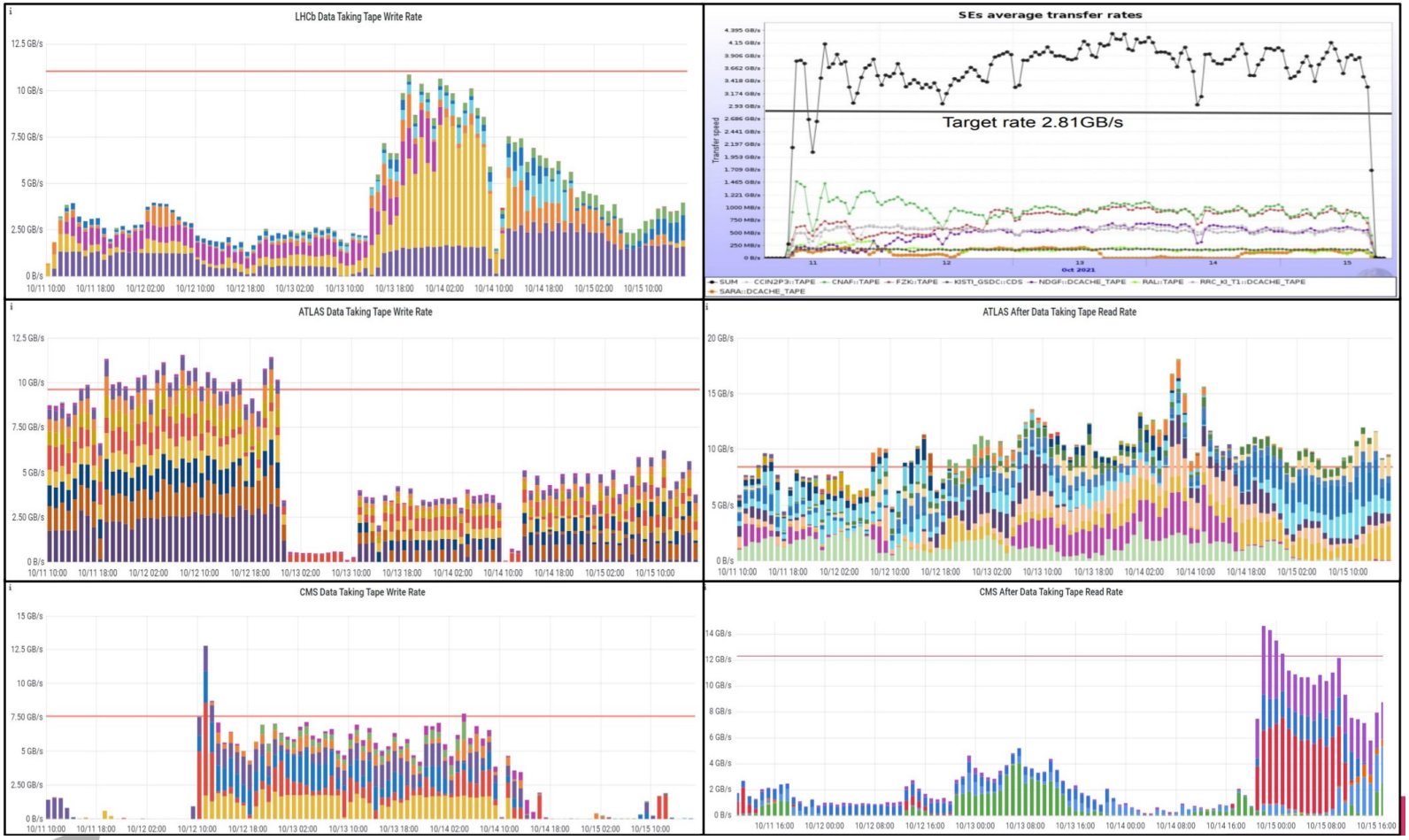


T.Maeno et al iDDS talk

Joint Tape Challenge Tests (LHC experiments)

- The first round was just for ATLAS and CMS (February-March 2021)
  - 300 TB of data
  - Distributed Data Management System (Rucio) for data transfer in both experiments
  - Two Tier-1s (PIC and KIT) - simultaneous data staging requests for both VOs
  - Constant network monitoring (FTS dashboard)
- All four LHC experiments have participated in the second round (October 2021)
  - Tape-driven workflow exercised by all experiments on all T1s at the LHC Run3 scale, *first of its kind*
    - *Two modes of operations : Data Taking Mode and After Data Taking Mode*
  - Tape I/O was monitored centrally
  - Common approach used in tape staging by ATLAS and CMS experiments
    - Common topics on smart writing between LHC experiments
    - File size vs I/O performance. VOs using bigger files generally obtained better tape I/O performance
      - ~10GB/file sizes are optimal for today's tape technologies
  - Central tape I/O monitoring across VOs is a crucial for success
  - more discussions on smart writing and bulk staging strategies between VOs and sites
  - ATLAS proposed and implemented sites staging profiles
- Next joint test in Q1 2022

Joint Tape Challenge Tests (LHC experiments)



October 2021 Tape Challenge. Tape Throughputs of all Experiments (colors represent Tier-1s)

Summary and Data Carousel Today

- We successfully and quickly passed the R&D project phases involving LHC experiments, 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. Major workflows (Monte-Carlo production, data reprocessing and derivation analysis objects production) are ran in Data Carousel mode.
- 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 LHC experiments tests were conducted in 2021 for all WLCG Tier-1s
- The major ATLAS 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. Files grouping on tapes is important to get the best data staging performance

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.*