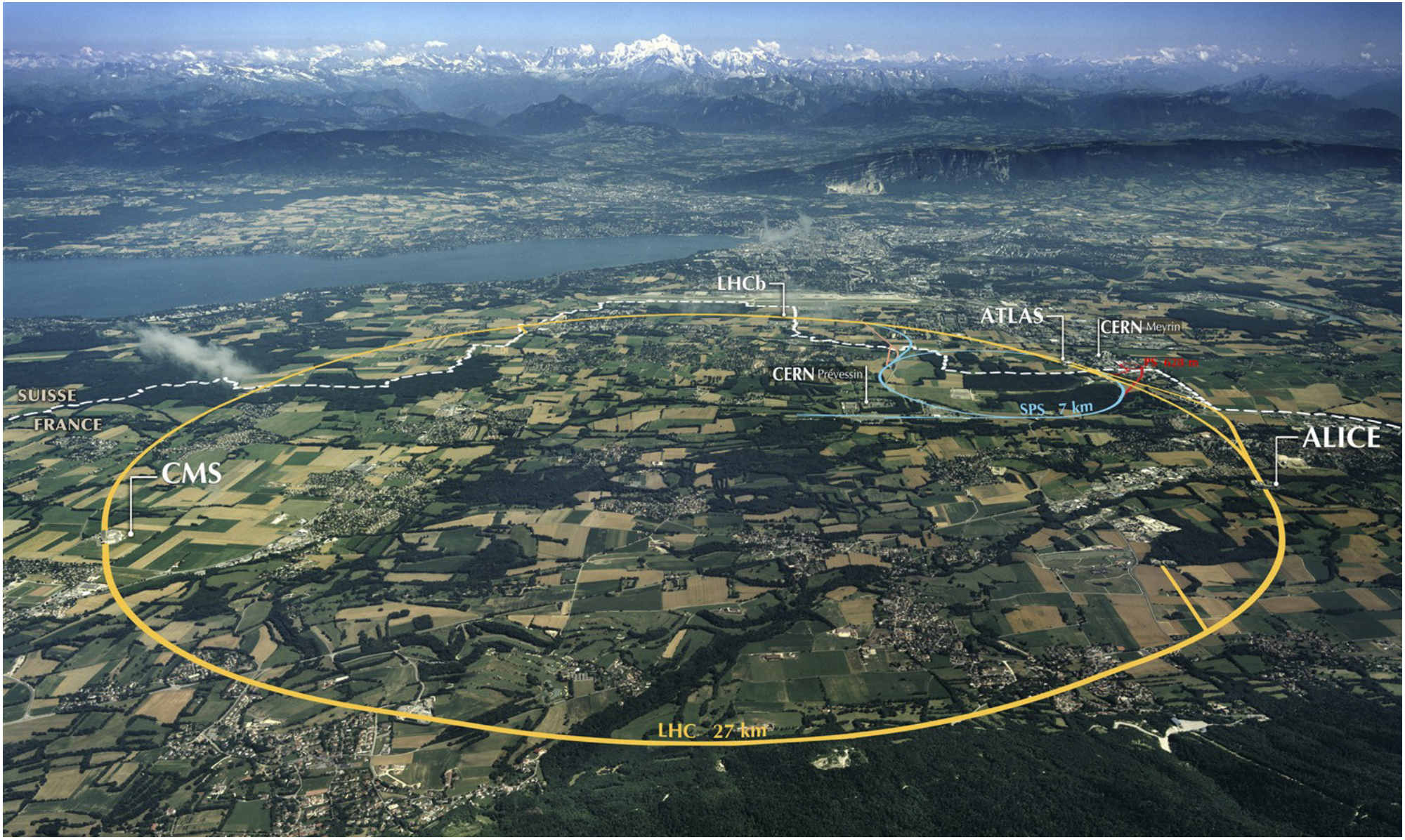
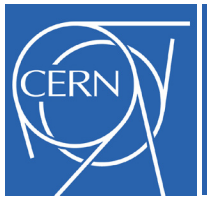

CERN

CERN Data Archive Challenges



Vladimír Bahyl



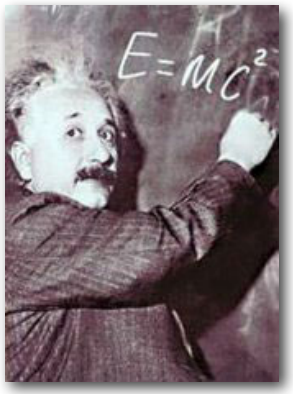
Agenda

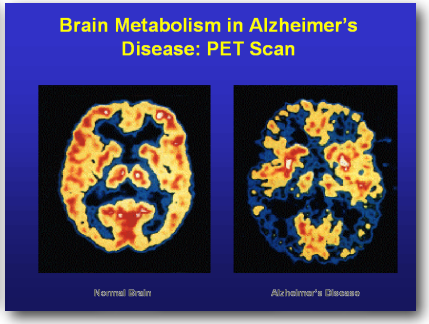
- What is CERN and what it does?
- Role of tape @ CERN
- Challenges
  - Performance
  - Data Protection
  - Lifecycle
- Conclusion


CERN: founded in 1954: 12 European States
"Science for Peace"
Today: 22 Member States
~ 3310 staff
~ 13580 scientific users
Budget (2017): ~1142 MCHF
Member States: Austria, Belgium, Bulgaria, Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Israel, Italy, Netherlands, Norway, Poland, Portugal, Romania, Slovak Republic, Spain, Sweden, Switzerland, United Kingdom
States in accession to Membership: Cyprus, Serbia, Slovenia
Associate Membership: India, Pakistan, Turkey, Ukraine
Observers to Council: Japan, Russia, USA; European Union, JINR, UNESCO

The Mission of CERN

- **Push forward** the frontiers of knowledge
  - E.g. the secrets of the Big Bang…
  - What was the matter like within the first moments of the Universe's existence?
- **Develop** new technologies for accelerators and detectors
  - Information technology –
  - Medicine – diagnosis
- **Train** scientists
- **Unite** people from different countries and cultures





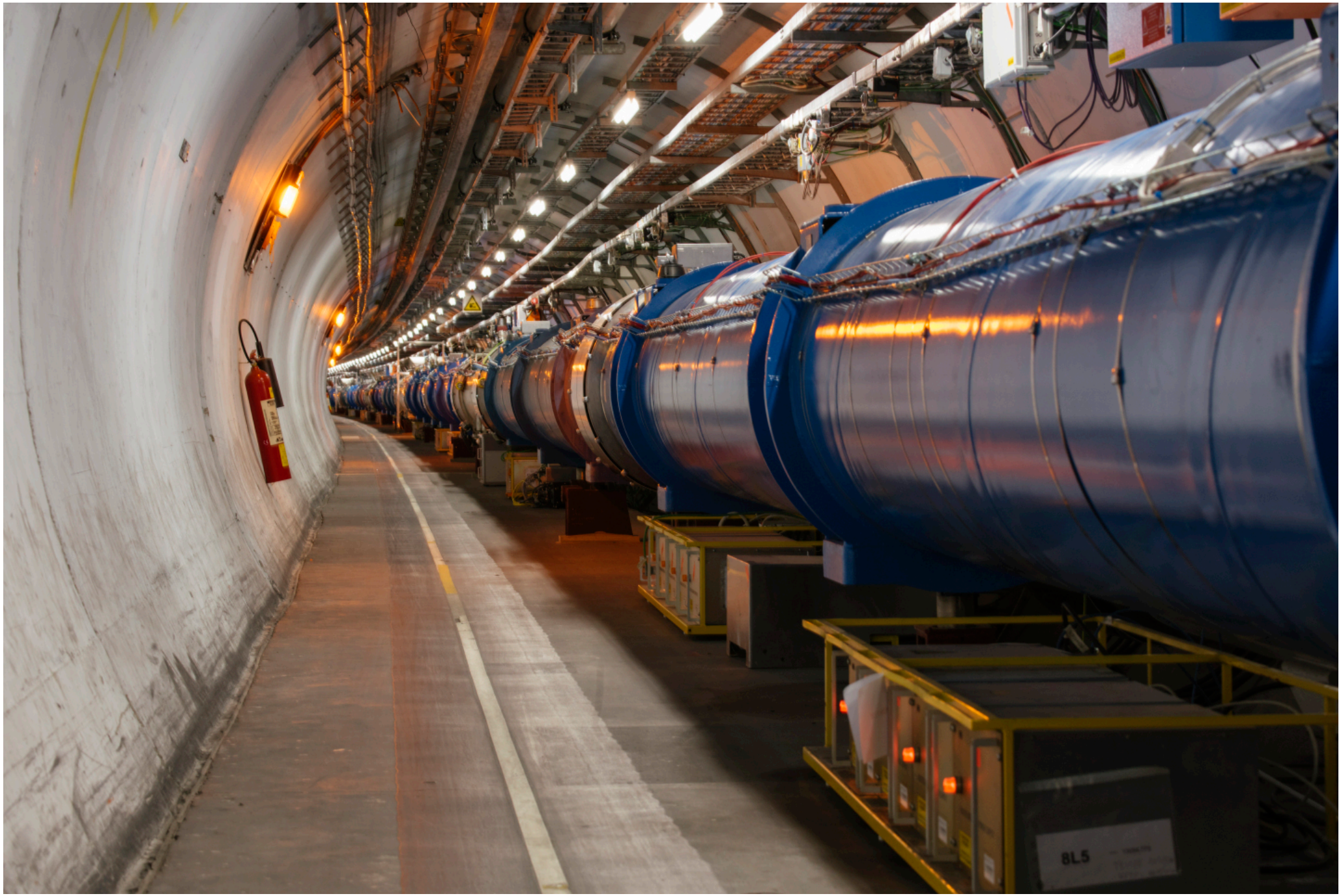
Innovation

Education

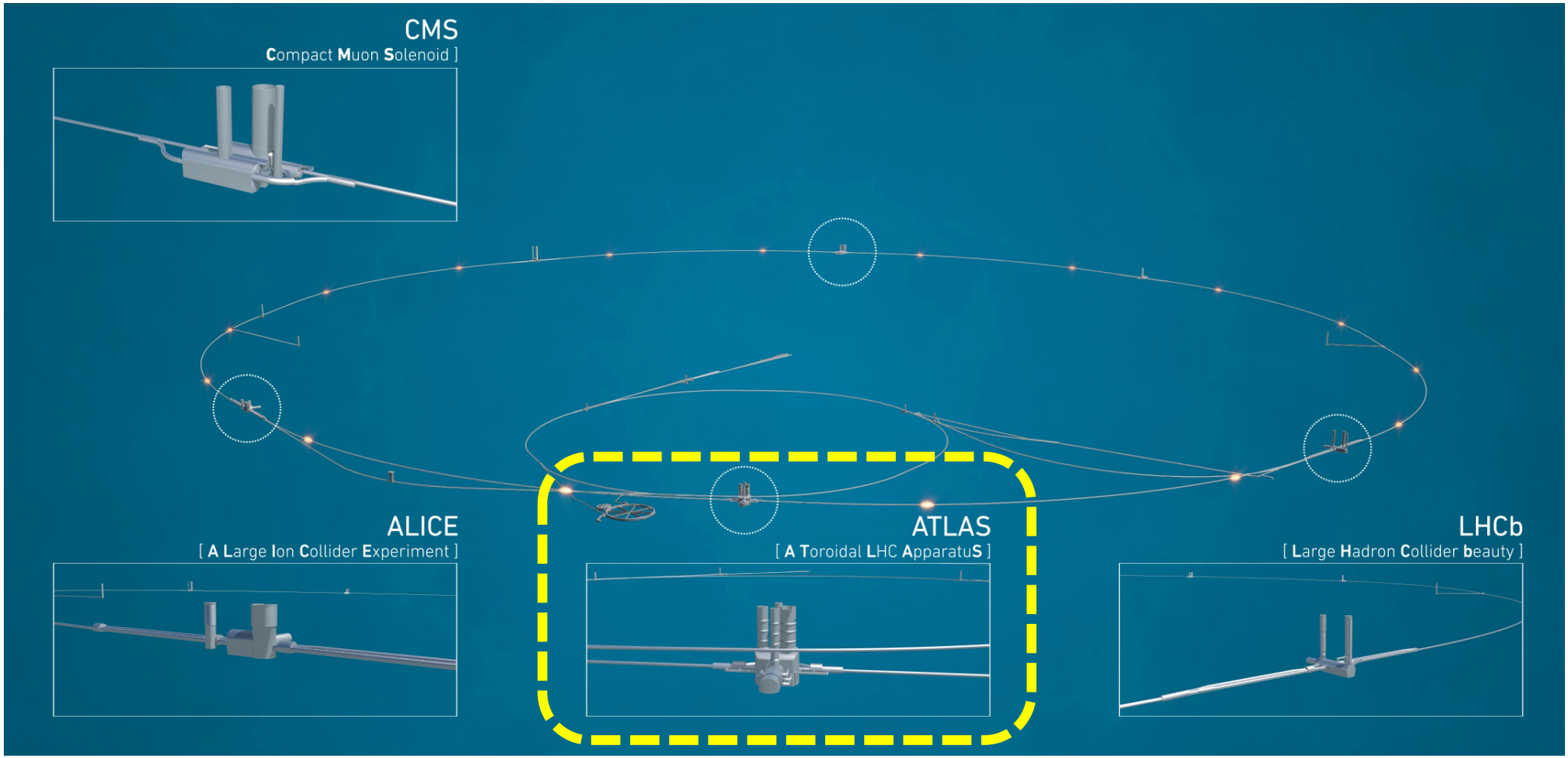
CERN

uniting people

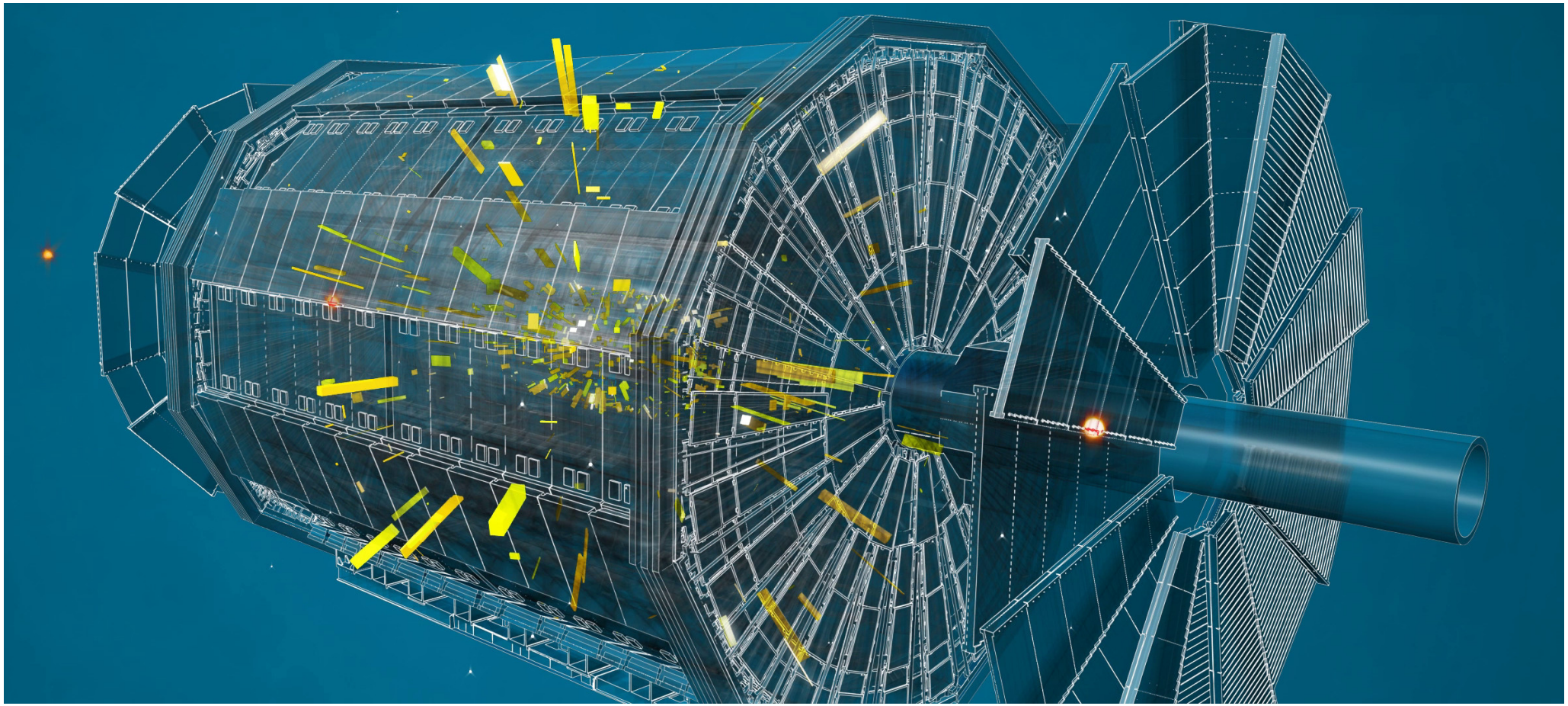
Research


8L5

Creation of the data in the detectors 1/2



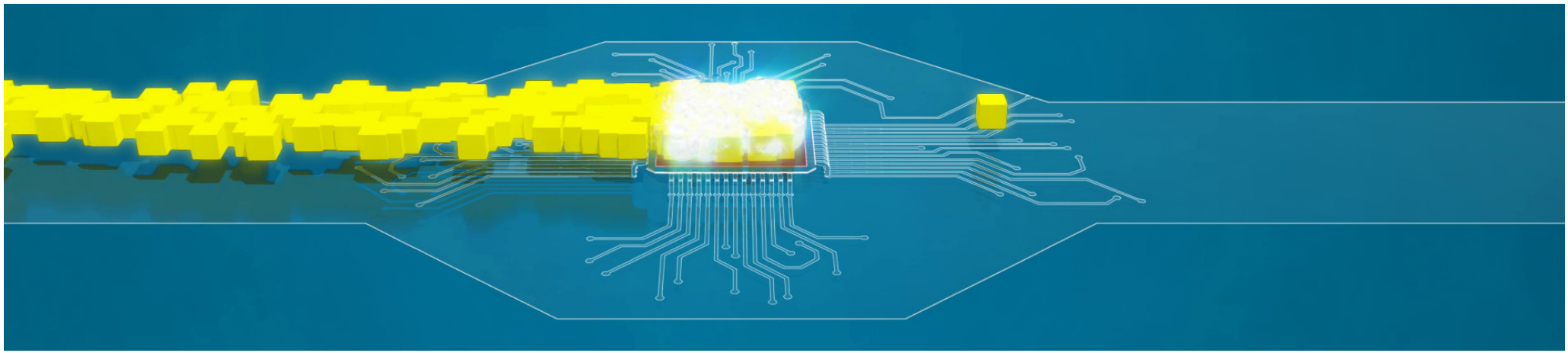
Creation of the data in the detectors 2/2



Filtering – 1st level

X PB/s

XX TB/s

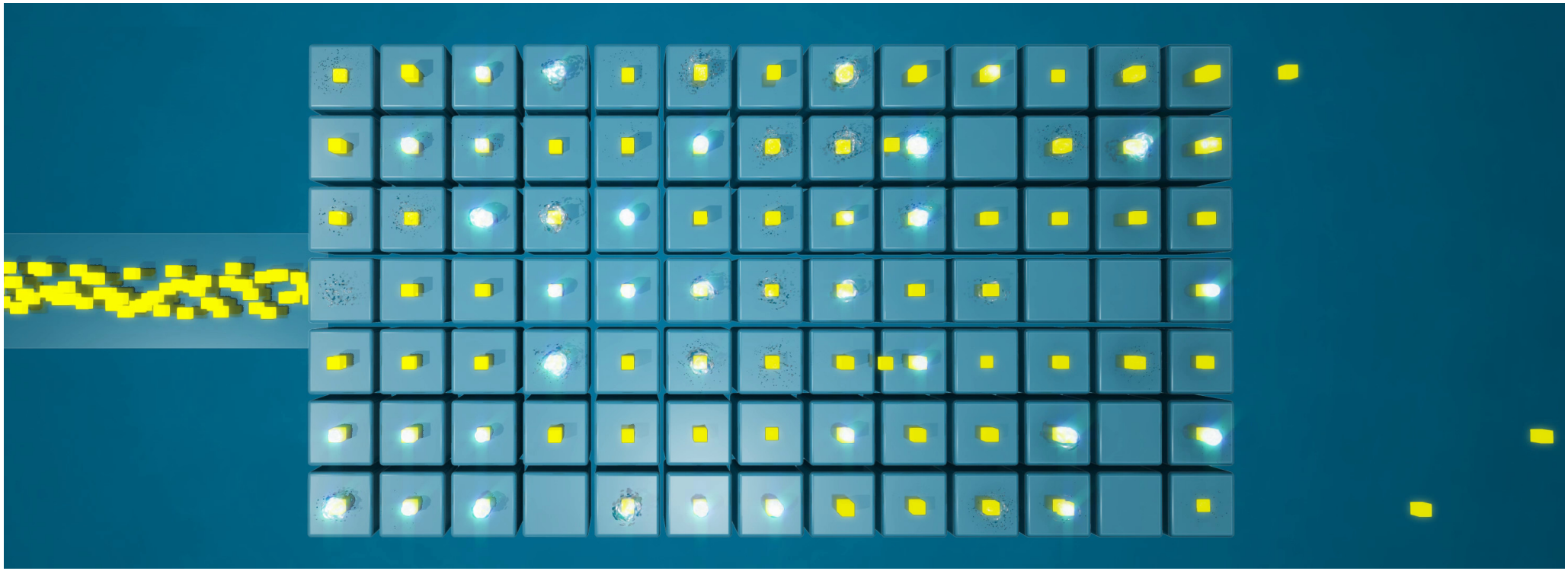


Level-1 Trigger
(hardware based)

Filtering – 2nd & 3rd level

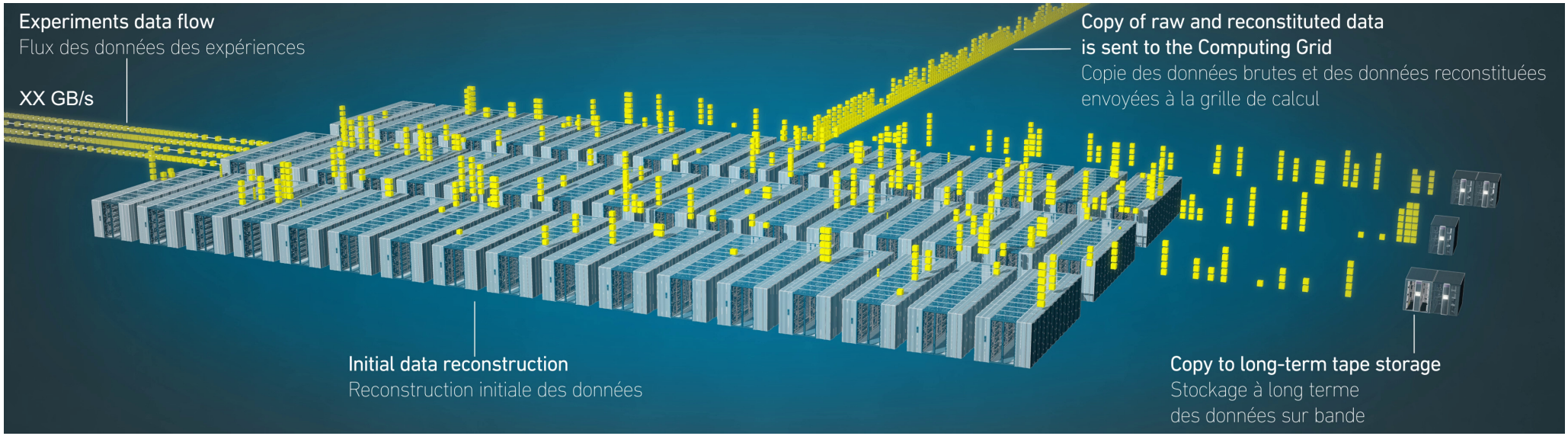
XX TB/s

XX GB/s

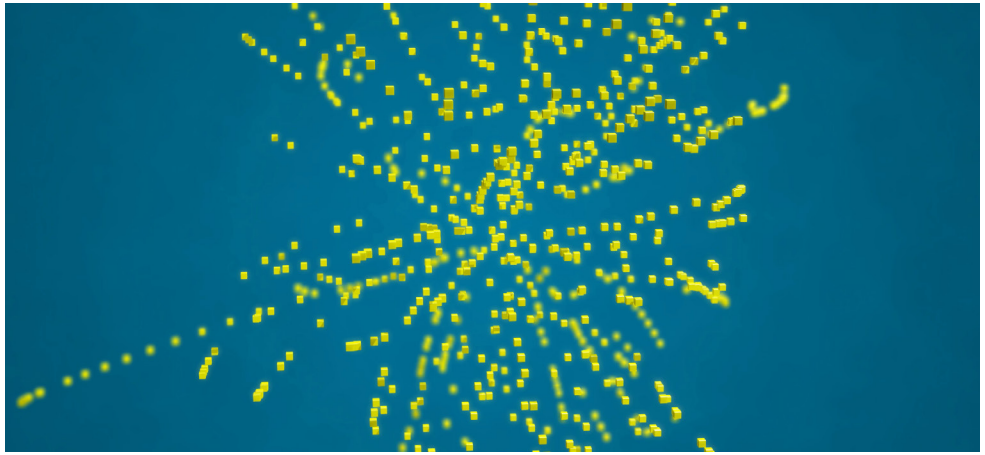


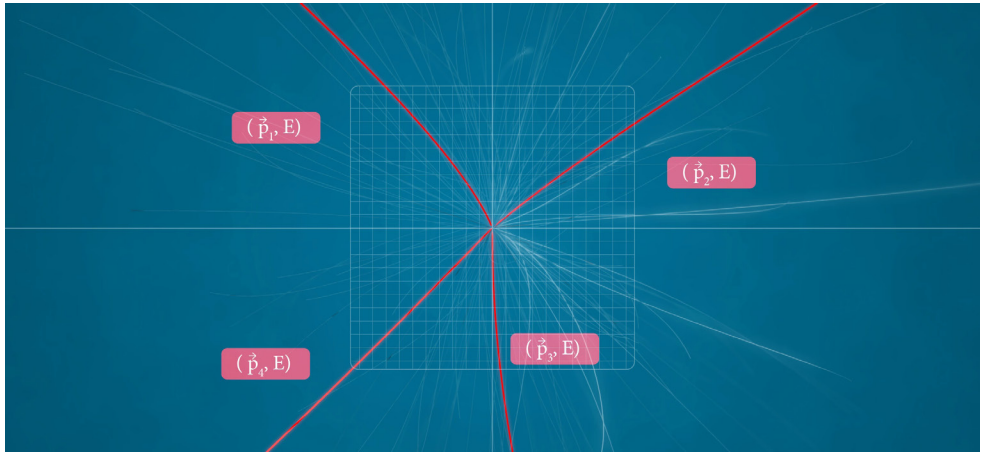
Level-2 & Level-3 Trigger
(computing cluster close to the detector)

Data distribution and archival

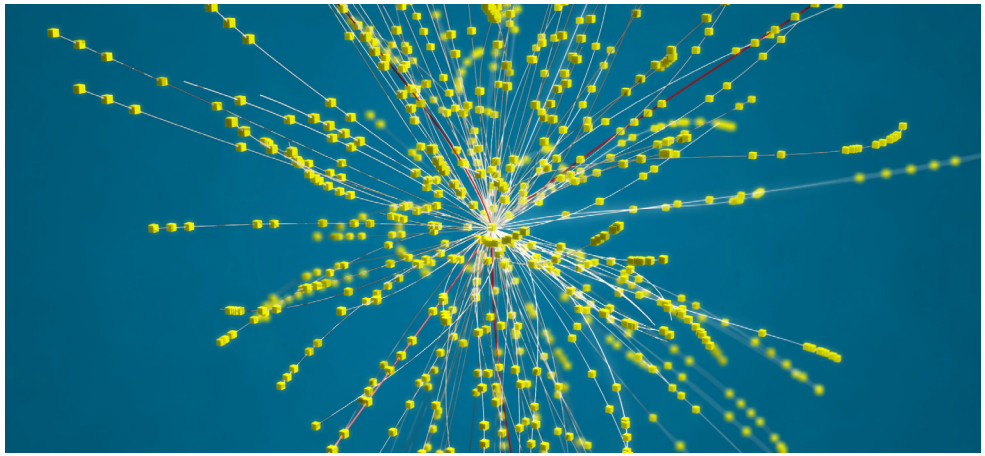


Reconstruction





Track Finder



Track Fitter

Reconstructing particle trajectories

Collaboration with the world – WLCG

CERN Tier-0 (Geneva & Budapest):

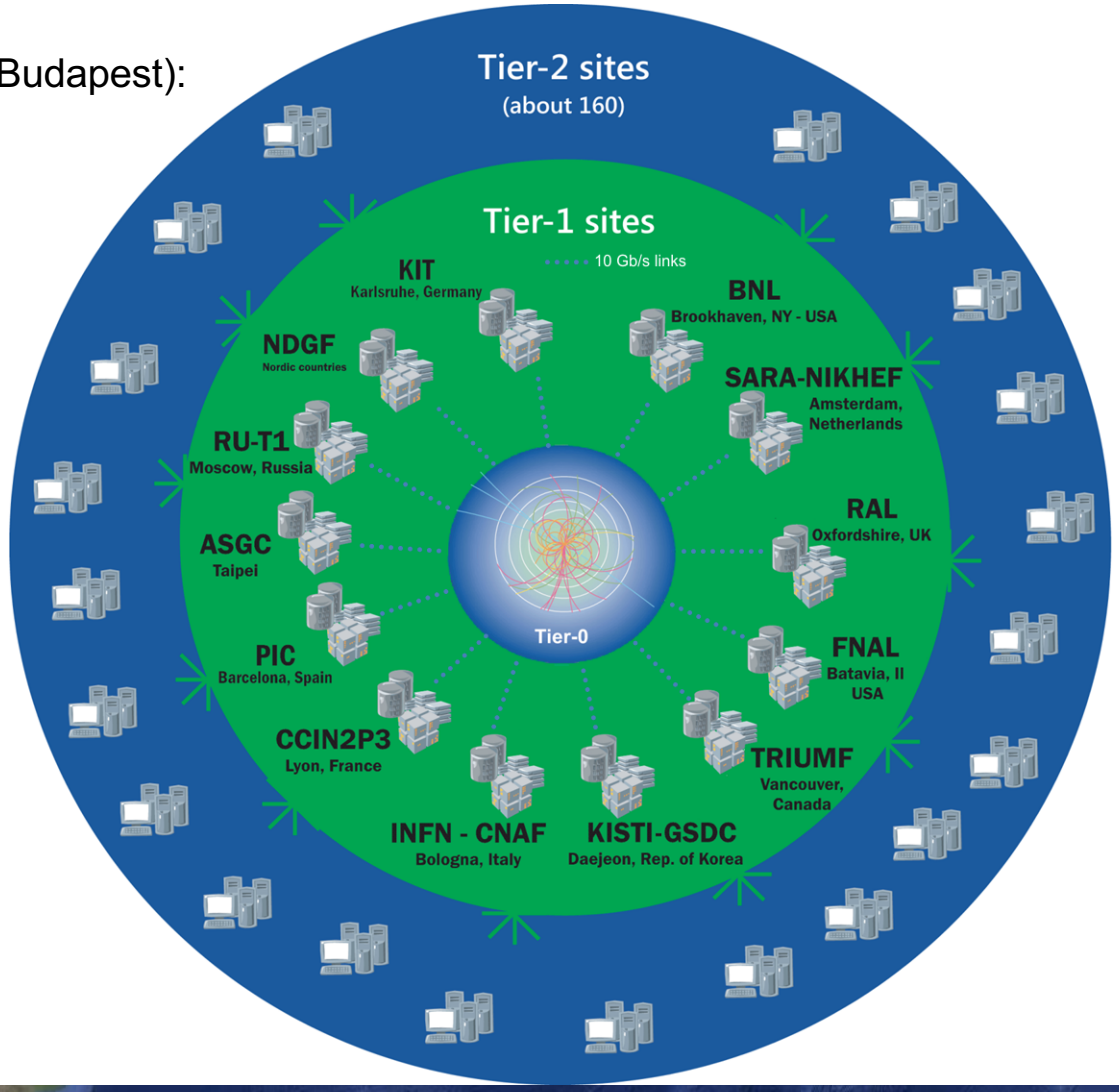
Data recording
Reconstruction
Distribution

Tier-1:

Permanent storage
Reprocessing
Analysis

Tier-2:

Simulation
End-user analysis



40 countries
~170 sites

~500k CPU cores
~500 PB of storage

> 2 million jobs/day

10-100 Gb links

An international collaboration to distribute and analyse LHC data.

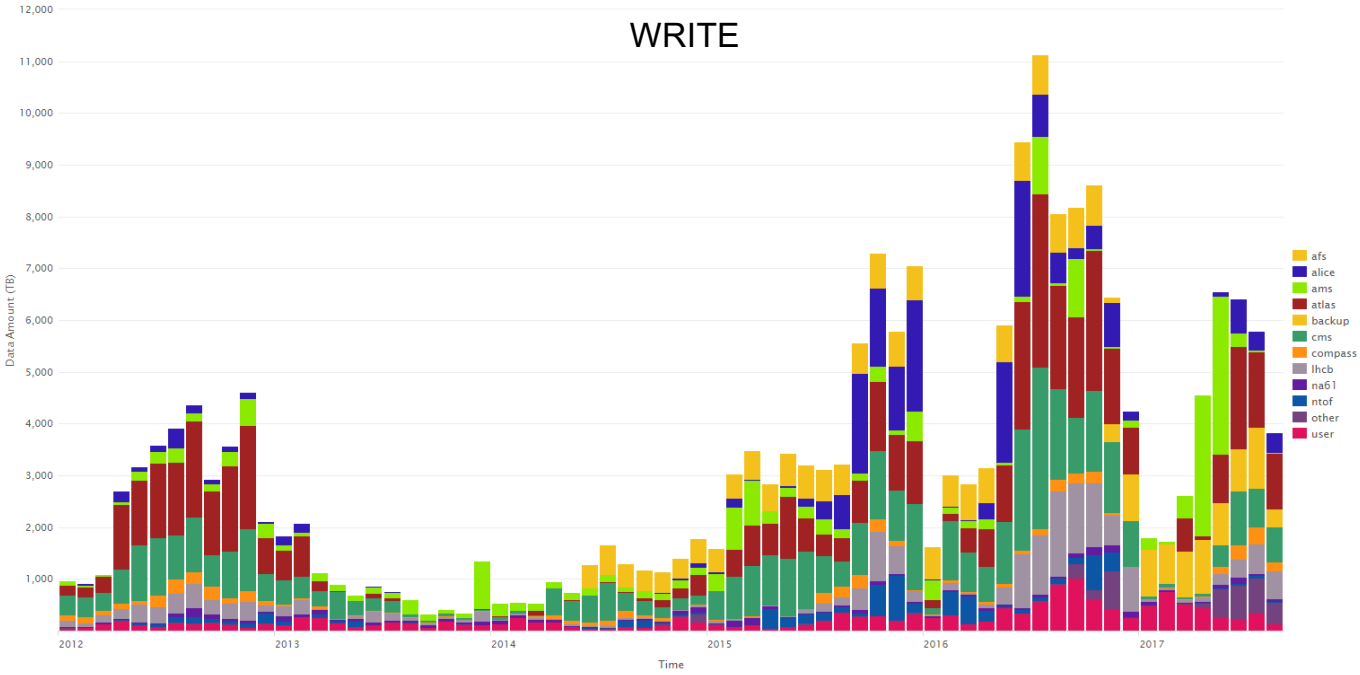
Integrates computer centres worldwide that provide computing and storage resource into a single infrastructure accessible by all LHC physicists.

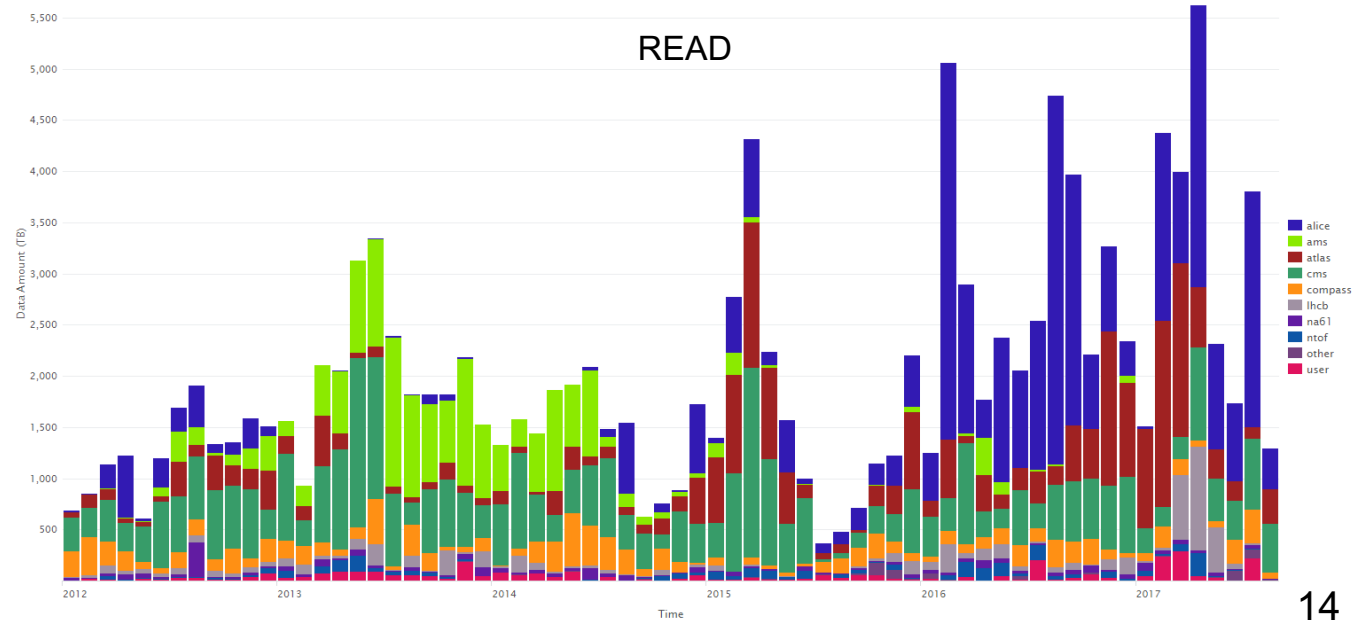
Role of Tape @ CERN

CASTOR archive:
- IBM
  - 1 x TS4500, 1 x TS3500
  - 46 x TS1150
  - 10000 x JD media (10 TB)
    6000 x JC media (7 TB)
- Oracle
  - 4 x SL8500
  - 40 x T10000D
  - 10000 x T2 media (8 TB)
- 10 PB disk cache
- ~200 PB of data on tape
  ~30 PB of free space
- Over 7 PB of new data per month
- Peaks of up to 7 GB/s to tape
- Lifetime of data: infinite

TSM backup:
- IBM
  - 2 x TS3500
  - 55 x TS1140
  - 200 x JC, 12000 x JB
- 8 PB; ~2300 M files
- 18 x TSM 7.1.4 servers

Main file or object storages:
- AFS 450 TB, 3 G files
- EOS 40 PB, 450 M files
- Ceph 360 TB, 80 M objects





Challenges

Performance

- Writing
- Reading

Data Protection

- Media Verification
- Logical Block Protection
- Failure Prediction

Lifecycle

- Media Migration
- Environment

Writing performance (data taking)

- Disk buffer made of commodity hardware
  - SATA drives have large capacity but low transfer rates
  - Replacing RAID soutions with RAIN
    - Inter-node traffic can be an issue
  - Using flash for special use cases (repack)
- Tape backhitch issue with small files
  - Drives have various optimizing features as well as multiple matching speeds
  - Best is to modify the application
    - Writing file marks with immediate bit set so the drive does not stop
    - Buffer data on the disk layer until the synchonizing file mark
- The goal is to find the sweet spot
  - Purchase only the minimum number of drives
  - … but what about the repack needs?

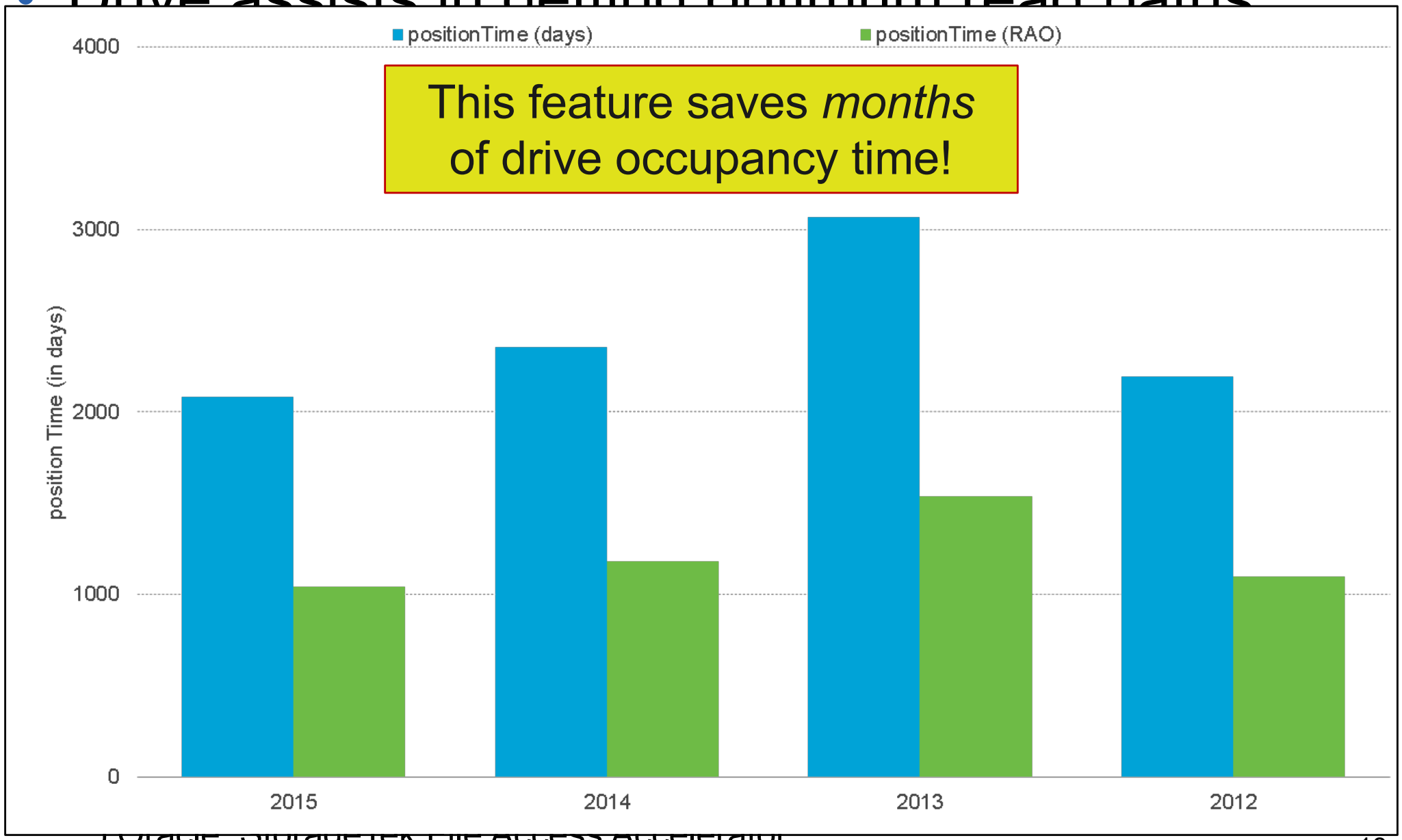
Reading performance (analysis)

- Data sets with multi-million files are spread over hundreds of tapes
  - Tapes may get (re-)mounted many times for few files

- Developed “traffic lights” to throttle and prioritise tape mounts
  - Thresholds for minimum volume and wait time, concurrent drive usage, group related requests
  - Separated archiving from end-user processing
  - End-users running low-latency jobs with high file access are migrated to a separate disk-only system

Recommended Access Ordering*

- Drive assists in getting optimum read paths



) Oracle: StorageTek File Access Accelerator

Challenges

Performance

- Writing
- Reading

Data Protection

- Media Verification
- Logical Block Protection
- Failure Prediction

Lifecycle

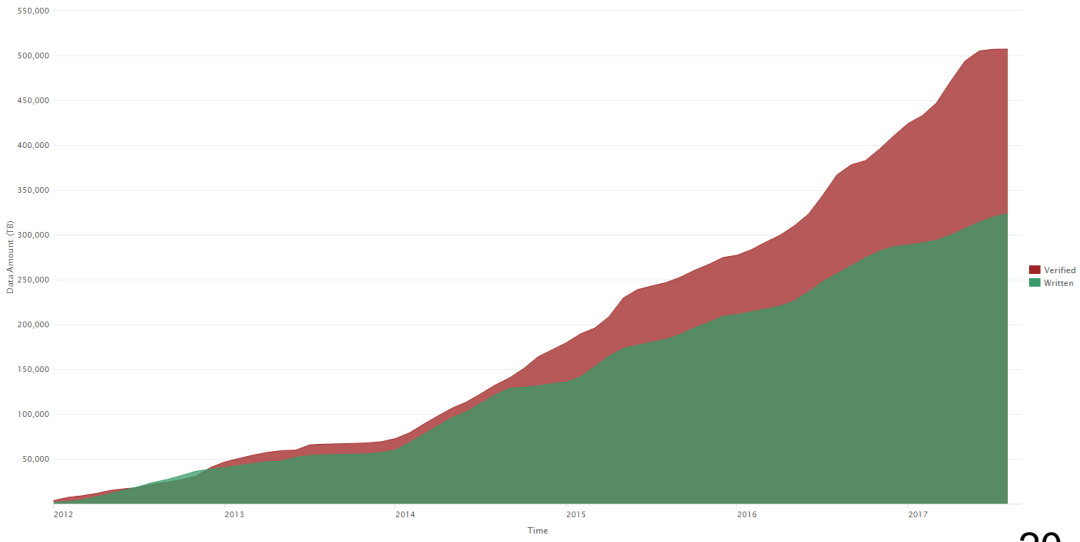
- Media Migration
- Environment

Media Verification

- Only ~20% of the data written to tape is read out by users.
  But data in the archive cannot just be written and forgotten about:
  - User: I have difficulties retrieving my file. Can you help?
  - Admin: Let me check… err, sorry, it seems we lost it.

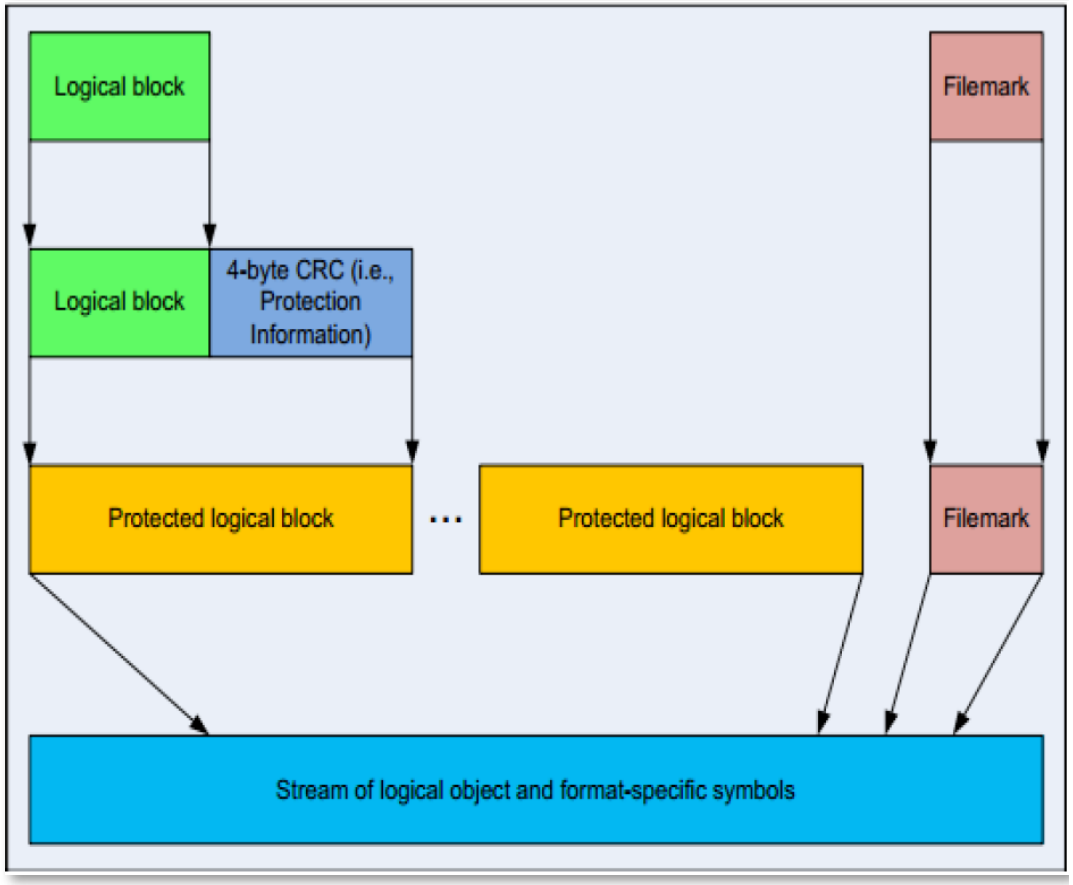
- A proactive and regular verification of the complete data archive is required to:
  - Ensure cartridges can be mounted
  - Check data can be read + verified against *our metadata* (checksum, file size, …)

- Two verification modes are supported:
  - *Full*: once tape is completely filled, or whenever it hasn't been accessed for a long time
  - *Partial*: Immediately after a tape has been written to, checking critical areas (beginning/end of tape)



Logical Block Protection

- Support for SCSI-4 Logical Block Protection (LBP)
- Protect against link-level errors eg. bit flips
- Data Blocks shipped to tape drive with pre-calculated CRC
- CRC re-calculated by drive (read-after-write) and stored on media; CRC checked again on reading
- Minimal overhead (<1%)
- Tape drive can do fast media verification autonomously
- Supported by newer LTO and enterprise tape drives



Failure prediction

- Re-engineered Tape Incident System
  - Taking full advantage of the SCSI tape alerts
  - Automated problem identification: tape vs. drive vs. library
  - Better detection of root cause → catch problems and disable faulty elements earlier
  - Enhanced low-level media repair tools

- Tried to exploit low-level tape system information
  - Transient/internal drive read/write/mount stats at SCSI level; library low-level logs
  - Assess the state of the drive and forecast a potential failure before it actually happens
  - Differences between Oracle and IBM – needs homogenization

tape="**I41398**" driveManufacturer="**IBM**" driveType="**03592E08**" firmwareVersion="**47A4**"
lifetimeBOTPasses="2495" lifetimeMOTPasses="2236" lifetimeVolumeMounts="327"
lifetimeVolume**RecoveredReadErrors**="**304**" lifetimeVolume**RecoveredWriteErrors**="**38**"
lifetimeVolume**UnrecoveredReadErrors**="**6**" lifetimeVolume**UnrecoveredWriteErrors**= "**2**"
volumeManufacturingDate="20110605"

  - Large log analysis did not lead to satisfactory conclusions

Challenges

Performance

- Writing
- Reading

Data Protection

- Media Verification
- Logical Block Protection
- Failure Prediction

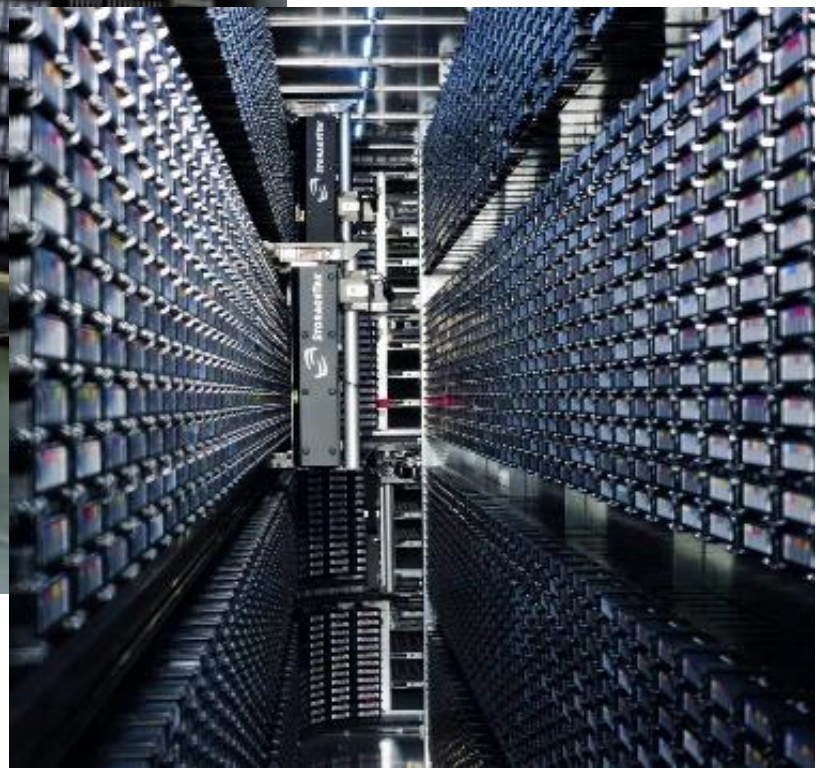
Lifecycle

- Media Migration
- Environment

Media migration (history)

1970's – GB's

1990's – TB's



Today – PB's

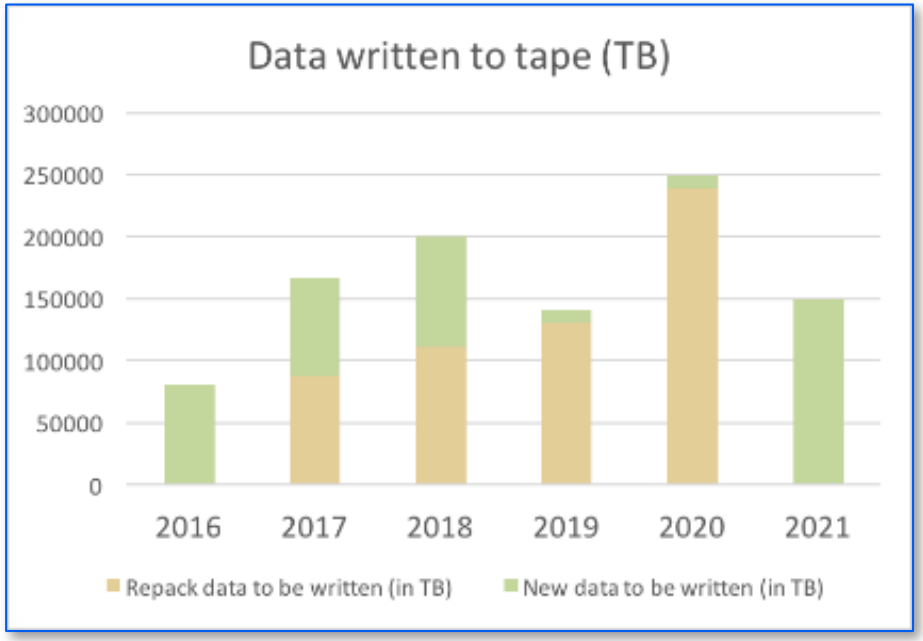
Media migration (repack)

- Challenge:
  - ~85 PB: 2013 – 51000 tapes → 2015 – 17000 tapes
  - Verify all data after write
    - 3 x (255PB!) pumped through the infrastructure (read->write->read)
  - Liberate library slots for new cartridges
    - Decommission ~35 000 obsolete tape cartridges

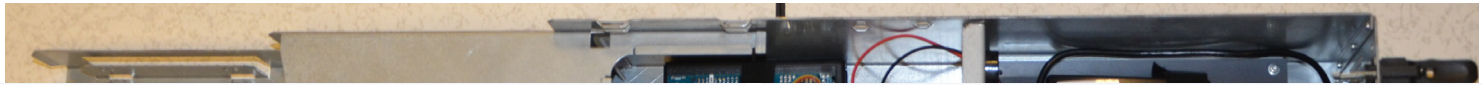
- Constraints:
  - Be transparent for user/experiment activities
  - Preserve temporal collocation
  - Finish before LHC run 2 start

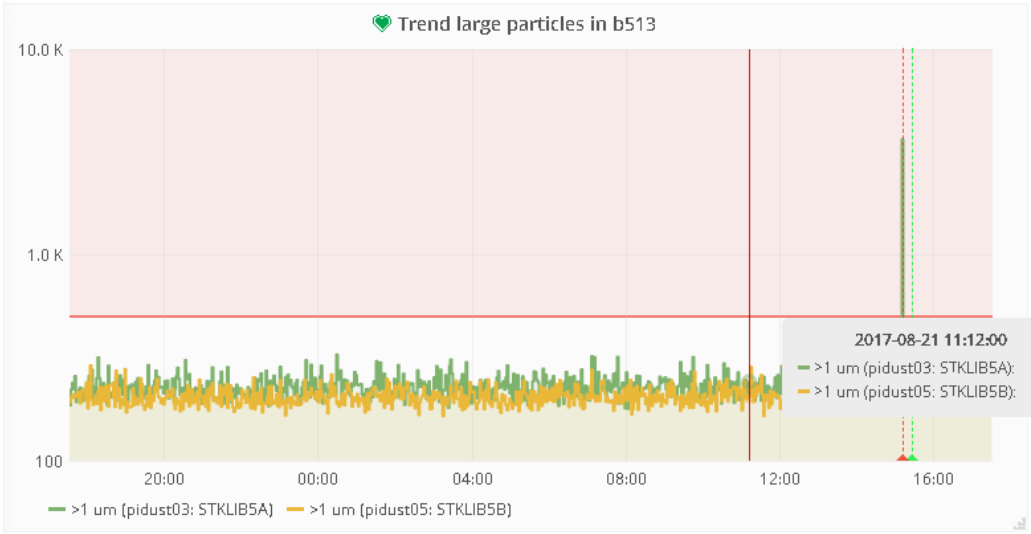
- LS2 (2019 – 2020) ideal moment for next repack (media replacement)
  - >350PB to migrate
  - New drives would advance the move

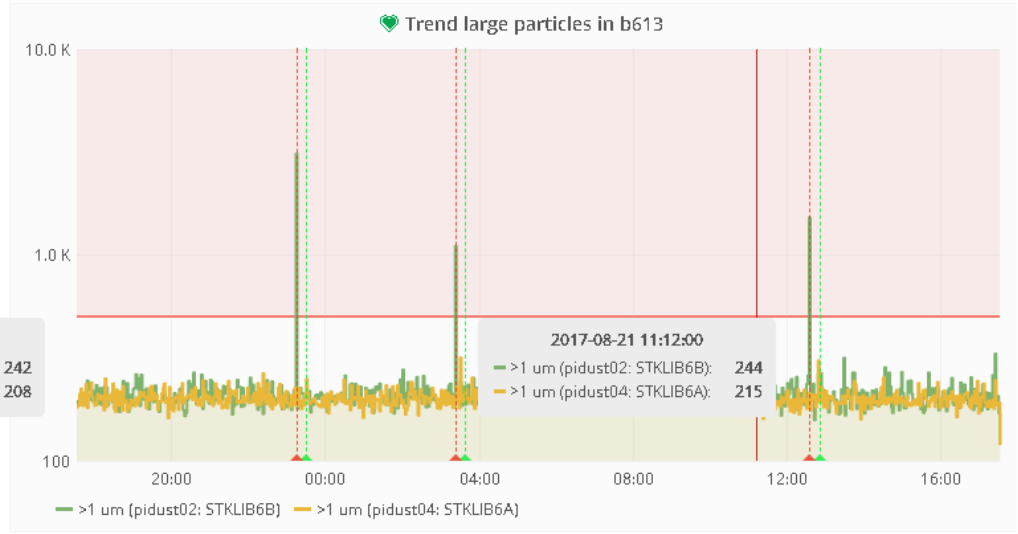
- Significant $aving$ to be made despite all this effort!

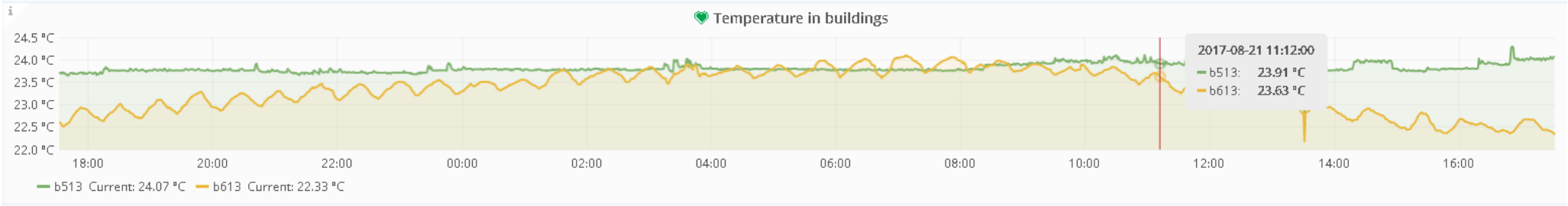


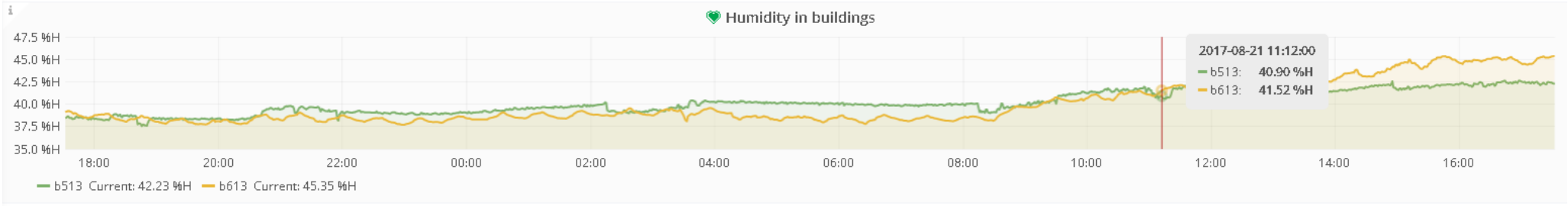
Environment











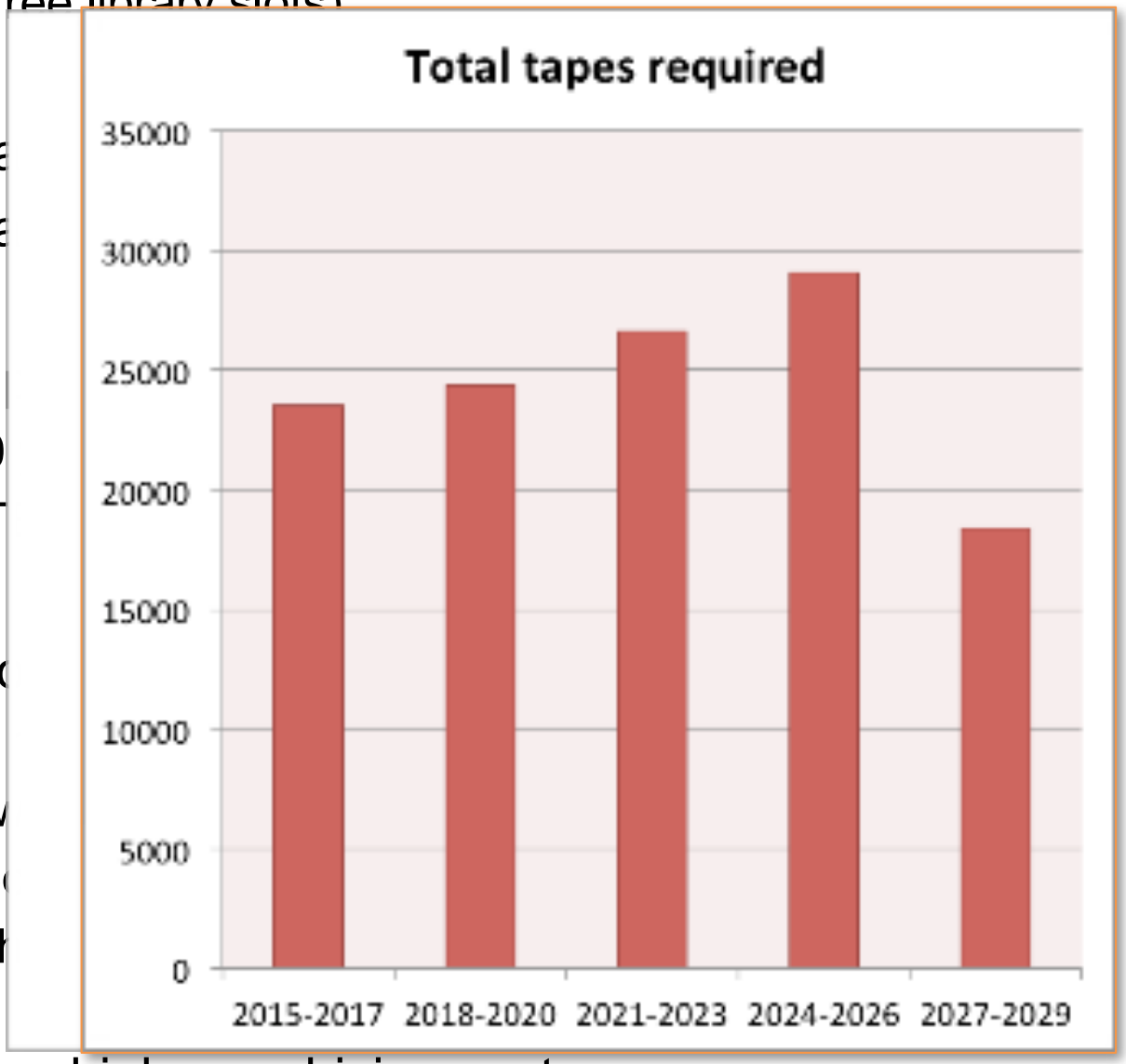
- Goal here is to report anomalies not to be 100% accurate

Future outlook

- Remaining Run-2 (–2018): 60-80PB/year of new data (LHC + non-LHC)
  - +5K tapes / year (~35'000 free library slots)
- 2019 – 2020: LHC upgrade
- Run-3 (– 2022): >150PB/yea
- Run-4 (2023 –): ~600PB/yea

- Tape technology roadmaps
  - ~30% CAGR for at least 10
  - Confirmed by recent ~330T

- Market evolution is difficult t
  - Low number of tape media
  - Disk market is basically dov
  - Cloud storage solutions? C
  - Disk capacity slowdown (wh products
- Storage capacity slowdown → higher archiving costs



Wishlist

- Enable RAO feature in LTO tape drives
  - Capacities grow faster than filesizes = retrieve many files = many seeks
- Improve the quality of failure prediction metrics
  - Higher capacity cartridges lose more data in an incident
- Focus on improving the products instead of the support processes
  - Use remote diagnostics
  - Consider the support engineers as your brand ambassadors
- Cloud providers should share more their tape solutions
  - They are the market leaders now