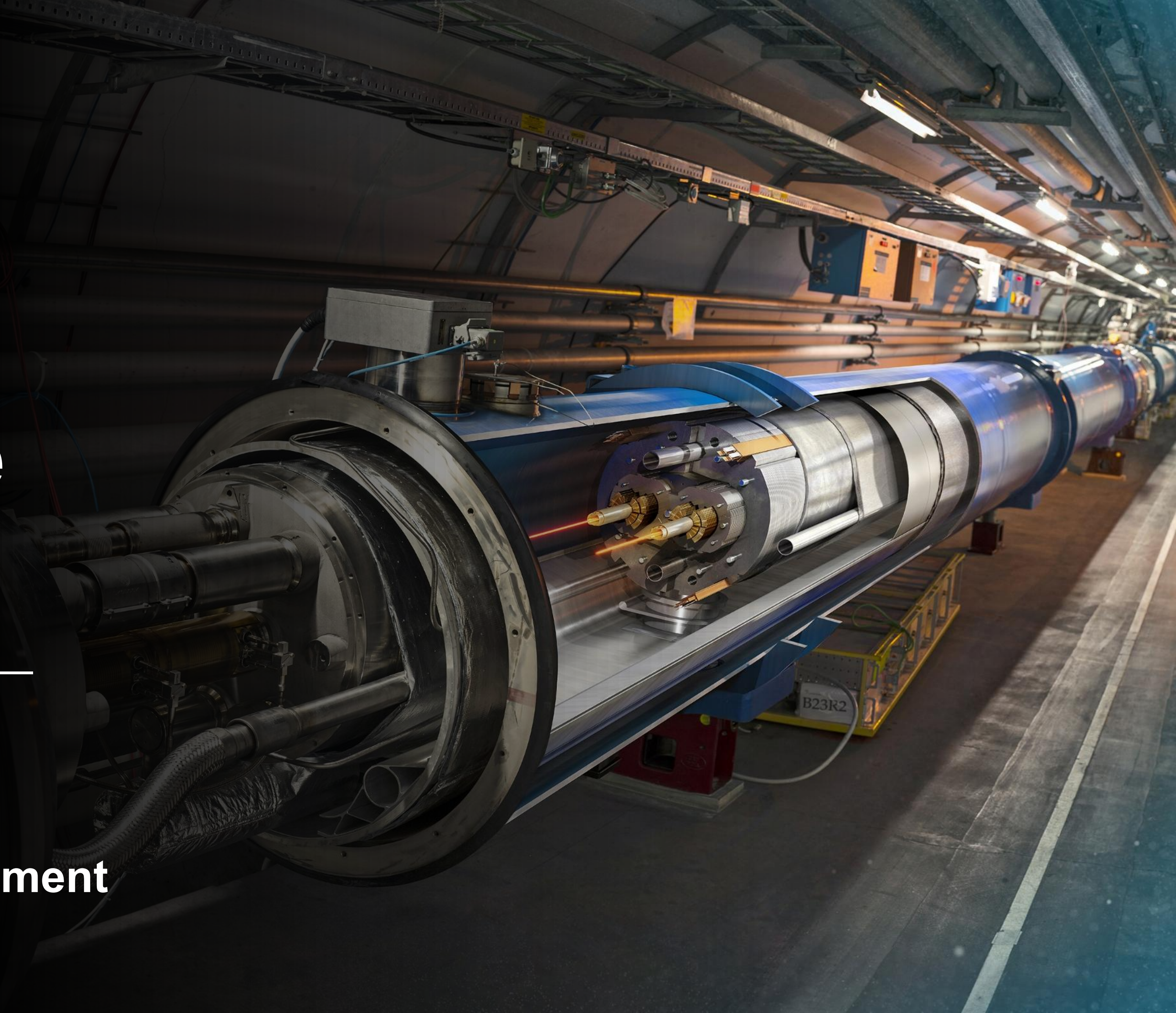
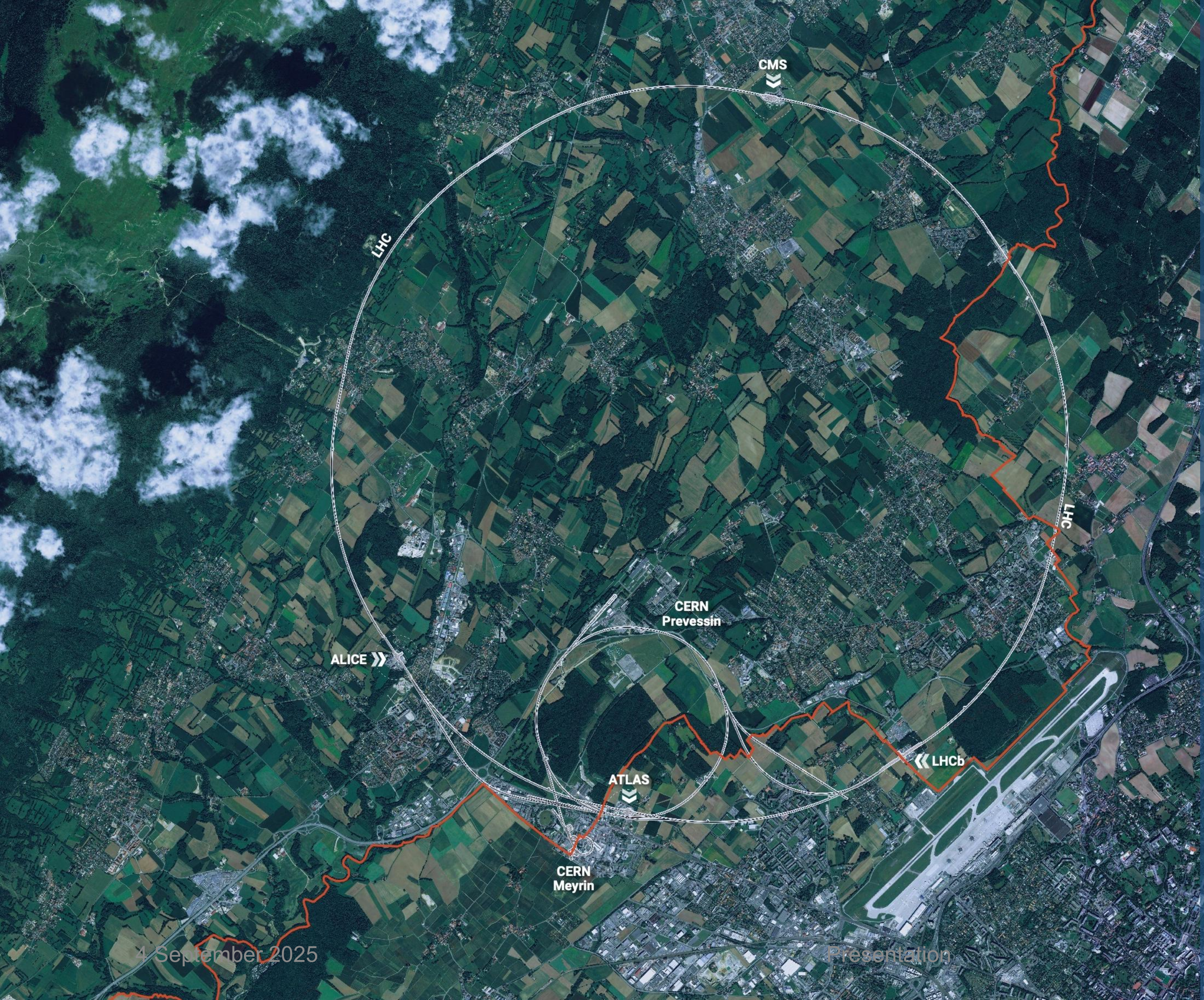
CERN

Accélérateur de science

Challenges of the Data Storage at CERN IT

Vladimír Bahyl
CERN IT Department
Storage and Data Management





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

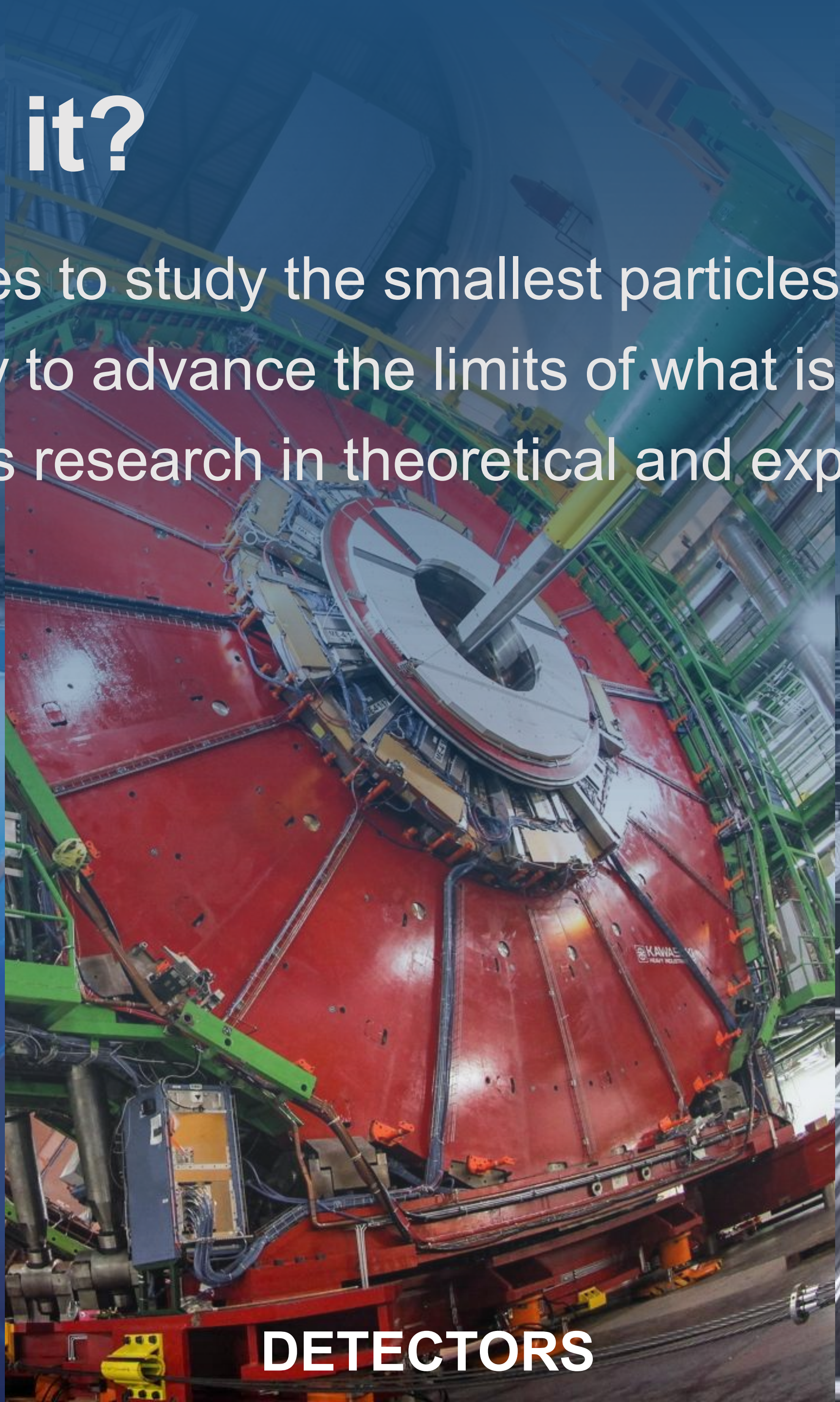
Our goal is to understand the most fundamental particles and laws of the universe.

Located near Geneva on either side of the Swiss French border

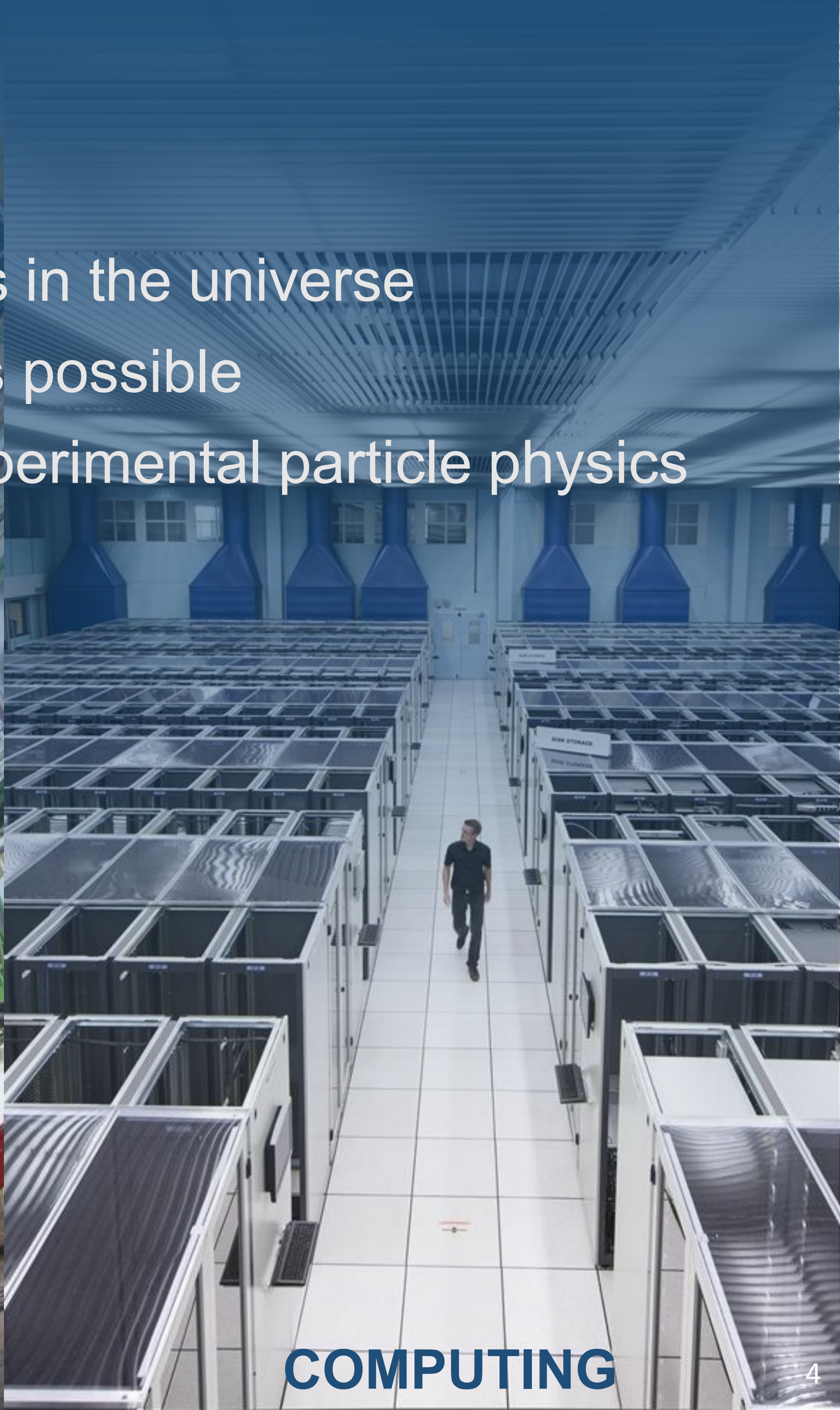
How do we do it?

- We build large 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

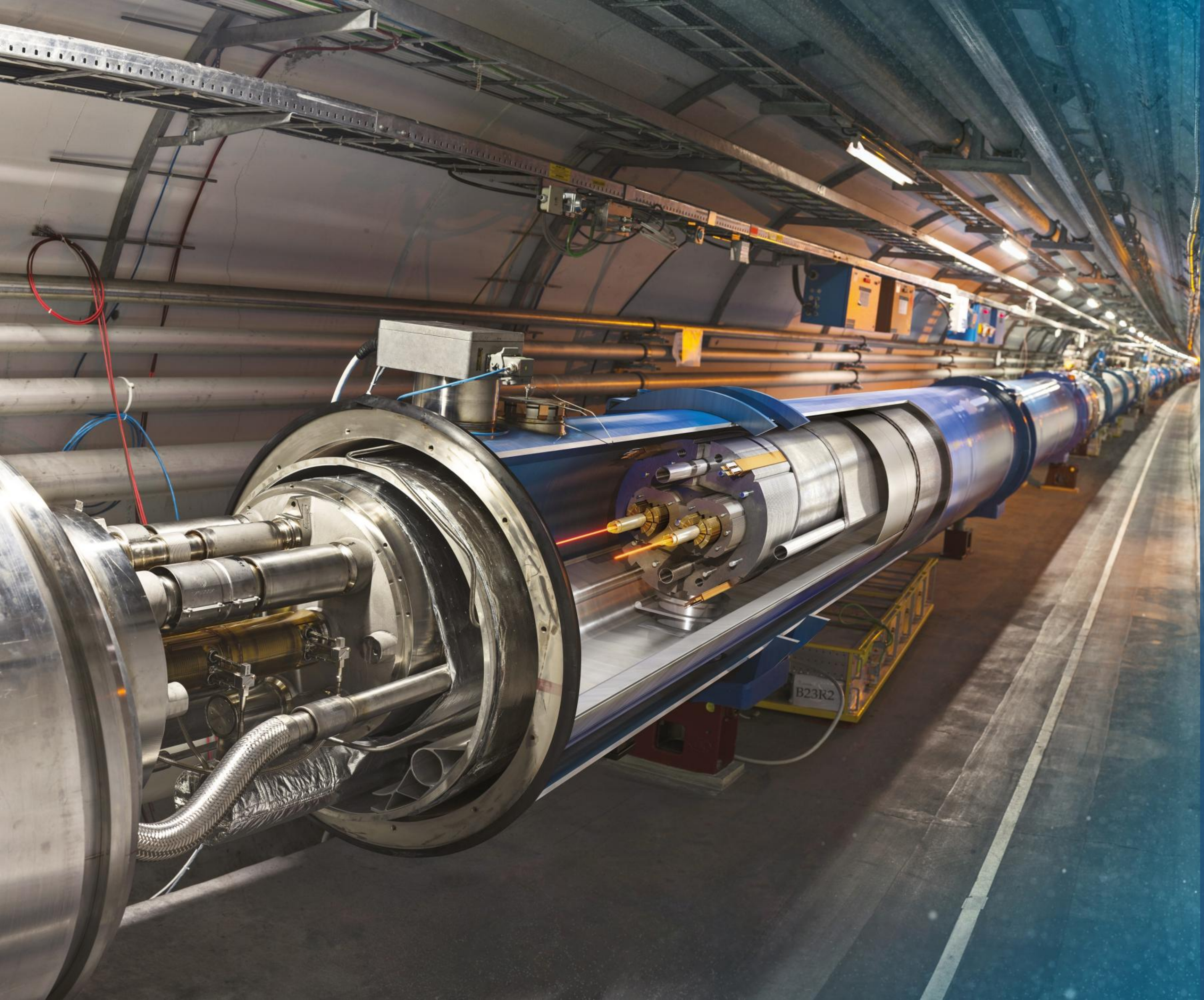
ACCELERATORS



DETECTORS



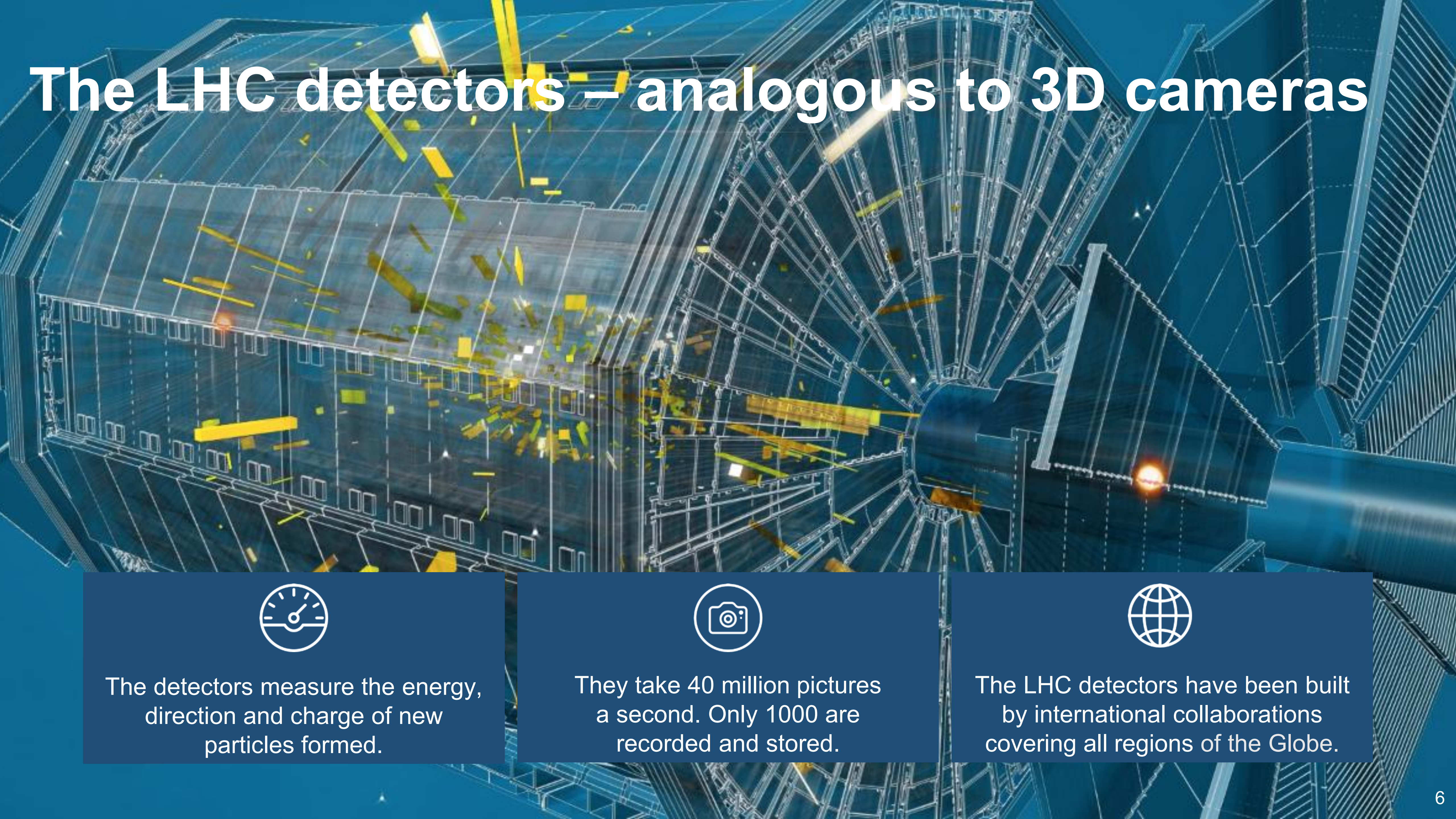
COMPUTING



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

The LHC detectors – analogous to 3D cameras

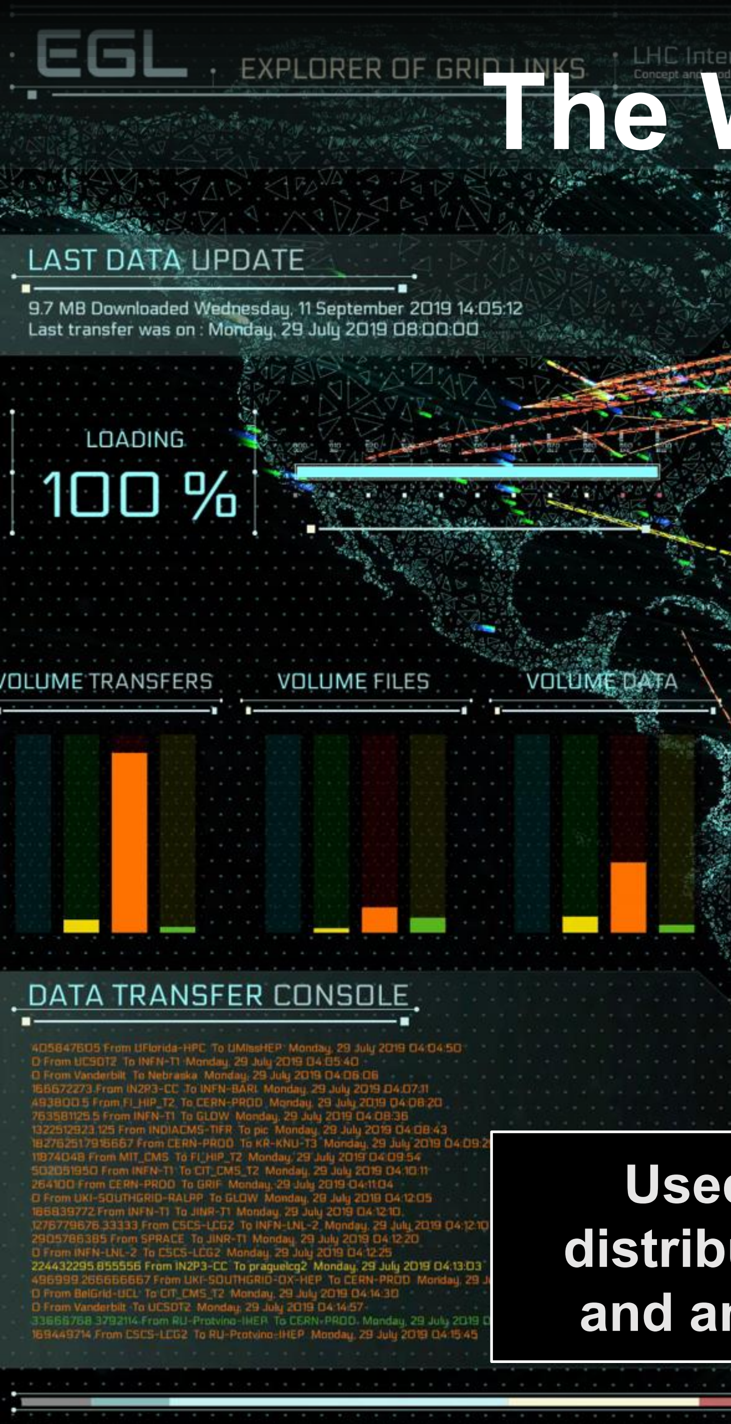


The detectors measure the energy, direction and charge of new particles formed.

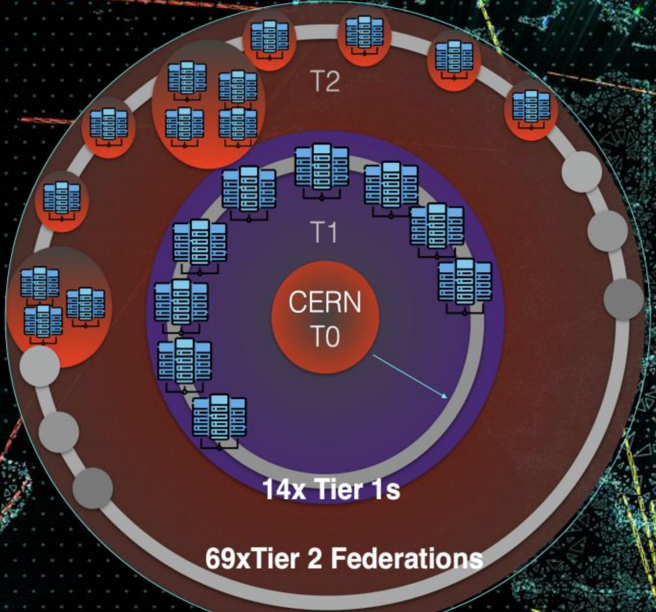
They take 40 million pictures a second. Only 1000 are recorded and stored.

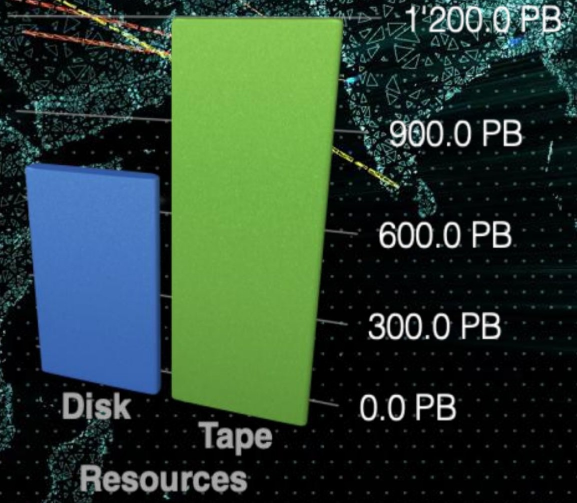
The LHC detectors have been built by international collaborations covering all regions of the Globe.

The Worldwide LHC Computing Grid (WLCG)



```
EGL  EXPLORER OF GRID LINKS   LHC Interactive Tunnel                    COMMA

LAST DATA UPDATE
9.7 MB Downloaded Wednesday, 11 September 2019 14:05:12
Last transfer was on : Monday, 29 July 2019 08:00:00

LOADING
100 %

VOLUME TRANSFERS   VOLUME FILES   VOLUME DATA

DATA TRANSFER CONSOLE
405847605 From UFlorida-HPC  To UMissHEP  Monday, 29 July 2019 04:04:50
0 From UCSDT2  To INFN-T1  Monday, 29 July 2019 04:05:40
0 From Vanderbilt  To Nebraska  Monday, 29 July 2019 04:06:06
166672273 From IN2P3-CC  To INFN-BARI  Monday, 29 July 2019 04:07:11
493800.5 From FI_HIP_T2  To CERN-PROD  Monday, 29 July 2019 04:08:20
763581125.5 From INFN-T1  To GLOW  Monday, 29 July 2019 04:08:36
1322512923.125 From INDIACMS-TIFR  To pic  Monday, 29 July 2019 04:08:43
1827625179.16667 From CERN-PROD  To KR-KNU-T3  Monday, 29 July 2019 04:09:2
11874048 From MIT_CMS  To FI_HIP_T2  Monday, 29 July 2019 04:09:54
502051950 From INFN-T1  To CIT_CMS_T2  Monday, 29 July 2019 04:10:11
264100 From CERN-PROD  To GRIF  Monday, 29 July 2019 04:11:04
0 From UKI-SOUTHGRID-RALPP  To GLOW  Monday, 29 July 2019 04:12:05
186839772 From INFN-T1  To JINR-T1  Monday, 29 July 2019 04:12:10
1276779676.33333 From CSCS-LCG2  To INFN-LNL-2  Monday, 29 July 2019 04:12:10
2905786385 From SPRACE  To JINR-T1  Monday, 29 July 2019 04:12:20
0 From INFN-LNL-2  To CSCS-LCG2  Monday, 29 July 2019 04:12:25
224432295.855556 From IN2P3-CC  To praguelcg2  Monday, 29 July 2019 04:13:03
496999.266666667 From UKI-SOUTHGRID-OX-HEP  To CERN-PROD  Monday, 29 J
0 From BelGrid-UCL  To CIT_CMS_T2  Monday, 29 July 2019 04:14:30
0 From Vanderbilt  To UCSDT2  Monday, 29 July 2019 04:14:57
33666768.3792114 From RU-Protvino-IHEP  To CERN-PROD  Monday, 29 July 2019 0
169449714 From CSCS-LCG2  To RU-Protvino-IHEP  Monday, 29 July 2019 04:15:45
```





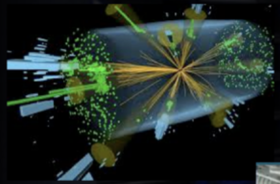
Used to store, distribute, process and analyse data.

1 million processing cores in about 120 data centres and 42 countries.

More than 1000 Petabytes of CERN data stored world-wide.

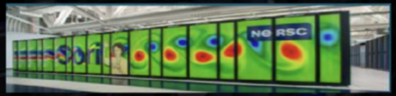
Storage in High Energy Physics

Archival & Backup Storage



Storage for Data Acquisition

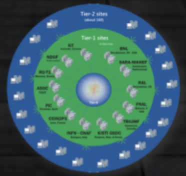
Storage for Home Directories



Storage for HPC

Storage for Applications

Private Cloud Storage



Storage for GRID Computing

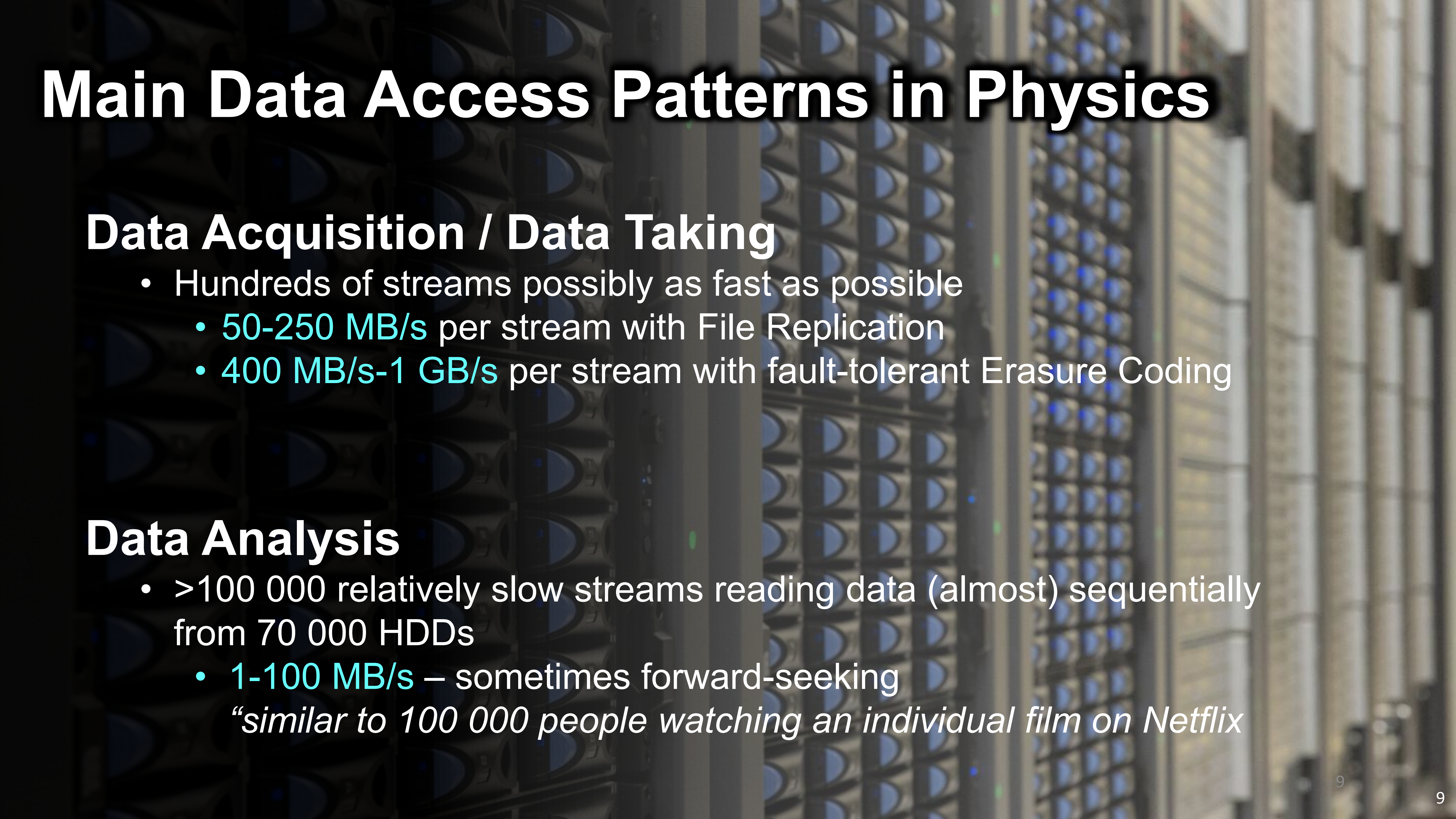
Public Cloud Storage

Storage for Software Distribution

Storage for Data Analytics

Storage for Physics Analysis

Storage for Sync&Share



Main Data Access Patterns in Physics

Data Acquisition / Data Taking

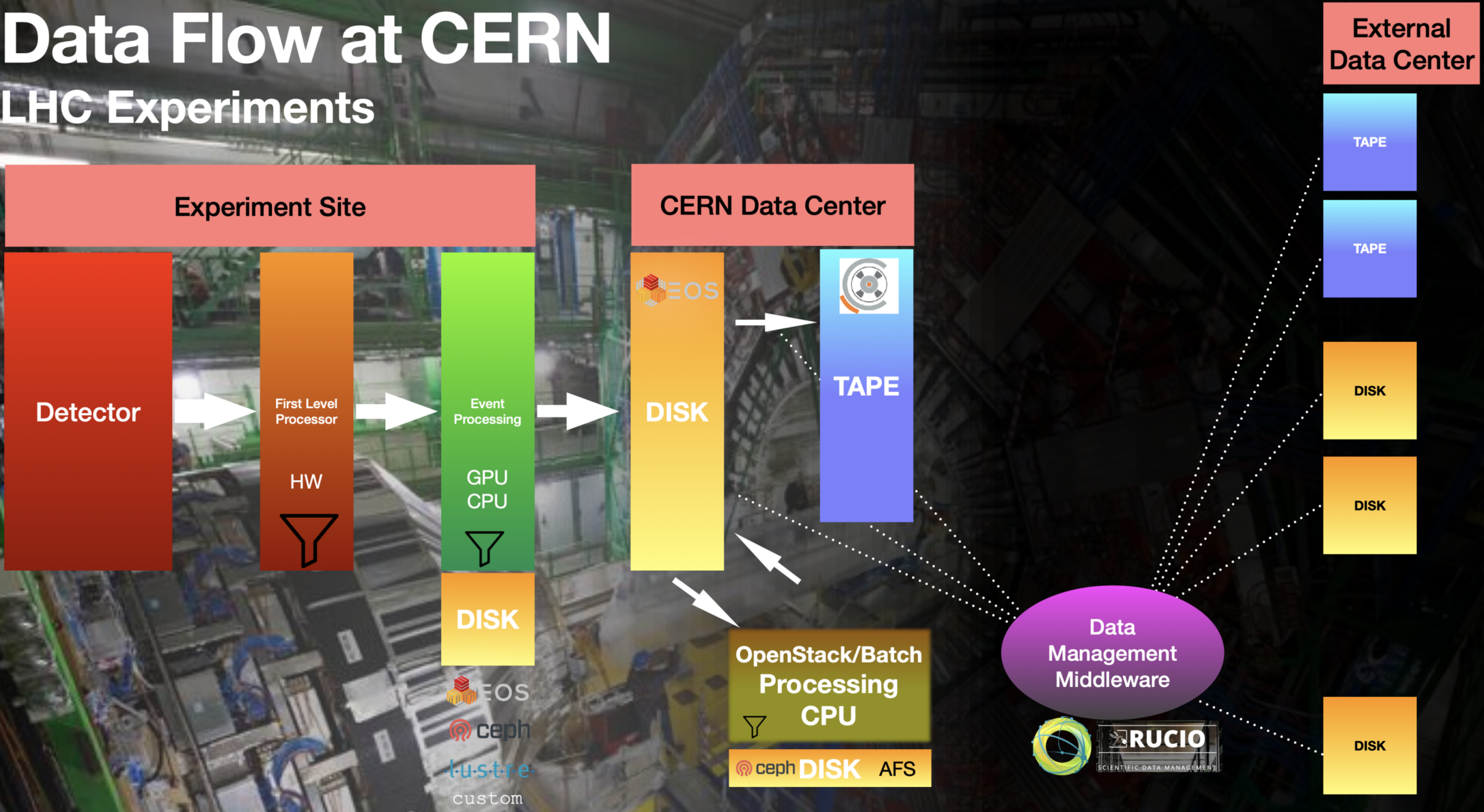
- Hundreds of streams possibly as fast as possible
  - 50-250 MB/s per stream with File Replication
  - 400 MB/s-1 GB/s per stream with fault-tolerant Erasure Coding

Data Analysis

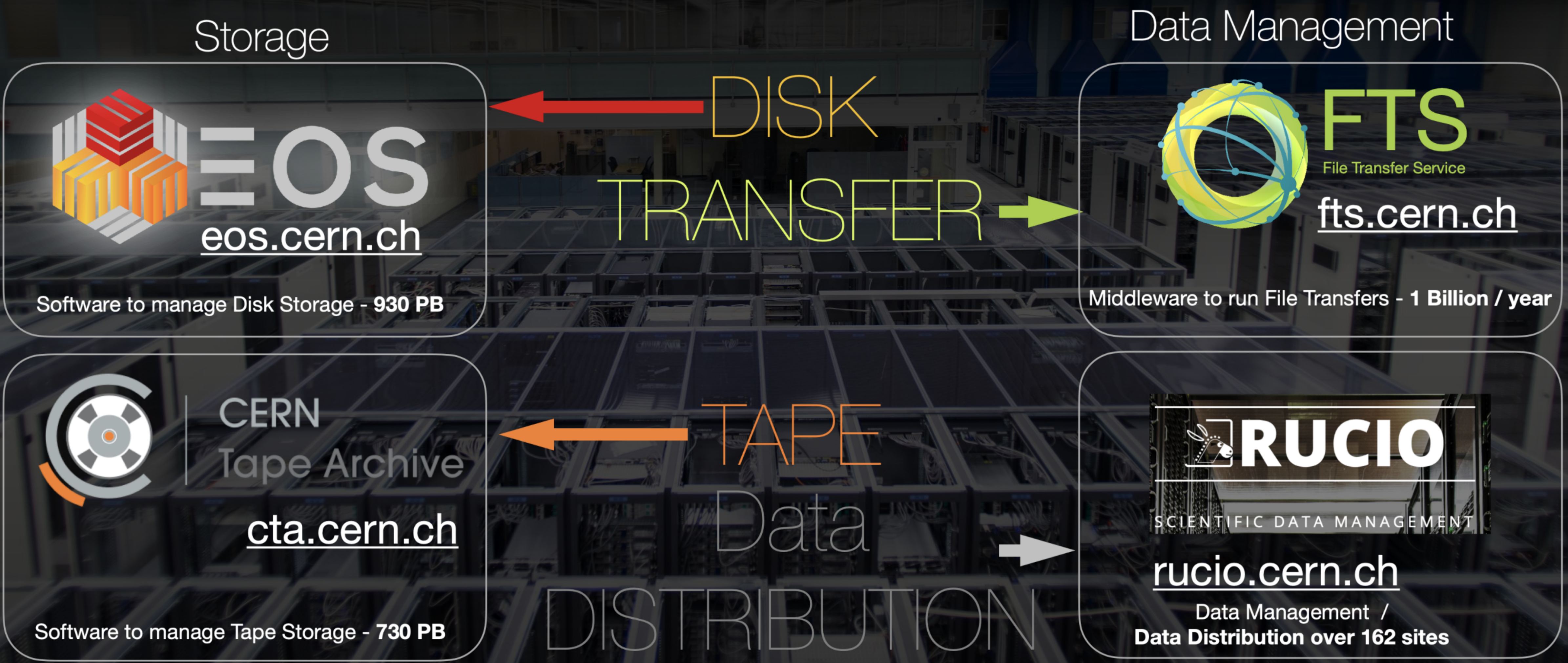
- >100 000 relatively slow streams reading data (almost) sequentially from 70 000 HDDs
  - 1-100 MB/s – sometimes forward-seeking
    “similar to 100 000 people watching an individual film on Netflix

Data Flow at CERN

LHC Experiments



Physics Storage and Data Management Services



Storage

eos.cern.ch

Software to manage Disk Storage - **930 PB**

cta.cern.ch

Software to manage Tape Storage - **730 PB**

Data Management

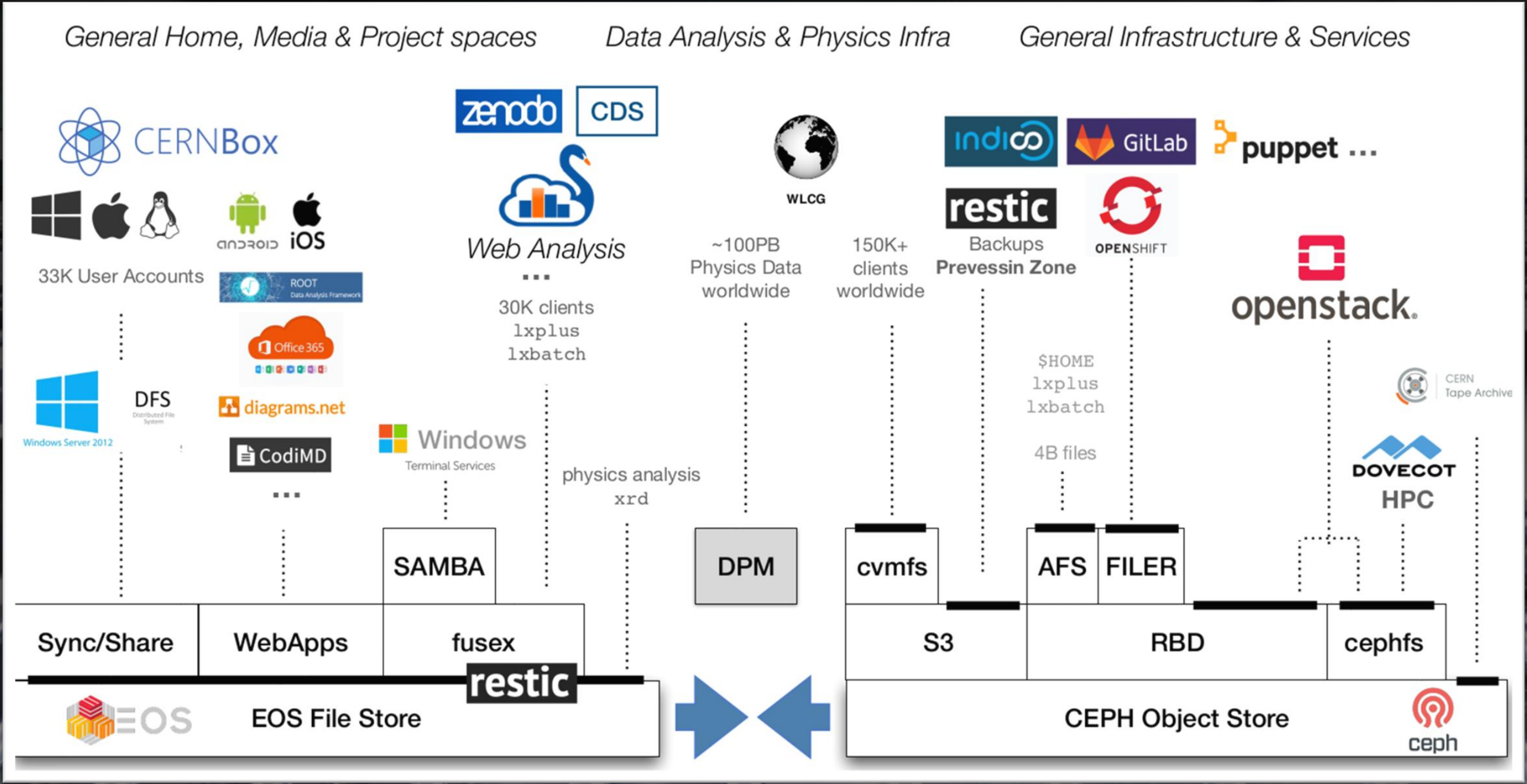
fts.cern.ch

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

rucio.cern.ch

Data Management /
Data Distribution over 162 sites

General Storage Services



General Home, Media & Project spaces

Data Analysis & Physics Infra

General Infrastructure & Services

CERNBox

33K User Accounts

android iOS

ROOT Data Analysis Framework

Office 365

diagrams.net

CodiMD

...

Windows Server 2012

DFS Distributed File System

zenodo CDS

Web Analysis

...

30K clients
lxplus
lxbatch

Windows Terminal Services

physics analysis
xrd

WLCG

~100PB Physics Data worldwide

150K+ clients worldwide

indico GitLab puppet ...

restic
Backups
Prevessin Zone

OPENSHIFT

openstack.

$HOME
lxplus
lxbatch

4B files

CERN Tape Archive

DOVECOT
HPC

SAMBA

DPM

cvmfs

AFS

FILER

Sync/Share

WebApps

fusex

S3

RBD

cephfs

restic

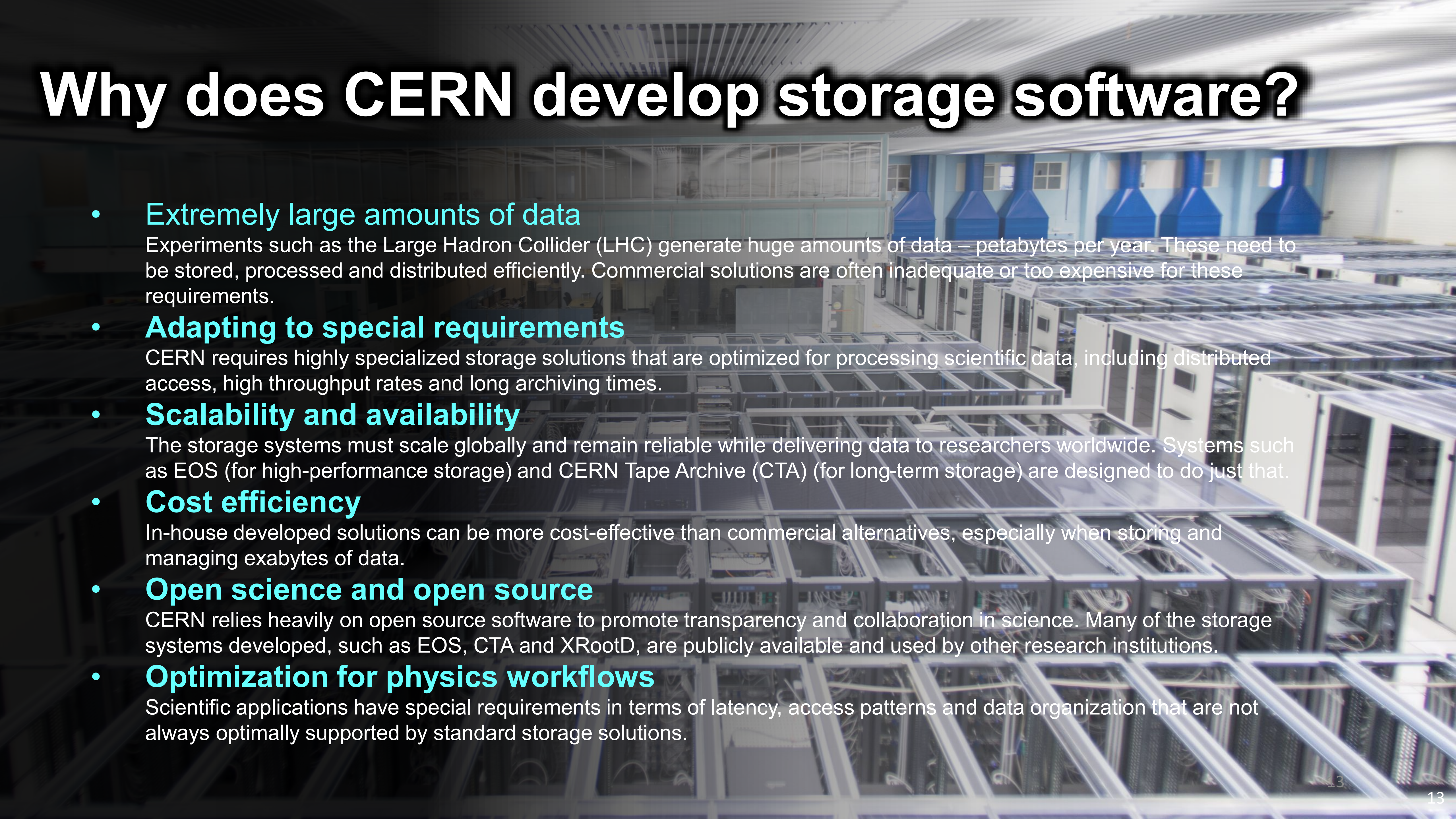
EOS

EOS File Store

CEPH Object Store

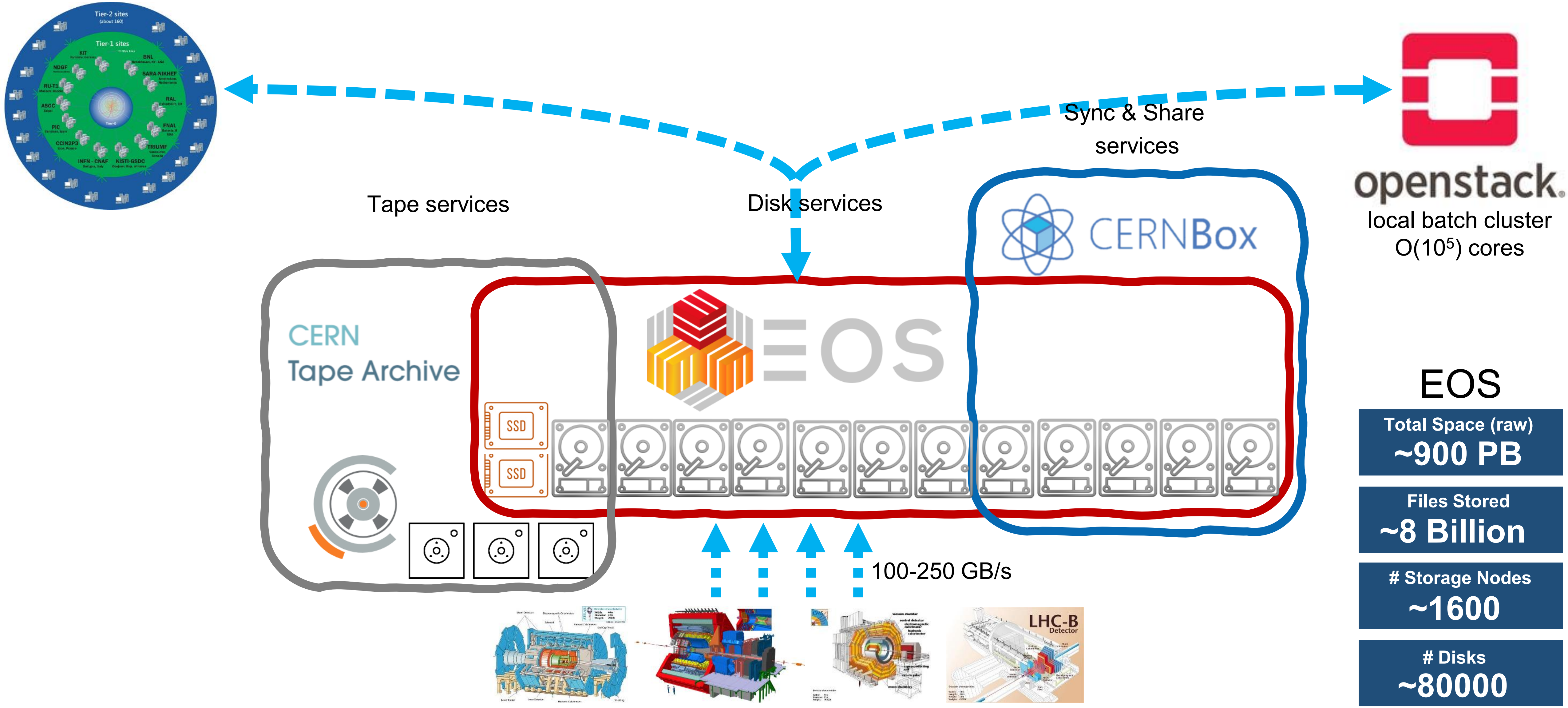
ceph

Why does CERN develop storage software?



- Extremely large amounts of data
  Experiments such as the Large Hadron Collider (LHC) generate huge amounts of data – petabytes per year. These need to be stored, processed and distributed efficiently. Commercial solutions are often inadequate or too expensive for these requirements.
- **Adapting to special requirements**
  CERN requires highly specialized storage solutions that are optimized for processing scientific data, including distributed access, high throughput rates and long archiving times.
- **Scalability and availability**
  The storage systems must scale globally and remain reliable while delivering data to researchers worldwide. Systems such as EOS (for high-performance storage) and CERN Tape Archive (CTA) (for long-term storage) are designed to do just that.
- **Cost efficiency**
  In-house developed solutions can be more cost-effective than commercial alternatives, especially when storing and managing exabytes of data.
- **Open science and open source**
  CERN relies heavily on open source software to promote transparency and collaboration in science. Many of the storage systems developed, such as EOS, CTA and XRootD, are publicly available and used by other research institutions.
- **Optimization for physics workflows**
  Scientific applications have special requirements in terms of latency, access patterns and data organization that are not always optimally supported by standard storage solutions.

CERN IT Data Storage Services



Sync & Share services

Tape services

Disk services

local batch cluster
O(10^5) cores

100-250 GB/s

EOS

Total Space (raw)
~900 PB

Files Stored
~8 Billion

# Storage Nodes
~1600

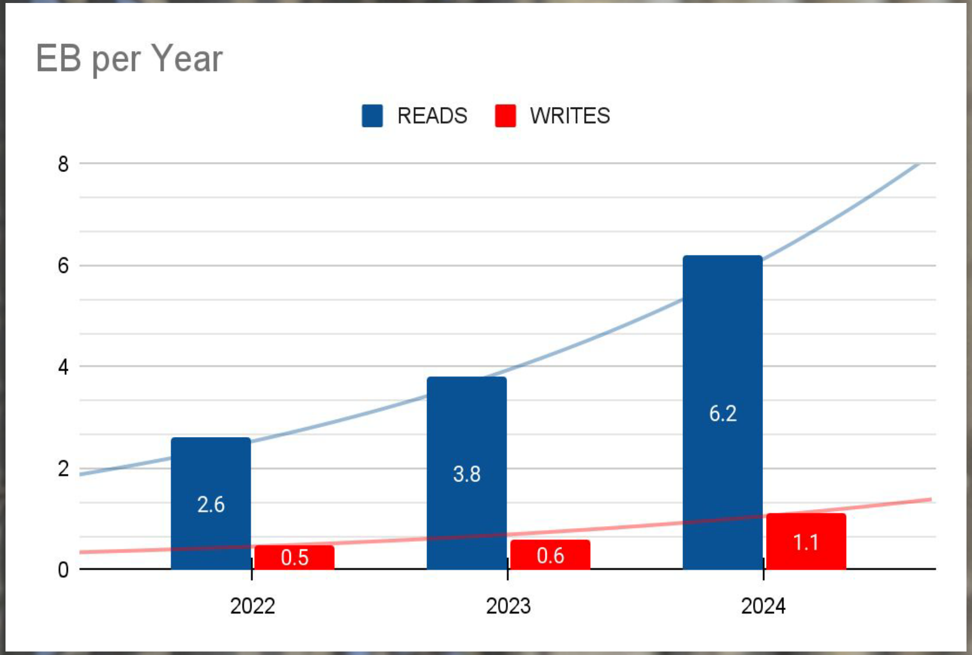
# Disks
~80000

EOS Open Storage – CERN Disk Storage

- **What is EOS?**
  → Highly scalable distributed storage system for large amounts of data at CERN
- **Development & use:**
  → Developed at CERN for high-energy physics experiments (e.g. LHC)
  → Optimized for high performance & low latency
- **Features:**
  - Software-defined storage with POSIX-like access
  - High scalability for petabyte to exabyte data volumes
  - Replication & erasure coding for data security
  - Low latency, optimized for many parallel I/O operations
  - Integrated with **CERNBox** ("cloud storage for physicists")
  - Cost-efficient
- **Areas of application:**
  → Data storage for LHC experiments - over 20 instances at CERN
  → Storage of scientific analysis data
  → Research & development in distributed storage (WLCG) - dozens of deployments worldwide

- **Technical details:**
  - Open-source software
  - Supports multiple backends (hard drives, SSDs, tape drives)
  - Interfaces for Linux (FUSE), HTTP, DAV, CIFS

Usage trend – INGRES & EGRES



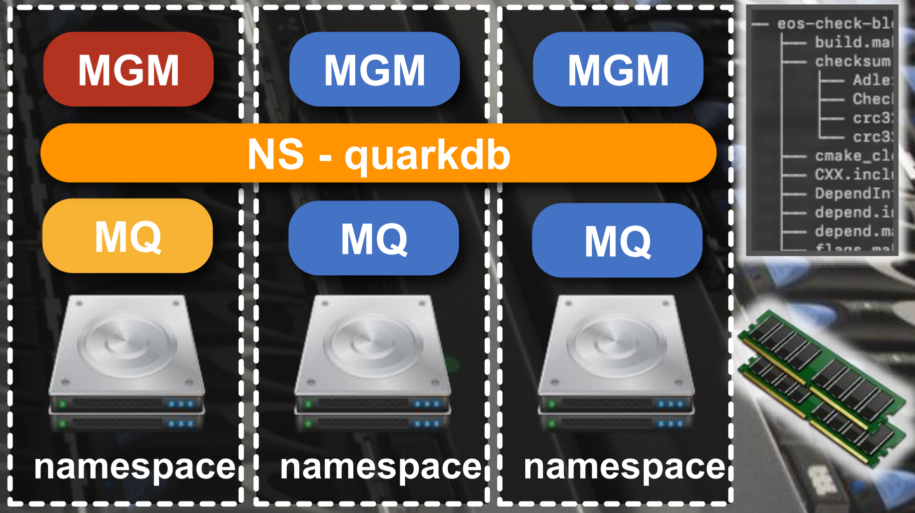
EOS Architecture

Highly-available and low latency namespace:

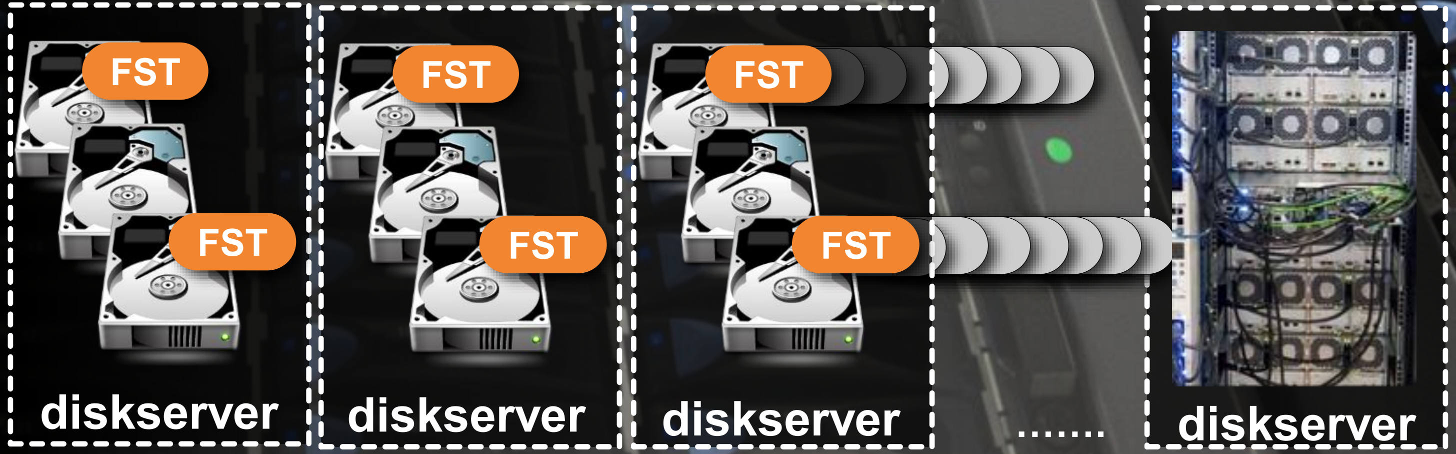
- Namespace persisted on a distributed key-value store
- Working entries cached in-memory

Highly-available and reliable file storage, based on (unexpensive) JBODs:

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

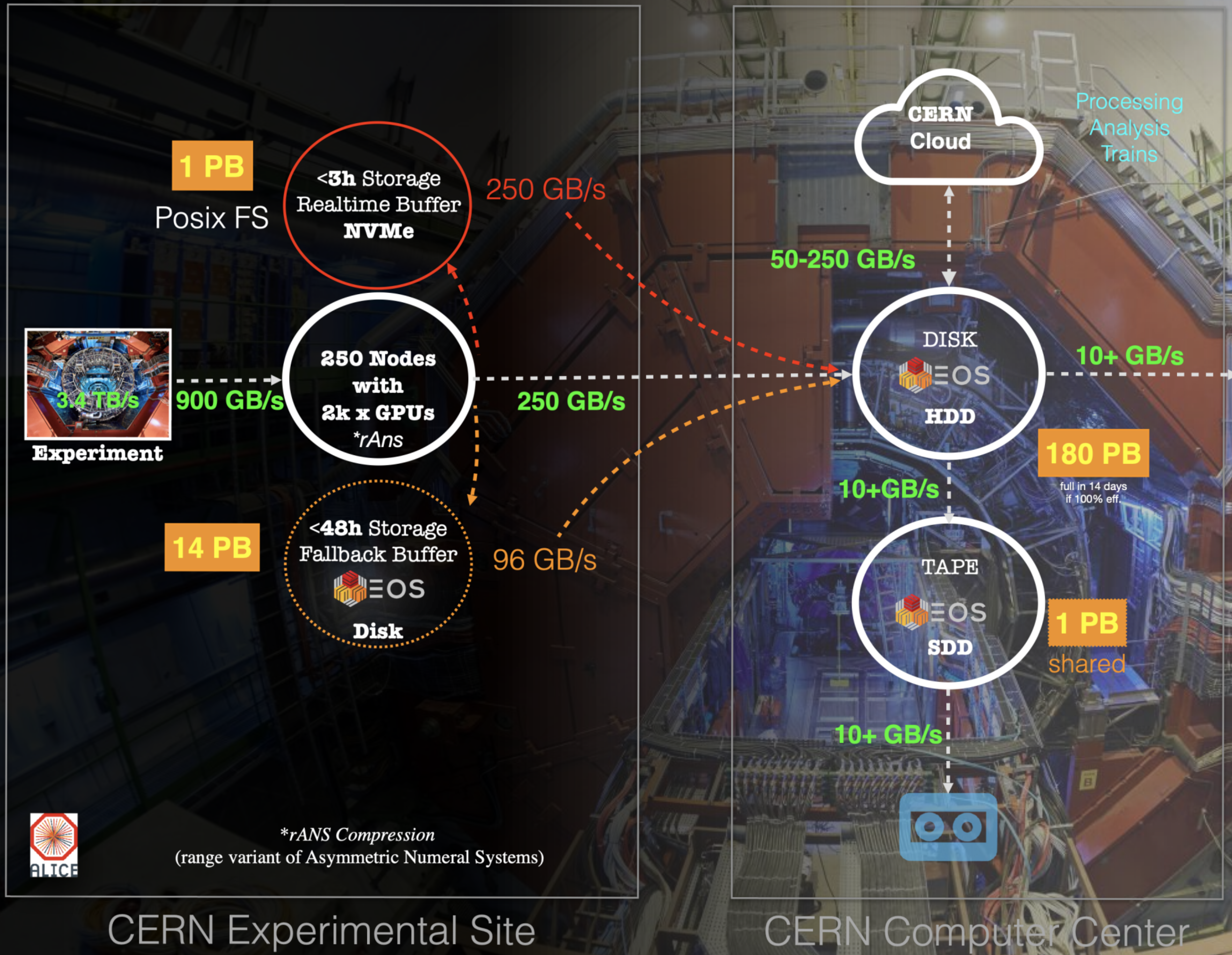


MGM : meta data server
MQ : message queue
NS : persistent namespace
FST : file storage server

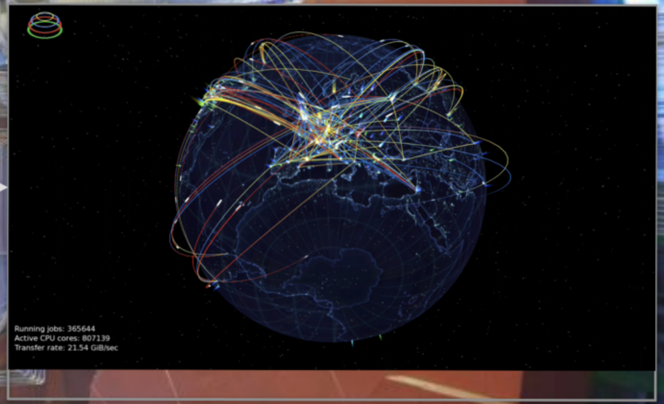


Dataflow & Storage
ALICE LHC Experiment

O2

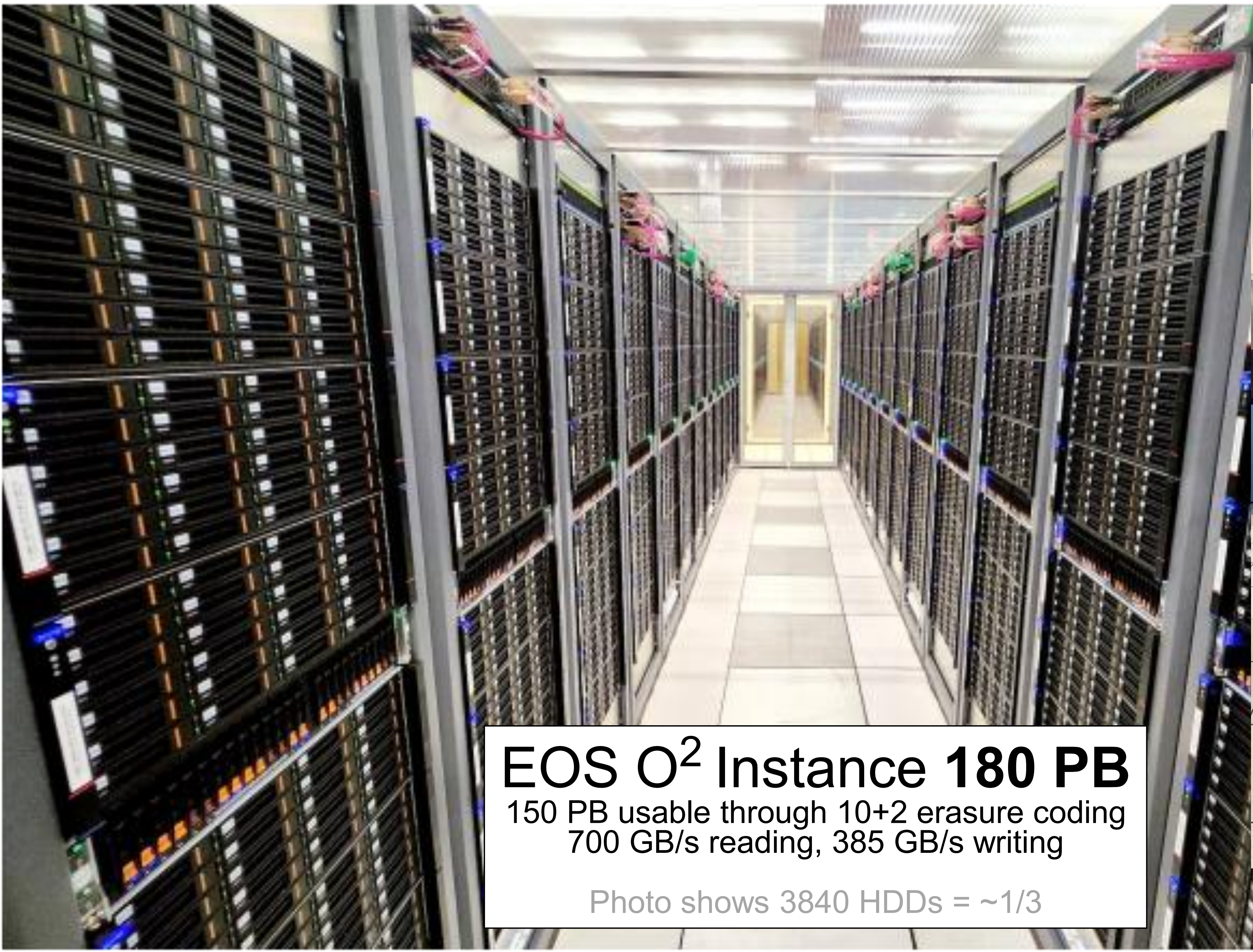


*rANS Compression
(range variant of Asymmetric Numeral Systems)



Running jobs: 365644
Active CPU cores: 807139
Transfer rate: 21.54 GiB/sec

Worldwide LHC
Computing GRID

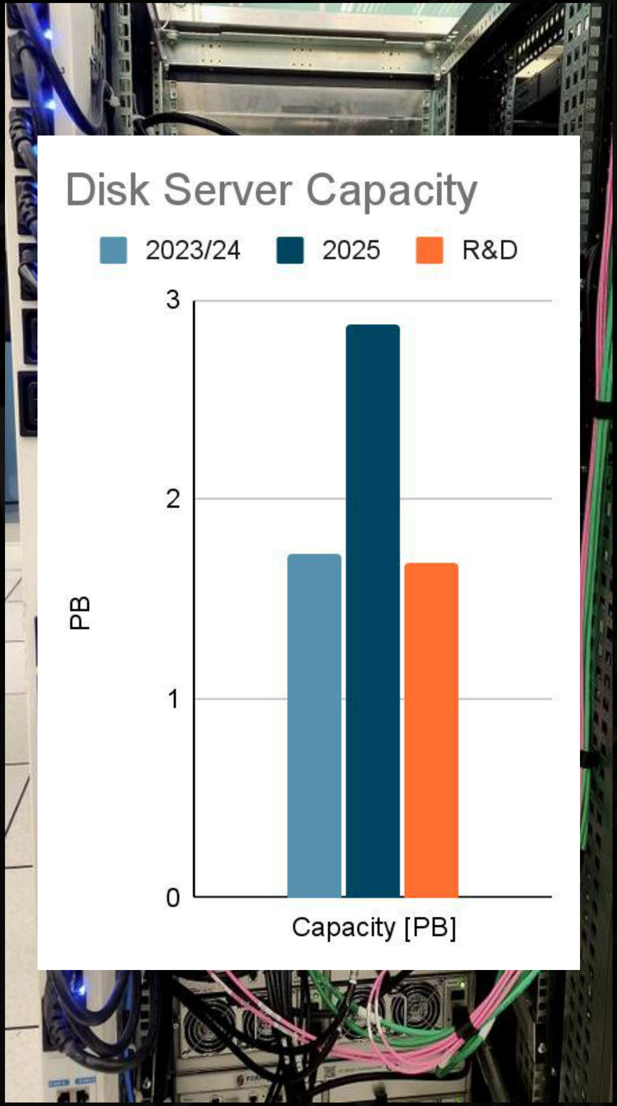


EOS O^2 Instance **180 PB**

150 PB usable through 10+2 erasure coding
700 GB/s reading, 385 GB/s writing

Photo shows 3840 HDDs = ~1/3

Disk Storage challenges and evolution



Points of Concern

- HDDs size growing
  - 50-100TB by 2030
- Performance-Capacity Ratio is going down
- *#streams* per HDD expected to go up
  - Reduces HDD bw
  - EC increases **#streams** by up to 10x
- Fewer servers needed to provide capacity
  - Need to increase network connectivity per server
  - Need to reduce number of disks per front-end server

Server Evolution

~~2022~~ → 2025 → R&D

~~100~~ → 100 →**200/400GE**

~~18~~ → 22 → **28 TB HDDs**

~~96~~ → 120 → **60 HDDs/node**

~~Rep(2)~~ →mix→ **EC(10,2)++**

~~Quad~~ →Quad→ **Pizza Box**

R&D

Platform

Arm/Intel/AMD

CMR/SMR/HAMR

NVME/ Low-cost Flash

Tiering

Hybrid Flash/HDD

Hybrid Disk/Tape

Tape Infrastructure

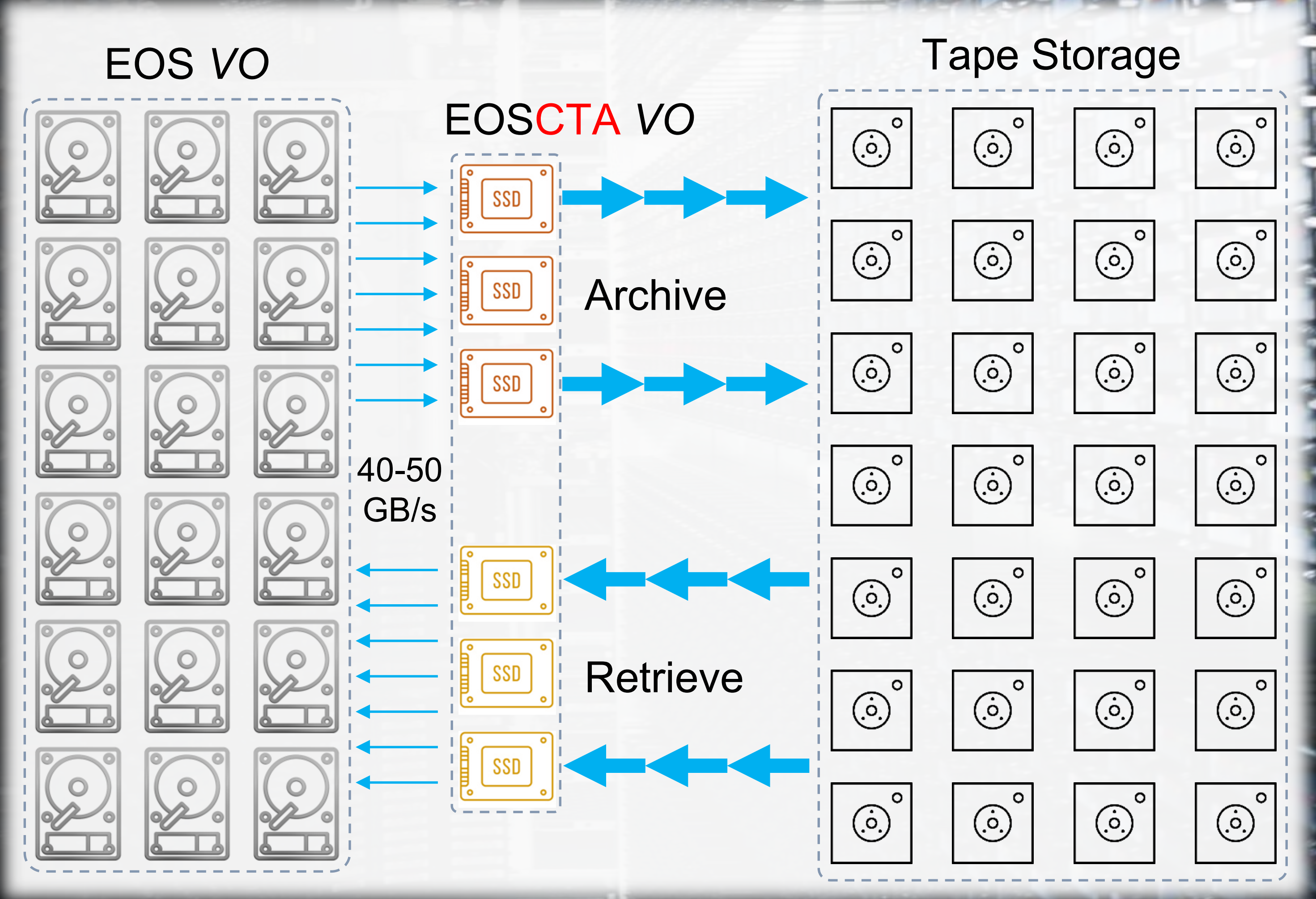
CTA (CERN Tape Archive) Architecture

EOS is natively used as a namespace and disk pool manager

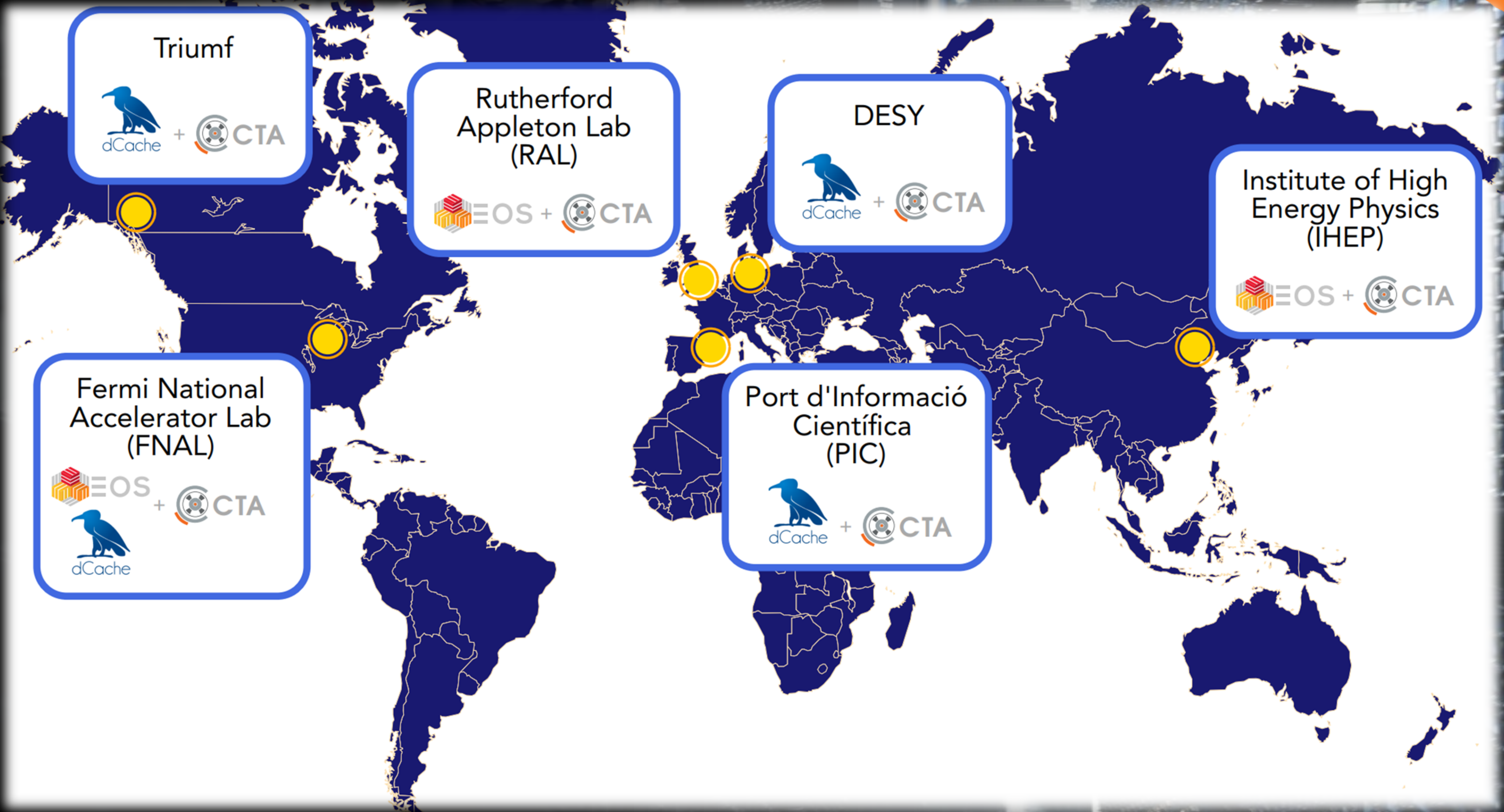
A pure SSD EOS instance with tape backend

Conceived as a fast buffer to the tape system

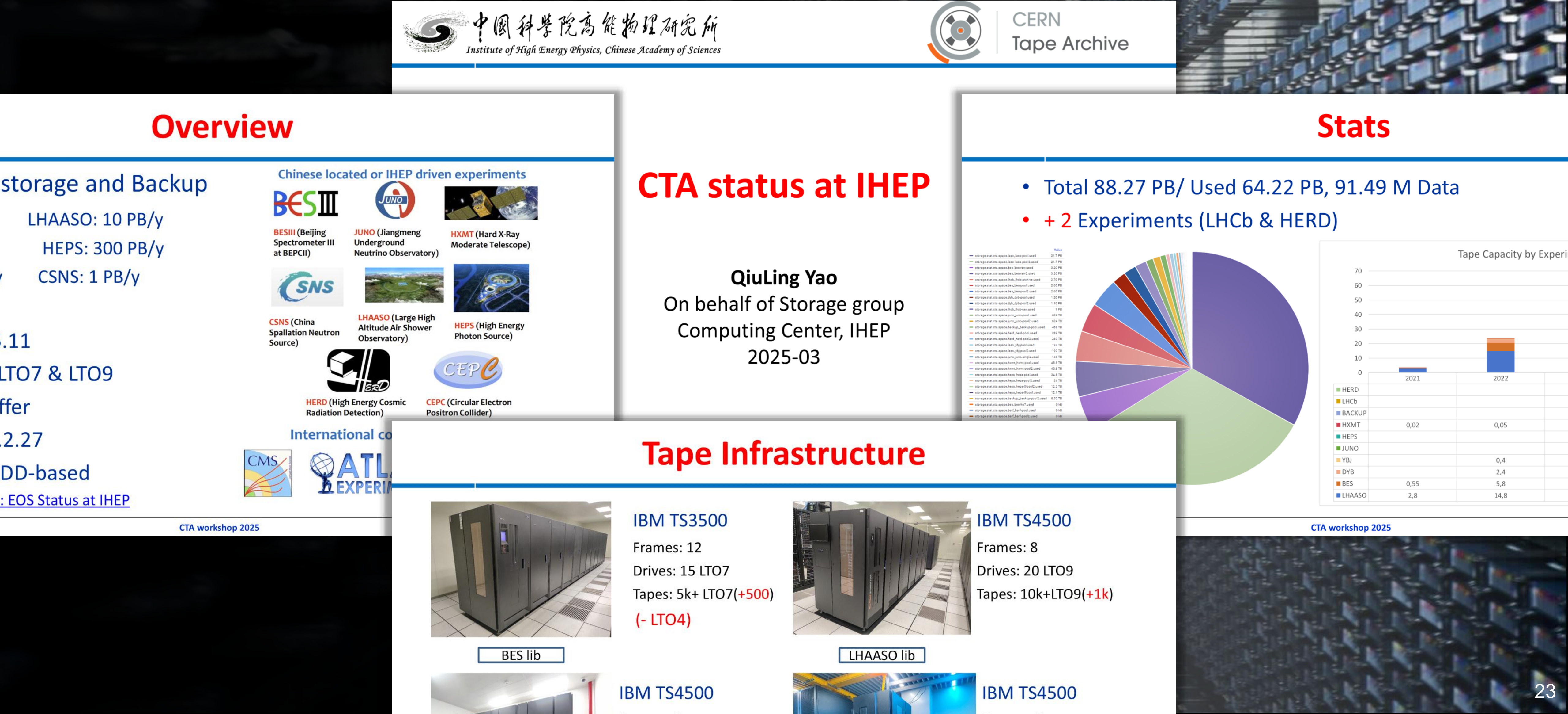
- File residency on the SSD disk is transitory
- A tape copy is an offline file for EOS
- Intended to meet the requirements of Run3 and Hi-Lumi LHC



CTA (CERN Tape Archive) Collaboration


Triumf
dCache + CTA
Rutherford Appleton Lab (RAL)
EOS + CTA
DESY
dCache + CTA
Institute of High Energy Physics (IHEP)
EOS + CTA
Fermi National Accelerator Lab (FNAL)
EOS + CTA
dCache
Port d'Informació Científica (PIC)
dCache + CTA

CERN – IHEP CTA Collaboration



中国科学院高能物理研究所
Institute of High Energy Physics, Chinese Academy of Sciences

CERN
Tape Archive

CTA status at IHEP

QiuLing Yao
On behalf of Storage group
Computing Center, IHEP
2025-03

Overview

storage and Backup

LHAASO: 10 PB/y

HEPS: 300 PB/y

/ CSNS: 1 PB/y

.11

LTO7 & LTO9

ffer

.2.27

DD-based

: EOS Status at IHEP

Chinese located or IHEP driven experiments

BESIII (Beijing Spectrometer III at BEPCII)

JUNO (Jiangmeng Underground Neutrino Observatory)

HXMT (Hard X-Ray Moderate Telescope)

CSNS (China Spallation Neutron Source)

LHAASO (Large High Altitude Air Shower Observatory)

HEPS (High Energy Photon Source)

HERD (High Energy Cosmic Radiation Detection)

CEPC (Circular Electron Positron Collider)

International co

CTA workshop 2025

Stats

- Total 88.27 PB/ Used 64.22 PB, 91.49 M Data
- + 2 Experiments (LHCb & HERD)

	Value
storage.stat.cta.space.laso_laso-pool.used	21.7 PB
storage.stat.cta.space.laso_laso-pool2.used	21.7 PB
storage.stat.cta.space.bes_bes-raw.used	3.20 PB
storage.stat.cta.space.bes_bes-raw2.used	3.20 PB
storage.stat.cta.space.lhcb_lhcb-archive.used	2.70 PB
storage.stat.cta.space.bes_bes-pool.used	2.60 PB
storage.stat.cta.space.bes_bes-pool2.used	2.60 PB
storage.stat.cta.space.dyb_dyb-pool.used	1.20 PB
storage.stat.cta.space.dyb_dyb-pool2.used	1.10 PB
storage.stat.cta.space.lhcb_lhcb-raw.used	1 PB
storage.stat.cta.space.juno_juno-pool.used	624 TB
storage.stat.cta.space.juno_juno-pool2.used	624 TB
storage.stat.cta.space.backup_backup-pool.used	468 TB
storage.stat.cta.space.herd_herd-pool.used	289 TB
storage.stat.cta.space.herd_herd-pool2.used	289 TB
storage.stat.cta.space.laso_ybj-pool.used	192 TB
storage.stat.cta.space.laso_ybj-pool2.used	192 TB
storage.stat.cta.space.juno_juno-single.used	146 TB
storage.stat.cta.space.hxmt_hxmt-pool.used	45.8 TB
storage.stat.cta.space.hxmt_hxmt-pool2.used	45.8 TB
storage.stat.cta.space.heps_heps-pool.used	34.5 TB
storage.stat.cta.space.heps_heps-pool2.used	34 TB
storage.stat.cta.space.heps_heps-l9pool2.used	12.2 TB
storage.stat.cta.space.heps_heps-l9pool.used	12.1 TB
storage.stat.cta.space.backup_backup-pool2.used	8.50 TB
storage.stat.cta.space.bes_bes-lto7.used	0 kB
storage.stat.cta.space.bsrf_bsrf-pool.used	0 kB
storage.stat.cta.space.bsrf_bsrf-pool2.used	0 kB

Tape Capacity by Experi

	2021	2022
HERD		
LHCb		
BACKUP		
HXMT	0,02	0,05
HEPS		
JUNO		
YBJ		0,4
DYB		2,4
BES	0,55	5,8
LHAASO	2,8	14,8

CTA workshop 2025

Tape Infrastructure

IBM TS3500
Frames: 12
Drives: 15 LTO7
Tapes: 5k+ LTO7(+500)
(- LTO4)

BES lib

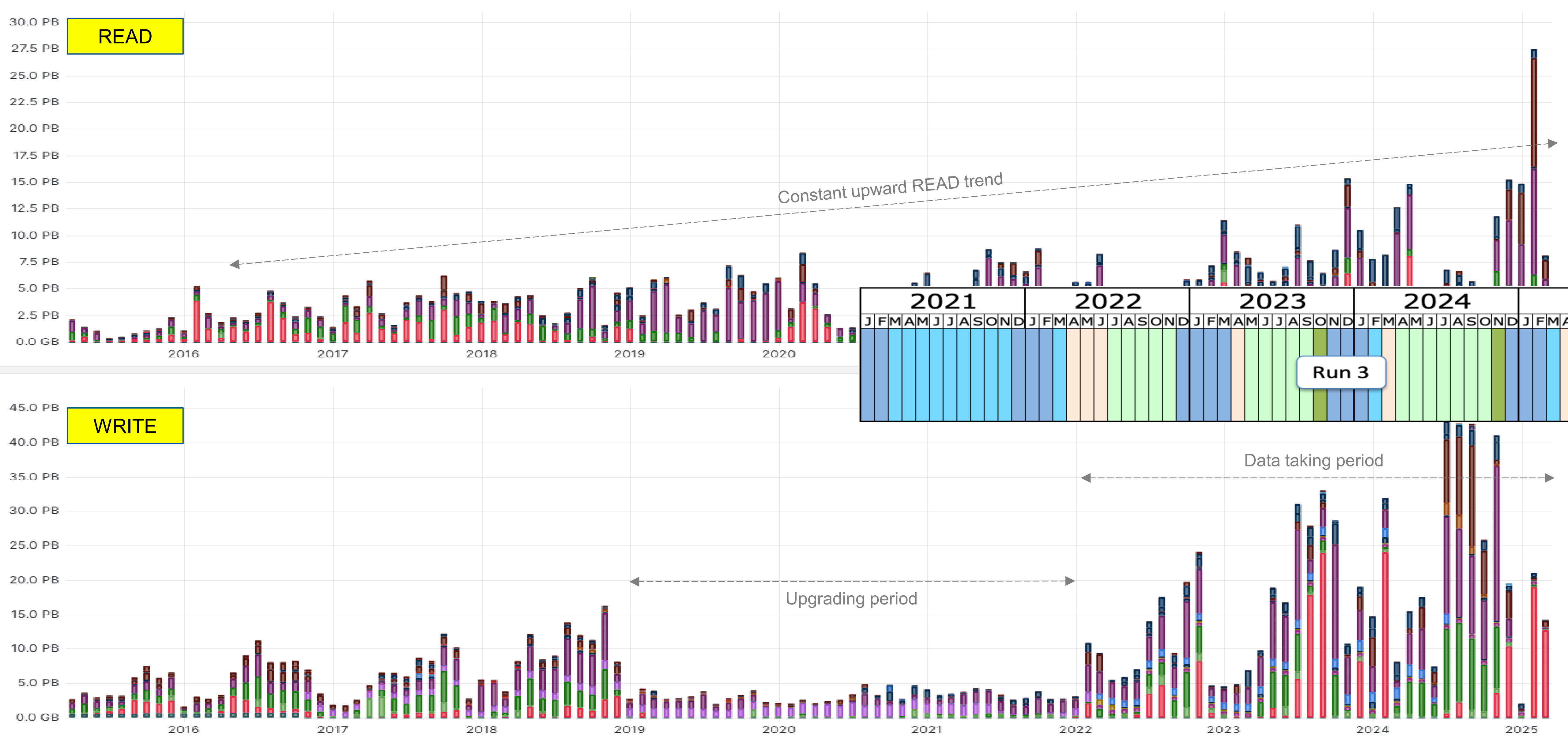
IBM TS4500
Frames: 8
Drives: 20 LTO9
Tapes: 10k+LTO9(+1k)

LHAASO lib

IBM TS4500

IBM TS4500

Monthly tape usage at CERN – past 10 years



Tape Storage challenges

Market dominated by hyperscalers

Vendor consolidation
- Monopoly on the tape drive manufacturing
- Duopoly on the tape media manufacturing

Lack of competition = Price increases

Absence of features available in the past:
- Backward compatibility
- Media upformating

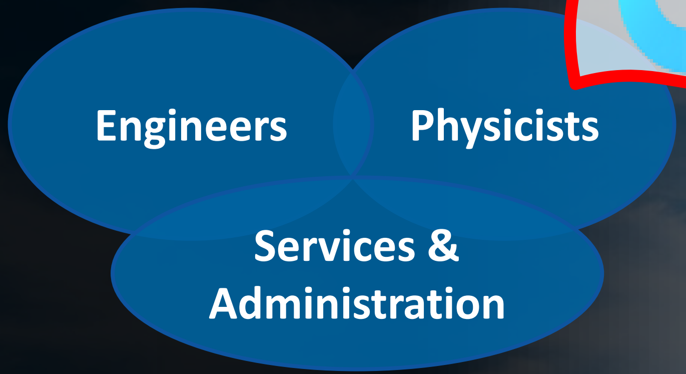
Focus on capacity growth not throughput

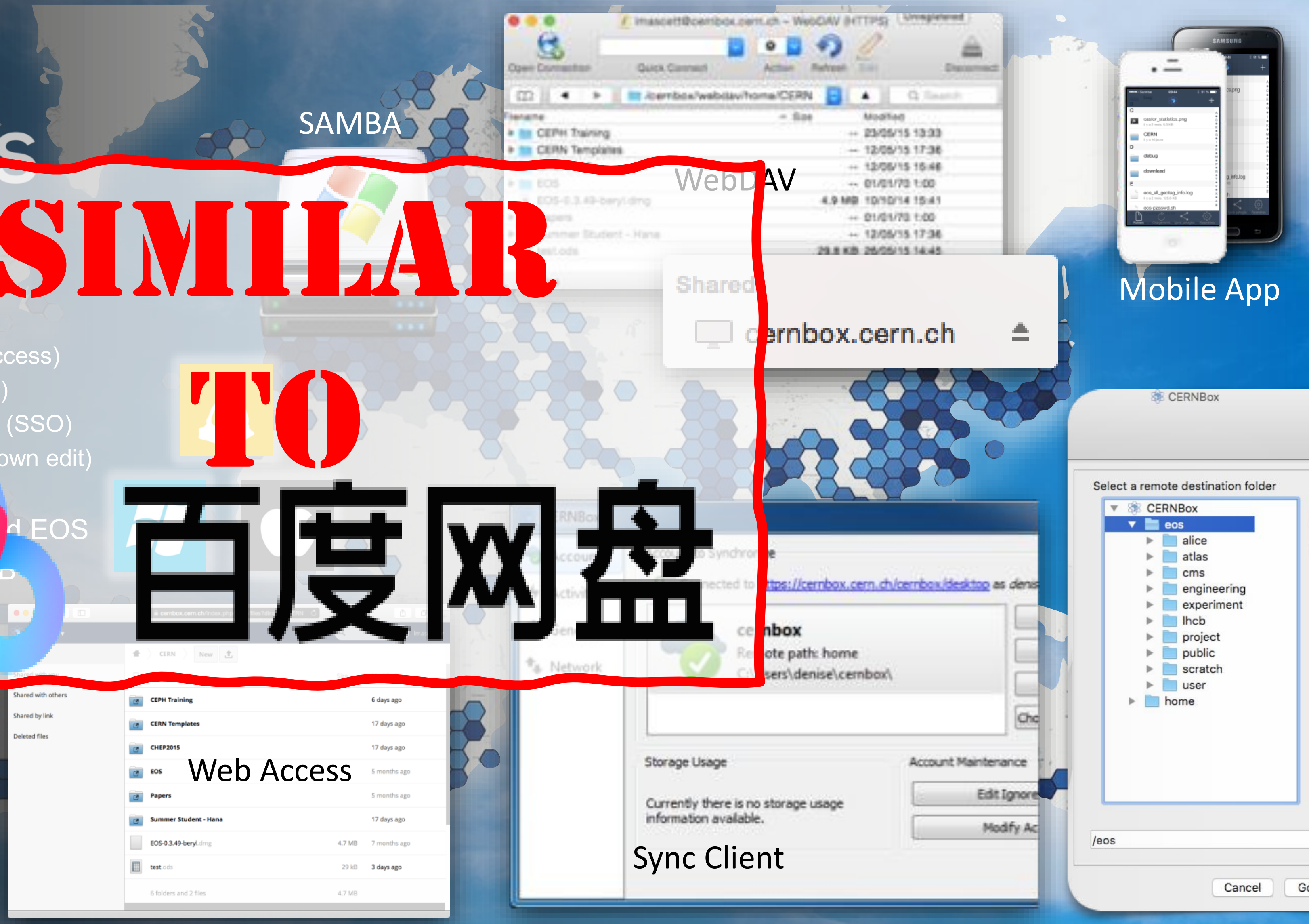
Consequences:
- Need to keep older technology for longer
- Data migration take longer and require resources (drive and cartridge slots)
- Loss of technological competitiveness

CERNBox – CERN Sync & Share platform

CERNBox *powered by* EOS

- Central Hub to access for CERN data on EOS
- Main features:
  - Storage and Synchronization (multiple OS/devices)
  - File Sharing and Collaboration (users control the access)
  - Versioning and File History (recover from accidents)
  - Security (encrypt data in transit and Authentication (SSO)
  - Integration with CERN Services (Office 365, [illegible]kdown edit)
  - Scientific Computing (SWAN, LxPlus, Lx[illegible]ch)
- Built upon the open-source software ownCloud and EOS
- Quotas: personal 1 TB; project space up [illegible] B

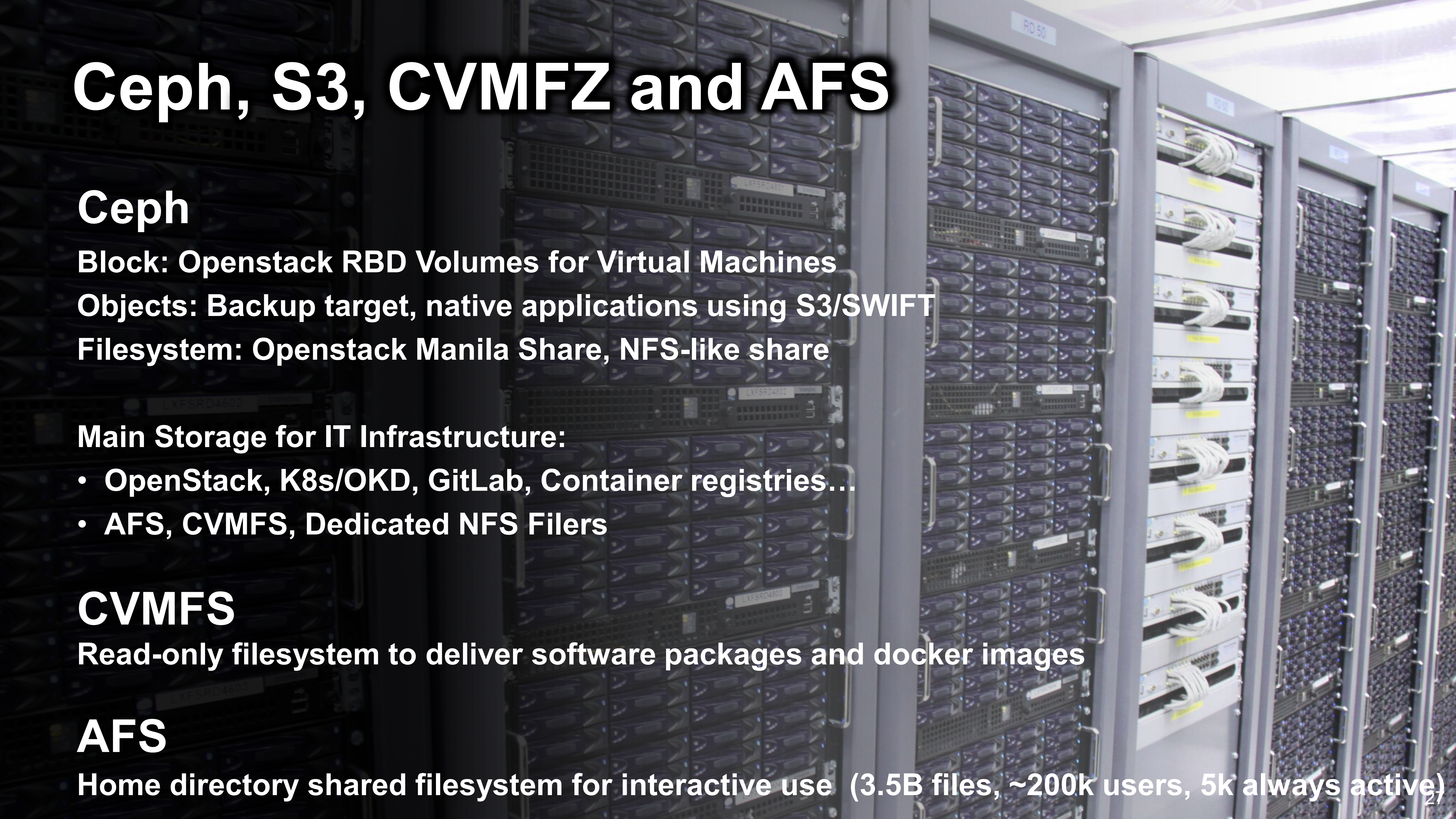




SIMILAR

TO

百度网盘



Ceph, S3, CVMFZ and AFS

Ceph

Block: Openstack RBD Volumes for Virtual Machines

Objects: Backup target, native applications using S3/SWIFT

Filesystem: Openstack Manila Share, NFS-like share

Main Storage for IT Infrastructure:

- OpenStack, K8s/OKD, GitLab, Container registries…
- AFS, CVMFS, Dedicated NFS Filers

CVMFS

Read-only filesystem to deliver software packages and docker images

AFS

Home directory shared filesystem for interactive use (3.5B files, ~200k users, 5k always active)

IT Provided Databases

ORACLE

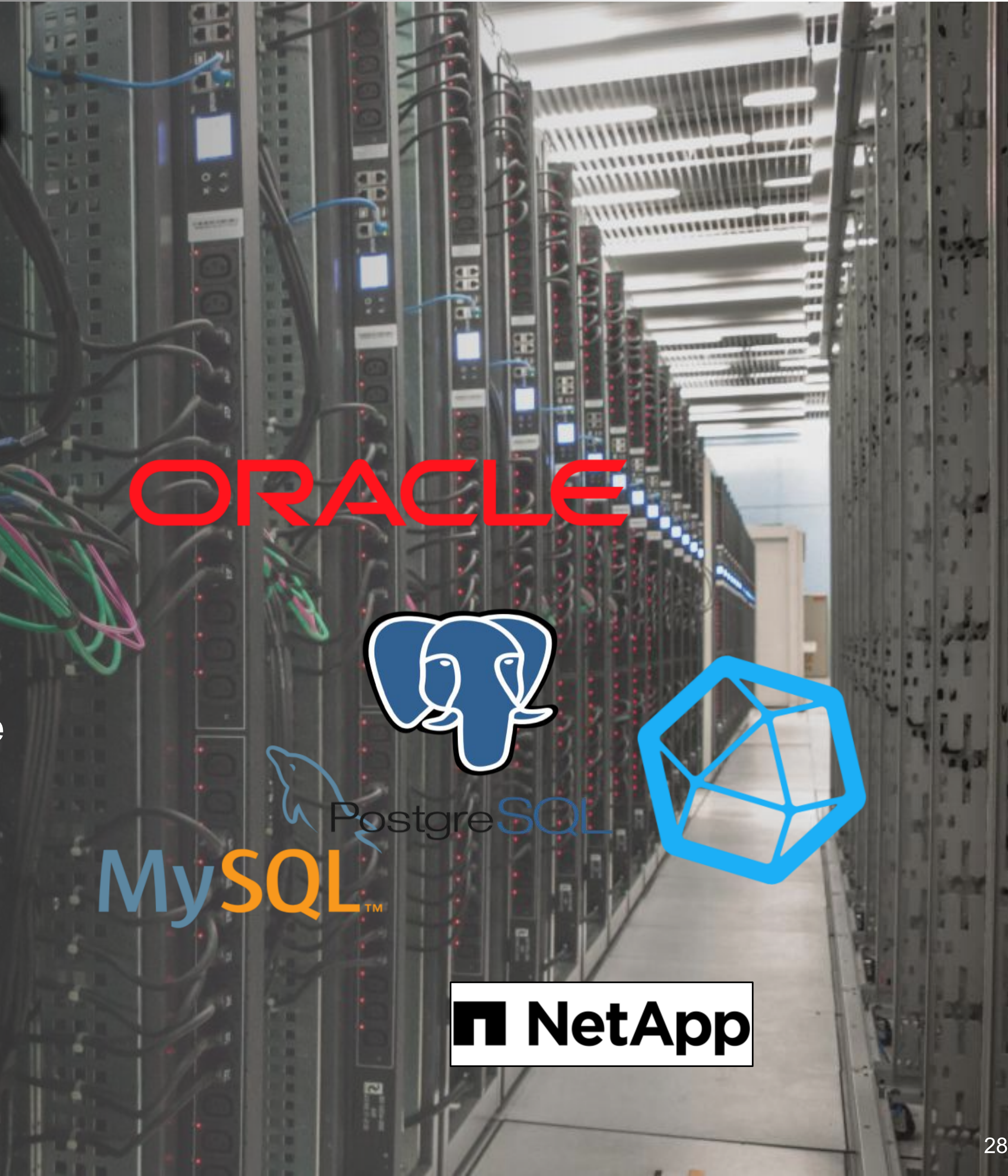
- 124 instances / ~1200 schemas
- +1000 data volumes
- 2.7PB physical space / 6.5PB logical space

MySQL / PostgreSQL / InfluxDB

- + 1500 instances
- + 3000 data volumes
- 165TB physical space / 245TB logical space

7 NetApp clusters

- NFS exported to database hosts
- Logical isolation per use case
- Physical isolation per criticality



CERN **Storage Volume** comparison

Status Q4 2024

raw capacity

0.9 EB	Physics	EOS
90 PB	On-premise Cloud	ceph
36 PB	Sync & Share	CERNBox
20 PB	HDFS	hadoop HDFS
1.5 PB	AFS Home Directories	OpenAFS

raw capacity

1.1 EB

CERN Tape Archive

IBM Storage Protect tape backup **Licensed 15 PB**

CERN **Throughput / Cost** comparison

1 TB/s

Physics

EOS

CERN

Tape Archive

60 GB/s

Cost ratio of **disk : tape** is around **4-5 : 1** of ¥€$ per TB

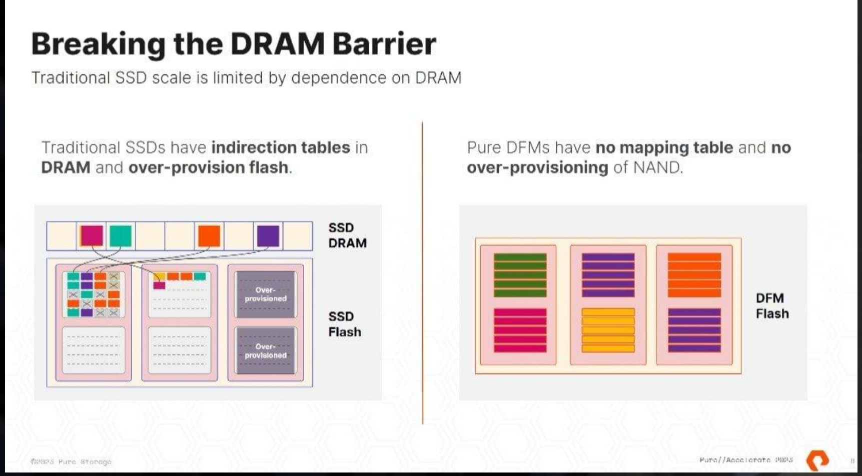
Why do we have so much *disk* vs. *tape* capacity?

Data **analysis** requires *disk* storage with **high bandwidth**

Data **archiving** requires with *tape* storage at **low cost**

CERN openlab innovation platform

PURESTORAGE®



Breaking the DRAM Barrier

Traditional SSD scale is limited by dependence on DRAM

Traditional SSDs have **indirection tables** in **DRAM** and **over-provision flash**.

Pure DFMs have **no mapping table** and **no over-provisioning** of NAND.

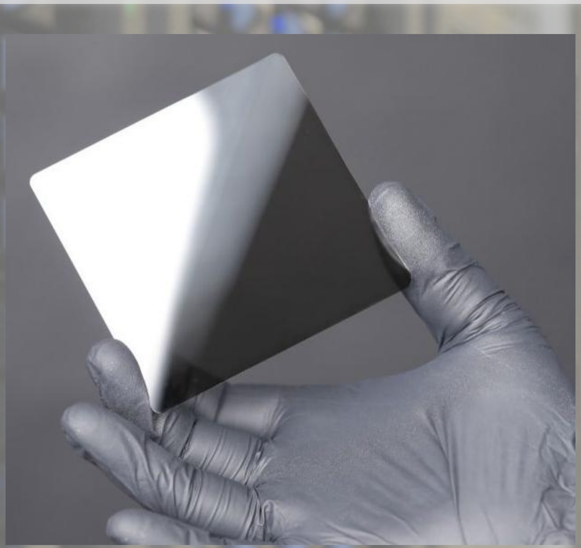
©2025 Pure Storage

Pure//Accelerate 2025

8

- **Next-Generation Exascale Flash Storage**
  - Investigate the latest flash technologies to build a high performance and environmentally-friendly scalable solution
  - Integrate DirectFlash technology into CERN's storage system (EOS) to increase the scalability and its efficiency
  - Evaluate DirectFlash and HDDs EOS auto-tiering model
  - Demonstrate viability and value in HEP environment, evaluate possible solutions that might replace HDDs

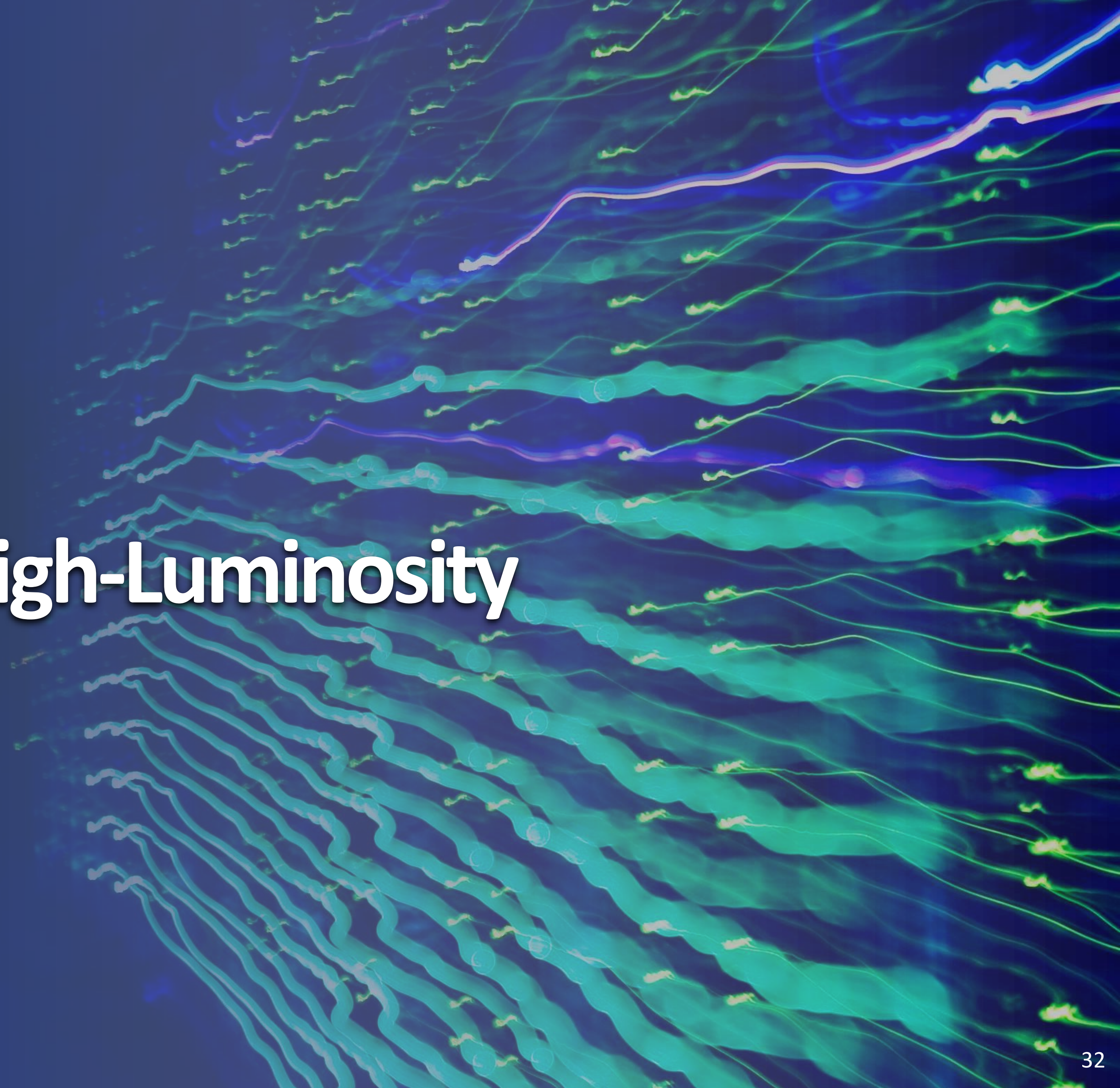
cerabyte



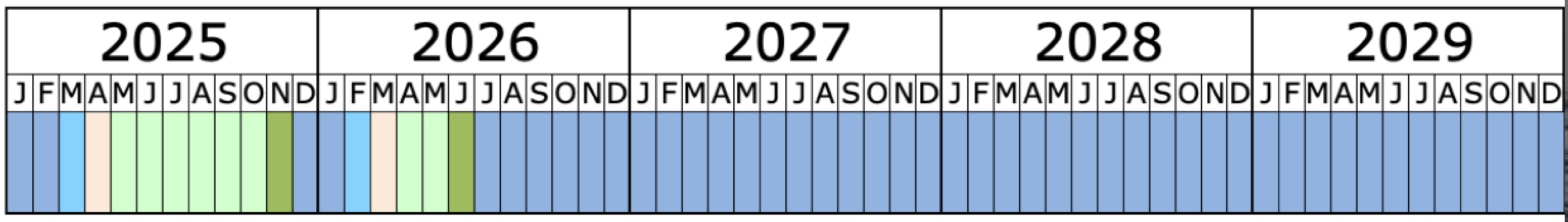
- **Evaluating new materials for data archiving**
  - Glass sheets of 9 cm x 9 cm x 100 µm; coated with a 10 nm thin dark ceramic nano layer; multiple sheets per cartridge
  - Writing: laser beam matrix permanently ablates the ceramic nano-layer; Reading: microscope optics
  - Data encoded in QR codes; arranged in matrix
  - Potential to offer alternative to tape technology

CERN LHC Future …

… towards High-Luminosity



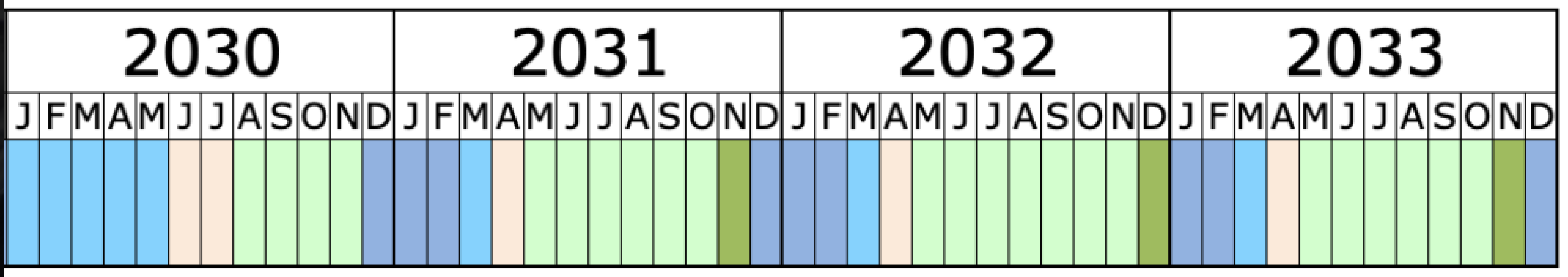
Hi-Luminosity LHC Schedule



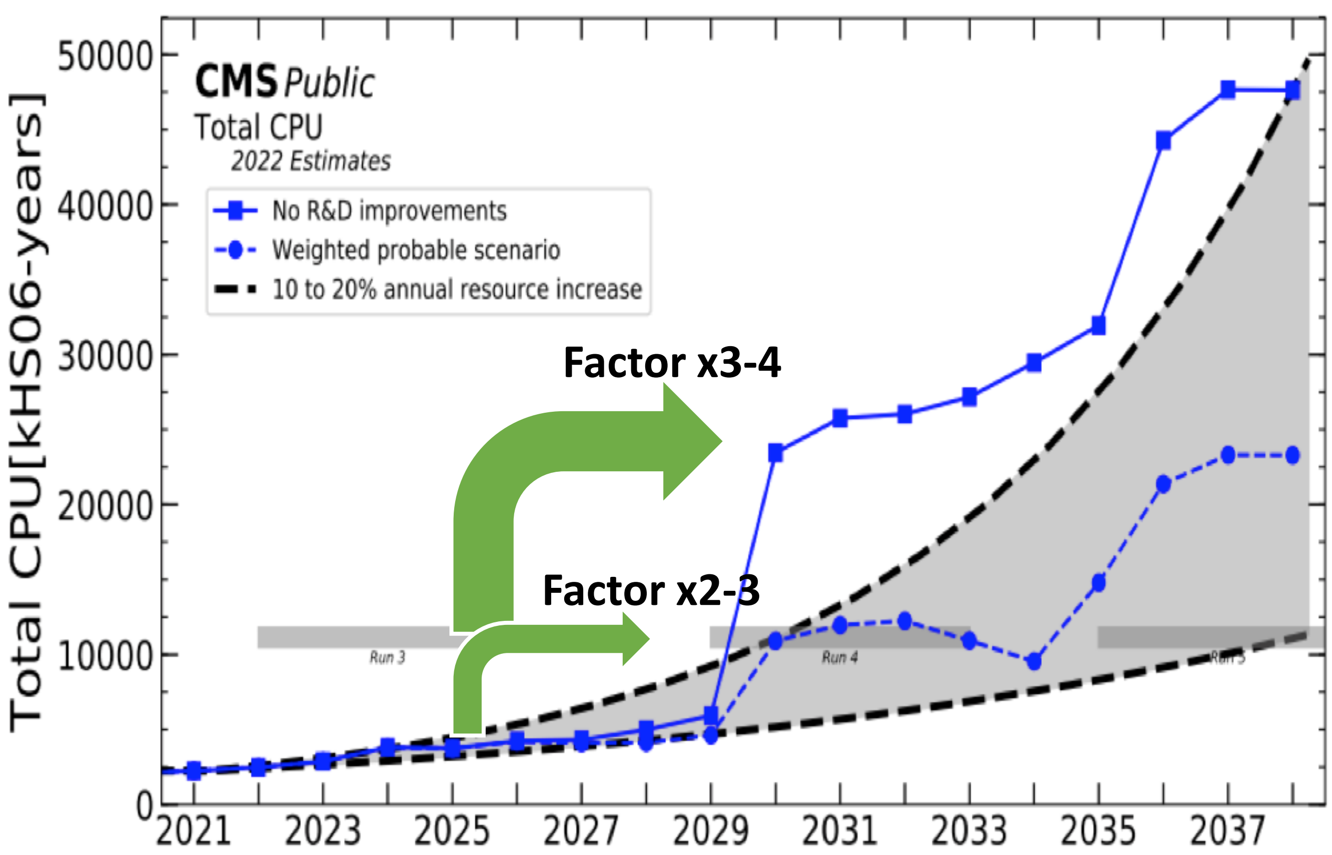
Extended LHC Run3
Operations

Long Shutdown 3
Accelerator Complex and Experiments Upgrades

High-Luminosity LHC Operations
Run4

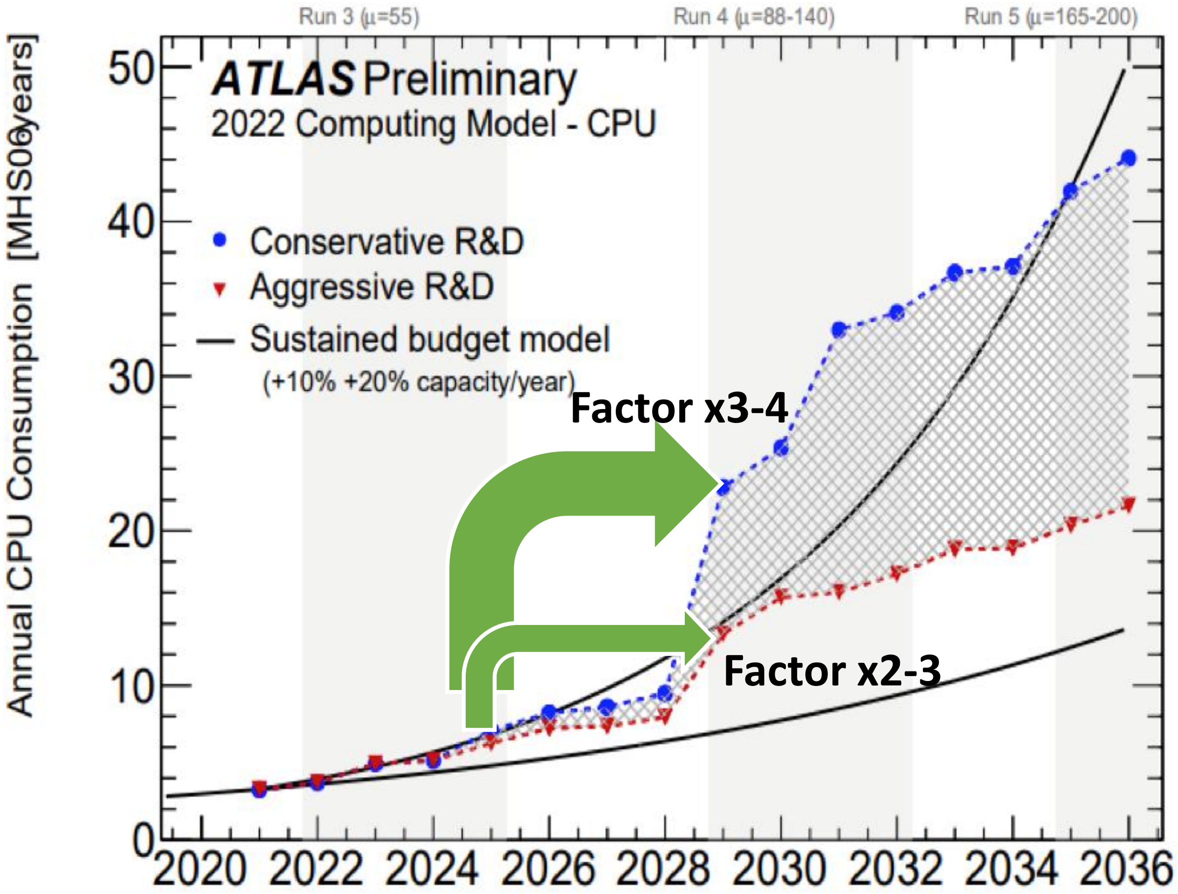


Experiments’ Computing Model



Factor x3-4

Factor x2-3



Factor x3-4

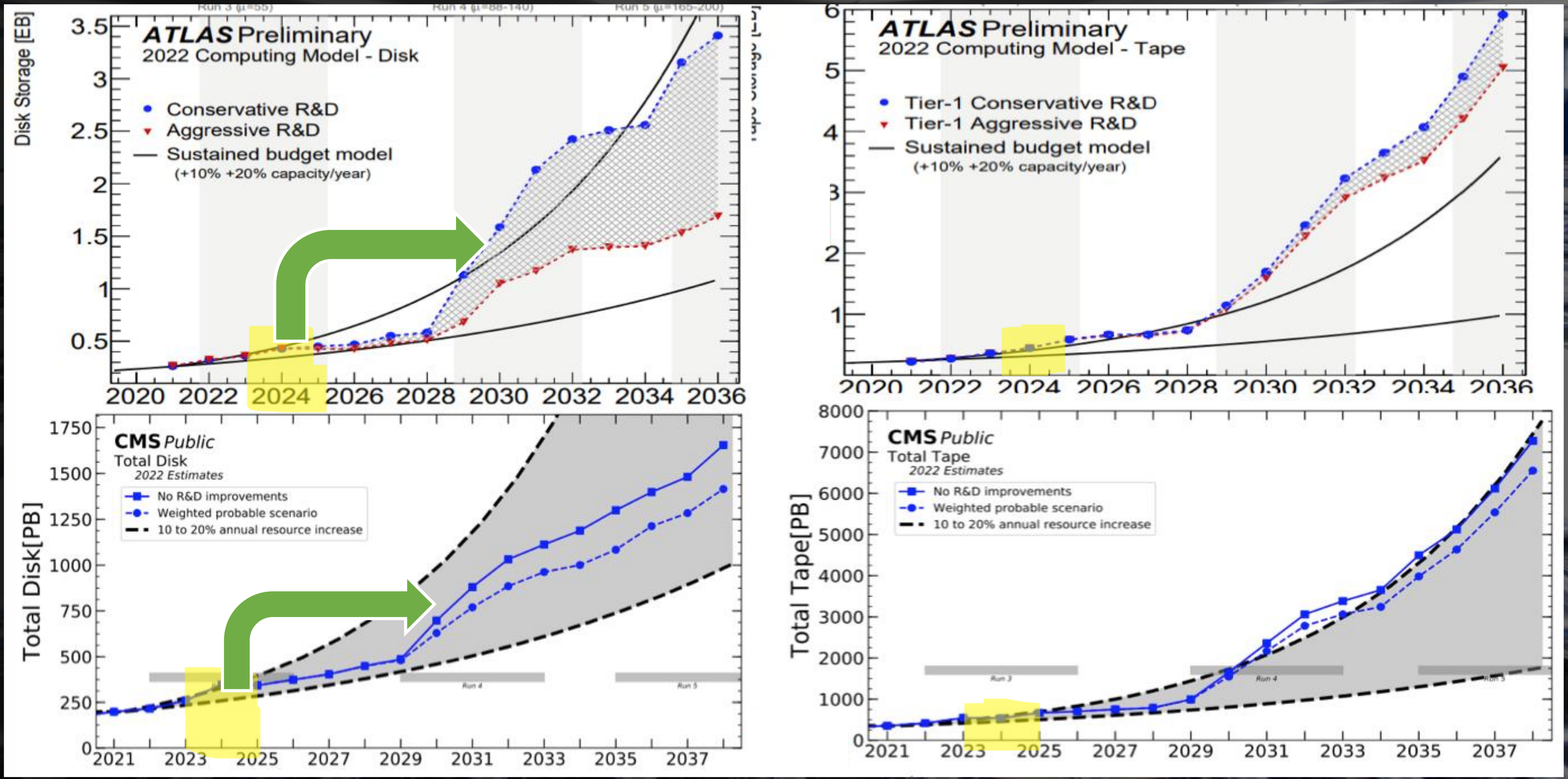
Factor x2-3

With ~7M HS06 (CMS+ATLAS) in WLCG we see 220k parallel streams on CERN Physics storage system (EOS)

  - How many parallel stream should we expect with 50 MHS06?

- In Q4-2024 CMS demonstrated remote reconstruction from WLCG Tier-1s reading directly from CERN
  - Will this be a general trend for the future?

Experiments’ Storage Model



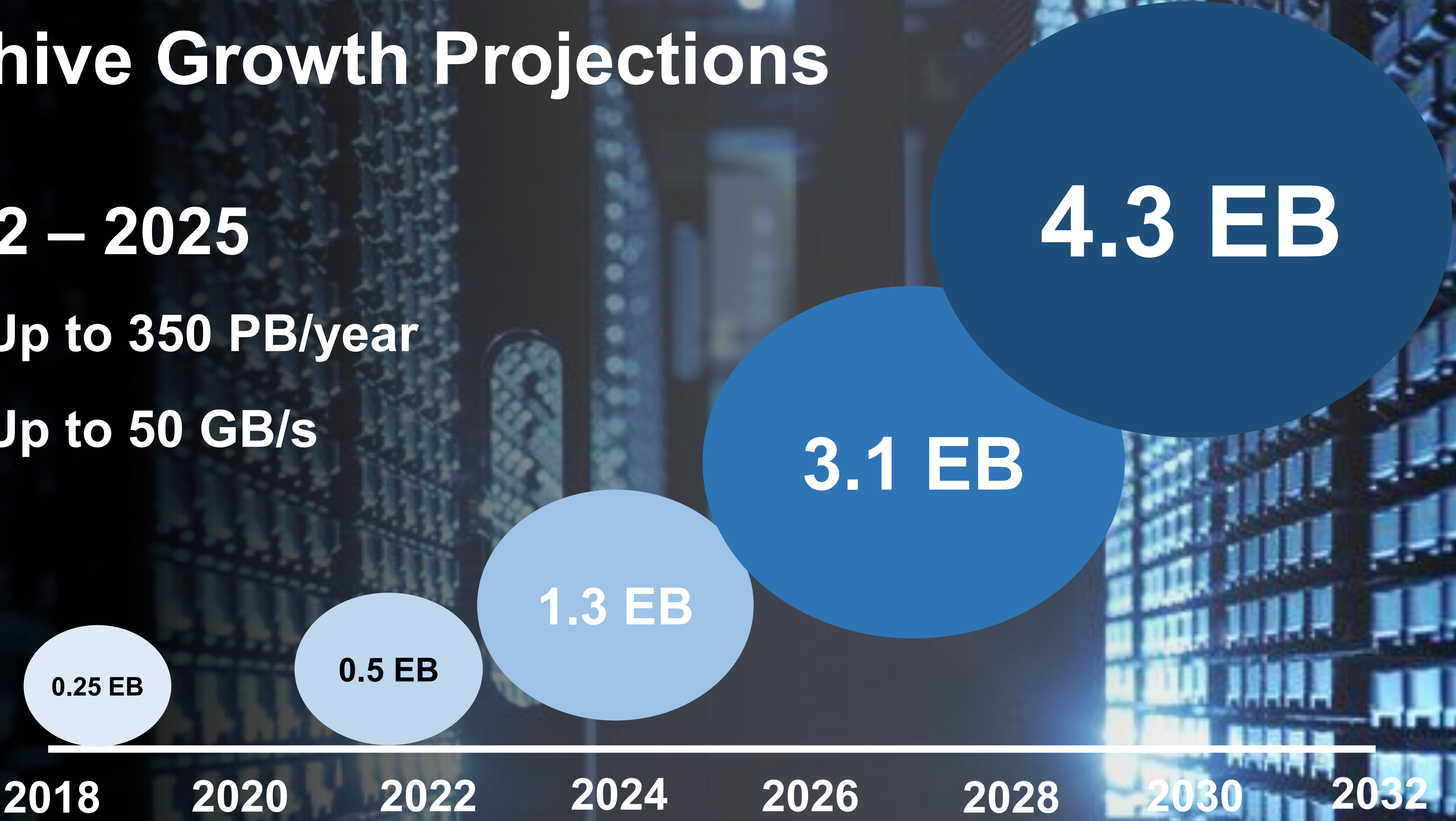
In ~2030 CERN IT will need to provide ~3-4x more of the current storage capacity to LHC Experiments

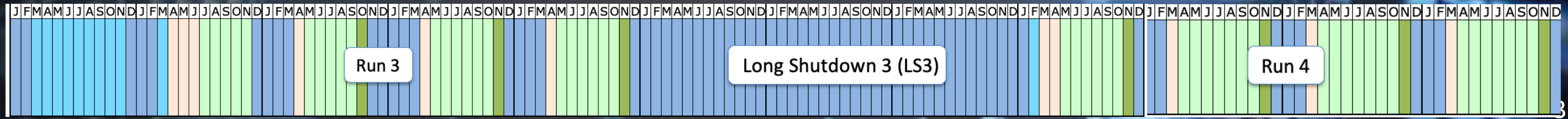
Note: Capacity does not automatically translate into required performance!

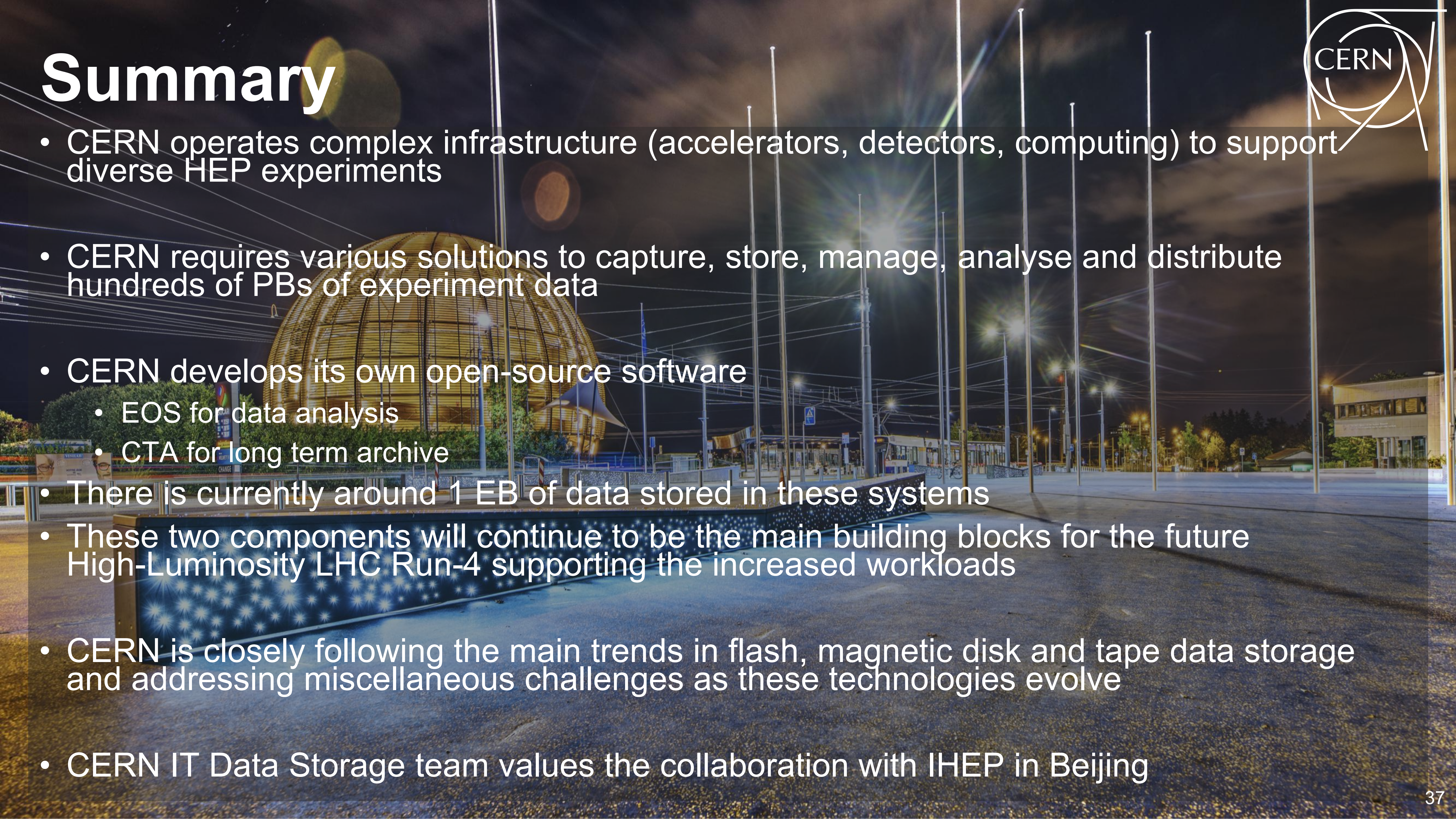
Archive Growth Projections

2022 – 2025

- Up to 350 PB/year
- Up to 50 GB/s







Summary

- CERN operates complex infrastructure (accelerators, detectors, computing) to support diverse HEP experiments
- CERN requires various solutions to capture, store, manage, analyse and distribute hundreds of PBs of experiment data
- CERN develops its own open-source software
  - EOS for data analysis
  - CTA for long term archive
- There is currently around 1 EB of data stored in these systems
- These two components will continue to be the main building blocks for the future High-Luminosity LHC Run-4 supporting the increased workloads
- CERN is closely following the main trends in flash, magnetic disk and tape data storage and addressing miscellaneous challenges as these technologies evolve
- CERN IT Data Storage team values the collaboration with IHEP in Beijing

Thank you for your attention

感谢您的关注