CEPH AND OPENSTACK

Current integration, roadmap and more!
OpenStack Summit Vancouver
May 2015

$ whoarewe

Sébastien Han

Senior Cloud Architect

Blogger

http://sebastien-han.fr/blog

Josh Durgin

Senior Software Engineer

RBD lead

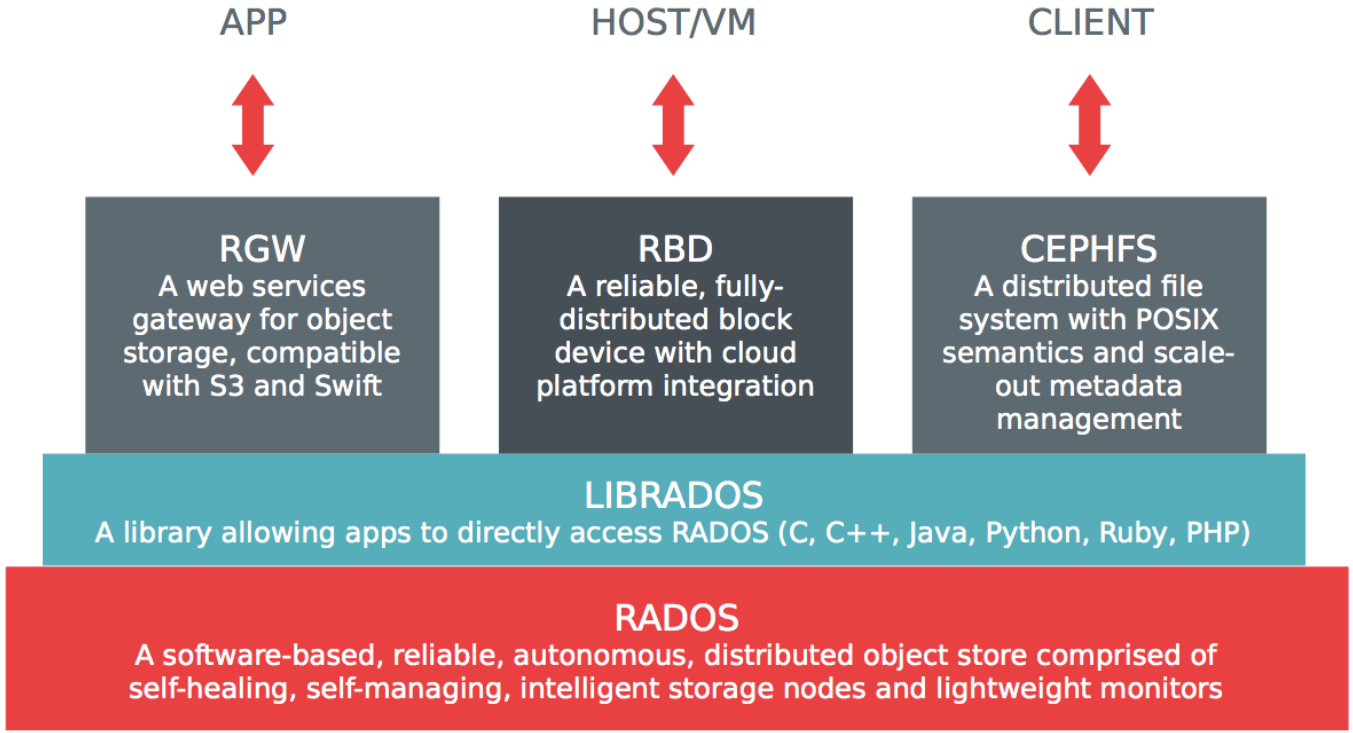
Agenda

1. Ceph?
2. Ceph in Kilo, Liberty and beyond
3. What's new in Ceph?
4. Get the best cloud configuration

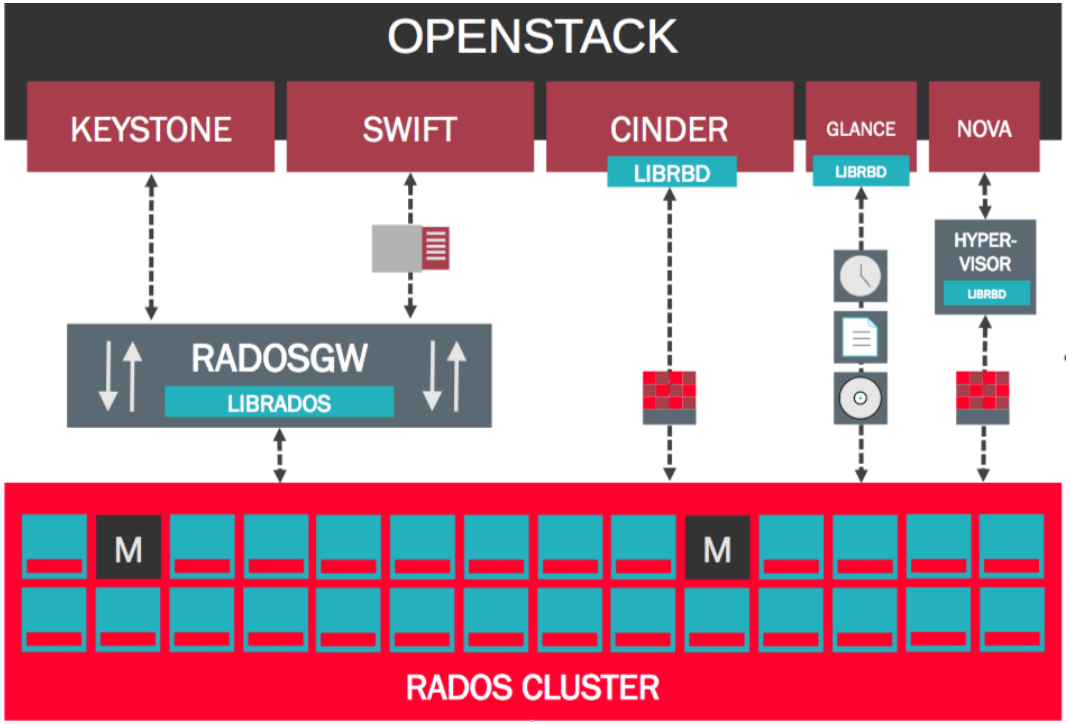
Ceph?

Unified, distributed, replicated open source storage solution

CEPH


APP
HOST/VM
CLIENT
RGW
A web services gateway for object storage, compatible with S3 and Swift
RBD
A reliable, fully-distributed block device with cloud platform integration
CEPHFS
A distributed file system with POSIX semantics and scale-out metadata management
LIBRADOS
A library allowing apps to directly access RADOS (C, C++, Java, Python, Ruby, PHP)
RADOS
A software-based, reliable, autonomous, distributed object store comprised of self-healing, self-managing, intelligent storage nodes and lightweight monitors

CEPH IN OPENSTACK



CEPH IN KILO, LIBERTY AND BEYOND

KILO - FEATURES

- Partially implements RBD snapshots instead of QEMU snapshots
- Basic image conversion in Glance
- DevStack Ceph and remote cluster
- Ceph CI in the gate
- Ceilometer integration for RGW
- Retype to change QoS for Cinder volumes
- Future-proofing for new RBD features

KILO - FIXED BUGS

- nova evacuate
- cinder clone volume size
- nova instance resize revert
- nova disk space stats (hypervisor)
- cinder create-from-image with kernel and ramdisk

LIBERTY - NOVA

- Implement force detach-volume
- QEMU throttling for ephemeral disks
- Multi-attach for RBD
- Nova snapshots for RBD-backed ephemeral disks

LIBERTY - CINDER

- Volume migration
- Support retype volume
- Import/export snapshots and RBD
- Update Ceph backup driver to support differential backup API
- Differential backup orchestration (scheduled)
- Multi-attach for RBD

M...

- Consistency groups in Cinder
- RBD mirroring integration
- Integrate with Manila:
  - RBD and NFS export
  - CephFS with Ganesha
  - See Sage's talk tomorrow:

    Keeping OpenStack storage trendy with Ceph and containers

    Room 109 2:40pm

WHAT'S NEW IN CEPH?

HAMMER - CEPH

- Many optimizations
  - cache hints, allocation hints, readahead during boot…
- Copy-on-read
- Exclusive locking
  - improved infrastructure, handled automatically by rbd
  - groundwork for future single-writer features
- Object Map
  - better performance for clones
  - efficient space tracking

INFERNALIS - CEPH

- Per-image metadata
  - enable rbd config options in the image itself
- Deep flatten
- Faster diff (with object map)
- Dynamically enable new features on an image
- rbd du
- Groundwork for rbd mirroring

GET THE BEST CLOUD CONFIGURATION

ceph.conf - On your hypervisors

```
[client]
rbd cache = true
rbd cache writethrough until flush = true
rbd concurrent management ops = 20
admin socket = /var/run/ceph/$cluster-$type.$id.$pid.$cctid.asok
log file = {{ rbd_client_log_file }}

# MAKE SURE BOTH LOG AND SOCKET PATHS ARE QEMU WRITABLE AND ALLOW SELinux
```

GLANCE - glance-api.conf

- Disable local cache:
  - `s/flavor = keystone+cachemanagement/flavor = keystone/`
- Expose images URL:
  - `show_image_direct_url = True`

GLANCE - Images type

- Use RAW images:
  - convert with qemu-img directly in Ceph if the image is already in Ceph
- If not in Ceph:

```
$ qemu-img convert -f qcow2 -O raw fedora21.img fedora21.raw
```

- If in Ceph:

```
$ qemu-img convert -O raw rbd:$pool/$uuid rbd:$pool/$uuid
$ rbd --pool images snap create --snap snap $uuid
$ rbd --pool images snap protect --image $uuid --snap snap
```

GLANCE - Images metadata

- `hw_scsi_model=virtio-scsi # for discard and perf`
- `hw_disk_bus=scsi`

NOVA - nova.conf

Configure discard in your nova.conf with:

```
[libvirt]
hw_disk_discard = unmap # enable discard support (be careful of perf)
inject_password = false # disable password injection
inject_key = false # disable key injection
inject_partition = -2 # disable partition injection
disk_cachemodes = "network=writeback" # make QEMU aware so caching works
live_migration_flag="VIR_MIGRATE_UNDEFINE_SOURCE,VIR_MIGRATE_PEER2PEER,
VIR_MIGRATE_LIVE,VIR_MIGRATE_PERSIST_DEST"
```

NOVA

- Configure hypervisors with a specific Ceph backend
  - Add “`AggregateInstanceExtraSpecsFilter`” scheduler filter
  - Host aggregates exposed through AZ and pinned to flavors
    - One AZ per virtual machine backend

CINDER - `cinder.conf`

- Use Glance API v2:
  - `glance_api_version = 2`

- Use the multi-backend and apply QoS policies per backend

CINDER BACKUP

Only one valid use case at the moment:

- 1 Openstack installation
- 2 Ceph clusters

Known issue:

- Importing metadata and restoring a volume on another Openstack installation doesn’t work

GUESTS

Configure the qemu-guest-agent along with fsfreeze hooks so you can perform consistent snapshots/backups.

This works with the Glance metadata `os_require_quiesce=yes`

DOCUMENTATION

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

THANK YOU

COME SEE US AT THE OPEN SOURCE LOUNGE

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