Update on Bit Preservation, HEPiX WG and Beyond

Germán Cancio, IT-DSS-TAB
CERN

DPHEP Collaboration Workshop
CERN, 8/9 June 2015

Outline

- CERN Archive, current numbers
- Large scale media migration (repack) and outlook
- Environmental hazards
- Current reliability and improvements
- Client access evolution (aka The Demise Of RFIO)
- HEPiX WG status

CERN Archive current numbers

Data:

- ~105 PB physics data (CASTOR)
- ~7 PB backup (TSM)

Tape libraries:

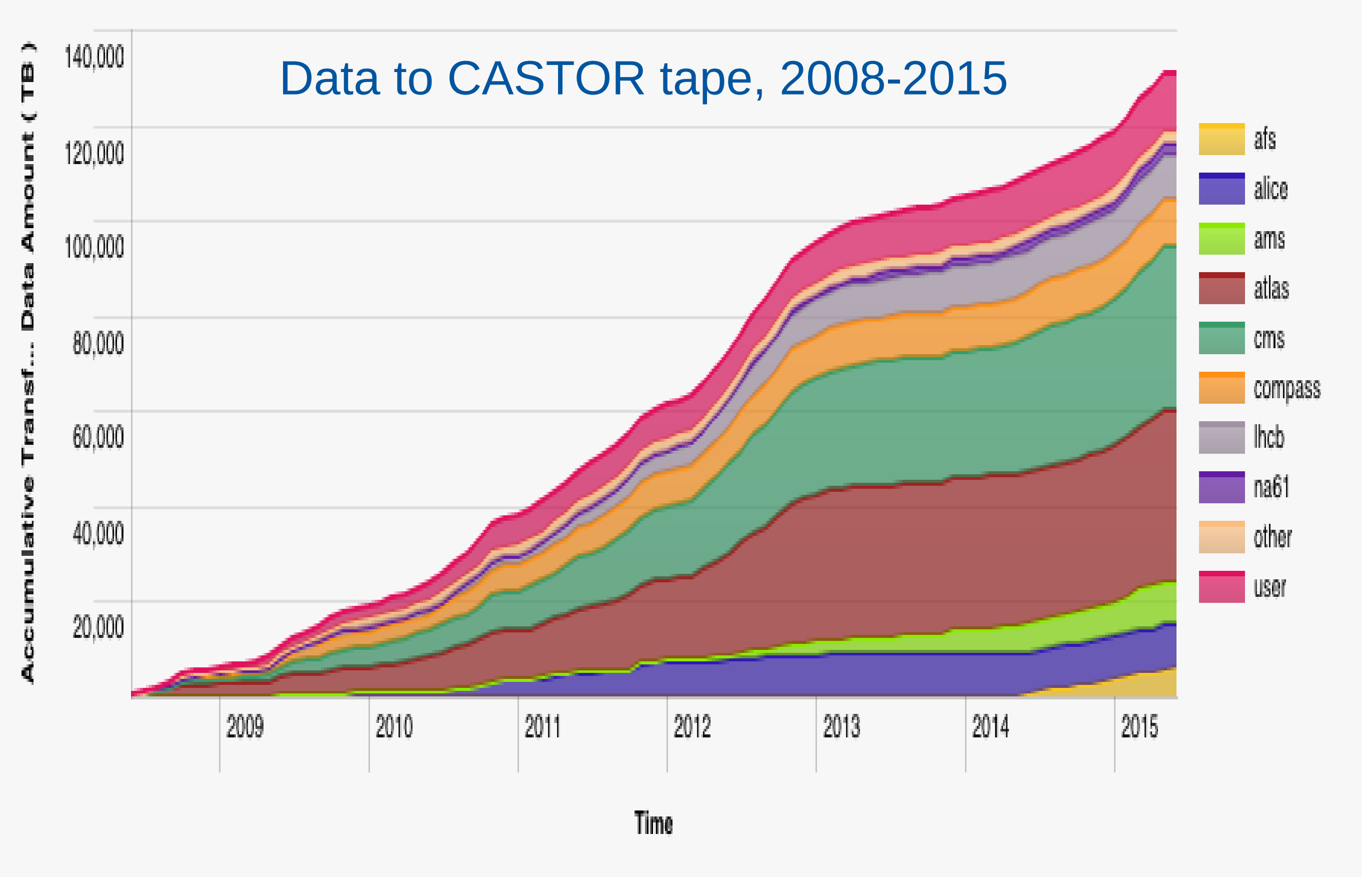
- IBM TS3500 (3+2)
- Oracle SL8500 (4)

Tape drives:

- ~100 archive

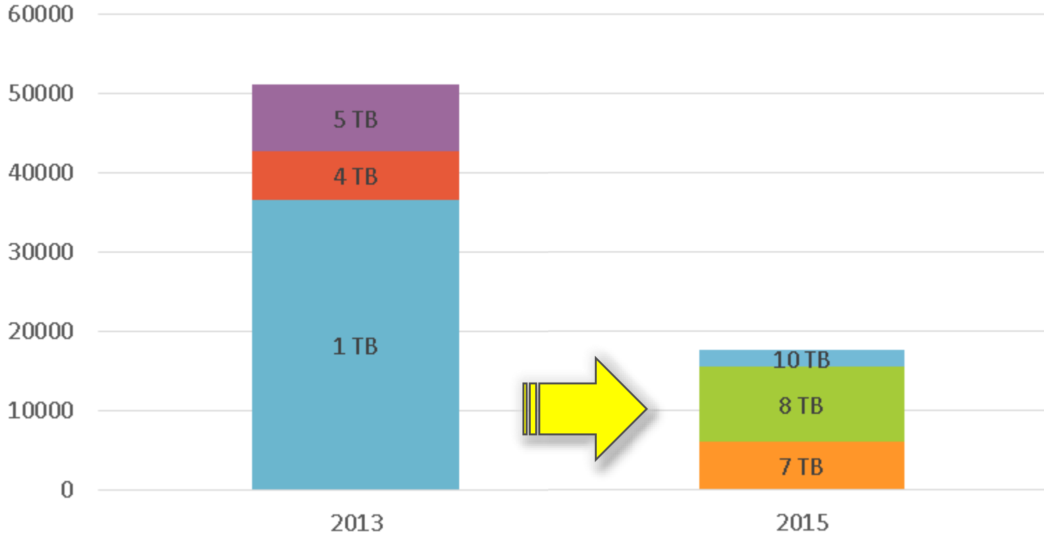
Capacity:

- ~70 000 slots
- ~25 000 tapes



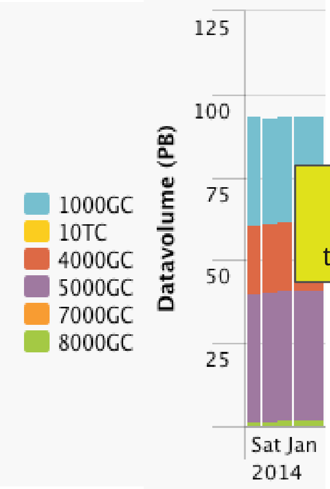
Large scale media migration

- Challenge:
  - ~85 PB of data
  - 2013: ~51 000 tapes
  - 2015: ~17 000 tapes
  - Verify all data after write
    - 3x (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

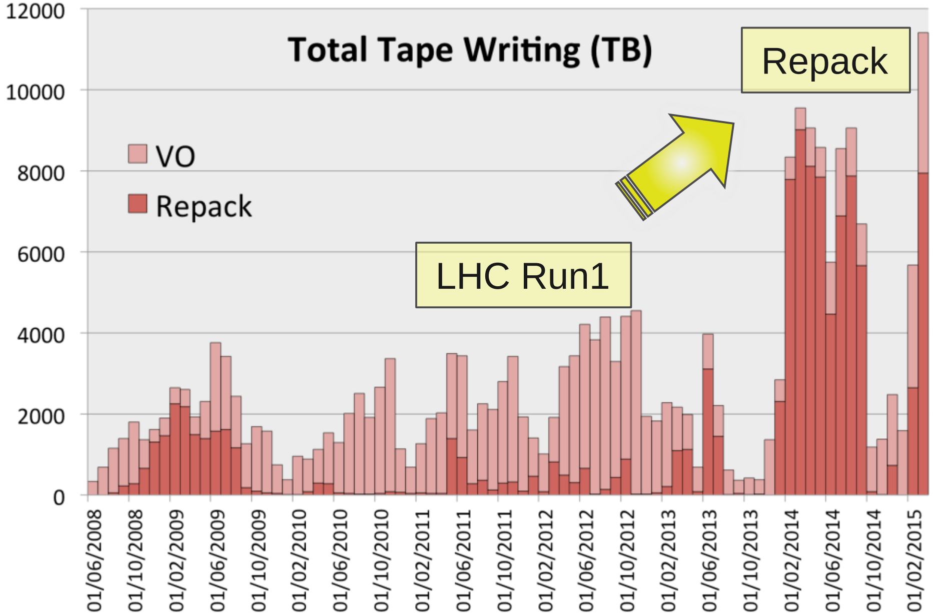
Large media migration: Repack



Part 1:
Oracle 5->8TB
then empty 1TB

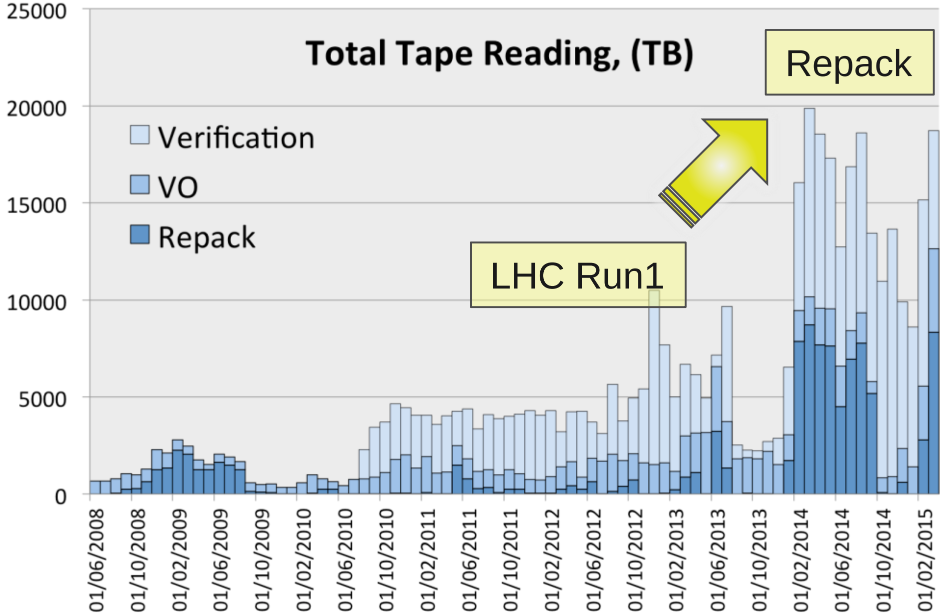
Part 2:
IBM 4->7TB
then 1TB

Completed
last week!



Repack

LHC Run1

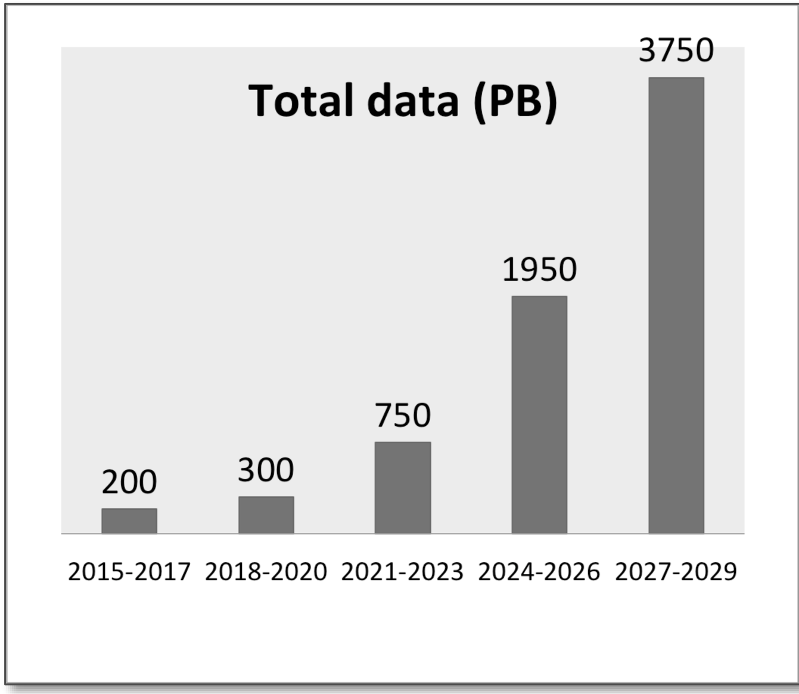


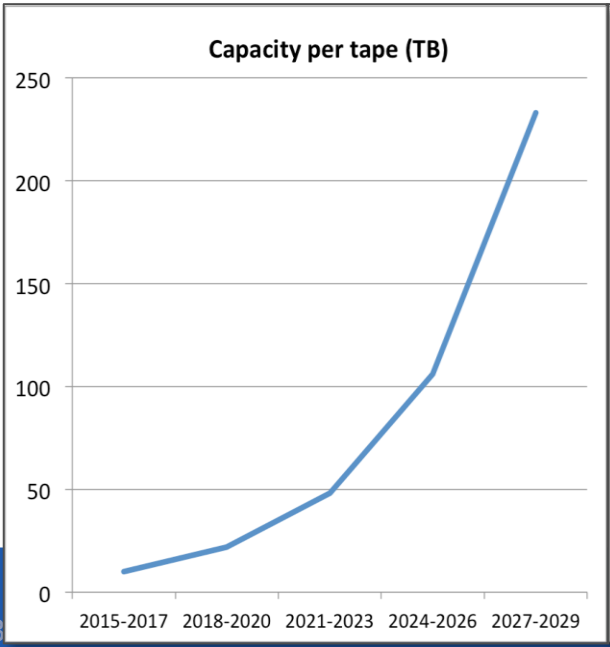
Repack

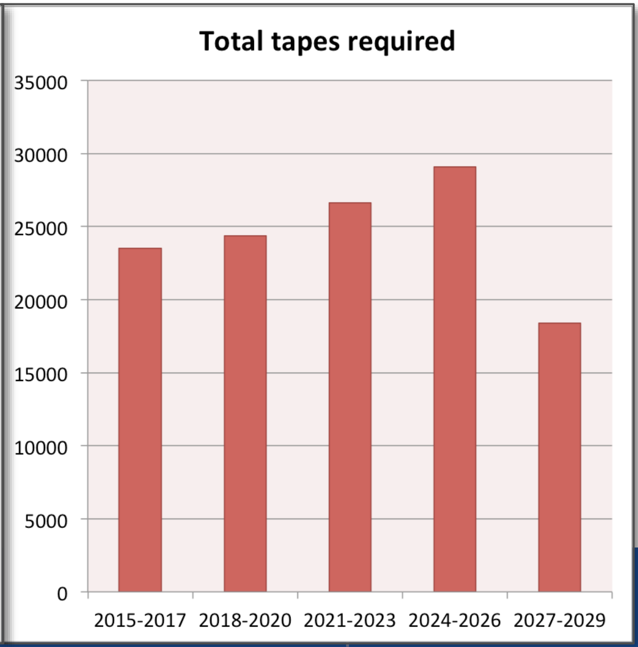
LHC Run1

Future…

- Run-2 (2015-2018): Expecting ~50PB/year of new data (LHC + non-LHC)
  - +7K tapes / year (~35’000 free library slots)
- Run-3 (-2022): ~150PB/year. Run-4 (2023 onwards): 600PB/year..
- .. tape technology grows faster
  - tape roadmaps at 30% CAGR for at least 10 years
  - demo for 220TB tape by IBM/Fujifilm in April
- … but: market evolution is difficult to predict
  - Tape media: monopoly in shrinking market
  - disk: “duopoly”
  - Cloud storage solutions
  - Disk capacity slowdown (HAMR) .. may slowdown tape products!
  - storage slowdown == higher archiving costs

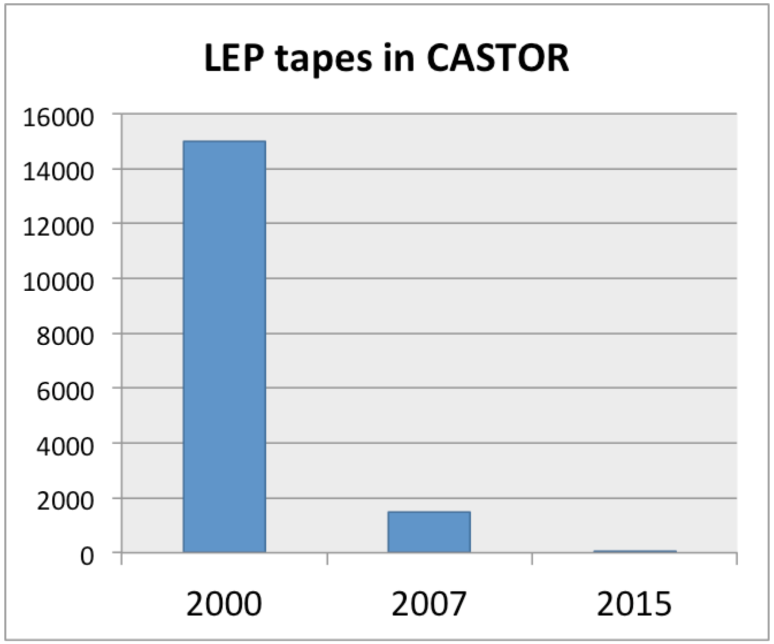






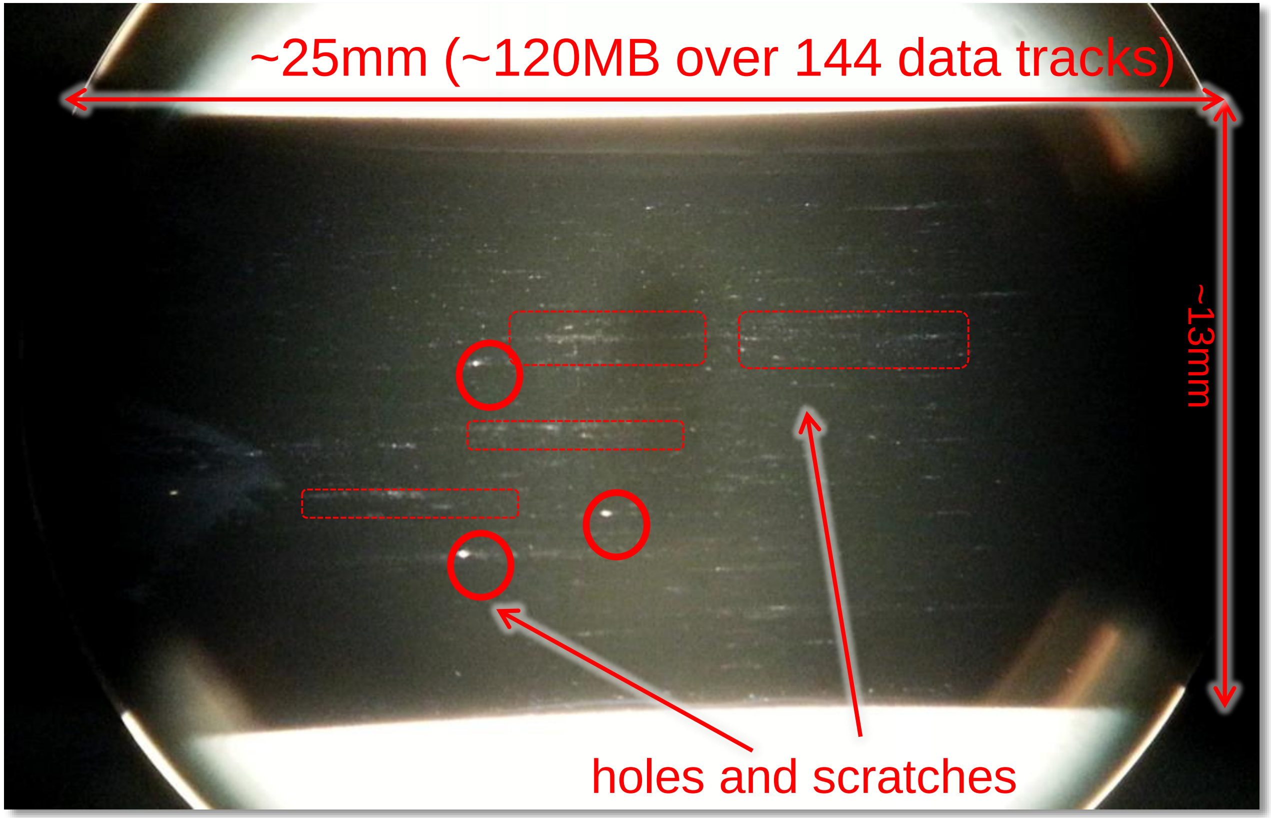
… and the past

- LEP-era data: ~370TB
- 2000:
  - ~ 15’000 tapes
- 2007:
  - ~ 1500 tapes
- 2015:
  - 30 tapes… x 2 (replicated in separate buildings)
  - Cost:



Tape contamination 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

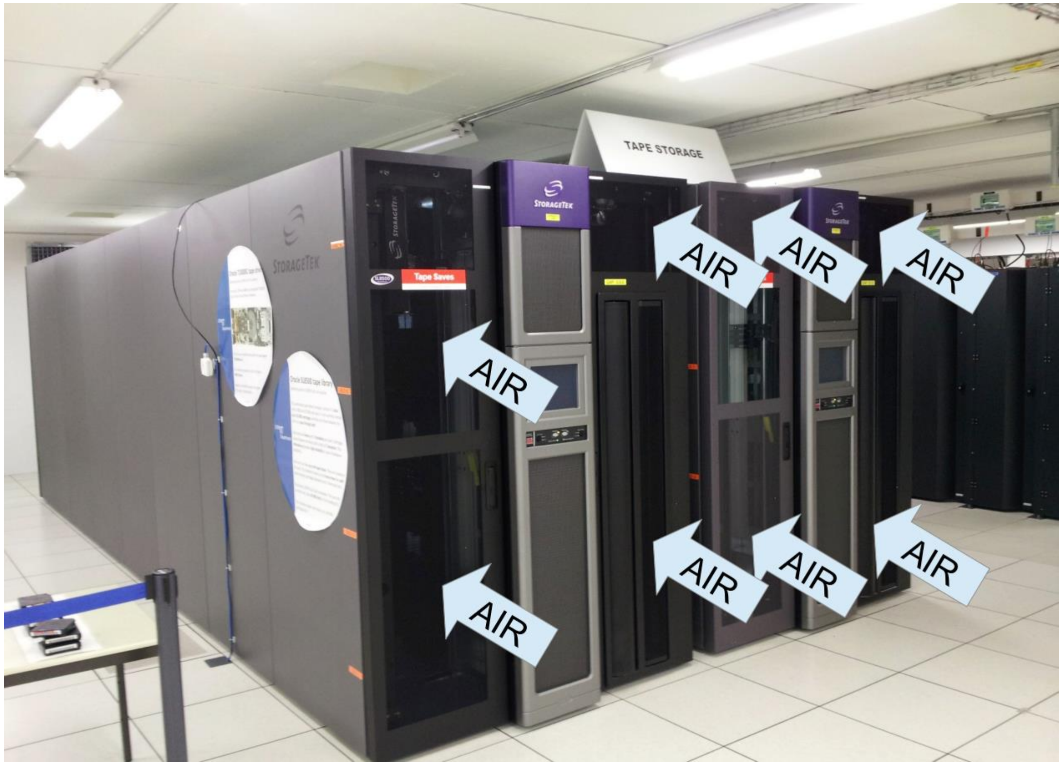


Airflows in tape libraries

- (Our) tape libraries are not sealed nor filtered
- Over 30m^3/min of airflows per library
  - (Home vacuum cleaner: ~2m^3/min)
  - On top of already existing strong CC airflows
- Operating environment required for new-generation drives: ISO-14644 Class 8 (particles / m^3):

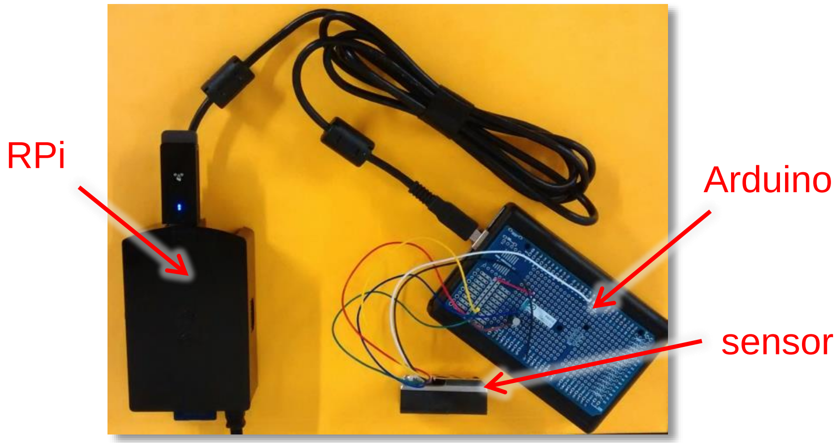
Class	>0.5 um	>1 um	>5 um
8	3 520 000	832 000	29 300

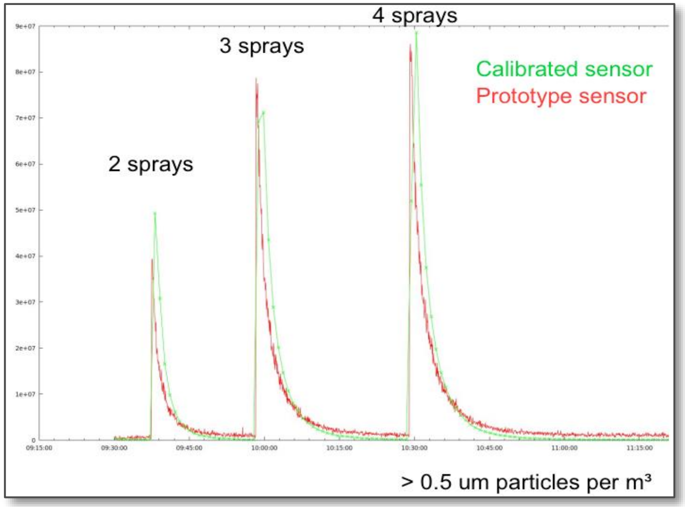
- Environmental sensitivity will continue increasing with newer drives as tape bit density grows exponentially



Environmental protection

- Fruitful exchanges with other HEP tape sites on CC protective measures (access and activity restrictions, special clothing, air filters etc)
- Sampling by external company and corrective actions taken at CERN-CC (air filters)
- Library cleaning by specialist company in June
- Prototyped a set of environmental sensors to be installed inside libraries, using cheap commodity components, achieving industrial precision and reaction time
  - Measure+correlate dust, temperature, humidity
  - Raise alert in case of anomalies
  - Can be integrated inside libraries
  - Done in coordination with vendor, potential for built-in solutions
  - Details: HEPiX Spring 2015 presentation

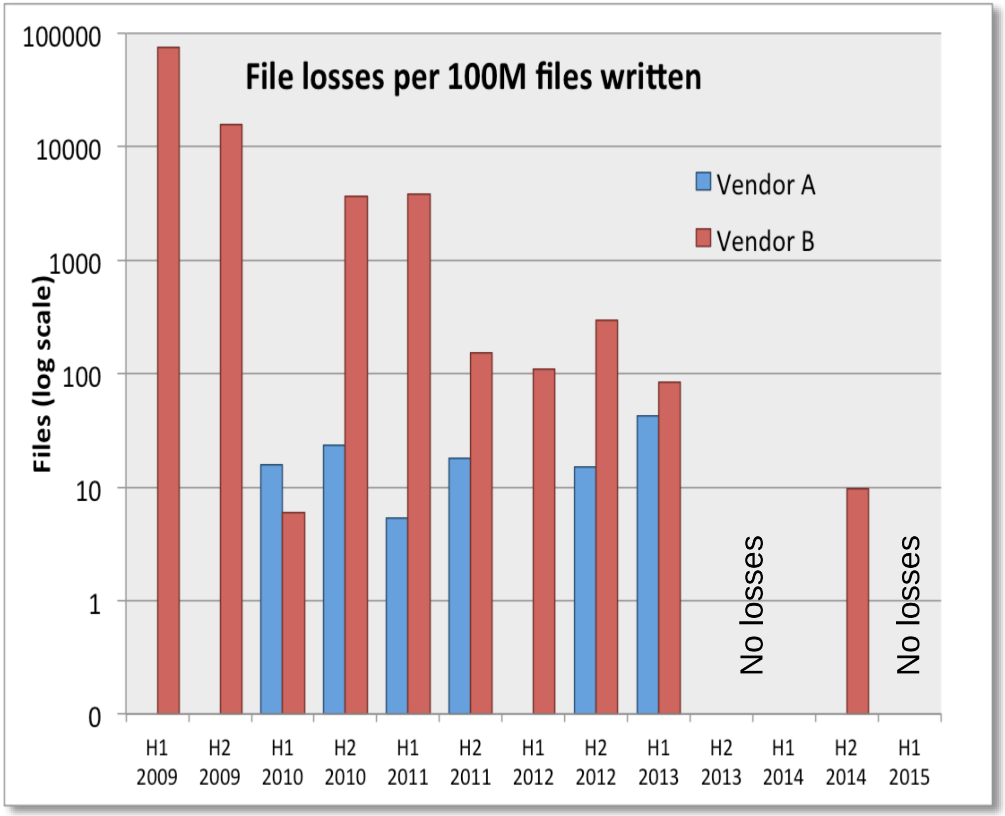




CERN Archive Reliability

Ongoing activity to improve archive reliability

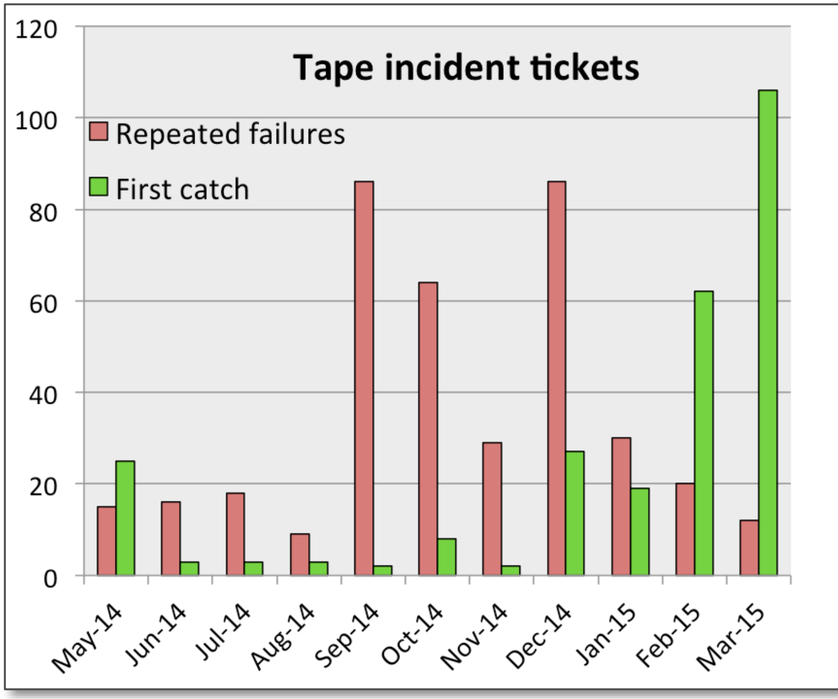
- Continued systematic verification of freshly written + “cold” tapes
- Less physical strain on tapes (HSM access, buffered tape marks)
- With new hardware/media, differences between vendors getting small
- For smaller experiments, created dual copies on separated libraries / buildings



Reliability improvements (1)

- 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

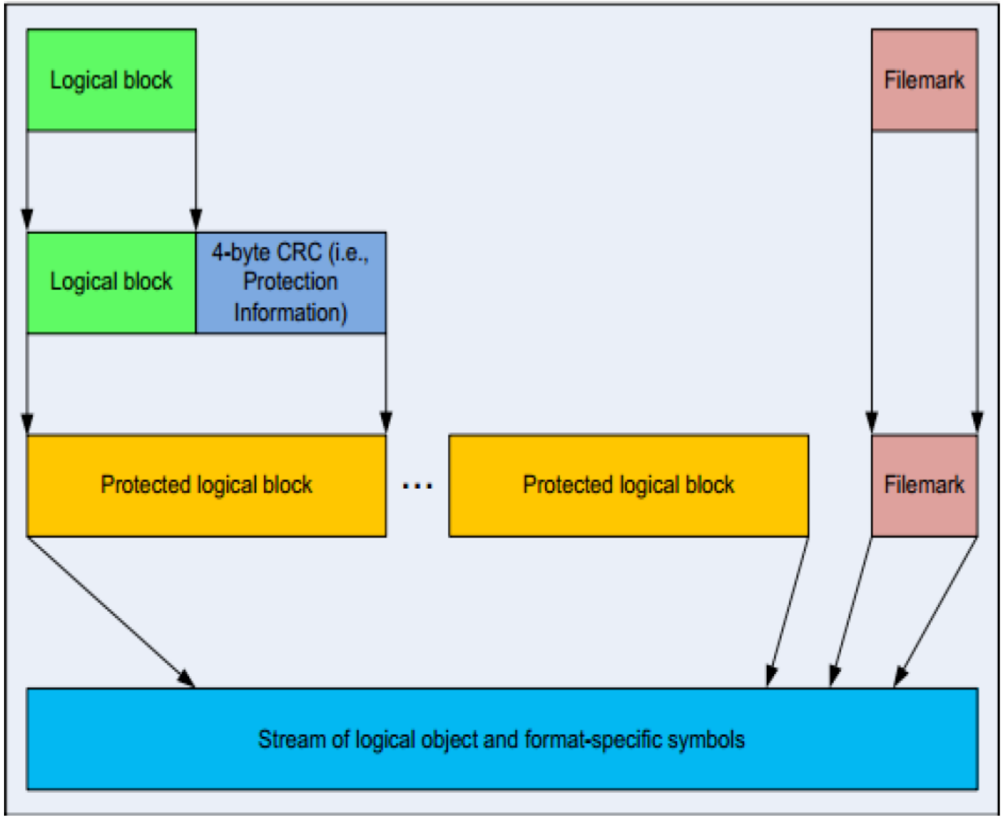
- Re-engineered Tape Incident System
  - Taking advantage of full SCSI tape alerts
  - Automated problem identification: tape vs. drive vs. library
  - Better detection of root cause -> catch problems and disable faulty elements earlier
  - Enhanced low-level media repair tools



- Still much unexploited systems level information
  - Transient/internal drive read/write/mount events at SCSI level; library low-level logs
  - Work area in 2015

Reliability improvements(2)

- 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%)
  - Tape drive can do fast media verification autonomously
  - Supported by newer LTO and enterprise tape drives
  - To be integrated in CASTOR in H2'15


Logical block
Filemark
Logical block
4-byte CRC (i.e., Protection Information)
Protected logical block
...
Protected logical block
Filemark
Stream of logical object and format-specific symbols

Data access evolution

- We migrate and protect your bit streams… but can you still access them?

- Venerable RFIO client/server protocol reaching its end of life and to pass away “soon”, consolidate on de-facto standard (XROOT)
  - RFIO: outdated code base, flaky security model

- Dependencies onto RFIO (and other CASTOR commands) within LEP SW need to be understood, isolated and removed
  - Direct and indirect dependencies (e.g. via CERNLIB/ZEBRA, BOS, …)
- Replace HSM by local file access
  - First copy files from CASTOR to a local file system
  - Then access/process from there via standard open/close/read()... POSIX calls
- Should a RO replica of the LEP era data be made available on as locally accessible files on PLUS/BATCH?
  - RO replica provided by EOS+FUSE
  - All data would continue to be kept in the CASTOR archive

HEPiX bit-preservation WG (1)

- Set up during summer 2013
- Co-chairs: D. Ozerov/DESY and myself
- Mandate:
  - Collect/share bit preservation knowledge across HEP (and beyond)
  - Provide technical advice to DPHEP
  - Recommendations for sustainable HEP archival storage
- w3.hepix.org/bit-preservation

HEPiX bit-preservation WG (2)

- Survey on large HEP archive sites
  - 19 sites; areas such as archive lifetime, reliability, access, verification, migration
  - Overall positive status but lack of SLA’s, metrics, common best practices, long-term costing impact
  - Presented at HEPiX Fall’2013
- Defined a simple, customisable model for helping establishing the long-term cost of bit-level preservation storage
  - Approximate cost of generic (tape-based) data archive over 10-30 years
  - Factors such as media, hw, maintenance cost, media capacity growth rate, etc
  - Different base scenarios (“frozen” to exponentially growing)
  - presented at HEPiX Spring’2014 (and DPHEP cost of curation WS)

HEPiX bit-preservation WG (3)

- WG “frozen” since ~summer 2014
  - D. Ozerov left DESY, no replacement nominated
  - Reduced interest from HEPiX community for active collaboration at WG level…
  - … good exchange of tape sites elsewhere (e.g. workshops such as LTUG, dedicated HEP mailing list, HEPiX!) covering broader topics; MSS-specific forums/workshops for tech implementations (CASTOR, HPSS, etc)
- Still some potential work in the pipeline
  - Complete (unfinished) recommendations for bit-preservation best practices
    - archive protection/verification/auditing/migration, reliability definition, etc.
    - concentrate on “what” rather “how” to do it, align with OAIS ref model
  - Expand and refine cost model
    - active I/O, LTO tapes (non-reusable media), disk-based archiving, manpower, etc.
    - compare to other models such as LIFE, KRDS, California Digital Library model)
  - Exchanges with non-HEP sites (and with related groups eg NDSA Infrastructure WG)

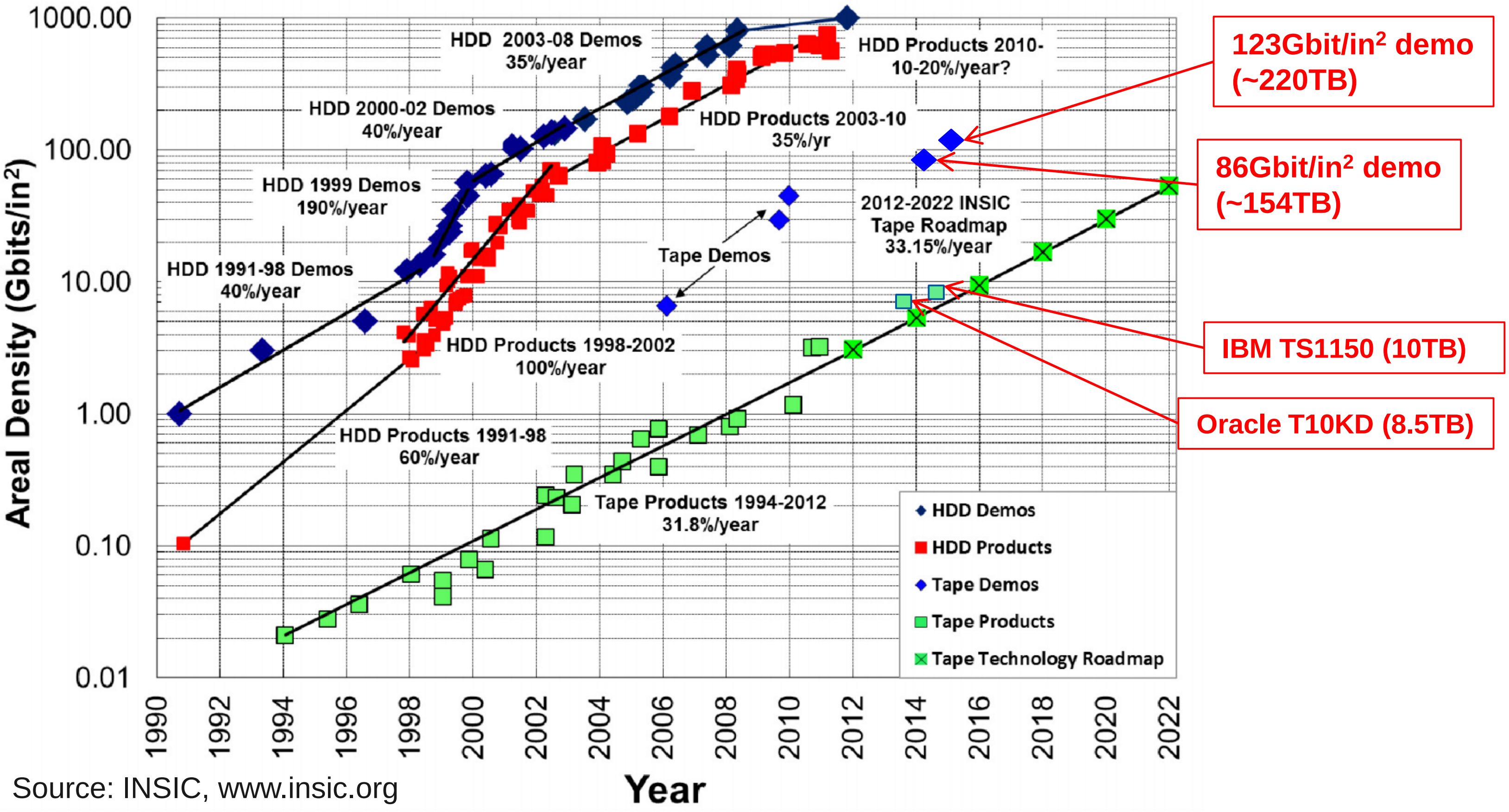
Summary

- CERN’s physics data archive has completed a large media migration during the Long Shutdown
- Assuming a continued market presence, tape still a perfect match for HEP/XXL-scale archiving
- Ensuring longevity of data has become a key and long-term activity: improve reliability, perform bit-level data preservation, ensure environmental conditions
- Applications need to adapt to changing client access protocols
- Collaboration with HEP sites continues also outside HEPiX WG, look outside HEP

CERN

Technology forecast

- +30% / yr tape capacity per $ (+20%/yr I/O increase)
- +20% / yr disk capacity per $



123Gbit/in^2 demo (~220TB)

86Gbit/in^2 demo (~154TB)

IBM TS1150 (10TB)

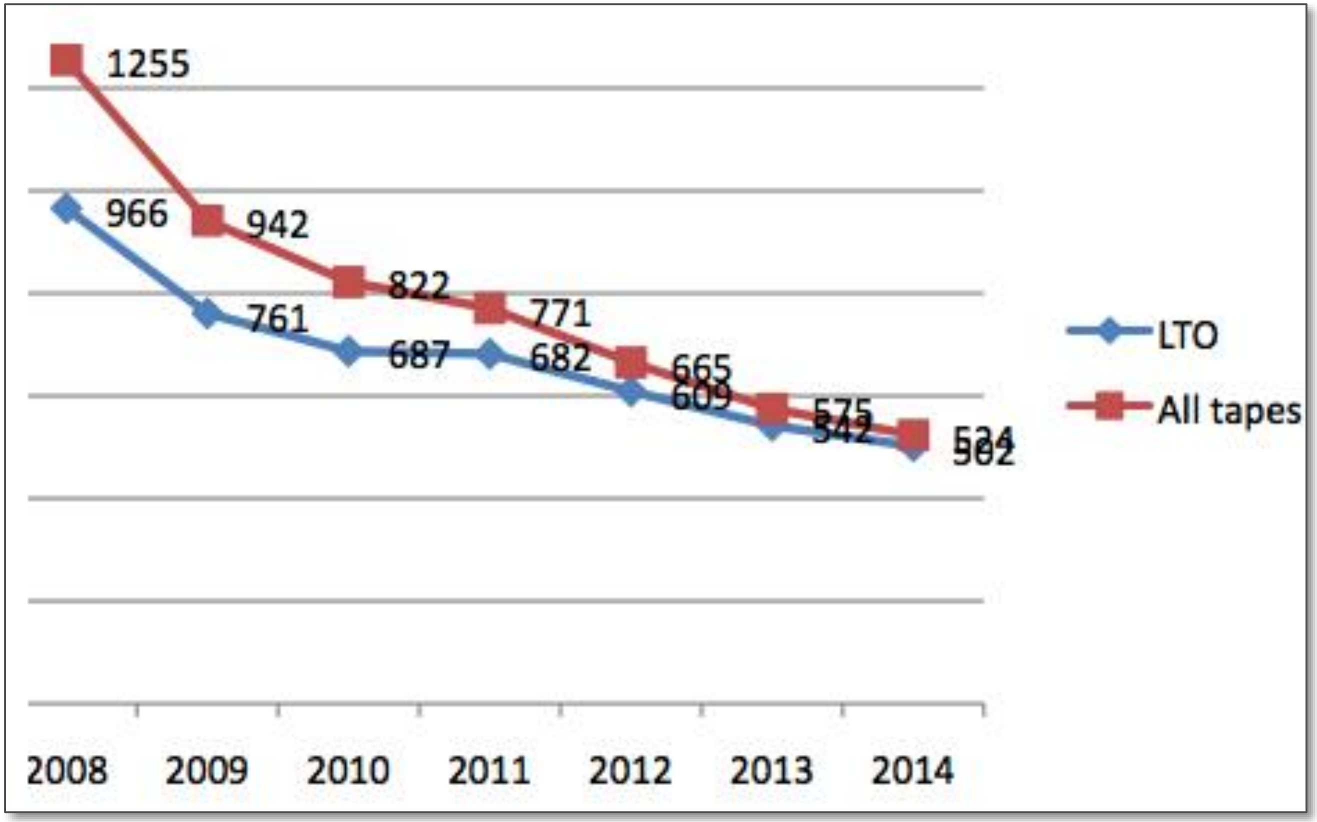
Oracle T10KD (8.5TB)

Source: INSIC, www.insic.org

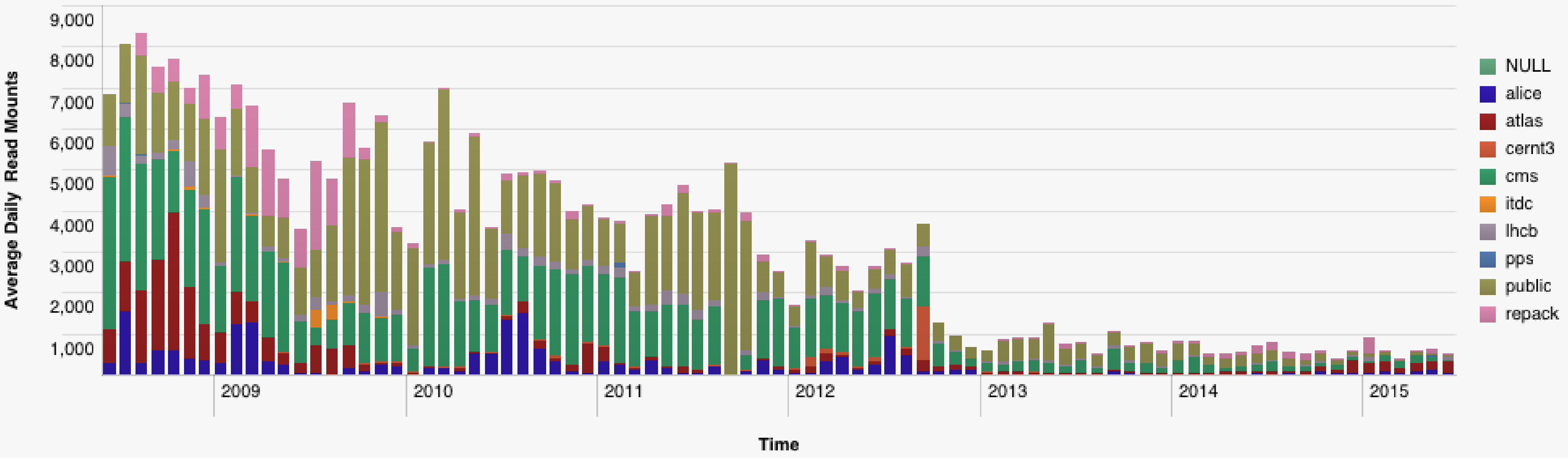
Tape Market

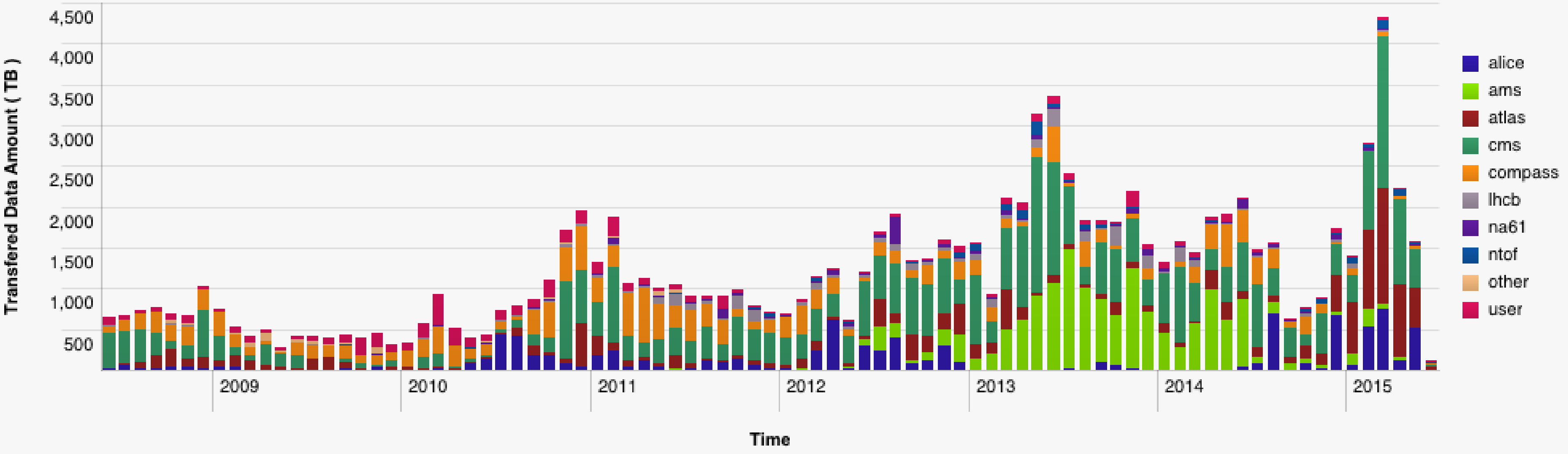
Tape cartridge revenue in M USD

(source: Santa Clara Consulting Group)



CERN Tape Read Mounts vs Volume





Bit preservation best practices(1)

- **R1: Define the list of offered storage classes and their characteristics with regard to supported access protocols, throughput, and access latency.**
  - *Related OAIS function: Receive Data (Archival Storage)*
  - The chosen technology for implementing the archive should be transparent to the customer (disk, tape, or a combination of both; number and physical location of replicas, etc.). The supported access protocols for read/write access on the archive and metadata lookup (such as namespace lookups) should be explicitly listed. Average and peak throughput and access latency rates should be defined for both reading and writing.
- **R2: Define the lifetime of the archived data, including the data retention period and/or expiration policy.**
  - *Related OAIS function: Manage Storage Hierarchy (Archival Storage)*
  - For each archive storage class or category, include a specific definition of how long the data is to be stored, and under what circumstances archive data is deleted (for example: experiment to be discontinued; user leaving the organization; etc). Budget constraints and planning should be taken into account for establishing these definitions.

Bit preservation best practices(2)

- **R3: Define expected data loss rates and associated metrics.**
  - *Related OAIS function: Error Checking (Archival Storage)*
  - For the sake of uniformity between HEP sites, we recommend defining two data loss metrics for each archive storage class/category: a) the numbers of bytes lost divided by the number of bytes written per year on one hand, and b) the number of bytes lost by year divided by the total number of bytes stored. These metrics should reflect data loss as from the customer perspective (not including failures which can be recovered from, such as failures on redundant media / storage). For each archive storage class / category, historical evidence/statistics should be collected and made available to customers.
- **R4: Provide mechanisms for file-level integrity checking via checksums.**
  - *Related OAIS function: Error Checking (Archival Storage)*
  - All files stored in the archive should have an associated checksum such as MD5, SHA-256, Adler32. This checksum should be calculated, stored and verified by the archive system; customers may also provide a pre-set checksum value against which the calculated checksum should be compared.

Bit preservation best practices(3)

- **R5: Provide mechanisms for regularly verifying the validity of the stored files within the archive.**
  - *Related OAIS function: Error Checking (Archival Storage)*
  - Define the frequency and policy associated with the archive verification such as scope (random sampling, complete archive scans, checking of tape cartridge beginning/end, verification after filling a cartridge, non-recently accessed files/tapes etc.). Verification results should be used for determining the archive reliability as in R3.
- **R6: Provide a workflow for contacting file owners in case of corruptions leading to data loss.**
  - *Related OAIS function: Error Checking (Archival Storage)*
  - Keep up-to-date contact details for customers owning archive files. Provide a workflow to inform users in case of temporary data unavailability (ie. caused by broken tapes sent for repair) or persistent data loss. Allow users to delete or re-populate corrupted/lost files.
- **R7: Provide a migration policy for media refreshments and/or storage technology upgrades.**
  - *Related OAIS function: Replace Media (Archival Storage)*
  - Upgrades to newer-generation storage technology (such as migration to new disk arrays, tape repacking to higher media densities or cartridges) should be in principle transparent to the customer. In particular, data access and archive contents should not be impacted by such migrations.

NDSA levels of digital preservation

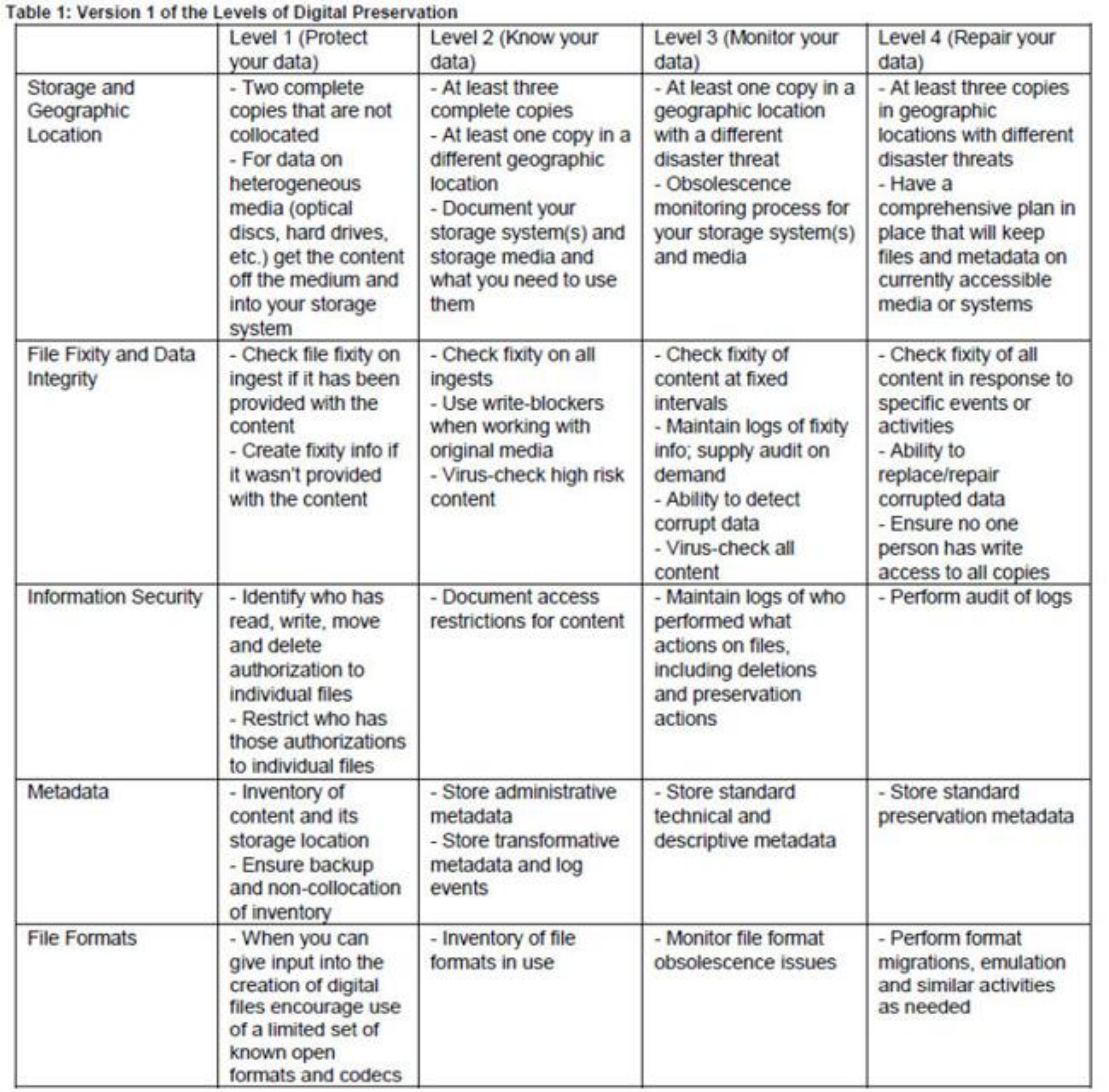


Table 1: Version 1 of the Levels of Digital Preservation

	Level 1 (Protect your data)	Level 2 (Know your data)	Level 3 (Monitor your data)	Level 4 (Repair your data)
Storage and Geographic Location	- Two complete copies that are not collocated - For data on heterogeneous media (optical discs, hard drives, etc.) get the content off the medium and into your storage system	- At least three complete copies - At least one copy in a different geographic location - Document your storage system(s) and storage media and what you need to use them	- At least one copy in a geographic location with a different disaster threat - Obsolescence monitoring process for your storage system(s) and media	- At least three copies in geographic locations with different disaster threats - Have a comprehensive plan in place that will keep files and metadata on currently accessible media or systems
File Fixity and Data Integrity	- Check file fixity on ingest if it has been provided with the content - Create fixity info if it wasn't provided with the content	- Check fixity on all ingests - Use write-blockers when working with original media - Virus-check high risk content	- Check fixity of content at fixed intervals - Maintain logs of fixity info; supply audit on demand - Ability to detect corrupt data - Virus-check all content	- Check fixity of all content in response to specific events or activities - Ability to replace/repair corrupted data - Ensure no one person has write access to all copies
Information Security	- Identify who has read, write, move and delete authorization to individual files - Restrict who has those authorizations to individual files	- Document access restrictions for content	- Maintain logs of who performed what actions on files, including deletions and preservation actions	- Perform audit of logs
Metadata	- Inventory of content and its storage location - Ensure backup and non-collocation of inventory	- Store administrative metadata - Store transformative metadata and log events	- Store standard technical and descriptive metadata	- Store standard preservation metadata
File Formats	- When you can give input into the creation of digital files encourage use of a limited set of known open formats and codecs	- Inventory of file formats in use	- Monitor file format obsolescence issues	- Perform format migrations, emulation and similar activities as needed