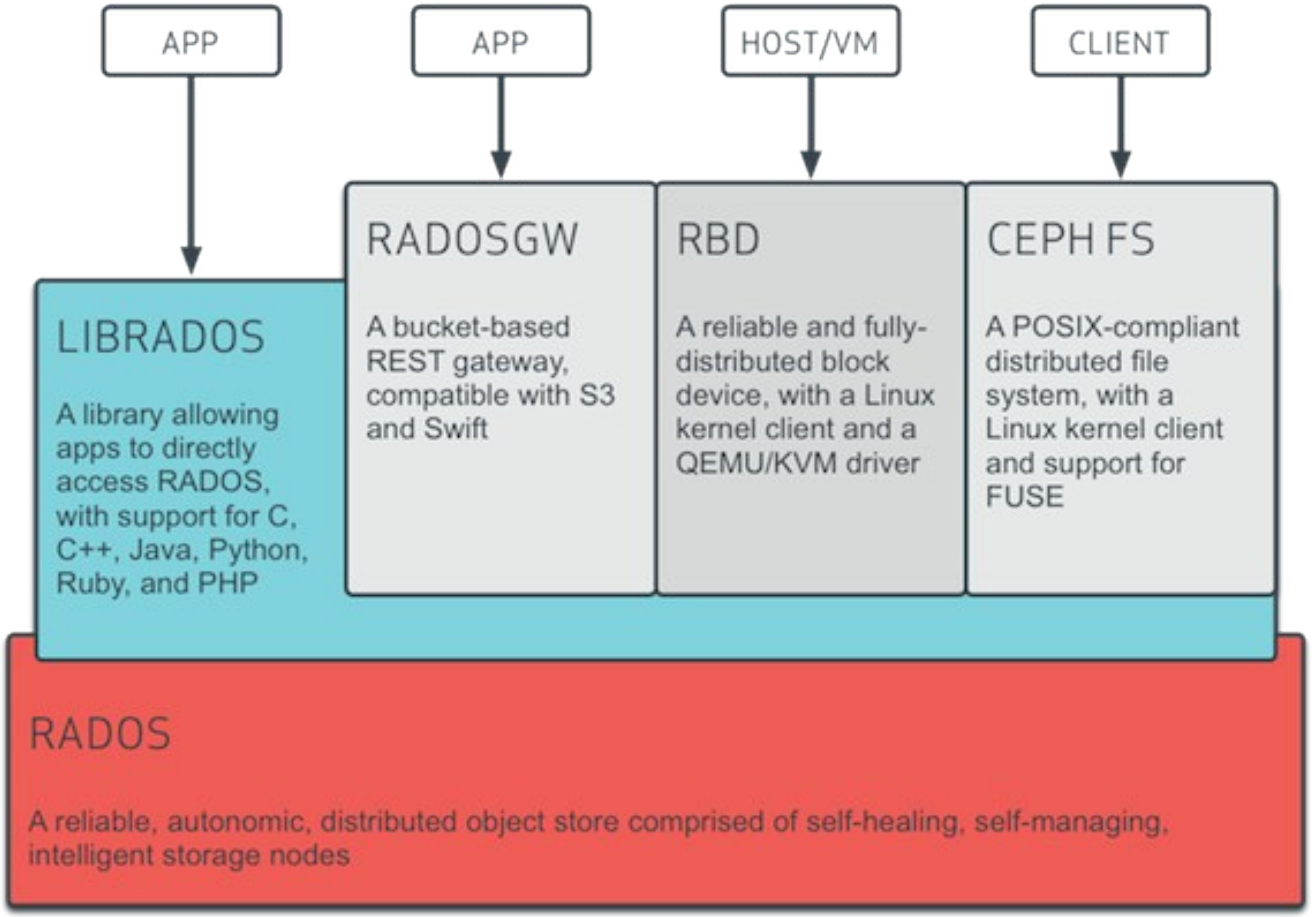
Red Hat Ceph Storage (RHCS), An Intro.

Vimal A.R
vimal@redhat.com

RHCS/Ceph Introduction

- Open source distributed storage solution
- Highly scalable, supports both scale-up and scale-out
- Built around the CRUSH algorithm, by Sage Weil
  - http://ceph.com/papers/weil-crush-sc06.pdf
- Inktank, acquired by Red Hat in April 2014
- Supports multiple access methods [File, Block, Object]

Ceph Architecture



RHCS Interfaces

- Rados Block Device (RBD) [Block device interface]
  - Kernel driver – krbd
  - Libvirt integration – librbd
- Rados Gateway (RGW) [ReST Interface for S3/Swift]
- CephFS [Filesystem interface]
- Librados API (Supports several language bindings)
  - Supports custom applications

RHCS Components

- Monitor nodes [MONs]
- Object Storage nodes [OSD]
  - Minimum of 1 MON node and 3 OSD nodes

RHCS Components (MONs)

- Maintains the state of the entire cluster, via maps.
- PG, OSD, MON, CRUSH, and MDS maps
- Serve the maps to all nodes in the cluster.
- Pushes updated maps to all nodes, including clients.
- Clients first connect to the MONs to get the maps.
- MONs use leveldb to store the maps
  - https://github.com/google/leveldb
- Uses Paxos protocols to agree upon the cluster maps
  - https://goo.gl/R89auQ
- MONs rely on the MONmaps to find other MONs
- Co-located MONs and OSDs – not suggested/supported.
- OSDs update MONs with their status, as well as their peers.

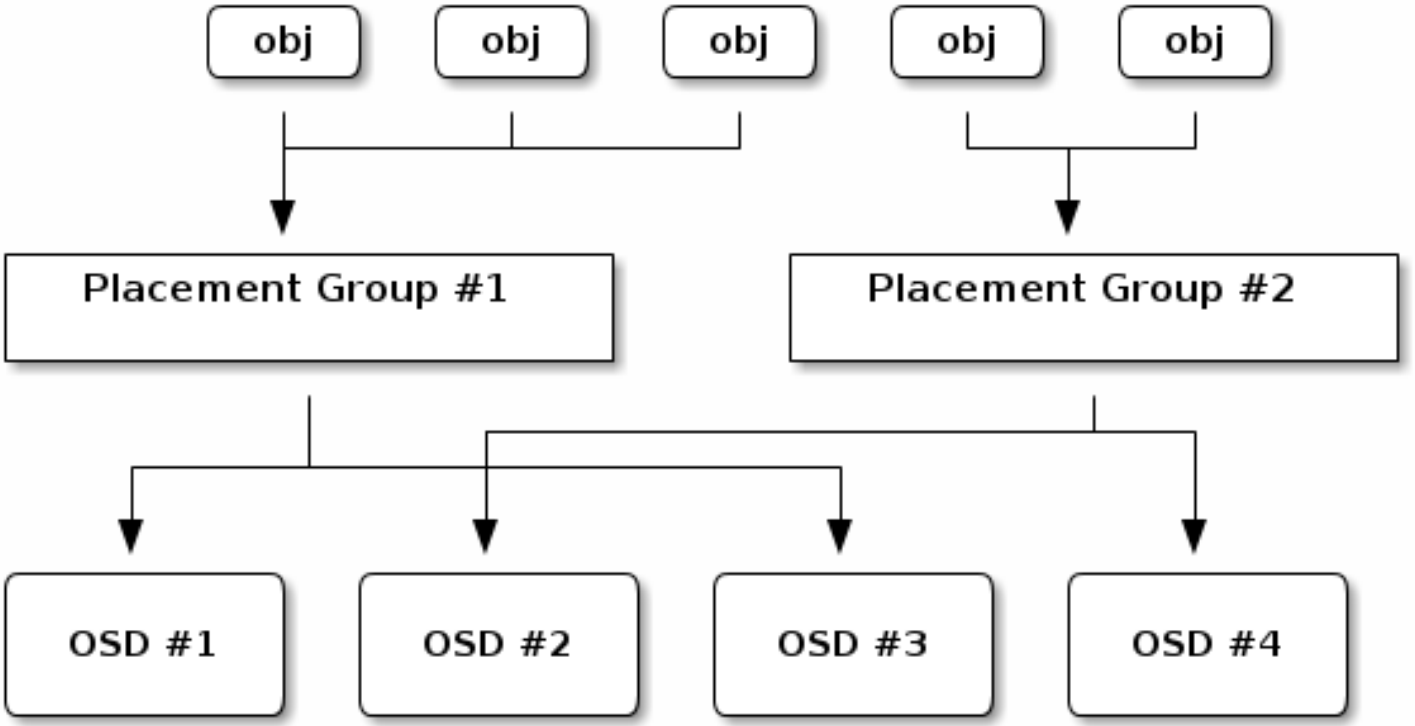
RHCS Components (OSDs)

- Serves the backend storage, a single 'ceph-osd' process for each disk
- Storage disk → File system → `ceph-osd` process
- Disks mounted at /var/lib/ceph/osd/$(cluster)-OSD#/
- OSDs use journals, the same OSD disk by default.
  - SSD disks in production, for speed.
  - Upto 6 OSDs (supported) can use an SSD disk, as its journal.
- Heartbeat check runs every 6 seconds between OSDs
- Replicates the data between other OSDs, as well as peering.
- Daily and weekly data scrub

RADOS Components

- Pools
  - Logical bucket where data is written to and read from.
- Placement Groups (PG)
  - Logical grouping of objects
- Object Storage Daemons/Disks
  - A trio of OSD process, File system, and Storage

RADOS Components



Client interaction

- A client connection needs:
  - The Ceph configuration file (/etc/ceph/ceph.conf)
  - The pool name
  - A keyring file created with proper pool permissions
- How a client connects/writes to the cluster
  - Contacts the MONs listed in the Ceph conf file
  - Gets the map sets [CRUSH, MON, PG, OSD, MDS maps]
  - Finds the placement groups responsible in the pool
  - Finds the primary OSD for the Pgs
  - Writes directly to the Primary OSD
  - The primary OSD replicates to the secondary/tertiary OSDs
  - Acks the client.

CRUSH map

- Defines the architecture of the cluster
- Defines failure/performance domains
- Define weights for specific buckets/hosts/OSDs
- Example:-

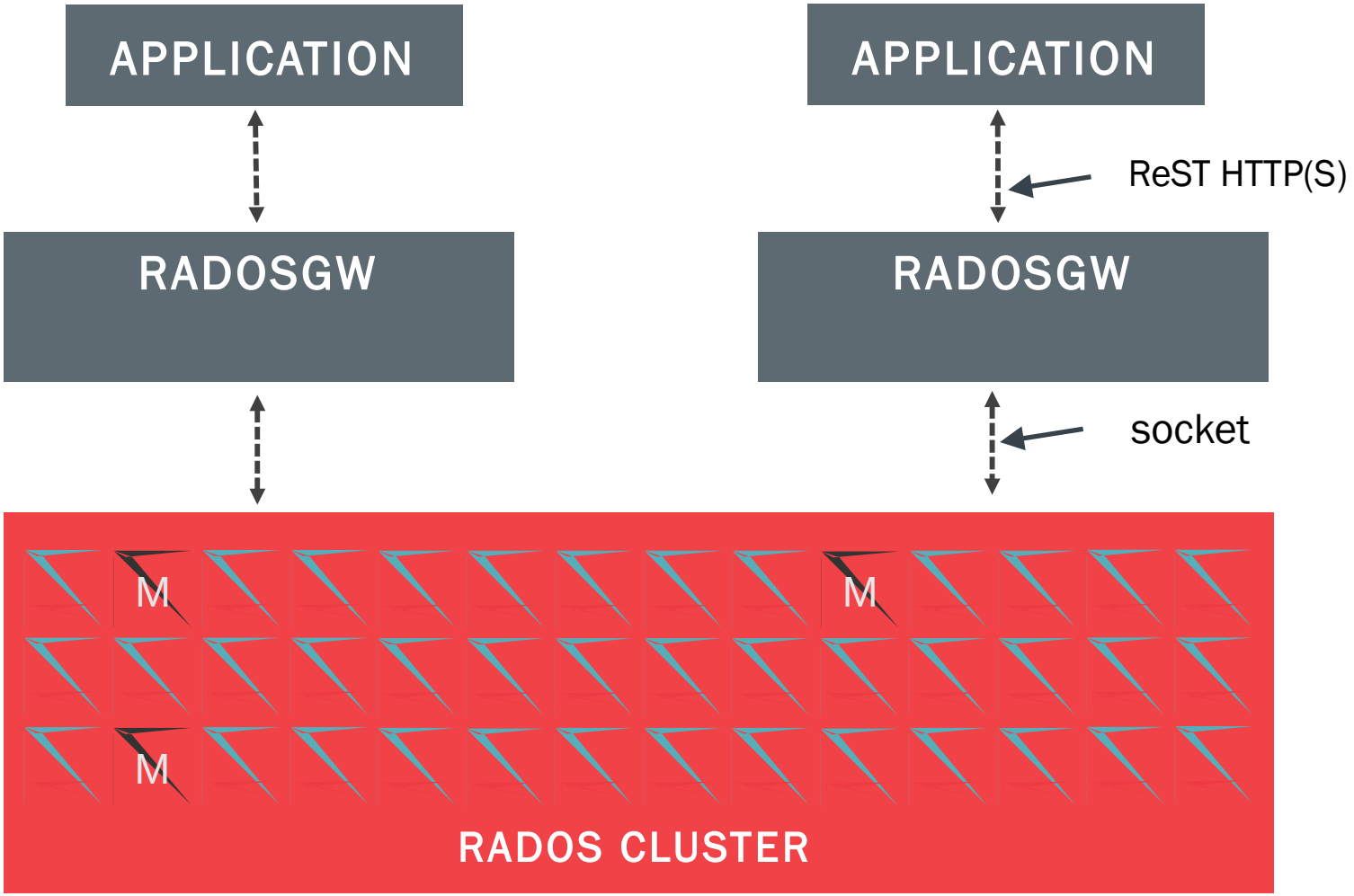
Rados Block Device [RBD]

- Serves block storage to the clients

Rados Gateway [RGW]

- Object storage interface emulating both Amazon S3 and Openstack Swift.
- Accessible through a ReST-ful HTTP interface
- ReST APIs for Amazon S3 and OpenStack Swift protocols
  - http://docs.ceph.com/docs/master/radosgw/s3/
  - http://docs.ceph.com/docs/master/radosgw/swift/
- Supports Regions, Zones, Users, ACLs, Quotas etc.. similar to S3/Swift
- Flickr's RGW Object store case study (Not RHCS) http://goo.gl/uS5V3I

Rados Gateway (Continued)



Rados Gateway (Continued)

Calamari

- Monitoring interface for Ceph clusters
- Being replaced with 'Unified Storage Management' interface in RHCS2.0.

Further info

- Upstream documentation : http://docs.ceph.com/docs
- Red Hat documentation : https://access.redhat.com/documentation/en/red-hat-ceph-storage/

Thank you!