DSS

Data & Storage Services

CERN IT Department

Building an organic block* storage service at CERN with Ceph

**** and object and POSIX***

Dan van der Ster
Arne Wiebalck

CHEP 2013, Amsterdam
14 October 2013

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

CERN

DSS Big Data at CERN

CERN IT Department

Physics Data on CASTOR/EOS

- LHC experiments produce ~10GB/s, 25PB/year

User Data on AFS/DFS

- Home directories for 30k users
- Physics analysis dev't
- Project spaces (applications)

Service Data on AFS/NFS

- Databases, admin applications

Tape archival with CASTOR/TSM

- RAW physics outputs
- Desktop/Server backups

Service	Size	Files
AFS	240TB	1.9B
CASTOR	87.7PB	317M
EOS	19.8PB	160M

CERN developed CASTOR & EOS because until very recently our storage reqs were globally unique.

Following the Google / Amazon / Facebook innovations, we are now trying to leverage community solutions

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

IT (R)evolution at CERN

Cloudifying CERN's IT infrastructure ...

- Centrally-managed and uniform hardware
  - *No more service-specific storage boxes*
- OpenStack VMs for most services
  - *Building for 100k nodes (mostly for batch processing)*
- Attractive desktop storage services
  - *Huge demand for a local Dropbox, Google Drive …*
- Remote data centre in Budapest
  - *More rack space and power, plus disaster recovery*

… brings new storage requirements

- Block storage for OpenStack VMs
  - *Images and volumes*
- Backend storage for existing and new services
  - *AFS, NFS, OwnCloud, Zenodo, ...*
- Regional storage
  - *Make use of the new data centre in Hungary*
- Failure tolerance, data checksumming, easy to operate, security, ...

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

Possible Solutions

GlusterFS

- Cloud team at CERN found it wasn't stable enough
- Doesn't offer block device for physical machines

NFS (NetApp)

- Expensive
- Vendor lock-in

Ceph

- Interesting architecture (on paper)
- Offers almost all features we needed

Early 2013 we started investigating Ceph ...

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

DSS

Never heard of Ceph?

CERN IT Department

Ceph is a distributed, open-source storage system.

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

CERN

Never heard of Ceph?

Ceph is a distributed, open-source storage system.

Scalability

TeraBytes to ExaBytes
10s to 10'000 machines
Grow and shrink

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

DSS

Never heard of Ceph?

CERN IT Department

Ceph is a distributed, open-source storage system.

Scalability

TeraBytes to ExaBytes
10s to 10'000 machines
Grow and shrink

Reliability

No SPOF
Commodity hardware
Configurable Replication

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

CERN

DSS

Never heard of Ceph?

CERN IT Department

Ceph is a distributed, open-source storage system.

Scalability

TeraBytes to ExaBytes
10s to 10'000 machines
Grow and shrink

Reliability

No SPOF
Commodity hardware
Configurable Replication

Manageability

Self-organizing
Self-healing
CRUSH

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

CERN

Never heard of Ceph?

Ceph is a distributed, open-source storage system.

Scalability

TeraBytes to ExaBytes
10s to 10'000 machines
Grow and shrink

Reliability

No SPOF
Commodity hardware
Configurable Replication

Manageability

Self-organizing
Self-healing
CRUSH

Unification

Object store
Block store
File system

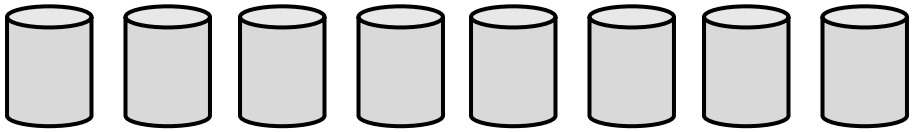
CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

DSS

Ceph's architecture

CERN IT Department

OSDs control a device on the hosts in a Ceph cluster



Object Storage Daemons

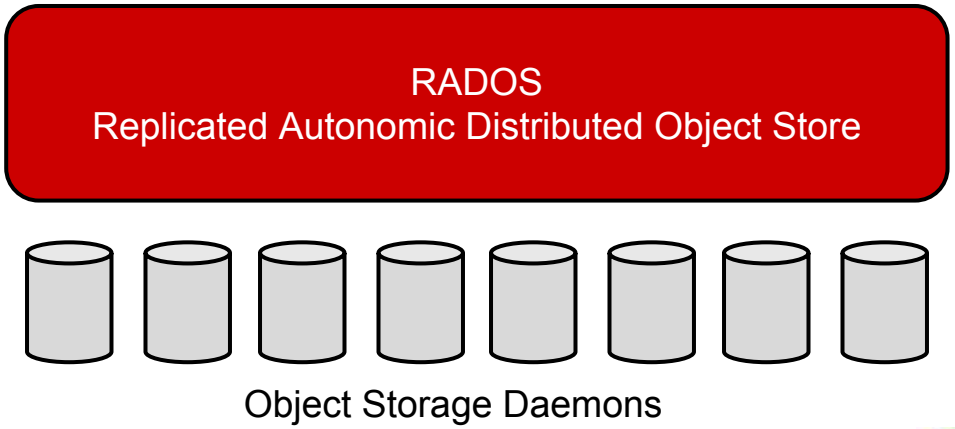
CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

CERN

DSS Ceph's architecture

CERN IT Department

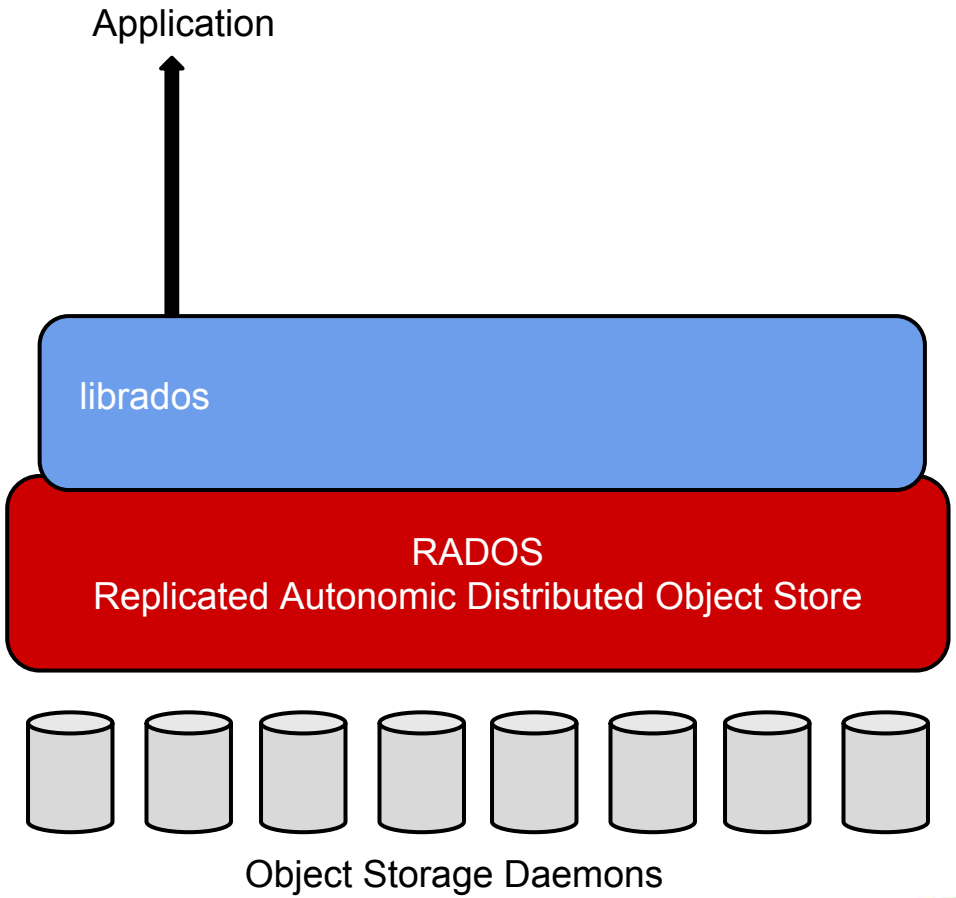
Reliable, self-managing, self-healing **object** store



CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

Ceph's architecture

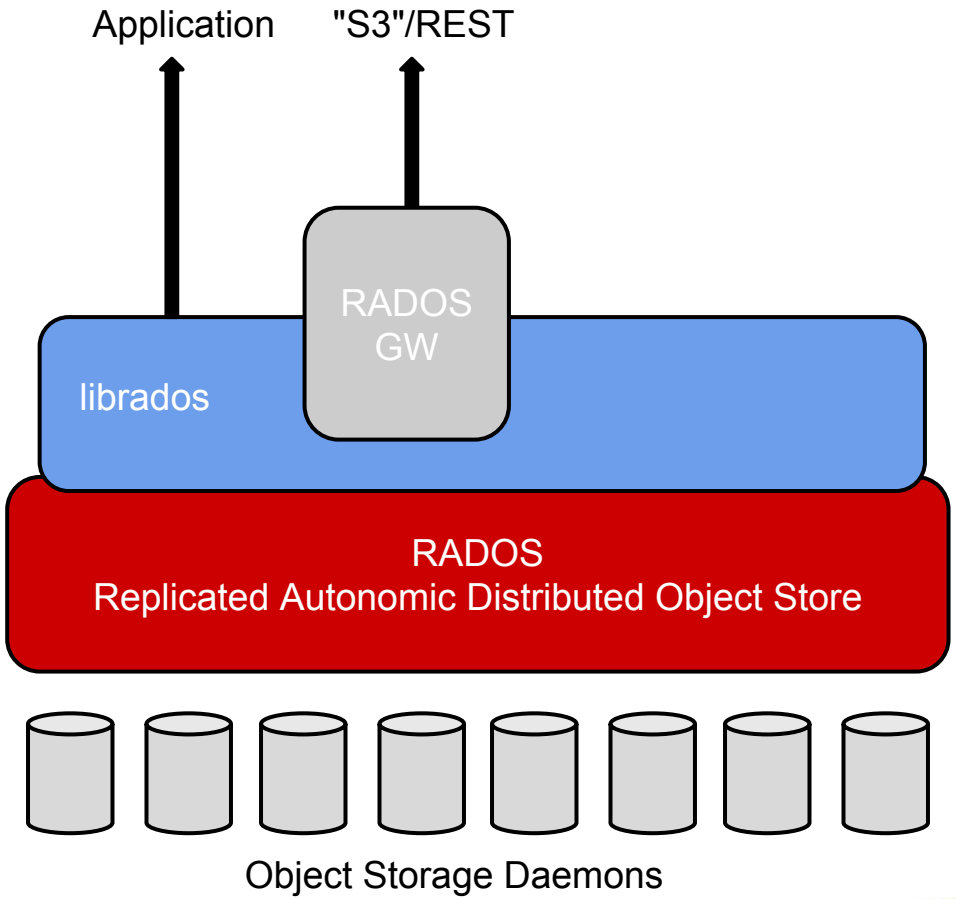
Library for RADOS access from C, C++, Java, Python, ...



CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

Ceph's architecture

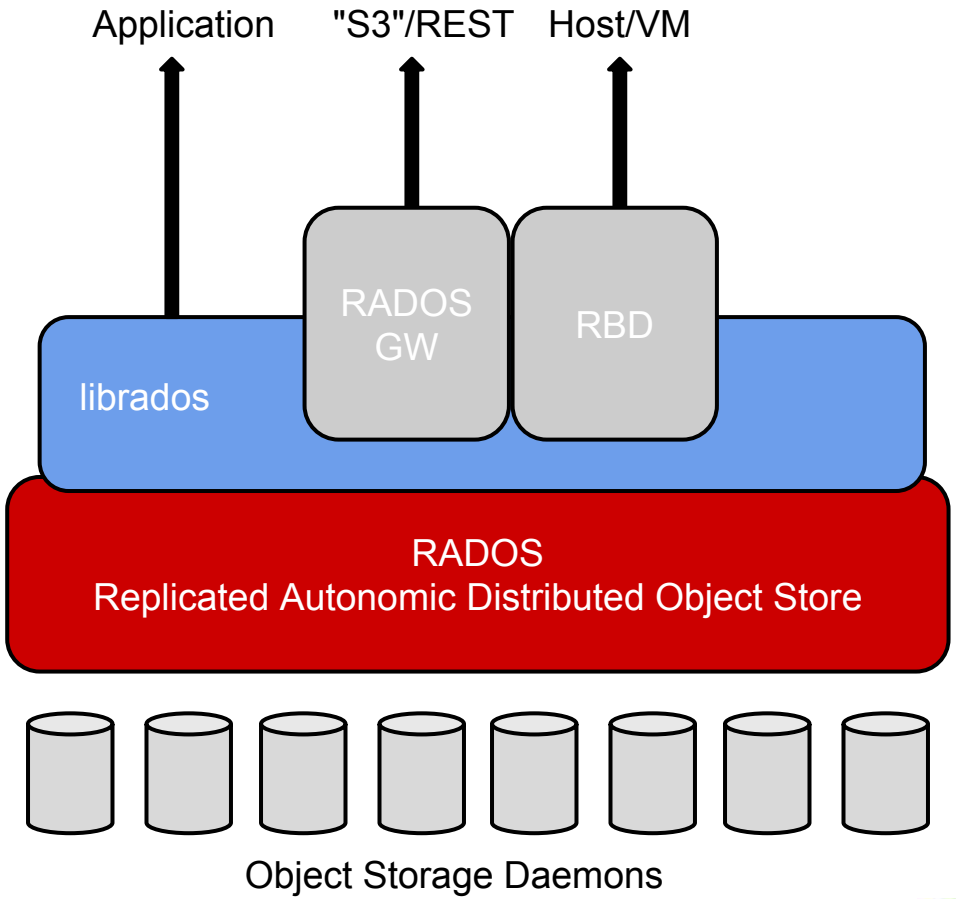
REST gateway compatible with (a large subset of) Swift and Amazon's S3.



CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

Ceph's architecture

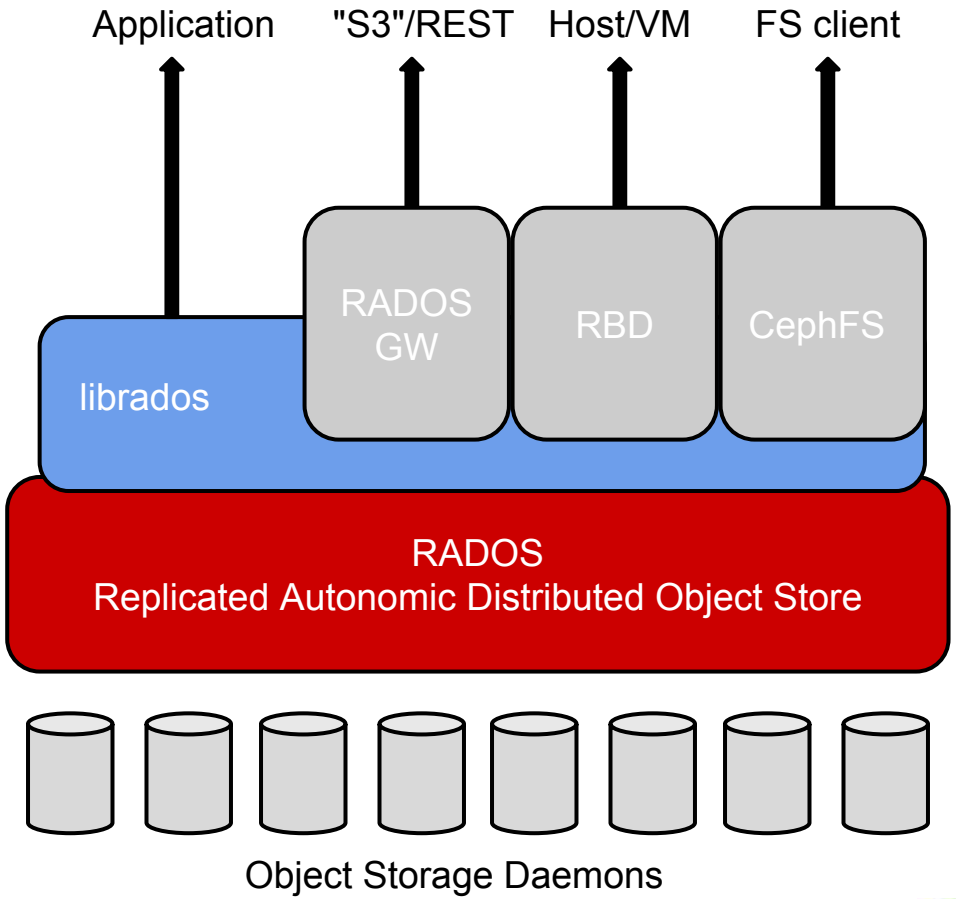
Thinly provisioned distributed block device to be used from VMs or hosts in general; Linux kernel module or KVM/QEMU/libvirt+librbd.



CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

Ceph's architecture

POSIX-compliant distributed file system that ships with the Linux kernel since 2.6.34; usable via kernel module (FUSE available as well).



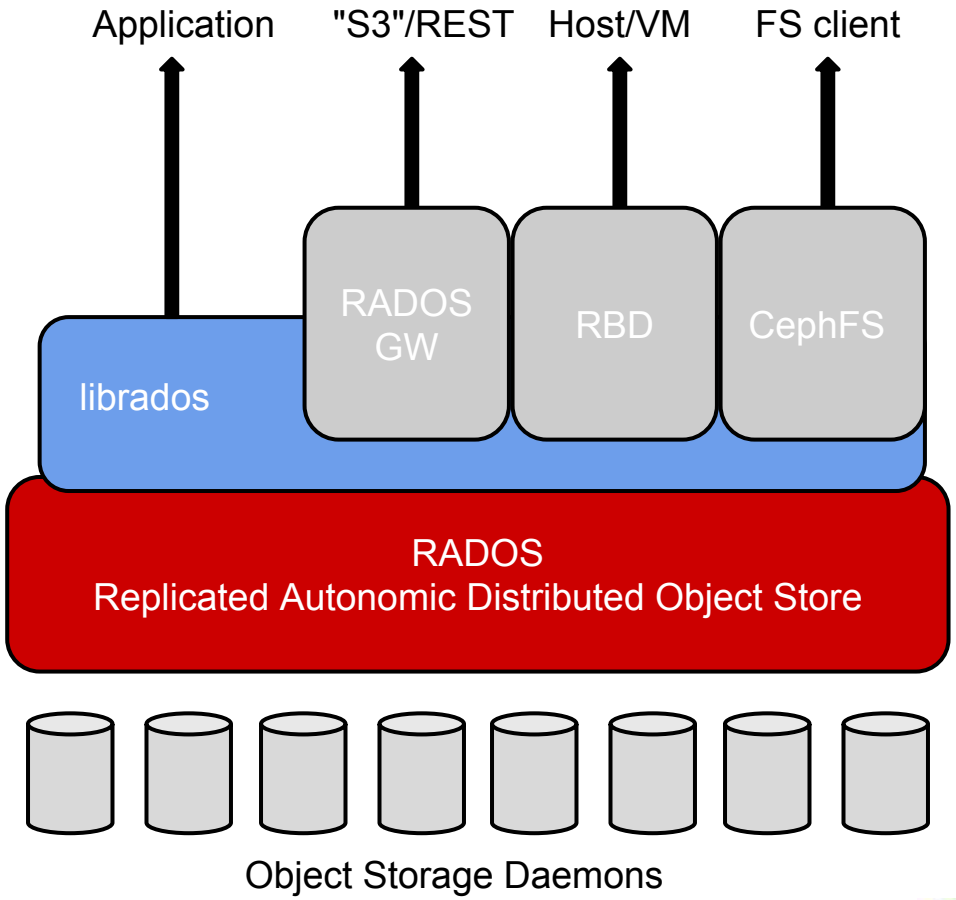
CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

Ceph's architecture

POSIX-compliant distributed file system that ships with the Linux kernel since 2.6.34; usable via kernel module (FUSE available as well).

Each of the grey boxes stripe their data for performance

... break large files into xMB objects and distribute across many disks



CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

CRUSH ...

Controlled Replication Under Scalable Hashing

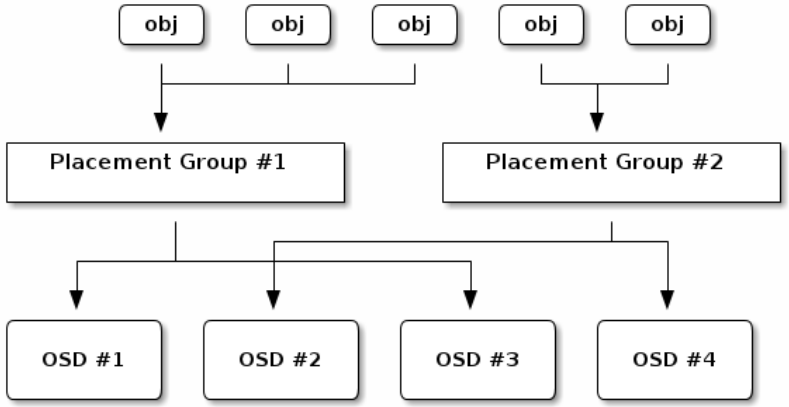
- algorithmic data placement
  - no central meta data server
  - clients compute where data is
  - based on CRUSH map (which is "gossip'ed")
  - stable mapping
- infrastructure-aware
  - centers, buildings, rooms, row-of-racks, racks, hosts
  - define placement rules (e.g. each replica in a different rack)
  - address correlated failures
  - allows weighting OSDs

Read more about CRUSH: http://ceph.com/papers/weil-crush-sc06.pdf

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

Data placement

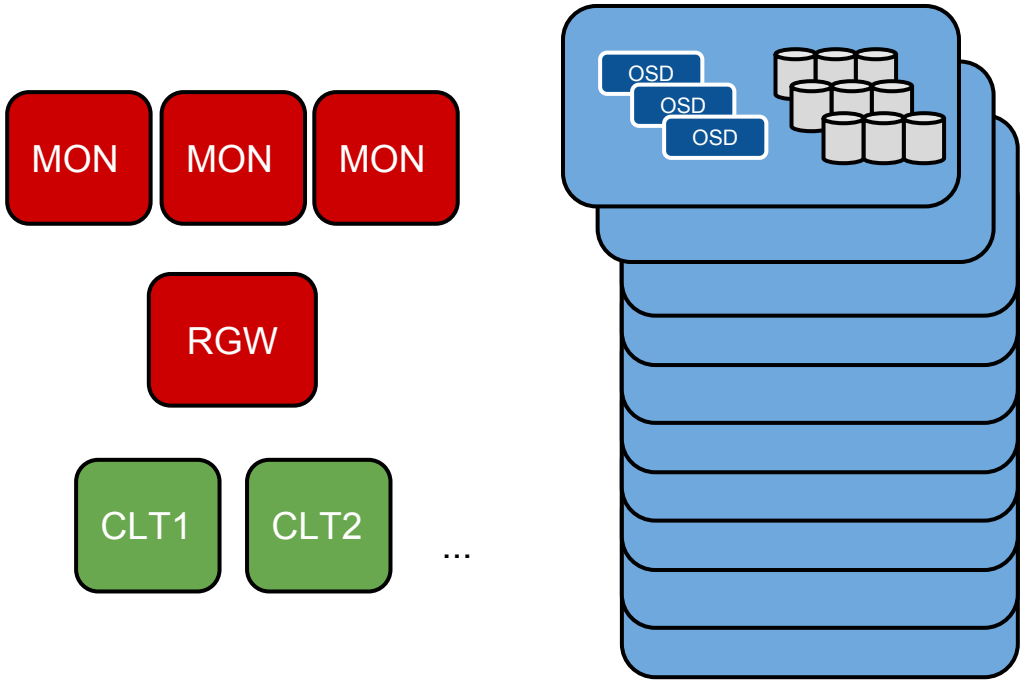
- **A Placement Group (PG) aggregates a series of objects into a group, and maps the group to a series of OSDs**
- Data is organized in pools; each pool has M replicas, N placement groups (PGs)
  - e.g. pool "data" with ID 1, 2 replicas, 512 PGs
- Placement groups are physically stored as a directory on an OSD
  - e.g. /osd.1/1.fe0/ and /osd.3/1.fe0/
- Object name is hashed to a PG
  - e.g. object "myfile" -> PG 1.fe0
- *CRUSH* map is used to lookup which servers/OSDs hold that PG



CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

First steps

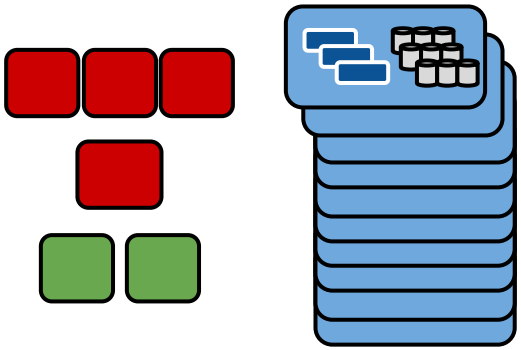
- **Set up a small scale test cluster**
  - 3 MON servers, 1 RADOS gateway (all VMs)
  - 8 OSD hosts with 4-5 disks each (ex-CASTOR)
  - Ceph 0.56.4 installed via *yum install ceph* on SLC6.4
  - Various clients: kernel rbd driver, OpenStack, AI monitoring, ...



CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

Early testing

- **Setup was easy**
  - ~2 days for our 250TB testbed
- **Passed our (simple) interface tests**
  - RADOS, RBD, RADOS GW, CephFS
- **Passed our first functional tests**
  - remove OSD, change replication size, delete object in PG, corrupt object in PG, …
  - OpenStack/Cinder
- **Passed our performance tests**
  - rados bench
- **Passed our community expectations**
  - very quick and helpful responses to issues we encountered



CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

Early testing

- **Setup was easy**
  - ~2 days for our 50TB testbed
- **P**
- **I**
- **Passed our performance tests**
  - rados bench
- **Passed our community expectations**
  - very quick and helpful responses to issues we encountered

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

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

DSS 12 racks of disk server quads

CERN IT Department

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

DSS Our 3PB Ceph Cluster

CERN IT Department

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
```

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

CERN

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
}
```

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

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
...
```

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

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

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

DSS Service Monitoring

CERN IT Department

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 managers: Arne Wiebalck, 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**

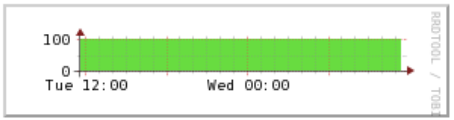
last update: 11:16:09, 2 Oct 2013 (13 minutes ago)
expires after: 15 minutes

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

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

DSS Initial Benchmarks

CERN IT Department

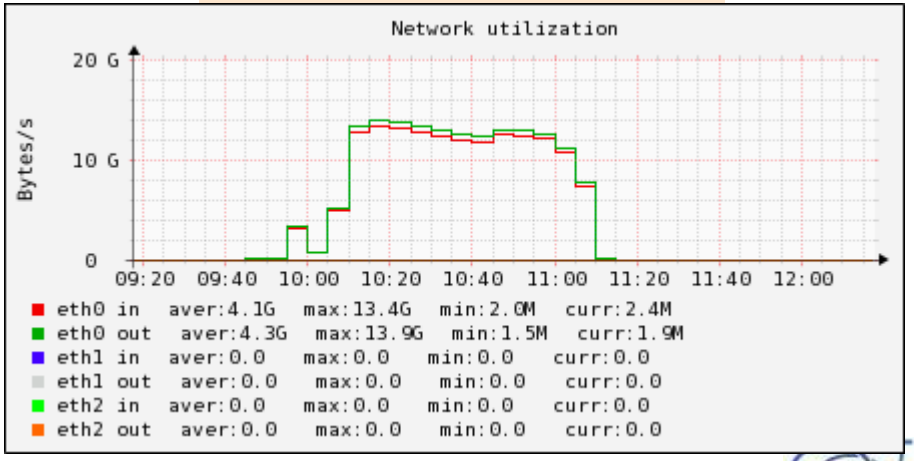
basic rados bench - *saturate the network*

```
[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



CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

Our Users

A few early adopters are helping us evaluate Ceph:

- **OpenStack**: usage for Glance images and Cinder volumes
- **AFS/NFS**: backend RBD storage for these commonly used fs's
- **CASTOR**: high performance buffer of objects to be written to tape
- **DPM**: backend RBD storage for this high-energy-physics fs
- **OwnCloud**: S3 or CephFS backend for desktop synchronisation
- **Zenodo**: backend storage for data and publications sharing service

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

Looking forward...

librados is very powerful:

- could be interesting for physics data

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

- Would help us get rid of NetApp boxes
- Dare we dream that it may one day replace AFS?!

CephFS is advertised as not yet production quality, so we don't yet advertise it to our users

- "Nearly Awesome" -- Sage Weil

To be generally usable we'd need:

- HA and load balanced (for AFS we get accessed at 75kHz)
- All the goodies we get from AFS: quotas, ACLs, krb5, ...

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

Conclusions

We are attracting various use-cases

- OpenStack images and volumes
- RBD backends for other storage services (AFS/NFS/DPM)
- Object storage for novel applications: (tape buffer, Zenodo, OwnCloud)

We have very high hopes for Ceph at CERN!

- the design is *interesting*
- the performance so far is adequate
- operationally it is very attractive

Everybody wants a filesystem: CephFS will be crucial

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

BACKUP SLIDES

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

DSS

Our puppet-ceph changes

CERN IT Department

- 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/

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

CERN

Puppet-ceph TODO/Wish-list

We have some further puppet work in mind:

- Add arbitrary ceph.conf options
- Move the OSD journal to a separate partition
- SSD OSD journals
- Use the udev triggers for OSD creation

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

Current Issues

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 (but we're worried they'll wear out)

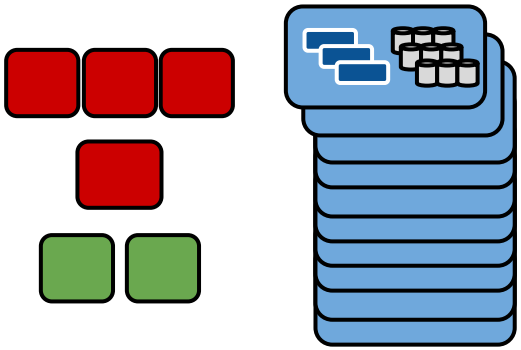
ulimits:

- 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 useful if ceph was somehow less process greedy.

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

Issues during early testing

- **ceph-deploy did not work for us at the time**
- **“2 rooms - 3 replicas - problem”**
- **“re-weight apocalypse”**
  - wrong ratio of RAM to OSDs



- **“flaky” server caused Ceph timeouts and constant re-balancing**
  - taking out the server “fixed” the problem
  - root cause not understood (can slow server slow down the cluster?)
- **qemu-kvm RPM on RHEL derivative SLC needs patching**
  - RPM provided by Inktank

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

Openstack / Ceph Testing

We are still validating the OpenStack / Ceph integration

- Being a RedHat shop, 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

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

Ceph as a Tape Buffer?

CERN IT Department

CASTOR holds much of our physics data

- 90PB total, 75PB on TAPE

Tapes write at 250MB/s; without striping CASTOR diskservers cannot supply data at that rate.

Idea: put a Ceph buffer between the disk servers and tape drives

but… single threaded read performance

```
[root@p05151113471870 ~]# rados bench 10 -p test seq -t 1
Total reads made:     612
Read size:            4194304
Bandwidth (MB/sec):    244.118
Average Latency:       0.0163772
```

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

Ceph as a Tape Buffer?

So our colleage Andreas Peters prototyped a striping RADOS object client: **cephcp**

```
Upload:
[root@p05151113471870 ~]# ./cephcp -p test -i admin -n 64 file:
/root/1G.dat ceph:/root/1G.dat
[cephcp] 1073741824 bytes copied in 1137.89 ms [ 943.63 MB/s ]
```

```
Download
[root@p05151113471870 ~]# ./cephcp -p test -i admin -n 64 ceph:
/root/1G.dat file:/dev/null
[cephcp] 1073741824 bytes copied in 1022.40 ms [ 1050.22 MB/s ]
```

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

Service Monitoring

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

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it