Examining the Impact of Data Layout on Tape on Data Recall Performance for ATLAS

Shigeki Misawa[1,*] on behalf of the ATLAS Collaboration

[1]Scientific Data and Computing Center, Brookhaven National Laboratory, Upton, New York 11973

Abstract. Increases in data volumes are forcing high-energy and nuclear physics experiments to store more frequently accessed data on tape. Extracting the maximum performance from tape drives is critical to make this viable from a data availability and system cost standpoint. The nature of data ingest and retrieval in an experimental physics environment make achieving high access performance difficult given the inherent limitations of magnetic tape. Tailoring the layout of data on tape is one key to improving read performance. This paper highlights the work in progress to characterize ATLAS data ingested in the tape system, understand how data layout, i.e. file co-location on tape and file distribution over tapes, affect read performance and how optimal data layout might be achieved in a production environment.

ATL-SOFT-PROC-2023-026
14 September 2023

1 Introduction

The ATLAS experiment [1] is expected to generate 100s of petabytes per year when the High Luminosity LHC starts running in 2029. The volume of data that ATLAS expects to actively use is well beyond what can be economically stored on disk [2]. To mitigate this problem, ATLAS is increasing its use of magnetic tape as an active, near-line store for less frequently used data. However, the cost effectiveness of tape can be compromised and data retrieval performance can be severely affected if the inherent limitations of tape are not taken into account.

2 Limitations of Tape

The two primary drawbacks of tape systems are long seek times (measured in seconds) and long tape cartridge mount and dismount times (up to two minutes for the combined operations [3]). The time spent mounting, dismounting and seeking is time not spent reading data. Equation 1 quantifies the impact of long seek and mount times on the effective tape drive read bandwidth.

$$\text{Fraction of Max Bandwidth} = \frac{1}{1 + T_{Move}/T_{Read}} \tag{1}$$

T_{Move} is the mount/dismount and positioning time, roughly 120 seconds if head repositioning time after initial seek is zero. It will be larger if the data to be read are not contiguous on

*e-mail: misawa@bnl.gov

tape. T_{Read} is the time spent reading data, i.e. data volume read divided by the read speed of the tape drive, the latter being approximately 400 MBytes per second for current generation tape drives. Figure 1 graphs effective drive read bandwidth as a function of the data volume read per tape mount.

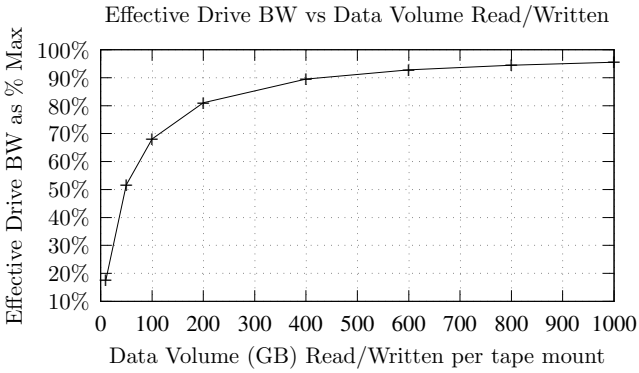

Figure 1. Effective tape drive bandwidth, as a percentage of the maximum theoretical drive bandwidth (400 MB/s for current generation drives), as a function of data volume being read or written per tape mount

3 ATLAS Environment

The limitations of tape systems would be a non issue if all data on a tape were read back as written on the tape in a single tape mount. However, this is unrealistic given the heterogeneous characteristics of the data ATLAS writes to tape.

3.1 Data Characteristics

The data that ATLAS writes to tape are either data from the detector or Monte Carlo data. It includes, but is not limited to "raw" data, output from various stages of simulation or analysis, log files, and conditions data. The "temperature" of the data ranges from cold, i.e. files that are never read, to warm data, i.e. files that might be accessed one or more times per year. Hot (frequently accessed) data are also written to tape, but a copy is likely to remain on disk obviating the need to access them from tape. The fundamental quantum of data in the ATLAS is the file. Files in turn are logically grouped into datasets by ATLAS. This grouping is reflected in ATLAS's Rucio based data management system [4]. Requests for data storage and retrieval are by dataset, but this information is lost at the tape system level as they see requests by individual file.

3.2 Data Read/Write Profile

At any given instance ATLAS is writing (and reading) files from multiple datasets into tape systems. Files from different dataset will get interleaved on tape by default. The long transfer

time for a datasets, i.e. the time it takes to move all files in a dataset to the site with the tape system, will aggravate this interleaving. A cumulative histogram of dataset transfer times is shown in Figure 2. Even if serially written to tape(s), the small size of ATLAS datasets, shown in Figure 3, can still result in low effective tape drive bandwidth. If only one dataset is read per tape mount, the effective tape bandwidth for data from the detector ("real" data) would be below 50% of maximum bandwidth for 70% of datasets. For Monte Carlo data, it would be below 50% for 85% of datasets. This situation is made worse by the fact that a dataset may end up on more than one tape, as multiple tapes are typically written at once to satisfy aggregate ingest bandwidth requirements. At the US ATLAS Tier 1 facility at Brookhaven, 25 tapes are written simultaneously during LHC data taking while 5 tapes written at once at other times.

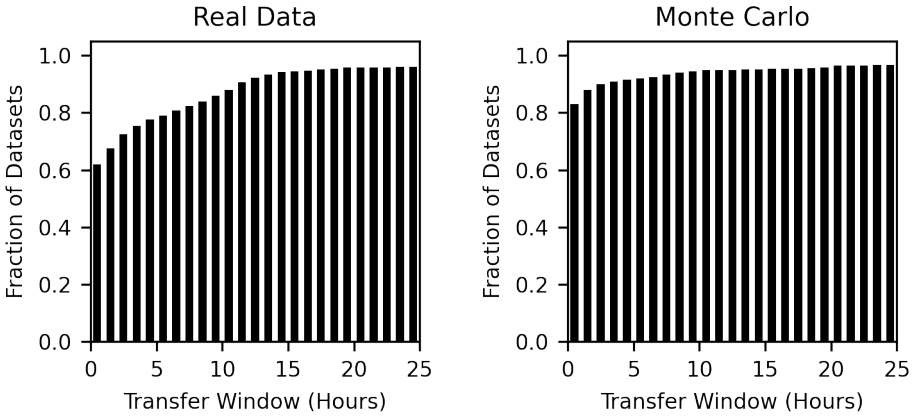

Figure 2. Cumulative histogram of ATLAS dataset transfer times for data from the detector (Real) and Monte Carlo data. From 988 real experiment and 8588 Monte Carlo datasets written between May 24, 2022 and September 14, 2022 at Brookhaven.

4 Optimizing Strategies

4.1 Reads

Given the limitations of tape systems, the clear strategy for enhancing reads is to read as much data as possible per tape mount and to read the data as it is physically laid out on tape. The former can be achieved at the expense of increased access latency by aggregating a large number of file read requests and sorting the requests by tape. The latter is more difficult due to the serpentine nature of magnetic tape. Recommended Access Order (RAO) technology, available in the latest generation of tape technology, is necessary to optimized the order of files read from tape. [3]. However, there is a limit to the effectiveness of read optimization as it cannot overcome the sparse distribution of files in a dataset on tape and over a large number of tapes.

4.2 Writes

Achieving operational efficiencies beyond what is attainable from optimizing read requests alone requires arranging the layout of files on tapes to match the expected read patterns.

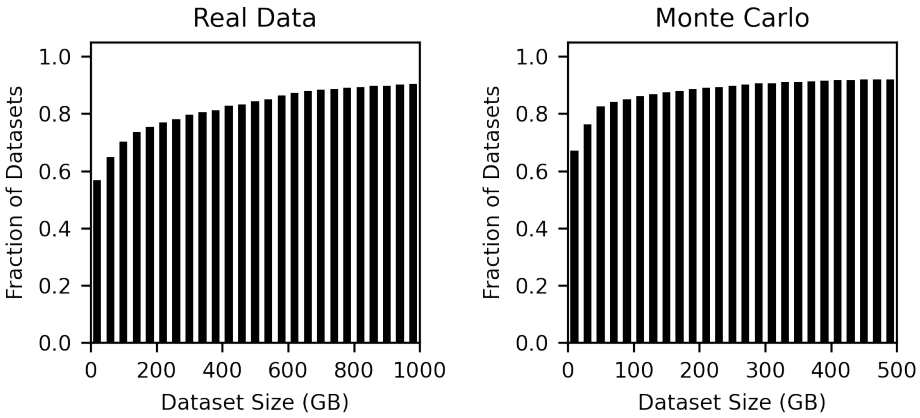

Figure 3. Cumulative histogram of ATLAS dataset sizes for real experiment data and Monte Carlo data. From 988 real experiment and 8588 Monte Carlo datasets written between May 24, 2022 and September 14, 2022 at Brookhaven

4.2.1 Group by Dataset

As datasets are the fundamental unit of data retrieval, collocating files in a dataset sequentially on tape is the obvious first step. Achieving this requires identifying the files in a dataset and verifying that all files have been received before writing them to tape. The first piece of information is in the ATLAS data management system, but isn't explicitly relayed to the tape system. The total size of the dataset must also be provided as it is needed to assign the correct number of tape drives to move the dataset to tape at the required bandwidth. Note that this has consequences on reads, as the number of tapes containing a dataset determines the upper limit on the access bandwidth (# tapes x 400MB/s) to the dataset, therefore there is tension between spreading a dataset over too few and too many tapes. Dataset size is also necessary to allow the tape system to manage the space where datasets are staged until completely received.

4.2.2 Segregate Datasets by Access Class

Another mechanism for increasing read efficiencies is to segregate datasets by access temperature. Putting warm (more likely to be accessed) and cold (rarely or never accessed) datasets on separate tapes would increase the density of warm datasets on a given tape and hopefully increase the amount of data being read per tape mount. There are four issues with this optimization. First, determining data access temperature ideally requires information from ATLAS, but analysis of historical access logs might illuminate categories of datasets that are likely to be cold in the absence of information from ATLAS. Second, dataset access temperature must be known before it is transferred to the tape site. Third, this additional dataset "metadata" must be created (either "by hand" or algorithmically from other dataset metadata like the dataset name), stored (most likely in Rucio), and communicated to the tape site. Finally, warm datasets on a tape may not be retrieved at the same time, despite being more likely to be accessed.

4.2.3 Group by Correlated Datasets

The final possible write optimization is the collocation of correlated datasets, i.e. placing datasets that may be read together onto the same set of tapes. Compared to the previous two techniques, this one is significantly harder. The identification of correlated datasets is likely to be more difficult. A more detailed understanding of dataset retrieval patterns will be necessary. As with segregation by access class, this information might be available from ATLAS or potentially derivable from the analysis of historical access logs. Logistically, the larger size of datasets, in comparison to files, will make collocation of correlated datasets more difficult in practice than the collocation of files within a given dataset.

5 Achieving Optimal Layout

Attaining optimal data layout on tape is a multi-step process:

- Identify files/datasets to collocate
- Create collocation metadata for the data
- Distribute collocation information to the tape system
- Have the tape system tailor the placement of data on tape based on the metadata

This will require changes to the ATLAS distributed data management system and data storage systems at each tape site. With this in mind, a cost benefit analysis is helpful to determine which specific collocation mechanisms are worth investing in. Given the differences in tape systems at the ATLAS tape sites, the results of this cost benefit analysis will vary from site to site.

5.1 Limitations

Although simple in principle, enhancing read performance through selective layout data on tape has distinct limitations. First, limits to the amount of disk capacity available to hold data targeted for collocation will restrict collocation to data that exhibit temporal locality, i.e., are received closely in time. Second, segregating different classes of data to separate tapes and using separate pools of tapes to work around limits to collocation disk storage capacity both increase tape mounts when writing data to tape. This will reduce the effective write bandwidth of tape drives in the same way that it affects reads. This overhead also limits the granularity to which data segregation can be applied. Third, data collocation depends on the ability to control how a tape system writes data to tape. This control may not be absolute or available at all on some tape systems. This in turn will reduce the level of collocation that can be achieved in practice. Fourth, excessive collocation can also be problematic, as there may not be enough tapes with desired data to keep all available tape drives operating, resulting in the under utilization of drive resources. The optimal level of collocation will likely depend on the operational environment, which will vary over time. Finally, optimizing reads through selective data placement assumes that future data access patterns can be articulated before data is written to tape for a large fraction of the data that will be accessed.

6 Future Work

Work is ongoing within the ATLAS distributed computing and data management community to enhance the systems in the data distribution pipeline to enable more optimized placement of data on tape and to identify and classify datasets according to their access characteristics.

The development of a tape system simulator is also of interest to quantify the efficiency gains achieved with specific optimization techniques. This would be done by replaying real file read and write logs on the simulator with different optimization techniques enable.

7 Conclusions

Paying attention to how data is written onto tape can potentially improve tape system file recall performance. Collocation of files to be read together sequentially on a limited number of tapes reduces the overhead of tape mounts and tape head seeks when reading data from tape, but requires prior knowledge of the files that need to be grouped together. Determining the appropriate level of collocation is in need of additional investigation. Finally, achieving the optimal layout of data on tape requires modification to tape systems and other parts of the ATLAS data distribution pipeline.

References

[1] The ATLAS Collaboration, Journal of Instrumentation **3**, DOI 10.1088/1748-0221/3/08/S08003 (2008)

[2] The ATLAS Collaboration, Tech. Rep. CERN-LHCC-2022-005 ; LHCC-G-182, CERN. Geneva. The LHC experiments Committee ; LHCC (2022)

[3] O. Asmussen, R. Beiderbeck, H. Hörhammer, K. Ngo, J. Rolon, F. Villarreal, L. Coyne, *IBM Tape Library Guide for Open Systems* (IBM, 2022)

[4] M. Barisits, T. Beermann, F. Berghaus, B. Bockelman, J. Bogado, D. Cameron, D. Christidis, D. Ciangottini, A.D. Girolamo, G. Dimitrov et al., Computing And Software For Big Science **3**, DOI 10.1007/s41781-019-0026-3 (2019)