The Data Storage Services (DSS) Strategy at CERN

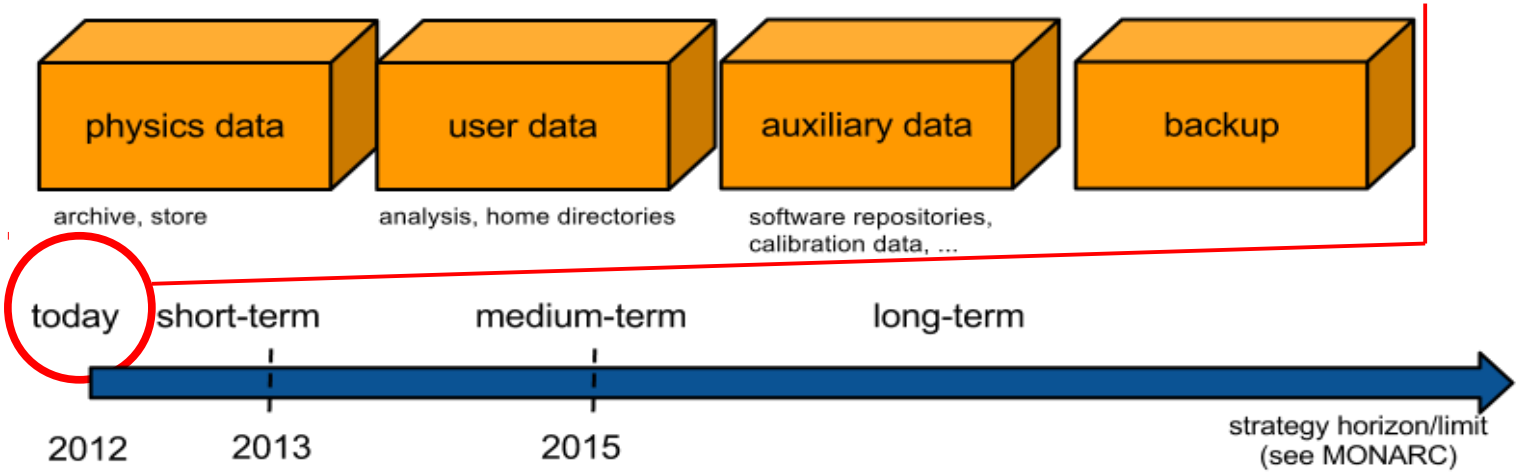
Jakub T. Moscicki

(Input from J. Iven, M. Lamanna
A. Pace, A. Peters and A. Wiebalck)

HEPiX Spring 2012 Workshop
Prague, April 2012

CERN IT Department
CH-1211 Genève 23
Switzerland

The big picture



LHC schedule

long shutdown

Users: experiments/WLCG, projects, individuals

analysis activity rises

requirements and use-patterns: evolving and unclear
rich feature set creates a lock-in situation

CERN/IT infrastructure

Service Improvement (ITIL), Business Continuity (new CC aka T0'),
Cloud Computing (Agile Infrastructure Project)

Technology and market

Hardware (solid state disks, LTO-5, storage appliances: e.g. NetApp)
Software: "simple" protocols (e.g. S3) and "simple" services (e.g. dropbox)

Situation today and short-/mid-term

CERN IT Department

AFS
- low-latency, general purpose file system
- 75TB, 1E9 files, 30kHz, 10ms
- daily backups (6 months)
- global namespace
- evolve and grow x 10
- provide large user workspace, possibly for end-user analysis

CASTOR
- long-term bulk storage
- 60PB (tape), 14PB (disk), 340M files, 200Hz, 1s..24h
- focus on T0, T0-T1 transfer
- focus on data integrity/durability
- become “the” archive for physics at CERN

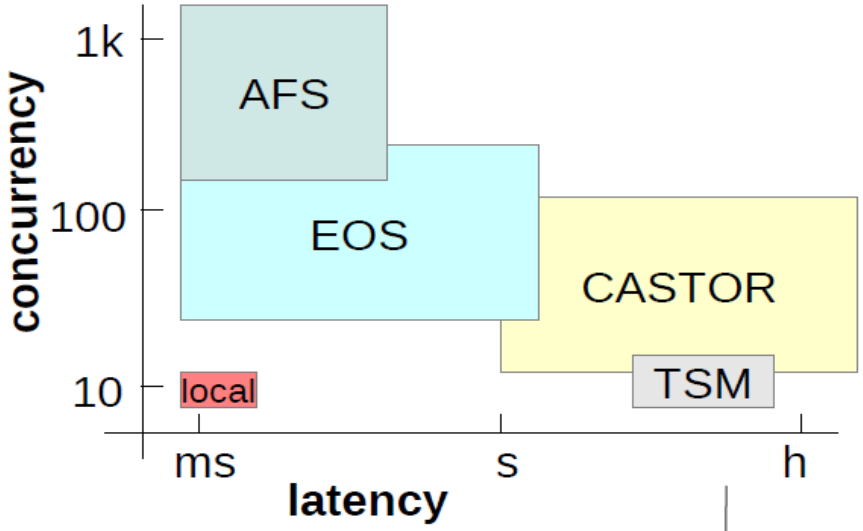
EOS
- low-latency storage for data analysis
- 9.3 PB (disk), 17M files 200Hz, 20ms
- data replication for QoS control
  - reliability, availibility, perf.
- fix, improve and scale up
- become “the” disk pool at CERN (storage for analysis facility)

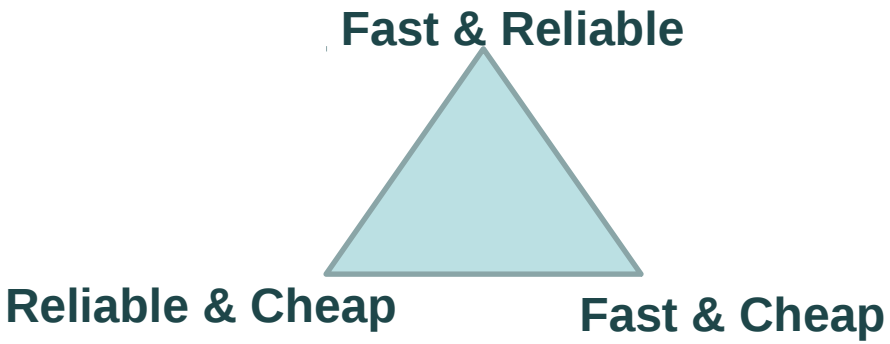
TSM
- high-latency, pure backup
- ~ 5PB (tape)
- extreme number of files (1E12)
- follow and adapt to tape technology

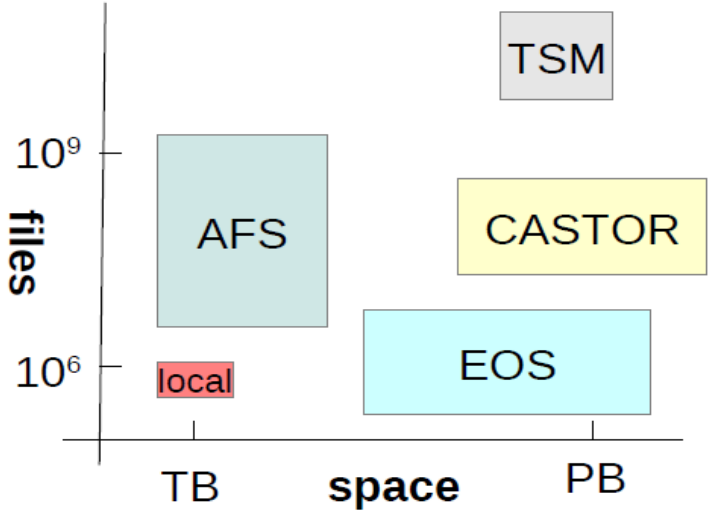
Other storage providers:
CERNVMFS, DBs NetApps, Experiments, ...

CERN IT Department
CH-1211 Genève 23
Switzerland

Storage Services at a glance





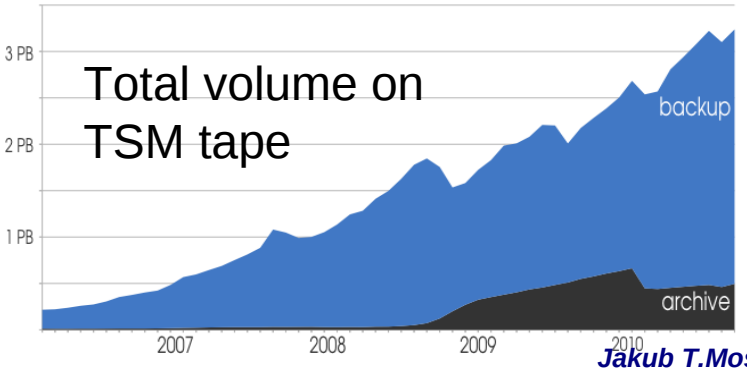


CERN IT Department
CH-1211 Genève 23
Switzerland

HW Technology

- Disk
  - JBOD, NLSAS storage (unexpensive) + SSD caches
    - scale up but maintain service quality
    - see next talk on AFS

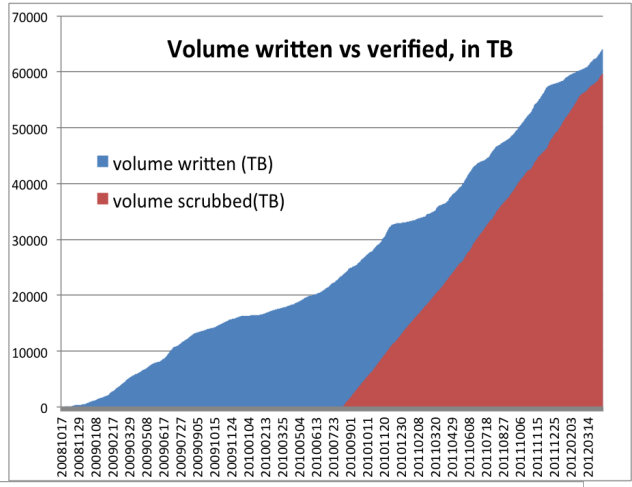
- Tape
  - Try multiple vendors (SpectraLogic)
    - → avoid/reduce vendor-lock in
  - lots of performance/storage headroom in the tape technology
    - LTO, faster drives, buffered tape marks, …
    - see next talks on TSM and Castor

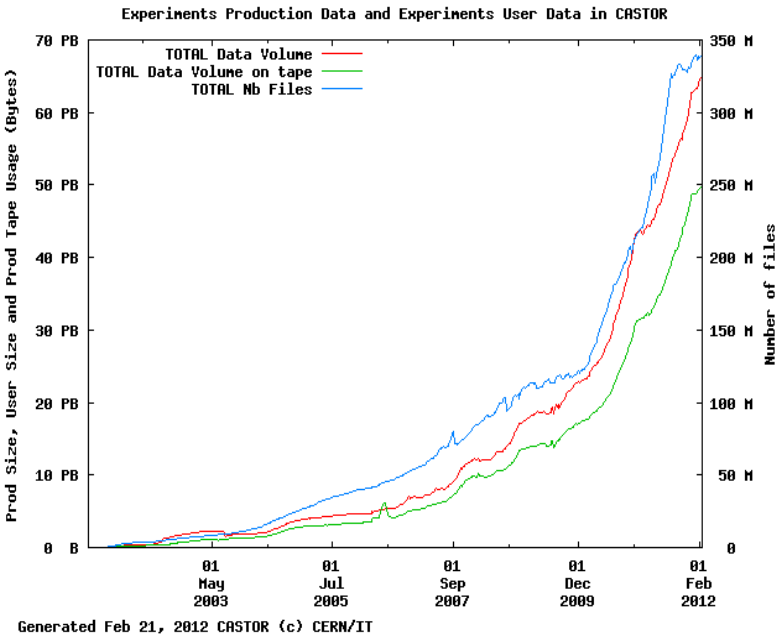


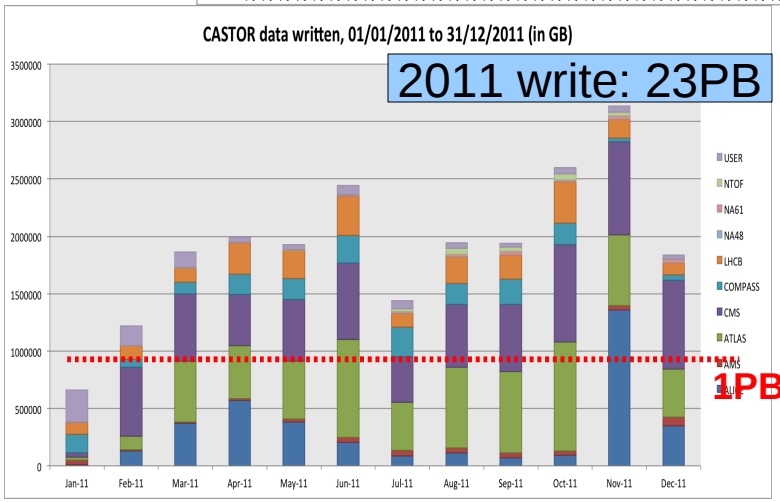
CERN IT
Department
CH-1211 Genève
23
Switzerland

CASTOR Tape

- Ensure we can read back the data
  - check all data at least once every 2 years
  - ~64PB verified so far
  - unrecoverable data: 37.2GB (62ppm)







60 PB of data, 208M files on tape

Avg file size 260 MB, 120 tape drives

Peak writing speed: 6.9 GiB/s (Heavy Ion run, 2011)

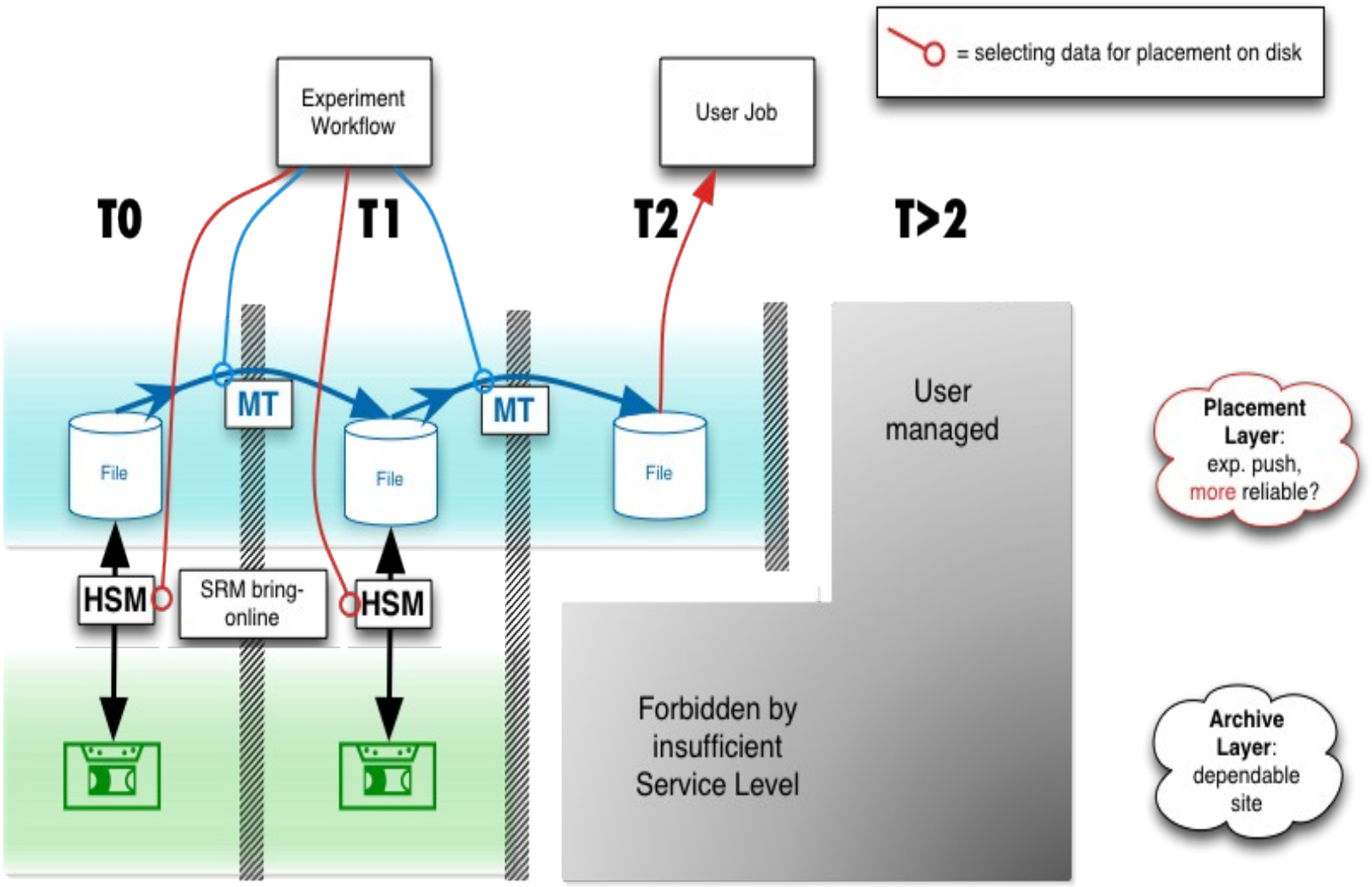
CERN IT Department
CH-1211 Genève 23
Switzerland

Areas of investigation

- Appliances (cloud storage)
  - Huawei (Openlab)
    - Reliable key-value storage
      - S3, Linux NBD, FUSE, ...
      - Test h/w: 768TB/384 nodes, 2.5 racks
    - reduce total cost of ownership?
      → just power-off disks, replace the whole storage unit after 3 years → intervention-less operation
  - SWIFT/OpenStack
    - “a drop-in replacement for S3”
      - private/public cloud federation on demand?
    - role of reduced-feature protocols/interfaces in physics?
      - S3: large-scale stress-tests with experiment apps
      - is it really needed / would it really work?

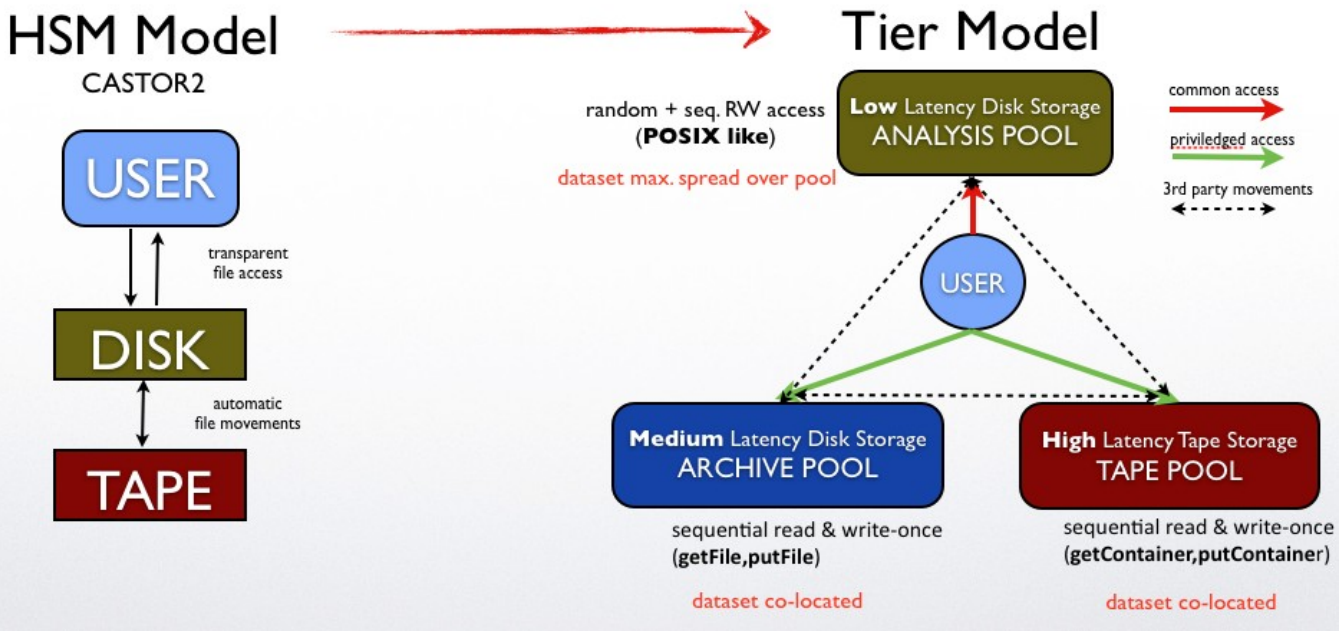
CERN IT Department
CH-1211 Genève 23
Switzerland

Physics storage model and...



CERN IT Department
CH-1211 Genève 23
Switzerland

...its evolution



- End of HSM

  → specialized storage pools, data placement managed by experiments frameworks

- Split: **Castor** → T0 TAPE
  EOS → Analysis Facility DISK

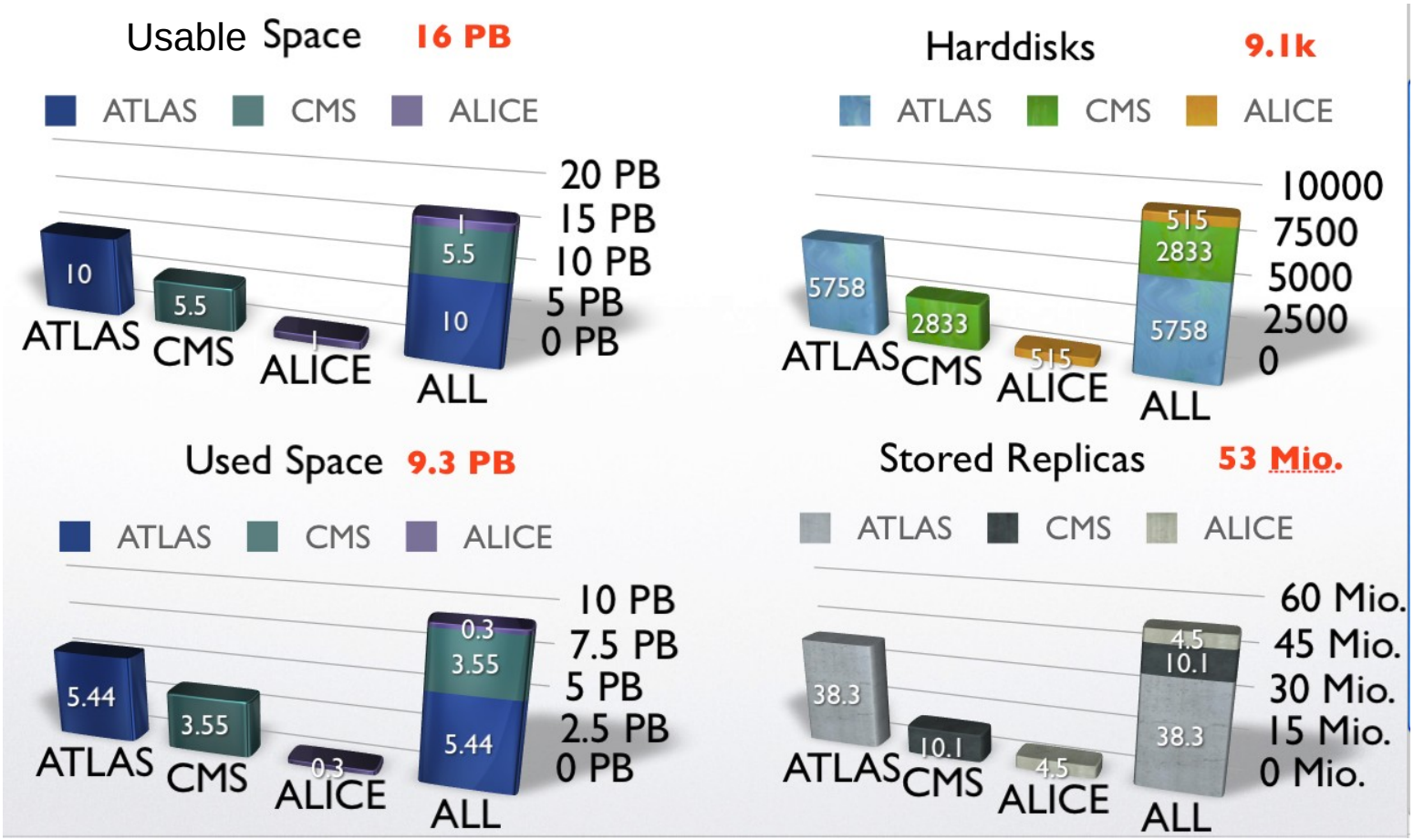
CERN IT Department
CH-1211 Genève 23
Switzerland

Analysis facility disk pool

- EOS is the key of our strategy for analysis
  - … although new AFS workspaces may also provide an alternative for end-user analysis
- EOS comes with features required for a very large Analysis Facility
  - integrated with a popular protocol in the experiments (xrootd)
  - fast metadata access (10x compared to Castor)
    - few ms read/write/open latency
  - designed for minimal operation cost and scalability
  - data replicated at the file level across independent storage units (JBODs)
    - automatic rebalancing of data replicas in case of hardware removal/addition → see example later
    - lost replicas are automatically recreated by the system, no loss of service by the client
    - operations simplified: power-off a disk or machine at any time without loss of service (service availability)
  - known issues: in-memory metadata server (not a problem for now, on a todo list)
  - EOS is freely available (GPL) but not packaged/supported off-the-shelf by us

CERN IT Department
CH-1211 Genève 23
Switzerland

EOS: Current Status



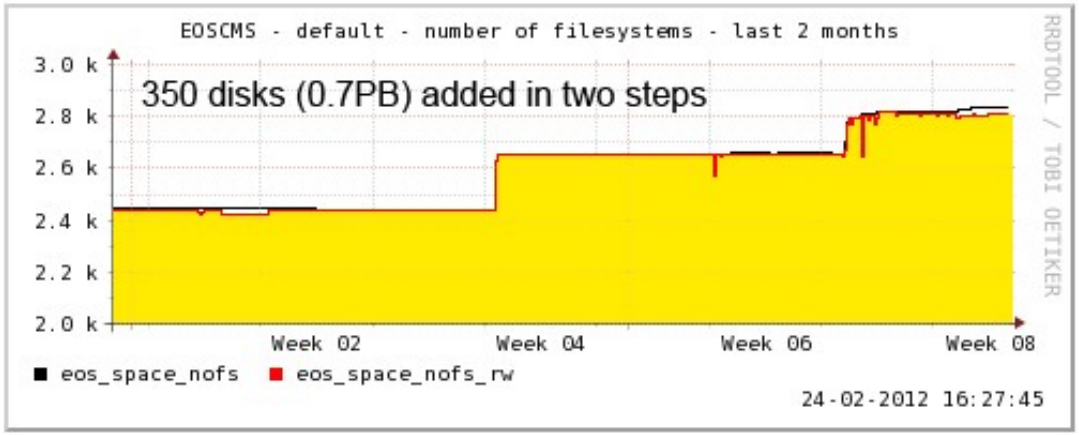
CERN IT Department
CH-1211 Genève 23
Switzerland

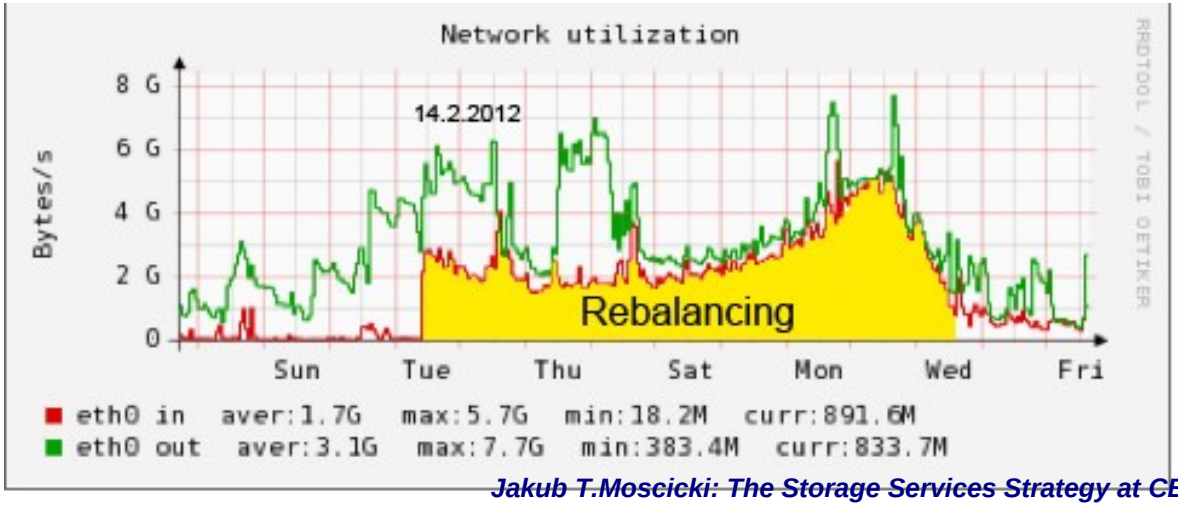
DSS

Rebalancing example

CERN IT Department

- February 2012: automatic rebalancing in EOSCMS. 1.6 PB added and rebalanced in 1 week internally (avg. ~2.7 GB/s), disk usage under 5% (~2.800 disks).





CERN IT Department
CH-1211 Genève 23
Switzerland

Future: overhead, reliability, performance

- Future: m+n block replication
  - for any set of “m” chunks chosen among the “m+n” you can reconstruct the missing “n” chunks
  - chunks stored on independent storage units (disk/servers)
  - different algorithms possible
    - double, triple parity, LDPC, Reed Solomon (space-optimal, at the theoretical limit), …
- This will possibly allow to define/adjust the quality of service parameters per file container (directory)
  - tune storage overhead vs reliability vs performance
    - simple replication: 3 copies, 200% overhead, 3 x streaming performance
    - 10+3 replicaton: can lose any 3, remaining 10 are enough to reconstruct, 30 % storage overhead
    - more possibilities … different QoS classes
- This will be phased in carefully into production
  - adding block-replicas to existing file-replicas first

CERN IT Department
CH-1211 Genève 23
Switzerland

Summary

- We expect all major systems to be there in 2015...
  … and much improved!
  - **AFS,CASTOR,EOS,TSM**
- Paradigm shift at CERN is a fact
  - analysis disk pools separated from the archive pools (no HSM)
- Evolution is constrained
  - many external drivers influence the strategy
  - large data volumes create inertia
  - rich features sets create inertia
- Cloud storage:
  - … we are ready
  - … no pressure from the users (yet?)
  - … price tag not there yet (?)

CERN IT Department
CH-1211 Genève 23
Switzerland