CASTOR development status and deployment experience at CERN

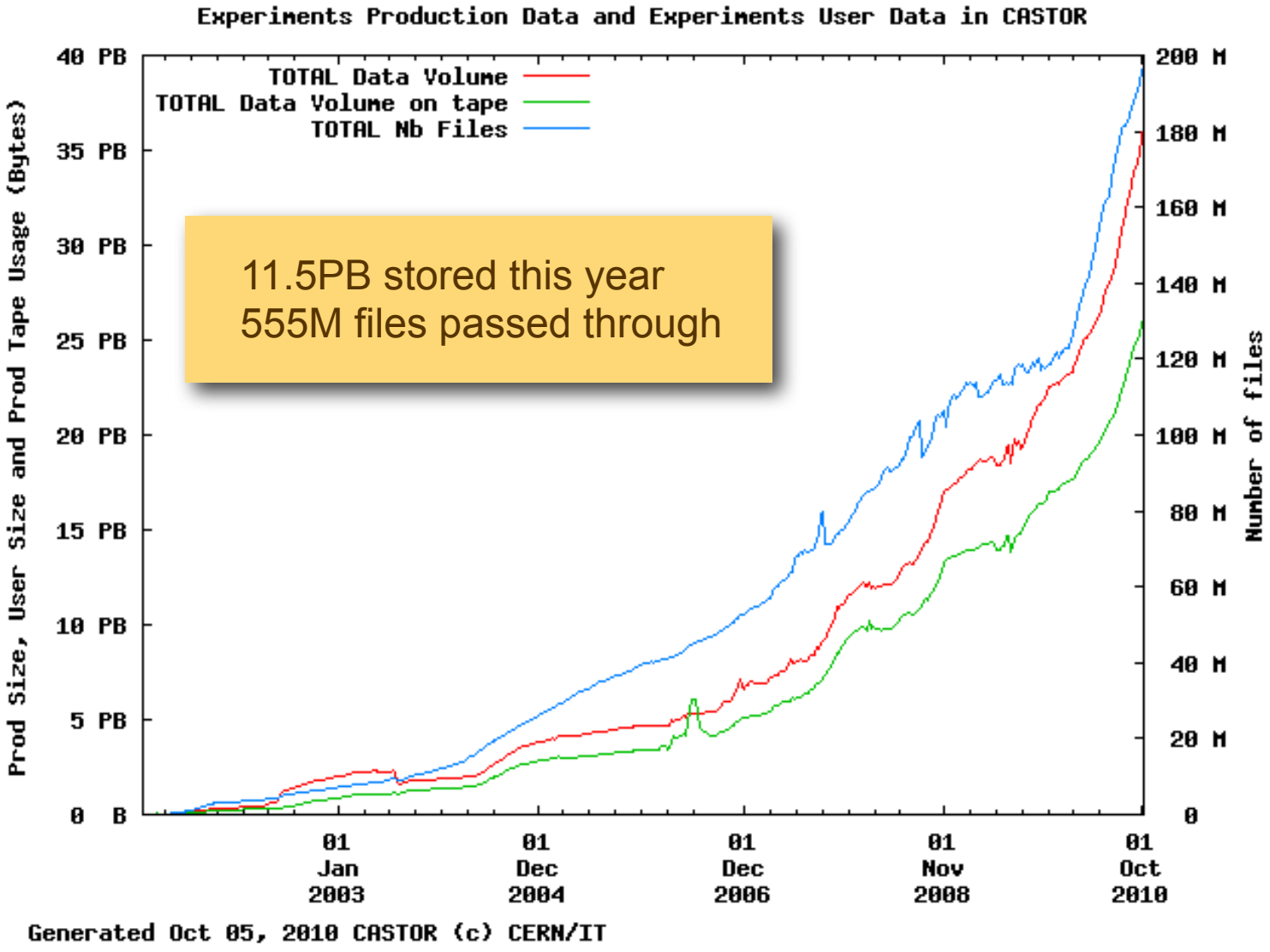
Łukasz Janyst
on behalf of the CERN IT-DSS group

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

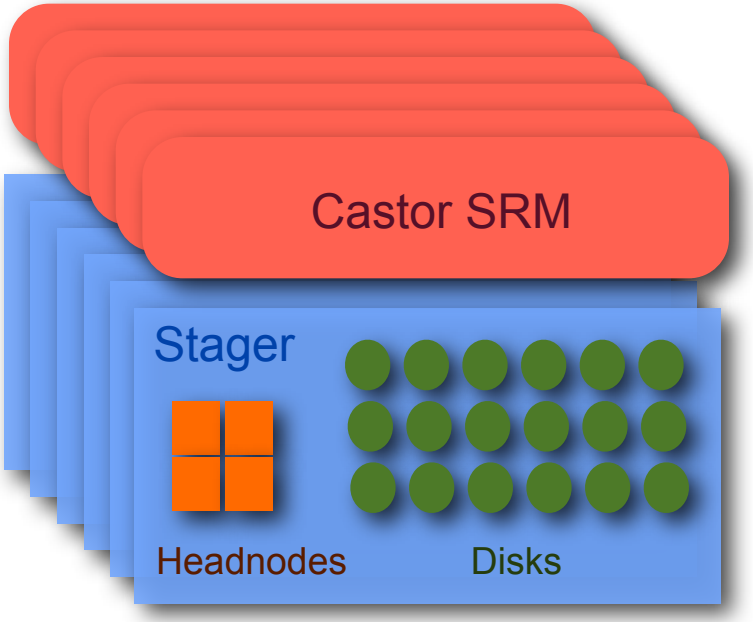
Outline

- CASTOR and the data it keeps now
- Recent improvements
- Performance during the data taking
- Plans for the future

Amount of data in CASTOR



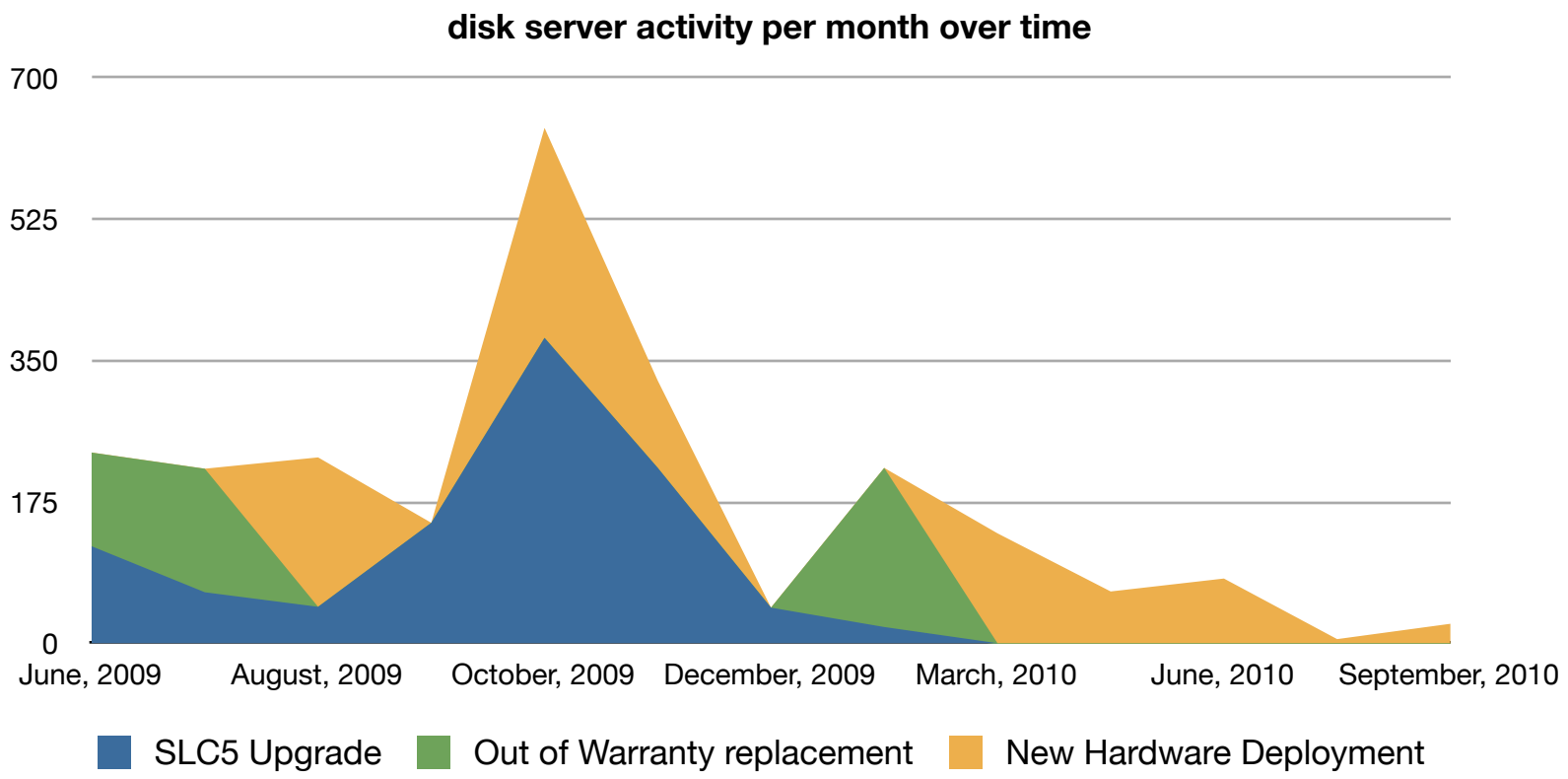
CASTOR setup at CERN



- 6 production stagers
- 1630 disk servers
- 30K disks
- 16.5PB disk space in RAID1

- 7 libraries
- 63K slots, 46K 1TB tapes
- 120 enterprise drives

Standard Changes - Servers



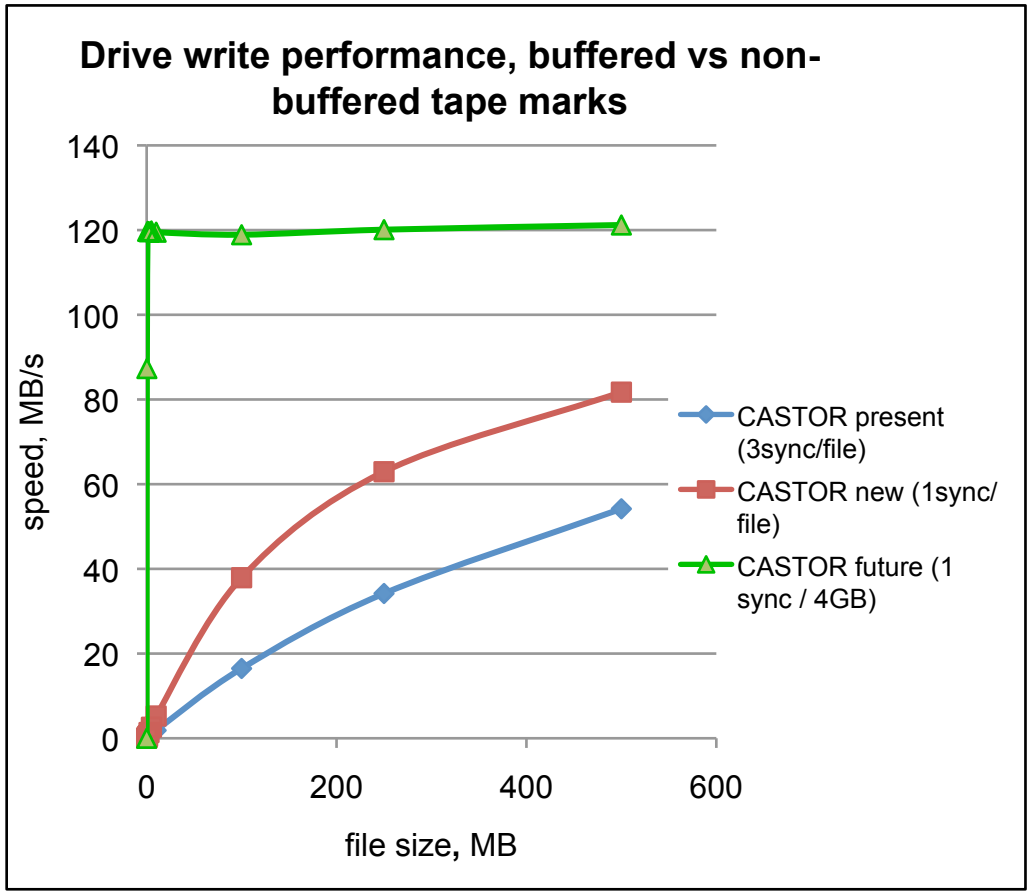
Improvements

- During the run of 2010 the following issues have been identified and addressed:
  - Tape marks
  - Tape (re)mounting
  - SRM bottlenecks
- General code maintenance and consolidation is an on-going process:
  - Source code analysis with Coverity
  - Enhanced test suite
  - Improved build system

Tape writing improvements

- Writing is well understood and aggregate write speed of 3-5 GB/s is not a problem
- However low per-drive performance
  - Disk cache IO and network contention
  - Per file tape format (ANSI AUL) - high overhead for writing small files (3 tape marks per file
- Solution: bulk data transfers using “buffered” tape marks
  - “Buffered” means no sync and tape stop - increased throughput and reduced tape wear
  - Part of the SCSI standard however not exposed by the Linux kernel driver
  - Worked with the tape driver maintainer to support them (to be released with Linux 2.6.37

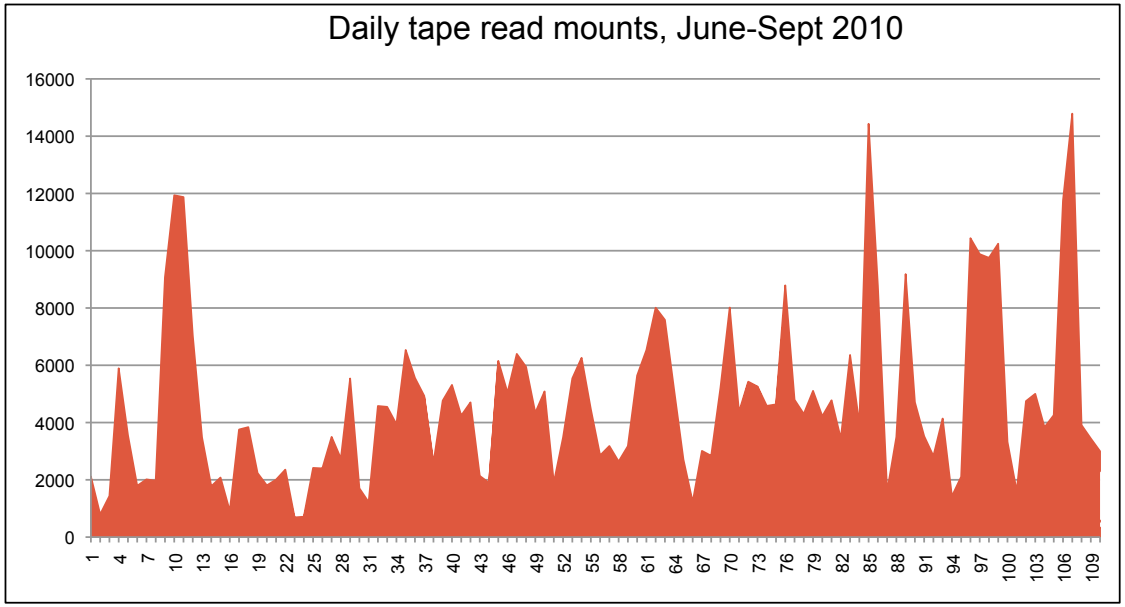
Tape writing improvements



- Released with new Castor tape module writing one synchronized tape mark instead of three
- Plan to reach native drive speed by writing synchronized tape marks every n GB

Tape reading

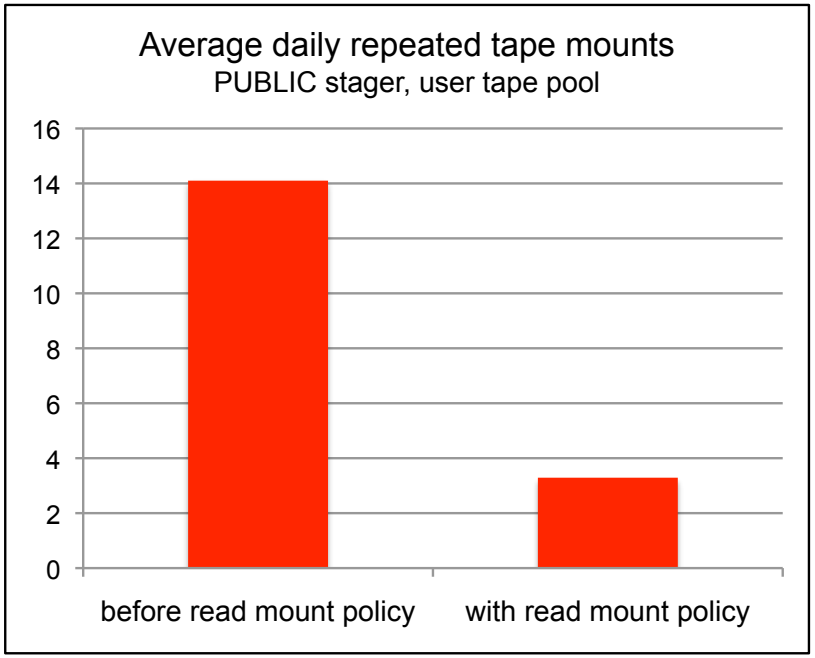
- CASTOR is a classic HSM. If the requested file is not on disk then recall it from tape ASAP
  - many tapes get remounted but average number of files read is very low (data sets spread over many tapes)
  - every mount is a wasted drive time (~2 minutes)
  - (Un)Mounting is the highest risk operation for tapes, drives and robotics and mechanical failure can affect access to large amount of media



Tape reading

- Current improvements:
  - Identify “heavy” users and educate them to use pre-staging
  - Verify adequateness and increase the size of disk caches (mount rates can be huge for even 1% of cache-misses)
  - Policy driven tape mounting: group requests and define thresholds for mounts (different for production and individual users)

- The future:
  - remove tape read rights from individual users
  - regular recalling of entire data sets by production managers
  - long term: increase tape storage granularity from files to data sets



SRM - recent improvements

- Activity ramped up with real data in Dec '09
- A change in the usage pattern unveiled a bottleneck in the system
  - Individual users were able to affect the production activity
  - Addressed with a re-engineering of the backend component
  - Deployed before the 2010 run
- Smooth operations since then
  - Order of 10 improvement in request rate w.r.t. 2009 activity

SRM - plans

- The CASTOR SRM interface will be object of only small developments in the future:
  - Support for VOMS roles
  - Improved support for aborting ongoing requests
- Some activity foreseen for the EMI project
  - In particular, moving from Globus GSI to pure SSL-based authentication

Data consistency checks

- Both the disk cache and the tape archive actively verify the data
- Opportunistic scanning when resources are available
- Verify correctness of size and checksums
- Rescan the all the stored data starting with the least recently accessed files (stager) or tapes (archive)
- Read back all newly written tapes

After 2 months ~5.5PB (5.5K tapes) have been verified.

Key Performance Indicators

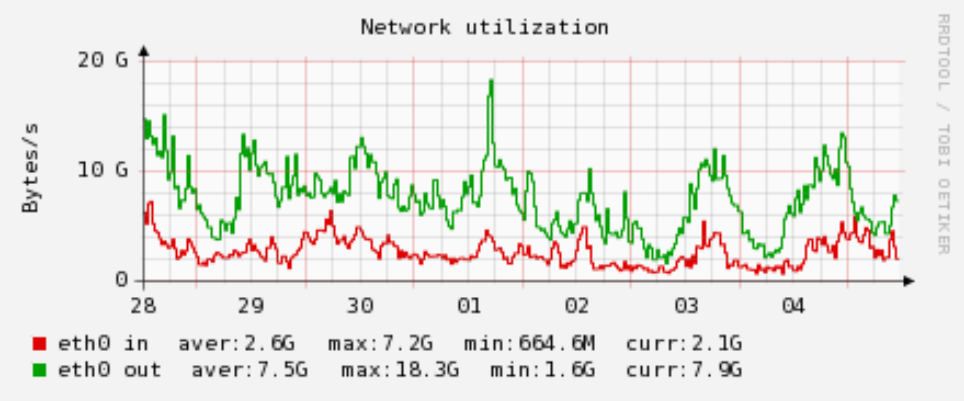
Current list of Key Performance Indicators:

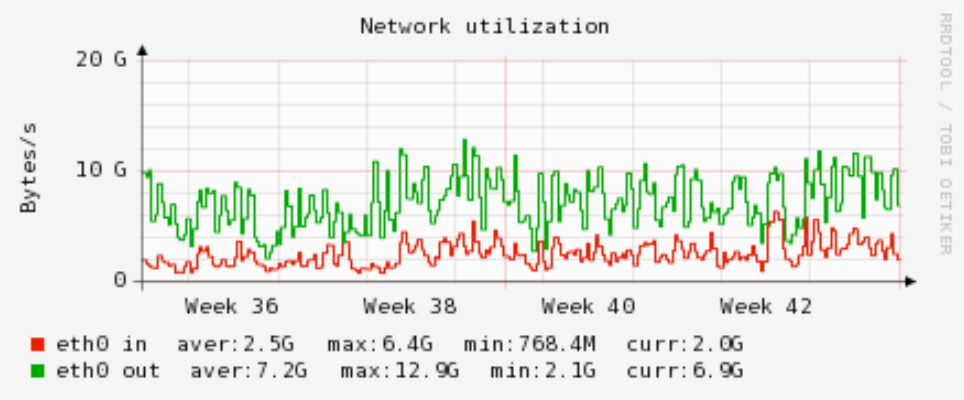
CASTOR Stager	Scope	Target
read/write latency (time to first byte (TTFB) for online files)	Transfer	TBD
time to migrate (time to tape)	Svcclass	4-8h cdr, 24h t1d0, t1d1 negotiated
time to recall file (TTFB for offline files)	User	TBD
number of parallel transfers	Svcclass	negotiated, depends on HW
number of queued transfers	Svcclass	negotiated, depends on HW
network throughput	Svcclass	negotiated, depends on HW
transfers start rate (open/s)	Svcclass	negotiated (max 20Hz per instance)
total space	Svcclass	negotiated, depends on HW
free space	Svcclass	min 5-10% on GCed classes

CASTOR Tape	Scope	Target
time waiting in the tape queue to write	VO	4h
time waiting in the queue to read	VO	4h
Number of days of free tape capacity at current consumption	All	60d

Each of these values is monitored in Lemon

Network traffic





Castor:

- Accepts data at average of 2.5 GB/s (peaks >7GB/s)
- Serves data at average of 7.2GB/s (peaks >18GB/s)

Recent Heavy Ion test

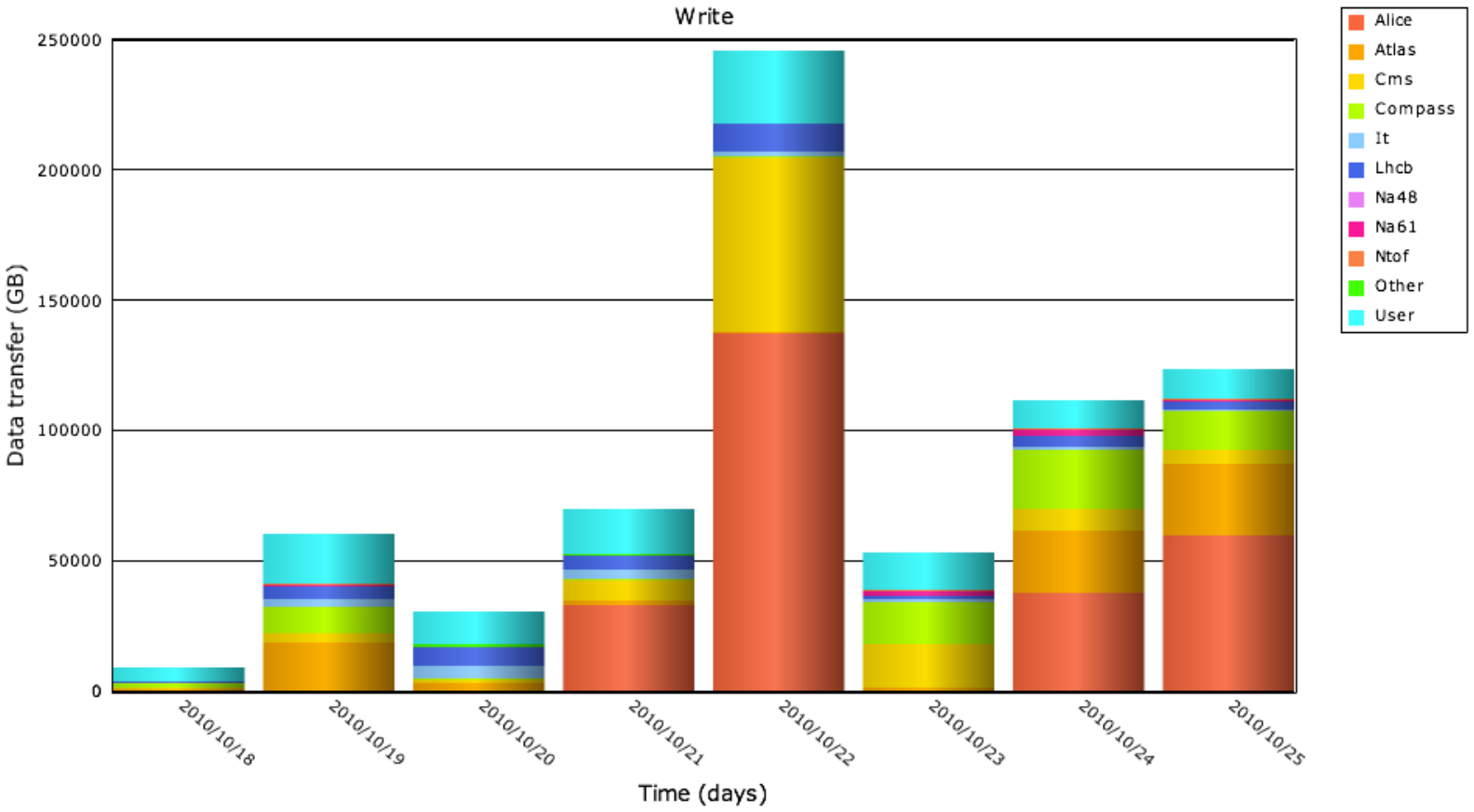
Started on Oct 22 and lasted (only) 24 hours (during the CHEP conference.)

- ALICE:
  - Test lasted 24 hours
  - Sustained data rate of around 2.5-3 GB/s, with peaks at 7 GB/s
  - Average file size of 3 GB
- CMS:
  - Test lasted only 6-7 hours
  - Data rate of 1.2 GB/s, with peaks at 2.5GB/s
  - Average file size of 30 GB

CASTOR has been able to handle successfully all the data without any indication of bottlenecks or scalability issues

A new record for the tapes

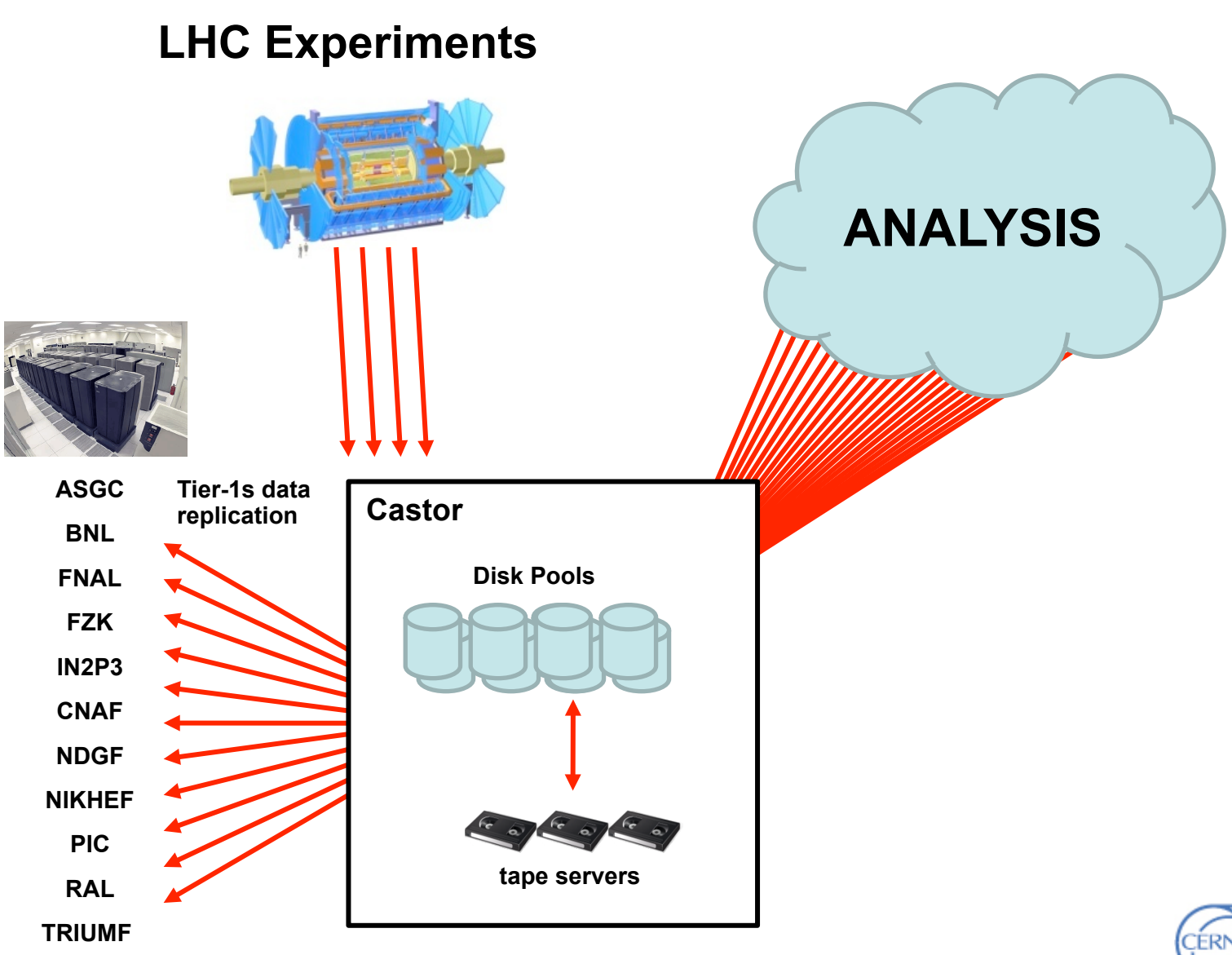
250 TB of data written in a single day!



Strategy

- Use CASTOR in the way it was designed to be used and what it is good at
  - Continuation of developments to improve tape efficiency
  - Restrict changes to maintenance only modifications to reduce risk on T0 operation
- Validate “simpler” alternate solutions to CASTOR disk pools to offer improved services necessary for analysis (EOS demonstrator)

Strategy



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

LHC Experiments

ASGC
BNL
FNAL
FZK
IN2P3
CNAF
NDGF
NIKHEF
PIC
RAL
TRIUMF

Tier-1s data replication

Castor

tape servers

Managed repli

Disk Pools

XROOT or mountable file system

ANALYSIS

Scalable, secure, accountable, globally accessible, manageable

Allow to choose service level for
- availability
- reliability
- performance

decoupled from HW

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

Conclusions

- CASTOR release 2.1.9 has performed well during 2010:
  - It has been able to handle successfully all the data sent to it without any indication of bottlenecks or scalability issues
  - Several years of stable operation ahead
- The EOS disk pool demonstrator is currently being tested with experiment users
  - Subject of another talk