ATL-SOFT-PROC-2025-014
21 January 2025

ATLAS HL-LHC Demonstrators with Data Carousel: Data-on-Demand and Tape Smart Writing

Mikhail Borodin[1], *Alexei* Klimentov[2], *Tatiana* Korchuganova[3], *Mario* Lassnig[4], *Julien* Leduc[4], *Tadashi* Maeno[2], *Haykuhi* Musheghyan[5], *Doris* Ressmann[5], and *Xin* Zhao[2,*]
on behalf of the ATLAS Computing Activity

[1]The University of Iowa, USA
[2]Brookhaven National Laboratory, Upton, NY, USA
[3]The University of Pittsburgh, USA
[4]European Organization for Nuclear Research (CERN), Geneva, Switzerland
[5]Karlsruhe Institute of Technology, Karlsruhe, Germany

Abstract. The High Luminosity upgrade to the LHC (HL-LHC) is expected to deliver scientific data at the multi-exabyte scale. To tackle this unprecedented data storage challenge, the ATLAS experiment initiated the Data Carousel project in 2018. Data Carousel is a tape-driven workflow in which 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. Put in ATLAS production before Run3, Data Carousel continues to be our focus for seeking new opportunities in disk space savings, and enhancing tape usage throughout the ATLAS Distributed Computing (ADC) environment. These efforts are highlighted by two recent ATLAS HL-LHC demonstrator projects: data-on-demand and tape smart writing. In this paper, we will discuss the recent studies and outcomes from these projects. The research was conducted together with site experts at CERN and Tier-1 centers.

1 Data Carousel and ATLAS HL-LHC demonstrators

Data Carousel [1] is an R&D project that the ATLAS experiment [2] started in June 2018, to address the data storage challenge of the High Luminosity upgrade to the LHC [3]. It 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. Since 2020, Data Carousel has been utilized in production for major ATLAS managed campaigns, including RAW data reprocessing, derivation, and Monte Carlo simulation. Figure 1 shows the recalled volume from ATLAS Tier-0 and Tier-1 sites since 2020.

The ATLAS HL-LHC demonstrators [7] are initiatives designed to test and validate new technologies and solutions across various areas, including software and computing, to meet the challenges of the HL-LHC upgrade. Results from these projects will be incorporated into

*e-mail: xzhao@bnl.gov

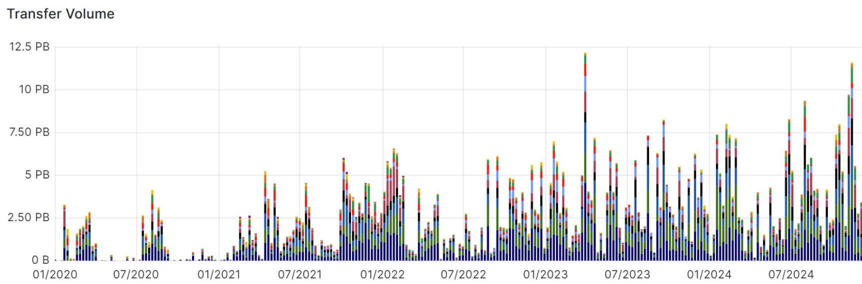

Figure 1: Data volume recalled from Tier-0/Tier-1 sites by ATLAS Data Carousel since 2020, from the ATLAS DDM dashboard (weekly bin size). Different colors represent different source countries/regions.

the ATLAS HL-LHC Technical Design Report (TDR). In this paper we present the progress and results from two HL-LHC demonstrators that are from the ATLAS Data Carousel activity: Data-on-demand and Tape smart writing.

2 Data-on-demand

The current practice in ATLAS is to use disk to store derived analysis object data (DAOD), which are inputs to user analysis jobs. However many DAOD datasets remain unused. In this demonstrator, we explore different scenarios to delete the unused DAOD datasets from disk and recreate them on demand later if needed. This will not only save disk space, but also reduce operational load on both the ATLAS Distributed Data Management (DDM) team and physics groups, because currently they need to collect and approve lists of unused DAOD datasets that users want to keep on disk(primarily out of concern for potential future revisions during paper approval or journal interactions), and put them in an exception list during each ATLAS deletion campaign, which is a labor-intensive procedure.

This demonstrator evaluates two scenarios: (1) reproducing DAOD datasets by rerunning jobs and (2) archiving DAOD datasets to tape and recalling them as needed. Two metrics are used for evaluation. The first is Time to Completion (TTC), defined as the time elapsed from a user's request submission to the availability of DAOD dataset(s) on disk for access. This process includes various steps such as task submission, jobs queuing and execution time, recalling data from tape etc. For requests involving multiple datasets (bulk mode), the TTC is measured on a per dataset basis, and the final result is calculated as the average time per dataset. The second metric assesses the additional CPU and storage resources required for each scenario.

The tests were based on a data sample chosen from the DAOD exception lists of 2023 ATLAS deletion campaigns. When selecting the sample datasets, the following were excluded: (1) not accessed for more than 12 months; (2) non-production related datasets, such as validation datasets; and (3) datasets smaller than 50GB.

For the rerunning jobs scenario, PanDA, the ATLAS Production and Distributed Analysis system [5], was used. Because this scenario often necessitates recalling analysis object data(AOD) from tape as job input, the tests are divided into two parts: first, job submission and execution using AOD datasets already available on disk; and second, evaluating recall performance for parent AOD datasets, which is combined with the DAOD tape recall tests in the archival scenario.

Tape recall tests were performed under varying conditions, considering the volume of parallel recall requests and multiple operational modes:

- Single dataset recalls, for both DAOD and AOD datasets individually.
- Paired recalls, a DAOD dataset and its parent AOD dataset were recalled simultaneously to compare performance.
- Bulk mode recalls. Large-scale recalls were tested at KIT Tier-1, partly combined with the tape smart writing demonstrator, which is discussed in the next section.

To minimize potential bias in the results due to specific tape site configurations and conditions, the tape tests were conducted at two ATLAS Tier-1 sites: the Karlsruhe Institute of Technology (KIT, also known as FZK) in Germany and the Rutherford Appleton Laboratory (RAL) in the UK.

After evaluating the two scenarios, we found all of them give comparable TTC:

$$\begin{aligned}\langle \text{TTC of recalling DAOD from tape} \rangle &= 13.1 \pm 5.6 \text{ hours} \\ \langle \text{TTC of recalling AOD from tape} \rangle &= 12.4 \pm 7.2 \text{ hours} \\ \langle \text{TTC of recreating DAOD by jobs} \rangle &= 7 \pm 3 \text{ hours}\end{aligned}$$

Figure 2 shows the TTC distribution of recalling DAOD and AOD datasets respectively, from the two Tier-1 sites. We did not observe correlations between the TTC and dataset size, as shown in Figure 3. As explained earlier, many factors can contribute to the final TTC, not just dataset size.

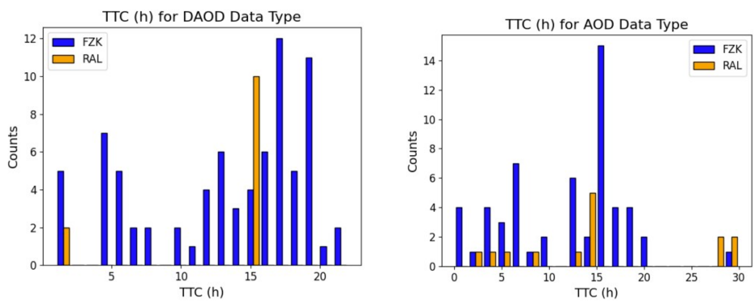

Figure 2: Measurements of Time-to-Completion (TTC) of recalling DAOD and AOD datasets respectively, from two ATLAS Tier-1 sites (FZK and RAL).

Since DAOD datasets are relatively small compared to AOD datasets, there was concern that recalling DAOD from tape would be less efficient. However, bulk test results show that by grouping files on tape, recalling DAOD can be as efficient as for the other data types. This will be discussed in greater detail in the next section.

To estimate the additional CPU and storage resources required for each scenario, we studied the access pattern of DAOD datasets listed in recent deletion campaign exception lists (Figure 4). It shows that while the volume of DAOD datasets users request to keep on disk has increased over the years, these datasets are rarely accessed afterwards. Based on this pattern, the additional resource load for reproducing them is deemed negligible.

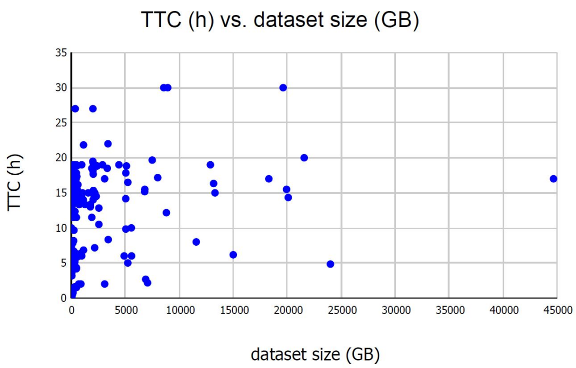

Figure 3: No correlation between Time-to-Completion (TTC) and size of datasets.

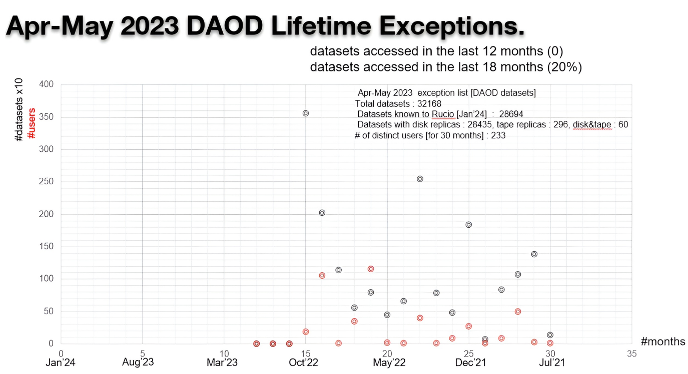

Figure 4: Access pattern to DAOD datasets after they are put into lifetime model exception list.

3 Tape smart writing

Data Carousel involves the frequent, high-throughput staging of files from tape, which represents a shift from the traditional "archival-only" tape model. Optimal tape utilization is essential to ensure the long-term success of the Data Carousel. Our key strategy to achieve this goal is by co-locating files on tape that are likely to be recalled together, so called "smart writing".

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. While the term "smart reading" is not used, effective tape usage inherently involves reading data in alignment with how it was written. Writing and reading are two sides of the same coin, while writing plays an even more important role because it sets the foundation for reading performance.

Since the inception of the Data Carousel project, various ATLAS tape sites have put in a lot of effort and made great progress implementing site level smart writing solutions. In this demonstrator, we run tape exercises with KIT Tier-1 site, to evaluate their solution and gain insights from their experiences.

In 2023, the KIT Tier-1 transitioned its tape system from TSM to HPSS. The new HPSS system [8] groups files from the same dataset together on tape. To ensure good write and read performance for large datasets, the number of tape drives allocated for writing a dataset can be dynamically adjusted based on the dataset's size. This size information is provided as metadata for each file by the ATLAS DDM system.

Figure 5: Tape recall rate vs number of tape drives in use for RAW,AOD and DAOD datasets respectively (from the KIT monitoring). 400MB/s is the native (max) transfer rate of the TS1160 tape drive. All test results show 80% or higher tape bandwidth utilization.

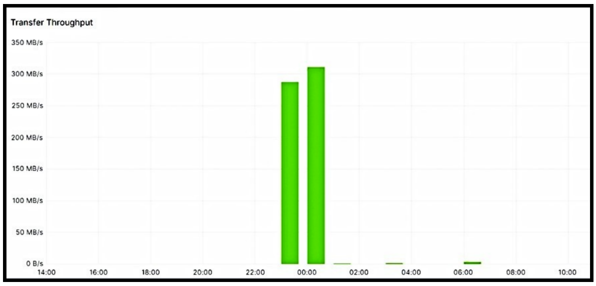

Figure 6: Transfer rate of a single AOD dataset (2TB in 295 files) out of KIT tape (from the ATLAS DDM dashboard).

This tape exercise was conducted during KIT's transition from the TSM to the HPSS tape system. During this period, both systems were operational, allowing us to write data samples to both systems, recall them, and compare performance.

We selected data samples for three data types: raw data (RAW), analysis object data (AOD) and derived analysis object data (DAOD), with approximately 100TB of sample data for each type. The dataset size distribution in the samples roughly mirrors the distribution in the real data. Additionally, a couple of datasets larger than 40TB were deliberately included to evaluate performance on large datasets.

The recall tests were conducted in a controlled environment, minimizing other parallel recall requests (production Data Carousel requests to KIT were paused during the test) to allow for a clear observation of the recall performance for each data type. The primary metric used in this test was tape bandwidth utilization, defined as the percentage of a tape drive's nominal bandwidth achieved during recall operation, calculated as (observed bandwidth)/(nominal bandwidth). Monitoring for this exercise was performed using both KIT's internal tape monitoring system and the ATLAS DDM dashboard.

Figure 5 shows the total recall rate vs the number of tape drives in use, from the RAW, AOD and DAOD tests respectively. The bandwidth utilization in all cases reached 80% or higher, a factor of two improvement over the old TSM system, which did not have file grouping in place. Figure 6 shows the transfer rate of a particular dataset, which has 295 files with a total size of 2TB. With dataset-level file grouping, this behavior of a single dataset recall is expected – when the corresponding tapes are mounted, all the files from the same dataset are recalled at a high rate within a short period of time. This also helps mitigate the long tail effect, where a small number of files in a dataset take significantly longer to be recalled compared to the rest.

4 Summary

Both ATLAS HL-LHC demonstrators have made substantial progress, with promising outcomes. The Data-on-demand demonstrator introduces a new way of handling disk replicas of unused DAOD datasets, potentially leading to more disk space savings and reduced operational load. The tape smart writing exercise highlights how grouping files on tape can greatly optimize tape bandwidth utilization. Looking ahead, we aim to collaborate with more sites to assess their solutions as they become available.

Acknowledgments

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] 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

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

[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] D. South et al, ATLAS Distributed Computing Evolution: Developments and Demonstrators Towards HL-LHC, EPJ Web of Conference 295, 04034 (2024) DOI: https://doi.org/10.1051/epjconf/202429504034

[8] H. Musheghyan et al, Efficient interface to the GridKa tape storage system, EPJ Web of Conference 295, 01020 (2024) DOI: https://doi.org/10.1051/epjconf/202429501020