ATLAS Data Carousel

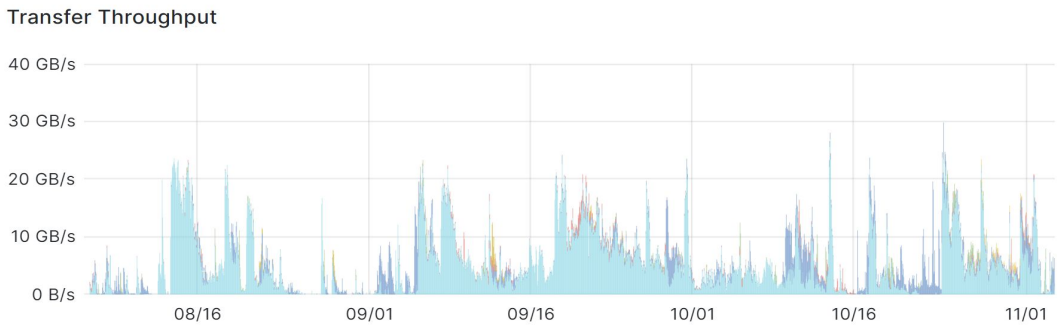
Alexei Klimentov (BNL), Mario Lassnig (CERN), Xin Zhao (BNL)
pre-GDB, November 7th, 2023

Outline

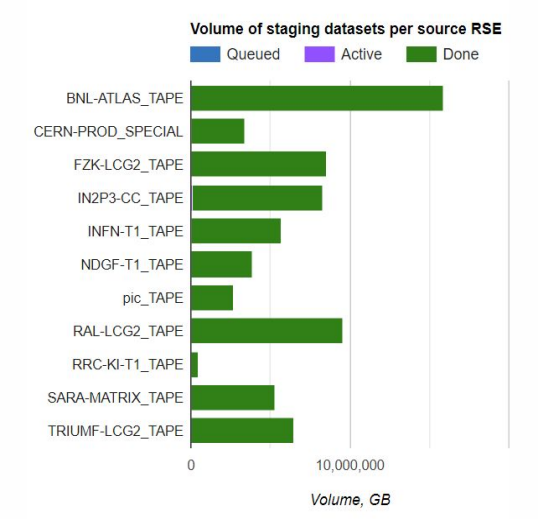
- Overview of the ATLAS Data Carousel
- Current activities
  - Two HL-LHC demos
    - DAOD-on-demand
    - Tape smart writing
  - Metadata of data grouping on tape
  - Tape @Tier-2

ATLAS Data Carousel

- Tape-driven workflow management system
  - To address the storage challenge of HL-LHC, ATLAS started the Data Carousel project in 2018, let jobs get inputs directly from tape.
- Tape resources integrated into the ATLAS distributed computing (ADC) system, new services and protocols developed and implemented in major ADC components (ProdSys2/PanDA, Rucio)
- In production since 2021
  - By now all major ATLAS production campaigns, including reprocessing, derivation datasets production and Monte-Carlo simulation are running in Data Carousel mode.
- Several R&D projects ongoing (next slides)
  - Part of the ADC HL-LHC demonstrators — R&D projects aimed to provide inputs to the HL-LHC Technical Design Report (TDR)



Overall throughput from T0/T1 tape endpoints since 2023-08 (grouped by activities)



Volume recalled per site in the year of 2023 (* only counts unique datasets, and not including non-Data-Carousel recalls)

DAOD-on-demand HL-LHC demonstrator (1/2)

- So far the data types ATLAS put on tape are mainly RAW, simulated detector HITS, and reconstructed AOD. There are a lot rarely used DAOD datasets (inputs for user analysis) on disk. The idea is to remove them from disk and reproduce them on demand, if needed later.
- Two possible scenarios under consideration, both involve tape.
  - DAOD recreation – rerunning jobs to recreate the DAODs (recall AOD from tape as input)
  - DAOD on tape — archiving DAOD to tape and recalling them back when necessary
- This demonstrator is conducted in steps
  - Start with small scale tests, then move to bulk mode.
  - Multiple tape sites (Tier-1s) participated.
- Metrics
  - Walltime to reproduce DAOD dataset
  - CPU, disk and tape resources needed

DAOD-on-demand HL-LHC demonstrator (2/2)

- Tests done so far on both scenarios, using data sample from the recent ATLAS data deletion campaign, at two Tier-1s (FZK and RAL).
- Preliminary results
  - Comparison of TTC (Time To Completion) among different scenarios

Data type	# datasets	#files	Size (GB)	Action	<TTC> per dataset (h)	Source (tape) site	Time stamp
AOD	13	31627	107047	Staging	19 +/- 9	FZK/RAL	July~Sep 2023
DAOD	11	1555	7284	Staging	3 +/- 4	FZK/RAL	July~Sep 2023
DAOD	5	1158	5459	recreation	7 +/- 3	N/A	July~Sep 2023

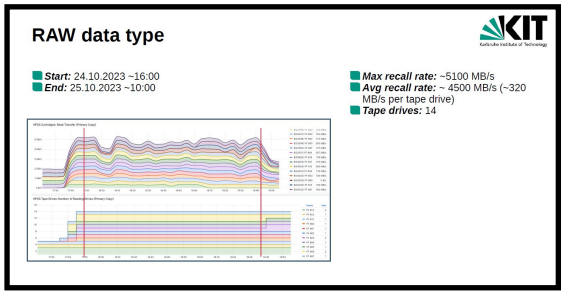
- Bulk mode tests are ongoing, of which the results will be used to estimate both the TTC at scale and the extra load on the tape resources.

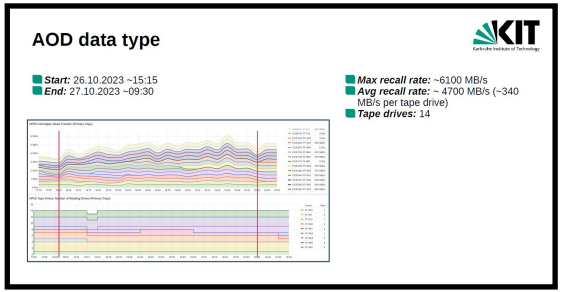
Tape smart writing HL-LHC demonstrator (1/2)

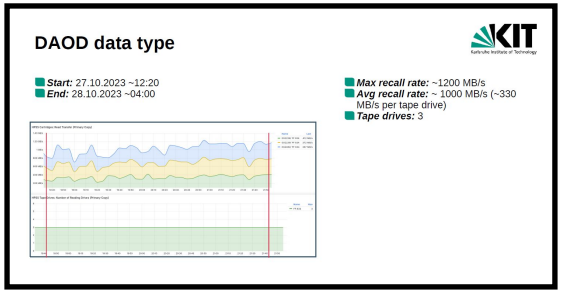
- “Smart writing” — to optimize tape usage by co-locating files on tape that will be recalled together in the future.
  - A “catch all” phrase that encompasses the multitude of techniques that are possible to intelligently lay out data on tape to enhance recall performance. Not referring to any specific implementation, which is under the control of sites.
  - Tape exercises by ATLAS and other WLCG experiments have shown good performance (delivered vs nominal throughput) from sites that group files on tape
- In this demonstrator
  - We run tape exercises with certain sites, to demonstrate and assess the smart writing solutions.
  - Proposal from some sites to do tape system simulation to evaluate possible optimization techniques.
- Starting with FZK Tier-1
  - File grouping in FZK HPSS tape system
    - Files are grouped by dataset — file is assigned to a file family (FF) based on its dataset info (dataset name and size). FF number (integer) is calculated on the fly and reused within the same VO (experiment). When flushing to tape cartridges, HPSS wraps files from the same directory (same dataset) into aggregates (to a certain size limit). 1 tape drive per FF in archiving.
    - When reading from tape, the entire aggregate is recalled (Full Aggregation Recall mechanism)
    - More details please refer to FZK talk and site experts.
  - Dedicated tape tests using ATLAS RAW/AOD/DAOD data samples individually, ~100TB each, in a relatively clean environment (production recall is blocked, a small portion of recalls from user jobs in parallel).
- Other sites are welcome to join this demonstrator.

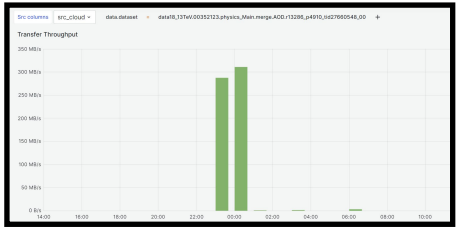
Tape smart writing HL-LHC demonstrator (2/2)

- Preliminary results from the FZK exercise
  - Very good performance, factor of two improvements on throughput per tape drive, over their old TSM tape system (all use TS1160 tape drive w/ 400MB/s nominal rate)

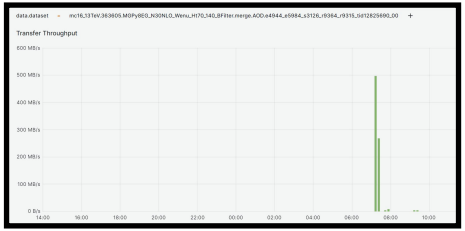








Transfer rate for a 2TB/295 files AOD dataset



Transfer rate for a 466GB/187 files AOD dataset

FZK HPSS monitoring (*courtesy of Haykuhi Musheghyan from FZK*)

ATLAS DDM dashboard

Metadata of data grouping on tape

- In order for sites to group files on tape, experiments like ATLAS need to provide grouping hints as metadata when writing files to tape
- Current status
  - To pass the metadata (Rucio → FTS → site SE)
    - Temporary solution in place for FZK Tier-1, using URL parameters in transfer command to pass simple metadata (dataset name & dataset size)
    - Long term solution to pass metadata in json in the http header
      - Archival metadata supported in FTS
  - Metadata format
    - A flexible format proposed by CTA/dCache team

Archive metadata proposal example for DAQ file

File: `data23_13p6TeV.00452799.physics_Main.daq.RAW._lb0777._SFO-19._0001.data`

```json
archive_metadata = {
    "scheduling_hints": {
        "archive_priority": "100"          # highest priority
    },
    "collocation_hints": {
        "0": "data23_13p6TeV",                                  # project
        "1": "RAW",                                             # datatype
        "2": "00452799",                                        # runnumber
        "3": "data23_13p6TeV.00452799.physics_Main.daq.RAW",    # dataset
    },
    "optional_hints": {
        "activity": "T0 Tape",          # Tier-0/DAQ
        "3": {                          # dataset level
          "length": "19123",                # total number of files at specified level
          "bytes": "80920799318456"         # total size of files at specified level
        }
    }
}...
```

Sample template from the CTA/dCache proposal

- What metadata to pass ?
  - ATLAS has provided a ggdoc describing ATLAS data access pattern to tape
  - Internal discussion ongoing about the hierarchy of the metadata to provide
    - dataset is a good starting point (as the basic grouping unit)
    - Above dataset levels ? project name/data type/production step/container/request ID/…
    - What sites want ? prescriptive vs descriptive ?
  - Site inputs are welcome and appreciated, we will continue to work out details with sites.

Tape@Tier-2

- DESY Tier-2 has offered to host tape outside of their pledge
  - DESY-HH_MCTAPE has been integrated into ATLAS DDM
    - ~2PB
    - Will be used for additional replicas of EVNT and HITS Monte-Carlo data.
- More development is needed in ATLAS DDM
  - RPG (Replication Policy on the Grid) machinery
  - Updates in Rucio subscriptions
- Other Tier-2s such as NET2 may also contribute tape later