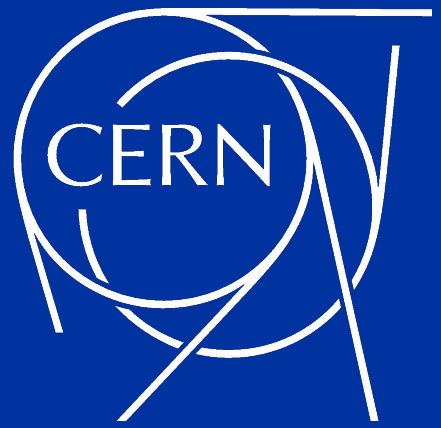


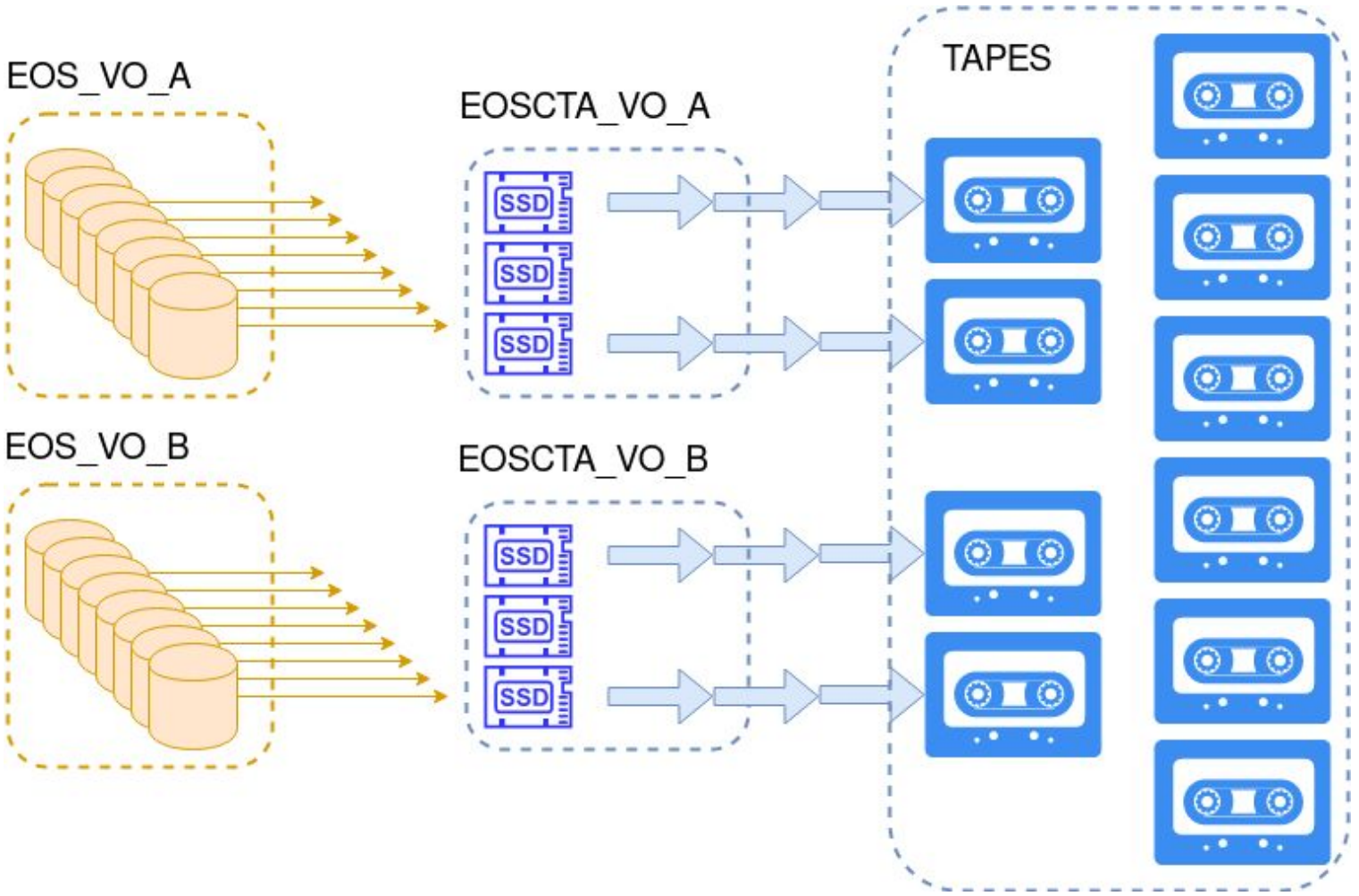
CERN Tape Archive status and plans

Julien Leduc
on behalf of the CTA team

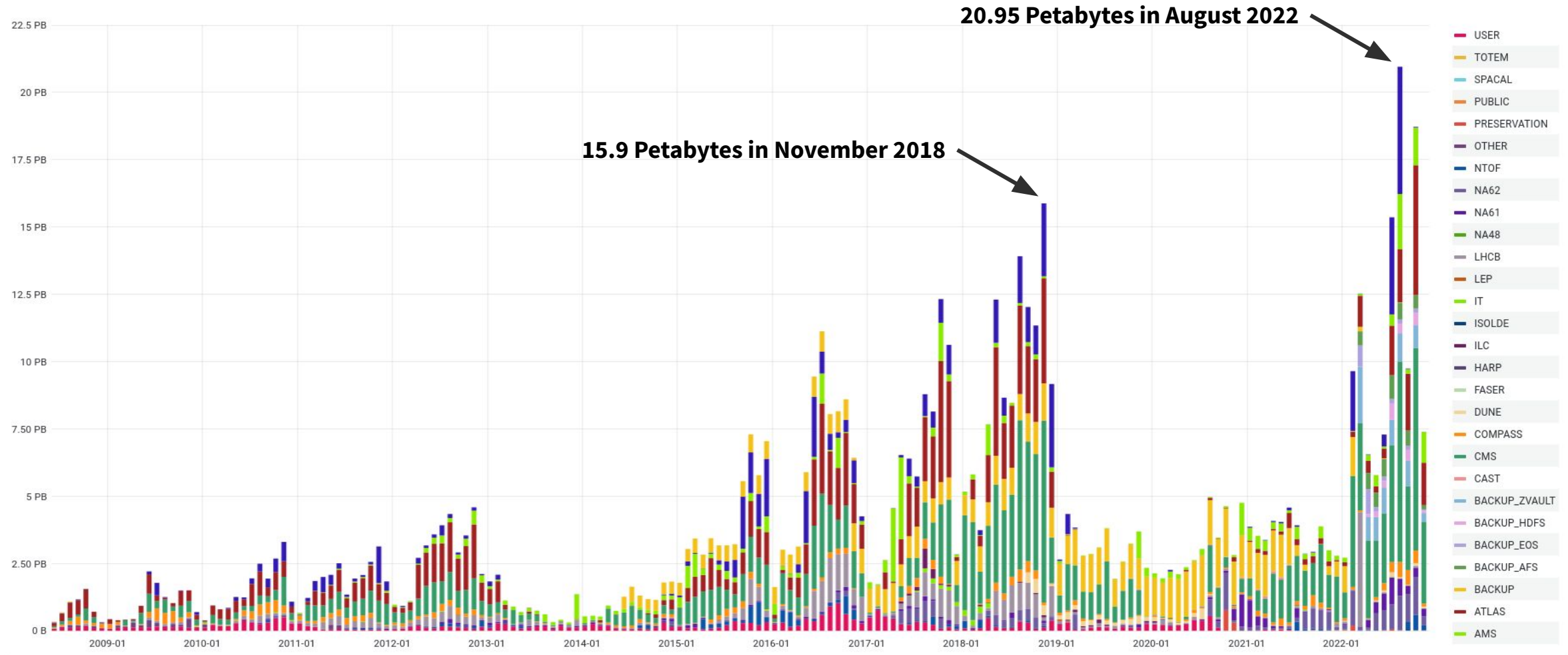
10/11/22

EOS+CTA Architecture

- **EOS+CTA is a pure tape system.**
- **Disk cache duty consolidated in main EOS instance.**
- **Operating tape drive at full speed full time efficiently requires a SSD based buffer: EOSCTA**



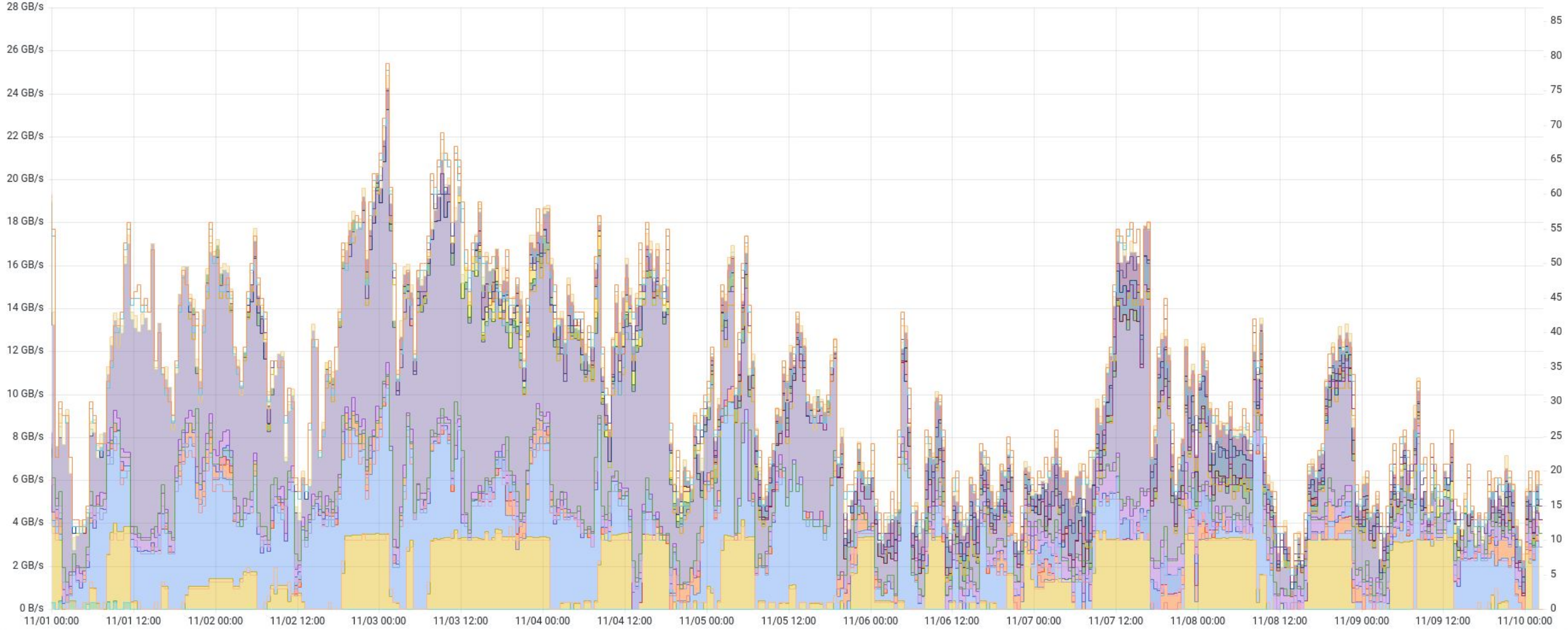
Beginning of Run3



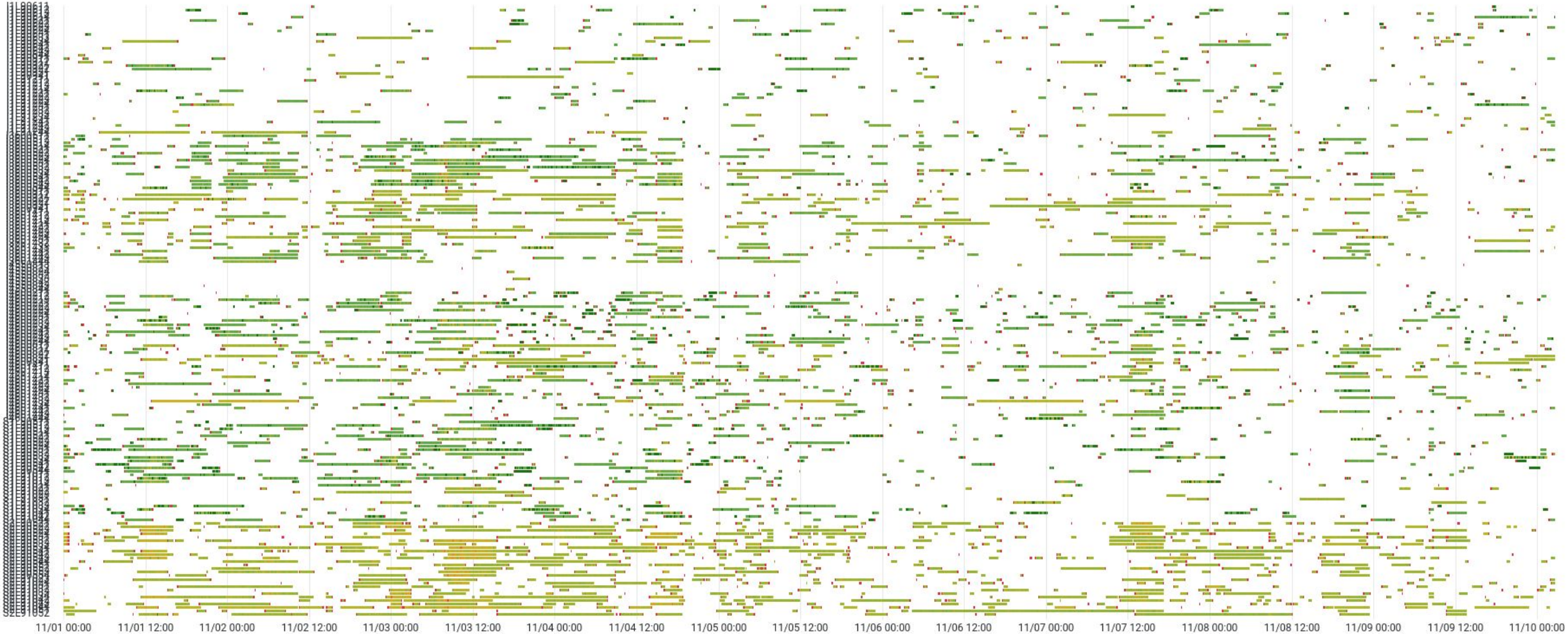
Beginning of Run3

- **Demonstrated archive performance:**
  - Nominal performance/efficiency as designed and tested during data challenges
    - Archival at 10 GB/s per LHC experiment consistently reached
    - CTA service wrote a new record of 20.9PB of data to tape in August 2022
- **Consolidation of tape service at CERN**
  - CASTOR service officially retired on 31/10/2022
  - CTA public rpms published on 27/10/2022
    - rpm installation instructions: https://eoscta.docs.cern.ch/install/install_public_rpms
- **Improving workflows with the experiments**
  - Publishing weekly namespace dumps for experiment data
    - What CTA service knows is safe on tape
    - **Compare with RUCIO NS?**
  - **Improve per directory tape collocation on tape**
    - **CMS split of MC, 2022 data**

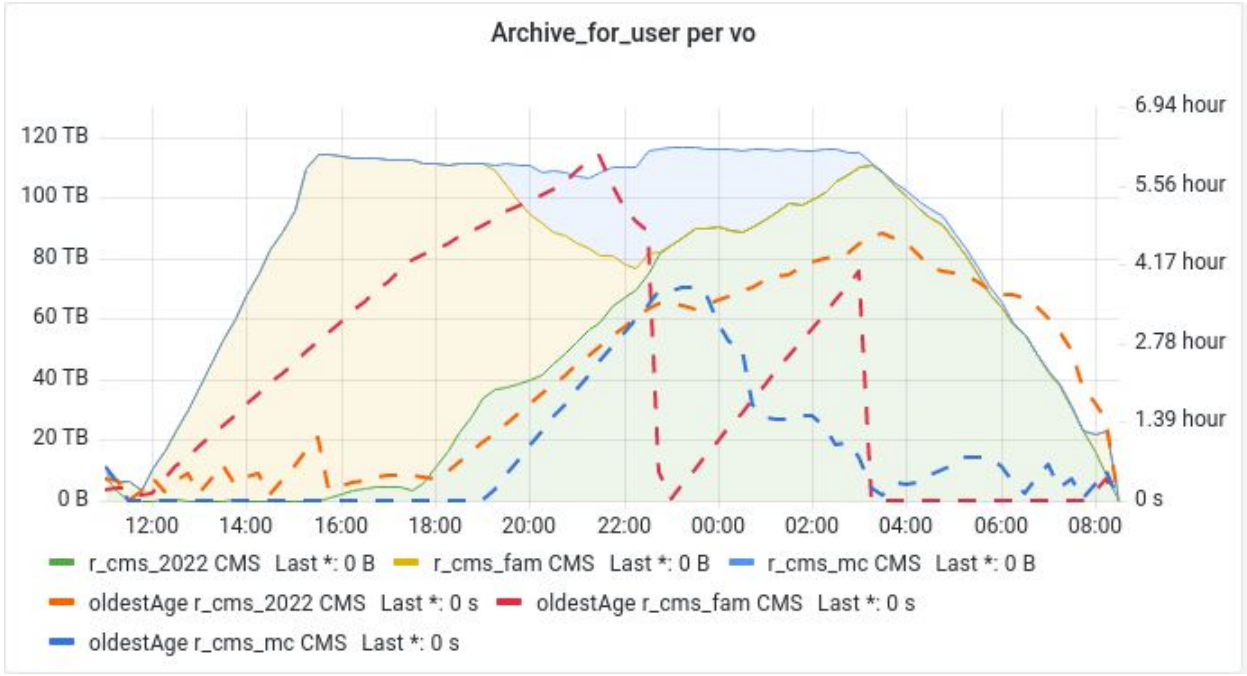
Cumulated archive speed this month

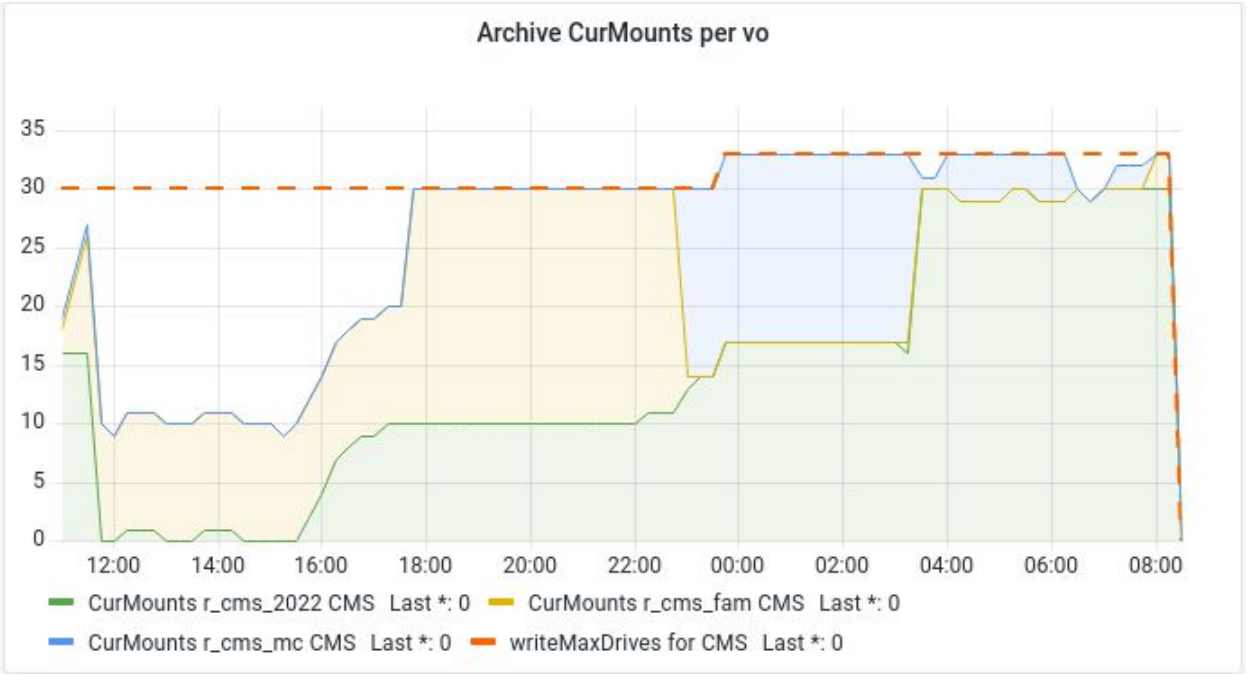


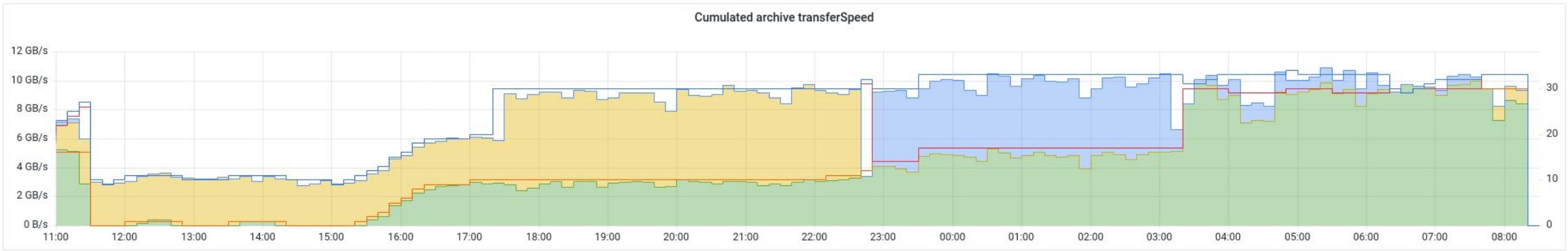
Archive efficiency this month



CMS archiving to multiple tapepools (aka tape families)







Short term plans

CTA service at T0

- **Deploy HTTP tape REST API by end of 2022 at T0**
  - Available outside CERN beginning of January
- **Deploy EOS5 and xrootd5 in CTA service**
  - CTA service will run the same version of EOS software as CERN EOS for physics service

Finalize/progress on some ongoing discussions

- **Storage endpoint accounting file for CTA service?**
  - ERROR is OK accorin
- **Interest / efforts in validating what is really on tape?**
  - Amount of dark data? Files on broken tapes?
  - CMS `Consistency Enforcement Toolbox`:
    - `xrdfs ls -l` is not enough for tape sites: should consider files that are on tape
- **Writing to tape: Destination file exists and overwrite is not enable #4447**
  - Use case still appearing
  - Use `Consistency Enforcement Toolbox` and CTA namespace dump to validate files on tape for failed transfers where dest file exists?

Long term plans: Improve tape read performance

Strictly mapping experiment directory structure to tape pools reaches some limitations:

- **Some type of data are orthogonal to experiment directory structure: CMS parking data for example**
- **Practical limitations of strict mapping**
  - 30 free tapes needed per tapepool…

Softer rules for file collocation on tape are needed

- **For example KIT file families prototype for ATLAS**
- **Requires additional metadata: dataset name, dataset total size, dataset file count**

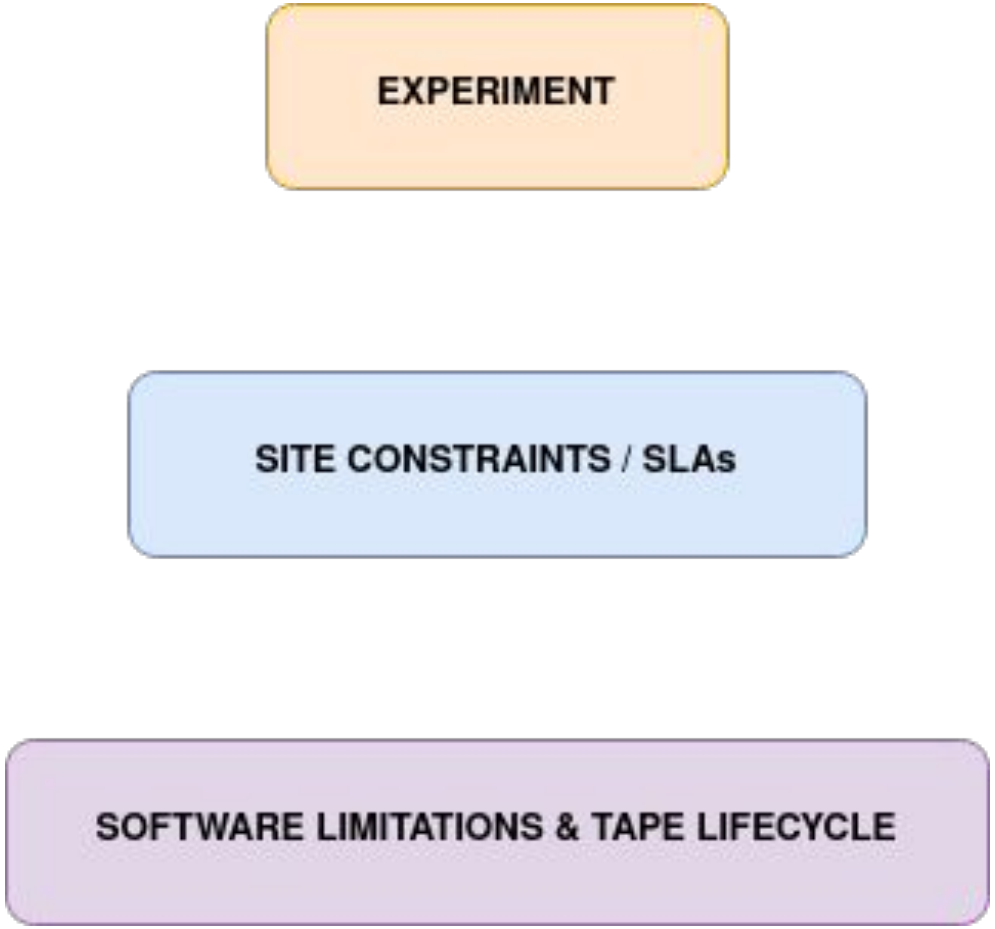
We need to standardize archive metadata and work together on tape collocation at various levels

Long term plans: Improve tape read performance

SEPARATE CONCERNS

- **Experiment**
  - knowledge of recall workflows
  - knows all file metadata
  - retrieve priority/archive priority?
- **Site**
  - constraints for:
    - T0 on tape ASAP, dataset not finished, multiple experiments
    - T1 dataset well defined
- **Software limitations and tape lifecycle**
  - Not coded overnight: metadata stored per file as hint for storage endpoint monitoring initially
  - Collocation must improve with tape repack

Metadata as a common language to define distance between files



Improve metadata for refinements

Archive metadata proposal

- **HTTP protocol only**
- **best effort style: every site gets it and can consume it/decide to do something with it**
- **Initially CTA will just store it along the file**
- **Interest expressing archive priority metadata**
  - Fail less critical workflows: if pledged BW is exceeded/tape infrastructure incident?
    - Fail MC archive storage operations think DAQ SLA could be affected

Retrieve metadata separate of concerns/cleanup/refresh?

- **storage endpoint should not reuse activity *as is* for tape read mount priority**
  - RUCIO should pass a specific *read_priority* weight computed per file depending on rucio activity, rucio priority inside activity, number of retries, rucio rule owner…

Outlook

- **CTA delivers nominal archival performance for Run3 with significant efficiency improvements**
  - Additional efficiency gains for large files will come too
- **Ongoing consolidation of data workflows, protocols and operations workflows**
  - migration from CASTOR allowed a first pass cleanup
  - Separation of concerns should be enforced in the RUCIO/FTS/STORAGE ELEMENT stack
    - Developer point of view VS operations concerns: transient ERRORs are OK VS ERRORs look bad…
      - Better differentiate transient ERRORs from CRITICAL ones to lower SNR?
- **Tape and protocol consolidation ongoing on the grid**
  - Opportunity to consolidate tape data workflows we should not missed
- **NEXT STEP clearly oriented toward monitoring and improving tape data reads**