CEPH AND OPENSTACK

Current integration, roadmap and more!
OpenStack Summit Tokyo
Oct 2015

$ whoarewe

Sébastien Han

Senior Cloud Architect

Blogger

http://sebastien-han.fr/blog

Josh Durgin

Senior Software Engineer

RBD lead

Agenda

1. Ceph?
2. Ceph in Liberty and beyond
3. What's new in Ceph?
4. Mitaka preview

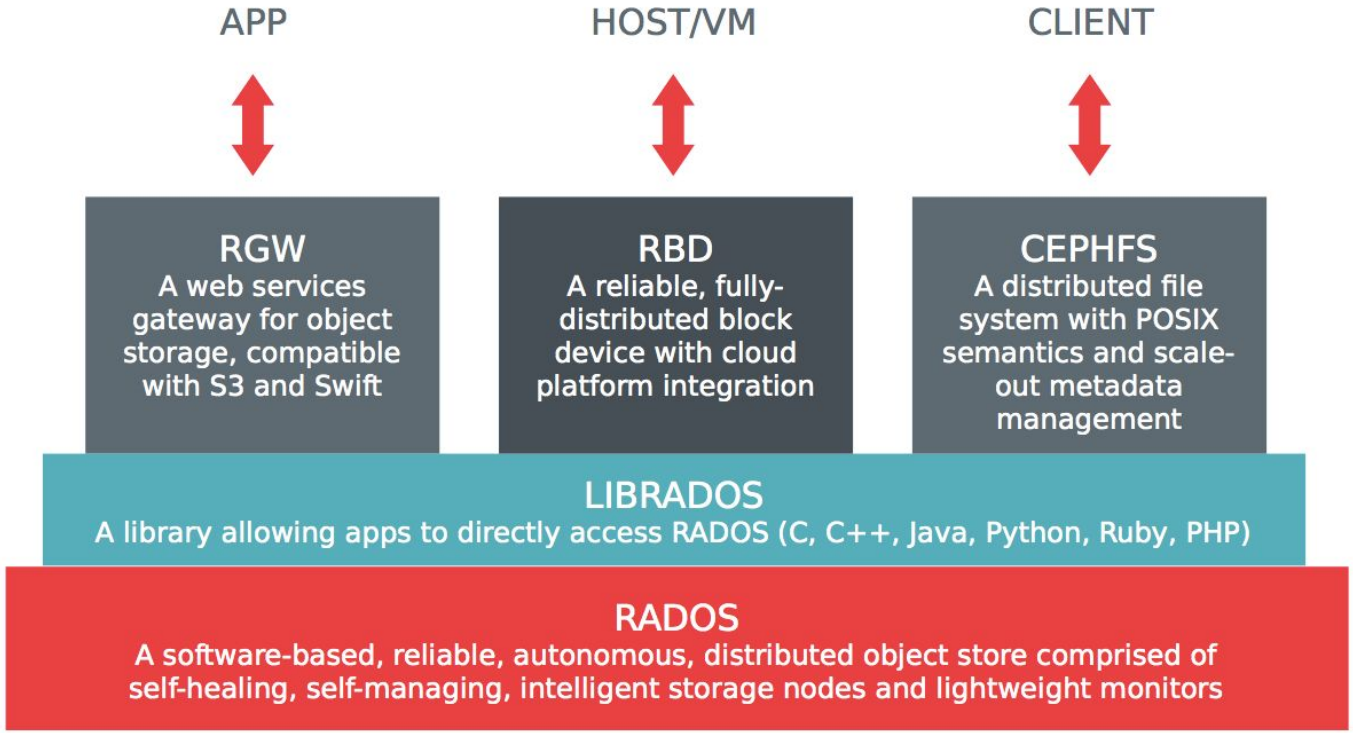
Ceph?

Unified, distributed, replicated open source software defined storage solution

CEPH MOTIVATING PRINCIPLES

- All components must scale horizontally
- There can be no single point of failure
- The solution must be hardware agnostic
- Should use commodity hardware
- Self-manage wherever possible
- Open Source (LGPL)
- Move beyond legacy approaches
  - client/cluster instead of client/server
  - Native rather than ad hoc HA

CEPH OVERVIEW



RADOS

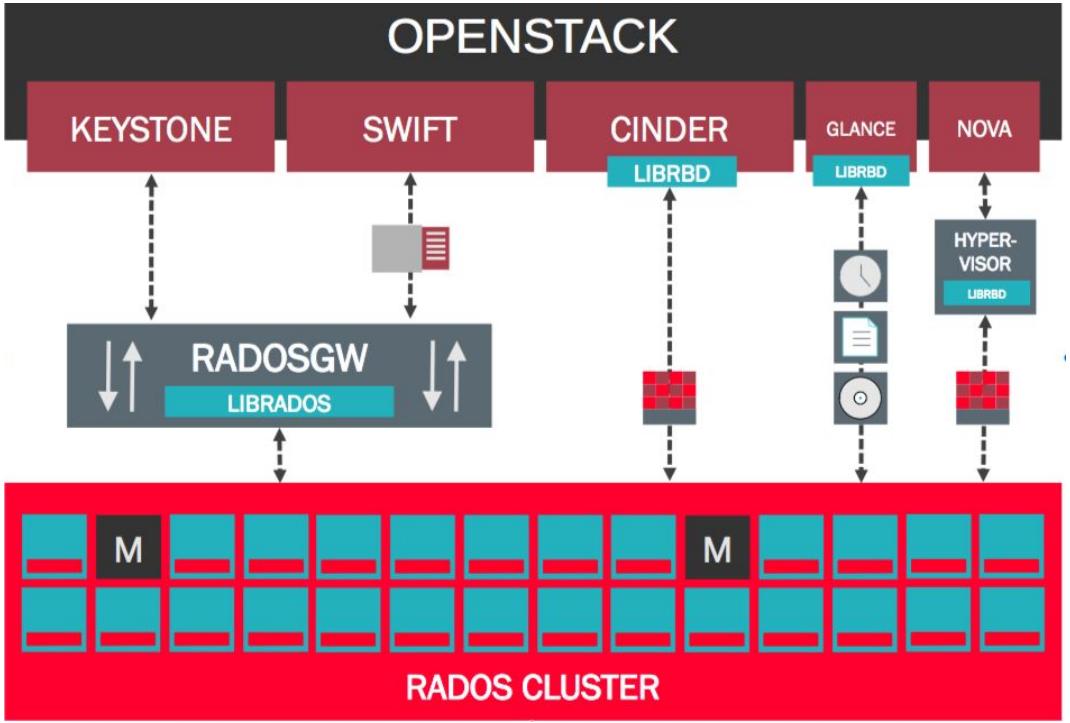
- Monitors
  - maintain cluster map
  - provide consensus for distributed decision making
  - should have an odd number (usually 3 or 5)
  - not in the data path
- Object Storage Daemons (OSDs)
  - one per disk
  - serve stored data (objects) to clients
  - intelligently coordinate to maintain data integrity and replication level

CRUSH

Controlled **R**eplication **U**nder **S**calable **H**ashing:

- Pseudo-random data placement algorithm
- Statically uniform distribution
- Rule-based configuration
- Topology aware

CEPH IN OPENSTACK



CEPH IN LIBERTY AND BEYOND

LIBERTY - FIXED BUGS

- Fix QoS for Nova ephemeral disks
- Add retries to delete a volume in the RBD driver
- Fix backup metadata import missing fields
- Handle config drives being stored on rbd
- Fix restore point if backup base is diff-format in ceph
- Long-running rbd calls moved to separate threads
- Cinder max_clone_depth option fixed

LIBERTY - FEATURES AND IMPROVEMENTS

- Support for Cinder volume migration
- Add ability for Cinder backend to report discard/unmap/trim (spec)
- use rbd_default_features from ceph.conf
- Ceph driver support retries on rados_connect_timeout
- rbd driver in cinder tries all glance image locations
- Cinder support for custom cluster names

CINDER VOLUME MIGRATION

- Using ‘`cinder retype --migration-policy on-demand`’
- Works in the following cases:
  - LVM to LVM (available and in-use)
  - LVM to/from NFS (available and in-use)
  - NFS to/from Ceph (available)
  - Ceph to LVM (in-use)
  - LVM to Ceph (available)
  - Ceph to Ceph (available)
- Do not use with attached devices

WHAT'S NEW IN CEPH?

INFERNALIS - CEPH

- Per-image metadata
  - enable rbd config options in the image itself (stripe, readahead)
- Deep flatten
  - snapshots flatten
  - parent images can be deleted more easily
- Faster diff (with object map)
- Dynamically enable new features on an image
- rbd du
- Groundwork for rbd mirroring

INFERNALIS - CEPH

- RGW supports the swift api for object expiration
- internal buffer and mutex tuning, other perf gains
- proxy writes for cache pools
- systemd support (still upstart on trusty)
- SHEC erasure coding plugin
- better defaults for recovery settings
- unified queue for client I/O and internal tasks
- improved pool quota and cluster full handling

JEWEL - CEPH

- RBD mirroring
- Easier to use and active/active multi-site RGW
- Cephfs fsck and repair
- prototype for client QoS
- stabilizing async messenger
- performance improvements - esp. writes

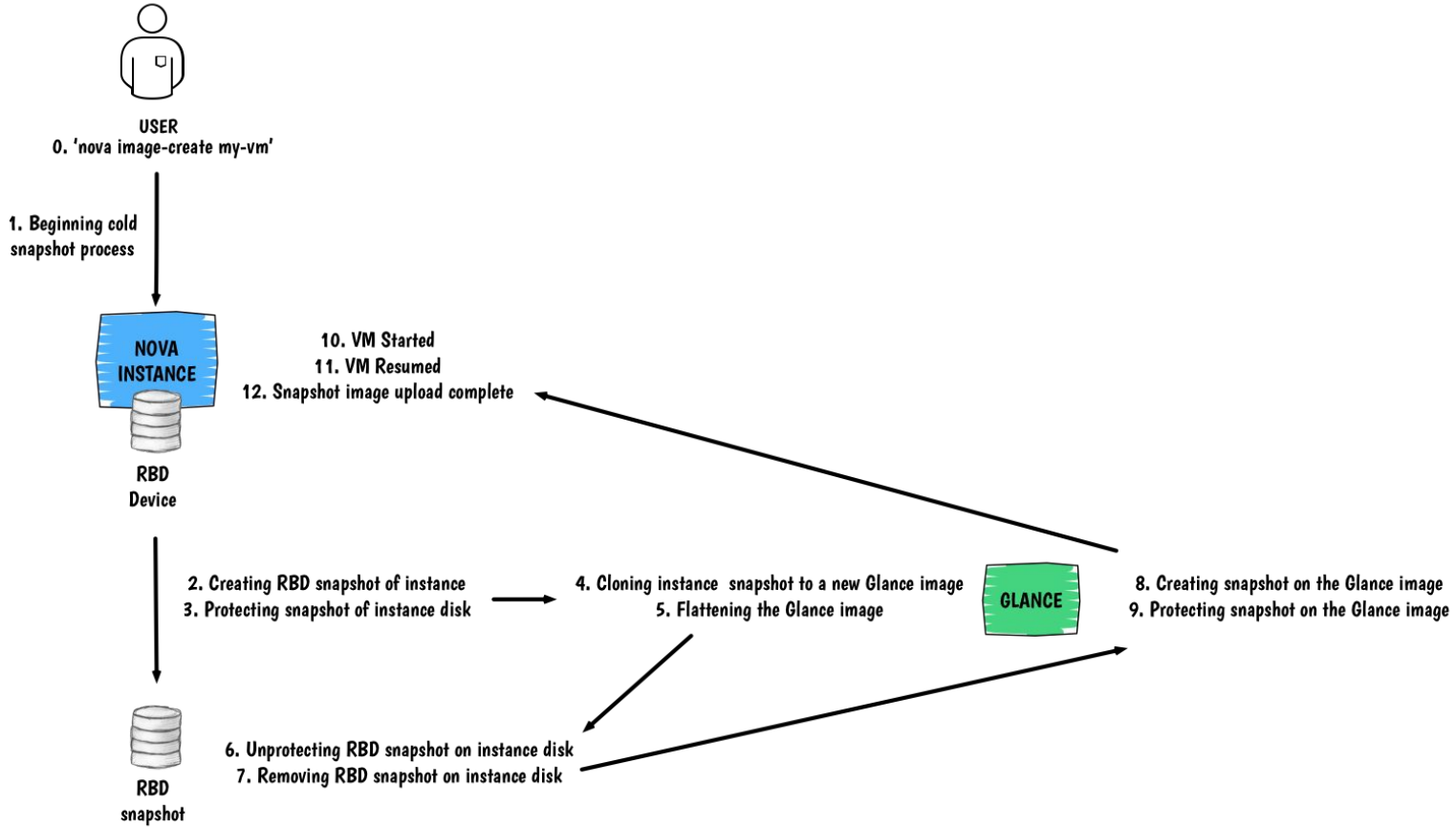
MITAKA PREVIEW

NOVA EPHEMERAL SNAPSHOTS

- Crucial for Public Clouds
- Final barrier to diskless compute nodes
- If not configured right, we fallback to the original method on local disk

```
[libvirt]
snapshots_directory = /fail/safe/path
```

NOVA ROOT EPHEMERAL RBD SNAPSHOTS



FUTURE OPENSTACK IMPROVEMENTS

- Attach the same volume to multiple instances
- Optimize volume migration and creating images from volumes
- Thin provisioning reporting
- Force detach support
- Online volume migration from ceph to ceph
- Volume encryption via qemu

DOCUMENTATION

http://ceph.com/docs/master/rbd/rbd-openstack

THANK YOU

COME SEE US AT THE RED HAT BOOTH

Sébastien Han | seb@redhat.com | @sebastien_han | leseb on irc
Josh Durgin | jdurgin@redhat.com | jdurgin on irc

PERFORMANCE TUNING

OS tuning

- Disable osd directory parsing by updatedb
- Disable transparent hugepage
  - tiny allocations won’t benefit from that
- Kernel values:
  - kernel.pid_max, value: 4194303
  - fs.file-max, value: 26234859 (clients only)
  - vm.zone_reclaim_mode, value: 0 (numa and page cache)
  - vm.vfs_cache_pressure, value: 50 (mitigate kernel’s behaviour)

Tuning for more IOPS

- Disable in-memory logging
- `max_open_files` to 131072 or higher
- For all-ssd setups, change osd settings:

```
filestore_op_threads = > default of 2 (hardware dependent)
filestore_max_sync_interval = 1
filestore_min_sync_interval = 0.01 (default)
throttler_perf_counter = false
osd_enable_op_tracker = false
```