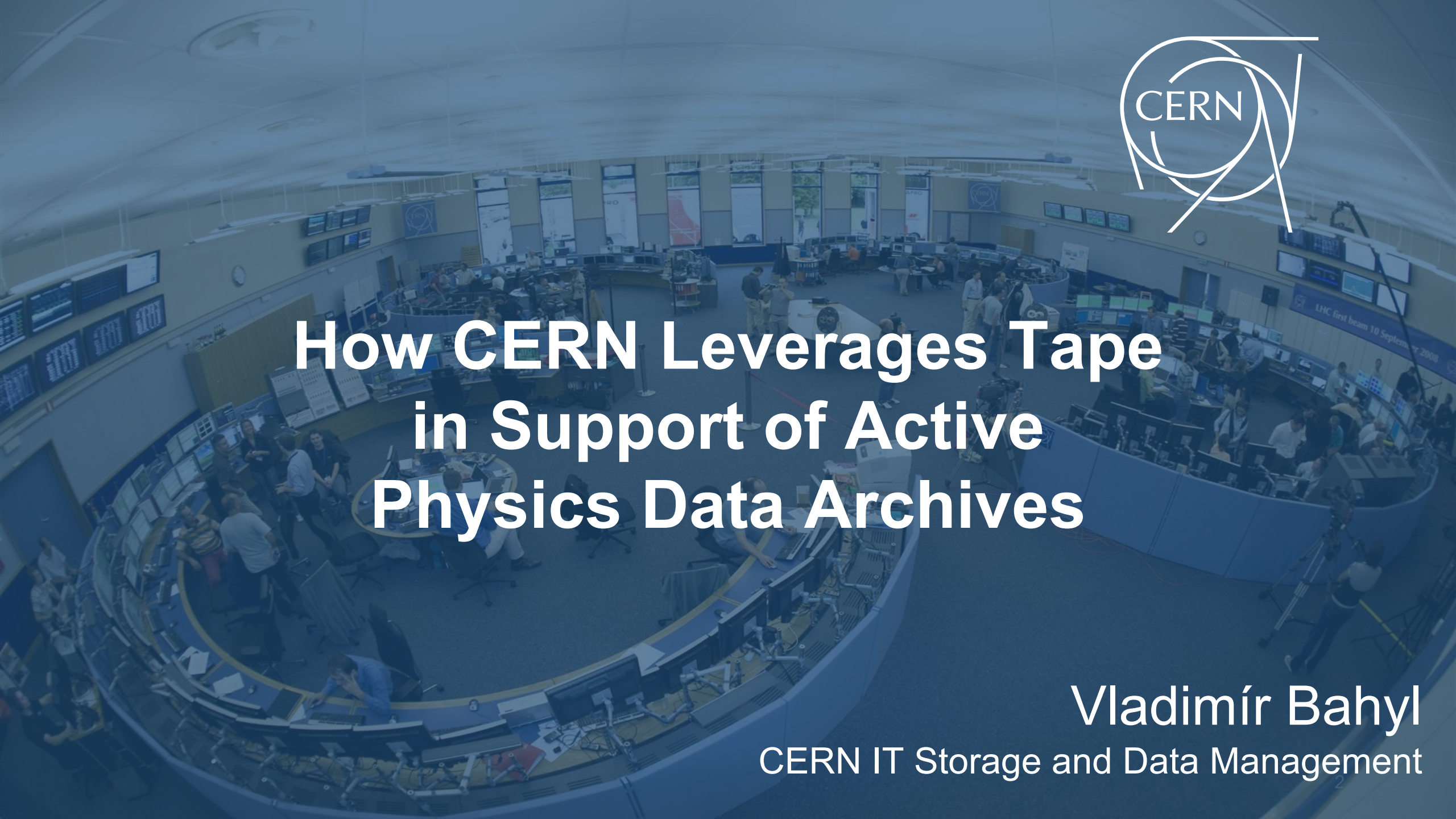

CERN
Accélérateur de science



How CERN Leverages Tape in Support of Active Physics Data Archives

Vladimír Bahyl
CERN IT Storage and Data Management

CERN

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

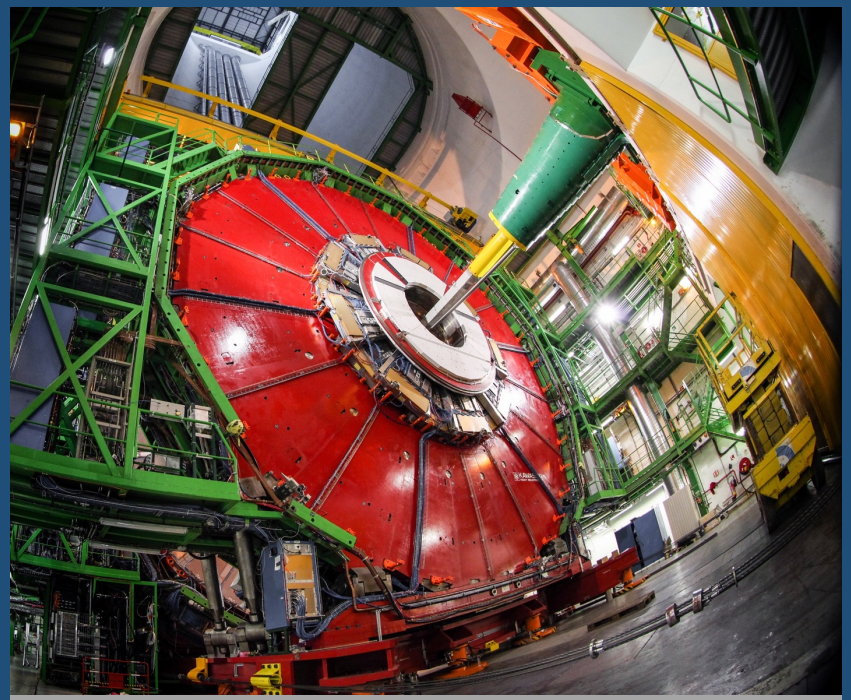
Our goal is to understand the most fundamental particles and laws of the universe.

Located near Geneva on either side of the Swiss French border

How do we do it?

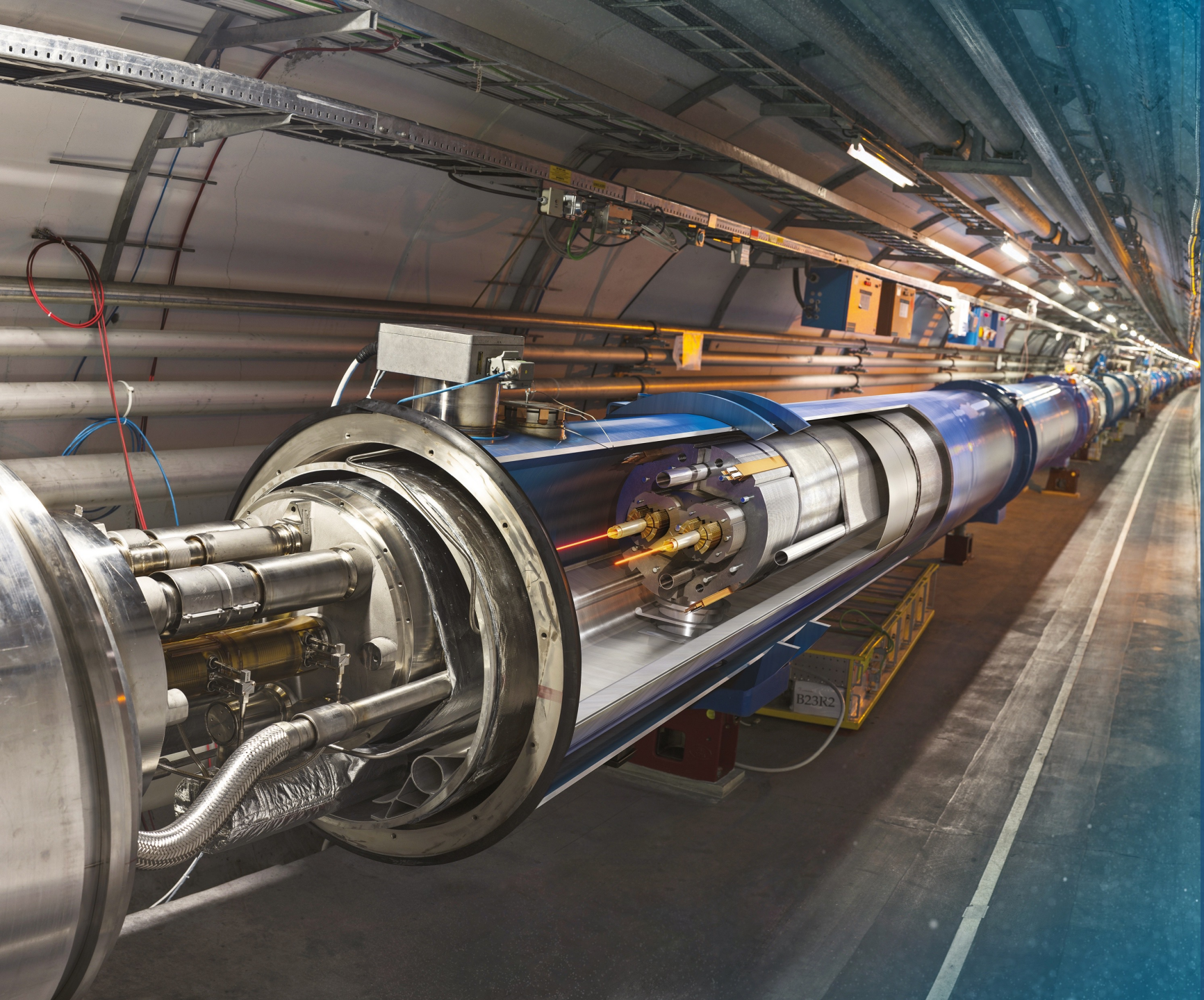
- We build large machines to study the smallest particles in the universe
- We develop technology to advance the limits of what is possible
- We perform world-class research in theoretical and experimental particle physics

ACCELERATORS



DETECTORS

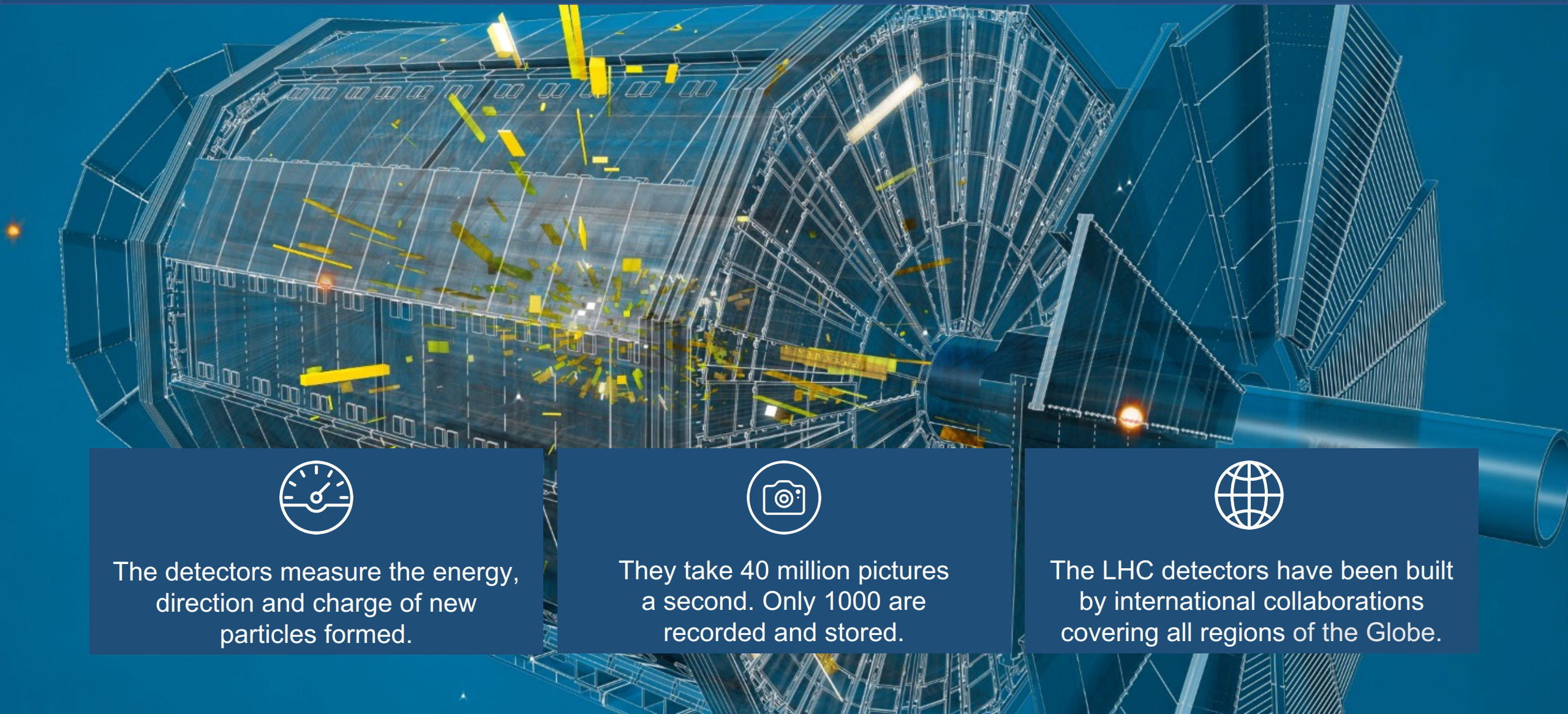
COMPUTING



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

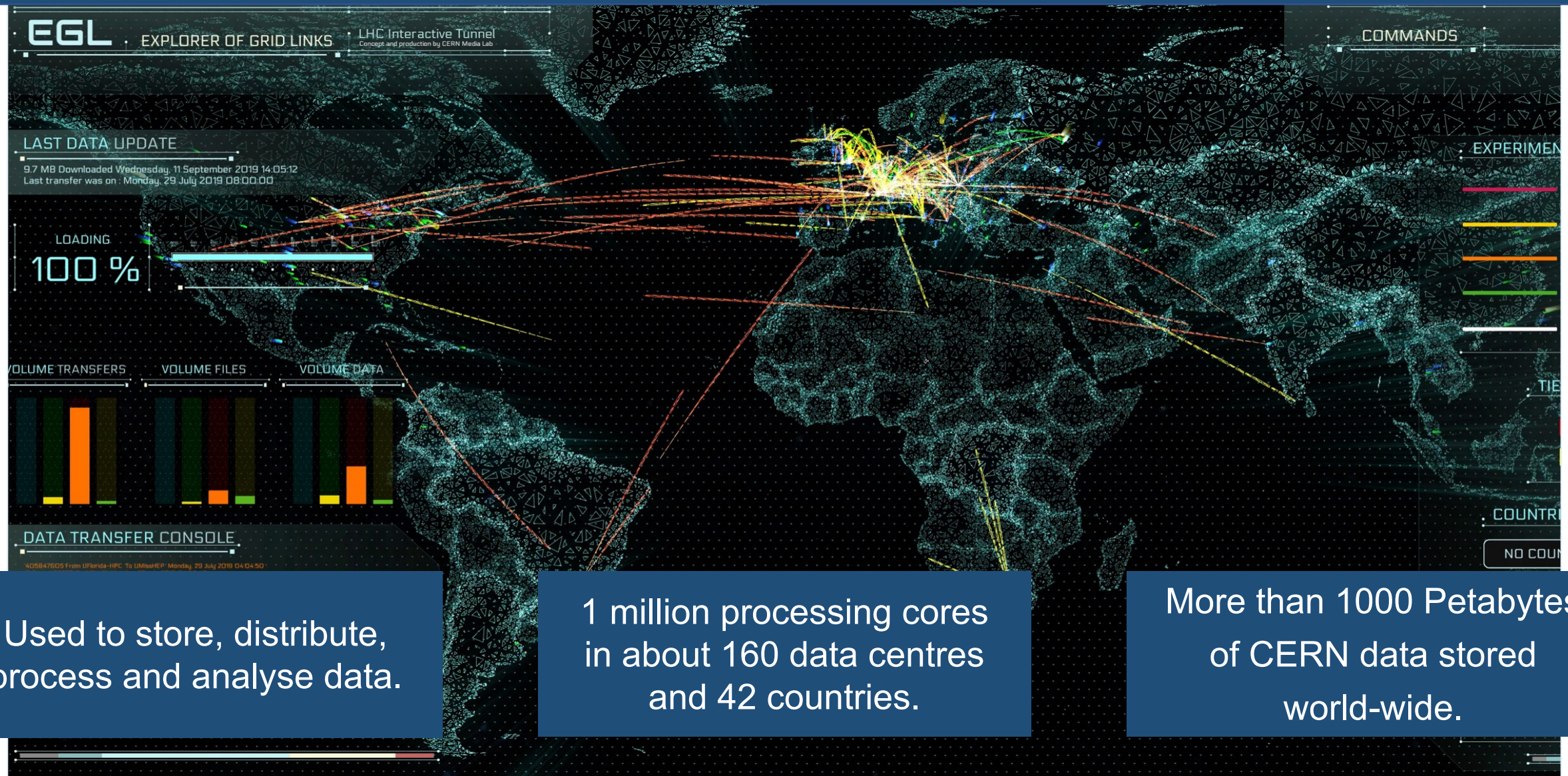


The detectors measure the energy, direction and charge of new particles formed.

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)

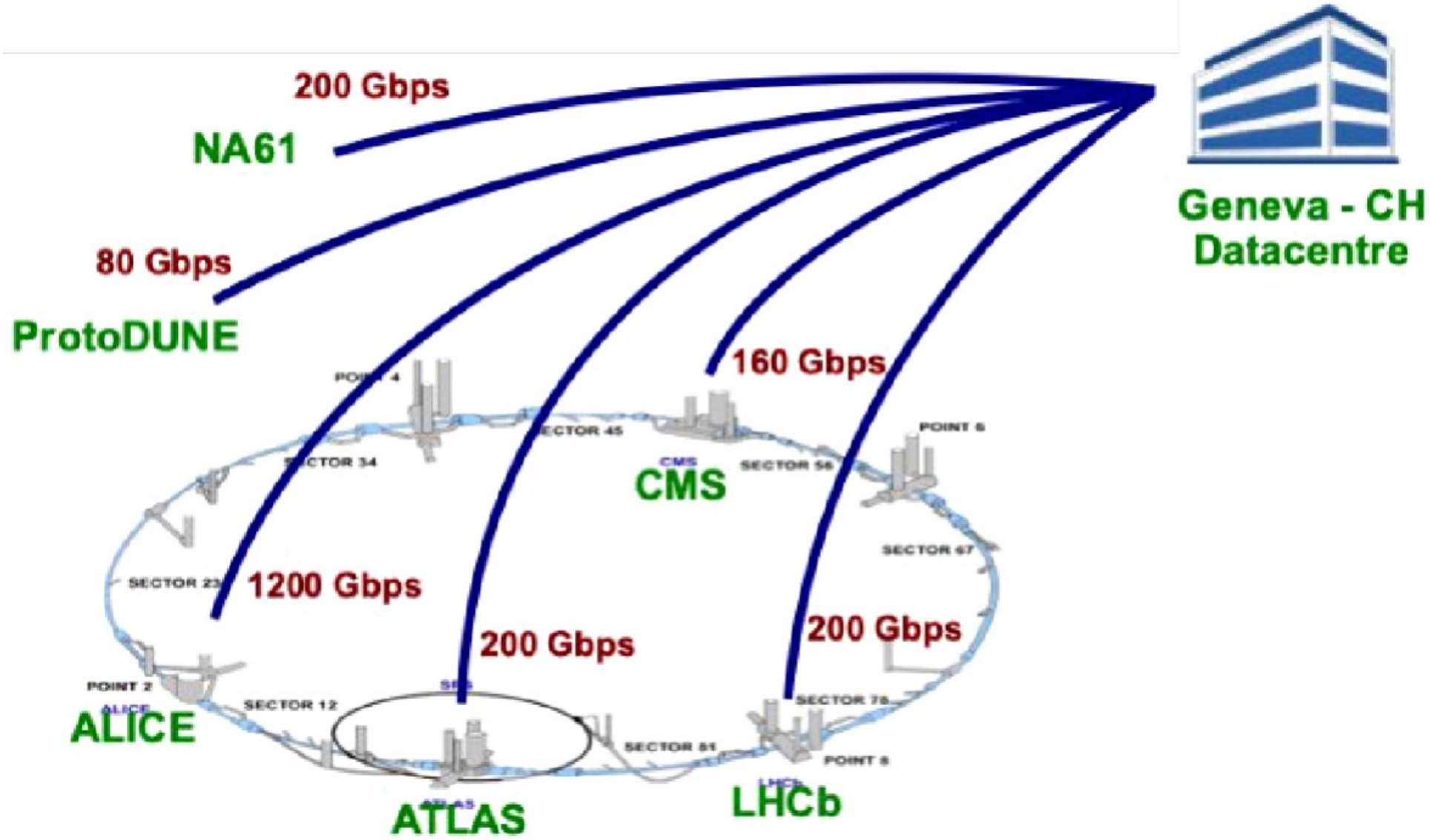


Used to store, distribute, process and analyse data.

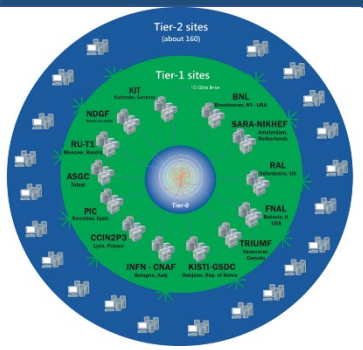
1 million processing cores in about 160 data centres and 42 countries.

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

CERN Tier-0 Data Rates (2022 – 2026)



CERN IT Data Storage Services



Tape services

Physics services

SWAN

Sync & Share services

openstack

local batch cluster
$O(10^5)$ cores

CERNBox

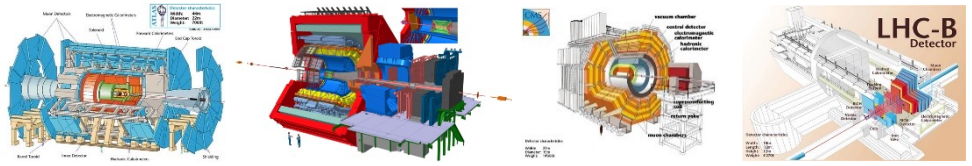
CERN
Tape Archive

EOS

SSD

SSD

30-40 GB/s



EOS

Total Space
600 PB

Files Stored
~7 Billion

 # Storage Nodes
~1600

 # Disks
~80000

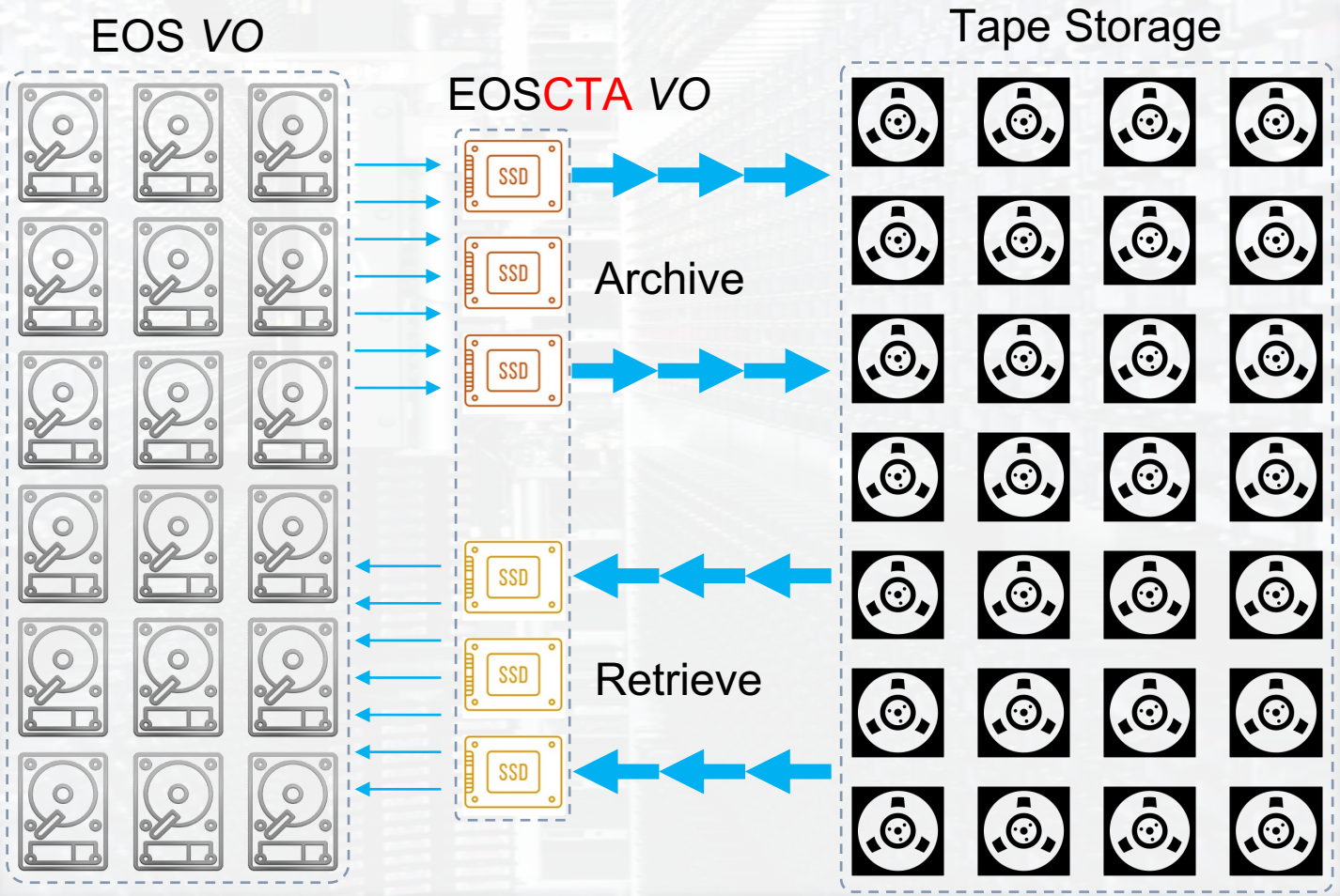
CERN Tape Archive (CTA) architecture

EOS is natively used as a namespace and disk pool manager

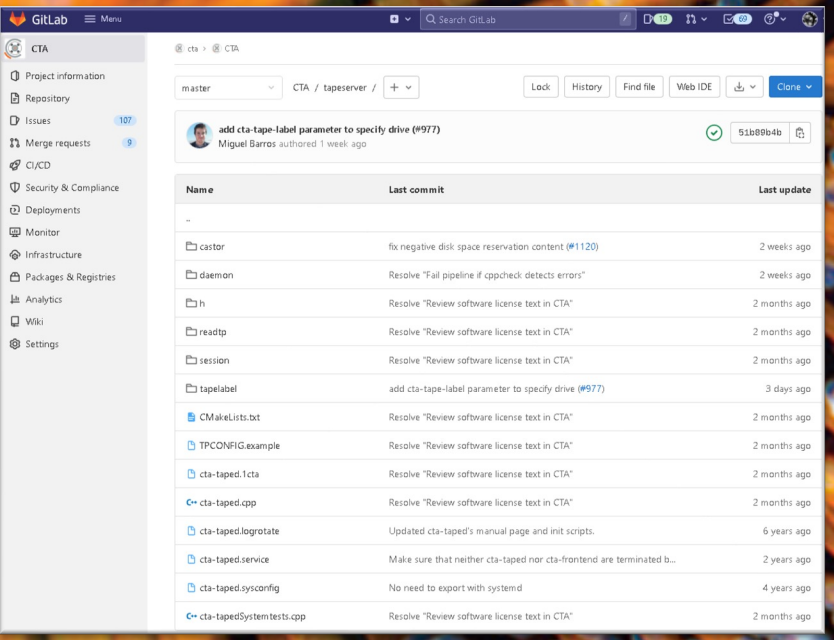
A pure SSD EOS instance with tape backend

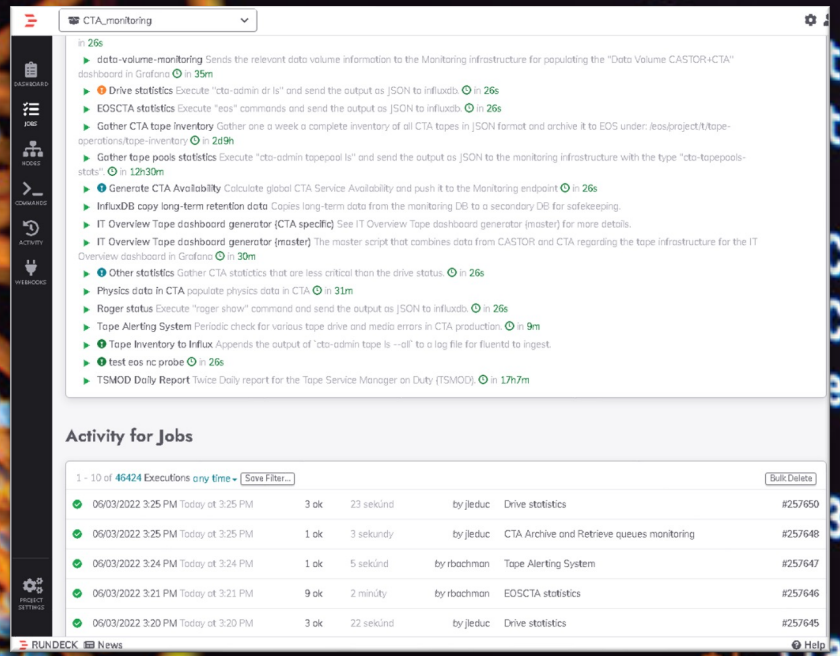
Conceived as a fast buffer to the tape system

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



Integrated with Open Source tools





Development using GitLab

Operations helped by Rundeck

Monitored with InfluxDB / Grafana

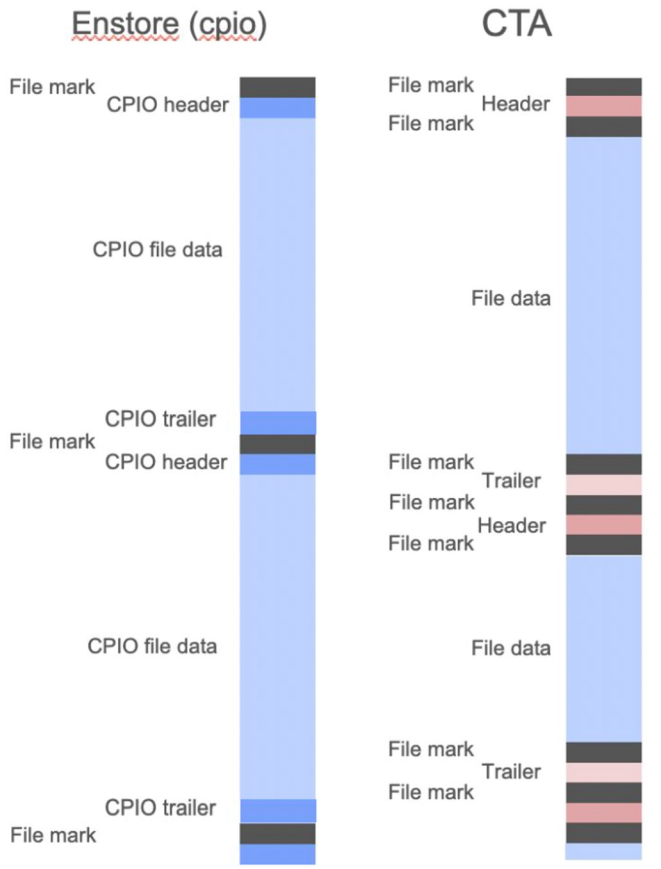
International collaboration

Deployment at DESY

CTA Evaluation at Fermilab

Compatibility with CPIO Tape File Format

- Data at Fermilab is largely stored in CPIO tape format
  - Enstore also supports CERN tape format used in CTA
    - This format is used by CMS and for large files (8GB+)
    - We and our colleagues at PIC find CPIO much more performant in Enstore
  - However, CTA currently has no support for CPIO format
    - Much of our dev effort focused on implementing this support
    - Planning to implement read, not necessarily write, functionality for this format
      - Contact: ewv@fnal.gov



Fermilab

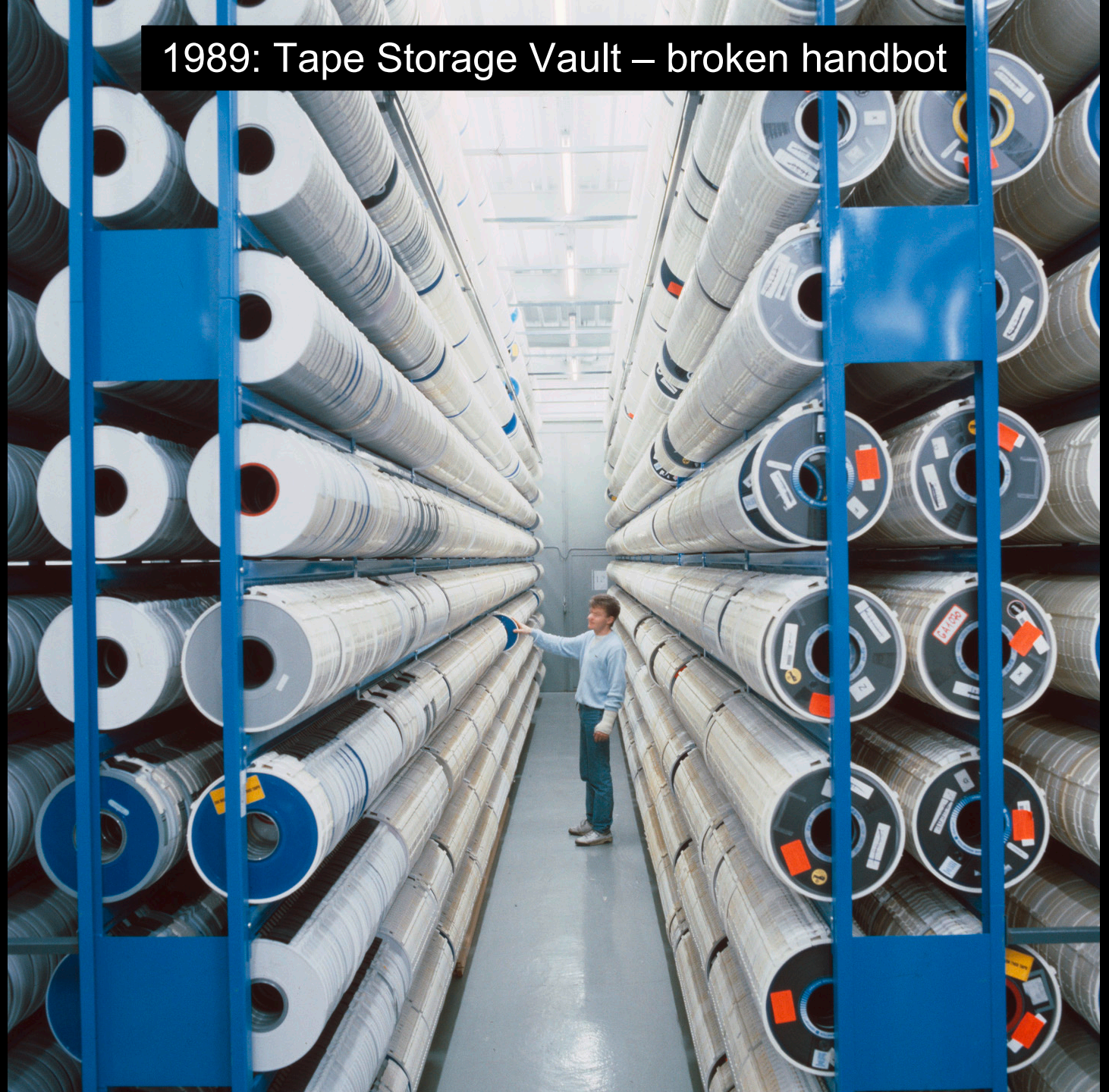
CTA Tra
- Files
- CTS
- EOS
- Just
- Work
- New

Yujiang Bi, IHEPCC, EOS W

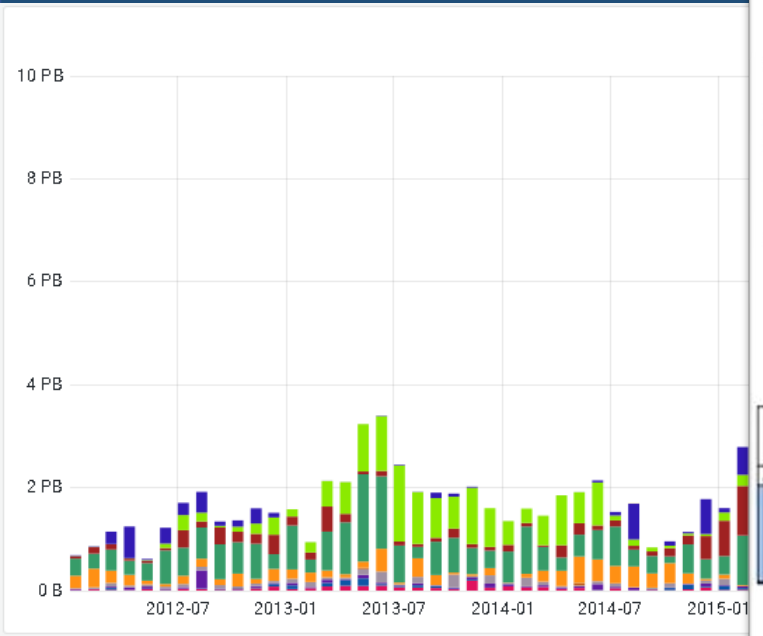
Antares

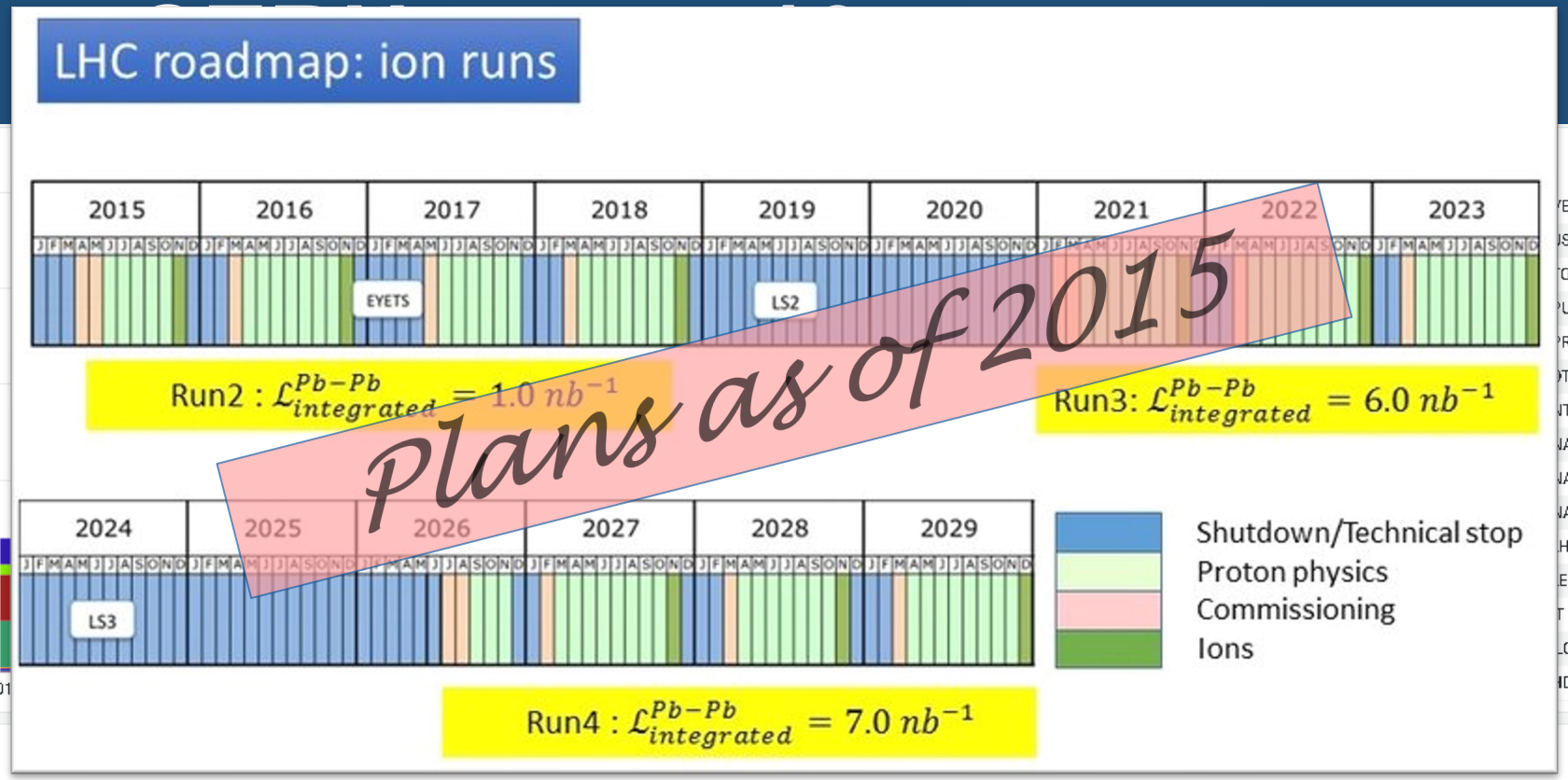
GridPP UK Computing for Particle Physics

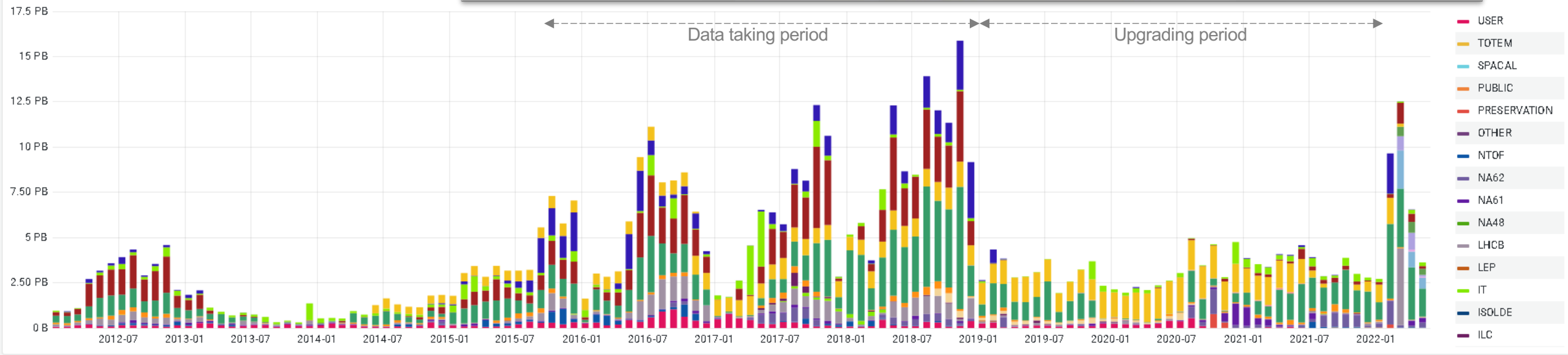
1974: Tape Storage Vault – when robots were human


1989: Tape Storage Vault – broken handbot

Tape usage


10 PB
8 PB
6 PB
4 PB
2 PB
0 B
2012-07
2013-01
2013-07
2014-01
2014-07
2015-01


LHC roadmap: ion runs
2015
2016
2017
2018
2019
2020
2021
2022
2023
EYETS
LS2
Run2 : $\mathcal{L}_{integrated}^{Pb-Pb} = 1.0\ nb^{-1}$
Run3: $\mathcal{L}_{integrated}^{Pb-Pb} = 6.0\ nb^{-1}$
Plans as of 2015
2024
2025
2026
2027
2028
2029
LS3
Shutdown/Technical stop
Proton physics
Commissioning
Ions
Run4 : $\mathcal{L}_{integrated}^{Pb-Pb} = 7.0\ nb^{-1}$


17.5 PB
15 PB
12.5 PB
10 PB
7.50 PB
5 PB
2.50 PB
0 B
Data taking period
Upgrading period
2012-07
2013-01
2013-07
2014-01
2014-07
2015-01
2015-07
2016-01
2016-07
2017-01
2017-07
2018-01
2018-07
2019-01
2019-07
2020-01
2020-07
2021-01
2021-07
2022-01
USER
TOTEM
SPACAL
PUBLIC
PRESERVATION
OTHER
NTOF
NA62
NA61
NA48
LHCB
LEP
IT
ISOLDE
ILC

Tape Infrastructure (June 2022)

- Archive of the physics data
- Provisioned capacity: ~520 PB

- Libraries:
  - 3 x IBM TS4500
  - 2 x Spectra Logic TFinity
- Drives:
  - 76 x IBM1160, 10 x IBM TS1155
  - 98 x LTO9, 10 x LTO8
- Media:
  - 84 PB on 3592JE, 227 PB on 3592JD, 34 PB on 3592JC
  - 83 PB on LTO9, 29 PB on LTO8, 62 PB on LTO7M

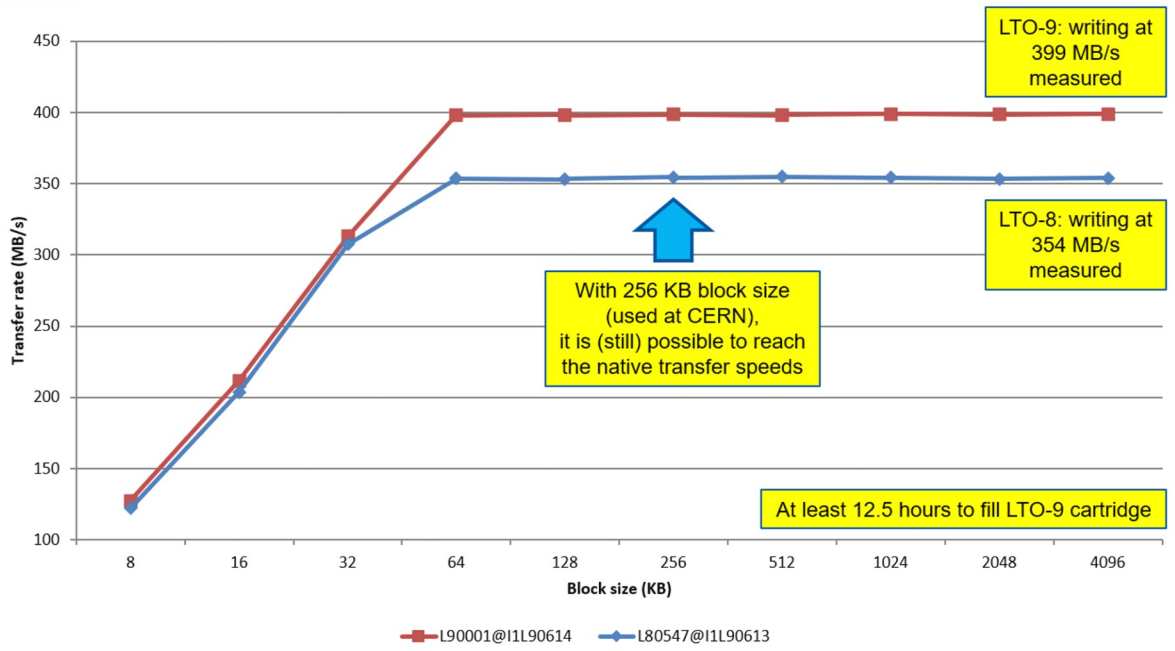
- Backup of the business data
- Licensed capacity: ~15 PB

- Libraries:
  - 1 x IBM TS4500 (partitioned)
  - 1 x Spectra Logic TFinity (partitioned)
- Drives:
  - 10 x IBM TS1155
  - 10 x LTO8
- Media:
  - 9.7 PB on 3592JC
  - 9 PB on LTO7M

Tape Infrastructure (June 2022)

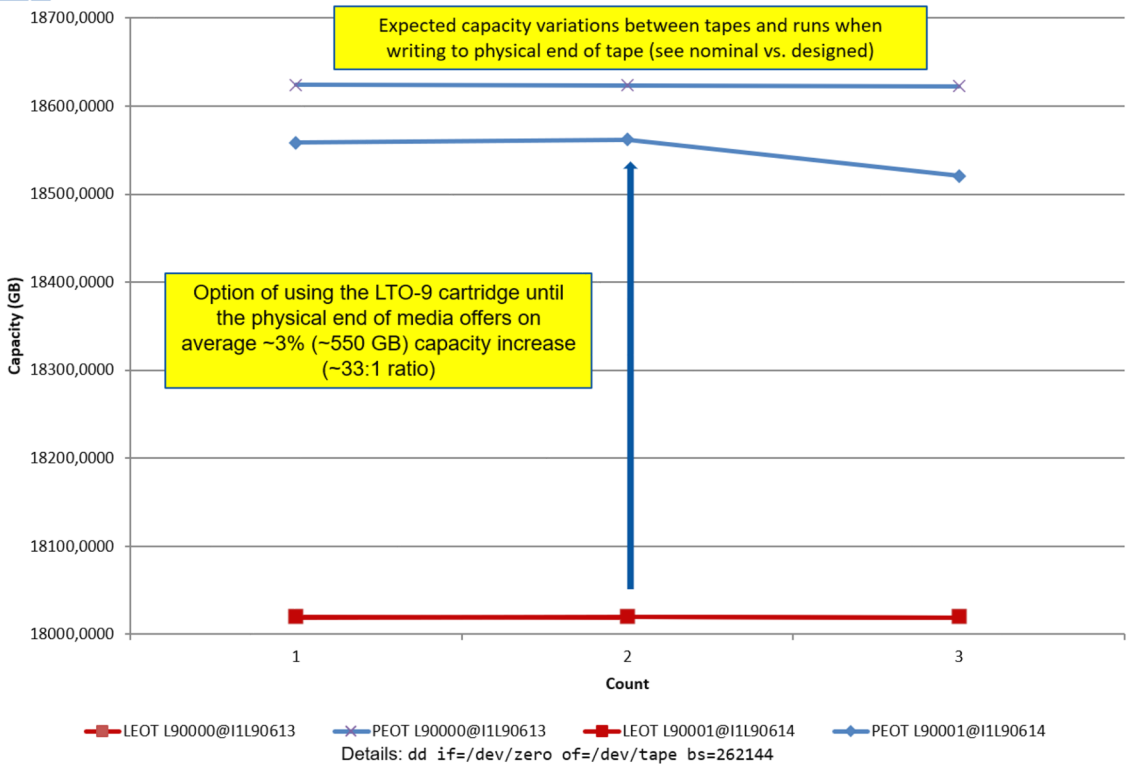
Selected as β test site

Performance – block transfer rates



Details: `dd if=/dev/zero of=/dev/tape bs=(8,16,32,64,128,256,512,1024,2048,4096)k`

LTO-9 capacity – logical vs. physical

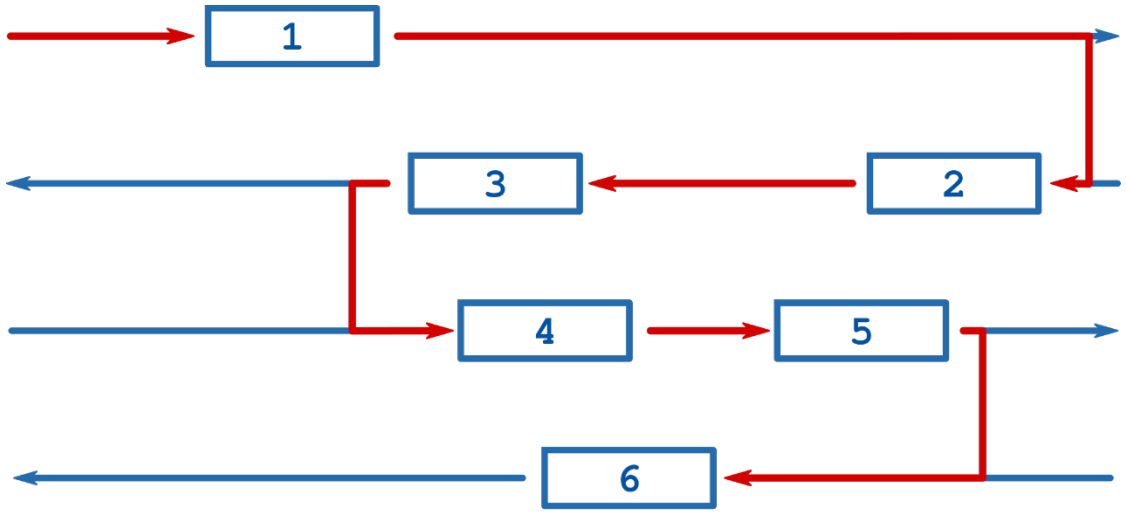


Details: `dd if=/dev/zero of=/dev/tape bs=262144`

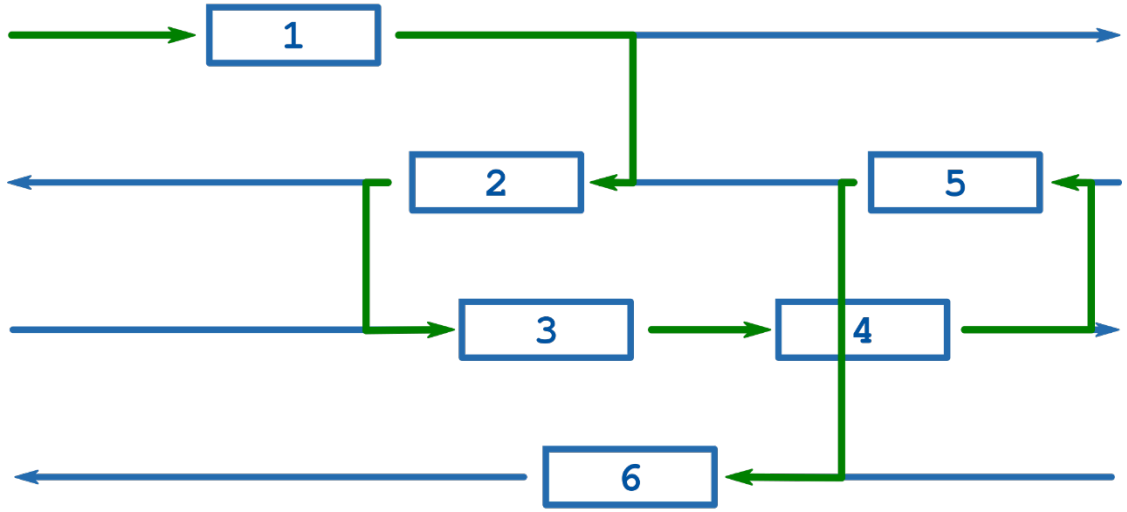
Serpentine layout

Recommended Access Order

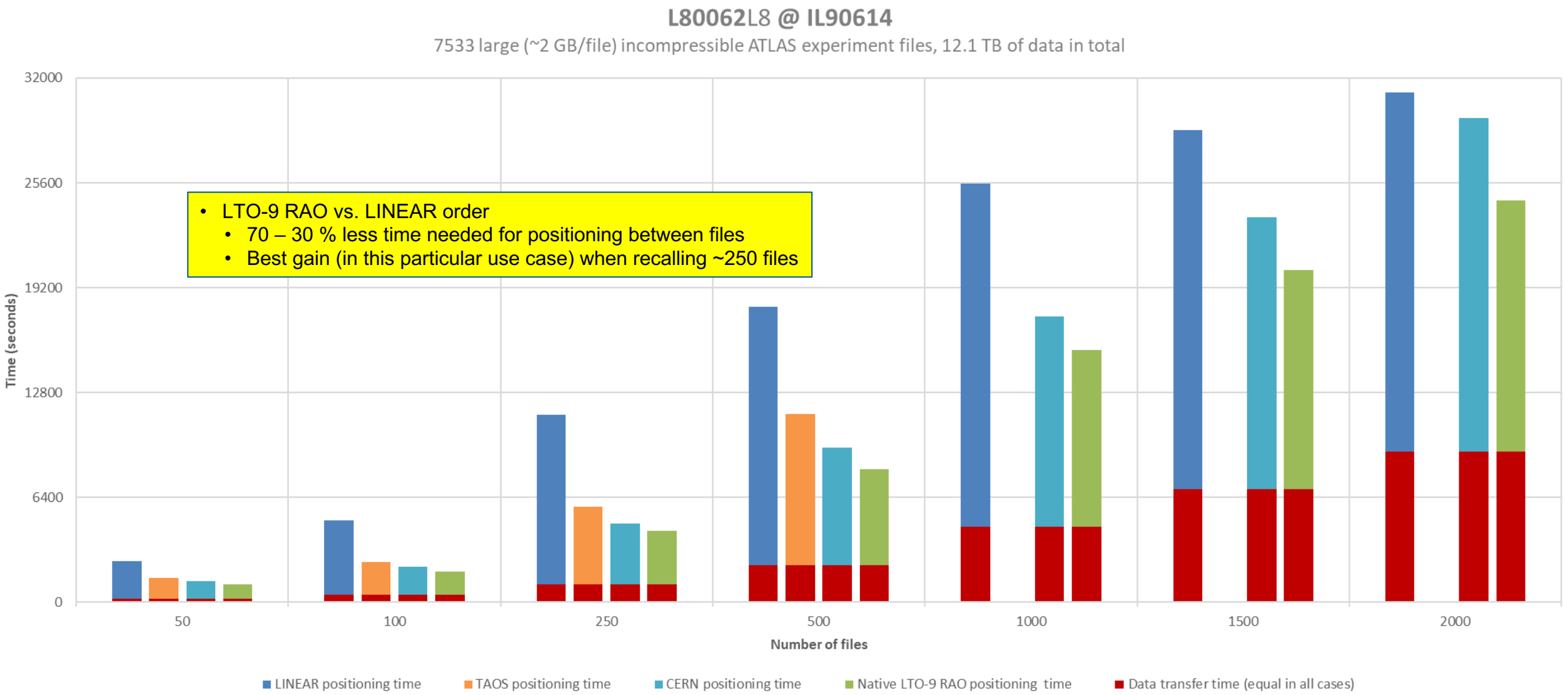
LINEAR



RAO



LTO-9 RAO comparison test results



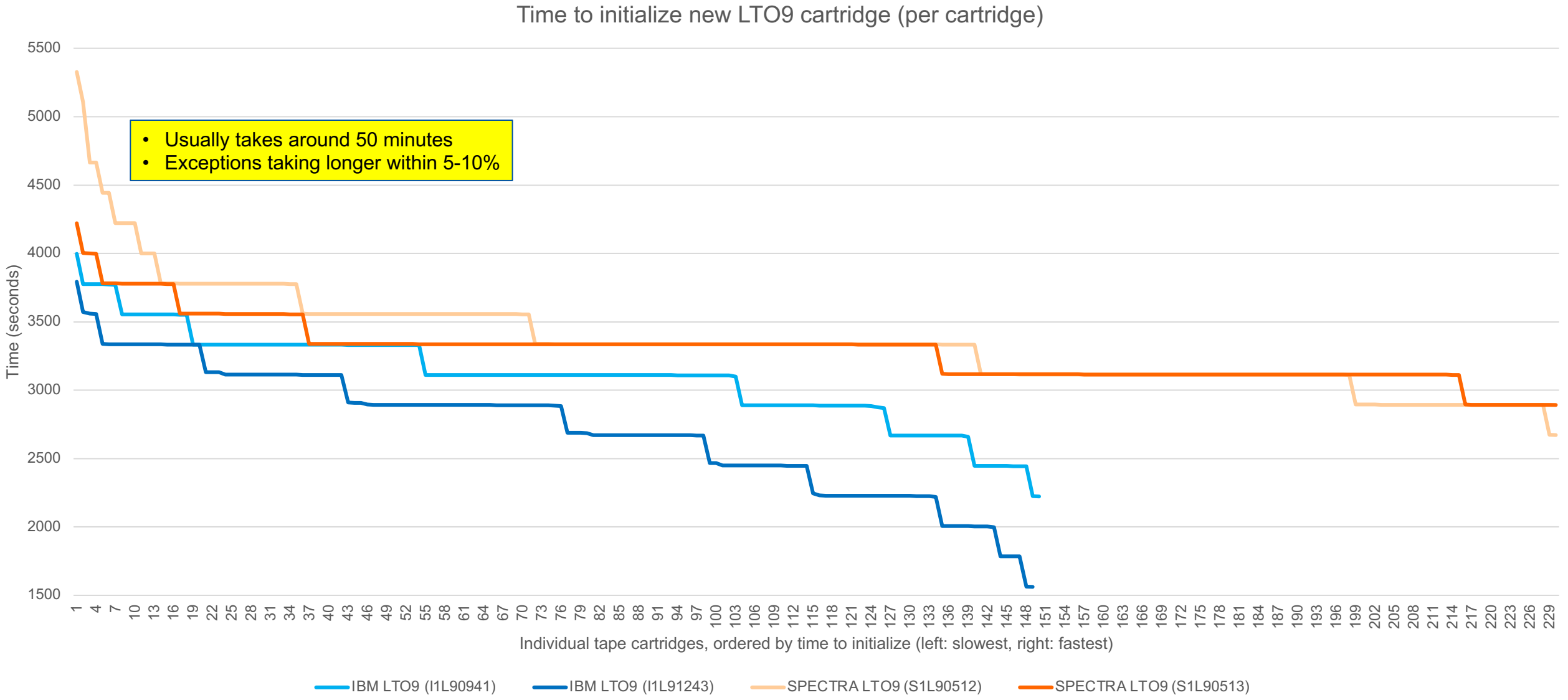
Hardware features vs. Software updates

IBM quote: *Backups are important, but Restores are essential.*

- RAO available since at least 2017
- Benefits demonstrated by CERN (CASTOR), BNL (HPSS) and others

- What about the backup software providers?

LTO-9 media initialization / calibration



Media initialization vs. Other industries

LTO9 media initialization has a cost

- Higher purchase price
- Device not immediately available when delivered

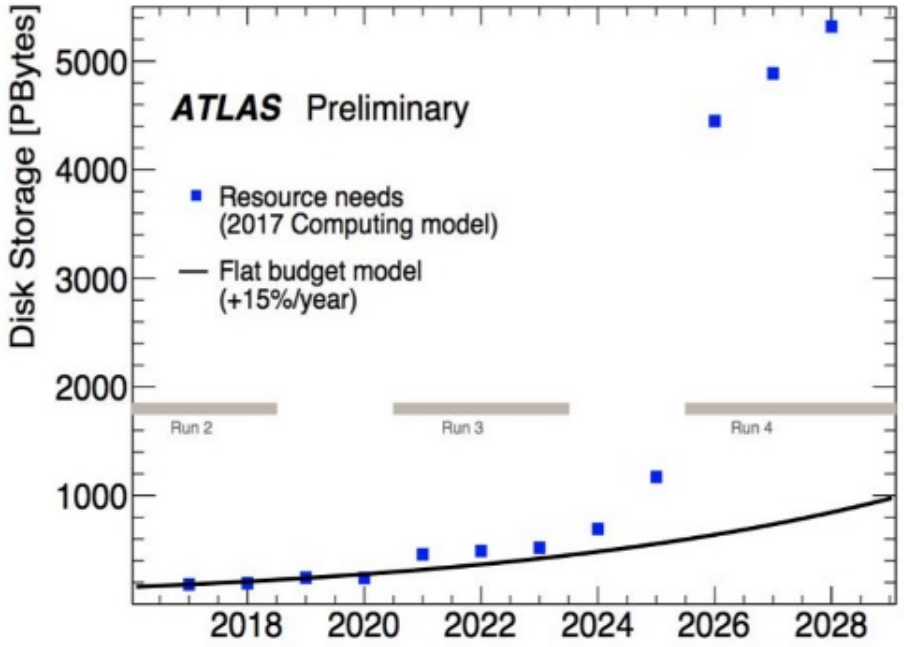
Engineering effort should be found to eliminate it

Store data for *online analysis* on tape

What is ‘data carousel’ and why ?

Data storage challenge of HL-LHC :

- ➔ ‘Opportunistic storage’ basically doesn’t exist
- ➔ Format size reduction and data compression are both long-term goals, require significant efforts from the software and distributed computing teams
- ➔ Tape storage is 3~5 times cheaper than disk storage, increasing tape usage is a natural way to cut into the gap of storage shortage for HL-LHC

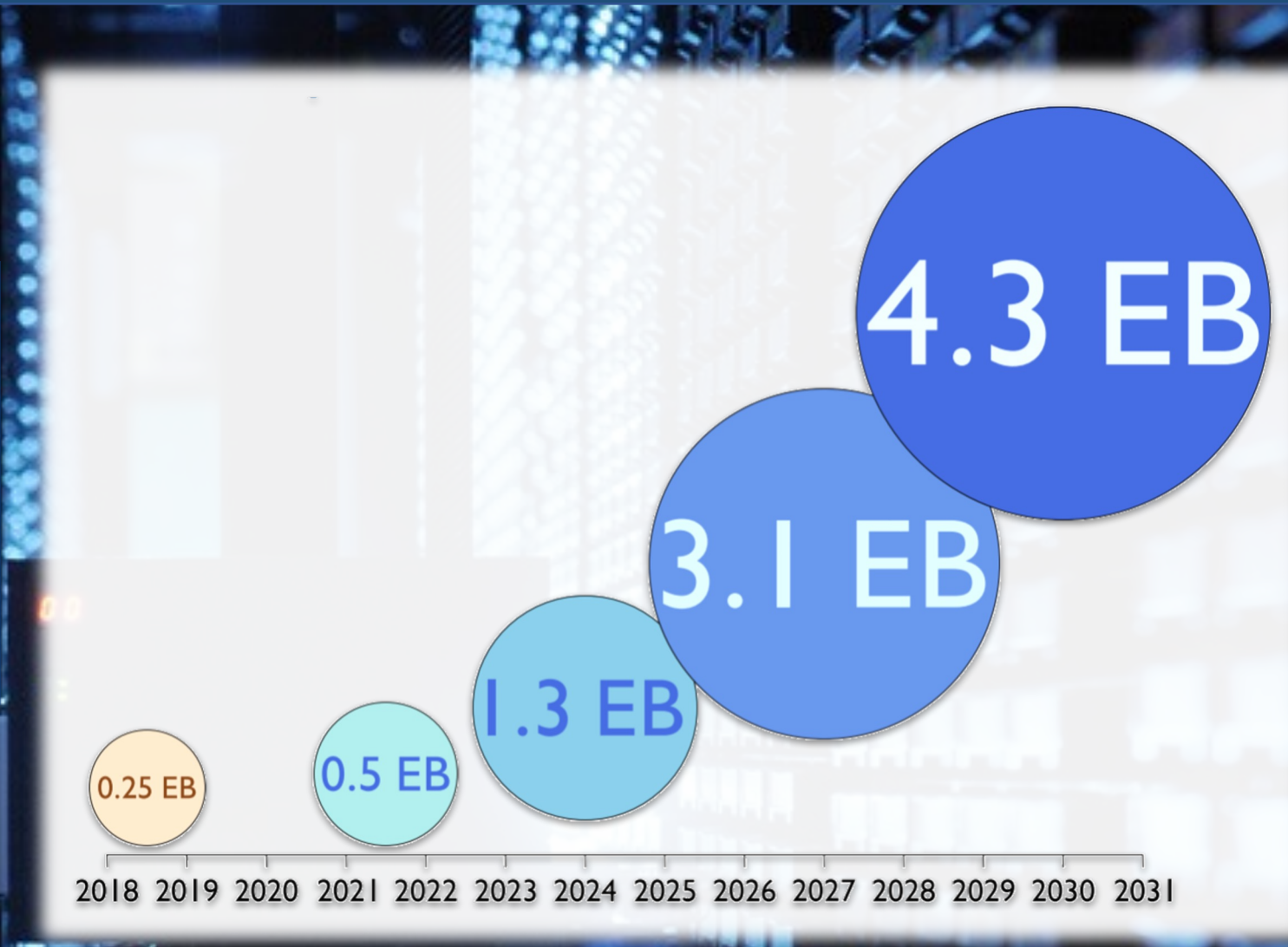


‘Data Carousel’ R&D → to study the feasibility to use tape as the input to various I/O intensive workflows.

Archive Growth Projections

2022 – 2025

- Up to 180 PB/year
- Up to 40 GB/s



Summary

- Tape is the best currently-available technology for archival storage, in terms of reliability, stability over long periods of time and cost
- CERN is investing in tape as its primary archival storage medium for LHC Run–3 (2022–26) and HiLumi LHC (2029–32)
- Storage needs are growing but budgets are flat
  - The CERN physics archive is ~520 PB but will soon grow to 1 EB
  - Data retrievals already exceed 1 EB/year
  - The storage demands of HL-LHC will mean more data on tape and new tape workflows

Support the CERN & Society Foundation with a donation of €30 or more and get an authentic LHC Data Tape souvenir today!

CERN & Society Foundation

*LIMITED TIME OFFER FOR DISTRIBUTION IN EUROPE