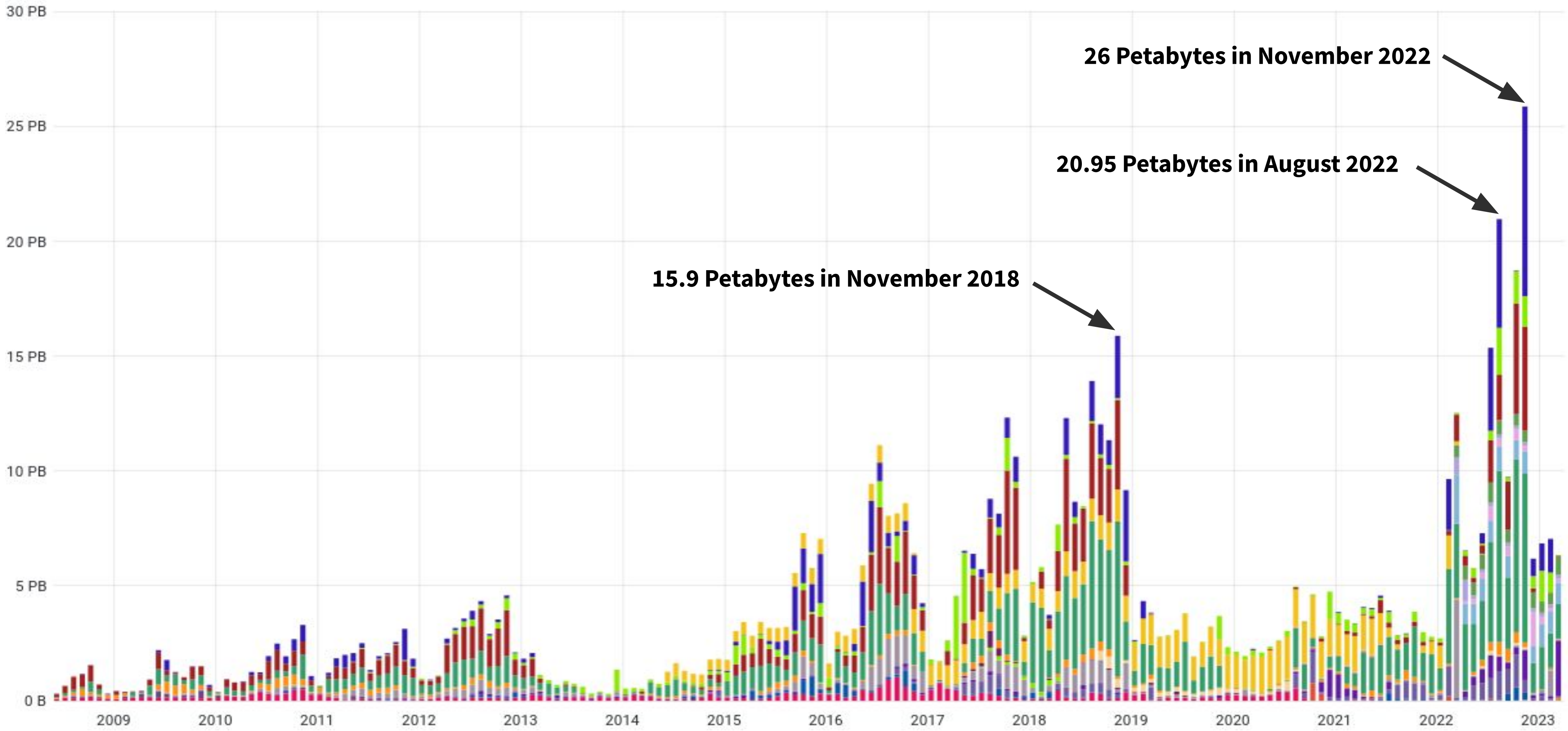
Status of CERN Tape Archive operations during Run3

Julien Leduc
on behalf of IT-SD-TAB

30/03/23

Monthly archive volume records for CTA T0

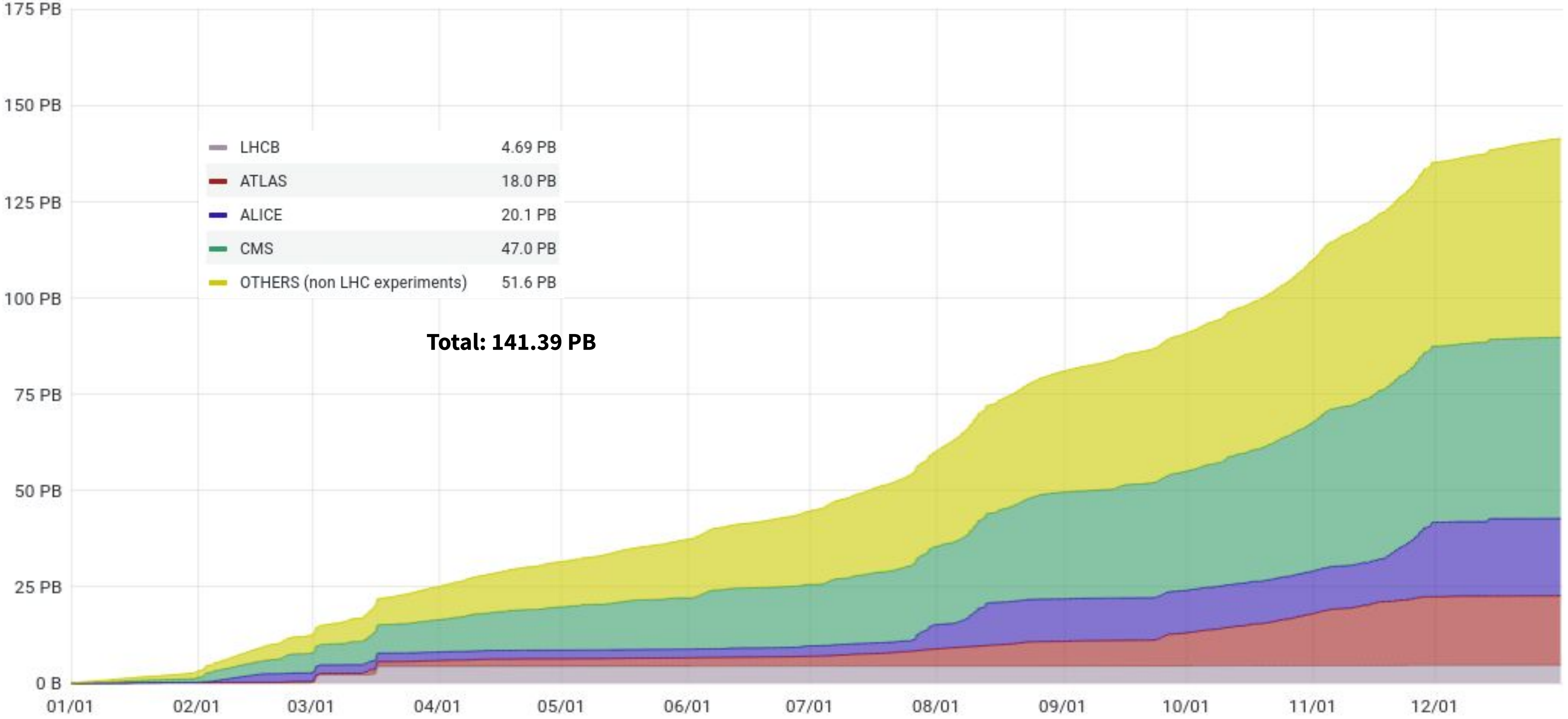


26 Petabytes in November 2022

20.95 Petabytes in August 2022

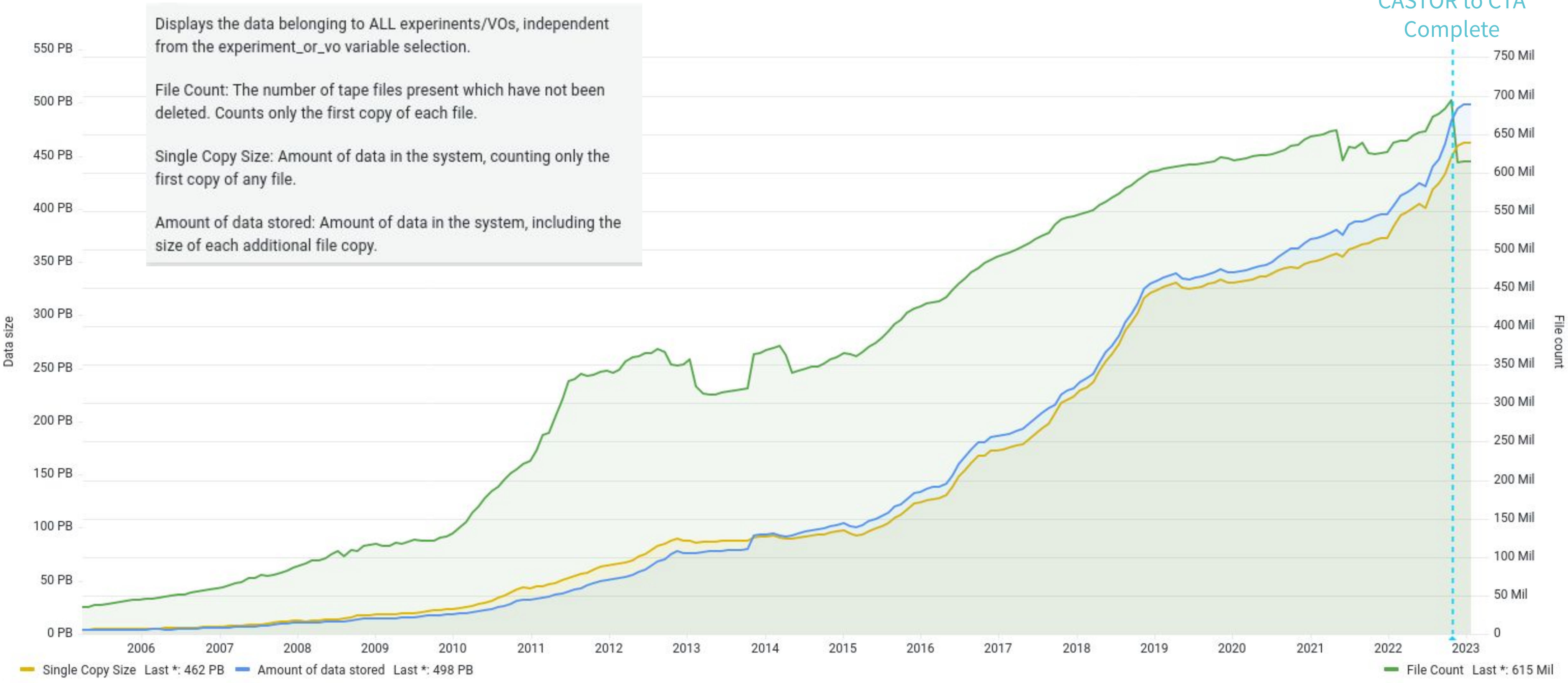
15.9 Petabytes in November 2018

Cumulated archive volume in 2022 for CTA T0



Total: 141.39 PB

Tape namespace statistics


Displays the data belonging to ALL experinents/VOs, independent from the experiment_or_vo variable selection.

File Count: The number of tape files present which have not been deleted. Counts only the first copy of each file.

Single Copy Size: Amount of data in the system, counting only the first copy of any file.

Amount of data stored: Amount of data in the system, including the size of each additional file copy.

Migration from CASTOR to CTA Complete

CTA T0 Production Tape infrastructure

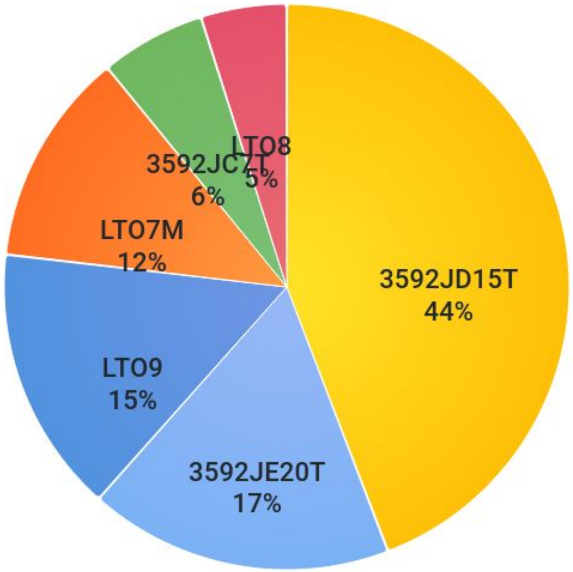
5 tape libraries:

- **3 x IBM T4500 (1 LTO + 2 Enterprise)**
- **2 x Spectra Logic TFinity (LTO)**

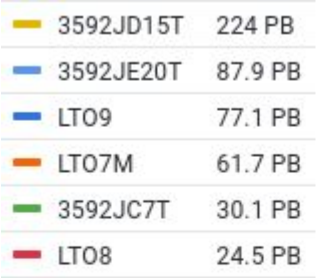
184 tape drives:

- **9 LTO8**
- **93 LTO9**
- **8 TS1155**
- **74 TS1160**

Tape volume distribution

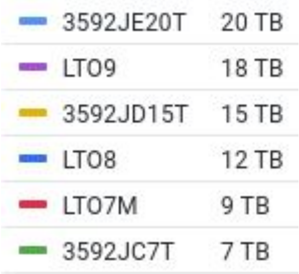


Capacity distribution per media type



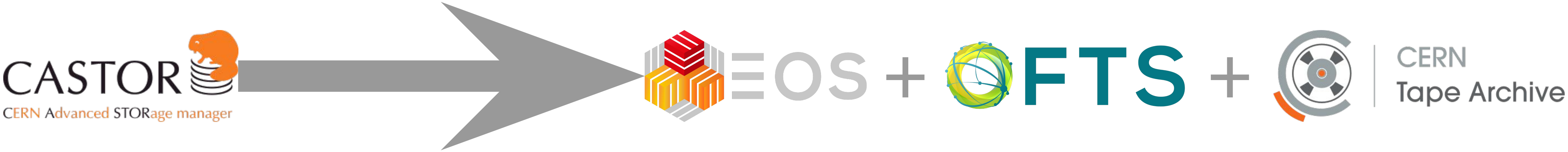
🟨	3592JD15T	224 PB
🟦	3592JE20T	87.9 PB
🟦	LTO9	77.1 PB
🟧	LTO7M	61.7 PB
🟩	3592JC7T	30.1 PB
🟥	LTO8	24.5 PB

Capacity helper



🟦	3592JE20T	20 TB
🟪	LTO9	18 TB
🟨	3592JD15T	15 TB
🟦	LTO8	12 TB
🟥	LTO7M	9 TB
🟩	3592JC7T	7 TB

CASTOR to CTA DT workflows migration



- **CTA is a pure tape system: DATA IS SAFE WHEN IT IS ON TAPE**
  - Compulsory for all DT workflows to use FTS CheckOnTape feature (or equivalent)
    - **supported by xrootd AND http**
- **Disk cache duty consolidated in the main EOS instance**
  - Separate disk and tape concerns
- Operating tape drives at full speed full time **efficiently requires a SSD based buffer: EOSCTA**
  - CTA cannot afford redundancy on SSDs
    - files corrupted/lost in the tape buffer are quickly marked as failed transfers by CheckOnTape
    - transfer must be retried from main EOS

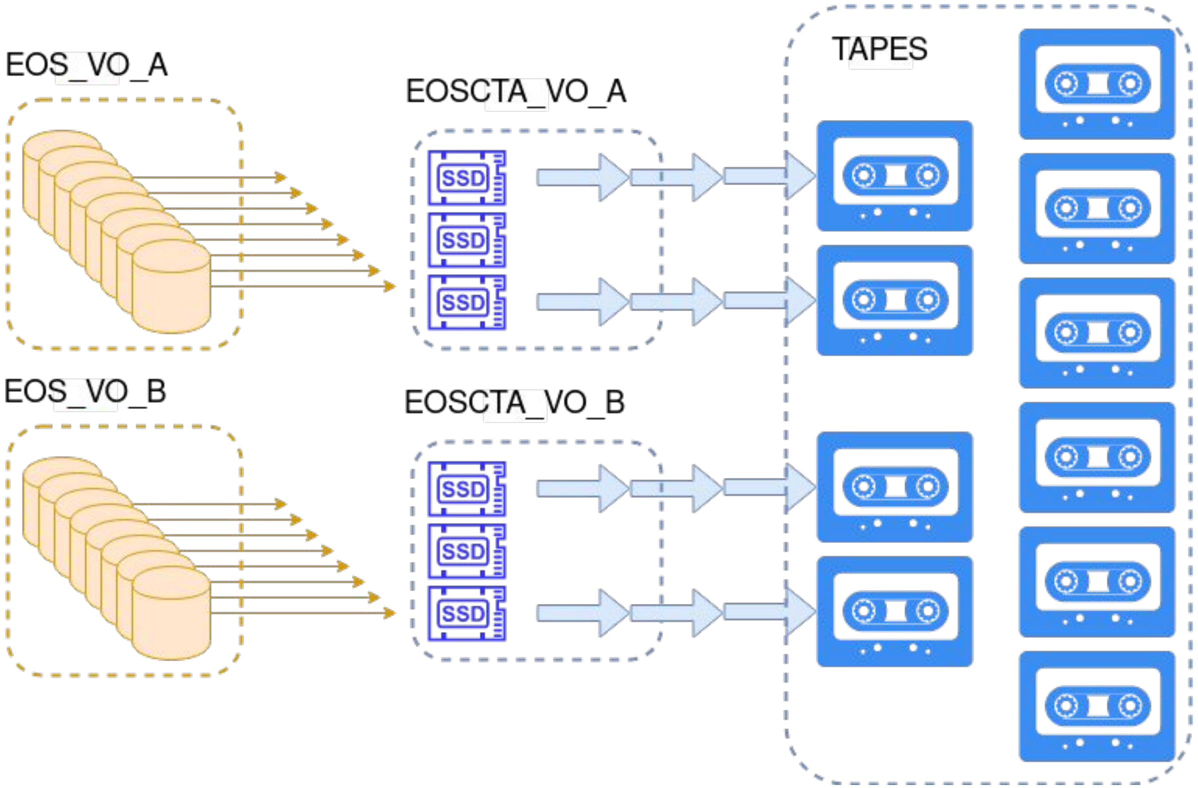
EOSCTA tape buffer characteristics

EOSCTA tape buffer hardware:

- 64 x hyper-converged servers
  - 16 x 2TB SSDs
  - 25Gb/s Ethernet
- 4:3 blocking factor connectivity to CERN CC router
  - 1.2Tb/s or 150GB/s of full duplex buffer bandwidth

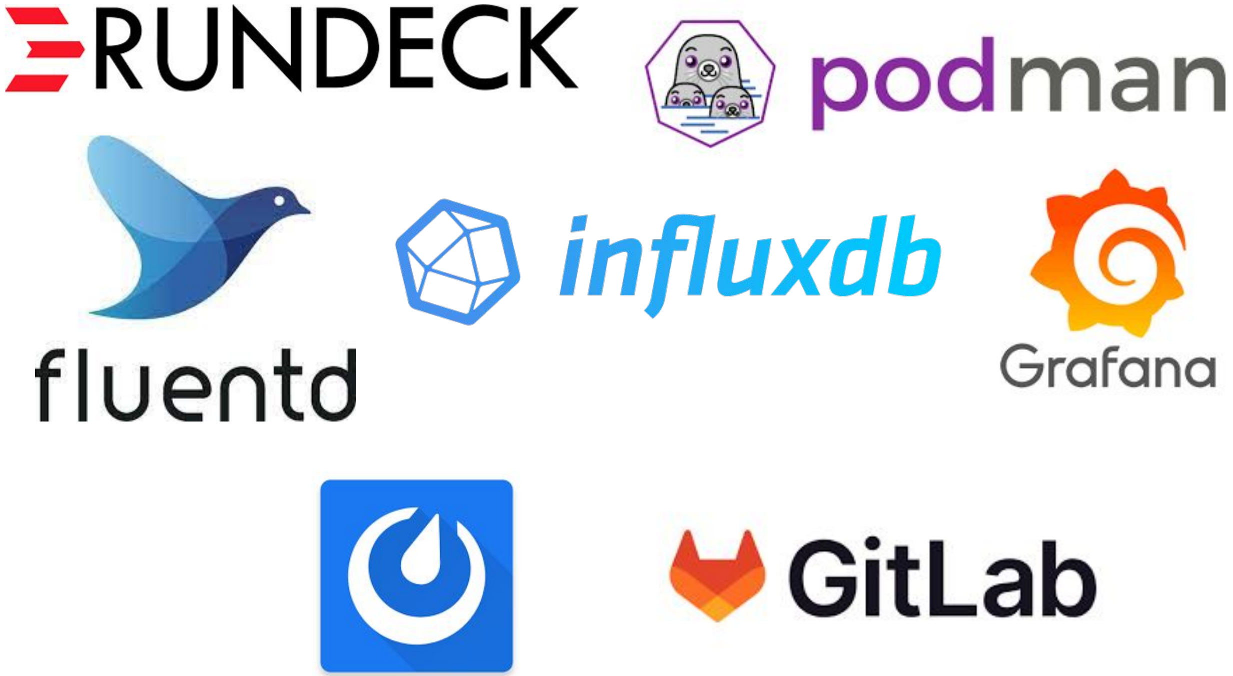
EOSCTA tape buffer properties:

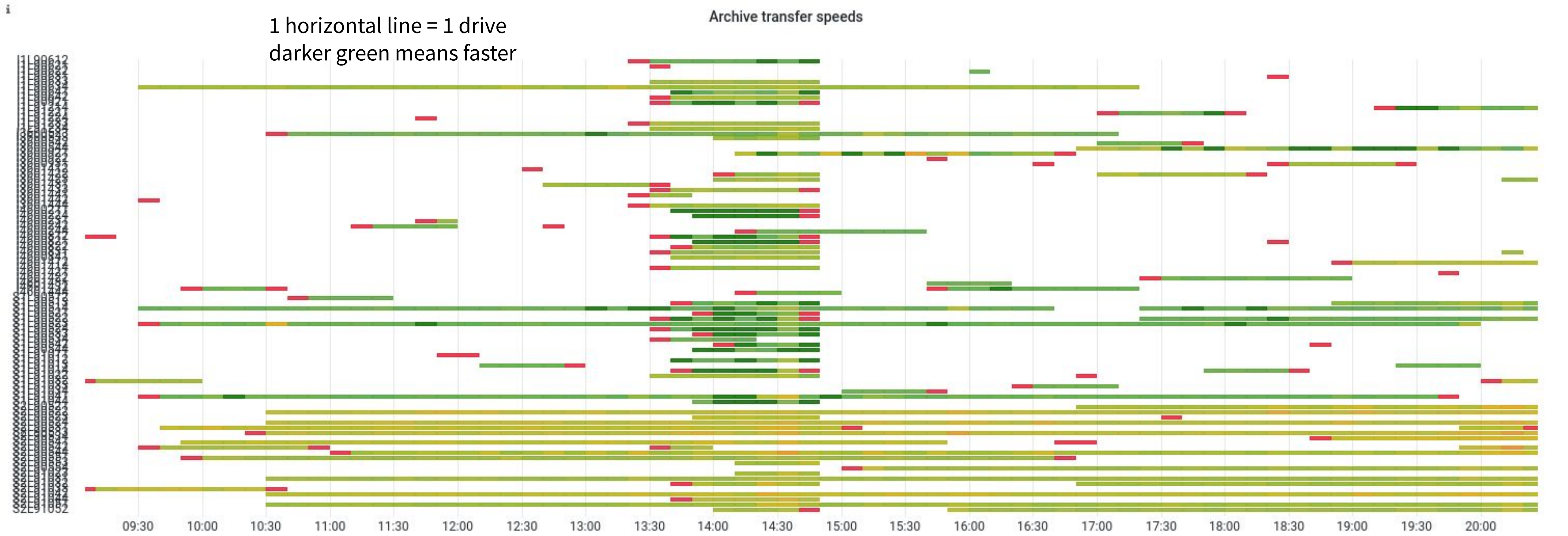
- Conservative setup *evolved*
  - tape buffer separated from tape infrastructure
  - up to 8 hours of buffer to tape
- Move files to/from tape
- Not part of the pledge: **not available for physics jobs**
- Files are *evicted* as soon as they are safely archived on tape
  - or copied on “Big EOS” for retrieves
- Efficiency first
  - **Cannot afford redundancy**
- Early failure notification for retries



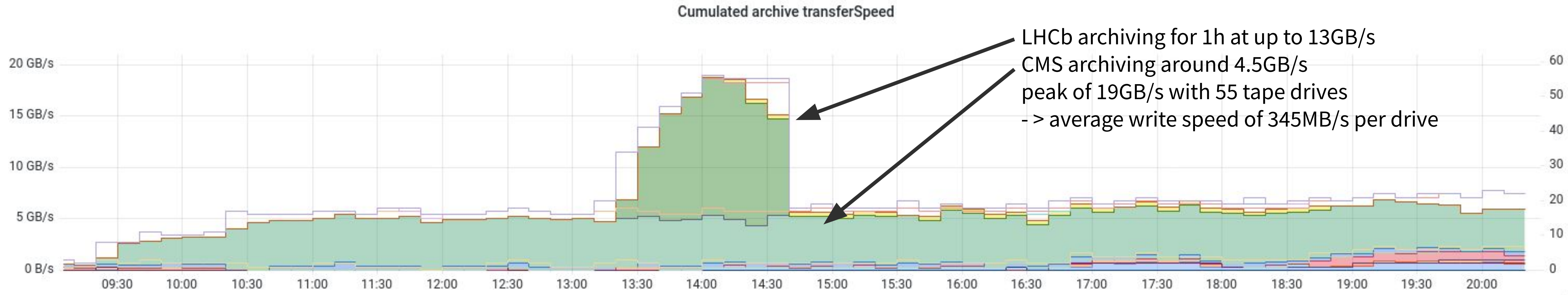
Continuous improvement of EOSCTA operations

- **Operations monitoring** CTA@EOS WORKSHOP 2023
  - real time, short lived, wipe and replace
  - sends alerts in mattermost
- **Operations issues in gitlab**
  - tracking incidents, specific activities, postmortem
  - follows up, dev_ticket needed,...
  - Reviewed once a week at CTA operations meeting
    - minutes, rota calendar in gitlab wiki
- **Operations procedures** CTA@EOS WORKSHOP 2023
  - automated workflows in rundeck scheduled jobs or containers
  - CTA catalogue upgrade container
  - Weekly EOSCTA namespace dump per vo
    - json list of **healthy files on tape**/**files on BROKEN tapes**





1 horizontal line = 1 drive
darker green means faster



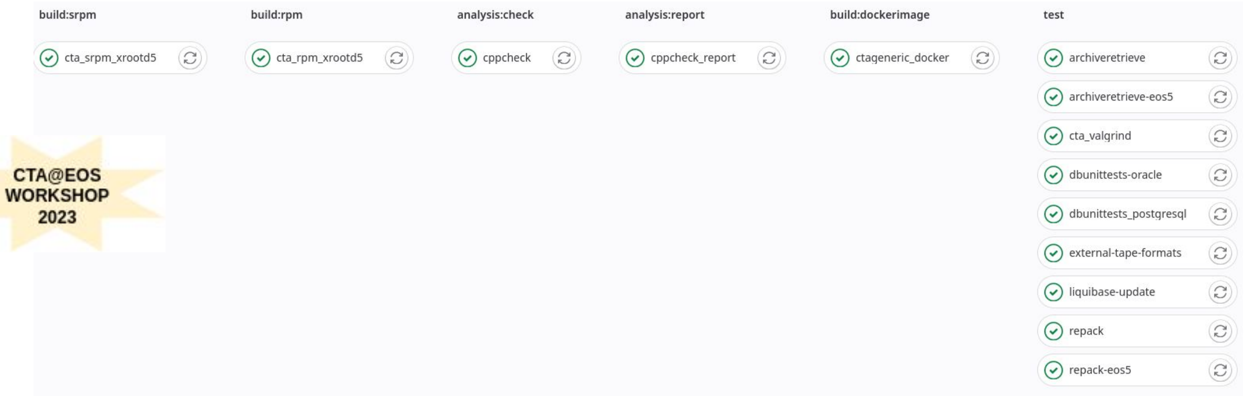
LHCb archiving for 1h at up to 13GB/s
CMS archiving around 4.5GB/s
peak of 19GB/s with 55 tape drives
- > average write speed of 345MB/s per drive

Release often with confidence

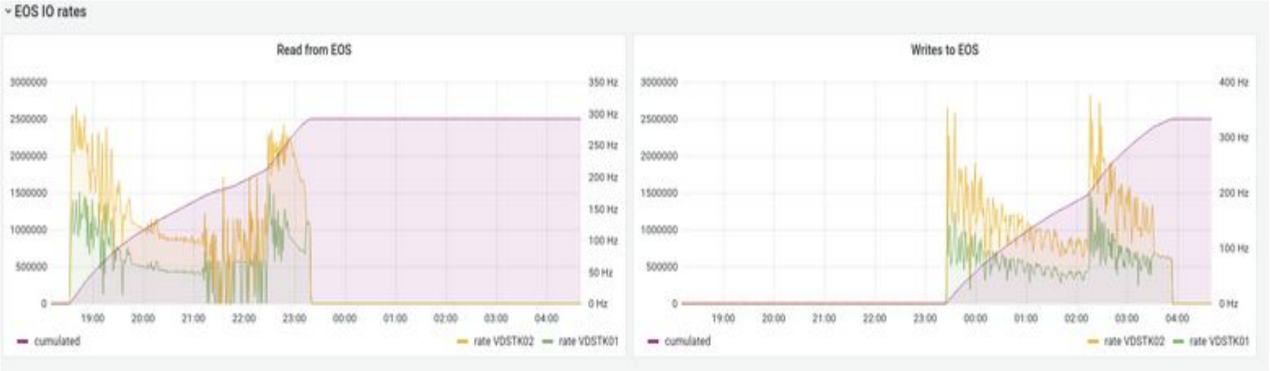
- **CTA developer develop and test on their dev box**
  - test in their private CI eoscta instance running in kubernetes
- **Pushed code is built and tested in gitlab CI**
  - deployed on dedicated runners that run series oı system tests on every commit
- **Release commit are stress tested**
  - CI instance on steroids that archive and retrieve 2.5M files in less than 10 hours
- **Tagging publishes rpms internally**
- **Release deployed on preproduction**
- **Following day it is deployed in production**
  - Not on Fridays

See [Tagging a new CTA release]

CTA@EOS
WORKSHOP
2023



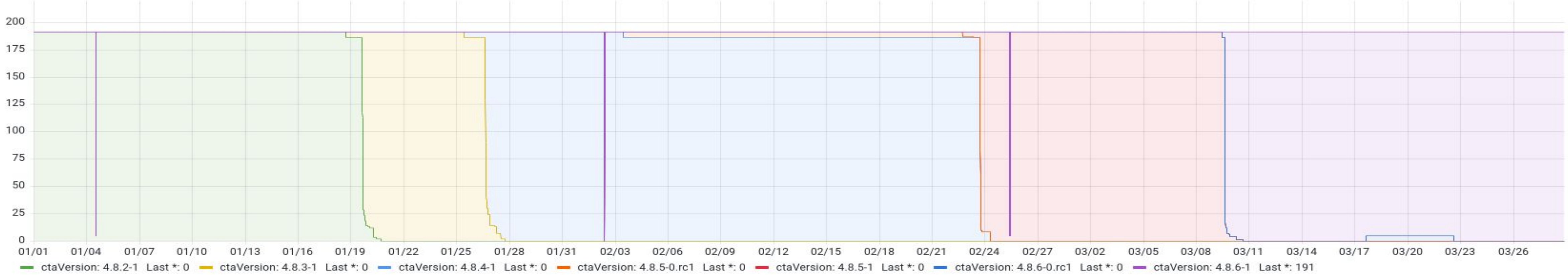
build:srpm	build:rpm	analysis:check	analysis:report	build:dockerimage	test
cta_srpm_xrootd5	cta_rpm_xrootd5	cppcheck	cppcheck_report	ctageneric_docker	archiveretrieve
					archiveretrieve-eos5
					cta_valgrind
					dbunittests-oracle
					dbunittests_postgresql
					external-tape-formats
					liquibase-update
					repack
					repack-eos5



Deploy often

- **5 releases of CTA deployed in production since 1/1/2023**
  - 2 additional rc deployed in preproduction for specific tape hardware tests
- **Possible thanks to:**
  - CTA development release policy
    - next CTA release must be compatible with previous release
      - cta-frontend compatible with tape servers on previous version finishing ongoing tape sessions
  - Automated rundeck upgrade procedure
    - upgrade cta-frontend code
    - put drives DOWN with *upgrading* reason, upgrade cta-taped on DOWN drives, put *upgrading* drives up

NO VISIBLE USER DOWNTIME DURING CTA SOFTWARE UPGRADES



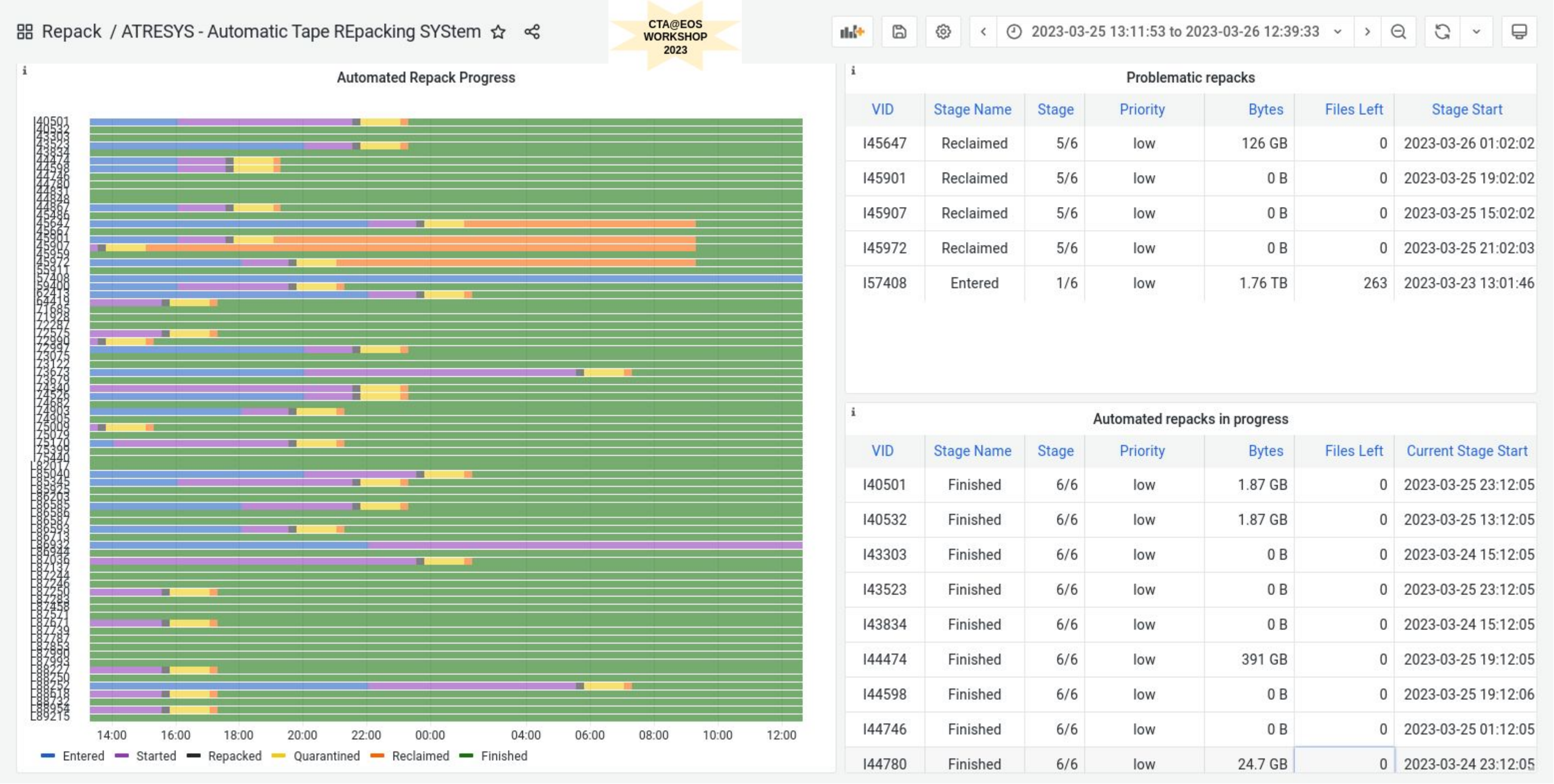
CTA versions in preproduction and production since 1/1/2022

Consolidating operations workflows: repack

CTA@EOS WORKSHOP 2023

- **Repack is key for issues with production issues reading tapes**
  - regular tape defragmentation: clean up backup use cases in CTA
- **Ramping up repack automation**
  - Lot of development efforts formalising and implementing tape state and tape lifecycle state machine with tape operations (in CTA v4/5.8.x)
    - user states: ACTIVE, DISABLED
    - operations states: REPACKING, BROKEN, EXPORTED
    - Additional transition states, and transition rules
  - Simplification of the repack retry logic to read the source tape
    - tapes likely to be problematic to read
  - ATRESYS - Automatic Tape REpacking SYStem
    - Manage queues of tapes that need to be repacked
    - Move tape to the next step in the tape lifecycle workflow
    - Provides tape lifecycle monitoring in standard grafana

CTA@EOS WORKSHOP 2023



Problematic repacks

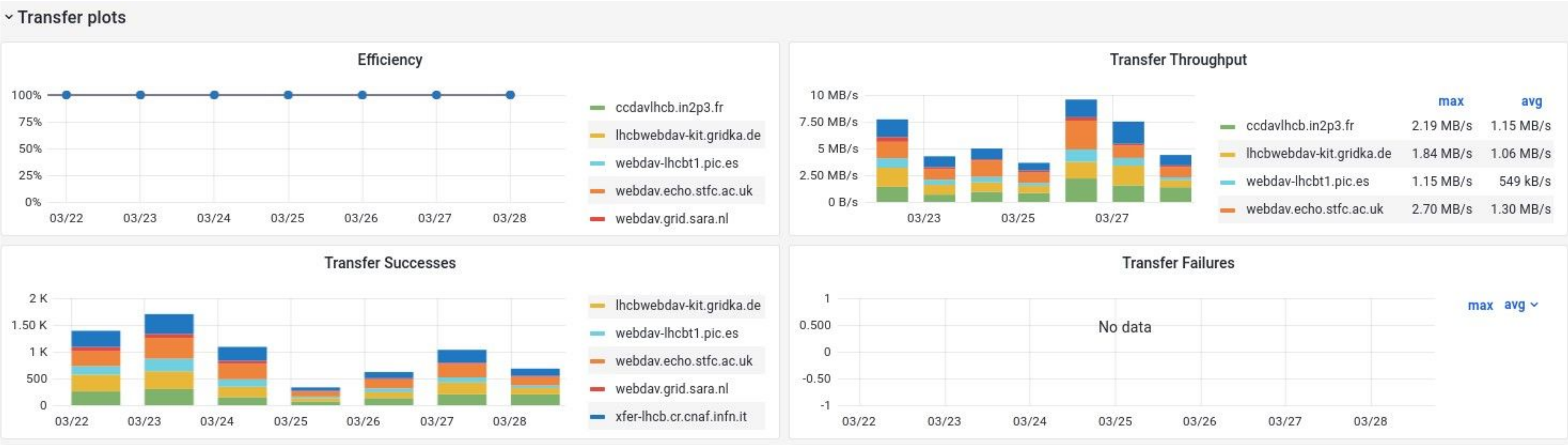
VID	Stage Name	Stage	Priority	Bytes	Files Left	Stage Start
I45647	Reclaimed	5/6	low	126 GB	0	2023-03-26 01:02:02
I45901	Reclaimed	5/6	low	0 B	0	2023-03-25 19:02:02
I45907	Reclaimed	5/6	low	0 B	0	2023-03-25 15:02:02
I45972	Reclaimed	5/6	low	0 B	0	2023-03-25 21:02:03
I57408	Entered	1/6	low	1.76 TB	263	2023-03-23 13:01:46

Automated repacks in progress

VID	Stage Name	Stage	Priority	Bytes	Files Left	Current Stage Start
I40501	Finished	6/6	low	1.87 GB	0	2023-03-25 23:12:05
I40532	Finished	6/6	low	1.87 GB	0	2023-03-25 13:12:05
I43303	Finished	6/6	low	0 B	0	2023-03-24 15:12:05
I43523	Finished	6/6	low	0 B	0	2023-03-25 23:12:05
I43834	Finished	6/6	low	0 B	0	2023-03-24 15:12:05
I44474	Finished	6/6	low	391 GB	0	2023-03-25 19:12:05
I44598	Finished	6/6	low	0 B	0	2023-03-25 19:12:06
I44746	Finished	6/6	low	0 B	0	2023-03-25 01:12:05
I44780	Finished	6/6	low	24.7 GB	0	2023-03-24 23:12:05

HTTP protocol consolidation on tape

- **Remove few sub-optimal data flows**
  - xrootd TPC with delegation transfers
  - 1 gridftp use case in CTA T0 (low priority)
- **Experiments moving to HTTP protocol on WLCG**
  - HTTP TAPE REST API version 1.0 specifications implemented in EOSCTA software stack in CTA 5/4.8.7-1
    - **Critical for check on tape** (implemented with fileinfo method in GFAL2)
  - Deployed at T0 on HTTP oriented EOSCTA LHC instances earlier this month
    - tested with RUCIO ATLAS team in preproduction
    - **archive transfers to eosctalhcb ongoing in production for LHCb using checkOnTape**



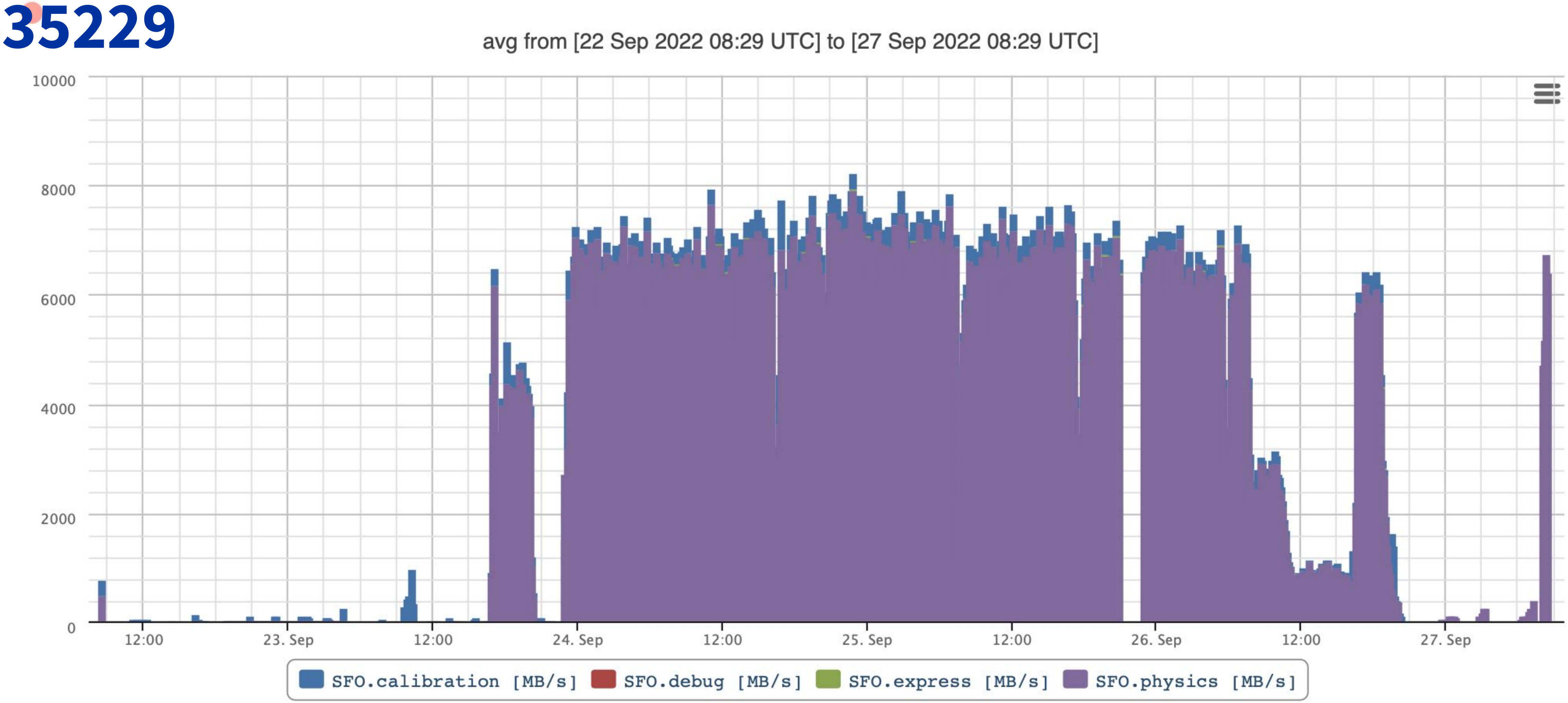
	max	avg
ccdavlhcb.in2p3.fr	2.19 MB/s	1.15 MB/s
lhcbwebdav-kit.gridka.de	1.84 MB/s	1.06 MB/s
webdav-lhcbt1.pic.es	1.15 MB/s	549 kB/s
webdav.echo.stfc.ac.uk	2.70 MB/s	1.30 MB/s

Beginning of Run3: *legacy* placement tweaks

- CTA maps tape family with directory
  - *CASTOR legacy*
- Improving written data placement with the experiments
  - Improve per directory tape collocation on tape
    - CMS split of MC, 2022 data, 2023 data
- Several limitations as
  - CTA directory structures is dictated by experiment namespace
    - no directory/file remapping takes place in CTA tape buffer

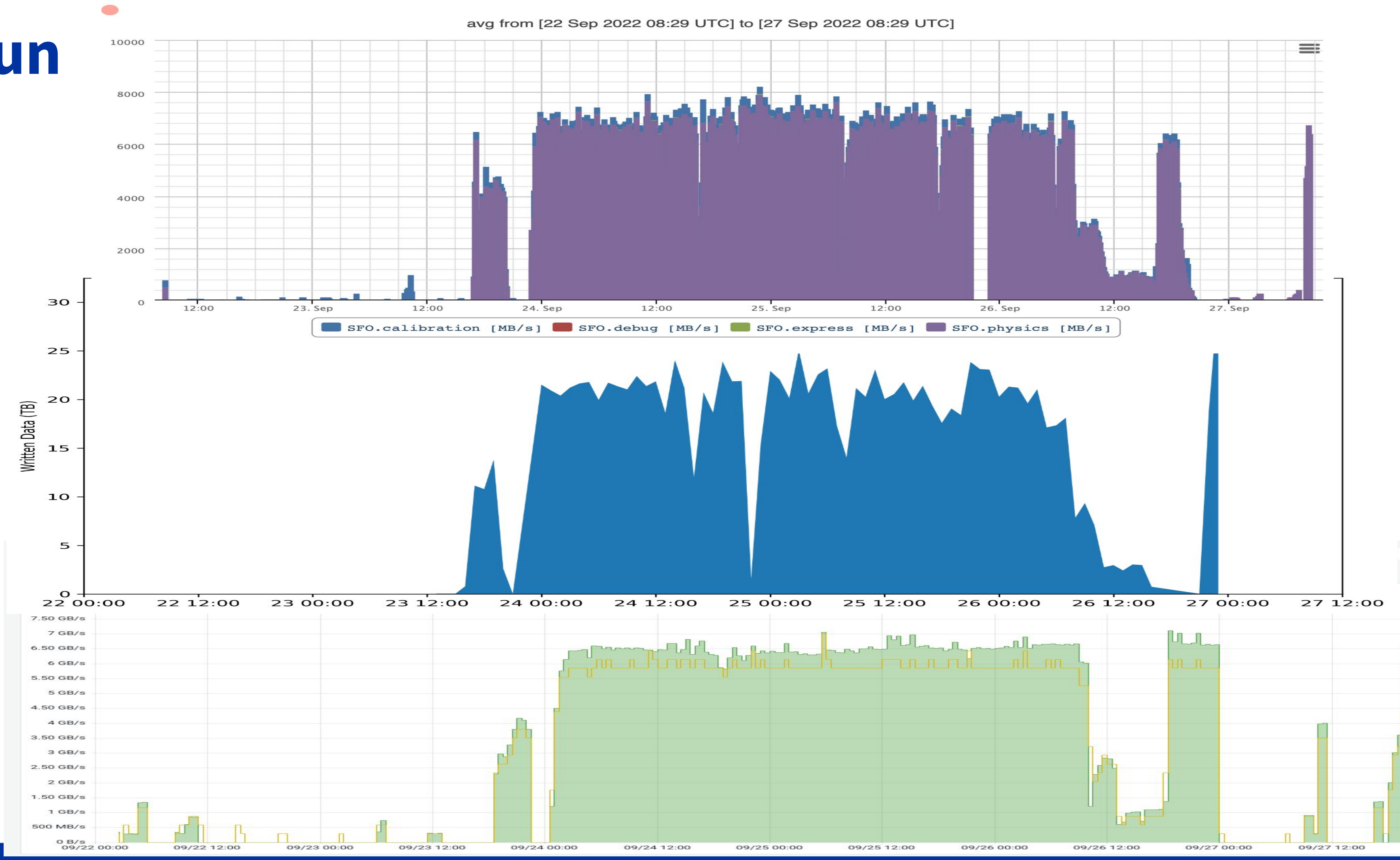
In tape T0 team we were convinced that time based collocation and low latency from DAQ was enough to ensure good enough read performance using a FIFO tape scheduler…

ATLAS Run 435229



ATLAS Run 435229

ATLAS DAQ to T0 tape latency is 21 minutes



ATLAS Run 435229

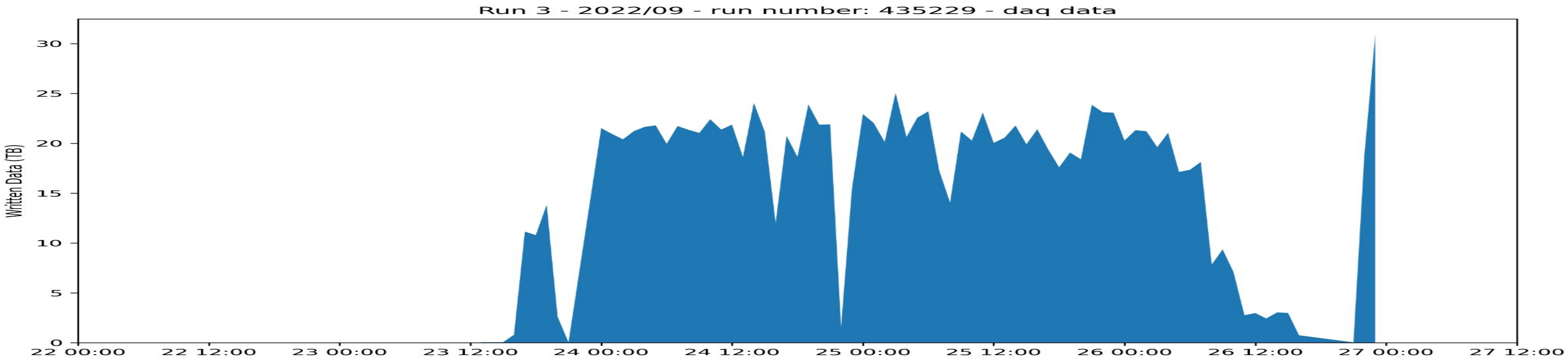
LHC stable beam -> huge datasets

1.8B events, 1.3PB

> 12 other smaller datasets sent in parallel

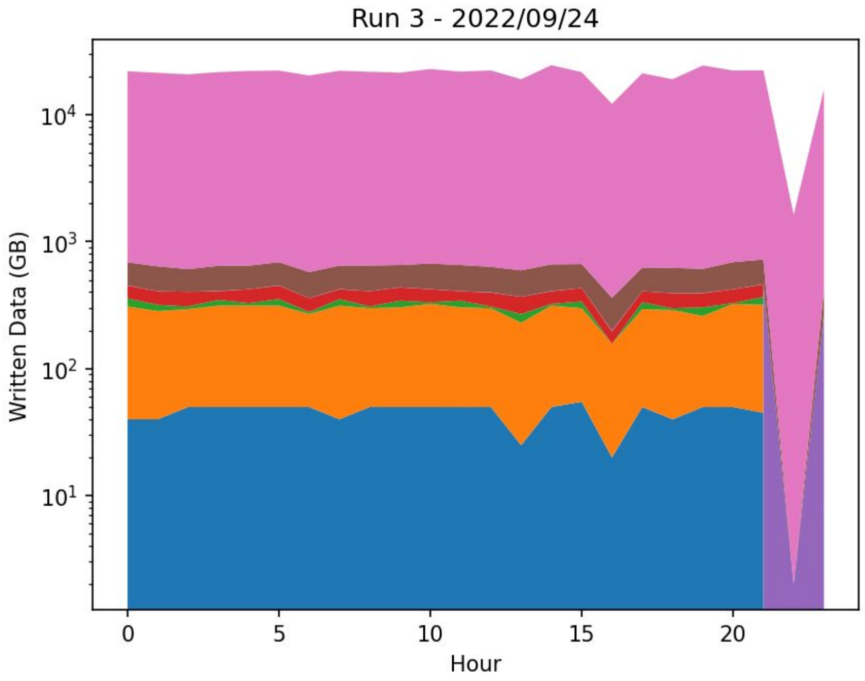
# of parallel datasets sent per run?

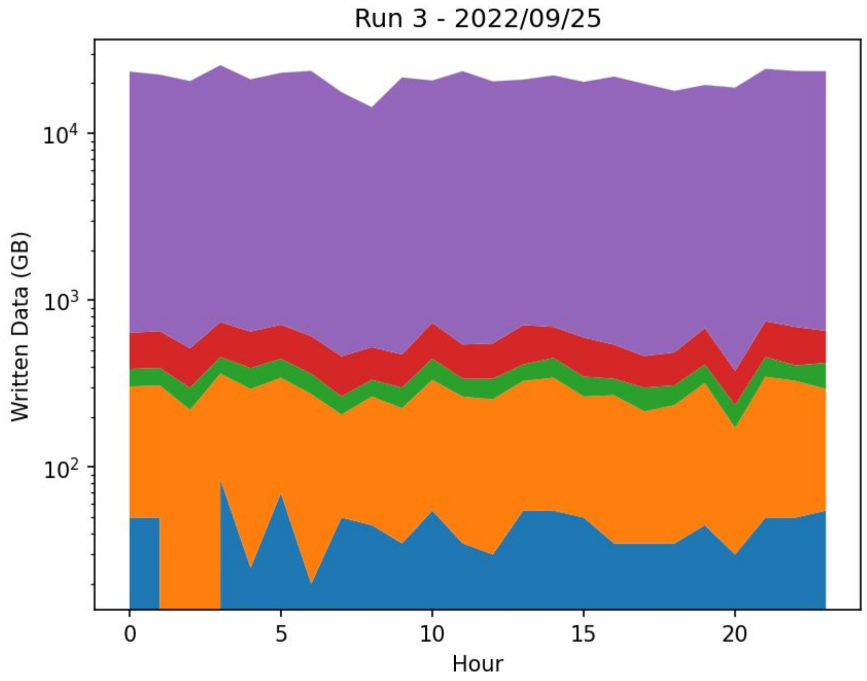
BAD FOR CTA DATA PLACEMENT ON TAPE!!



```
                                                name               wb
10      data22_13p6TeV.435229.physics_MinBias.daq.RAW.  508882192115748
2   data22_13p6TeV.435229.calibration_LArCells.daq...    6029697160416
9         data22_13p6TeV.435229.physics_Main.daq.RAW.    5633626551664
7   data22_13p6TeV.435229.calibration_ZDCCalib.daq...    2060844419964
0   data22_13p6TeV.435229.calibration_AFPCalib.daq...    1095598880584
3   data22_13p6TeV.435229.calibration_LArCellsEmpt...     618731659276
1   data22_13p6TeV.435229.calibration_CostMonitori...     199619777636
6     data22_13p6TeV.435229.calibration_Tile.daq.RAW.     157313259108
5   data22_13p6TeV.435229.calibration_LArPEBDigita...      52435401824
4   data22_13p6TeV.435229.calibration_LArPEB.daq.RAW.      52432509416
8    data22_13p6TeV.435229.calibration_lucid.daq.RAW.      23718811248
11   data22_calib.435229.calibration_MuonAll.daq.RAW.       6442449660
```

```
                                                name               wb
10      data22_13p6TeV.435229.physics_MinBias.daq.RAW.  533995407337536
2   data22_13p6TeV.435229.calibration_LArCells.daq...    6150297188928
9         data22_13p6TeV.435229.physics_Main.daq.RAW.    6010032822852
7   data22_13p6TeV.435229.calibration_ZDCCalib.daq...    2229594391816
0   data22_13p6TeV.435229.calibration_AFPCalib.daq...    1116667277768
3   data22_13p6TeV.435229.calibration_LArCellsEmpt...     576785520352
1   data22_13p6TeV.435229.calibration_CostMonitori...     210124626128
6     data22_13p6TeV.435229.calibration_Tile.daq.RAW.     199264141996
5   data22_13p6TeV.435229.calibration_LArPEBDigita...      83896391044
4   data22_13p6TeV.435229.calibration_LArNoiseBurs...      52431081832
8    data22_13p6TeV.435229.calibration_lucid.daq.RAW.      21562555680
11   data22_calib.435229.calibration_MuonAll.daq.RAW.       6442448020
```





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**
  - at T0 30 free tapes needed per tapepool…
  - cardinality of datasets written in parallel at T0 cannot accommodate 1 tape pool per dataset
    - expecting worse cardinality in T1 CTA sites…

Softer rules for file collocation on tape are needed

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

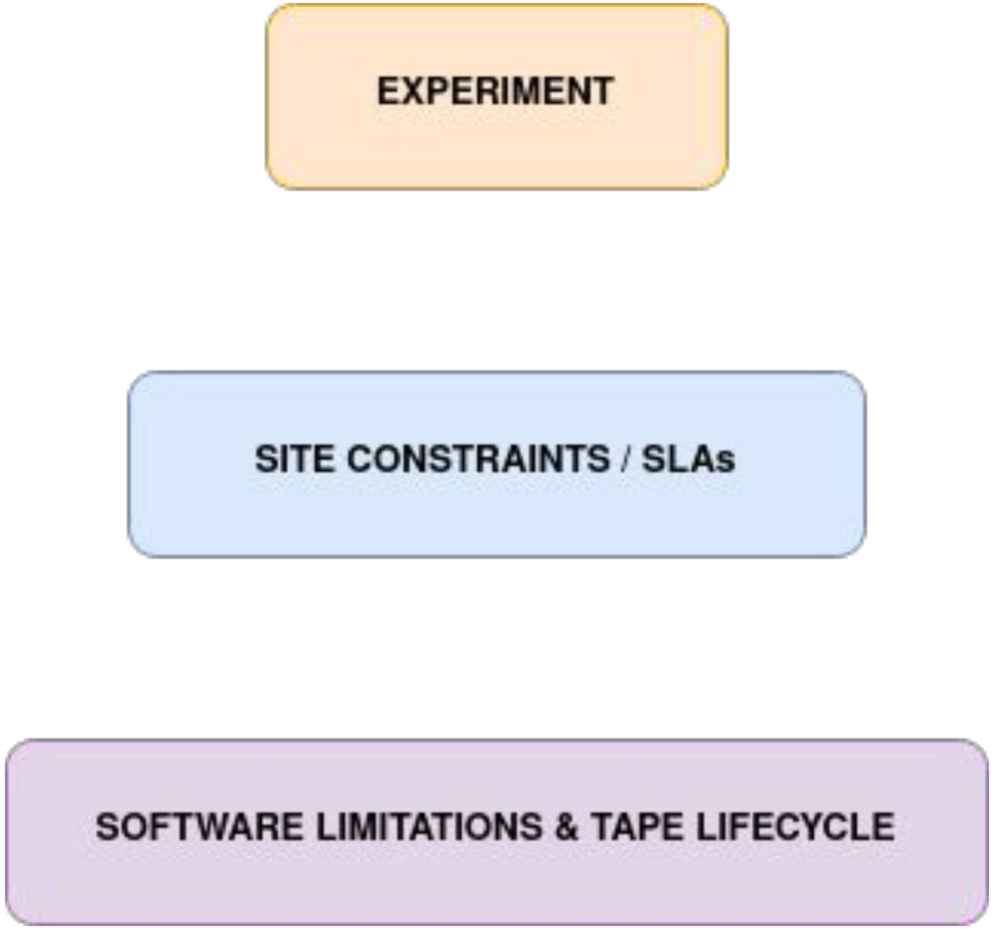
We need to standardize archive metadata and work together on tape collocation at various levels

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 datasets are well defined but all mixed
- **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



EXPERIMENT

SITE CONSTRAINTS / SLAs

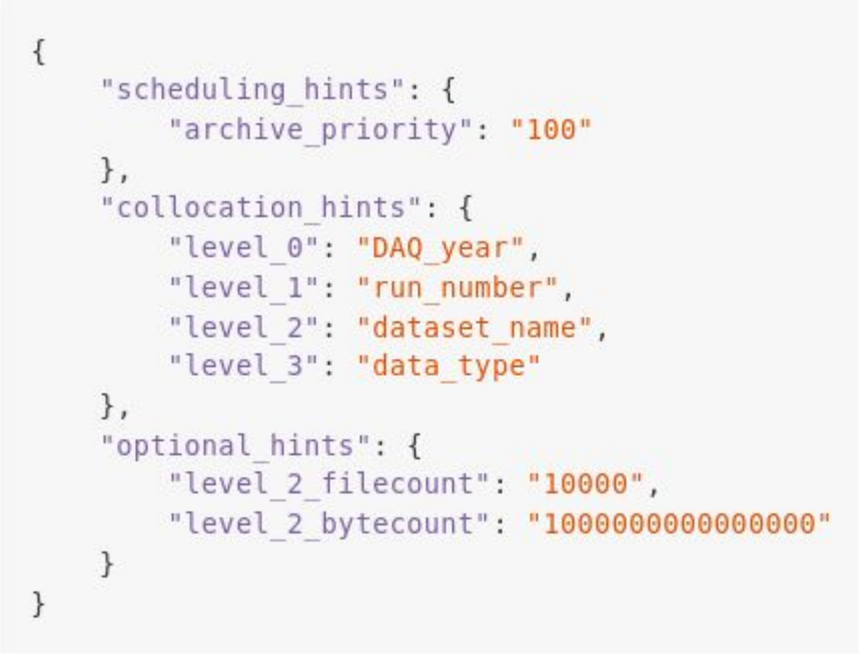
SOFTWARE LIMITATIONS & TAPE LIFECYCLE

Archive metadata early DRAFT

- **CTA/dcache development agreement**
  - up to 4 hierarchical levels for *collocation_hints*
- **Discussed with experiments**
- **FTS transparently encapsulates archive_metadata**
  - header in HTTP file transfer stream
  - this is only a hint tape sites are free to ignore
  - initially targeting placement monitoring
    - measure file distance on tapes
- **CTA team starting work on new tape scheduler**
  - targeting tape placement improvements

CTA@EOS WORKSHOP 2023

Example archive metadata content



```
{
    "scheduling_hints": {
        "archive_priority": "100"
    },
    "collocation_hints": {
        "level_0": "DAQ_year",
        "level_1": "run_number",
        "level_2": "dataset_name",
        "level_3": "data_type"
    },
    "optional_hints": {
        "level_2_filecount": "10000",
        "level_2_bytecount": "1000000000000000"
    }
}
```

Conclusion

- **CTA delivers nominal archival performance for Run3 with significant write efficiency improvements**
  - with initially limited data placement features inherited from CASTOR
- **NEXT STEP clearly oriented toward monitoring and improving data placement for tape data reads**
  - HTTP only
- **Tape and protocol consolidation ongoing on the grid**
  - Opportunity to consolidate tape data workflows should not be missed

Do not miss EOS workshop 2023: 24–27 Apr 2023 at CERN!

CERN

home.cern