

TECH WEEK STORAGE 24



EOS Open Storage

Large Scale Disk Storage at CERN

Dr. Andreas-Joachim Peters

for the CERN IT - Storage Group

Council Chamber - CERN

14.03.2024

- What is EOS?
- Physics Data Storage at CERN
 - Use-cases
 - A brief history about EOS
- Architecture
 - Evolution
 - Deployment Model
- Software
 - Team, Version Highlights, R&D Activities
- Fundamental Concepts
- EOS for Physics at CERN and external Usage
- Summary & Outlook
- EOS Community and “how to join”



EOS

Born 2010
Open Source Storage System written in C++

About EOS

EOS provides a service for storing large amounts of physics data and user files, with a focus on interactive and batch analysis.

Flexible



EOS is a storage solution for central data recording, analysis and processing++

Adaptable and Scalable



EOS supports thousands of clients with random remote I/O patterns with multi protocol support
WebDAV, CIFS, FUSE, XRootd, GRPC.

Over 900 PB at CERN



Designed for high capacity and low latency.



Security

EOS offers a variety of authentication methods:KRB5, X509, OIDC, shared secret, and JWT and proprietary token authorisation.



Sync & Share

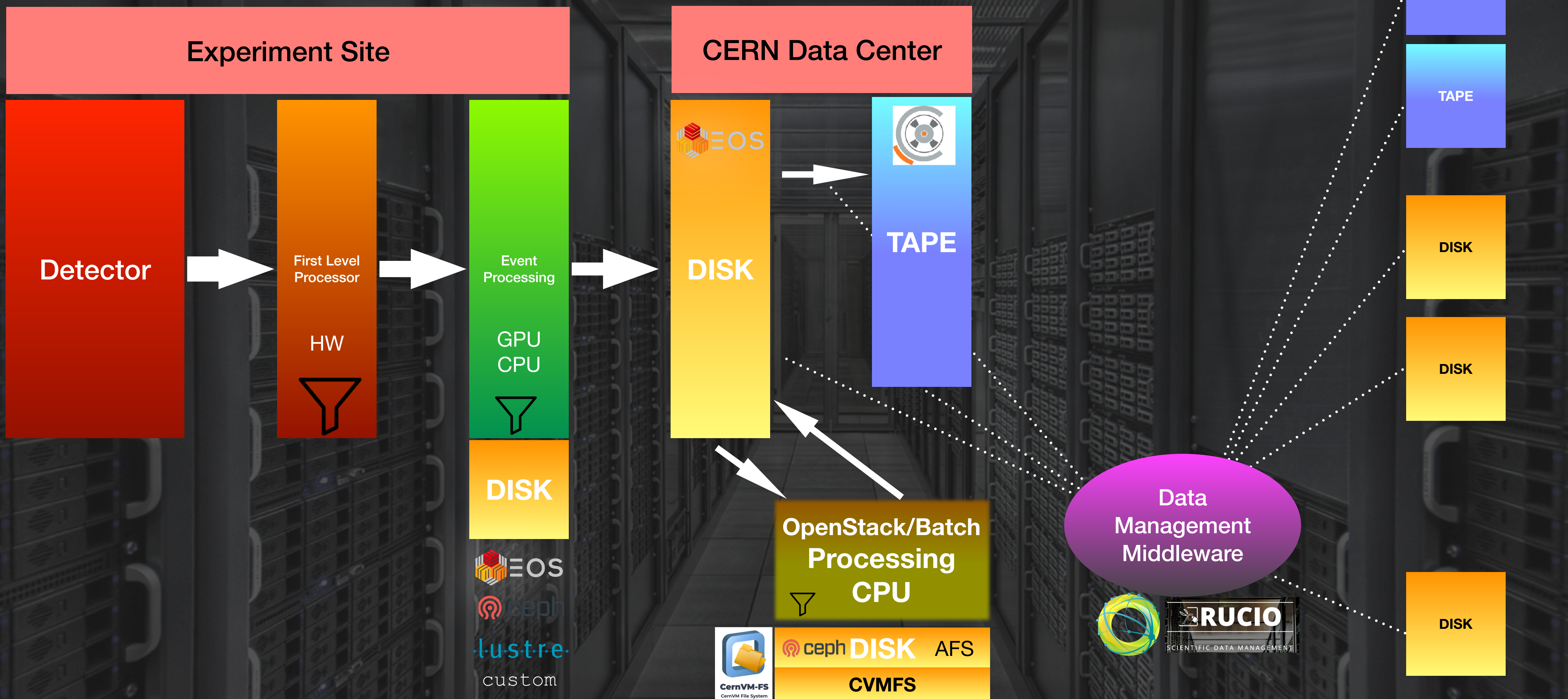
EOS provides Sync&Share functionality for the **CERNBox** front-end services.



Tape Storage

EOS includes tape storage in combination with the **CTA** Cern Tape Archive software.

EOS for Physics at CERN





Largest Disk Storage System at CERN 180 PB HDD space

12.000 HDDs

126 server

100 GE

EC 10+2

150 PB usable

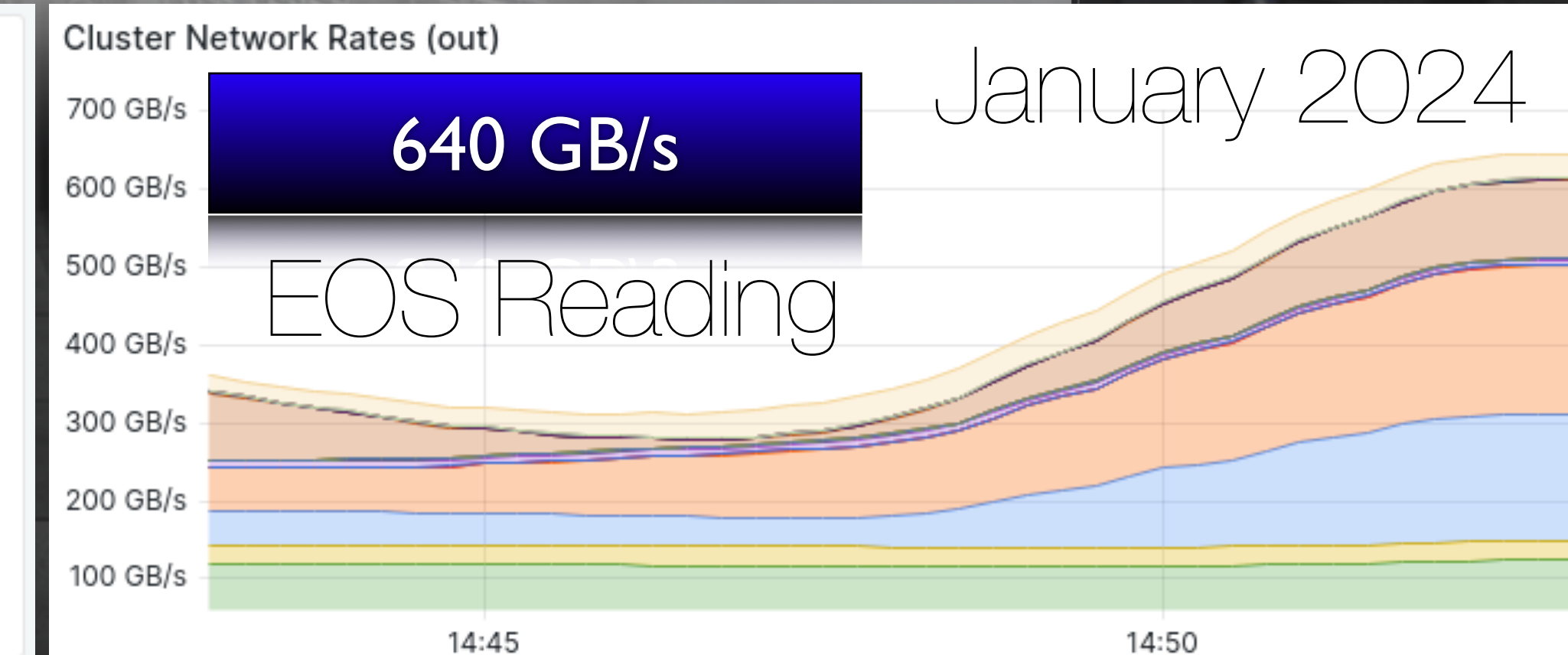
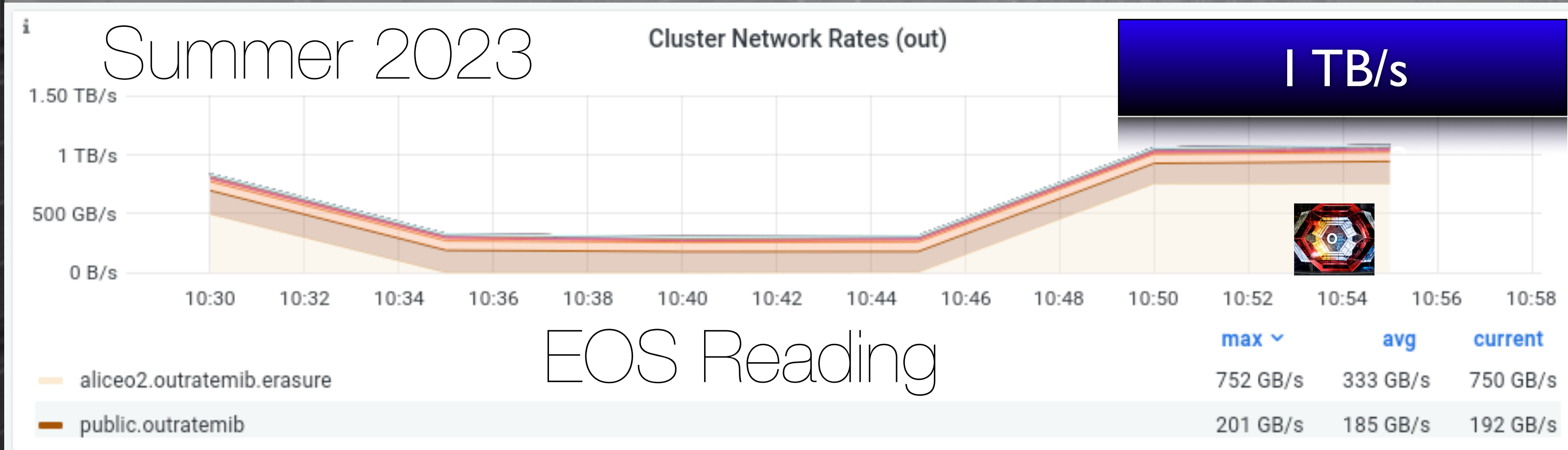
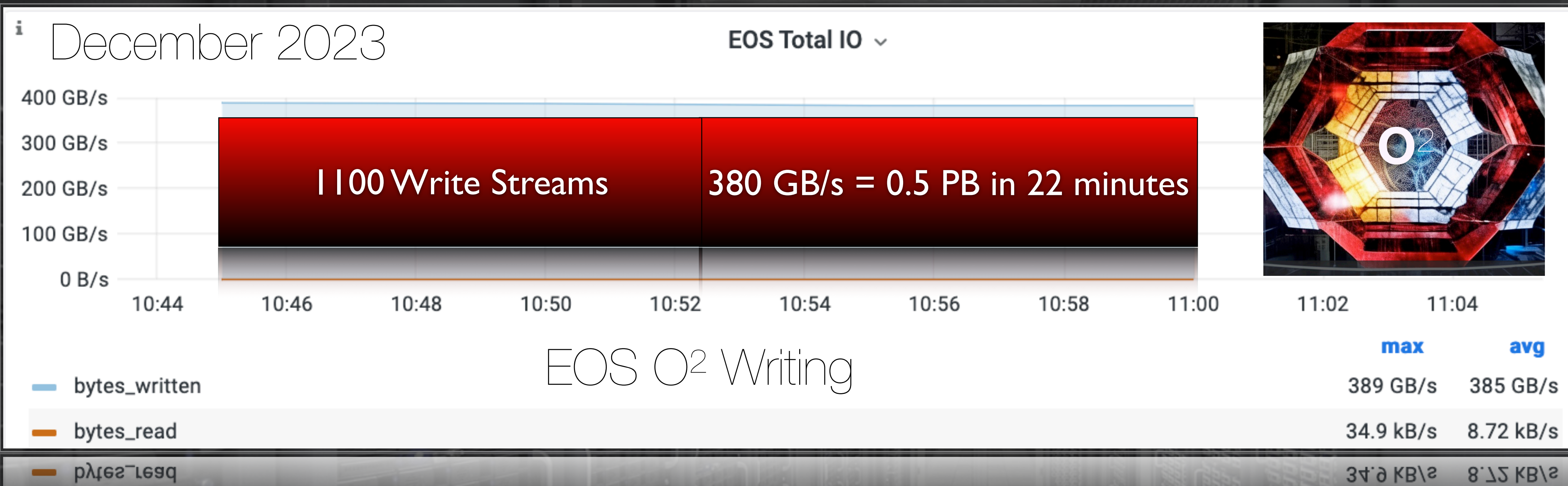
O₂

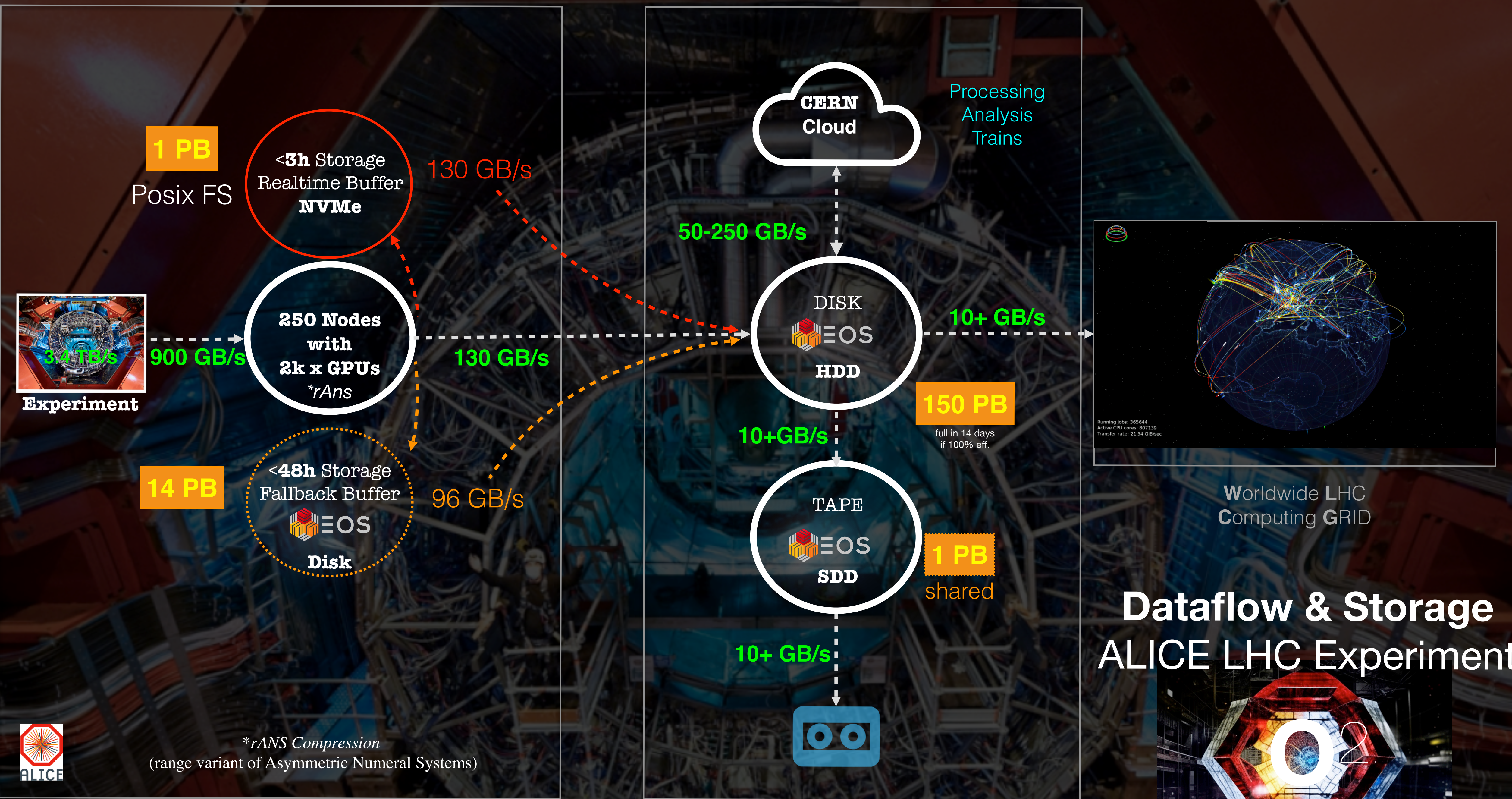


O² Instance **180 PB**

Photo shows 3840 HDDs = ~ 1/3

96 HDDs x 18 TB
per node, 1x **100GE**





*rANS Compression
(range variant of Asymmetric Numeral Systems)

CERN Experimental Site

CERN Computer Center



Production/Online SYSTEMS

GRID JOBS

BATCH JOBS

Interactive VMs

Personal Devices

Swan

CERNBox

WEB Services for **Jupyter Notebooks**

WEB Services for **Sync&Share**

EOS

24 individual instances
8 Physics 8 CERNBox 8 CTA

CTA
 EOS
CERN Tape Archive

>200 Tape Drives
720 PB

SSD
1 PB

70k HDDs - 930 PB



EOS Disk Storage at CERN

Files Stored
8 Bil

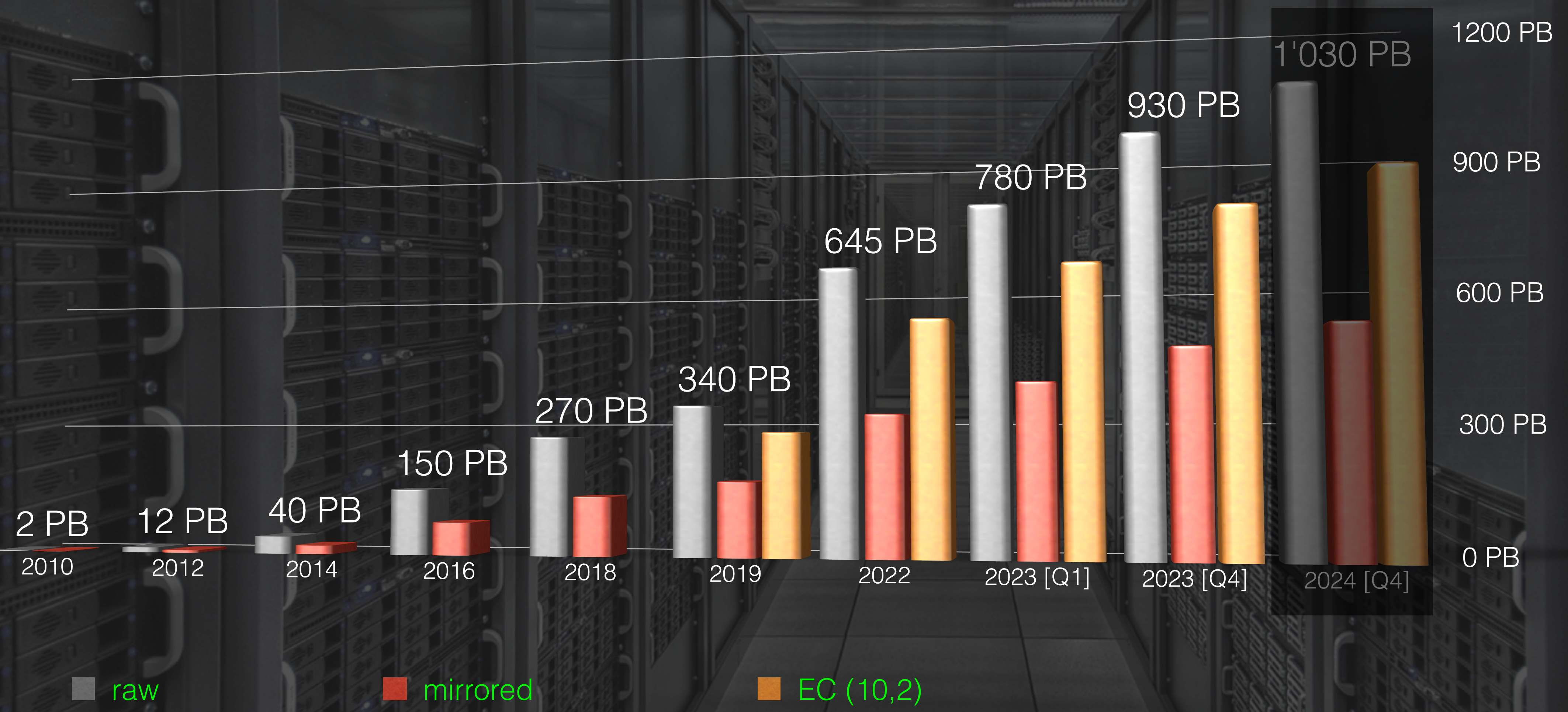
Storage Nodes
850

Hard Disks
70k

Raw Space
930 PB

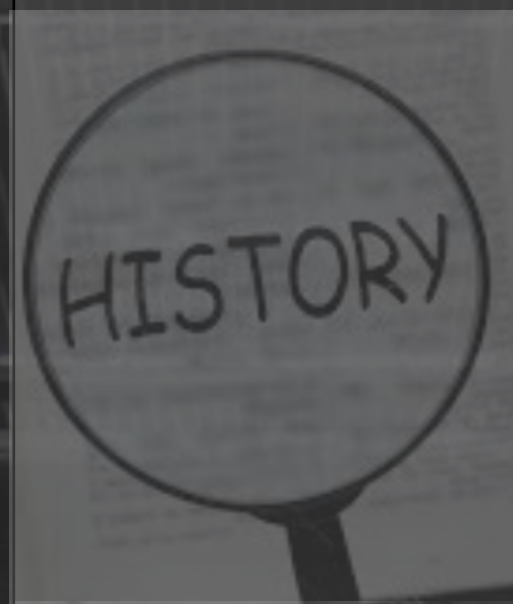
Single Instance
180 PB

IO Streams
>100k





How did we get there?





Introduction

A brief history about EOS architecture

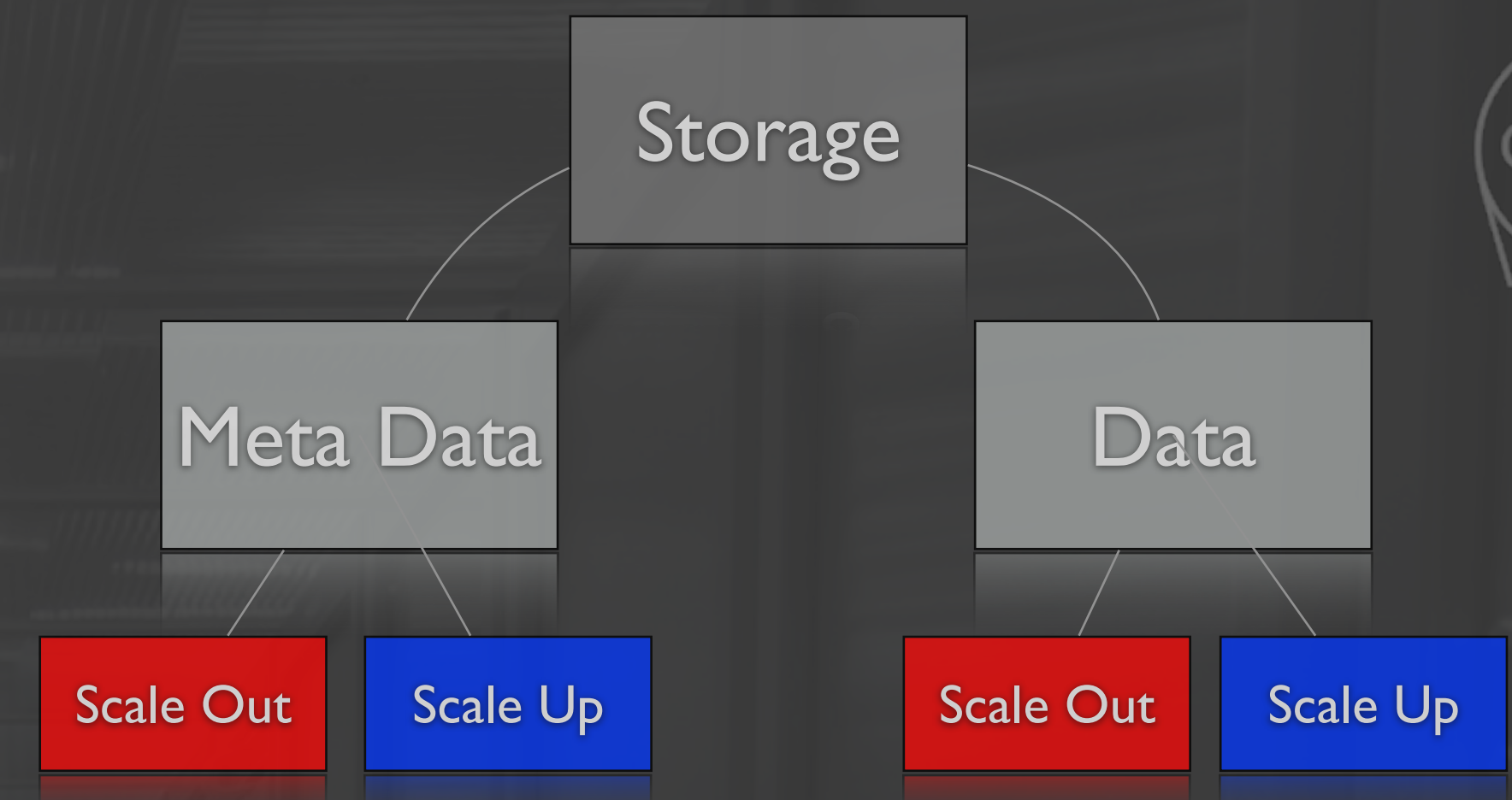
Scalability in Storage Systems



Scale Up = Bigger



Scale Out = More



- meta-data
 - hierarchically organised
 - small ~ TB
- data
 - non-hierarchically organised
 - large ~ PB-EB

In EOS we profit from both options!



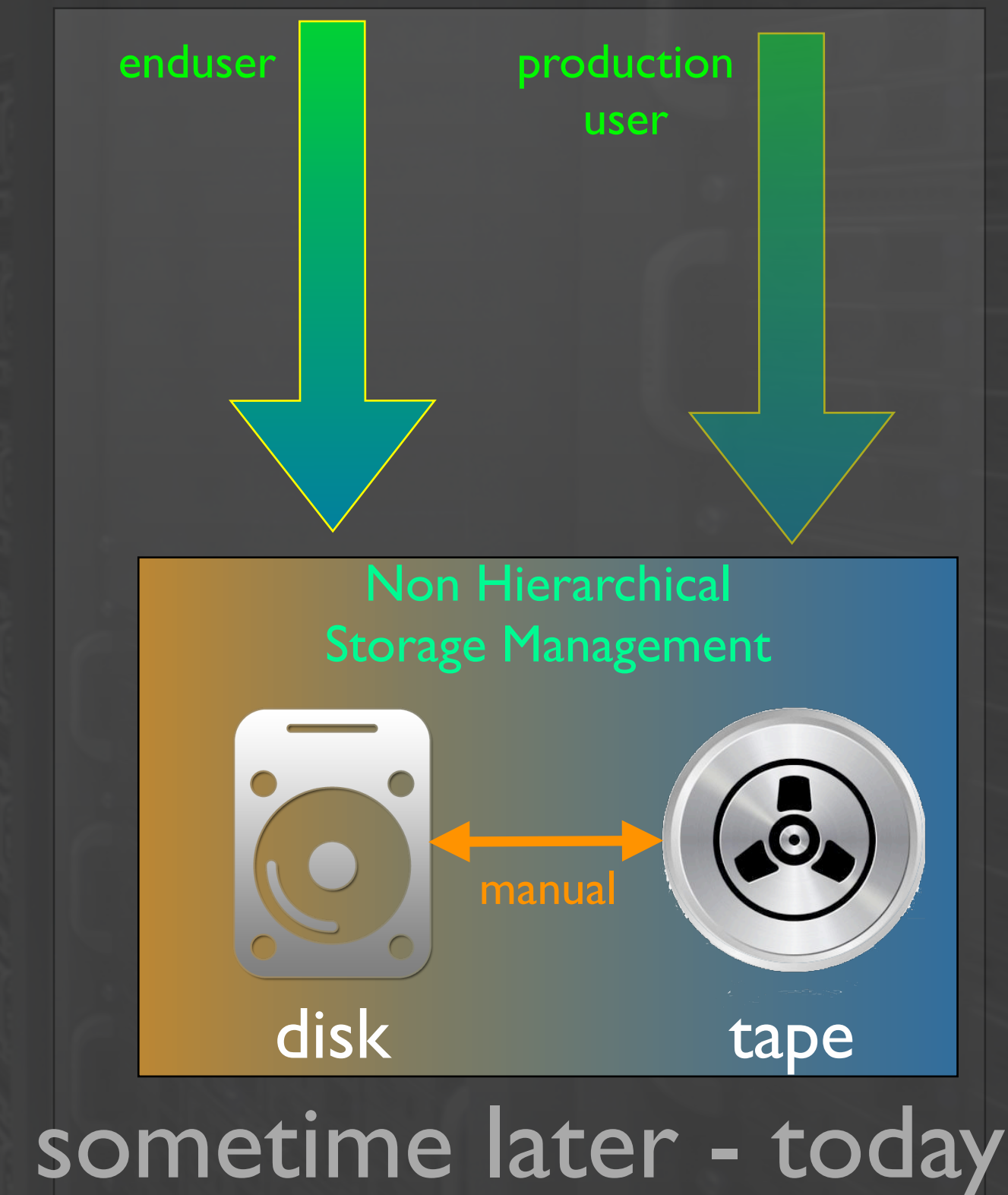
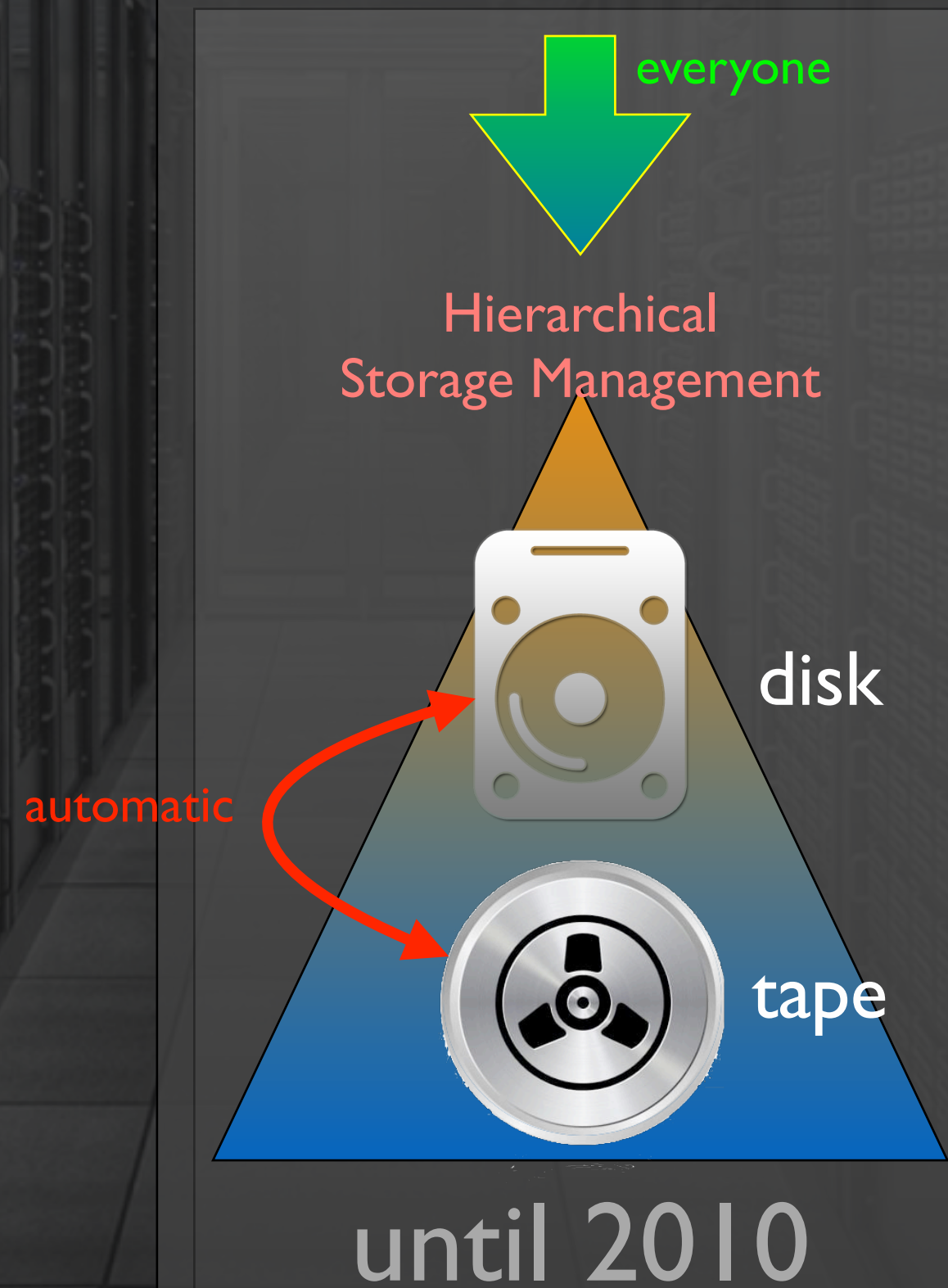
Introduction

A brief history about EOS architecture

Outsourcing Hierarchical Storage Management



- EOS was started in 2010 with the rationale to solve the analysis use-case for LHC physics
 - The **hierarchical storage** architecture used also for analysis was broken into two parts
 - Low-latency disk storage - **EOS**
 - High-latency tape storage - later **CTA**





What else did we need/want?

- an extremely **cost effective** storage system
 - minimal \$/TB - storage HW under 1CHF/month/TB with EC10,2
- a storage system allowing to **share efficiently resources** with thousands of users
 - SECURITY
 - QUOTAS
 - ACLs for sharing
 - QOS for meta-data and data
- **remote accessible** storage infrastructure
 - High Energy Physics computing model includes over 150 computing sites - originates from the funding model
 - efficient protocols for LAN & WAN
- a storage system suitable for **physics analysis use cases** and data formats
 - 100k Netflix movies watched at the same time and people might skip forward





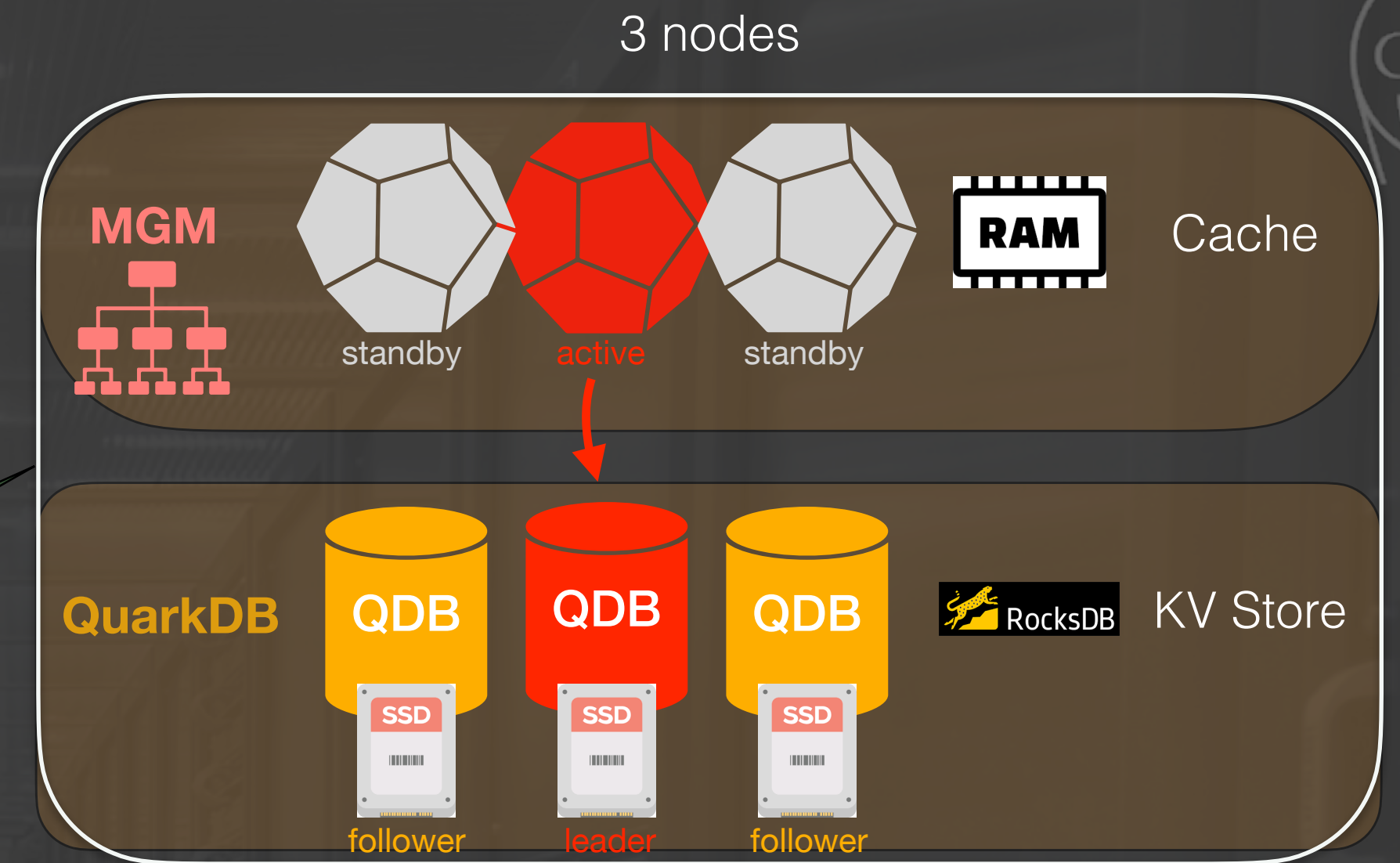
Service Architecture Today

High-available and low latency namespace

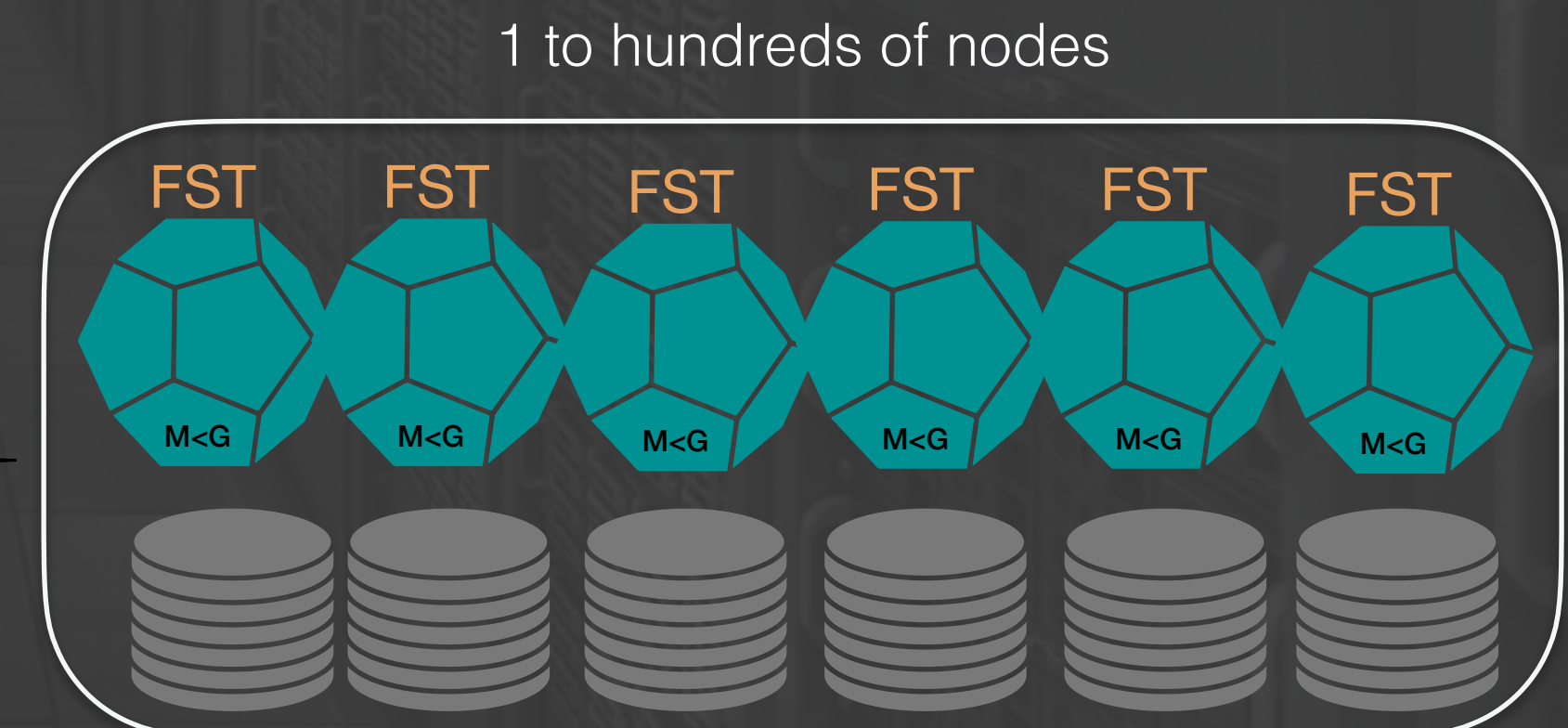
- namespace persisted on a key-value store
 - **QuarkDB** - REDIS protocol - developed at CERN
- used entries cached in-memory (LRU)

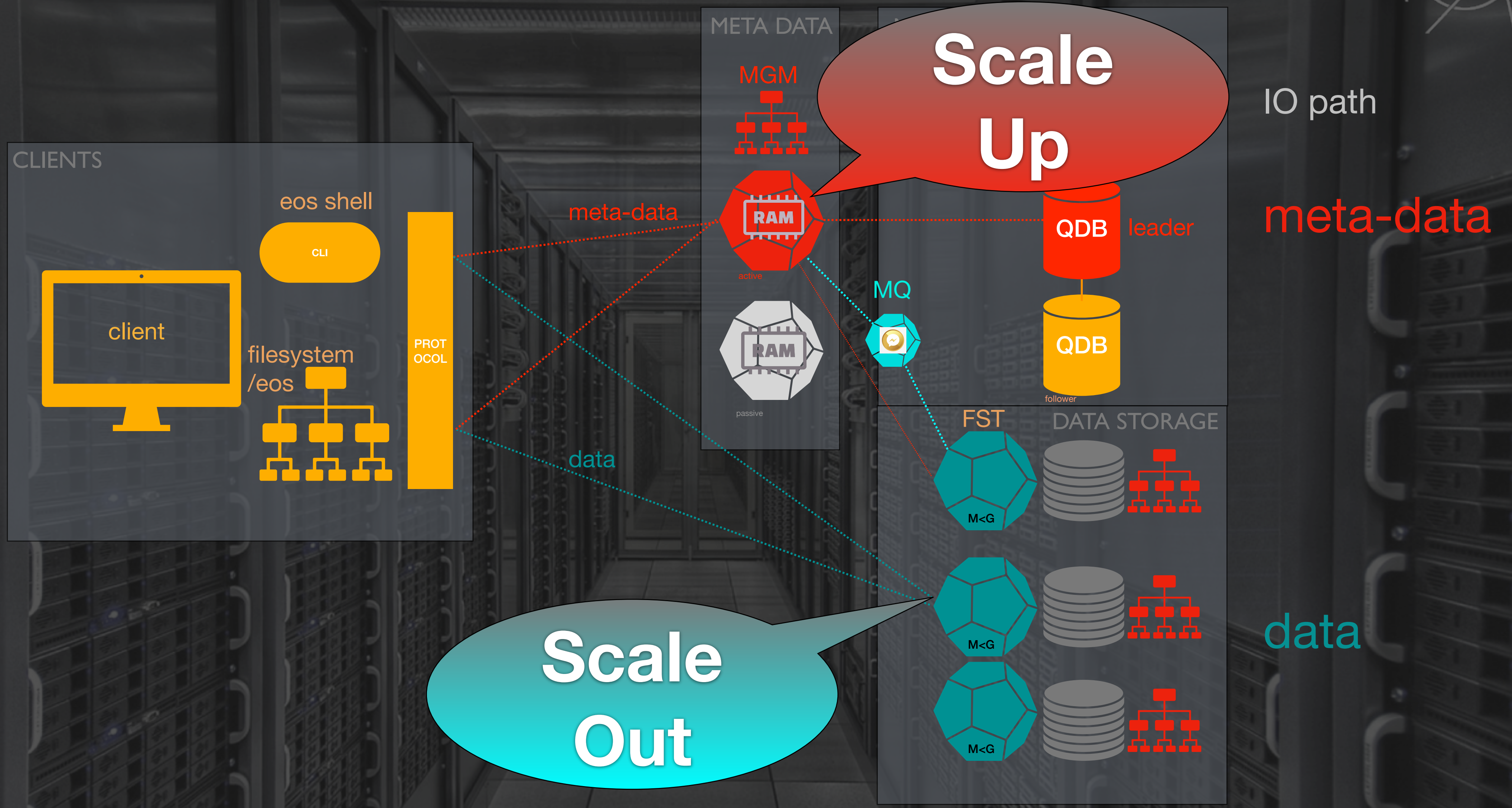
High available and reliable file storage, based on (cheap) JBODs and RAIN:

- File replication across independent nodes and disks
- Erasure coding to optimize costs and data durability



EOS is implemented by three daemons



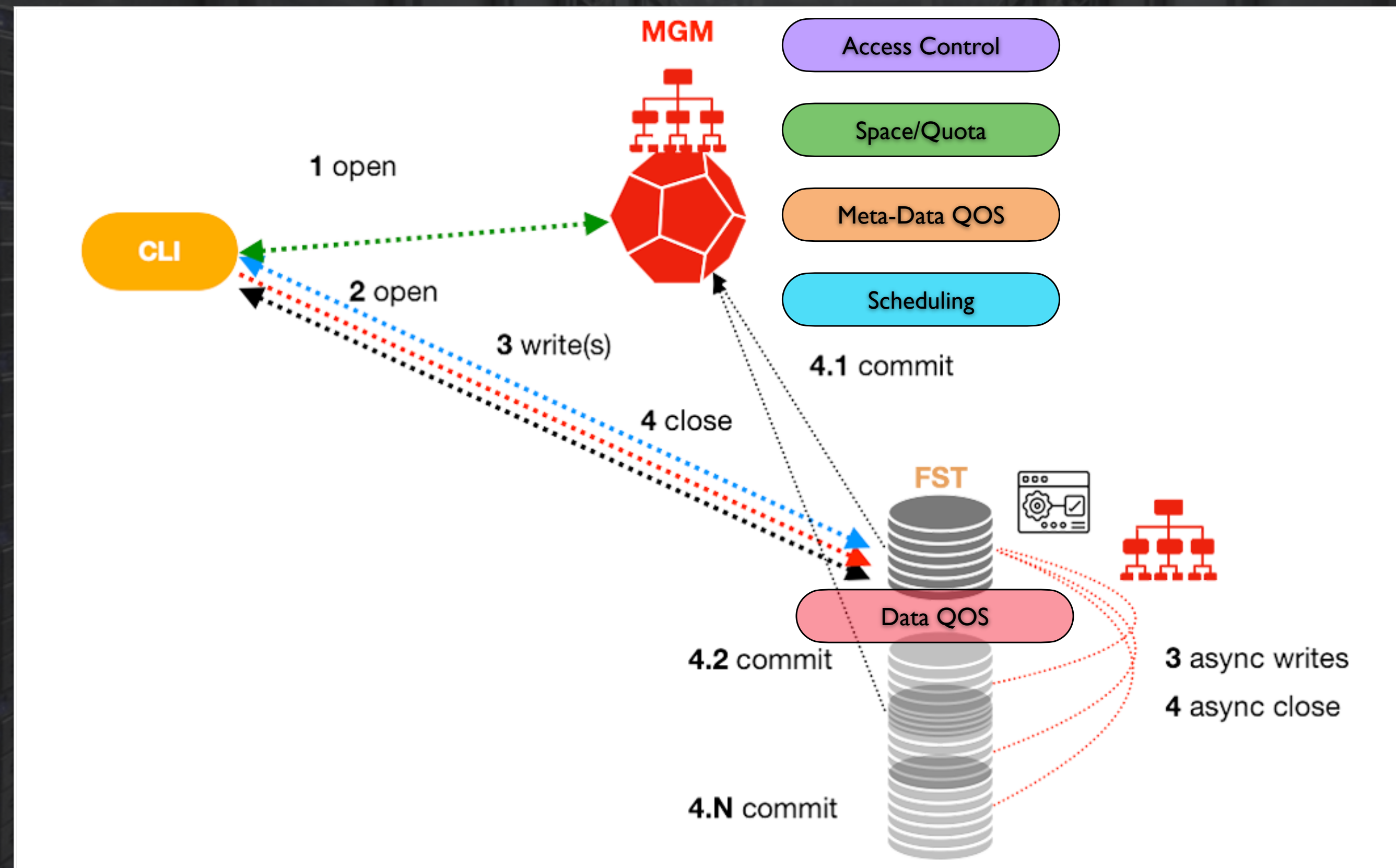


- framework
- XRootD
- components
- CLIENTs
- MGM
- MQ
- FST
- QuarkDB

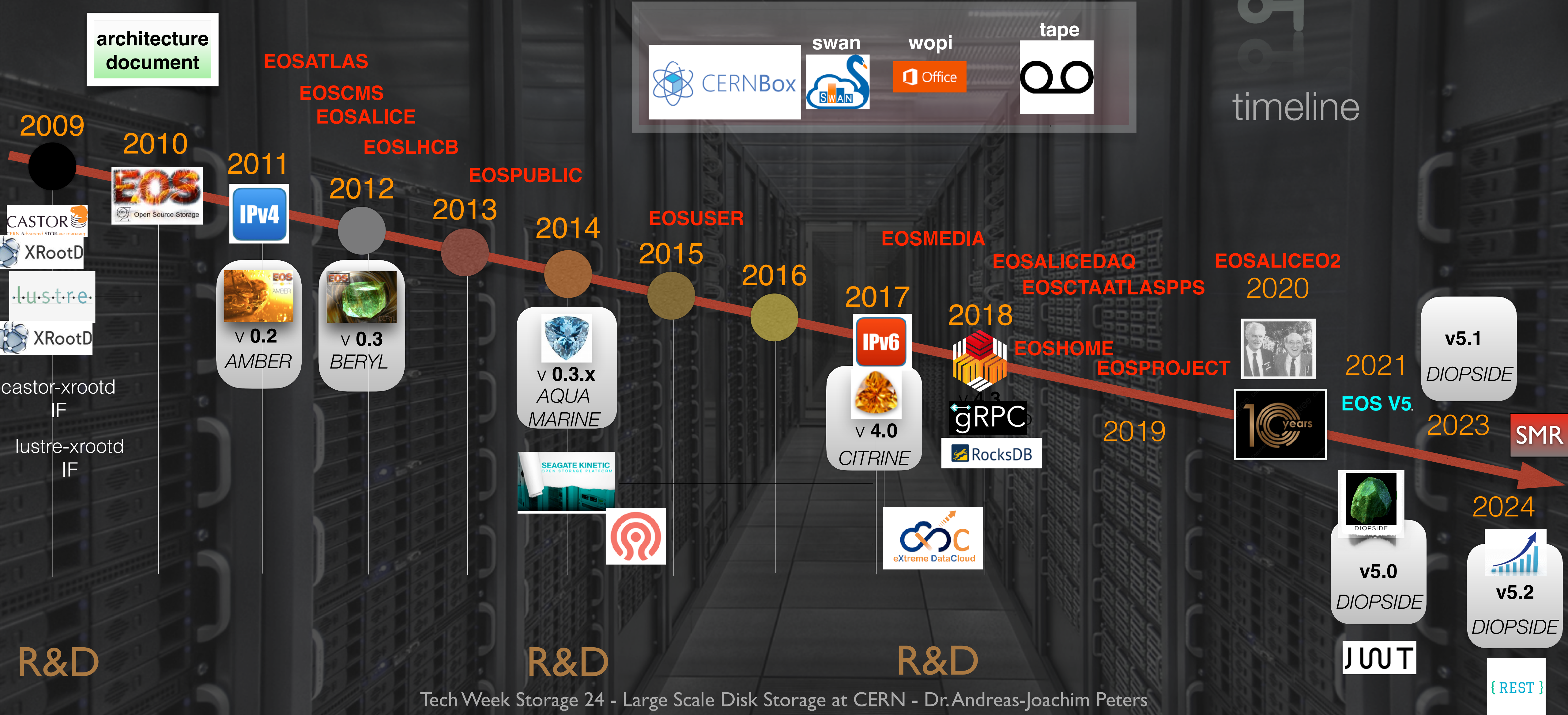
MGM meta-data server FST storage server MQ messaging server QuarkDB meta-data persistency

File Transaction Model

- EOS follows a file transaction model with server-side authorisation
 - Security is enforced always on server side, clients are not trusted



EOS Software Timeline



swan
 wopi
 tape

CERNBox
 Office

timeline

- EOS is written using the **XRootD** framework
 - something like curl, libcurl + NGINX in one framework written in C++ providing root(s):// and http(s) protocol
- **XRootD** protocol
 - provides POSIX-like API, request redirection + third party copy functionality
 - provides many authentication methods UNIX, KRB5, X509, JWT, shared secrets ...
- **XRootD HTTP(S)** plug-in
 - provides HTTP and HTTPS protocol with X509 authentication and JWT auth/authz
 - provides a third-party copy implementation based on COPY verb (not standardised)

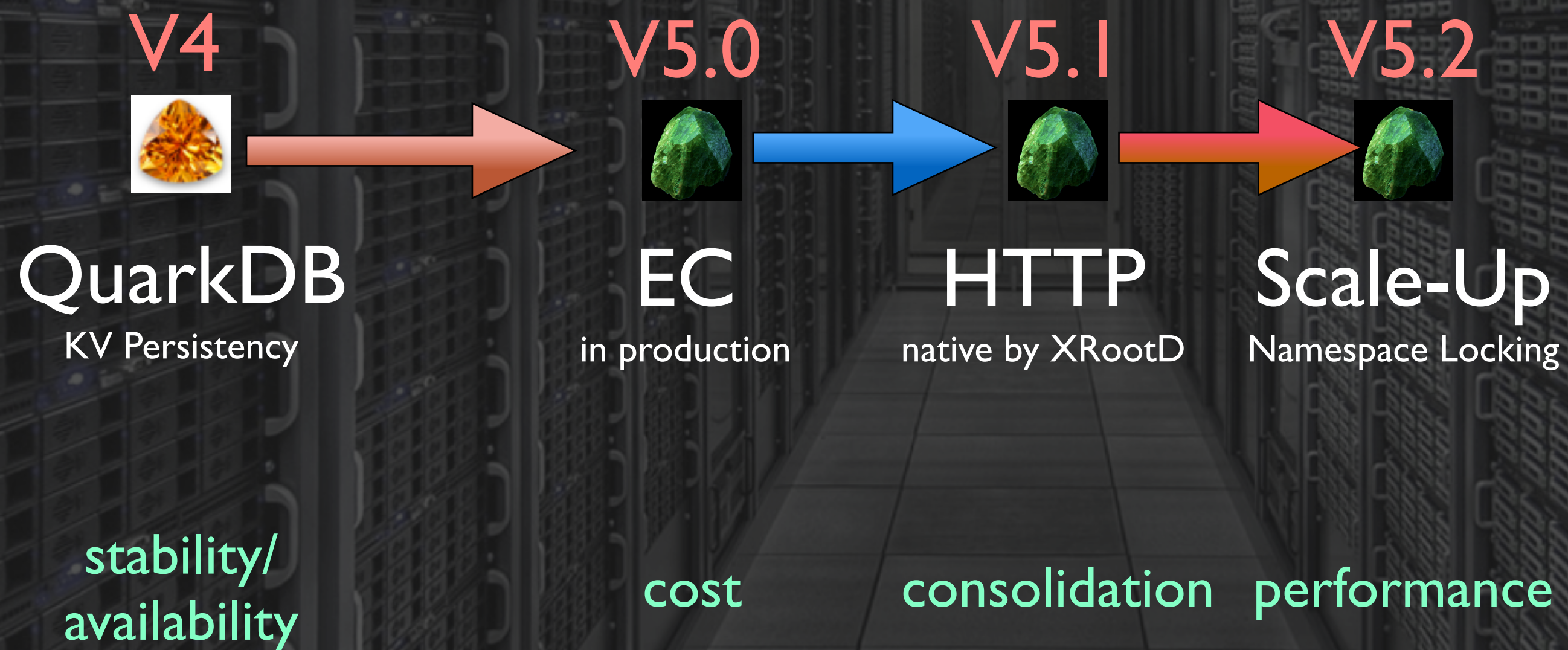


C/C++ Lines of Code (cloc)



- EOS source code has contributions from 42 authors
- 10 active contributors during the last month
- Mainstream platforms: **CentOS7, ALMA8, ALMA9**
- Almost weekly testing releases - **agile** release procedure
 - CI pipelines with 1034 system tests, many more unit/component tests

Version Highlights





R&D Activities

Current & Upcoming R&D Activities

- Shingled Magnetic Recording SMR

- basic support has been added
- waiting for larger testbed



- HAMR

- will work out of the box
- currently no hardware available

- Low-cost flash storage

- OpenLab collaboration in preparation



- ARM platform

- Evaluation of storage servers build on ARM architecture



100+ PB

Bare Metal Container

1 PB

1++
Disk
Sever

10 PB

10++
Disk
Sever

**Production
Setups**

100 ++
Disk
Sever

Favoured Redundancy

RAID, Replication,
Erasure Coding - maybe

RAID, Replication, Erasure Coding

Replication, Erasure Coding

Virtualized/
Kubernetes

VM

VM
Cluster

Kubernetes
Cluster

Redundancy

Shared Filesystem,
Block Device

Shared Filesystem,
Block Devices

Shared Filesystem,
Block Devices

many



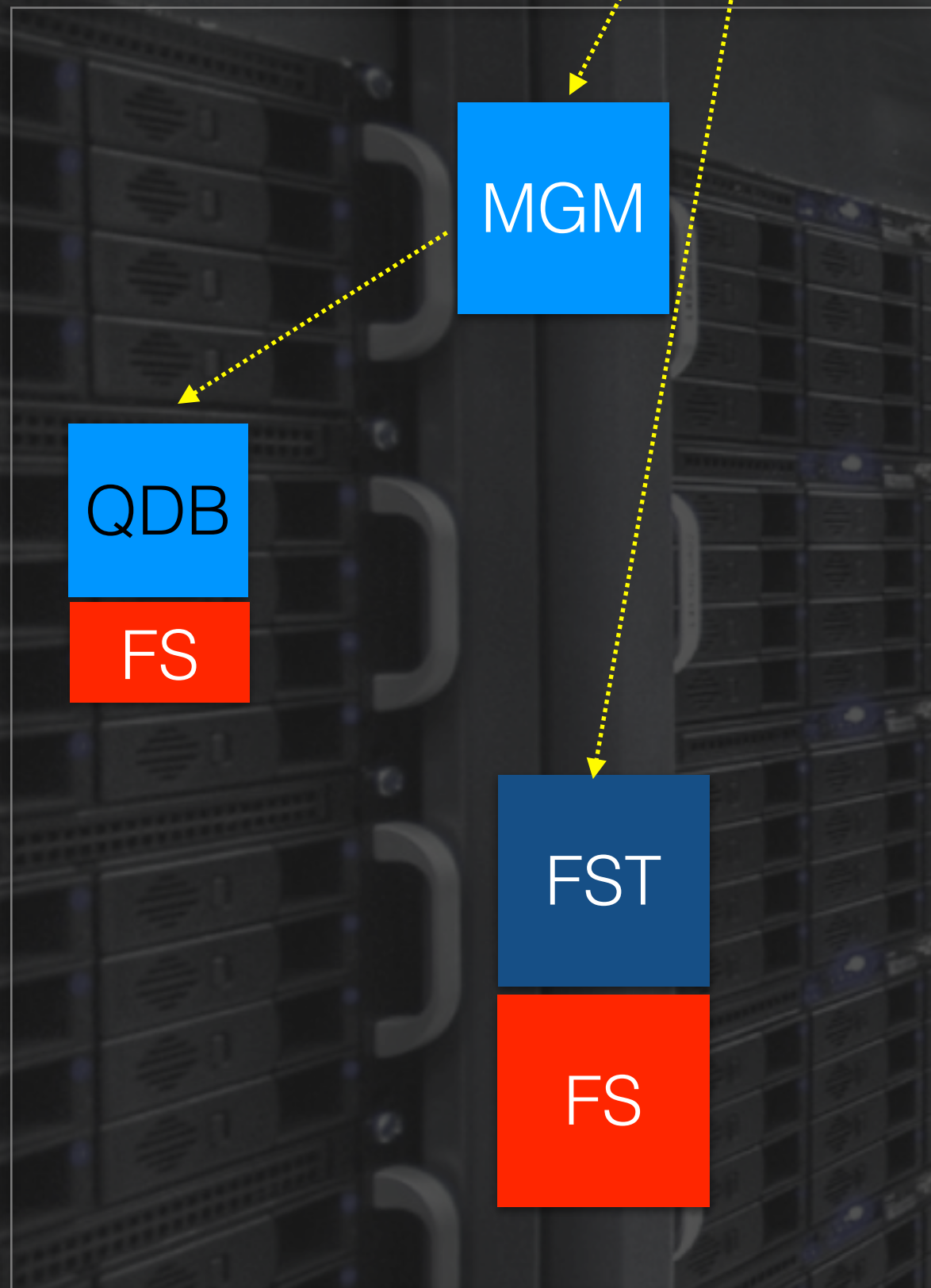
few

EOS Deployment

Single Node Deployment



node



All daemon in one physical box:
QDB, MGM, FST

Hardware Requirements:

QDB: **SSD/NVMe**
0.1-0.2 GB/Million Entries

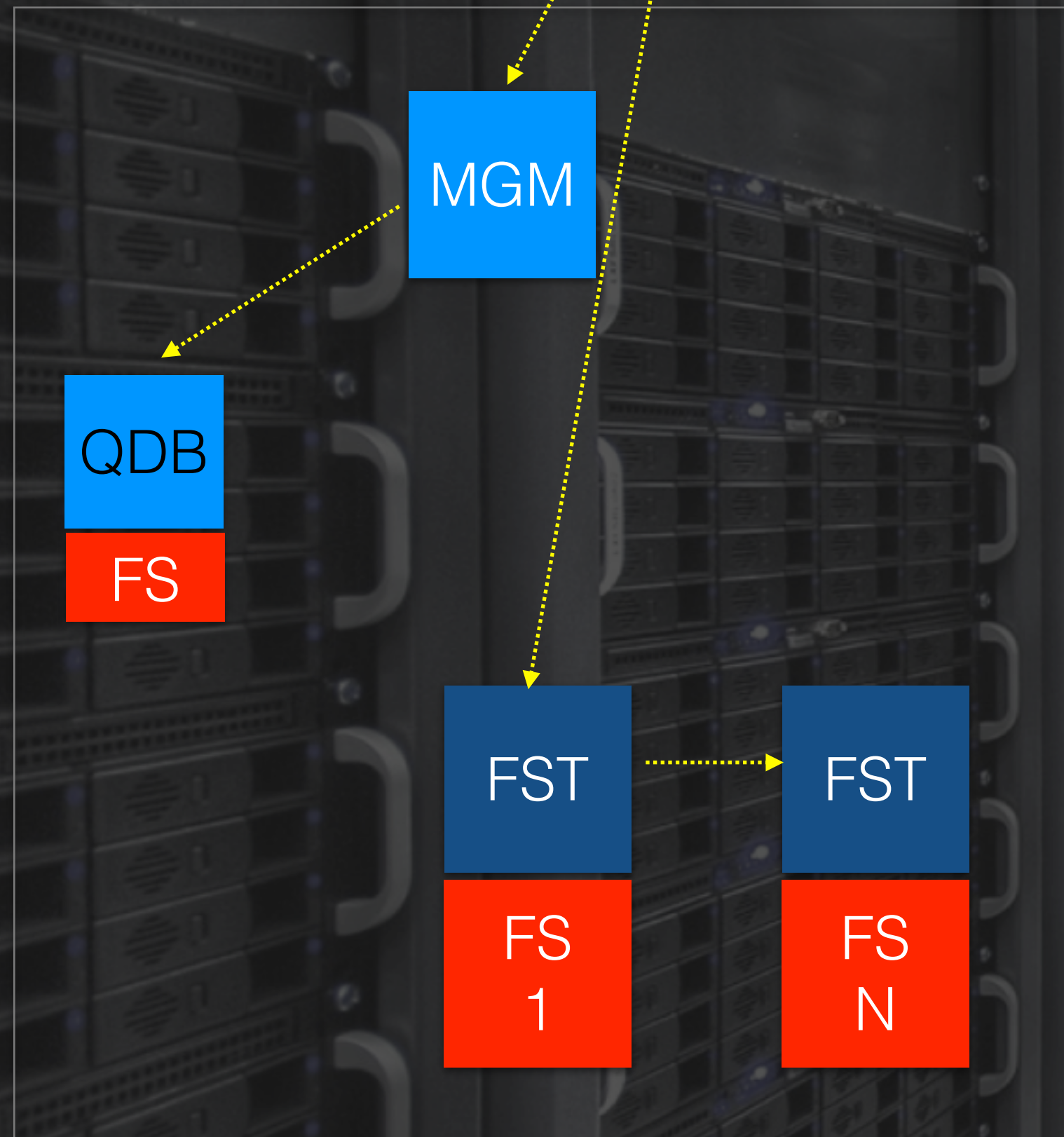
MGM:
4 core - min. 8 GB

FST:
4 core - min. 8 GB
1 GB RAM / HDD
HDD FS: XFS+XAttr

EOS Deployment

Single Node Deployment

node



All daemon in one physical box:
QDB, MGM, FST1-N

Hardware Requirements:

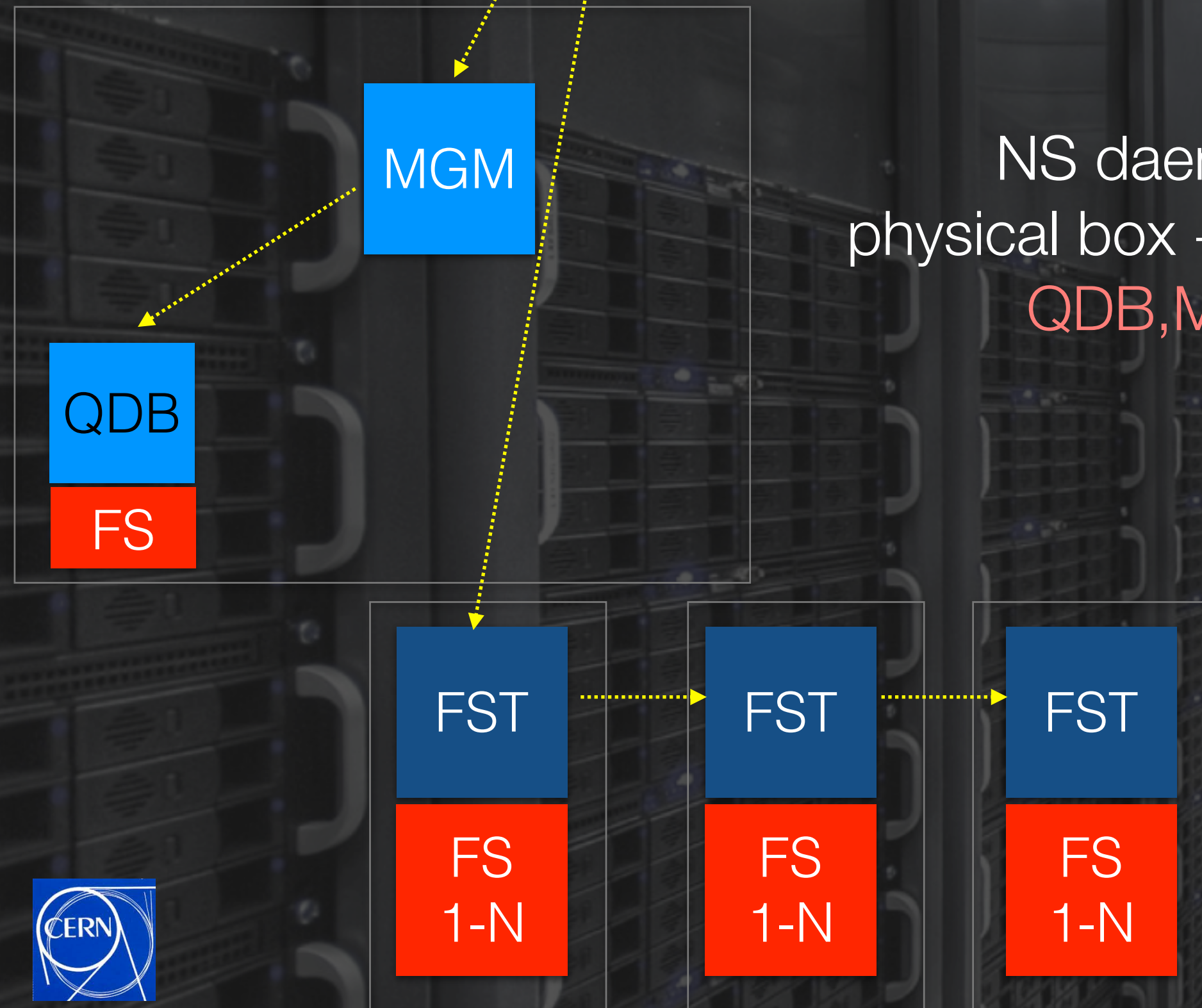
QDB: **SSD/NVMe**
0.1-0.2 GB/Million Entries

MGM:
4 core - min. 8 GB

FST:
4 core - min. 8 GB
1 GB RAM / HDD
HDD FS: XFS+XAttr



node



NS daemon in one
physical box + N FST Nodes:
QDB, MGM, FST

Hardware Requirements:

QDB: **SSD/NVMe**
0.1-0.2 GB/Million Entries

MGM:
4 core - min. 8 GB

FST:
4 core - min. 8 GB
1 GB RAM / HDD
HDD FS: XFS+XAttr

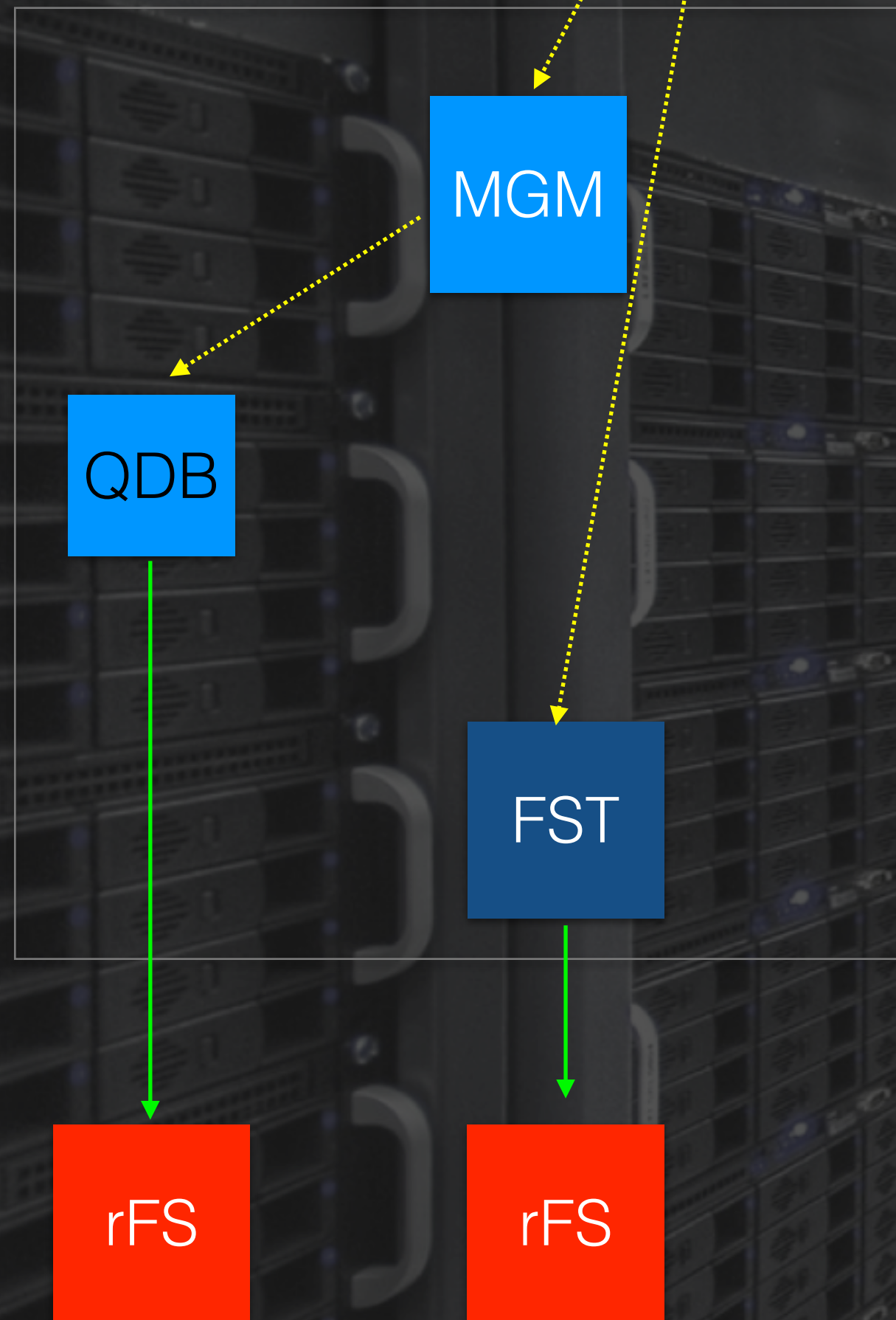


EOS Deployment

Virtual Single Node Deployment



node



All daemon in one virtual box:
QDB, MGM, FST

Hardware Requirements:

QDB: **HIGH IOPS Virtual Disk**
0.1-0.2 GB/Million Entries

MGM:
4 core - min. 8 GB

FST:
4 core - min. 8 GB
1 GB RAM / HDD

HDD FS: remote FS with XAttr Support

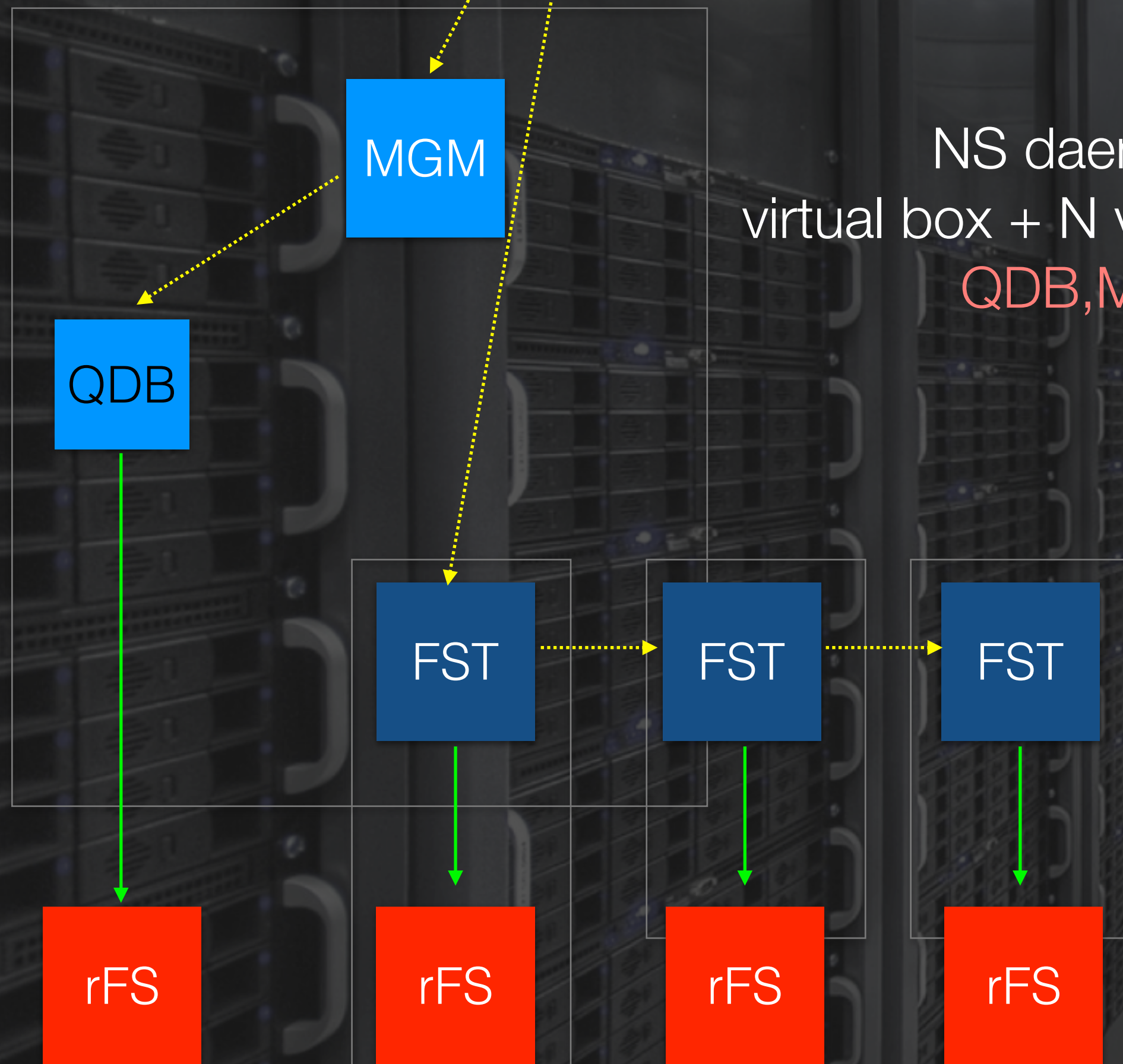
High IOPS

High BW

remote



node



NS daemon in one virtual box + N virtual FST Nodes:
QDB, MGM, FST

Hardware Requirements:

**QDB: HIGH IOPS Virtual Disk
0.1-0.2 GB/Million Entries**

**MGM:
4 core - min. 8 GB**

**FST:
4 core - min. 8 GB
1 GB RAM / HDD
HDD FS: remote FS with XAttr Support**

High IOPS

High BW

remote

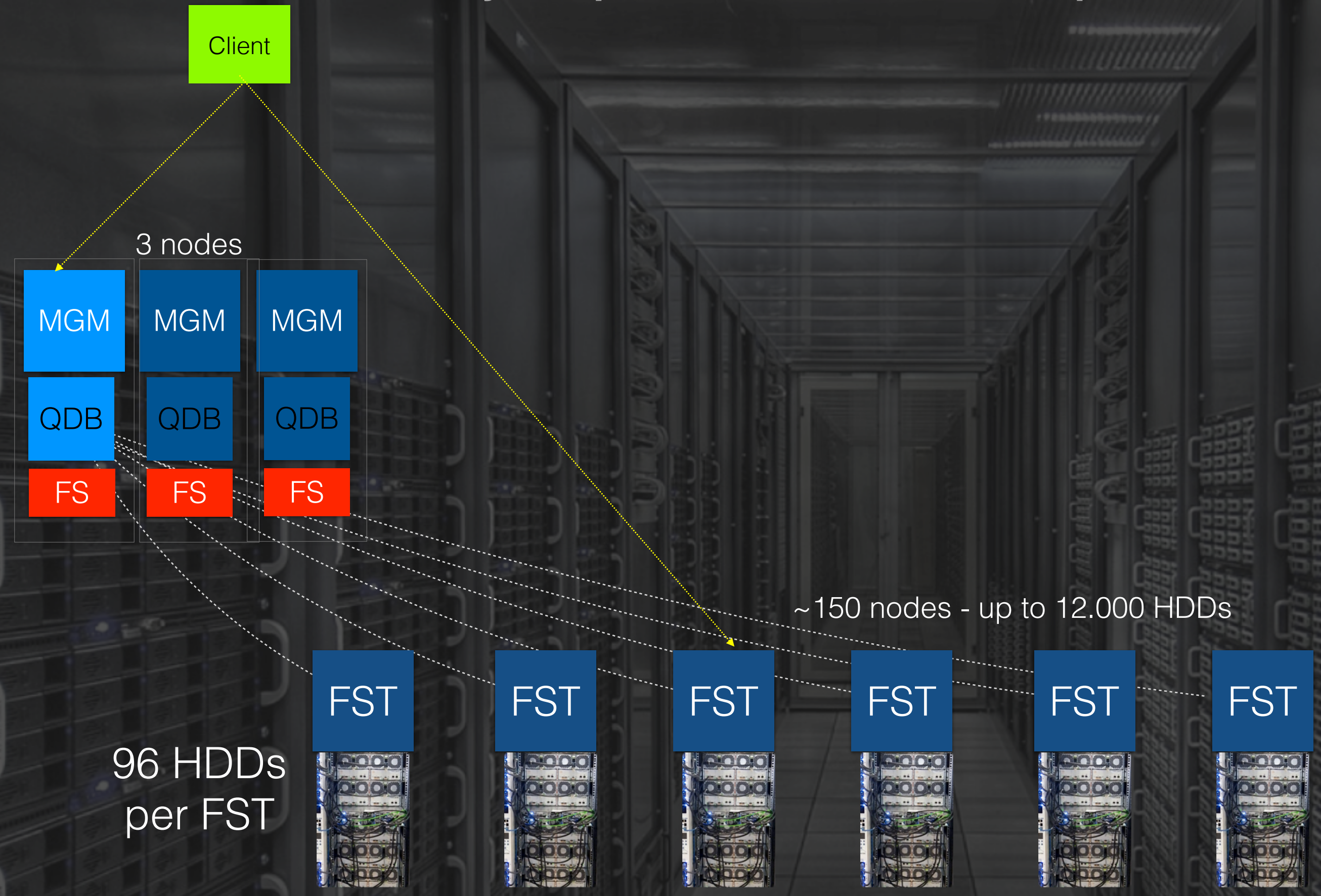
EOS Deployment

Physics production instance setup

Meta-Data Service

KV Store

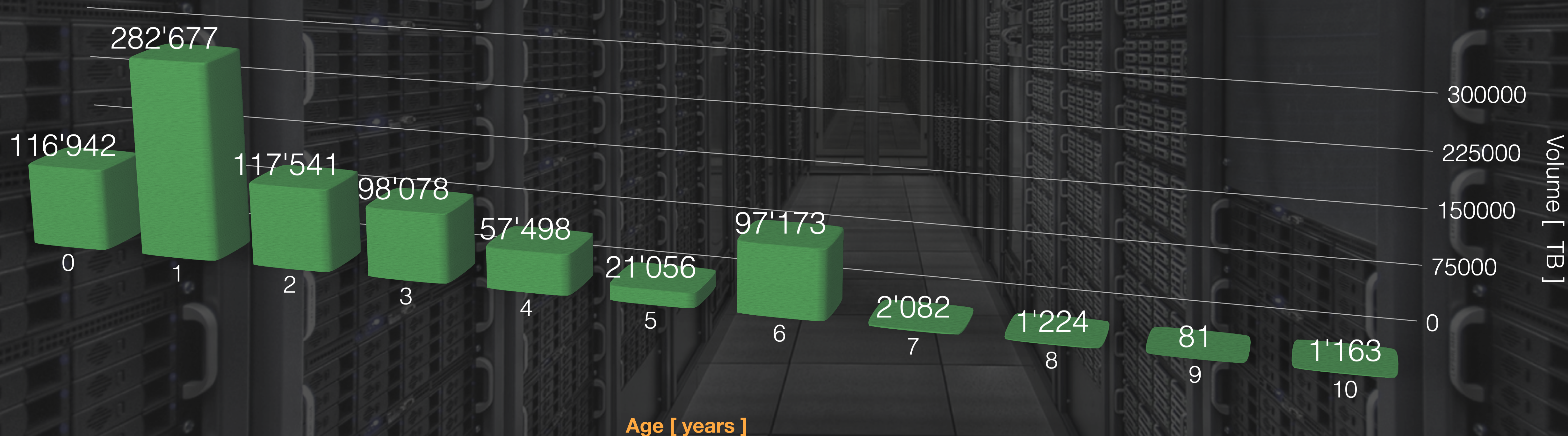
Data Store



- EOS allows to **mix many hardware generations** within an instance
- EOS service oldest HDDs have **10.8** years, average age in 2023 **3.8** years,
- annual failure rate was **1% in 2022**

Example: 1TB HDD space costs on the consumer market 15 CHF - assume 5y lifetime 3 CHF/year

Volume vs Disk Runtime [2023]



Instance Scale-out

- running over 24 instances at CERN with independent namespaces
 - gives you 24x meta-data performance to compensate scale-up architecture of meta-data service



/eos/

/eos/atlas

/eos/cms

/eos/lhcb

/eos/user

/eos/project

/eos/media



eosatlas



eoscms



eoslhcb



eosuser



eosproject



eosmedia

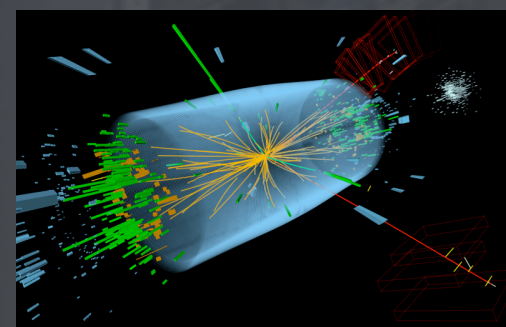


eoshome-i00,01,02,03,04



eosproject-i00,01,02

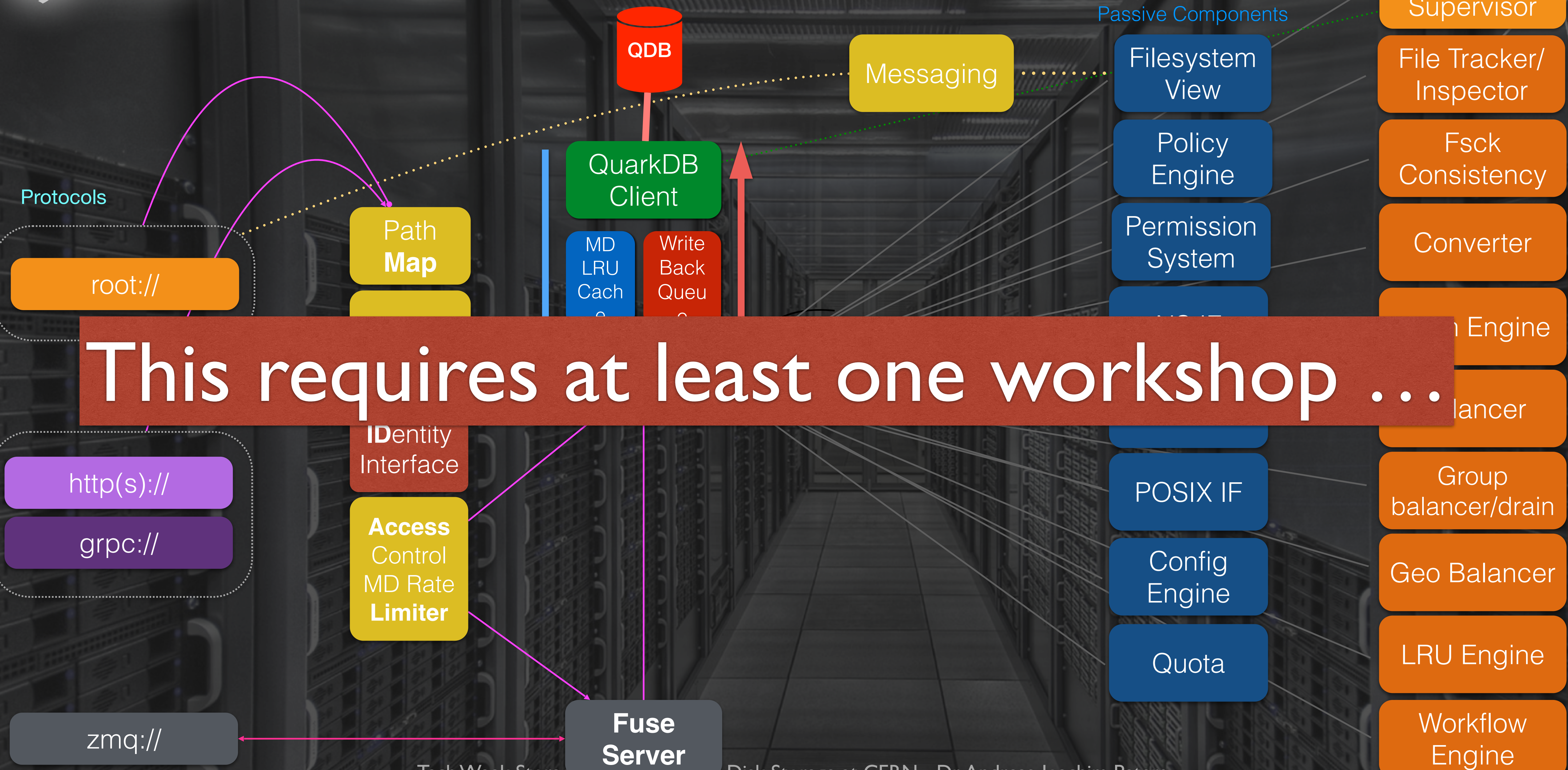
EOS for Physics



Namespace Server Components

Active Components

Passive Components





Fundamental Concepts

File Storage & File Layouts

EOS is a **file storage system**, not block storage!

EOS supports **Replication** and **Erasure Coding** layouts with 1,2,3,4 parities with up to 255 stripes.

```
EOS Console [root://localhost] |eos/aliceo2/raw/2024/LHC24aa_ZDC/547905/raw/1450/> file info o2_ctf_run00547905_orbit000000288_tf000000001_epr
t
File: '/eos/aliceo2/raw/2024/LHC24aa_ZDC/547905/raw/1450/o2_ctf_run00547905_orbit000000288_tf000000001_epr308.root'  Flags: 0644
Size: 602172639
Status: healthy
Modify: Wed Mar  6 14:52:03 2024 Timestamp: 1709733123.336500000
Change: Wed Mar  6 14:52:00 2024 Timestamp: 1709733120.834302209
Access: Wed Mar  6 14:52:00 2024 Timestamp: 1709733120.834302701
Birth: Wed Mar  6 14:52:00 2024 Timestamp: 1709733120.834302209
Cuid: 13798 CGid: 1395 Fxid: 23897e70 Fid: 596213360 Pid: 671828 Pxid: 000a4054
XStype: adler  XS: c2 88 7f c0  ETAGs: "160044805164892160:c2887fc0"
Layout: raid6 Stripes: 12 Blocksize: 1M LayoutId: 20640b42 Redundancy: d3::t0
#Rep: 12
```

no.	fs-id	host	schedgroup	path	boot	configstatus	drain	active	geotag
0	12235	st-096-100gb022.cern.ch	erasure.28	/data41	booted	rw	nodrain	online	9a417ecb
1	10361	st-096-o2-192a7b.cern.ch	erasure.28	/data10	booted	rw	nodrain	online	17db40c5
2	11947	st-096-100gb019.cern.ch	erasure.28	/data43	booted	rw	nodrain	online	6d750011
3	11562	st-096-100gb015.cern.ch	erasure.28	/data04	booted	rw	nodrain	online	29e8ee40
4	11370	st-096-100gb013.cern.ch	erasure.28	/data49	booted	rw	nodrain	online	7e4083e9
5	2078	st-096-100gb002.cern.ch	erasure.28	/data41	booted	rw	nodrain	online	e86ebe02
6	12043	st-096-100gb020.cern.ch	erasure.28	/data32	booted	rw	nodrain	online	89ede750
7	2655	st-096-100gb008.cern.ch	erasure.28	/data10	booted	rw	nodrain	online	17dc5b9f
8	11275	st-096-100gb012.cern.ch	erasure.28	/data55	booted	rw	nodrain	online	739633a9
9	1966	st-096-100gb001.cern.ch	erasure.28	/data22	booted	rw	nodrain	online	127e4662
10	2269	st-096-100gb004.cern.ch	erasure.28	/data05	booted	rw	nodrain	online	f9fbf9f1
11	10073	st-096-o2-191da7.cern.ch	erasure.28	/data06	booted	rw	nodrain	online	9a3e665c

EOS supports various **file checksum algorithms** **ADLER32, MD5, CRC32C, BLAKE, SHA ...**

Every file has a checksum after CLOSE if configured.

EOS erasure coding uses **4k block checksumming** to identify data corruption.



Fundamental Concepts

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

Access Control



EOS has an **access control interface** to allow/ban users, groups, nodes, domains after connection

EOS provides very rich **ACL language** to grant permissions on a directory bases

- ACLs **similar to NFS4 ACLs**
- ACLs are defined by POSIX user/group or E-GROUP expressing **GRANT & DENY**
- CERN provides E-GROUP interface to create **dynamic groups of people** on a web page - these groups can be referenced in EOS ACLs

```
$ eos attr ls /eos/mypath  
sys.acl="u:99999:rw,egroup:mygroup:rw"  
#
```




Fundamental Concepts

Policies



- **Placement Policies**
 - File Layouts e.g two replica or erasure coding
 - Physical hardware to use
 - Geographic location for replicas
- **Checksumming Policies**
 - File Checksum Algorithm e.g. `adler32`, `md5`
 - Block Checksum Algorithm e.g. `crc32c`
- **Conversion/Cleanup Policies**
 - **Automatically convert files** e.g. move all files bigger than 1GB from SSD to HDDs
 - **Cleanup** all log files after 1 month
 - **Cleanup** empty directories after 6 month
- **IO Policies / Data QOS**
 - **Read/Write** via **BC** or direct **IO**
 - **Limit the bandwidth** of a streams to max **N MB/s** by user, group or application
 - **Use local LINUX IO scheduling priorities** for certain use-cases e.g. `REALTIME0` for highest priority



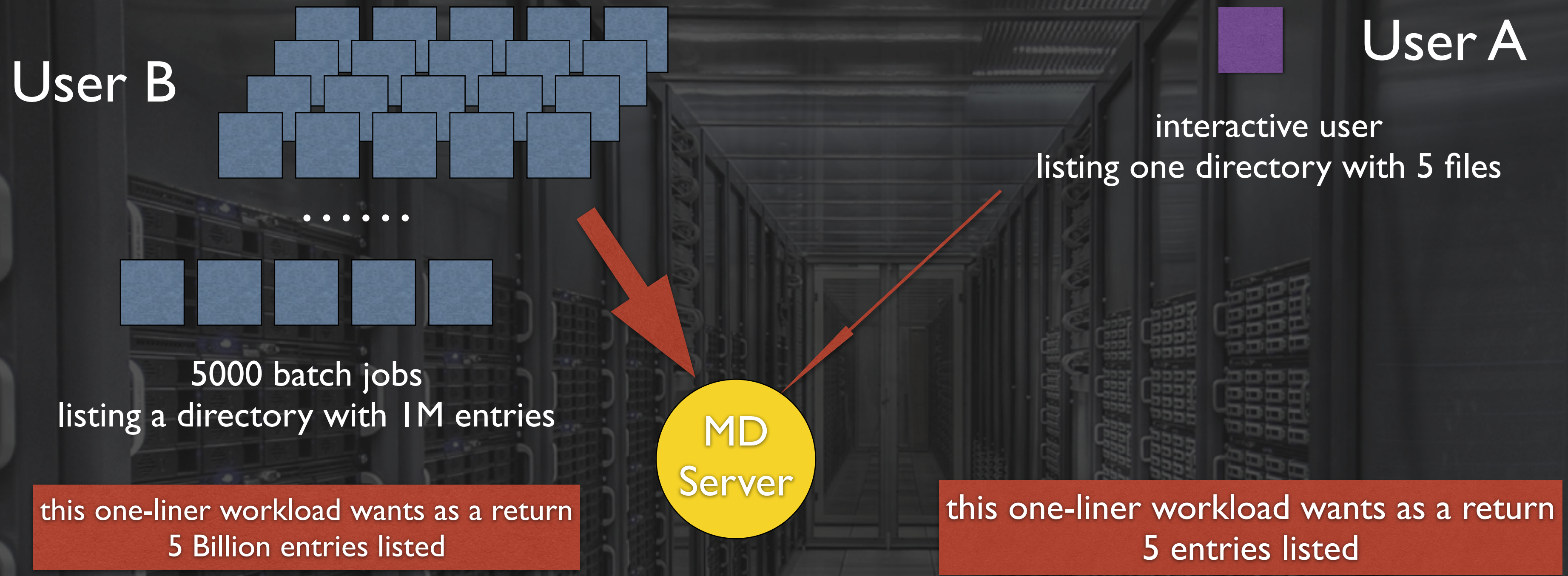
Fundamental Concepts

Quality of Service Meta-Data QOS



- Meta-Data operation throttling
 - max file open at N Hz by user, group
 - max listing with M Hz entries by user, group
- User Thread-pool limits
 - allow a single user to use maximum N threads in the meta-data service
- Meta-Data operation hard limits
 - stall users when throttling and thread-pool limits are hit
- Global Thread-pool limit
 - don't run more than N threads for all users

an (almost) real case



We stall or delay requests
of B making room for A to be served



Quota

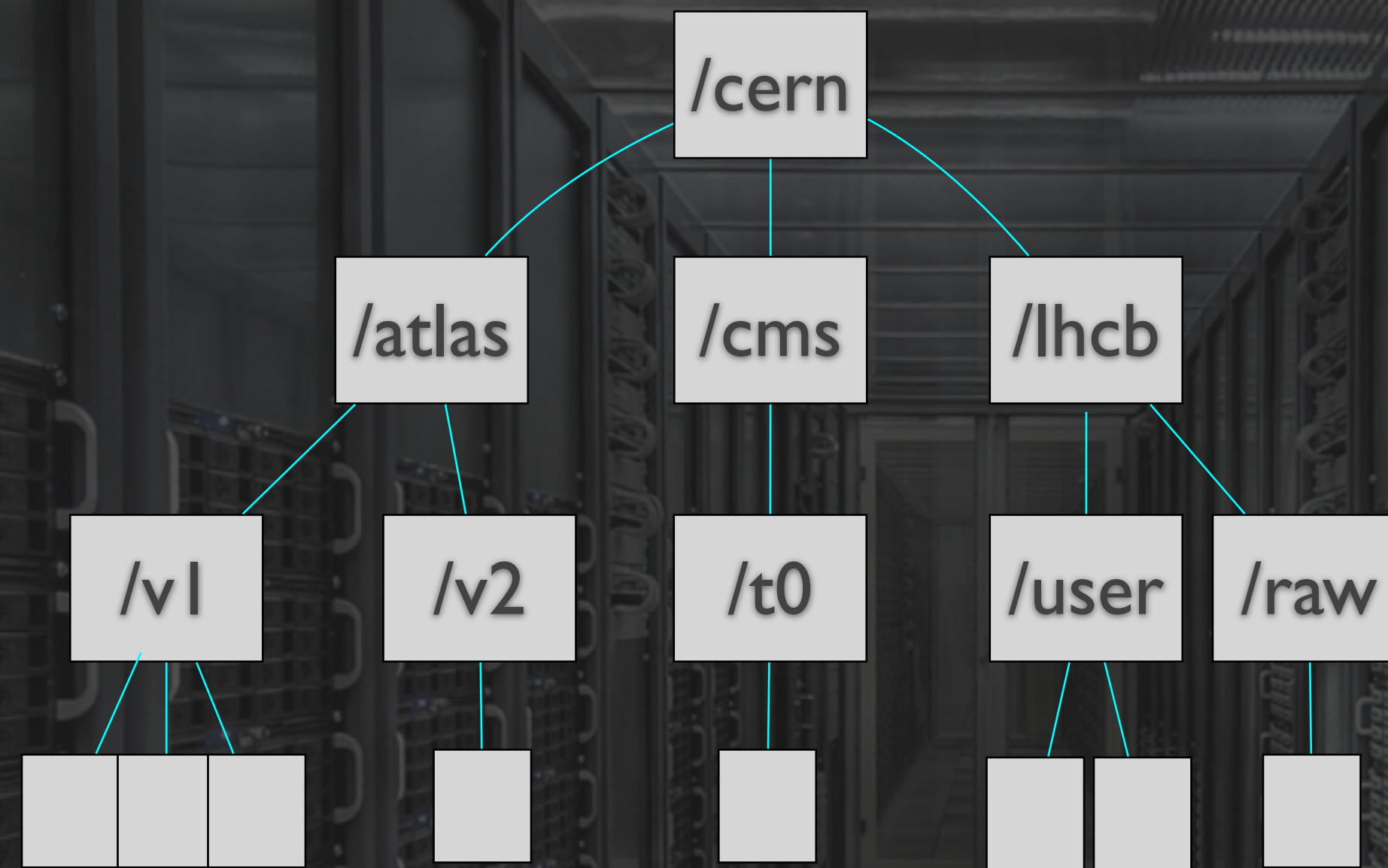
Quota Listing in EOS EOS supports user, group & project quota

```
➔ Quota Node: /eos/ams/user/
```

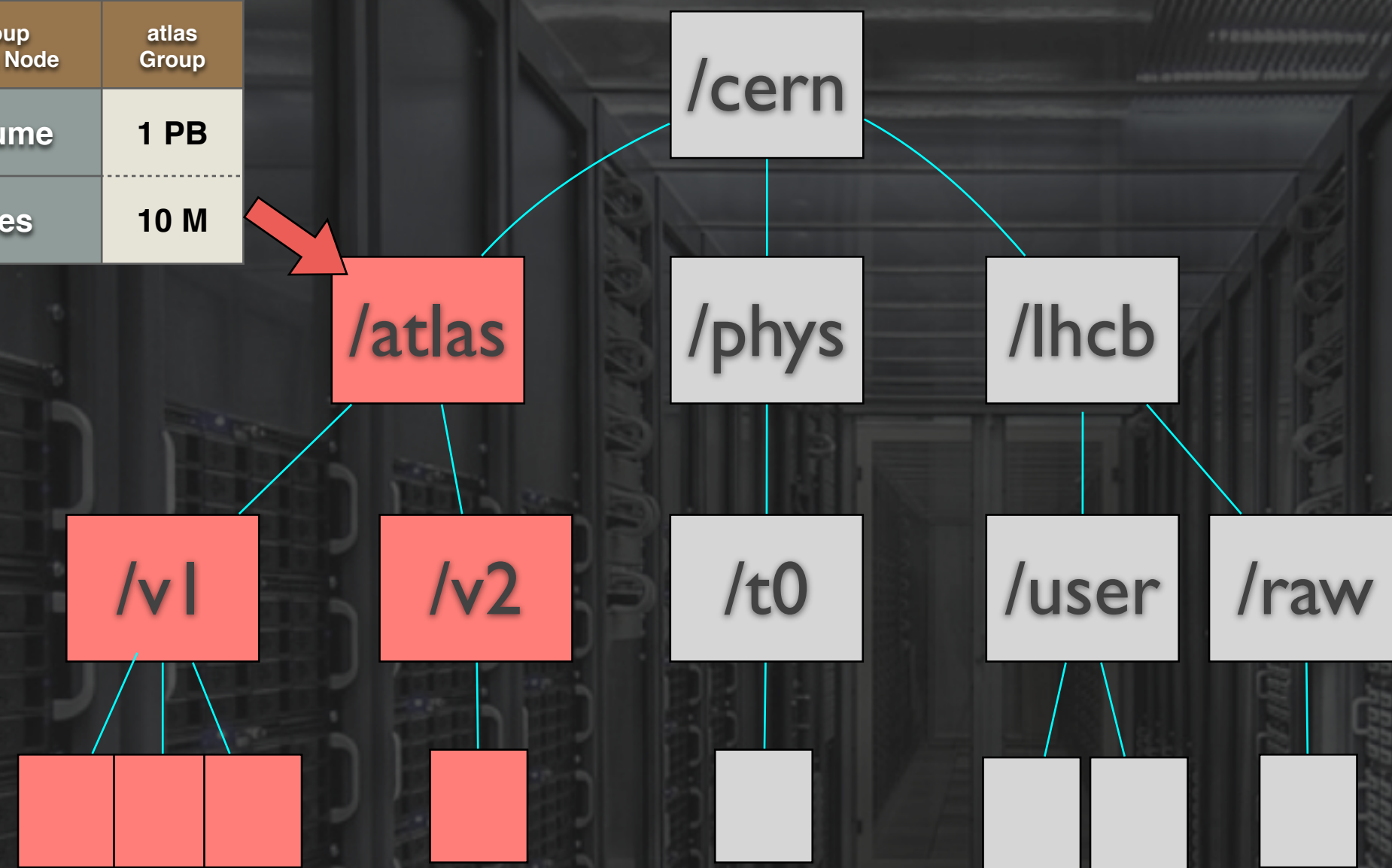
user	used bytes	logi bytes	used files	aval bytes	aval logib	aval files	filled[%]	vol-status	ino-status
103634	79.98 GB	39.99 GB	7.14 K	4.00 TB	2.00 TB	300.00 K	2.00 %	ok	ok
107153	0 B	0 B	0	2.00 TB	1.00 TB	100.00 K	0.00 %	ok	ok
118088	388.48 GB	194.24 GB	20.91 K	1.00 TB	500.00 GB	100.00 K	38.85 %	ok	ok
120935	605.13 GB	302.57 GB	78.25 K	2.00 TB	1.00 TB	100.00 K	30.26 %	ok	ok
125271	0 B	0 B	0	15.00 TB	7.50 TB	1.00 M	0.00 %	ok	ok
21716	1000.00 GB	500.00 GB	8.27 K	1.00 TB	500.00 GB	100.00 K	100.00 %	exceeded	ok
22146	0 B	0 B	0	2.00 TB	1.00 TB	200.00 K	0.00 %	ok	ok
24027	0 B	0 B	0	1.00 TB	500.00 GB	100.00 K	0.00 %	ok	ok
27043	359.73 GB	179.87 GB	26.29 K	2.00 TB	1.00 TB	200.00 K	17.99 %	ok	ok
29911	888.97 MB	444.49 MB	1	2.00 TB	1.00 TB	200.00 K	0.04 %	ok	ok
30141	823.32 GB	411.66 GB	116.08 K	2.00 TB	1.00 TB	200.00 K	41.17 %	ok	ok
33601	17.00 GB	8.50 GB	4	1.00 TB	500.00 GB	100.00 K	1.70 %	ok	ok
3482	1.07 TB	532.82 GB	298.42 K	2.00 TB	1.00 TB	300.00 K	53.28 %	ok	exceeded
34872	1.28 TB	642.40 GB	16.55 K	2.00 TB	1.00 TB	100.00 K	64.24 %	ok	ok
3500	1000.00 GB	500.00 GB	4.61 K	2.00 TB	1.00 TB	200.00 K	50.00 %	ok	ok
35224	169.55 GB	84.77 GB	12.15 K	2.00 TB	1.00 TB	200.00 K	8.48 %	ok	ok
35324	10.07 TB	5.03 TB	78.69 K	12.00 TB	6.00 TB	1.00 M	83.92 %	ok	ok
39256	39.58 TB	19.79 TB	111.03 K	50.00 TB	25.00 TB	800.00 K	79.16 %	ok	ok
40105	1.31 TB	653.58 GB	4.95 K	4.00 TB	2.00 TB	400.00 K	32.68 %	ok	ok
41198	1.21 TB	606.41 GB	768	4.00 TB	2.00 TB	300.00 K	30.32 %	ok	ok
41244	319.39 GB	159.69 GB	56.47 K	2.00 TB	1.00 TB	200.00 K	15.97 %	ok	ok
43632	833.29 MB	416.64 MB	1	20.00 TB	10.00 TB	0	0.00 %	ok	ignored
44056	2.44 GB	1.22 GB	33	2.00 TB	1.00 TB	100.00 K	0.12 %	ok	ok
4459	0 B	0 B	0	1.00 TB	500.00 GB	100.00 K	0.00 %	ok	ok
45870	0 B	0 B	0	2.00 TB	1.00 TB	100.00 K	0.00 %	ok	ok
47663	47.05 GB	23.52 GB	31.30 K	2.00 TB	1.00 TB	200.00 K	2.35 %	ok	ok
49162	1.36 TB	679.02 GB	64.57 K	0 B	0 B	0	100.00 %	ignored	ignored

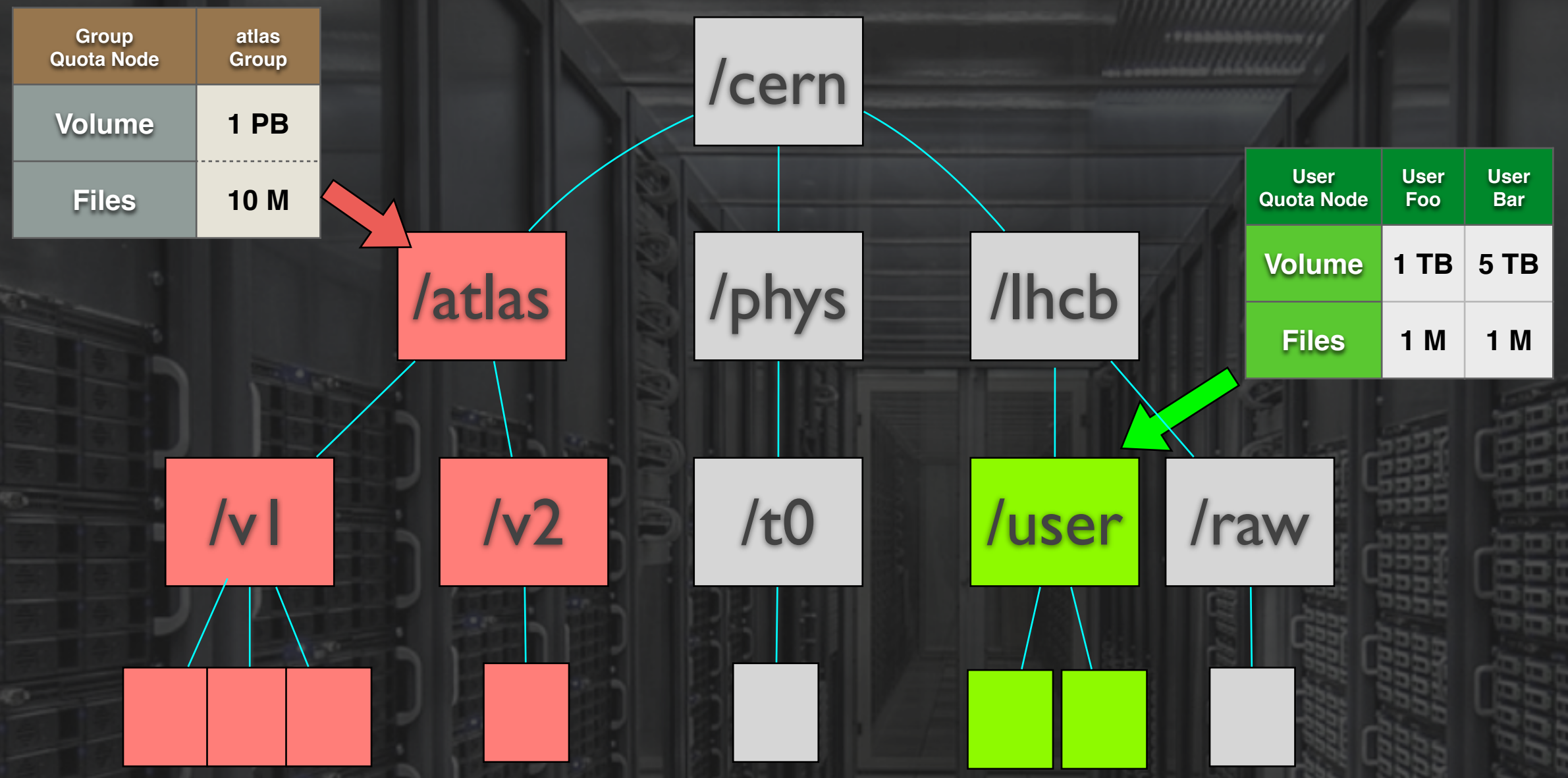
Physical Space Listing in EOS nominal quota per space

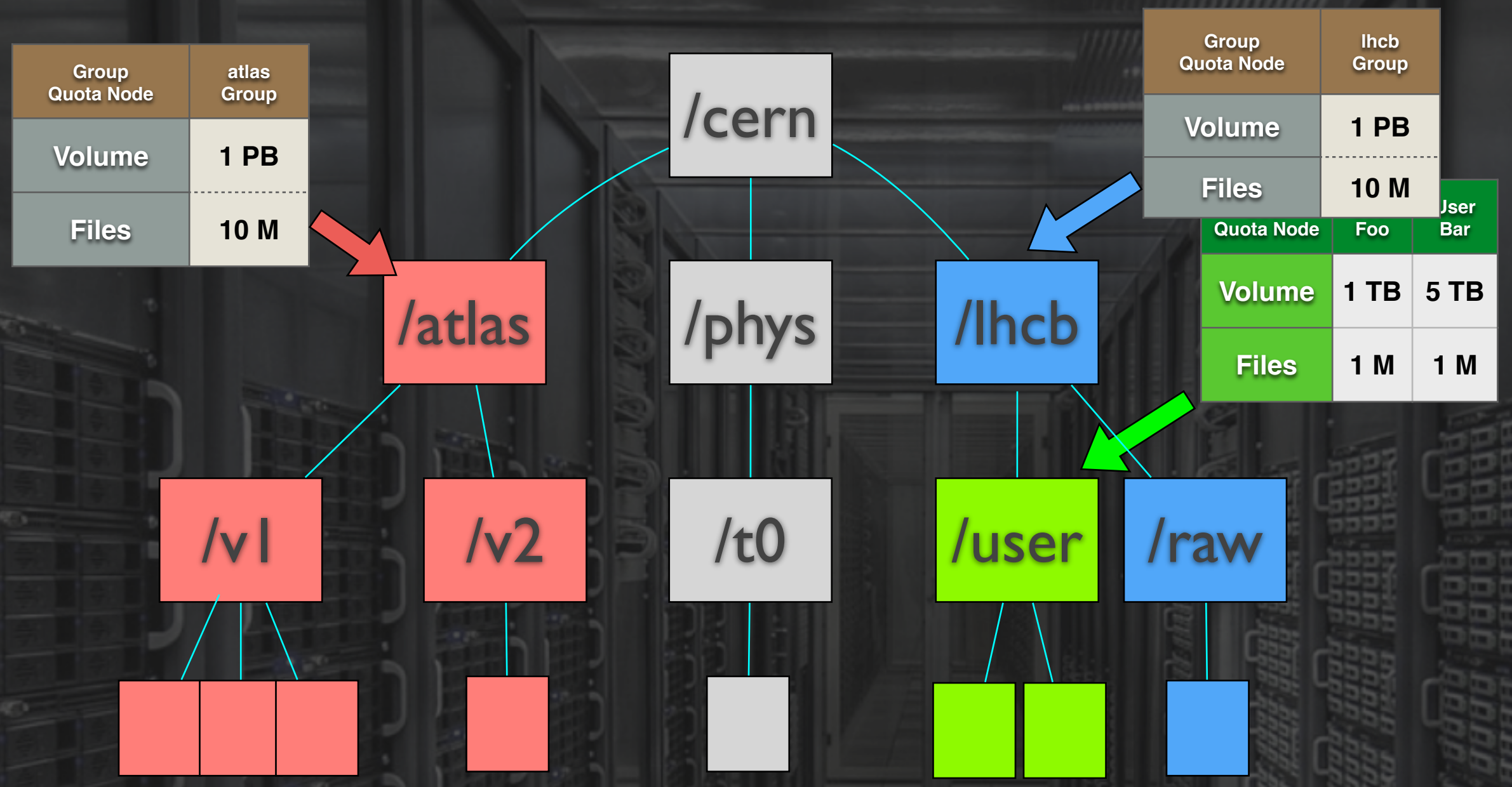
type	name	groupsize	groupmod	N(fs)	N(fs-rw)	sum(usedbytes)	sum(capacity)	capacity(rw)	nom.capacity	sched.capacity	usage	quota	balancing	threshold	converter	ntx	active	wfe	ntx	active	intergroup
spaceview	erasure	30	384	12066	12061	163.85 PB	181.53 PB	180.85 PB	180.00 PB	17.03 PB	91.37	off	off	6	on	400	0	off	1	0	on

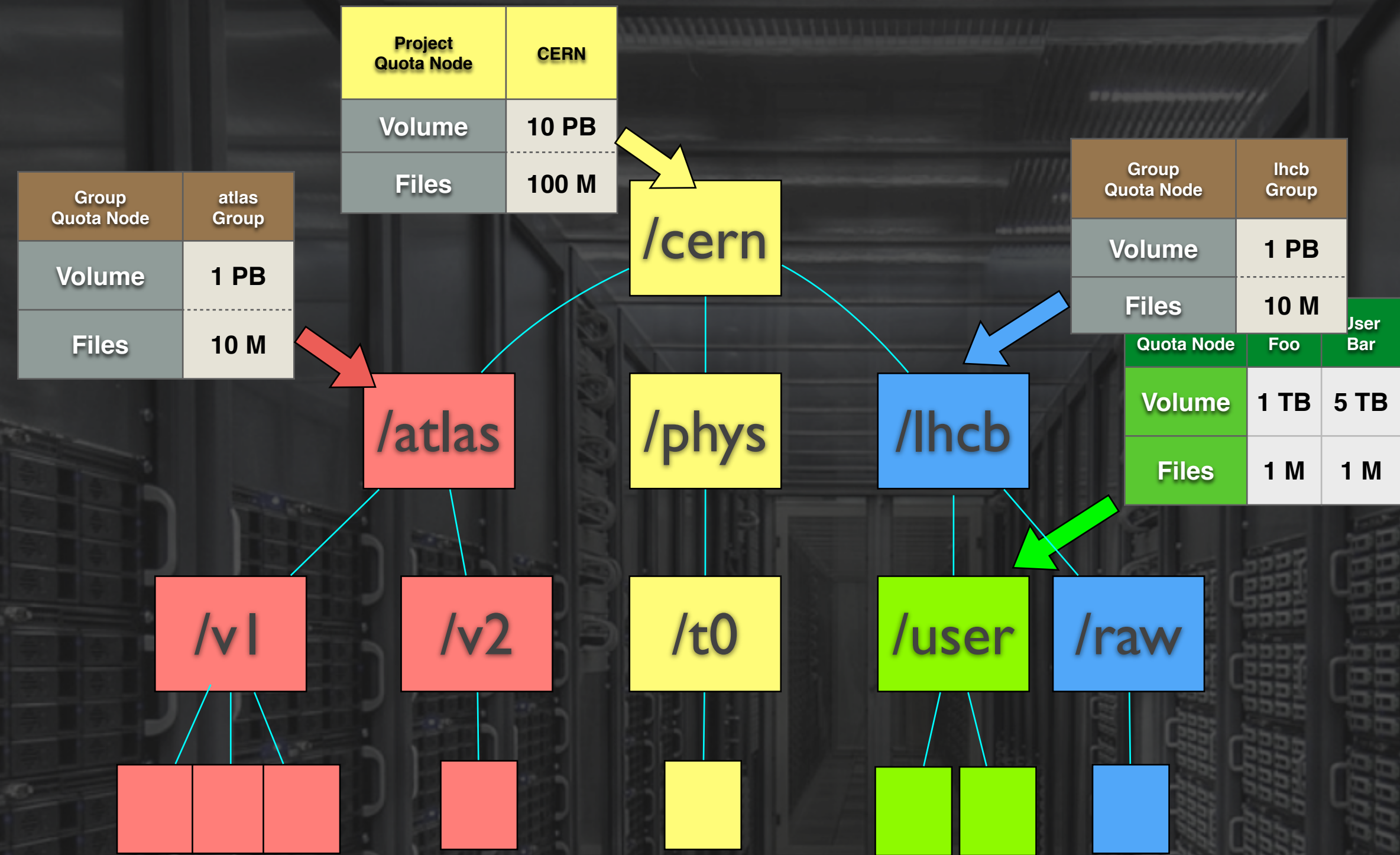


Group Quota Node	atlas Group
Volume	1 PB
Files	10 M









EOS

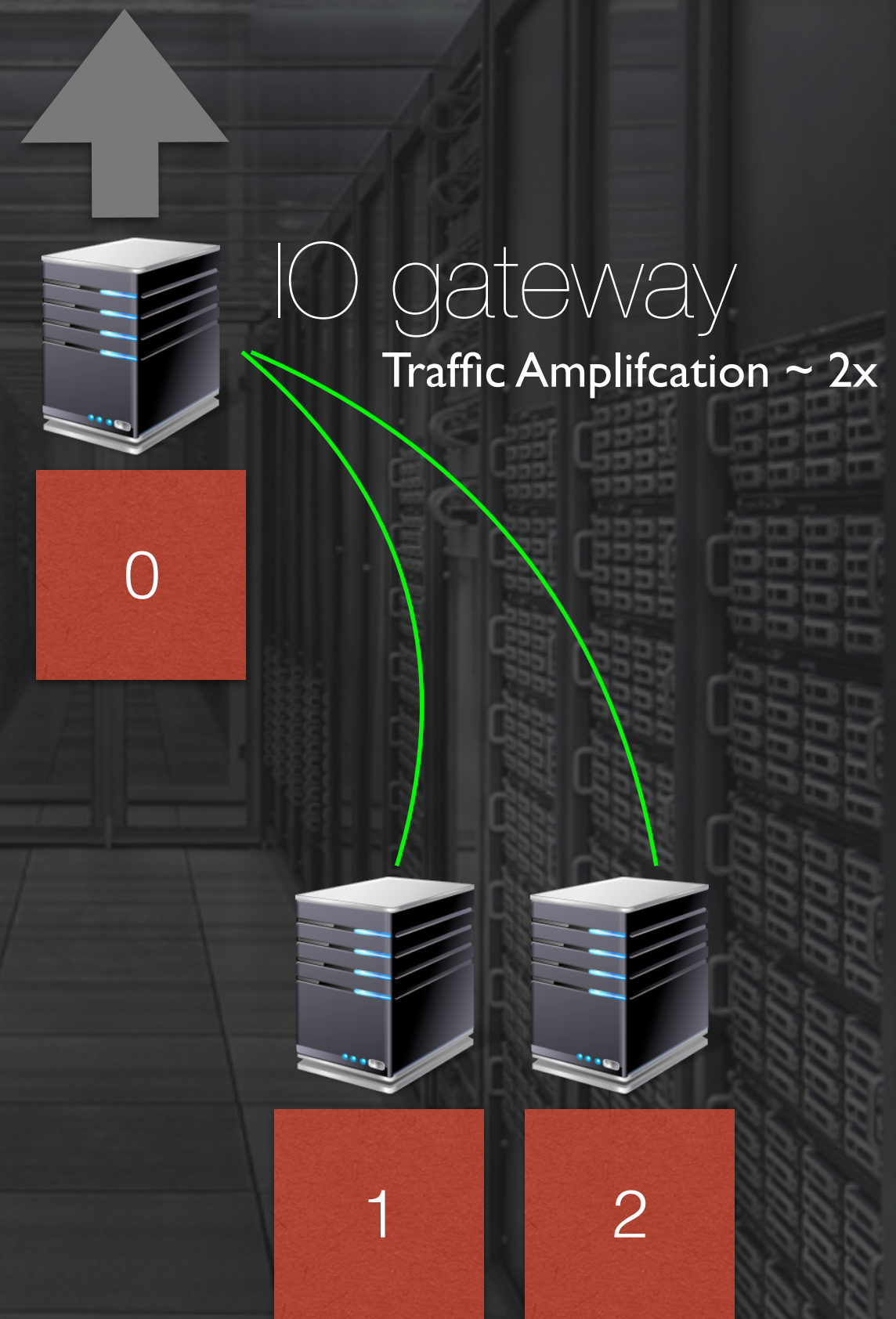
Erasure Coding

Erasure Coding in EOS is implemented using the JERASURE open-source library

EOS Erasure Coding IO path



Read



Read



EOS EOS4Physics Usage at CERN 2023



Total amount of files read

15.8 Bil

Total amount of bytes read

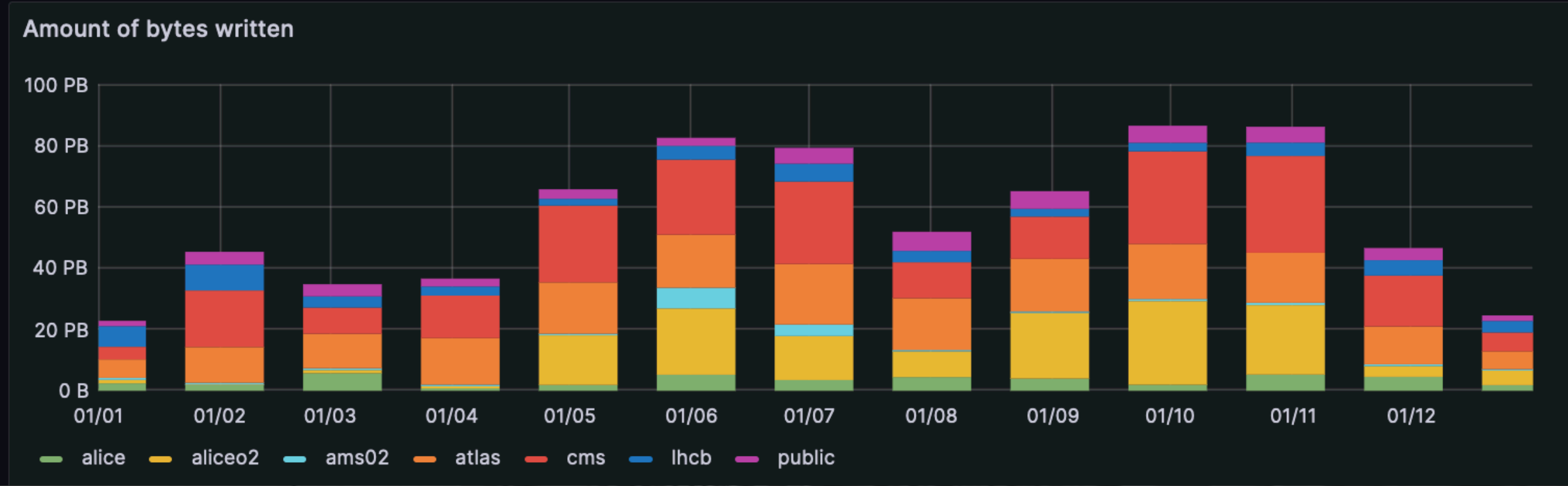
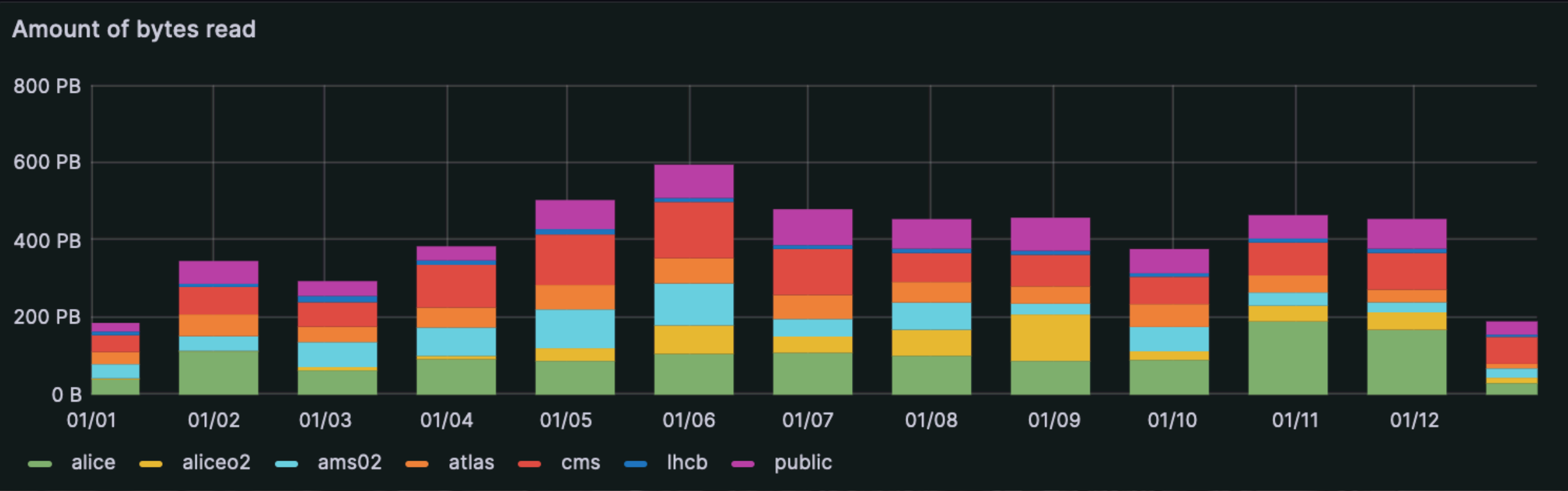
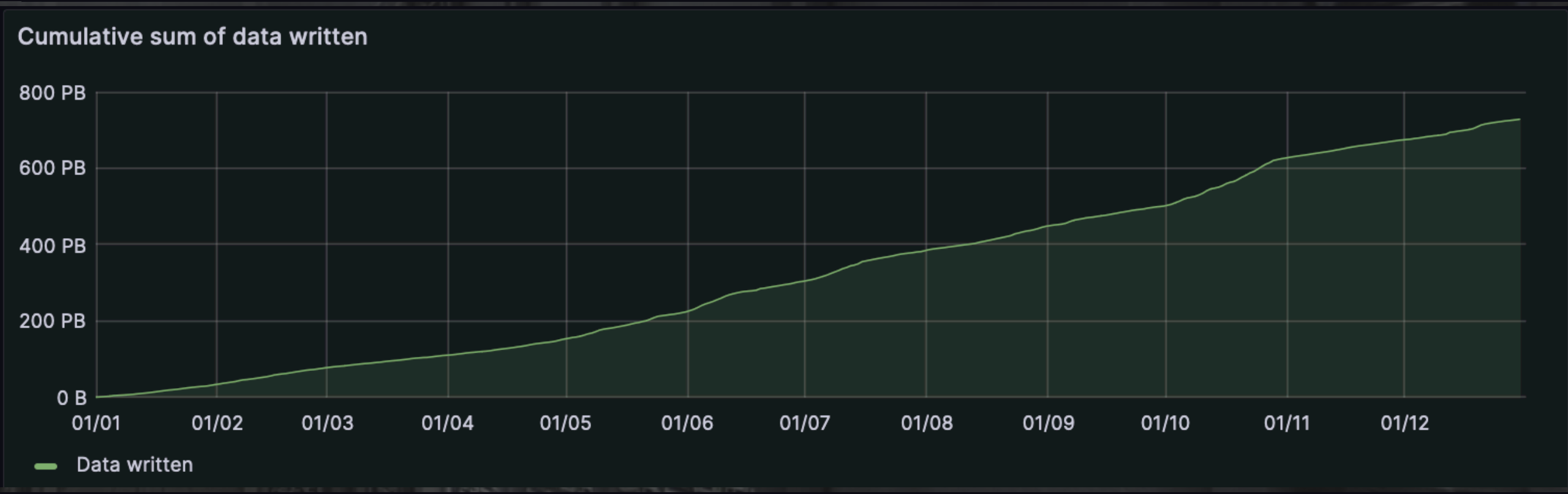
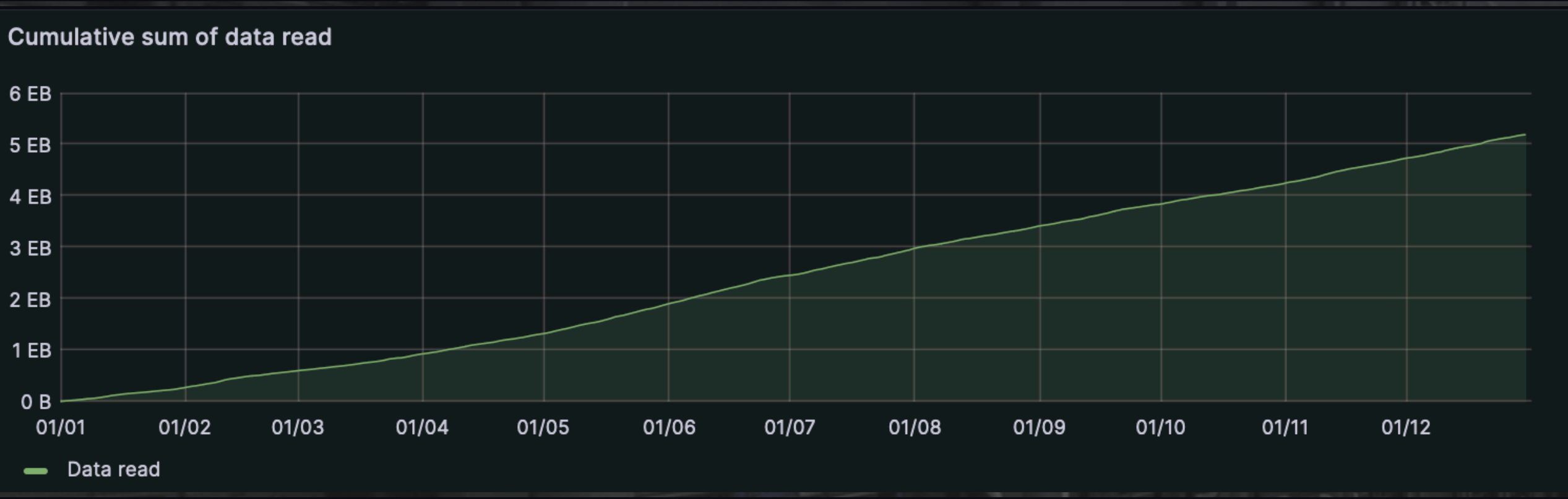
5.13 EB

Total amount of files written

1.48 Bil

Total amount of bytes written

668 PB



EOS EOS4Physics Usage at CERN 2023

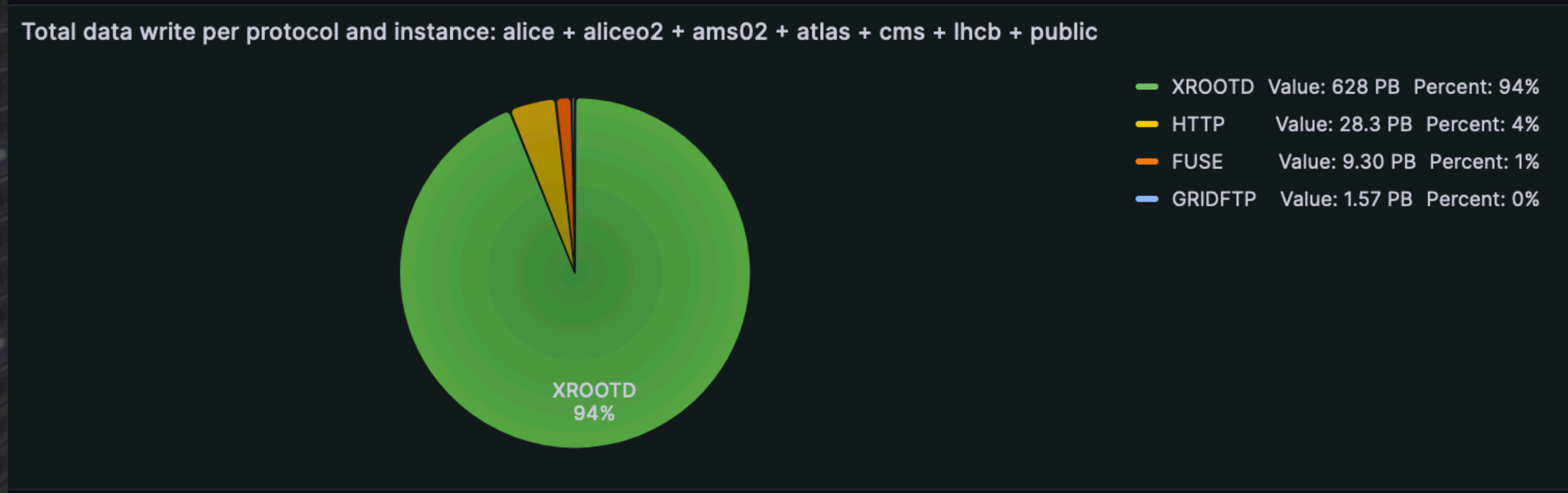
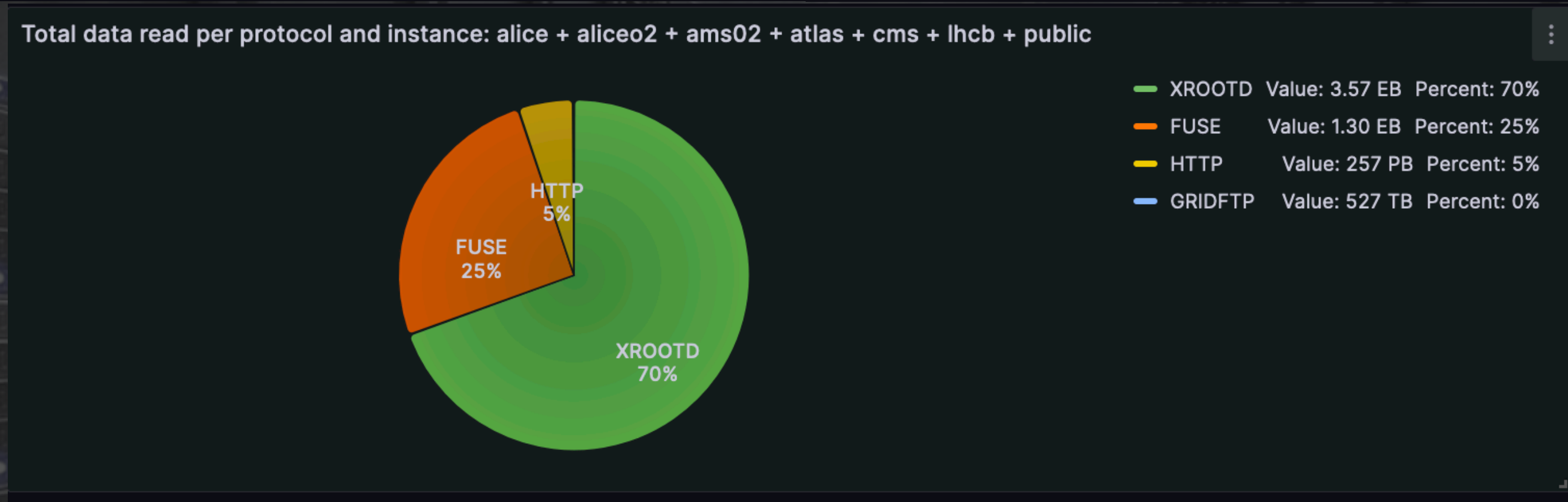


Total amount of files read
15.8 Bil

Total amount of bytes read
5.13 EB

Total amount of files written
1.48 Bil

Total amount of bytes written
668 PB



High Energy Physics

- **EOS** is used by dozens of sites in WLCG - we don't have exact numbers because we do not ship telemetry in software installations - few examples:
 - **EOS** is used in the **KISTI Tier-1** centre (Korea) as tape system replacement using Erasure Coding with 4 parities
 - **EOS** is used at **IHEP** (China) with many installations
 - **EOS** is used at **Fermilab** (US)



Other

- **EOS** is used by the Joint Research Centre JRC as Big Data Analytics Platform

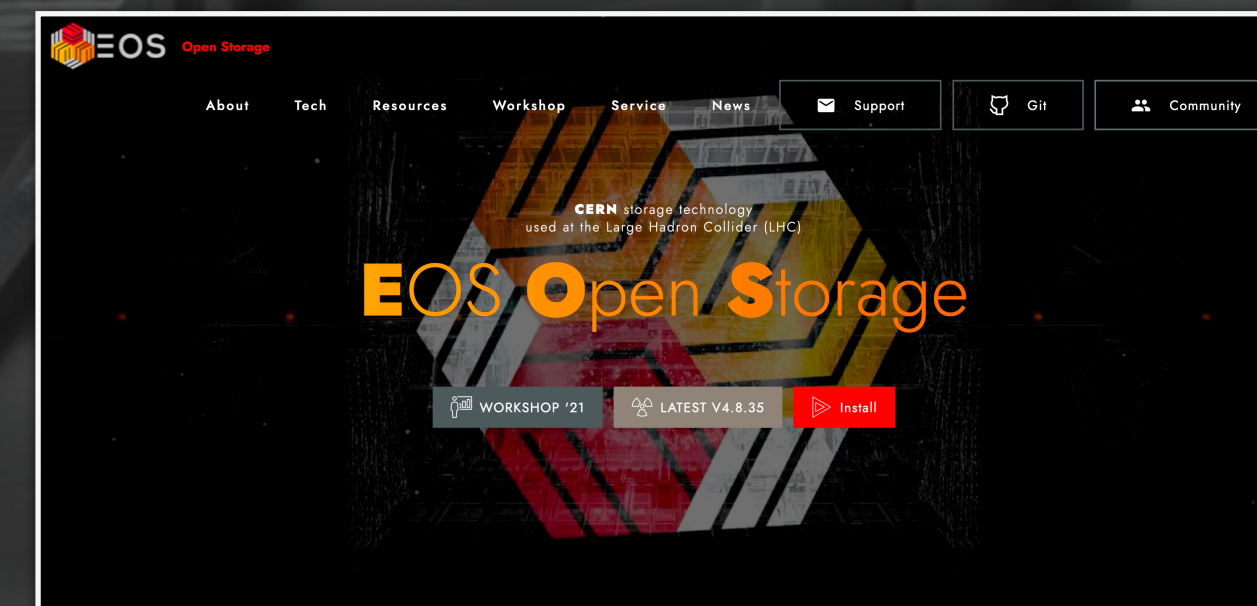


- **EOS** is **used and developed at CERN** as software to provide Large Scale Disk Storage for Physics use-cases & beyond
 - this spans from data acquisition systems to end-user analysis
 - involved in almost every physics analysis done at CERN
- **EOS** is particularly **appreciated by users** due to the fact that the same data repository is accessible from almost everywhere remotely, as a filesystem or via web and Sync&Share applications
- With LHC Run-4 starting in 5 years **new storage challenges** are upcoming
 - we are constantly improving **EOS** and try to prepare for the future
- You will get many **more details tomorrow** during the 8th EOS workshop!



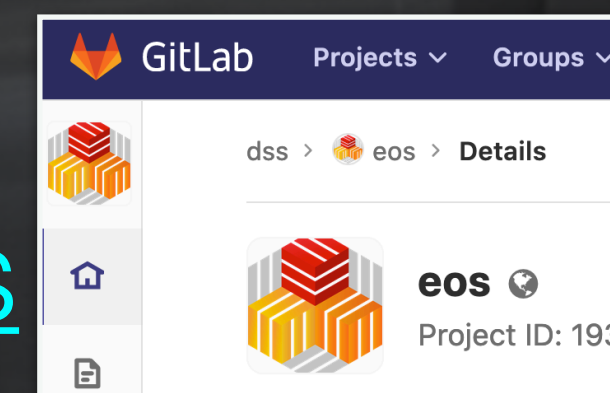
Joining EOS

Web Page <https://eos.cern.ch>



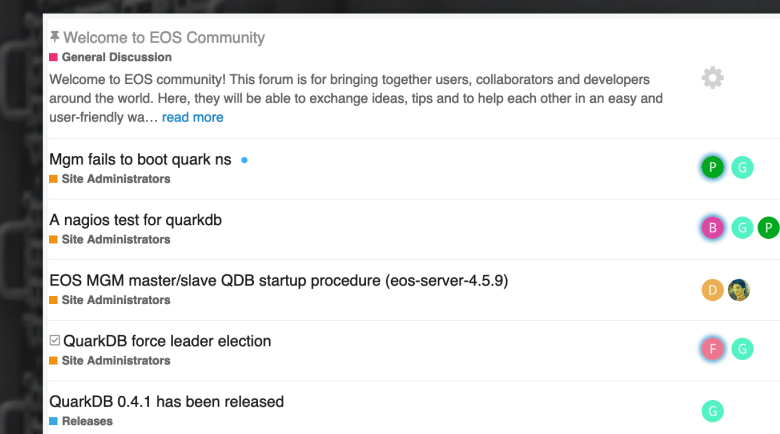
GITLAB Repository <https://gitlab.cern.ch/dss/eos>

GITHUB Mirror <https://github.com/cern-eos/eos>



Community Forum <https://eos-community.web.cern.ch/>

email: eos-community@cern.ch



Documentation <http://eos-docs.web.cern.ch/eos-docs/>

Support email: eos-support@cern.ch



Thank you for your attention!
Questions?



