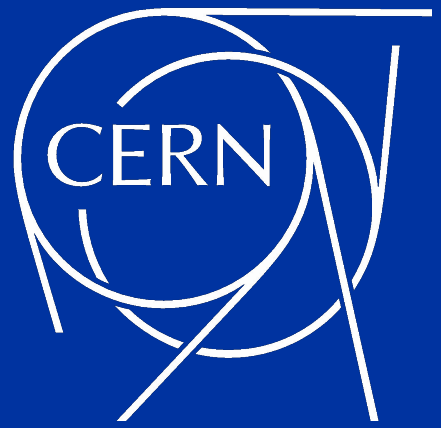


CERN Tape Archive status and plans

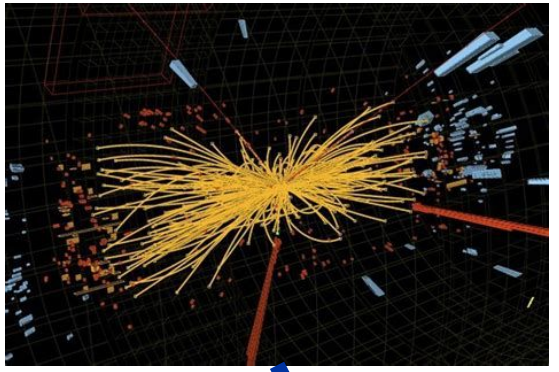
Julien Leduc
on behalf of the CTA team

2025-12-10

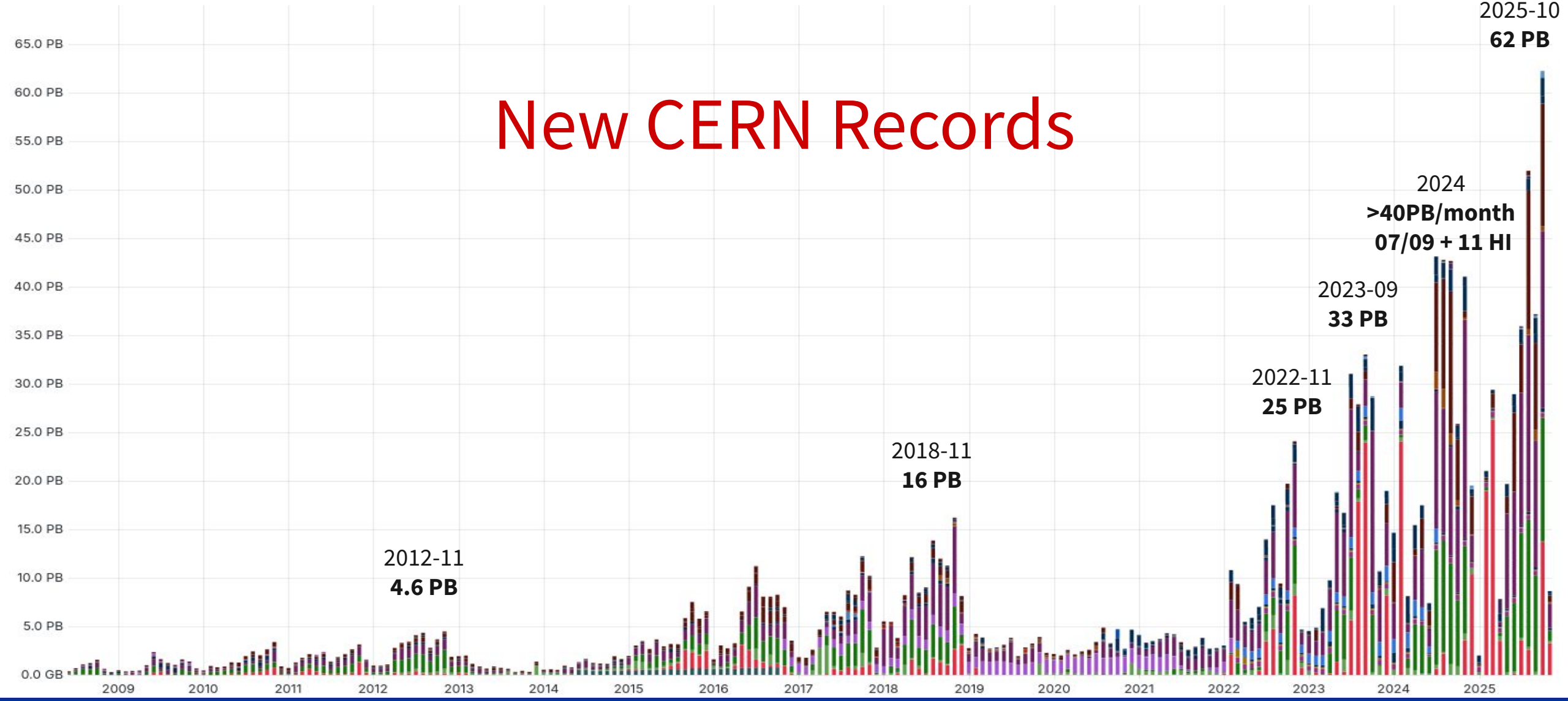
What storage needs/constraints do you face or your existing storage solution address?

What criteria are used to choose it over other available solutions?

CERN *oversimplified*



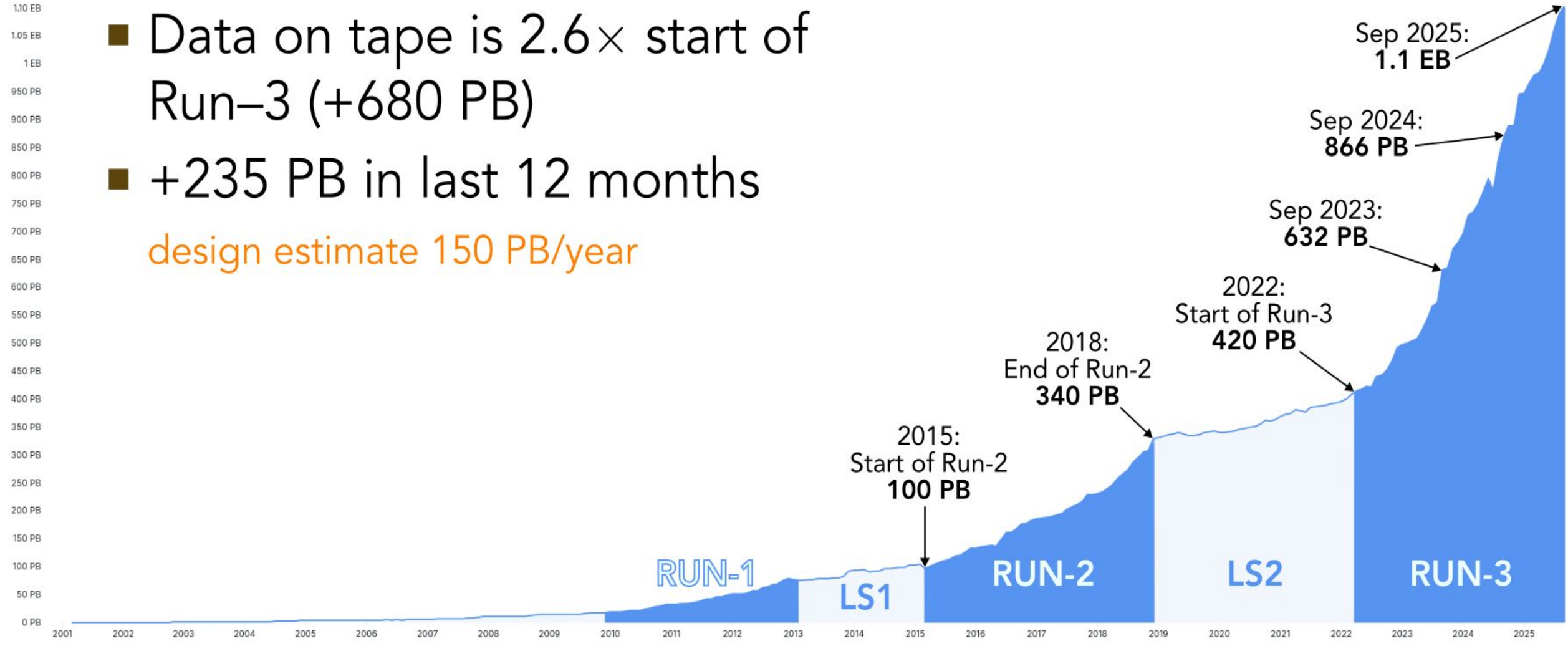
Monthly written data over the past 17 years



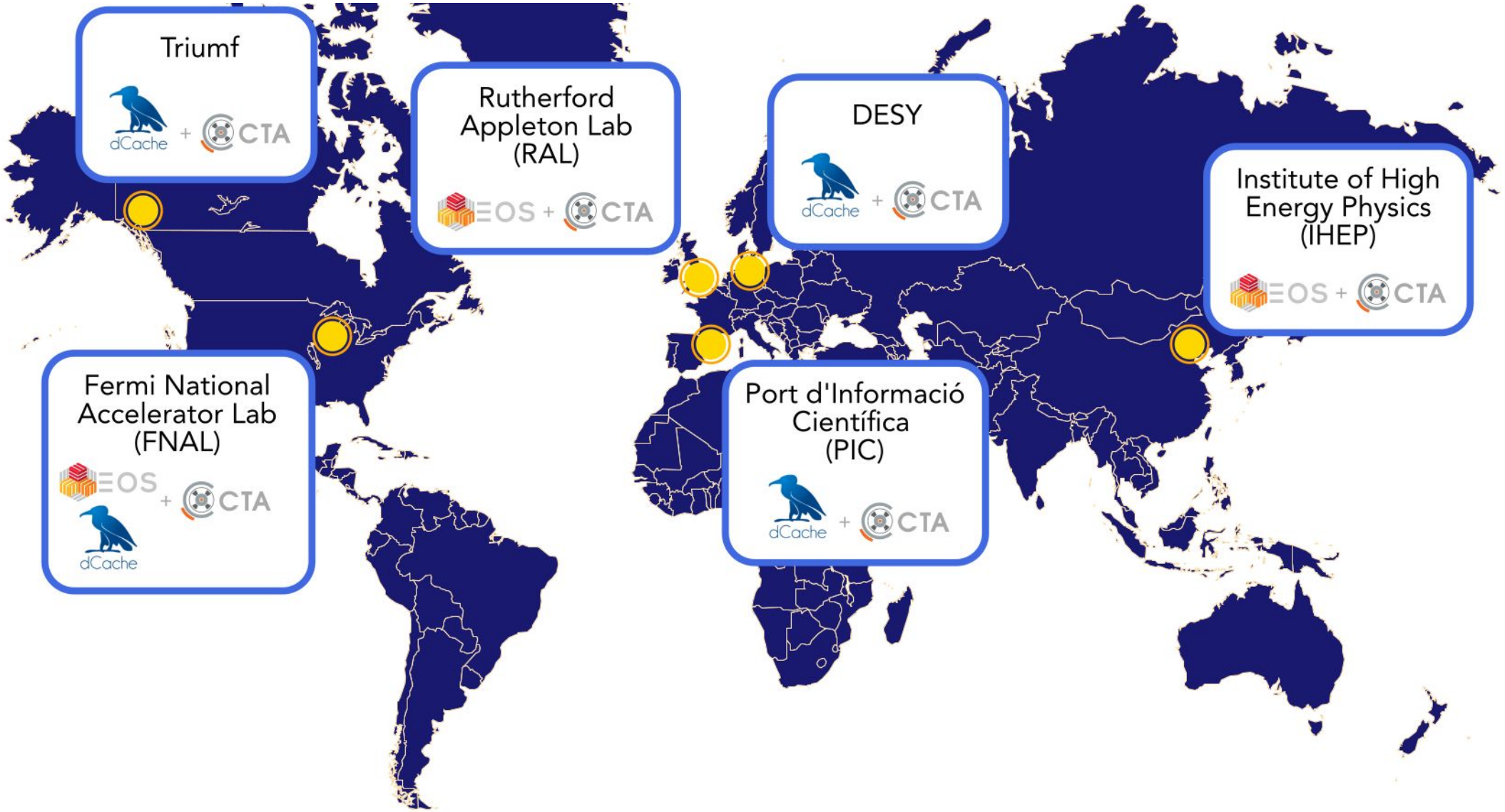
Tape namespace statistics

- Data on tape is 2.6× start of Run–3 (+680 PB)
- +235 PB in last 12 months

design estimate 150 PB/year



CTA Community



What are the advantages and disadvantages of your storage solution?

What does it do well, and where does it need improvement, operationally-speaking?

LHC Run-3 (2022-2026): EOS + CTA (tape buffer)

EOS is experiment facing:

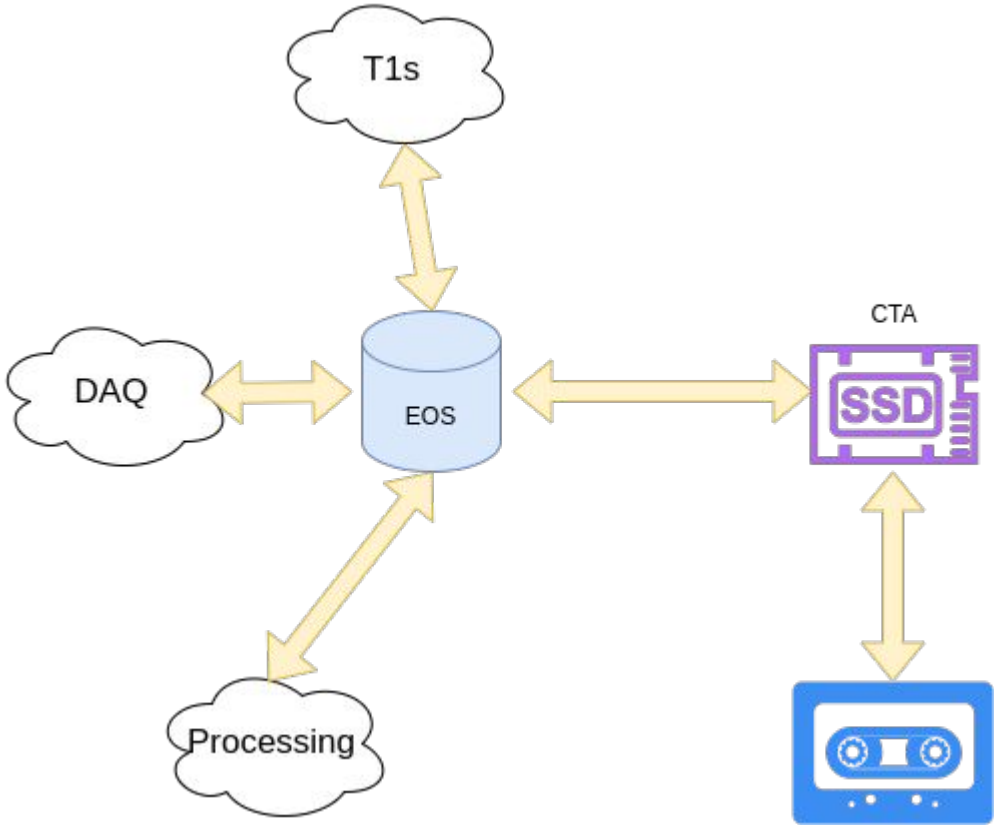
- **one Namespace per experiment**
  - **contains all disk capacity for an experiment: huge bandwidth as a by product**
  - **For example:**

CTA is a pure tape endpoint

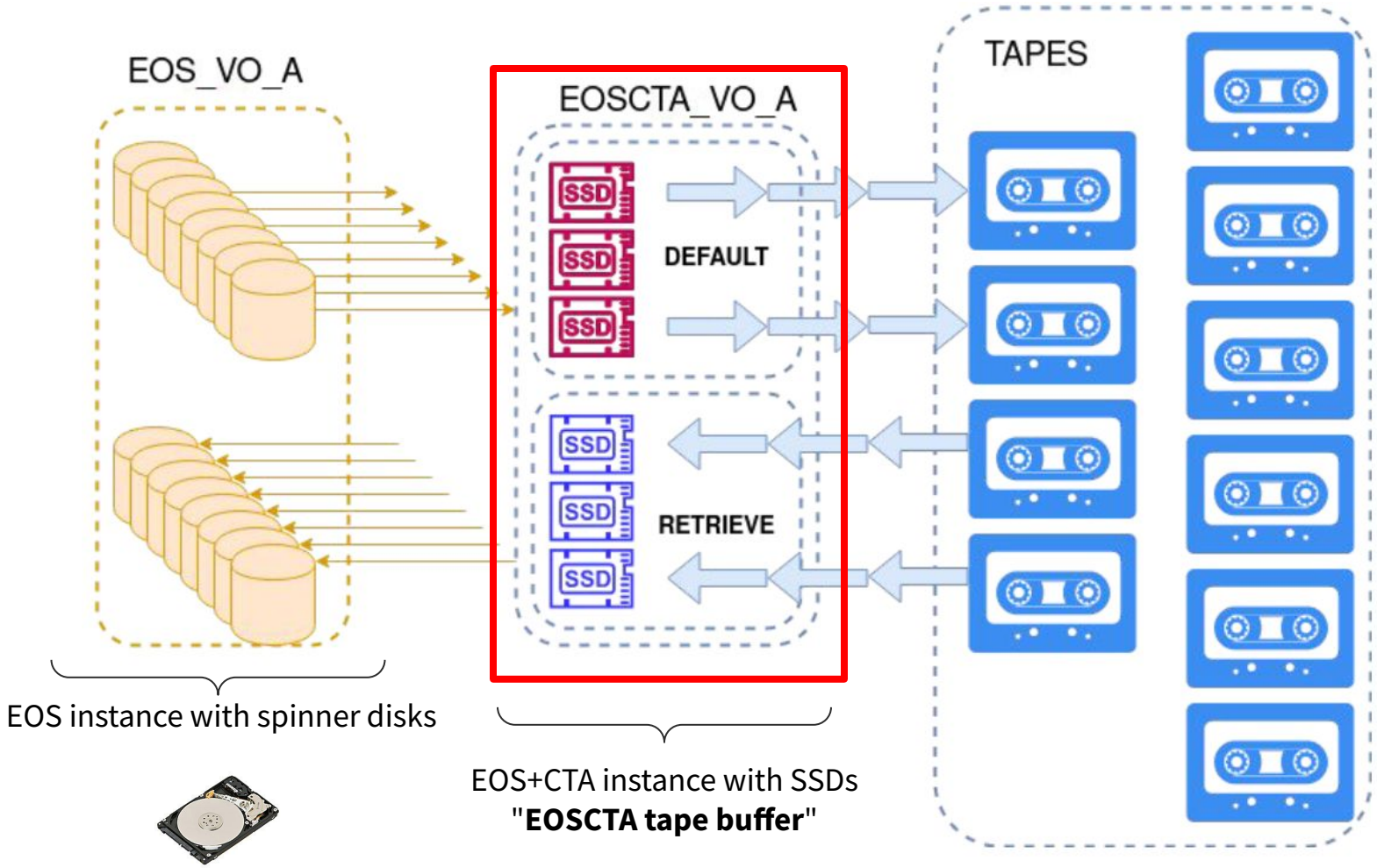
- **Shared tape catalogue for all tapes and their file index**
- **Tape drives fed from SSD buffer outside of pledge**

Experiment larger files 10GB per file and drive increased throughput prevents any meaningful in memory buffering.

Transition toward Run-4 requirements (o(100GB/s), o(100GB) perf file) required to move to this CTA architecture for Run3.



EOS + CTA architecture @ CERN



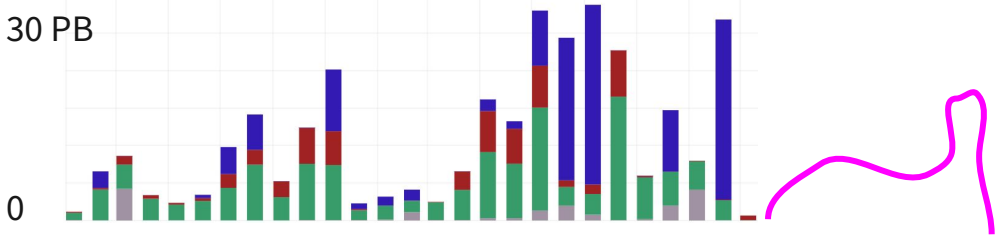
Archive/Staging bandwidth allocation

Standard LHC eoscta instance: 10GB/s archival SLA for CTA T0

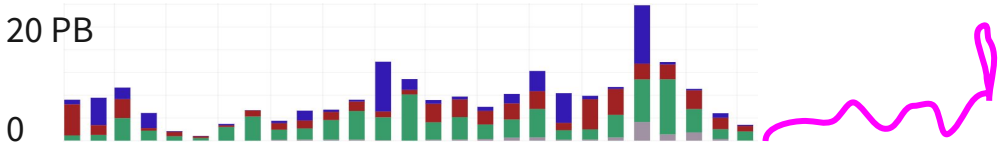
- **Archive boost needed during data taking, tape flushing, Heavy Ion run**
- **Staging boost during Year End Technical Stop (YETS) Heavy Ion data duplication to T1s/T2s**
- **Change eoscta bandwidth allocation accordingly**

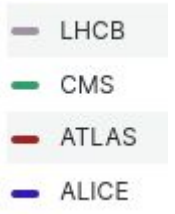
Run3 LHC planning

Archived volume per month



Staged volume per month



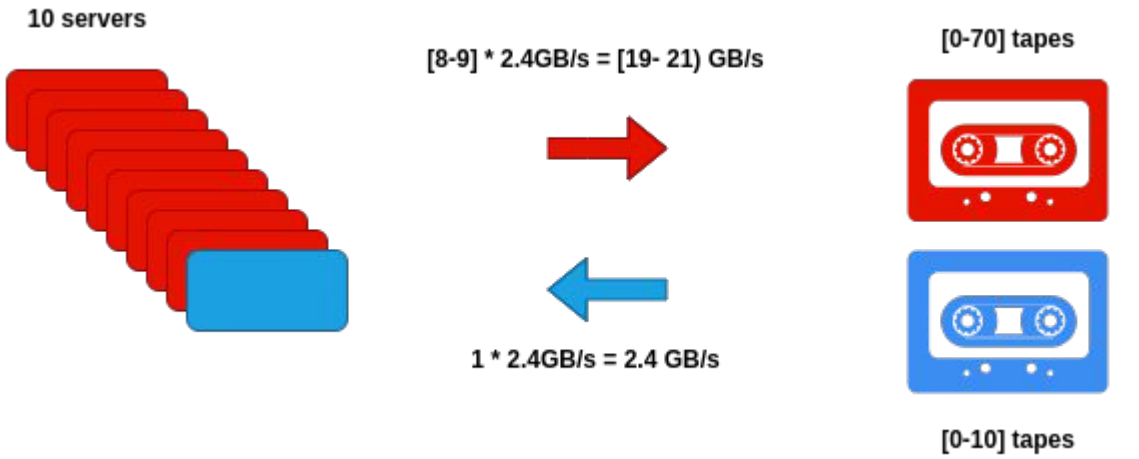


AI/quantum computing generated interpolation

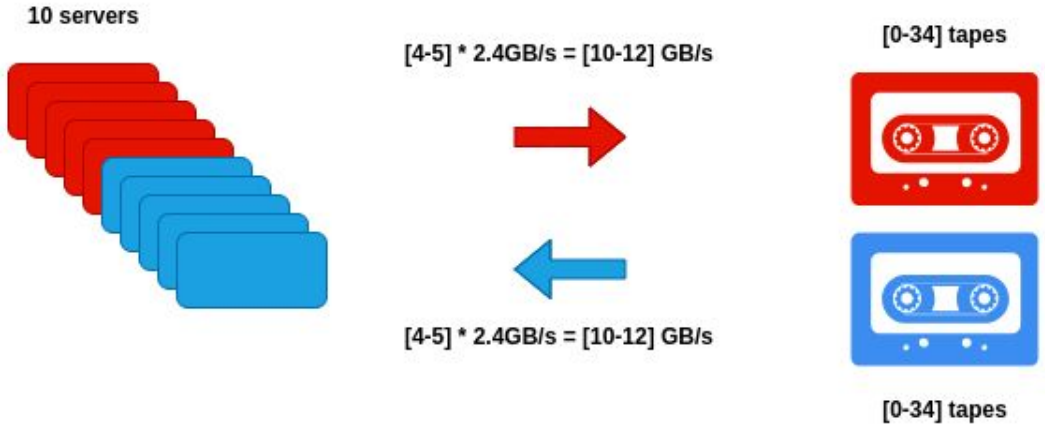
Archive/Retrieve bandwidth allocation

Standard LHC eoscta instance: 10GB/s archival SLA for CTA T0

- **10 SSD servers**
- **Archive boost during data taking, tape flushing, Heavy Ion run**
- **Staging boost during Year End Technical Stop (YETS) HI data duplication to T1s/T2s**
- **Configure CTA ALICE VO writemaxdrives, readmaxdrives accordingly**
- **Measure bandwidth to/from tape buffer AND INDIVIDUAL TAPE DRIVE EFFICIENCY**

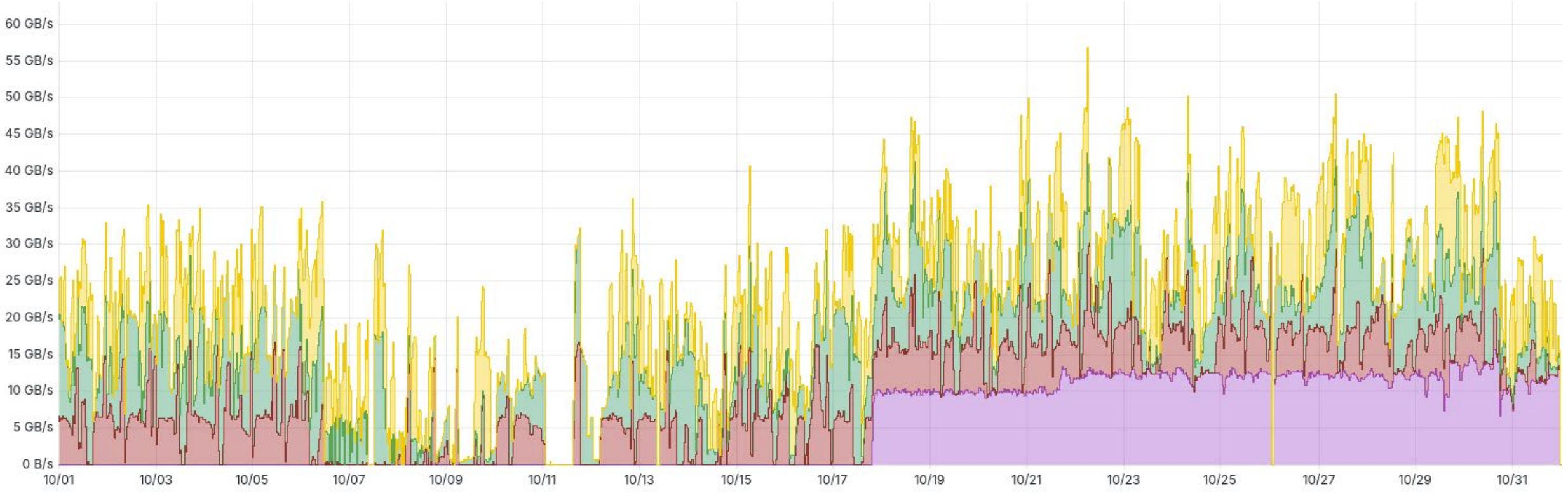


Data taking configuration

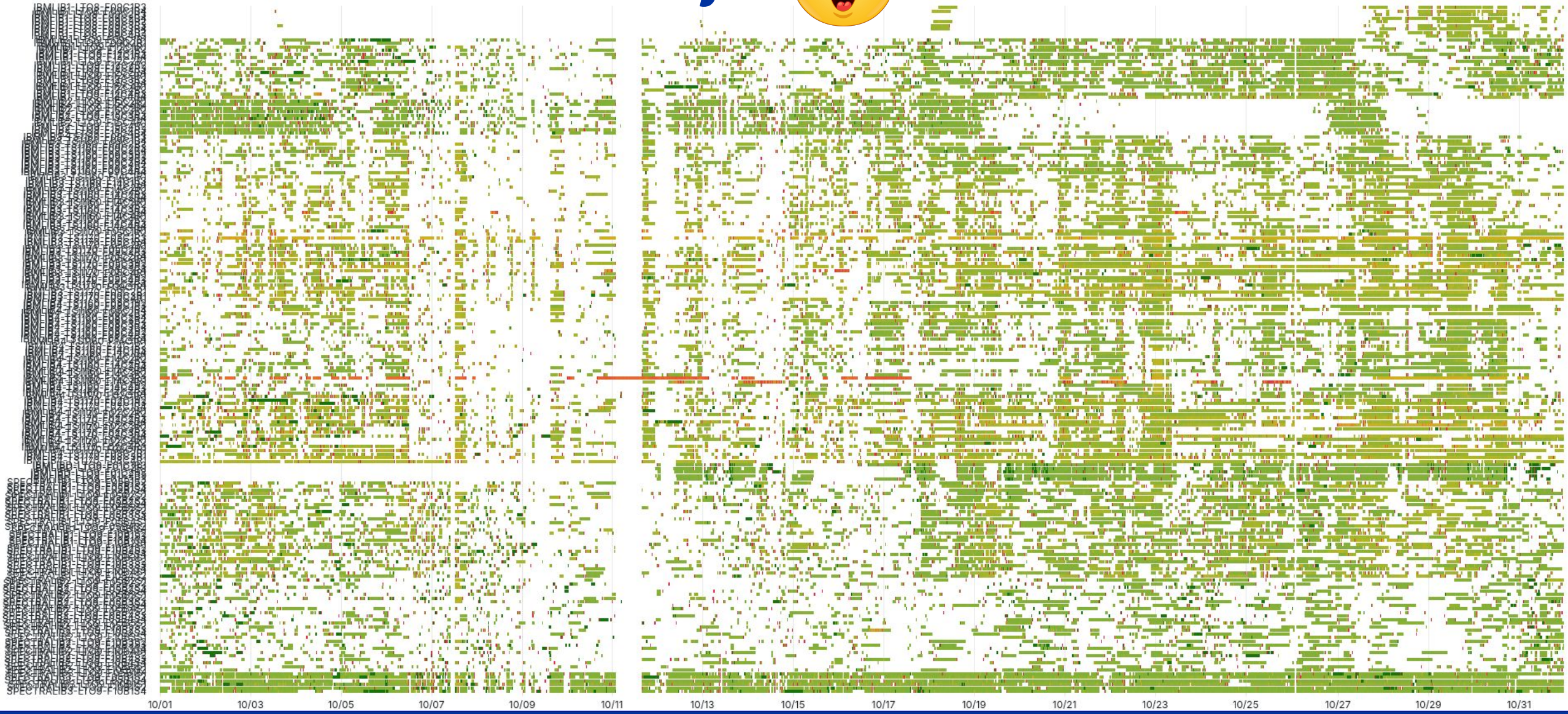


YETS configuration

2025-10 Cumulated archive speed



2025-10 Archive efficiency



What storage challenges do you foresee in the near future and does your solution scale up to meet these challenges?

If not, what upgrades/modifications are planned?

Run-3 traffic evolution

- **Beginning of Run-3: improve tape write efficiency**
  - 250 PB archived in last 12 months (+45% compared to 2024)
- **End of Run-3: more reads from tape**
  - YETS 2024 work on tape activities and priority mapping using Staging Metadata at T0
  - 170 PB retrieved in last 12 months (+100% compared to 2024)

DATA CAROUSEL IS COMING WRITE/READ RATIO IS EVOLVING

- **TAPE: CAPACITY / DISK: CACHES FOR PROCESSING AND TRANSFERS**
  - Clear goal: work on tape read efficiency improvements

HTTP Staging metadata: prioritize tape reads

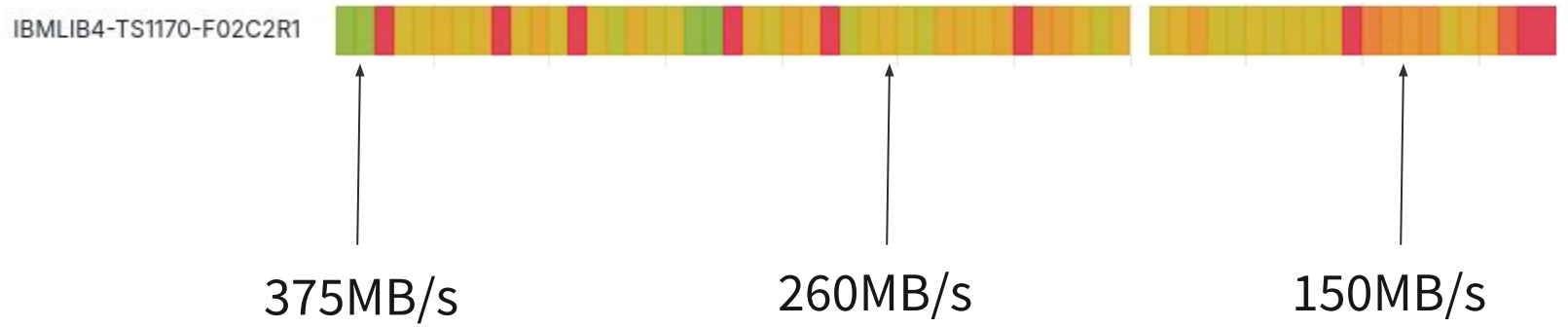
Prioritize rucio tape reads depending on rucio activity

Initial Tape REST API implementation was too complex to configure, impractical

- **Simplified by implementing default priority**
  - FTS team updated Tape REST API to include default metadata
  - Rucio 37.0 Dungeons & Donkeys released in 2025-04
  - EOS 5.3.9 released in 2025-05
- **Provided guidance to ATLAS and CMS to re-label tape activities for data carousel model**
- **Rucio Activity is now mapped to three tape priorities:**
  - **High**: Interactive T0 staging (“express”)
  - **Default**: Default staging priority
  - **Low**: Best-effort long-running background staging campaigns

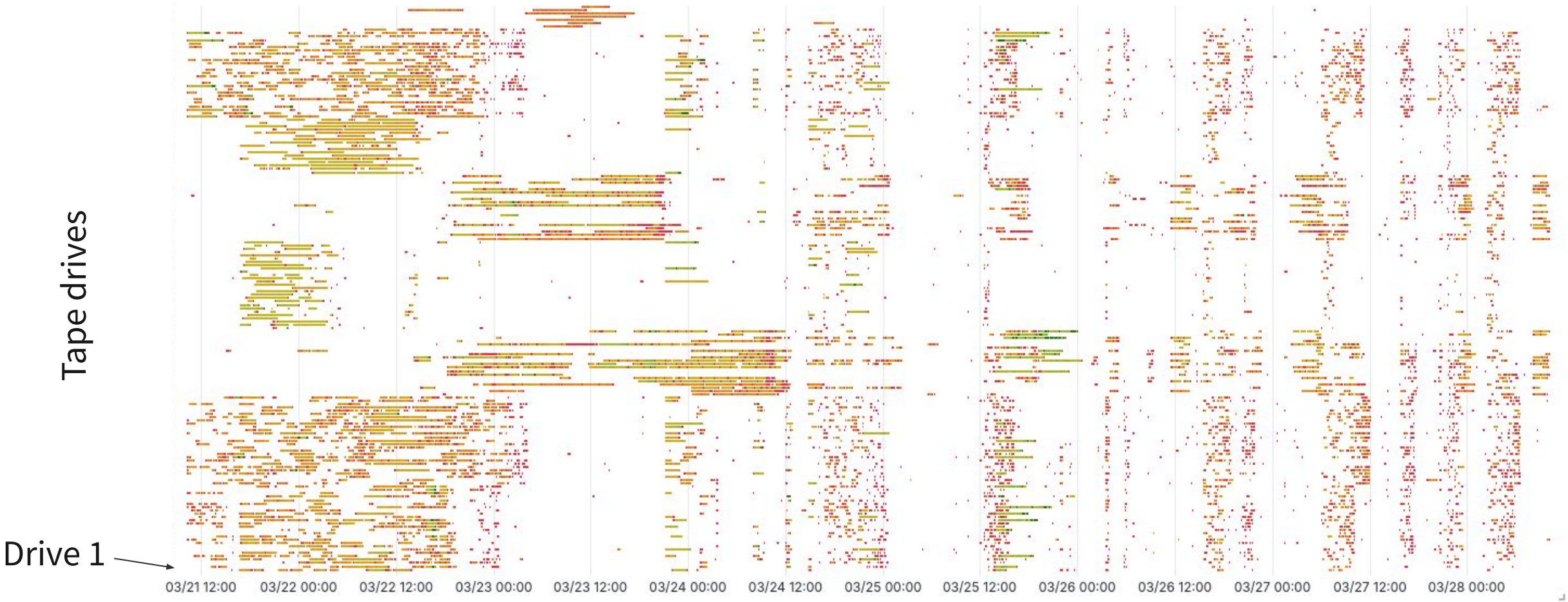
User Staging efficiency

- **Single long enough tape session efficiency is more or less OK**
  - Various tricks like RAO help there



- **Tape drive efficiency in practice**
  - 3 minutes mount and position, read, 3 minutes rewind and eject
  - achieving 80% drive efficiency on reads require at least 24 minute long mount
    - minimize #tapes / dataset
  - On top: minimize seeks to maximize read throughput
    - low entropy per (dataset, tape)

Production staging over 7 days



Archive Metadata: monitor data placement

- **Coordinated development and tests to pass per-file Archive Metadata across the full chain**
  - Rucio → FTS → EOS → Monitoring
- **Archive Metadata collection started with ATLAS during 2024 HI run**
- **End of Run-3:**
  - CTA: use experiment provided per file Archive Metadata to monitor rucio dataset placement on tape
  - Experiments: refine Rucio -> Archive Metadata translation
    - Data popularity analysis, data carousel evaluation, per-datatype AM tree definition, data management strategies,...

LATER: IMPROVE PLACEMENT WITH AM DRIVEN TAPE SCHEDULER DURING LS-3 USING RUN-3 AM COLLECTION TO SIMULATE DATA PLACEMENT

Archive Metadata collection in practice

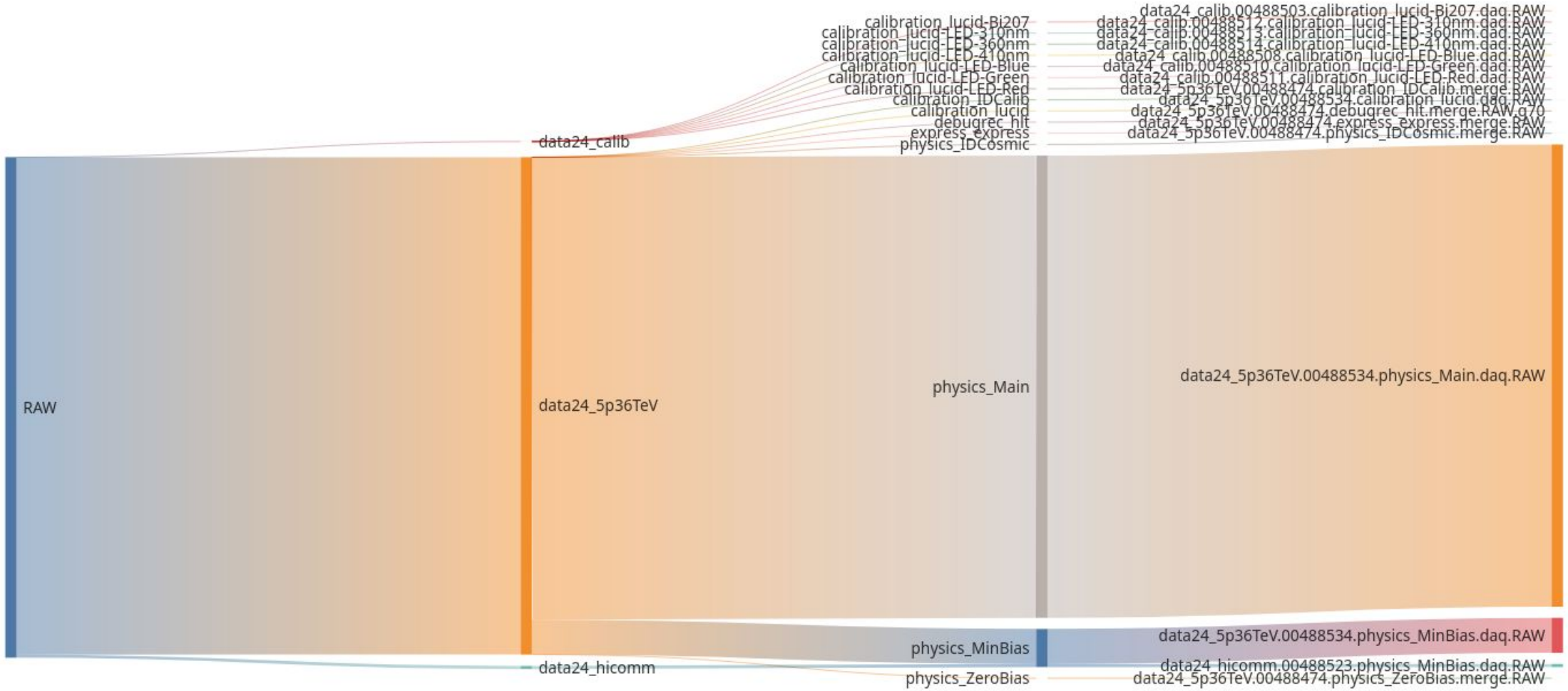
- **How this works in practice?**
  - Rucio extracts and generates Archive metadata via a Rucio plugin
    - Base Archive Metadata plugin released in Rucio 34.0
    - ATLAS specific plugin implemented and deployed in production via *atlas-rucio-policy-package*
  - Passes file specific AM for every FTS transfer where destination **RSE is tape AND RSE property archive_metadata=True**
    - in *–archive-metadata* FTS option
  - eosctaatlas instance receives and collect ArchiveMetadata for later tape placement analysis
    - in eosreport *tape_create* lines

log=xxxxxxxx-xxxx-xxxx-xxxx-xxxxxxxxxxxx&path=/eos/ctaatlas/archive/grid/atlas/rucio/raw/data24_hi/physics_MinBias/00489961/data24_hi.00489961.physics_MinBias.daq.RAW/data24_hi.00489961.physics_MinBias.daq.RAW._lb0117._SFO-19._0001.data&ruid=10763&rgid=1307&td=atlas003&host=&ts=1732060983&tns=61902751&fid=433541814 7&fxid=102693b23&eos.btime=1&archivemetadata=eyJjb2xsb2NhdGlvbl9oaW50cyI6IHsiMCI6ICJSQVciLCAiMSI6ICJkYXRhMjRfaGkiLCAiMiI6ICJwaHlzaWNzX01pbkJpYXMiLCAiMyI6ICJkYXRhMjRfaGkuMDA0ODk5NjEucGh5c2ljc19NaW5CaWFzLmRhcS5SQVcifSwgImFkZGl0aW9uYWxfaGludHMiOiB7ImFjdGl2aXR5IjogIlQwIFRhcGUiLCAiMyI6IHsibGVuZ3RoIjogMTMwLCAic2l6ZSI6IDYxNDE3NTM1NTZ9fSwgImZpbGVfbWV0YWRhdGEiOiB7InNpemUiOiAyMTQwNDQ4OCwgIm1kNSI6IG51bGwsICJhZGxlcjMyIjogImU0NzJiYjE1In0sICJzY2hlbWFfdmVyc2lvbiI6IDEsICJzY2hlZHVsaW5nX2hpbnRzIjogeyJwcmlvcml0eSI6IDEwMH19&sec.app=tape_create

Archive Metadata: schema_version 1

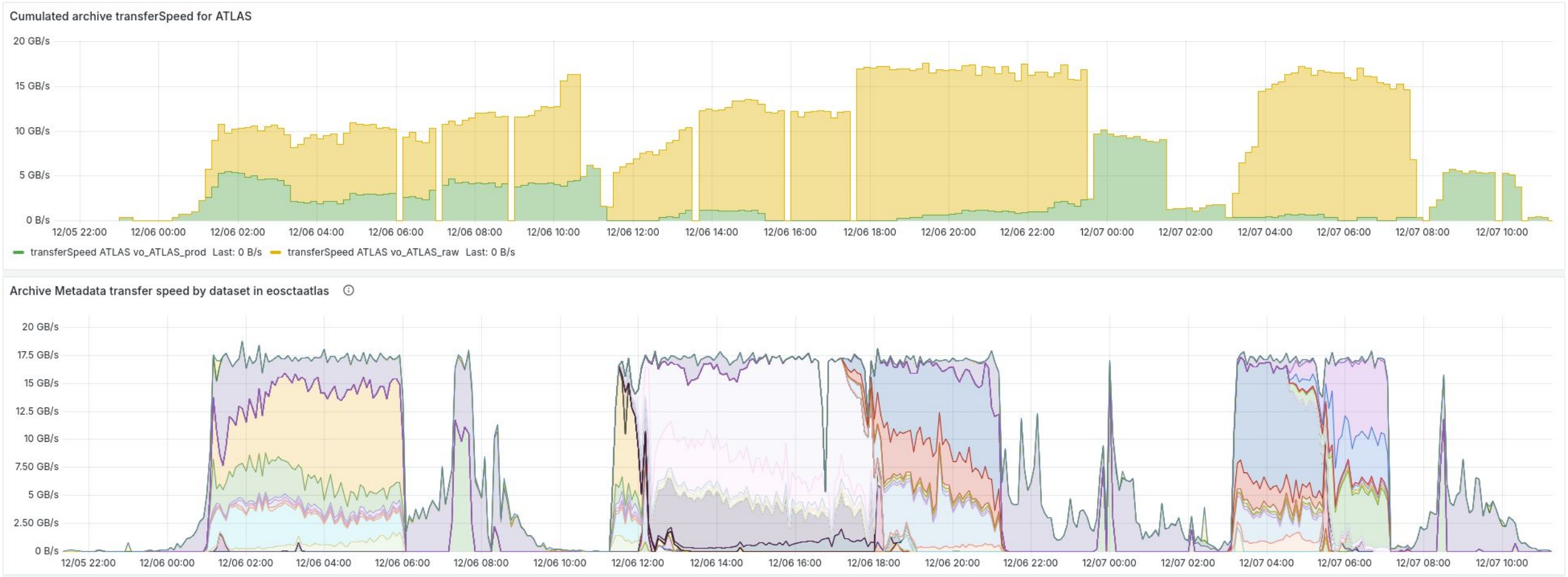
```
{
  "collocation_hints": {
    "0": "RAW",                                                  # datatype
    "1": "data24_hicomm",                                        # project
    "2": "physics_MinBias",                                      # [calibration|physics]_streamName
    "3": "data24_hicomm.00488523.physics_MinBias.daq.RAW"        # dataset
  },
  "additional_hints": {
    "activity": "T0 Tape",                                       # Tier-0/DAQ
    "3": {                                                       # dataset level
      "length": 340,                                             # fileCount at specified level
      "size": 69032083700                                        # fileSizeSum at specified level
    }
  },
  "file_metadata": {                                             # file metadata
    "size": 184019848,
    "md5": null,
    "adler32": "3509f7e2"
  },
  "schema_version": 1,                                           # ArchiveMetadata schema version
  "scheduling_hints": {
    "priority": 100                                              # priority 0-100
  }
}
```

Archive Metadata at T0



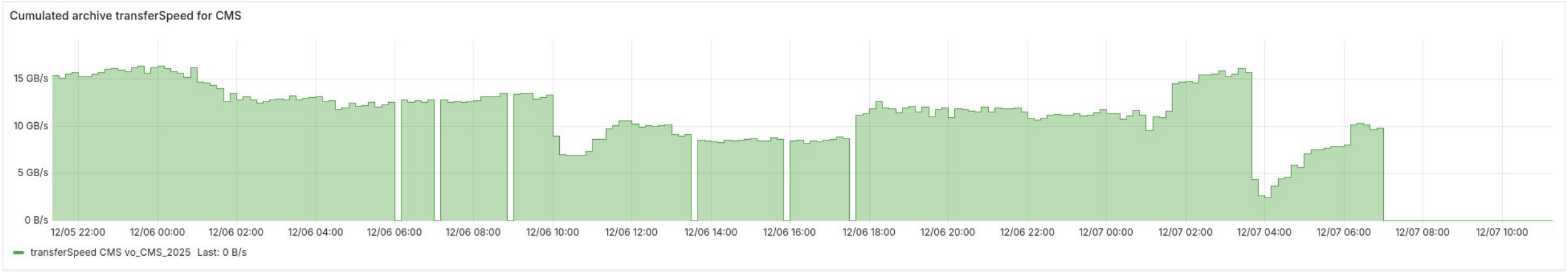
/0:datatype /1:scope /2:stream /3:ds

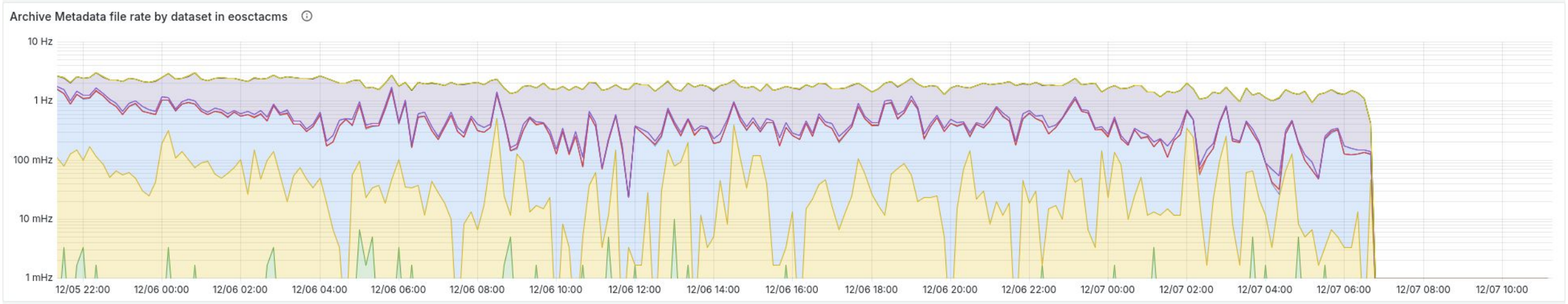
Live AM monitoring in Production during HI 2025



For all ATLAS data types

Live AM monitoring in Production During HI 2025



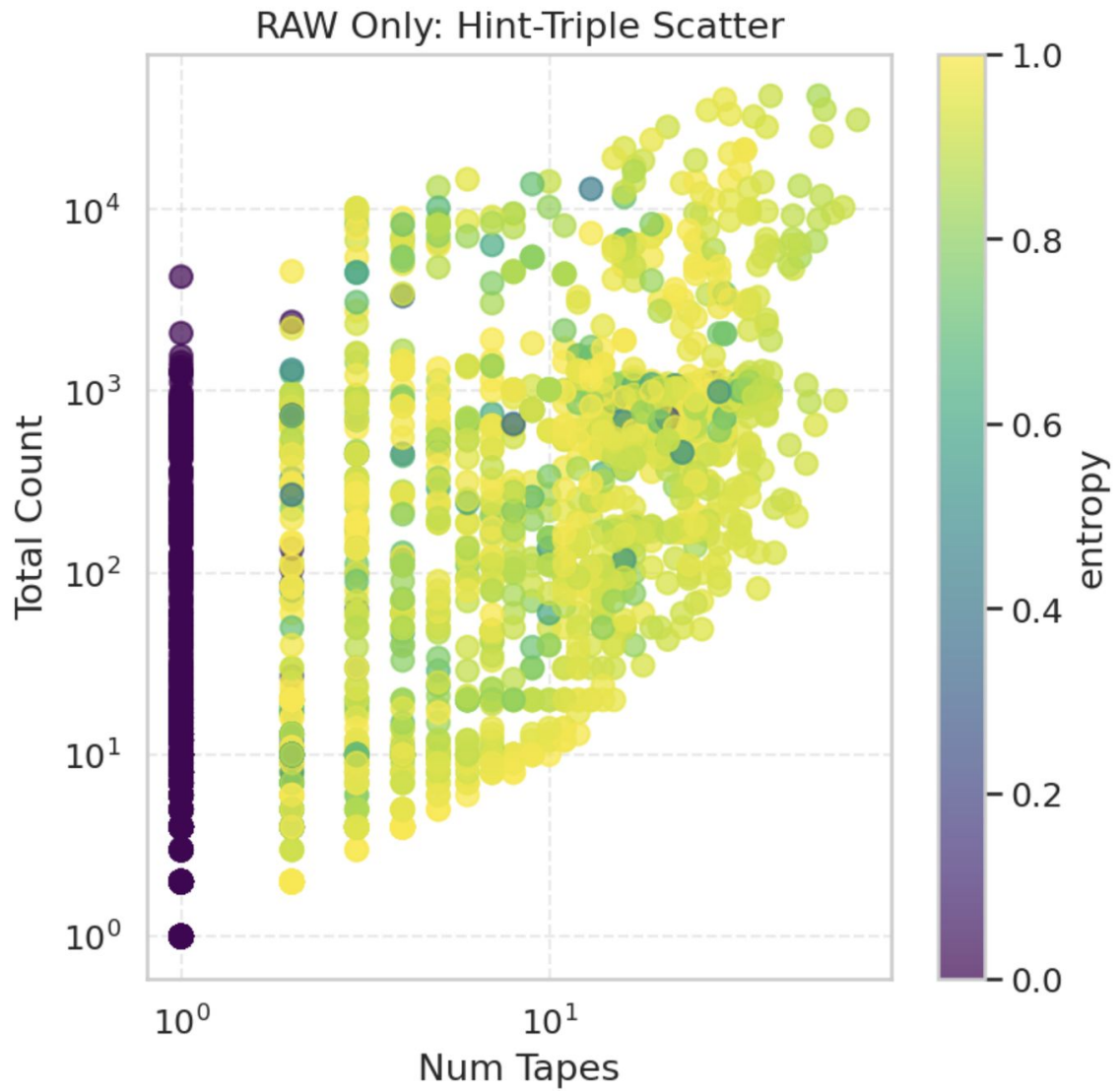


And for CMS too

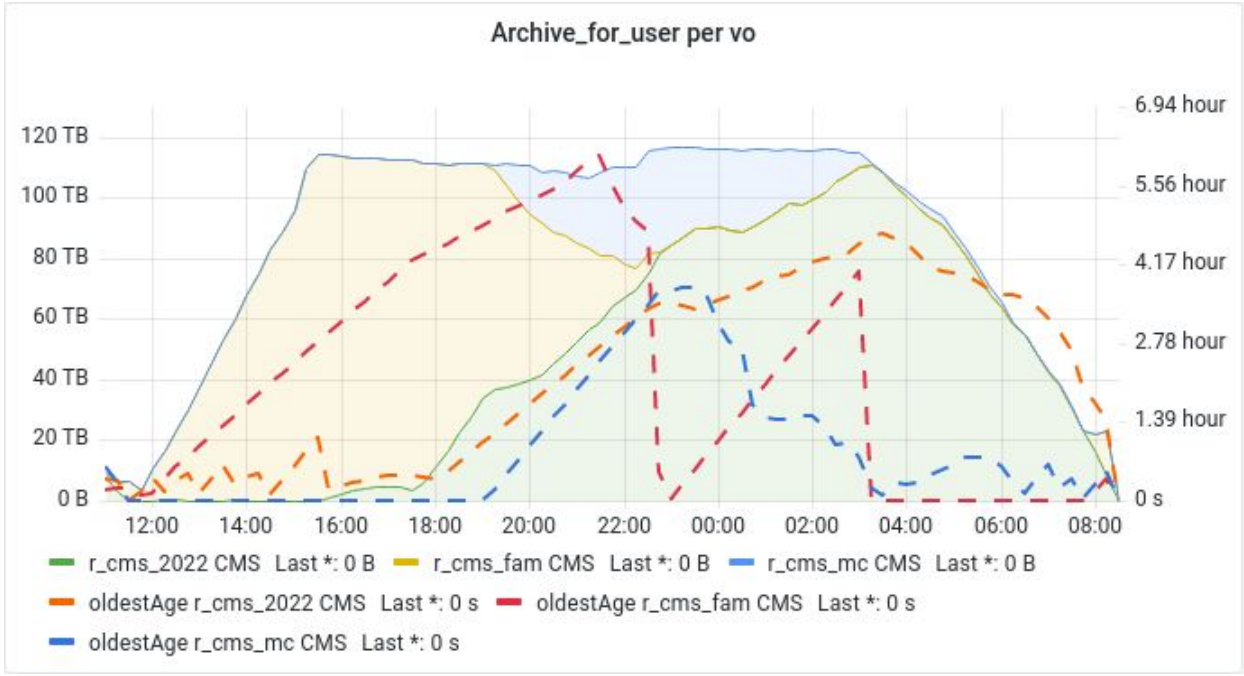
Current tape placement analysis

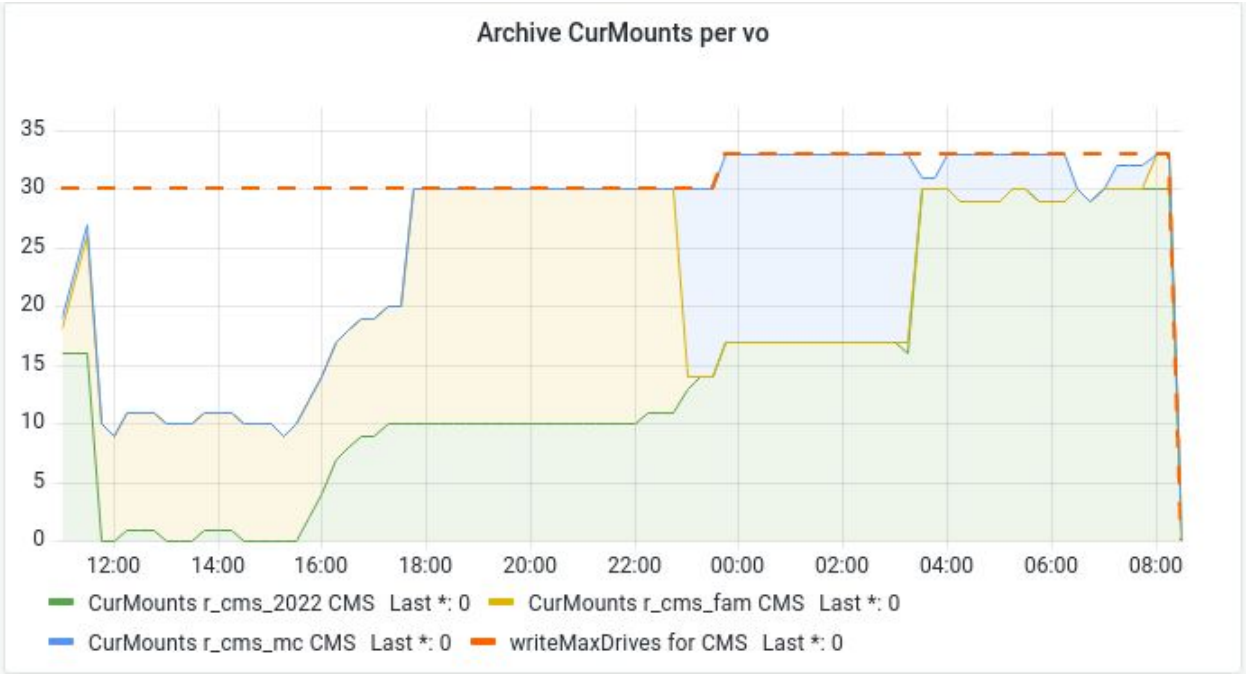
- **2025-08: ATLAS provided full rucio metadata dump**
  - translated into AM for all CTA ATLAS tapes for RAW
  - entropy measurement of ATLAS dataset file placement per tape count

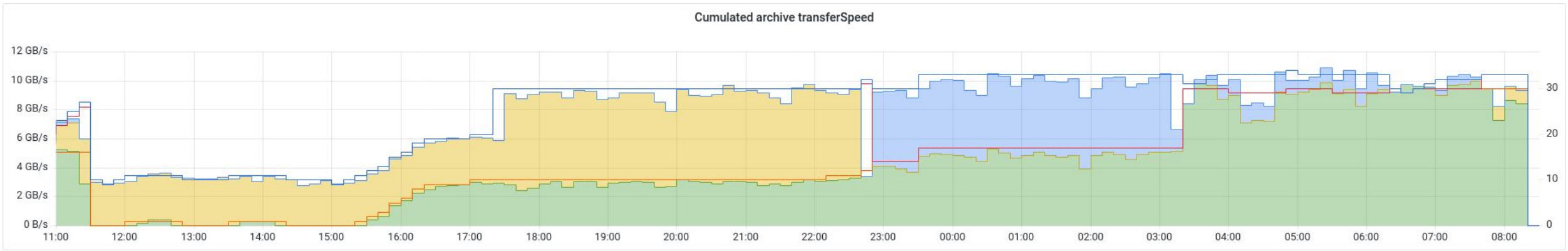
There is room for improvement



Archive Metadata: to backpressure low priority data

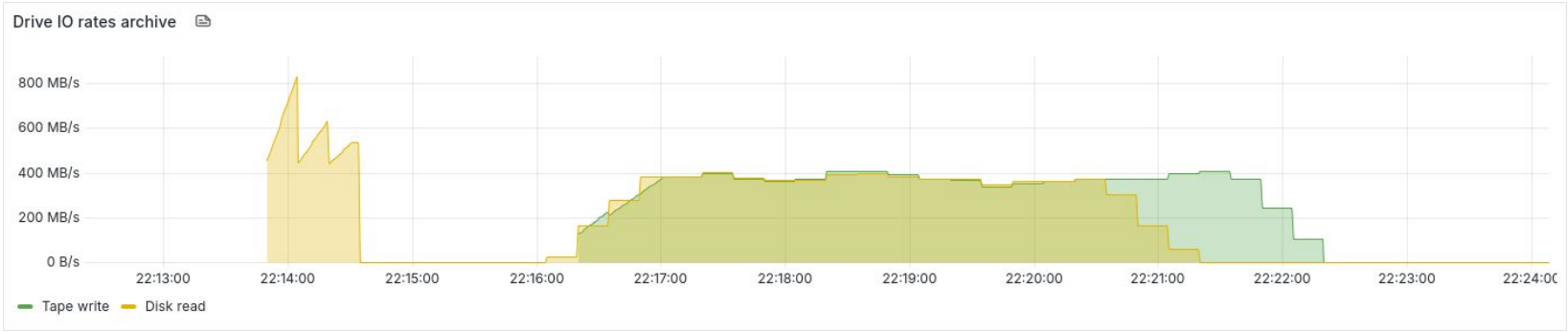




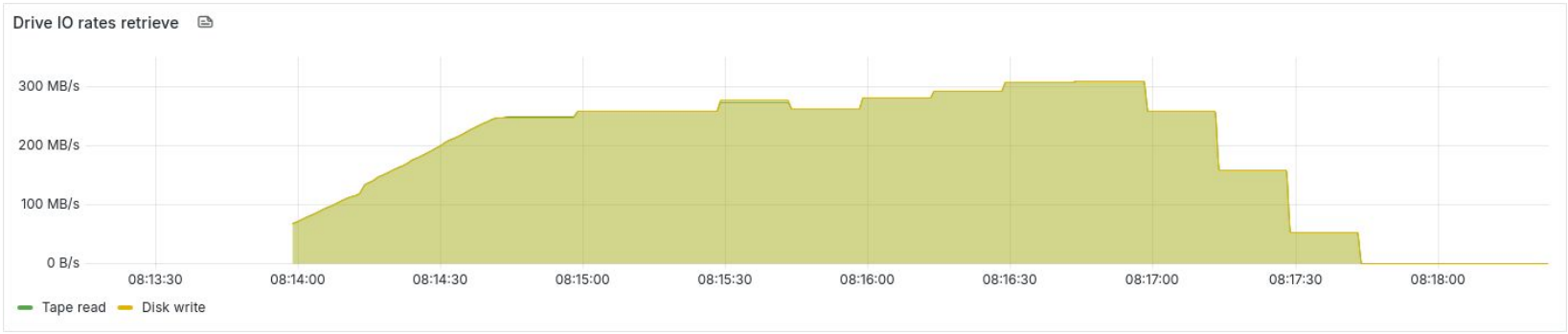


CTA internal improvements

- **Experimental feature: open telemetry metrics sent by CTA processes**
  - sample low level performance indicators and publish
  - examples:
    - archival separating disk read thread and tape write thread counters:



    - staging separating tape read thread and disk write thread counters:



Outlook

- **CTA delivers nominal archival performance for Run3 with significant efficiency improvements**
- **Archive Metadata will shape CTA for Run-4 for T0 and T1 specific needs**
  - continue work with experiment data management teams
  - continue work with CTA community and WLCG tape direction
- **Model tape staging performance on various tape models**
- **Collect and associate Archive Metadata with CTA files for interested VOs**

Use this collocation information to improve file tape placement for the next CTA T0 massive repack during LS-3

Use it for experiment data placement during Run-4