(Physics) Archival Storage Status and Experiences at CERN

Joint DASPOS / DPHEP7 Workshop

22 March 2013

Germán Cancio

Tapes, Archives and Backup

Data Storage Services Group – IT-CERN

(*) with input from J. Iven / L. Mascetti / A. Peters for disk ops

Agenda

- Overview of physics storage solutions
  - CASTOR and EOS
  - Reliability
- Data preservation on the CASTOR (Tape) Archive
  - Archive verification
  - Tape mount rates, media wear and longevity
  - Multiple tape copies
  - Other risks
- Outlook
  - Tape market evolution
  - Media migration (repacking)
  - R&D for archiving
- Conclusions

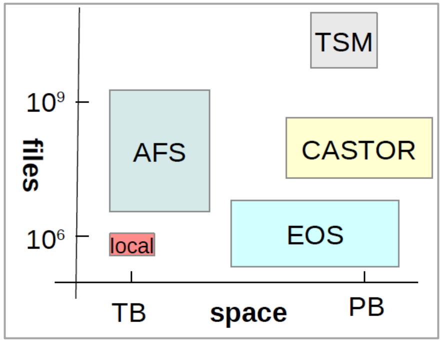
Physics Storage Solutions

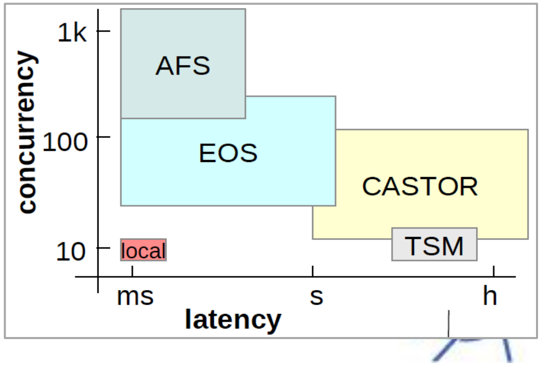
Two complementary services:

- CASTOR
  - Physics data storage for LHC and non-LHC experiments – active or not
    - COMPASS, NA48, NA61/2, AMS, NTOF, ISOLDE, LEP
  - HSM system with disk cache and tape backend
  - Long-lived and custodial storage of (massive amounts of) files
  - In prod since 2001, many incarnations, data imported from previous solutions (ie. SHIFT)

- EOS
  - Low-latency, high-concurrency disk pool system deployed in 2011
  - Physics analysis for O(1000) (end-)users
  - Tunable reliability on cheap HW – multiple copies on disk (no tape) – no “unique” data
  - Quota system – no “endless” space
  - “Disk only” pools moving from CASTOR to EOS

- Other storage solutions
  - AFS/DFS, Backup/TSM
  - R&D: Hadoop, S3,..

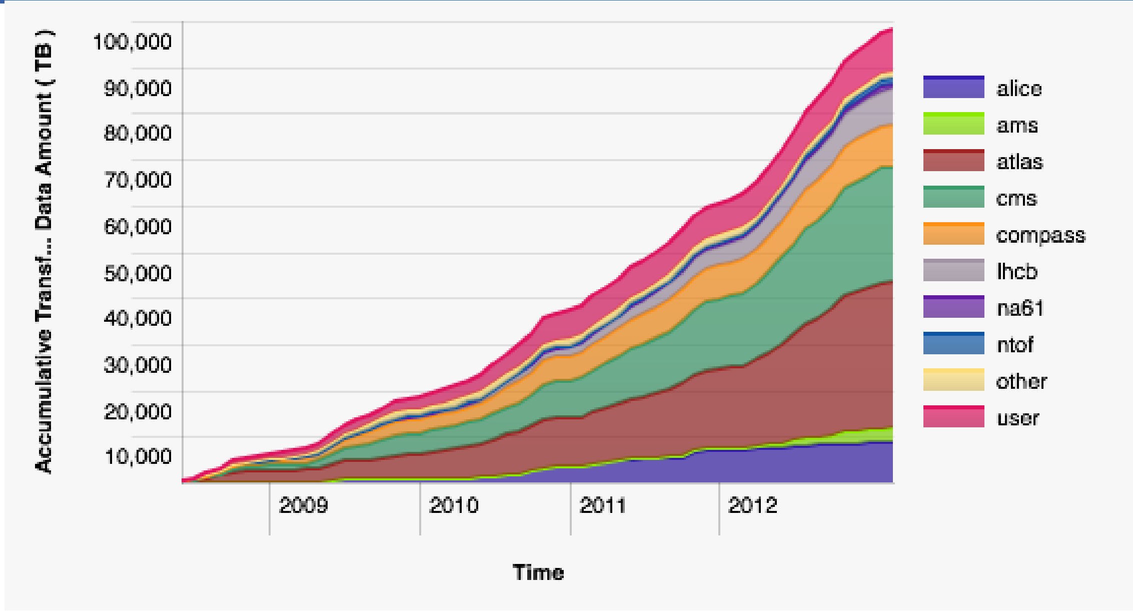




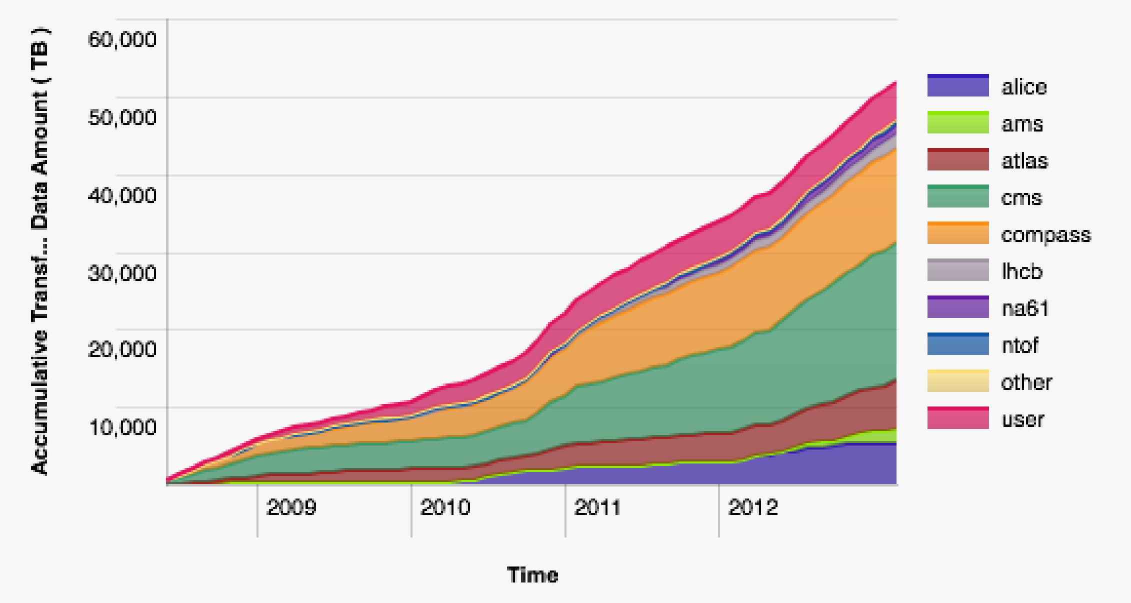
CASTOR archive in Numbers

Data:
88PB (74PiB) of data on tape; 245M files over 48K tapes
Average file size ~360MB
1.5 .. 4.6 PB new data per month
Up to 6.9GB/s to tape during HI period

Lifetime of data: infinite



Infrastructure:
~ 52K tapes (1TB, 4TB, 5TB)
7 libraries (IBM and Oracle) – 65K slots
90 production + 20 legacy enterprise drives
15PB disk cache (staging + user access) on ~750 disk servers

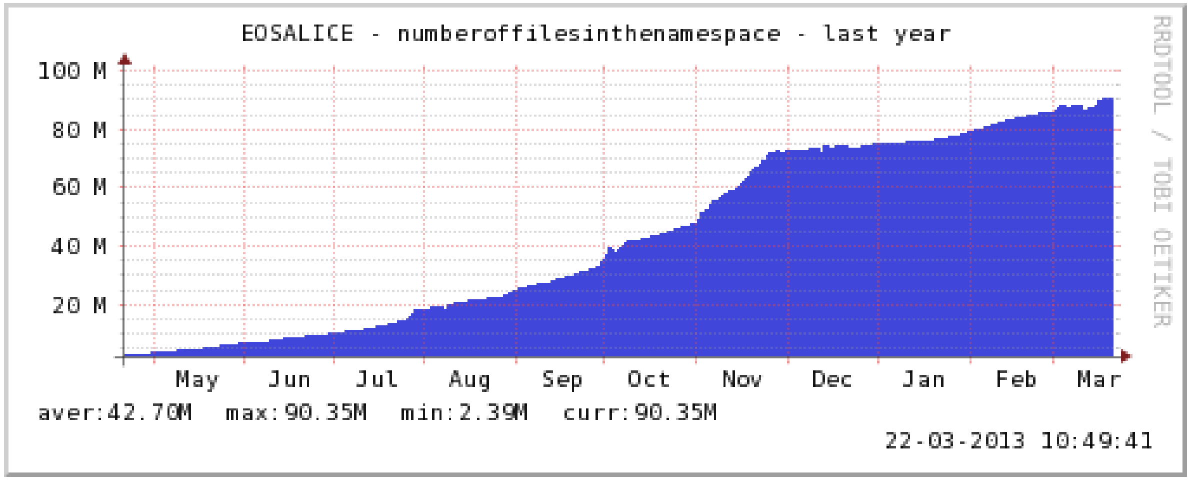


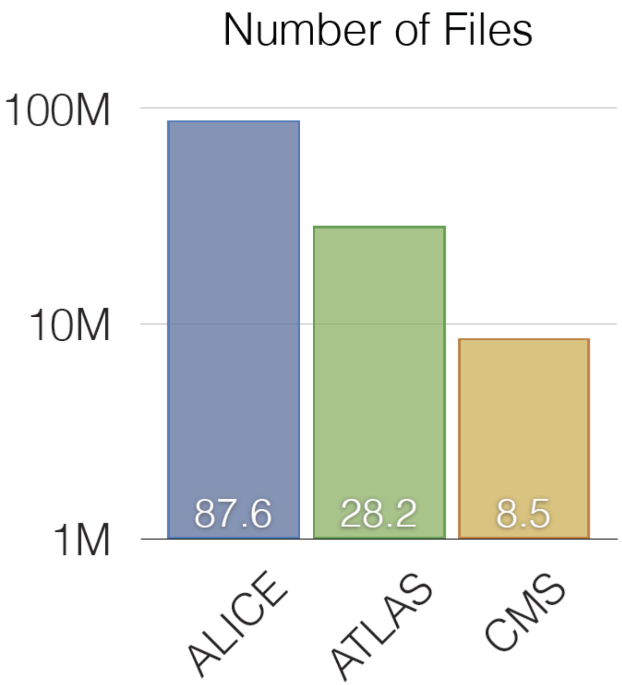
CASTOR
CERN Advanced STORage manager

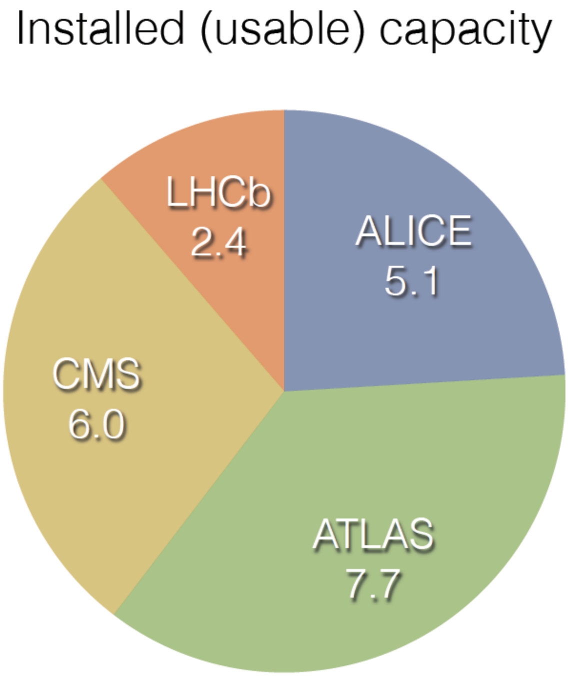
EOS in Numbers

Data:
~15 PB of data stored
~ 125M files
Average file size ~120MB
~8K-25K concurrent clients

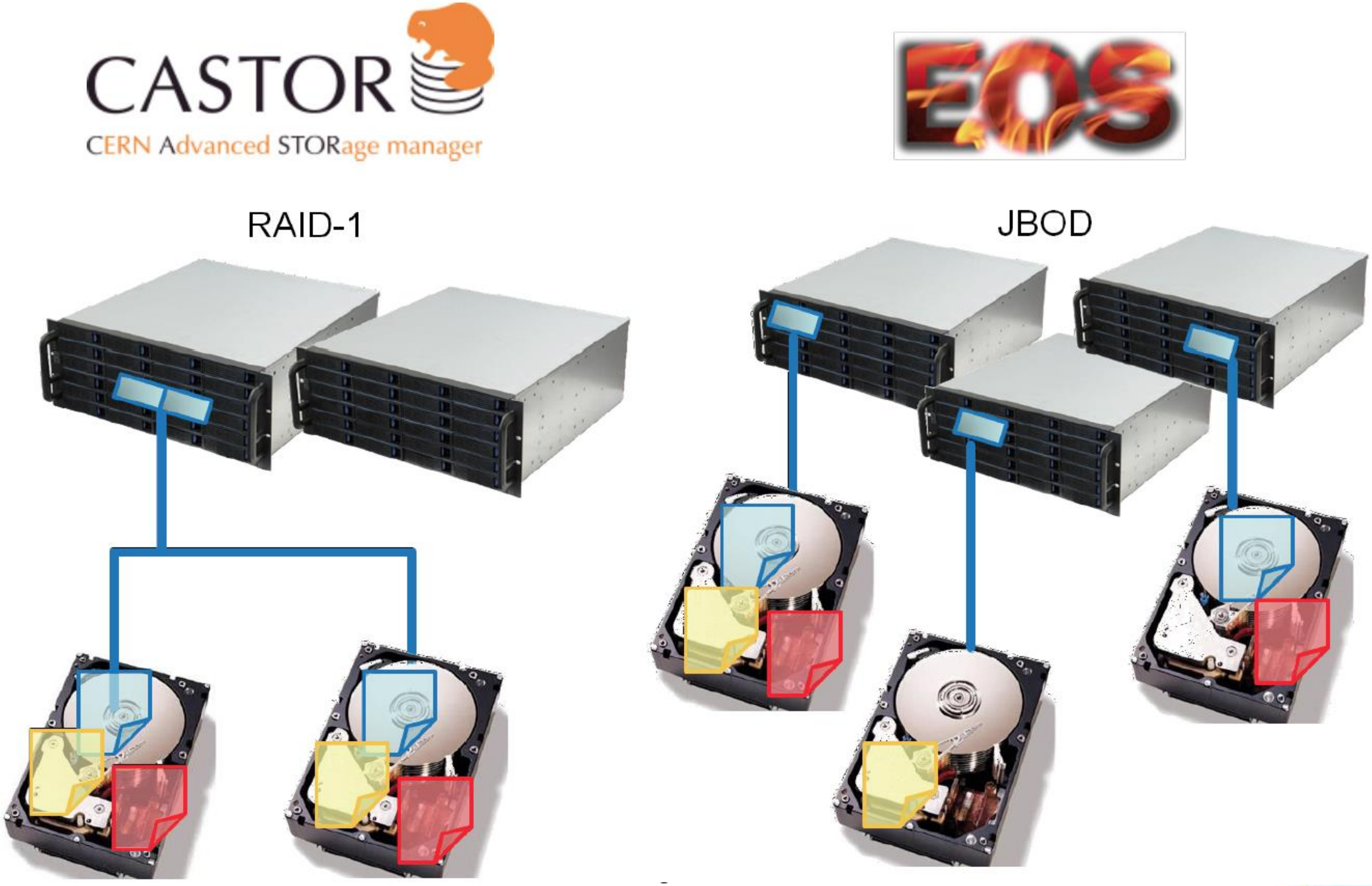
Infrastructure:
~ 850 disk servers
Installed raw disk capacity:
~40PB (usable: ~20PB)







Disk server setup differences

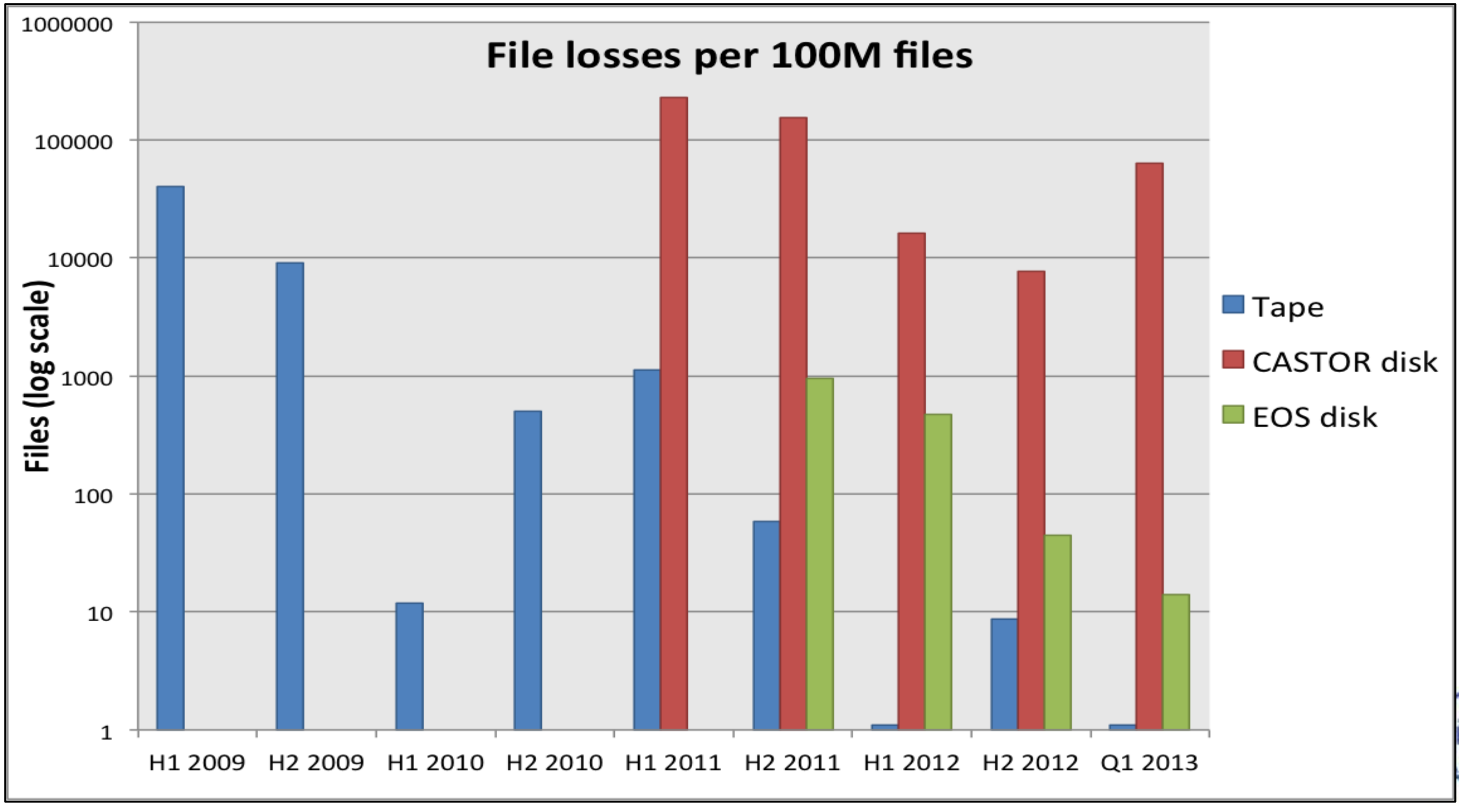


RAID-1

JBOD

Reliability

- File loss is unavoidable and needs to be factored in at all stages
- Good news: it has been getting better for both disk and tape
- Disk storage reliability greatly increased by EOS over CASTOR disk
  - RAID-1 does not protect against controller or machine problems, file system corruptions and finger trouble
- Tape reliability still ~O(1) higher than EOS disk
  - Note: single tape copy vs. 2 copies on disk



Agenda

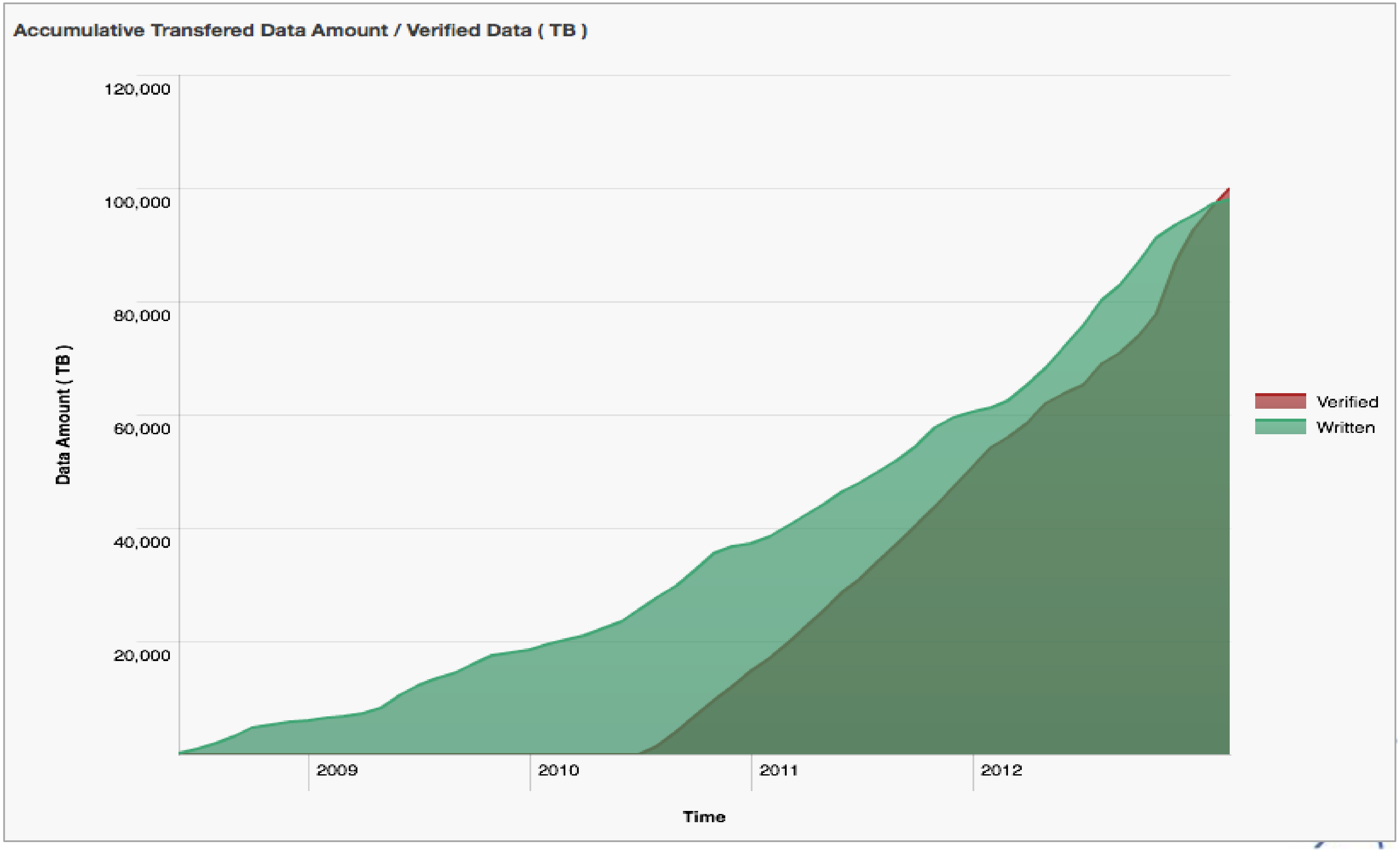
- Overview of physics storage solutions
  - CASTOR and EOS
  - Reliability
- Data preservation on the CASTOR (Tape) Archive
  - Archive verification
  - Tape mount rates, media wear and longevity
  - Multiple tape copies
  - Other risks
- Outlook
  - Tape market evolution
  - Media migration (repacking)
  - R&D for archiving
- Conclusions

Tape archive verification

- Data in the archive cannot just be written and forgotten about.
  - Q: can you retrieve my file?
  - A: let me check… err, sorry, we lost it.
- Proactive and regular verification of archive data required
  - Ensure cartridges can be mounted
  - Check data can be read+verified against metadata (checksum/size, …)
  - Do not wait until media migration to detect problems
- Several commercial solutions available on the market
  - Difficult integration with our application
  - Not always check *your* metadata
- In 2010, implemented and deployed a background scanning engine:
  - Read back all newly filled tapes
  - Scan the whole archive over time,
starting with least recent accessed tapes

Verification: first round completed!

- Up to 10-12 drives (~10%) for verification @ 90% efficiency
- Turnaround time: ~2.6 years @ ~1.26GB/s
- Data loss: ~ 65GB lost over 69 tapes



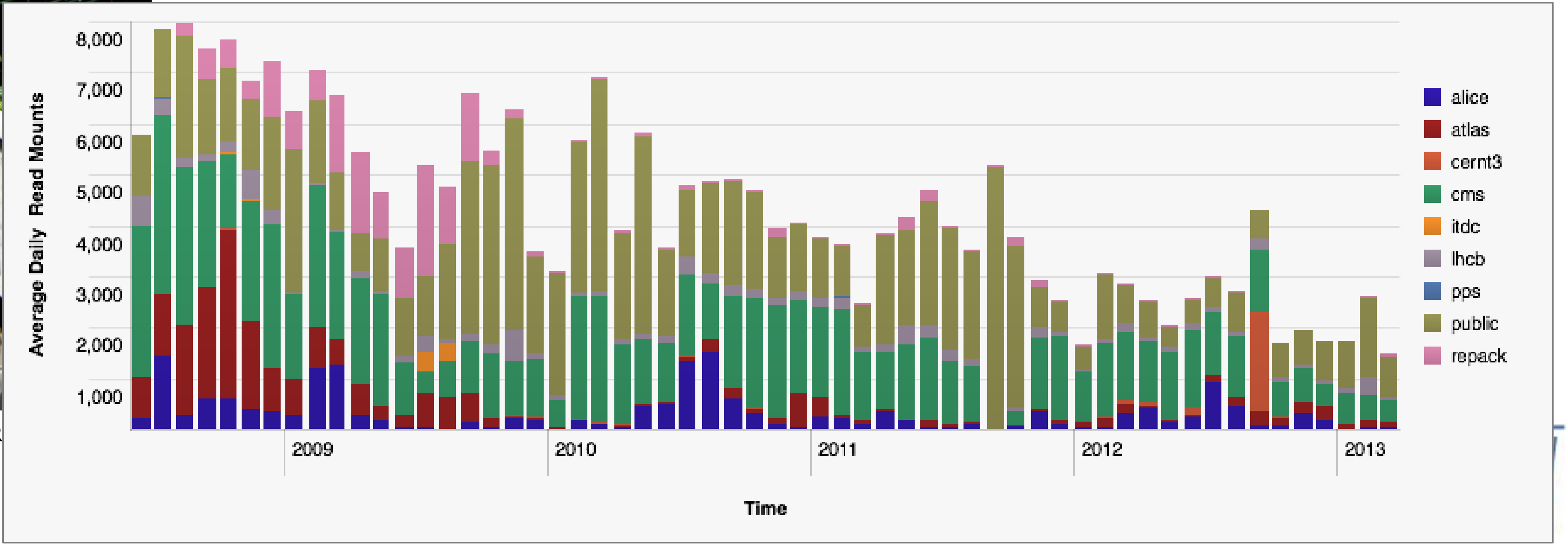
Increasing media / robotics longevity

- 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

DSS

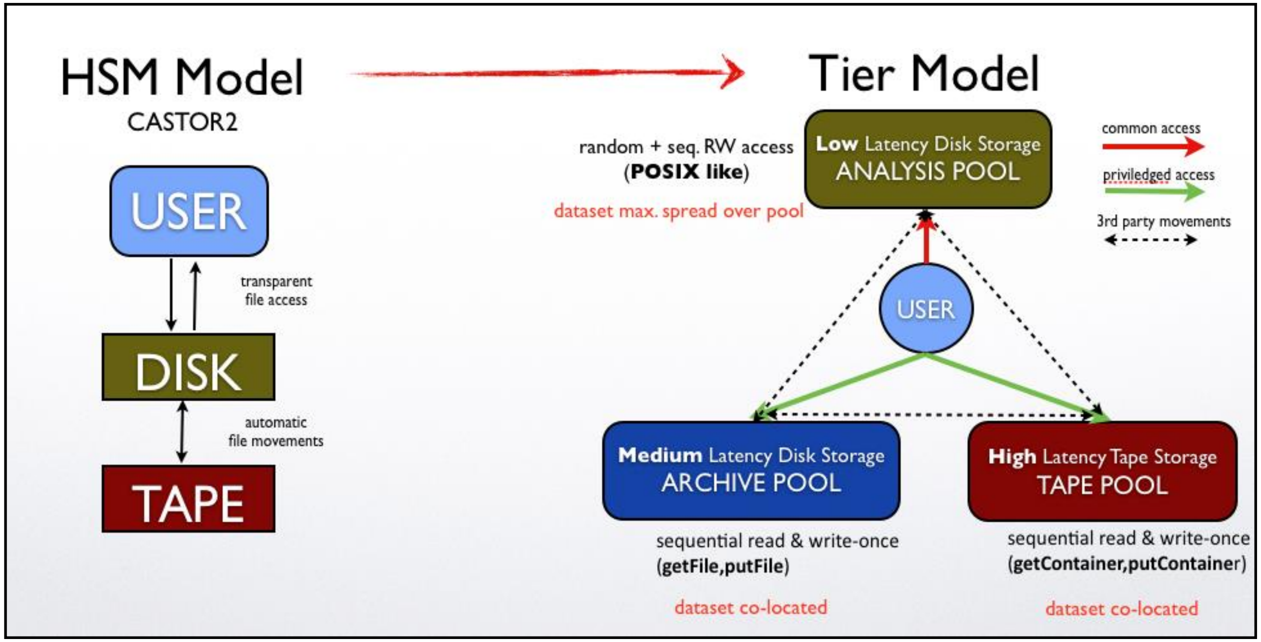
Tape mount rate reduction

- 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



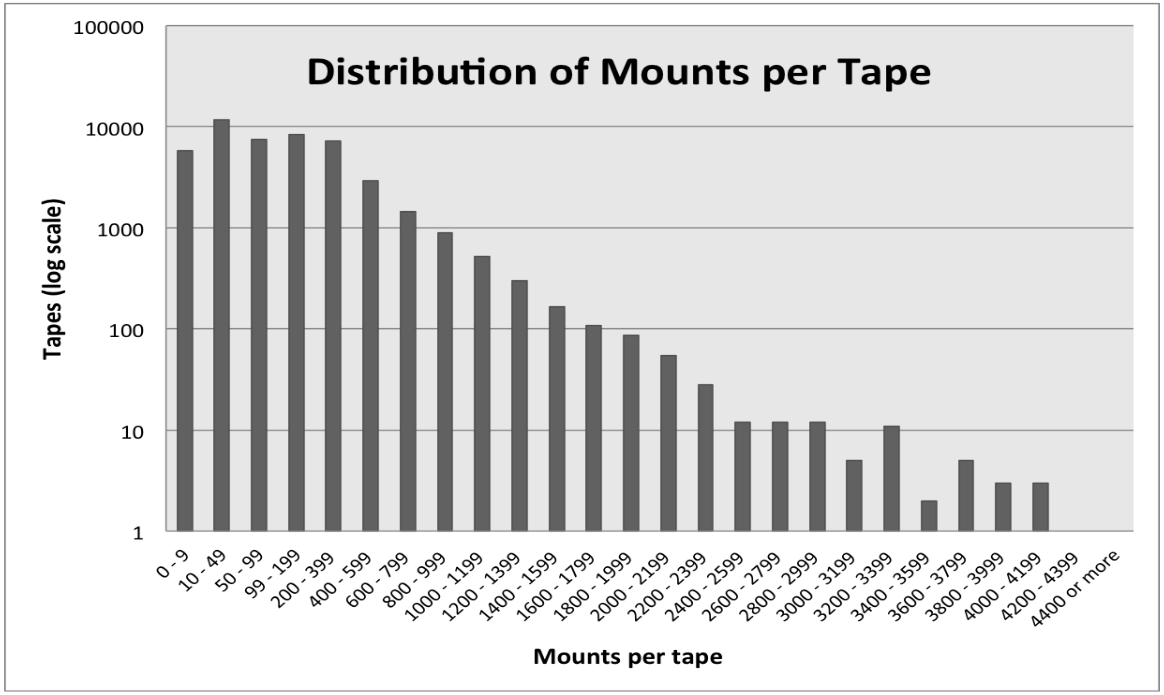
HSM model limitations

- 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



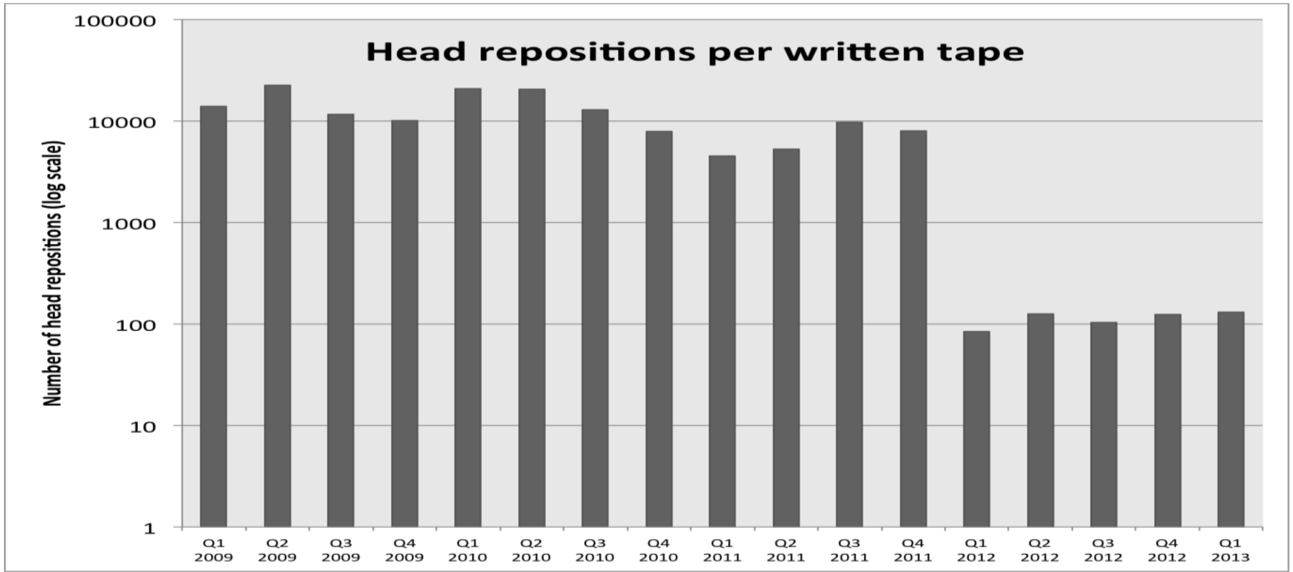
Addressing media wear

- 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



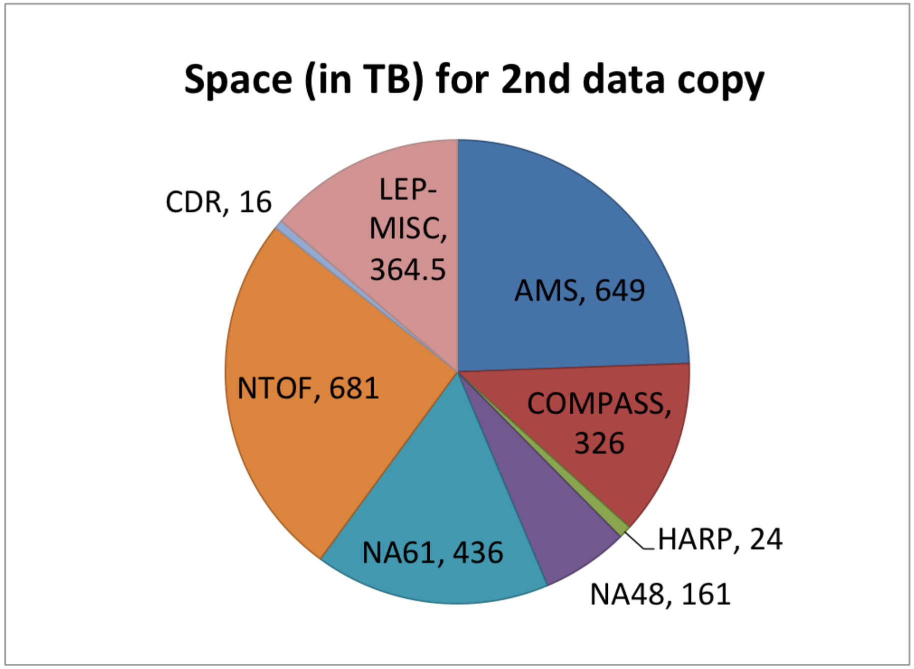
Avoiding “shoe-shining”

- 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



Multiple file copies

- By default, only one copy of a file is stored on tape.
- If justified, second copies can be generated on different tapes (or even different libraries)
- Typically the case for experiments where data is stored only at CERN and/or legacy experiments
- Around 2.6PB of additional space (3% of total space)



Other risks…

Many other risks for data integrity to be aware of:

- Security break-ins
  - Strong authentication deployed on CASTOR… eventually
- Finger trouble
  - `nsrm -rf /castor/cern.ch/opal/rawd/ test/blahblah`
  - If noticed “quickly”, metadata can be restored (manual work)
- Bugs, misconfigurations, devops misalignment
  - ALICE incident 2010: routing production files to test tape pools being recycled
  - Meta(data) was restored, but some tapes had been recycled -> data loss
  - Test tape pool recycling decommissioned since
  - Stopped automated media repacking (defragmentation)
- Disasters affecting CC equipment integrity
  - Planes crashing in (none so far…)
  - Water leaks (had one exactly over a STK silo in 2004)
- etc…

Agenda

- Overview of physics storage solutions
  - CASTOR and EOS
  - Reliability
- Data preservation on the CASTOR (Tape) Archive
  - Archive verification
  - Tape mount rates, media wear and longevity
  - Multiple tape copies
  - Other risks
- Outlook
  - Tape market evolution
  - Media migration (repacking)
  - R&D for archiving
- Conclusions

Outlook: Tape market evolution

- Tape technology getting a push forward
  - Drive generations last released

Vendor	Name	Capacity	Speed	Type	Date
LTO consortium(*)	LTO-6	2.5TB	160MB/s	Commodity	12/2012
Oracle	T10000C	5.5TB	240MB/s	Enterprise	03/2011
IBM	TS1140	4TB	240MB/s	Enterprise	06/2011

  - Vendor roadmaps exist for additional 2-3 generations, up to 20TB / tape (~2016-17) (+70% capacity / year) – new generations expected 2013/14
  - 35/50TB tape demonstrations in 2010 (IBM/Fuji/Maxell); 125-200TB tapes being investigated by IBM
- Tape market evolving from NEARLINE to ARCHIVING
  - Increased per-tape capacity and transfer speed
  - Little or no increases for mounting/positioning – unsuitable for random access
  - Small-to-medium backup market shrinking (de-duplication, disk-only)
  - Large-scale archive/backup market building up (legal, media, cloud providers - Google: ~6-10EB?)

(*) LTO consortium: HP/IBM/Quantum/Tandberg (drives); Fuji/Imation/Maxell/Sony (media)

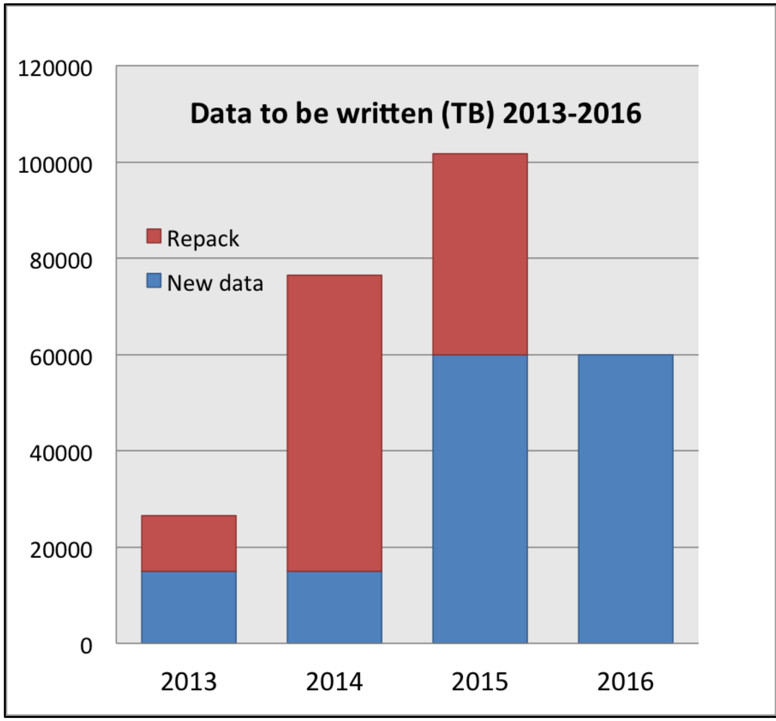
Outlook: Media repacking

- Mass media migration or “repacking” required for
  - Higher-density media generations, and / or
  - Higher-density tape drives (enterprise media rewriting)
  - Liberating tape library slots
- Media itself can last for 30 years, but not the infrastructure!
- Repack exercise is **proportional** to the **total size of archive** - and **not** to the fresh or active data

- Next Repack run (expected): 2013/4 - 2016
  - New drive generations appearing “soon”
  - ~100PB to migrate from over 50’000 cartridges
- Data rates for next repack will exceed LHC data rates…
  - Over 3 GB/s sustained
  - Cf . LHC proton-proton tape data rates : ~1-1.5GB/s

- …. but we need to share the drives –
  which become the bottleneck

- Will compete with up to 60PB/year data taking after LS1



- Infrastructure, software and operations must sustain writing up to 0.1EB in 2015 (+ reading!)

Outlook: Research and Development

- Older tape data getting “colder” (excluding repacking/verification)
  - Only ˜14PB read from tape in 2012; 20K tapes not mounted at all in 12 months (25PB)
  - Excluding data written in 2012 still leaves ~40PB of data not being read
  - Trend likely to continue as “freshest” data being most relevant
  - Not all data requires to be online and/or directly visible
- Fits into the from-HSM-to-Tier model strategy
- Market solutions appearing for cold data archiving
  - Notably Amazon Glacier
  - Service price not competitive for the time being (0.01$/GB/month storage, 0.1$/GB retrieval)
  - .. but this may change in the future
- Appealing approach and API
  - “stripped down S3” WS-based RESTful interface
  - Immutable files, minimal metadata and operations, synchronous upload but asynchronous (high latency) data retrieval
- Investigate potential as simple tape front-end interface
  - Archiving of physics and non-physics data
- Many questions to be addressed (client access, namespace handling, efficient transfer, load balancing, data import and migration, verification etc)

Agenda

- Overview of physics storage solutions
  - CASTOR and EOS
  - Reliability
- Data preservation on the CASTOR (Tape) Archive
  - Archive verification
  - Tape mount rates, media wear and longevity
  - Multiple tape copies
  - Other risks
- Outlook
  - Tape market evolution
  - Media migration (repacking)
  - R&D for archiving
- Conclusions

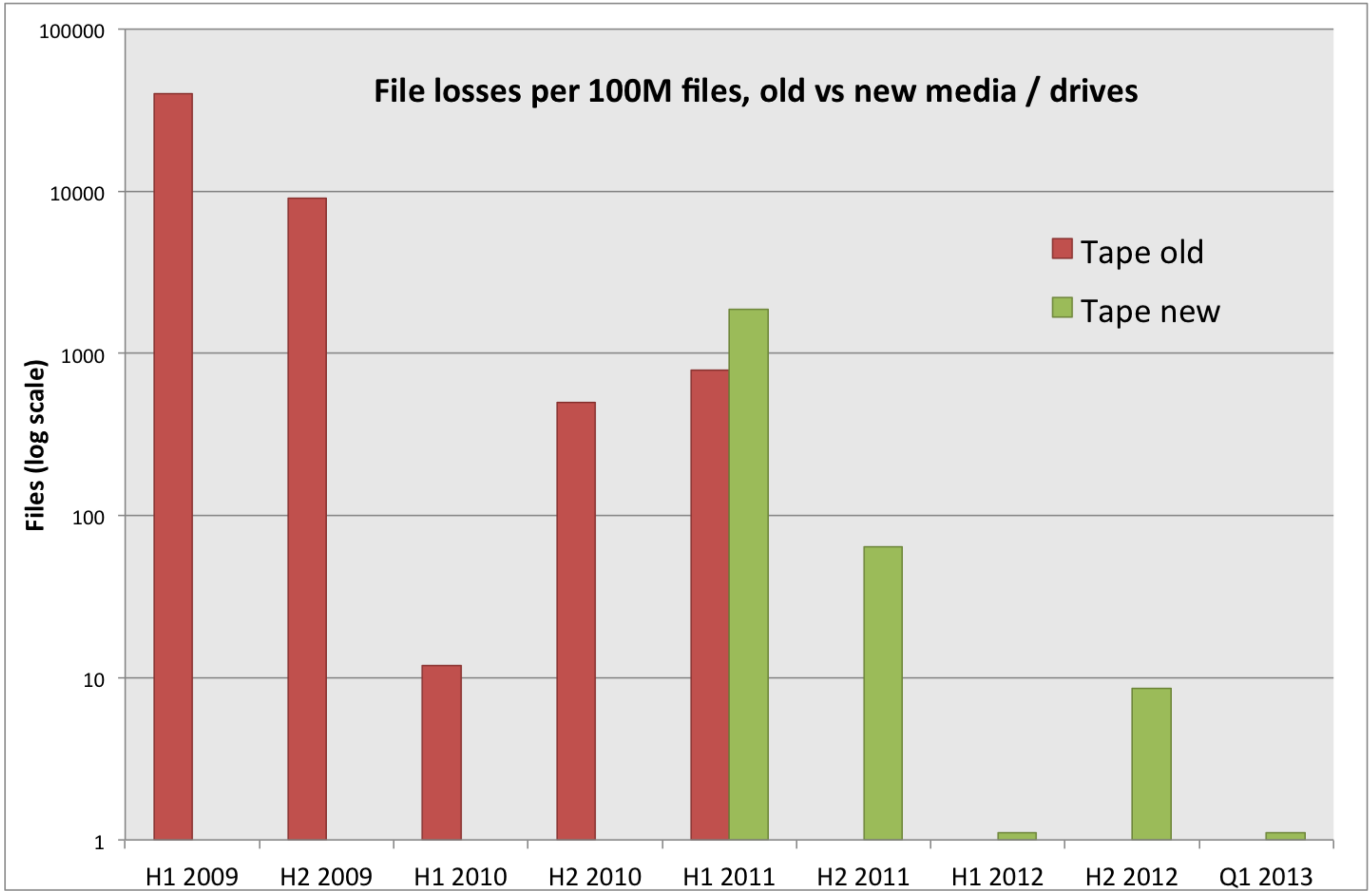
Conclusions

- Managing a large, PB-scale, tape-backed archive is an active task. The effort is proportional to the total archive size.
- A non-negligible fraction of resources need to be allocated for housekeeping such as migration and verification.
- Tape has a not-so-large *effective* lifetime requiring regular media migration to new generations.
- Reliability and performance requires to separate end-user access from archiving. Continue moving to what tape is really built for: bulk archiving and streaming access.

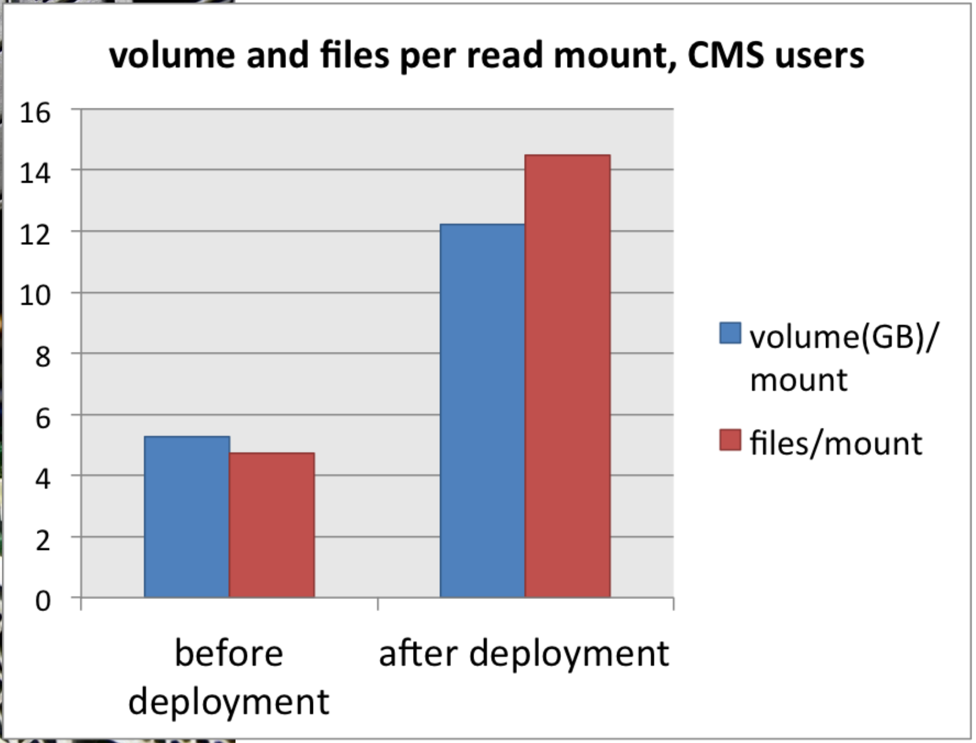
Discussion

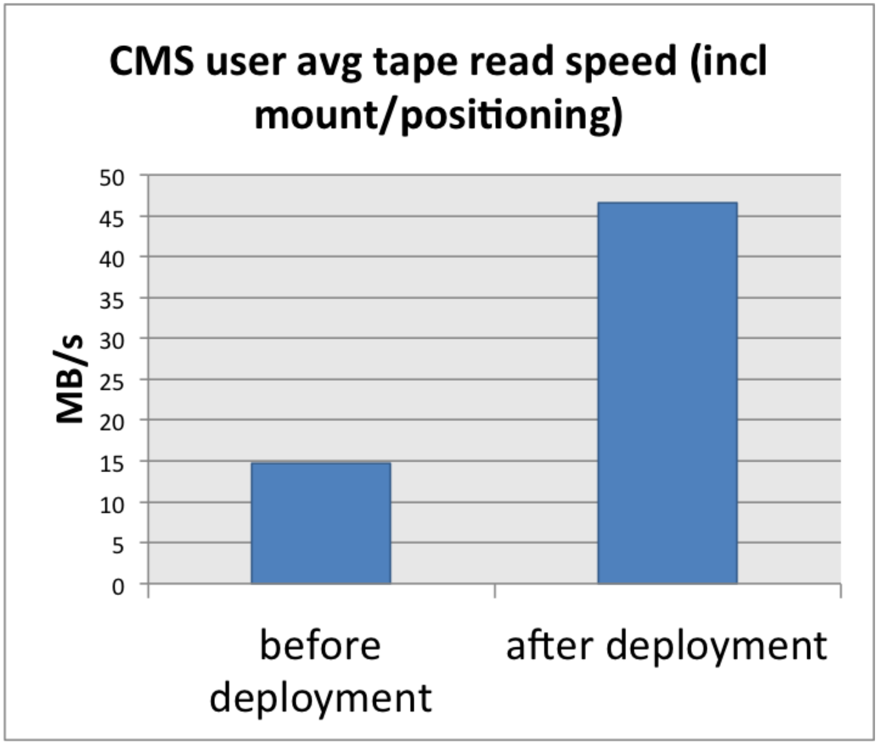
Reserve slides

New drive/media generations



Tape Reading optimization results





3x files / volume per mount -> 3x increase in effective tape access speed
~50% less tape mounts (~7K to 3.5K mounts per day)

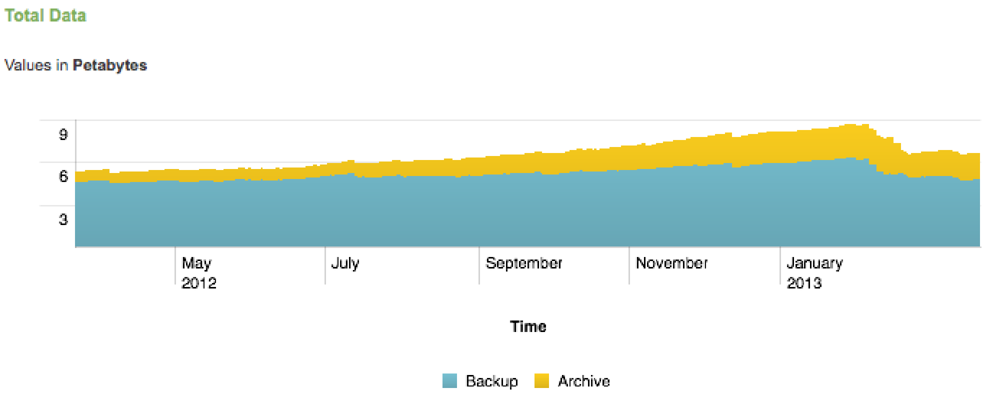
Market: Enterprise vs. LTO

- Tape usage at CERN was heavy-duty requiring enterprise-class tape equipment from IBM and Oracle
  - With far less demand in terms of “small” file writes and read mounts, “commodity” tape (LTO) becomes a serious option, i.e. for “dusty” archived data which is infrequently accessed
- Market share: LTO (~90%) vs. enterprise media (~2%)
- Completed field testing of a LTO SpectraLogic T-Finity library (max 120 drives, 30K slots)
  - Test drives, library, and vendor support – storing 2nd copies of experiment data
  - Test configuration: 5-10 LTO-5 drives, 1000 cartridges (1.5PB)
  - Necessary CASTOR adaptations coded and released
- Satisfactory results in general

TSM key numbers

Data:

- ~ 6.6 PB of data
  - 4.7 PB backup
  - 1.9 PB archives
  - 8K tapes
- Daily traffic: ~75TB
- 2.2B files (112M archive)
- 1400 client nodes
  - Servers, VM’s
  - AFS/DFS
  - Exchange
  - DB



Infrastructure:

- 13 new-generation TSM servers (RHES6, TSM6)
  - 2 server + 2 SAS expanders setup
- 6 legacy TSM5 being decommissioned
  - SAN-disk setup
- 2 IBM TS3500 libraries
  - 24 TS1140 drives
  - 32 TS1130 drives