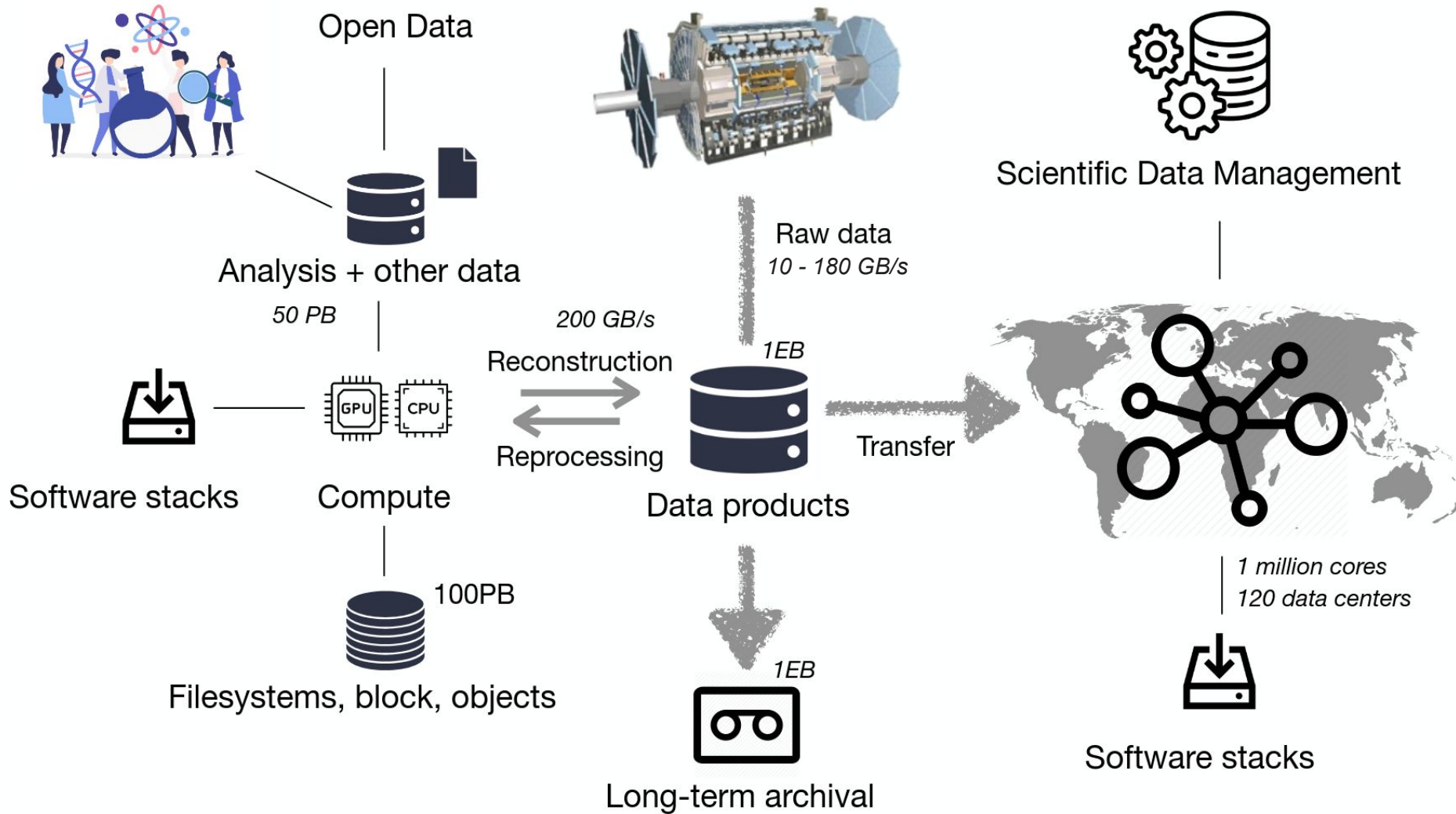


# EOS operational experiences and future challenges



*Andreas J. Peters, Cedric Caffy,  
Elvin A. Sindrilaru, Luca Mascetti*



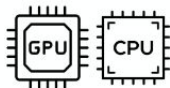




Open Data



Analysis + other data



Software stacks

Compute

Reconstruction



Reprocessing



Data products

Transfer



XRootD



CVMFS

Software stacks



Raw data



Scientific Data Management



Filesystems, block, objects

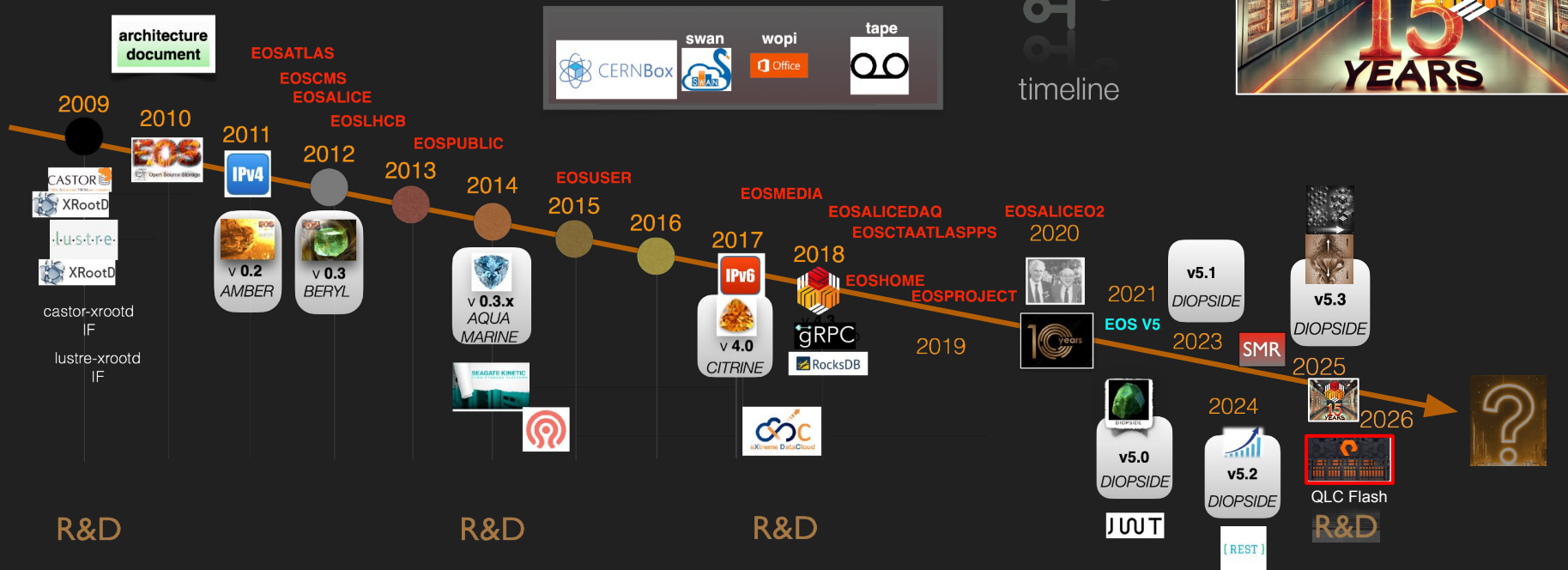


CERN  
Tape Archive

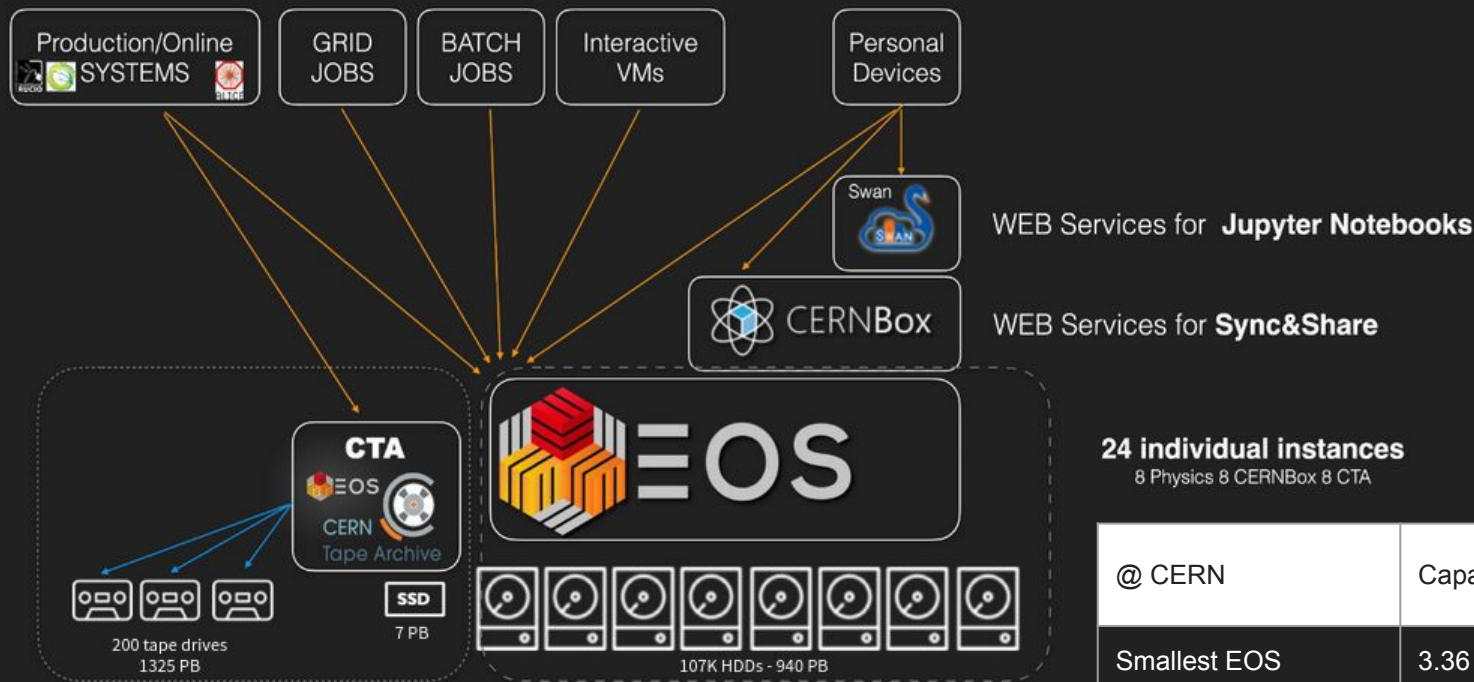
Long-term archival

# The EOS Project Timeline

15 years of history



# EOS Service in Numbers



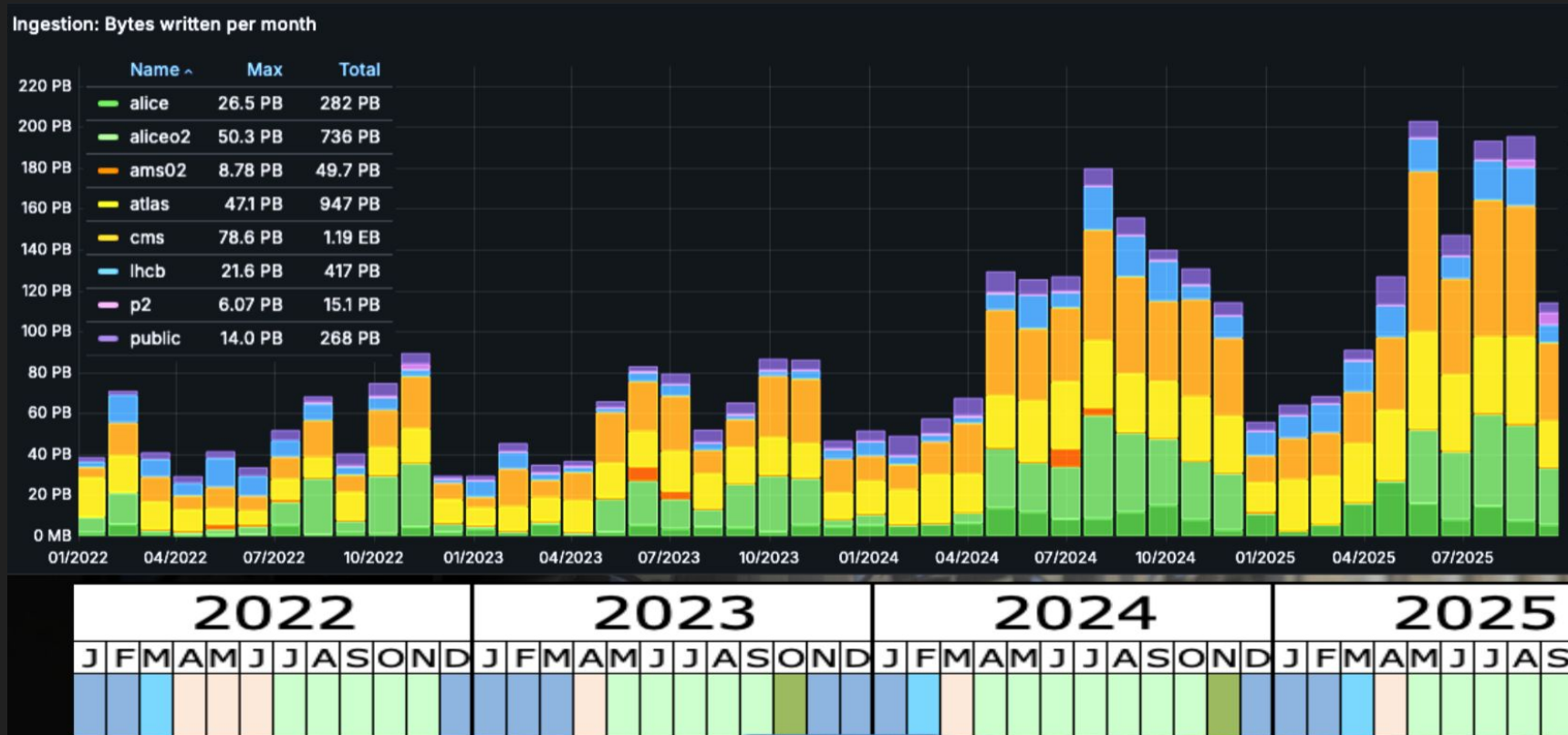
How is EOS used?

## 24 individual instances

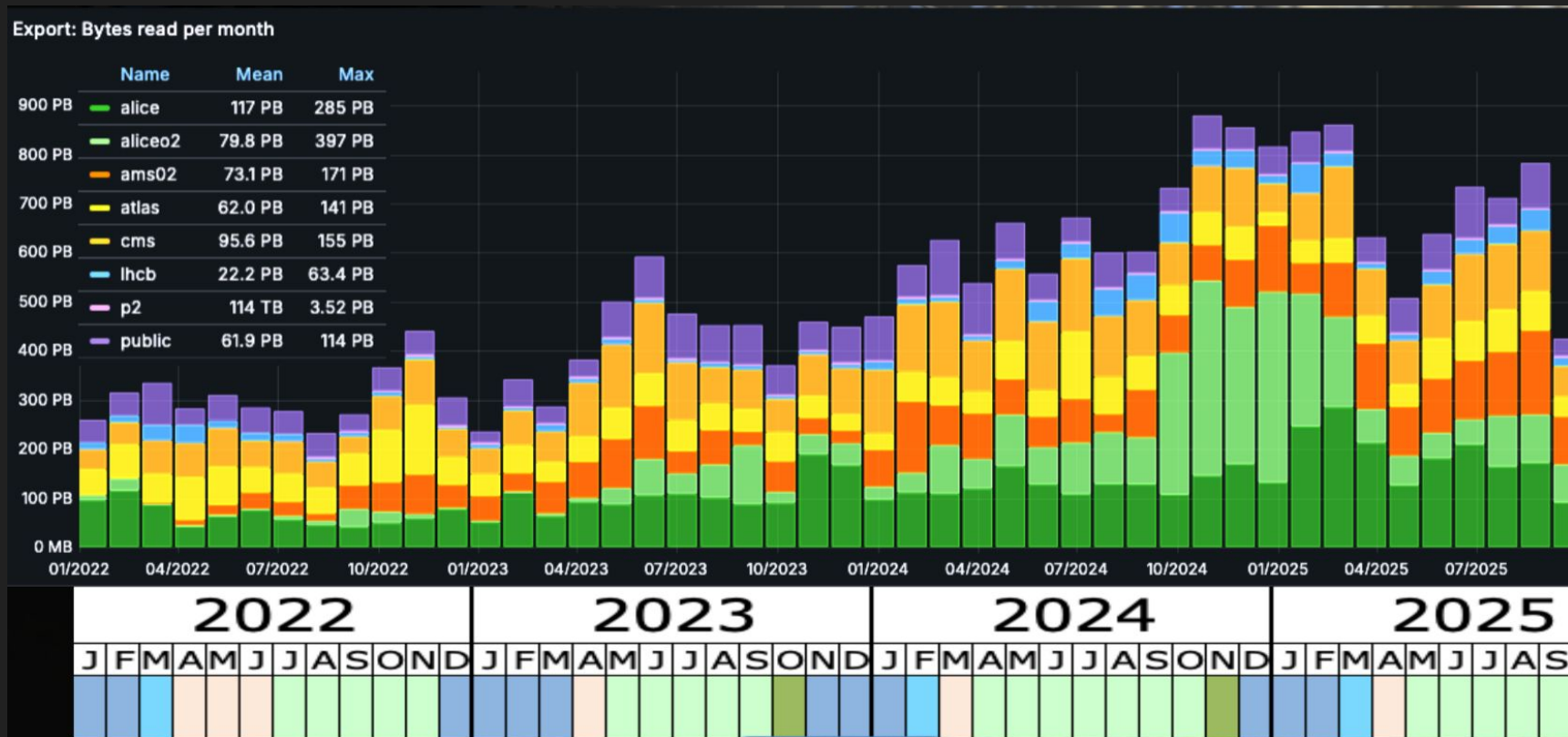
8 Physics 8 CERNBox 8 CTA

@ CERN	Capacity	Disks	Storage servers
Smallest EOS instance (Project-i02)	3.36 PB	270	6
Biggest EOS instance (ALICE02)	181 PB	12 000	128

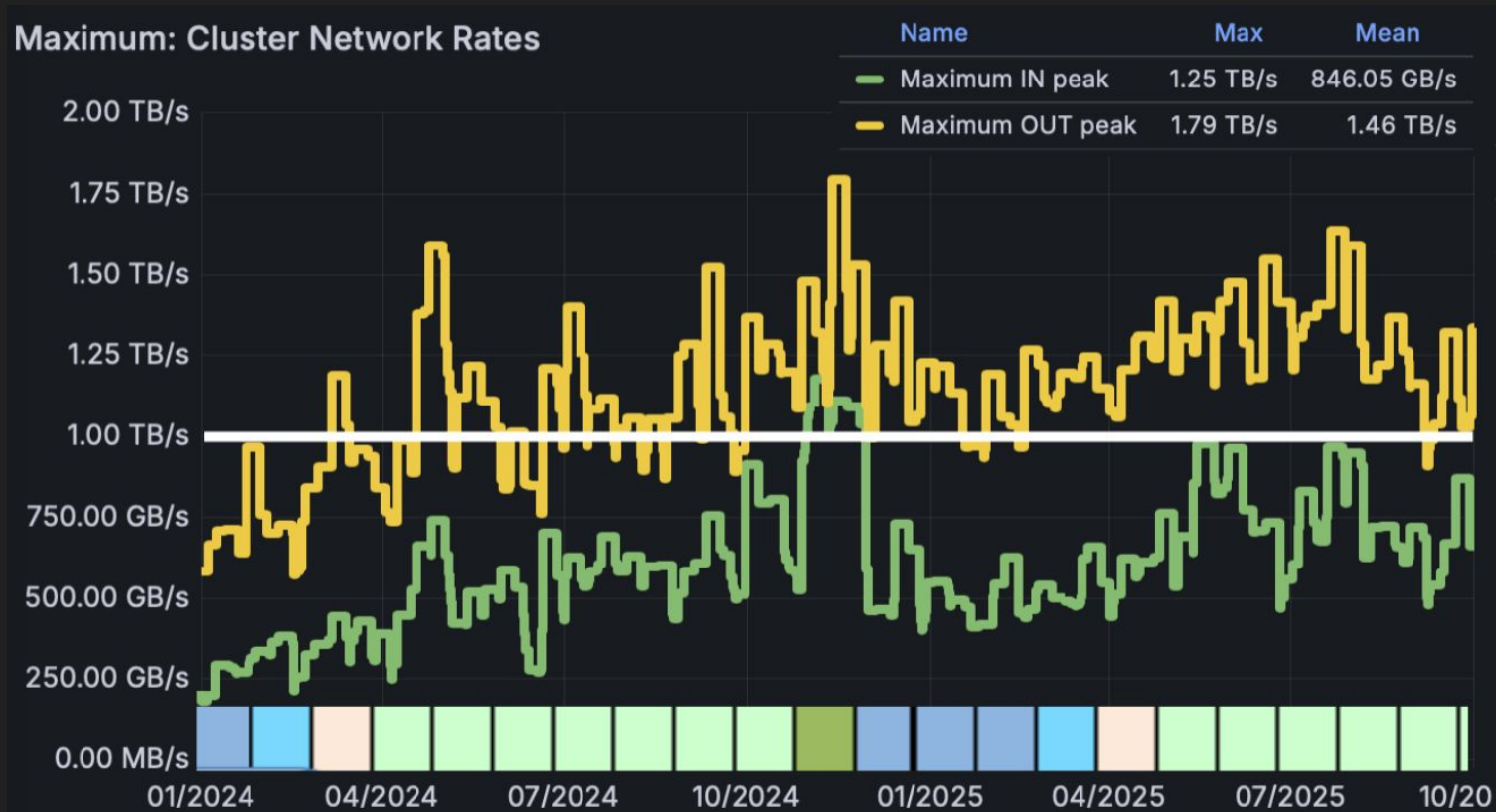
# EOS for physics since beginning of LHC RUN 3 - WRITES



# EOS for physics since beginning of LHC RUN 3 - READS



# EOS for physics 2024 and 2025 Network rates



# The public cloud storage equivalent of **EOS** at CERN

... has provided over **2.9 Million TB-years** of storage at a significantly lower cost!

The main goal of the EOS project is to deliver high-performance storage  
minimising the overall Total Cost of Ownership (TCO)



# At CERN - EOS Disk Data Storage

**959**

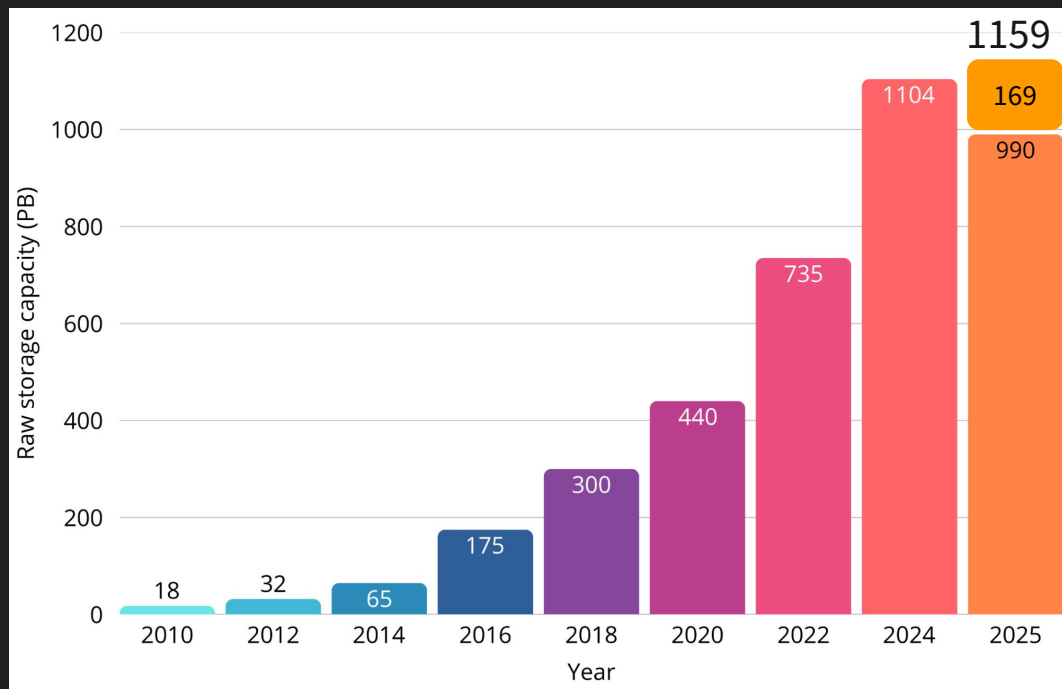
Disk storage servers

**111 000**

Disks

**990 PB**

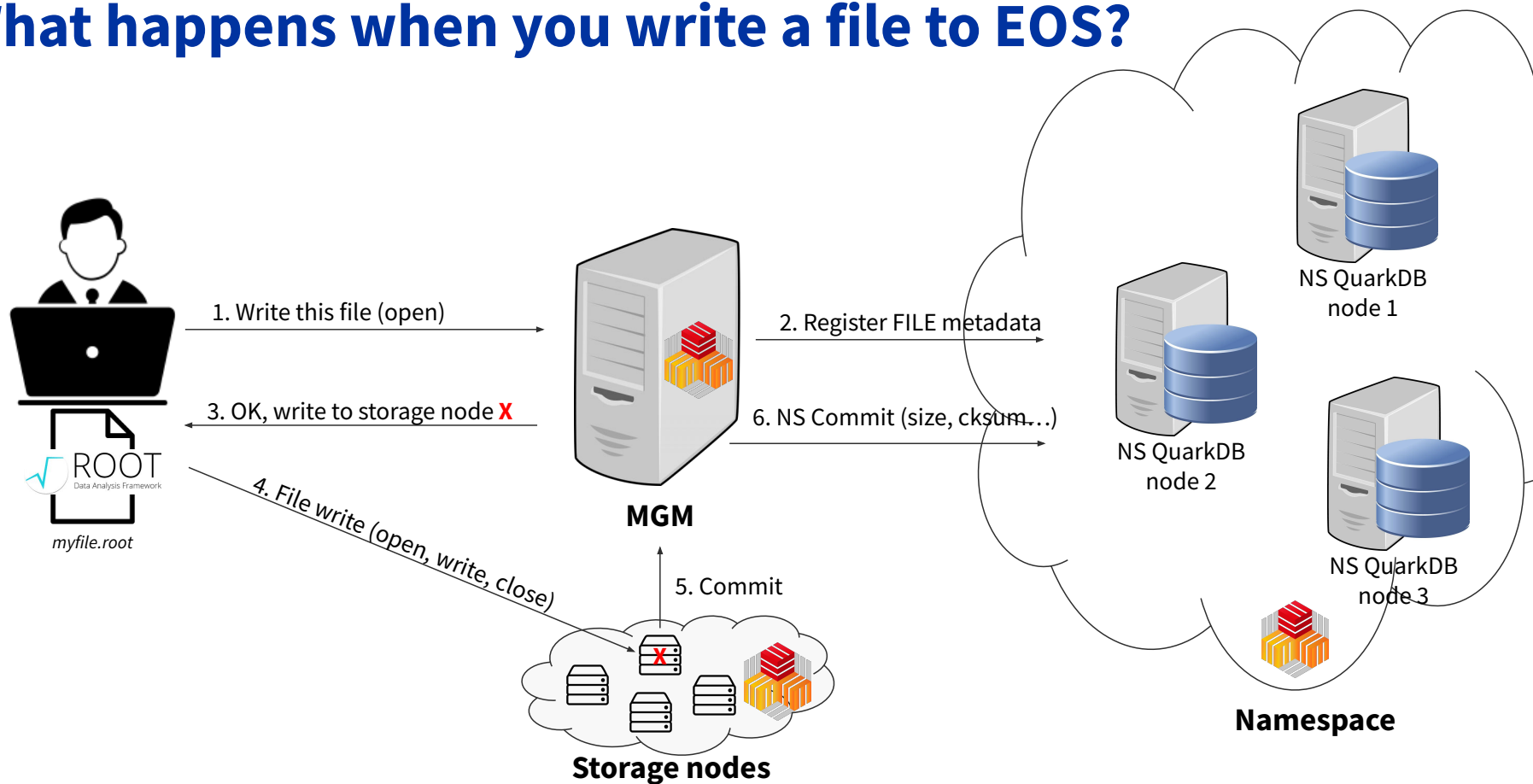
Raw capacity deployed



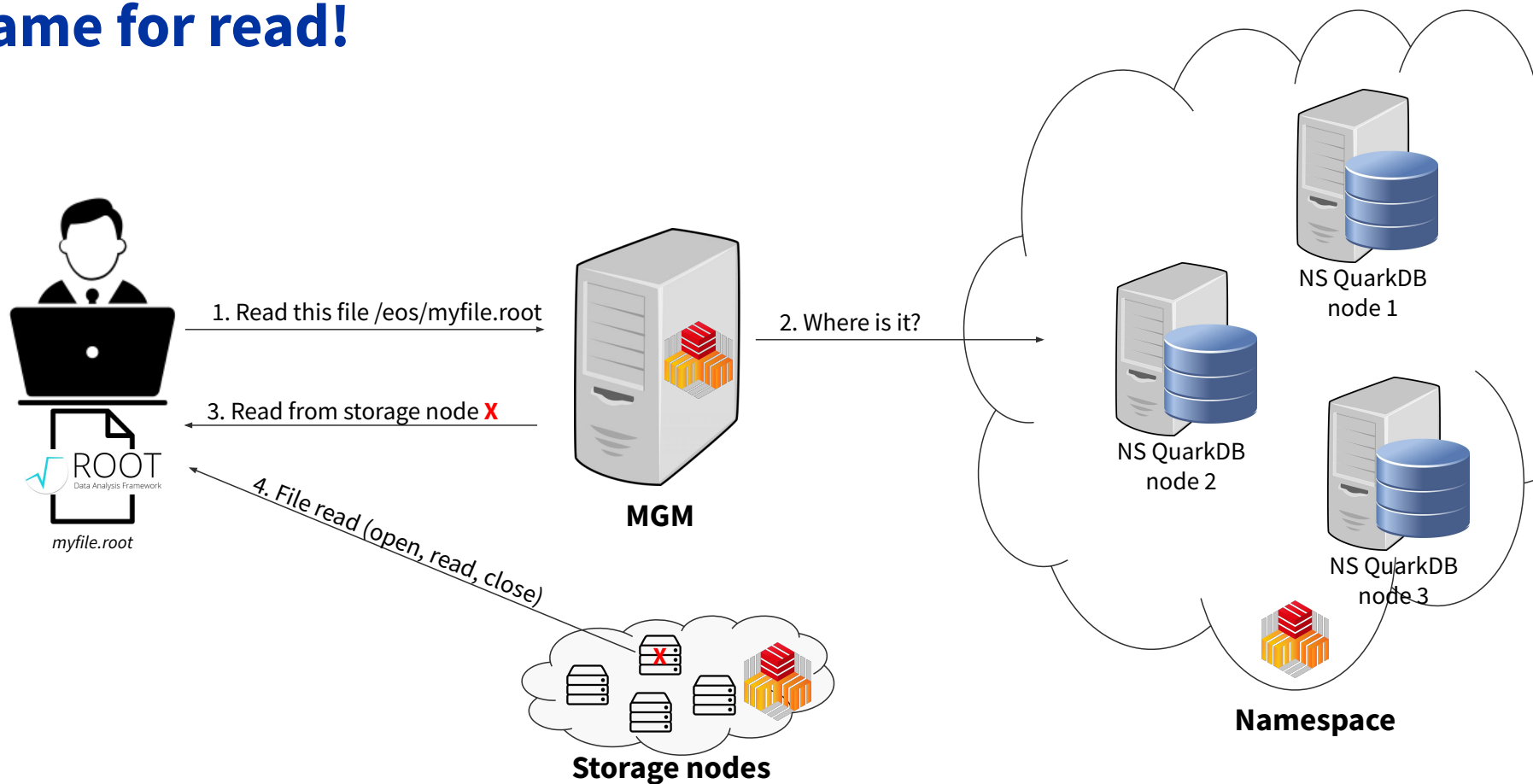
New hardware  
delivery in  
december!

**Raw storage capacity deployed over years @ CERN**

# What happens when you write a file to EOS?



# Same for read!



# EOS Architecture



- 3 entities to operate
  - MGM - The "brain" of EOS
  - FST - The physical data storage servers
  - QuarkDB - The namespace metadata key-value store

# Protocols



- Data transfer protocol supported

- XRootD

- Write: `xrdcp /etc/passwd root://eos.bnl.gov/eos/bnl/passwd`
- Read: `xrdcp root://eos.bnl.gov/eos/bnl/passwd /tmp/passwd`

- HTTPS

- Write: `curl --upload-file /etc/passwd https://eos.bnl.gov/eos/bnl/passwd`
- Read: `curl https://eos.bnl.gov/eos/bnl/passwd -o /tmp/passwd`

- FUSE

- EOS mounted directly on the local machine (`cp`, `cat`, `ls...`)

# Transfer data between two datacenters



- HTTP Third Party Copy (TPC)

- `curl -X COPY -H 'Source: https://eos.cern.ch/eos/cern/passwd'`  
`https://eos.bnl.org/eos/bnl/passwd`
- No data streaming via the client (see next slide)

# HTTP Third Party Copy (TPC)

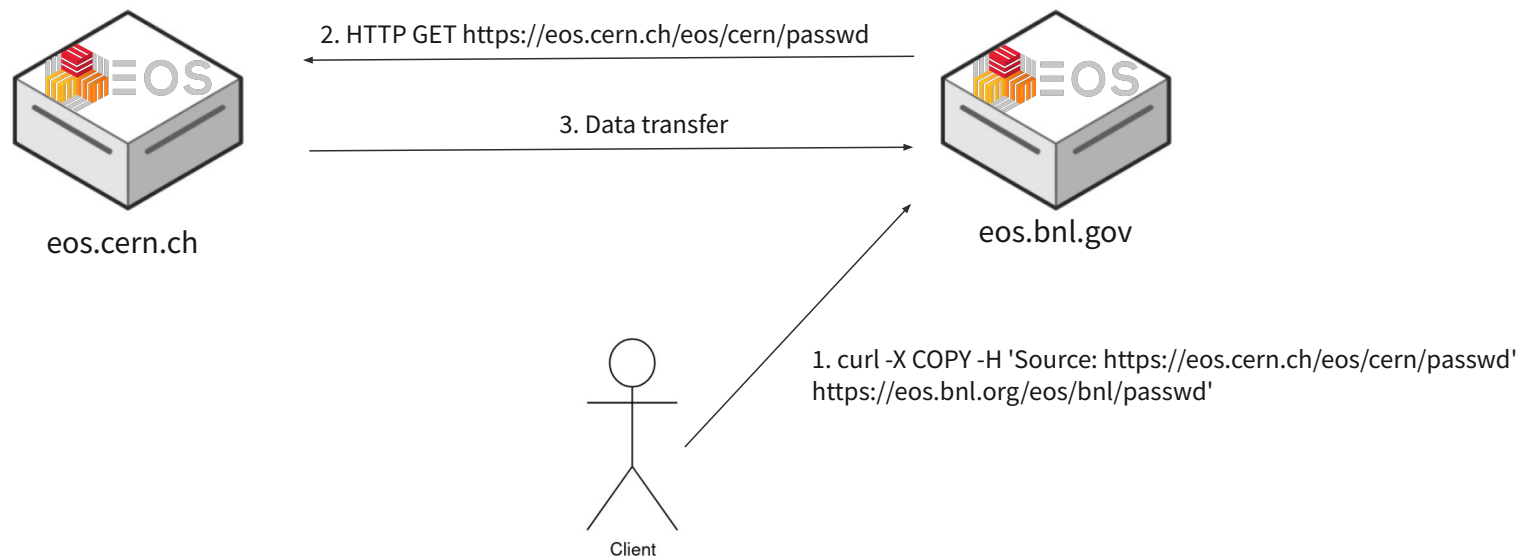


image: Flaticon.com

# EOS Authentication



- X509 certificate authentication
  - DN to username mapping
  - VOMS role to username mapping
- Kerberos - only for XRootD transfers
  - kinit [username@BNL.GOV](#) (only for xrootd transfers)
- Tokens
  - Scientific tokens (SciTokens)
  - EOS-native tokens

# EOS Authorization



- POSIX

```
EOS Console [root://localhost] |/> cd /eos/dteam001
EOS Console [root://localhost] |/eos/dteam001/> ls -alhrt
drwx----- 1 dteam001 dteam001    104.86 M Sep  1 11:42 .
drwxrwxr-x  1 root      root        418.57 M Nov 25 09:24 ..
drwx----- 1 dteam001 dteam001    104.86 M Sep 26 14:40 http_test
-rw-r----- 1 dteam001 dteam001      1491 May 30  2025 passwd
-rw-rw-r--  1 dteam001 dteam001     2514 Jun  3  2025 passwd2
```

# EOS Authorization



- Access Control List (ACL)

Example: `sys.acl="g:1470:rx!d,u:111522:rwx+d,egroup:lhcb-datamanagement:rwx+d"`

tag	definition
r	grant read permission
w	grant write permission
x	grant browsing permission
m	grant change mode permission
!m	forbid change mode operation
!d	forbid deletion of files and directories
+d	overwrite a 'd' rule and allow deletion of files and directories
!u	forbid update of files
+u	overwrite a 'u' rule and allow updates for files
q	grant 'set quota' permissions on a quota node
c	grant 'change owner' permission on directory children
i	set the immutable flag
a	grant archiving permission

# EOS Quota system



- Set files/storage limit on a specific tree for users and groups

```
└─> Quota Node: /eos/fuse/
```

user	used bytes	logi bytes	used files	aval bytes	aval logib	aval files	filled[%]	vol-status	ino-status
ccaffy	212.33 MB	212.33 MB	10	0 B	0 B	0	100.00 %	ignored	ignored

group	used bytes	logi bytes	used files	aval bytes	aval logib	aval files	filled[%]	vol-status	ino-status
99	0 B	0 B	0	0 B	0 B	0	100.00 %	ignored	ignored
ccaffy	212.33 MB	212.33 MB	10	20.00 GB	10.00 GB	2.00 M	2.12 %	ok	ok

summary	used bytes	logi bytes	used files	aval bytes	aval logib	aval files	filled[%]	vol-status	ino-status
All users	212.33 MB	212.33 MB	10	0 B	0 B	0	100.00 %	ignored	ignored
All groups	212.33 MB	212.33 MB	10	20.00 GB	10.00 GB	2.00 M	2.12 %	ok	ok

# Ensure data availability on EOS



- Configurable data storage layout to ensure data availability
  - Replica layout (1-replica, 2-replica, ..., N-replica)
    - Create N copies of the same file on different data storage servers
  - Erasure-coding (RAIN - Redundant Array of Inexpensive Nodes)
    - Split, encode and create parity stripes every  $<N>$  MB block of a file
    - Distribute all stripes across different storage servers
    - Ex: EC(10+2): 10 data stripes, 2 parities distributed across 12 servers

# Ensure data integrity on EOS



- Checksumming

- adler32, crc32c, md5, blake3, sha-1, sha-256
- Multiple checksums for same file support with EOS > 5.4 (2026)

# Ensure data integrity on EOS

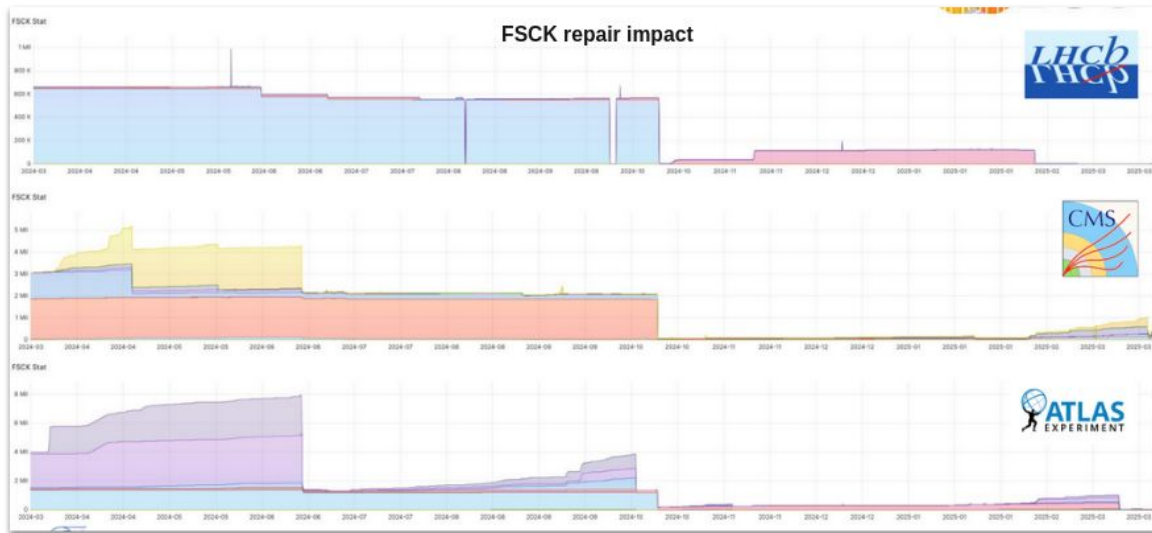


- EOS FSCK (FileSystem Check)

- Collection: Background task scanning files on all data storage servers
- Repair: Early fix inconsistency error to avoid data loss

```
[root@eosams02-ip563 (mgm:master) ~]$ eos fsck stat
Info: collection thread status -> enabled
Info: repair thread status -> enabled
Info: repair category -> all
Info: best effort -> true
250328 00:02:41 1743116561.474107 Start error collection
250328 00:02:41 1743116561.474116 Filesystems to check: 4408
250328 00:02:41 1743116561.497339 d_cx_diff : 4
250328 00:02:41 1743116561.497356 d_mem_sz_diff : 45
250328 00:02:41 1743116561.497362 m_cx_diff : 3
250328 00:02:41 1743116561.497379 m_mem_sz_diff : 14
250328 00:02:41 1743116561.497386 rep_diff_n : 2408
250328 00:02:41 1743116561.497387 rep_missing_n : 30
250328 00:02:41 1743116561.497391 stripe_err : 175
250328 00:02:41 1743116561.497392 unreg_n : 26
250328 00:02:41 1743116561.497394 Finished error collection
250328 00:02:41 1743116561.497404 Next run in 5 minutes
```

FSCK collection summary



# Operating EOS



- EOS CLI interface

- Administrate data storage nodes (FST), disks, namespace operation
- `eos node ls`, `eos fs ls`, `eos ns ...`
- EOS was conceived to help storage operators, plenty of monitoring information available at hand

- GRPC interface

- CERNBox and CTA integration

# Operating EOS



- EOS microservices - managed internally
  - Balancers: balance the volume usage on disks
  - Converter engine: Change data storage layout
  - LRU engine: namespace cleanup or file conversion
    - E.g: create a temporary space where files get automatically cleaned up after 30 days

# Operating EOS @ CERN



- File-transfer probe

- Is EOS up and running?
- EOS probe
  - Copy, download, delete a file to EOS
  - In case of failure
    - SMS to EOS operation team
    - Mattermost notification
    - Email notification

Gitlab link to the probe: <https://gitlab.cern.ch/eos/ops/probe>

# Operating EOS @ CERN



- Automate tasks using rundeck
  - Automated Data Servers gradual rolling EOS upgrade
  - Gracefully shutdown data storage server for hardware repair
  - EOS Logs Backups
  - EOS Namespace Backups

# Operating EOS @ CERN



- Anomaly Detection: EOS Winston
  - Integrated to Mattermost Chat

Command	Description
/winston diagnostic [instance]	The diagnostic for a specific instance. If the option is not there, it will run a diagnostic on all instances
/winston eosversionsummary	EOS Version summary
/winston eosclientsummary	EOS Client summary
/winston eosclientversionsummary	EOS Clients Version summary
/winston eosaccesssummary	User Access summary

# Operating EOS @ CERN



- Anomaly Detection: EOS Winston

- Integrated to Mattermost

eos-winston BOT 2:46 PM  
⚠ Processing the request "/winston diagnostic eoslhcb"...please wait...

eos-winston BOT 2:46 PM

⚠ EOSLHCB

**Quota issues. Please review:**

- 🚫 The amount of used bytes reported by the quota view is different from the one reported by the space view by 4.55 PB, maximum permitted is 1.00 PB

eos-winston BOT 2:46 PM  
⚠ Processing the request "/winston diagnostic eoscms"...please wait...

eos-winston BOT 2:46 PM

⚠ EOSCMS

**Filesystem issues. Please review:**


- 🚫 Total count of faulty FS on Instance: 26 (Except spare or drained FS and restarting FSTs)
- 🚫 Node: st-096-100gb-ip308-69665.cern.ch has 24/95 disks not in "booted rw" (25.26% > 10.00% threshold). Disks in "spare" are ignored.


Status	Count
Opseerror	26

# Operating EOS @ CERN



- Anomaly Detection: EOS Winston
  - Integrated to Mattermost

 eos-winston BOT 2:47 PM  
⚠ Processing the request "/winston eosversionsummary"...please wait...

 eos-winston BOT 2:47 PM

---

**Latest EOS reports as of 10 Dec 2025 14:47**

---

**Instance: eosalice**  
Headnode version: eos-server-5.3.22-1.el9.x86\_64

- 106 nodes have version 5.3.22-1.

---

**Instance: eosatlas**  
Headnode version: eos-server-5.3.26-1.el9.x86\_64

- 105 nodes have version 5.3.22-1.

---

**Instance: eoscms**  
Headnode version: eos-server-5.3.25-1.el9.x86\_64

- 119 nodes have version 5.3.22-1.

---

**Instance: eoslhcb**  
Headnode version: eos-server-5.3.26-1.el9.x86\_64

- 72 nodes have version 5.3.25-1.

# Operating EOS @ CERN



- AI chatbot - PoC

- Integrated to Mattermost
- Available on the [eos-community](#) forum



accgpt BOT 10:27 AM (Only visible to you)

Got it! Processing your question: `*"How do I display all storage nodes in EOS?..."`



accgpt BOT 10:27 AM

Question: How do I display all storage nodes in EOS?

Answer: To display all storage nodes in EOS, you can use the command: `eos node ls`. This command will list the nodes along with their details such as host, port, id, path, schedgroup, and status. For more detailed information, you can use additional output options like `-l` or `--io` with the command, for example: `eos node ls -l`.


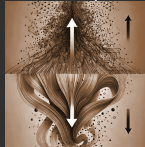




## Sources

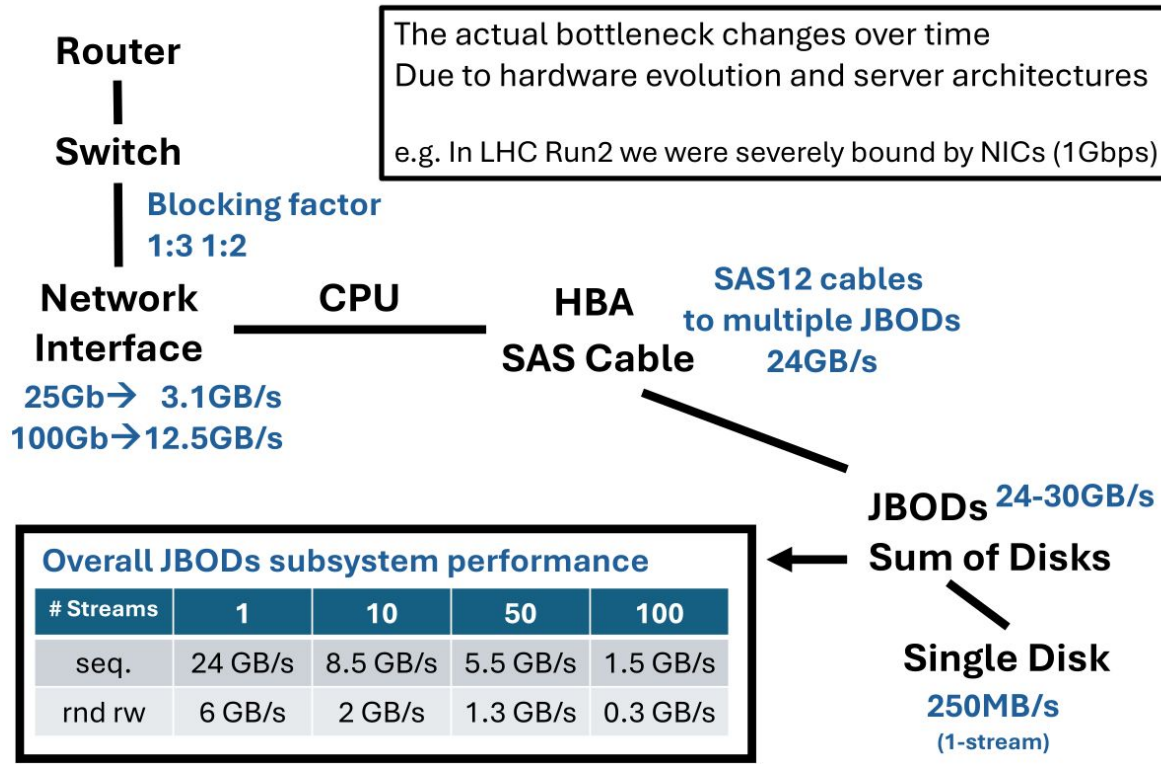
1. [eos-docs.web.cern.ch](#) (Confidence: 60.6%)
2. [eos-docs.web.cern.ch](#) (Confidence: 59.5%)

# EOS Future



- The evolution of **EOS** is still ongoing:
  - **Simplification and Performance Improvement** of the overall architecture 
  - Tiered-Storage: **EOS-Flow** 
  - **EOS-R&D** of the **EOS-Connect** architecture merging disk & tape deployments 
  - **EOS-R&D** with new technologies like QLC flash 

# Tracking of Disk Storage HW Bottlenecks



# Disks are bigger, but not faster

## Latest WD on Market

All latest technological advancements

- 11-platters
- ePMR
- Triple-stage Actuator
- Helio Sealed
- UltraSMR
- OptiNAND
- ArmorCache

[Link](#)



	30TB SATA	30TB SAS	32TB SATA	32TB SAS
Model Number	WSH723200ALxxyz	WSH723200ALxxyz	WSH723220ALxxyz	WSH723220ALxxyz
Formatted capacity <sup>1</sup>	30TB	30TB	32TB	32TB
Recording Technology	SMR	SMR	SMR	SMR
Interface	SATA 6 Gb/s	SAS 12 Gb/s	SATA 6 Gb/s	SAS 12 Gb/s
Format: Sector size (bytes)	512e/512   4Kn: 4096	512e/512   4Kn: 4096	512e/512   4Kn: 4096	512e/512   4Kn: 4096
Areal density (Gbits/sq. in.)	1385	1385	1485	1485
<b>Performance</b>				
<b>Max ~270 MB/s (1-stream)</b>				
Data buffer <sup>2</sup> (MB)	512	512	512	512
Rotational speed (RPM)	7200	7200	7200	7200
Latency average (ms)	4.16	4.16	4.16	4.16
Interface transfer rate (MB/s, max)	600	1200	600	1200
Sustained transfer rate* (MB/s, max) / (MiB/s, max)	260 / 248	260 / 248	269 / 256	269 / 256

## Comparison with other products

### WD Gold® Enterprise Class SATA HDD

#### Specifications

**~280-300 MB/s (1-stream)**

Model Number	WD261KRYZ	WD242KRYZ	WD241KRYZ	WD221KRYZ	WD203KRYZ	WD202KRYZ
Formatted capacity <sup>1</sup>	26TB	24TB	24TB	22TB	20TB	20TB
Form factor	3.5-inch	3.5-inch	3.5-inch	3.5-inch	3.5-inch	3.5-inch
Interface	SATA 6 Gb/s	SATA 6 Gb/s	SATA 6 Gb/s	SATA 6 Gb/s	SATA 6 GB/s	SATA 6 Gb/s
512n / 512e user sectors per drive <sup>5</sup>	512e	512e	512e	512e	512e	512e
OptiNAND™ technology	Yes	Yes	Yes	Yes	No	Yes
ArmorCache™	Yes	Yes	Yes	Yes	No	No
RoHS compliant <sup>6</sup>	Yes	Yes	Yes	Yes	Yes	Yes
<b>Performance</b>						
Data transfer rate* (max sustained)	285MB/s	279MB/s	298MB/s	291MB/s	285MB/s	285MB/s
RPM	7200	7200	7200	7200	7200	7200
Cache (MB) <sup>7</sup> *	512MB	512MB	512MB	512MB	512MB	512MB

# Servers hold more data, but deliver less speed!

100 PB comparison over hardware generations

	HW Generation 2017 10Gbps 48x6TB	HW Generation 2018-2019 25Gbps 96X12TB	HW Generation 2020-2022 100Gbps 96x18TB	HW Generation 2024-2025 100Gbps 120x24TB
Number of servers	347	87	58	35
Sum of NICs	433 GB/s	272 GB/s	725 GB/s	425 GB/s
Disk Speed (est.)	150MB/s per disk	200MB/s per disk	250MB/s per disk	250MB/s per disk
Sum of Disk BW (1-stream)	~2.5TB/s	~1.67TB/s	~1.39TB/s	~1TB/s
Sum of Disk BW (10-streams)	~890 GB/s	~600 GB/s	~500 GB/s	~360 GB/s

What can we do about it?

# New functionality: Storage Tiering

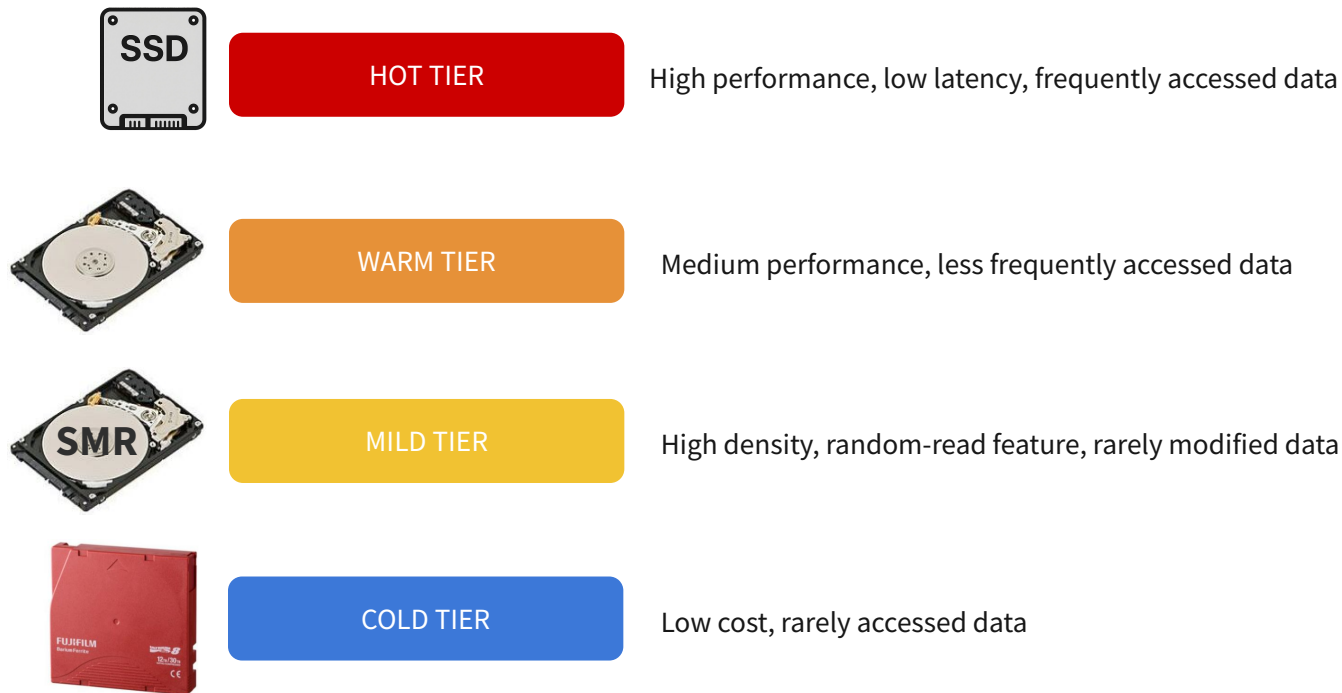
**Categorize and store data on  $\neq$  types of media**

Based on

- Performance
- Costs
- Access frequency

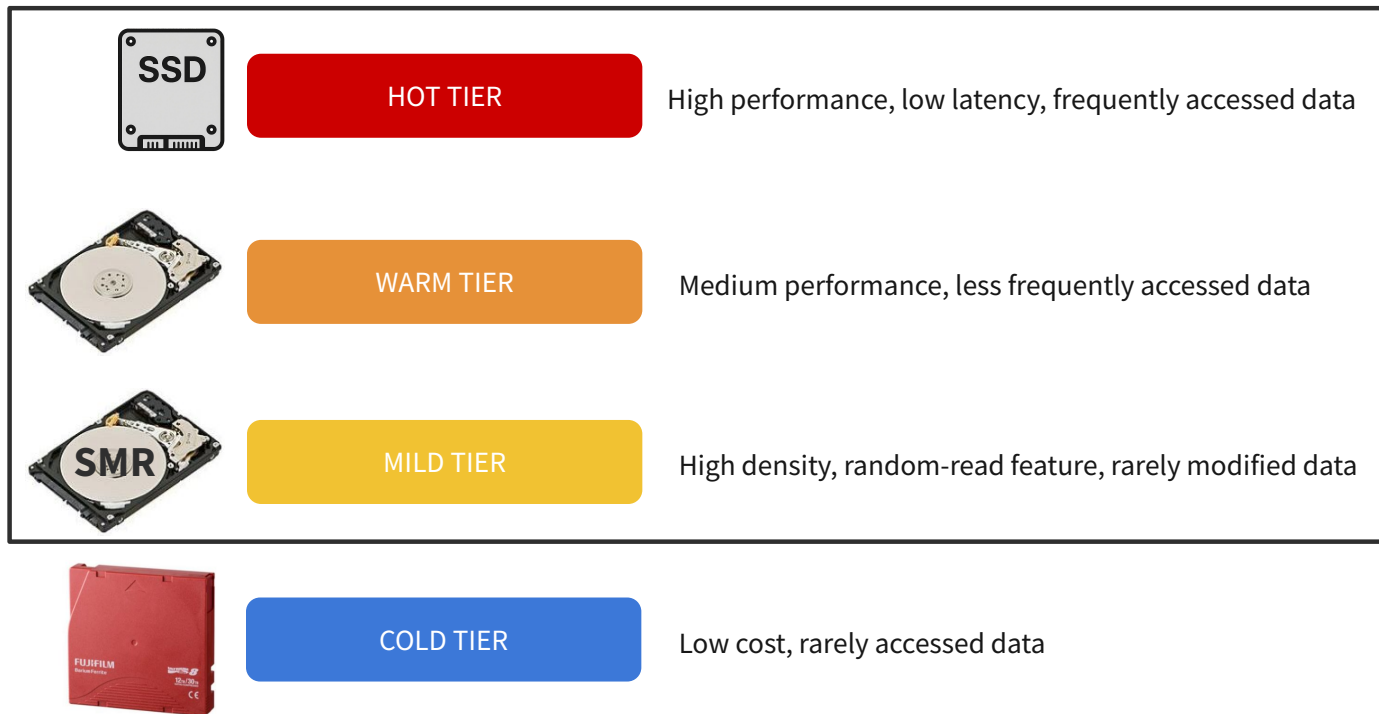
# New functionality: Storage tiering

## Media-based storage tiering



# New functionality: Storage tiering

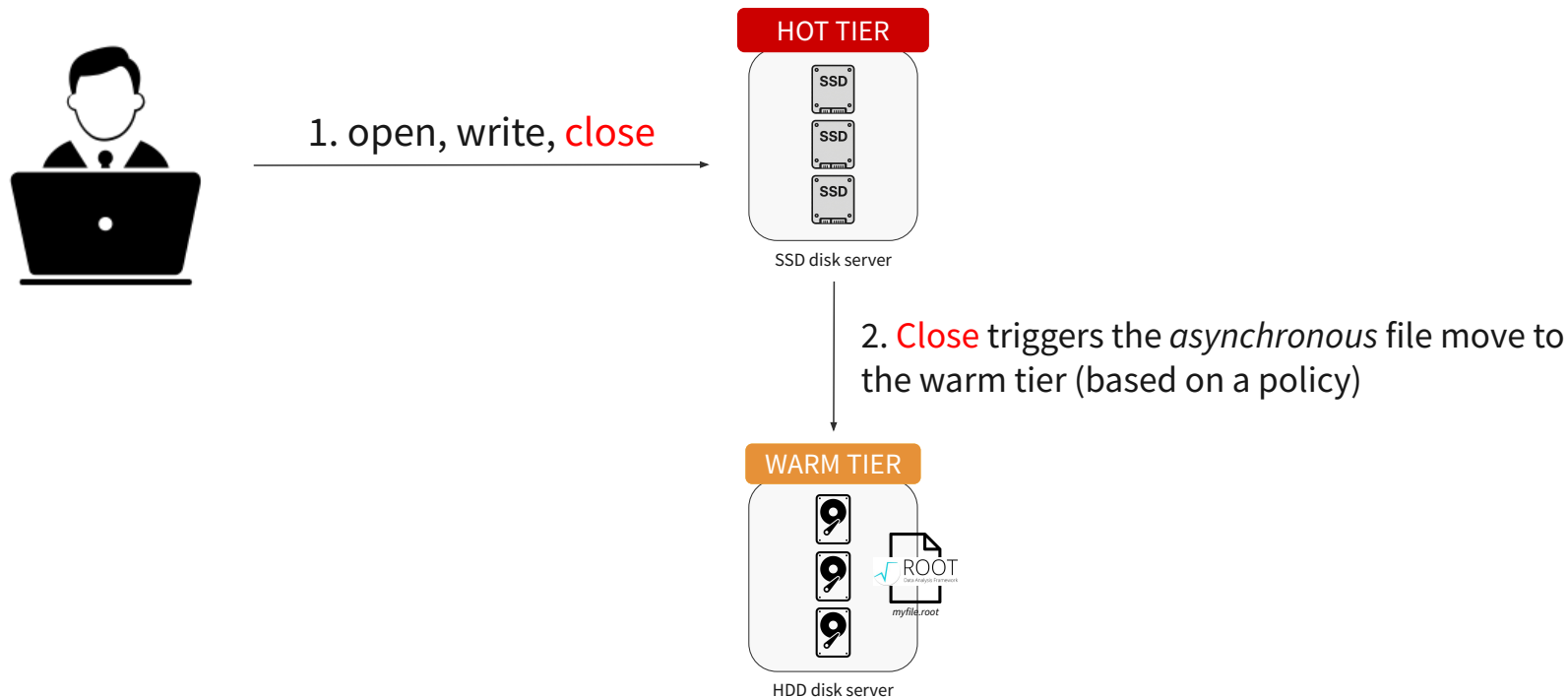
## Media-based storage tiering



# New functionality: Storage tiering



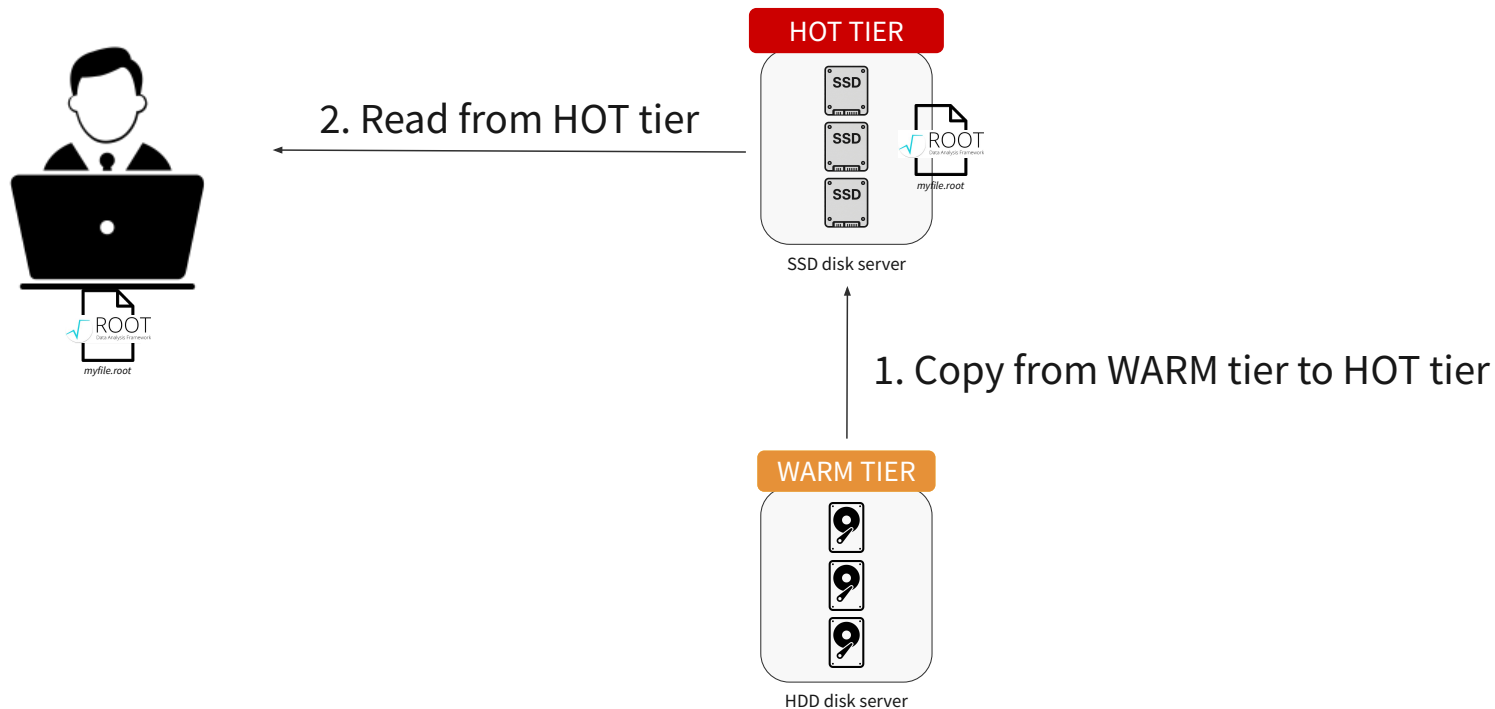
## Media-based storage tiering - file write



# New functionality: Storage tiering



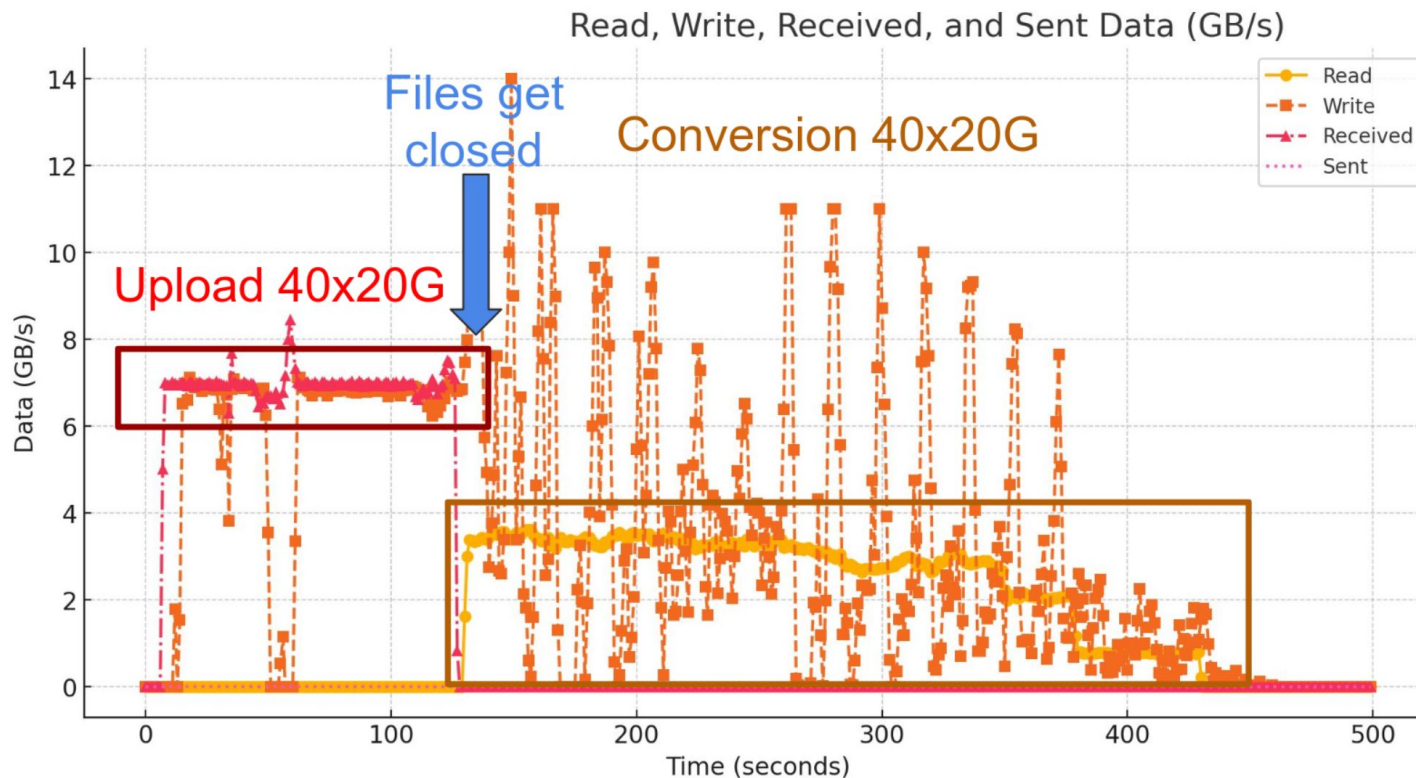
## Media-based storage tiering - file read



# Storage tiering benchmark



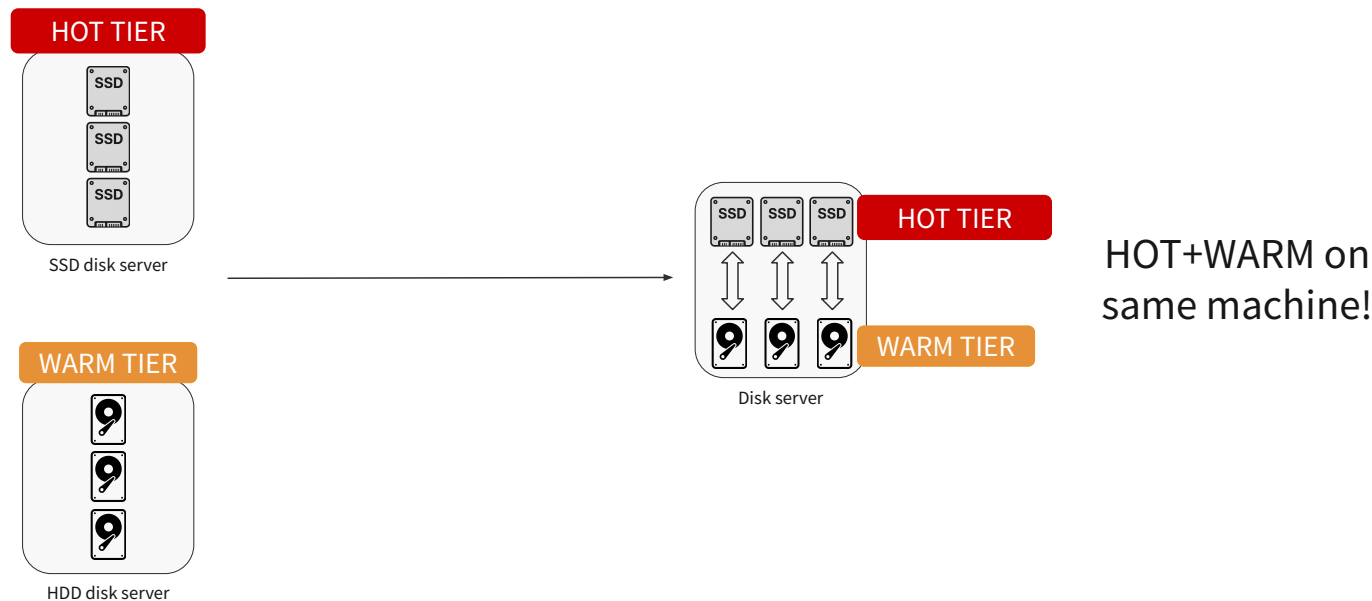
## Result



# New functionality: Storage tiering



More scalable approach - local tiering - WIP



More infos: <https://eos-docs.web.cern.ch/diopside/manual/interfaces.html#storage-tiering>



# New Functionality: Notifications / Notification Workflow

Example Use Case: transparent Rucio for small experiments

- avoids RUCIO/FTS integration in DAQ systems





## Supports four notification transports:

- http(s) (POST)
- grpc (Notify rpc)
- activeMQ (Message)
- redis (PUBLISH)

```
# REDIS rpush notification - the port is given as an extra
argument
eos attr set
sys.workflow.closew.default=notify:redis|localhost|6349|notification|2000
```

```
# GRPC notification - the port is given as an extra argument,
# <queue> is empty
eos attr set
sys.workflow.closew.default=notify:grpc|localhost|55100||2000
```

# New Functionality: Audit Logging

## Documentation



AUDIT LOGGING

AUDIT LOGGING

**Audit:** a new feature on the MGM logging all modifications of the namespace and on demand even all inspections of the namespace (readdir, stat)

### Use case:

Who modified this?

When?

What was changed?

(meta-data only)

```
{
  "timestamp": 1730985600,
  "path": "/eos/user/a/alice/data/file.txt",
  "operation": "WRITE",
  "client_ip": "192.0.2.10",
  "account": "alice",
  "auth": { "mechanism": "krb5", "attributes": {"principal":
"alice@EXAMPLE.ORG"} },
  "authorization": { "reasons": ["uid-match"] },
  "trace_id": "srv-abc123",
  "uuid": "550e8400-e29b-41d4-a716-446655440000",
  "tid": "cli-xyz789",
  "app": "eoscp",
  "svc": "mgm",
  "before": { "ctime": 1730980000, "mtime": 1730981000, "uid": 1000, "gid":
1000, "mode": 420, "mode_octal": "0100644", "size": 1024, "checksum":
"a1b2..." },
  "after": { "ctime": 1730980000, "mtime": 1730985600, "ctime_ns":
"1730980000.000000000", "mtime_ns": "1730985600.123456789", "uid": 1000,
"gid": 1000, "mode": 420, "mode_octal": "0100644", "size": 4096, "checksum":
"dead..." },
  "src_file": "mgm/FuseServer/Server.cc",
  "src_line": 2600,
  "version": "<eos-version>"
}
```

# New Functionality: ZSTD compressed log files and reports

## ZSTD Logs

## ZSTD Reports

## ZSTD Audits

**Audit** Logs are always ZSTD compressed, **Log files** are optionally ZSTD compressed, same for **Report logs**.

One difference is, that there is no **need for external log rotation anymore** and the most recent file is a symlink to the time labeled log file.

We provide an executable **zstdtail**, which behaves like the LINUX tail command to follow logfiles for parsing etc.

By default ZSTD comes with **zstdcat** and **zstdless**.

- Stream current audit records:

```
zstdcat <logdir>/audit/audit.zstd | jq '.'
```

- Follow audit logs across rotations (like `tail -F`):

```
zstdtail <logdir>/audit/audit.zstd
```

```
# Or with filtering:
```

```
zstdtail <logdir>/audit/audit.zstd -- jq 'select(.operation == "DELETE")'
```

# Save the date! EOS Workshop 2026

9th - 11th of March 2026 at CERN

Thank you! Questions?

# Backup

# Backup slide - questions from Tony W.

- What storage needs/constraints do you face or your existing storage solution address? What criteria are used to choose it over other available solutions?
  - Storage needs
    - Store PetaBytes of expensive physics data efficiently and reliably at the lowest cost possible
    - Ensuring full control and complete insight into the software to enable rapid incident response and resolution
    - Ensuring the timely addressing of new requirements from CERN's experiments
    - 24/7 data availability for physics use-cases at low operational costs
  - Constraints
    - Fixed budget, very heterogeneous hardware (~10 years old data storage servers)
    - Must interact with the Worldwide LHC Computing Grid (data created at CERN is also stored all around the globe)
    - Support multiple physics use cases from DAQ data storage to data reconstruction, data reprocessing, data analysis and ensure these activities don't interfere with each other

# Backup slide - questions from Tony W.

- What are the advantages and disadvantages of your storage solution? What does it do well, and where does it need improvement, operationally-speaking?
  - Advantages
    - Scale-out data storage architecture
    - Multi-protocol and Multi-authentication (each protocol see the same set of data)
    - Integration with CTA and CERNBox
    - EOS was conceived to streamline and simplify service operations
      - Many tunable to protect (or limit) dedicated activities or users
  - Disadvantages
    - EOS is not providing block devices to end-users (RBD, DRBD, ...)
    - Special configuration can be tricky in some cases
      - This is why our team is there to advise and help :)
    - Performance depends on hardware used and its configuration
  - Possible operational improvements
    - In the future, we plan to provide native NFS access instead of using FUSE.

# Backup slide - questions from Tony W.

- 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?
  - Storage challenge for the future
    - HL-LHC (Run 4) expectations
      - Disk storage capacity: 3 EB of usable storage
      - Bandwidth: 4-5TB/s
      - Transfer rate: Avg 2.2M parallel transfer at all times
    - Incoming hardware has more capacity per machine → less bandwidth per storage!
      - Storage tiering will help in this respect
  - We are always evolving EOS to increase its performance and meet experiments needs