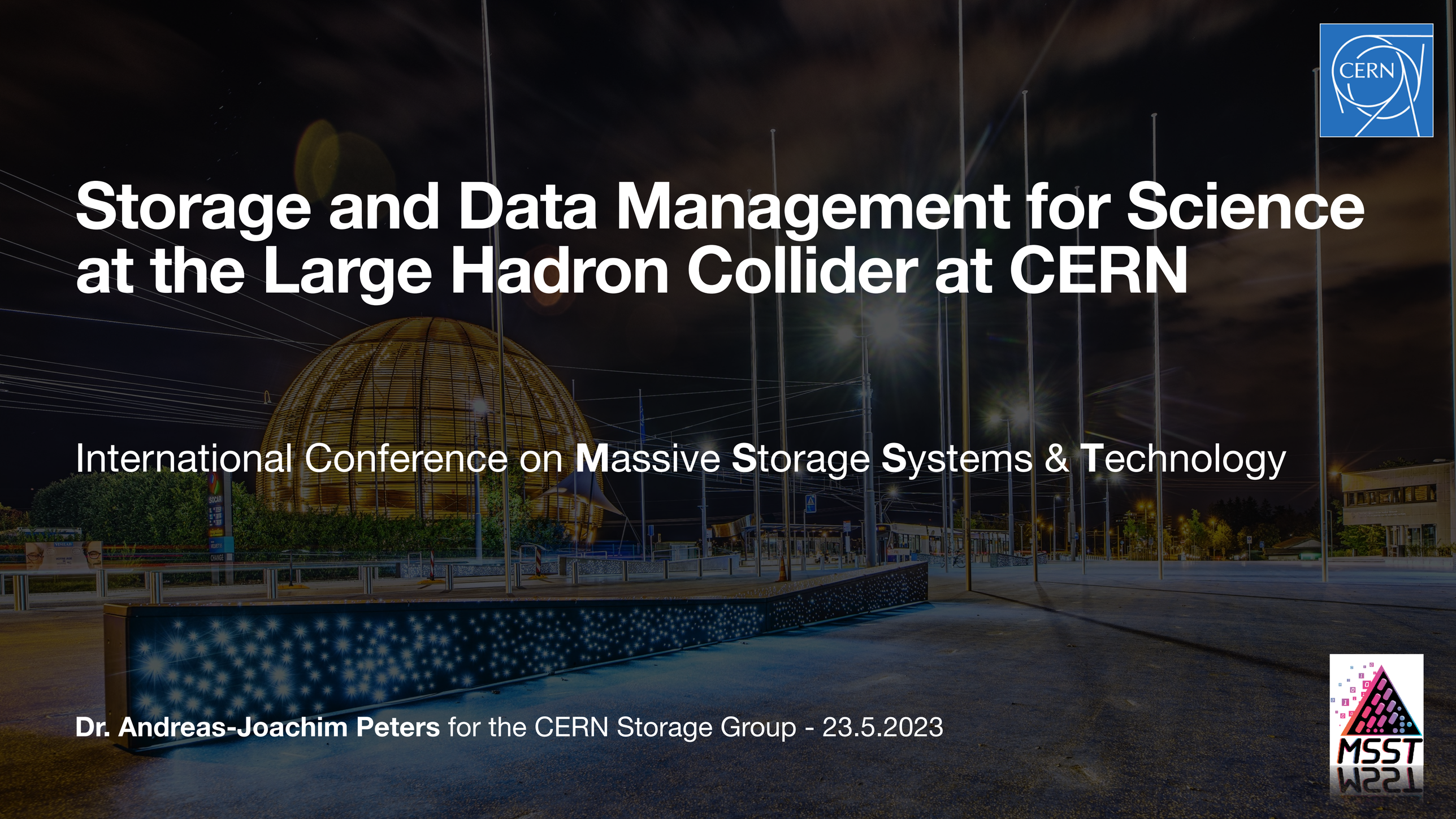
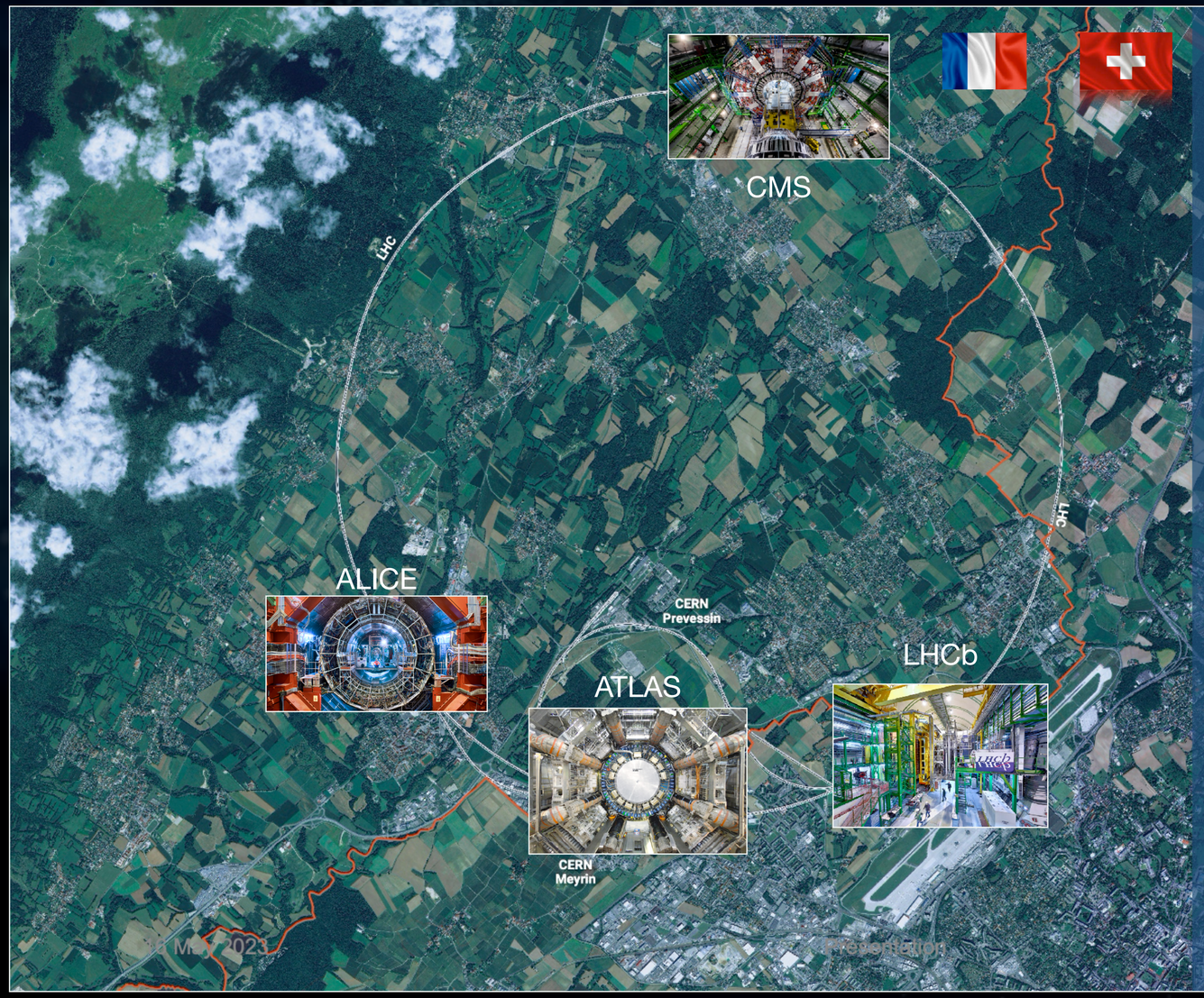


Storage and Data Management for Science at the Large Hadron Collider at CERN

International Conference on **M**assive **S**torage **S**ystems & **T**echnology

Dr. Andreas-Joachim Peters for the CERN Storage Group - 23.5.2023

About CERN



CERN is the world's biggest laboratory for particle physics.

Our goal is to understand the most fundamental particles and laws of the universe.

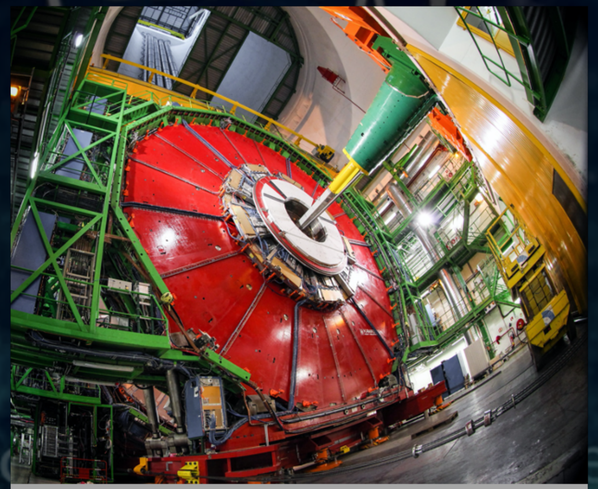
Located near Geneva on either side of the Swiss French border

Science at CERN

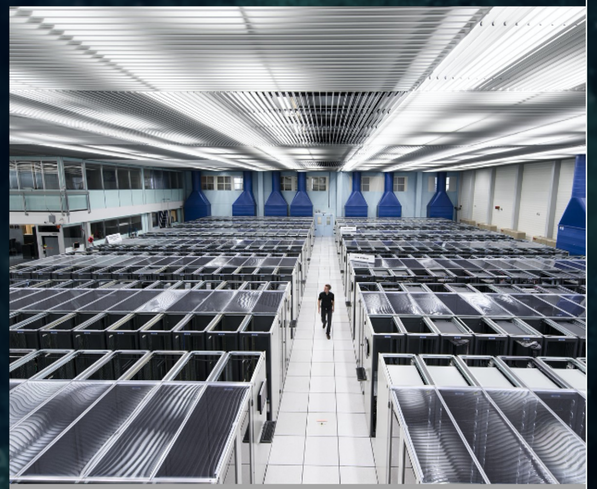
How do we do it?

- We build the largest machines to study the smallest particles in the universe
- We develop technology to advance the limits of what is possible
- We perform world-class research in theoretical and experimental particle physics

ACCELERATORS

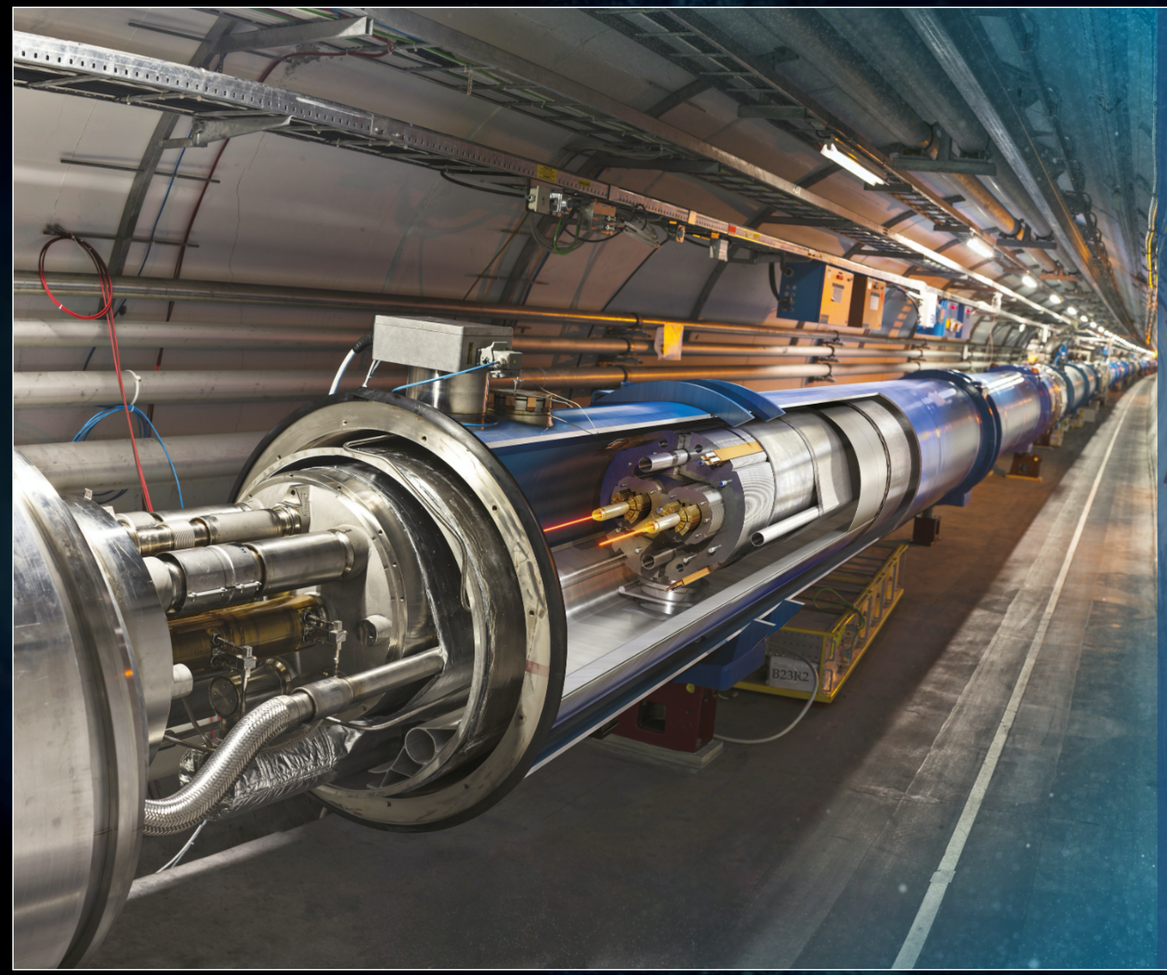


DETECTORS



COMPUTING

The Large Hadron Collider LHC

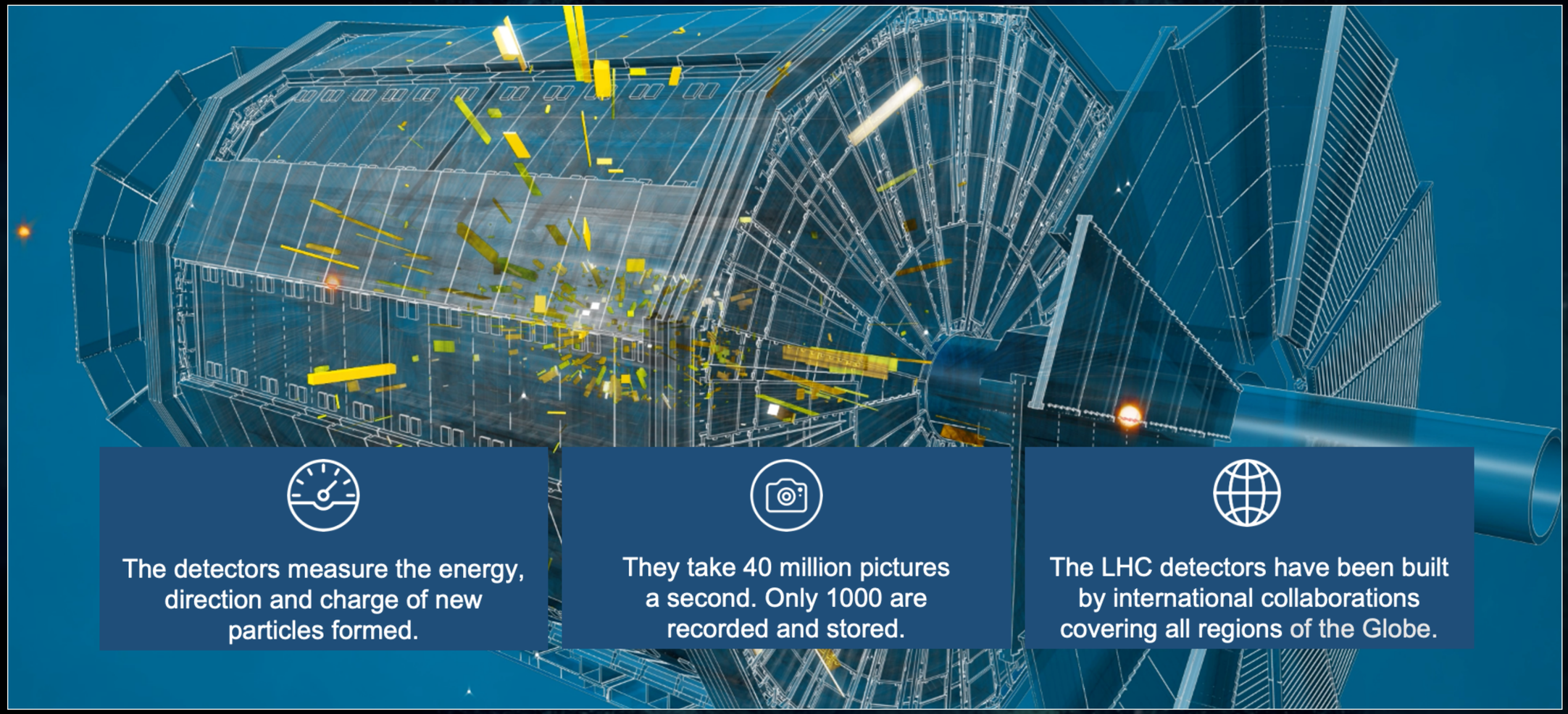


Large Hadron Collider (LHC)

- 27 km in circumference
- About 100 m underground
- Superconducting magnets steer the particles around the ring
- Particles are accelerated to close to the speed of light

Detectors

The LHC detectors are analogous to 3D cameras



The detectors measure the energy, direction and charge of new particles formed.

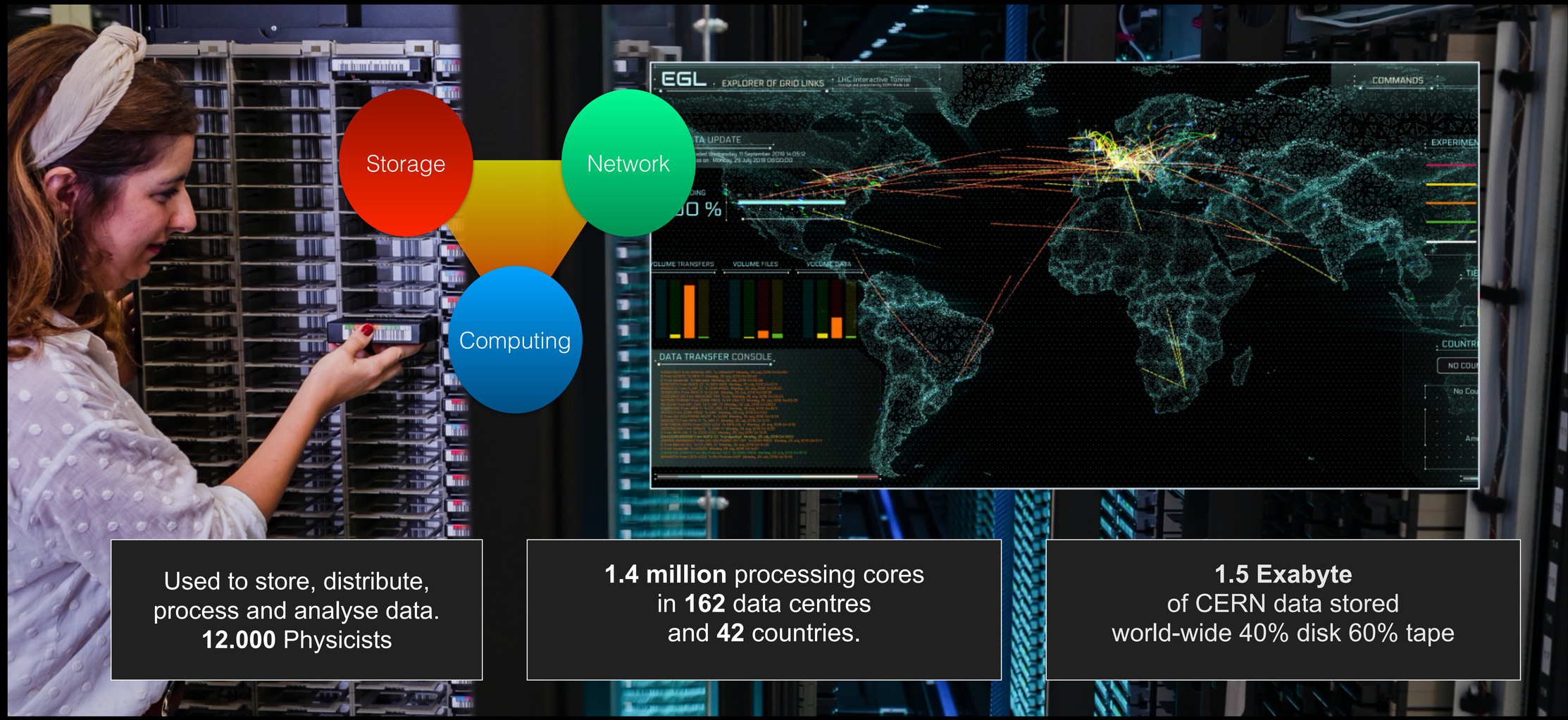
They take 40 million pictures a second. Only 1000 are recorded and stored.

The LHC detectors have been built by international collaborations covering all regions of the Globe.

Computing

The Worldwide LHC Computing GRID - WLCG

providing the distributed computing infrastructure for the needs of LHC experiments



Used to store, distribute, process and analyse data.
12.000 Physicists

1.4 million processing cores
in **162** data centres
and **42** countries.

1.5 Exabyte
of CERN data stored
world-wide 40% disk 60% tape

Open Source Storage & Data Management

Software developed at CERN / Physics Community

Storage

Data Management

EOS

eos.cern.ch

Software to manage Disk Storage - **780 PB**

DISK TRANSFER

FTS

File Transfer Service

fts.cern.ch

Middleware to run File Transfers - **1 Billion / year**

CERN Tape Archive

cta.cern.ch

Software to manage Tape Storage - 600 **PB**

TAPE Data DISTRIBUTION

RUCIO

SCIENTIFIC DATA MANAGEMENT

rucio.cern.ch

Data Management /
Data Distribution over 162 sites

Open Source Storage & Data Management

Software developed at CERN / Physics Community

Data Access

XRootD

xrootd.org

Client/Server Framework
root:// + http(s):// access protocol

Remote Access

- provides **crucial** data management **features**
  - **vector operations** for WAN access
  - **third party** storage-to-storage **copy** implementation for transfers with ROOT and HTTPS protocol
  - **checksumming, caching, encryption** …
  - strong authentication & token AUTHZ

SOFTWARE

- FUSE mounted read-only FS
- on-site caching
- provides **software distribution** to 1M cores
  - experiment software frameworks
  - docker/singularity images
  - calibration data and more …

Software Distribution

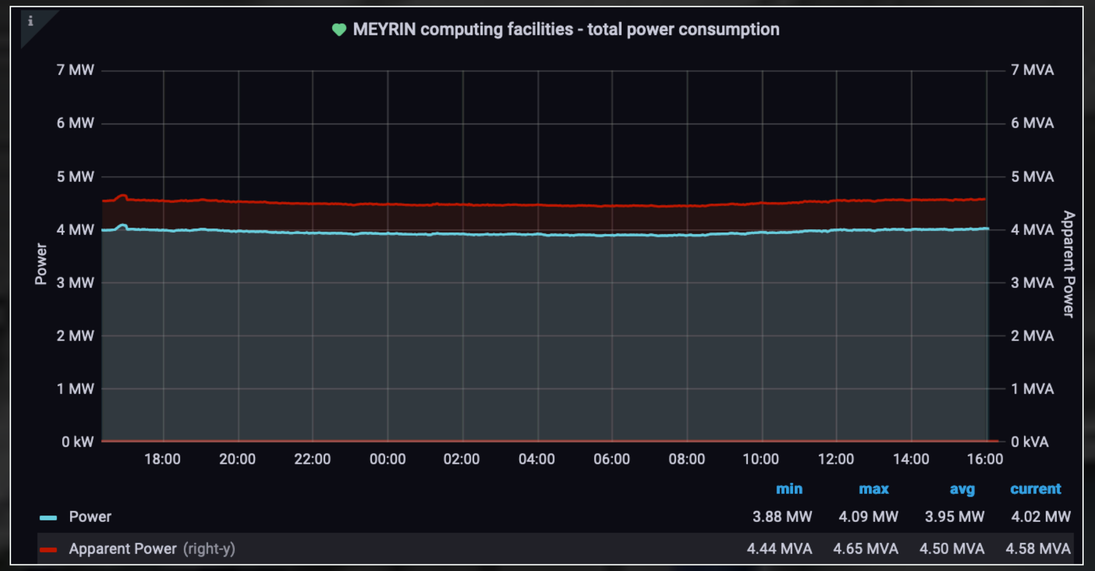
CernVM-FS
CernVM File System

cvmfs.web.cern.ch

presented at **MSST 2019**

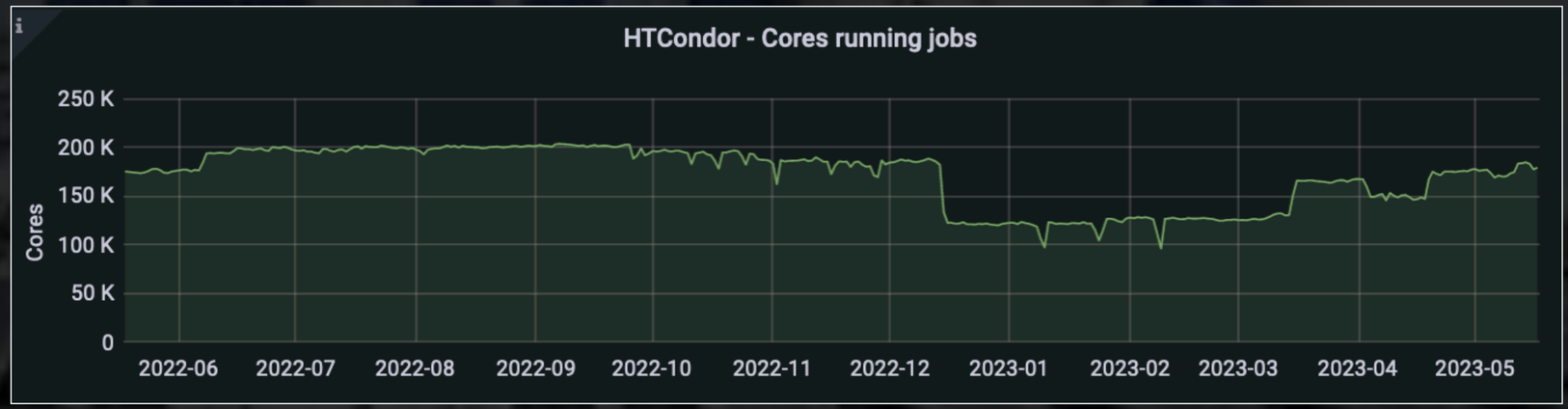
CDN for software/image distribution

CERN Data Center (Meyrin)



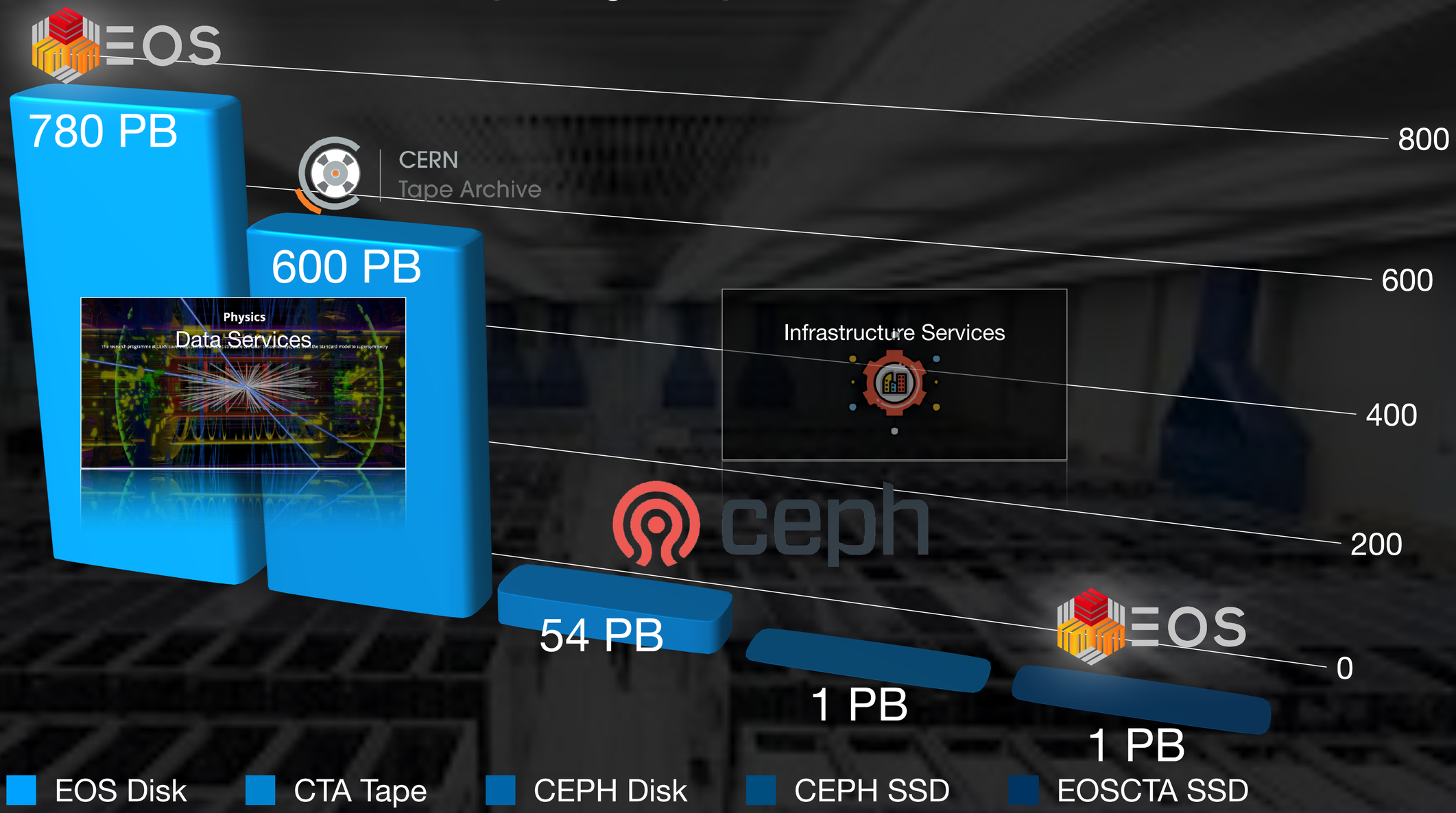
1 Exabyte Disk+Flash **+ 600 PB** Tape Media

~482.000 CPU Cores



- **CERN** uses **OpenStack** to manage infrastructure *IaaS*
- **HTCondor** as Batch system + Kubernetes & OKD4

CERN Data Center (Meyrin)



New CERN Data Center (Prevessin)

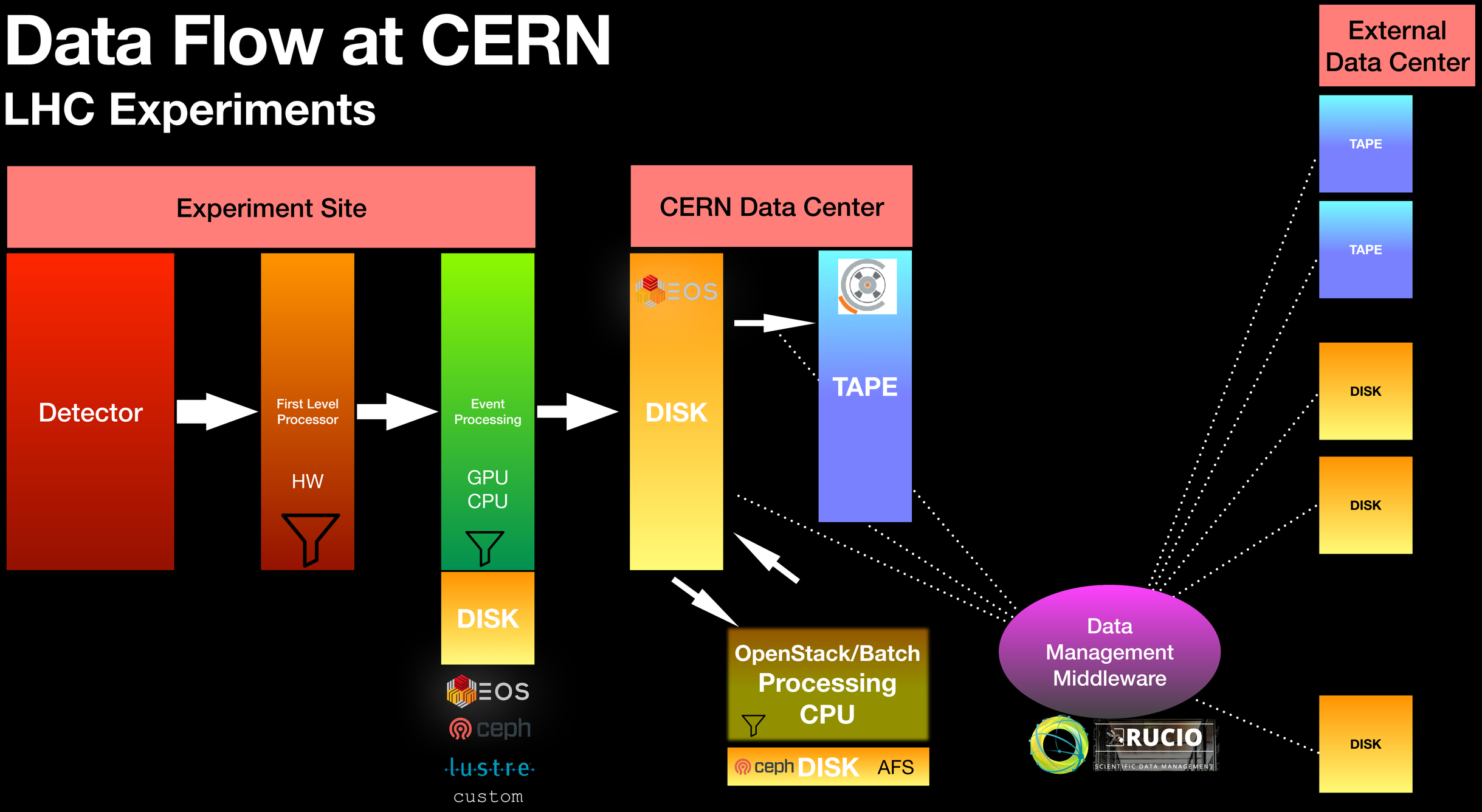
- will be online end of 2023
- **Meyrin** centre hosts most of **STORAGE**, **Prevessin** centre focus is **CPU**
- second computer centre allows to implement better business continuity for mission critical services

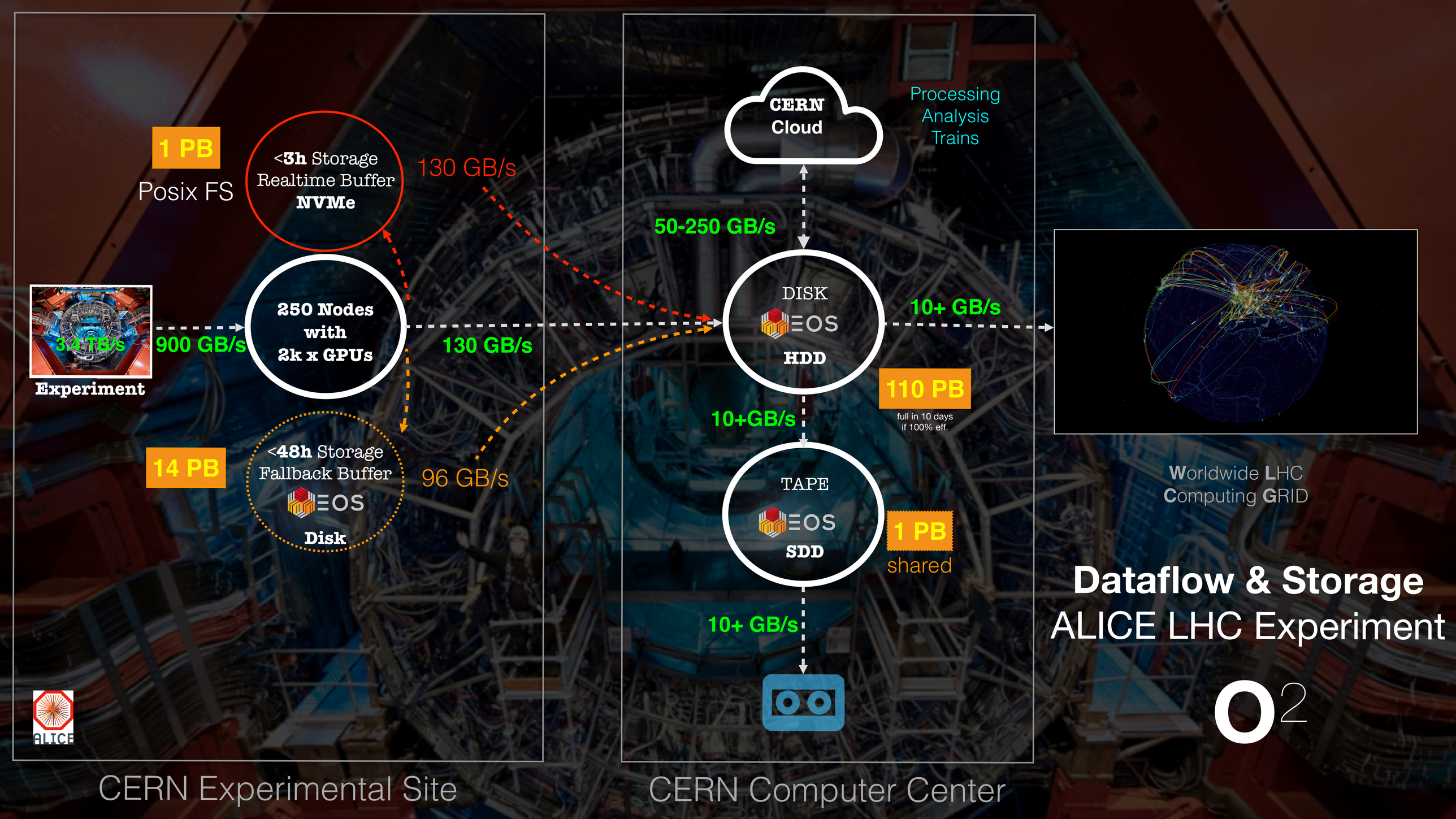
Meyrin DC PUE 1.5 - new Prevessin DC PUE 1.1

4MW installation - capable of 12MW

Data Flow at CERN

LHC Experiments




1 PB
Posix FS
<3h Storage
Realtime Buffer
NVMe
130 GB/s
3.4 TB/s
900 GB/s
250 Nodes
with
2k x GPUs
130 GB/s
Experiment
14 PB
<48h Storage
Fallback Buffer
EOS
Disk
96 GB/s
CERN
Cloud
Processing
Analysis
Trains
50-250 GB/s
DISK
EOS
HDD
10+ GB/s
110 PB
full in 10 days
if 100% eff.
10+GB/s
TAPE
EOS
SDD
1 PB
shared
10+ GB/s
Worldwide LHC
Computing GRID
ALICE
CERN Experimental Site
CERN Computer Center
Dataflow & Storage
ALICE LHC Experiment
O2

EOS Disk Service for Physics

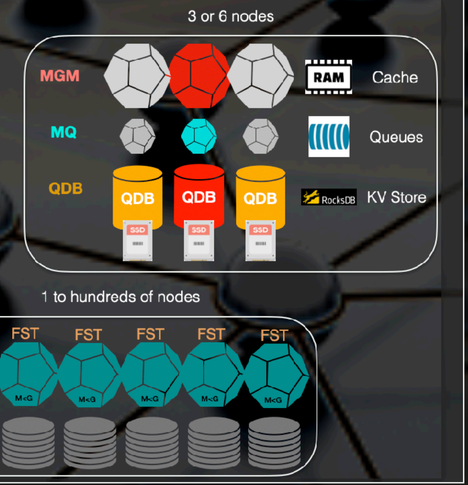
eos.web.cern.ch

EOS

EOS Architecture

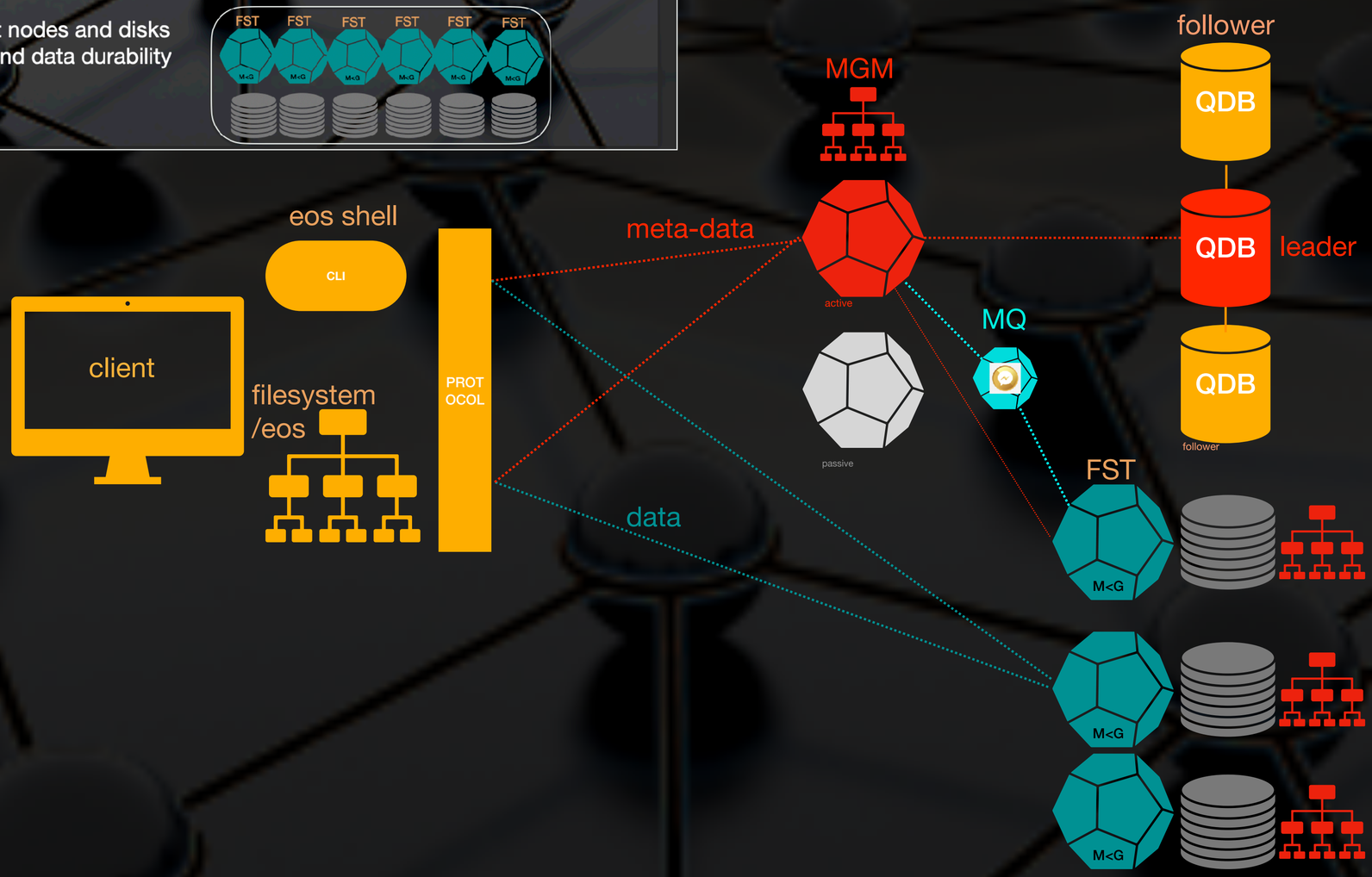
High-available and **low latency namespace**
- namespace persisted on a high available key-value store
- working entries cached in-memory

Highly available and **reliable file storage**,
based on (cheap) JBODs:
- File replication across independent nodes and disks
- Erasure coding to optimize costs and data durability



framework
XRootD

components
CLIENTs
MGM
MQ
FST
QuarkDB



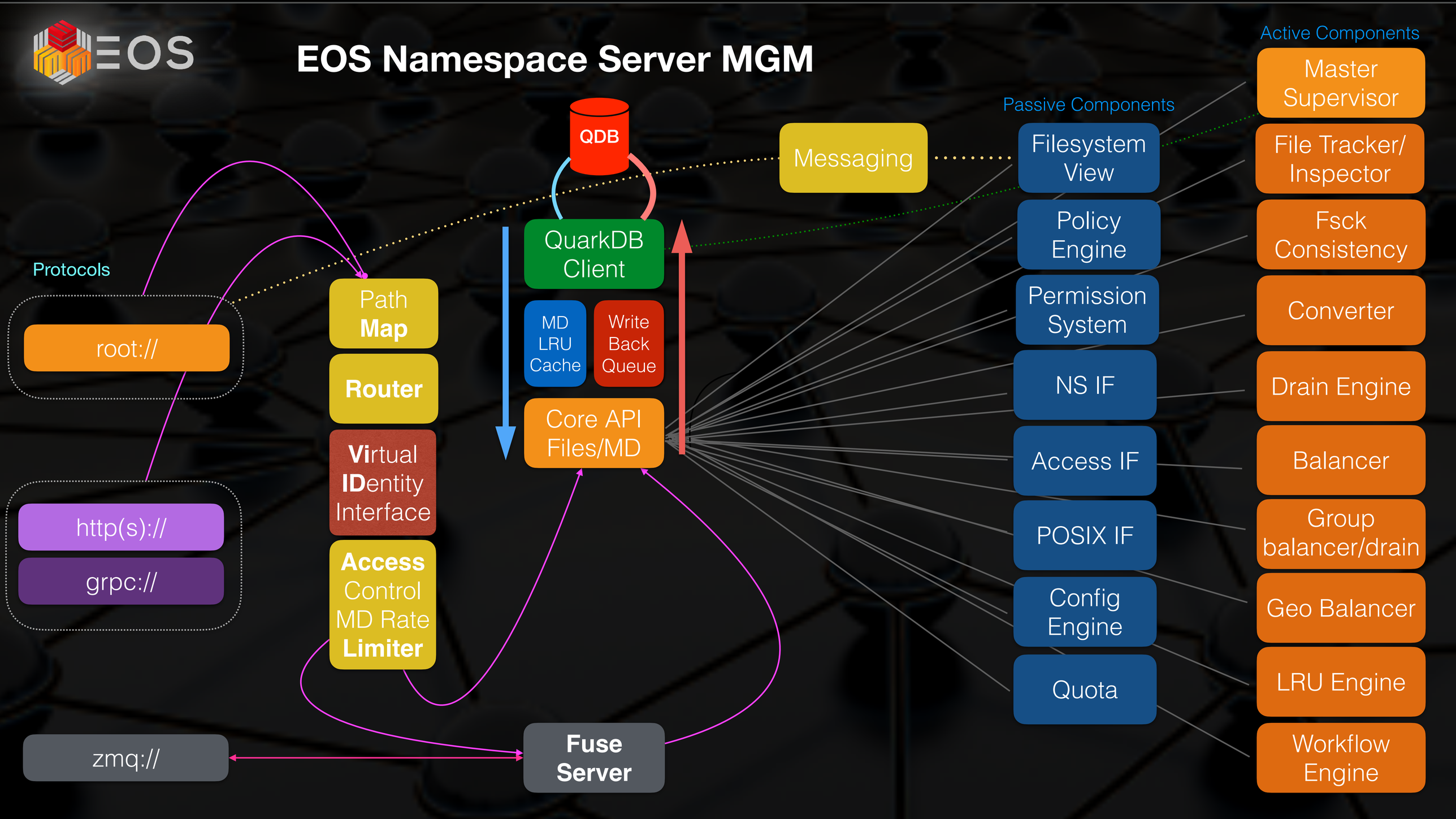
IO path

meta-data

data

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

EOS Namespace Server MGM



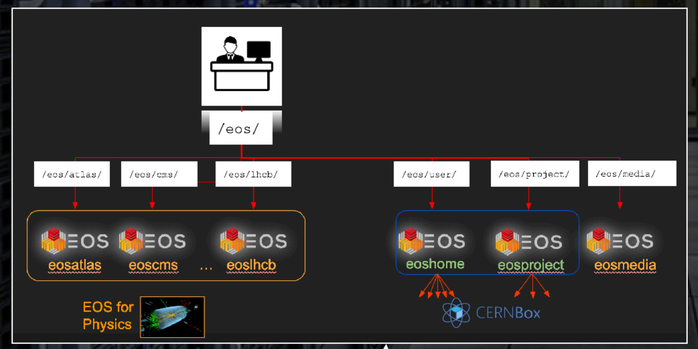
EOS Disk Service for Physics

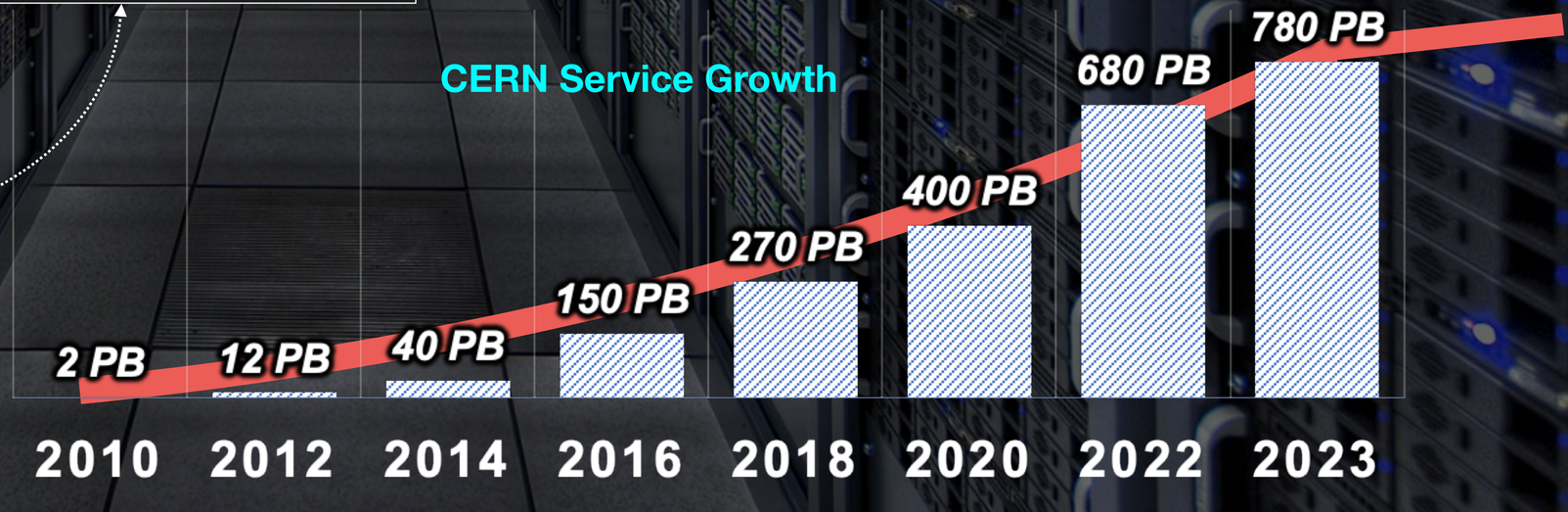
4 Exabyte read
in last 12mo

EOS 2023
CERN Service in numbers

- **780** PB
- **73k** HDD
- **1300** server
- **7.5 B** files
  0.61 B directories
- **24** instances

- smallest EOS instance has no storage at all
- largest EOS instance uses **RS(10+2)** Erasure Coding and provides **110 PB** usable space





EOS Disk Storage Server

JBODs with XFS filesystems

- Profiting from ***economy of scale***
  - minimise price per TB
- System Unit:
  - 1-2 CPUs: 8-16 physical cores 64-256GB RAM
  - disk-tray of 24 x **4-6-10-12-14-18** TB HDDs

- Running different generations
  - 2 trays per system unit - 48 disks
  - **4 trays per system unit - 96 disks**
  - 8 trays per system unit - 192 disks
  - 10/25/40/**100 GE** ethernet

2023 Server Configuration

96 x 18 TB HDDs
per server - **1.7 PB**

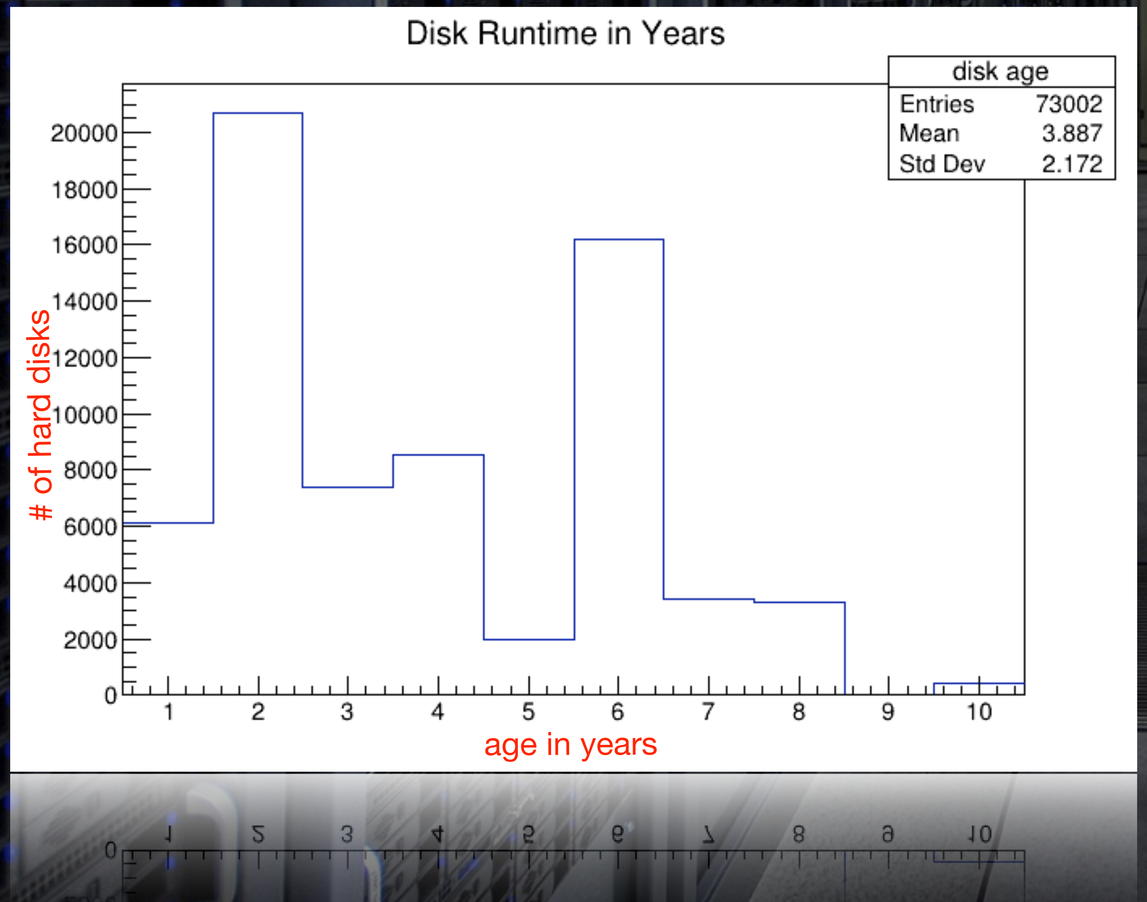
2 x 16-core CPU
256 GB Memory
100 GE Ethernet

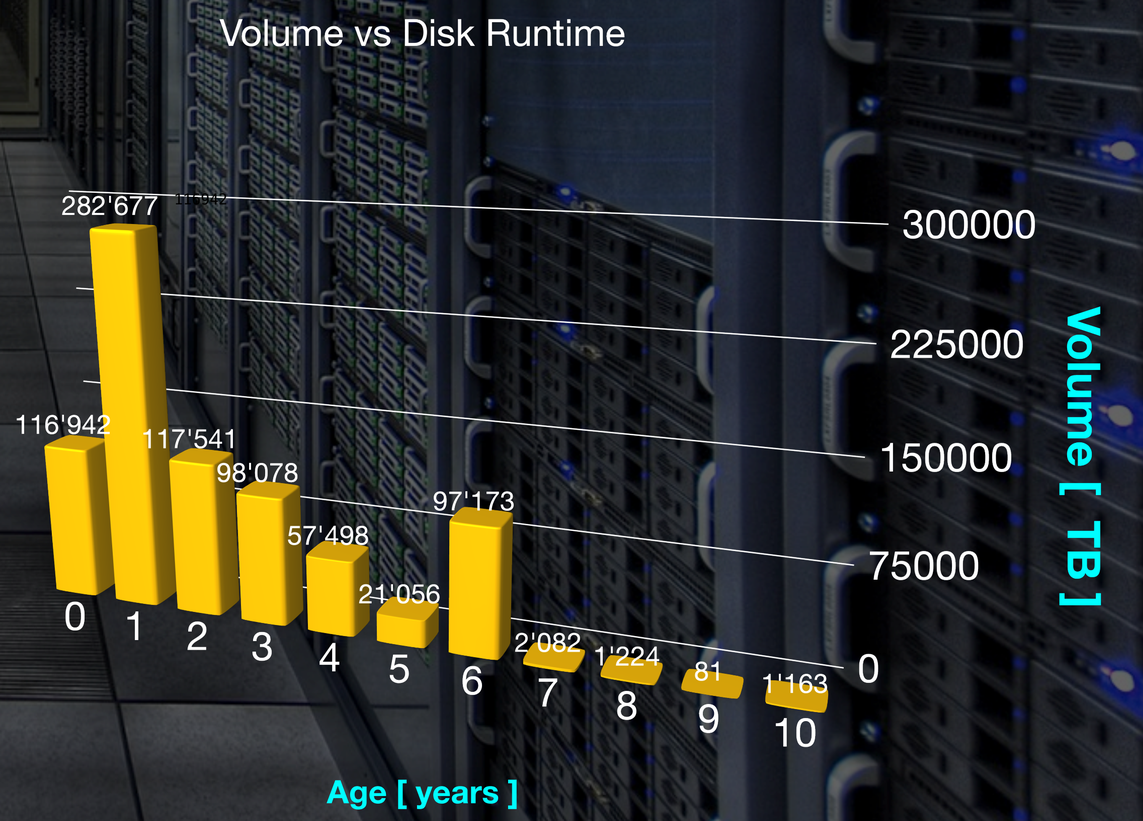
XFS Filesystems
on JBODS

EOS Disk Service for Physics

~1/4 of capacity is running out of warranty
1.5% of HDDs are 10 years in production

Average Age of HDDs **3.8** years

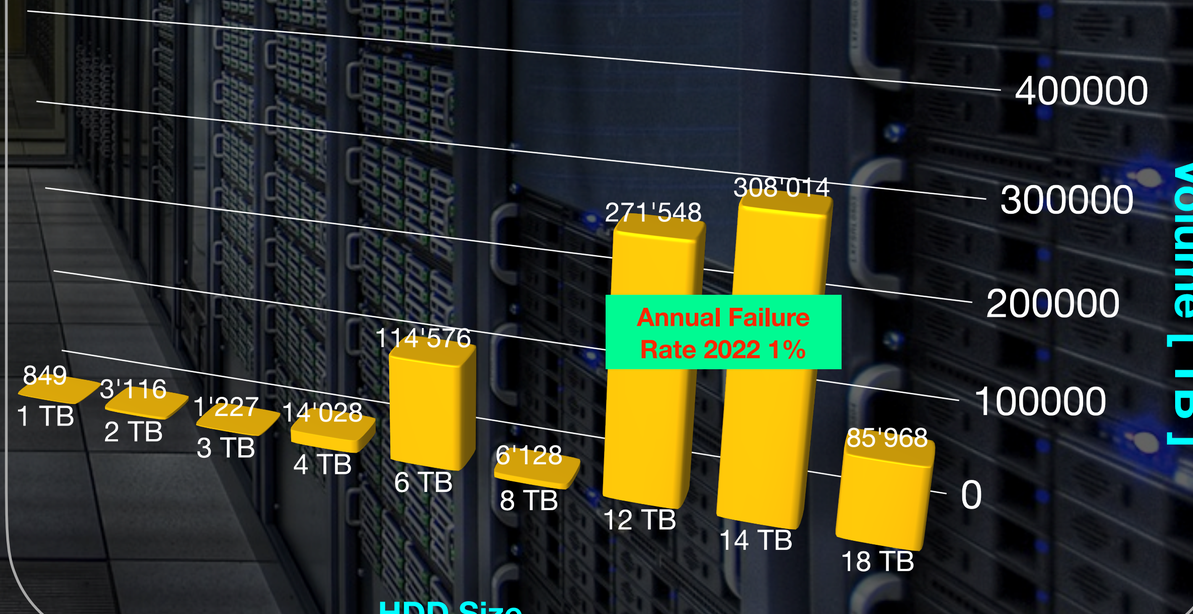




EOS Disk Service for Physics

HDD Size Distribution [TB]

Volume by HDD Size [TB]

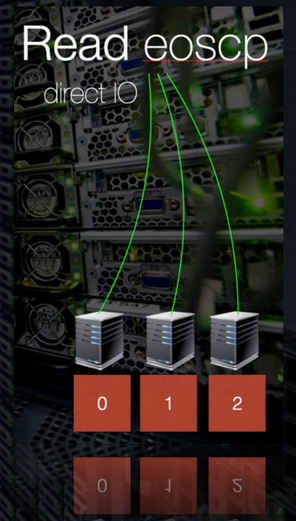


EOS O^2 Storage Benchmark

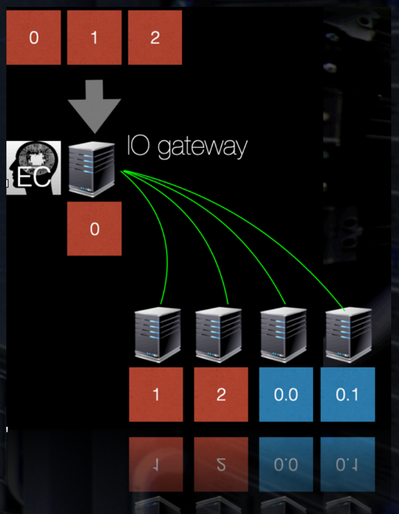
EOS ALICE O^2 is the largest EOS installation at CERN which got expanded recently to 9600 HDDs hosted by 100 storage nodes with 100GE technology using Reed Solomon 10+2 Erasure Coding - capacity **137 PB** [eff. 110]

EOS supports client-side EC for reads to avoid traffic amplification!

Client-side EC

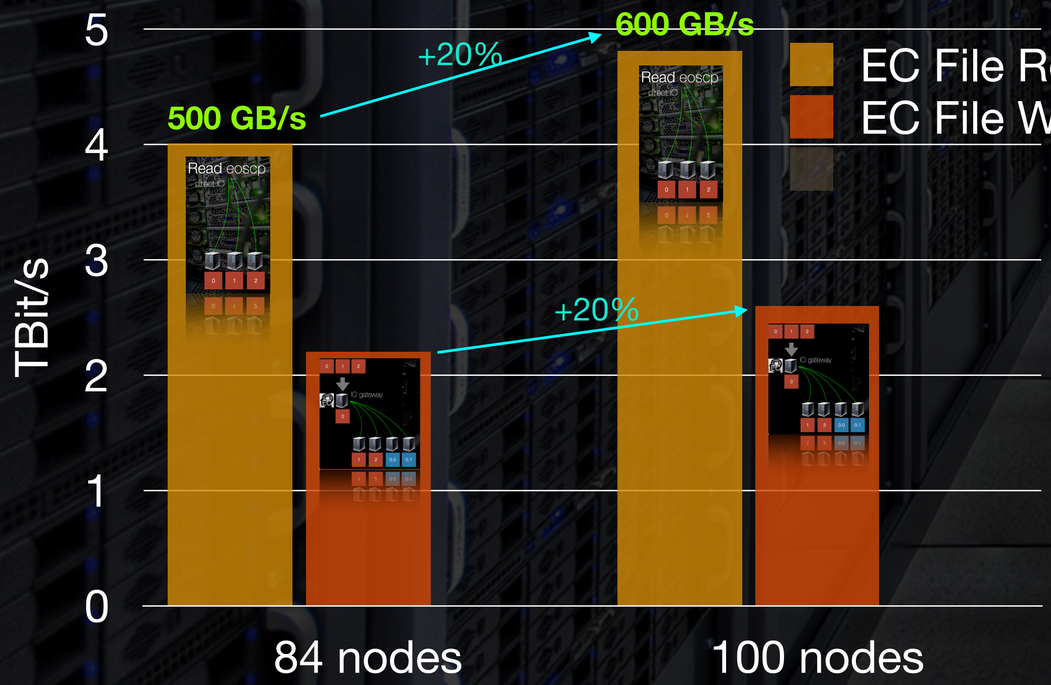


Server-side EC



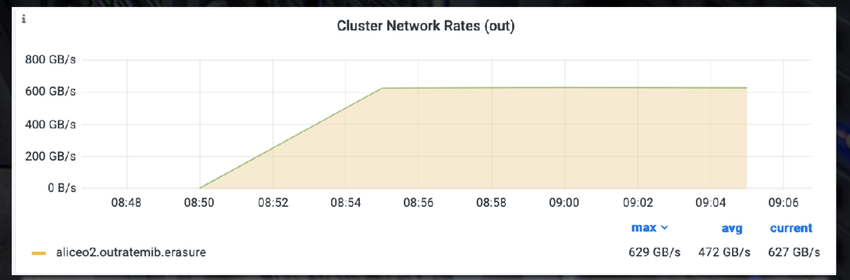
READ or WRITE

6800/7400 cp Streams 20G files



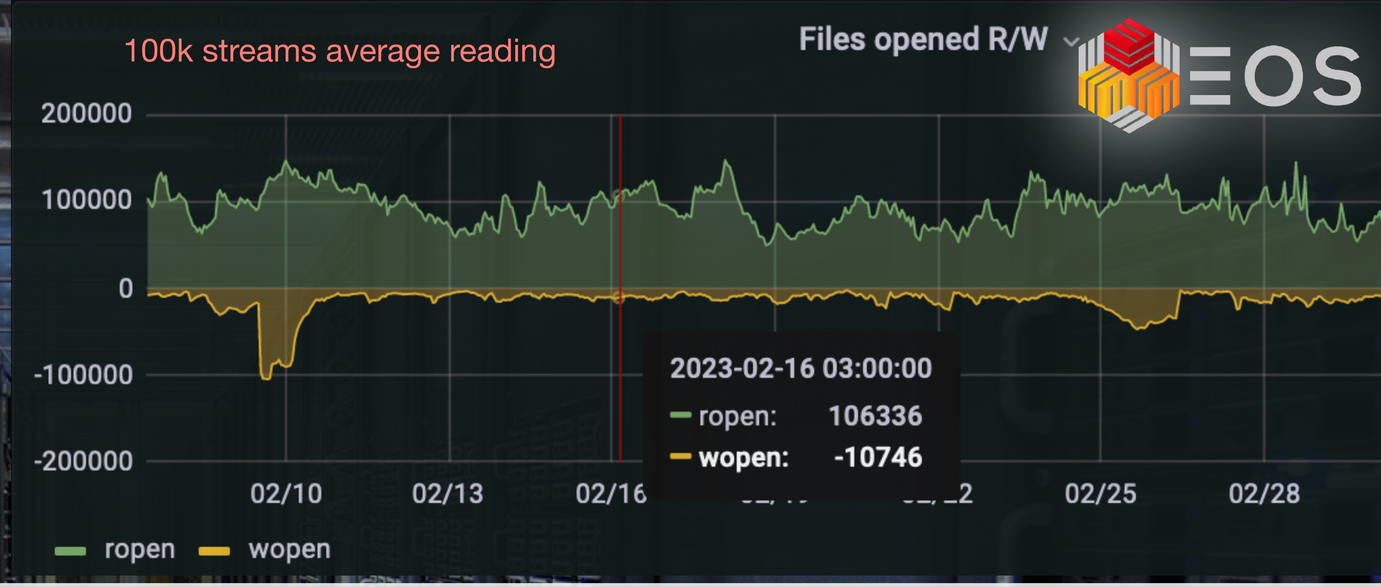
Perfect BW scaling for 20% capacity increase!

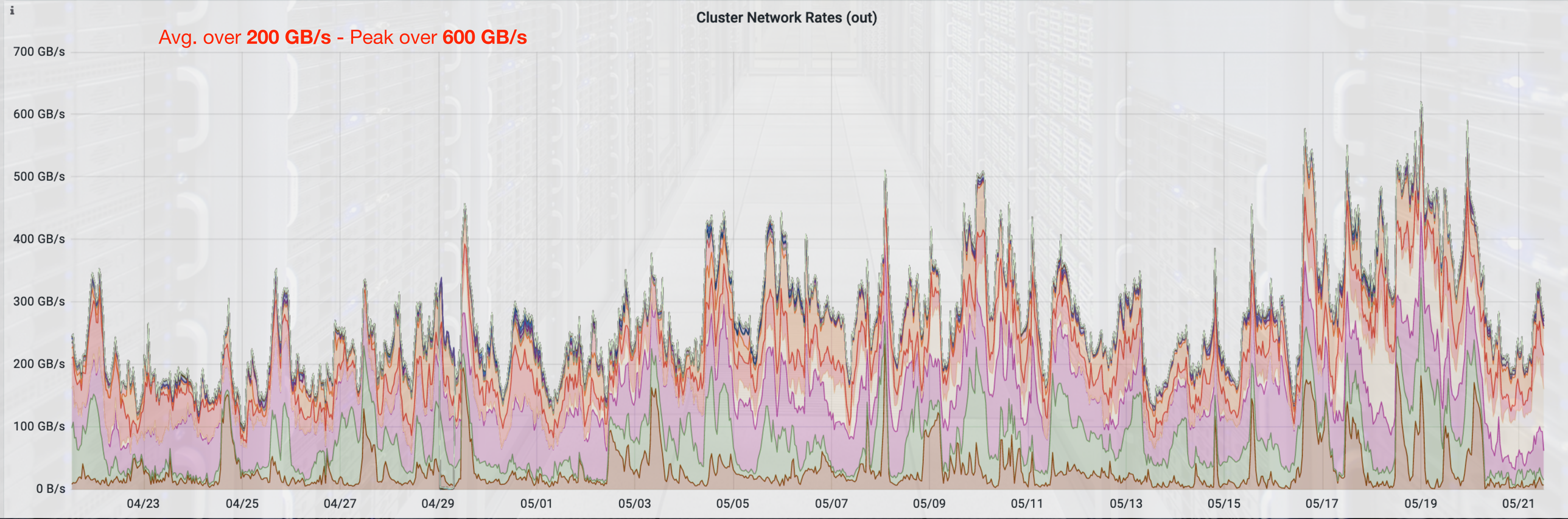
600 GB/s network traffic



EOS Read Activity

All Production CERN Services



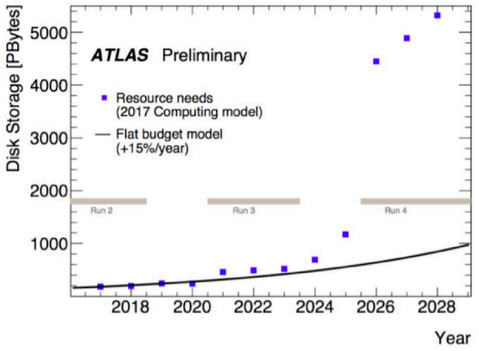


CERN
Tape Archive

Why TAPE is important for LHC

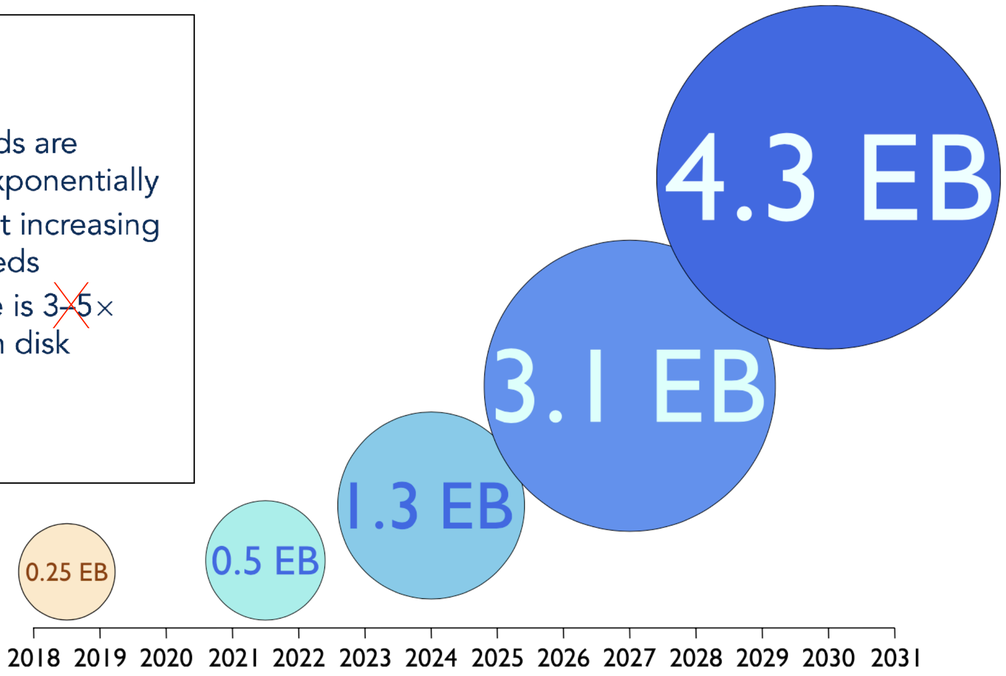
Cost optimisation using storage tiering in WLCG

Advantages of Tape : Cost!



- Storage needs are increasing exponentially
- Budget is not increasing to match needs
- Tape storage is 3–5× cheaper than disk storage

Predicted Tape Archival Storage Needs



CERN Tape Archive

Archival System for a custodial copy of all data at CERN

CERN Tape Archive

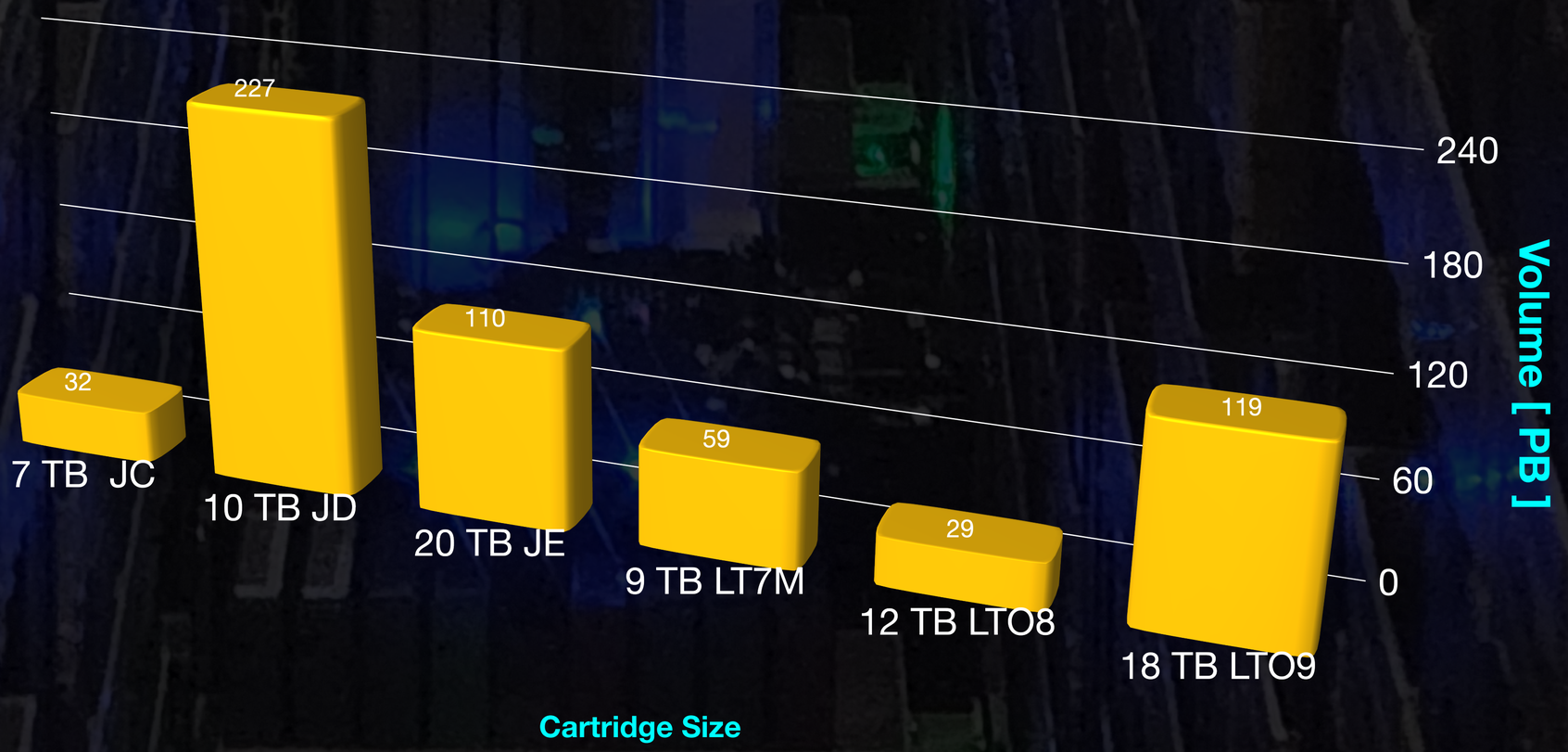
600 PB Media

520 PB stored

3 x IBM TS4500

2 x Spectra Logic TFinity

86 IBM **drives**
108 LTO **drives**

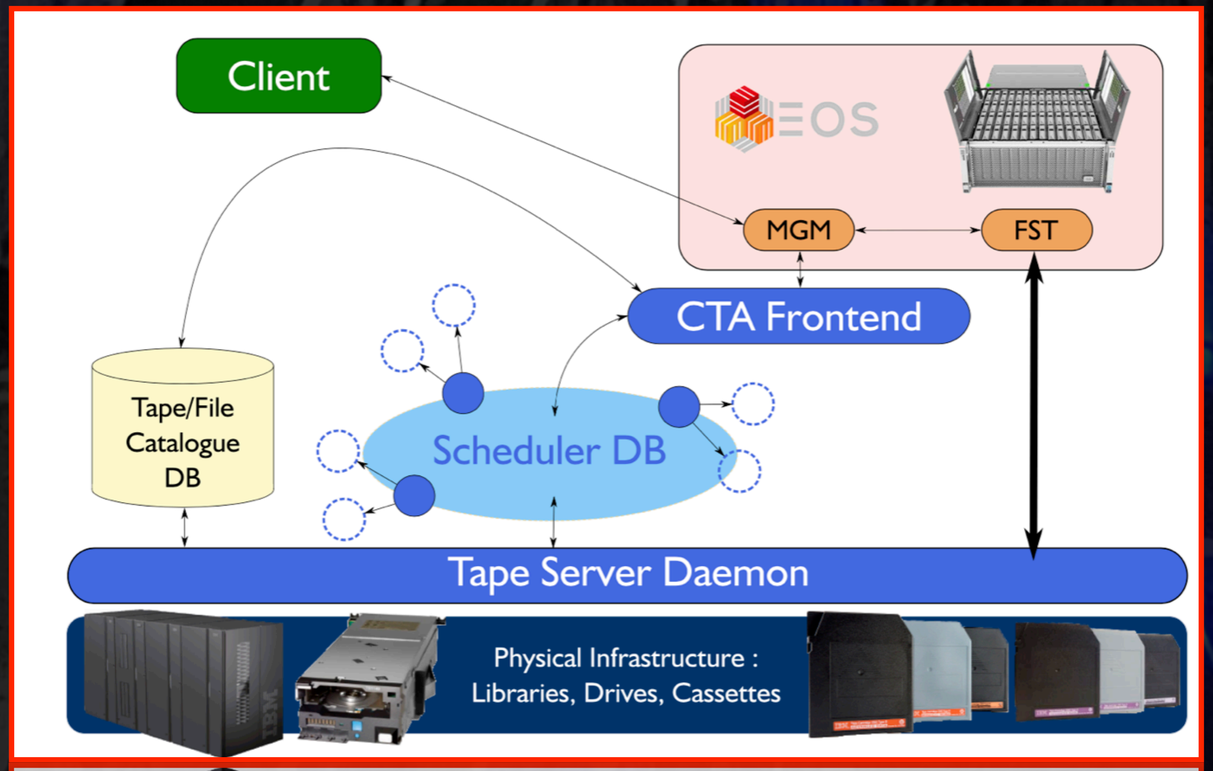


CTA Architecture & Deployment Model

CERN Tape Archive

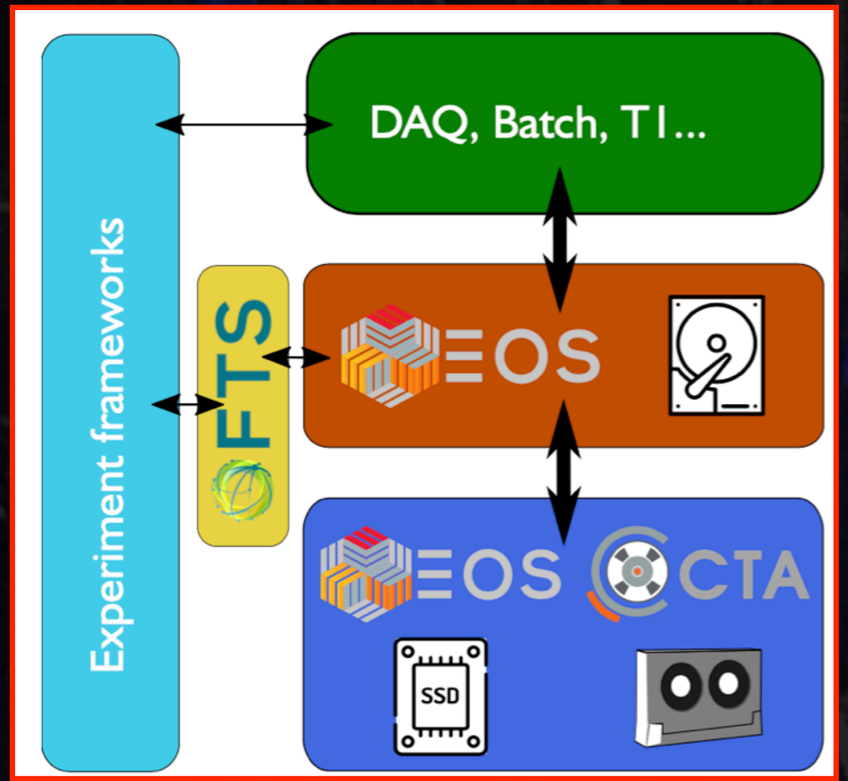
CTA Architecture

CTA uses EOS and few specific extensions to implement migration/staging workflows + tape copy tracking.

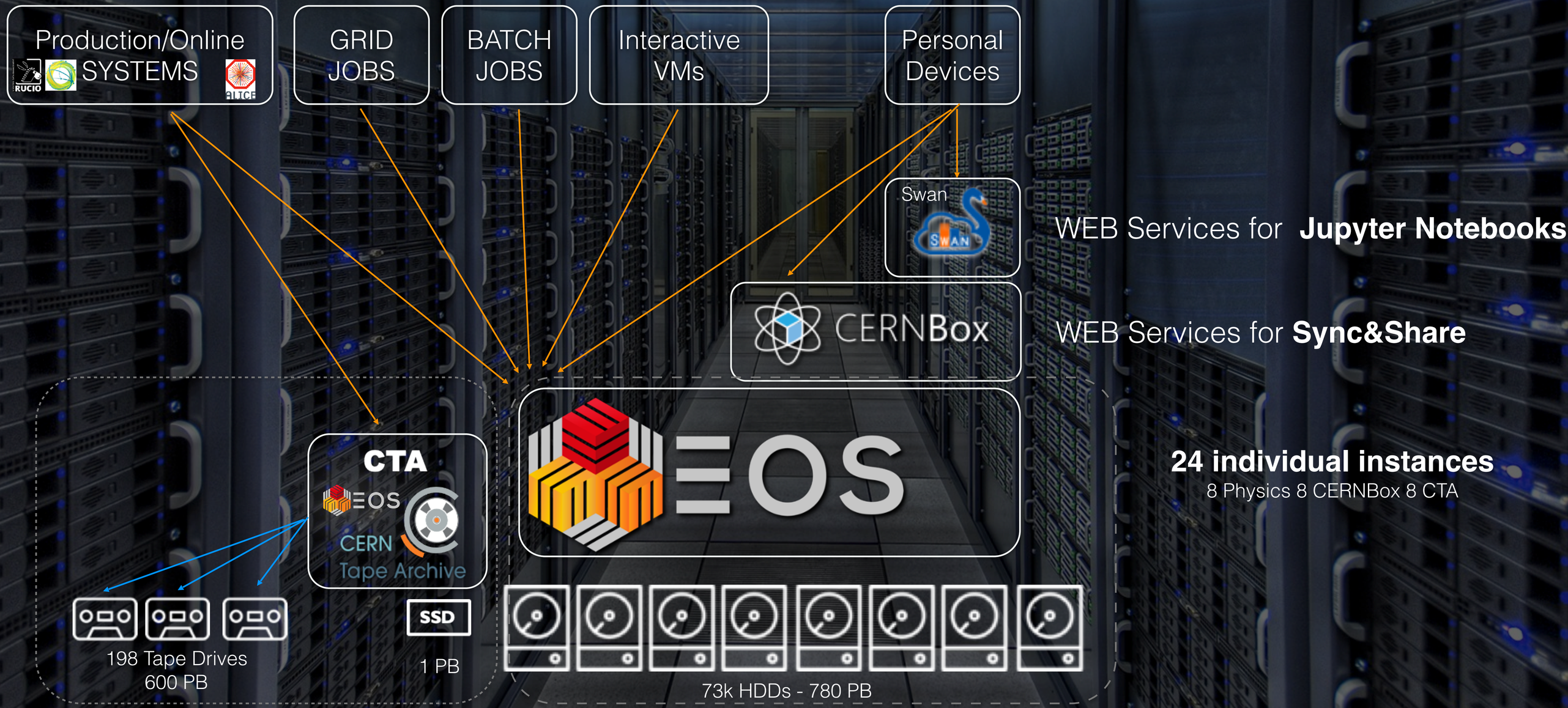


CTA Deployment Model

WLCG dropped HSM model - files in disk buffer in CTA are short lived until they are transferred to tape or to another disk storage system



Storage Services for Physics



WEB Services for **Jupyter Notebooks**

WEB Services for **Sync&Share**

24 individual instances

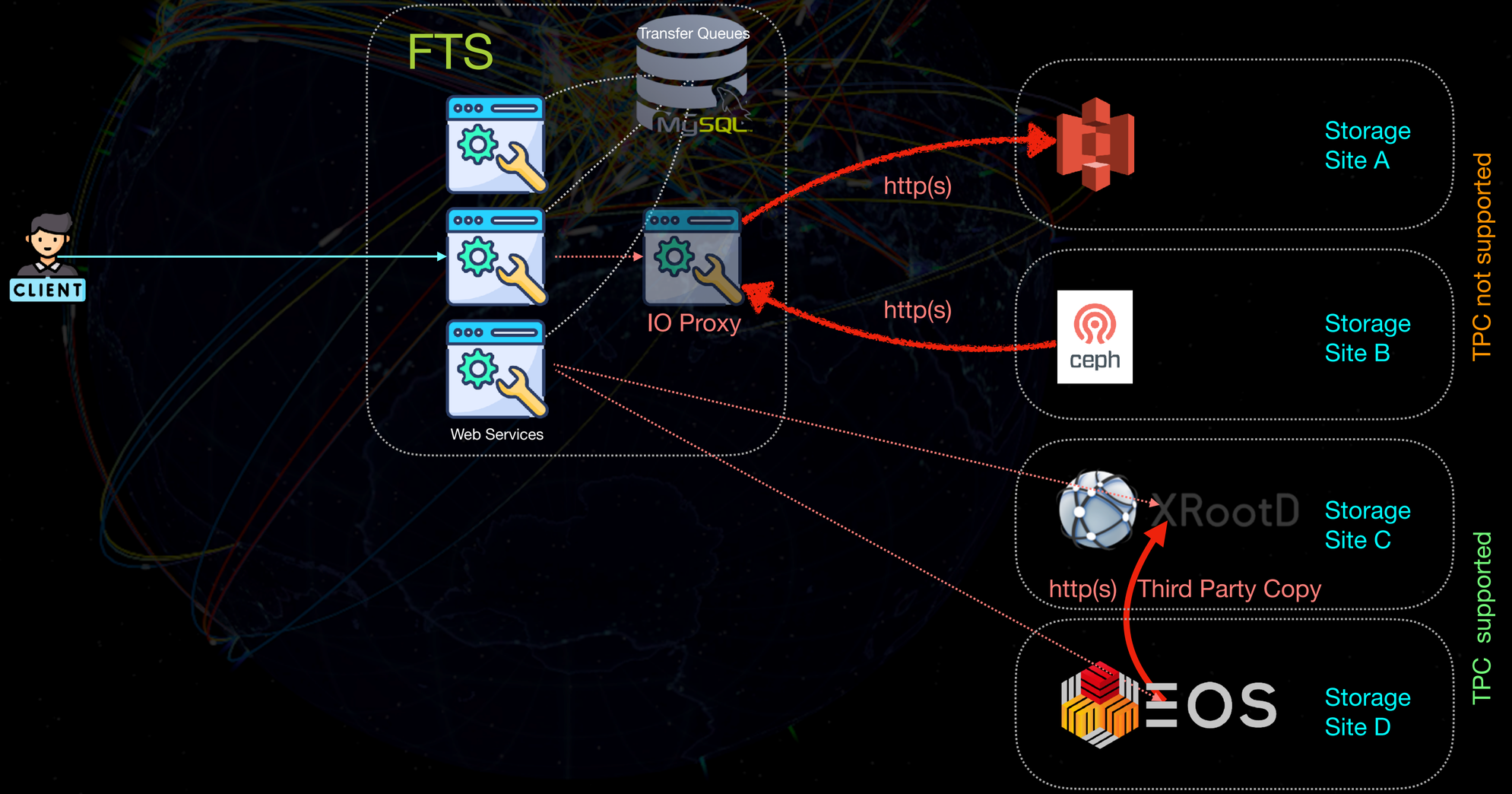
8 Physics 8 CERNBox 8 CTA

File Transfer Service - FTS

Third Party Copy Transfer

FTS

fts.cern.ch

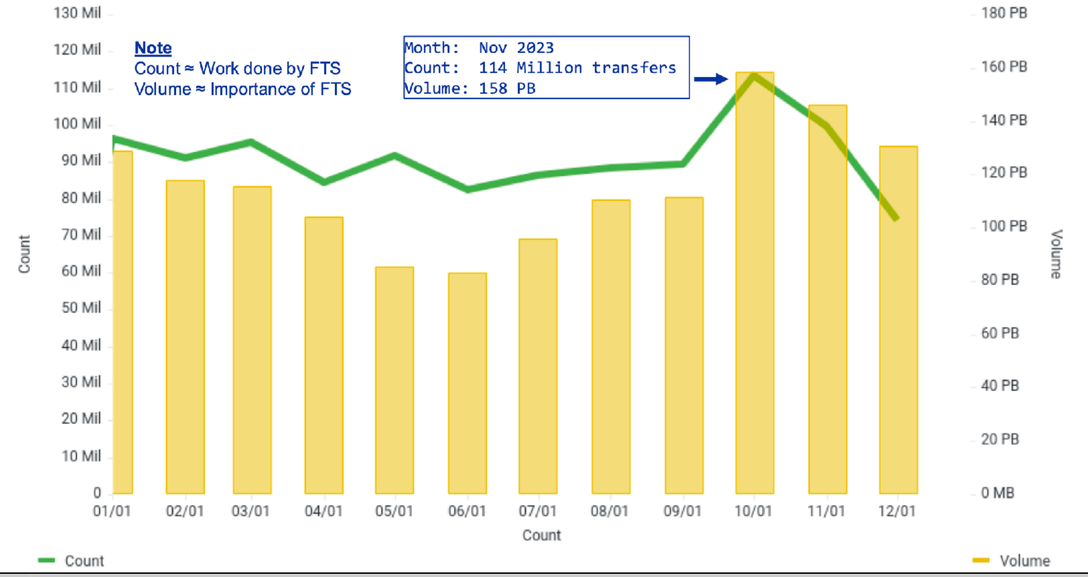


Transfers in WLCG using FTS

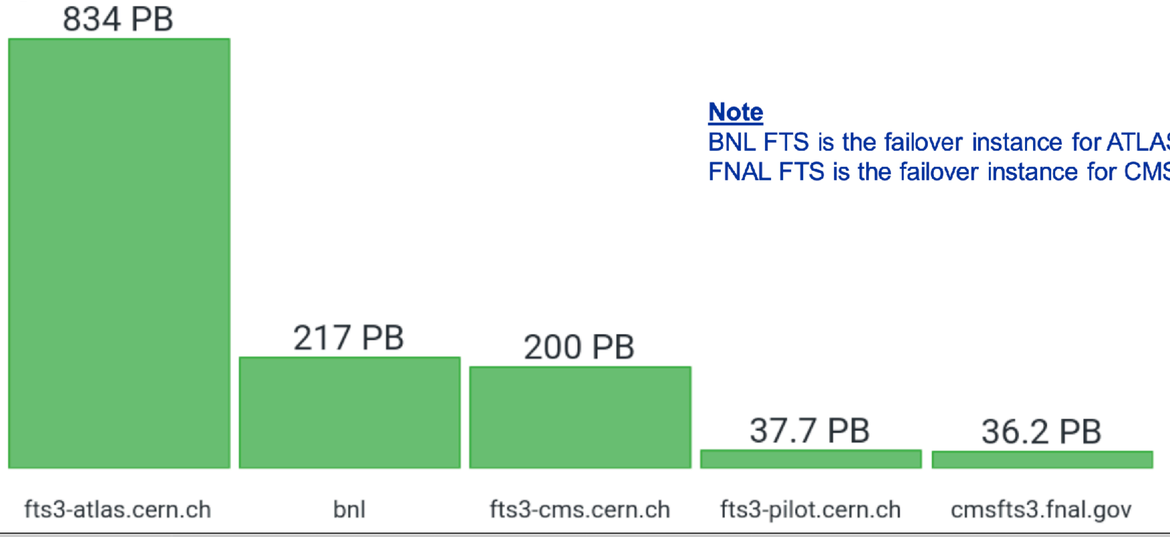
~ 1 Billion transfers in 2022

> 1 EB in 2022

Successful WLCG FTS file transfers per month - 2022



Data volume transferred during 2022
Top 5 WLCG FTS instances

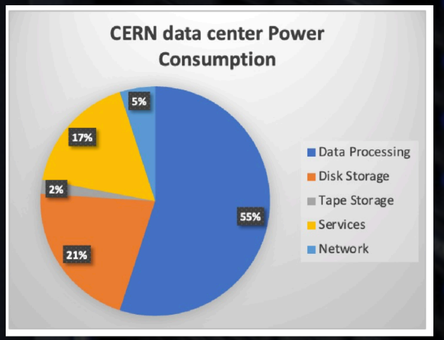


Trends in Storage & Data Management

@CERN and/or WLCG for physics

- Transition from Replication to Erasure Coding
  - benefits: reliability, cost, per file bandwidth - supported in EOS, CephFS, Ceph S3 and others
- Online space is HDDs, Archive is (mostly*) TAPE - FLASH not competitive/needed
  *Korean site has replaced tape storage with erasure coded disk storage with lower cost
  - for LHC Run-4 2030 experimenting with data carousels shuffling data between TAPE and DISK for analysis
- EOS throttles meta-data per user, bandwidth per stream, working on global per user bandwidth real-time regulation
- Moving from X509 to JWT based authentication/authorization schema
- CERN storage group starting evaluation of SMR technology for EOS
- WLCG looking at power & cost of infrastructure
  - 22% for disk storage - 2% for tape storage power consumption at CERN
- Deployment of small storage sites in WLCG as regional caches
  [XCache technology]

CERN 1.25 TWh per year
Computing < 5%



Trends in Storage & Data Management

@CERN and/or WLCG for physics

- HPC/SC used for simulations (CPU), for analysis it is difficult (slow) to get data into these systems
- CephFS replacing NFS Filer services (if not enforced by application support license)
- Introduction of new columnar data format [RNTuple]
  - better performance, smaller data files due to byte transposition and modern compression algorithms like ZSTD
    - interest in QAT ZSTD plug-in
  - columns splitting into low and high-resolution bits - smaller dataset for low resolution datasets
  - possibility of lossy compression (dangerous)
  - evaluation of object storage/S3 for columnar data - possibility to have cold columns on tape
- Users love POSIX filesystem access - we avoid to use it without users knowing it
  - transparent protocol redirection from filesystem access to remote protocol root:// if application supports it
    [EOS is mounted by 30k FUSE clients at CERN but also from US sites …]
  - considering bulk open API [vector open] for analysis uses cases
- CERN considers to decommission AFS as home directory service (2026+)
  - all evaluated file systems ran into similar problems (meta-data or data DOS) - no free alternative to support 30k+ mount clients from large batch/kubernetes cluster - possible future: only interactive trusted nodes with home directories on shared filesystem

Summary

- **CERN** is operating successfully a **large exabyte storage infrastructure**, which is optimised for physics use cases and cost minimisation
  - **infrastructure** is not targeting few high-performance clients but optimised for the maximum throughput for tenthousands of concurrent clients with defined priorities and at minimal cost
  - **price** for storage at CERN is still competitive to hyper scaler cloud offerings
    - that might change - physics frameworks have already integrated cloud compute and storage APIs to include transparently temporary cloud resources
    - main challenge for the next LHC Run-4 in 2030 is affordability of computing & storage - less a performance issue: ~~TB/s~~ but TB/$ and TB/s/$
- **CERN/WLCG** is/was steering or contributing to Open Source Storage projects AFS, EOS, CTA ,FTS, RUCIO, XRootD, CVMFS, CEPH which enabled physics discoveries in high energy physics during the last decade


Thanks for your attention!
SOCAR
VESTAR
CHANGE