

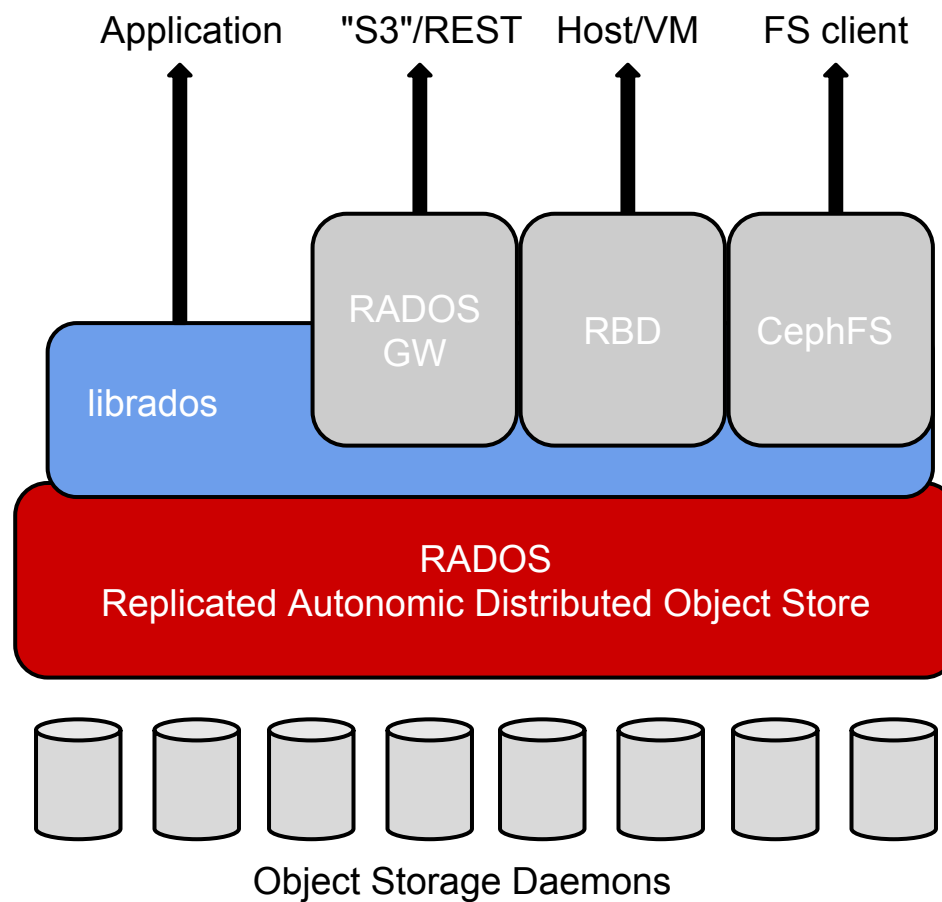


Building an organic storage service at CERN with Ceph

Arne Wiebalck
Dan van der Ster

HEPiX Autumn Meeting 2013
Ann Arbor (MI), U.S.
31 October 2013

DSS Recap: Ceph Basics

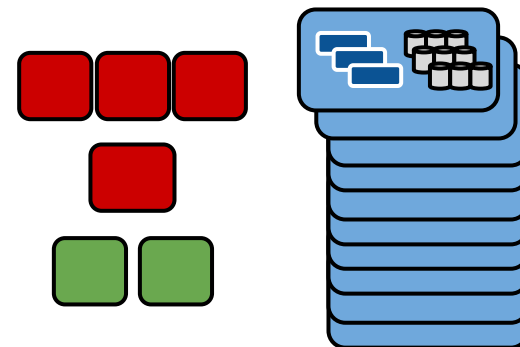


- **Some out-of-warranty CASTOR disk servers**

- 8 OSDs, 3 MONs, 1 RGW, 1 MDS, clients
- 250TB up and usable in 2 days

- **Passed our initial tests**

- RADOS, RBD, RADOS GW, CephFS
- remove OSD, change replication size, delete object in pg, corrupt object in pg, ...
- OpenStack/Cinder
- radosbench
- community support: quick and helpful responses to issues we encountered



- **Some minor issues**

- “2rooms-3replicas” problem, “reweight apocalypse”, “flaky” h/w affects cluster
- RHEL qemu-kvm RPM needed patching

- **Some out-of-warranty CASTOR disk servers**

- 8 OSDs, 3 MONs, 1 RGW, 1 MDS, clients
- 250TB up and usable in 2 days

The results of this initial testing allowed us to convince management to support a more serious Ceph prototype ...

- **Some minor issues**

- “2rooms-3replicas” problem, “reweight apocalypse”, “flaky” h/w affects cluster
- RHEL qemu-kvm RPM needed patching



DSS 12 Racks of Disk Server Quads



DSS Our 3PB Ceph Cluster

48 OSD servers

Dual Intel Xeon E5-2650
32 threads incl. HT
Dual 10Gig-E NICs
Only one connected
24x 3TB Hitachi disks
Eco drive, ~5900 RPM
3x 2TB Hitachi system disks
Triple mirror
64GB RAM

5 monitors

Dual Intel Xeon L5640
24 threads incl. HT
Dual 1Gig-E NICs
Only one connected
3x 2TB Hitachi system disks
Triple mirror
48GB RAM

```
[root@p01001532971954 ~]# ceph osd tree | head -n2
# id weight  type name up/down  reweight
-1  2883 root  default
```




DSS Fully Puppetized Deployment

Fully puppetized deployed

- Big thanks to eNovance for their module!
<https://github.com/enovance/puppet-ceph/>

Automated machine commissioning

- Add a server to the hostgroup (osd, mon, radosgw)
- OSD disks are detected, formatted, prepared, auth'd
- Auto-generated ceph.conf
- Last step is manual/controlled: service ceph start

We use mcollective for bulk operations on the servers

- Ceph rpm upgrades
- daemon restarts



DSS Our puppet-ceph Changes

- Yum repository support
- Don't export the admin key
 - *our puppet env is shared across CERN*
 - *(get the key via k5 auth'd scp instead)*
- New options:
 - *osd default pool size, mon osd down out interval, osd crush location*
- RADOS GW support (RHEL only)
 - *https to be completed*
- /dev/disk/by-path OSDs
 - *better handle disk replacements*
- Unmanaged osd service
 - *manual control of the daemon*
- Other OSD fixes: delay mkfs, don't mount the disks, ...

Needs some cleanup before pushing back to enovance

<https://github.com/cernceph/puppet-ceph/>



DSS Ceph Configuration

11 data pools with 3 replicas each

- mostly test pools for a few different use-cases
- 1-4k pgs per pool; 19584 pgs total

Room/Rack in ceph.conf:

```
osd crush location = room=0513-R-0050  
                    rack=RJ35
```

Rack-wise replication:

```
rule data {  
    ruleset 0  
    type replicated  
    min_size 1  
    max_size 10  
    step take 0513-R-0050  
    step chooseleaf firstn 0 type  
    rack  
    step emit  
}
```




DSS Ceph Configuration

11 data pools with 3 replicas each

- mostly test pools for a few different use-cases
- 1-4k pgs per pool; 19584 pgs total

Room/Rack in ceph.conf:

```
osd crush location = room=0513-R-0050  
                    rack=RJ35
```

```
-1 2883 root default  
-2 2883      room 0513-R-0050  
-3 262.1          rack RJ35  
-15 65.52                host p05151113471870  
-16 65.52                host p05151113489275  
-17 65.52                host p05151113479552  
-18 65.52                host p05151113498803  
-4 262.1          rack RJ37  
-23 65.52                host p05151113507373  
-24 65.52                host p05151113508409  
-25 65.52                host p05151113521447  
-26 65.52                host p05151113525886  
...
```

DSS Service Monitoring

Service information

full name: **Ceph Storage Service**

short name: Ceph


group: IT/DSS


site: CERN

email: ceph-admins@cern.ch


web site: <https://twiki.cern.ch/twiki/bin/viewauth/DSSGroup/CephP...>

alarms page: <http://cern.ch/ceph/alarms.html>

service Arne Wiebalck 

managers: Dan van der Ster 

Part of (subservice of):

 IT/DSS services

Subservices

none / not declared

Clusters, subclusters and nodes

cluster **ceph_beesly_mon**


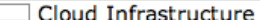
cluster **ceph_beesly_osd** 

Depends on

none / not declared

Depended on by

services that depend on this service:

  Cloud Infrastructure

Service availability [\(more\)](#)


availability: 

percentage: 100%

status: **available**

last update: 11:16:09, 2 Oct 2013
(13 minutes ago)

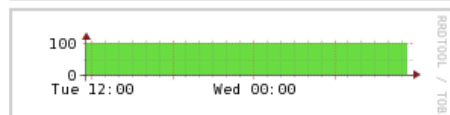
expires after: 15 minutes

 [rss feed with status changes](#)

how is availability measured or estimated:

Availability is 100% when Ceph reports HEALTH_OK, otherwise it is the percentage placement groups which can actively accept IOs.

availability in the last 24 hours [\(more\)](#):



Additional service information [\(more\)](#)

Num Mons:	5
Num Mons in Quorum:	5
Num Pools:	12
Num OSDs:	1,056
Num OSDs Up:	1,056
Num OSDs In:	1,056
Num PGs:	19,584
Num PGs Active:	19,584
OSD Gigabytes Total:	2,949,955
OSD Gigabytes Used:	13,371
OSD Gigabytes Avail:	2,936,583
PG Gigabytes:	762
Num Objects:	134,787
Num Object Copies:	404,359
Num Objects Degraded:	0
Num Objects Unfound:	0
Total Read (GB):	3,501
Total Write (GB):	6,064

Service information

full name: **Ceph Storage Service**

Part of (subservice of):

IT/DSS services

A few monitoring helper scripts

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

ceph-health-cron:

- report on the ceph health hourly

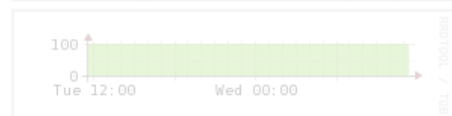
cephinfo:

- python API to the ceph JSON dumps

cern-sls:

- example usage of cephinfo.py
- compute and publish ceph availability and statistics

availability in the last 24 hours (more):



Num Objects:	134,767
Num Object Copies:	404,359
Num Objects Degraded:	0
Num Objects Unfound:	0
Total Read (GB):	3,501
Total Write (GB):	6,064



DSS Initial Benchmarks

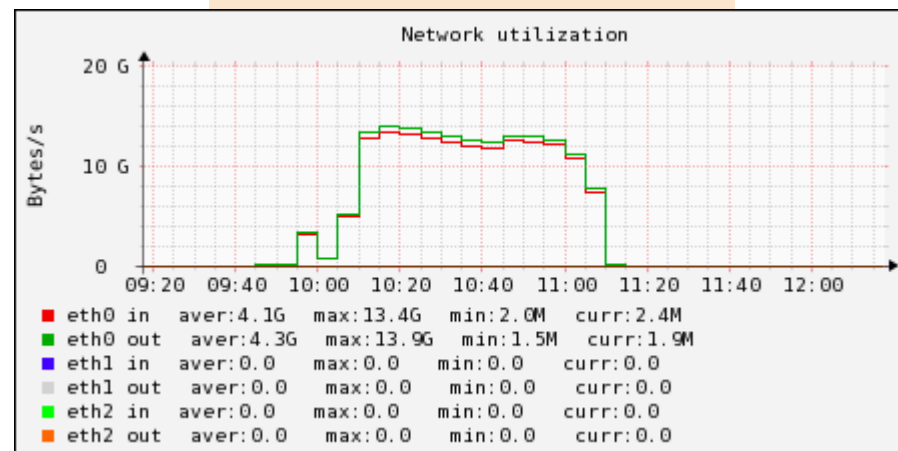
basic rados bench - saturate the network

```
[root@p05151113471870 ~]# rados bench 30 -p test write -t 100
Total writes made:      7596
Write size:             4194304
Bandwidth (MB/sec):     997.560
Average Latency:        0.395118
[root@p05151113471870 ~]# rados bench 30 -p test seq -t 100
Total reads made:       7312
Read size:              4194304
Bandwidth (MB/sec):     962.649
Average Latency:        0.411129
```

120M file test

Wrote 120 million tiny files into RADOS to measure scalability by that dimension. No problems observed. Then we added one OSD server, and the rebalance took ages (~24hrs) which is probably to be expected.

all-to-all rados bench





DSS Our Users

A few early adopters are helping us evaluate Ceph:

- **OpenStack:** usage for Glance images and Cinder volumes
- **AFS/NFS:** backend RBD storage
- **DPM:** backend RBD storage
- **OwnCloud:** S3 or CephFS backend for desktop synchronisation
- **Zenodo:** backend storage for data and publications sharing service



DSS Openstack / Ceph Testing

We are still validating the OpenStack / Ceph integration

- We require the version of qemu-kvm patched by Inktank to support RBD
- Our workloads benefit from striping:
 - Gary McGilvary developed and pushed some patches to allow configurable striping via the OpenStack UI
- Our grizzly cluster is using RBD
 - Small problem related to ulimit, see coming slide...
- For Cinder usage we are currently blocked:
 - Deployed Grizzly with *cells* to divide our large facilities
 - Grizzly cells don't support Cinder
 - Belmiro Moreira backported the Havana code for Cinder/Cells; currently under test



Latency:

- Our best case write latency is presently 50ms
 - 1 replica, journal as a file on the OSD
- We tested an in-memory OSD and saw ~1ms latency
 - So our high latency comes from our journal
- We need to put our journals on the blockdev directly (should get ~12ms writes) or use SSDs

ulimits:

- With more than >1024 OSDs, we're getting various errors where clients cannot create enough processes to connect to the OSDs
 - failed ceph tell, failed glance image uploads
- Our clients have been informed to increase ulimit -u to 4096, but it would be useful if ceph was somehow less process greedy.



DSS Summary and Outlook

Following the successful initial testing, we've set up a PB scale Ceph cluster that is currently being used or evaluated by various use cases

- OpenStack
- AFS/NFS
- DPM and others ...

The killer app for Ceph would be to build upon it a general purpose network file system

- “Nearly awesome” -- Sage Weil (Ceph Day 2013, London)
- Used in production by some users
- Some features are missing
- CERN will start testing it



DSS

Thanks!