redhat.

INTRODUCTION TO CEPH

Orit Wasserman
Red Hat
August Penguin 2017

CEPHALOPOD

A cephalopod is any member of the molluscan class Cephalopoda. These exclusively marine animals are characterized by bilateral body symmetry, a prominent head, and a set of arms or tentacles (muscular hydrostats) modified from the primitive molluscan foot. The study of cephalopods is a branch of malacology known as teuthology.

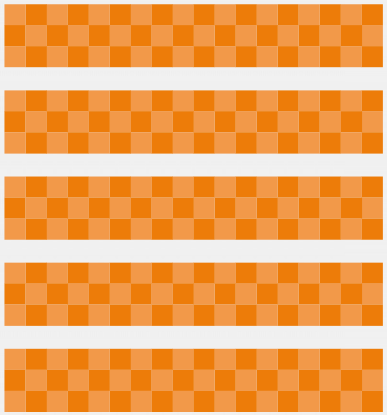
OPEN SOURCE

- Code: https://github.com/ceph
- IRC: OFTC #ceph,#ceph-devel
- Mailing list:
  - ceph-users@ceph.com
  - ceph-devel@ceph.com
- Documentation: http://docs.ceph.com/docs/master/

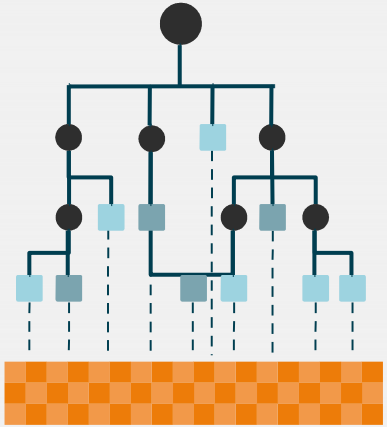
SOFTWARE DEFINE STORAGE

PROPRIETARY HARDWARE	**Common, off-the-shelf hardware**	Lower cost, standardized supply chain
SCALE-UP ARCHITECTURE	**Scale-out architecture**	Increased operational flexibility
HARDWARE-BASED INTELLIGENCE	**Software-based intelligence**	More programmability, agility, and control
CLOSED DEVELOPMENT PROCESS	**Open development process**	More flexible, well-integrated technology

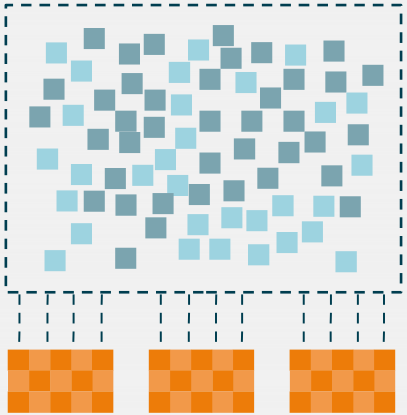
UNIFIED STORAGE



BLOCK STORAGE
Physical storage media appears to computers as a series of sequential blocks of a uniform size.



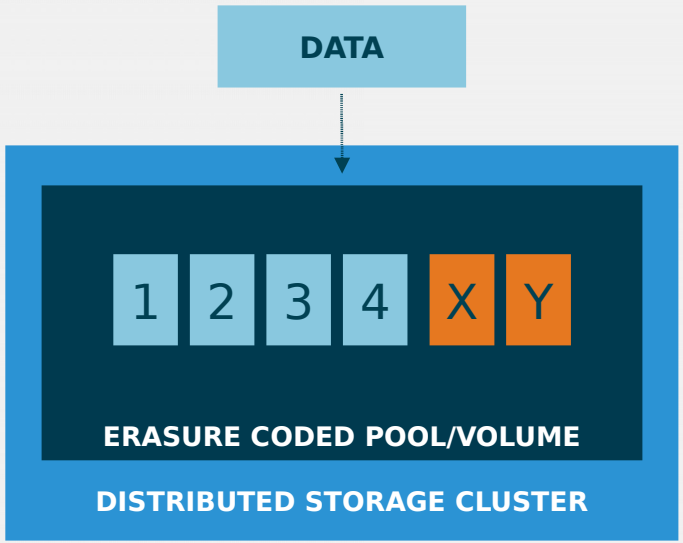
FILE STORAGE
File systems allow users to organize data stored in blocks using hierarchical folders and files.

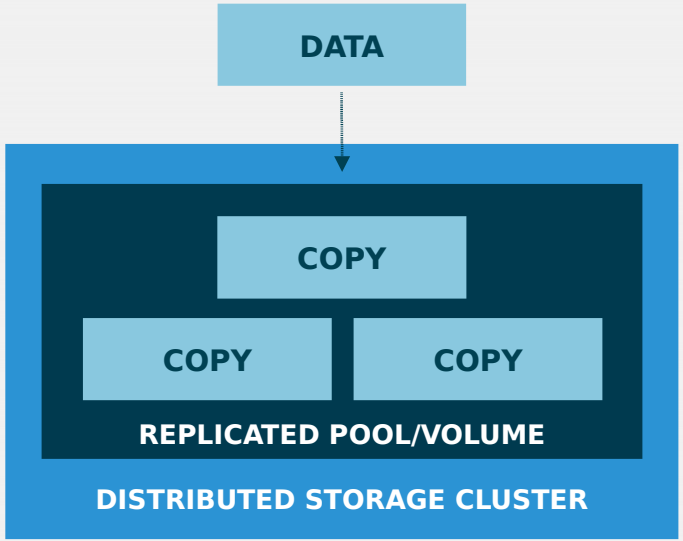


OBJECT STORAGE
Object stores distribute data algorithmically throughout a cluster of media, without a rigid structure.

DISTRIBUTED AND MASSIVE SCALE

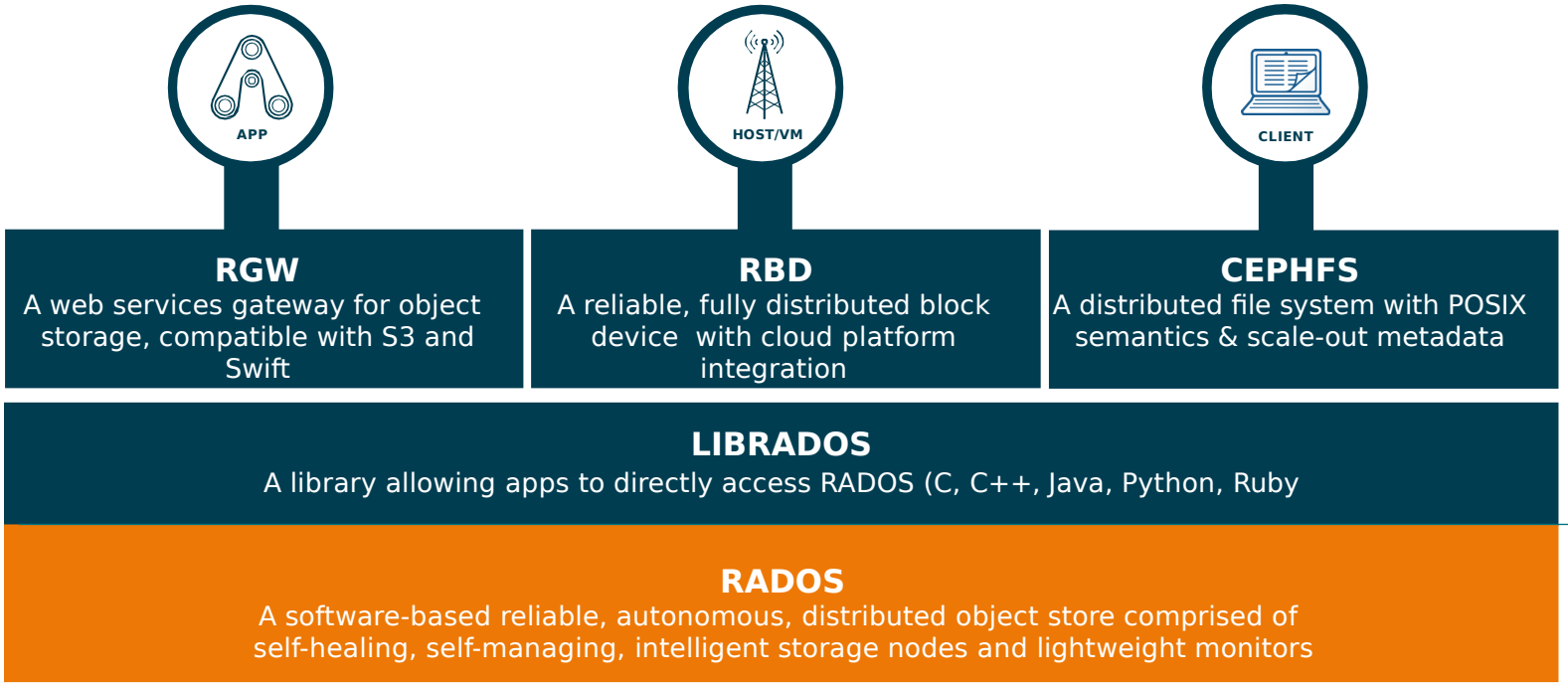
- No single point of failure
- Self healing





CEPH ARCHITECTURE

ARCHITECTURAL COMPONENTS



RADOS

- Reliable Autonomic Distributed Object Storage
- Replication and Erasure coding
- Flat object namespace within each pool
  - Different placement rules
- Strong consistency (CP system)
- Infrastructure aware, dynamic topology
- Hash-based placement (CRUSH)
- Direct client to server data path

OSD

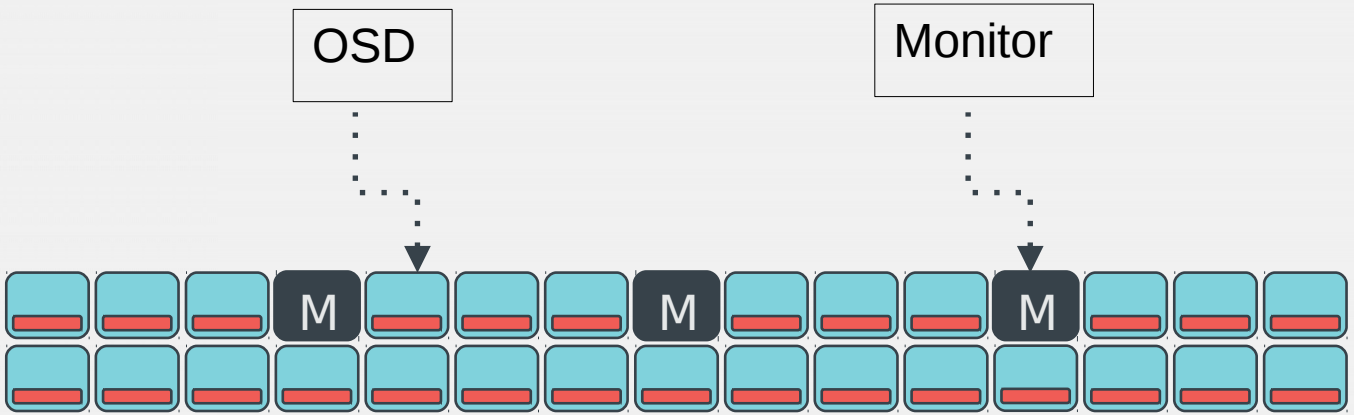
- 10s to 10000s in a cluster
- One per disk (or one per SSD, RAID group…)
- Serve stored objects to clients
- Intelligently peer for replication & recovery

MONITOR

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

RADOS CLUSTER

- 3-5 Monitors
- 1000s of OSD

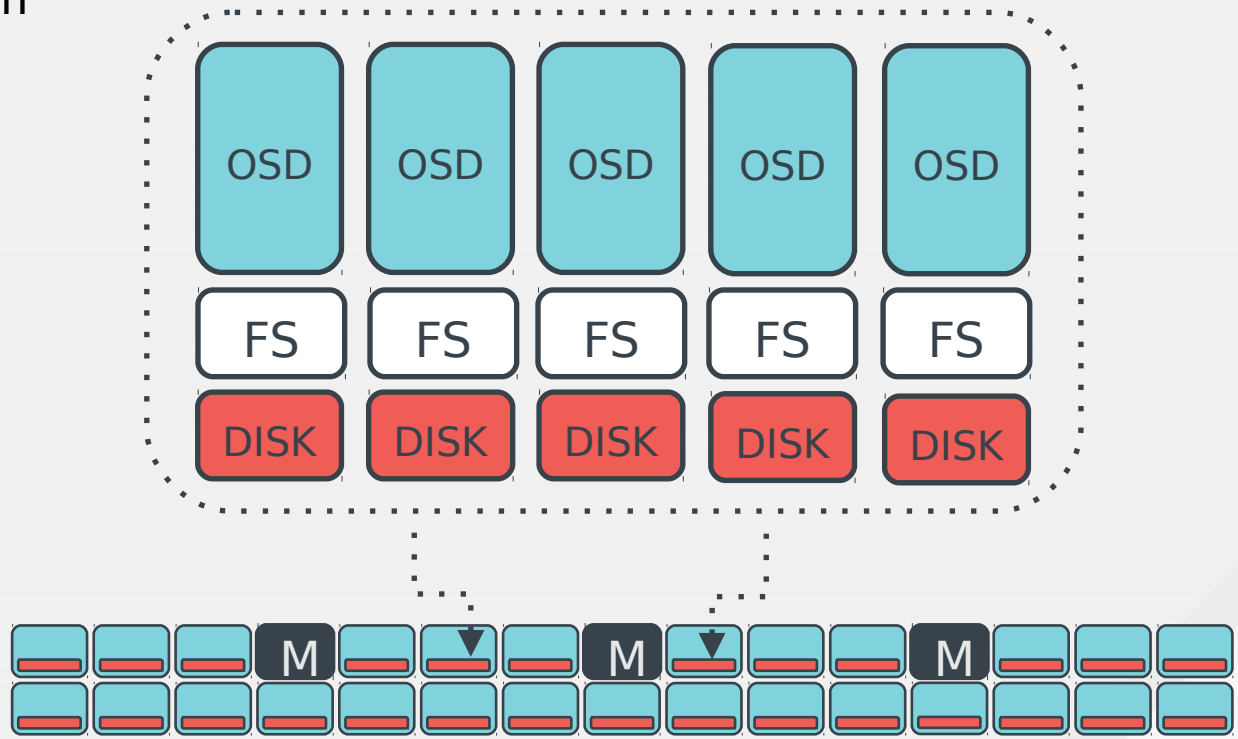


Librados API

- Efficient key/value storage inside an object
- Atomic single-object transactions
- update data, attr, keys together
- atomic compare-and-swap
- Object-granularity snapshot infrastructure
- Partial overwrite of existing data
- Single-object compound atomic operations
- RADOS classes (stored procedures)
- Watch/Notify on an object

FILESTORE

- Ceph backend based on local filesystem



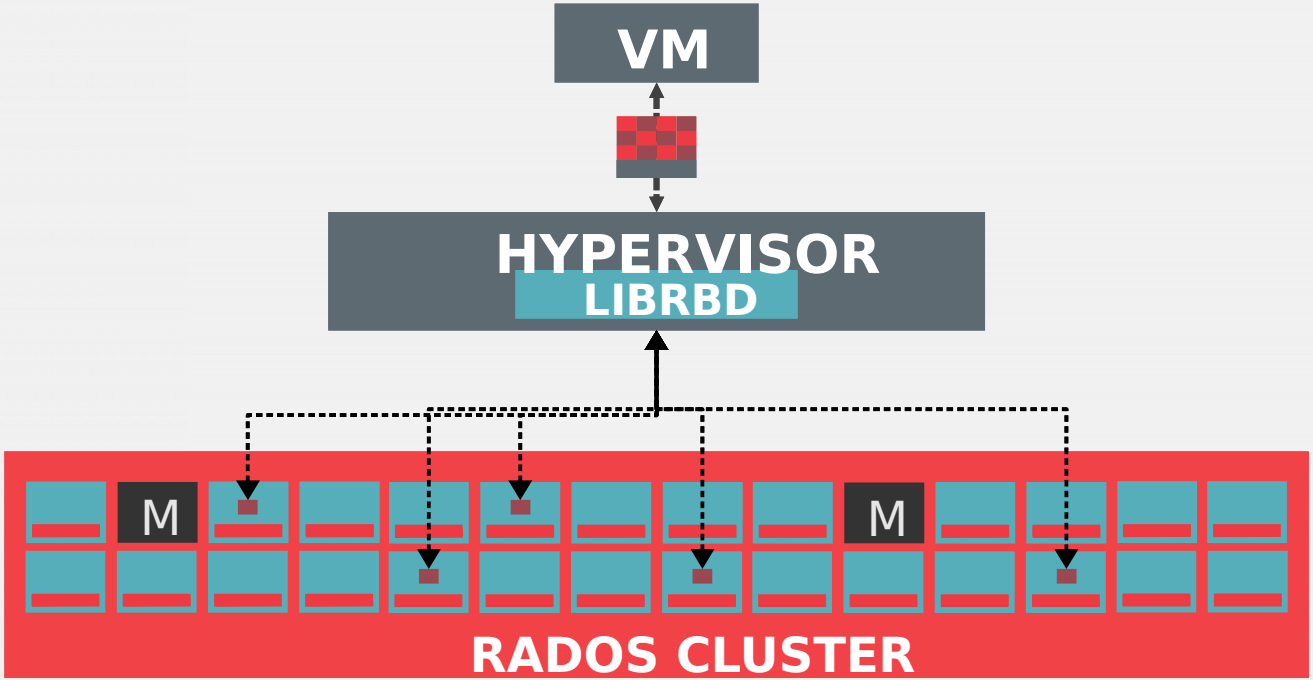
BLUESTORE

BlueStore is a new Ceph storage backend optimized for modern media

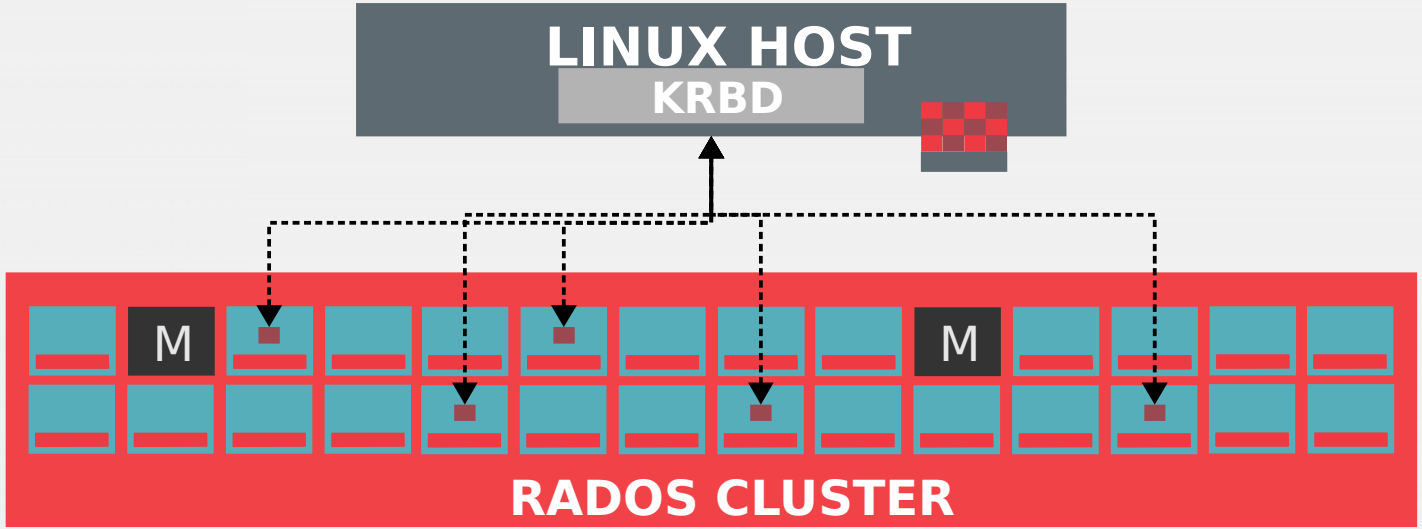
- Replaces FileStore, which was designed for HDDs
- Supports flexible media topologies (flash, K/V drives, persistent memory)
- Eliminates the need for an underlying filesystem or dedicated journal device
- Provides a 2-3X performance boost

RBD

LIBRBD



KRBD



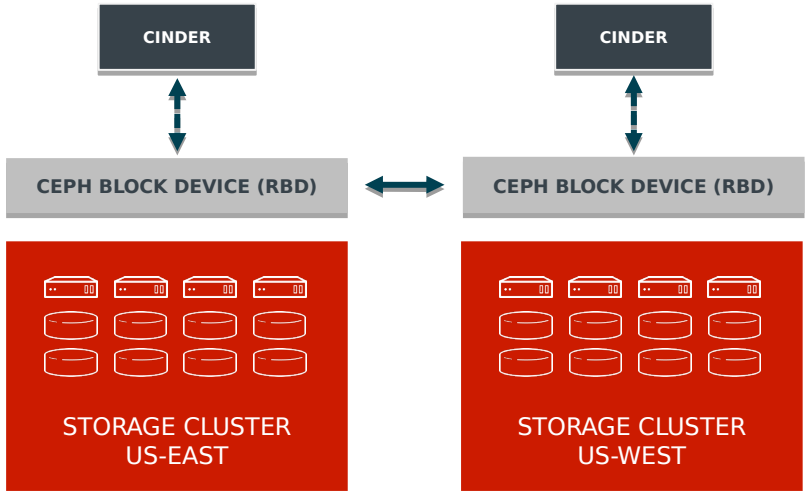
RBD

- Stripe images across entire cluster (pool)
- Read-only snapshots
- Copy-on-write clones
- Broad integration
  - QEMU, libvirt
  - Linux kernel
  - iSCSI (STGT, LIO)
  - OpenStack, CloudStack, OpenNebula, Ganeti, Proxmox, oVirt
- Incremental backup (relative to snapshots)

RBD MIRRORING

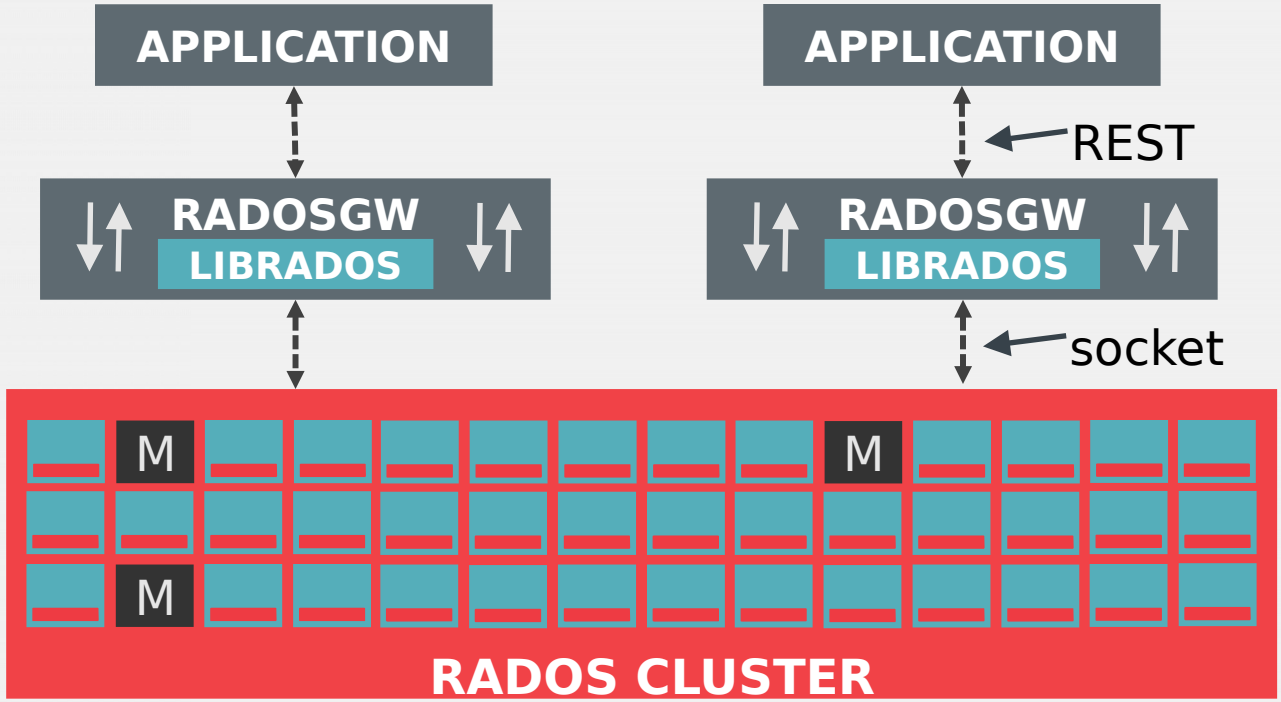
Multi-site replication for block devices (RBD Mirroring)

- Replicates virtual block devices across regions
- Designed for disaster recovery and archival
- Integration with Cinder Volume Replication (OSP-10)



RGW

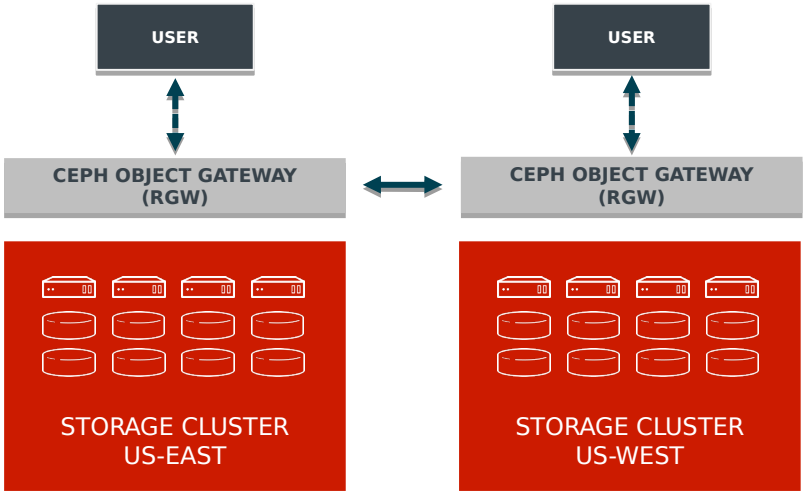
RADOS GATEWAY



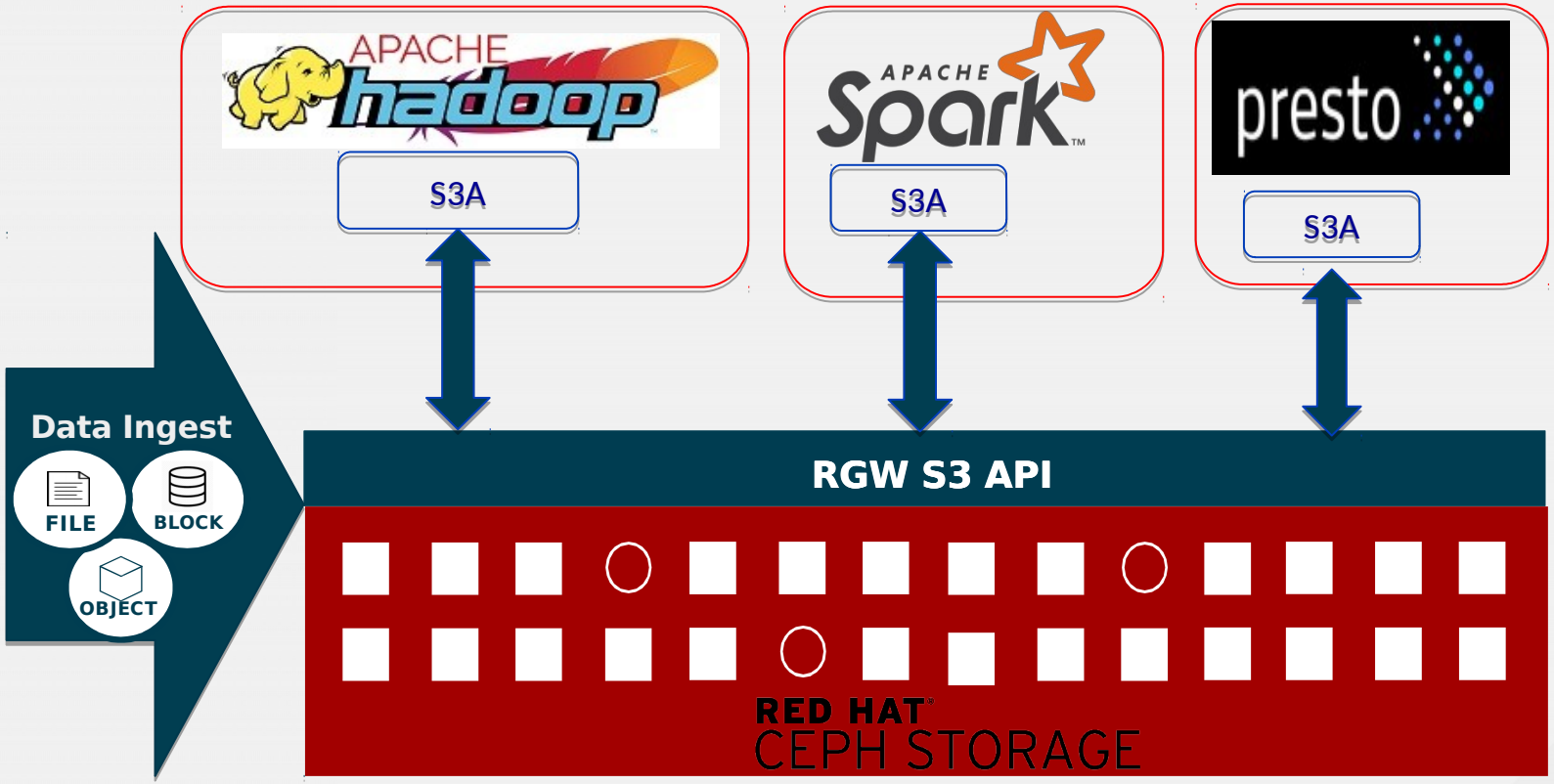
OBJECT GEO REPLICATION

Global object storage clusters with a single namespace

- Enables deployment of clusters across multiple geographic locations
- Clusters synchronize, allowing users to read from or write to the closest one

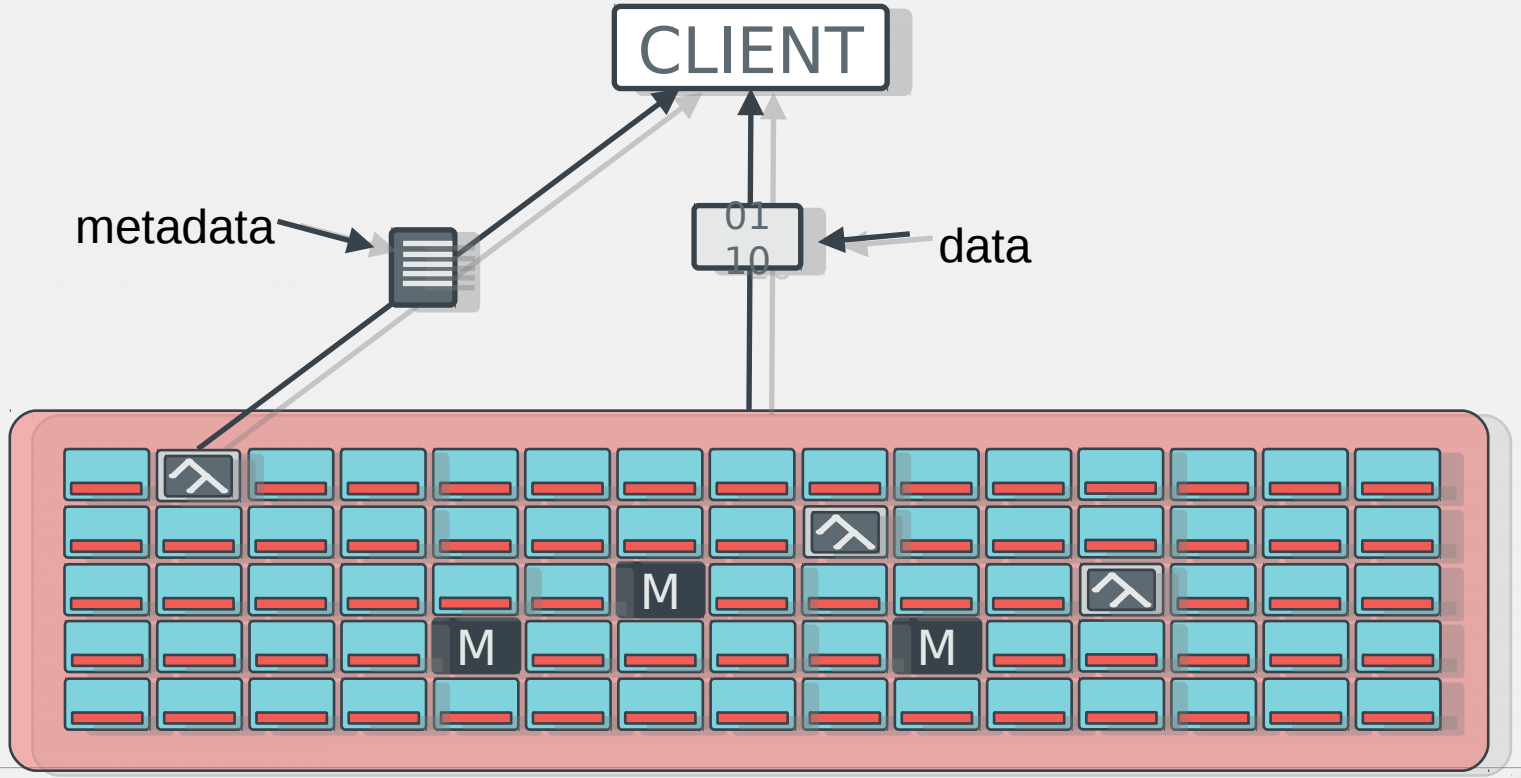


COMPATIBILITY WITH HADOOP S3A CLIENT



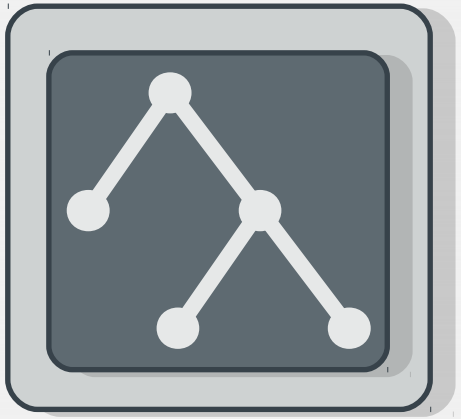
CEPHFS

MetaDataServer

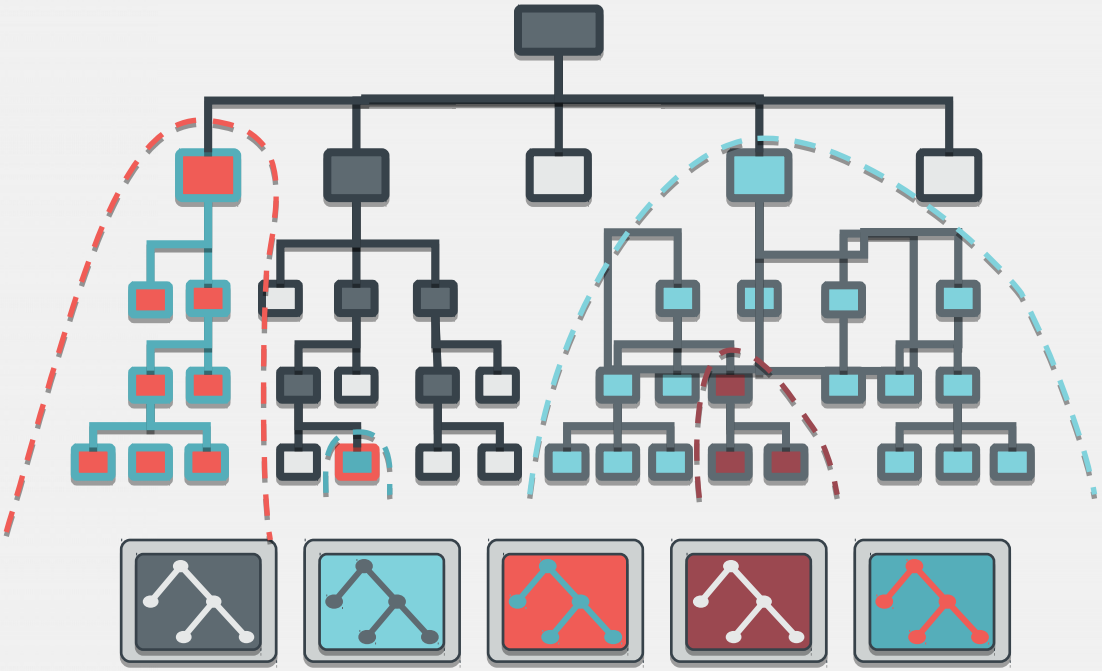


MDS

- manages metadata for POSIX shared file system
- directory hierarchy
- file metadata (size, owner, timestamps)
- stores metadata in RADOS
- does not serve file data to clients
- only required for the shared file system



DYNAMIC SUBTREE PARTITIONING



DYNAMIC SUBTREE PARTITIONING

DYNAMIC SUBTREE PARTITIONING

- scalable
  - arbitrarily partition metadata
- adaptive
  - move work from busy to idle servers
  - replicate hot metadata
- efficient
  - hierarchical partition preserve locality
- dynamic
  - daemons can join/leave
  - take over for failed nodes

CONCLUSIONS

- Ceph is an open, massive scalable, elastic and adaptive storage
- It has a unified storage interface: block, object and file
- We would like you to join our user and developers community!
- Email: owasserm@redhat.com
- owasserm@IRC
- oritwas@twitter

BACKUP

CEPH CONTAINERIZATION

- Alternative vehicle for deploying Red Hat Ceph Storage
- Single container image of product available on Red Hat Container Registry
- Delivers same capabilities as in traditional package format
- Supports customers seeking to standardize orchestration and deployment of infrastructure software in containers with Kubernetes

BENEFITS OF CEPH CONTAINERIZATION

- Enables installations, upgrades, and updates atomically
- Offers reduced complexity, easier management, and faster deployment for key use cases like telco, NFV, and mass scale edge computing
- Facilitates deployment and scale of complex architectures like OpenStack by containerizing individual services and managing deployments with Kubernetes scheduler