CERN

Storage at CERN

Hervé Rousseau — IT Department
hroussea@cern.ch

Table of Contents

Introduction

Storage for physics

Infrastructure storage

Wrap up

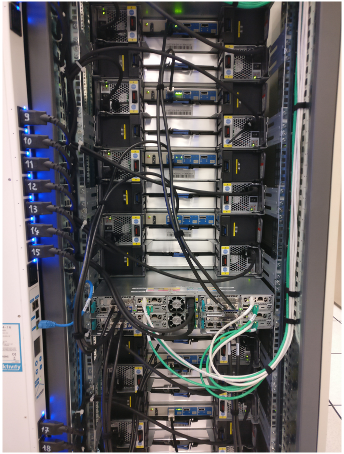
Unified building blocks

Storage node

- Compute node
- 10Gbit/s network interface
- SAS expander

Storage array

- Dummy SAS array
- 24x 6TB drives



Services Portfolio

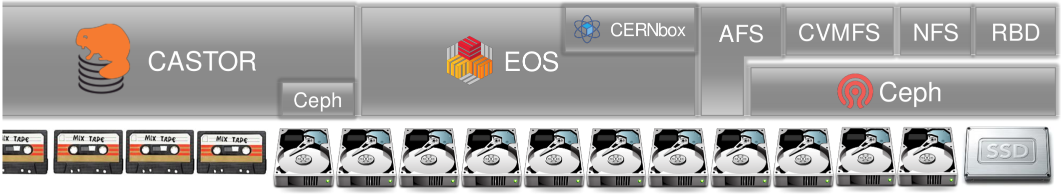


Table of Contents

Introduction

Storage for physics

Infrastructure storage

Wrap up

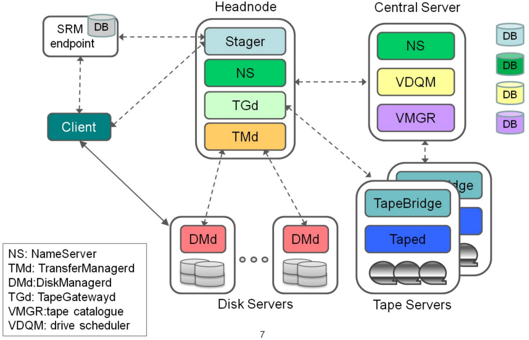
Castor

CASTOR
CERN Advanced STORage manager

Tape-backed storage system

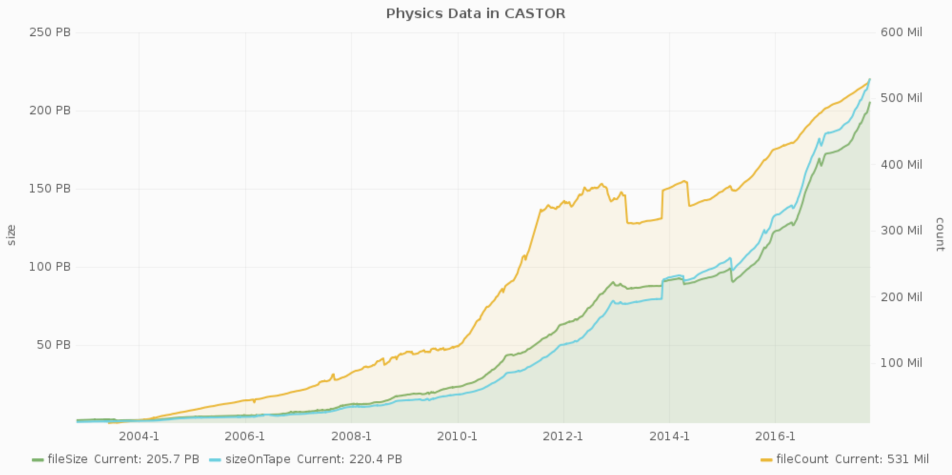
- Home-made HSM[a] system
- Users write data to disk
- Which gets migrated to tape

[a] Hierarchical storage management



NS: NameServer
TMd: TransferManagerd
DMd:DiskManagerd
TGd: TapeGatewayd
VMGR:tape catalogue
VDQM: drive scheduler

Castor

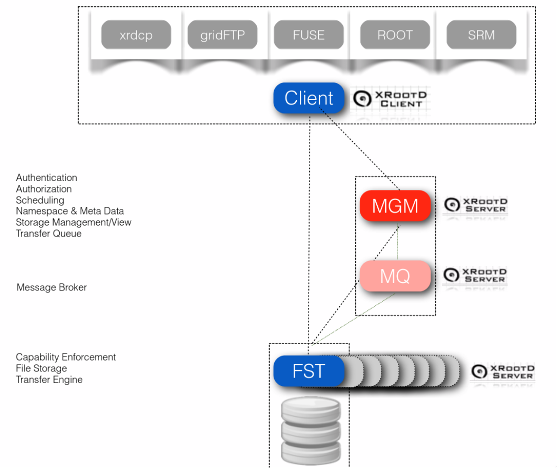

Physics Data in CASTOR
250 PB
200 PB
150 PB
100 PB
50 PB
size
600 Mil
500 Mil
400 Mil
300 Mil
200 Mil
100 Mil
count
2004-1
2006-1
2008-1
2010-1
2012-1
2014-1
2016-1
fileSize Current: 205.7 PB
sizeOnTape Current: 220.4 PB
fileCount Current: 531 Mil

EOS

Aggregated numbers

- ~ 1500 nodes
- ~ 55k drives
- ~ 220PB raw capacity

Spread over 6 instances



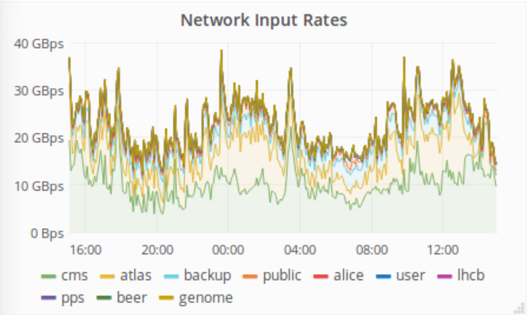
Authentication
Authorization
Scheduling
Namespace & Meta Data
Storage Management/View
Transfer Queue

Message Broker

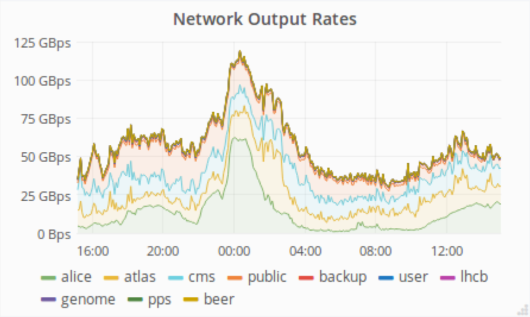
Capability Enforcement
File Storage
Transfer Engine

EOS

Write speed



Read speed



CERNBox — SWAN

CERNBox

- File sync and sharing
- Office tools integration
- Integration with ROOT[a]

[a] https://root.cern.ch

SWAN

- Jupyter based notebooks[a]
- Python, ROOT, R, Spark
- Nice CERNBox integration

[a] http://cern.ch/swan

Table of Contents

Introduction

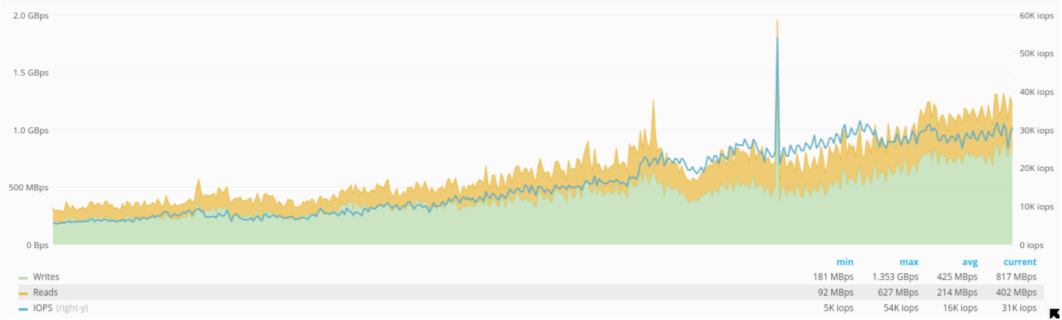
Storage for physics

Infrastructure storage

Wrap up

Ceph

- Openstack is Ceph's killer app: 4x usage in 2 years
- Not a single byte lost or corrupted

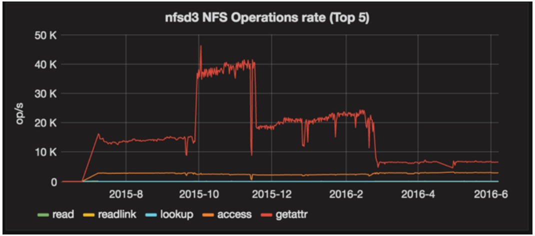


	min	max	avg	current
Writes	181 MBps	1.353 GBps	425 MBps	817 MBps
Reads	92 MBps	627 MBps	214 MBps	402 MBps
IOPS (right-y)	5K iops	54K iops	16K iops	31K iops

Ceph: NFS on RBD

Replace NetApps with VMs

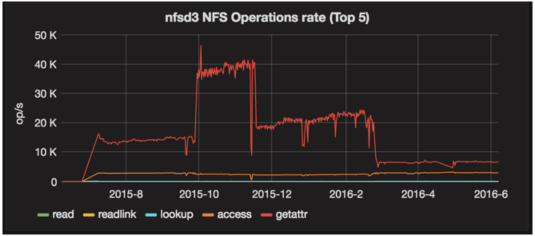
- ∼ 60TB across 30 servers
- Openstack VM + RBD vol.
- CentOS7 with ZFS
- Not highly-available, but. . .
- Cheap (thin-provisioning)
- Resizable



Ceph: NFS on RBD

Replace NetApps with VMs

- ∼ 60TB across 30 servers
- Openstack VM + RBD vol.
- CentOS7 with ZFS
- Not highly-available, but. . .
- Cheap (thin-provisioning)
- Resizable



puppet labs
GitLab
collaborate with TWiki
Platform Computing an IBM Company
BOINC
OPENSHIFT
elastic

Moving to Manila+CephFS very soon

CephFS for HPC

CERN is mostly a HTC lab

- Parallel workload, quite tolerant to relaxed consistency
- HPC corners in the Lab
  - Beams, Plasma simulations
  - Computation Fluid Dynamics
  - Quantum ChromoDynamics
- Require fill POSIX, read-after-write consistency, parallel IO

- $\sim$ 100 nodes HPC cluster accessing $\sim$ 1PB CephFS

Ceph: Scale testing

Bigbang scale tests mutually benefitting CERN Ceph

Bigbang I: 30PB, 7200 OSDs, Ceph `Hammer`

Found several `osdmap` limitations

Bigbang II: Similar size, Ceph `Jewel`

Scalability limited by OSD-MON traffic.

Lead to development of `ceph-mgr`

Bigbang III: 65PB, 10800 OSDs, Ceph `Luminous`

No major issue found

Table of Contents

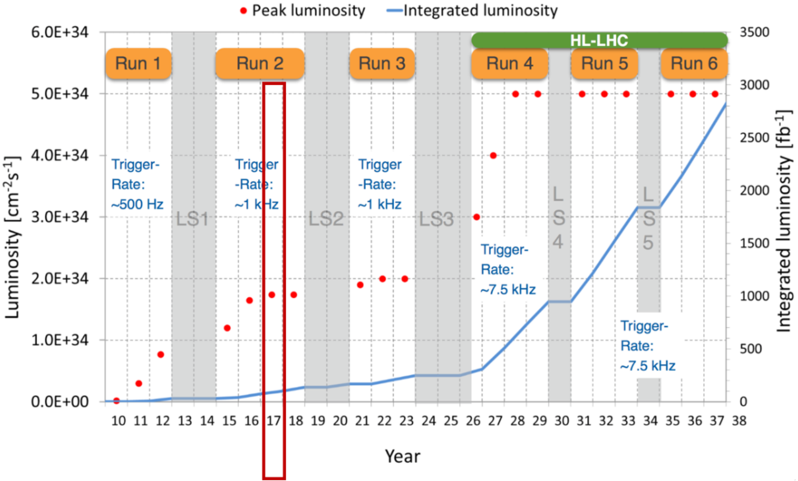
Introduction

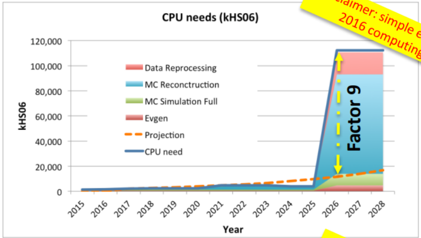
Storage for physics

Infrastructure storage

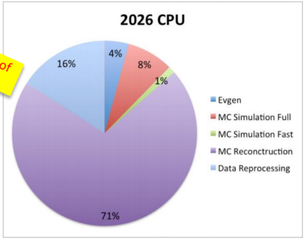
Wrap up

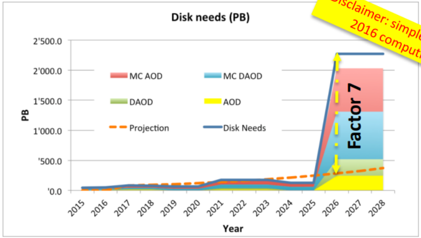
Next challenges



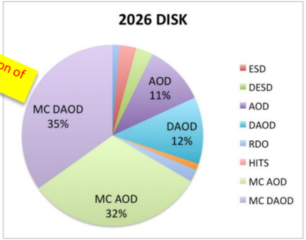


Disclaimer: simple extrapolation of 2016 computing model !





Disclaimer: simple extrapolation of 2016 computing model !



Wrap up

- Homegrown storage systems, augmented by open source
- “Data deluge” forecasted for 2026
- CentOS is powering a huge part of our services

www.cern.ch

References

- A. Peters: Present Future Solution for dta storage at CERN
- D. van der Ster: Building Scale-Out Storage Infrastructures with RADOS and Ceph
- S. Campana: The ATLAS Computing Challenge for HL-LHC