Updates on ATLAS Data Carousel

Xin Zhao (BNL), Alexei Klimentov (BNL), Mario Lassnig (CERN), Misha Borodin(UIowa)
DOMA, December 4th, 2024

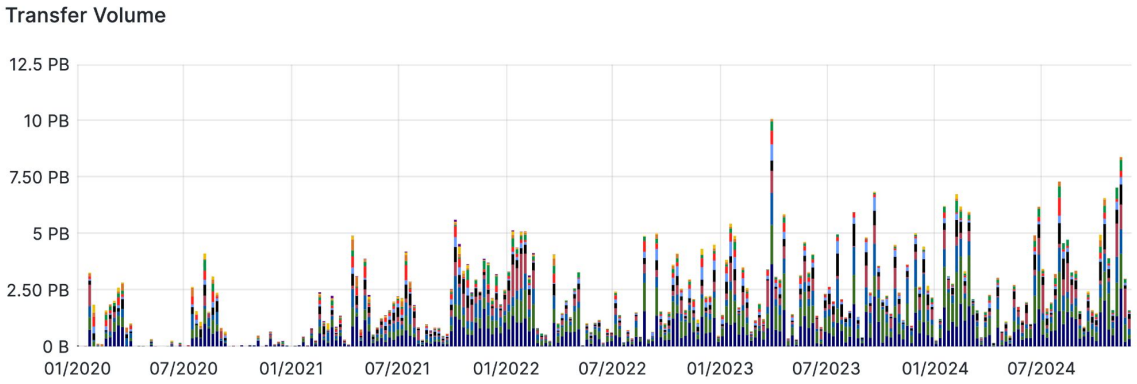
Outline

- Introduction
- Current activities on tape smart writing
  - Demo with selected sites
  - Ongoing discussions and open questions (archival metadata and beyond)

* Team effort --- WFMS team, DDM team, Ops team, many other ADC experts, all T0 and T1 site experts and various storage service provider groups

ATLAS Data Carousel (1/2)

- Tape driven workflow
  - Jobs can get inputs directly from tape
  - To address the storage challenge of HL-LHC
- In production since 2020
  - Today major ATLAS production campaigns(reprocessing, derivation, MC simulation etc)) run in Data Carousel mode



Data volume recalled from T0/T1 tape since 2020 (weekly bin size)

ATLAS Data Carousel (2/2)

- Operationally, continuously address issues encountered in production, e.g. :
  - Alarm for long tail requests (GGUS tickets to sites)
  - Holding “T0 export” traffic till the end of a run, so T1s can get dataset size metadata for RAW
- A recent example – in expectation of big runs/datasets (O(~PB)) coming out of the 2024-10 LHC p-p reference run, ADC had a plan in place to split big datasets among multiple T1s, to help release pressure on tape buffer at sites
  - This plan was not applied because the run didn’t produce big fills.
  - For the long term, one suggestion is to have FTS automatically adjust the tape writing stream based on backpressure from sites (under discussion)
- While mitigating current operational issues, always focus on the key to our long term success — optimal tape usage

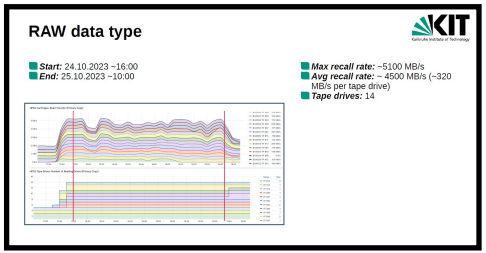
Tape Smart Writing

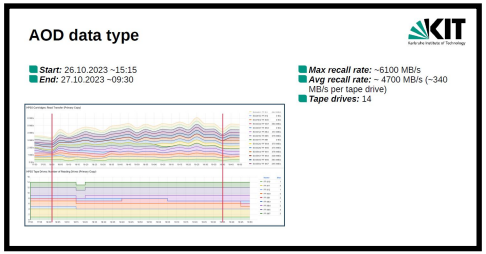
- How to optimize tape usage ?
  - to reduce tape (re)mounts and seek time
    - the “Reference” slide has an incomplete list of studies done by various sites/groups, from different perspectives, over the years on this topic
- Our key strategy to achieve optimal tape usage is to group files on tape according to access patterns — so called “smart writing”
  - “Smart writing” is a catch-all phrase, encompassing various techniques for optimizing data layout on tape to improve read performance.
  - Reading should match how data is written on tape, the other side of the same coin, although we don’t call it “smart reading”
- The following slides mainly focus on this topic …

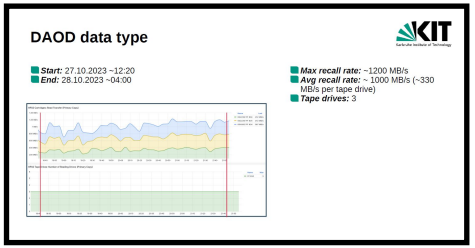
Aim to put our data on tape like a well-organized warehouse

Tape smart writing exercise with KIT (1/2)

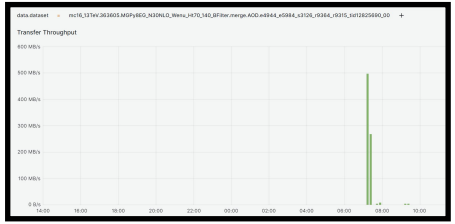
- Together with KIT site experts (Haykuhi Musheghyan etc), did a dedicated tape test
- Result shows 80%+ tape bandwidth utilization, a factor of two improvements over their old TSM tape system (all use TS1160 tape drive w/ 400MB/s nominal rate)







Transfer rate for a 21B/295 files AOD dataset



Transfer rate for a 466GB/187 files AOD dataset

KIT HPSS tape monitoring (*courtesy of Haykuhi Musheghyan from KIT*)

ATLAS DDM dashboard

Tape smart writing exercise with KIT (2/2)

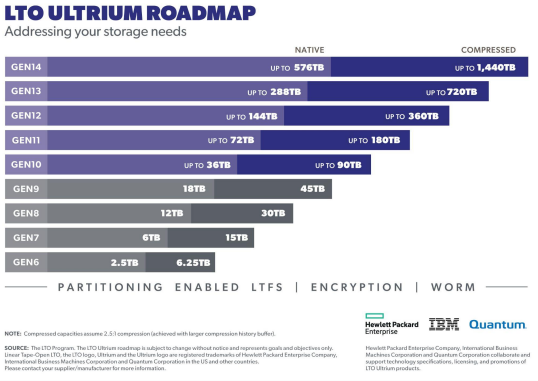
- KIT implementation of tape smart writing
  - Details on KIT presentations (link1, link2, link3)
- Some points of the KIT implementation I'd like to highlight
  - A flexible way to assign different number of tape drives to write a dataset to tape, depending on the size of the dataset
  - Information of dataset size is a metadata that ATLAS DDM (Rucio) passes along when transfer files to tape endpoint
    - Temporary solution from Rucio for passing metadata using URL parameters
- One discussion point about the KIT test result
  - How much of the factor two improvements (over the old TSM) is attributed to file grouping ?
    - No detailed measurements to determine contributions of each factor
    - But, theoretically 80%+ bandwidth utilization would not be possible without good file placement on tape

Next Steps

- Work with more sites, do demo exercises when they feel ready
- Provide sites with tape grouping hints, a.k.a. archival metadata

Data grouping unit(s) on tape

- Dataset is a natural grouping unit for ATLAS (for some other experiments as well)
  - ATLAS can provide additional information like “number of files” and “total size” of a dataset (as we have done for KIT)
- As tape capacity and speed continue to grow in the future, grouping levels above dataset will become necessary, in order to keep the bandwidth utilization high
  - c.f. BNL studies



Archival metadata

- A generic solution being developed
  - Using HTTP header (in json format) in the transfer request
  - A flexible format proposed by CTA/dCache group (1KB size limit enforced)
  - Experiments need to fill in the contents of the metadata
- ATLAS provides the first archival metadata template (draft) for RAW data type, to be tested by CTA@CERN
  - Rucio passes these metadata to CTA, via FTS, during the recent LHC Heavy Ion run.
  - ATLAS still needs to work on metadata templates for the other data types (AOD etc)

```
archive_metadata = {
    "scheduling_hints": {
        "archive_priority": "100"           # highest priority
    },
    "collocation_hints": {
        "0": "data23_13p6TeV",                                  # project
        "1": "RAW",                                             # datatype
        "2": "physics_Main",                                    # stream_name
        "3": "data23_13p6TeV.00452799.physics_Main.daq.RAW",    # dataset
    },
    "additional_hints": {
        "activity": "T0 Tape",          # Tier-0/DAQ
        "3": {                          # dataset level
           "length": "19123",               # total number of files at specified level
           "size": "80020799318456"         # total size of files at specified level
        }
    },
    "file_metadata": {                  # file content metadata
        "size":"193734404",
        "adler32":"379ebf71",
        "md5":"952c4c0dabc622a94f09b053d71d0dfb"
    }
}
```

Questions about Archival metadata templates

- What are a good grouping hierarchy for a data type ?
  - Ask experts (production managers, physics groups …)
    - Sometimes not easy to converge among experts
  - Ask data ?
    - Analyze historical recall logs
    - Rucio has the full recall history for all files and datasets with tape origin.
  - Ask machine ?
    - Run the historical recall logs through ML models, let AI/ML learn recall patterns (e.g. what datasets are likely to be recalled together ?)
- It's hard (if not impossible) to know the size of a grouping unit above dataset level
  - Size info is important, refer to the KIT implementation
  - Ideas floating around …
    - No need to know the real size of all RAW datasets belonging to a particular stream collected during 2024 run. Our purpose is to find grouping units that's big enough to ensure good bandwidth utilization in recall campaign
    - Rucio can create artificial retrieval groups within a level, e.g. put several physics_main stream (level 3) datasets into one container, and tell sites to co-locate them together.
      - we can call them "tape containers", a container type solely for tape grouping purpose
    - Definition of a "good size" is expected to grow as tape technology evolves, and may even be different per site.

Other open questions/discussions (1/2)

- Tape simulator
  - Proposed and planned by some sites
    - For example, to replay tape write history, through a particular file placement scenario; then replay tape read history, and tell what's the expected (theoretical) tape drive bandwidth utilization and overall throughput
  - Answer questions like :
    - which grouping scenario is better, under a certain condition, e.g. one dataset on one (or few) tape or stripped grouping among multiple tapes ?
    - how much performance improvements (theoretically) is expected from one grouping scenario over the others ?
    - what's the ideal size of grouping units, assuming certain conditions and tape technology ?
    - may point out things to improve also on the way tape write/read requests are sent to sites

Other open questions/discussions (2/2)

- Expected data volume and size, throughput targets etc for Run4 ?
  - These will come from experiments, closely related to what we do here.
  - They set the goal for any optimization we do
    - e.g. if a site feels comfortable with meeting the goals without changing the current tape operation model, it's perfectly fine.
  - They help provide guidance to the optimization
- Tape monitoring
  - Overall throughput delivered from tape
  - Bandwidth utilization
  - …
- Within ADC, we continue to evaluate our tape workflows, to leverage the strength of the tape system for optimal usage.

References

Below is an ***incomplete*** collection of various studies on optimizing tape usage (in no particular order)

1. https://iopscience.iop.org/article/10.1088/1742-6596/898/8/082024/pdf
2. https://indico.cern.ch/event/823340/contributions/3558591/attachments/1918104/3171992/ATLAS-CTA.pdf
3. https://indico.cern.ch/event/915292/contributions/3848357/attachments/2039058/3414671/TRIUMF_Tape_Carosal_20200514.pdf
4. https://www.epj-conferences.org/articles/epjconf/pdf/2020/21/epjconf_chep2020_04026.pdf
5. https://www.epj-conferences.org/articles/epjconf/pdf/2021/05/epjconf_chep2021_02016.pdf
6. https://indico.cern.ch/event/1212249/contributions/5128663/subcontributions/404547/attachments/2563622/4419225/OptWriting-TIM-Dec-2022.pdf
7. ……