Ceph & OpenStack

An Overview

OpenStack BLR Meetup, June 2015
Deepak C Shetty (deepakcs)
Why Ceph ?
Ceph in OpenStack - 2014

Block Storage Drivers

- Ceph RBD: 47%
- LVM (default): 29%
- GlusterFS: 21%
- NFS: 12%
- NetApp: 10%
- EMC: 9%
- HP 3PAR: 9%
- HP LeftHand: 6%
- VMware VMDK: 3%
- SolidFire: 3%
- Dell EqualLogic: 3%
- IBM GPFS: 4%
- IBM XIV/DS8000: 2%
- Nexenta: 2%
- SAN/Solaris: 2%
- All Others: 17%

Ceph in OpenStack - 2015

Source:
http://superuser.openstack.org/articles/openstack-users-share-how-their-deployments-stack-up
What is Ceph?

- Distributed storage system
- Algorithmic placement - CRUSH
- No single point of failure
- Self healing and self managing
- Runs on commodity hardware
  - No vendor lockin!
- Opensource
  - GPLv2 License
  - Community driven

Probably, one of the best examples of SDS (aka Software Defined Storage)
Storage that scales

PAST: SCALE UP

FUTURE: SCALE OUT
Ceph - Architecture

**APP**
- RGW
  - A web services gateway for object storage, compatible with S3 and Swift

**HOST/VM**
- RBD
  - A reliable, fully-distributed block device with cloud platform integration

**CLIENT**
- CEPHFS
  - A distributed file system with POSIX semantics and scale-out metadata management

**LIBRADIOS**
- 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

- Not Production ready yet!
RADOS COMPONENTS

OSDs:
- 10s to 10000s in a cluster
- One per disk (or one per SSD, RAID group...)
- Serve stored objects to clients
- Intelligently peer for replication & recovery

Monitors:
- Maintain cluster membership and state
- Provide consensus for distributed decision-making
- Small, odd number
- These do not serve stored objects to clients
OBJECT STORAGE DAEMONS

- btrfs
- xfs
- ext4
- zfs?
CRUSH IS A QUICK CALCULATION
CRUSH: DYNAMIC DATA PLACEMENT

CRUSH:
- Pseudo-random placement algorithm
  - Fast calculation, no lookup
  - Repeatable, deterministic
- Statistically uniform distribution
- Stable mapping
  - Limited data migration on change
- Rule-based configuration
  - Infrastructure topology aware
  - Adjustable replication
  - Weighting
ACCESSING A RADOS CLUSTER

Application

Librados

Object

Socket

Rados Cluster
LIBRADOS: RADOS ACCESS FOR APPS

LIBRADOS:
- Direct access to RADOS for applications
- C, C++, Python, PHP, Java, Erlang
- Direct access to storage nodes
- No HTTP overhead
THE RADOS GATEWAY

APPLICATION

RADOSGW
LIBRADOS

RADOSGW
LIBRADOS

APPLICATION

REST

socket

RADOS CLUSTER
RADOSGW MAKES RADOS WEBBY

RADOSGW:
- REST-based object storage proxy
- Uses RADOS to store objects
- API supports buckets, accounts
- Usage accounting for billing
- Compatible with S3 and Swift applications
STORING VIRTUAL DISKS

VM

HYPervisor

LIBRBD

RADOS CLUSTER

M

M
SEPARATE COMPUTE FROM STORAGE
KERNEL MODULE FOR MAX FLEXIBLE!

LINUX HOST

KRBD

RADOS CLUSTER
RBD STORES VIRTUAL DISKS

RADOS BLOCK DEVICE:
- Storage of disk images in RADOS
- Decouples VMs from host
- Images are striped across the cluster (pool)
- Snapshots
- Copy-on-write clones
- Support in:
  - Mainline Linux Kernel (2.6.39+)
  - Qemu/KVM, native Xen coming soon
  - OpenStack, CloudStack, Nebula, Proxmox
Ceph - Unified Storage for OpenStack
Ceph as Unified Storage for OpenStack - Adv.

- No storage silos
  - Deploy/Manage 1 cluster with diff. pools
- Create image from volume and vice-versa optimizations
- Nova boot from volume optimizations
- Live migration
- Volume retype/migrate optimizations possible (WIP)
- Cinder Backup optimizations
  - Full and differential
- Cinder Volume replication (DR) made efficient via RBD mirroring (WIP)
Ceph & OpenStack Storage - Summary

- Object Storage like Swift
  - Ceph RADOSGW as a drop-in replacement for OpenStack Swift
- Block Storage in Cinder
  - Ceph RBD pool for storing Cinder volumes
- Ephemeral Storage in Nova
  - Ceph RBD pool as backend for Ephemeral storage
  - Nova boot from volume
- Image Storage in Glance
  - Ceph RBD pool as a glance image store
- Backup target for Cinder-Backup
  - Ceph RBD pool as a backup target for Cinder
  - Backup / Restore cinder volumes to/from Ceph RBD pool
- File Storage in Manila (upcoming / future)
  - CephFS as a backend for Manila FS shares
References

• Get the best configuration for your cloud
  – Devil is in the details
  – http://ceph.com/docs/master/rbd/rbd-openstack

• Ceph and openstack, current integration, roadmap (Vancouver summit prez)
  – http://www.slideshare.net/Red_Hat_Storage/open-stack-ceph-liberty
• Disclaimer

– Most (if not all) content for this prez taken from slideshare, youtube videos, ceph.com docs & other publicly available presentations.