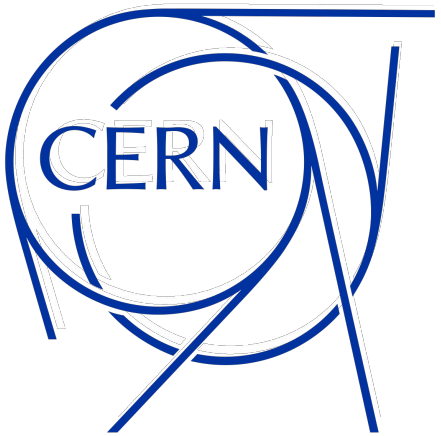


Enabling LHC Run 3 data storage workflows at CERN

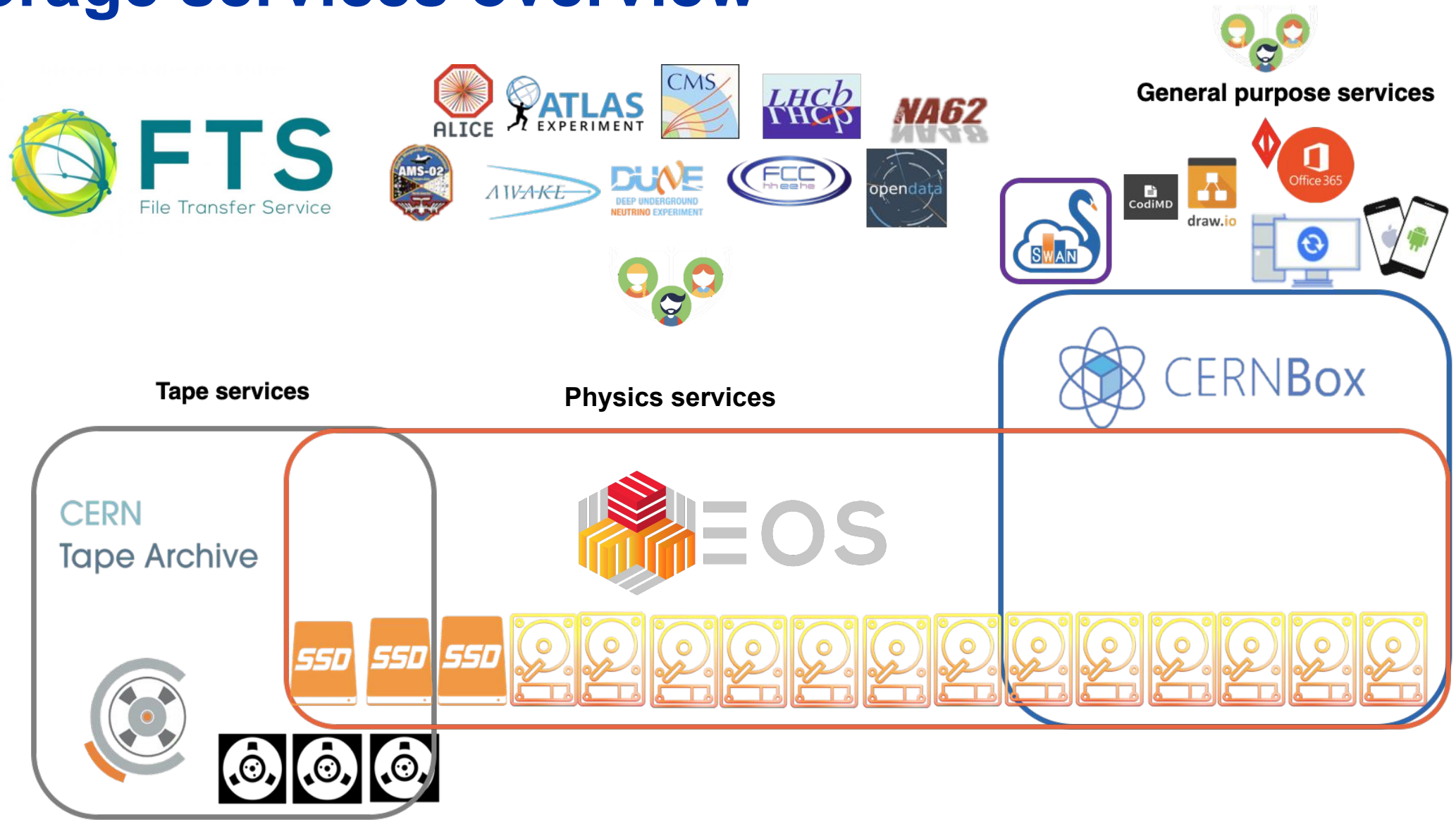
HEPiX Spring 2024, Paris

Elvin Sindrilaru
On behalf of the Storage and Data Management Group

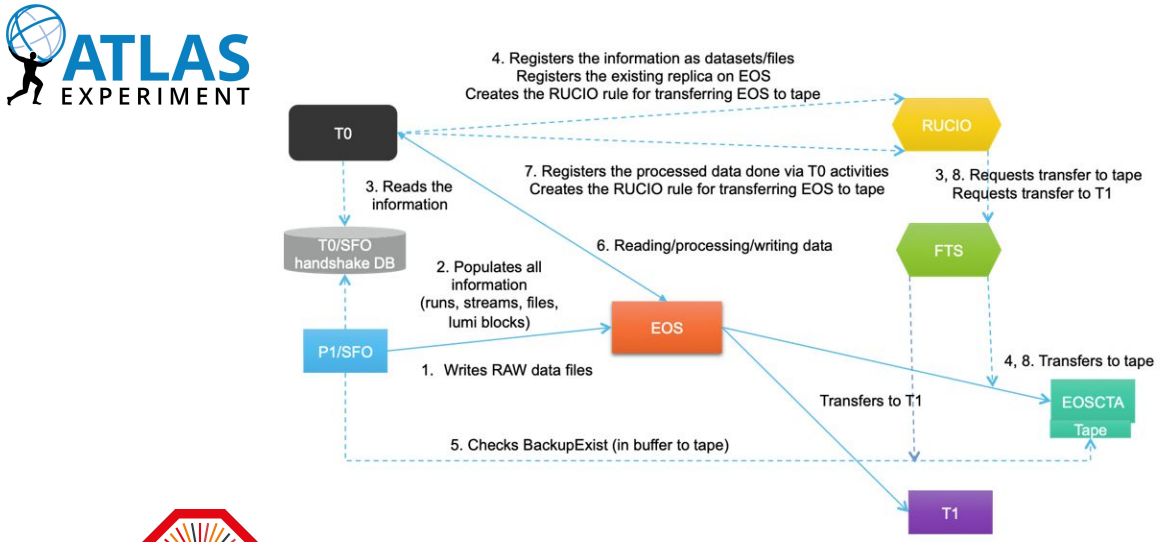
Outline

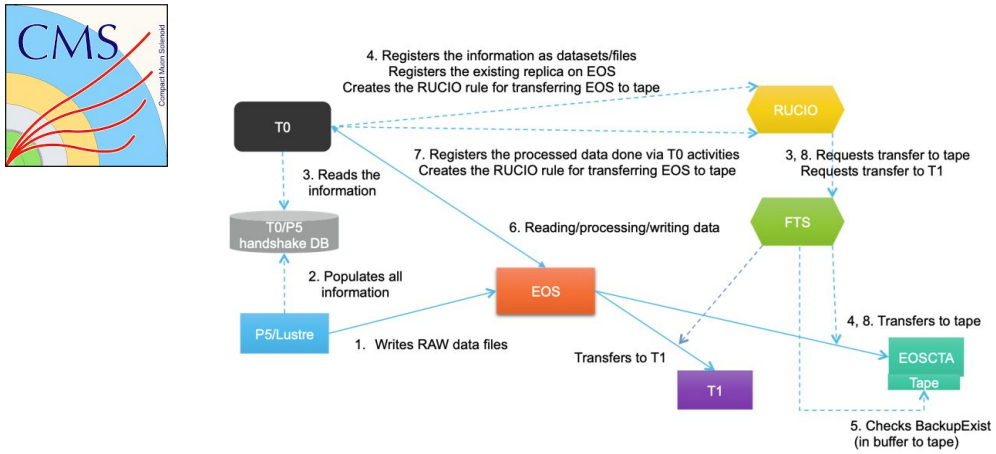
- Overview of the storage services and workflows
- FTS - File Transfer Service
- EOS - Foundational storage service
- CTA - CERN Tape Archive service
- Fun with OpenSSL and Alma 9

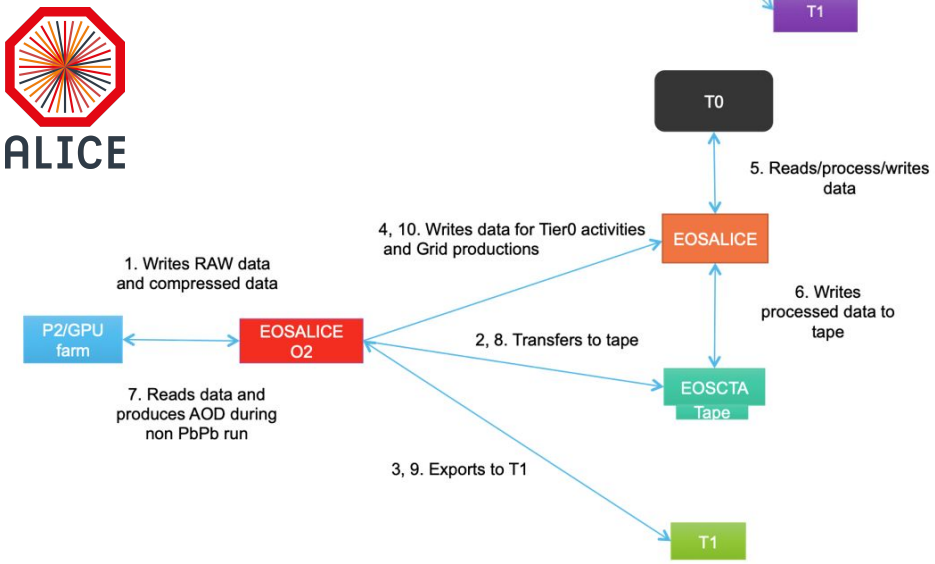
Storage services overview

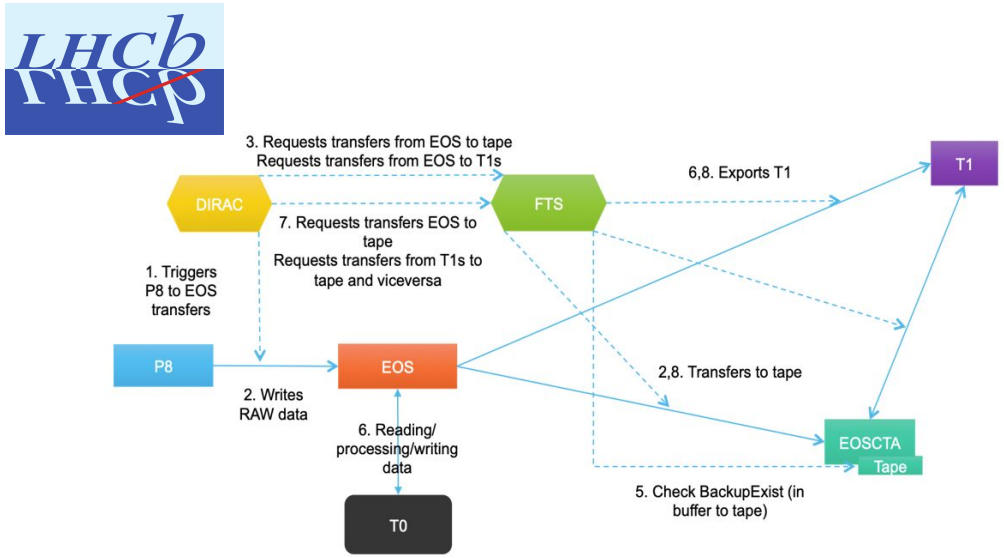


Experiment workflows overview





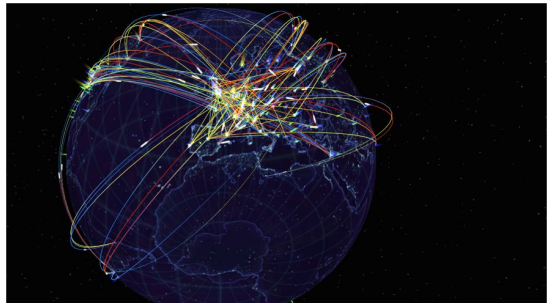




File Transfer Service (FTS)

- **Open-source software** designed for **large scale queueing** and **reliable execution of file transfers** - backbone of WLCG data transfer orchestration

- **Capabilities:**
  - orchestrator of **Third-Party-Copies (TPC)**
  - **streaming** support if TPC not possible
  - tape storage operations via HTTP **Tape REST API,** XRootD etc.
  - support for **Cloud** based storage
  - **X509** and **token** authentication



- **Diverse community of users**

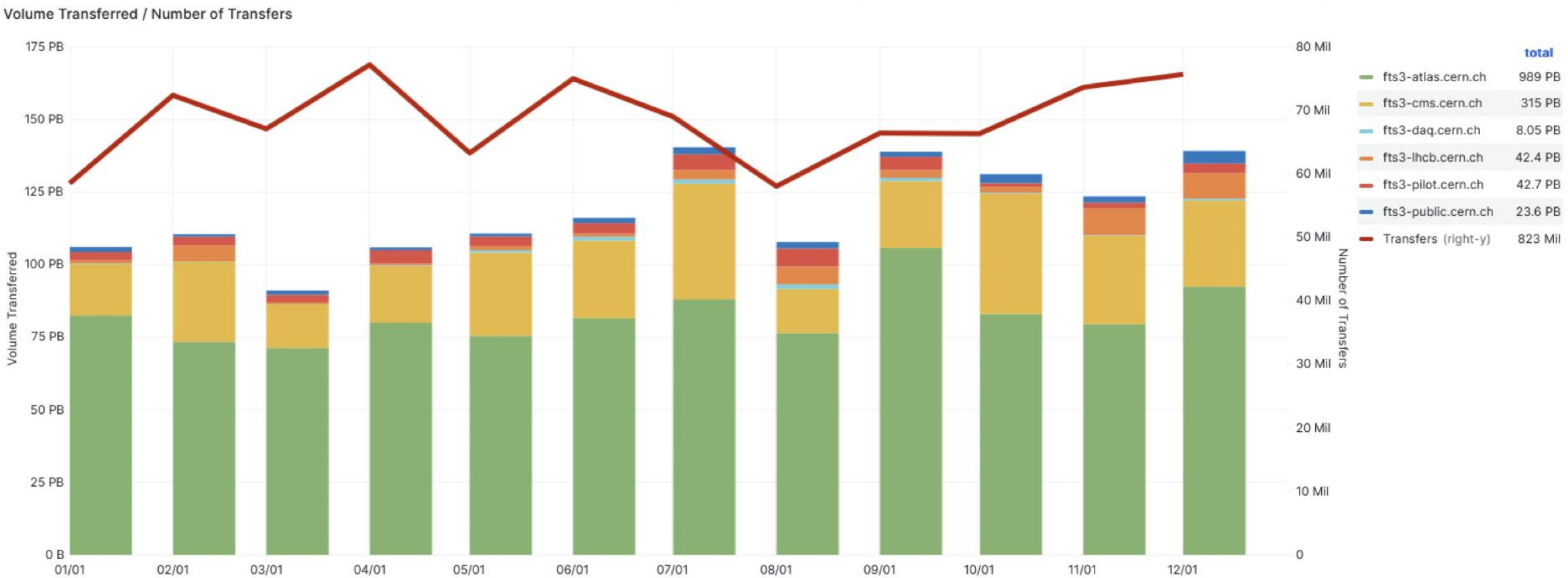
scientific experiments and communities

scientific collaborations

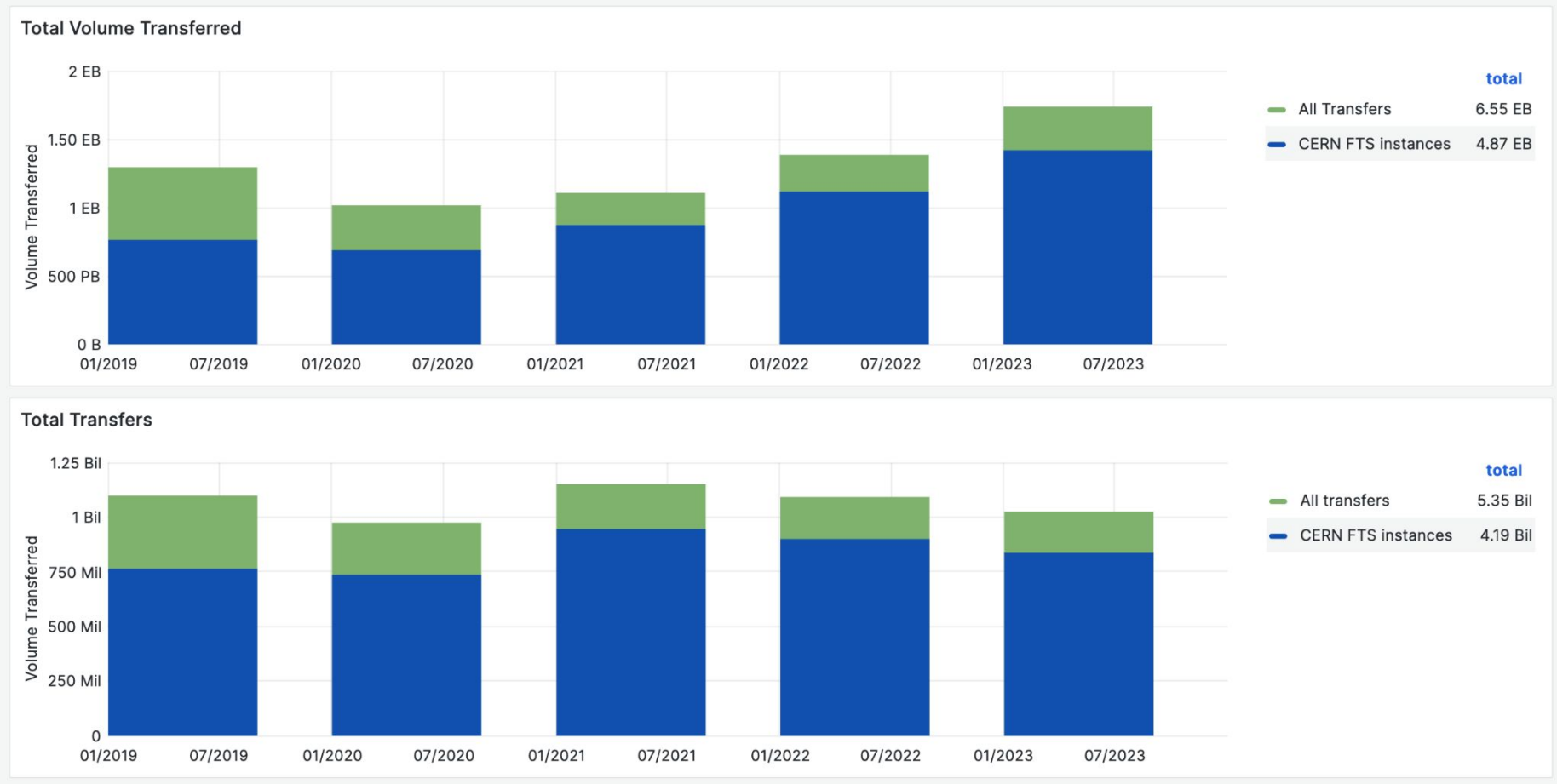
scientific frameworks

Data moved by CERN FTS (2023)

- 823 Million file transfers
- ~1.4 Exabytes of data



FTS multi-year trends



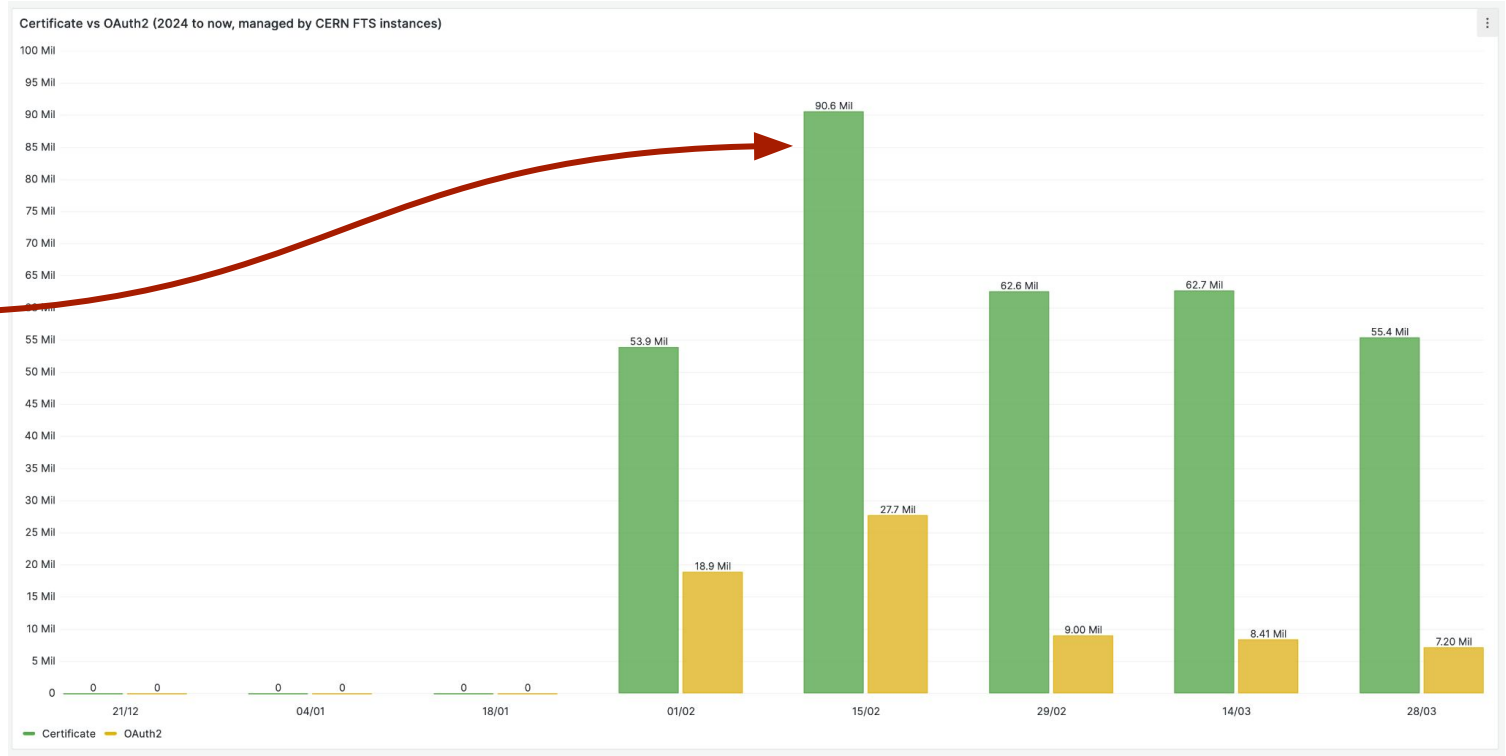
- Considerable **increase in data volume** transfers
- **Constant number of transfers** per year

Consequence of bigger raw data file sizes and RUN 3 restart

OAuth2/Scitokens in FTS

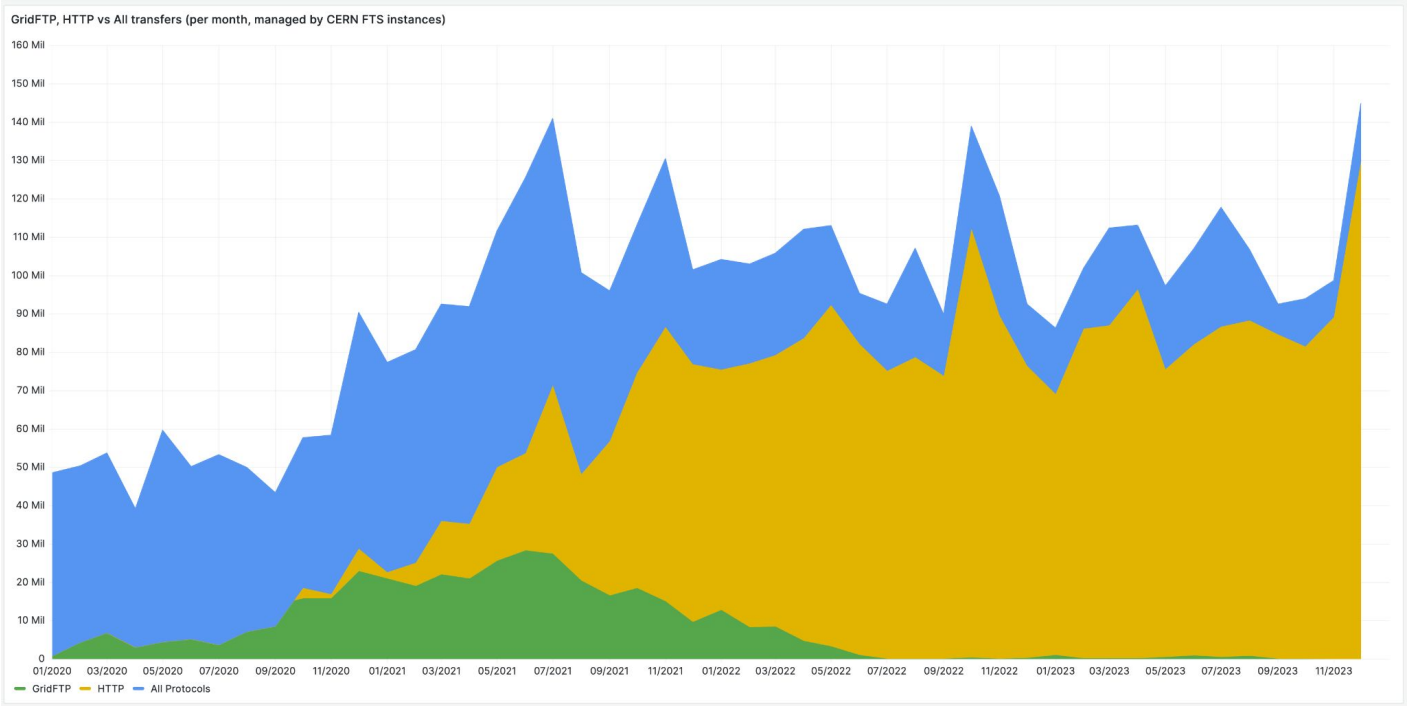
- **OAuth2/SciTokens support**
  - **Important development and deployment** effort across all services involved
    - **FTS, Identity and Access Management (IAM) and storage endpoints**
  - FTS supports both **X509** (certificates) and **tokens** as authorization mechanism
  - Demonstrated viability at scale during **Data Challenge ‘24** (Feb 2024)

DC ‘24



FTS HTTP dominance and GridFTP phase-out

- FTS will **switch off GridFTP support** for ATLAS and LHCb **by end of April**
- **HTTP** is now the **dominant transfer protocol**
  - Strongly coupled with **token support for TPC transfers**
- **Recommended FTS release v3.11.2**
- **Alma 9 support coming soon!**



EOS & CTA Services @ CERN

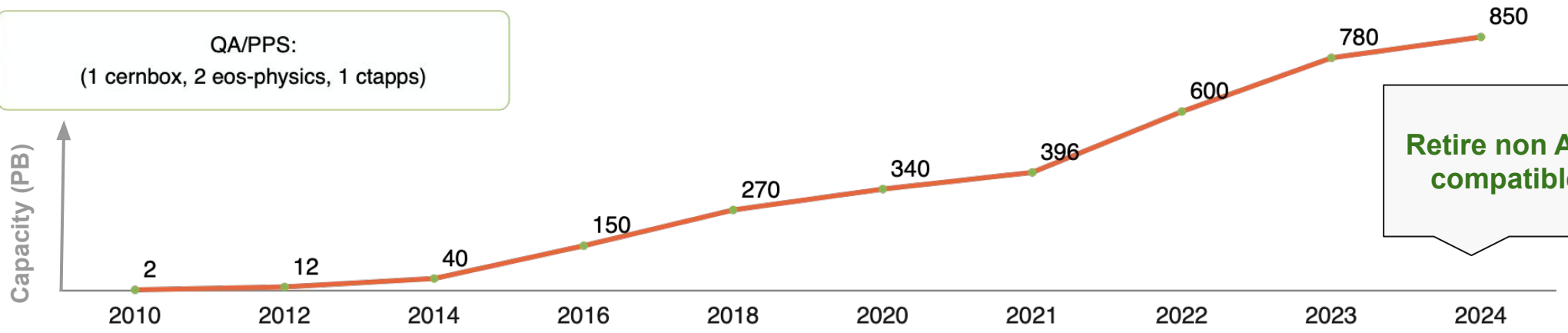
EOS instances

- Physics: 9 (6 LHC, 3 non-LHC)
- CTA: 8 (4 LHC, 2 non-LHC, 1 backup, 1 internal repack)
- CERNBox: 9 (5 homes + 3 projects + 1 media)
- QA/PPS: (1 cernbox, 2 eos-physics, 1 ctapps)

831 Storage Nodes (+4)

66K Disks (+10%)

4 LHC:
500 SSDs + 800 HDDs for 7.5PB (+0%)



PLANS 2024:
retirement ~20PB
new capacity +100PB

Retire non ALMA-9 compatible HW

EOS for Physics statistics and growth 2023

+15%

287 M
Number of directories

+14%

3.10 B
Number of files

+8%

> 700 PB
Total space

6 + 3 instances

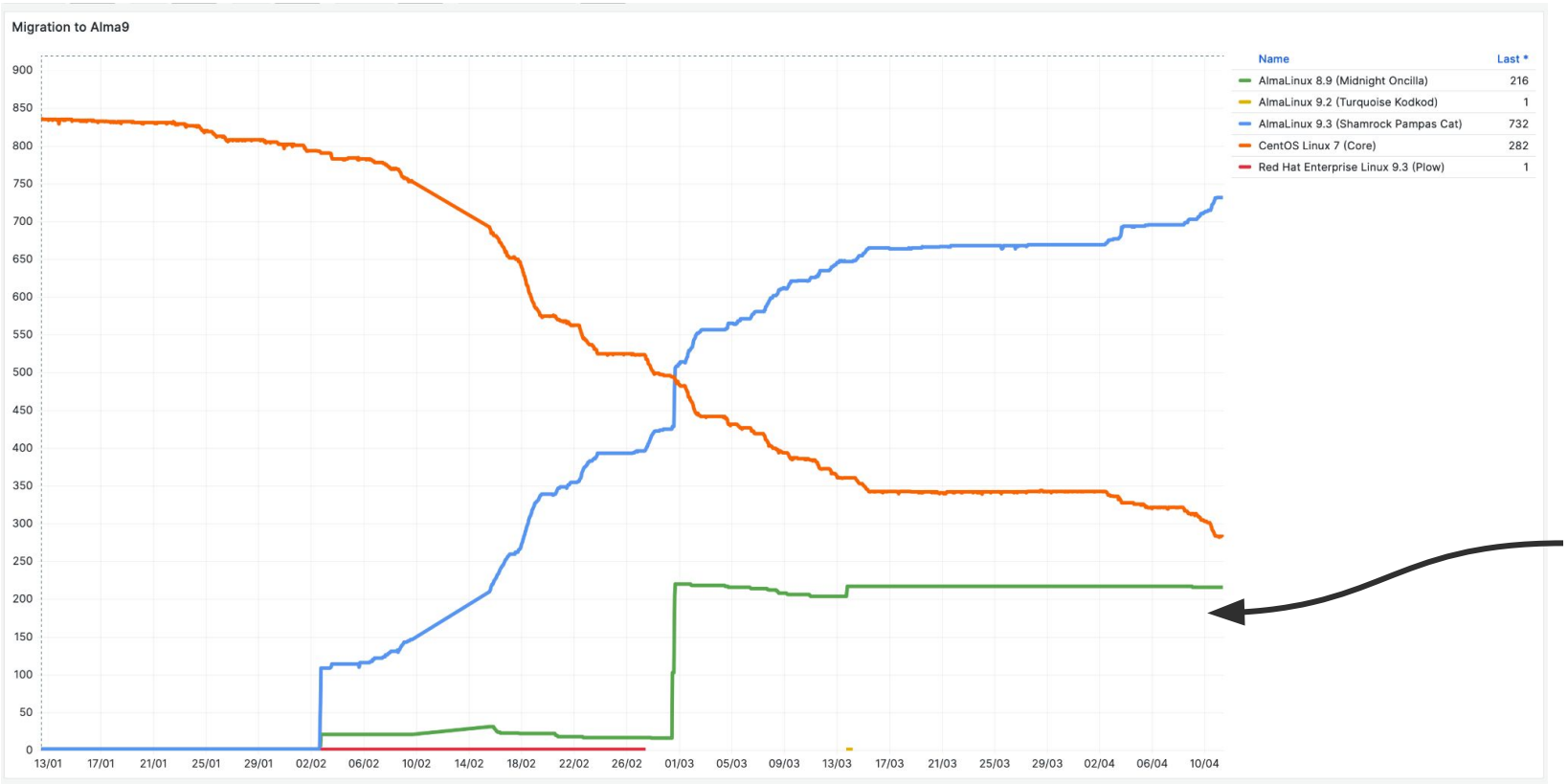
	Total space	Used space	Number of files
ATLAS	94.56 PB	79.58 PB	266 Mil
CMS	102.97 PB	77.67 PB	240 Mil
ALICE	116.73 PB	107.32 PB	840 Mil
LHCb	70.14 PB	44.43 PB	1.13 Bill
Public & AMS	134.25 PB	108.97 PB	589 Mil
ALICEO2	181.99 PB	162.66 PB	30.1 Mil
TOTAL	**700.64 PB**	**580.63 PB**	**3095.1 Mil**

Recommended deployment setup:

- **EOS version: 5.2.22**
- **XRootD version 5.6.9**
- **OS: Alma 9.3**

EOS transition to ALMA 9

- **RPM** packages also provided for:
  - **Alma Linux 9 - recommended, RHEL 9**
  - Alma 8 and Fedora 38 (opportunistically)
  - CC7 (until June 2024)

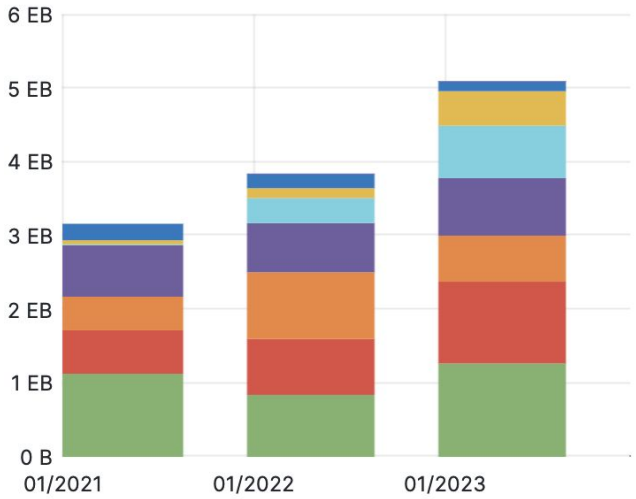


EOSALICEO2 installed before ALMA 9 available

EOS for Physics usage statistics

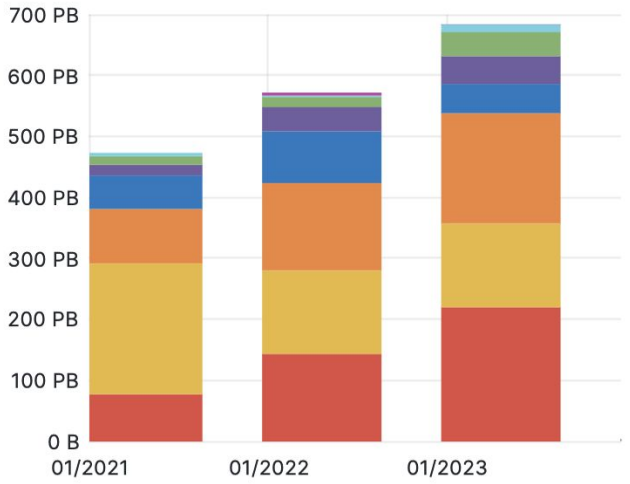
- **EOS** instances served **5 Exabytes (+34%)** of data and received **~0.7 Exabytes (+16%)**
- Charts below show **trends over the last 3 years**

Export: Amount of bytes read



	max	avg	current
alice	1.27 EB	811 PB	7.10 PB
cms	1.11 EB	615 PB	4.08 PB
atlas	904 PB	497 PB	1.14 PB
public	781 PB	536 PB	2.40 PB
ams02	711 PB	267 PB	229 TB
aliceo2	468 PB	165 PB	1.11 PB
lhcb	223 PB	138 PB	1.34 PB
p2	4.03 PB	1.47 PB	251 MB

Ingestion: Amount of bytes written

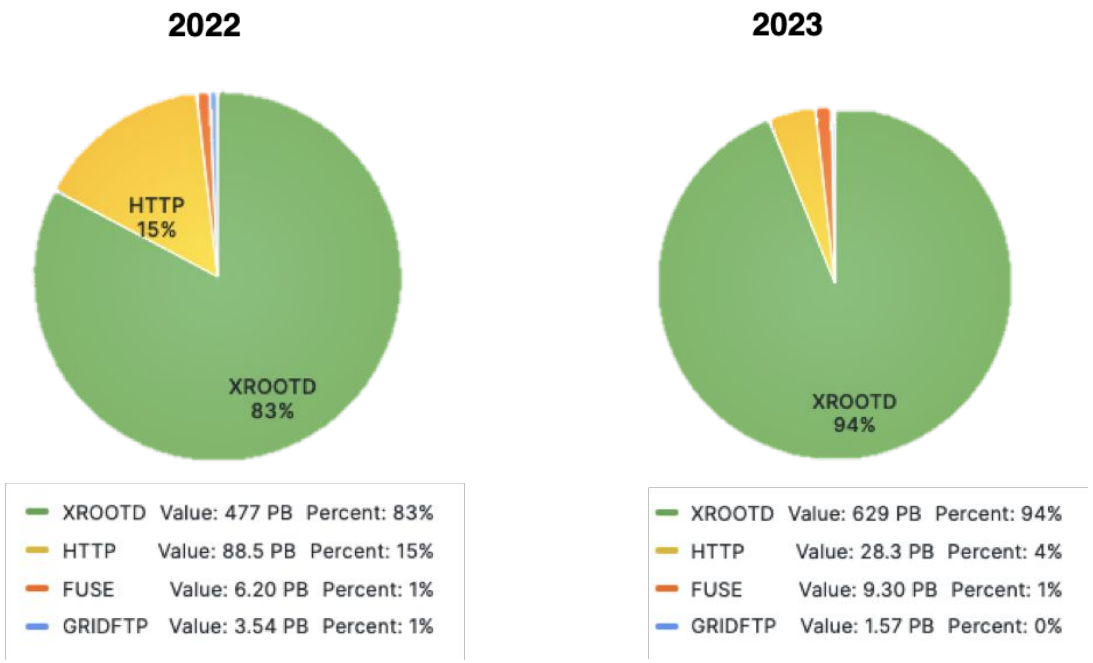


	max	avg	current
cms	221 PB	111 PB	412 TB
aliceo2	215 PB	123 PB	1.51 PB
atlas	181 PB	103 PB	560 TB
lhcb	84.7 PB	47.2 PB	1.22 PB
public	45.7 PB	25.7 PB	97.2 TB
alice	39.7 PB	17.6 PB	187 TB
ams02	12.4 PB	5.08 PB	16.4 TB
p2	4.80 PB	1.74 PB	251 MB

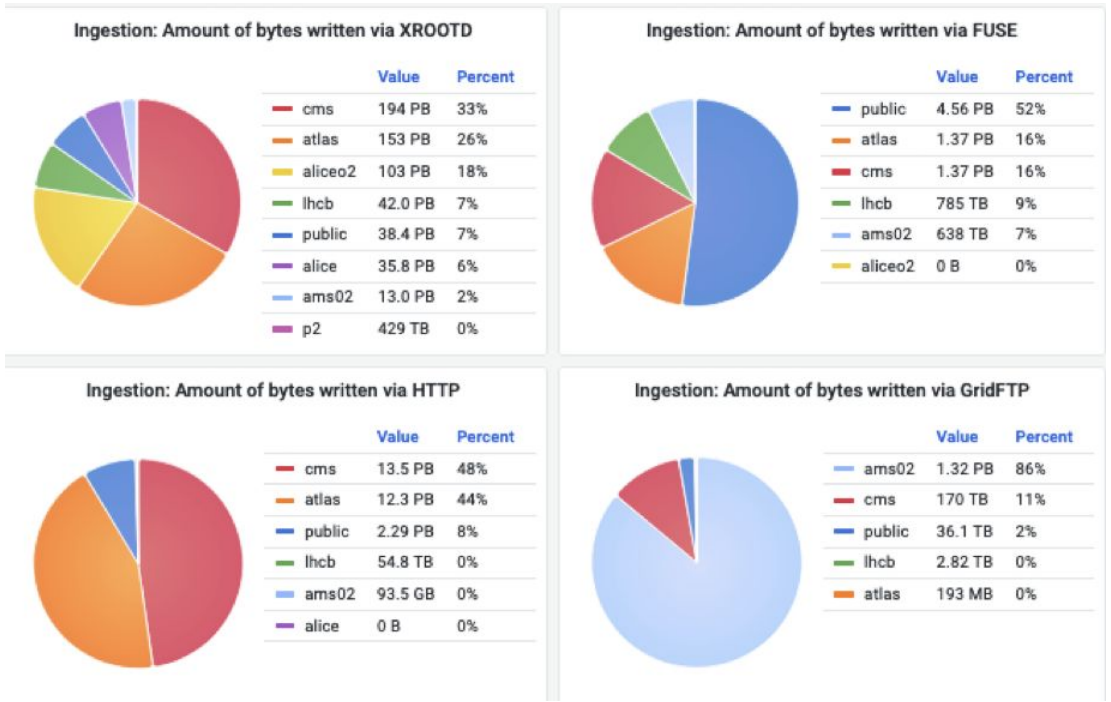
EOS ingress protocol statistics

Most used protocol for **writes in 2023: XRootD**

Total writes per protocol



Total writes per protocol and instance

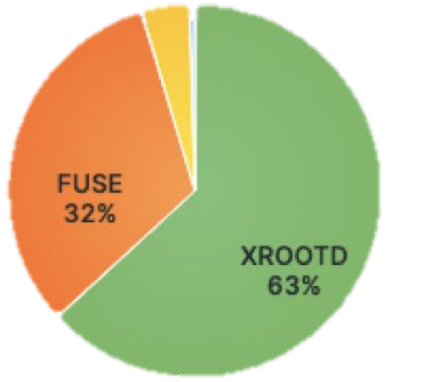


EOS egress protocol statistics

Most used protocol for **reads in 2023**: **XRootD** and **FUSE**

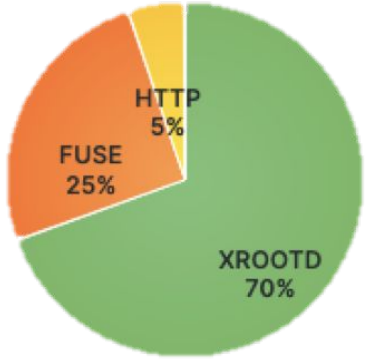
Total reads per protocol

2022



XROOTD	Value: 2.39 EB	Percent: 63%
FUSE	Value: 1.22 EB	Percent: 32%
HTTP	Value: 155 PB	Percent: 4%
GRIDFTP	Value: 14.6 PB	Percent: 0%

2023



XROOTD	Value: 3.57 EB	Percent: 70%
FUSE	Value: 1.30 EB	Percent: 25%
HTTP	Value: 257 PB	Percent: 5%
GRIDFTP	Value: 527 TB	Percent: 0%

Total reads per protocol and instance

Export: Amount of bytes read via XRootD

	Value	Percent
alice	1.15 EB	35%
cms	826 PB	25%
aliceo2	414 PB	13%
public	277 PB	9%
ams02	249 PB	8%
atlas	234 PB	7%
lhcb	92.0 PB	3%
p2	374 TB	0%

Export: Amount of bytes read via FUSE

	Value	Percent
public	437 PB	37%
ams02	428 PB	36%
atlas	200 PB	17%
cms	108 PB	9%
lhcb	17.8 PB	1%
aliceo2	851 MB	0%

Export: Amount of bytes read via HTTP

	Value	Percent
atlas	146 PB	62%
cms	80.3 PB	34%
lhcb	9.22 PB	4%
public	1.22 PB	1%
ams02	36.0 TB	0%
alice	453 MB	0%

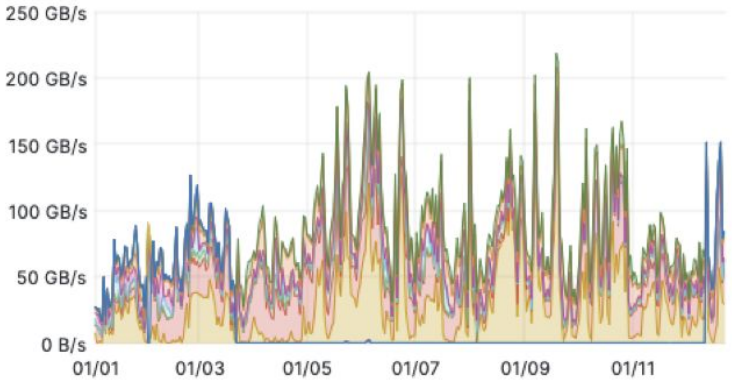
Export: Amount of bytes read via GridFTP

	Value	Percent
ams02	319 TB	61%
lhcb	96.3 TB	18%
public	95.9 TB	18%
cms	15.2 TB	3%
atlas	193 MB	0%

EOS for Physics network statistics

- **During 2023**, we were able to achieve more than **200 GB/s ingress** and **500 GB/s egress**
- **Overall delivered excellent performance and stability!**

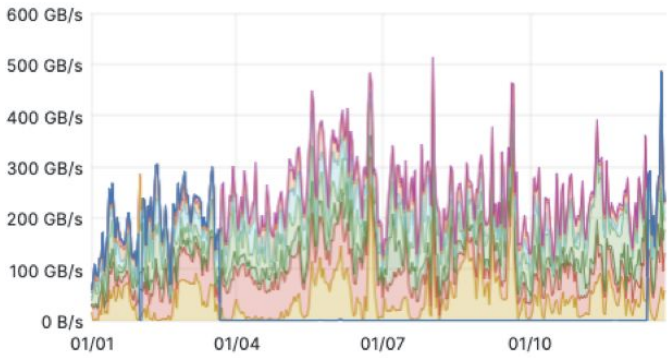
Cluster Network Rates (in)



	max	avg	current
aliceo2.inratemib.erasure	170 GB/s	28.9 GB/s	28.7 GB/s
cms.inratemib	64.5 GB/s	23.1 GB/s	27.7 GB/s
alice.inratemib	41.8 GB/s	5.36 GB/s	5.92 GB/s
ams02.inratemib	38.8 GB/s	3.89 GB/s	814 MB/s
lhcb.inratemib	35.7 GB/s	3.88 GB/s	1.32 GB/s
atlas.inratemib	30.4 GB/s	13.8 GB/s	15.6 GB/s
public.inratemib	20.6 GB/s	4.60 GB/s	4.53 GB/s
aliceo2.inratemib	4.80 GB/s	1.09 GB/s	0 B/s

Excellent EOSALICEO2 performance enabled by erasure coding layouts

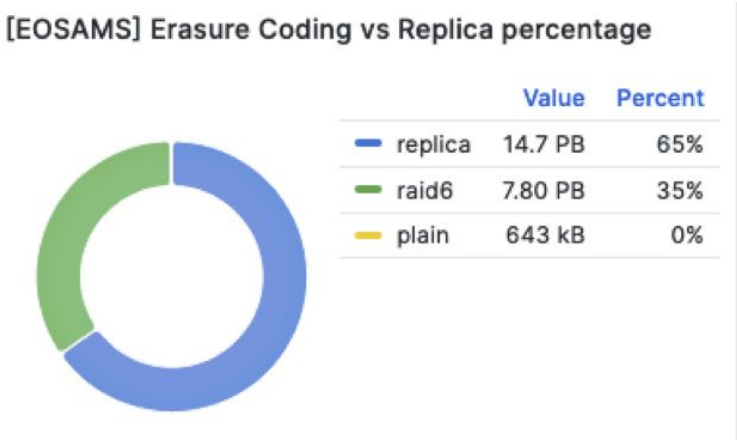
Cluster Network Rates (out)

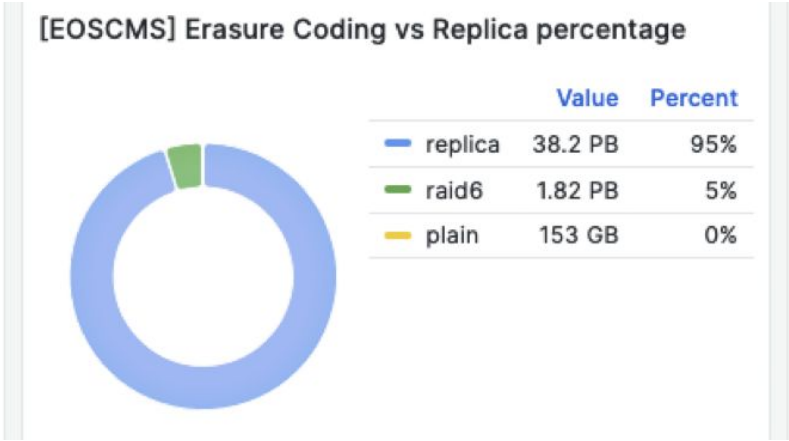


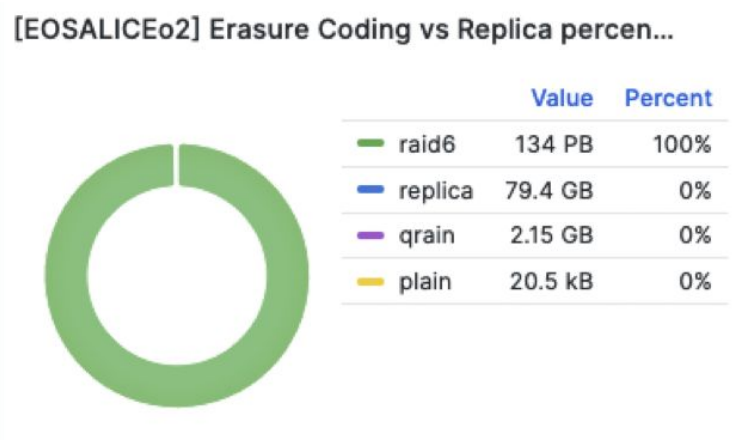
	max	avg	current
aliceo2.outratemib.erasure	340 GB/s	43.9 GB/s	59.6 GB/s
cms.outratemib	178 GB/s	54.7 GB/s	73.4 GB/s
public.outratemib	173 GB/s	30.0 GB/s	45.5 GB/s
alice.outratemib	157 GB/s	47.7 GB/s	13.6 GB/s
ams02.outratemib	94.7 GB/s	27.8 GB/s	14.4 GB/s
atlas.outratemib	66.2 GB/s	28.1 GB/s	19.3 GB/s
lhcb.outratemib	34.9 GB/s	6.65 GB/s	3.37 GB/s
aliceo2.outratemib	9.96 GB/s	2.12 GB/s	0 B/s

EOS space optimisation - erasure coding

- ***Game-changing*** adoption of **erasure coding layouts (a.k.a RAIN layouts)** bringing:
  - Raw storage space optimisation - **48 PB *saved* during 2023** in comparison to 2 replica layout!
  - Better transfer performance

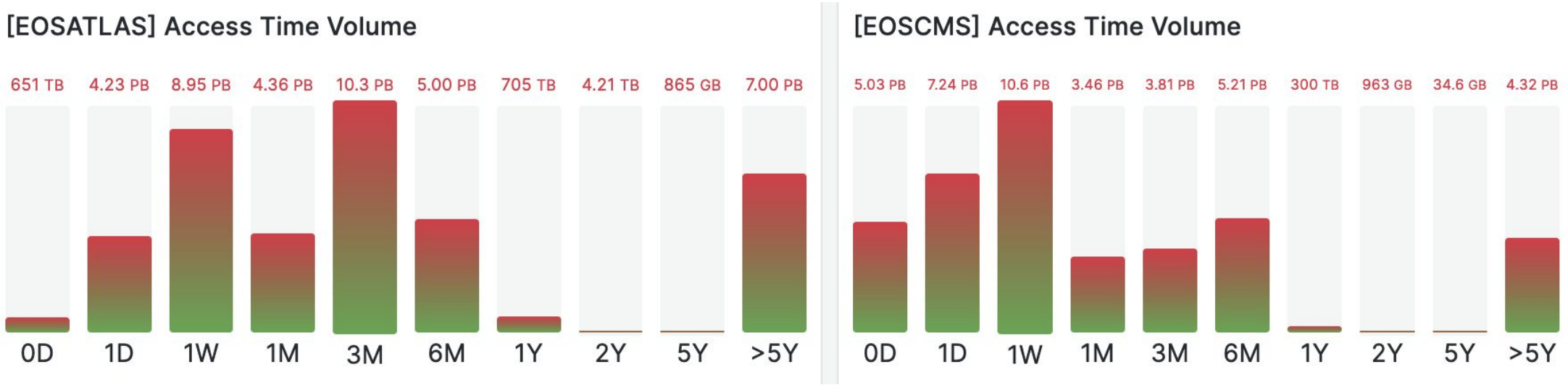






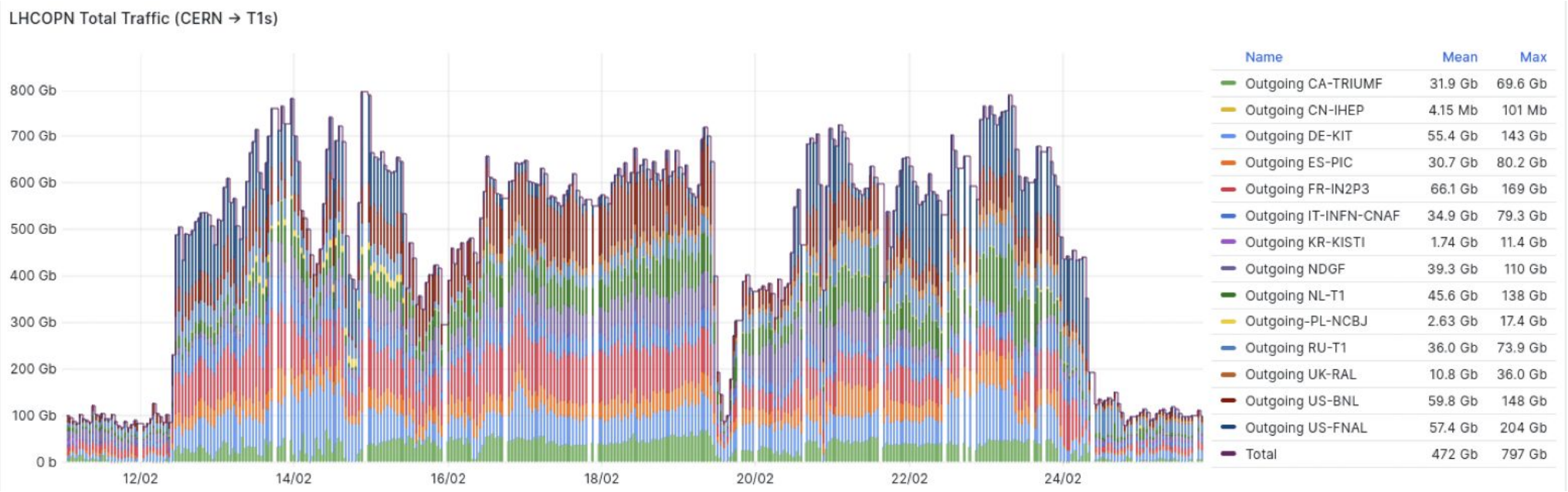
EOS space optimisation - identify “stale” data

- **Goal:** identify **data not accessed in a long time** as good candidates for **tape migration**
  - Information provided by the `eos file inspector` functionality



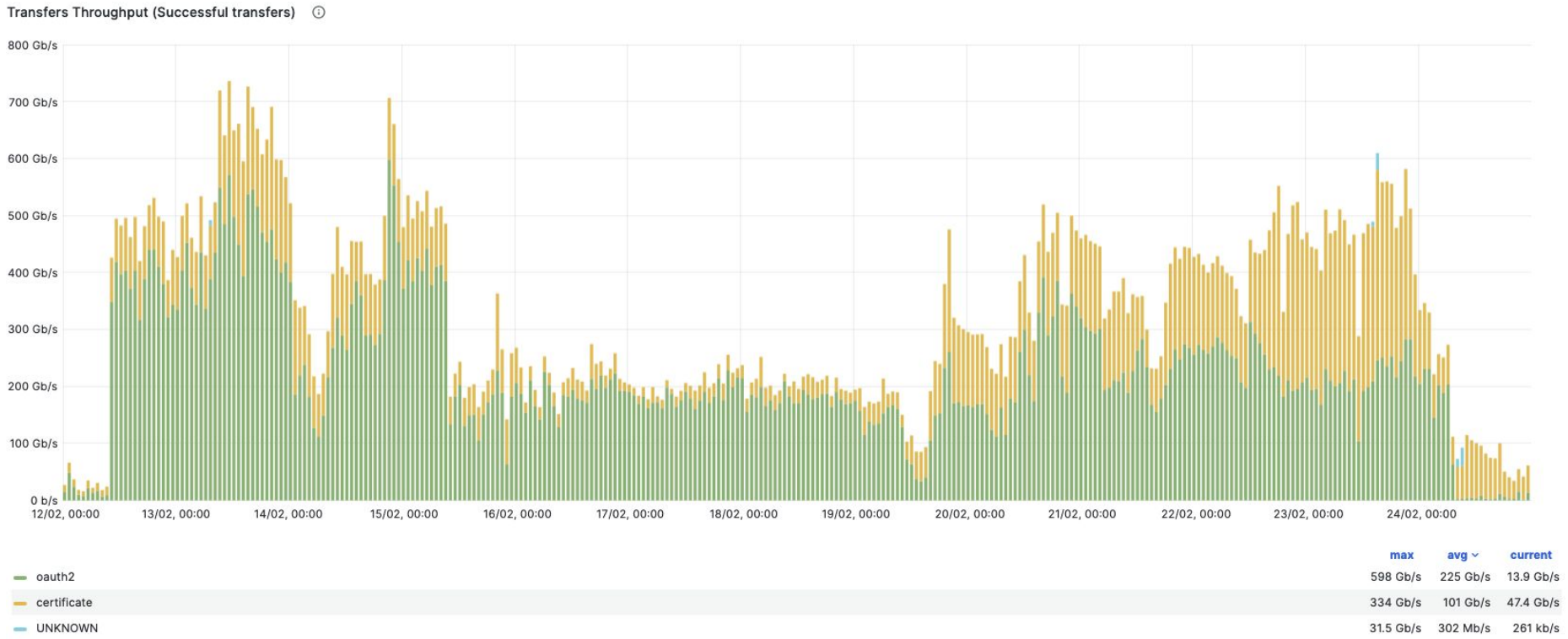
EOS contribution to Data Challenge ‘24

- **EOS targets during the Data Challenge:**
  - Ensure **stability**, **availability** and required **performance** for exports to **T1s**
    - On top of existing activities!
  - **Deployment and configuration of SciTags** for CMS instance



EOS/FTS contribution to Data Challenge ‘24

- **Deployment of HTTP tokens support** in all instances
  - **CERN EOS** instances handled the bulk of the traffic
  - FTS plot of transfer reports using **X509** or **tokens** during **DC ‘24**

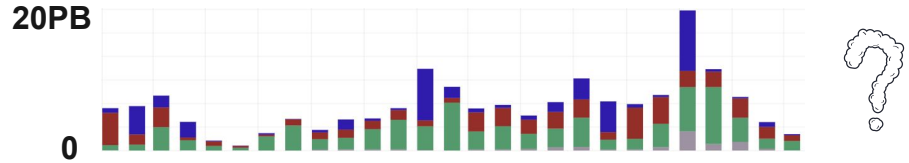


CTA - Tape Storage

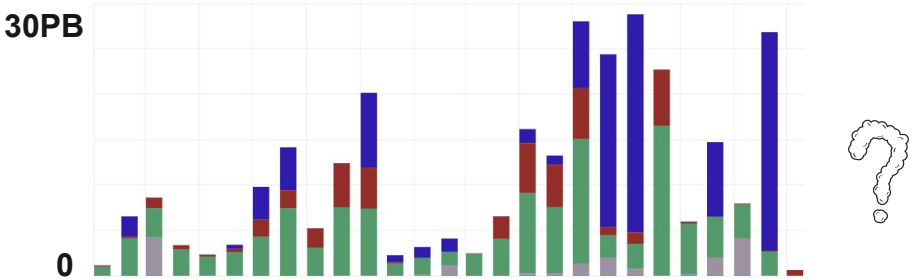
CERN Tape Archive

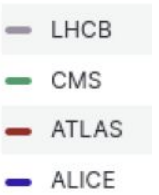
- **LHC EOSCTA instances: 10GB/s archival SLA for T0**
  - **Archive boost** during Heavy-Ion Run
  - **Staging boost** during Year End Technical Stop (YETS) - data duplication to T1/T2s

Staged vol. per month

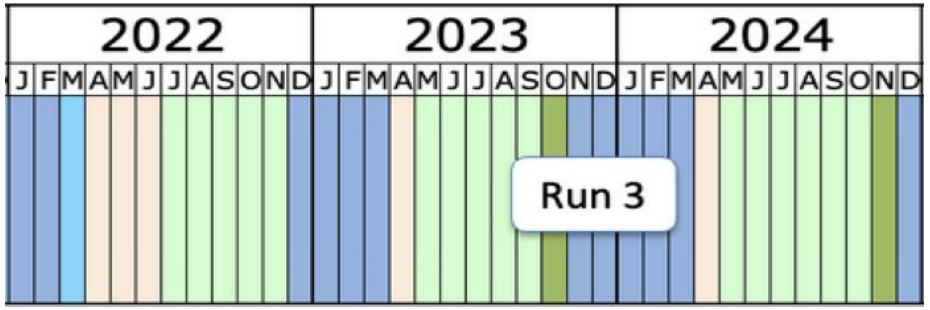


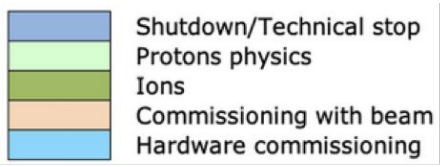
Archived vol. per month





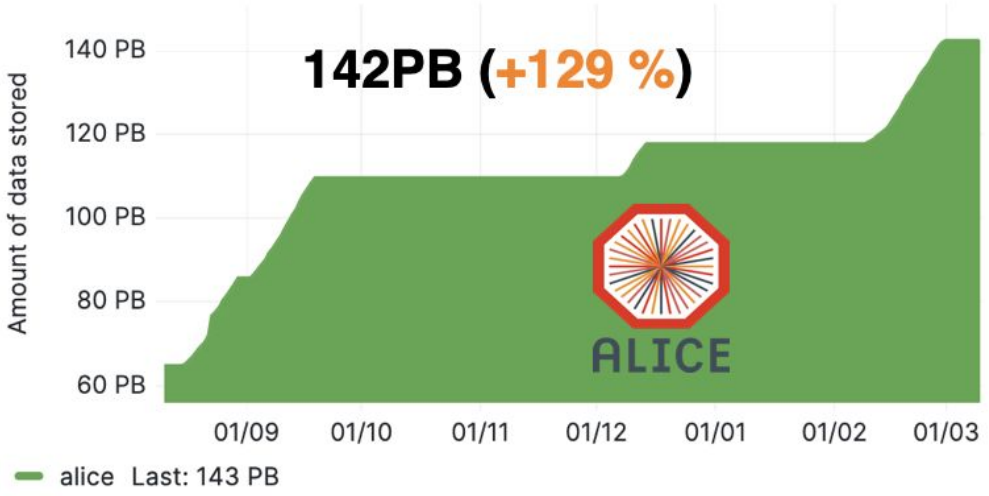
RUN3 LHC planning

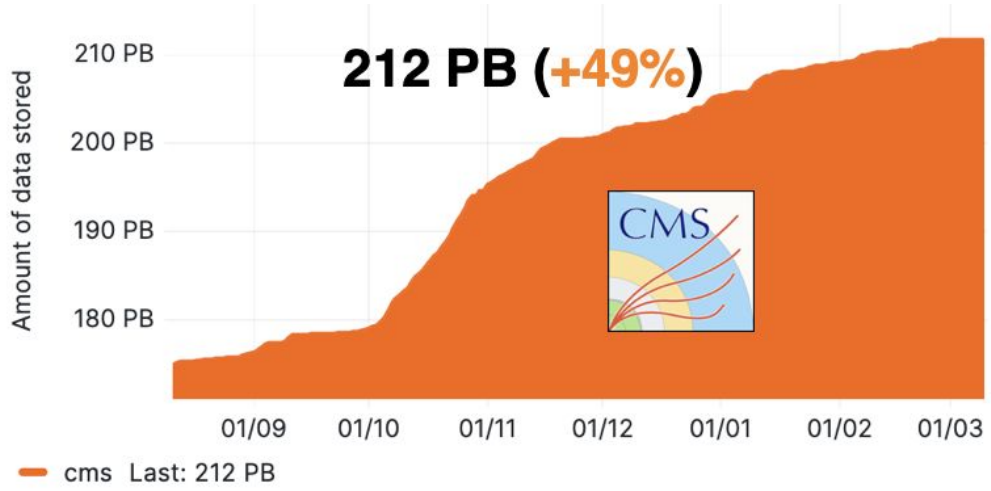


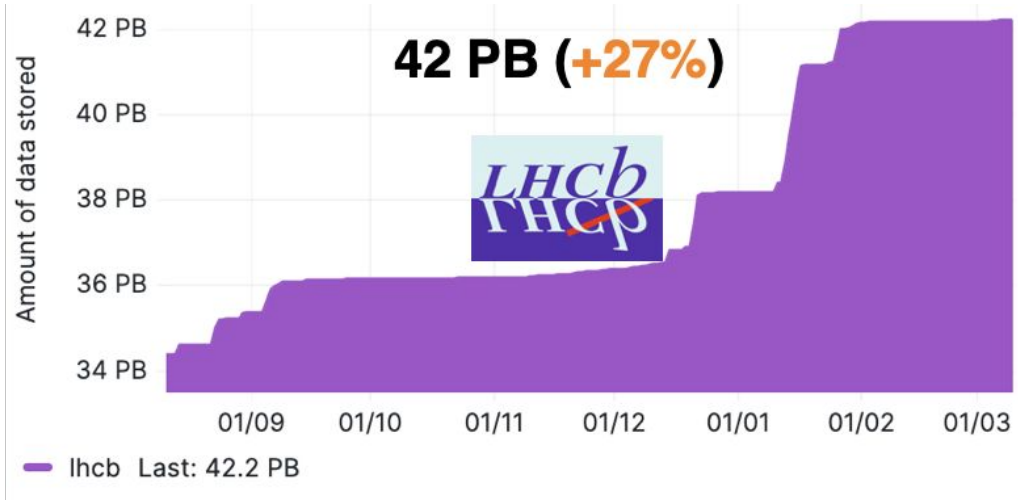


CTA - Run 3 data volumes

CERN Tape Archive



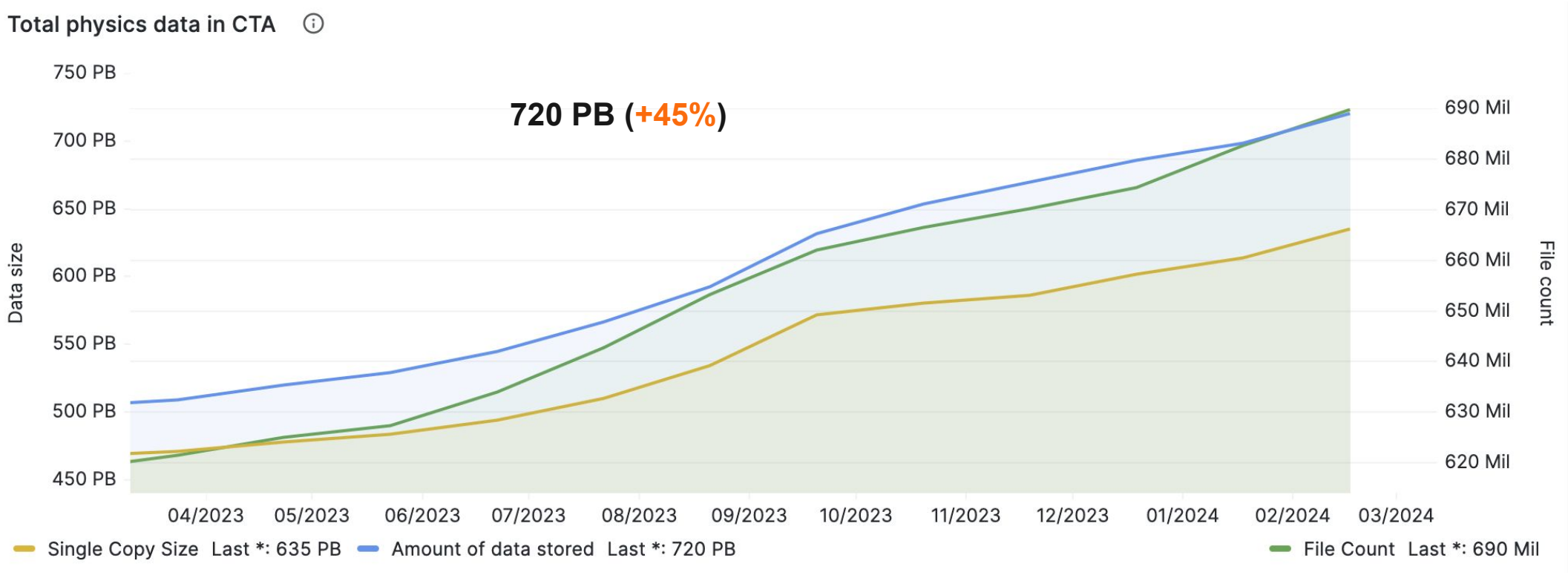




CTA - Physics data volume

CERN Tape Archive

- **CTA primary propose: provide reliable, long-term archival storage for physics experiments**



ALMA 9 & OpenSSL 3 issues

- **Slow with Alma9 (openssl 3.0.7)**
  - Seen to be much slower for **GSI** connections
    - 0.15s to ~1.9s for an **xrdcp** of a tiny file
  - **High CPU consumption**
  - Call in **OpenSSL 3 (DH_check) ~10 times** slower than in **OpenSSL 1**
  - DH_check validates some **Diffie-Hellman exchange parameters** used during the **GSI handshake**
    - computationally intensive check if a number is prime (Typically a 3072 bit number p and also (p-1)/2 are checked)
- **OpenSSL3 adjusted its prime checking functions (see openssl #9272) to be easier to use & harder to misuse**

```
$ perf report --call-graph --stdio
.
.
.
#
# Children      Self  Command  Shared Object          Symbol
# ........  ........  .......  ......................  ..............................................
#
    99.34%     0.00%  xrdcp    libXrdSecgsi-5.so      [.] XrdSecProtocolgsi::getCredentials
            |
            ---XrdSecProtocolgsi::getCredentials
               |
                --98.97%--XrdSecProtocolgsi::ParseClientInput
                          |
                          |--93.35%--XrdSecProtocolgsi::ClientDoCert
                          |          |
                          |           --93.13%--XrdCryptosslFactory::Cipher
                          |                     XrdCryptosslCipher::XrdCryptosslCipher
                          |                     |
                          |                      --92.30%--EVP_PKEY_param_check
                          |                                evp_pkey_param_check_combined
                          |                                try_provided_check
                          |                                evp_keymgmt_validate
                          |                                dh_validate
                          |                                DH_check_ex
                          |                                DH_check
                          |                                BN_check_prime
                          |                                ossl_bn_check_prime
                          |                                bn_is_prime_int
                          |                                |
                          |                                 --92.17%--ossl_bn_miller_rabin_is_prime
```

- **Solution provided on the XRootD side**
  - Recent xrootd-based servers will use the **same DH parameters**. If the client detects an exact match no specific DH_check is done.

Conclusions

- **Storage and Transfer Services performance requirements for 2023 Run 3 have been successfully met!**

- **FTS service**
  - Demonstrated viability of the **token authorization** during DC ‘24
  - Pioneered support for **SciTags** during DC ‘24
  - Transferred more than **1.7EB** during 2023

- **EOS service**
  - Achieved and exceeded experiment requirements
  - **Erasure-coded files** support proven in production (ALICEO2)
  - **HTTP protocol** gaining traction and **GridFTP phase-out**

- **CERN Tape Archive**
  - Proved its **versatility** in addressing both **archiving and staging ops.**
  - Proper **bandwidth allocation and configuration** can efficiently absorb peaks in demand

home.cern