Our Ceph Clusters

**Beesly** (5 PB + 433 TB, v0.94.9):
- Cinder (block storage volumes)
- Glance (images repo)
- Rados GW (object storage interface: S3, Swift)

**Dwight** (0.5 PB, v10.2.3):
- Preprod cluster for development (client side)
- Testing, upgrades and crazy ideas

**Erin** (4.2 PB, v10.2.3):
- New cluster for CASTOR: disk buffer/cache in front of tape drives

**Flax** (0.4PB, v.10.2.3 - early stage):
- Ceph-FS HPC cluster for QCD studies

**Gabe** (1PB, v.10.2.3):
- New S3 Object Store IPV6 only

**Bigbang** (~30 PB, master):
- Playground for short term scale tests
- Usually when we receive new hardware
Beesly cluster

OpenStack is our killer app: doubled usage in the past year

Very stable, almost 100% uptime* and no data durability issues

*Libnss kvm crashes & 0.94.6->7 broke our record
NFS on RBD

~50TB across 28 servers
OpenStack VM + RBD CentOS 7
with ZFS for DR

Not highly-available, but...
cheap, thinly provisioned, resizable,
trivial to add new filers

Example: ~25 puppet masters reading
node configurations at up to 40kHz
/cvmfs on RBD

Read Only POSIX File System to deliver software over the WAN

How it works: stratums of preloaded HTTP servers + CDN of Squids + CVMFS FUSE client

We use the same architecture as our NFS filers: ZFS on RBD
Beesly cluster

S3 for Volunteer Computing

LHC@Home uses BOINC for volunteer computing

Donate your home CPU cycles to LHC data processing
>10000 volunteer’s cores running in parallel

Data stage-in/out with our Ceph radosgw via Dynafed

Dynafed: The Dynamic Federation project
Exposé via HTTP and WebDAV a dynamic name space from remote endpoints

Redirect GET/PUT requests to the nearest copy

Auth with pre-signed URLs: keep secrets off the desktops
Ops stories
hardware
deliveries

Or: how
to replace
smoothy
1000
disks
(3PB)
Ops stories
hardware
deliveries

Or: how to replace smoothy 1000 disks (3PB)
Hardware Replacement

Need to replace 960 x3TB OSDs with 1152 x6TB drives

How **not** to do it: add new OSDs and remove old OSDs all at once

Lead to massive re-peering, re-balancing and utterly unacceptable IO latency (and cluster was in full prod)


**Strategy followed**: rack-by-rack with dynamic re-weighting based on latency measurements
Total capacity changing as we added/drained OSDs
Hardware Replacement

Activity during replacement campaign

15’000 user IOPS

Write latency stayed under ~10ms
Hardware Replacement

Filling and draining OSDs
To really make an impact, we need Ceph to scale to many 10s of PB’s

At OpenStack Vancouver we presented a 30PB Ceph test. It worked, but had various issues:
- osd to mon’s big messaging volume
- pool creation/deletion
- osdmap churn
- memory usage

We later worked with Ceph devs on further scale testing

Ceph jewel incorporates these improvements
Bigbang II is a second 30PB test we’ve been running during May 2016
  Previous issues all solved
  Benchmarking: ~30GB/s seems doable

New jewel features:
  `ms type = async` (new messaging layer)
  Fewer threads, no tcmalloc thrashing, lower RAM usage

![BigBang II Throughput (GB/s) (5588 OSDs, 32 PB)](chart)
HPC on CEPH

Evaluated CephFS for a shared scratch space on 50 HPC nodes

Strong POSIX needs and 100% uptime, O(month) wall clock jobs

Stable operation over three months

Continue testing plus evaluate OpenStack Manila layer for self-service management

400TB, 3x replication

Two client cache inconsistency (network issue)

ceph-fuse bug quota code crashing (fixed 10.2.4)
CEPH-S3 for ATLAS

Buckets: **atlas_eventservice** & **atlas_logs**: up to ~40TB used, ~50 million objects

5x 10Gbit S3 gateways to Ceph, though not heavily used

Issues with contention in the S3 bucket index:
Fixed by upgrade to Ceph jewel and recreating sharded buckets
Testing "indexless" buckets (no namespace, clients must GET by the exact object name)

Plans: continue testing with new 500TB usable cluster
CEPH-S3 for ATLAS
The Broader HEP Community

Ceph is gaining popularity across the Worldwide LHC Grid

Many OpenStack/RBD deployments + growing usage for physics

U Chicago ATLAS Tier2 (http://cern.ch/go/6T9q): running CephFS + RBD

OSiRIS Project (http://cern.ch/go/F6zS): Three U’s in Michigan building distributed Ceph infrastructure

Orsay/Saclay in France have a similar distributed Ceph project

STFC/RAL in the UK: 12PiB cluster for WLCG Tier1

Meeting monthly to discuss Ceph in HEP

ceph-talk_at_cern.ch ML for discussions: http://cern.ch/go/H6Zh

Thanks to Alastair Dewhurst (STFC/RAL) for the initiative
New Tools

https://github.com/cernceph/ceph-scripts

**ceph-gentle-reweight**
Gradually add or remove OSDs from a cluster

**ceph_osds_in_bucket.py**
Module to find OSDs in a CRUSH bucket

**crush-reweight-by-utilization**
Updated to reweight OSDs in a CRUSH bucket

**ceph-leader**
Tool which exits 0 if the current machine is the mon leader (useful for crons)