DSS

Data & Storage Services

CERN IT Department

Ceph: A scalable, organic option for Storage-as-a-Service at CERN

Arne Wiebalck
Dan van der Ster

HEPiX 2013 Spring Meeting
Bologna, 19 April 2013

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

CERN

Disclaimer

This is a report on work in progress.

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

DSS Current DSS services and demands

CERN IT Department

- AFS as a distributed file system
- CASTOR for tape storage / archive
- EOS for analysis
- NetApp filers for special cases new
- Block storage for AI VMs new
- Hadoop new
- "dropbox" new

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

CERN

DSS Current DSS services and demands

CERN IT Department

- AFS as a distributed file system
-
-
-
-
- Hadoop new
- "dropbox" new

The new demands motivate a consolidated, generic storage system ...

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

Never heard of Ceph?

Ceph is a distributed, open-source storage system.

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

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

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

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

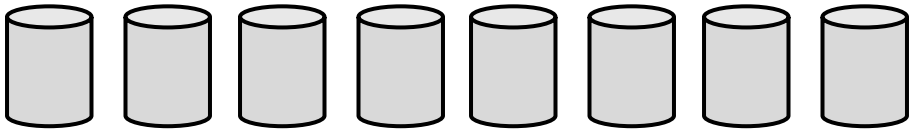
Unification

Object store
Block store
File system

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

Ceph's architecture

OSDs control a device on the hosts in a Ceph cluster



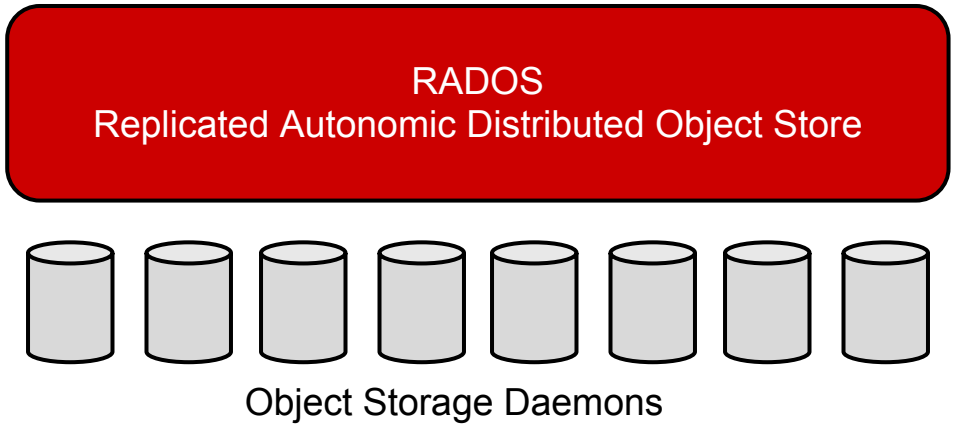
Object Storage Daemons

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

DSS Ceph's architecture

CERN IT Department

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

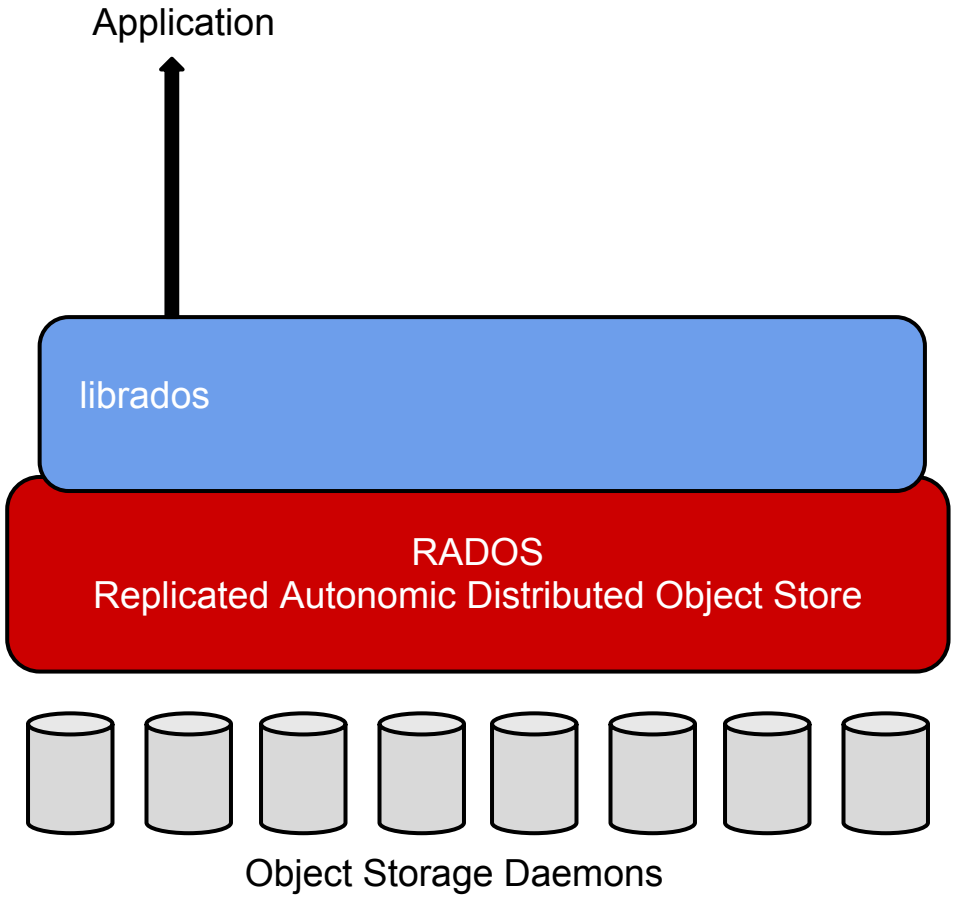


CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

Ceph's architecture

CERN IT Department

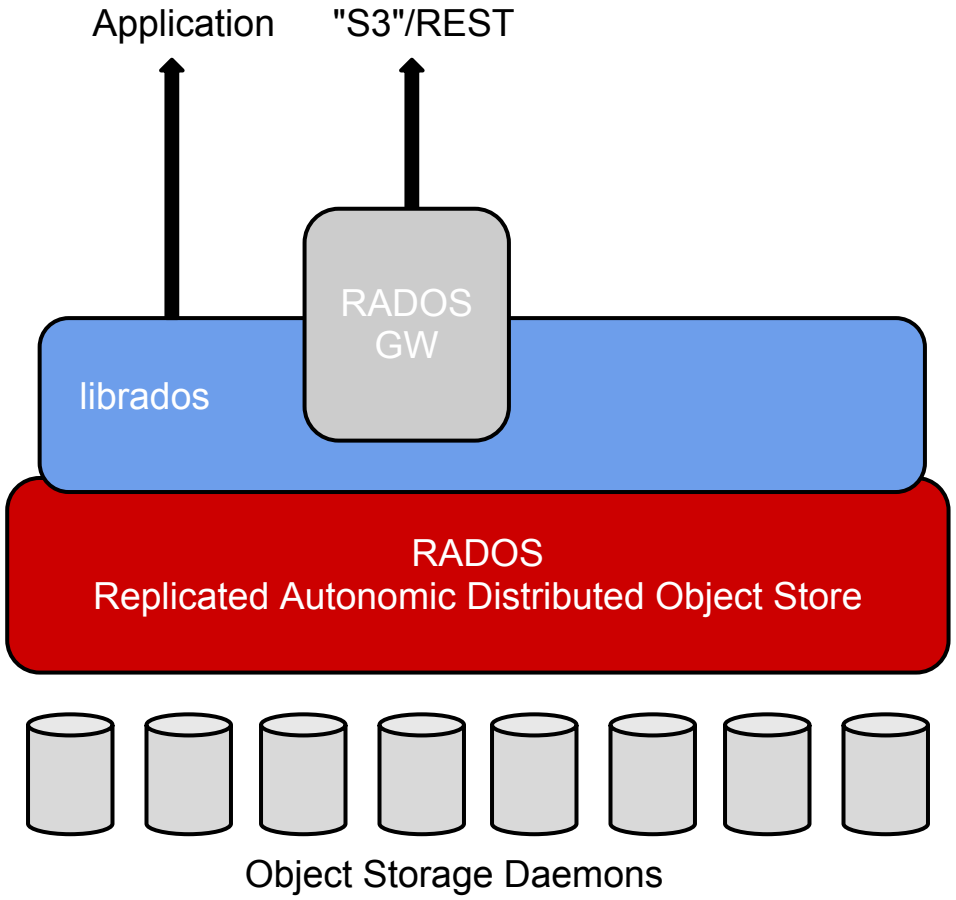
Library for RADOS access from C, C++, Java, Python, ...



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

Ceph's architecture

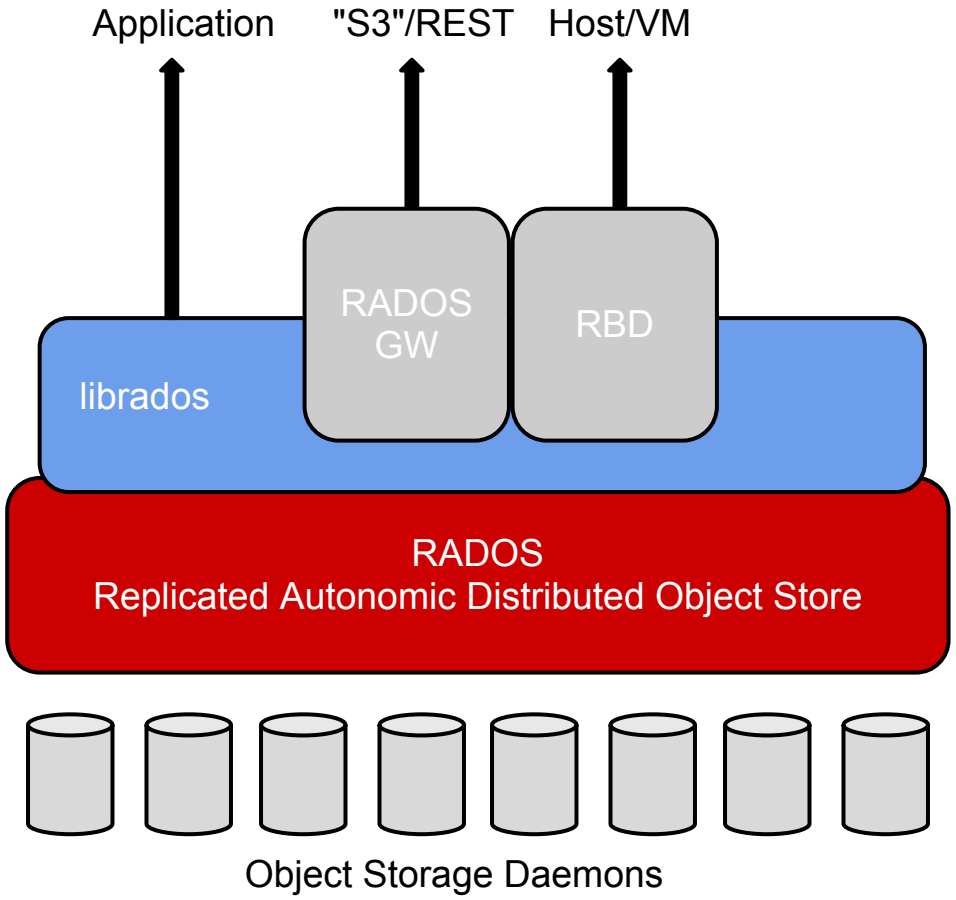
REST gateway compatible with (a large subset of) Swift and Amazon's S3.



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

Ceph's architecture

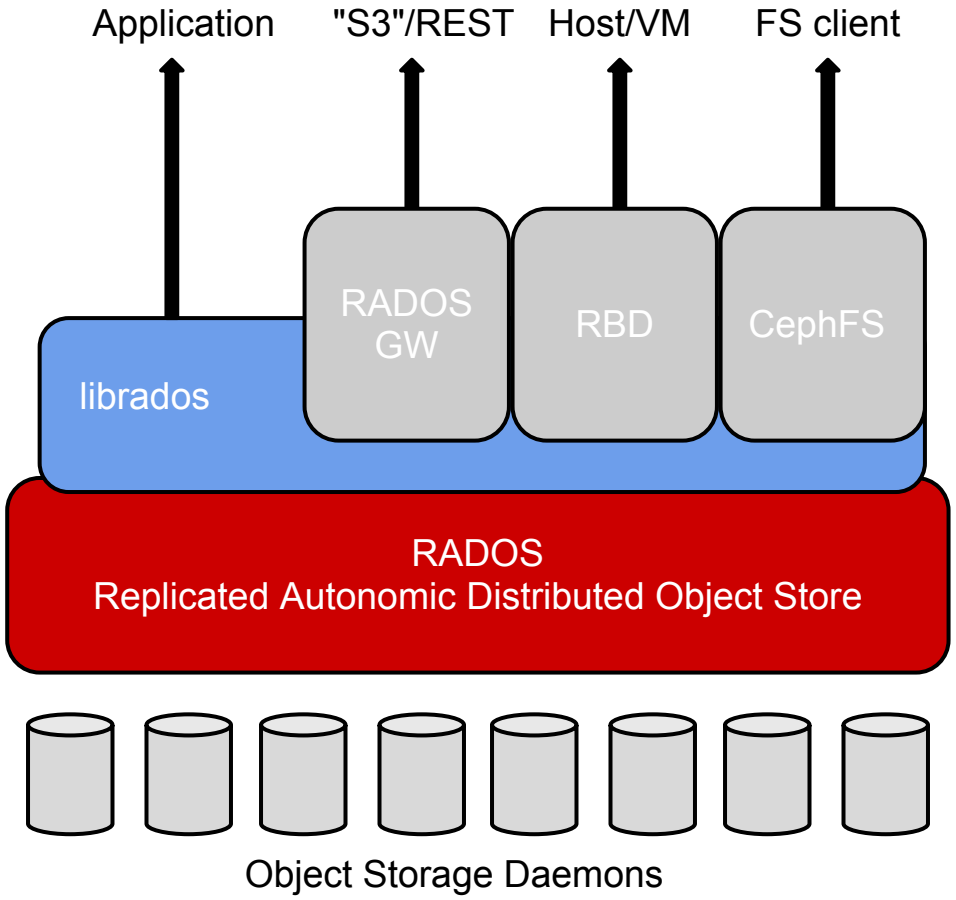
Thinly provisioned distributed block device to be used from VMs or hosts in general; Linux kernel module or KVM/QEMU/libvirt+librbd.



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

One word about 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")
- infrastructure-aware!
  - centers, buildings, rooms, row-of-racks, racks, hosts
  - define placement rules (e.g. all replicas in different racks)
  - address correlated failures

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

How does it fit with what we need?

- Provide block storage for AI VMs
  - OpenStack/Cinder & Glance compatible (few options)
  - GlusterFS: investigated, not there yet
  - NetApp: works, but vendor lock-in/closed source
  - Ceph: supported, among main use cases
- Further simplify AFS (and NFS)
  - Rely on safe backend
  - rbd (or radosgw & S3)
- Future options
  - Object backend for "dropbox"-like service?
  - CephFS as next distributed file system?

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

Our plan

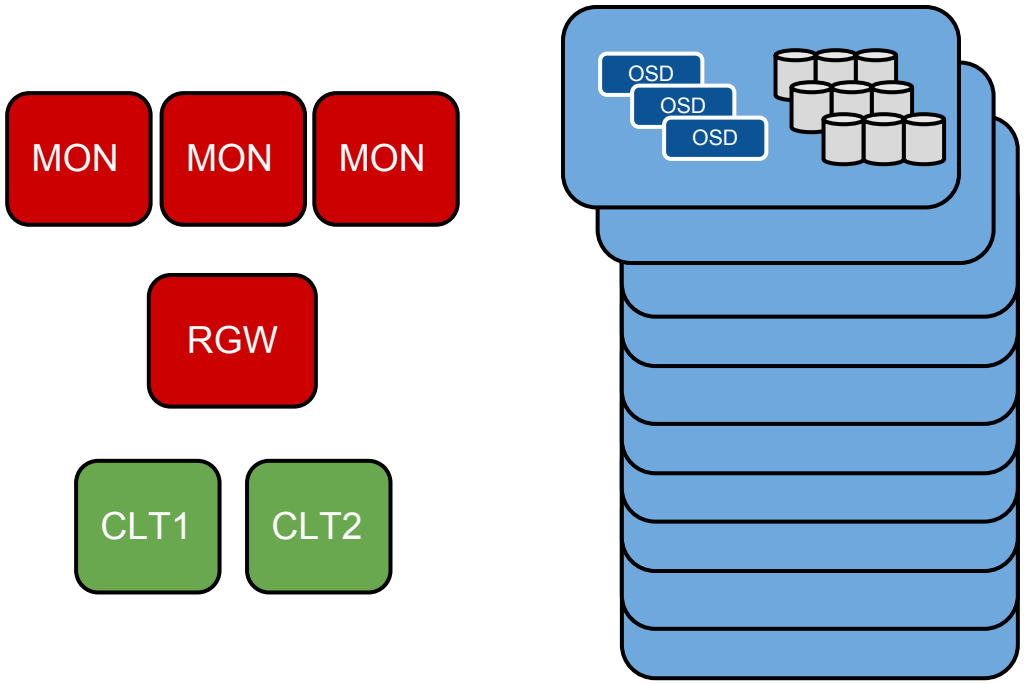
Evaluate Ceph as an option for our use cases:

- Build a (serious) prototype
  - 1 PB usable
- Test it for our use cases
  - OpenStack/Cinder
  - AFS backend
  - OwnCloud
  - CephFS (too early: not prod quality yet, SLC kernel)

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

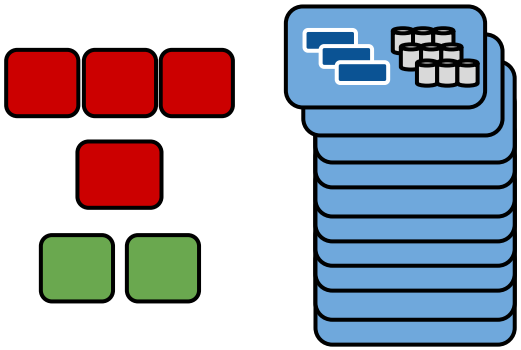
What we have done so far (1)

- Set up a small scale test cluster
  - 3 MON servers, 2 clients (on 3.8.2), 1 RADOSGW (all on VMs)
  - 10 OSD hosts with 22 disks each (ex-CASTOR)
  - Ceph 0.56.3 installed via RPMs on SLC6.4



What we've done so far (2)

- Setup was easy
  - 250TB cluster up in ~2 days
  - manual (ceph-deploy did not work for us)
  - Puppet modules exist, but not tested yet
  - RADOSGW a little more tricky

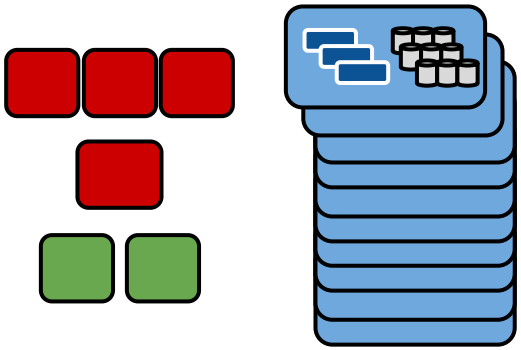


- Interface tests
  - RADOS: create/delete/list objects
  - RBD: on 3.8.2 client, mkfs.etx4 + mount
  - CephFS: on 3.8.2 client
  - RADOSGW: CyberDuck (list, create, delete buckets)

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

What we've done so far (3)

- Functional testing
  - add/remove OSDs
  - increase/decrease replication size
  - delete object in placement group
  - corrupt object in placement group

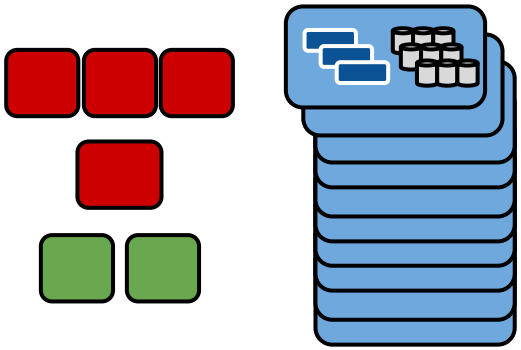


- Performance/load tests
  - up to 900MB/s seq read from real machine on 10GE with fio (30 threads, 4M)
  - rand read 512K: 800MB/s, 4K: 3500IO/s, writes slower: 130MB/s seq, 200IO/s rand
  - 60 fio threads on rbd device plus 60 threads from RADOS bench
  - s/w upgrade under load from 0.56.3 to 0.56.4
    (worked for us, but there upgrade problem reports on ML)
- Community support
  - quick and helpful
  - newbie questions and kernel BUG as examples

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

What we've done so far (4)

- Grizzly OpenStack/Cinder (volumes)
  - create/delete volume
  - create snapshot
  - create vol from snapshot
  - attach fails (qemu/libvirt version!)

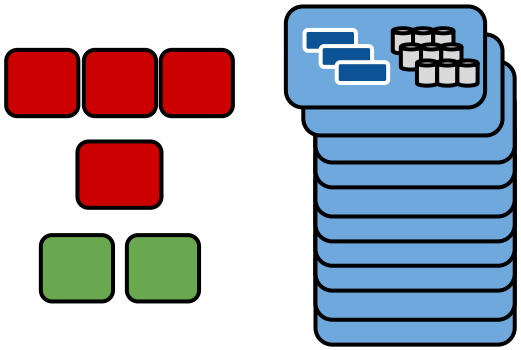


- AFS backend testing
  - rbd device as vicep partition
  - volume served from there
- OwnCloud backend
  - connected with a patched ownCloud version (S3 libs)
  - dir creation worked, everything else failed ...

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

Issues identified so far

- ceph-deploy did not work for us
- CRUSH complexity
  - "2 rooms - 3 replica" problem



- peering is heavy (new OSDs)
- "re-weight" apocalypse
  - too little memory per OSD
  - too many simultaneous weight changes
- flaky server caused Ceph timeouts and constant rebalancing
  - slow server can slow down the cluster
- "budgeting problem"
  - some planning required as placement groups changes difficult

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

Summary

- The feature-rich, open-source Ceph storage system is an interesting candidate to address some of our existing and emerging storage needs.
- The initial results of our investigation look promising and justify further investments.

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