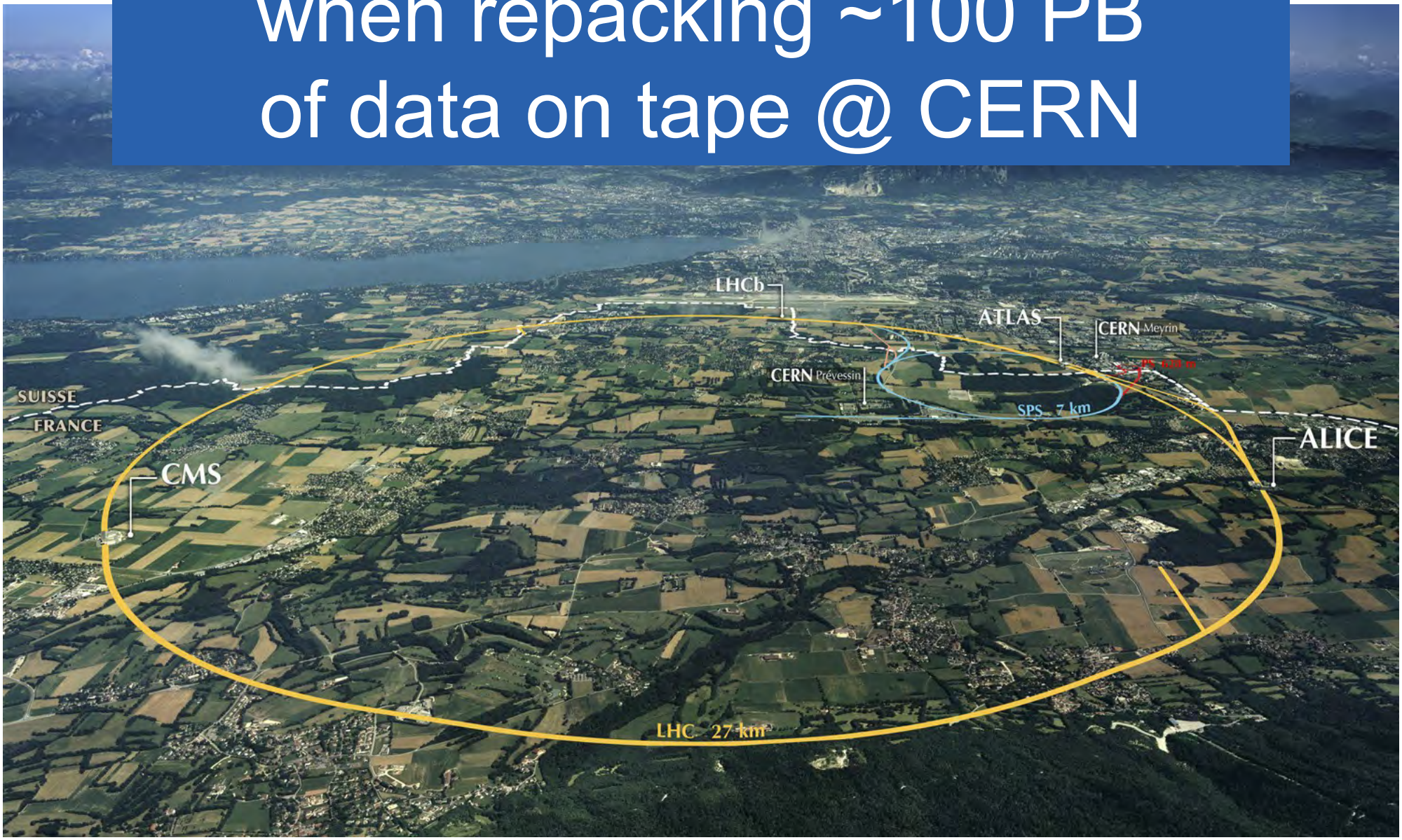

CERN

Benefits of using SSDs when repacking ~100 PB of data on tape @ CERN



Vladimír Bahyl

Agenda

- Quick CERN introduction
- Role of tape @ CERN
- Data migration:
  - Why doing it?
  - Constraints
  - Infrastructure details
  - Issues and solutions
  - Results
- Conclusion

CERN: founded in 1954: 12 European States

"Science for Peace"

Today: 22 Member States

~ 3440 staff

~ 14092 scientific users

Budget (2018): ~1148 MCHF

Member States (22): 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 (3): Cyprus, Serbia, Slovenia

Associate Membership (5): India, Lithuania, Pakistan, Turkey, Ukraine

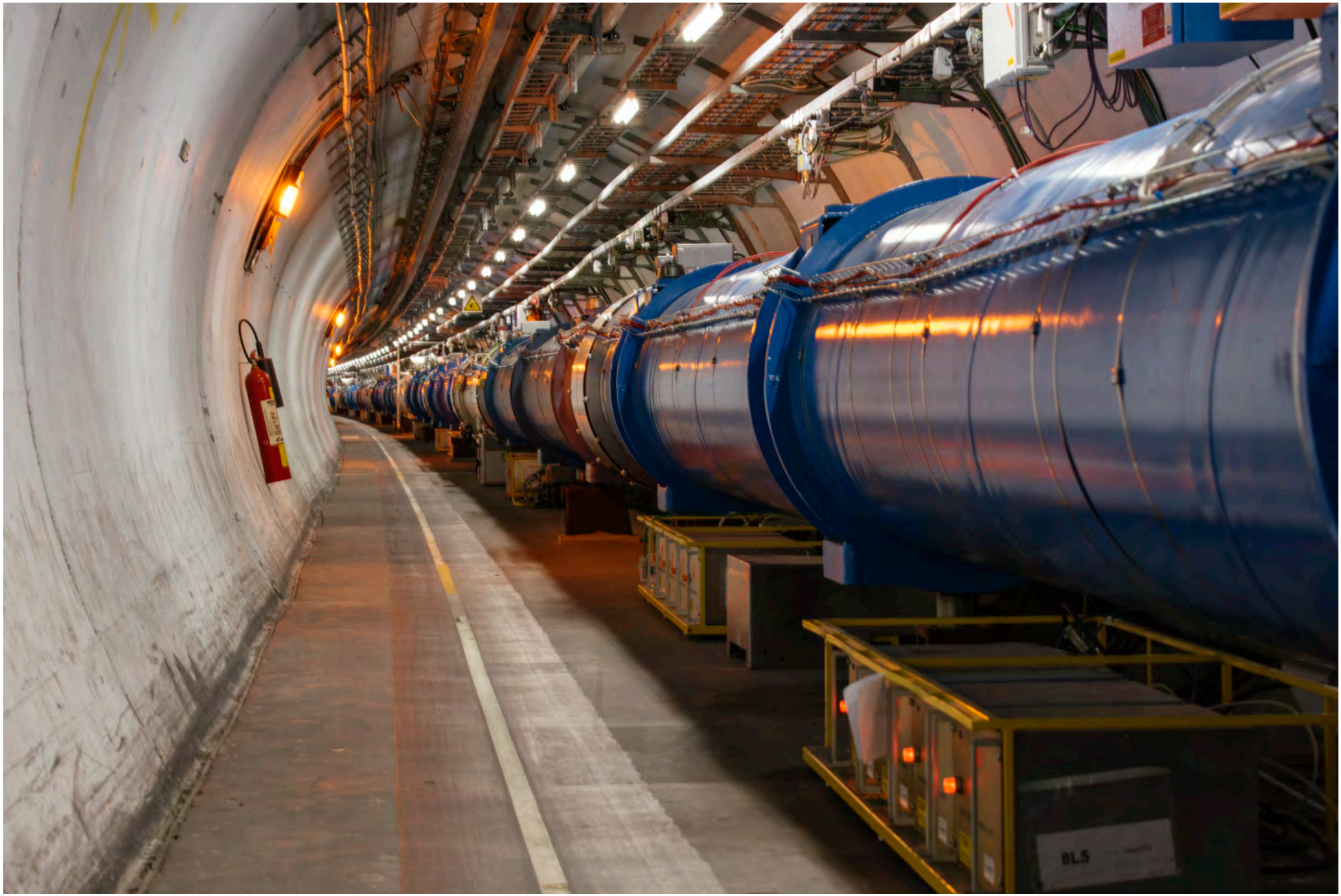
Observers to Council (6): Japan, Russia, USA; European Union, JINR, UNESCO

Research programme

Accelerator-based experiments	
LHC	7+7 TeV Large Hadron Collider, 27 km in circumference 7 active experiments: ALICE, ATLAS, CMS, LHCb, LHCf, MoEDAL and TOTEM
SPS	450 GeV Super Proton Synchrotron, 6.9 km in circumference 6 active experiments: COMPASS, NA61/SHINE, NA62, NA63, NA64, UA9; 2 data analysis: OPERA and ICARUS
PS	28 GeV Proton Synchrotron 2 active experiments: CLOUD and n_TOF (2 experimental areas); 1 data analysis: DIRAC
ISOLDE	Booster-ISOLDE isotope separator 86 active experiments, 45 in preparation
AD	100 MeV/c Antiproton Decelerator 5 active experiments: AEgIS, ALPHA, ASACUSA, ATRAP and BASE; 1 in preparation: GBAR

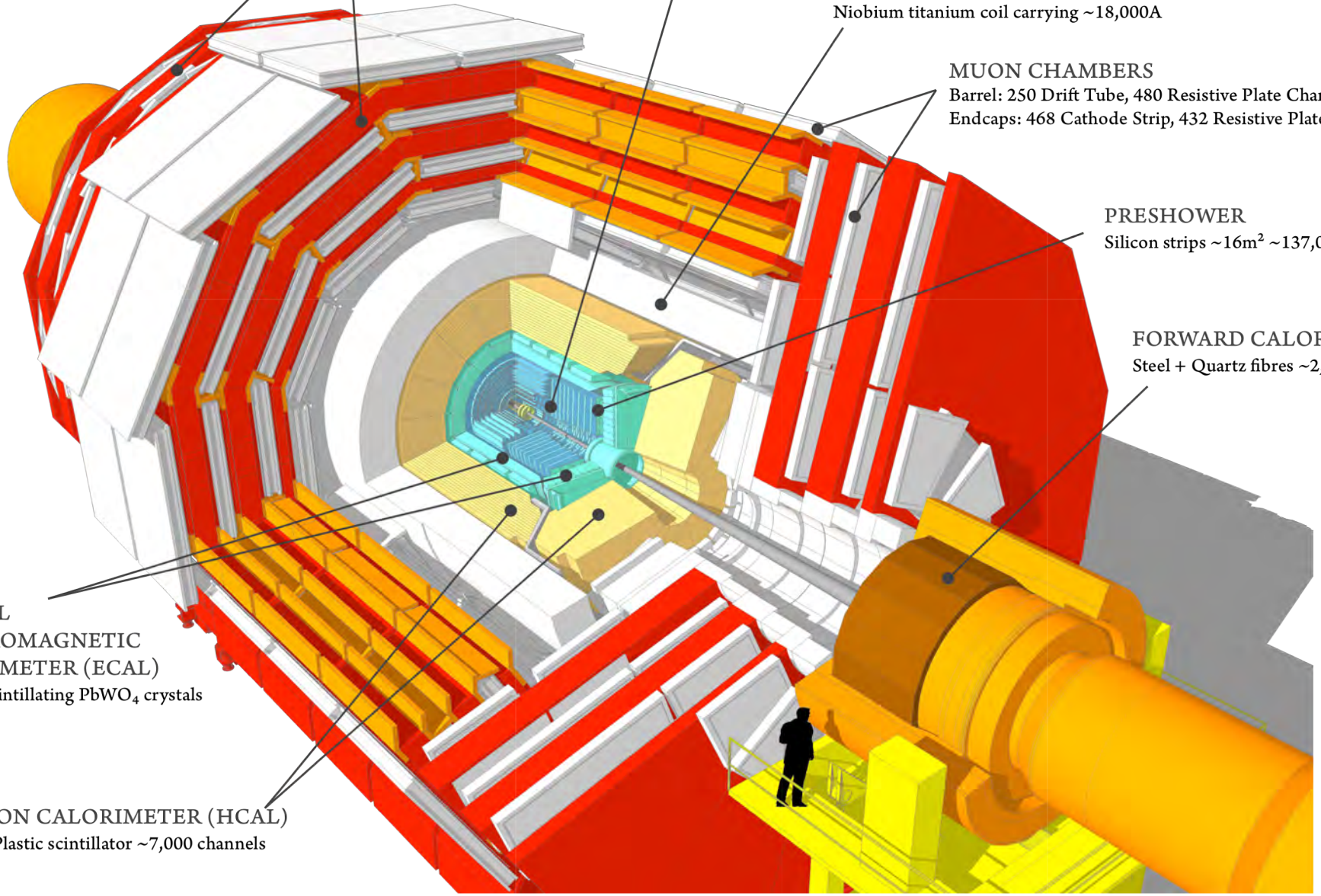
Non-accelerator experiments and detector developments
CAST, OSQAR and CERN Neutrino Platform (WA104, WA105, NP02, NP04, NP05)

Advanced accelerator development	
CTF3	150 MeV CLIC test facility electron beam: Accelerator R&D for future linear collider
AWAKE	At the SPS: Using 400 GeV protons to drive plasma wakefield acceleration


BLS

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

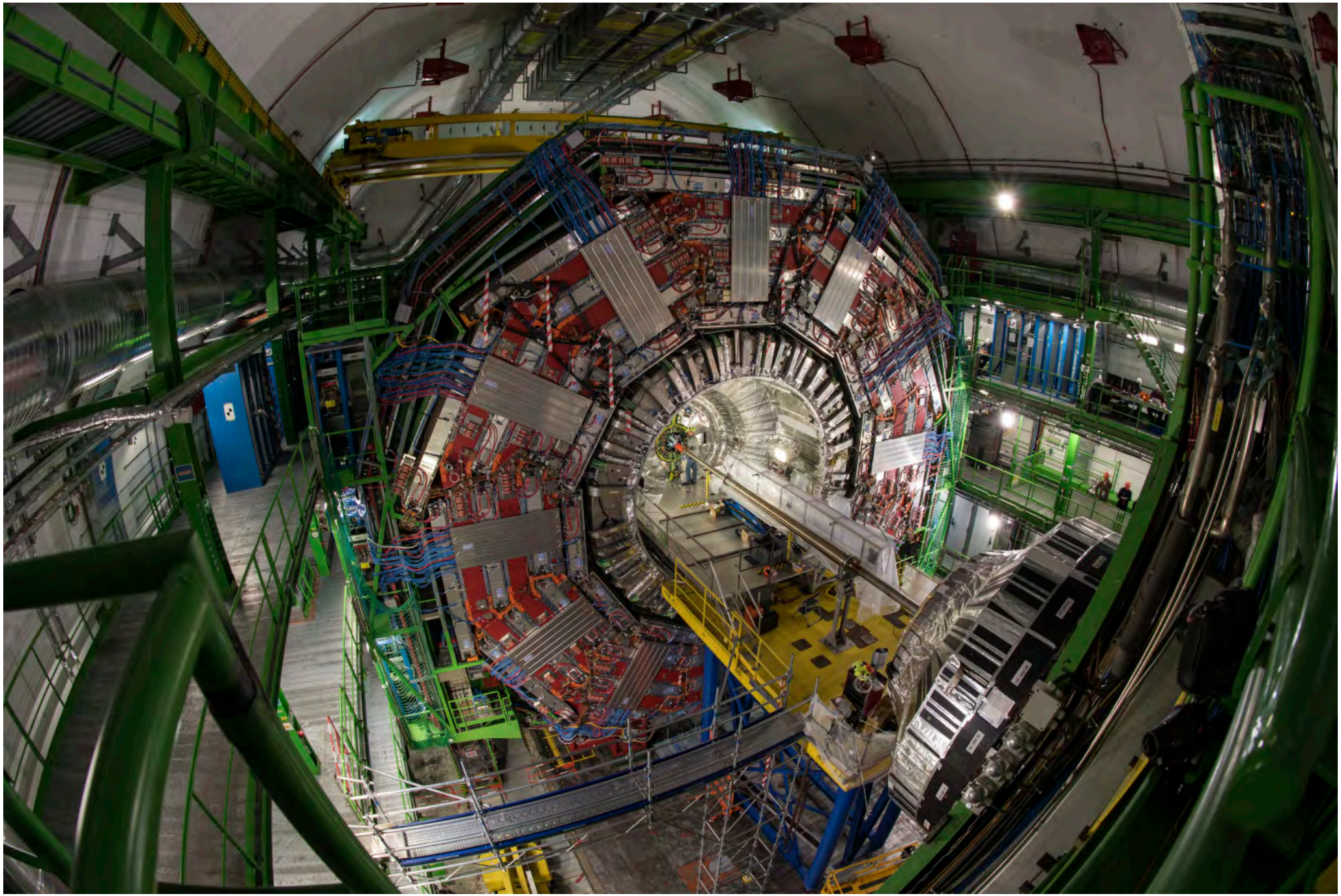
MUON CHAMBERS
Barrel: 250 Drift Tube, 480 Resistive Plate Chambers
Endcaps: 468 Cathode Strip, 432 Resistive Plate Chambers

PRESHOWER
Silicon strips ~16m² ~137,000 channels

FORWARD CALORIMETER
Steel + Quartz fibres ~2,000 Channels

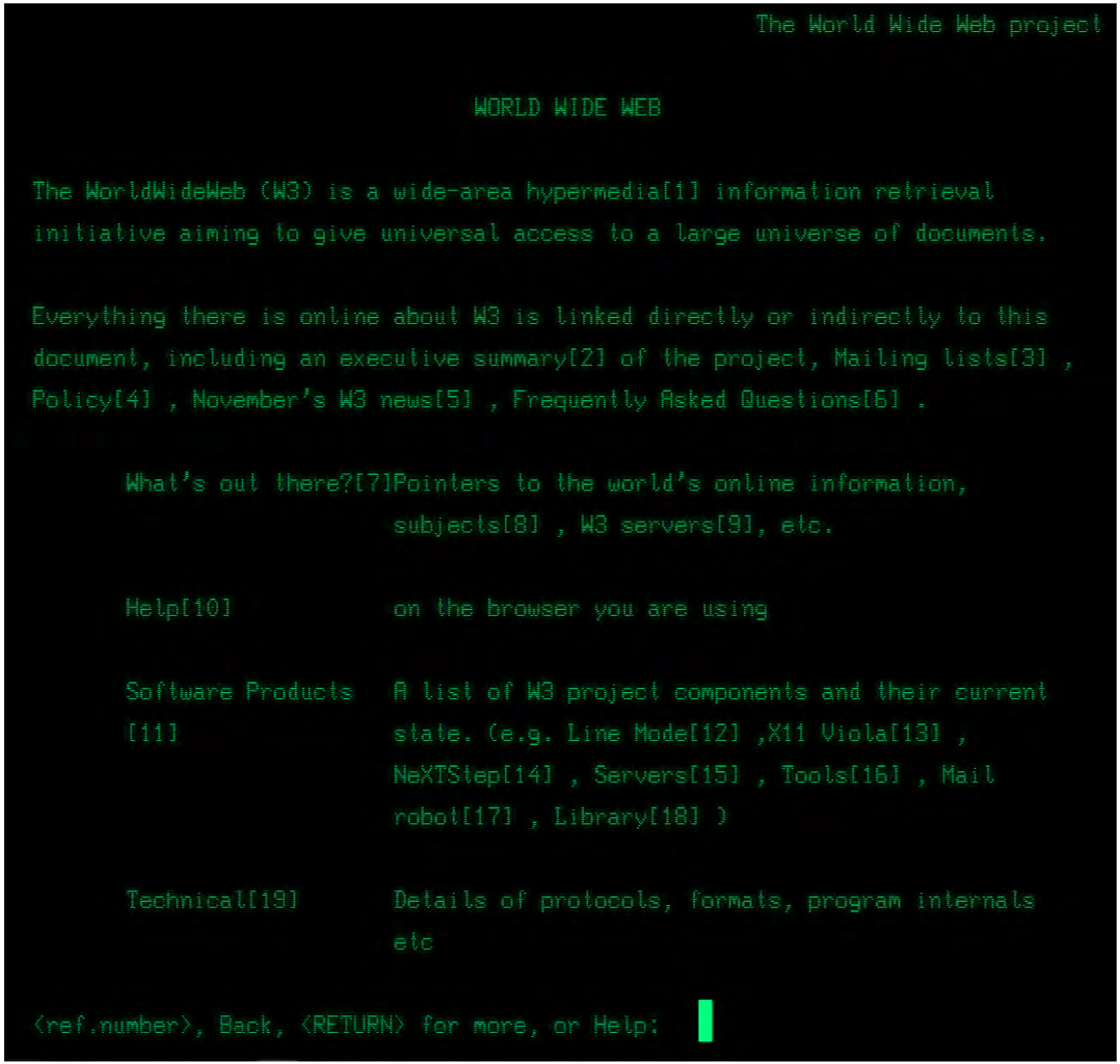
CRYSTAL ELECTROMAGNETIC CALORIMETER (ECAL)
~76,000 scintillating $PbWO_4$ crystals

HADRON CALORIMETER (HCAL)
Brass + Plastic scintillator ~7,000 channels



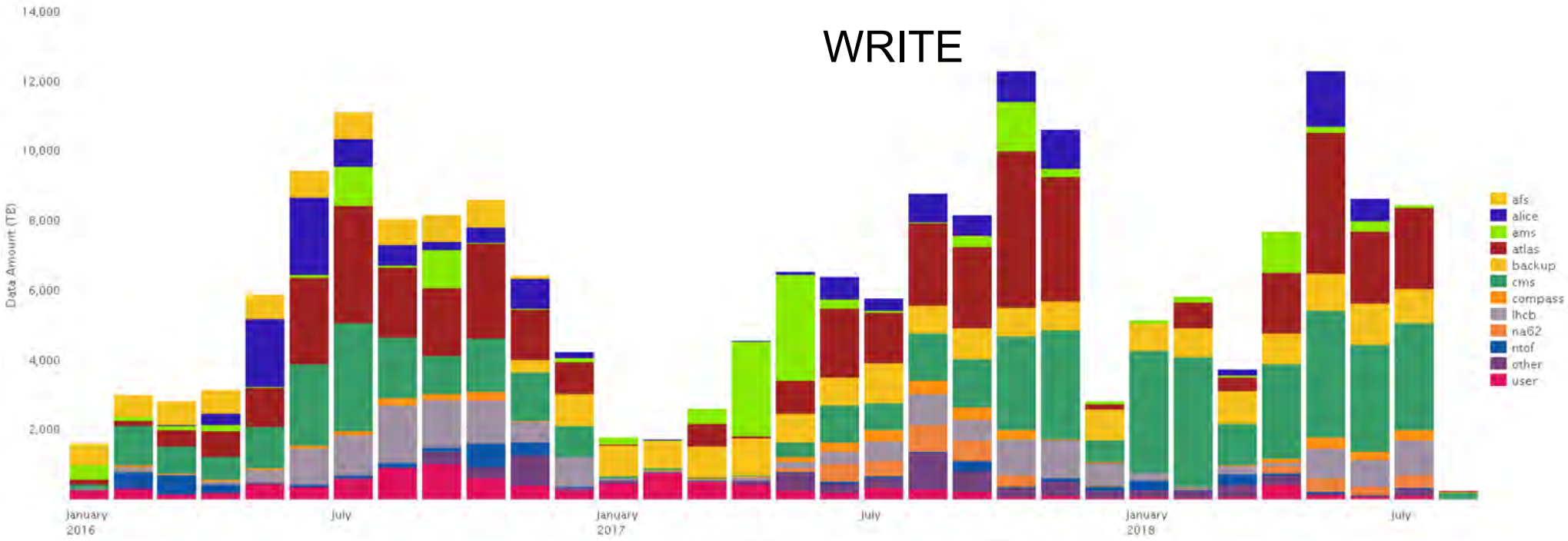
Real proton-proton collision event at 13 TeV in the CMS detector in which two high-energy electrons (green lines), two high-energy muons (red lines), and two-high energy jets (dark yellow cones) are observed. The event shows characteristics expected from Higgs boson production via vector boson fusion with subsequent decay of the Higgs boson in four leptons, and is also consistent with background standard model physics processes.

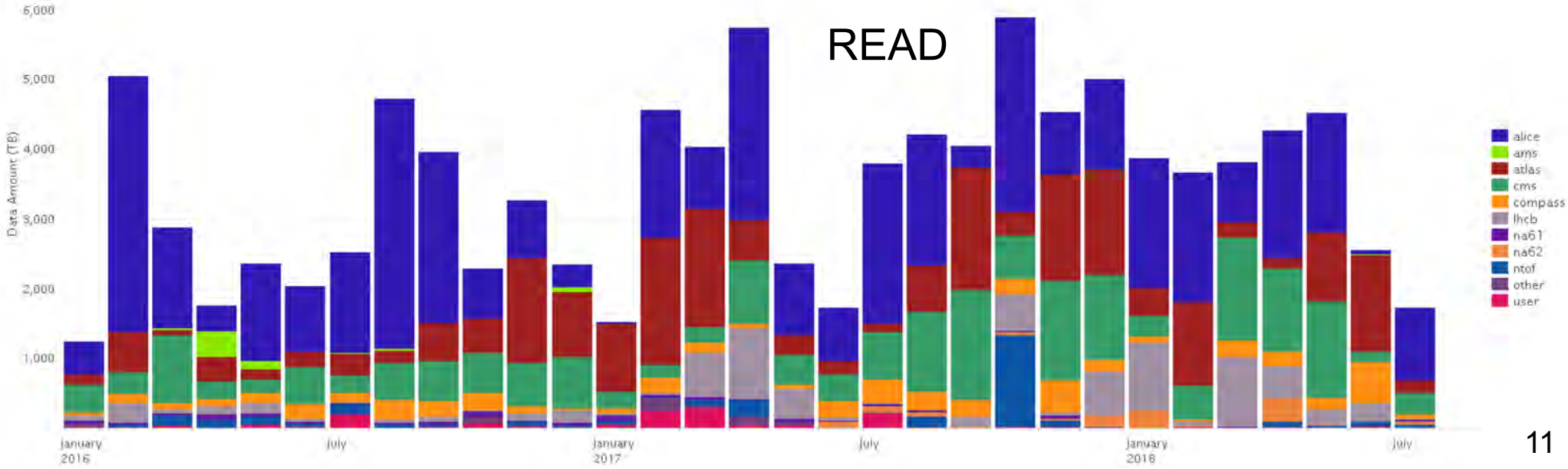
The rest is on the Web ...



... thanks to its invention by Tim Berners-Lee and Robert Cailliau around 1990 while working at CERN.

Tape Usage (August 2018)

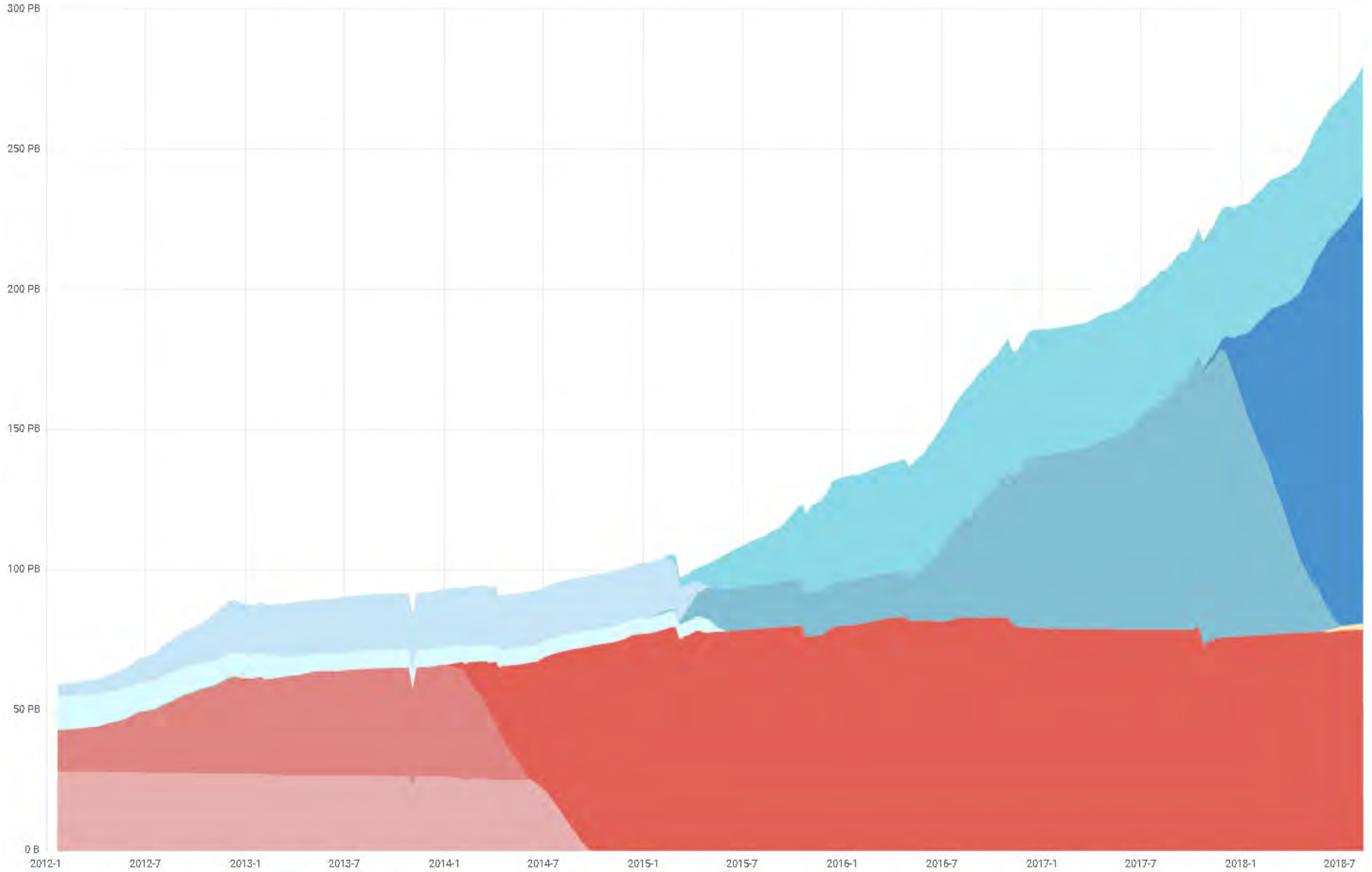




Tape Infrastructure (August 2018)

- CASTOR archive:
  - IBM : 2 x TS4500, 1 x TS3500
    - 46 x TS1155, 20 x LTO-8
    - 14000 x JD media (15 TB),
      6000 x JC media (7 TB),
      500 x LTO-7M media (9 TB),
      160 x LTO-8 media (12 TB)
  - Oracle : 2 x SL8500
    - 20 x T10000D
    - 10000 x T2 media (8 TB)
  - 10 PB disk cache
  - ~300 PB of data on tape
    ~70 PB of free space
- TSM backup:
  - IBM : 2 x TS3500
    - 55 x TS1140
    - 200 x JC media, 12000 x JB media
  - 8 PB; ~2300 M files
  - 18 x TSM 7.1.4 servers

Data Volume vs. Tape Technology



Data migration – why doing it?

- **Increase capacity of the tape archive by reformatting (certain types of) tape cartridges at higher density**
  - Reclaim fragmented space
  - Move data out of problematic media
  - Verify readability of the data
  - Liberate library slots

- IBM TS1155 tape drive introduced @ CERN in 2017
  - Primarily for cloud providers who wanted to squeze the maximum capacity out of the existing tape cartridges
  - Allows to gain 50% more space on 3592JD media
  - Reformatting from 10 TB to 15 TB
- We had ~9500 10 TB cartridges to repack
  - Repack should have liberated ~50 PB of free space
  - = ~700 000 CHF in savings if we would be buying new tapes

Constraints

- Expecting ~100 PB of new data in 2018
  - ~70 PB of LHC; ~30 PB of non LHC

- Transparent to users
- Preserve temporal collocation
- Neither starve nor overflow the disk buffer
- Verify that the data is correctly written

- Be a step ahead (of the data taking): Need to liberate *as much space as possible* before the new data arrives

Infrastructure – 2014/2015

Data: ~100 PB

Large (~1-2 PB), but slow(ish) disk pool

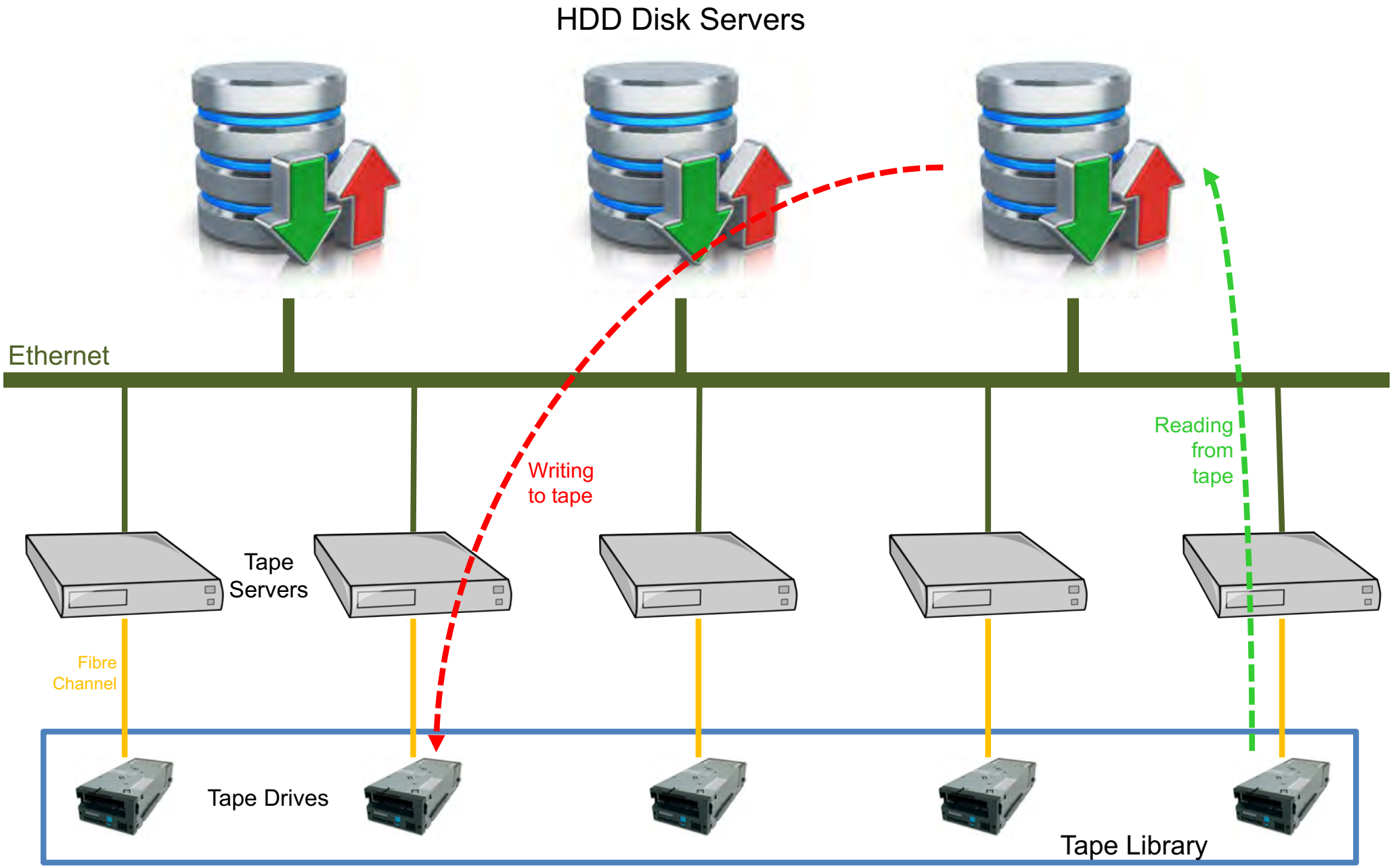
Infrastructure – 2017/2018

Data: ~100 PB

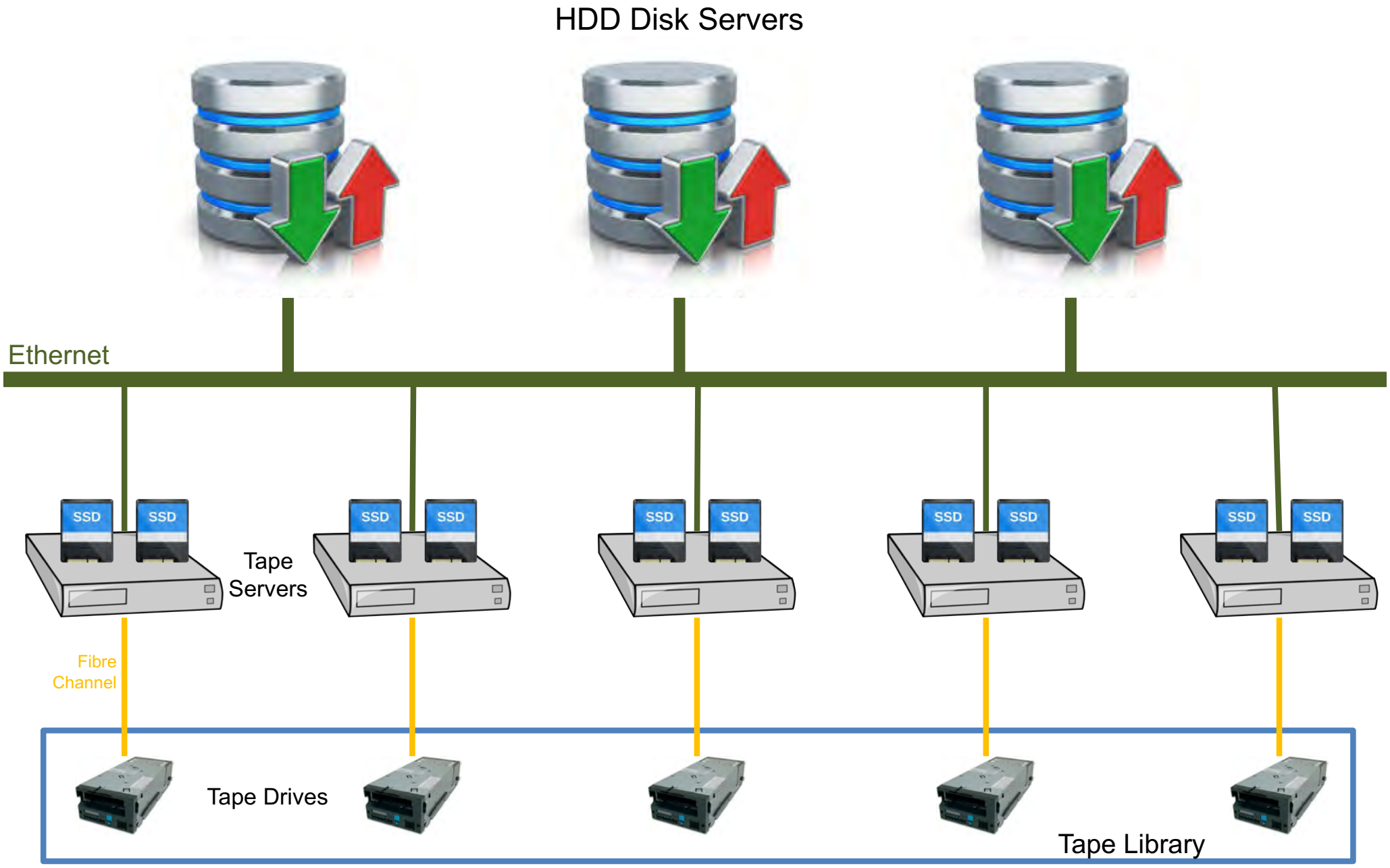
Tiny (~200 TB) but
very efficient disk pool

Hardware setup

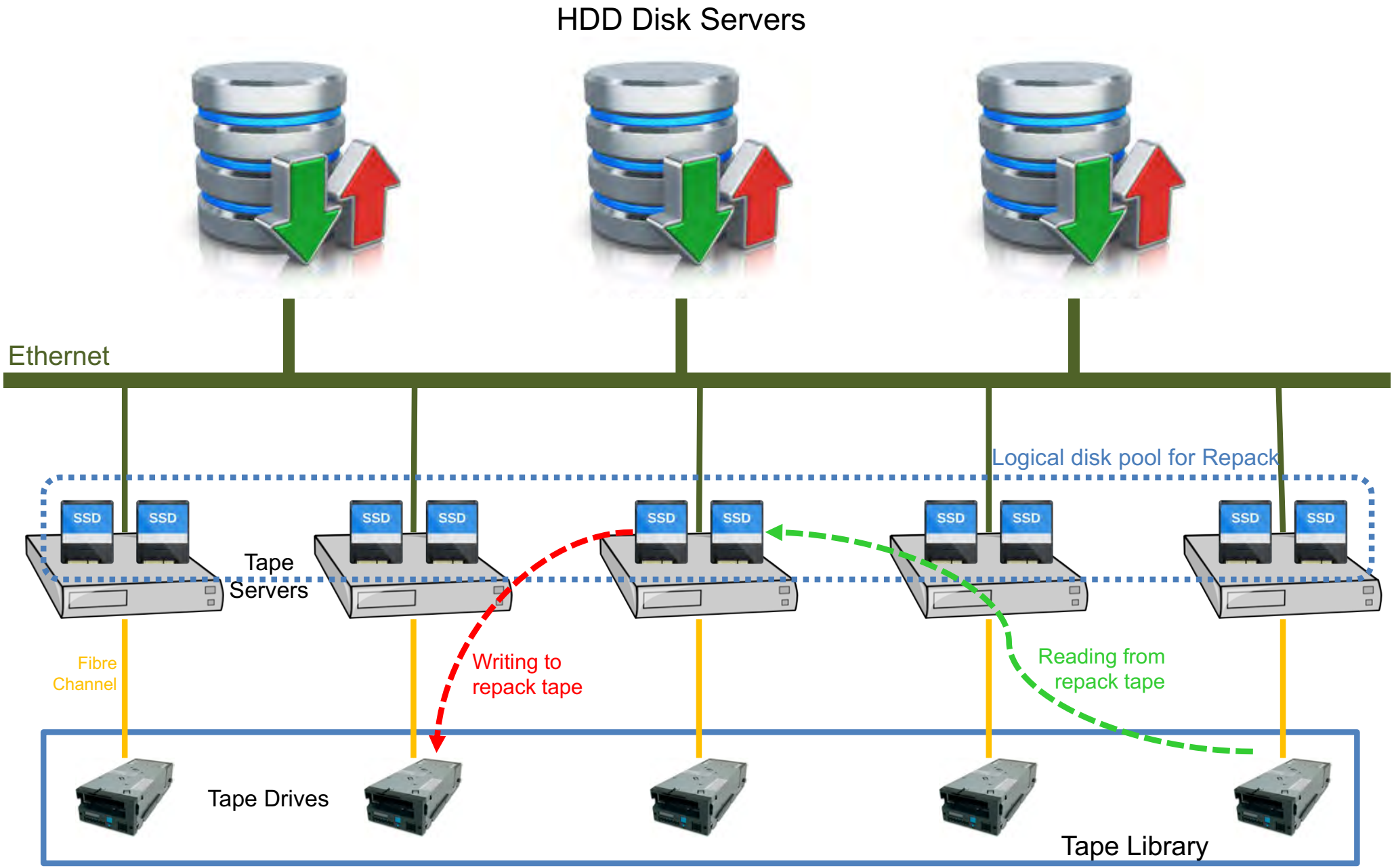
Usual Data Flow (including earlier repack)



Adding SSDs


HDD Disk Servers
Ethernet
SSD
SSD
SSD
SSD
SSD
SSD
SSD
SSD
SSD
SSD
Tape Servers
Fibre Channel
Tape Drives
Tape Library

New Data Flow – Repack only



Setup details

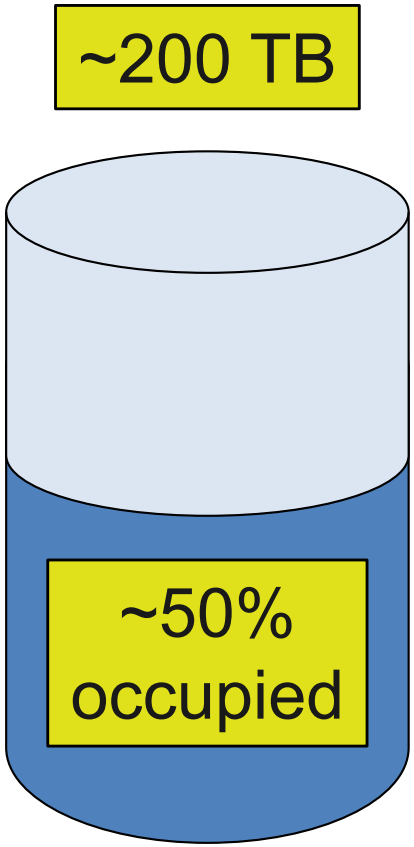
- Tape servers sharing network switch(-es):
  - Faster transfers thanks to using the inter-connects
  - Backbone network traffic offloaded

Model	160 x Samsung MZ7KM960HAHP-00005	128 x Intel (added later) SSDSC2BB960G7
Capacity	960 GB	960 GB
Sequential Read (up to)	510 MB/s	450 MB/s
Sequential Write (up to)	485 MB/s	380 MB/s
Endurance Rating (Lifetime Writes)	1400 TBW	1750 TBW
Mean Time Between Failures (MTBF)	2 Million hours (228 years)	

- Infrastructure Endurance Rating (Lifetime Writes): ~400 PBW

- Used RAID 0 (striping) to combine the disks to get the maximum space and transfer rates
  - No worries about the data loss as data on tape anyway

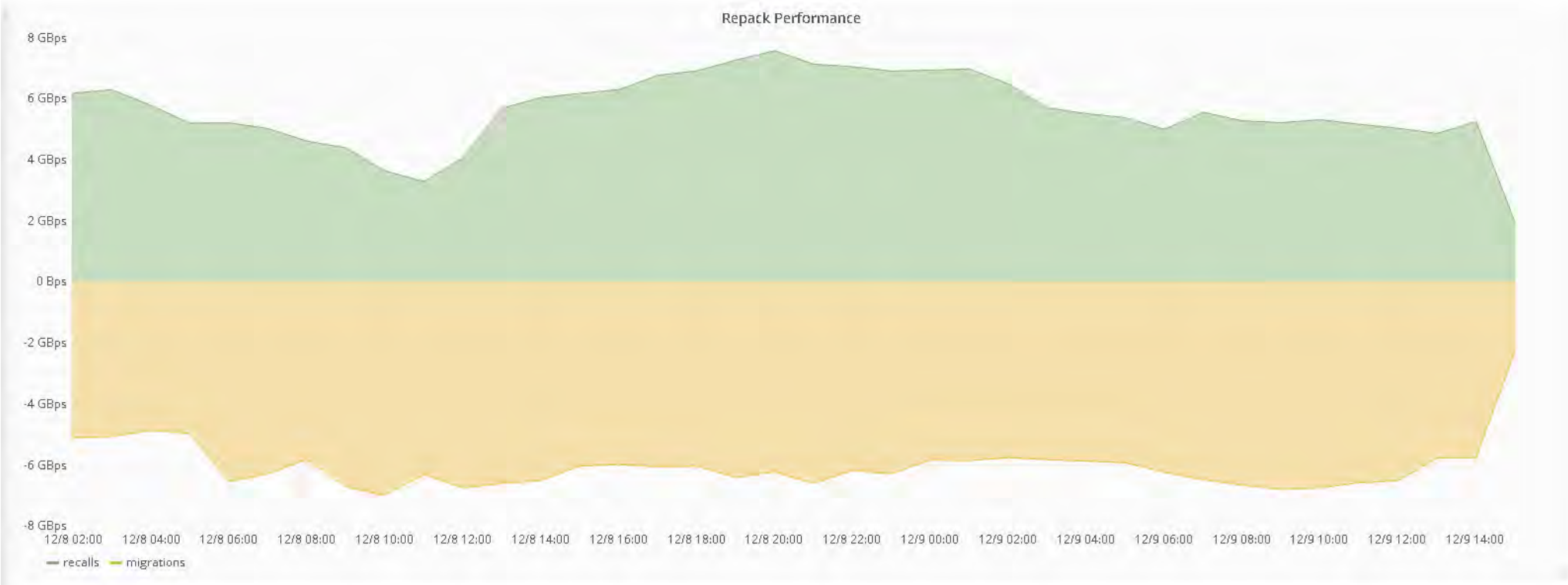
Objectives



- ~100 PB / 4 months = ~800 TB / day
- 800 TB / day = ~ 9 GB/s full duplex = 18 GB/s
- Tape drive counts to meet the throughput (each can transfer data at ~350 MB/s):
  - >27 for writing (will have 46)
  - >27 for reading (will have 24)
    - Will need to use some new drives for reading
- ~200 TB buffer, keep it at ~50% full

Startup phase performance

- Can sustain over 6 GB/s (incomplete setup)
- Challenge to keep the line(s) flat while sharing drives with experiments



Reality check – libraries

- IBMLIB3 – TS4500
  - Recently upgraded
    - More drive slots
  - 24 new TS1155 (read/write)
  - 22 old TS1150 (read)
  - ~3800 tapes
  - ~83 tapes / drive
- IBMLIB4 – TS3500
  - Historically longer
    - Lot of cartridges
  - 22 new TS1155 (read/write)
  - 2 old TS1150 (read)
  - ~5700 tapes
  - ~237 tapes / drive

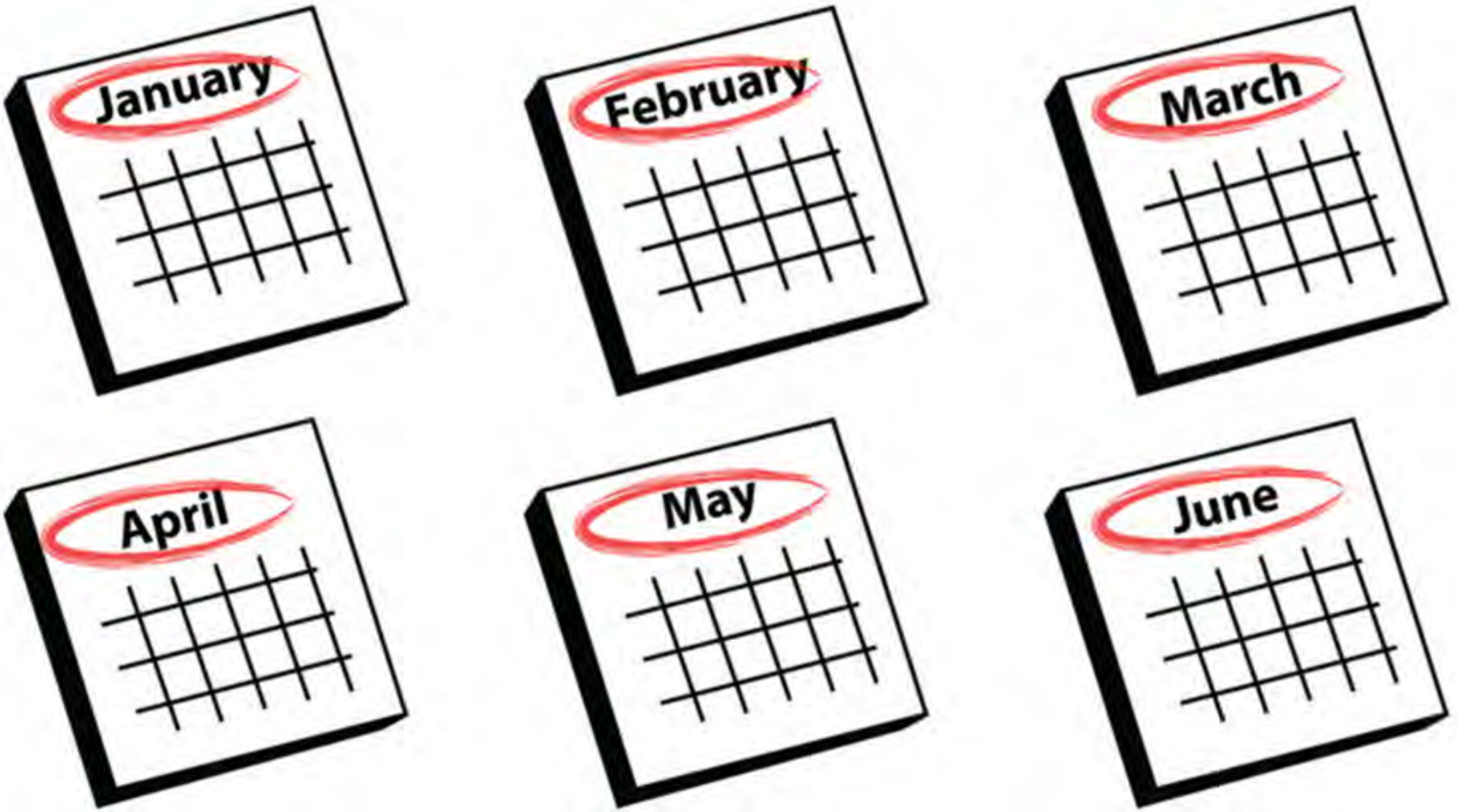
Reality check – users

- CMS: >1100 tapes to recall
- ATLAS: >1300 tapes to recall
- RQF0904697: „I maybe need to work on being more zen ... strategy for running during the holiday period …“
- Load follows the data
  - As soon as critical mass of data was migrated, load will increase on the new drives
- Would help:
  - Inform experiments and ask them to delete unnecessary data so there is less data to move

Solution(s)

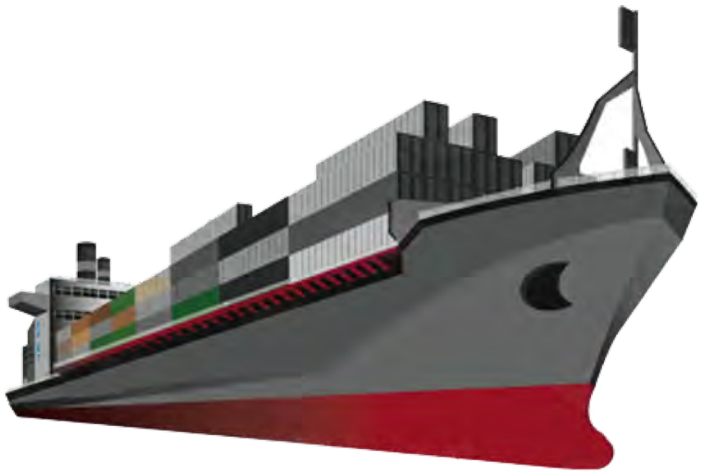
- Need to move tapes where the drives are
  - A human can relocate at least 2 PB from IBMLIB4 into IBMLIB3 in a day
- Tapes ordered by last written to
  - Less interference with experiments who want to read recent data
- Leave few drives dedicated to experiments
  - To avoid user jobs being stuck in the queue
- Full verifications replaced by lite (but LBP in production now)
- Emptied tapes are moved to a special reclaim pool awaiting pre-labeling at higher densities
  - This is to increase the time between: reading old data – verification – destructive upformat of old cartridges at higher density

Around 6 months later ...



Disk buffer management

- Repack READ jobs
  - Occupy tape drives for a long time
  - Should not be interrupted
    (otherwise you do no advance)
- Repack WRITE jobs
  - Usually (slightly) faster than READ
  - Multiple parallel streams
  - Should have lower priority than user WRITE jobs



- Two mechanisms:
  - READ jobs throttling based on the disk pool occupancy:
    < 35% – maximum throughput
    > 35% but < 55% – slowing down
    > 55% – stop
  - Flexibly dedicate tape drives to users as their queue grows

Issues encountered

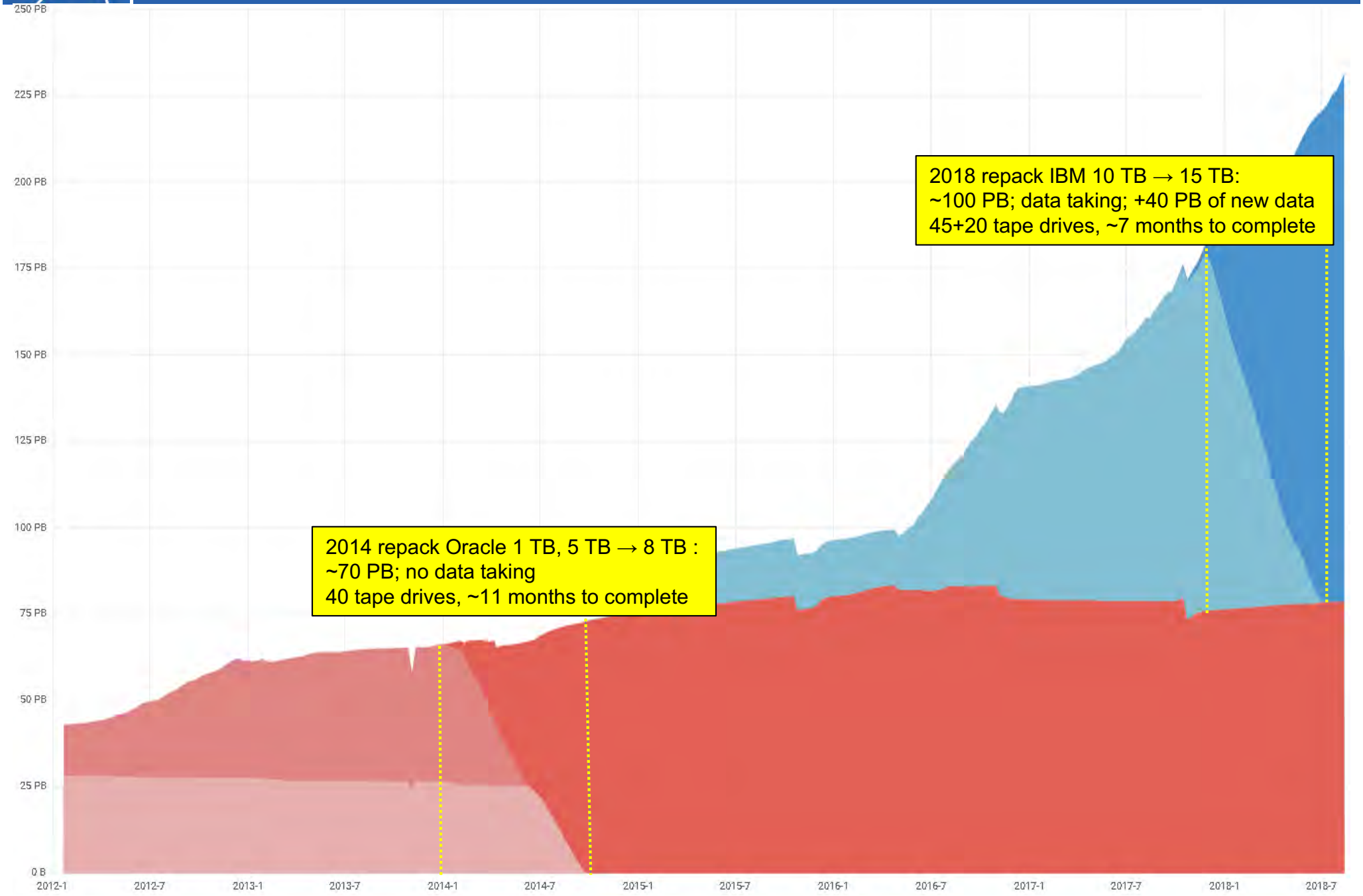
- Mostly smooth ride ....

- Some 10TB tapes had ~100TB of data (each)
  - Highly compressible data need to reserve initial size on disk for de-compression
  - With a small buffer, many parallel streams were needed to consume the data

- One cartridge with user data was overwritten (during the labeling process)
  - Library firmware problem / missing check in our application
  - Data recovered after involving IBM Germany & Japan
    - Complications due to density change

Any Data Loss in those ~100PB?

Comparing Repacks



Conclusion

- Getting SSDs closer to tape can increase the overall throughput of the infrastructure
  - Even a small(-ish) SSD (buffer) can make a difference
- Proper buffer input/output management is crucial
- We observed 99.99999%* reliability of the IBM 3592JD (Barium Ferrite) media
  - Needed to use IBM TS1155 drives to read some problematic tapes
- We managed to liberate sufficient space in time for the new data from the experiments

* Calculated as 10 GB out of 100 PB effectively lost.