CASTOR status and development

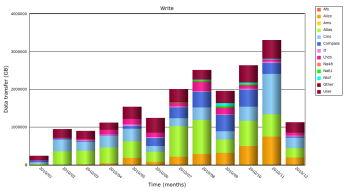
HEPiX Spring 2011, 4th May 2011
GSI, Darmstadt

Eric Cano
on behalf of CERN IT-DSS group

CASTOR
CERN Advanced STORage manager

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

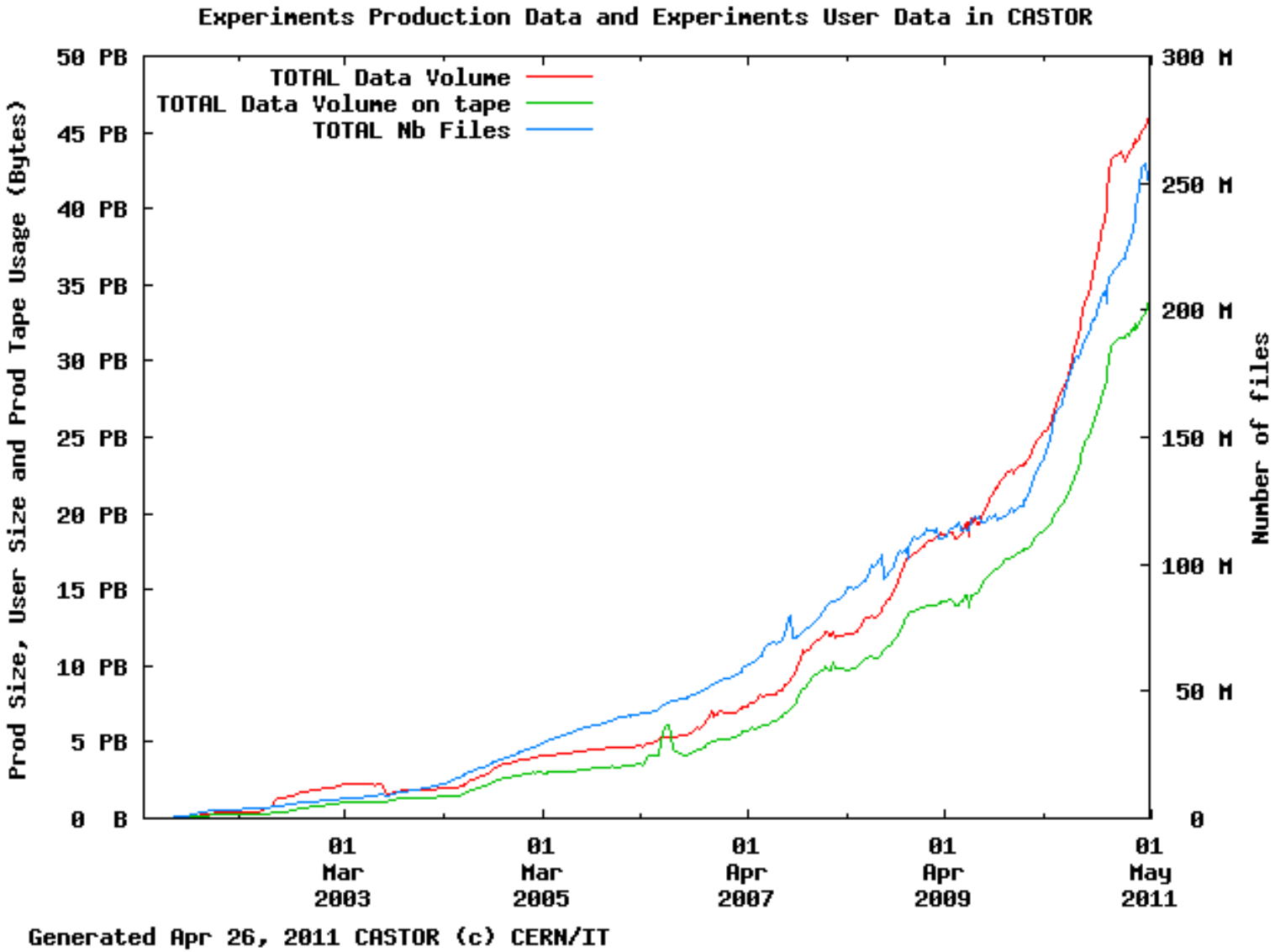
Outline



CASTOR
CERN Advanced STORage manager

- CASTOR's latest data taking performance
- Operational improvements
  - Tape system performance
  - Custodial actions with in-system data
- Tape hardware evaluation news
  - Oracle T10000C features and performance
- Recent improvements in the software
- Roadmap for software

Amount of data in CASTOR

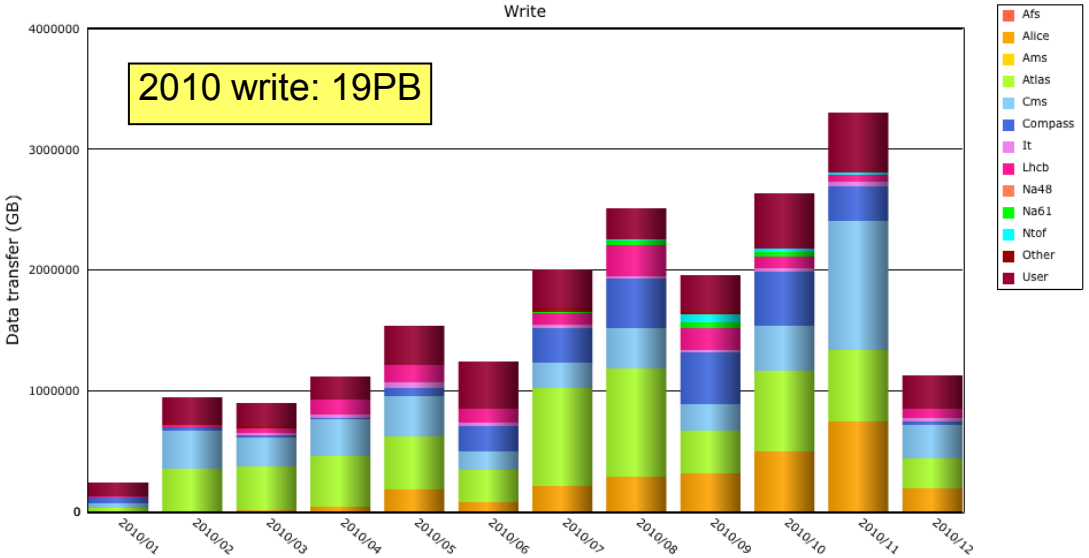


Key tape numbers

39 PB of data on tape

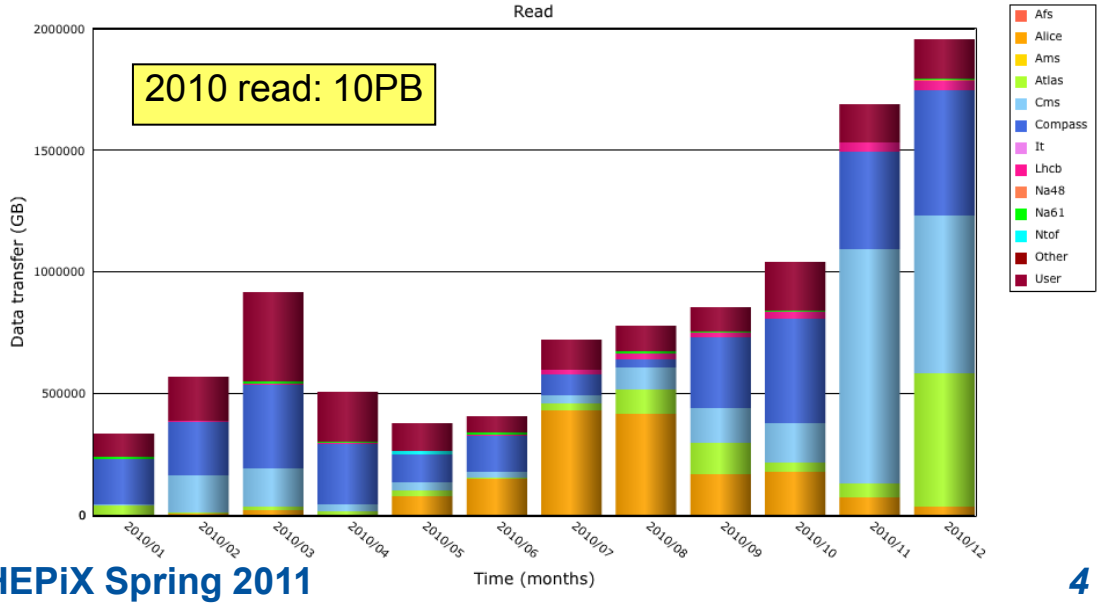
170M files on tape

Peak writing speed: 6GB/s (HI)



Infrastructure:

- 5 CASTOR stager instances
- 7 libraries (IBM+STK), 46k 1TB tapes
- 120 enterprise drives (T10000B, TS1130, installing T10000C)

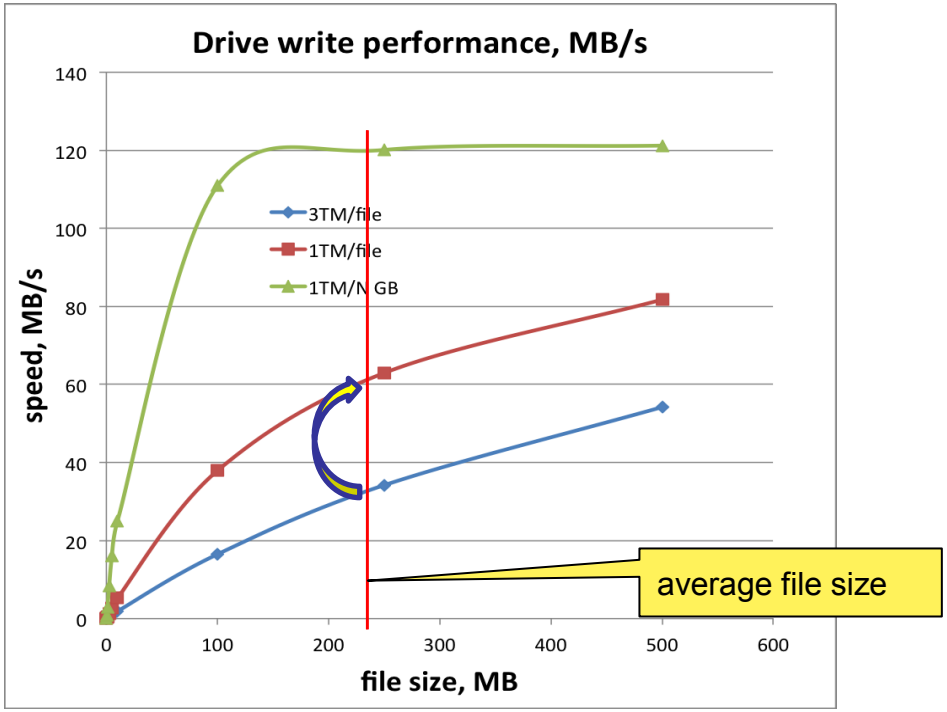


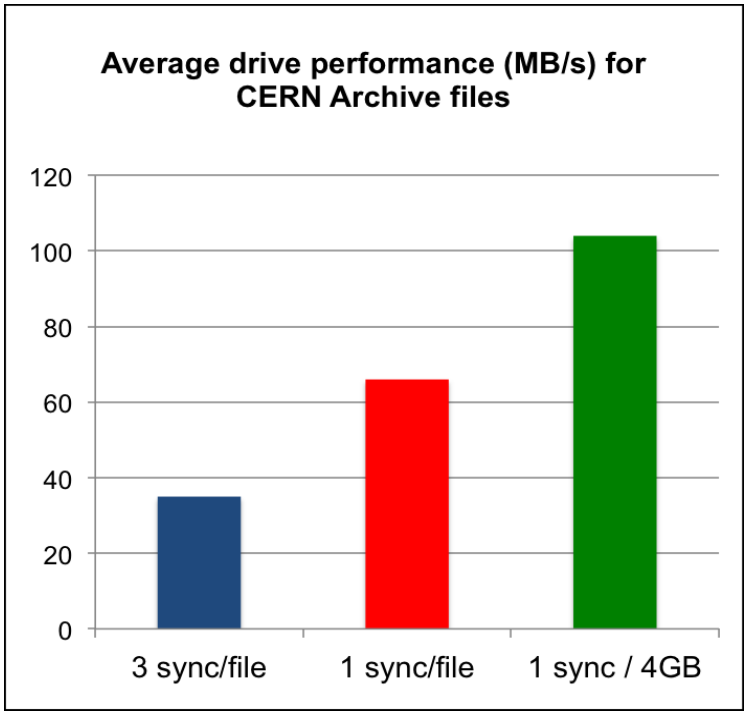
Areas for improvement

- Improve tape write performance
  - Get closer to the native speed
  - Improve efficiency of migration to new media
- Discover problems earlier
  - Actively check the data stored on tape
- Improve tape usage efficiency
  - Minimize overhead of tape mounts
  - Encourage efficient access patterns from users

Writing: Improvements

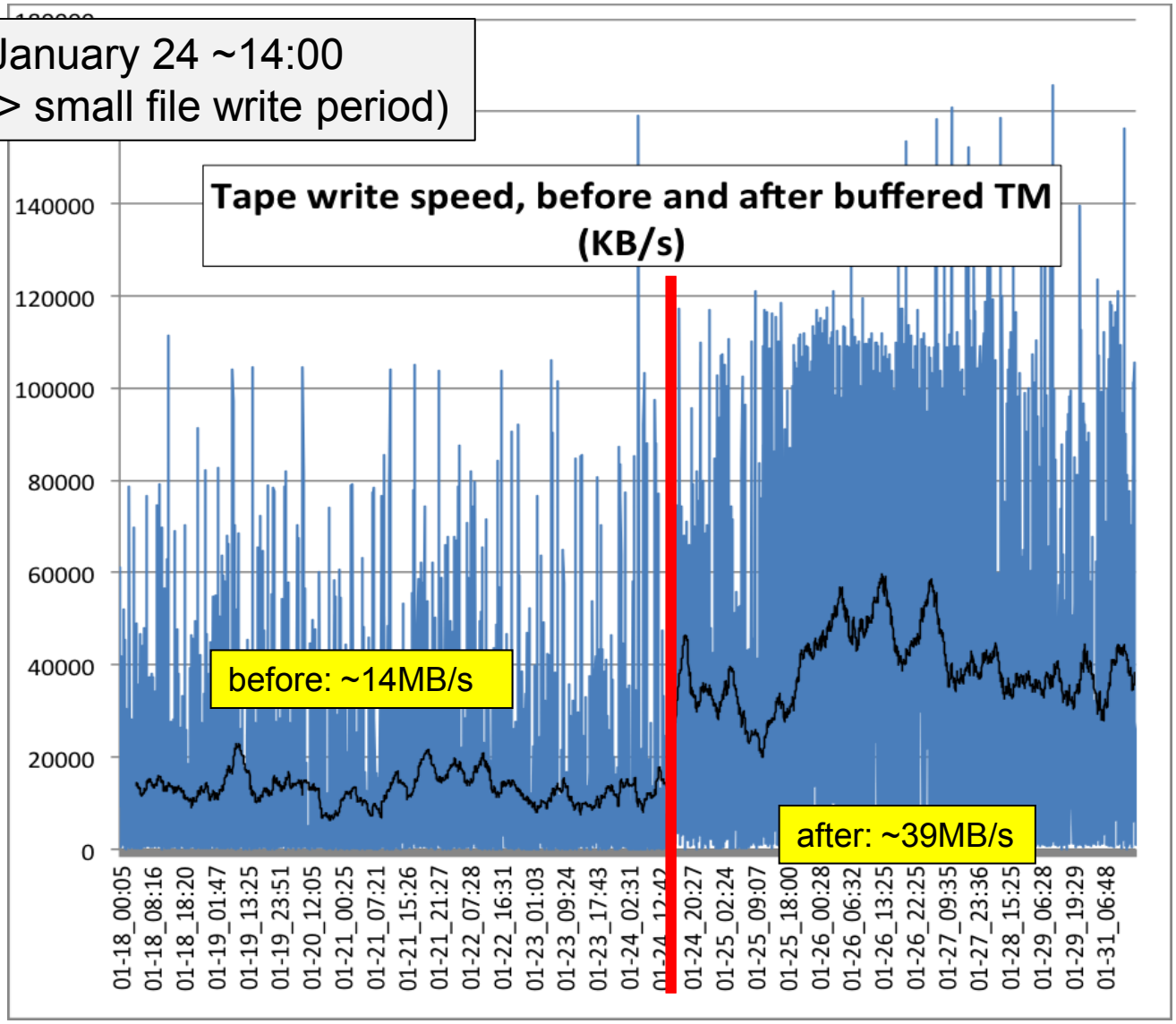
- Using buffered tape marks
  - Buffered means no sync and tape stop – increased throughput, less tape wear
  - Part of SCSI standard, available on standard drives, was missing in Linux tape driver
  - Now added to the mainstream Linux kernel, based on our prototype
- New CASTOR tape module writing one synchronizing TM per file instead of 3
  - Header, payload and trailer: synchronize trailer's TM only
- More software changes needed to achieve native drive speed by writing one synchronizing TM for several files





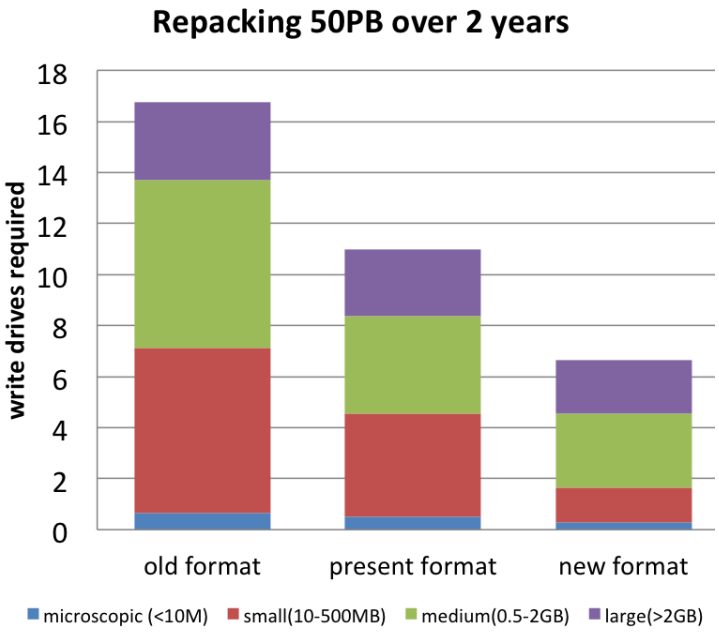
Buffered TM deployment

Deployment on January 24 ~14:00
(no data taking -> small file write period)



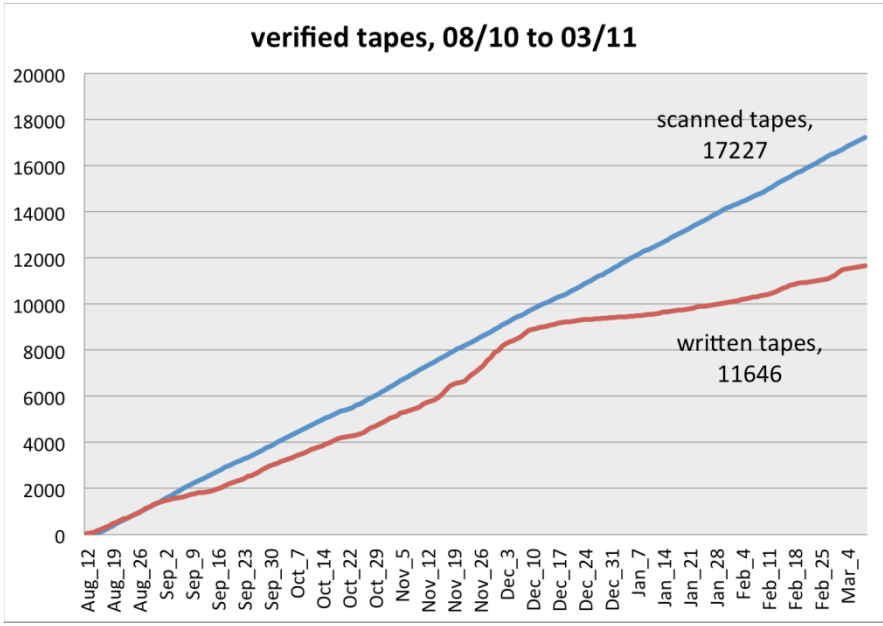
Media migration / repack

- Write efficiency is critical for migration to new media (aka “repacking”)
- Repack exercise is proportional to the total size of archive
  - Last repack (2009, 18PB): 1 year, 40 drives,1FTE: ~570MB/s sustained over 1 year
  - Next repack (50PB): ~800MB/s sustained over 2 years
  - Compare to p-p LHC rates: ~700MB/s
  - Efficiency will help prevent starvation on drives
- Turn repack into a reliable, hands-off, non-intrusive background activity
  - Opportunistic drive and media handling
  - Ongoing work for removing infrastructure and software bottlenecks/limitations



Media and data verification

- Proactive verification of archive to ensure that:
  - cartridges can be mounted
  - data can be read and verified against metadata (checksum, size, ..)



- Background scanning engine, from "both ends"
  - Read back all newly filled tapes
  - Scan the whole archive over time, starting with least recently accessed tapes

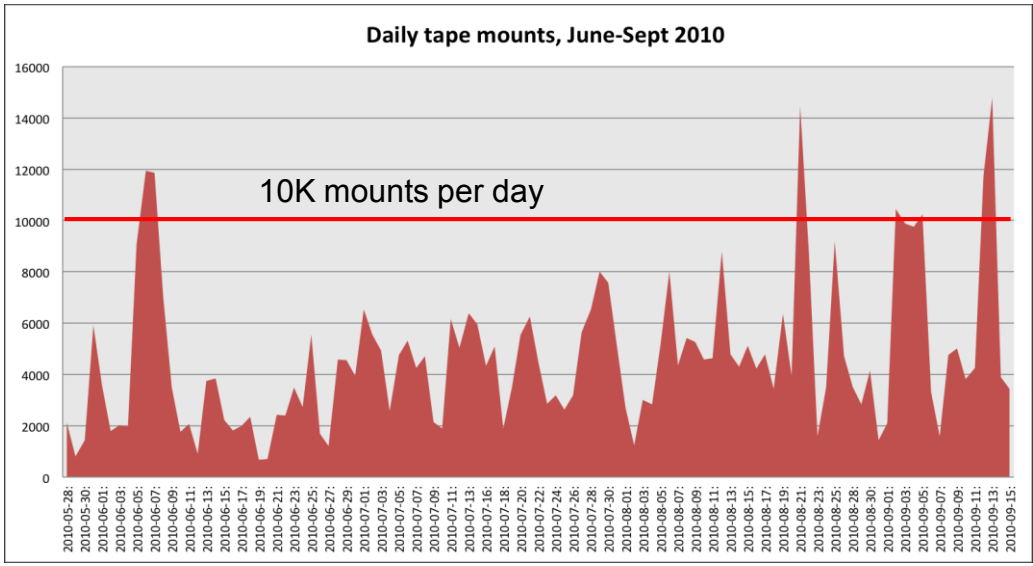
- Since deployment (08/2010-03/2011), verified 19.4 PB, 17K tapes
  - Up to 10 drives @ native drive speed (~8% of available drives)
  - Required but reasonable resource allocation, ~90% efficiency
  - About 2500 tapes/month, or 16 month for the current 40k tapes.

Verification results

- ~ 7500 “dusty” tapes (not accessed in 1 year)
  - ~40% of pre-2010 data
- ~ 200 tapes with metadata inconsistencies (old SW bugs)
  - Detected and recovered
- ~ 60 tapes with transient errors
- 4 tapes requiring vendor intervention
- 6 tapes containing 10 broken files

Reading: too many mounts

- Mount overhead: 2-3 minutes
- Recall ASAP policy lead to many tape mounts for few files
  - Reduced robotics and drive availability (competition with writes)
  - Equipment wear out
  - Mount / seek times are NOT improving with new equipment
  - Spikes above 10K over all libraries

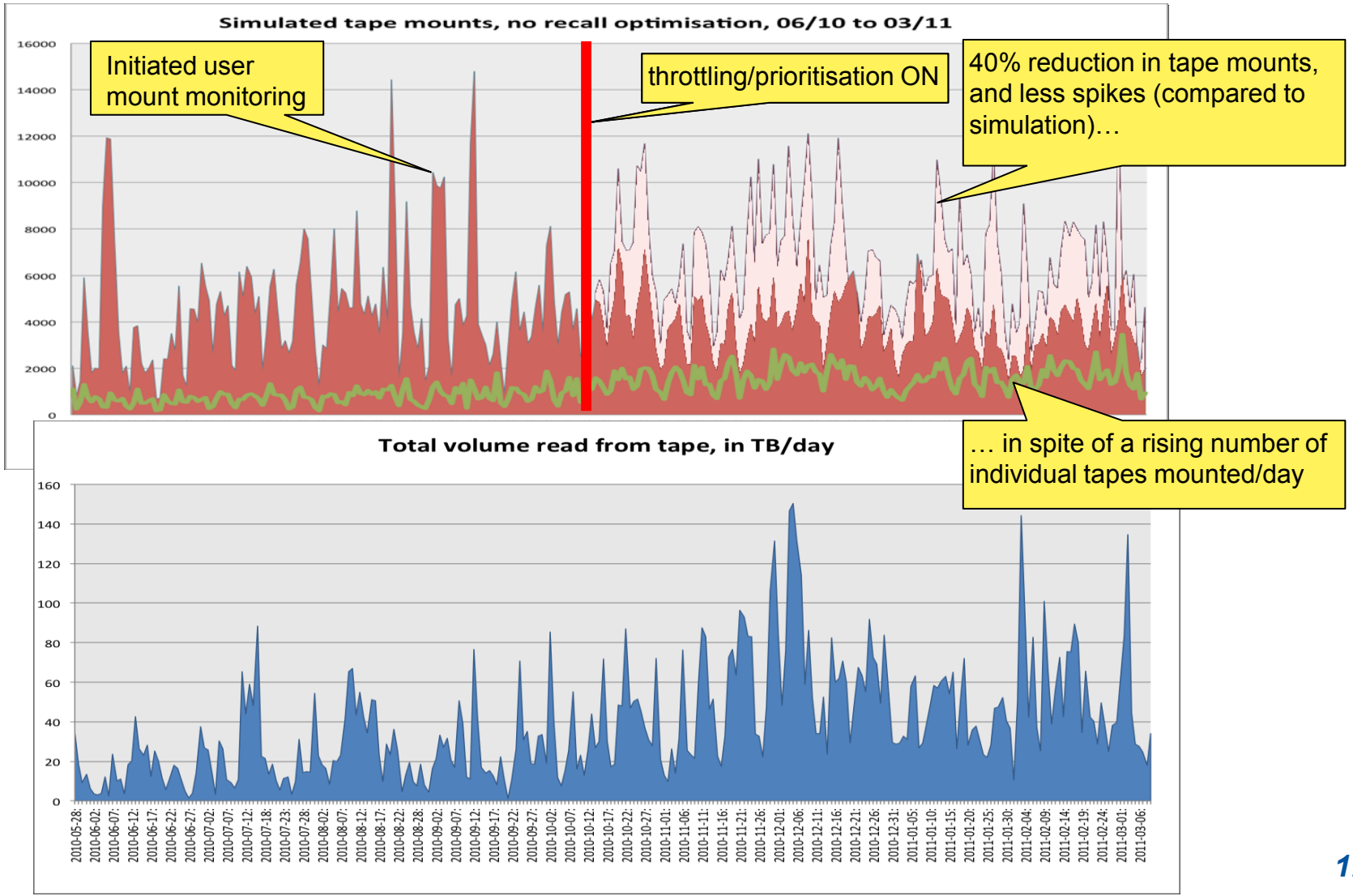


4 month: May – Sept 2010

- Actions taken
  - Developed monitoring for identifying inefficient tape users
  - User education: inefficient use detected and bulk pre-staging advertised to teams and users
  - Increased disk cache sizes where needed to lower cache miss rate
  - Updated tape recall policies: minimum volumes, maximum wait time, prioritization
    - Mount later, to do more

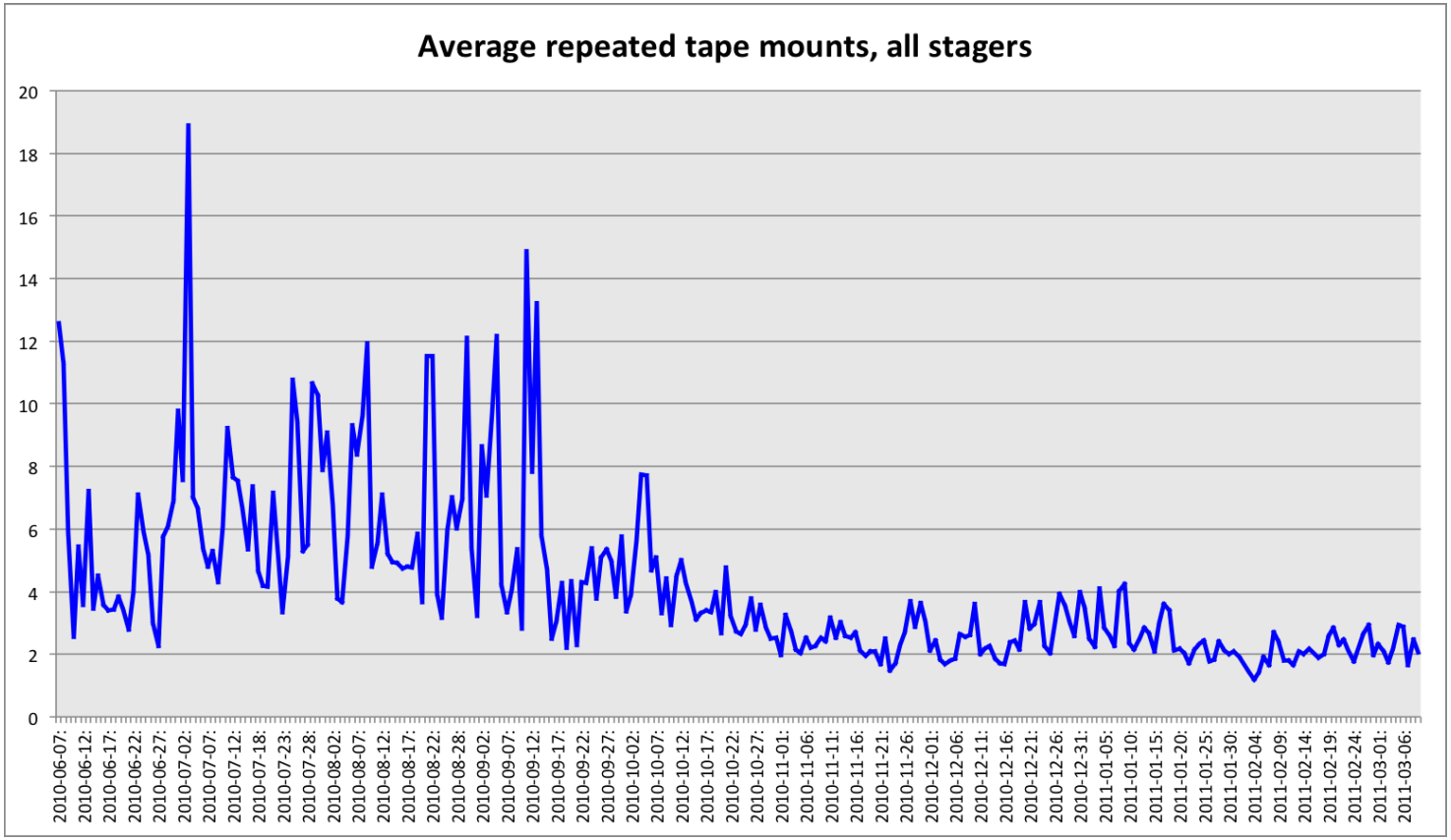
Reading: Operational benefits

Benefit for operations: less read mounts (and less spikes) despite increased tape traffic…



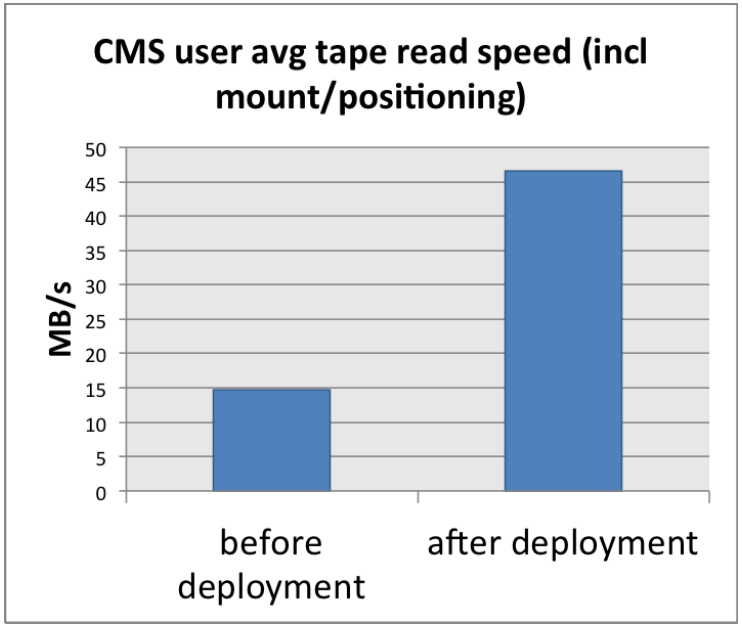
Reading: Operational benefits

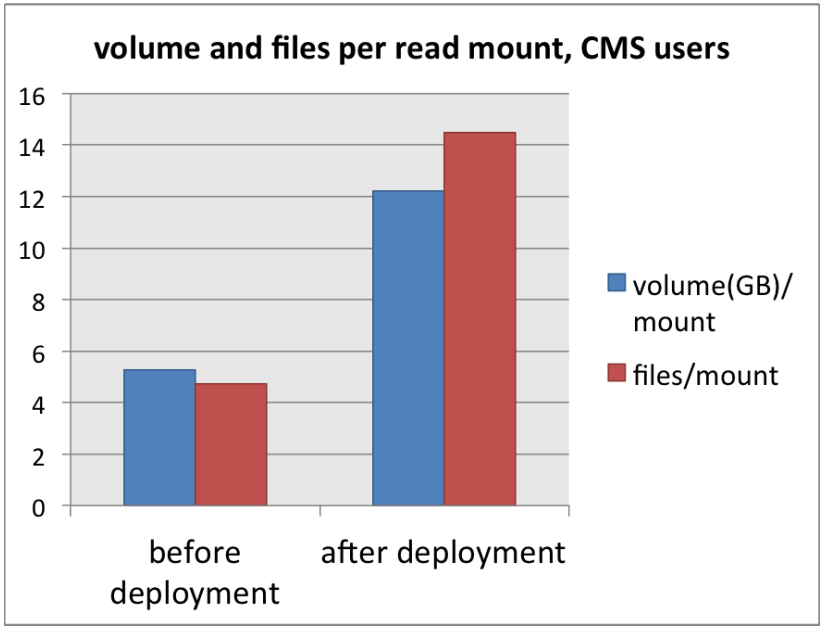
… also, reduction in repeatedly mounted tapes.



Reading: User benefits

- Do end users benefit? The case of CMS
  - From a tape perspective, CMS is our biggest client
    - >40% of all tape mounts and read volume
  - CMS tape reading is 80% from end users, and only 20% from Tier-0 activity
  - CMS analysis framework provides pre-staging functionality
  - Pre-staging pushed by experiment responsibles and recall prioritisation in place
- Average per-mount volume and read speed up by a factor 3





Oracle StorageTek T10000C

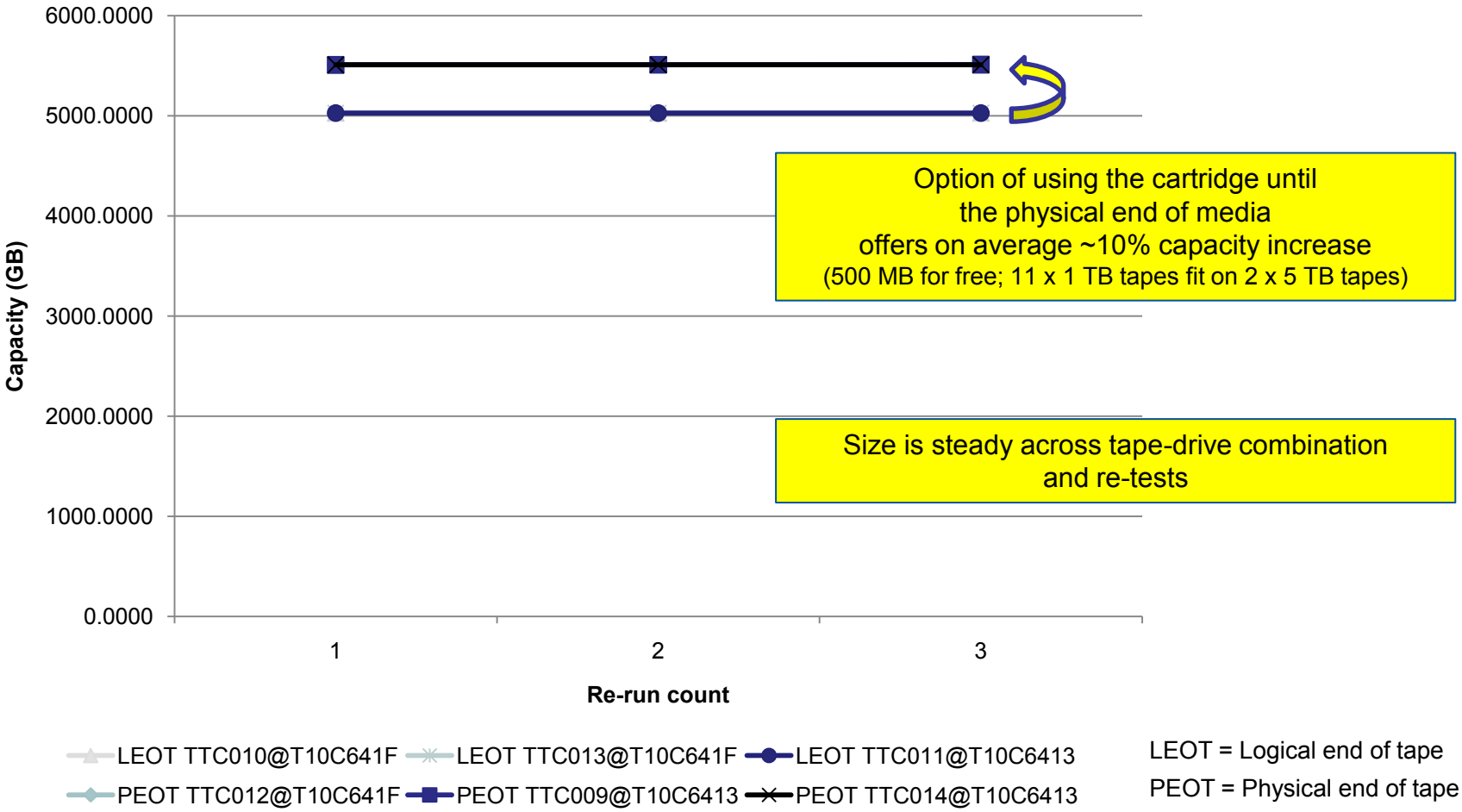
- Oracle StorageTek T10000C tape drive test results
  - Report from beta tests by Tape Operations team
- Hardware
  - Feature evolution
  - Characteristics measured
- Tests
  - Capacity
  - Performance
  - Handling of small files
  - Application performance

T10000C vs. T10000B

Extract of features relevant for CERN	T10000C	T10000B
Tape Format	Linear Serpentine	Linear Serpentine
Native cartridge capacity (uncompressed)	5 TB	1 TB
Head design	Dual heads writing 32 tracks simultaneously	Dual heads writing 32 tracks simultaneously
Native data rate performance (uncompressed)	Up to 240 MB/sec	Up to 120 MB/sec
Tape speed (low / high / locate)	3.74 / 5.62 / 10-13 m/s	2 / 3.7 / 8-12 m/s
Internal data buffer	2 GB	256 MB
Small files handling functionality	File Sync Accelerator	None
Max Capacity Feature	Yes	Yes

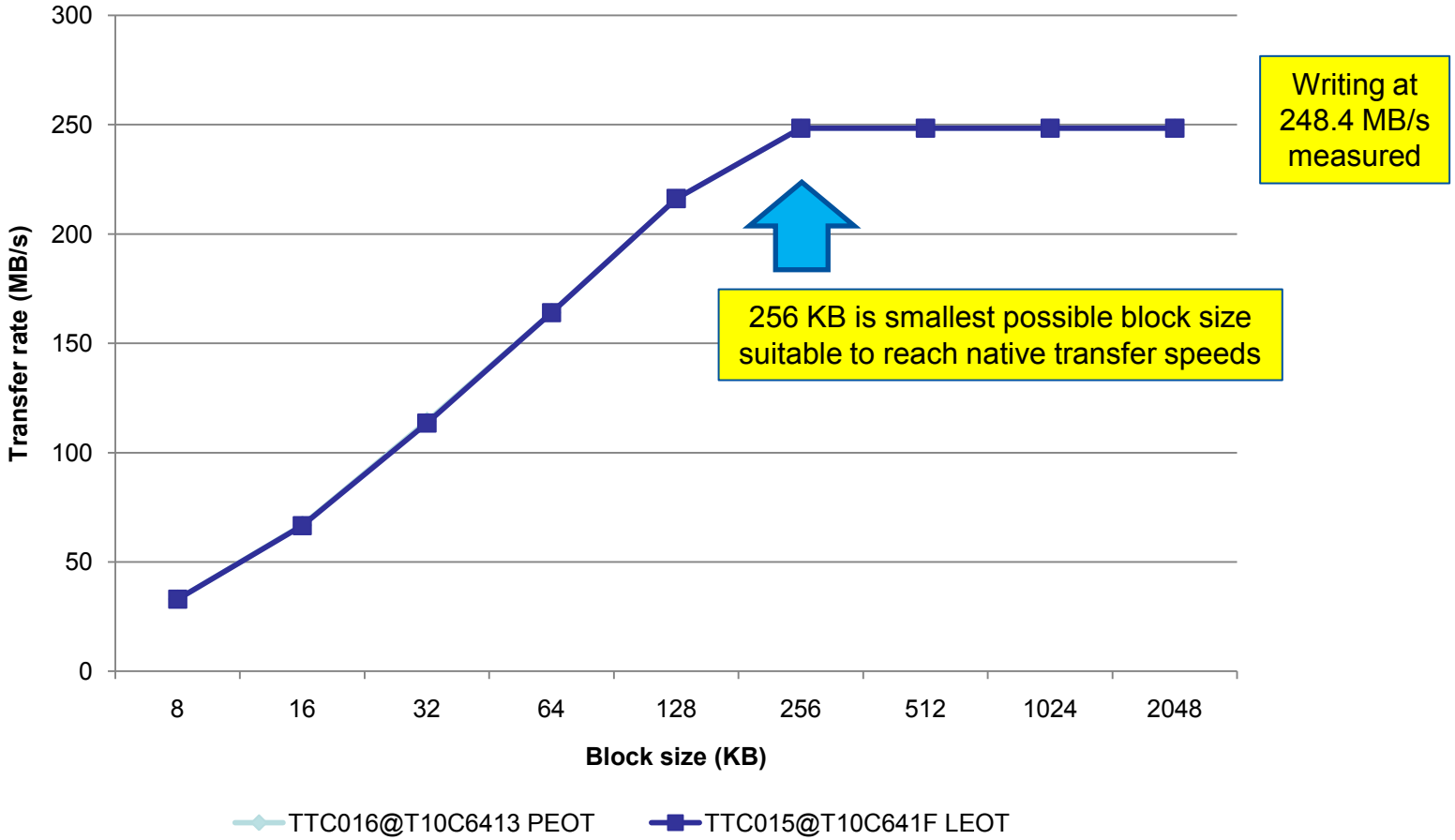
http://www.oracle.com/us/products/servers-storage/storage/tape-storage/t10000c-drive-ds-289750.pdf

Capacity – logical vs. physical



LEOT = Logical end of tape
PEOT = Physical end of tape

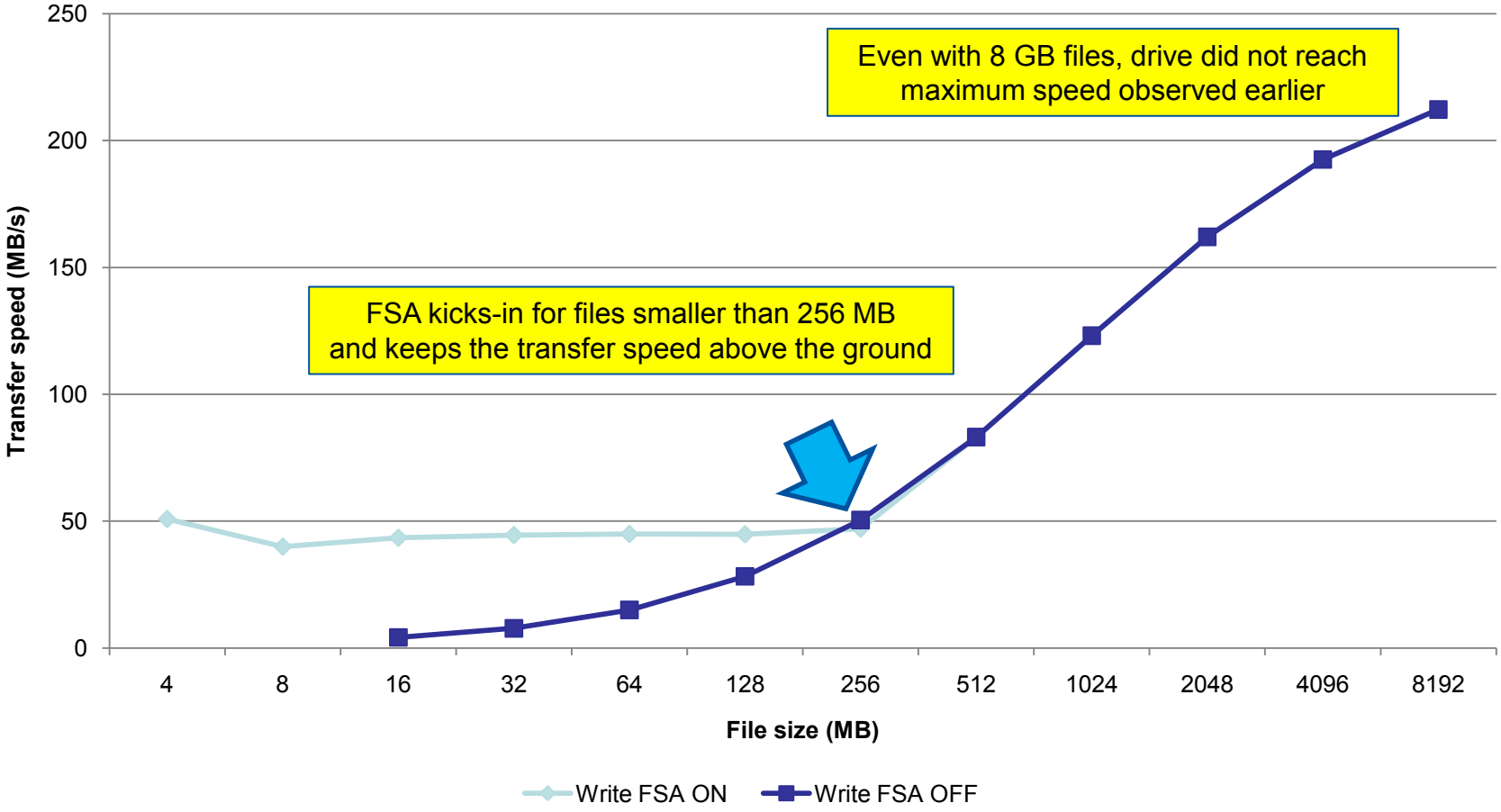
Performance – block transfer rates



Details:

- `dd if=/dev/zero of=/dev/tape bs=(8,16,32,64,128,256,512,1024,2048)k`

File Sync Accelerator – write 1TB

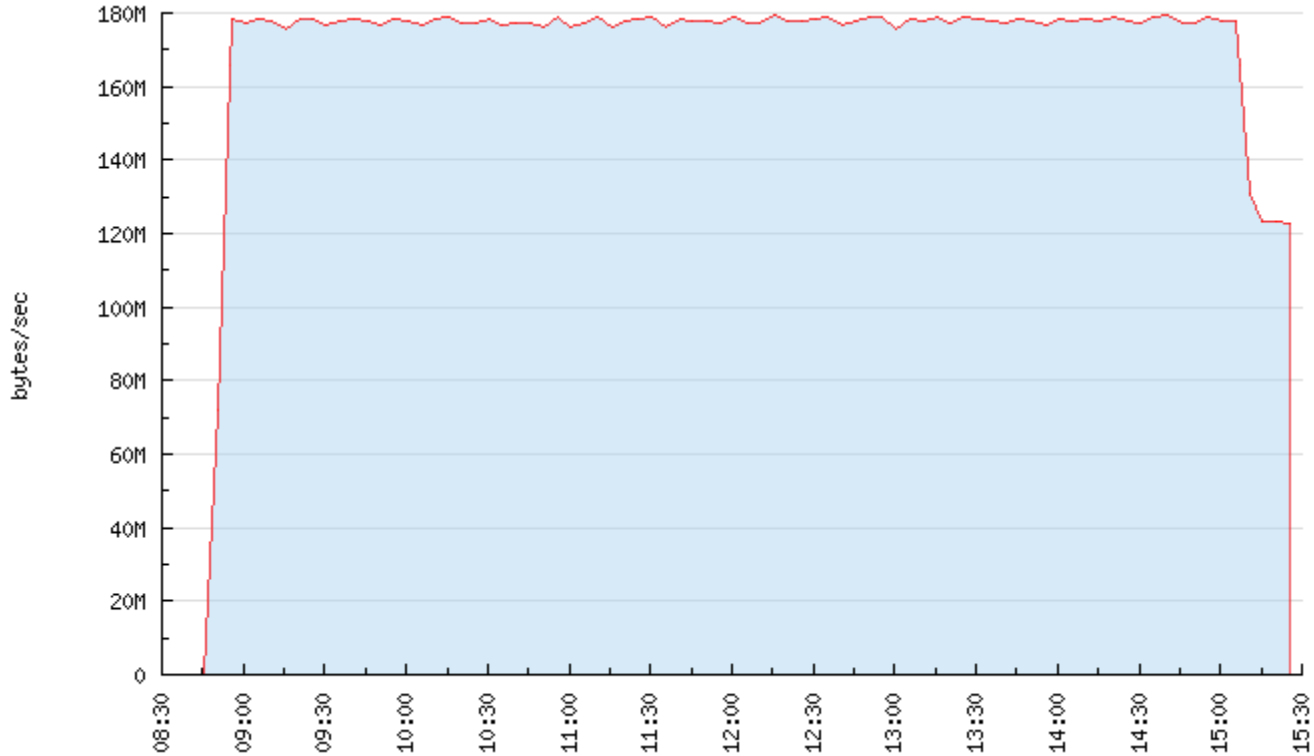


Details:

- Repeat following command until 1 TB of data written (measure the total time):
- `dd if=/dev/zero ibs=256k of=/dev/nst0 obs=256k count=(16,32,64,128,256,512,1024,2048,4096,8192,16384,32768)`
- FSA ON = `sdparm --page=37 --set=8:7:1=0 /dev/tape`
- FSA OFF = `sdparm --page=37 --set=8:7:1=1 /dev/tape`

Application performance

- 7 March 2011
  - Tape server tpsrv664 with 10 Gbps network interface
  - Receiving 2 GB files from several disk servers
  - Writing it to tape during almost 7 hours



- Achieved almost 180 MB/s = 75% of native transfer speed. Very good result!

Conclusions on T10000C

- The next generation tape drive is capable of storing 5 TB of data on tape
- It can reach transfer speeds up to 250 MB/s
  - Day-to-day performance however will likely be lower if files smaller than 8 GB (with tape marks in between)
- Requires less cleaning cycles than previous generation tape drive
- Need to adapt the CASTOR software to benefit from the improved performance

Further market evolution

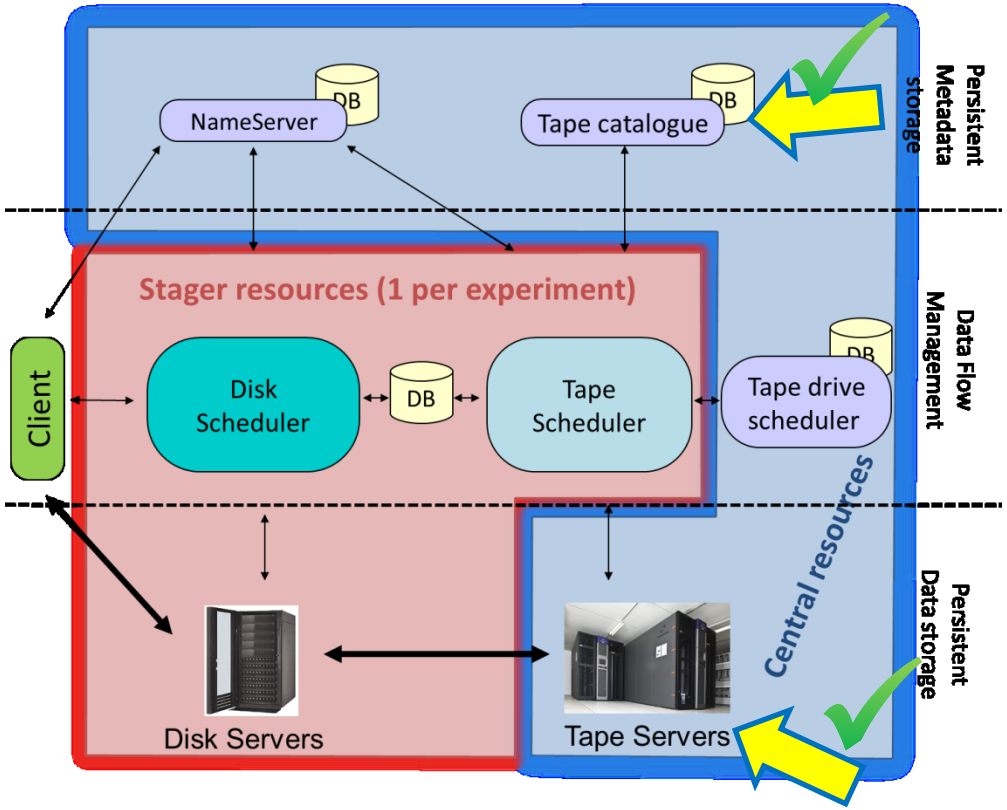
Vendor	Name	Capacity	Speed	Type	Date
IBM	TS1130	1TB	160MB/s	Enterprise	07/2008
Oracle (Sun)	T10000B	1TB	120MB/s	Enterprise	07/2008
LTO consortium(*)	LTO-5	1.5TB	140MB/s	Commodity	04/2010
Oracle	T10000C	5TB	240MB/s	Enterprise	03/2011
IBM	?	?	?	Enterprise	?
LTO consortium(*)	LTO-6?	?	?	Commodity	?

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

- Tape technology recently getting a push forward
  - New drive generation released
- CERN currently uses
  - IBM TS1130, Oracle T10000B, Oracle T10000C (commissioning)
- We keep following up enterprise drives and will experiment LTO
  - LTO good alternative or complement if low tape mount rate achieved
  - Will test other enterprise drives as they come.
- More questions? Join the discussion! <hep-tape-experts@cern.ch>

Software improvements

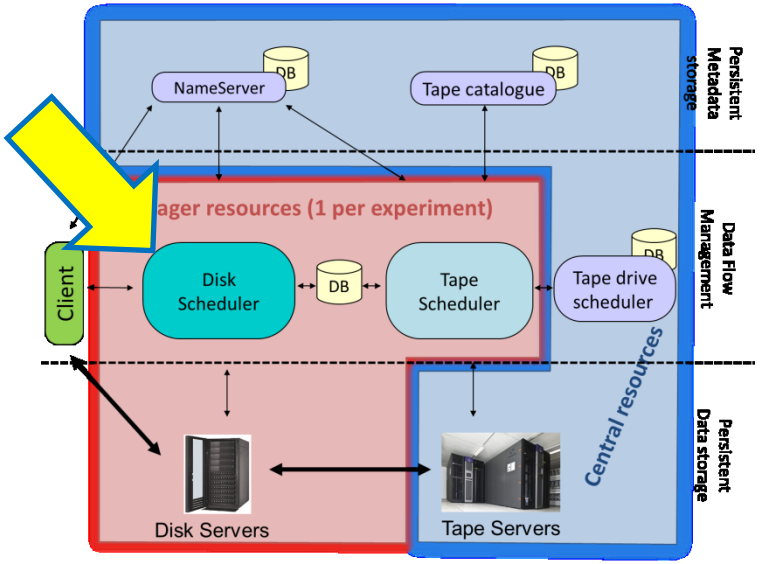
- Already in production
  - New front-end for tape servers
  - Tape catalogue update
    - Support for 64bits tape sizes (up to 16 EB, was 2TB)
- In the next release
  - New transfer manager
  - New tape scheduler
  - Transitional release for both
    - Switch both ways between old and new



- In the works
  - One synchronous tape mark over several files
  - Modularized and bulk interface between stager and tape gateway
    - More efficient use of the DB under high load

Transfer manager

- Schedules client disk accesses
- Replaces LSF (an external tool)
  - To avoid double configuration
  - To remove the queue limitation
  - And associated service degradation
  - To reduce the latency of scheduling
  - To replace failovers by redundancy
- Provides long-term scalability as traffic grows

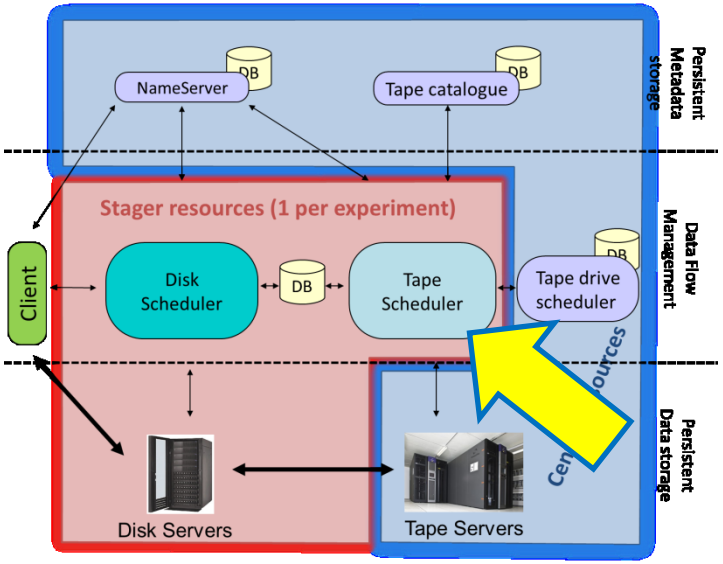


Transfer manager: test results

- Scheduling latency (few ms) (around 1s for LSF)
  - Latency from LSF's bunched processing of requests
- Scheduling throughput > 200 jobs/s (Max 20 in LSF)
- Scheduling does not slow down with queue length
  - Tested with > 250K jobs
  - LSF saturates around 20k jobs
- Active – Active configuration
  - No more failover of the scheduling (failover in 10s of seconds in LSF)
- A single place for configuration (stager DB)
- Throttling available (25 jobs/transfer manager by default)

Tape gateway and bridge

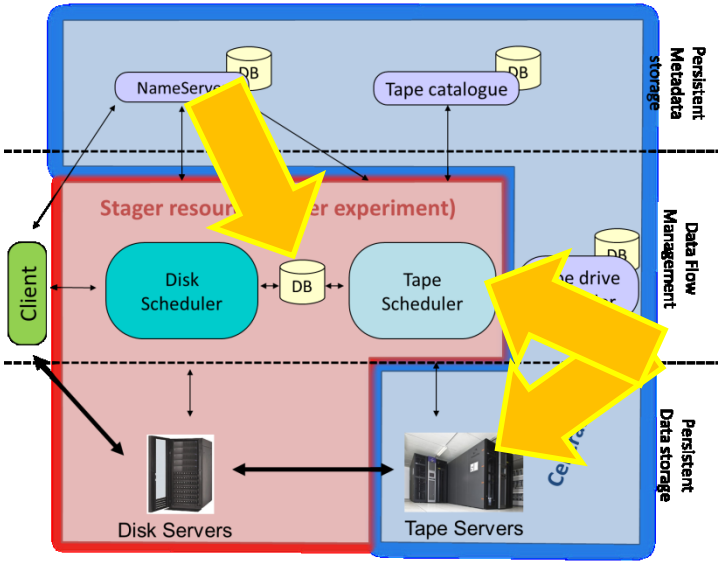
- Replacement for tape scheduler
  - Modular design
  - Simplified protocol to tape server
  - Reengineered routing logic



- Base for future development
  - Integration in the general CASTOR development framework

Future developments

- 1 synchronous tape mark per multiple files
  - Potential to reach hardware write speed
  - Performance less dependent on file size
  - Atomicity of transactions to be retained (data safety)
    - Bulk synchronous writes will lead to bulk name server updates



- Modularized and bulk interface between stager and tape gateway
  - Lower database load
  - Easier implementation of new requests

Conclusion

- CERN's CASTOR instances kept up with data rates of the heavy ion run
- New operational procedures improved reliability and reduced operational cost
- Moving to new tape generation in 2011
- Major improvements of software that will allow to tackle native hardware performance (100% efficiency)
  - and also improve the user-side experience
  - allow for shorter media migration
- New transfer manager and bulk DB usage will ensure long term scalability as usage grows