

# Update on CERN Tape Status

*HEPiX Spring 2014, Annecy  
German Cancio / CERN*



- Tape performance / efficiency
- Big Repack exercise
- Verification and reliability
- Hardware + software evolution
- Outlook

## Volume:

- ~100PB of data on tape
  - 94PB on CASTOR
  - 6PB on TSM
- Files: 274M (CASTOR) + 2.1B (TSM)

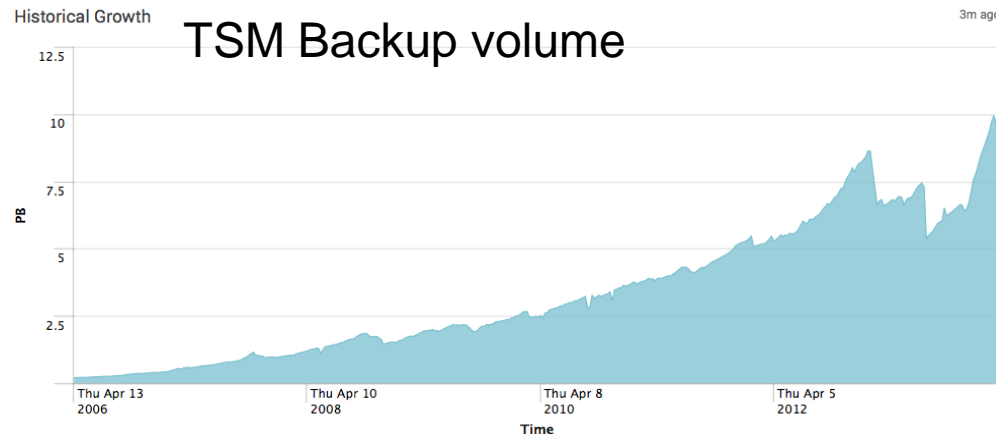
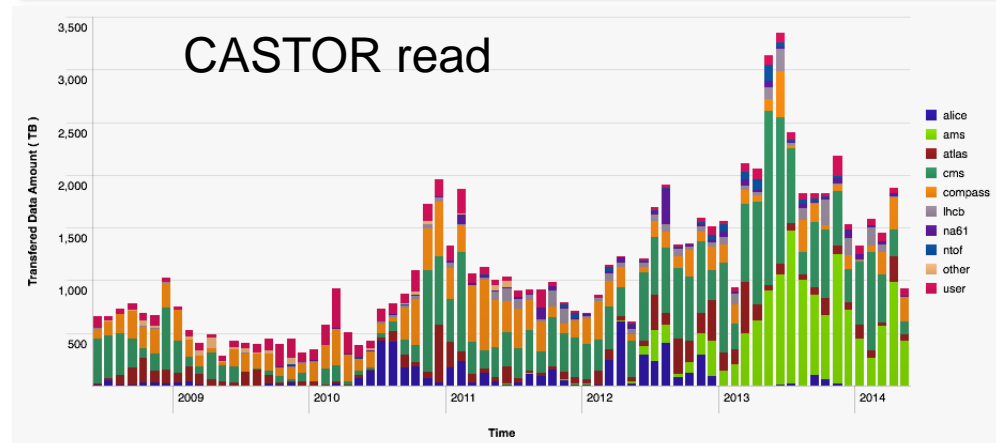
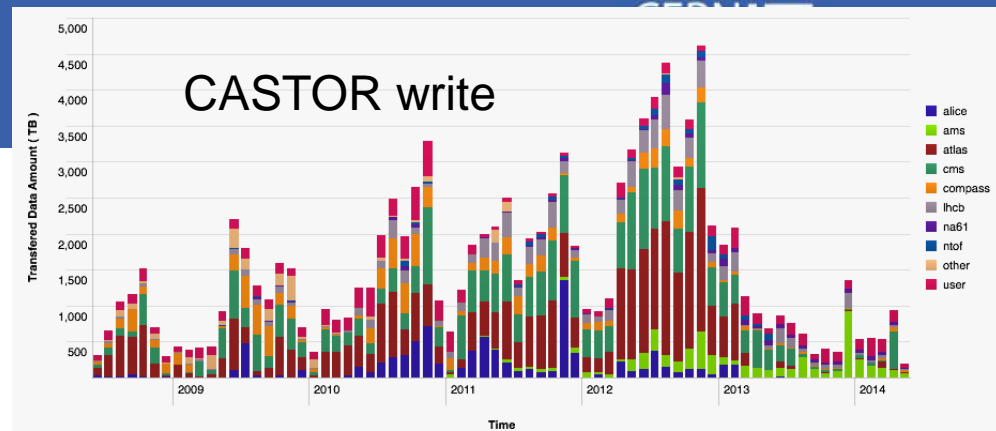
## Infrastructure:

- 60K tapes (1-8TB)
- 200 FC-attached tape drives
  - CASTOR: 80 production + 70 legacy
  - TSM: 50 drives
- 9 libraries (IBM TS3500 +Oracle SL8500)
  - 7 for CASTOR, 2 for TSM
- 150 Castor tape servers
- 12 TSM servers (~1100 clients)

## Manpower: 7 staff + fellow/students

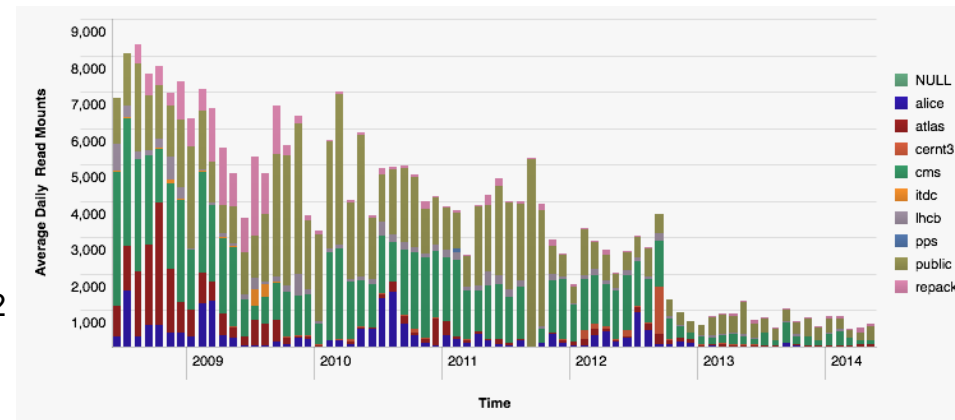
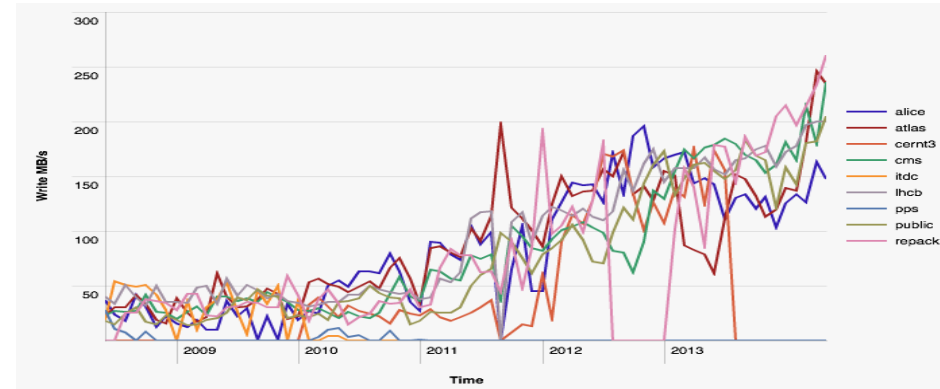
- ~3 FTE Tape Operations
- ~3 FTE Tape Developments
- ~2 FTE Backup Service

Shared operations and infrastructure (libraries, drives, media) for CASTOR and TSM Backup



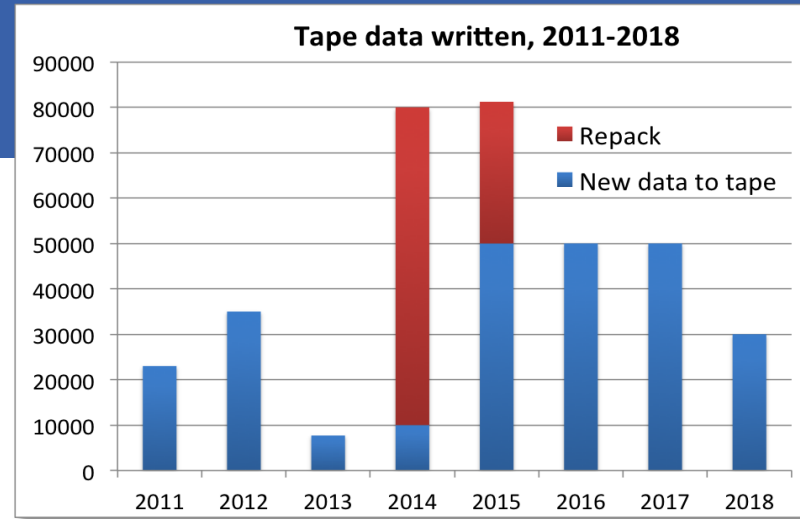
Increasing CASTOR tape efficiency has been a core activity over the last 2-3 years

- Writes: From ~30% to >80% of native drive speed, thanks to development of “buffered” tape marks and re-engineering of stager-tape middleware
  - Handling 4x nominal ALICE DAQ rates
- Reads: Reduction of tape mounts from >7K/day to 1-3K/day, despite increasing recall traffic
  - Introduction of recall policies (group recall requests until threshold), encourage pre-staging
  - (Ongoing) migration of end-users CASTOR->EOS
  - Avg files/mount from 1.3 to ~50; avg remounts/day: <2
  - From HSM to ARCHIVE
- And many other improvements
  - including optimization of head movements for sequential recalls, skip over failed recall files, drives in UNKNOWN state, ...
- Cost savings - reduction of production tape drives from ~120 to 80

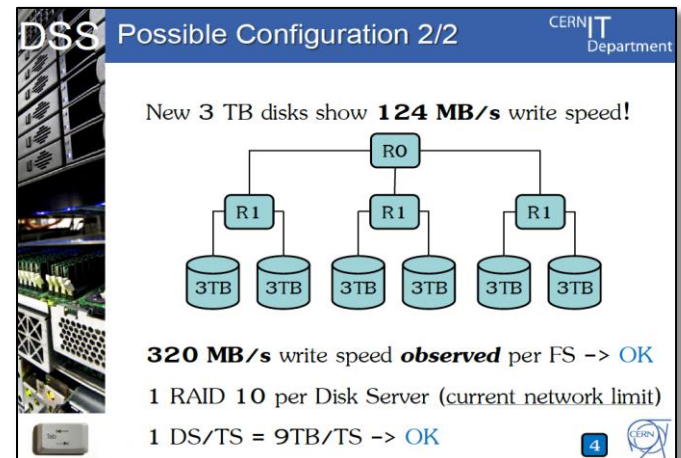


# Big Repack Exercise

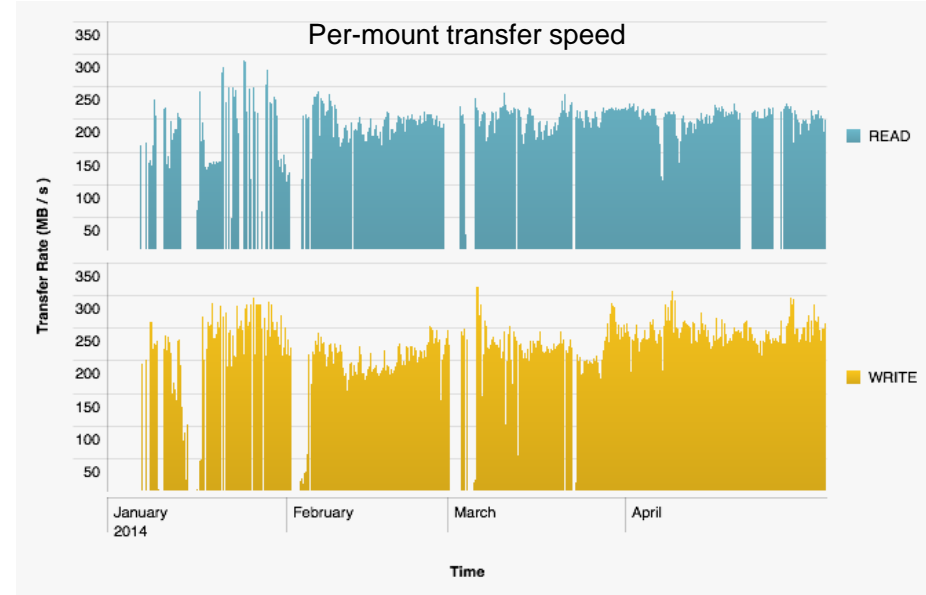
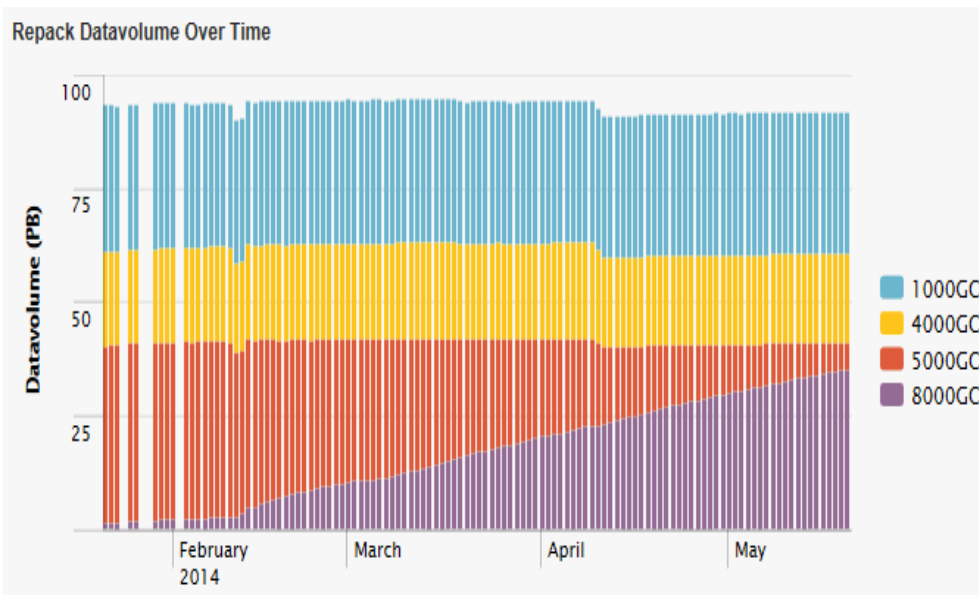
- End of January, started repack of 52K tapes
  - 92PB
  - 16K 4/5TB tapes to be repacked on higher density (7/8TB) [gain: ~35PB]
  - 36K 1TB tapes to be decommissioned



- Goal: repack as much as possible before end of LS1
  - Avoid competing with ~50PB/year data taking during Run2 (tape drives are the bottleneck!)
- Repack framework re-engineering in 2012/3
  - Repack application now a thin (and rock-solid) layer on top of the standard CASTOR stager
  - Workload engine (aka “feeder”) developed, with configurable policies, taking into account user/experiment activity and minimising interference
- Optimised repack disk for staging rather than caching
  - 40 disk servers (~0.9PB), RAID-10 based, reaching peaks of 300MB/s per server



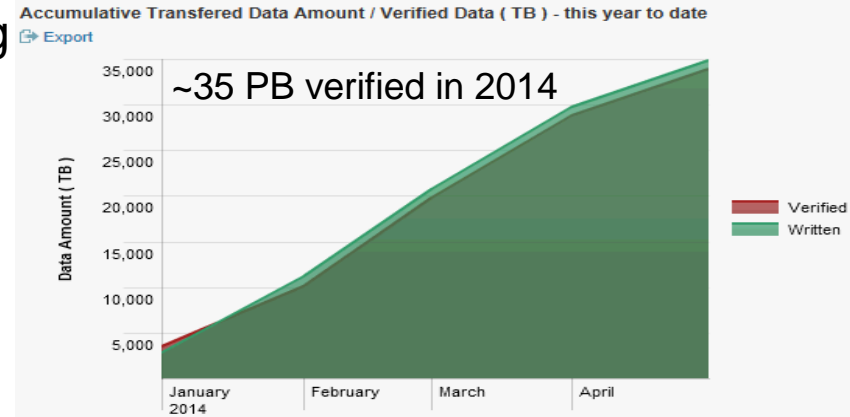
- Repacking ~2PB / week ==sustained ~3.4GB/s with 16 drives / avg (~206 MB/s per drive), write peaks up to 8.6GB/s
  - So far, no surprises found (all data was verified previously, and being re-verified after repack)



- With this performance sustained, repack could complete Q4 2014
  - ... but new drive generation unlikely to be available before Q4 2014 -> ~20PB to be done in Q1 2015
- Excellent validation case for CASTOR tape + stager software stack
  - CASTOR “bulk access” archiving use case more and more similar to repack
  - Run2 Pb-Pb data rates (~10GB/s): OK

- Systematic verification of archive data ongoing

- “Cold” archive: Users only accessed ~20% of the data (2013)
- All “historic” data verified between 2010-2013
- All new and repacked data being verified as well

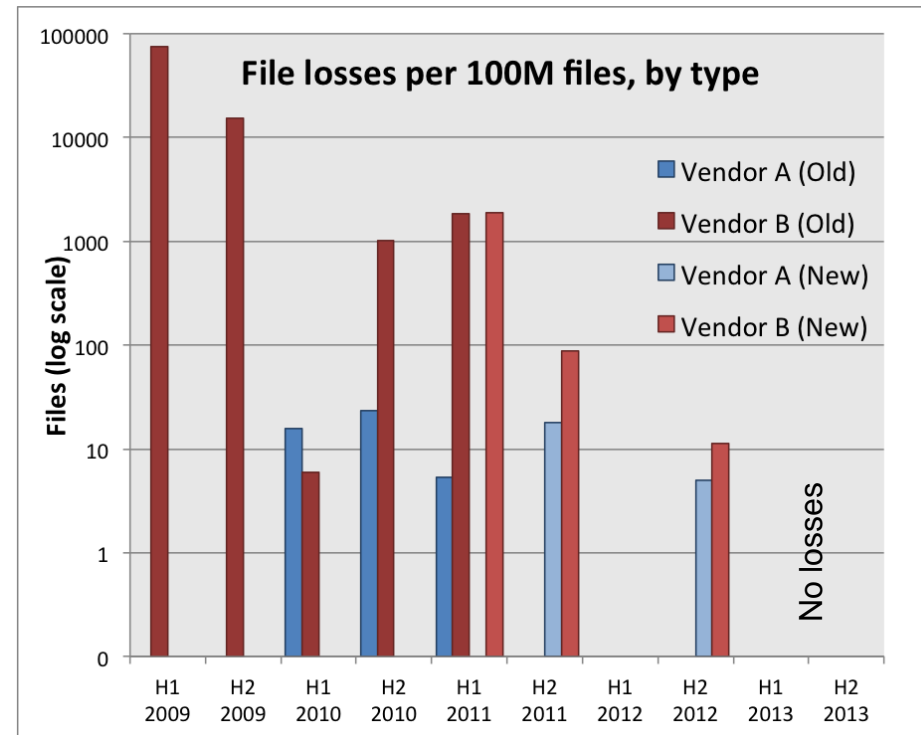


- Data reliability significantly improved over last 5 years

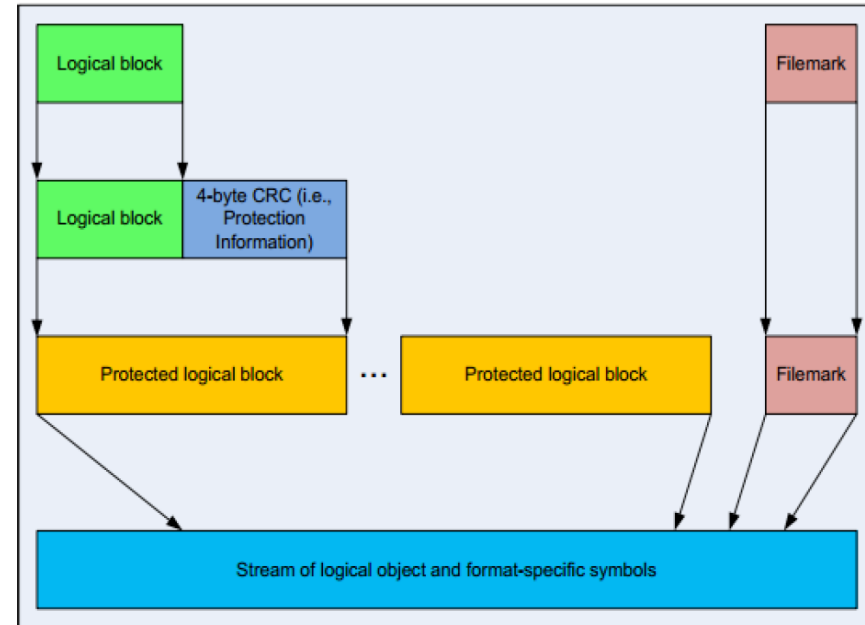
- From annual bit loss rates of  $O(10^{-12})$  (2009) to  $O(10^{-16})$  (2012)
- New drive generations + less strain (HSM mounts, TM “hitchback”) + verification
- Differences between vendors getting small

- Still, room for improvement

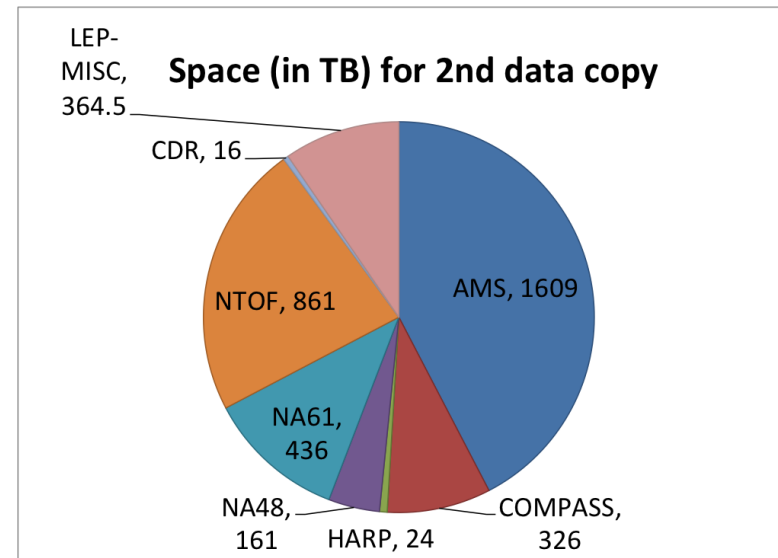
- Vendor quoted bit error rates:  $O(10^{-19..-20})$
- But, these only refer to media failures
- Errors (eg bit flips) appearing in complete chain



- Developing support for SCSI-4 Logical Block Protection
  - Data blocks shipped to tape drive with pre-calculated CRC
  - CRC checked by drive (read-after-write) and stored on media; CRC checked again on reading
  - Tape drive can do full media verification autonomously (and fast)
  - Supported by new-generation enterprise drives and LTO-5/6; marginal performance overhead



- Enabling dual-copy for non-LHC data (whenever justified)
  - Moving from “2 copies on 2 tapes” to different libraries
  - Around 3.8PB (4%) of additional space
  - Small experiments only (except AMS) – but everything gets “small”



# Hardware testing / validation

- Successful evaluation of SpectraLogic T-Finity library as 3<sup>rd</sup>-vendor option
- Oracle T10KD: 8.5TB, 250MB/s; 40 purchased and in prod

Extract of features relevant for CERN	T10000D	T10000C
Native cartridge capacity (uncompressed)	Up to <b>8.5 TB</b>	Up to 5(.5) TB
Bit Density	<b>4.93 Gb/in<sup>2</sup> / 449 kbp<i>i</i> / 4608 data tracks</b>	3.14 Gb/in <sup>2</sup> / 367 kbp <i>i</i> / 3584 data tracks
Head design	Dual heads writing 32 tracks simultaneously; GMR readers	Dual heads writing 32 tracks simultaneously; GMR readers
Native data rate performance (uncompressed)	Up to <b>252 MB/sec</b>	Up to 240 MB/sec
Fibre Channel interface	<b>16 Gb/s</b>	4 Gb/s
Tape speeds (m/s) (read/write, locate)	<b>5 speeds (2.75, 3.25, 3.75, 4.25, 4.75)</b> 10-13	2 speeds (3.74, 5.62) 10-13
Internal data buffer	2 GB	2 GB
Small files handling functionality	File Sync Accelerator ( <b>Small and Large files</b> ) <b>Tape Application Accelerator</b>	File Sync Accelerator
Max Capacity Feature	Yes (default: <b>ON</b> )	Yes (default: OFF)

+54%  
(same media)

+5%

not used  
in CASTOR

<http://www.oracle.com/us/products/servers-storage/storage/tape-storage/t10000d-ds-1991052.pdf>

Vladimír Bahyl : Oracle StorageTek T10000D tape drive test summary

3

- Only minor items seen during beta-testing (FSA performance, now fixed in MC)
- Issue with CASTOR SCSI timeouts settings discovered when already running Repack, also fixed
- Over 35PB written (+ re-read) without problems!

## Investigated alternatives to (parts of) CASTOR software stack

- Amazon Glacier: potential as simple tape front-end interface
  - “stripped down S3” WS-based interface; minimal metadata and operations
  - .. but in reality, coupled to S3 infrastructure; key functionality missing from API (redirection support, no staging concept, etc) ; modest interest from Amazon to share knowledge with CERN
- LTFS: abstraction layer (POSIX) on top of complex tape I/O
  - Shipped by IBM and Oracle; being adopted by film industry
  - High complexity and low maturity, incompatible with present ANSI format, diverging (and non-OSS) extensions for library management



## Strategy: re-engineer rather than replace CASTOR tape layer

- Replace CASTOR tape server codebase
  - Code aged (20+ years) , full of legacy OS/hardware, exotic tape formats and pre-CASTOR support
  - Replace 10+ daemons and executables by two: tape mounting and serving
  - Extensions such as Logical Block Protection and Ceph client support
- Review CASTOR drive queue / volume management services
  - Provide a single integrated service, take better into account reduced number of higher-capacity tapes
  - Avoid drive write starvation problems, better load-balancing, allow for pre-emptive scheduling (ie user vs verification jobs)

- New tape drives and media released or in pipeline

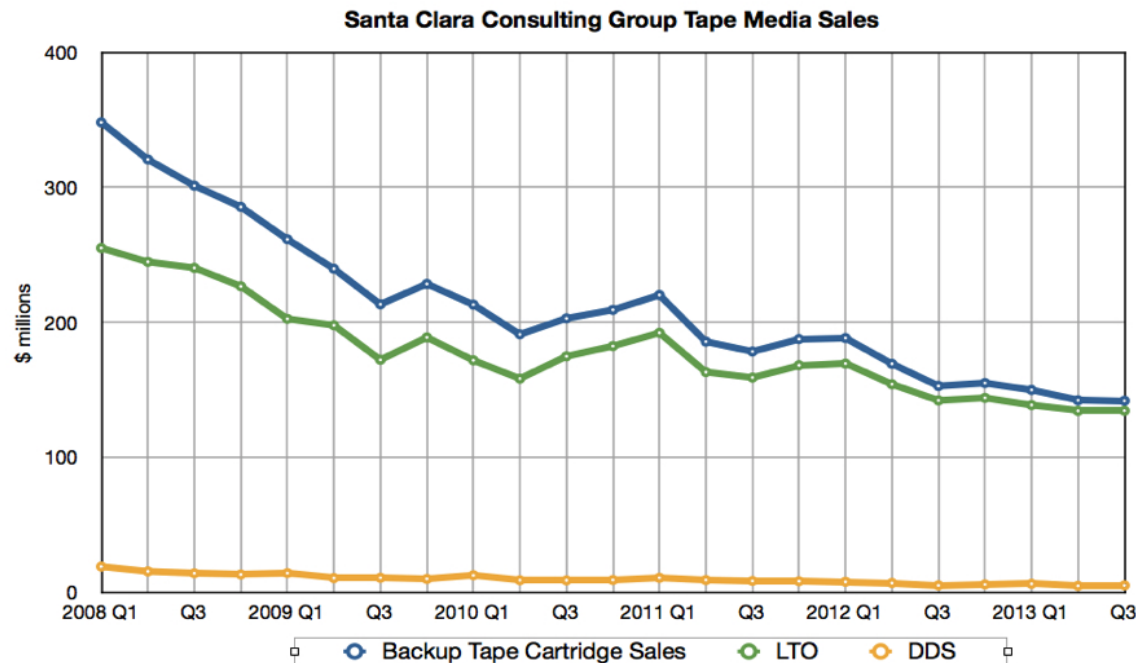
Vendor	Name	Capacity	Speed	Type	Date
IBM	TS1140	4TB	240MB/s	Enterprise	06/2011
LTO(*)	LTO-6	2.5TB	160MB/s	Commodity	12/2012
Oracle	T10000D	8.5TB	252MB/s	Enterprise	09/2013
IBM	???	???	???	Enterprise	???

(\*) : IBM/HP/Quantum (drives); Fuji/Maxell/TDK/Sony (media)

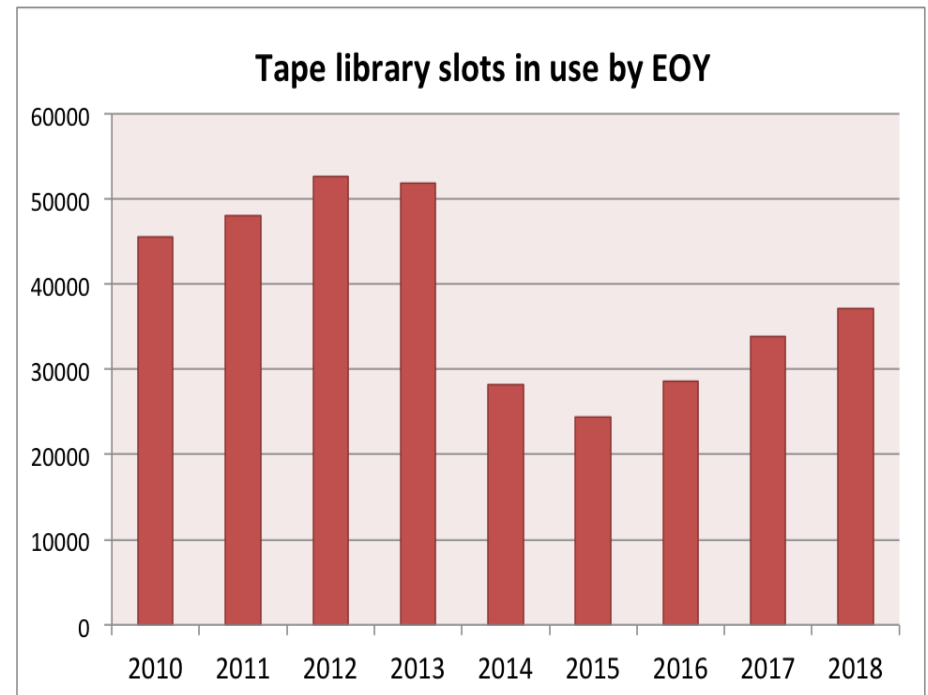
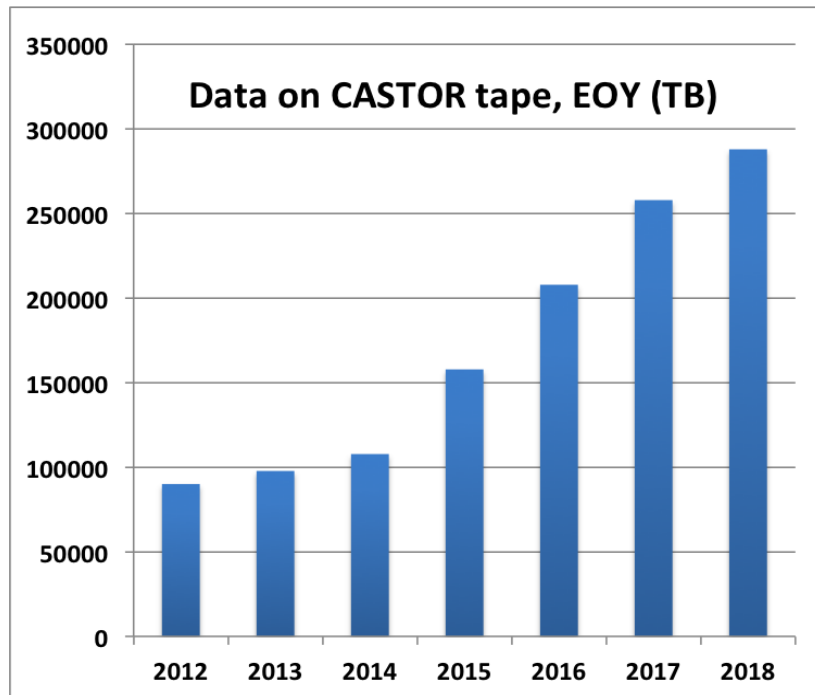
- R&D and Roadmaps for further evolution
  - Change from MP to BaFe media allowing finer particles and magnetisation
    - 45Gb/in<sup>2</sup> demo (~50TB tape)
    - 85.9Gb/in<sup>2</sup> demo by IBM/Fuji (~154TB tape) – announced this Monday!
  - Sony demonstration 4/2014: 125Gb/in<sup>2</sup> (~185TB) with sputtered CoPtCr
    - Cost of media production could be a concern
  - LTO Roadmap: LTO-7: 6.4TB (~2015), LTO-8: 12.8TB (~2018?)
  - Next enterprise drives generation? 2017? 15-20TB? (~2017)
  - Little / no improvements in tape loading/positioning

# Tape Market evolution (2)

- Commodity tape market is consolidating
  - LTO market share is > 90%; but market shrinking by ~5-10% / year (~600M\$ / yr in 2013)
  - Small/medium sized backups go now to disk
  - TDK & Maxell stopping tape media production; other commodity formats (DAT/DDS, DLT, etc) frozen
  - LTO capacity increase slower (~27% / year compared to ~40% / year for enterprise)
- Enterprise tape is a profitable, growing (but niche) market
  - Large-scale archive market where infrastructure investment pays off, e.g. Google (O(10)EB), Amazon(?), scientific (SKA – up to 1EB/yr), ISP's, etc
  - Will this suffice to drive tape research and production?
  - Competition from spun-down disk archive services ie Evault LTS2 (Seagate)

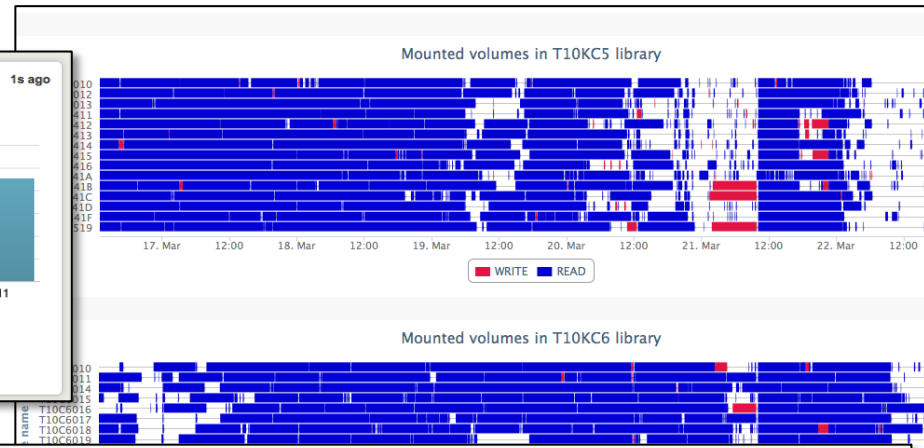
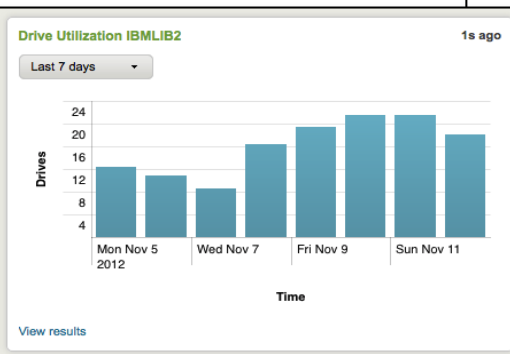
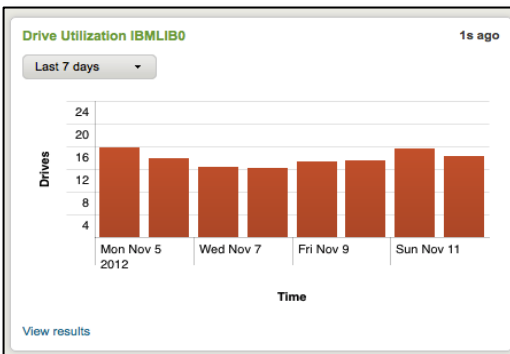


- Detailed capacity/cost planning kept for ~4y time window (currently, up to beginning of LHC LS2 in 2018)
  - Expecting ~50PB / year of new data
- Tape libraries will be emptier.. for some time
  - Decommissioned media will be sold or re-used for TSM
  - ~25K tapes after repack completes
  - + ~7K tapes / year with Run2
  - Will review library assets during LS2
- Next Big Repack likely to take place during LS2



- CERN Tape services, infrastructure in good running order and keeping up with media migration during LHC LS1
- Focus on developing, delivering and operating a performing reliable, long-term archive service
- Ensure scalability in terms of traffic, volume and cost for LHC Run 2

# Reserve material



## Tapelog Monitoring - Overview

German Cancio Melia | App - Manager Alerts Jobs

Overview Detailed Plots Failures Advanced Reports Searches Indexing Info Help About

Click on the graph. to access details !

TimeRange Year to date

#### Transferred Data Amount per Virtual Organization for Read Requests (Without repack, tape verification)

< 1m ago

Virtual Organization	Percentage
alice	17.58%
ams	5.79%
atlas	11.38%
compass	14.86%
cms	39.62%
hcb	3.62%
na48	0.06%
na61	4.64%
ntof	0.77%
other	0.0000%
user	1.66%

#### Transferred Data Amount per Virtual Organization for Write Requests (Without repack, tape verification)

< 1m ago

Virtual Organization	Percentage
alice	4.25%
ams	5.23%
atlas	39.88%
compass	5.56%
cms	27.41%
harp	9.63%
it	0.13%
na61	1.68%
ntof	1.57%
other	0.17%
repack	0.0000%
user	4.50%

#### Transferred Data Amount per Virtual Organization per Time for Read Requests (Without repack, tape verification)

< 1m ago

Legend: alice, ams, atlas, cms, compass, hcb, na48, na61, ntof, user

#### Transferred Amount By VO (Pie)

Total Data Amount Volume ( TB ) per VO ( in the last 30 days )

Legend: user, repack, other, ntof, na61, na48, lhcb, harp, compass, cms, atlas, alice, ams

#### Transferred Amount By Time

Total Data Amount Volume ( TB ) per Time ( in the last 30 days, scale=Auto )

Time: Wed Oct 10 2012, Wed Oct 24 2012

#### Transferred Amount By Time

Total Data Amount Volume ( TB ) per Time per VO ( in the last 30 days, scale=Auto )

Legend: alice, ams, atlas, cms, compass, lhcb, na61, ntof, repack, user

#### Transferred Amount By Device Group

Total Data Amount Volume ( TB ) per Device Group ( in the last 30 days )

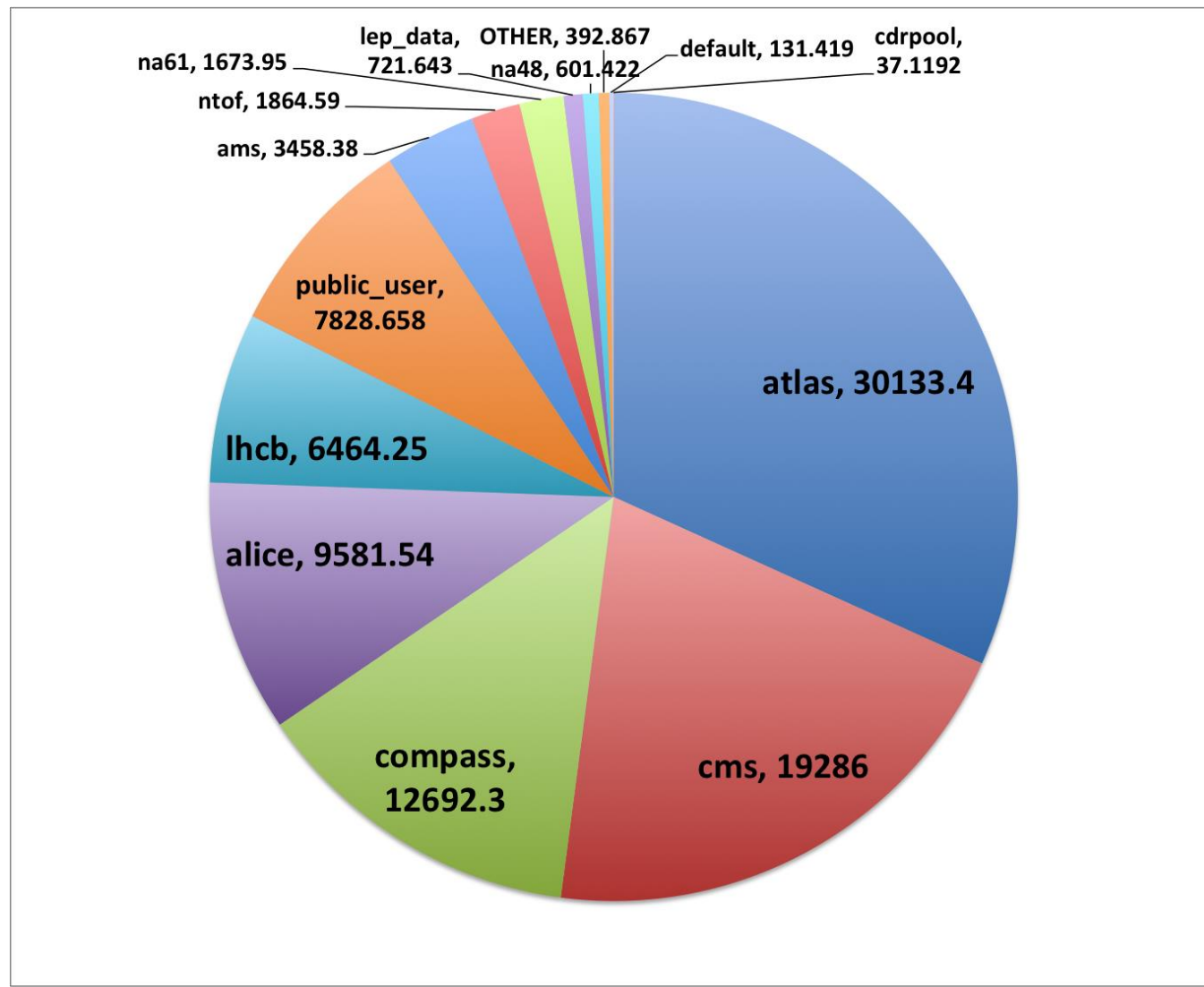
Legend: 3..1, 3..3, 3..4, N/A, S..5, T..5, T..6, T..5, T..6

#### Transferred Amount By Device Group

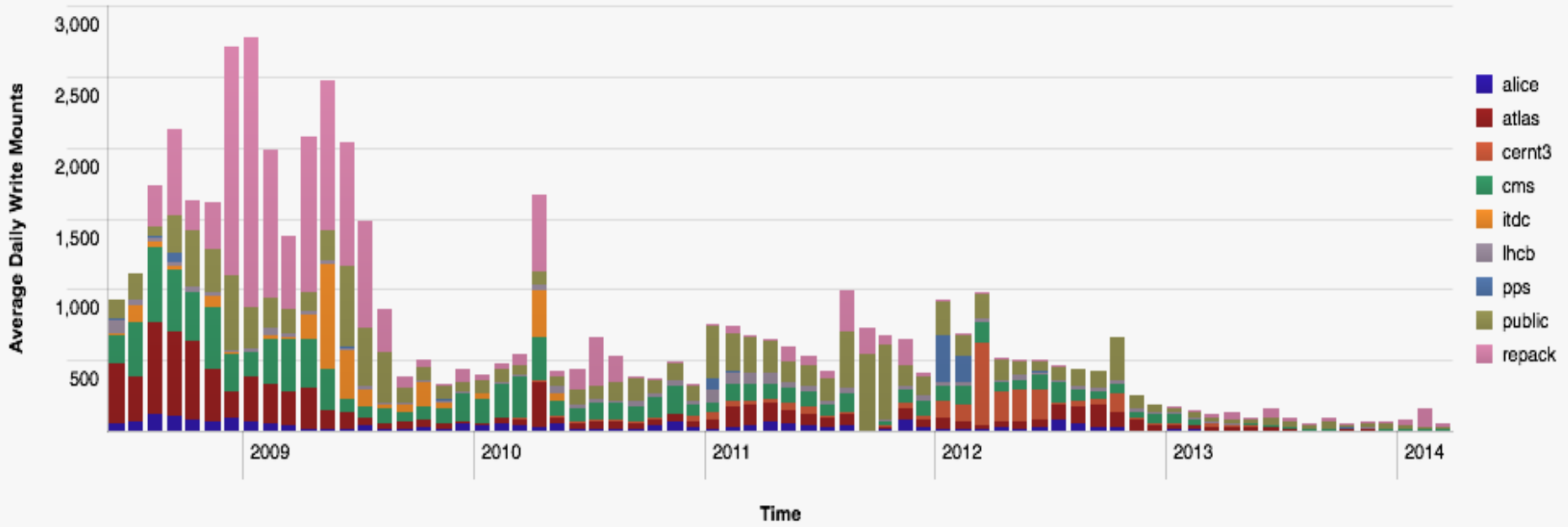
Total Data Amount Volume ( TB ) per Device Group per VO ( in the last 30 days )

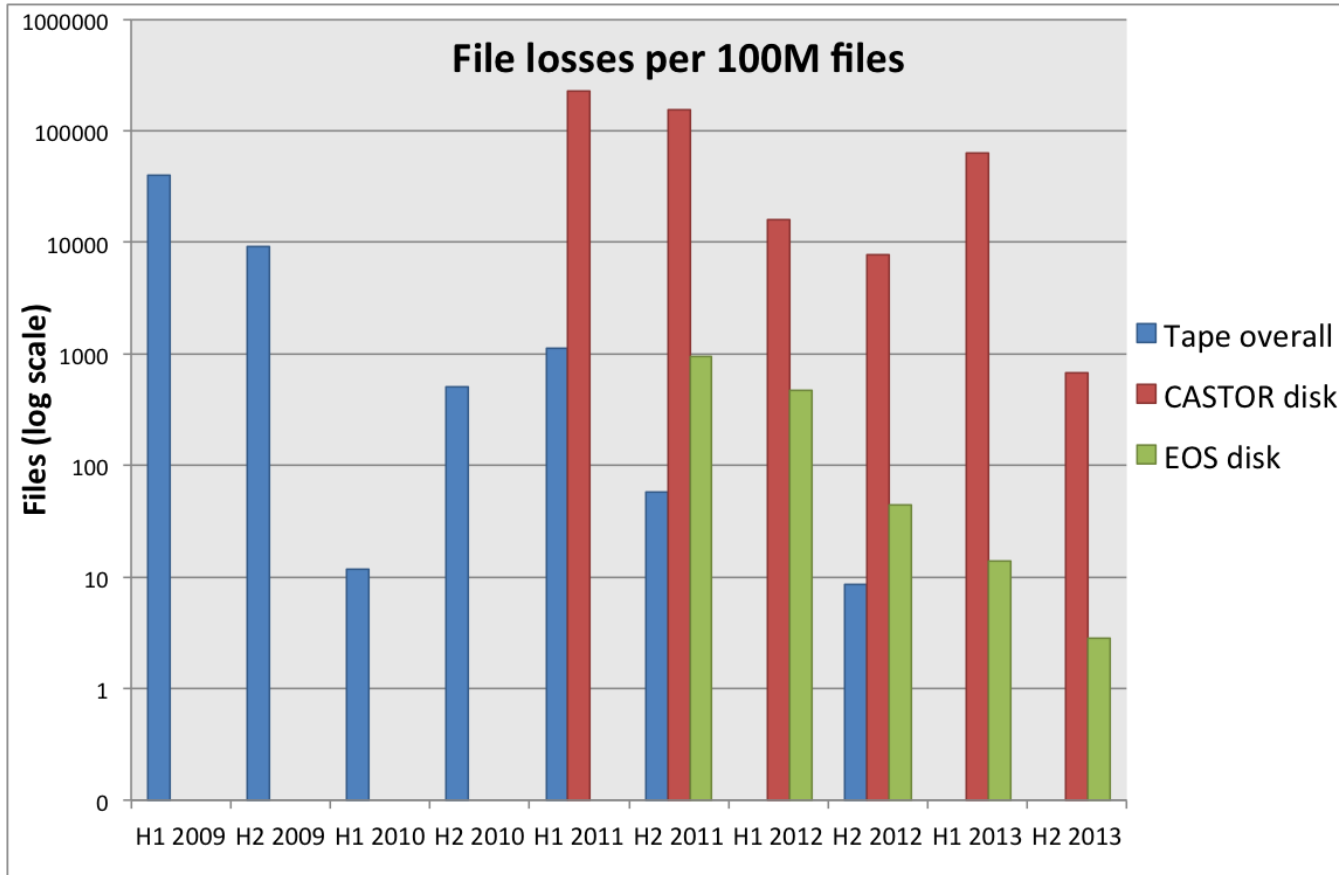
Legend: 359281, 359283, 359284, N/A, S1LT05, S1T0K85, T10K86, T10KC5, T10KC6

# CASTOR tape volume breakdown (TB)



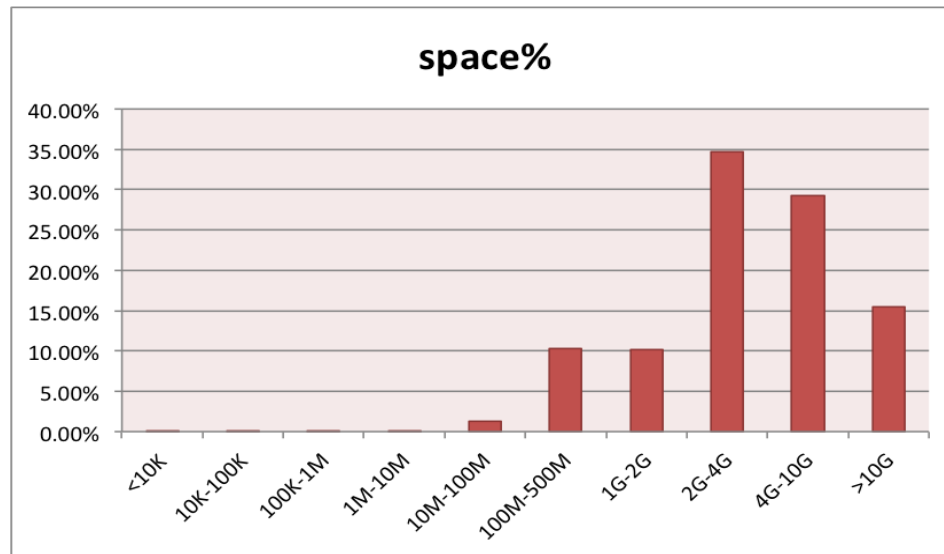
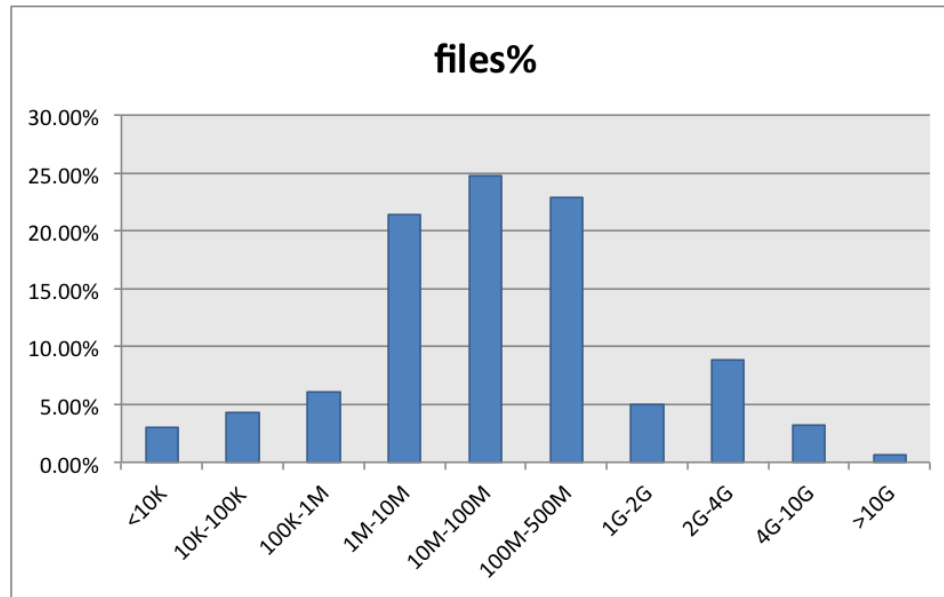
# CASTOR write tape mounts, 2009-2014





NB: 1 tape copy vs 2 disk copies (RAID1-CASTOR, JBOD-EOS)

# File size distribution, CASTOR tape



## RAID1 (CASTOR default)

Max controller throughput: ~350 MB/s

Synch time: ~1-2 minutes

1 stream 130 MB/s 100%

2 streams 107 MB/s 82%

4 streams 95 MB/s 73%

8 streams 87 MB/s 67%

## RAID10 (Repack)

Max controller throughput: ~400 MB/s

Synch time: ~3-4 seconds

1 stream 395 MB/s 100%

2 streams 304 MB/s 77%

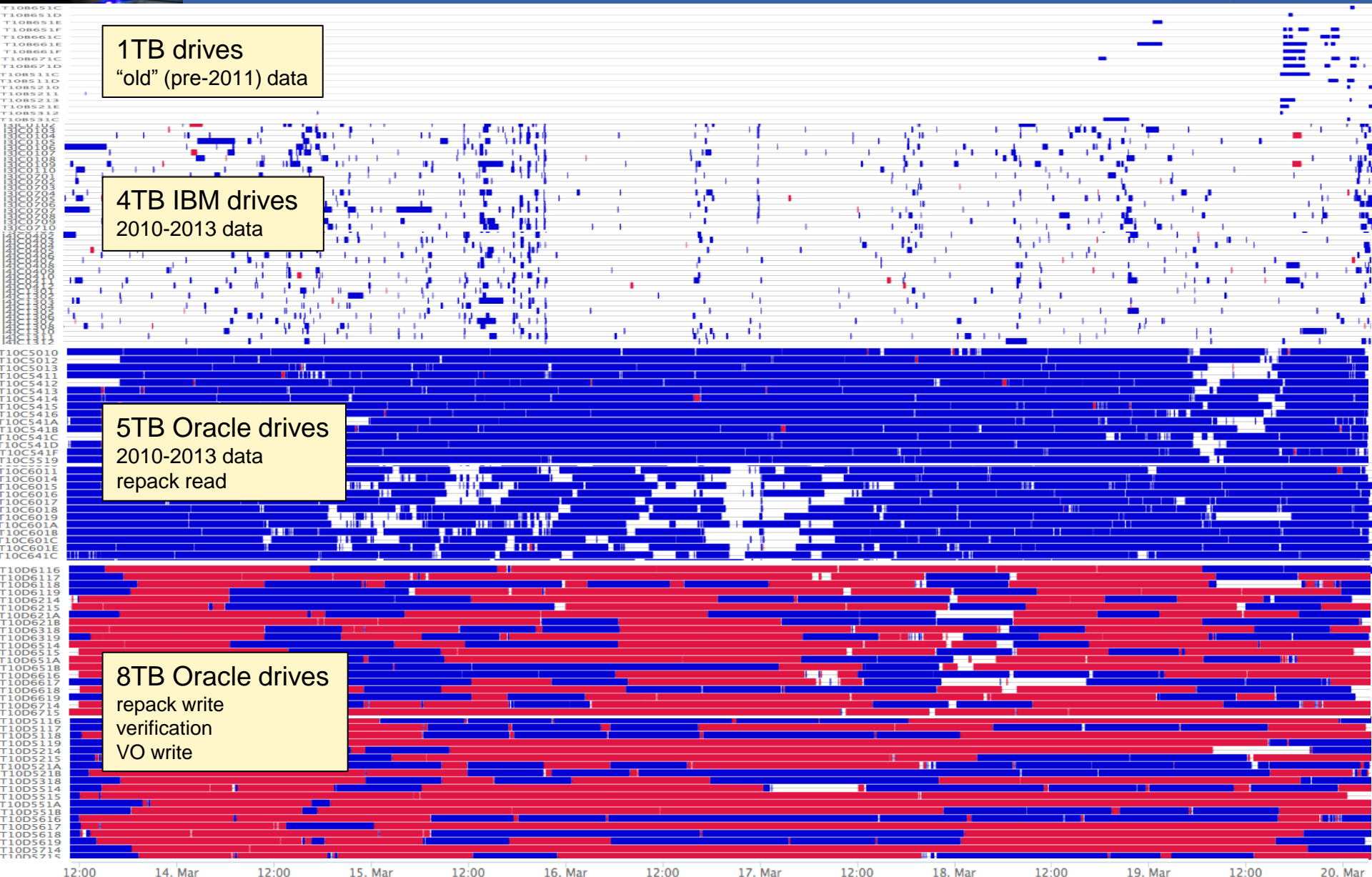
4 streams 272 MB/s 70%

8 streams 240 MB/s 61%

- **Throughput to disk** of parallel streams to the same filesystem measured **dd**'ing a 40GB file from /dev/zero
- **Max controller throughput** calculated with multiple streams to different filesystems



# Repack tape drive usage, 1w



1TB drives  
"old" (pre-2011) data

4TB IBM drives  
2010-2013 data

5TB Oracle drives  
2010-2013 data  
repack read

8TB Oracle drives  
repack write  
verification  
VO write

WRITE READ

# Drive comparison (T10KD: missing)

Tape Drive	LTO-6	LTO-5	TS1140	T10000C	T9940B	T9840D
Transfer rate Native/comp., MB/second	160 / 400	140 / 280	250 / 650	240 / 360	30 / 70	30 / 70
Capacity Native/comp.	2.5 TB / 6.25 TB*	1.5 TB / 3 TB	4 TB / 12 TB	5 TB / up to 5.5TB	200 GB / 400 GB	75 GB / 150 GB
Buffer, MB	1024	512	1000	2000	64	N/A
Speed match	14	14	12	2	N/A	N/A
Watts. max-avg.	27	29.5	50	67	61.7	82
MTBF Power-on, hours	250,000	250,000	237,000	N/A	290,000	290,000
Uncorrected Bit Error Rate	10 <sup>-17</sup>	10 <sup>-17</sup>	10 <sup>-19</sup>	10 <sup>-19</sup>	10 <sup>-18</sup>	10 <sup>-18</sup>
Head Life	60,000 hours	60,000 hours	N/A	5 years	8.5 years at 70% duty	
Load / First File Access / Rewind seconds	12 / 62 / 42	12 / 60 / 60	15 / 42 / 38	13.1 / 57 / 115	16.5 / 41 / 48	N/A / 9 / 8
Interfaces Gb/sec	8 FC / 6 SAS*	8 FC / 6 SAS	8 FC / FICON / ESCON	4 FC / FICON	2 FC / FICON / ESCON	
New features	LTFS	LTFS	LTFS	LTFS	No	No

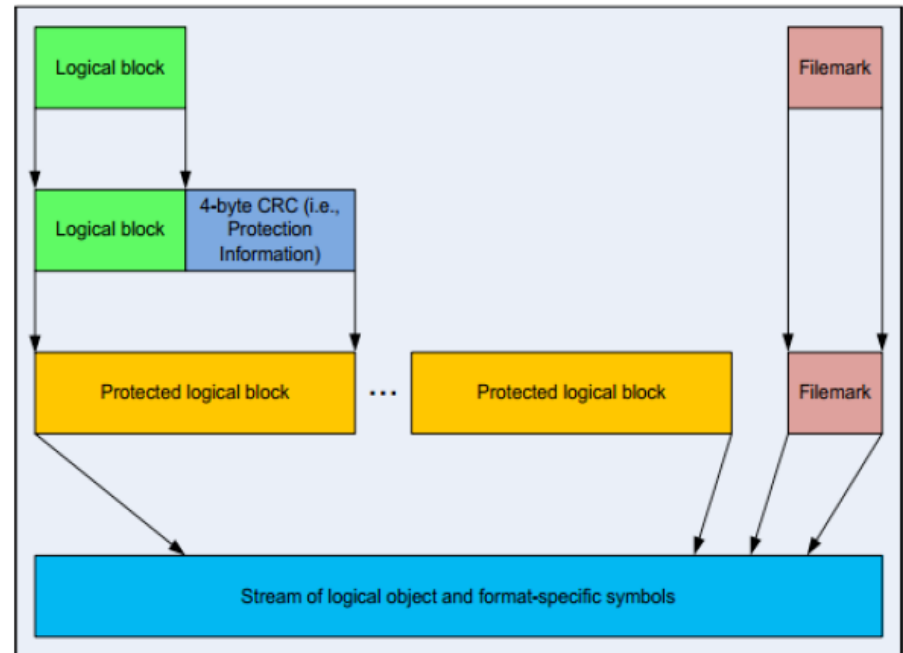
## Logical Block Protection

### Steps

1. Append CRC to each logical data block before writing to the tape.
2. Tape Drive calculates CRC of given data block and compares it with appended CRC.
3. If they match, block is processed. Otherwise is discarded.

### Why ?

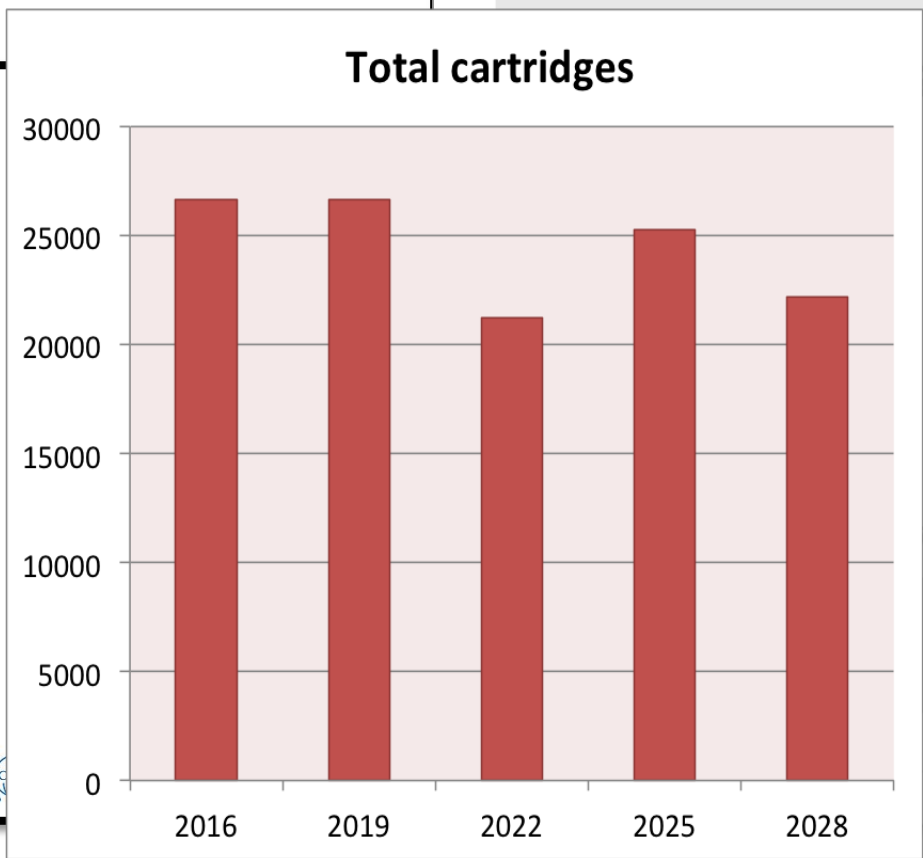
1. Ensure data consistency on writing / reading
2. Verify data on-drive ( no need to send data to tapeserver )



# Longer term?

- Beyond 2018?
  - Run 3 (2020-2022): ~150PB/year
  - Run 4 (2023-2029): ~600PB/year
  - Peak rates of ~80GB/s

### Total cartridges



### Number of disk servers required

