Updates to the ATLAS Data Carousel Project

Mikhail Borodin[1], *David* Cameron[2], *Alexei* Klimentov[3], *Tatiana* Korchuganova[456], *Mario* Lassnig[7], *Tadashi* Maeno[3], *Haykuhi* Musheghyan[8], *David* South[9], and *Xin* Zhao[3,*] on behalf of the ATLAS Computing Activity

[1]The University of Iowa, USA
[2]University of Oslo, Oslo, Norway
[3]Brookhaven National Laboratory, Upton, NY, USA
[4]Institute of System Programming, Russian Academy of Science, Moscow, Russia
[5]Plekhanov Economy University, Moscow, Russia
[6]University Andres Bello, Santiago, Chile
[7]European Organization for Nuclear Research (CERN), Geneva, Switzerland
[8]Karlsruhe Institute of Technology, Karlsruhe, Germany
[9]Deutsches Elektronen-Synchrotron, Hamburg, Germany

Abstract. The High Luminosity upgrade to the LHC (HL-LHC) is expected to deliver scientific data at the multi-exabyte scale. In order to address this unprecedented data storage challenge, the ATLAS experiment launched the Data Carousel project in 2018. Data Carousel is a tape-driven workflow whereby bulk production campaigns with input data resident on tape are executed by staging and promptly processing a sliding window to disk buffer such that only a small fraction of inputs are pinned on disk at any one time. Data Carousel is now in production for ATLAS in Run3. In this paper, we provide updates on recent Data Carousel R&D projects, including data-on-demand and tape smart writing. Data-on-demand removes from disk data that has not been accessed for a predefined period, when users request them, they will be either staged from tape or recreated by following the original production steps. Tape smart writing employs intelligent algorithms for file placement on tape in order to retrieve data back more efficiently, which is our long term strategy to achieve optimal tape usage in Data Carousel.

ATL-SOFT-PROC-2023-007
15 August 2023

1 ATLAS Data Carousel in production

The High Luminosity upgrade to the LHC [1] will deliver an unprecedented volume of scientific data at the exabyte scale. To tackle this data storage challenge, the ATLAS experiment [2] started the Data Carousel R&D project [3] since the fall of 2018. Data Carousel is a tape-driven workflow that allows jobs to get input data directly from tape by orchestrating the workflow management systems ProdSys2 [4] and PanDA [5], the distributed data management (DDM) system Rucio [6], and the tape services.

*e-mail: xzhao@bnl.gov

Since 2020, Data Carousel has been utilized in production for various ATLAS managed campaigns, including RAW data reprocessing, derivation, and Monte Carlo simulation. Figure 1 shows the overall staging throughput from ATLAS Tier-0 and Tier-1 sites during the first month of 2023, reaching a peak rate of 20~25 GB/s. Over the course of 2022, ATLAS staged 64PB of unique data from tape, and if counting for multiple stages of the same files, the total volume amounted to approximately 120PB. This led to significant disk space savings, such as the reduction of Analysis Object Data (AOD) stored on disk by over half.

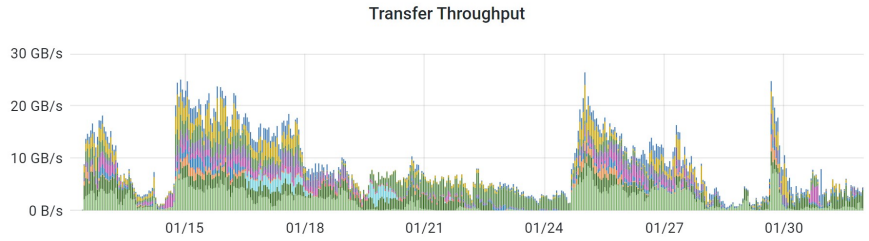

Figure 1: Total staging throughput (GB/s) from Tier-0/Tier-1 sites during Junuary 2023. Different color represents different Tier-0/Tier-1 site.

The ongoing Data Carousel R&D project is primarily focused on two activities. The first one, known as "data-on-demand", aims to explore new opportunities for saving disk space by incorporating new data types into the Data Carousel workflow. The second activity, called "tape smart writing", focuses on researching and developing new mechanisms for optimal utilization of tape resources, to enhance the efficiency and performance of tape operations within the Data Carousel workflow. Both of these activities are included in the list of the ATLAS HL-LHC demonstrator projects, aimed at the preparation of the HL-LHC Technical Design Report (TDR).

2 Data-on-demand

The current practice in ATLAS involves storing all Derived AOD on disk, which are inputs to user analysis jobs. Currently the total volume of DAOD is approximately 120PB. However many of the DAOD datasets remain unused. Following the ATLAS lifetime model, the ATLAS Distributed Data Management (DDM) team periodically identifies and deletes old and unused data based on predefined policies. For instance, if a DAOD dataset remains untouched for 6 months, it will be marked for deletion. Exceptions to the lifetime model can be granted through an application process, allowing datasets to be retained for extended periods. Nevertheless, further investigation reveals that only about 20% of the DAOD datasets in the exception list are accessed within the next 12 months. Additionally, the application and approval of lifetime model exceptions list is a very labor-intensive procedure for both the DDM operations team and physics user groups.

To address these issues, ATLAS is exploring the possibility of removing disk replicas of unused DAOD datasets and reproducing them on demand. Two scenarios are being considered: reproducing the datasets by rerunning jobs through the production system, or archiving the datasets to tape and staging them back when necessary. Each scenario has its pros and cons. The DAOD recreation approach is technically feasible, as the production system keeps historical records of datasets. However, it requires validation to ensure the reproduced data are identical to the originals, along with the development of new mechanisms and functionalities within the workflow and data management systems. In the archival DAOD scenario,

it will be straightforward to stage them using Data Carousel, similar to other data types like RAW and AOD. But DAOD are generally smaller and have more sporadic access pattern than the other data types, we need to evaluate the extra load on the tape system to regularly stage them out of tape. In both scenarios, one common concern is how quickly we can either reproduce them or stage them from tape. Note that in the recreation scenario, the input AOD datasets are most likely to come out of tape as well.

In this HL-LHC demonstrator, both scenarios will be covered, executed in sequential steps with assessments conducted after each step:

- The first step is for Proof of Concept. We will select a small data sample consisting of 4 to 5 DAOD datasets from a recent lifetime model exception list. Ensure that the parent AOD datasets for these DAOD datasets are available, either on disk or tape. Transfer the DAOD sample to tape at a Tier-1 site where both the DAOD and parent AOD samples are stored, allowing for a performance comparison of different data types during staging. Subsequently, recreate the disk replica of the DAOD sample by either reproducing them using the workflow management system or staging them from tape. In both cases, measure the time-to-completion (TTC).
- Next step(s) will be to repeat the above step, but with bigger data samples, e.g. 50∼100TB DAOD or beyond. This will give more realistic measurements and evaluation of the TTC and tape performance.

In order to minimize the potential influence of site-specific setups on the results, we have devised a plan to conduct these tests at multiple tape sites. Several Tier-1 sites have generously volunteered to participate in these tests. The success of the demonstrator heavily relies on the involvement and expertise of both our Tier-0 and Tier-1 sites.

In terms of the metrics and deliverables, as mentioned earlier, we will measure the time-to-completion (TTC). Furthermore, we will verify the fidelity of the reproduced DAOD datasets, confirming that they match the original ones. Additionally, we plan to perform a spreadsheet or Python notebook exercise to calculate the space savings for a non-exception lifetime model, the extra data volume to stage from tape, as well as the CPU resources needed to recreate the deleted DAOD datasets. We will also assess the additional tape bandwidth required in each scenario, taking into account extrapolations to the HL-LHC conditions. For instance, if staging DAOD from tape exhibits lower efficiency compared to the other larger data types, and if DAOD necessitates lower latency in data delivery, it may result in increased resource purchases at tape sites.

3 Tape smart writing

Data Carousel involves continuous high throughput staging of files from tape, a departure from conventional tape usage. Ensuring the long-term effectiveness of the Data Carousel relies on optimizing tape utilization. Tape exercises by ATLAS and other WLCG experiments [8] have shown that the key to achieving optimal tape usage lies in co-locating files on tape that are likely to be recalled together, so called "smart writing". Note that smart writing is a catch-all phrase that encompasses the multitude of techniques that are possible to intelligently lay out data on tape to enhance read performance. It doesn't refer to any particular implementation. Given the diversity of tape systems and operational models across different sites, it is anticipated that the smart writing solution will be highly specific to each site's requirements and capabilities.

Smart writing has been a long term strategy in Data Carousel. ATLAS tape sites have made various progress implementing site level solutions over the years. To further facilitate this process, we proposed a HL-LHC demonstrator that involves two main activities:

- Running exercises with selected sites, to demonstrate and assess the effectiveness of existing solutions.
- Development of a tape system simulator. The simulator will allow us to explore potential optimization techniques and assess the impact on tape system performance.

By combining these efforts, we aim to enhance our understanding of smart writing techniques and identify optimal strategies for tape utilization in the context of the HL-LHC data scale.Throughout the process, other deliverables are expected as well, including a mechanism to pass data co-location metadata from the experiment to site tape system, and improved tape monitoring at both site and experiment levels.

3.1 Smart writing exercise with KIT Tier-1

Karlsruhe Institute of Technology (KIT) is a Tier-1 site that has made significant advancements in data co-location on tape recently with their new HPSS tape system (Figure 2). We are initiating the exercise with KIT, and we anticipate the involvement of additional sites in the future.

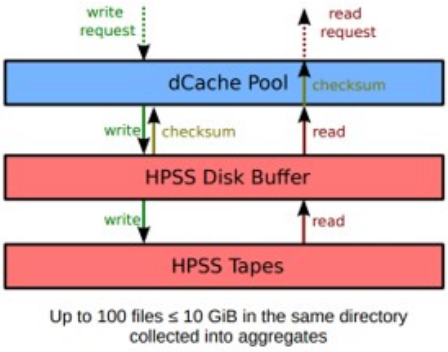

Figure 2: Diagram showing the tape system dataflow at KIT Tier-1

In the KIT HPSS tape system, files are grouped at dataset level, which is the basic unit for file transfers in ATLAS. The conventional file organization technique called File Families is utilized to group files belonging to the same dataset. However, at KIT, file families are a set of predefined random numbers, not associated with any particular namespace. This approach allows for a limited number of file families to be maintained in the system, rather than mapping to the actual number of datasets, which is at the O(1000) level. Additionally, File Aggregates are employed during migration to tape. One file aggregate can contain up to 100 files from the same directory, with each file being less than 10GiB in size. If any file within a file aggregate is recalled, the entire aggregate is staged from tape, ie. a Full Aggreation Recall Mechanism. For further details on the implementation, please refer to the KIT presentation [10].

We plan to conduct this exercise in two phases. The initial phase involves a functional test using a small data sample, to validate the full chain from the ATLAS DDM system to the KIT tape system, with FTS [9] and dCache [11] in the middle. During this phase, the HPSS tape test operations and monitoring will be internally managed by the KIT site experts. Once the new HPSS tape system is deployed for ATLAS production, we will proceed to tape write and

tape read tests using larger data samples. The larger scale tests will be conducted based on the ATLAS production and monitoring services, such as Data Carousel and DDM dashboards. The metrics for this exercise will include the overall tape write and read throughput, as well as the throughput per tape drive, to evaluate the performance of the system.

Success of this exercise is contingent on the availability of metadata, especially name and size of the datasets to which a file belongs. These metadata will be crucial in enabling the site to make dynamic decisions regarding file family assignments and determining the number of tape drives needed to migrate a dataset. Larger datasets require more tape drives in order to meet the throughput requirements while ensuring co-location of files from the same dataset on tapes. For passing the metadata from ATLAS DDM/Rucio to site tape systems, the long term plan is for Rucio to include the metadata as json dictionary when submitting FTS transfer jobs, FTS then transforms metadata into HTTP headers in the WebDav transfer requests sent to site tape storage endpoints. The specific format and implementation details of the metadata are still under discussion and development. For the time being, a short-term solution using URL parameters in the https third party transfer requests is in place, enabling the exercise to proceed while more comprehensive metadata handling mechanisms are being finalized.

3.2 Tape system simulator

Several ATLAS tape sites have proposed to develop a tape system simulator, in collaboration with the ATLAS Distributed Computing (ADC) team. This approach draws inspiration from the time-honored technique of evaluating NFS file system performance using simulators. We will start with exchanging and collecting information from both experiment and site sides, to identify data grouping opportunities. While dataset is an obvious grouping unit, we will also explore opportunities that go beyond dataset level.

I/O transaction logs at various levels (experiment, WLCG services like FTS, and site storage endpoints) will be used to analyze ATLAS tape write and read patterns. An evaluation framework will be implemented to replay these historical I/O transactions on the simulator, to assess different data placement strategies and their effectiveness and cost in achieving the desired data layout on tape.

The simulator will not only help more sites to find and evaluate their own solutions, but also to be used for comparison with real tape exercise results, facilitating the identification of further improvements on existing solutions.

4 Summary

Since 2020, Data Carousel has been running successfully in ATLAS production, resulting in significant savings in disk space. Two new R&D projects, namely "data on demand" and "tape smart writing", are currently underway. The data-on-demand project aims to explore opportunities for further disk space savings by moving less popular data to tape and recreating them on-the-fly when needed. The tape smart writing project is to address the long term tape usage optimization strategies, starting from tape writing, in preparation for the HL-LHC data scale. For all of these projects to be successful, collaboration between experiment and sites is crucial.

Acknowledgments

The work at Plekhanov University is funded by the Russian Science Foundation grant (project No.19-71-30008). The work at Brookhaven National Laboratory is funded by the U.S. Department of Energy, Office of Science, High Energy Physics contract No. DE-SC0012704

References

[1] LHC – The Large Hadron Collider, http://lhc.web.cern.ch/lhc/

[2] ATLAS Collaboration, The ATLAS Experiment at the CERN Large Hadron Collider, JINST **3** S08003 (2008), DOI: https://iopscience.iop.org/article/10.1088/1748-0221/3/08/S08003

[3] M Borodin et al, The ATLAS Data Carousel Project Status, EPJ Web of Conferences 251, 02006 (2021), DOI: https://doi.org/10.1051/epjconf/202125102006

[4] F. H. Barreiro et al, The ATLAS Production System Evolution: New Data Processing and Analysis Paradigm for the LHC Run2 and High-Luminosity, J. Phys.: Conf. Ser. 898, 052016 (2017), DOI: http://dx.doi.org/10.1088/1742-6596/898/5/052016

[5] F H Barreiro et al, PanDA for ATLAS distributed computing in the next decade, J. Phys.: Conf. Ser. 898 052002 (2017), DOI: http://dx.doi.org/10.1088/1742-6596/898/5/052002

[6] M. Barisits et al, Rucio: Scientific Data Management, Comput. Softw. Big Sci. (2019) 3: 11, DOI: https://doi.org/10.1007/s41781-019-0026-3

[7] J. Shiers, World LHC Computing Grid, Computer Physics Communications 177, 219–223, (2007), DOI: https://doi.org/10.1016/j.cpc.2007.02.021

[8] A. Forti et al, Tape Data Challenge 2021, DOI: https://doi.org/10.5281/zenodo.5780575

[9] The FTS project homepage – https://fts.web.cern.ch/fts/

[10] H. Musheghyan et al, Efficient interface to the GridKa tape storage system (2023), in the Proceedings of the 26th International Conference on Computing in High Energy and Nuclear Physics (CHEP2023)

[11] The dCache project homepage – https://dcache.org/