EOS - The CERN's distributed multi-petabyte disk storage system

Status and future evolution

Presented by Cedric Caffy on behalf of the EOS team

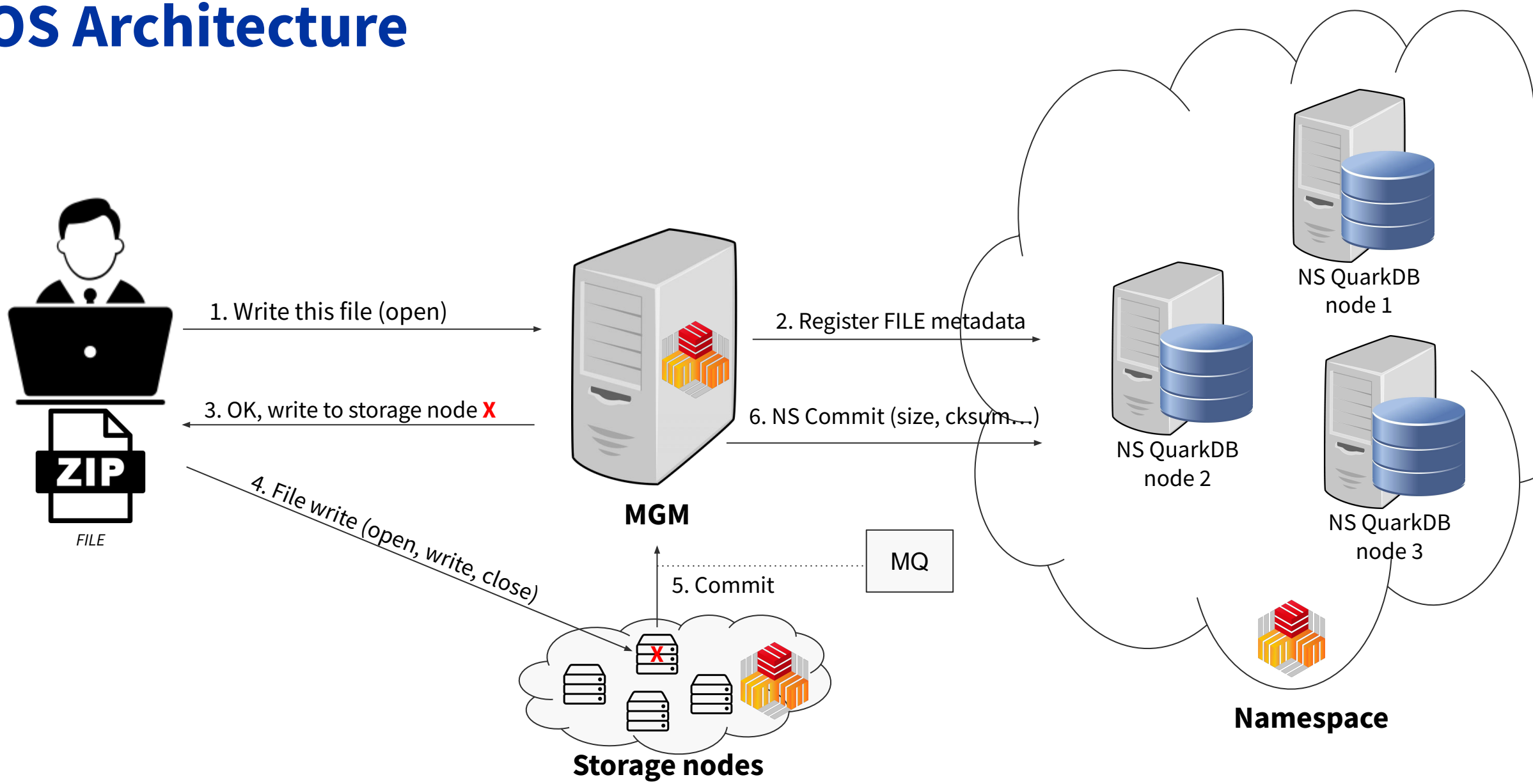
2023-07-24

What is EOS?

EOS means "EOS Open Storage"

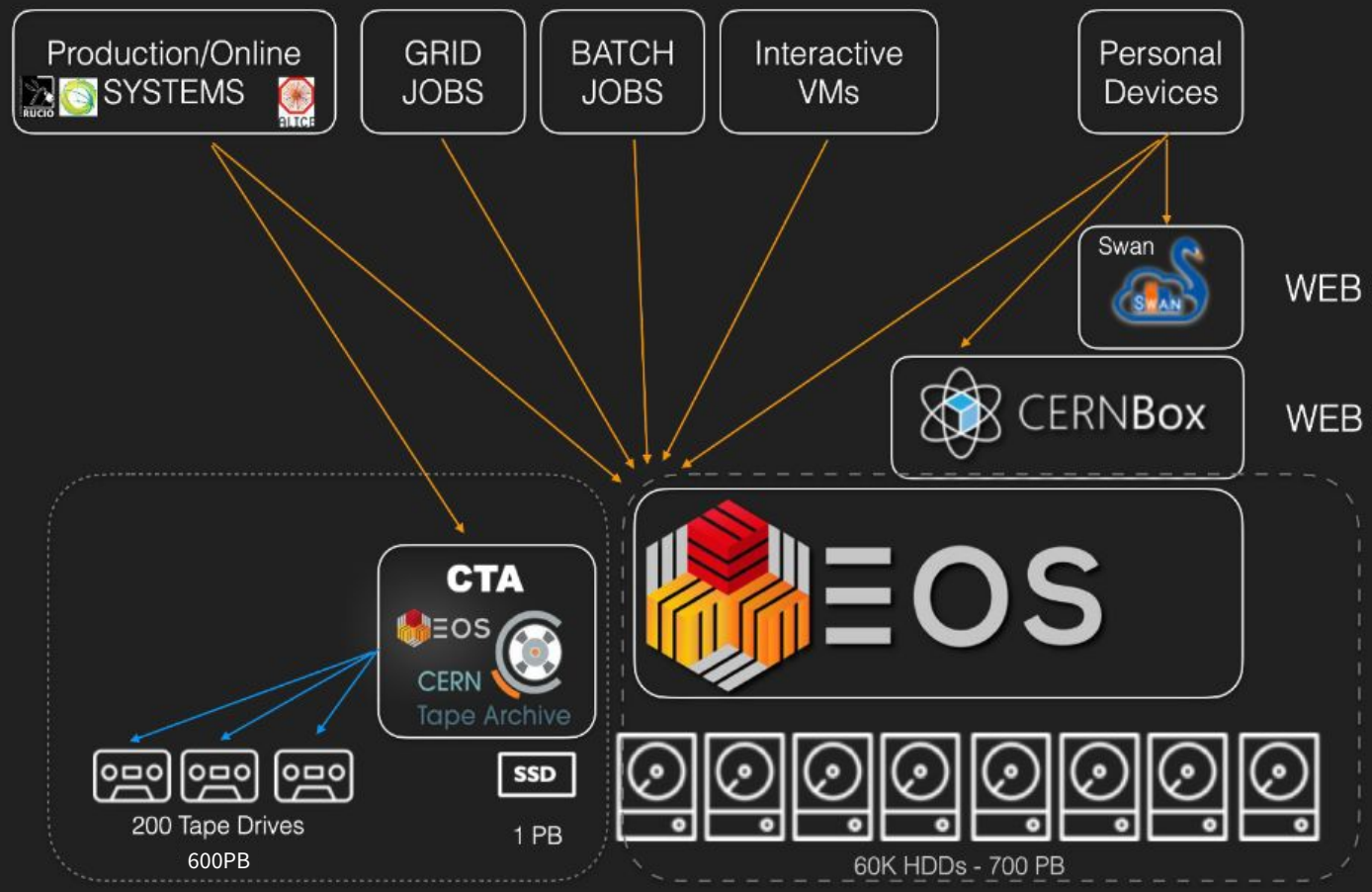
- A long story short: it's a distributed disk storage system
- A service for storing large amounts of physics data and user files, with a focus on interactive and batch analysis
- Filesystem-like hierarchical namespace with a feature-rich permission and quota system
- Designed to support several thousand users at the same time providing secure data storage access

EOS Architecture



EOS in Numbers

EOS Service in Numbers



WEB Services for **Jupyter Notebooks**

WEB Services for **Sync&Share**

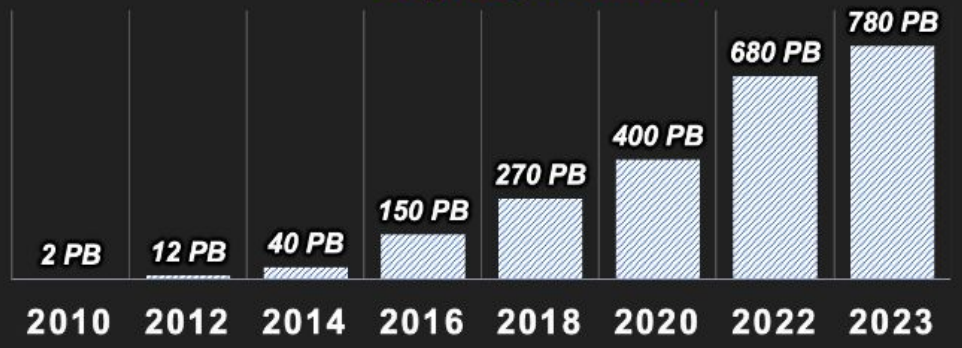
24 individual instances
8 Physics 8 CERNBox 8 CTA

How is EOS used?

2023 Targets

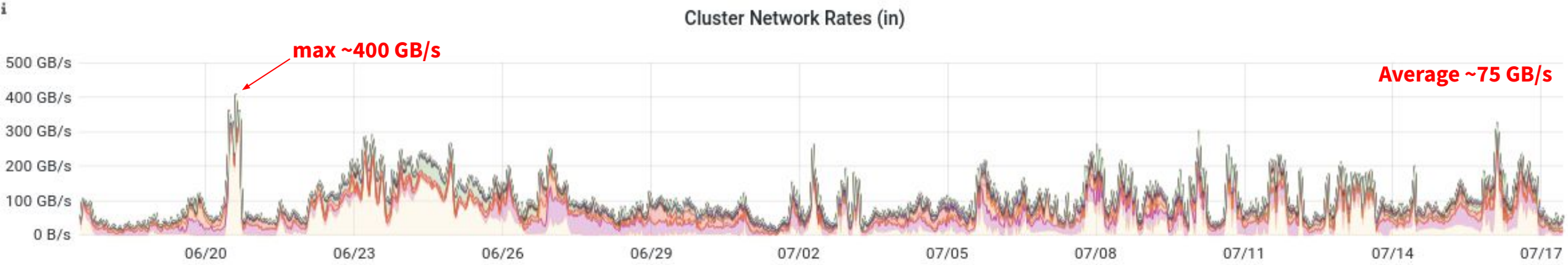
Total Space (raw)	**780 PB**
Files Stored	**~8 Bil**
# Storage Nodes	**~1300**
# Disks	**~60000**

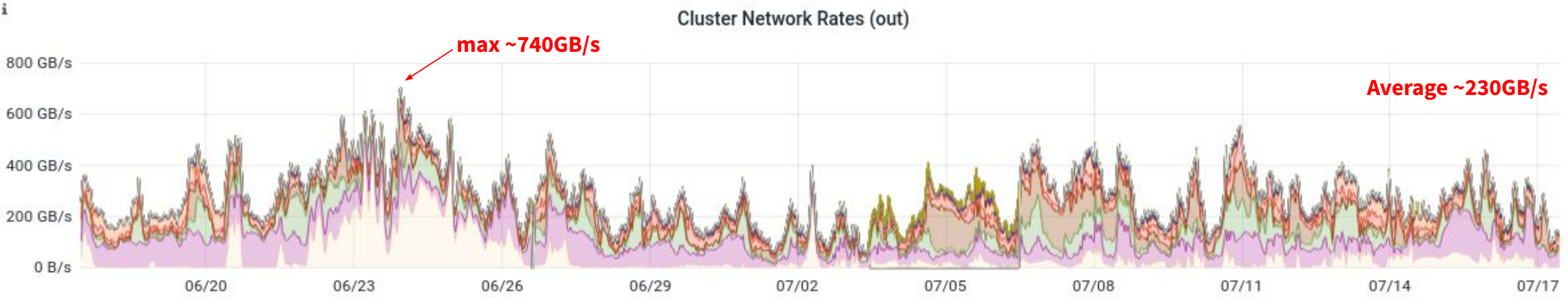
Capacity Evolution



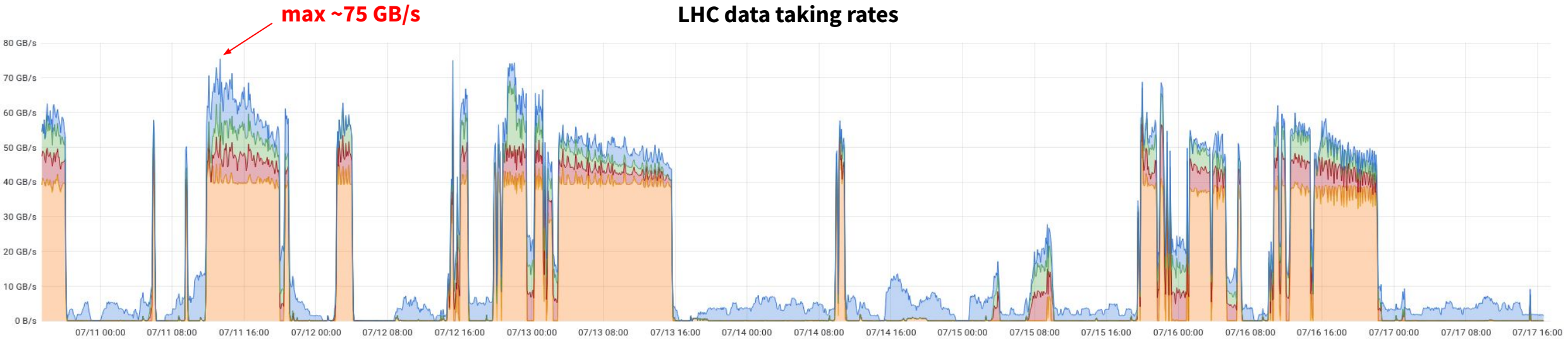
During the last 30 days…

Number of Files	Number of Directories	Used Space	Total Space	IOPS
7.92 Bil	637 Mil	551.17 PB	774.90 PB	85M io/s





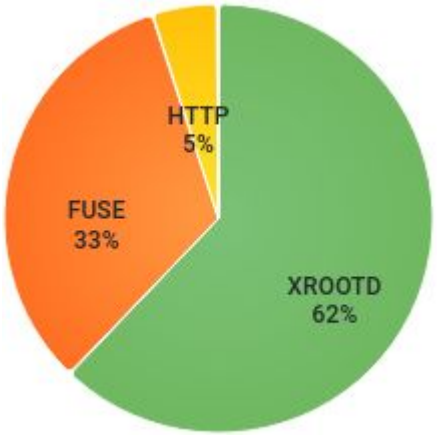
During the last 7 days…



	max	avg	current
ALICE O2 Point2	57.7 GB/s	12.1 GB/s	0 B/s
ALICE P2 Point2	0 B/s	0 B/s	0 B/s
ATLAS Point1	16.7 GB/s	2.30 GB/s	8.54 MB/s
CMS Point5	19.3 GB/s	2.13 GB/s	8.61 MB/s
LHCb Point8	13.6 GB/s	3.77 GB/s	1.33 GB/s

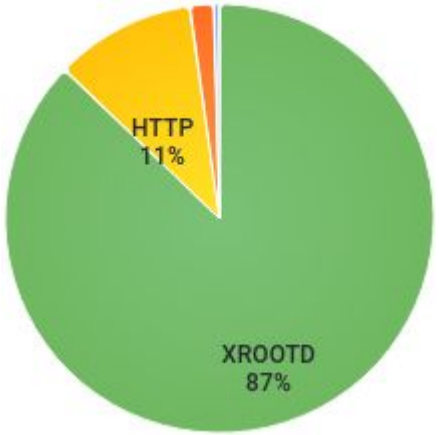
Since the beginning of Run 3 - all EOS instances

17.2 Billion files read for **3.97 EB**



XROOTD	2.47 EB
FUSE	1.30 EB
HTTP	192 PB
GRIDFTP	741 TB

3.23 Billion files written for **532 PB**

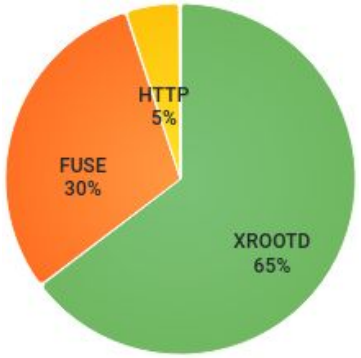


XROOTD	465 PB
FUSE	9.21 PB
HTTP	56.2 PB
GRIDFTP	1.82 PB

Since the beginning of Run 3 - EOS for Physics and CERNBox

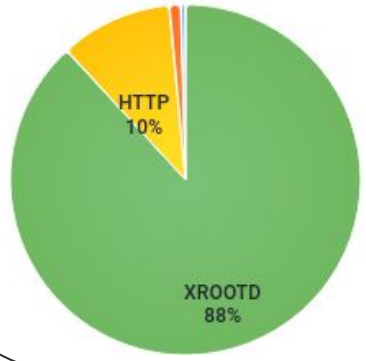
EOS for **Physics**

12.9 Billion files read for **3.79 EB**



XROOTD	2.46 EB
FUSE	1.15 EB
HTTP	186 PB
GRIDFTP	721 TB

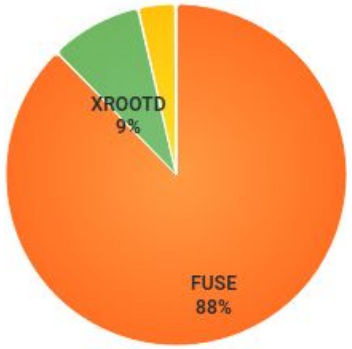
1.23 Billion files written for **525 PB**



XROOTD	464 PB
FUSE	5.13 PB
HTTP	54 PB
GRIDFTP	1.82 PB

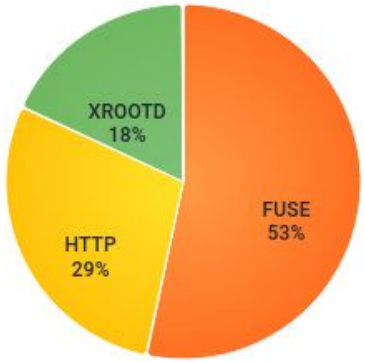
EOS for **CERNBox**

4.32 Billion files read for **177 PB**



XROOTD	15.3 PB
FUSE	156 PB
HTTP	6.13 PB
GRIDFTP	19.6 TB

2 Billion files written for **7.63 PB**



XROOTD	1.37 PB
FUSE	4.08 PB
HTTP	2.18 PB
GRIDFTP	1.62 TB

EOS service model at CERN

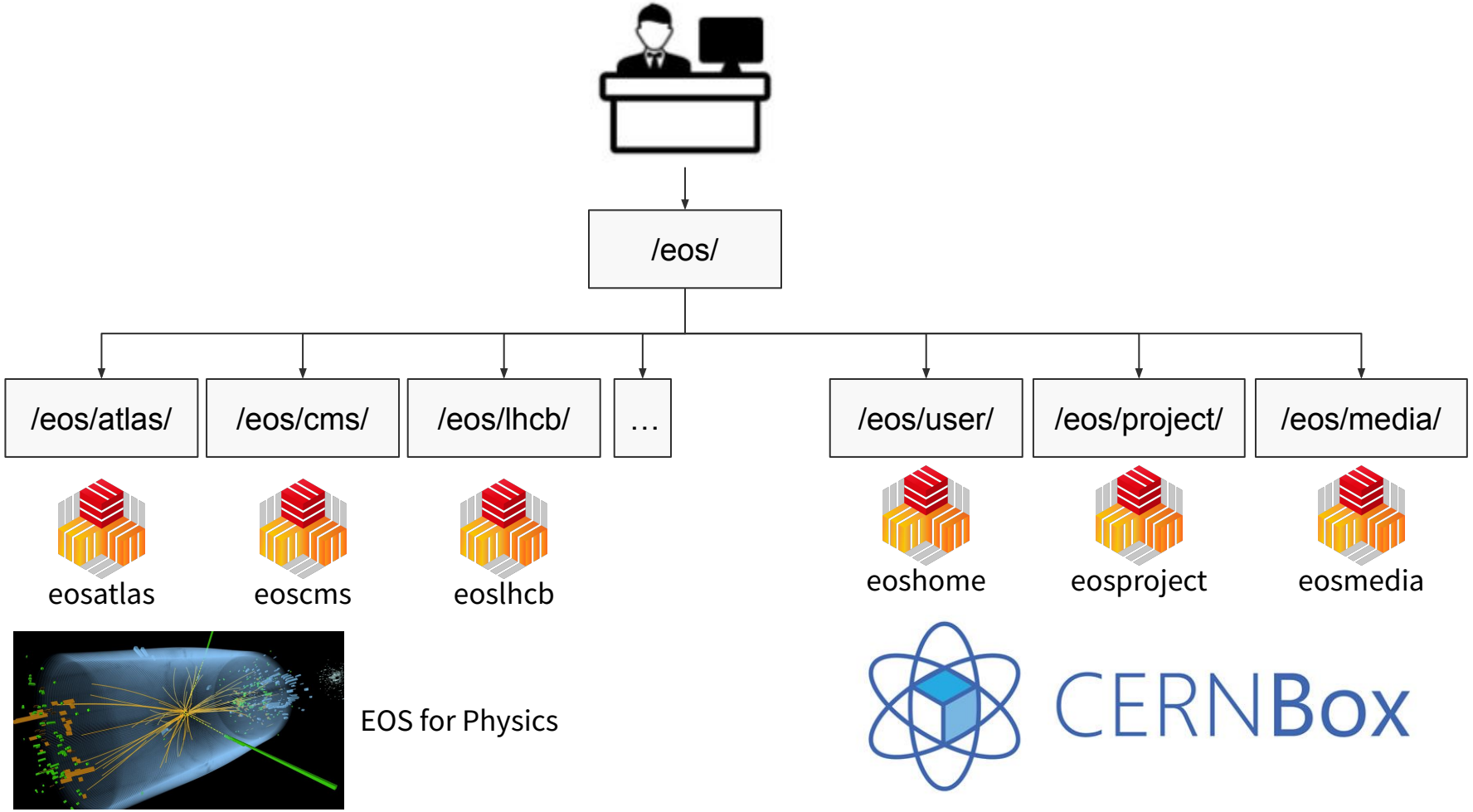
EOS Service model at CERN

Multiple EOS instances for various user communities

- Independent EOS instance for each LHC experiment + AMS (ex: EOSATLAS, EOSLHCb…)
  - One experiment activity does not impact other experiments
- Smaller experiments are hosted in the EOSPUBLIC instance
- CERNBox is split into
  - Instances for private user directories EOSHOME
  - Instances for shared project directories EOSPROJECT

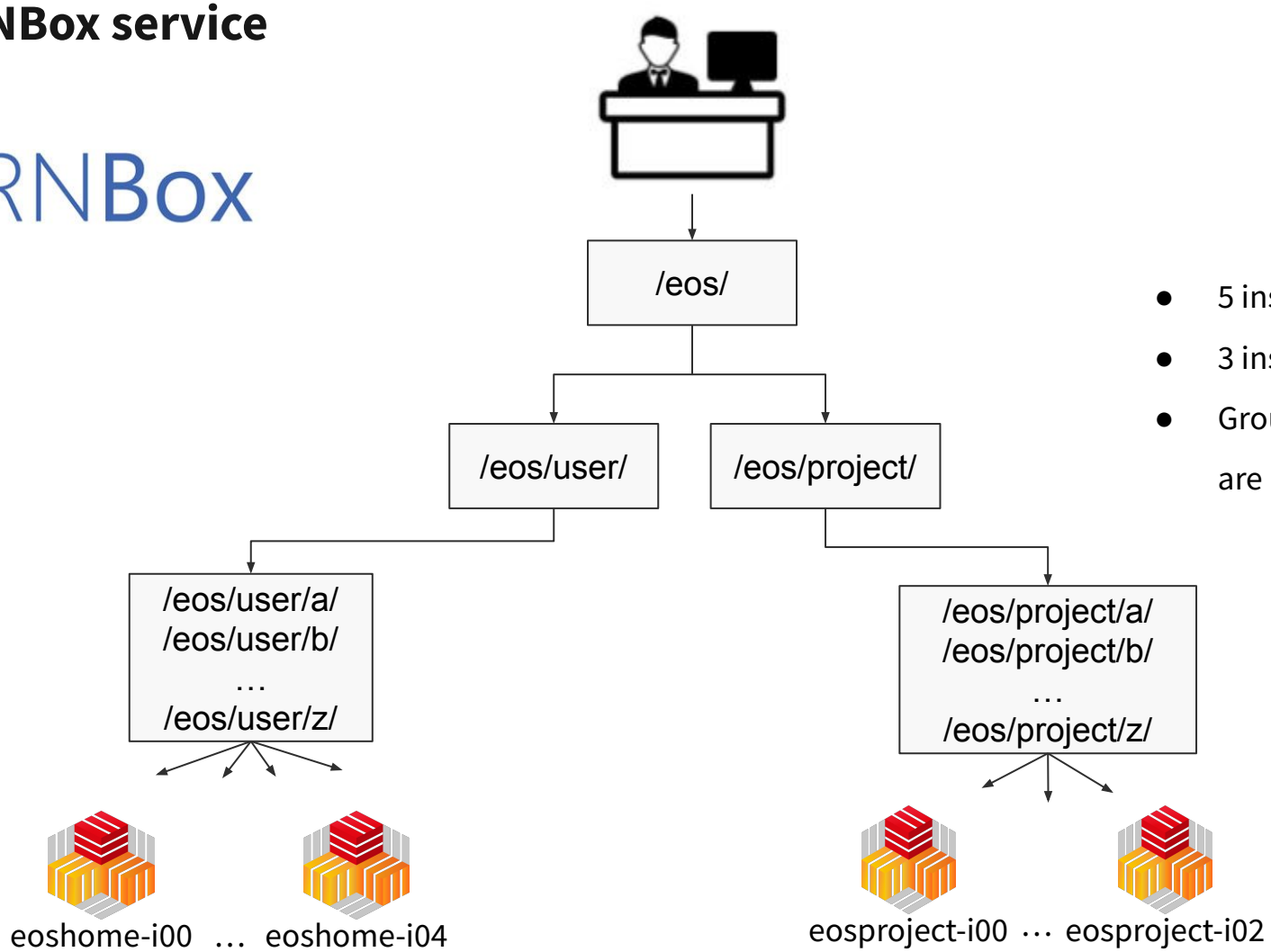
EOS Service model at CERN

Multiple EOS instances for various user communities



EOS Service model at CERN

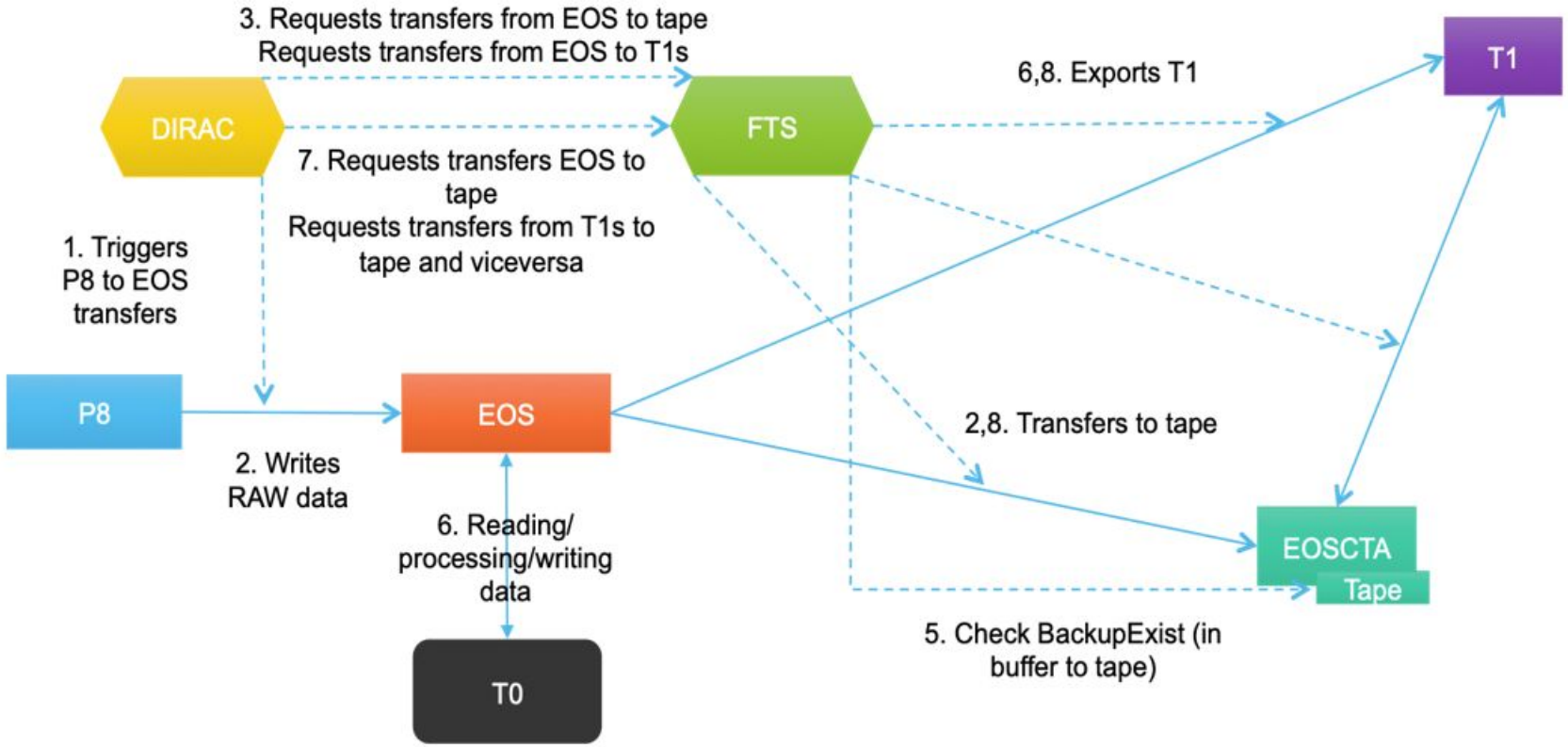
Split of the CERNBox service



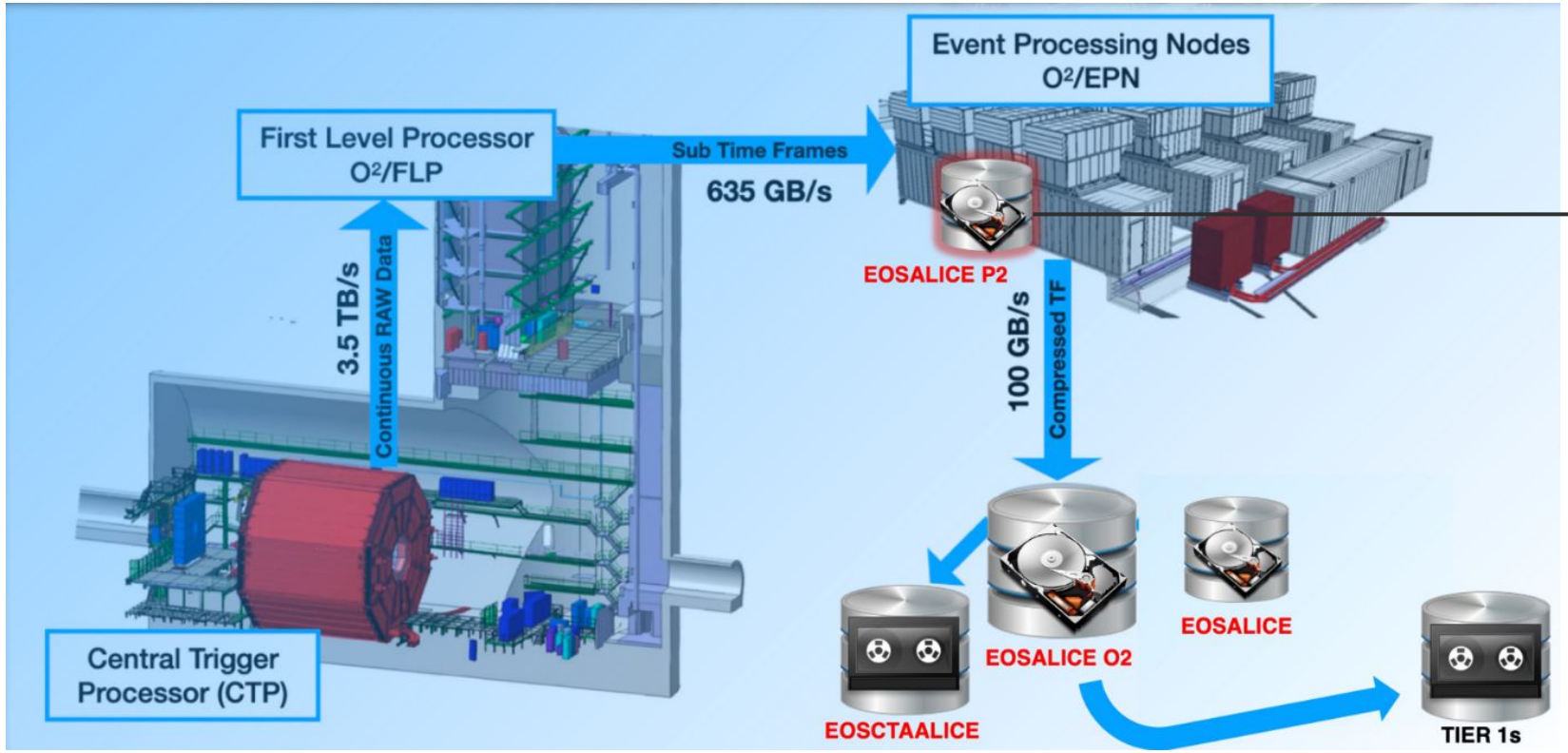
- 5 instances for /eos/user/…
- 3 instances for /eos/project/…
- Groups of letters of user and project names are mapped to 5+3 instances

LHC experiments workflows

LHC experiments workflows - example of LHCb



LHC experiments workflows - The ALICEO2 setup



- "Backup" in case of link failure between EPNs and CERN Data Center
- Sustains 100GB/s
- 13.5 PB → 1.5 day of buffer at 100GB/s

LHC experiments workflows - Run 3 expected throughput

	Experiments pits to T0 disks
ALICE	100GB/s (150GB/s)
ATLAS EXPERIMENT	8-10GB/s (12 GB/s)
CMS	13-17GB/s (20GB/s)
LHCb	10GB/s (20GB/s)

EOS software evolution enabling LHC Run 3

EOS Software evolution enabling LHC Run 3

What does that mean?

- New authentication / authorization mechanism
  - Token support
- Store more data with less hardware
  - Erasure coding
- I/O optimizations
  - I/O types, I/O priorities, bandwidth shaping
- Improving reliability
  - FSCK

EOS software evolution enabling LHC Run 3

Authn / Authz - Token support

Authentication / authorization - token support

EOS token - provided by the EOS MGM

```
$ eos root://eos.cern.ch// token --path /path/to/file.txt --expires 1681807613 --permission rwx
zteos64:MDAwMDAyMmN4nOP6z8jFXFReIfB348[....]>-4PQ%3d%3d
```


```
{
 "token": {
  "permission": "rwx",
  "expires": "1681807613",
  "owner": "ccaffy",
  "group": "it",
  "generation": "1",
  "path": "/path/to/file.txt",
  "allowtree": false,
  "vtoken": "",
  "voucher": "208f1f84-ddc6-11ed-84f6-fa163e6ca3c9",
  "requester": "[Tue Apr 18 10:50:53 2023] uid:112019[ccaffy] gid:2763[it]
tident:eosdev.25745:427@localhost name:ccaffy dn: prot:krb5 app: host:localhost
domain:localdomain geo: sudo:1",
  "origins": []
 },
 "signature": "[...]",
 "serialized": "[...]",
 "seed": 878494853
}
```

Authentication / authorization - token support

Macaroons - provided by the EOS MGM

- Request a macaroon using your X509 certificate

```
curl --cert [...] --key [...] --cacert [...] --capath [...] -X POST -H 'Content-Type: application/macaroon-request' -d '{"caveats":
["activity:UPLOAD,DELETE,LIST"], "validity": "PT3000M"}' https://eos.cern.ch//path/to/file.txt | jq -r '.macaroon'
```

```
location eosdev
identifier bc8bedfd-072c-4fea-b3bc-042cf73d8bb3
cid name:ccaffy
cid activity:READ_METADATA
cid activity:DOWNLOAD,UPLOAD,MANAGE
cid path:/path/to/file.txt
cid before:2020-01-29T15:13:35Z
signature
b8d9b5e4d09badbeb628222fc710e54a0af080c64a8c63eb3bb370c454302327
```

Token authorization (activity) = file permission (ACL) set on the file on EOS

Authentication / authorization - token support

sci-token - Provided by an IAM (Identity and Access Management) provider

- Using oidc-token tool


```
{
  "wlcg.ver": "1.0",
  "sub": "4d863cdd-5736-44a0-a03b-81ce144b5fe3",
  "aud": "https://wlcg.cern.ch/jwt/v1/any",
  "nbf": 1681818273,
  "scope": "openid profile storage.read:/ eduperson_entitlement wlcg
storage.create:/ offline_access eduperson_scoped_affiliation
storage.modify:/ email wlcg.groups",
  "iss": "https://wlcg.cloud.cnaf.infn.it/",
  "exp": 1681821873,
  "iat": 1681818273,
  "jti": "acd4f929-3294-4383-ba7a-3fadb30e8321",
  "client_id": "2002057c-bacc-4d5c-b79d-52d42a7a3596",
  "wlcg.groups": [
    "/wlcg",
    "/wlcg/xfers"
  ]
}
```

Authentication / authorization - token support

Usage - All tokens

HTTP

```
curl -x GET -H "Authorization: Bearer $TOKEN" https://eos.cern.ch//path/to/file.txt
```

XRootD

```
xrdcp ./file.txt root://eos.cern.ch//path/to/file.txt?authz=$TOKEN
```

Authentication / authorization - token support

Summary

Token type	Issuer	Permissions
EOS token	EOS MGM	Whatever is set by the token creator
Macaroon	EOS MGM (authentication with X509)	File/Parent directory permissions
WLCG scitoken	IAM Provider	Maps to a user - scopes limit permissions

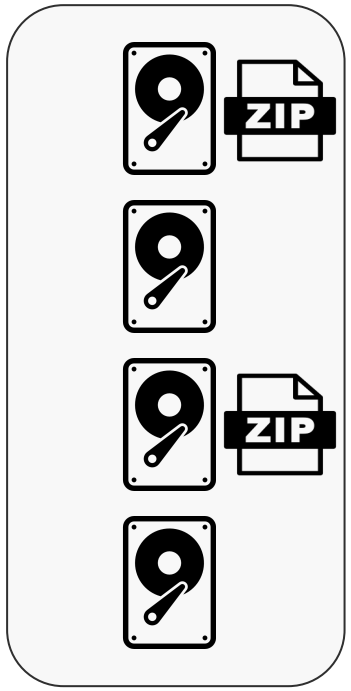
EOS software evolution enabling LHC Run 3

Store more data with less hardware

How do we ensure files availability in EOS?

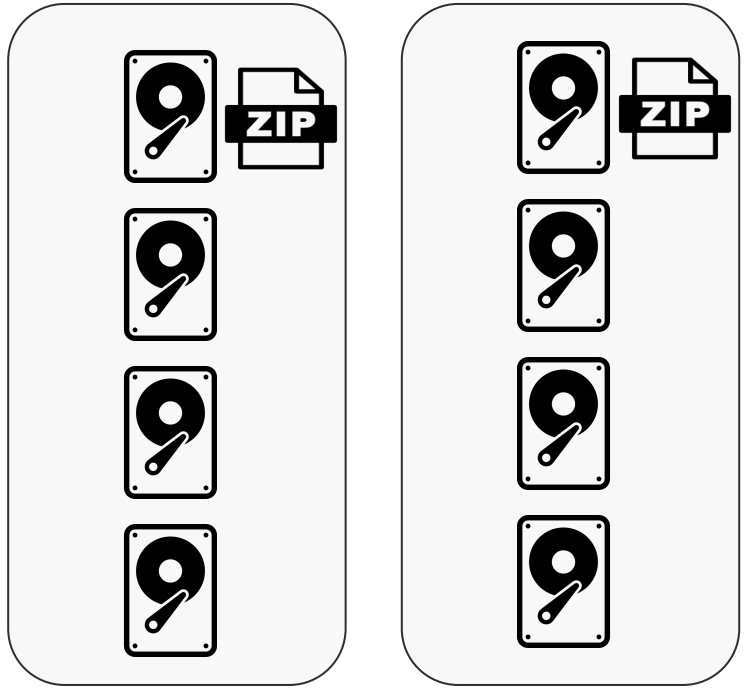
- **RAID vs RAIN**

RAID (Redundant Array of Inexpensive Disks)



ServerDisk1

RAIN (Redundant Array of Inexpensive Nodes)



ServerDisk1 ServerDisk2

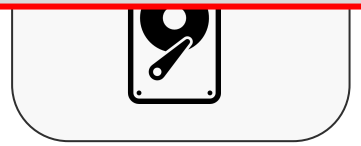
How do we ensure files availability in EOS?

- **RAID vs RAIN**

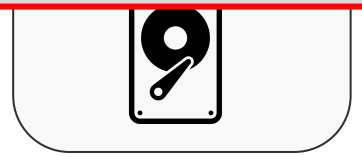
RAID (Redundant Array of Inexpensive Disks)

RAIN (Redundant Array of Inexpensive Nodes)

Storing 2 replicas is a good solution, but expensive!



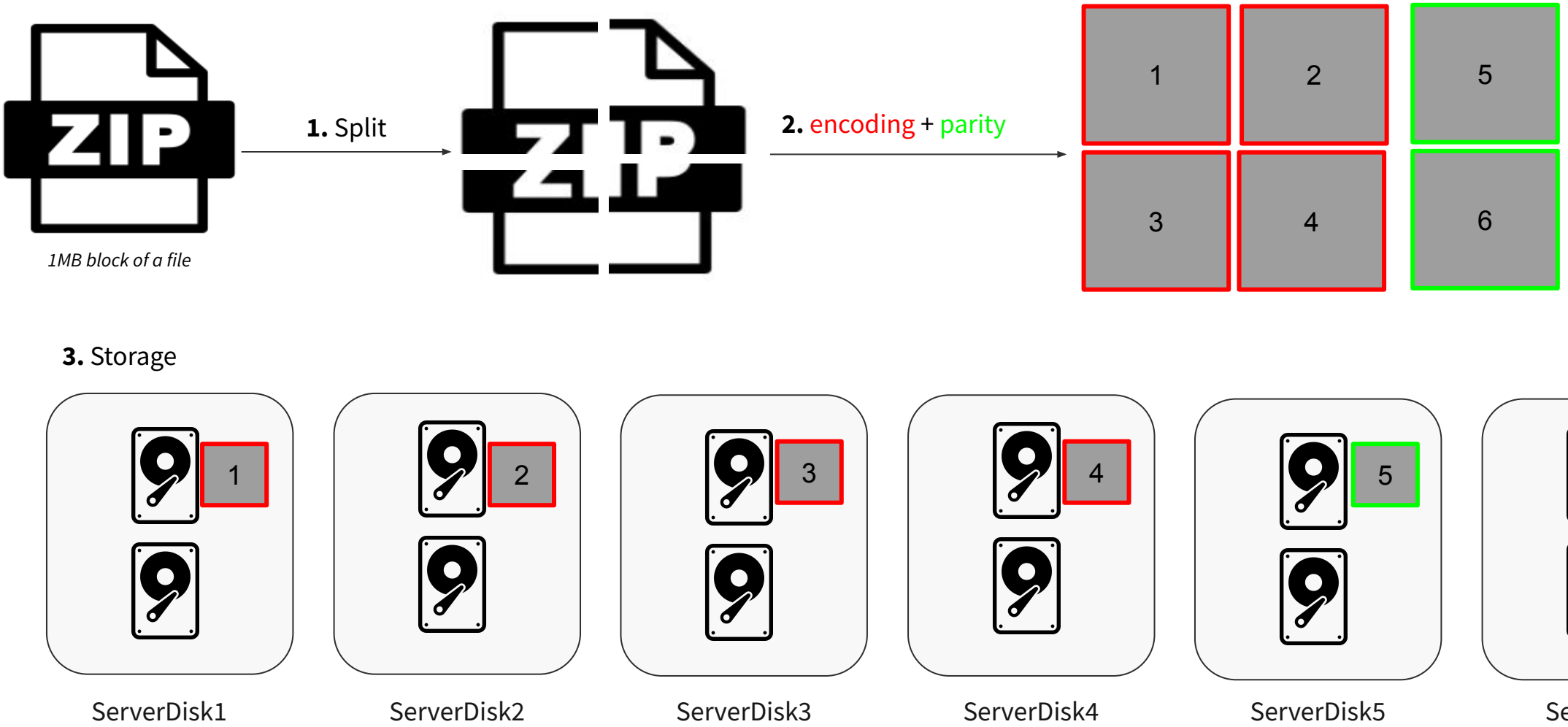
ServerDisk1



ServerDisk1

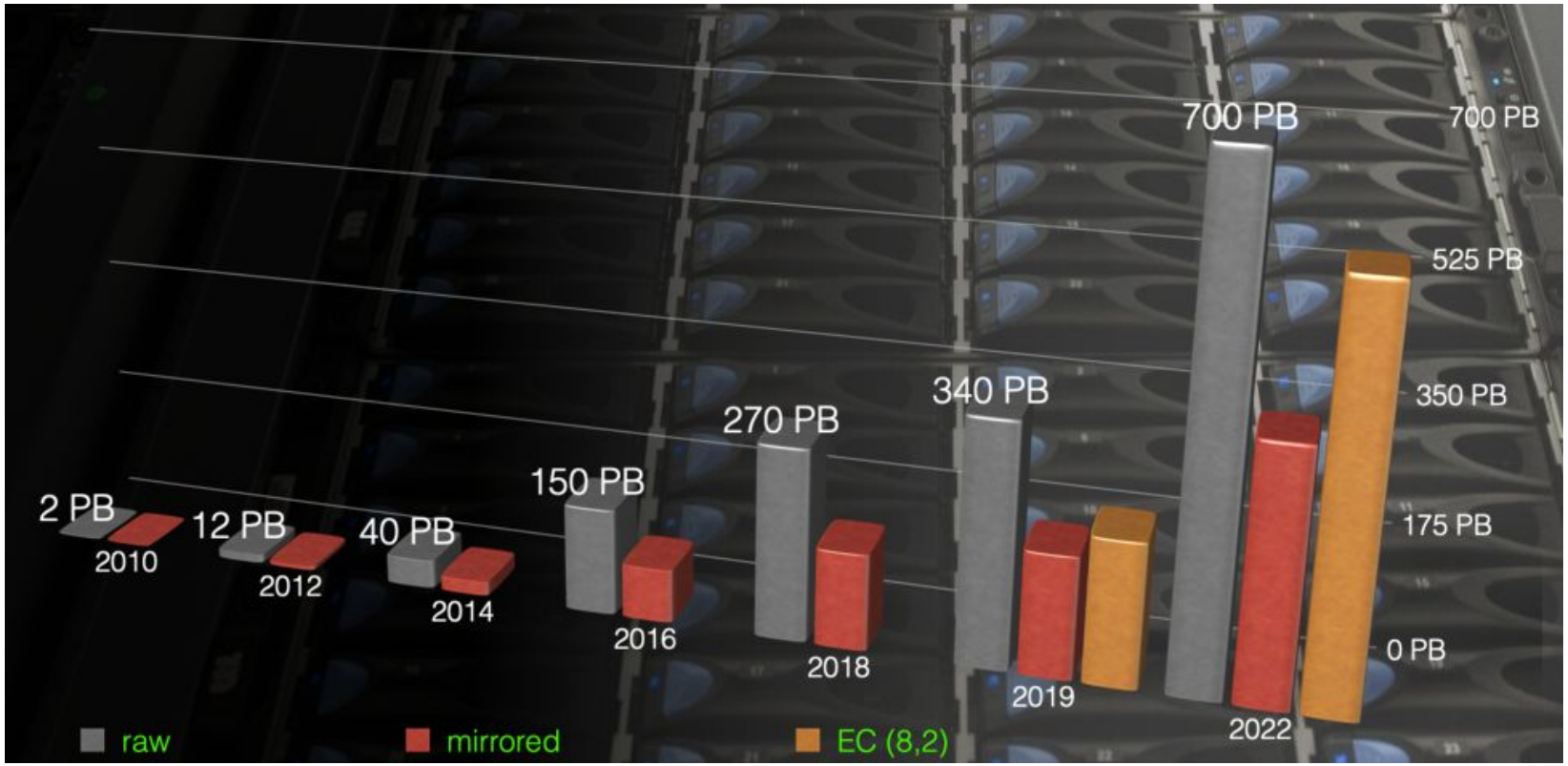
ServerDisk2

Erasure Coding

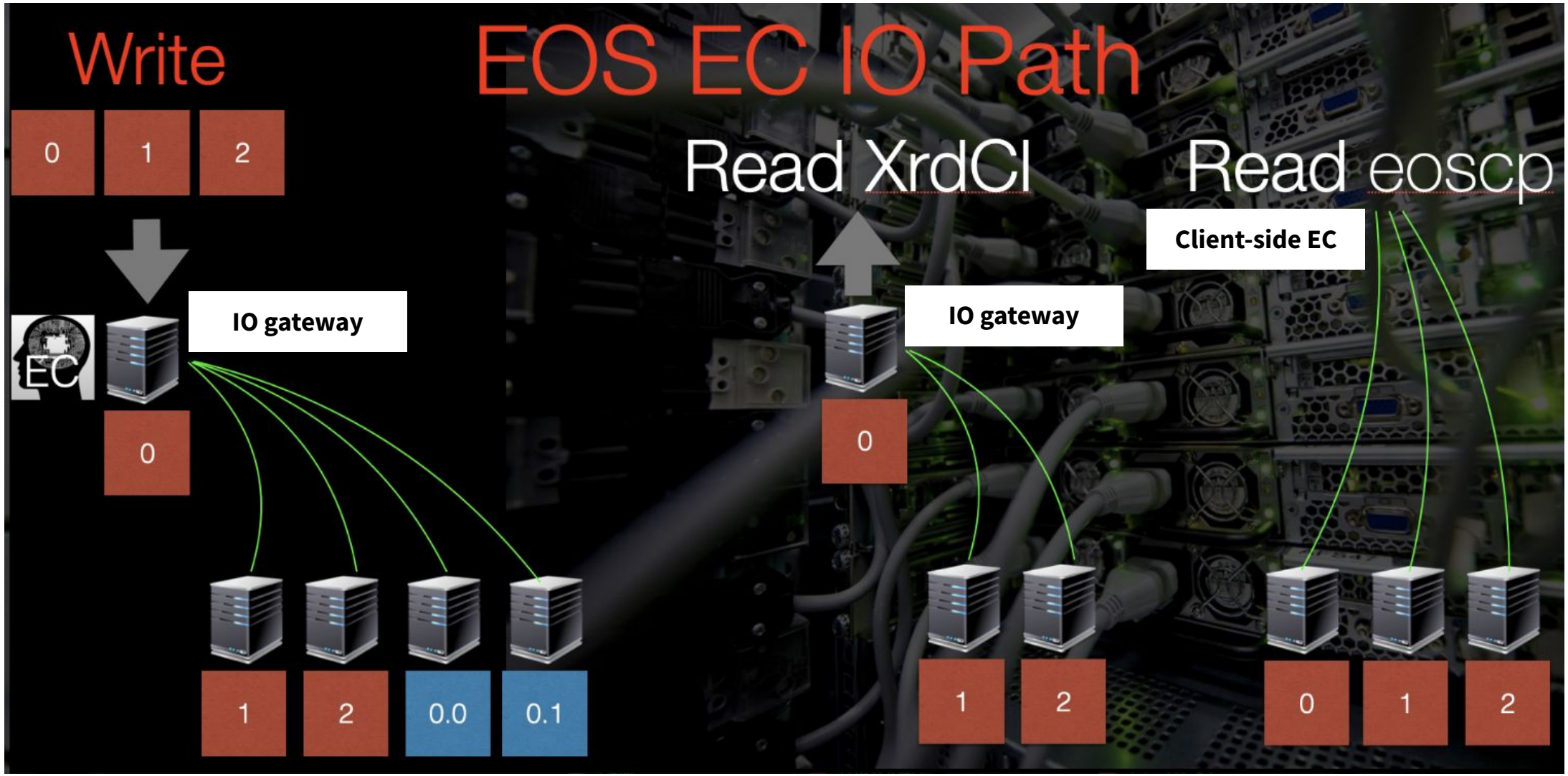


Storing more data with less hardware

Erasure coding



Storing more data with less hardware



Storing more data with less hardware

Performances of EC

- 2x traffic amplification for read using the gateway model compared to replication
- You benefit from the parallel transfers of the different disks involved in Erasure coding
  - Ex: Client-side EC read with buffered I/O is extremely performant!

Drawbacks

- CPU and network intensive
- Do not use for small files (< 100MB)

Storing more data with less hardware

Erasure coding @ CERN

Erasure Coding vs Replica percentage

	Value	Percent
Replica	11.5 PB	54%
Erasure Coding	9.83 PB	46%

AMS-02

Erasure Coding vs Replica percentage

	Value	Percent
Erasure Coding	69.1 PB	100%
Replica	114 GB	0%

ALICE AliceO2

Erasure Coding vs Replica percentage

	Value	Percent
Replica	31.6 PB	96%
Erasure Coding	1.39 PB	4%

CMS

Space recuperated

	Value
Aliceo2 Savings with Erasure Coding 10+2	55.3 PB
AMS Savings with Erasure Coding 8+2	7.37 PB
CMS Savings with Erasure Coding 10+2	1.12 PB

Total: 63.79PB

EOS software evolution enabling LHC Run 3

Improve I/O performance

Improving performance

I/O shaping

- Disk optimizations
  - I/O Types
  - I/O Priorities
- Bandwidth shaping

Improving performance

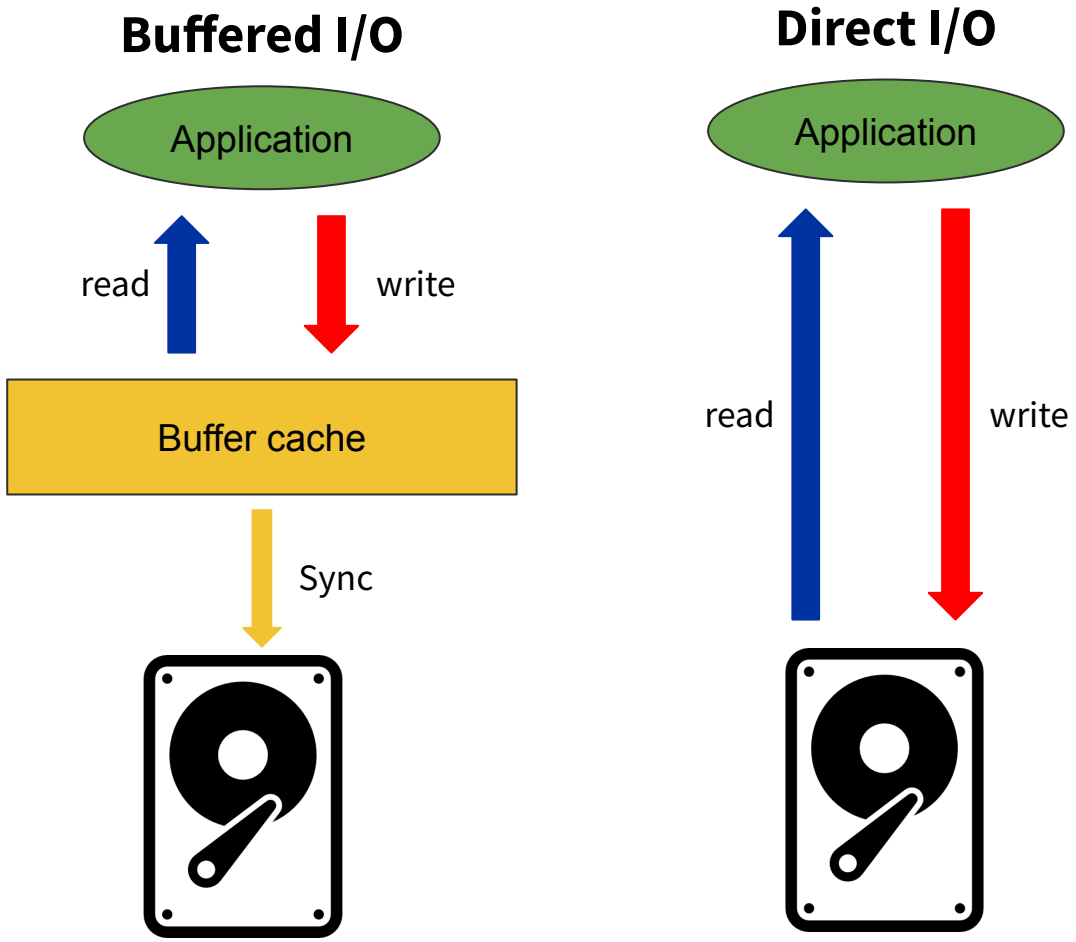
I/O types - Buffered I/O VS Direct I/O

read-ahead: anticipate file read

write-back: file is flushed from the cache on close (sync)

Allows asynchronous writes

If many files are competing for the cache, it may reduce its efficiency…



Reduces CPU/memory usage

Writes are more efficient

Read penalty as they don't benefit from the caching

Every op is synchronous

Improving performance

Direct I/O

Very good for **writes**

- Max perf of a standalone XRootD disk server increased from 7GB/s to 9GB/s
- Reduces perf tails
- Increases overall instance performance for write workloads

Not as good for reads…

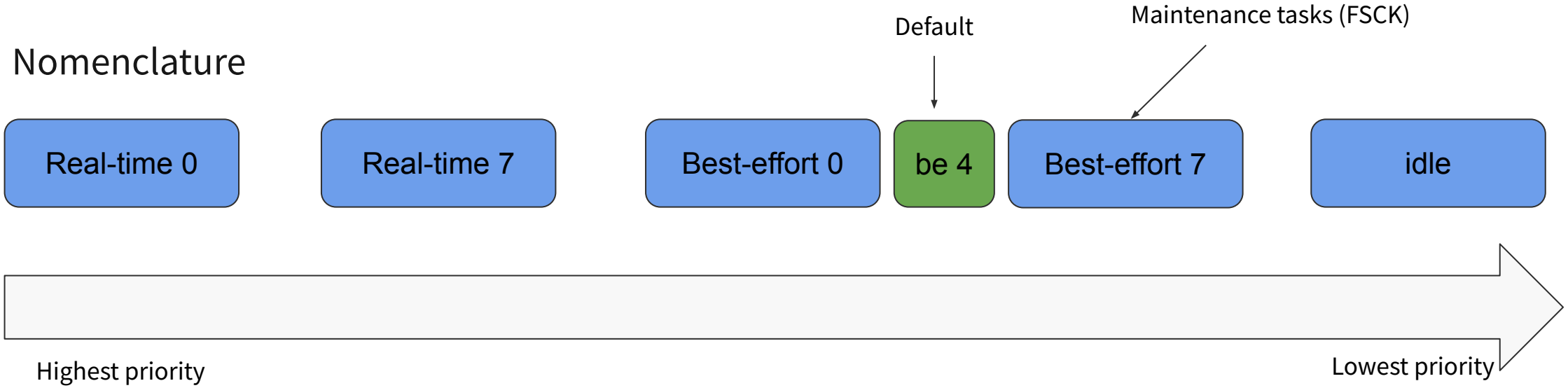
- You don't benefit from the cache

Documentation and configuration:
https://eos-docs.web.cern.ch/using/policies.html?highlight=iotype#setting-user-group-and-application-policies

Improving performance

I/O priorities

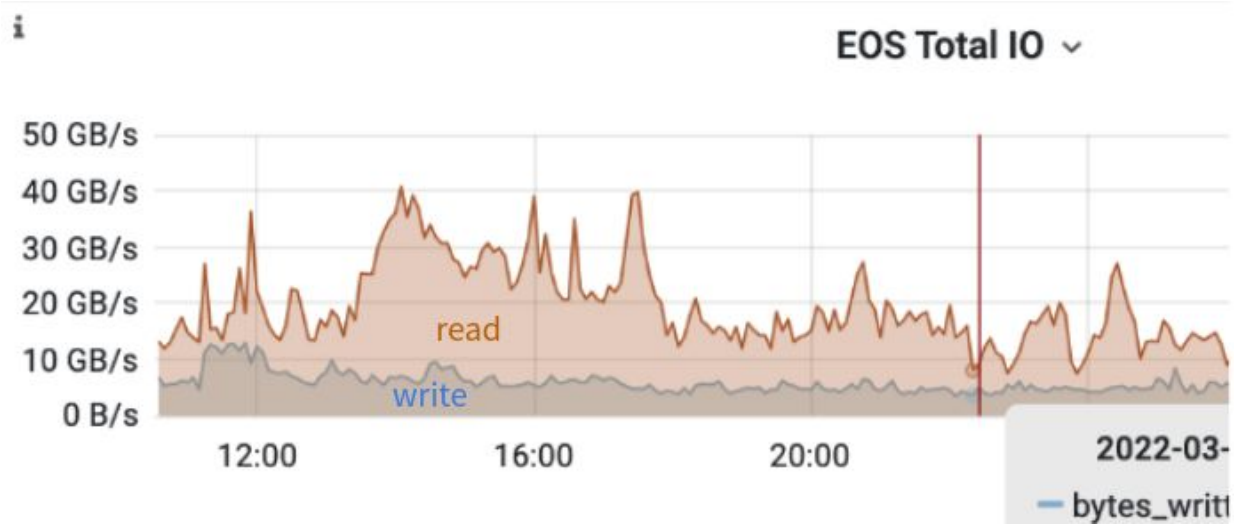
- Balance the needs for high-throughput by fairly sharing I/O requests among processes
  - Maintenance tasks can have lower priority than experiments transfers
- Works with **read** + **direct I/O writes** on devices using BFQ and CFQ scheduler
- Nomenclature



Configuration: https://eos-docs.web.cern.ch/using/priorities.html?highlight=iopriority

Improving performance

I/O priorities

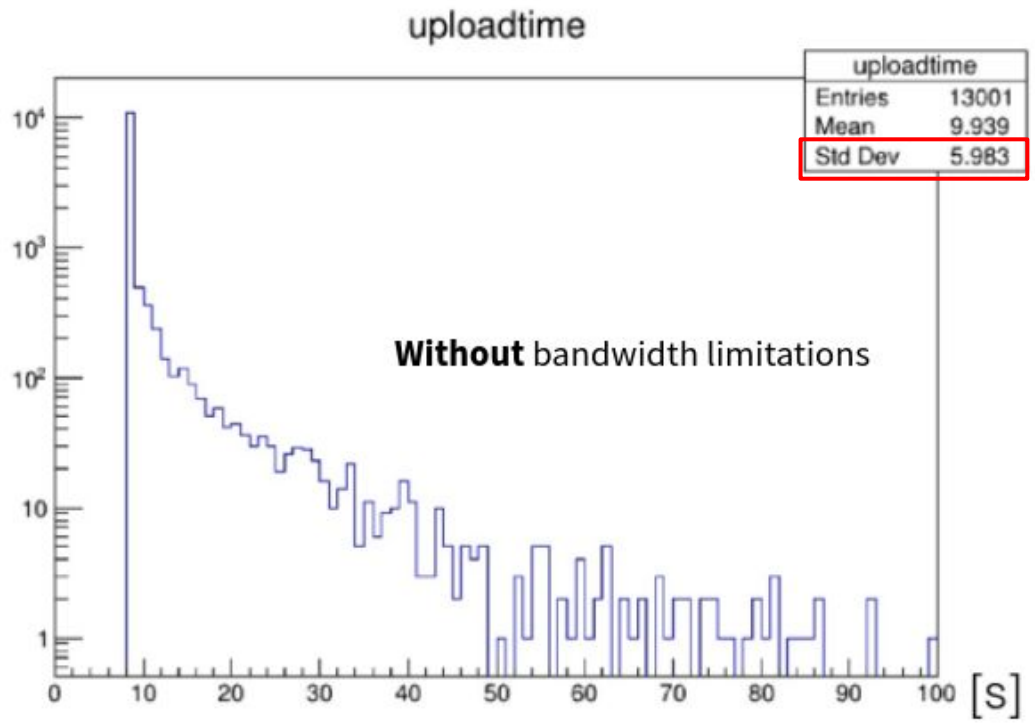


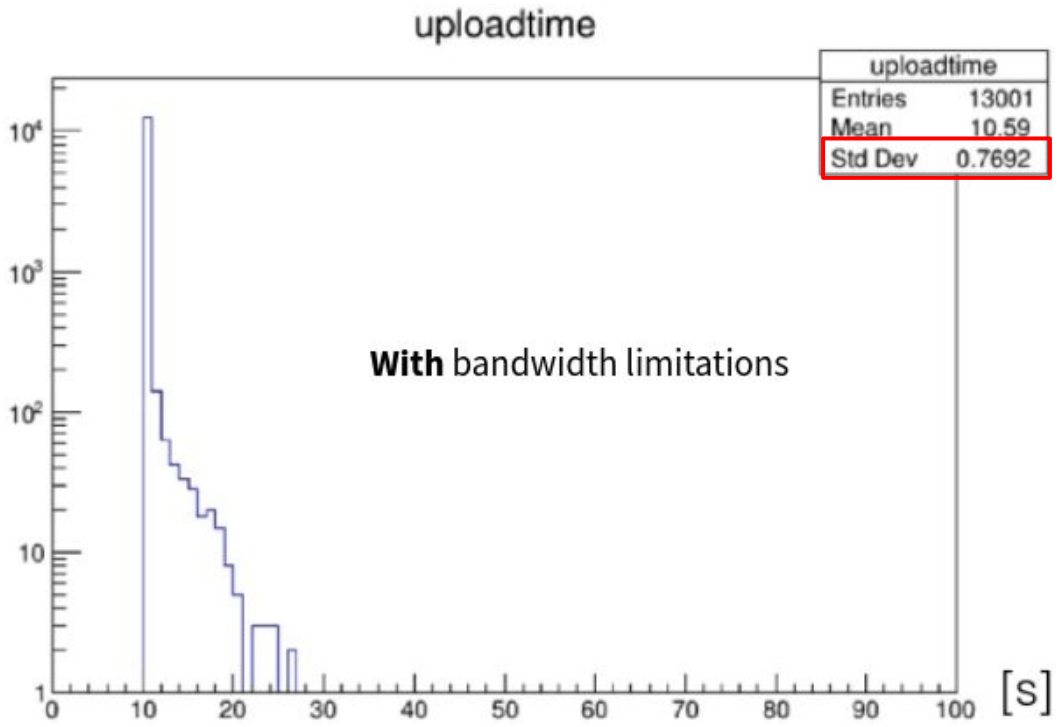
ATLAS prod workflows

Improving performance

Bandwidth regulation

- Benchmarks did show that I/O performance tails are reduced by limiting the bandwidth of clients





Configuration: https://eos-docs.web.cern.ch/using/policies.html?highlight=bandwidth

Very good for data taking use case!

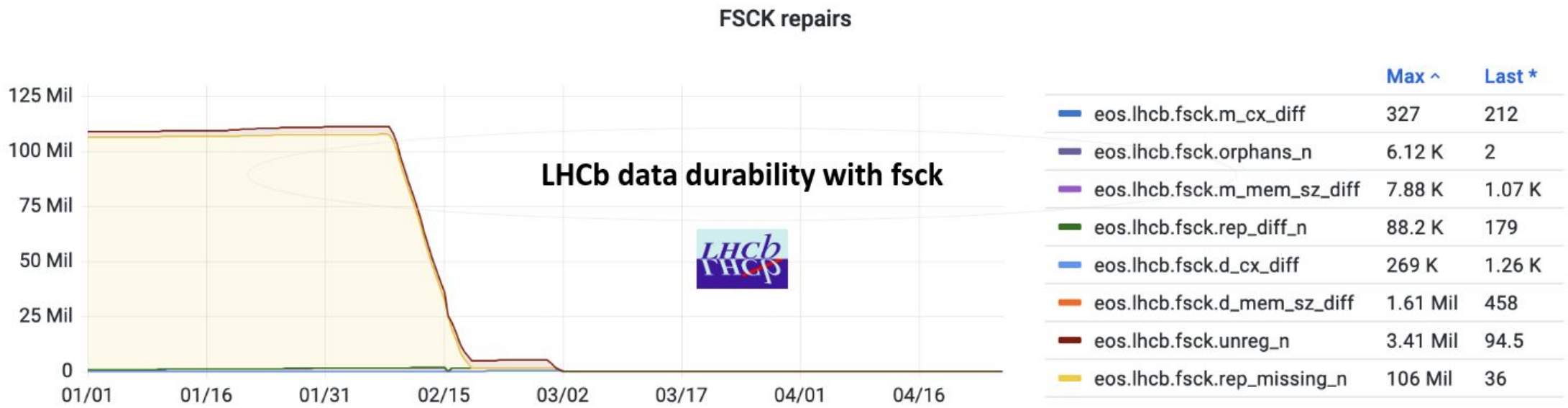
EOS software evolution enabling LHC Run 3

Improve reliability

Improving reliability

FSCK

- Automatic detection and repair of different type of recoverable errors
  - Background thread by filesystem, scanning all the files in the disk (*IO priority: be:7*)

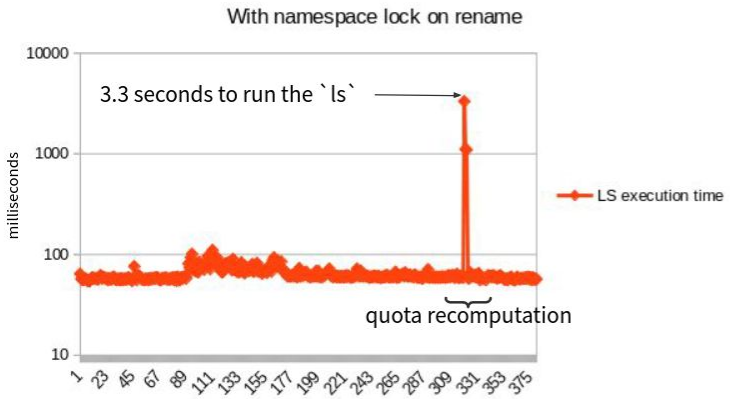


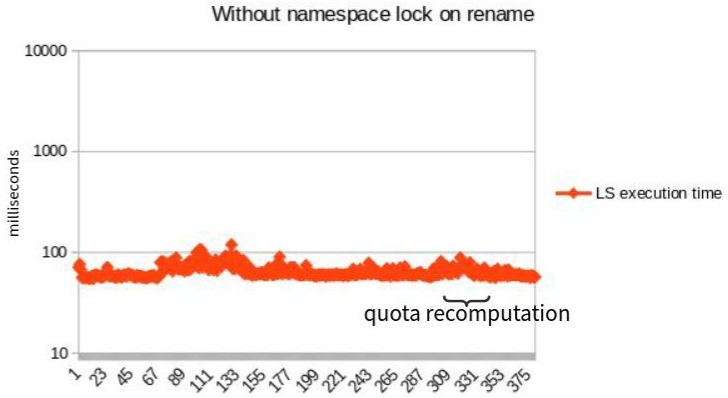
Future evolution

- **EOS 5.2**
  - What we want to drop
    - libmicrohttpd
    - LevelDB file metadata
    - Old FSCK reporting
    - Old Balancer
    - Transfer Queues/Multiplexer
    - MQ Daemon
    - Internal HTTP browser JS
  - What we will target
    - HA without MQ
    - MGM Latency Reduction (replacing global mutexes with local mutexes)
    - FLAT Scheduler
    - REST API MGM (summer project)
    - Merge SHARE API and permission system homogenisation
    - Possibly move to TPC processes instead of in-MGM multithreading with XrdCl
    - FUSE Performance
    - FST Gateway IO & scheduling for shared backends
    - EC Updates (when XRootD range-clone/copy-on-write functionality is available)

Future evolution

- **Start to think about Run 4!**
  - MGM performance improvements → finer-grained namespace locking
    - Goal: improve the parallelism of namespace operations





run `ls` in a infinite loop and compute the time it takes to execute. In the meantime, move a directory with 10k files to another place using `eos mv`

  - API evolution
    - Study feasibility/benefits of bulk API for opening files
  - Data format evaluation
    - Study RNTuple format with erasure coding
    - Study what range of file size are sustainable

Conclusion

- EOS is running smoothly at CERN for T0 activities during Run 3
  - Multiple EOS instances for various user communities
- The different evolution of the EOS software offers new ways to authenticate but also ways to improve I/O performance and the reliability
  - Token authn/authz
  - Erasure coding
  - I/O optimization
  - FSCK
- The EOS team is constantly improving the EOS software in order to meet future Run 4 needs!
  - Stay tuned!

CERN
Tape Archive

The CERN Tape Archive (CTA) : an efficient storage system for HEP data archival

Status and future evolution

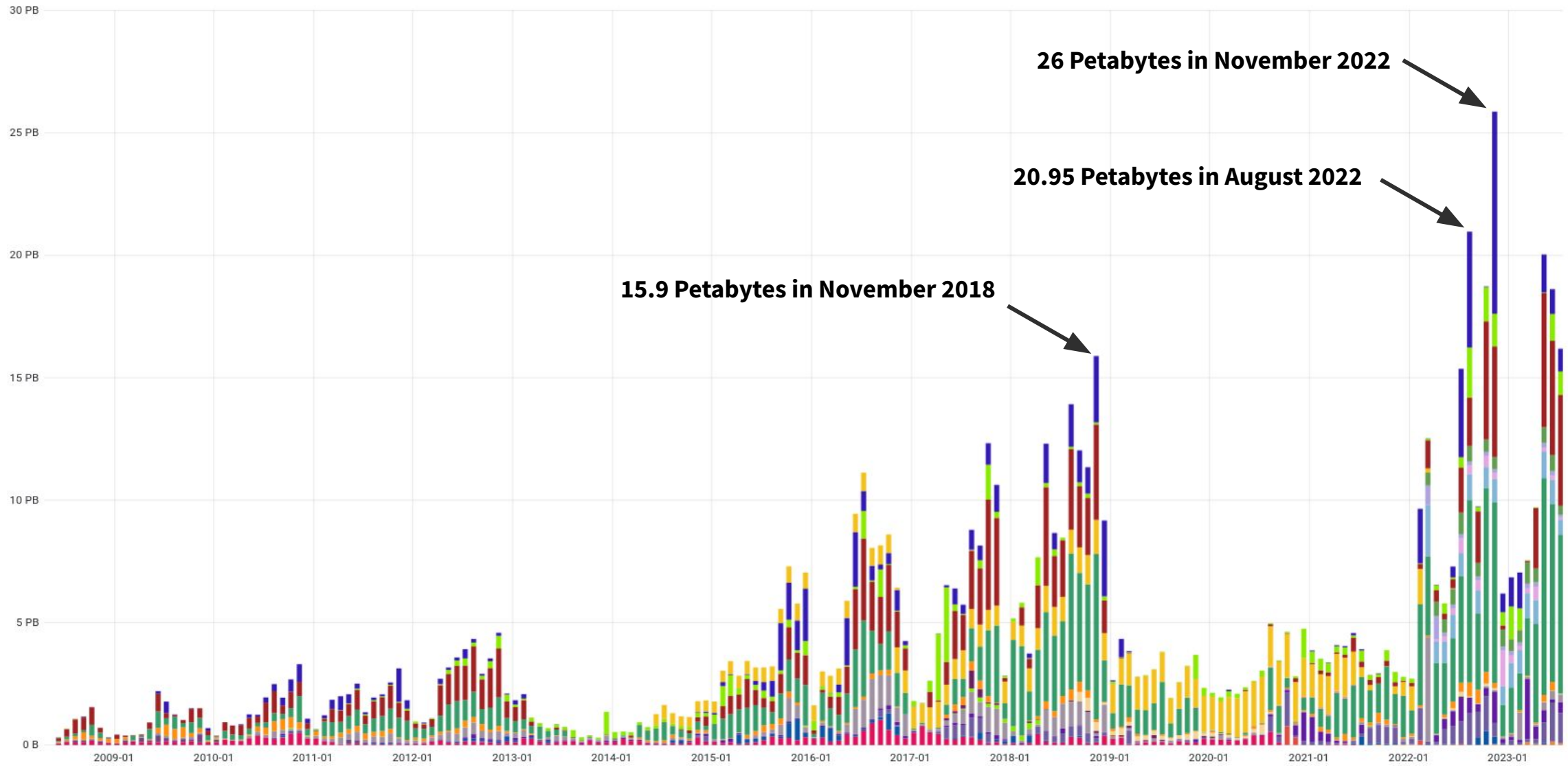
Presented by Cedric Caffy on behalf of the CTA team

2023-07-24

CTA - CERN Tape Archive

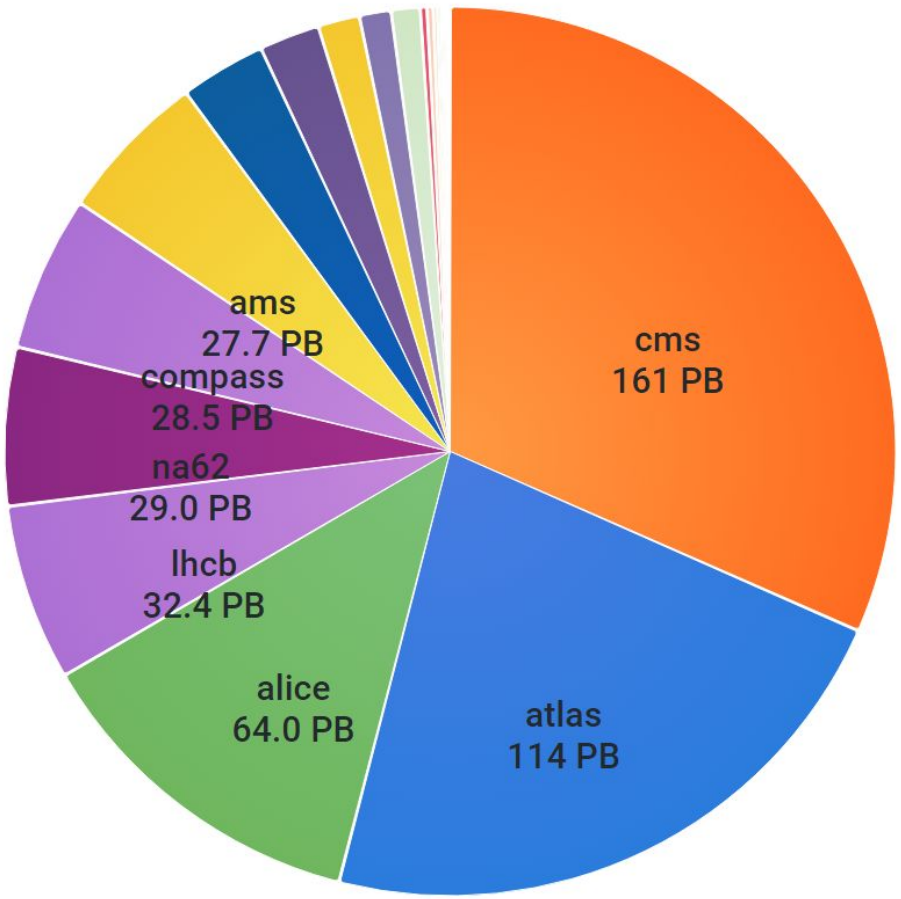
Tape backend to EOS

Monthly archive volume records for CTA T0



CERN T0 tape archival volume per VO

Data volume by VO: ~550 PB



	Value
cms	161 PB
atlas	114 PB
alice	64.0 PB
lhcb	32.4 PB
na62	29.0 PB
compass	28.5 PB
ams	27.7 PB
na61	15.5 PB
ntof	10.8 PB
ilc	7.31 PB
dune	5.47 PB
preservation	4.95 PB
cast	0.836 PB
totem	0.725 PB
public	0.403 PB
faser	0.120 PB
backup	0.0842 PB
it	0.0557 PB
isolde	0.0153 PB
spacal	2.62e-7 PB

CERN T0 production tape hardware

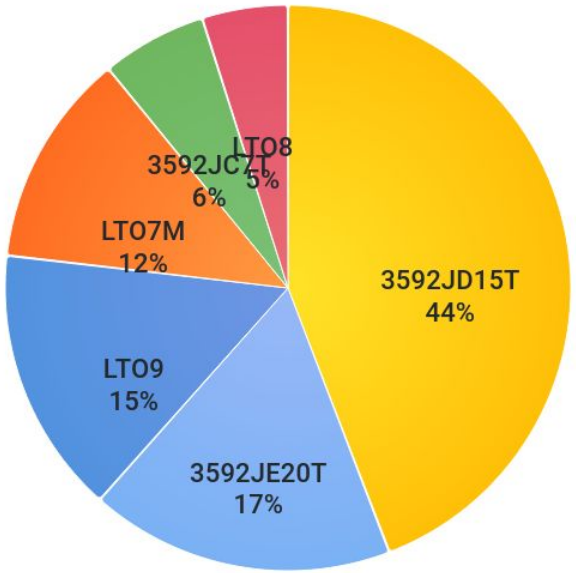
5 tape libraries:

- **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



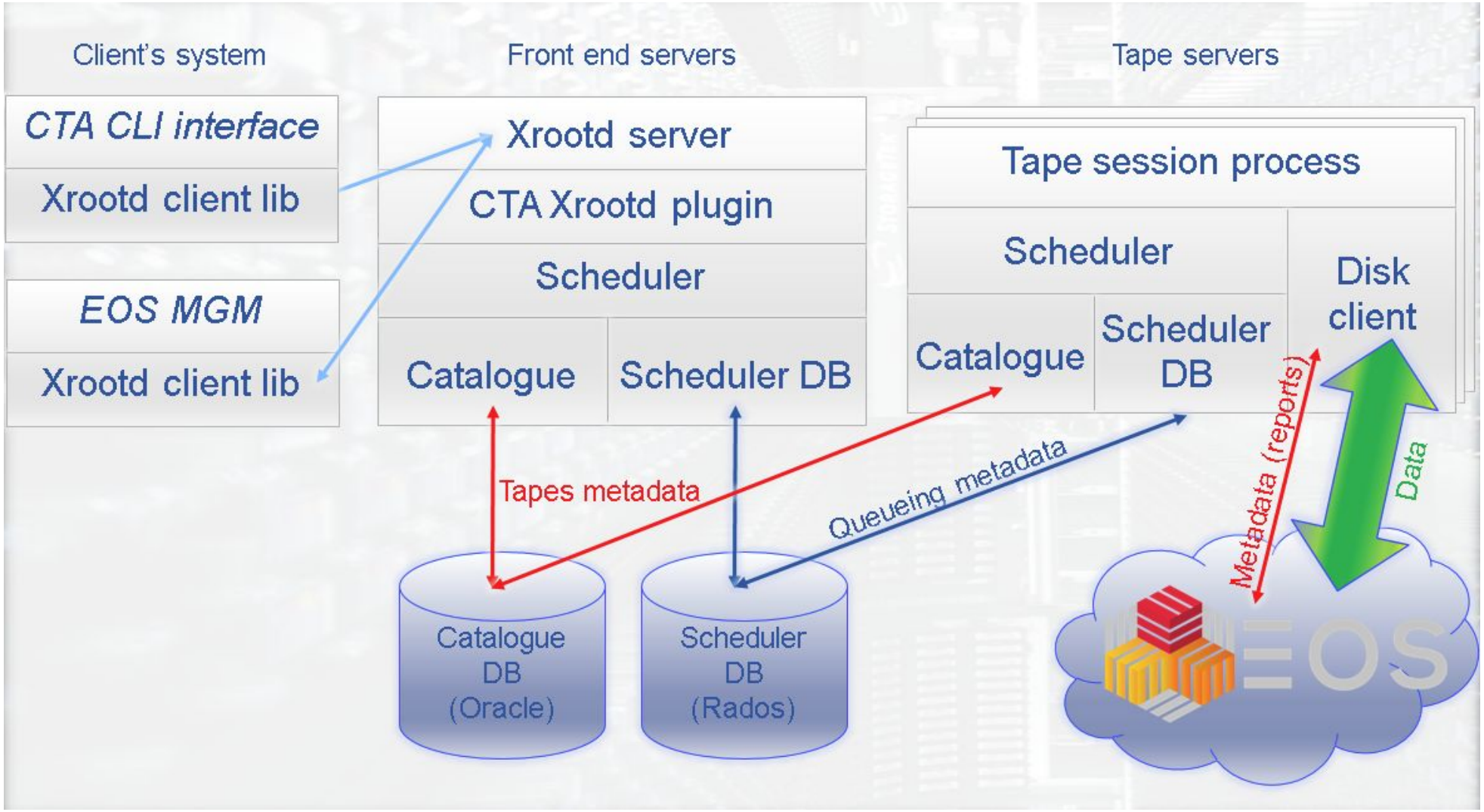
Capacity distribution per media type

3592JD15T	224 PB
3592JE20T	87.9 PB
LTO9	77.1 PB
LTO7M	61.7 PB
3592JC7T	30.1 PB
LTO8	24.5 PB

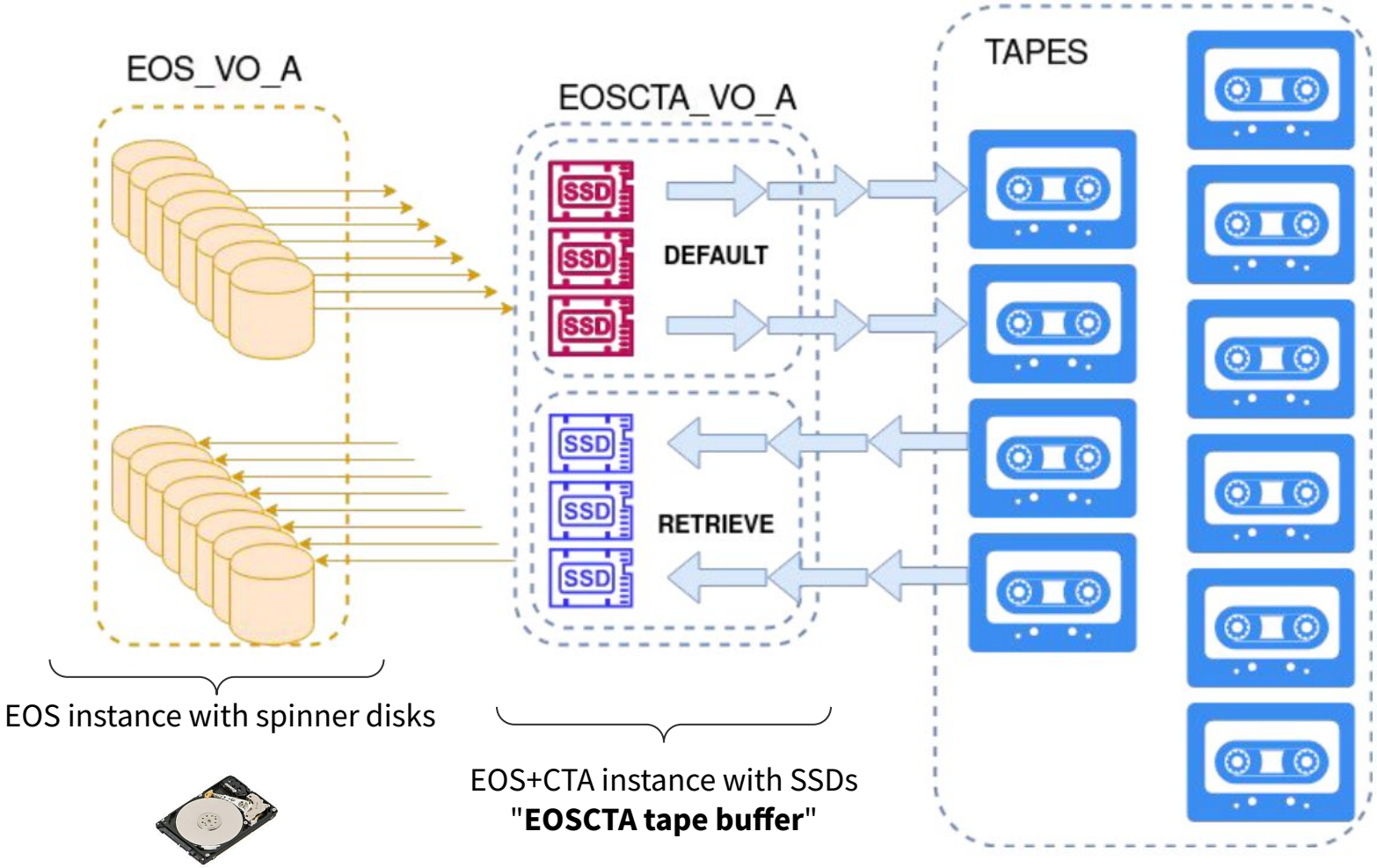
Capacity helper

3592JE20T	20 TB
LTO9	18 TB
3592JD15T	15 TB
LTO8	12 TB
LTO7M	9 TB
3592JC7T	7 TB

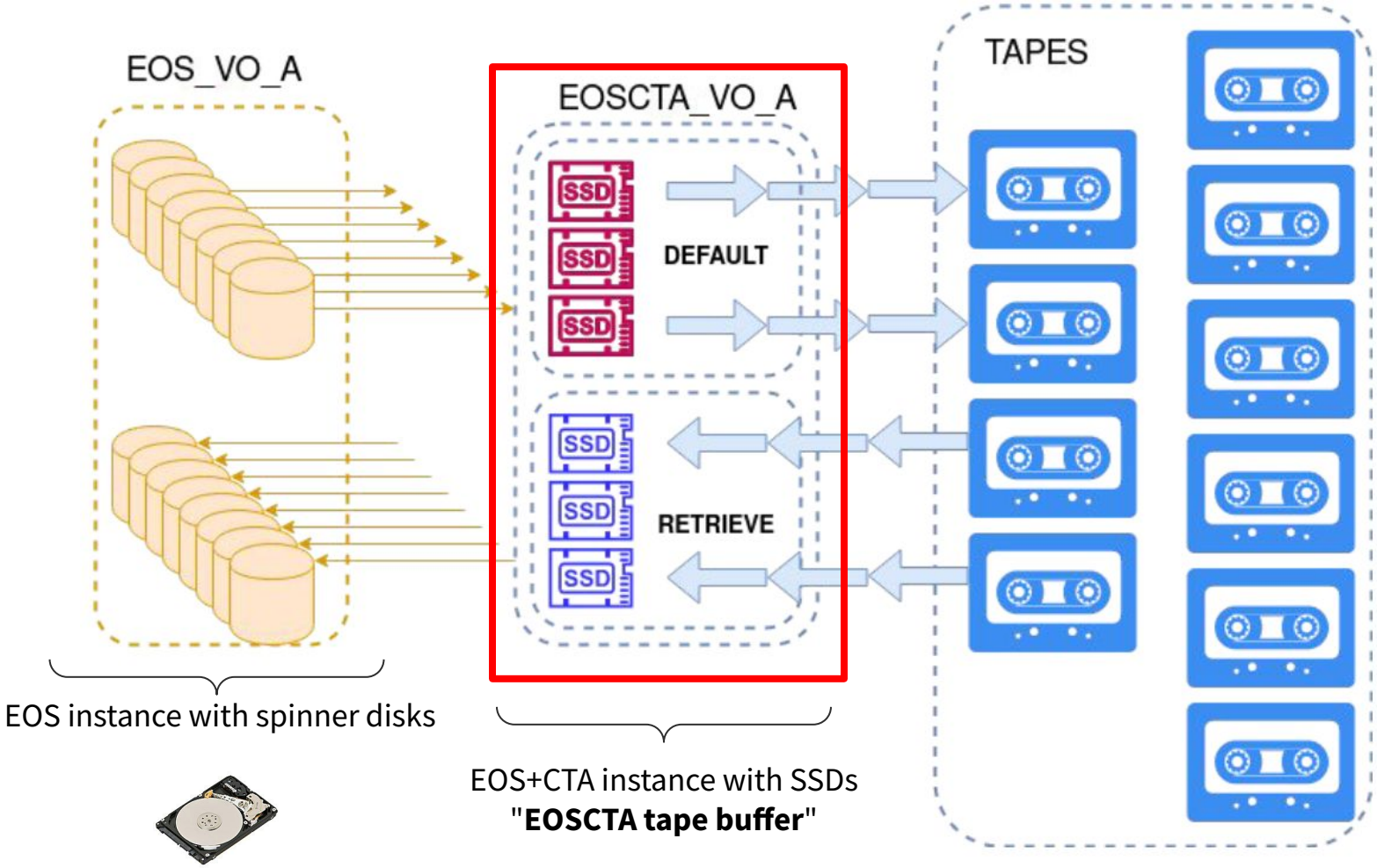
CTA architecture



EOS + CTA architecture @ CERN



EOS + CTA architecture @ CERN



The 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
  - Bandwidth: 1.2Tb/s or 150GB/s of full duplex

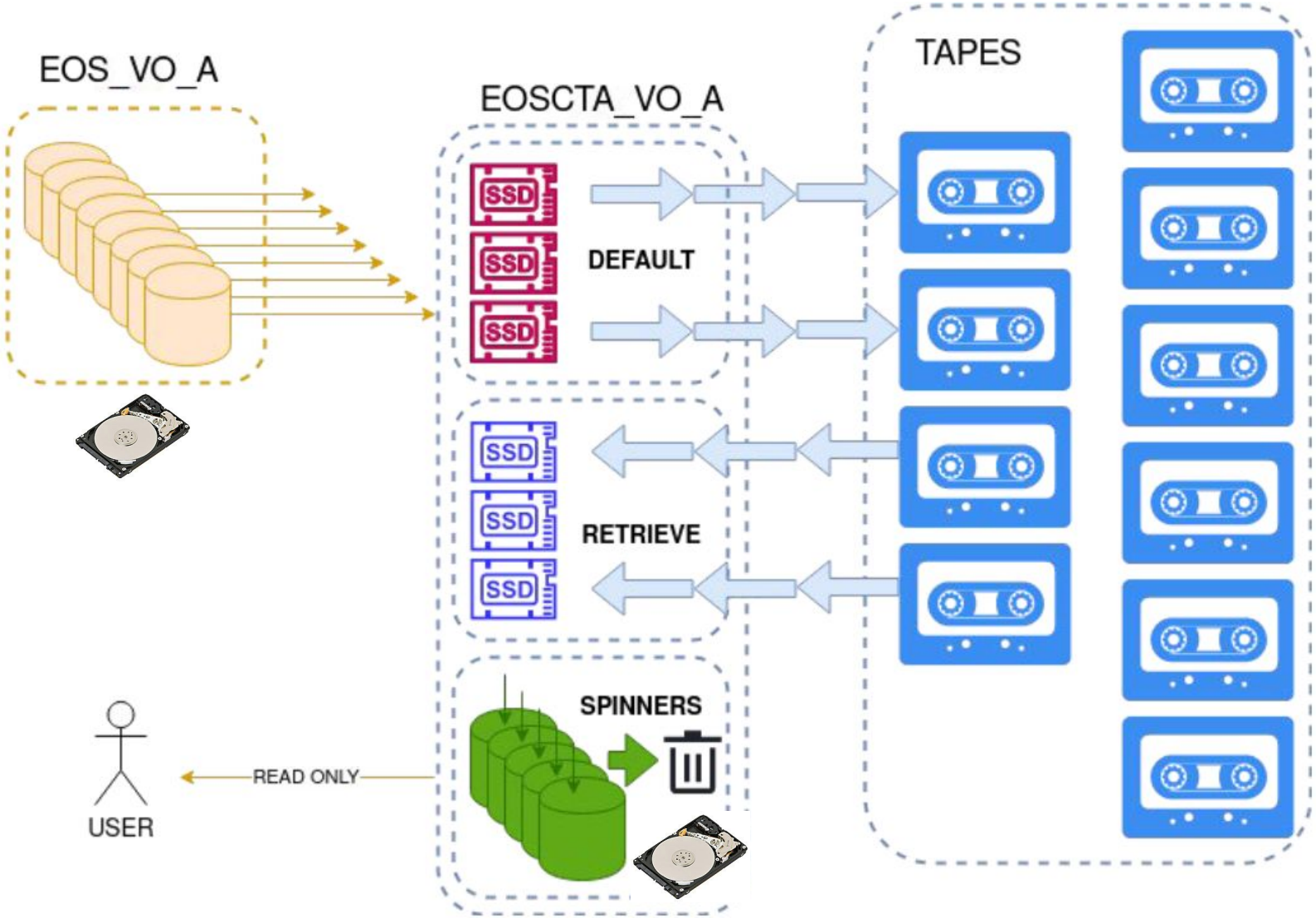
Total SSD buffer capacity = 2PB

The EOSCTA tape buffer

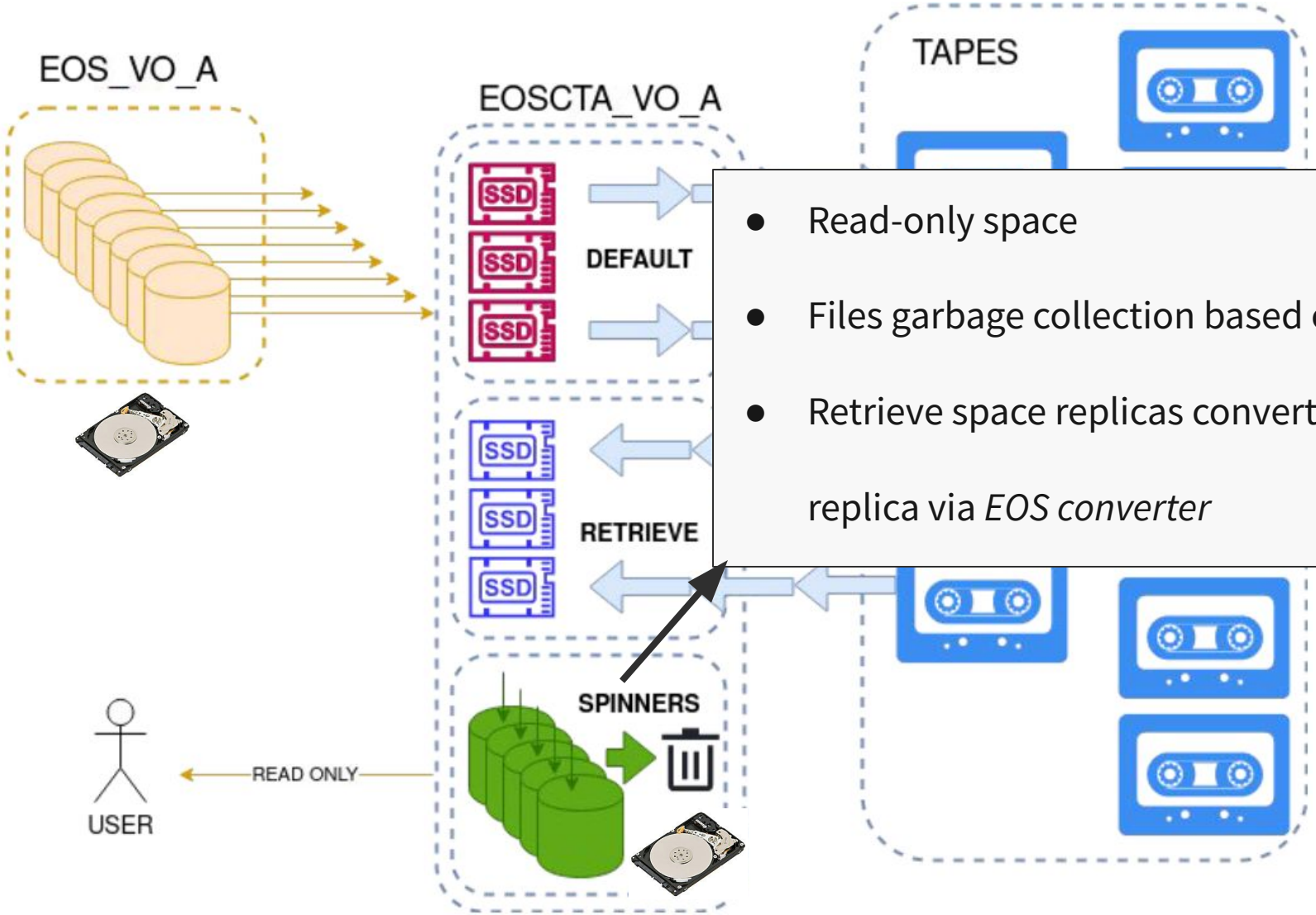
Properties

- Move files to/from tape
  - Very efficient → cannot afford data redundancy
- File eviction
  - Archive workflow: as soon as the data is safely archived on tape
  - Retrieve workflow
    - As soon as the data has been copied somewhere else (handled by FTS)
    - Garbage collection
- Up to 8h of buffer to tape

EOS + CTA architecture with *spinners* addon



EOS + CTA architecture with *spinners* addon



- Read-only space
- Files garbage collection based on LRU policy
- Retrieve space replicas converted to Spinners space replica via *EOS converter*

CTA Run 3 operations

- LHC experiments needs

	EOS to EOS + CTA T0 tapes
ALICE	10 GB/s
ATLAS EXPERIMENT	10 GB/s
CMS	10 GB/s
LHCb	10 GB/s

- Tape infrastructure **shared** between all experiments
- Bandwidth is ensured by allocating tape drives to each VO

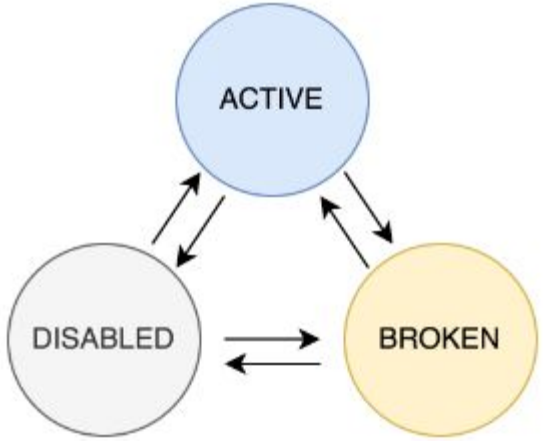
CTA Run 3 operations

Data transfer protocols available

- XRootD
- HTTP
  - Requires CTA version >= [4,5].8.7-1 and the configuration of EOS HTTP transfers with XrdHttp
  - Support for the WLCG HTTP tape REST API
    - Enabled in production for all LHC instances

CTA Run 3 operations

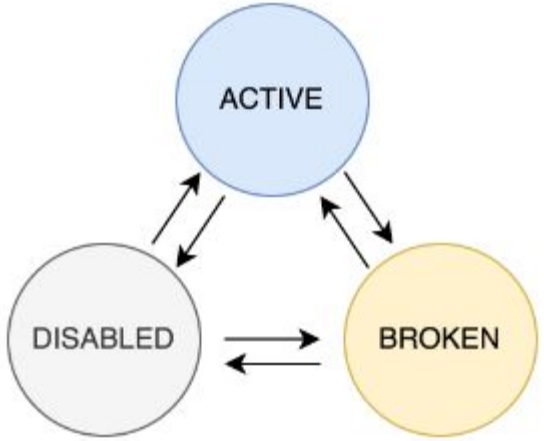
New tape lifecycle (CTA >= 4.8.0)

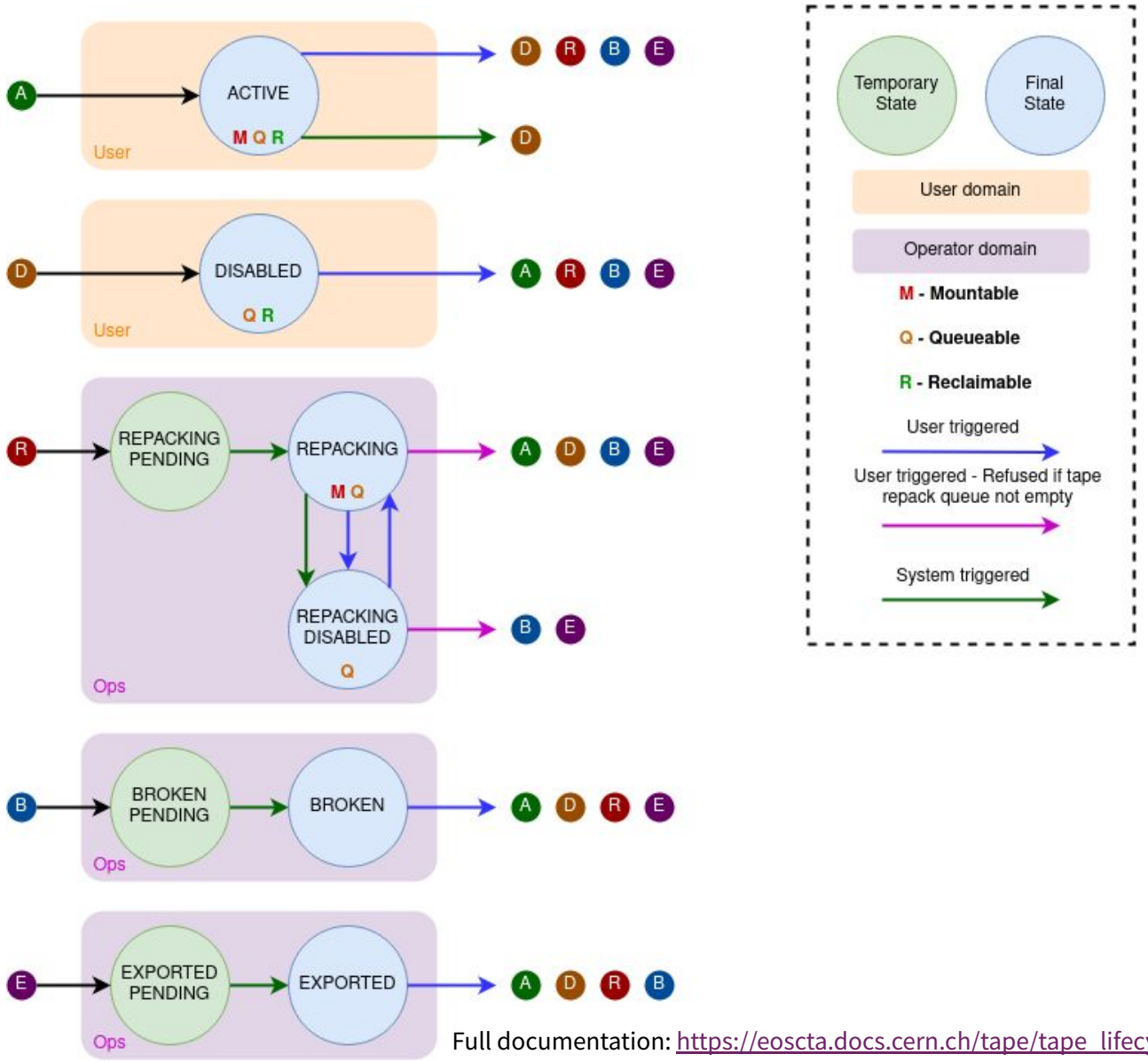


- No clear state for Repack
- Mix of user-requests and repack-requests
- User requests may be queued indefinitely
- No state for exported tapes

CTA Run 3 operations

New tape lifecycle (CTA >= 4.8.0)





- No clear state for Repack
- Mix of user-requests and repack-requests
- User requests may be queued indefinitely
- No state for exported tapes

Full documentation: https://eoscta.docs.cern.ch/tape/tape_lifecycle/

CTA Run 3 operations

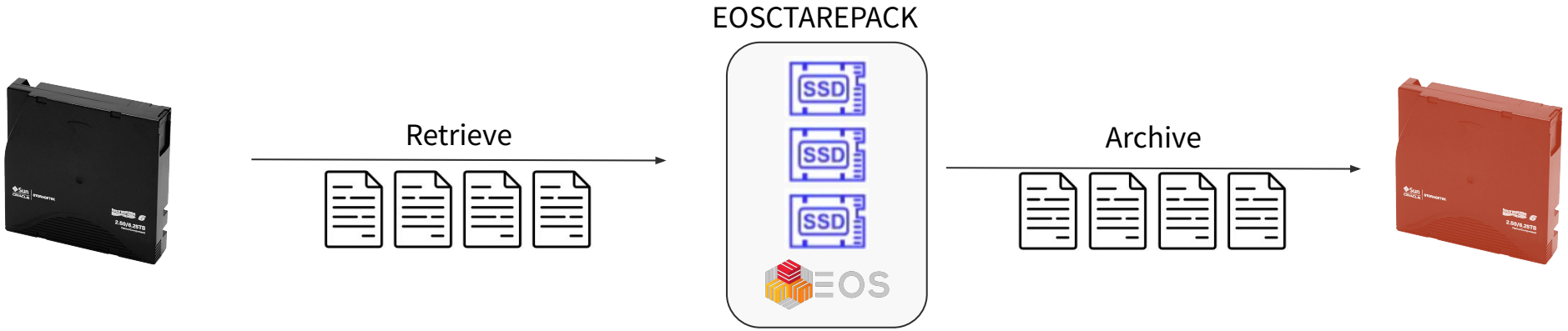
Repack

- Very important maintenance operation
- Use cases
  - Long-term data preservation - move data from old tapes to new ones to preserve the reliability of the storage
  - Recuperate free space after deletion
  - Archive manually-repaired files to CTA
  - Replicate files

CTA Run 3 operations

Repack

- Use the CTA repack engine
  - *cta-admin repack add --vid V01001*
- What does it do (simplified)?



CTA Run 3 operations

Repack

- CTA repack engine
  - Works well for few set of tapes :)
  - Problem
    - Missing other steps needed by operations → It only repacks tapes!
    - Limited ways to manage overload

CTA Run 3 operations

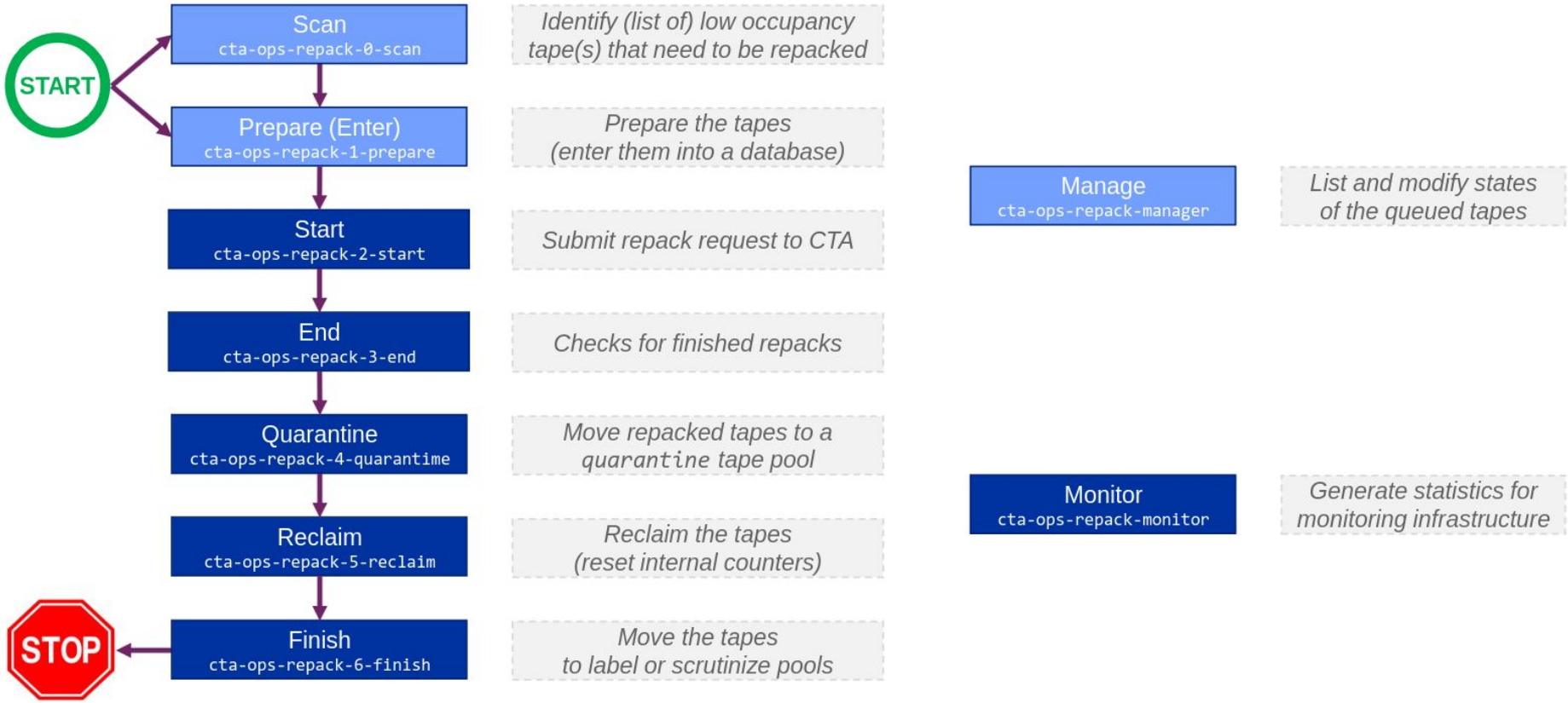
Repack - ATRESYS (Automated Tape REpacking SYStem)

- Identify what tapes need to be repacked
- Handles large number of tapes
  - Controls submission to CTA repack engine
    - Uses the new tape lifecycle!
- Manages tape quarantine
  - For future eventual reclaims (reset tape counters to 0)

CTA Run 3 operations

Repack - ATRESYS (Automated Tape REpacking SYStem)

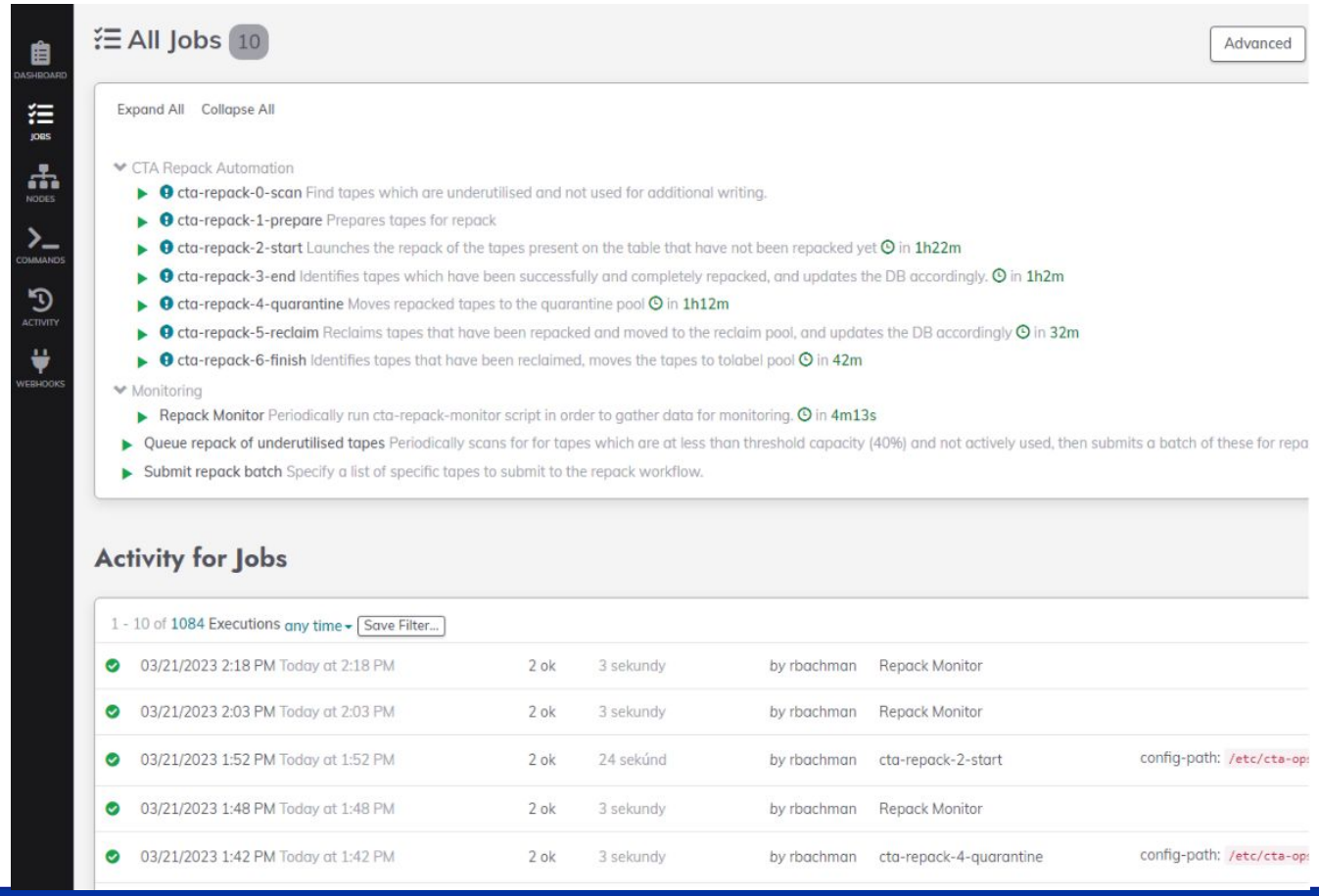
- Process architecture



CTA Run 3 operations

Repack - ATRESYS (Automated Tape REpacking SYStem)

- Orchestration with Rundeck
  - Runs every 2 hours



CTA Run 3 operations

Repack - ATRESYS (Automated Tape REpacking SYStem)

- cta-ops-repack-manager

```
[root@ctaproductionfrontend02 ~]# cta-ops-repack-manager ls
Tape    Status                          Media      Pool                 ytes    Usage  Last Written        Mode    Priority
I72148  6/6 Finished 2023-03-21 09:12   3592JE20T  tolabel_IBM3JE      3.49T      17%  2022-07-14 05:31    auto         low
L87414  6/6 Finished 2023-03-21 03:12   LTO9       tolabel_SPC1L9      1.36T       8%  2022-07-31 23:40    auto         low
L87119  6/6 Finished 2023-03-21 03:12   LTO9       tolabel_IBM1L       1.09T       6%  2022-07-21 20:01    auto         low
L86107  6/6 Finished 2023-03-20 21:12   LTO9       tolabel_SPC1L9      3.61T      20%  2022-08-29 21:47    auto         low
I74408  4/6 Quarantined 2023            3592JE20T  quarantine        331.45G       2%  2022-12-11 04:38    auto         low
I73297  2/6 Started 2023                JE20T      r_backup_afs_2      3.61T      18%  2022-07-24 10:23    auto         low
I73195  2/6 Started 2023                E20T       r_backup_afs_2    207.33G       1%  2022-06-09 22:59    auto         low
L87387  2/6 Started 2023-03-            LTO9       r_backup_afs_1      6.55T      36%  2022-07-08 11:34    auto         low
L85813  2/6 Started 2023-0 19 09:52     LTO9       r_backup_afs_1      6.98T      39%  2022-08-24 23:02    auto         low
L87457  2/6 Started 2023-03-16 14:53    LTO9       r_backup_afs_1       7.0T      39%  2022-08-27 00:02   error         low
L87993  1/6 Entered 2023-03-16 14:51    LTO9       r_backup_afs_1    793.74G       4%  2022-07-23 20:13    auto         low
L87571  1/6 Entered 2023-03-16 14:51    LTO9       r_backup_afs_1      4.13T      23%  2022-07-07 15:08    auto         low
L87131  1/6 Entered 2023-03-16 14:51    LTO9       r_backup_afs_1     86.19G       0%  2022-06-16 21:47    auto         low
I74430  1/6 Entered 2023-03-16 14:51    3592JE20T  r_backup_afs_2     76.12G       0%  2022-06-05 19:21    auto         low
I75403  1/6 Entered 2023-03-16 14:51    3592JE20T  r_backup_afs_2      6.99T      35%  2022-08-14 19:22    auto         low
I75399  1/6 Entered 2023-03-16 14:51    3592JE20T  r_backup_afs_2      7.13T      36%  2022-08-18 20:14    auto         low
I74375  1/6 Entered 2023-03-16 14:51    3592JE20T  r_backup_afs_2      6.27T      31%  2022-07-25 04:3     auto         low
I74346  1/6 Entered 2023-03-16 14:51    3592JE20T  r_backup_afs_2    291.11G       1%  2022-06-23 22:1     auto         low
I74310  1/6 Entered 2023-03-16 14:51    3592JE20T  r_backup_afs_2    720.69G       4%  2022-07-07 22:      auto         low
```

Tape in quarantine

Tapes transition between states

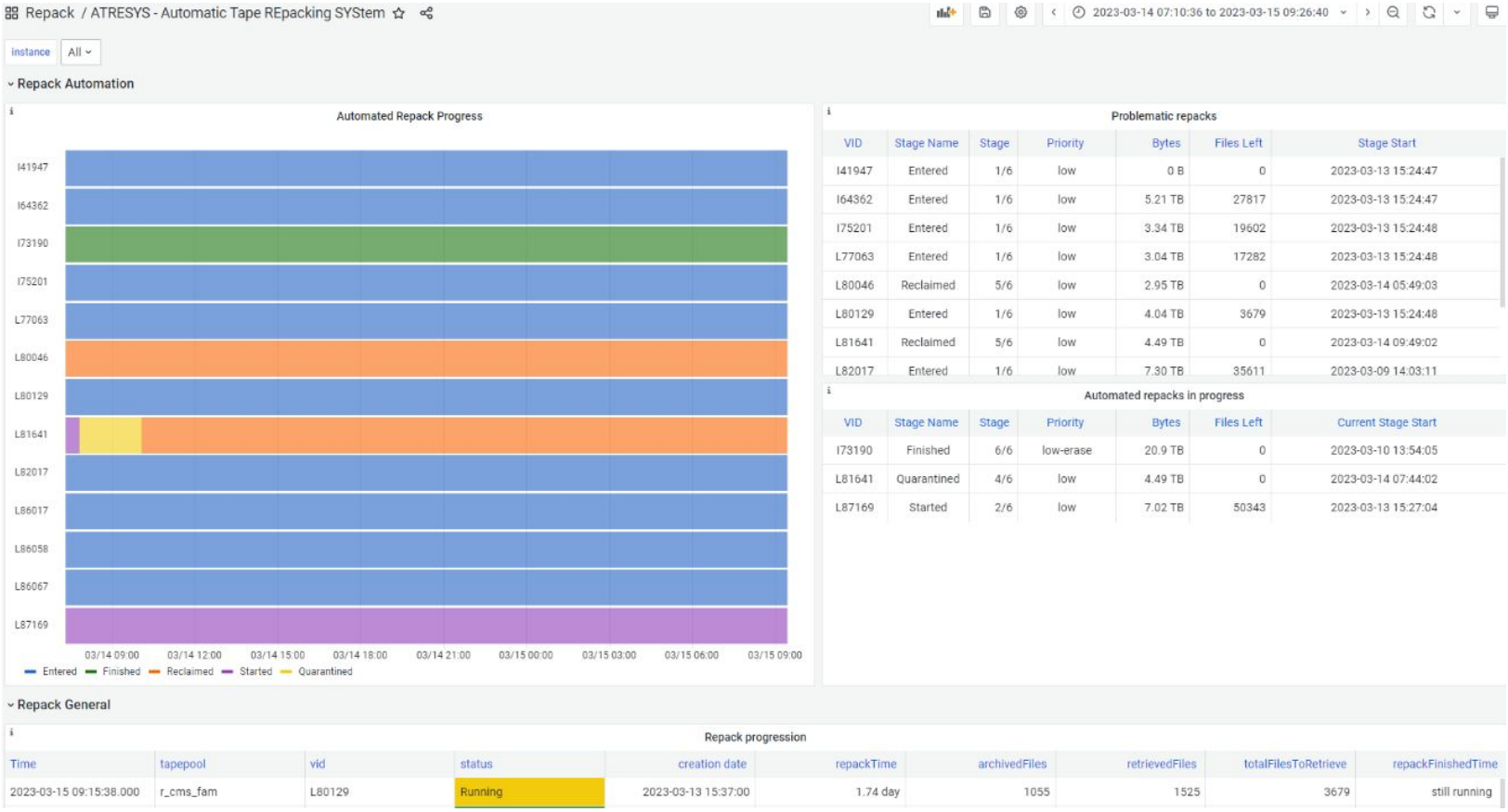
6 states in total

Error situation detected, but not blocking

CTA Run 3 operations

Repack - ATRESYS (Automated Tape REpacking SYStem)

- Monitoring in Grafana



CTA Run 3 operations

Repack - ATRESYS (Automated Tape REpacking SYStem)

- Deployed in production since March 2023
- Still in the process of improving it
  - Already available for the community
    - Link to Gitlab repo: https://gitlab.cern.ch/cta/cta-operations-utilities/-/tree/master/tools/pip/atresys

CTA Run 3 operations

Perform metadata operations on files stored on CTA

- Why?
  - Restore a file from the CTA recycle bin
  - Move a file from one EOSCTA instance to another
  - Migrating a file to CTA
  - Change the storage class of a file
    - Change its tape pool
    - Create new a copy of it

CTA Run 3 operations

Perform metadata operations on files stored on CTA

Operation	Tool	Documentation
Restore a file from the CTA recycle bin	*cta-restore-deleted-files*	https://eoscta.docs.cern.ch/lifecycle/Restoring/
Move a file from one EOSCTA instance to another	*cta-eos-namespace-inject*	https://eoscta.docs.cern.ch/lifecycle/NamespaceInjection/
Migrating a file to CTA		
Change the storage class of a file	*cta-change-storage-class*	https://eoscta.docs.cern.ch/lifecycle/ChangeStorageClass/

CTA Run 3 operations

Handle failed requests

- Failed requests are located in specific queues
  - *cta-admin fr ls*
- Requests are failed after
  - 2 in-mount retries for **archive** jobs
  - 3 in-mount retries x 2 mounts for **retrieve** jobs
  - no retry for **repack** jobs
- Failed-to-archive files stay in the SSD buffer
  - We need to deal with them!

CTA Run 3 operations

Handle failed requests

- How to recover?
  - Ask user to **re-transfer**
  - **Reinject**
    - send CLOSEW event to CTA to trigger the archival of the disk file
      - More details on EOS+CTA workflows: https://indico.cern.ch/event/985953/contributions/4238328/
  - **Ignore** (repack failed requests)

CTA Run 3 operations

Handle failed requests

- 2 steps

classify

reinject

1. Dump requests to a json file
2. Filter out files that:
   - were deleted by user
   - were overwritten by user
   - have wrong fileID (but same path and not on tape)
   - successfully written to tape
   - have valid request still ongoing
   - request the second copy
   - have no archiveID - <u>to be reinjected!</u>

CTA Run 3 operations

Handle failed requests

- 2 steps

classify.sh

reinject.sh

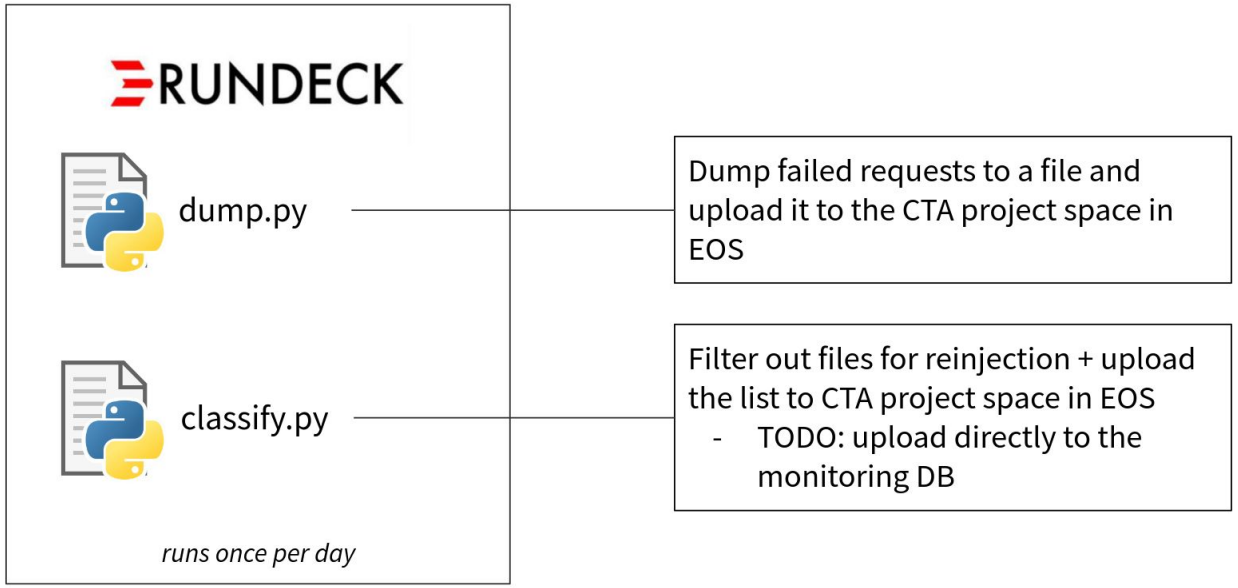
1. Sanity checks
2. Delete ‘`sys.cta.archive.objectstore.id`’ extended attribute
3. Send CLOSEW event (using the tool *cta-send-event*)

manually triggered by the operator on RUNDECK

CTA Run 3 operations

Handle failed requests

- Automated classification



CTA Run 3 operations

Backpressure mechanisms

- EOS+CTA SSD buffer can fill up if the reader is too slow to evict the data
  - archive - when write bandwidth to tape is too slow (library down, not enough free drives)
  - retrieve - EOS instance is slow (heavy experiment use, heavy disk operations (draining, balancing)

CTA Run 3 operations

Backpressure mechanisms

- Archive
  - Destination is full → user will fail to upload their file to the EOS+CTA buffer and will retry later
  - Will be improved by *archive metadata* later

CTA Run 3 operations

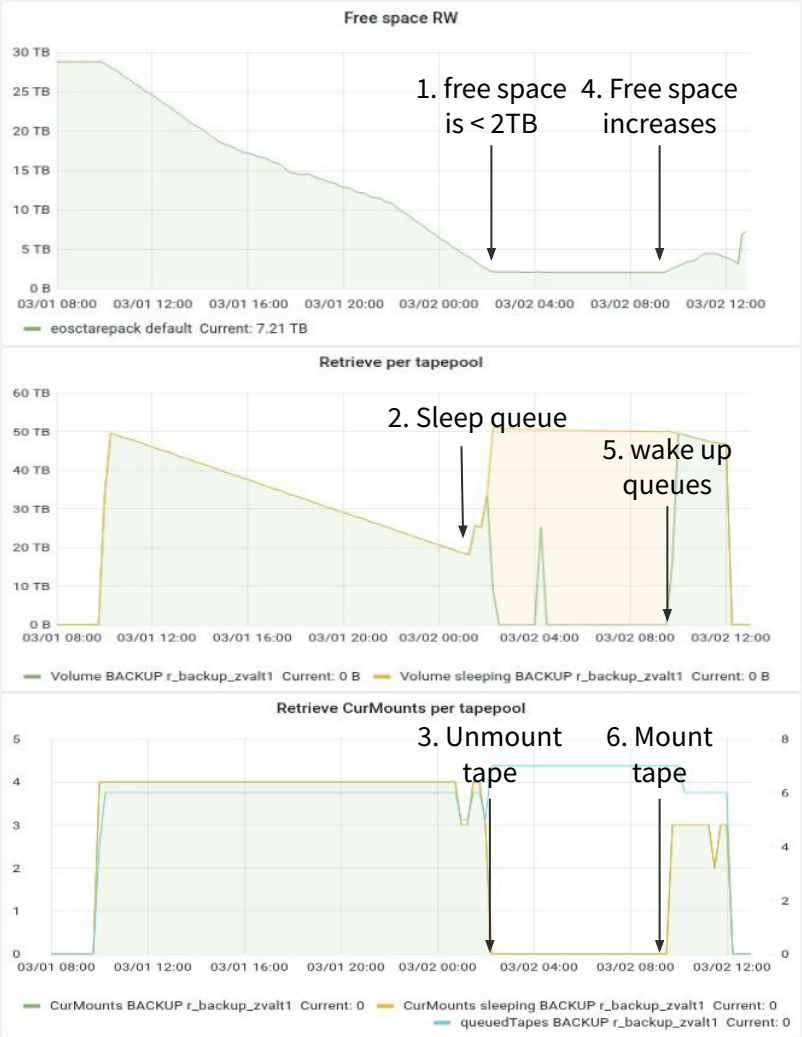
Backpressure mechanisms

- Retrieve
  - Dismount tapes and suspend VO retrieve queues for XX minutes
    - Configuration via *cta-admin disksystem*

Disk system configuration for eosctarepack instance:

```
"name": "eosctarepack",
"fileRegexp": "^root://eosctarepack.*",
"freeSpaceQueryUrl": "eos:eosctarepack:default",
"refreshInterval": "300",
"targetedFreeSpace": "2000000000000",
"sleepTime": "1800",
```

"sleep the retrieve queues for 1/2h if the free space on the SSD buffer is less than 2 TB"



CTA future evolution

Archive metadata

- 2 main goals
  - Improve data collocation on tape
  - Improve tape scheduling
    - Archive backpressure
    - Archive priority
- Will **only be available via HTTP**

CTA future evolution

Archive metadata - Improve data collocation on tape

- Experiments generally retrieve data by **dataset**
  - Improving data collocation on tape will allow better retrieve performances
- CTA archive request queueing is purely FIFO
  - Relies on the way T0 is queueing data for archival
  - Retries are mixed with different datasets

  Results in non-optimal data placement on tapes
- Common rules for tape collocation are needed for T0 and T1s
  - Discussions are ongoing with the experiments

CTA future evolution

Archive metadata - Improve tape scheduling

- Example: DAQ data must go to tape ASAP
- Experiments will be able to tell CTA which data should be archived to tape with a high priority
  - CTA scheduler will therefore make tape scheduling decision based on this information

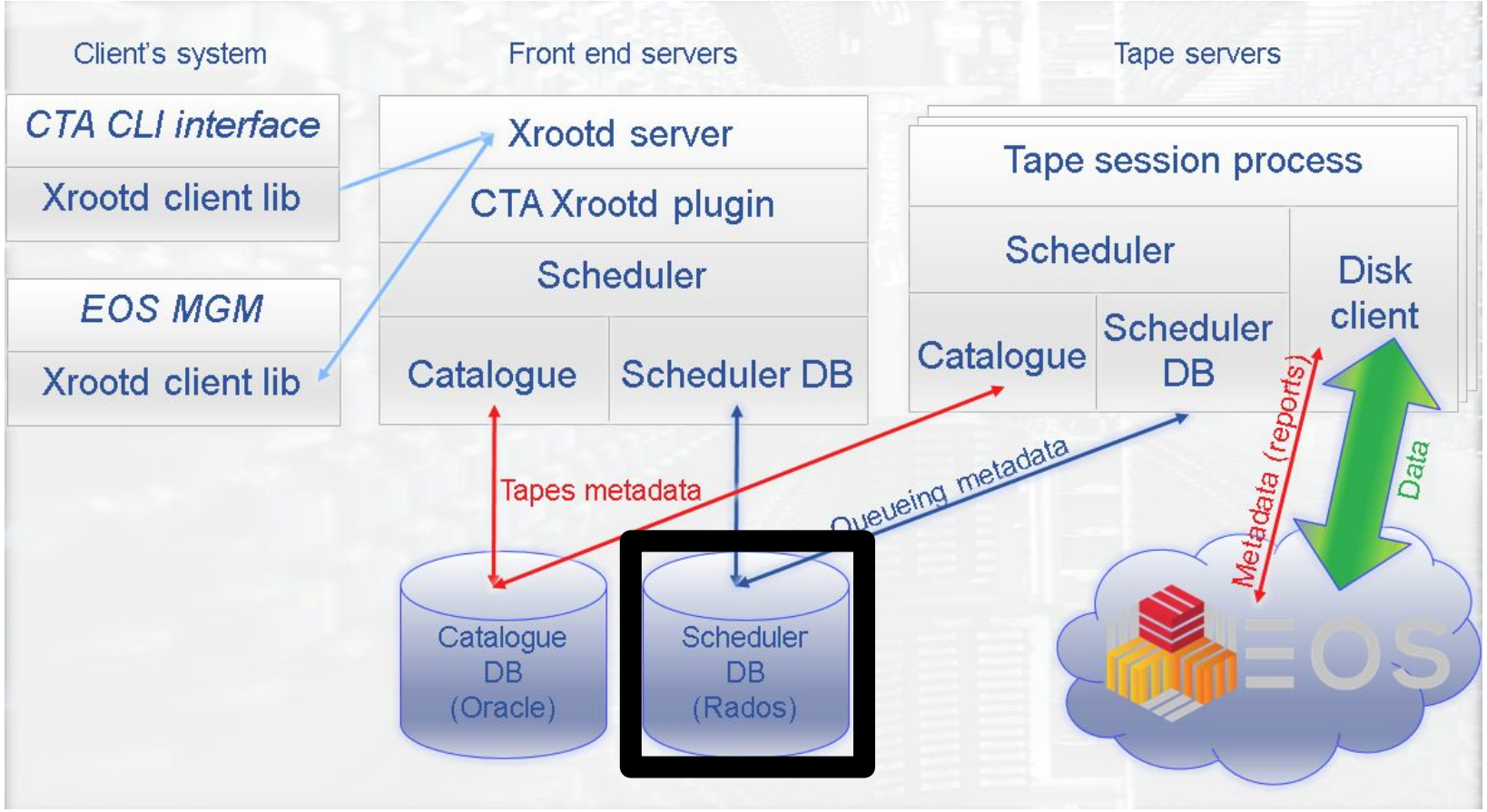
CTA future evolution

Archive metadata - format

Still under discussion…

CTA future evolution

Replacement of the Scheduler Database



CTA future evolution

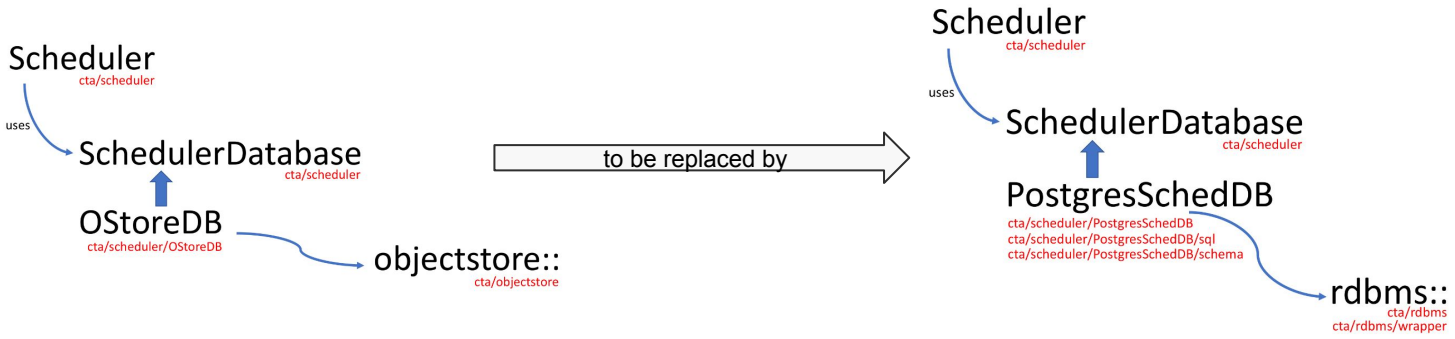
Replacement of the Scheduler Database

- Scheduler Database
  - Used by the scheduler to control the workflow and lifecycle of Archive, Retrieve, Repack requests
  - Is implemented by a CEPH RADOS objectstore
  - Works well for FIFO queueing
- Limitations of the objectstore
  - Constraint on CTA software development
  - Operational issues: difficult to change schema, trace problems, clean up
  - Additional software dependency
  - Additional technology for new team members to learn

CTA future evolution

Replacement of the Scheduler Database

- The objectstore will be replaced by a PostgreSQL database
  - Archive methods mostly done
  - Retrieve methods in progress
  - Additional functionality to do
    - Repack
    - Requests reporting
- Goal is to begin testing in 2H 2023
  - Repack as initial production use-case



CTA future evolution

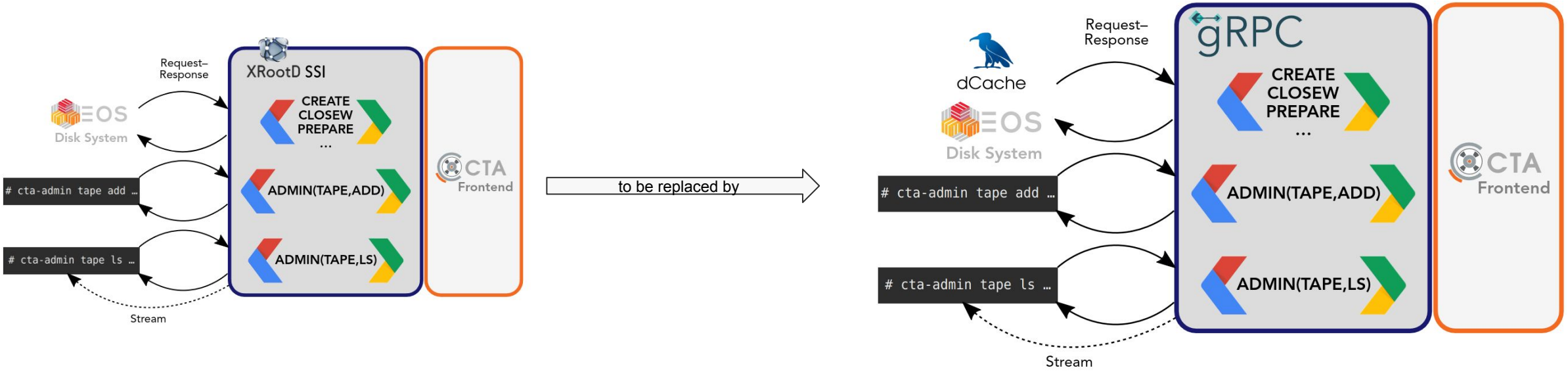
CTA Frontend Transport Protocol

- Requests to CTA frontend are serialized in Google Protocol Buffers
- Transport protocol is XrootD SSI (Scalable Service Interface)
  - Not supported by dCache client → CTA will be integrated to dCache
  - Additional non-standard dependency
  - gRPC is the native transport protocol for protobuf
- Goal is to replace SSI and use gRPC

CTA future evolution

CTA Frontend Transport Protocol

- gRPC Frontend implementation/PoC contributed by dCache team



CTA future evolution

CTA operations tools will be made available for the community

- A big work is going on to make available the different operation tools that we use at CERN to the community
  - ATRESYS already available
  - CTA Operations utilities - https://gitlab.cern.ch/cta/cta-operations-utilities
  - For the rest, stay tuned on the CTA community channel

Conclusion

- CTA meets the CERN Run 3 requirements in terms of Archive and Retrieve performances
  - Fast EOS+CTA SSD buffer in front of the tape infrastructure
    - Protected by file eviction and back pressure
  - Bandwidth regulated by adding/removing tape drives
- Different tools have been created to ease CTA operation
  - Repack - ATRESYS
  - Metadata management tools (recycle-bin restore, namespace file injection)
- Future evolutions are work in progress

Stay tuned on the CTA community channel

home.cern