Distributed Object Storage System Ceph in Practice

Dominik Joe Pantůček
dominik.pantucek@trustica.cz

Trustica

8.10.2016

Legal notice.

Distributed Object Storage System Ceph in Practice

Dominik Joe Pantůček
dominik.pantucek@trustica.cz

Trustica

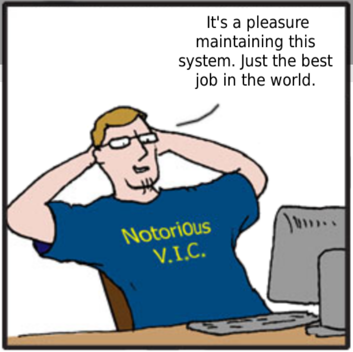
8.10.2016

Why?

- Why daily operations?
- We need convenient deployment of VMs: private IaaS cloud.
- Why cloud?
- It's convenient.
- Why private cloud?
- Negotiations with cloud service providers are tough.
- Why Ceph?
- Everything else failed miserably.

What is Ceph?

- Distributed,
- highly scalable,
- open source,
- object storage...
  - block storage,
  - file system
  - and other applications.


It's a pleasure maintaining this system. Just the best job in the world.
Notorious V.I.C.

A parallel universe, where, thanks to masterfully written Ceph, peace and prosperity prevail.

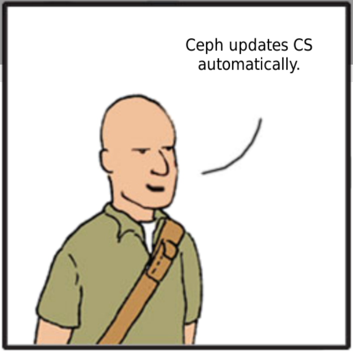
Ceph Architecture

- Daemons:
  - OSD - Object Data storage
  - Monitor
  - Metadata Server – MDS
- Nodes:
  - OSD, Monitor, MDS,
  - client,
  - admin.

Ceph OSD

- Object Storage Device
- Disk or at least part of it.
- The Ceph OSD daemon.
- Multiple OSDs in one physical node,
- managed in pools by monitors.

Ceph Monitor



- Cluster state – cluster map,
- consisting of maps:
    - monitor map,
    - OSD map,
    - placement group map,
    - CRUSH map,
    - MDS map.
- Only 1 monitor and 2 OSDs are needed for bare minimum cluster.

Ceph (OSD) Hierarchy

Thanks to Ceph I have picked up these two bimbos willing to do anything.

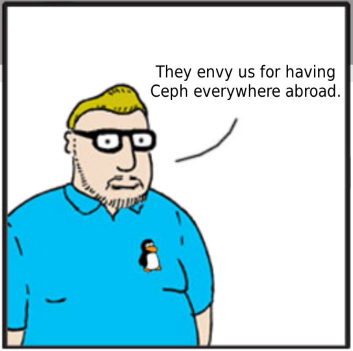
Cluster:

- Node
  - OSD
  - Monitor
  - MDS
  - Client

Objects placement

- No files or directories.
- Object storage stores only objects.
- Behaves like key:value store.
- The value can be really big.
- Find balance:
  - across all OSDs
  - and across nodes.

Placement groups



- Set of OSDs within a pool
- which can store an object.
- Size based upon pool number of replicas.
- Vast number of objects – compared to number of PGs.

CRUSH

- Controlled Replication Under Scalable Hashing.
- Each client computes which PGs to use.
- Describes cluster hierarchy as a weighted tree.
- Selects sets of disks based on deterministic criteria.
- Does not need any central authority.

If it wasn't for this super-tuned system I would never leave this school.

RADOS

- Set of protocols governing CRUSH,
- RBD and
- metadata.
- Reliable Autonomic Distributed Object Store.

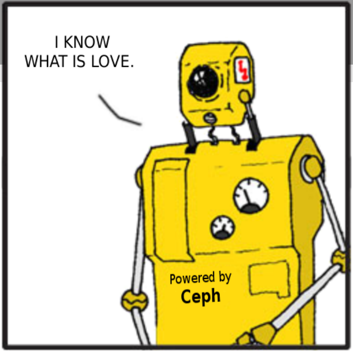
Thanks to Ceph they have legalized marijuana to assist us with our studies.

Monitors

- Cluster management,
- voting,
- osd handling,
- yes, everything.

Ceph File System

- Built on top of the underlying objects,
- POSIX file system with advanced features,
- directory size is reported immediately,
- virtually no limits on file sizes,
- data/metadata separate redundancy settings:
  - different pools.



Metadata Server

System Ceph be better than three young squaw in heated up teepee. Moonwalking Michael knowing what I talk about.

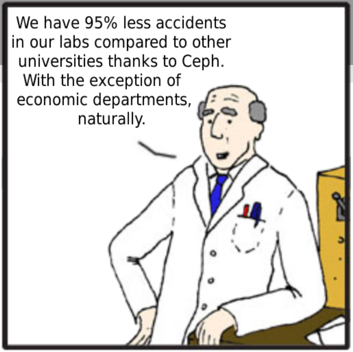
- Separate pool for metadata,
- MDS serves objects metadata,
- reasonable performance,
- required for the actual FS implementation.

In the cloud

Let a twenty-ton partridge fall upon me if there is anything better than our storage system.

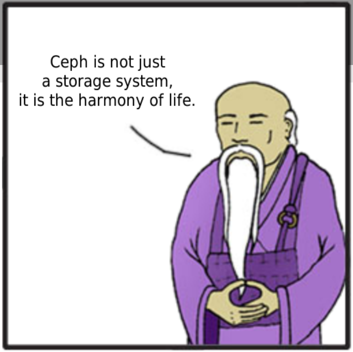
- Amazon Simple Storage Service (S3)?
- Openstack?
- Swift.
- Ceph backend.

RBD for disks


We have 95% less accidents in our labs compared to other universities thanks to Ceph. With the exception of economic departments, naturally.

- Block device,
- from multiple objects,
- naming conventions – prefix,
- provides caching,
- true concurrent access.

Ceph FS for shared data


Ceph is not just a storage system, it is the harmony of life.

- ISO images,
- configurations,
- shared data.

Ceph FS release

Perhaps only thanks to this excellent storage system I found this school and its teachers in perfect order.

- Stable in 10 series,
- available as FUSE.

Kernel support

That Ceph looks like my creation. It's a pity I don't remember making it.

- Both Ceph RBD and File Systems
- Support merged into 2.6.34 (released May 16, 2010)
- Convenient usage
- Ceph FS:

```
mount -t ceph mon1,mon2,mon3:/ /mnt/ceph
```

- RBD:

```
rbd map mypool/mydevice
mkfs -t ext4 /dev/rbd1
```

But...

RBD thin provisioning

- Provision 20PB,
- wait 8 days before de-provisioning finishes
- ...
- profit???

RBD locking

- `$ rbd map`
- On two nodes,
- on one mkfs on other mount,
- everything works,
- another time on single node...
- D indefinitely.
- `$ reboot`

Ceph-FS locking

- Processes in D state any time,
- no way to unmount,
- `$ reboot`

Log clogging

- libceph ...
- Does not prevent the locking bugs.
- Does not even log them ...
- `$ reboot`

Userspace RBD

- RBD as storage backend for VM
- in KVM – librbd in userspace (qemu-kvm)

Userspace Ceph filesystem

- In 10.x ceph-fuse implementation.
- Stable.
- Decent performance.

RBD cache

- Performance is an issue,
- commits are a problem,
- always use backup battery for the controllers.

References

- Ceph, http://ceph.com
- Bugemos — Jojin&HedgeHog: The Chronicles of KOS - Alternativní přítomnost, 2006, available online at http://www.bugemos.com/?q=node/357
- CRUSH: Controlled, Scalable, Decentralized Placement of Replicated Data, Sage A. Weil, Scott A. Brandt, Ethan L. Miller and Carlos Maltzahn, Storage Systems Research Center University of California, Santa Cruz, SC2006 November 2006, Tampa, Florida, USA 0-7695-2700-0/06, available online at http://ceph.com/papers/weil-crush-sc06.pdf

Questions

… and answers.

Thank you!