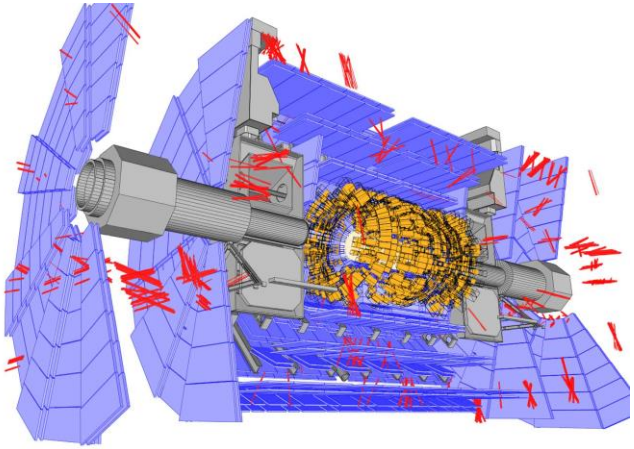
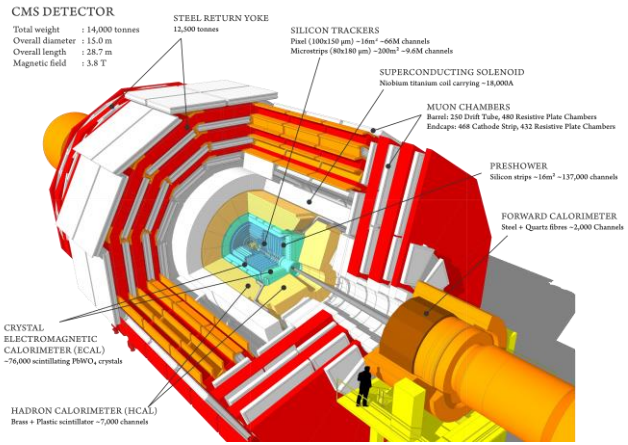
These slides at https://indico.cern.ch/event/666320/.

PV2020 @ CERN




CMS DETECTOR
Total weight : 14,000 tonnes
Overall diameter : 15.0 m
Overall length : 28.7 m
Magnetic field : 3.8 T
STEEL RETURN YOKE
12,500 tonnes
SILICON TRACKERS
Pixel (100x150 μm) ~16m² ~66M channels
Microstrips (80x180 μm) ~200m² ~9.6M channels
SUPERCONDUCTING SOLENOID
Niobium titanium coil carrying ~18,000A
MUON CHAMBERS
Barrel: 250 Drift Tube, 480 Resistive Plate Chambers
Endcaps: 468 Cathode Strip, 432 Resistive Plate Chambers
PRESHOWER
Silicon strips ~16m² ~137,000 channels
FORWARD CALORIMETER
Steel + Quartz fibres ~2,000 Channels
CRYSTAL ELECTROMAGNETIC CALORIMETER (ECAL)
~76,000 scintillating $PbWO_4$ crystals
HADRON CALORIMETER (HCAL)
Brass + Plastic scintillator ~7,000 channels

Goals

- Attract more scientific communities
- Broaden information exchange, sharing of experiences, tools and even services
- Keep in step with (or ahead of) funding agencies / policy makers in their push for LTDP & OD
- (Discussion at end)

Recent EU directive stresses:

- Open Access
- Open Data
- Preservation
- Rewards

- Full document: https://ec.europa.eu/digital-single-market/en/news/recommendation-access-and-preservation-scientific-information

Schedule

- Assume 2.5 day mid-week meeting as per “tradition”
- Co-located events: possible Monday, Thursday pm and Friday
- There are already some candidates!
- Define well in advance (travel, hotels etc.)

Possible co-located events

- EIRO / ESFRI workshop
- Show-casing of preservation project(s)
- …

Constraints

- 88th Geneva motor show early March
- “Summer time” Sunday 29th March (Europe), Sunday 8th March (US)
- Easter Sunday: April 12th (Orthodox April 19th)
- LHC Long Shutdown 2: 2019 / 2020

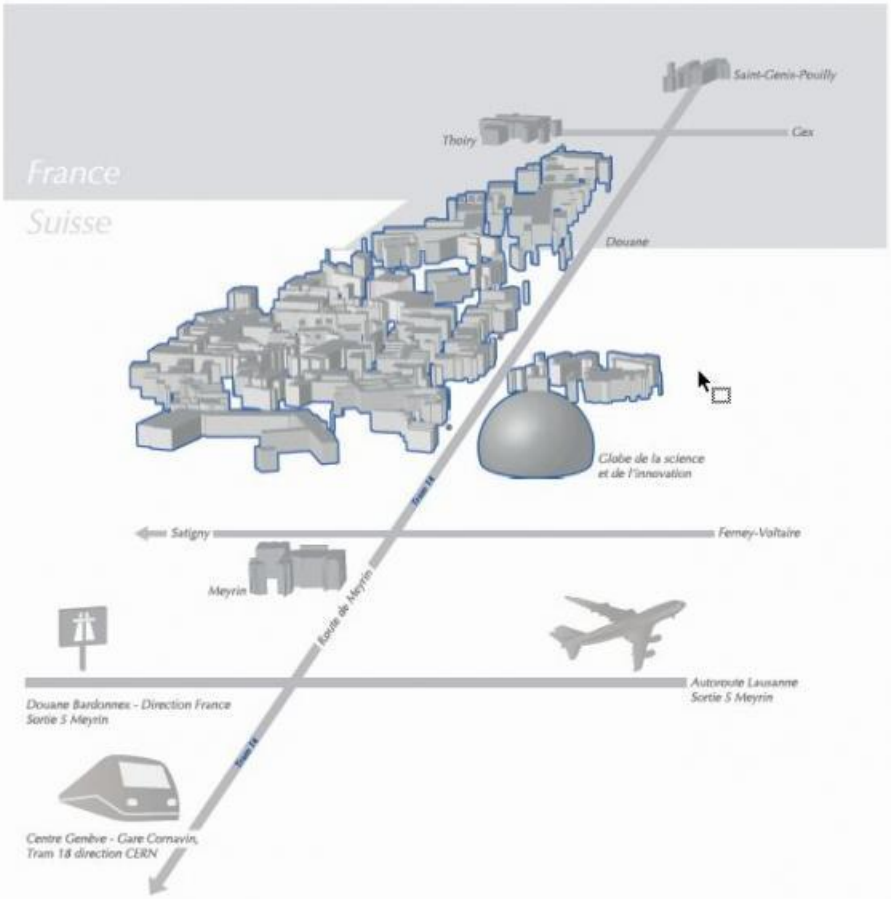
➢ **Opinions???**

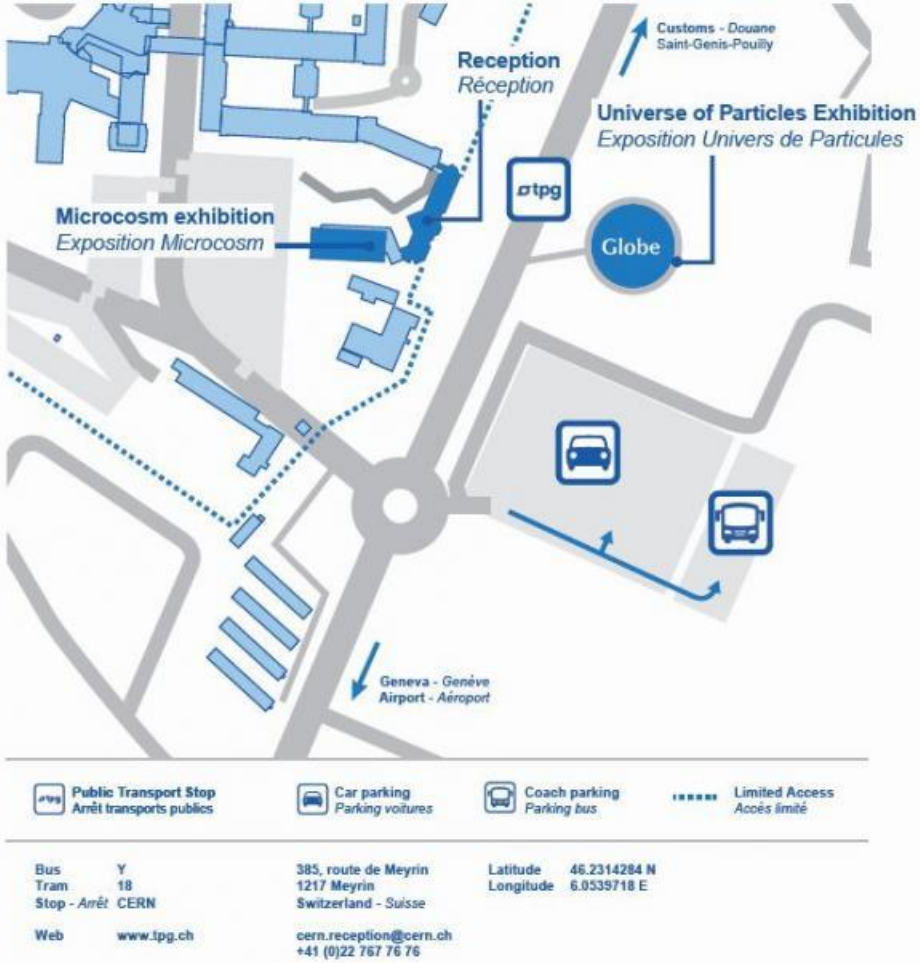
Travel etc.

- Geneva easily reached by air (& road & train)
- CERN easily reached by Tram 18 (FREE transport with your hotel IN GENEVA)
- Final stop: CERN opposite reception building

➢ **YOU WILL NEED YOUR PASSPORT TO GET YOUR BADGE WHICH MUST BE WORN**

How to get to CERN - *Comment se rendre au CERN*





Tram 18 from downtown to CERN

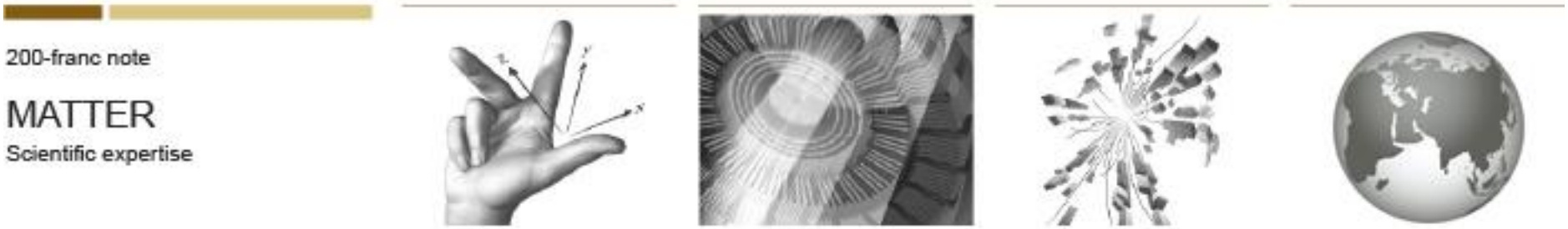
Network

- Eduroam and visitors' network available
- The former is strongly preferred!
- Your roaming mobile data contract will probably entail charges in CH (but F networks often available)

Money!

- Many places in CH take both Swiss Francs (CHF) as well as Euros – the latter often at favourable exchange rates!

- The reverse side of the 200 CHF note (from August 2018) will feature the LHC!



Key dates etc

- First need to agree / confirm conference dates then work back from there
- Conference committee, sessions, call for papers etc.
- Requirement for events at CERN to use Indico – including for registration / badges
- Can probably record / webcast sessions too (and archive all papers / presentations / videos)

Goals

- Attract more scientific communities
- Broaden information exchange, sharing of experiences, tools and even services
- Keep step (or ahead) of funding agencies / policy makers in their push for LTDP & OD

➢ **Discussion…**

Some background slides on CERN, LHC, storage, bit preservation
Borrowed from Germán Cancio on behalf of the CERN Storage Team.

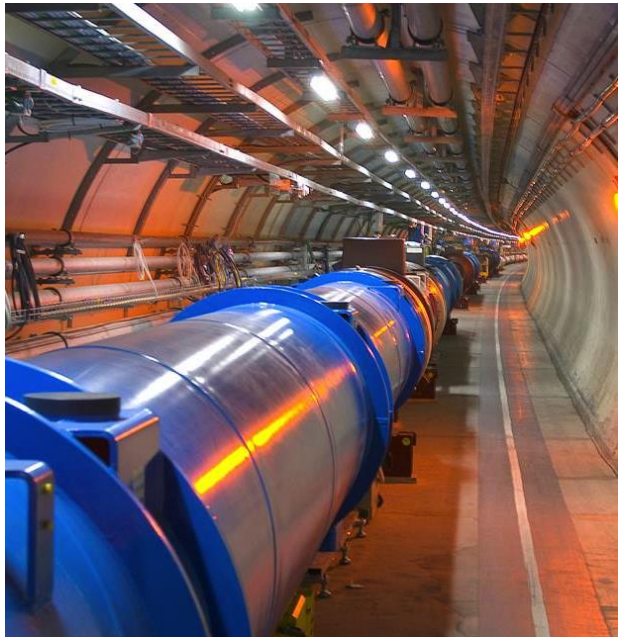
What is CERN?

- Conseil Européen pour la Recherche Nucléaire – aka European Laboratory for Particle Physics
- Between Geneva and the Jura mountains, straddling the Swiss-French border
- Founded in 1954 with an international treaty

- Our business is fundamental physics, what is the universe made of and how does it work…
  - How to explain particles have mass?
  - What is 96% of the universe made of? We can only see 4% of its estimated mass!
  - Why isn't there anti-matter in the Universe? Nature should be symmetric…
  - What was the state of matter just after the « Big Bang » ?

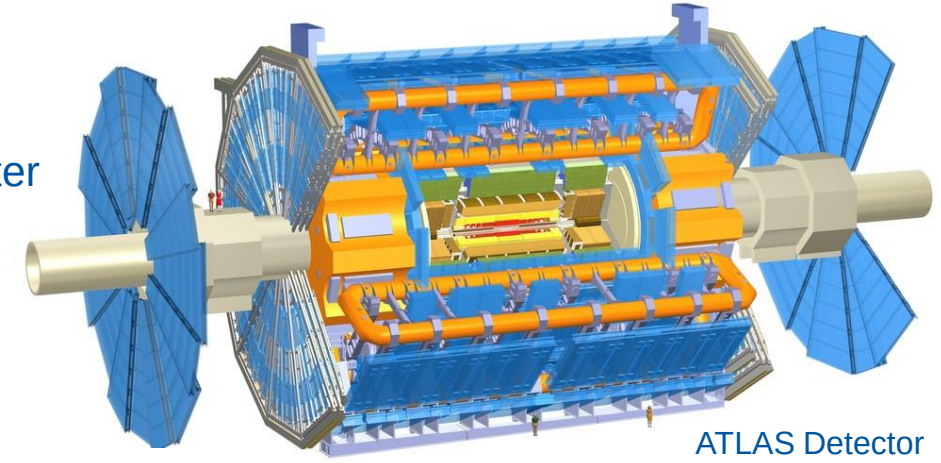
CERN Tools

- The world's most powerful particle accelerator: LHC
  - A 27 km long tunnel filled with high-tech instruments
  - Equipped with thousands of superconducting magnets
  - Accelerates particles to energies never before obtained



- 4 very large sophisticated Detectors ("Experiments")
  - Buried between 50-150m below earth
  - Sophisticated Data Acquisition systems

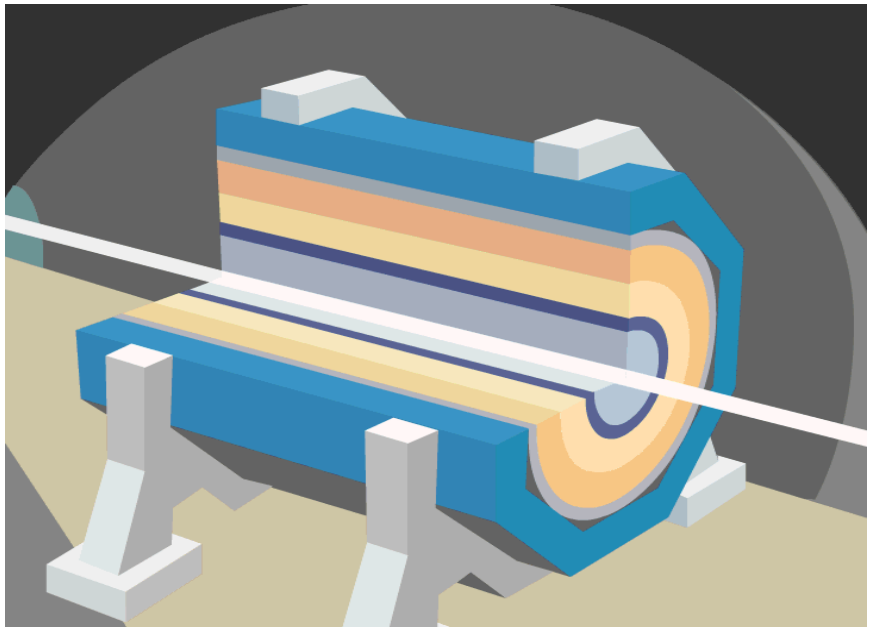
- Top level computing to distribute and analyse the data
  - Sufficient storage to handle massive amounts of data, making it available to thousands of physicists
  - A Computing GRID linking ~200 Computer Centres around the globe over WAN, as CPU capacity at CERN only 20% of total needed!



ATLAS Detector

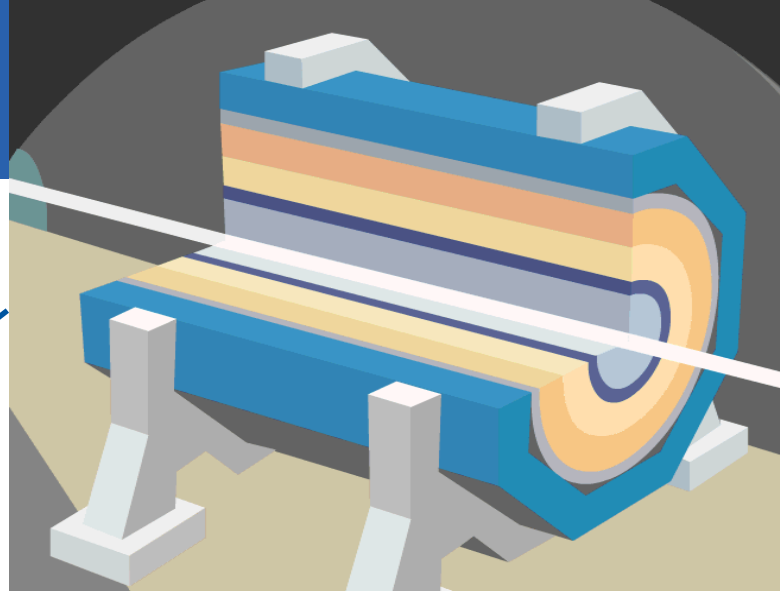
LHC Data

The accelerator generates 40 million bunch collisions (“events”) every second at the centre of each of the 4 experiments’ detectors

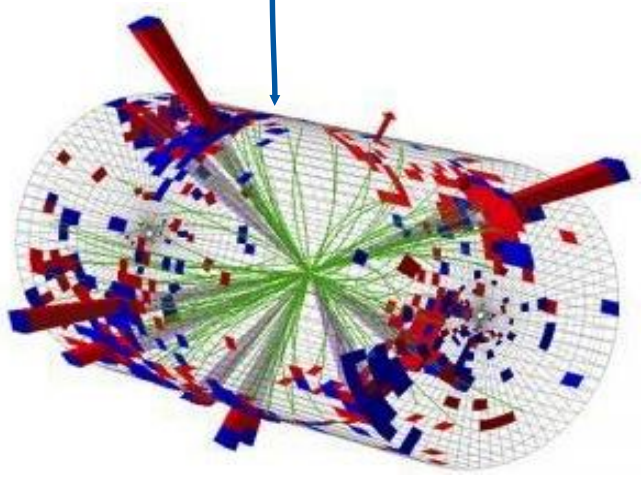


Each detector is equipped with ~ 150M sensors -> PB/s (!)

LHC Data



Reduced by online computers that filter out a few hundreds "good" events per second …

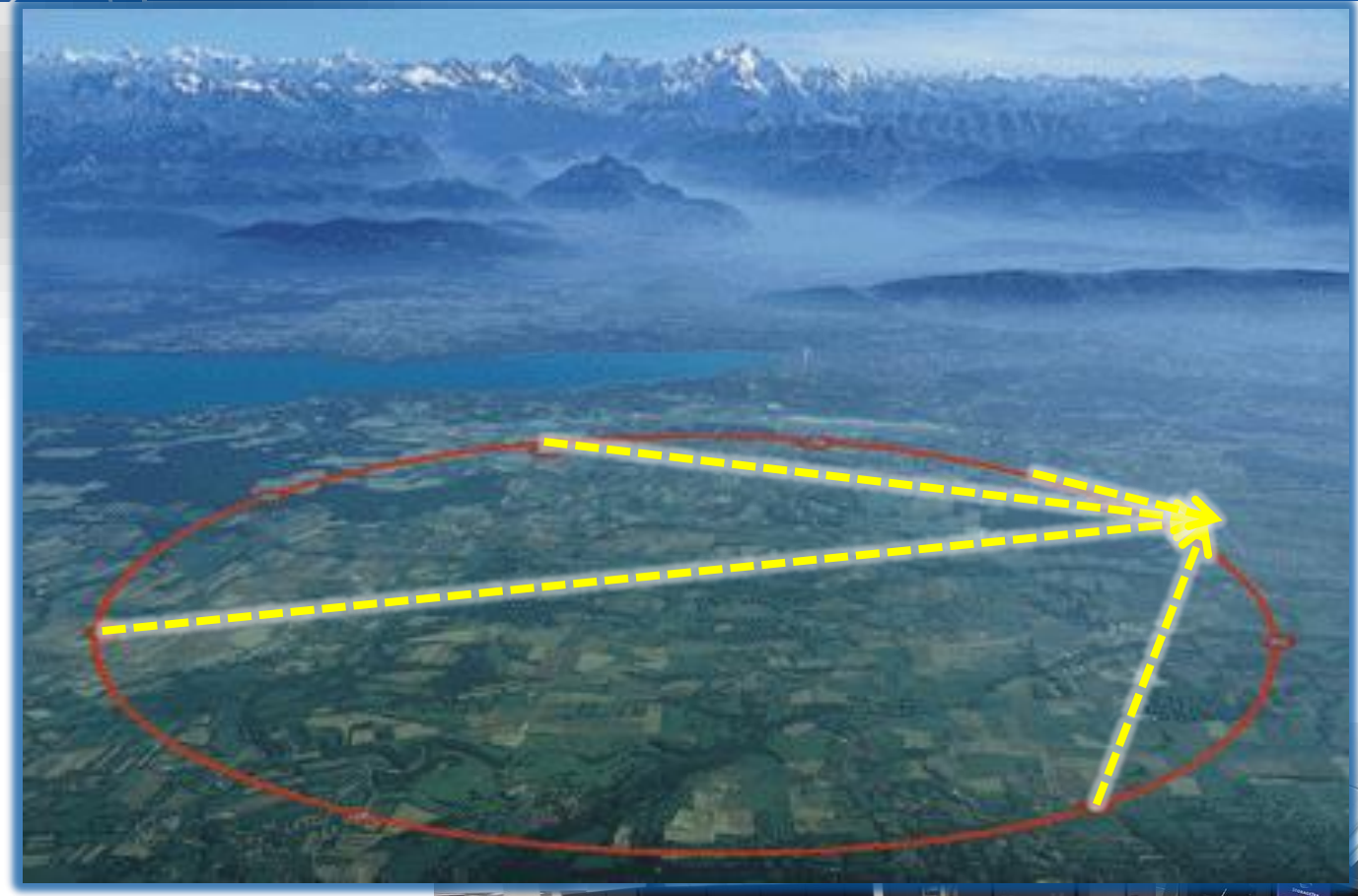


1 event = few Megabytes

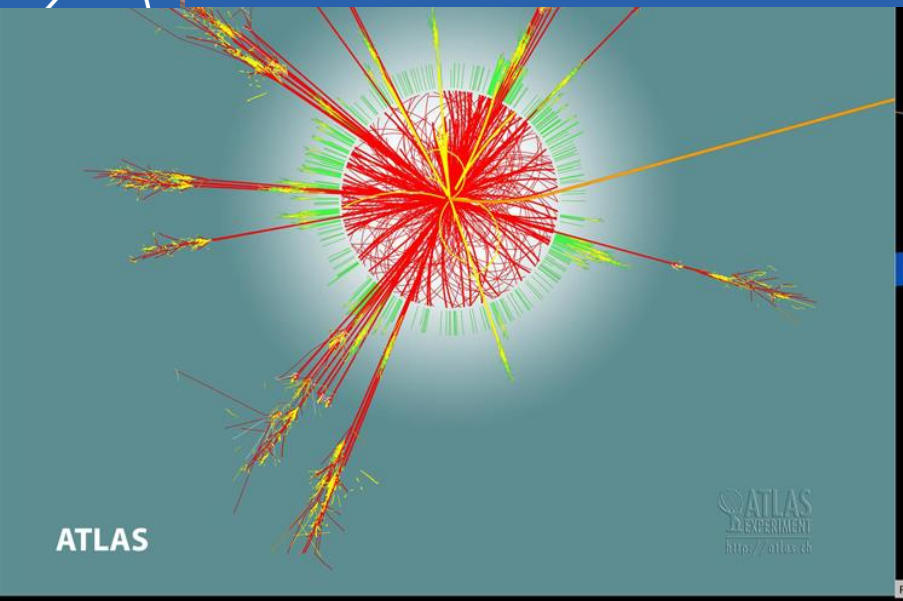
… which are recorded on Disk and Tape at up to 10GB/s

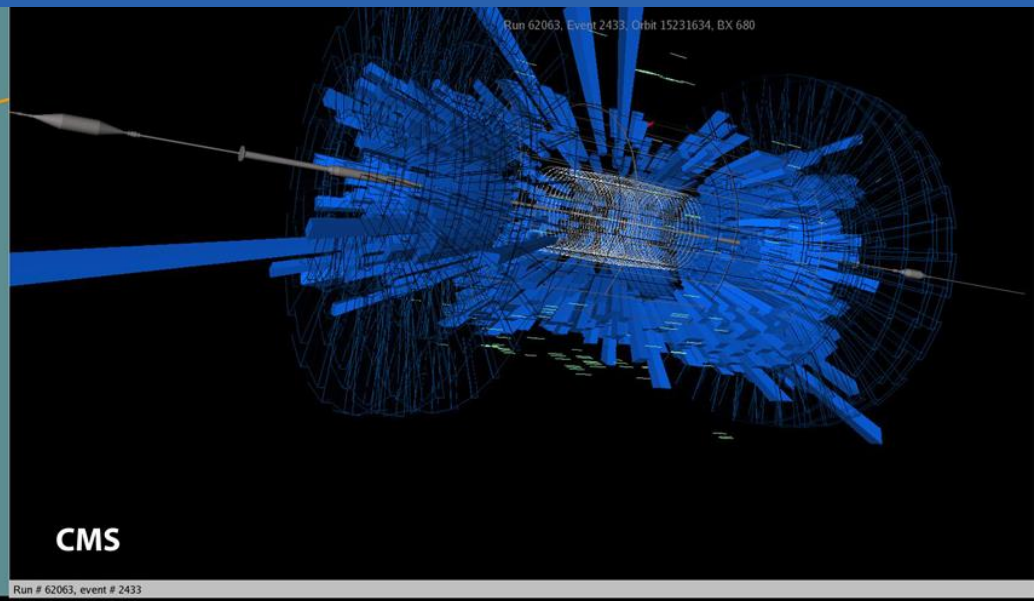
50 Petabytes per year (today) for four experiments

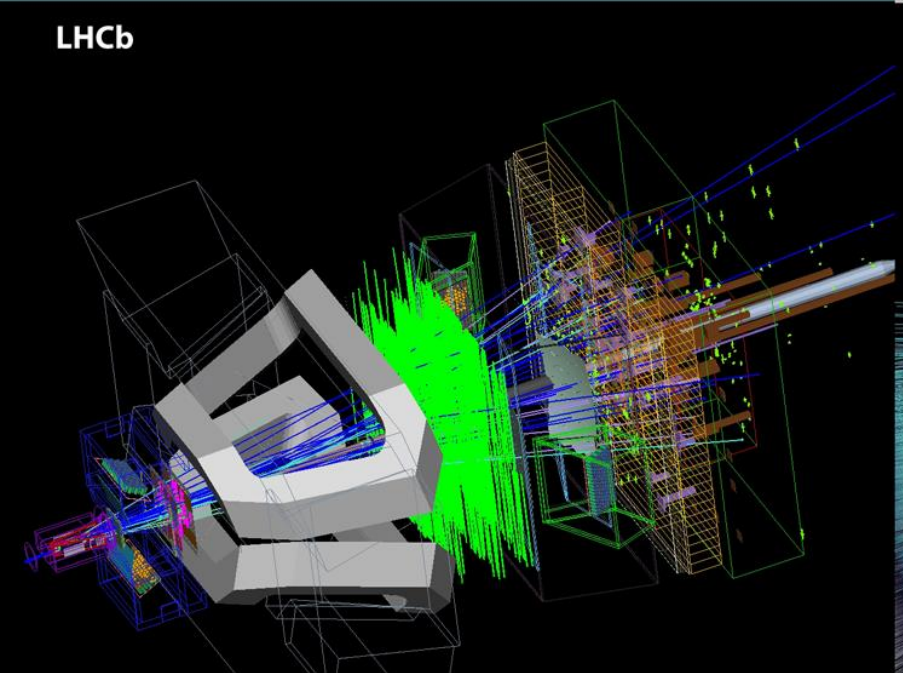
Data Transfer and Storage at the CC

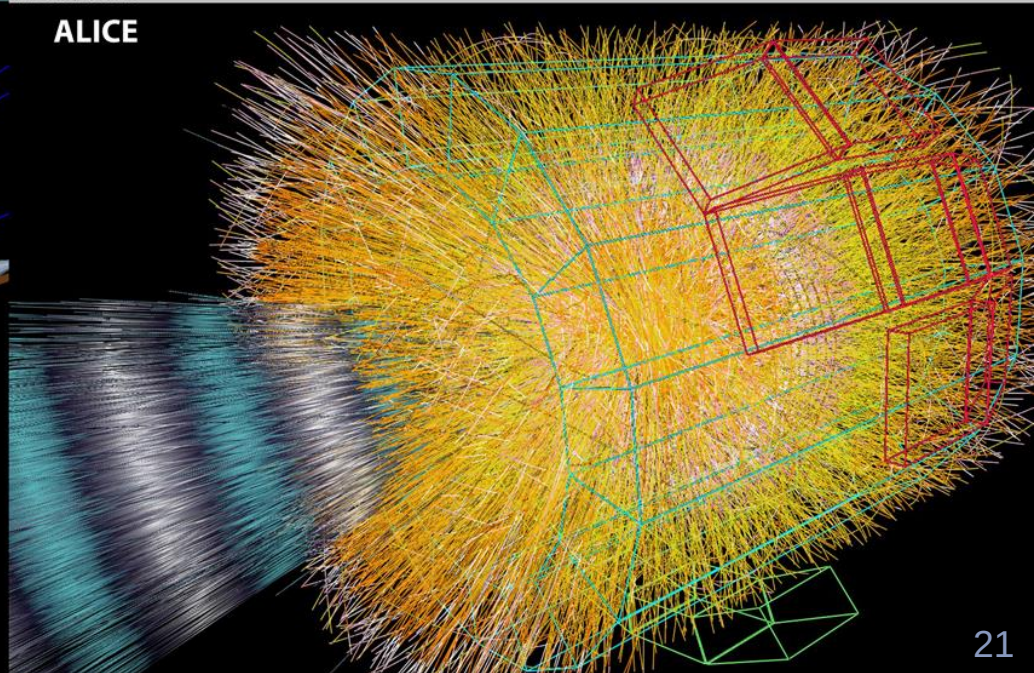


Reconstruction of Collision Events







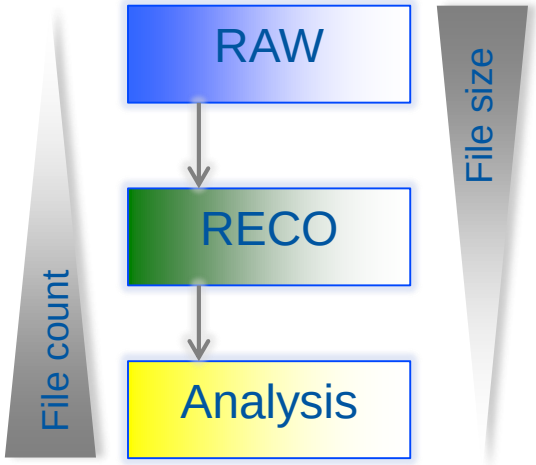


LHC Data Types and Access Patterns

1. Detector Data Acquisition (DAQ) raw data
   - Sensor results from collision events
   - Writing of continuous streams of RAW format data when LHC is running
   - Packed into large files (1-10GB) stored ad eternum

2. Event Reconstruction
   - Reading of RAW data files, mostly streaming
   - Writing of RECO data files, size ~1GB

3. WAN Data Export to Grid Laboratories
   - Sequential reading of RAW and RECO data sent via WAN over the whole world



4. Data Analysis
   - Analysis by O(1000) end-users, skimming over O(TB) RECO data sets, mostly random I/O
   - O(M) small files written, high volatility
   - Bulk activity supposed to happen outside CERN

Hardware Architecture at CERN

- Simple, flexible scale-out HW architecture
  - Easy to integrate mass market components
  - Easy evolution to new technologies
  - No need for fast interconnects ie InfiniBand

Tape storage

WAN in/out
DAQ Detectors
Secondary DC in Budapest
GRID Export world-wide

CPU servers
(via OpenStack / batch system)

8/16 Gb FC

10GbE

N*100Gb

40Gb/100GbE redundant router backbone

10/40GbE

tape servers

10GbE

disk servers (caching, analysis)

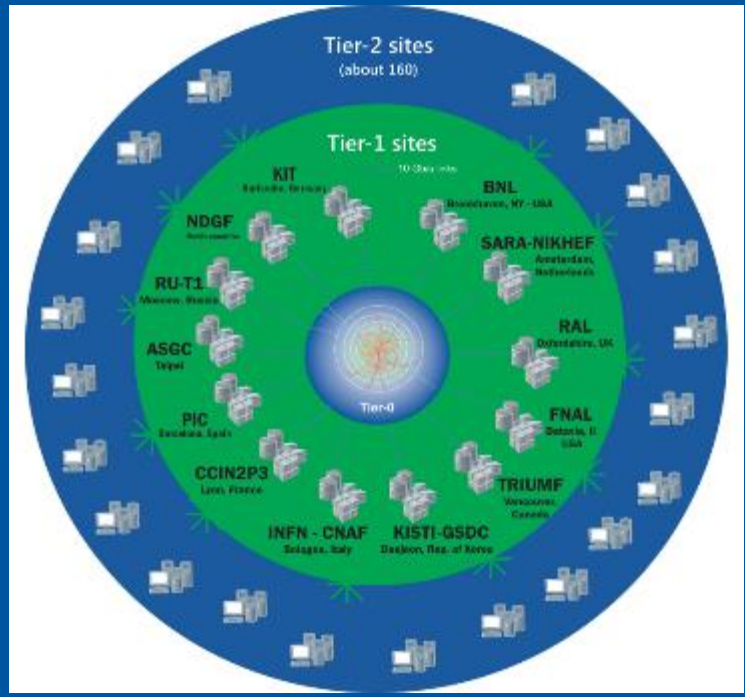
Infrastructure: The Worldwide LHC Computing Grid

A distributed computing infrastructure to provide the production and analysis environments for the LHC experiments

Managed and operated by a worldwide collaboration between the experiments and the participating computer centres

The resources are distributed – for funding and sociological reasons

Our task was to make use of the resources available to us – no matter where they are located



Tier-0 (CERN):
- Data recording
- Initial data reconstruction
- Data distribution

Tier-1 (12 centres):
- Permanent storage
- Re-processing
- Analysis

Tier-2 (~140 centres):
- Simulation
- End-user analysis

- **~ 160 sites, 35 countries**
- **300000 cores**
- **200 PB of storage**
- **2 Million jobs/day**
- **10 Gbps links**

WLCG Worldwide LHC Computing Grid

CERN Archive current numbers

Data:

- 230 PB physics data – CASTOR
- ~10 PB backup (TSM)

Tape libraries:

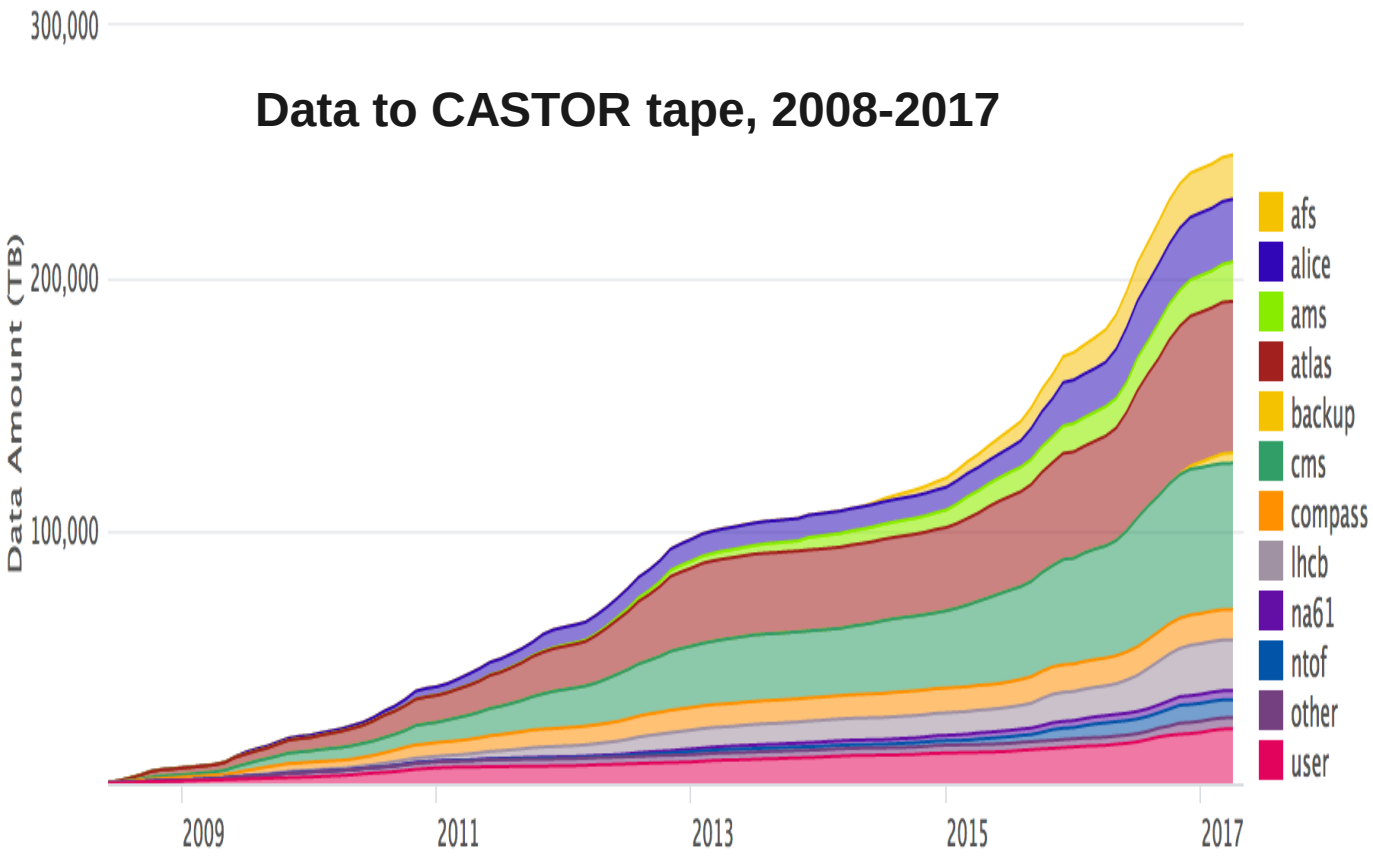
- IBM TS3500/4500 (3+2)
- Oracle SL8500 (4)

Tape drives:

- ~90 archive
  - TS1155/50+T10000D
- ~55 backup

Capacity:

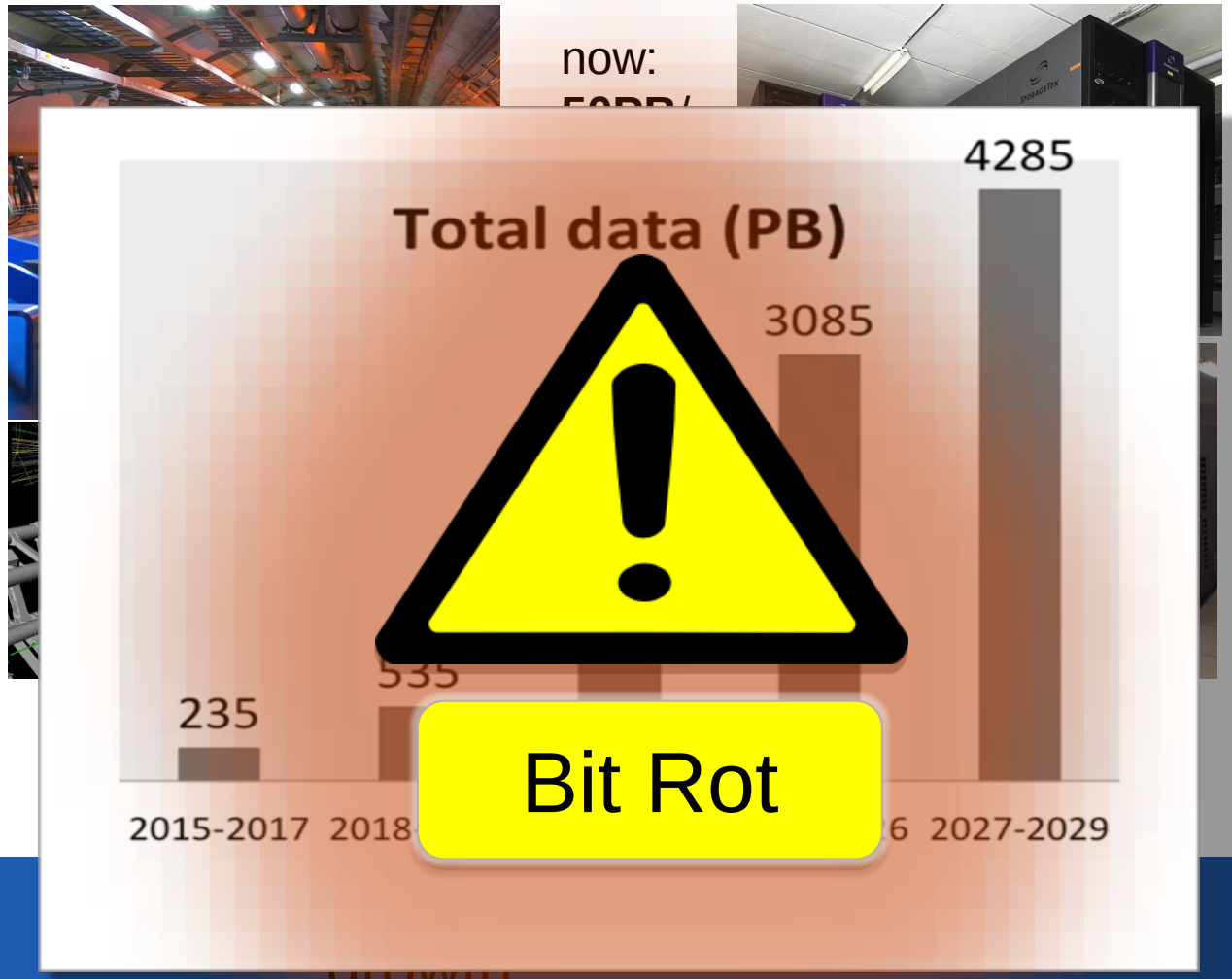
- ~70 000 slots
- ~30 000 tapes



The challenge at CERN: Preserving long-term data at the Exa-scale

LHC experiments

CERN Archive



The fight against Bit Rot

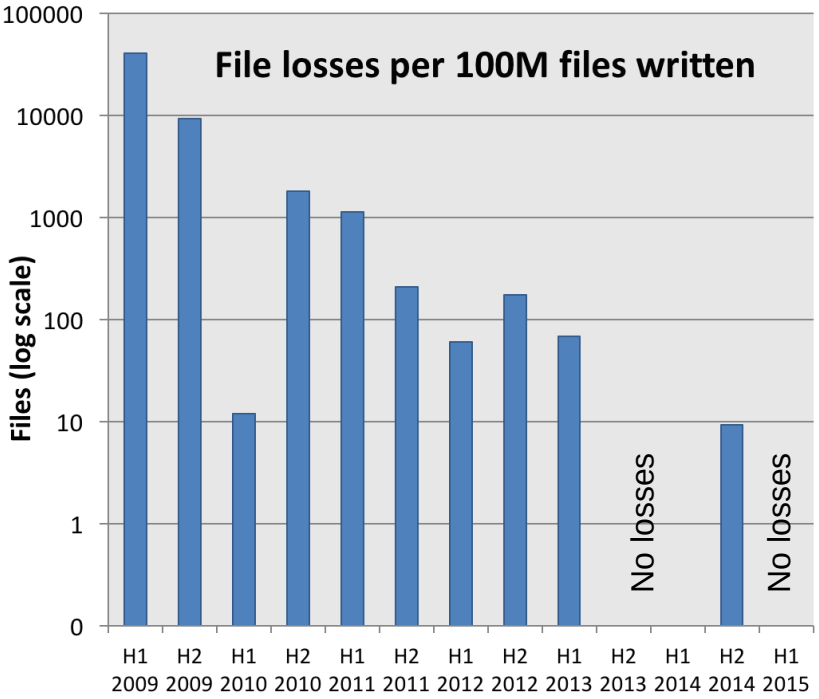
- ***The putative tendency of content in storage to become corrupt over time*** (Wiktionary)
- Bit rot is **unavoidable** at Exabyte scale and needs to be factored in at all stages. Reasons:

- Corruptions (OS, firmware issues)
- Undetected bit flips (data transmission, within storage device)
- Media wear out and breakage: disk head crashes, snapping tapes
- Hardware or media obsolescence (after 5-10 years max)
- Environmental elements (humidity, temperature, dust)
- Disasters such as fire, earthquakes, accidents etc.

Protecting the CERN Archive

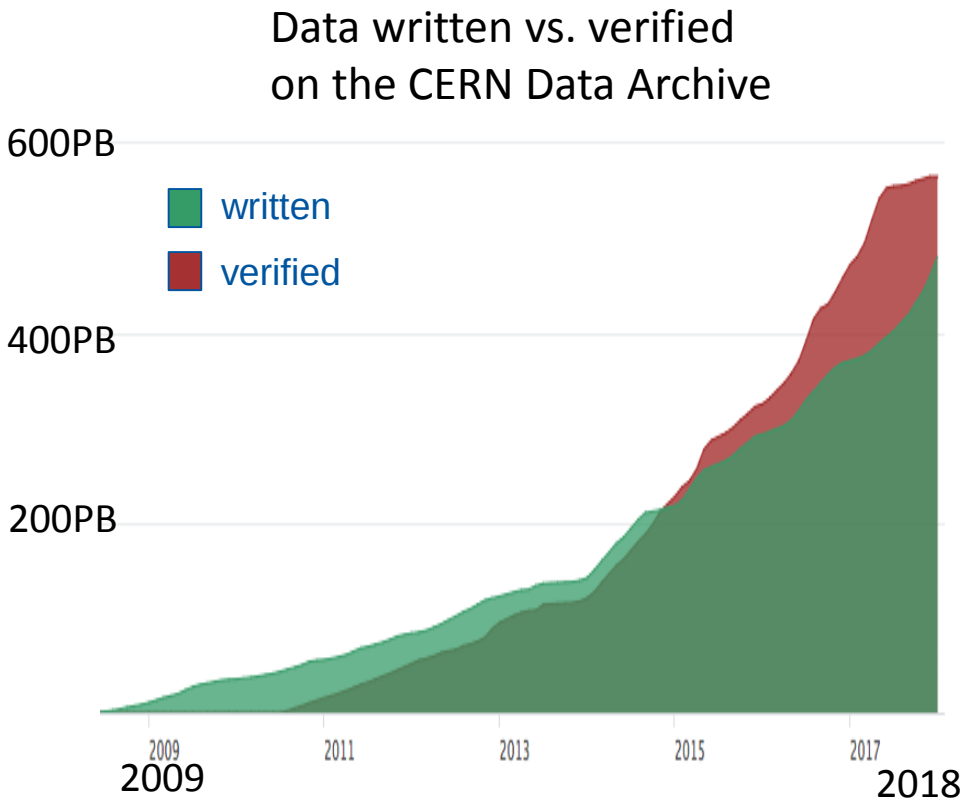
Ongoing activity to protect the data and improve reliability

- Systematic tape verification
- Reducing media wear
- Redundant data copies
- New techniques: SCSI alerts, Logical Block Protection, failure prediction
- Protecting the enviroment
- Migration to new media generations



Continuous 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:
- *Q: can you retrieve my file?*
- *A: let me check… err, sorry, 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 metadata (checksum/size, …)
- Two verification modes are supported:
- *Full*: once tape is completely filled, or whenever it hasn't been accessed for a long time
- *Light*: Immediately after a tape has been written to, checking critical areas (beginning/end of tape)



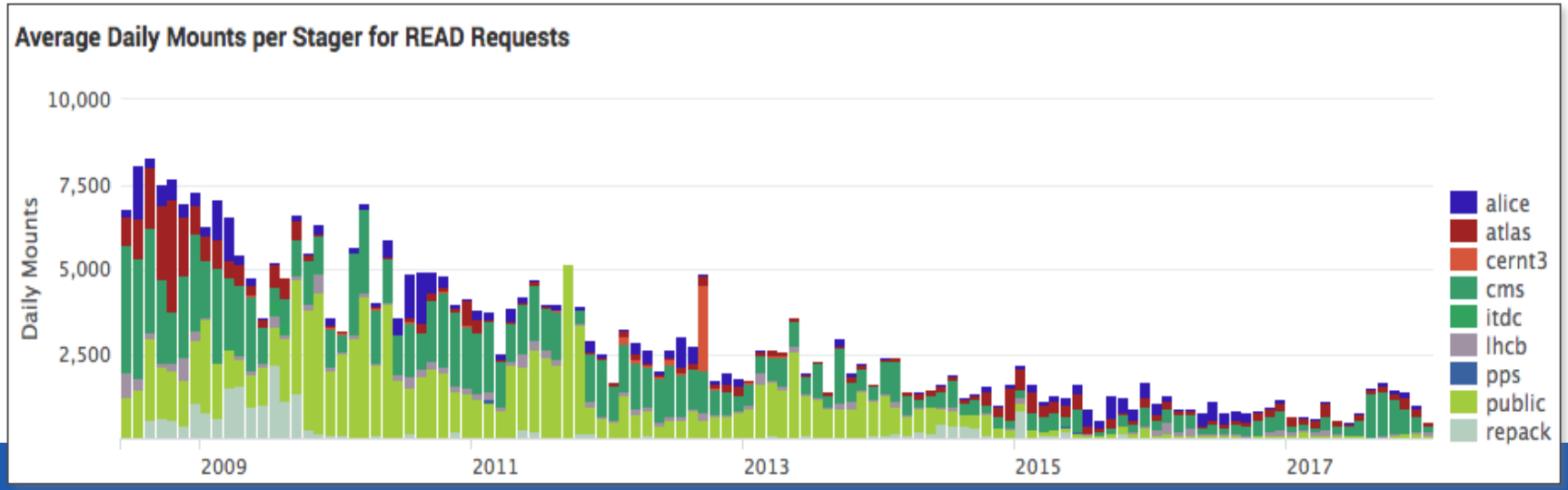
Increasing media / robotics longevity

- CASTOR was designed as a “classic” file-based HSM. If user file is not on disk -> recall it from tape ASAP
  - Experiment data sets can be spread over hundreds of tapes
  - Many tapes get (re)mounted but files read is very low (1-2 files)
  - Every mount is wasted drive time (~2 min for mounting / unmounting).
  - Mount/unmount times are *not* improving with new technology
  - Many drives used -> reduced drive availability (ie for writes)

- Mounting and unmounting is the highest risk operation for tapes, robotics and drives.
  - Mechanical (robotics) failure can affect access to a large amount of media.
- Technology evolution moves against HSM:
  - Bigger tapes -> more files -> more mounts per tape -> reduced media lifetime

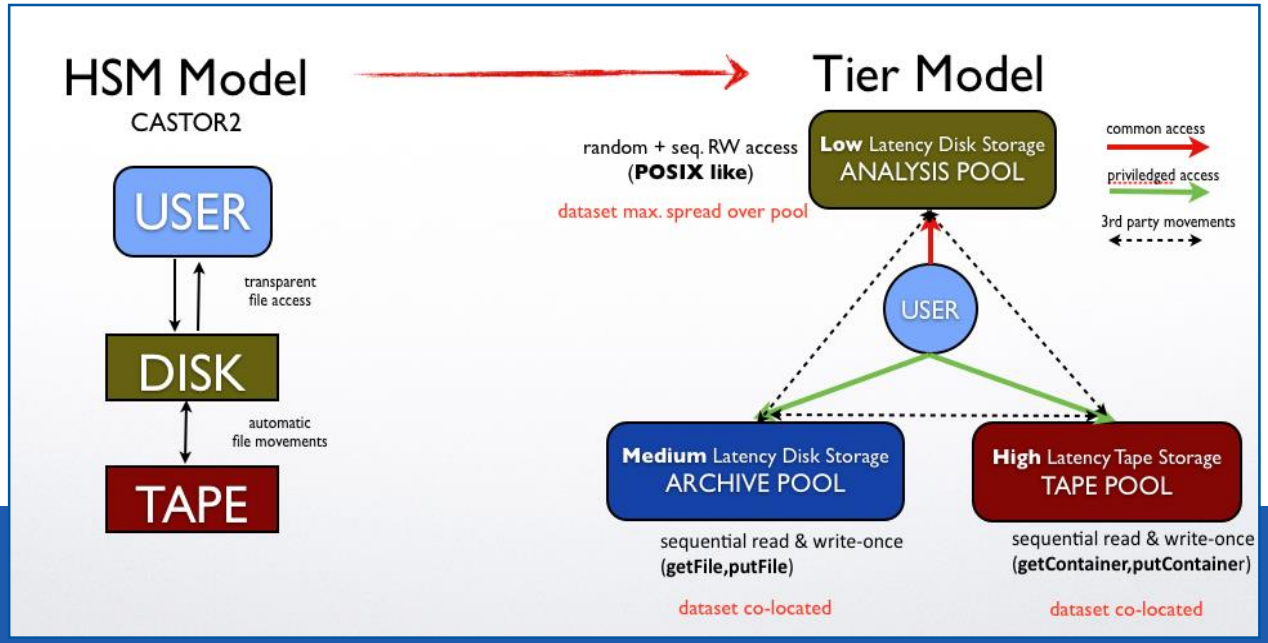
Tape mount rate reduction

- Deployed “traffic lights” to throttle and prioritise tape mounts
  - Thresholds for minimum volume, max wait time, concurrent drive usage, group related requests
- Developed monitoring for identifying inefficient tape users, encourage them to use bulk pre-staging on disk
- Work with experiments to migrate end-user analysis to EOS as mostly consisting in random access patterns
- Tape mount rates have decreased by over 50% since 2010, despite increased volume and traffic



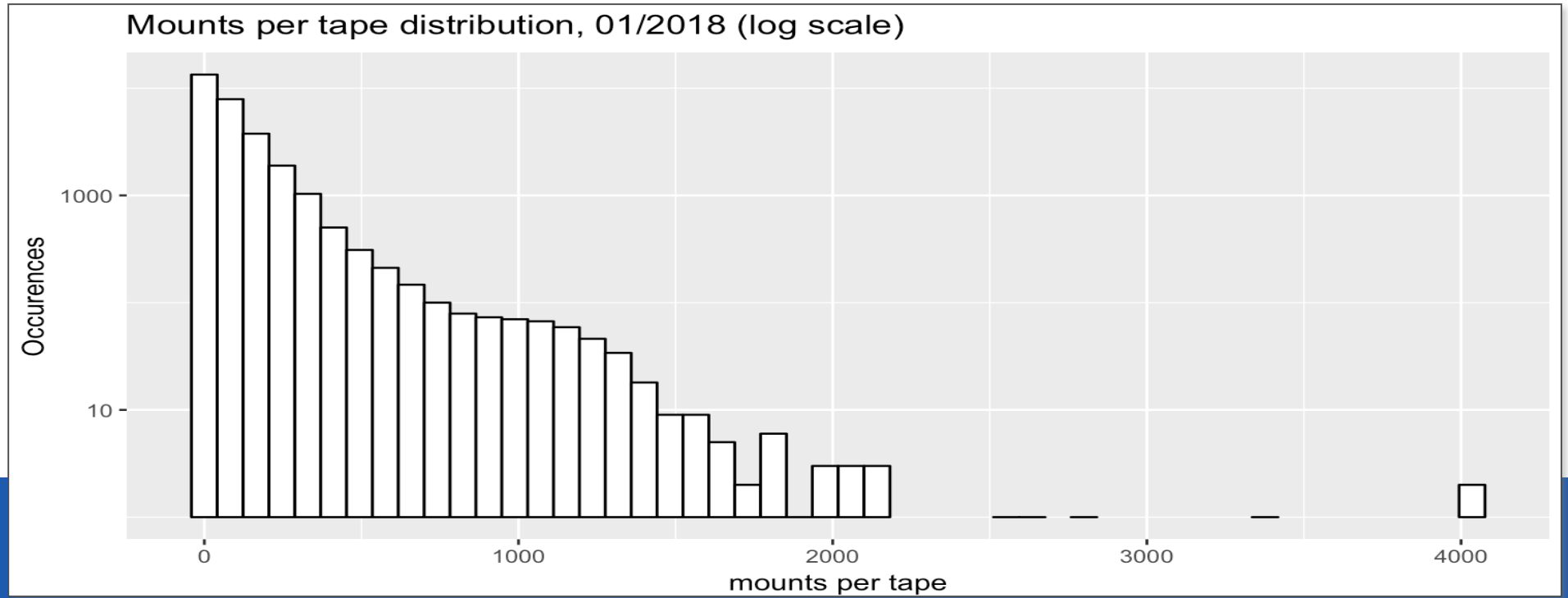
HSM model limitations

- HSM model showing its limits
  - Enforcing “traffic lights” and increasing disk caches not sufficient
  - … even if 99% of required data is on disk, mount rates can be huge for missing 1%!
- Ultimate strategy**: move away from “transparent”, file/user based HSM**
  - Remove / reduce tape access rights from (end) users
  - Move end users to EOS
  - Increase tape storage granularity from files to data (sub)sets (Freight-train approach) managed by production managers
- Model change from HSM to more loosely coupled Data Tiers
  - Using CASTOR == Archive, EOS == Analysis Pool



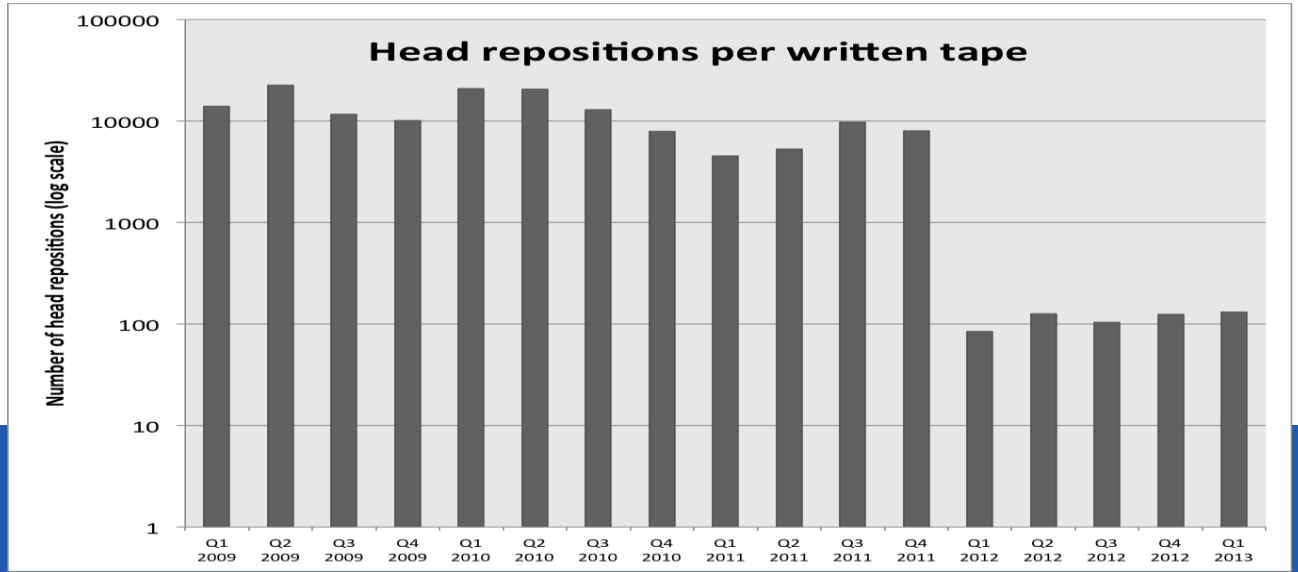
Addressing media wear

- With “traffic lights” in place, average daily repeated tape mount rates are down to ~2-3 / day.
  - Monitoring disables tapes mounted “too frequently” + operators notified.
- Also, introduced automated decommissioning of media mounted >= 5000 times
  - Tape gets disabled and ticket generated for media repacking
  - (but we rarely get there)



Avoiding “shoe-shining”

- Media wear also happens when writing small files to tape
  - By default, tape flushes buffers after close() of a tape file -> stop motion and rewind to end of last file (“head reposition”)
  - CASTOR uses ANSI AUL as tape format: 3 tape files per CASTOR file!
  - Performance (and media life time) killer in particular with new-generation drives (higher density -> more files)
- Can be avoided by using file aggregations (requires tape format change)
- Alternative found: logical (or “buffered”) tape marks
  - Prototyped by CERN, now fully integrated in Linux kernel
  - Synchronize only every 32GB worth of data
- Reduced number of head repositions from ~10000/tape to ~100/tape
  - Nowadays enterprise and LTO-8 support caching on tape – but much less efficient!



Redundant data copies

- By default, only one copy of a file (version) is stored on tape.
  - CASTOR does not have a versioning concept (archiving – not backup!)
- If justified, second copies can be generated on different tapes and buildings
  - Requested by user (can be proposed by operations)
  - Requires management approval for larger data sets
- Double copies are stored in separate buildings.

- Mostly critical data sets and ”small” or legacy experiments not having off-site copies
  - LHC experiments export and keep physics data copies outside CERN