Disk Storage at CERN

Presenter:
Luca Mascetti

Co-authors:
Belinda Chan, Xavier Espinal, Alessandro Fiorot, Hugo Gonzalez Labrador, Jan Iven, Massimo Lamanna, Giuseppe Lo Presti, Jakub Moscicki, Sebastien Ponce, Herve Rousseau, Dan van der Ster

CERN - IT/DSS
CERN Disk Storage Overview

Our goal

provide to experiments and users
storage solution and tools for
data management and data analysis,
and operate them...
# CERN Disk Storage Overview

<table>
<thead>
<tr>
<th></th>
<th>AFS</th>
<th>CASTOR</th>
<th>EOS</th>
<th>Ceph</th>
<th>NFS</th>
<th>CERNBox</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Raw Capacity</strong></td>
<td>3 PB</td>
<td>20 PB</td>
<td>140 PB</td>
<td>4 PB</td>
<td>200 TB</td>
<td>1.1 PB</td>
</tr>
<tr>
<td><strong>Data Stored</strong></td>
<td>390 TB</td>
<td>86 PB (tape)</td>
<td>27 PB</td>
<td>170 TB</td>
<td>36 TB</td>
<td>35 TB</td>
</tr>
<tr>
<td><strong>Files Stored</strong></td>
<td>2.7 B</td>
<td>300 M</td>
<td>284 M</td>
<td>77 M (obj)</td>
<td>120 M</td>
<td>14 M</td>
</tr>
</tbody>
</table>

AFS is CERN’s linux home directory service

CASTOR & EOS are mainly used for the physics use case (Data Analysis and DAQ)

Ceph is our storage backend for images and volumes in OpenStack

NFS is mainly used by engineering application

CERNBox is our file synchronisation service based on OwnCloud+EOS
Born in 1999
Common Namespace and **Database** Centric
Main Role: **Long Term Data Archive**
  Tape backed
Focus on Tape Performance
Protected Transfers from DAQ (Slot-based)
Secure Access with KRB5 and X509

6 CASTOR instances
  ALICE, ATLAS, CMS, LHCb, Public, Repack

Latest Release (v15)
  New Tapeserver daemon
  xroot as main protocol (Obsolete rfio)
Several stager performance improvement
Introduction of Data Pools (based on Ceph)
Getting ready for **LHC run2**

- complete removal of end-users access from our tape system (LHC exp.)
- decommission of disk-only pools (moved to EOS)
  - moving towards an archive system
- some “spring cleaning” is mandatory before the start of LHC
  - ATLAS ESD deletion campaign (16PB and 11M files)

**AMS “stressing” our tape migration queue**

Peak of 270TB waiting to be migrated
Born in 2010
Hierarchical **in-memory** namespace focus on very **low latency**
Main Role: **Disk-only** storage optimised for concurrency
Quota System for users & groups with secure auth (KRB5, X509)

6 EOS instances
   ALICE, ATLAS, CMS, LHCb, Public, **User**

New Functionality
   Location Awareness & GEO Scheduling
   Archive & Backup tools

![EOS Raw Disk Capacity Deployed graph](image)
The latest hardware delivery (Mar 2015) balanced the capacity installed in the 2 computer centres (~50% ~50%)

Experiments replicas are not yet spread equally between the 2 geolocation

Geo-balancing need to be activated and tuned to avoid filling the network links to Wigner

EOS is now optimised for managing efficiently data in different computer centre providing to our user a single site view

And in the future it will be possible to specify adhoc scheduling policy based on the namespace location
LHC Data Taking (Run2)
LHC Data Taking (Run2)
Commodity Hardware for All

- Uniform storage for all our services (same hardware)
  - AFS, CASTOR, EOS, Ceph
- System Unit:
  - 8 physical cores (16 virtual)
  - 64GB of RAM
  - 2x 2TB HDDs
  - 2x disk-tray of 24x 4TB HDDs
  - almost 200 TB building block unit
Coupling Storage Solution

Ceph virtual volumes (RBD) for AFS & NFS backend

- Expose storage volume via VM gateways
- Exploit ZFS features for snapshots and backups
- Disaster recovery server in Wigner

Prototype Virtualised AFS Server

Virtualised NFS Server

Virtualised NFS Server

Ceph Cluster (Meyrin)

Ceph Cluster (Wigner)
CASTOR version 15 introduced the concept of DataPool
- built on top of Ceph RADOS
- striping physics files into Ceph using “disk servers" as proxies
- enhance single tape stream performance (up to 500 MB/s)
- Tape server speed around 300 MB/s
- released as part of CASTOR v15 but deployment targeted after LHC RUN2
Coupling Storage Solution

CERNBox

- Fusion between OwnCloud and EOS
  - more details in my other talk: OwnCloud+EOS End-user storage for Science
Future

LHC RUN2!!!!

Exploit Wigner also for disaster recovery

Master strategic storage technologies