CERN

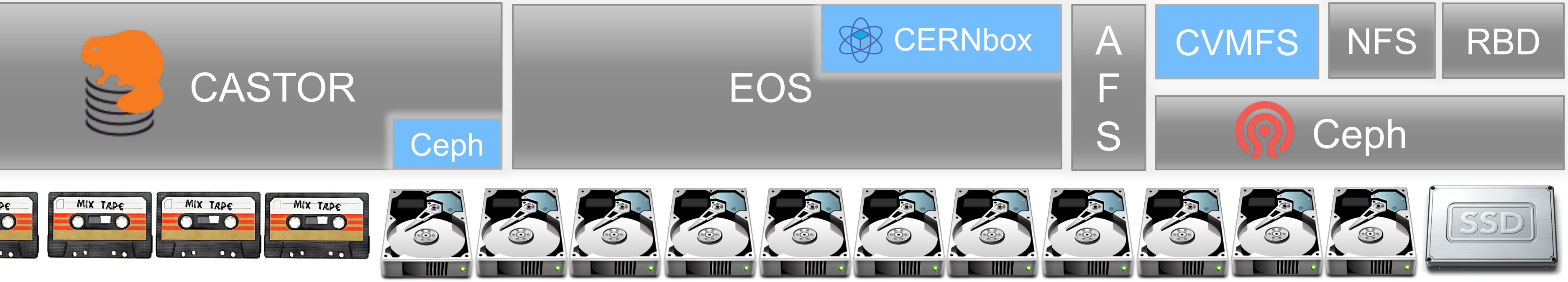
Ceph @ CERN: Status and Plans

Dan van der Ster, CERN IT Storage Group
daniel.vanderster@cern.ch

Red Hat @ CERN
17 January 2017

Storage for CERN and Particle Physics

- Huge data requirements (150PB now, +50PB/year in future)
- Worldwide LHC Grid *standards* for accessing and moving data
  - GridFTP, Xrootd to access data, FTS to move data, SRM to manage data



- Shrinking AFS, growing EOS (plus added CERNbox for sync)
- Ceph for the cloud and starting physics data and HPC

🟦 NEW KIDS ON THE BLOCK

Why Ceph?

- RBD Block Storage: an essential component of the private cloud (OpenStack *Images* and *Volumes*)
- RADOS: a reliable storage layer upon which we build high level (storage) services

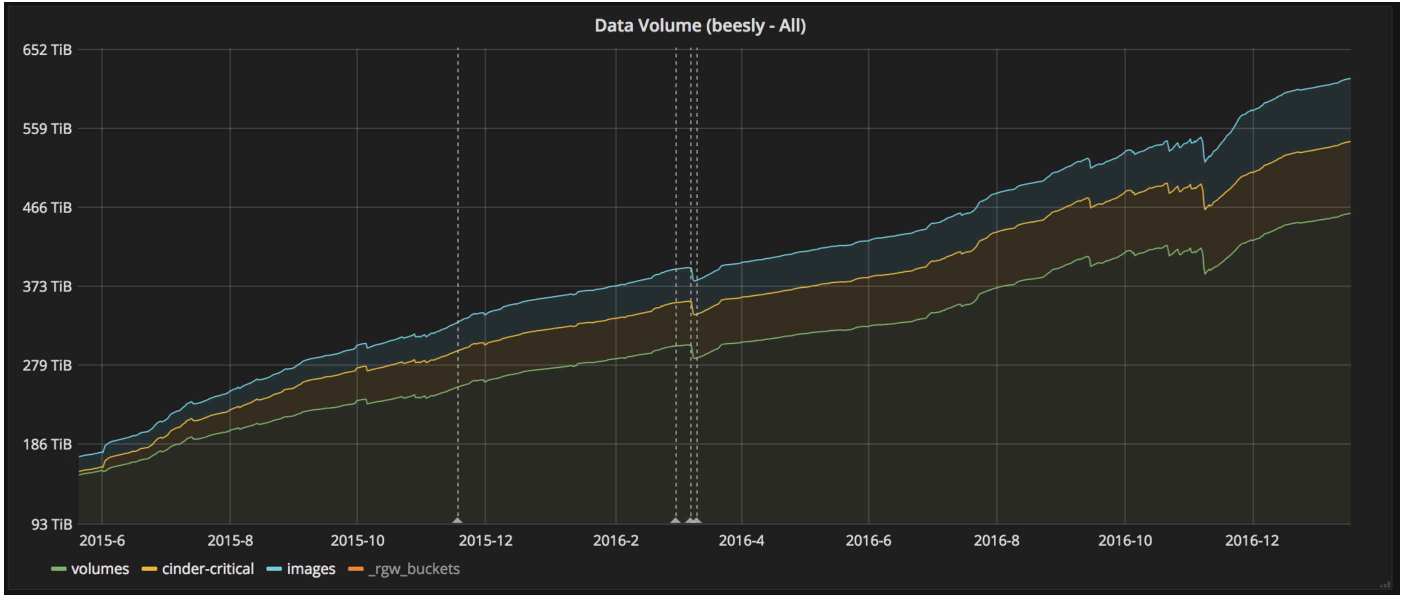
- Ceph has proven to be reliable and flexible in our deployments:
    - 2nd generation Ceph deployments (started in 2013)
    - >10 petabytes in production

Our Ceph Clusters

- ***Beesly + Wigner** (>5 PB, hammer v0.94.9):*
  - OpenStack Cinder (various QoS types) + Glance
- ***Dwight** (0.5 PB, jewel v10.2.5)*:
  - Pre-prod cluster for development (client side) & ops testing.
- ***Erin** (4.2 PB, jewel v10.2.5)*:
  - New cluster for physics: disk layer for our CASTOR tape system
- ***Flax** (400 TB, jewel v10.2.5)*:
  - CephFS cluster for HPC and OpenStack Manila
- ***Gabe** (800 TB, jewel v10.2.5)*:
  - S3 cluster for physics and BOINC volunteer computing

- ***Bigbang** (~30 PB, master)*:
  - Playground for short term scale tests whenever CERN receives new hardware.

Growing OpenStack Usage



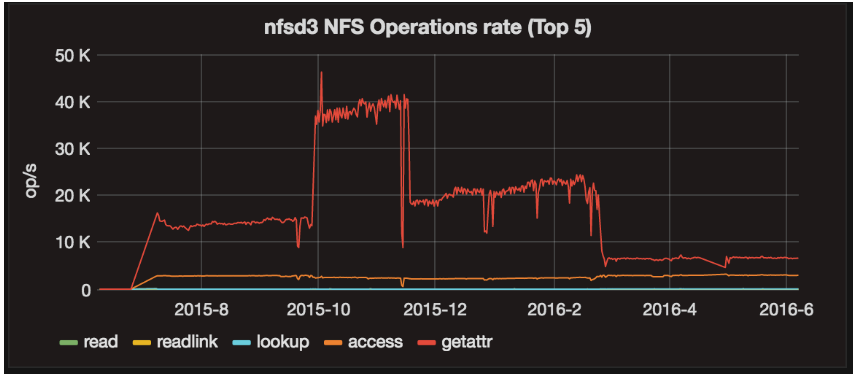
- OpenStack is still Ceph’s killer app. Usage is doubling every 12 months.
- Now ~4000 OS images and ~3000 attached volumes.

NFS on Ceph RBD

- ~50TB across 28 servers:
- OpenStack VM + RBD
- CentOS 7 with ZFSonLinux

- *Not highly-available, but…*

cheap, thinly provisioned, resizable, trivial to add new filers



Example: ~25 puppet masters reading node configurations at up to 40kHz

puppet labs

GitLab

collaborate with TWiki®

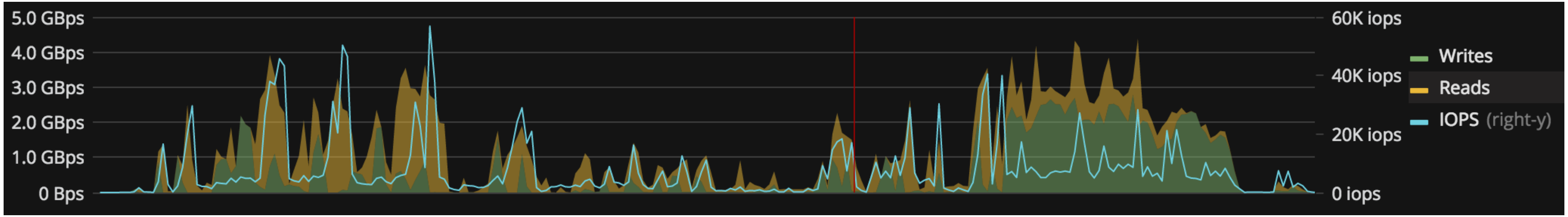
Platform Computing an IBM Company

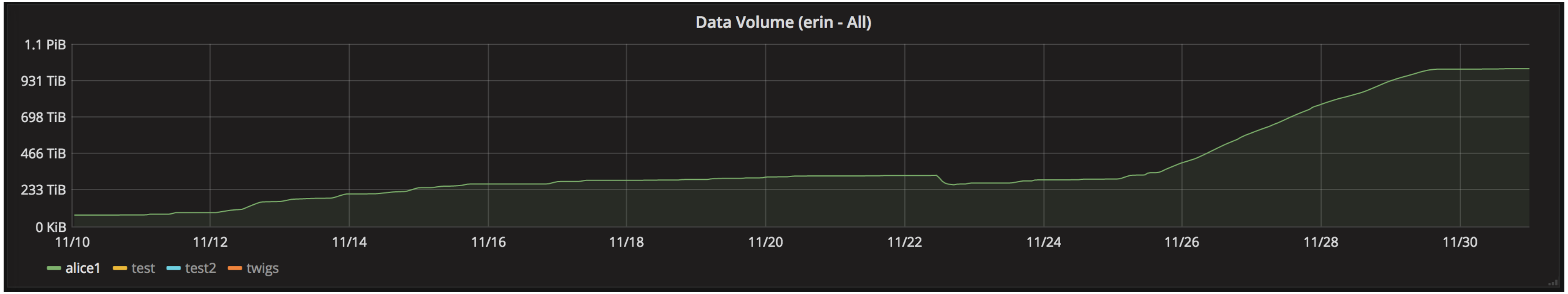
BOINC

OPENSHIFT

elastic

Ceph for Physics Data



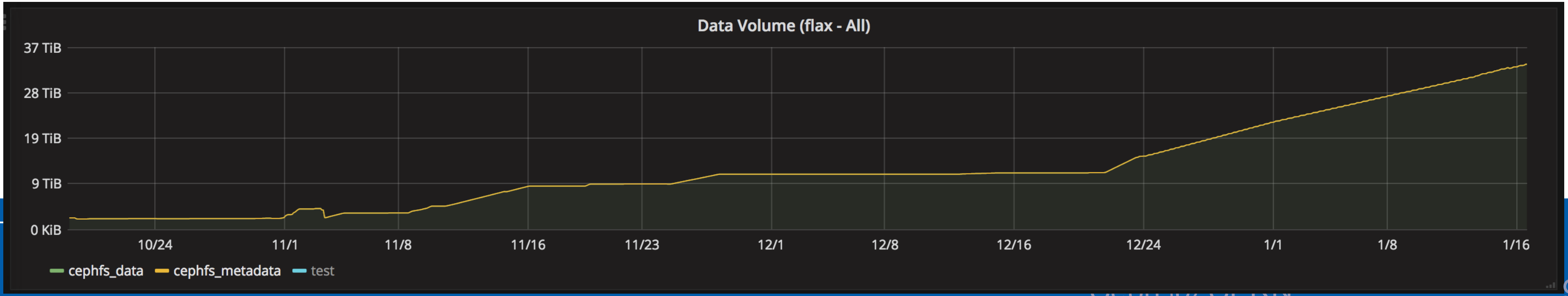


- Data taking for the ALICE experiment during Nov 2016
- Averaging several GB/s over 1-2 weeks, accumulated a total of 1PB

Steps toward CephFS

- Two emerging issues where CephFS can help:
  - CERN is starting to enter the HPC game: need a reliable parallel filesystem
  - Our small NFS Filer service is hitting some scaling & availability limitations

- HPC testing since summer 2016:
  - Looking for compatibility issues or data corruptions. Only a few small issues.
- OpenStack Manila and k8s testing:
  - Manila integration works quite well. Planning a 1000 node scale test.



CERN Contributions to Ceph

- Community participation and membership on the Ceph Advisory Board
- Development contributions: erasure coding and object striping
- Hardware contributions: 30 petabyte scale tests leading to software improvements

Summary

- Happy Ceph users for close to 4 years now
- Growing usage in several areas:
  - Cloud block storage is here to stay
  - Will evaluate new use-cases for physics data
  - CephFS needed for HPC and our Filer use-case

CERN