
CERN

21st International Conference on Computing in High Energy and Nuclear Physics **CHEP2015** Okinawa Japan: April 13 - 17, 2015

Experiences and Challenges running CERN's High-Capacity Tape Archive

Germán Cancio, Vladimír Bahyl **{German.Cancio,Vladimir.Bahyl}@cern.ch**
Daniele Kruse, Julien Leduc, Eric Cano, Steven Murray

Presented by Eric Cano **Eric.Cano@cern.ch**

Outline

- Tape at CERN
- Large-scale media migration
- Archive reliability
- Software and management tools
- LHC Run-2 and beyond
- Conclusion

Tape at CERN: Overview

Data:

- ~100 PB physics data (CASTOR)
- ~7 PB backup (TSM)

Tape libraries:

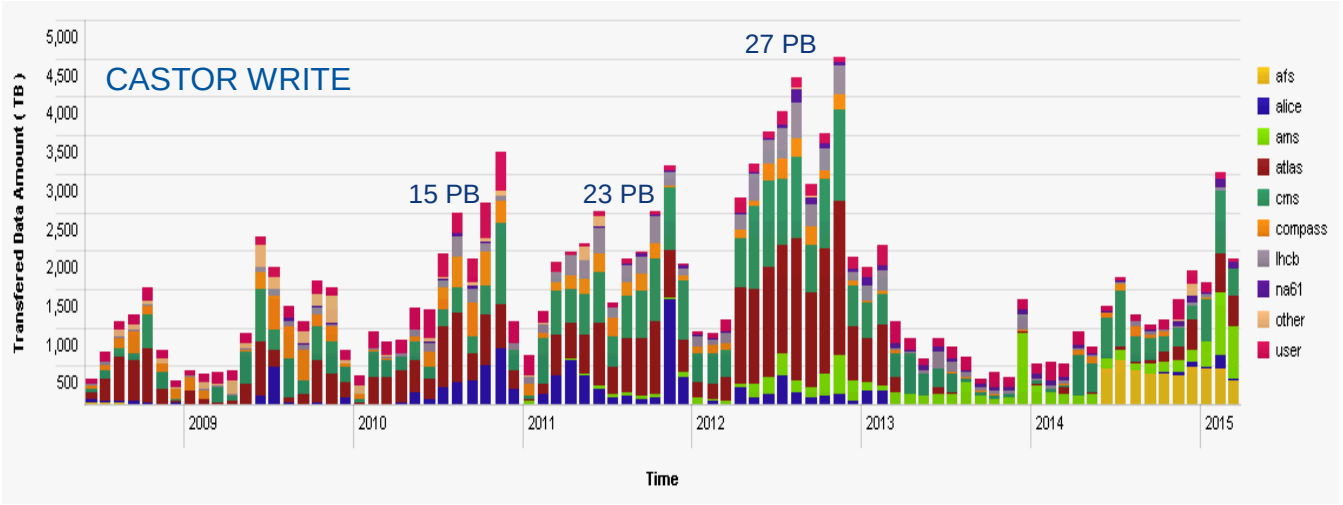
- IBM TS3500 (3+2)
- Oracle SL8500 (4)

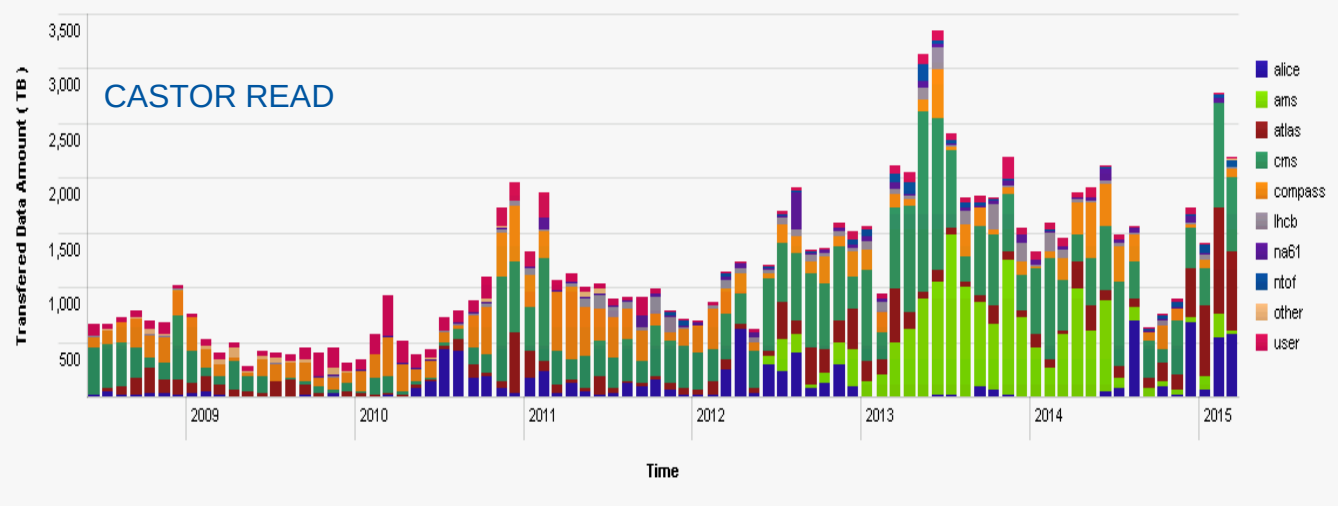
Tape drives:

- 100 archive
- 50 backup

Capacity:

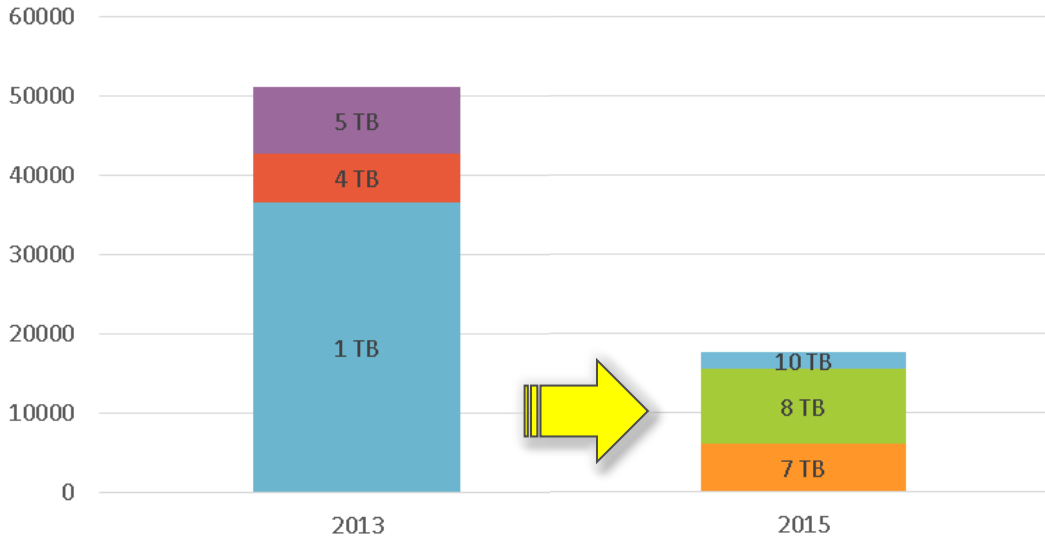
- ~70 000 slots
- ~30 000 tapes



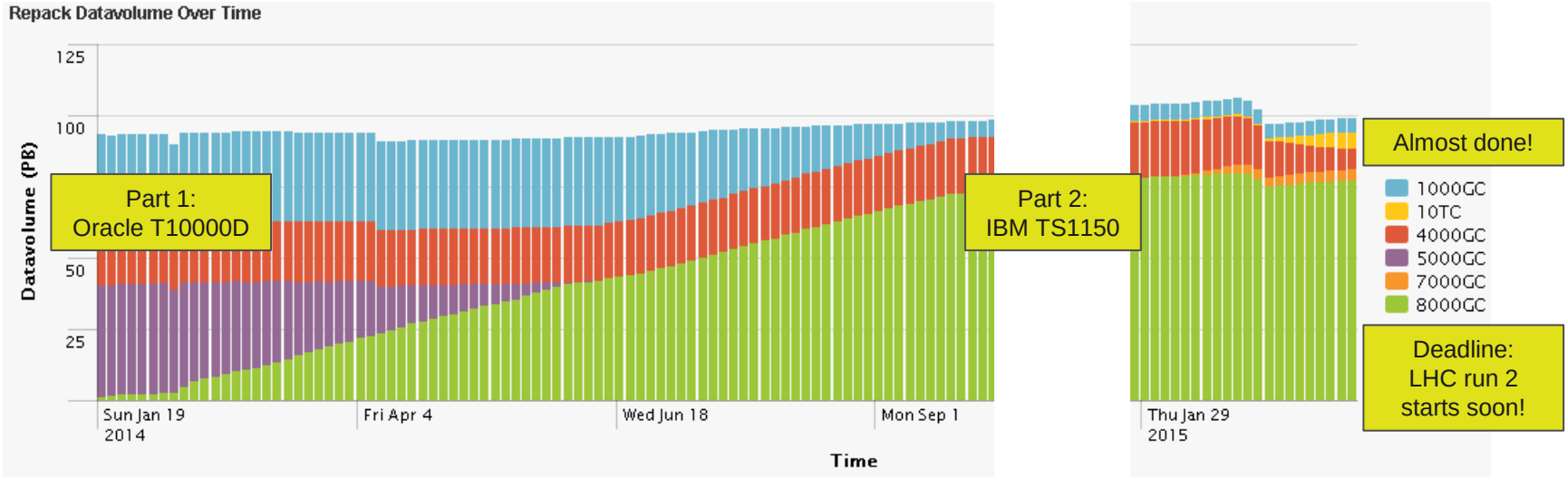


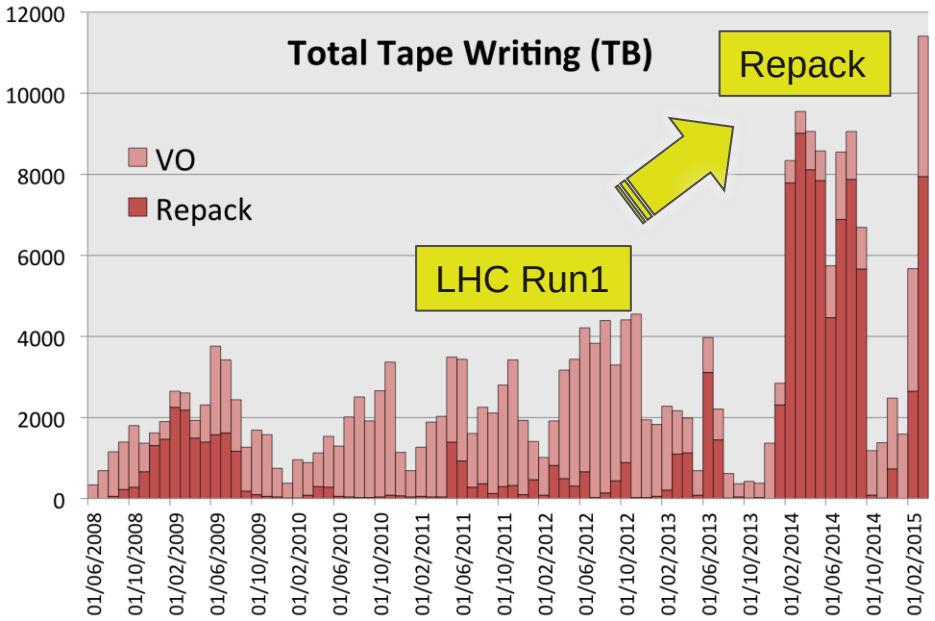
Large scale media migration

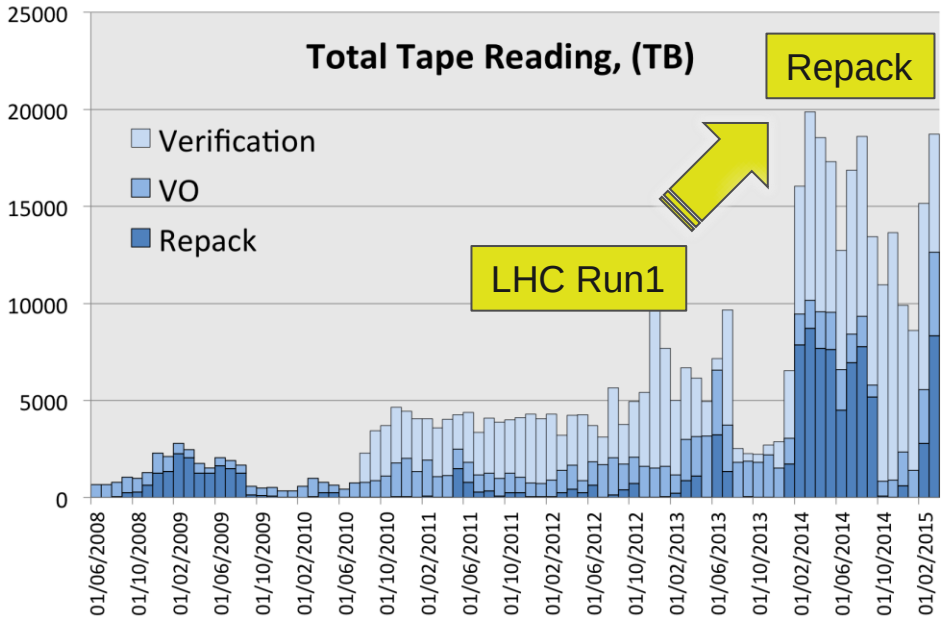
- Challenge:
  - ~100 PB of data
  - 2013: ~51 000 tapes
  - 2015: ~17 000 tapes
  - Verify all data after write
    - 3x (300PB!) pumped through the infrastructure (read->write->read)
  - Liberate library slots for new cartridges
    - Decommission ~33 000 obsolete tape cartridges
- Constraints:
  - Be transparent for experiment activities
  - Exploit the high speeds of the new tape drives
  - Preserve temporal collocation
  - Finish before LHC run 2 start



Large scale media migration

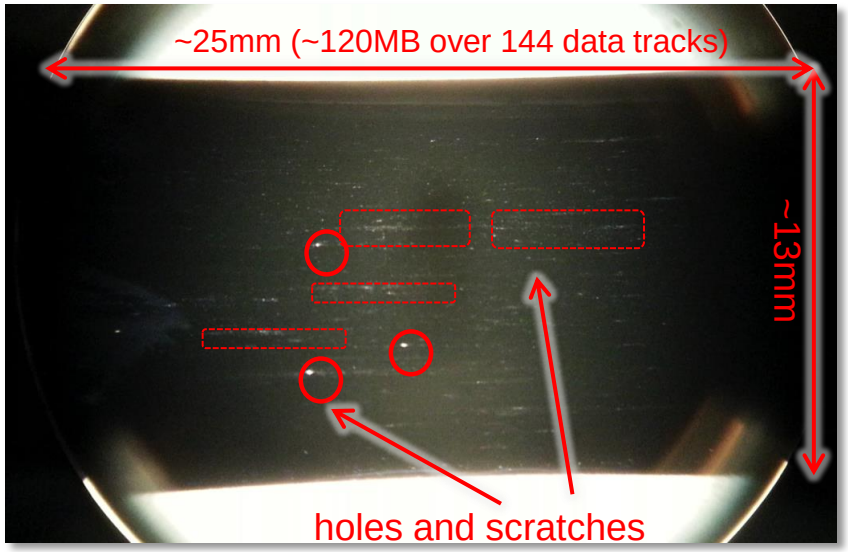




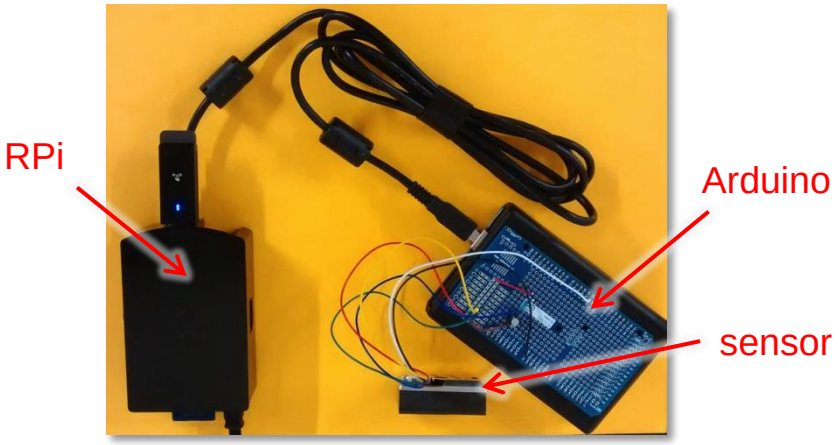


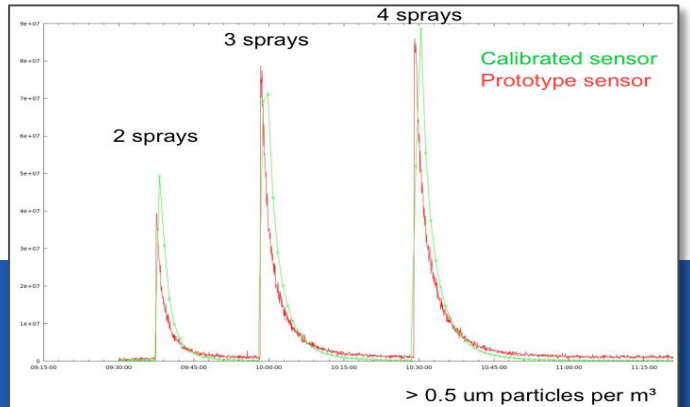
Dust incident

- Identified 13 tapes in one library affected by concrete or foam particles
- Isolated incident by verifying all other tapes in the building
- Recovered 94% files with custom low-level tools and vendor recovery; 113 files lost



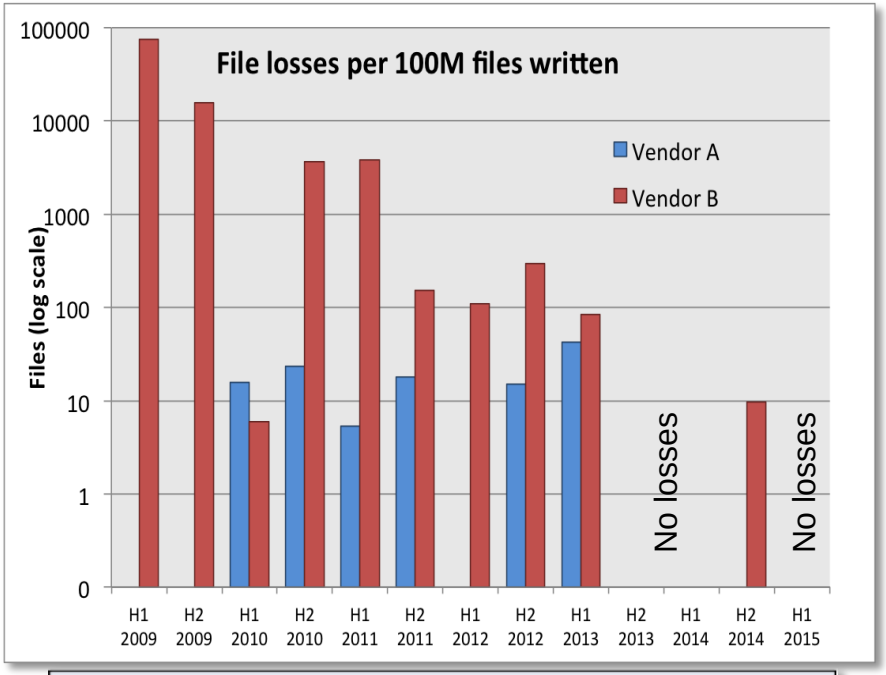
- Fruitful exchanges with other tape sites on CC protective measures (access and activity restrictions, special clothing, air filters etc)
- Library cleaning by specialist company envisaged
- Prototyped a dust sensor to be installed inside libraries, using cheap commodity components, achieving industrial precision and reaction time



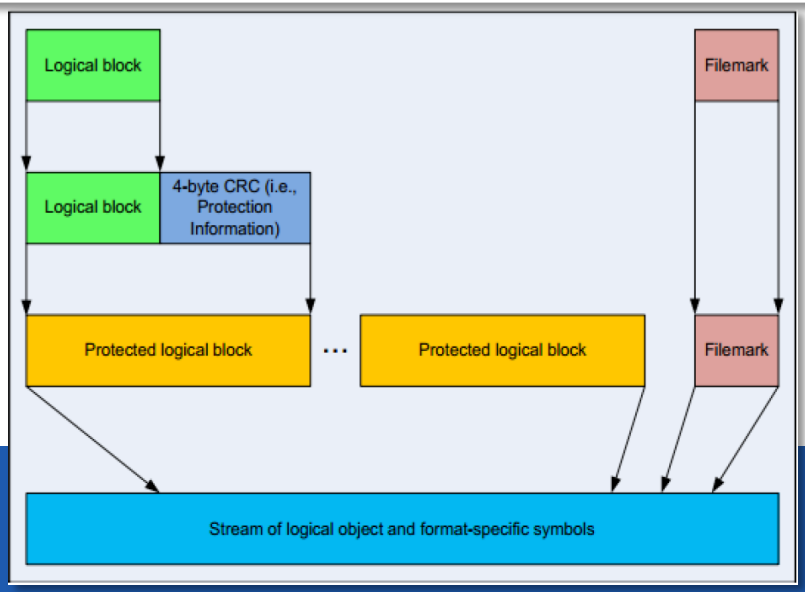


Archive Reliability

- Bit-preservation techniques to improve archive reliability
  - Annual 2012-2015 bit loss rate: O(10^{-16})
  - Systematic verification of freshly written + “cold” tapes
  - Less physical strain on tapes (HSM access, buffered tape marks)
  - With new hardware/media, differences between vendors getting small
  - For smaller experiments, creating dual copies on separated libraries / buildings



- Working on support for SCSI-4 Logical Block Protection
  - 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%)
  - Supported by LTO and enterprise drives

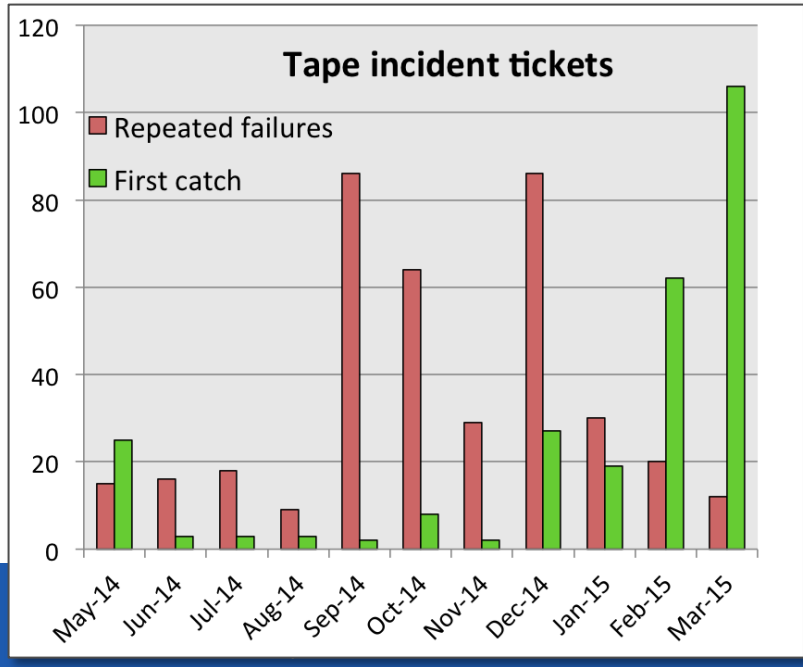


Software and management tools

- New CASTOR tape software developed and deployed in production
  - Completely redesigned architecture, moved from C to C++
  - Improved error detection / handling, full support for SCSI tape alerts, soon LBP
  - Support for multiple access protocols (RFIO, XROOT), soon Ceph
  - More details: cf poster by E. Cano in Session B

- Investigating direct-to-tape backend to EOS (avoid double disk layer)

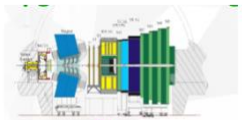
- Re-engineered Tape Incident System
  - Taking advantage of full SCSI tape alerts
  - Automated problem identification: tape vs. drive vs. library
  - Better detection of root cause -> catch problems and disable faulty elements earlier
  - Comprehensive media repair workflow



LHC Run-2 and beyond (1)

- Run-2 (2015-2018): Expecting ~50PB/year of new data (LHC + non-LHC)
  - +7K tapes / year. CERN has now ~35'000 free library slots
- Run-3 (-2022): ~150PB/year. Run-4 (2023 onwards): 600PB/year!
  - Peak rates of ~80GB/s

LHCb & ALICE @ Run 3



40 MHz

DAQ

40 MHz

LLT: μ,e/γ, hadrons

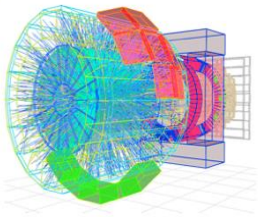
5-40 MHz

HLT: event recon.

20 kHz (0.1 MB/event)

Storage

2 GB/s



50 kHz

Reconstruction + Compression

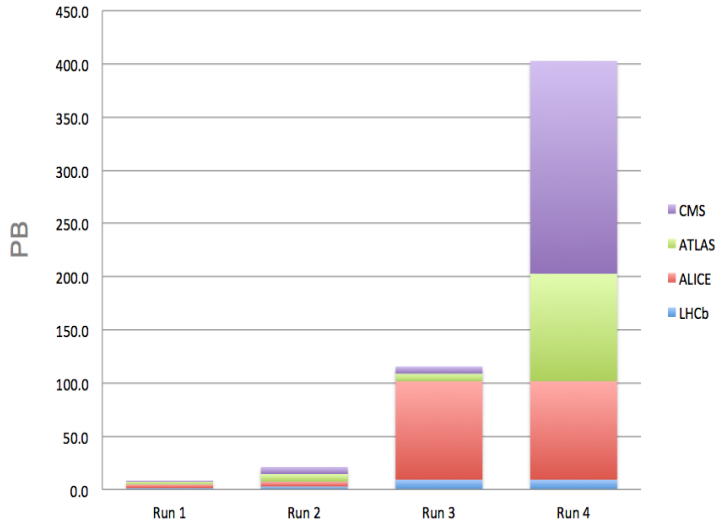
50 kHz

Storage

75 GB/s

← PEAK OUTPUT →

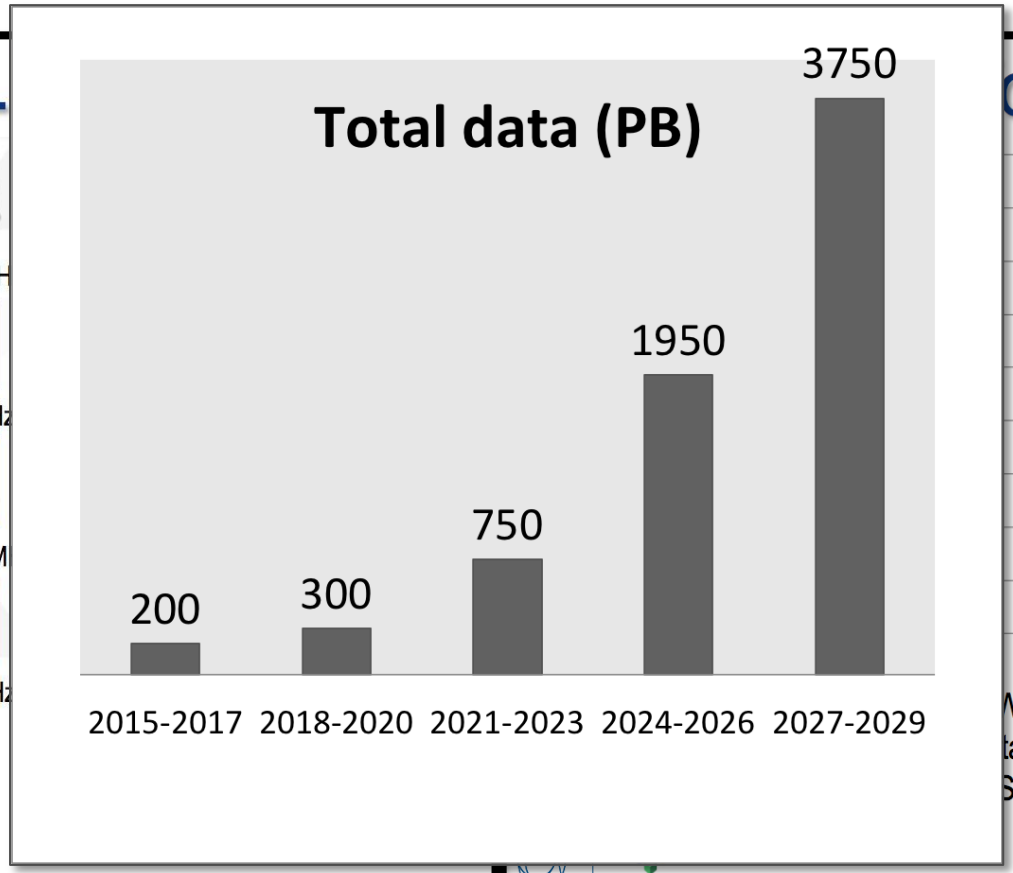
Data: Outlook for HL-LHC

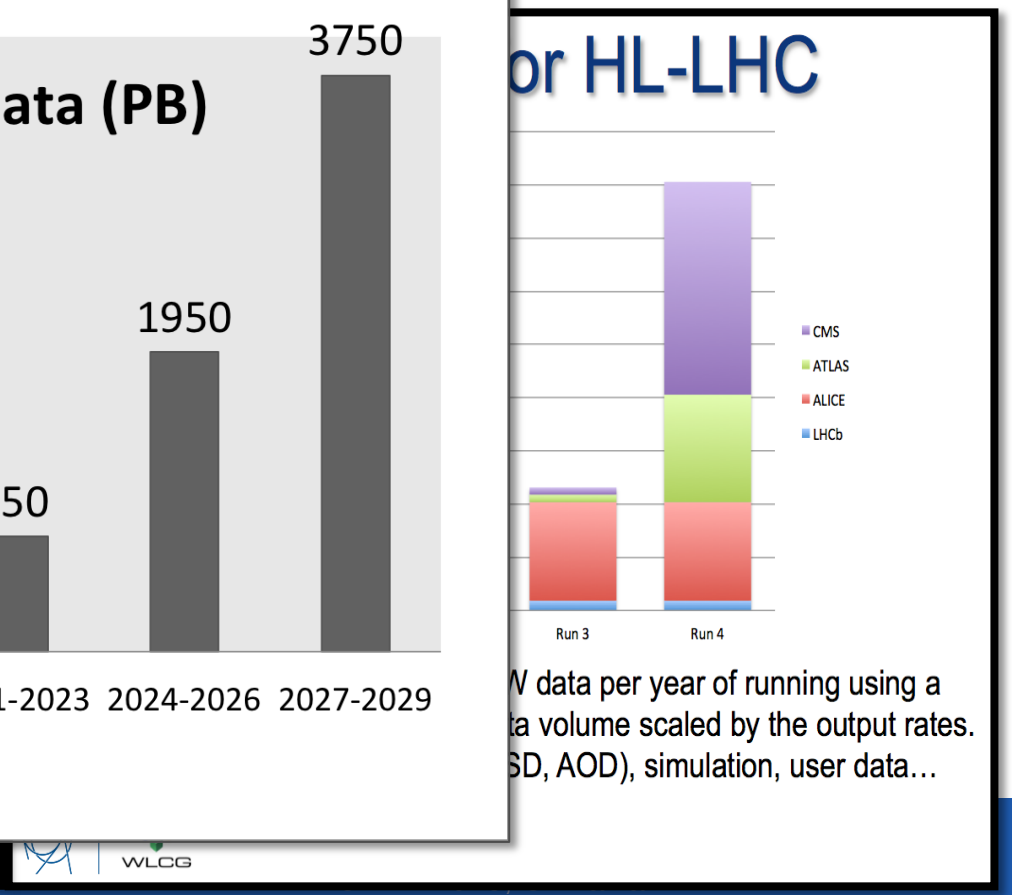


- Very rough estimate of a new RAW data per year of running using a simple extrapolation of current data volume scaled by the output rates.
  - To be added: derived data (ESD, AOD), simulation, user data…

LHC Run-2 and beyond (1)

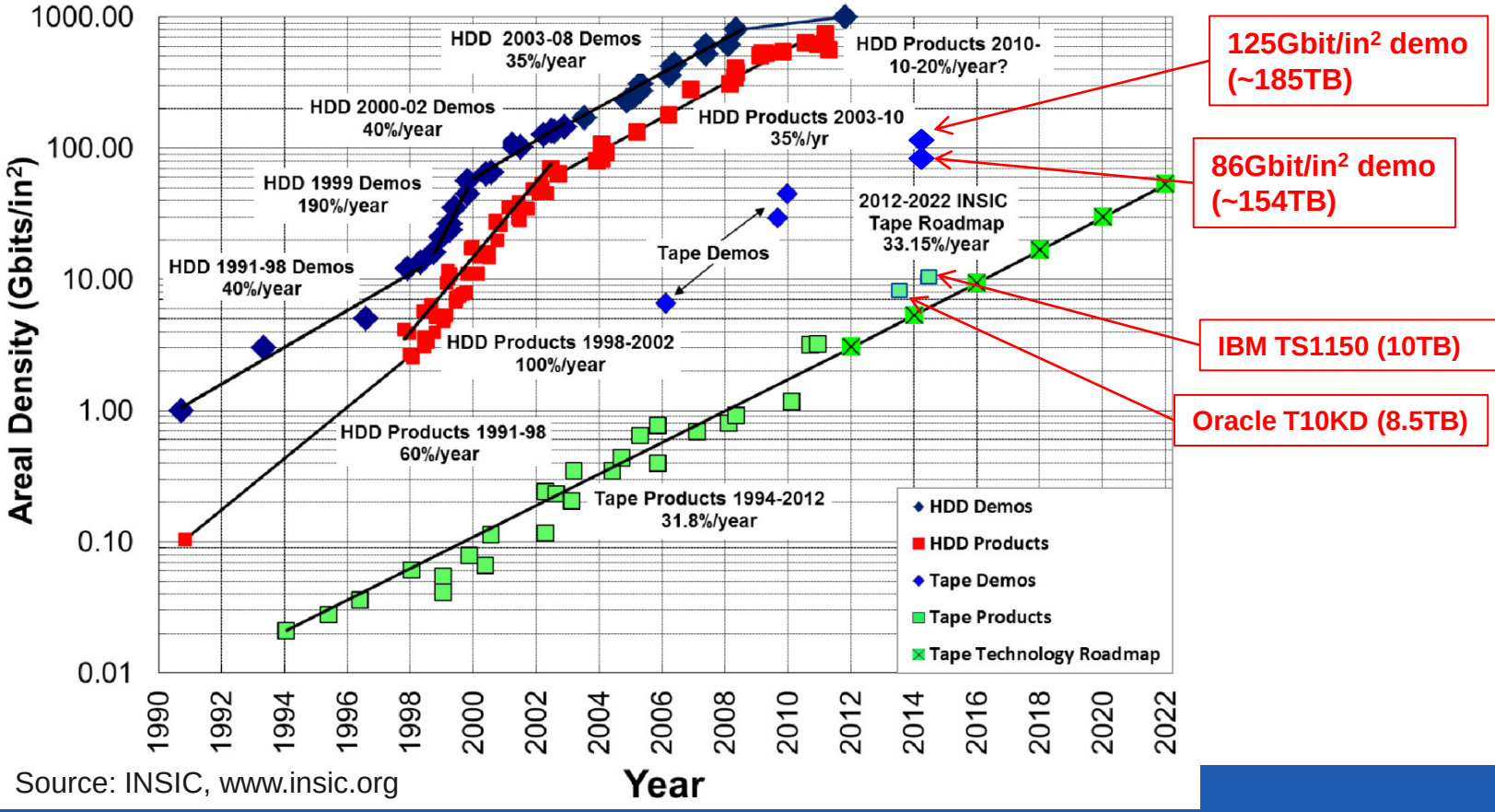
- Run-2 (2015-2018): Expecting ~50PB/year of new data (LHC + non-LHC)
  - +7K tapes / year. CERN has now ~35'000 free library slots
- Run-3 (-2022): ~150PB/year. Run-4 (2023 onwards): 600PB/year!
  - Peak rates of ~80GB/s





LHC Run-2 and beyond (2)

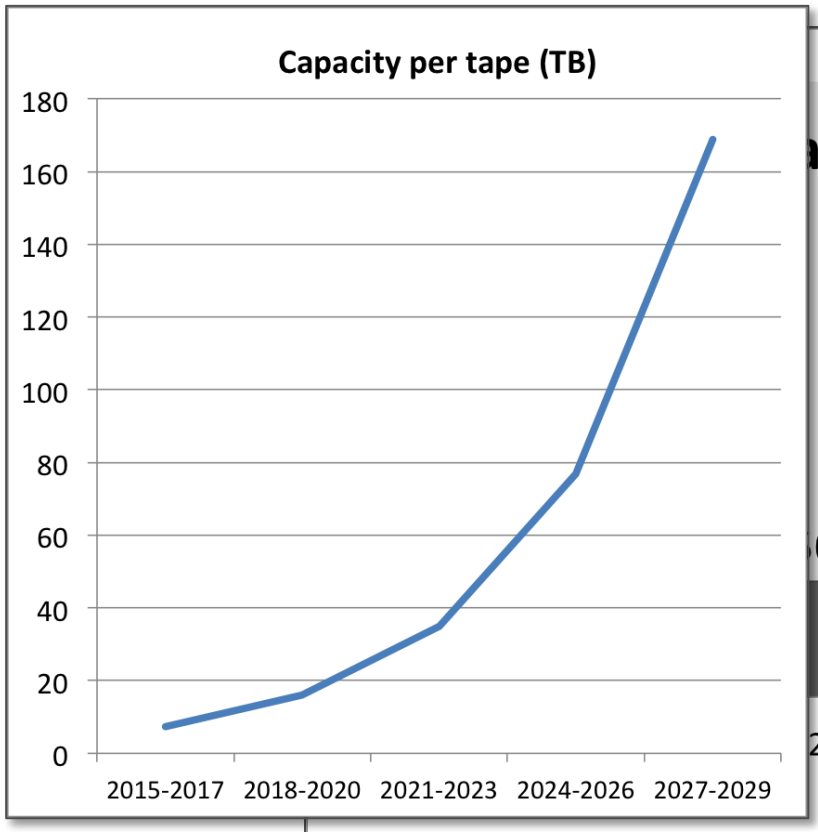
- Technology/market forecast (…risky for 15 years!)
- INSIC Roadmap:
  - +30% / yr tape capacity per $ (+20%/yr I/O increase)
  - +20% / yr disk capacity per $

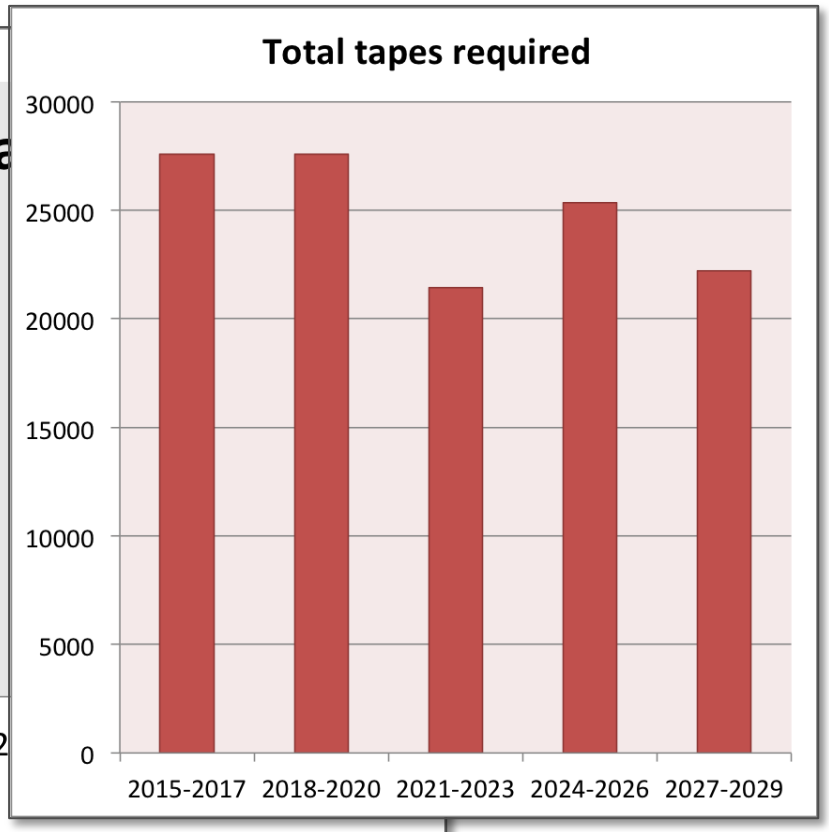


Source: INSIC, www.insic.org

LHC Run-2 and beyond (2)

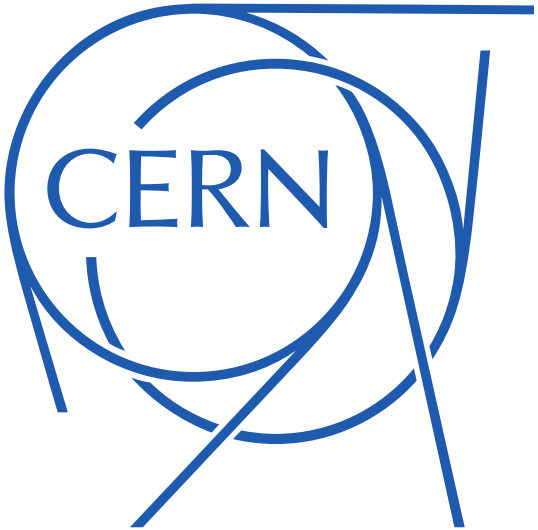
- Technology/market forecast (…risky for 15 years!)
- INSIC Roadmap:
  - +30% / yr tape capacity per $ (+20%/yr I/O increase)
  - +20% / yr disk capacity per $





Conclusion

- CERN's Tape Archive is at the core of physics data storage and archiving
- Successfully dealt with LHC Run-1 and a large media migration during the Long Shutdown
- Improving reliability and bit-level data preservation has become a key and long-term activity
- Focus on having archive infrastructure, software and tools ready and scalable for LHC Run-2 and beyond


CERN