CERN

Michal Simon
on behalf of CERN IT/ST group

Storage Development at CERN

Outline

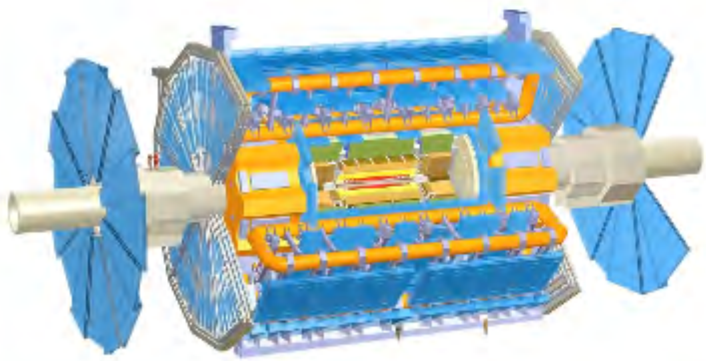
- Introduction: CERN
- Disk
- Tape
- Analytics
- Cloud / WLCG
- Summary

CERN: introduction

- An international laboratory situated between Geneva and the French Jura mountains
- The world's most powerful particle accelerator: LHC
- 4 very large detectors ('Experiments')
- Experiments register particle collisions at rates up to 40MHz (depending on the run and beam type)

CERN: the data

- Each detector is equipped with up to ~40M sensors -> PB/s!
- Reduced by online filtering farms that select few hundreds ‘good’ collisions per second



- Selected events are recorded on disk and tape at up to 10GB/s
- 50 Petabytes per year (today) for four experiments

CERN storage group

- Mandate:
  Ensures a coherent development and operation of storage services at CERN for all aspects of physics data.

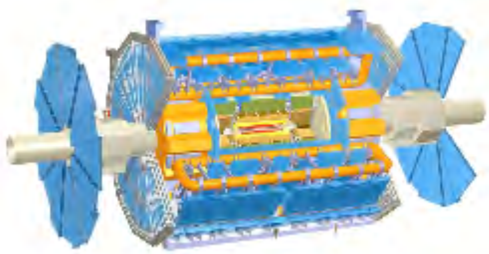
- Design and develop central storage services and their evolution.

CERN: data flow

Tape archive CASTOR
CERN Advanced STORage manager

5 – 10 GB/s

Local batch cluster
O(10^5) cores



LHC Detector

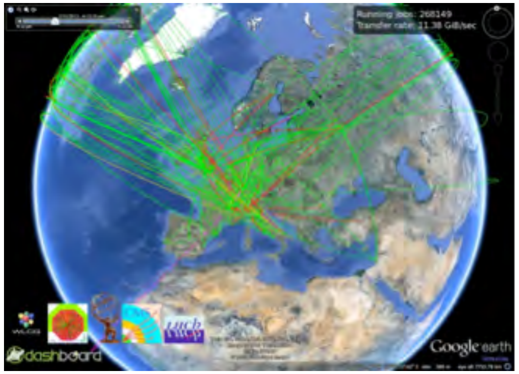
O(GB/s)

EOS
Open Source Storage

peak 80 GB/s

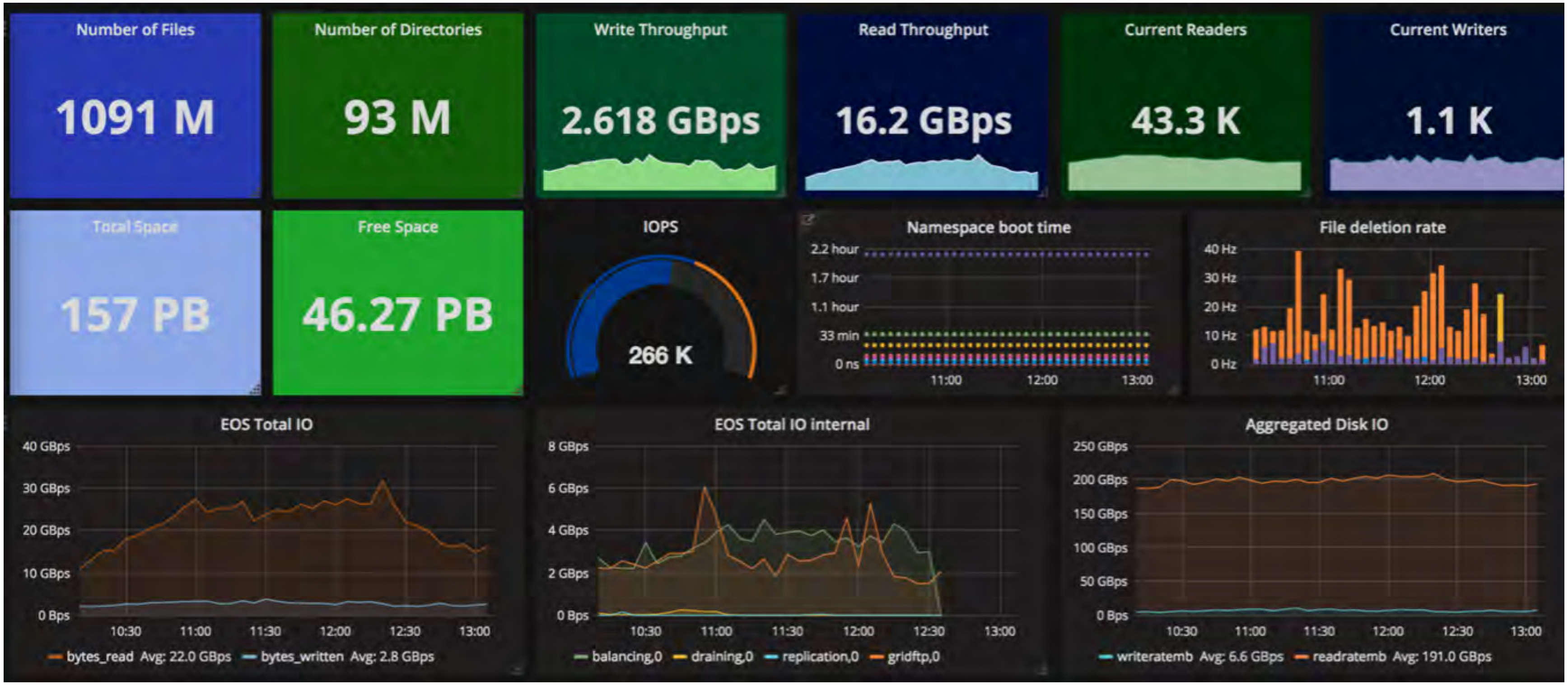
openstack™

5 – 10 GB/s



Data Export to WLCG

EOS statistics



EOS core: XRootD

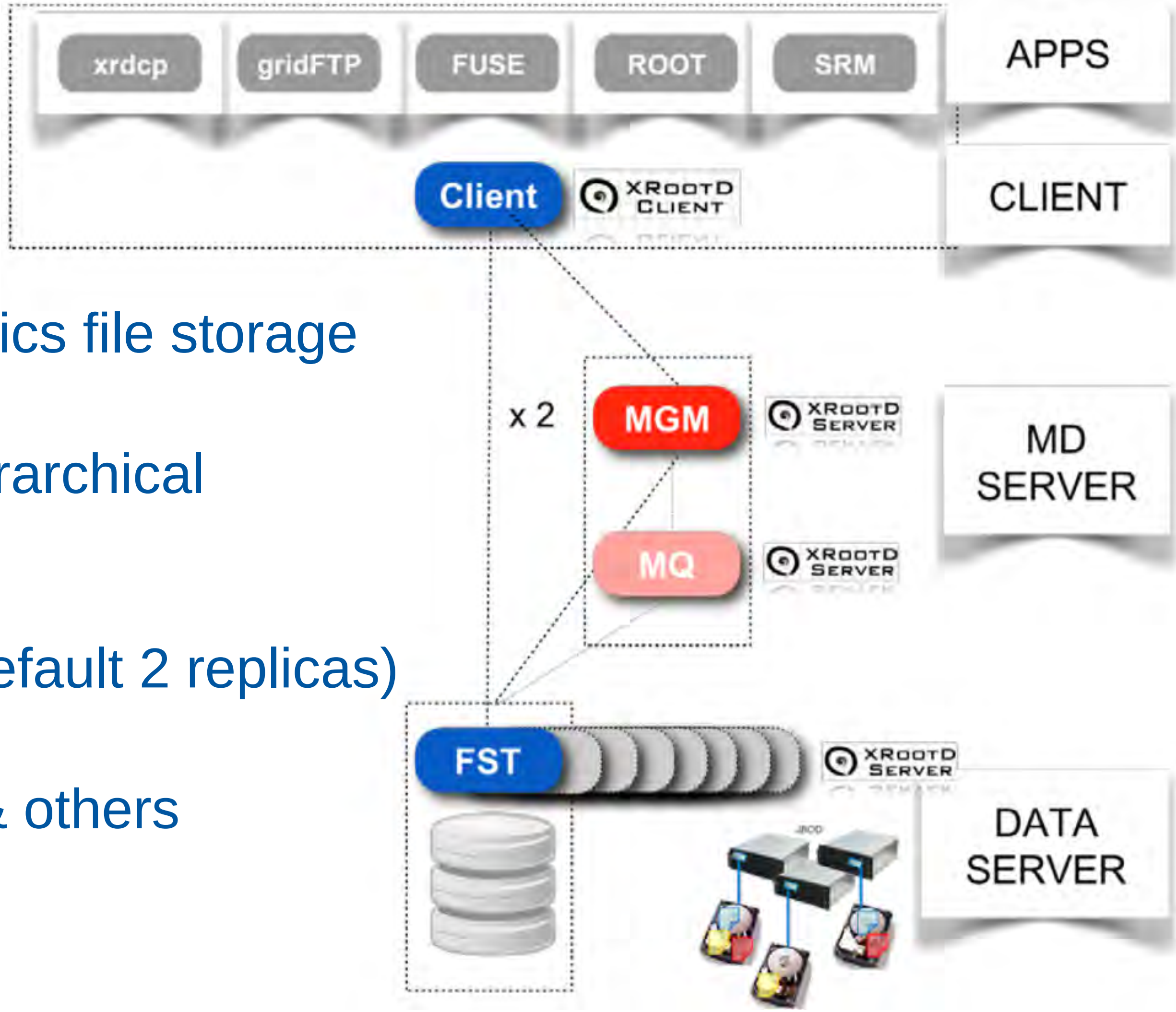
- XRootD originated at Stanford SLAC
- Collaboration members: SLAC, CERN, UCSD, JINR & Duke University
- The primary data access framework for the high-energy physics community
- Large Sky Synoptic Telescope (LSST)

EOS core: XRootD

- XRootD protocol designed for efficient remote file access in LAN/WAN
  - sync/async IO interfaces
  - 3rd party copy
  - storage clustering with hierarchical redirections
- Latest protocol enhancement: request signing

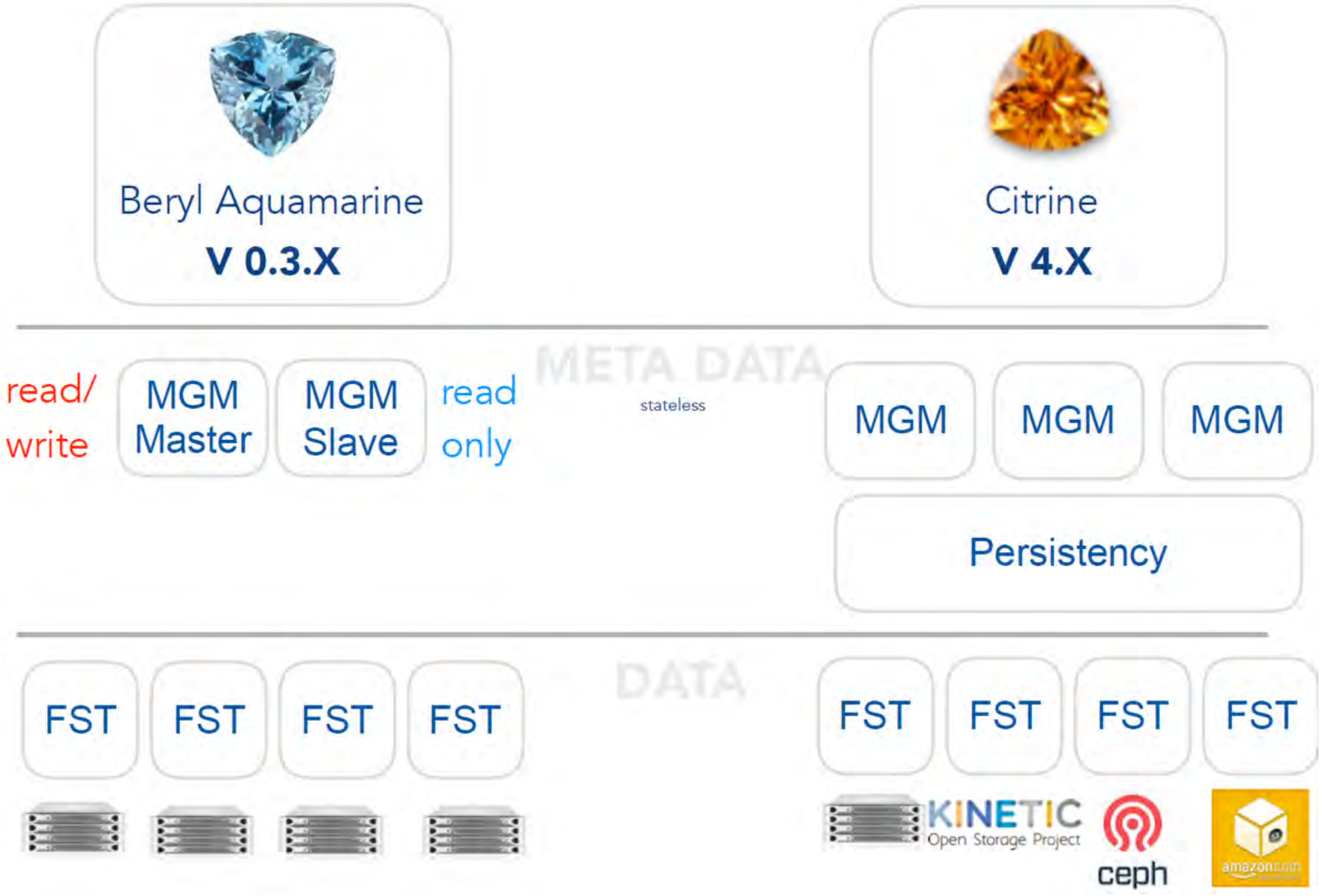
XRootD

EOS architecture



- Disk only physics file storage
- In memory hierarchical namespace
- File layouts (default 2 replicas)
- Physics data & others

EOS: architecture evolution



EOS namespace evolution

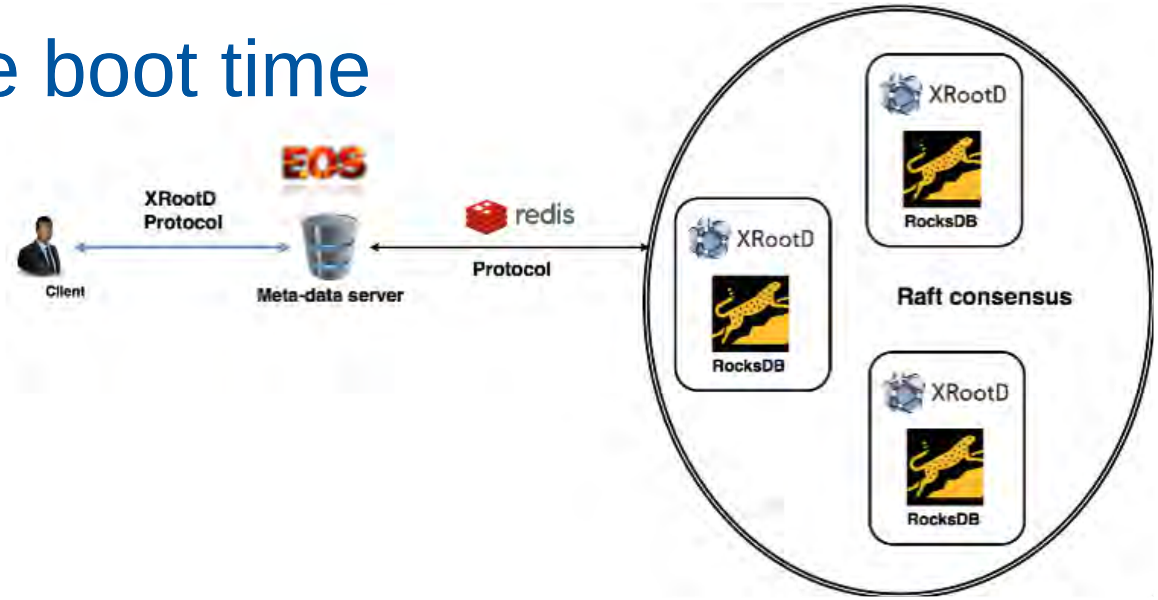
- Currently: C++ library used by the EOS MGM node
- Provides API for dealing with hierarchical collections of files

Pros	**Cons**
Using hashes all in memory (extremely fast)	For big instances it requires **a lot** of RAM
Every change is logged (low risk of data loss)	Booting the namespace from the change log takes long
Views rebuild at each boot (high consistency)	

EOS namespace evolution

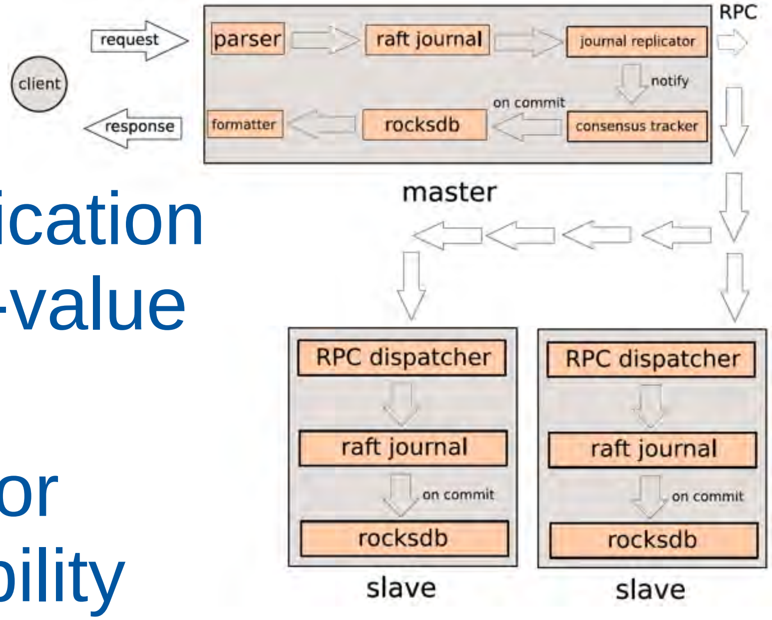
Goals:

- Still fast and consistent
- Scale-out solution to avoid one machine's memory limitation
- Reduce the boot time



EOS namespace evolution

- The new namespace persistence layer is code named QuarkDB
- RocksDB as the storage backend
- Translation of the communication protocol into RocksDB key-value transactions
- Raft consensus algorithm for replication and high-availability



EOS: interface evolution

2011

remote data store — XRootD

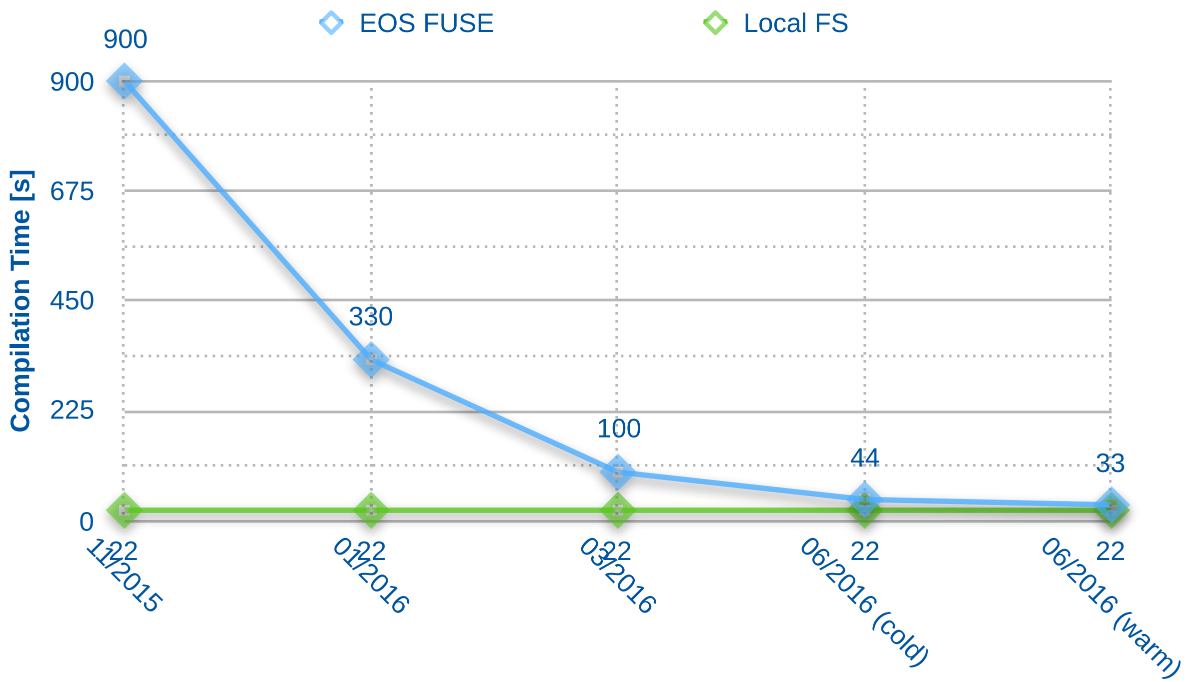
remote data store + file transactional storage — CERNBox

remote data store + file transactional storage + distributed filesystem behaviour — /eos

2017

EOS FUSE mount

- Current implementation (2nd generation):
  - Pure client side implementation
  - FUSE low level API
- Benchmark: (compilation)



EOS FUSE 3rd generation

- Motivation:
  - Limitations in consistency and performance
  - Help AFS retire gracefully

- Implementation
  - Dedicated server-side support
    - Async (bulk) communication, new locking model, file in-lining
  - Local meta-data & data caching

CERN Data Archive

Data:

- 190 PB physics data (CASTOR)
- ~7 PB backup (TSM)

Tape libraries:

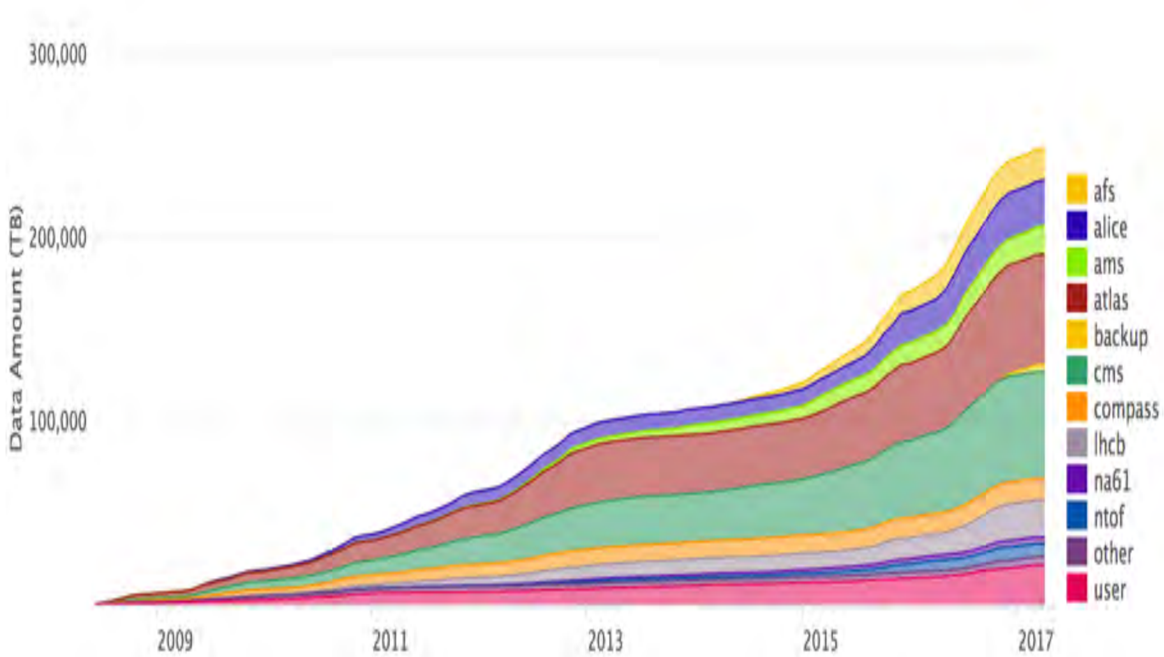
- IBM TS3500 (3+2)
- Oracle SL8500 (4)

Tape drives:

- ~90 archive
- ~55 backup

Capacity:

- ~70 000 slots
- ~25 000 tapes

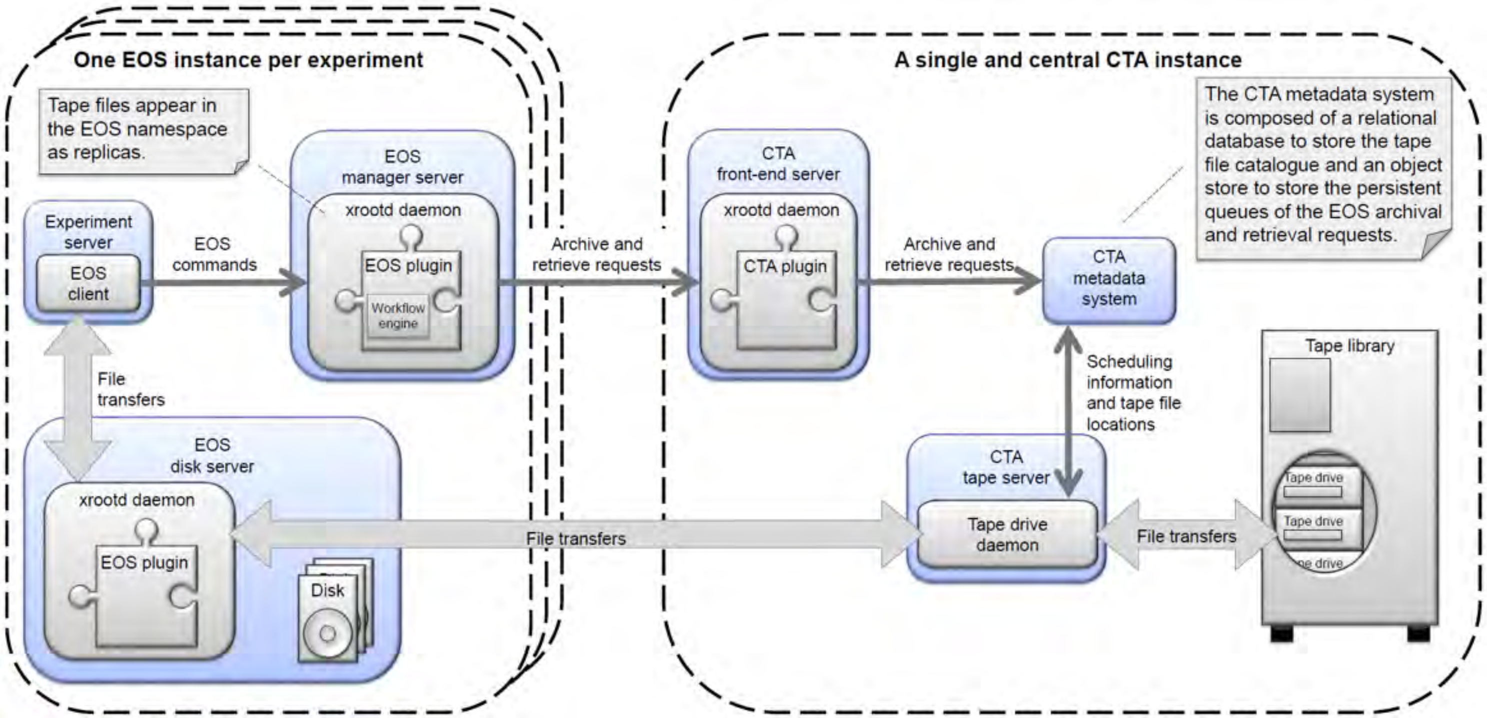


CERN Data Archive SW

- CASTOR (CERN Advanced STORage manager)
  - Hierarchical Storage management (HSM) system
    - Front-end disk and back-end tape layer
  - In production since 2001
  - Slowly being retired
- A new data archive solution (CTA) is being currently developed
  - Closely integrated with EOS
  - Designed to sustain Run3 (2021) expected data rate (150 PB per year)

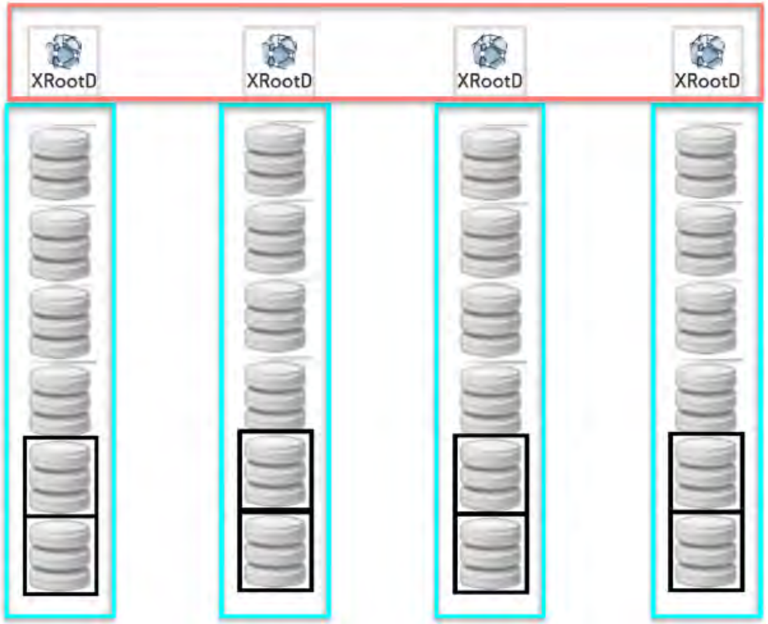
CERN Tape Archive (CTA)

- CTA is an EOS tape backend
- Archived files appear in the EOS namespace as file replicas



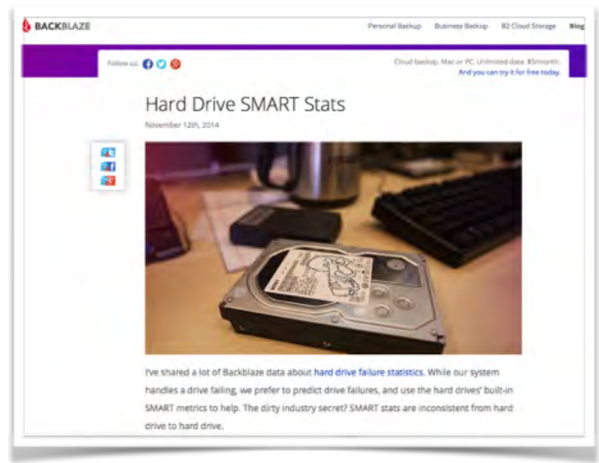
EOS: 2D erasure encoding

- Native XRootD plugin
- Dimension 1: EC over nodes
- Dimension 2: EC over many disks within a node
- Based on Intel ISAL library
- Possible alternative to tape archive



Analysis of Disk failures

- Failures on some 70k disks (order similar as blackblaze)
  - Failure impact on service performance
  - Comparison of enterprise and consumer disks
  - Predictive maintenance
- Using data from:
  - smart sensors
  - disk replacement logs
  - disk hardware repository
  - logs from EOS & Hadoop cluster

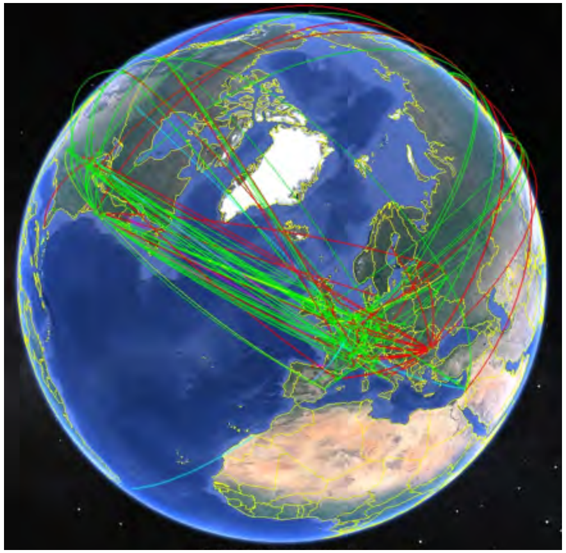


Cloud storage

- We are interested in tactical storage (or cache) to support CPU procurements
- We are not interested in long term cloud storage
- CERN is part of Helix-Nebula
- We are in the process of evaluating for which use cases cloud storage cloud be attractive

File Transfer Service

- Low-level transferring from 3 big experiments (LHCb, CMS, ATLAS)
  - Multi-level, fair-share transfer scheduler
  - Maximise resource usage & congestion avoidance
  - Multi-protocol support
  - Support for staging
- ~15 PB data transferred monthly



Summary

- XRootD: the framework of choice for our storage developments
- EOS: ~160 PB, ongoing development: namespace and FUSE mount
- Archive: ~200 PB, ongoing development: CTA
- Analytics: collaboration with UC Santa Cruz

Useful links

- XRootD: http://xrootd.org/
- EOS: https://eos.web.cern.ch/
- CASTOR: http://castor.web.cern.ch/
- FTS3: http://fts3-service.web.cern.ch/
- Helix-nebula: http://www.helix-nebula.eu/

Questions?