redhat.

# RED HAT CEPH STORAGE

Jacob Shucart

Sr. Cloud Storage Solution Architect

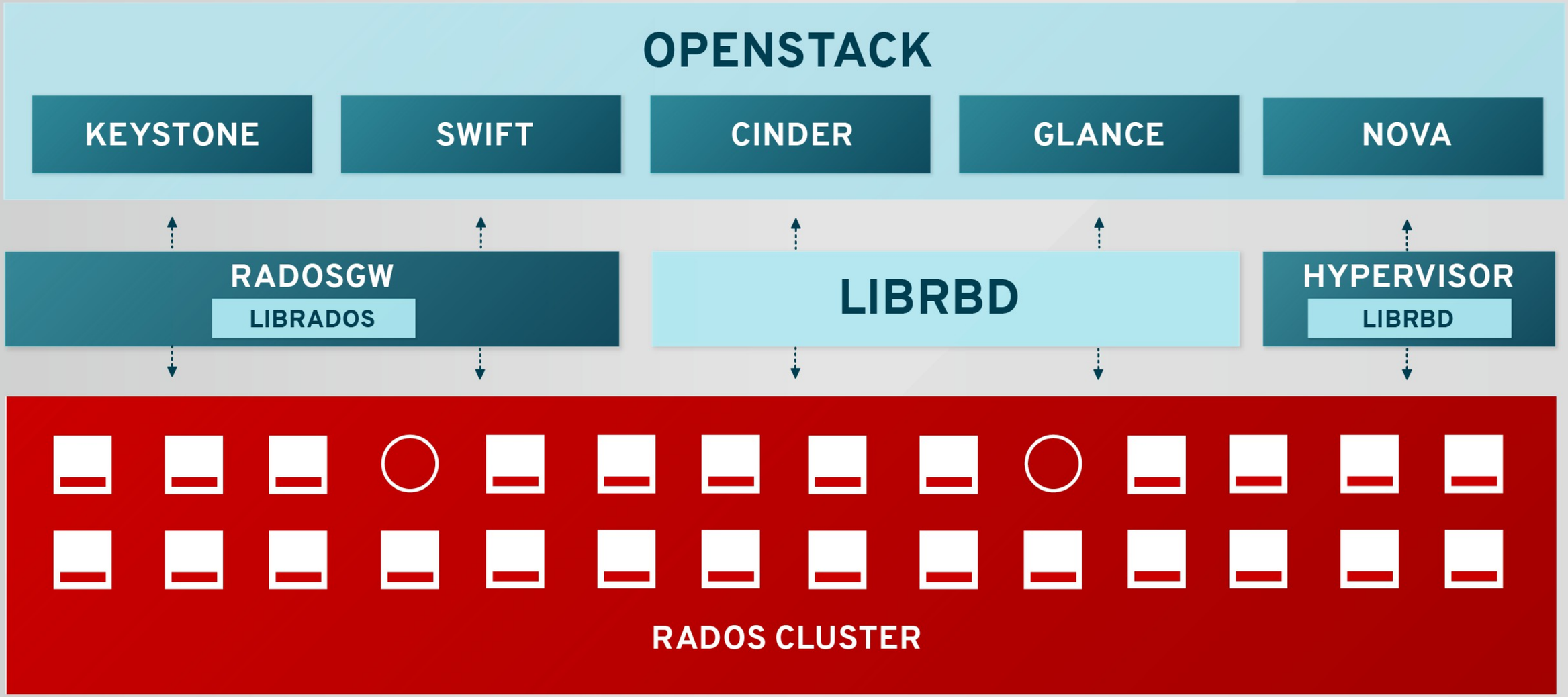
2016

# THE FUTURE OF STORAGE

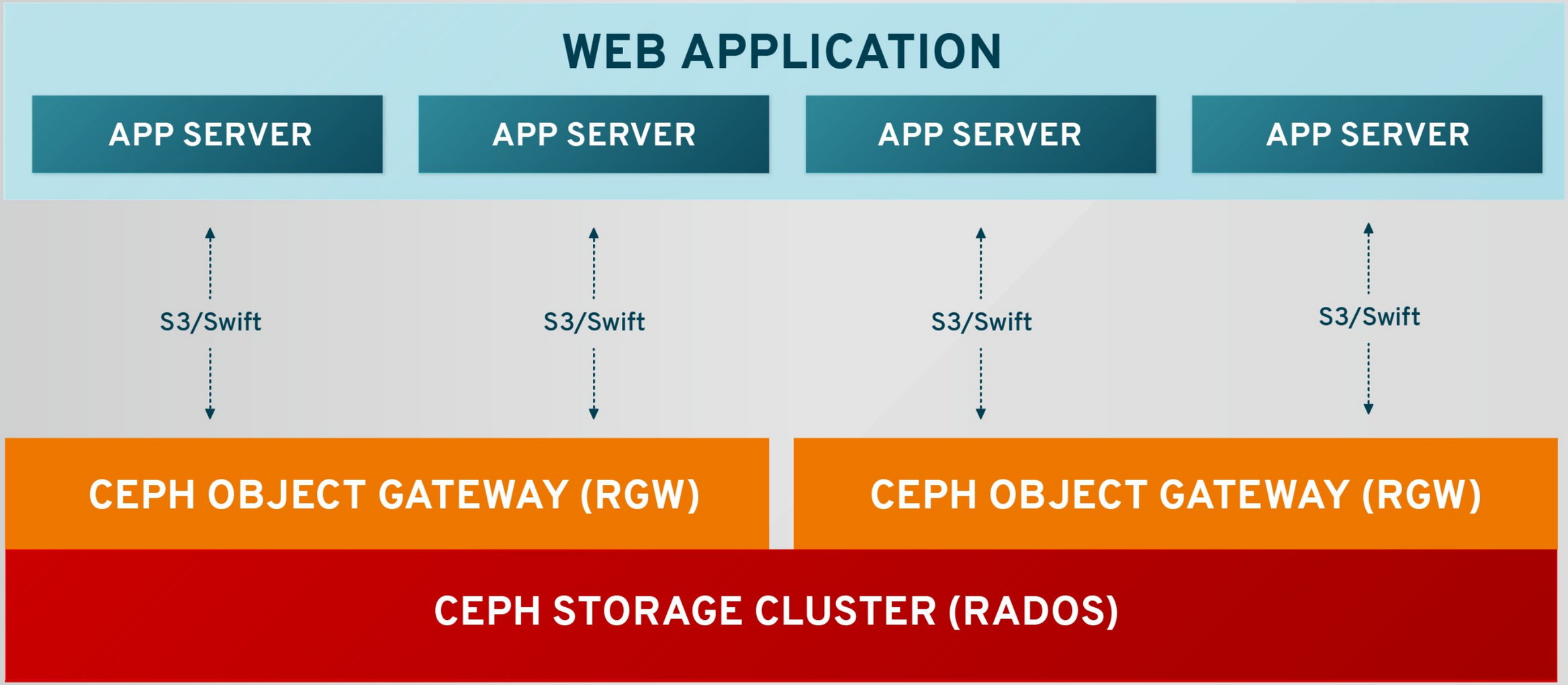
| | |
|---|---|
| **PROPRIETARY HARDWARE** | **Common, off-the-shelf hardware**<br>Lower cost, standardized supply chain |
| **SCALE-UP ARCHITECTURE** | **Scale-out architecture**<br>Increased operational flexibility |
| **HARDWARE-BASED INTELLIGENCE** | **Software-based intelligence**<br>More programmability,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


ARCHITECTURE

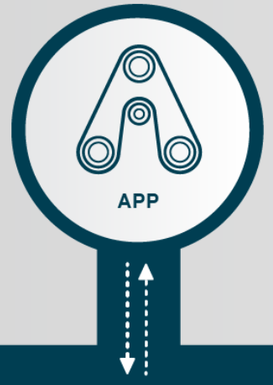
# CEPH AND OPENSTACK

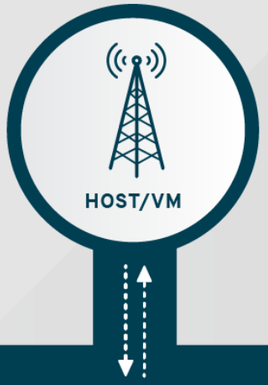


# WEB APPLICATION STORAGE



# ARCHITECTURAL COMPONENTS




## RGW

A web services gateway for object storage, compatible with S3 and Swift

## RBD

A reliable, fully distributed block device with cloud platform integration

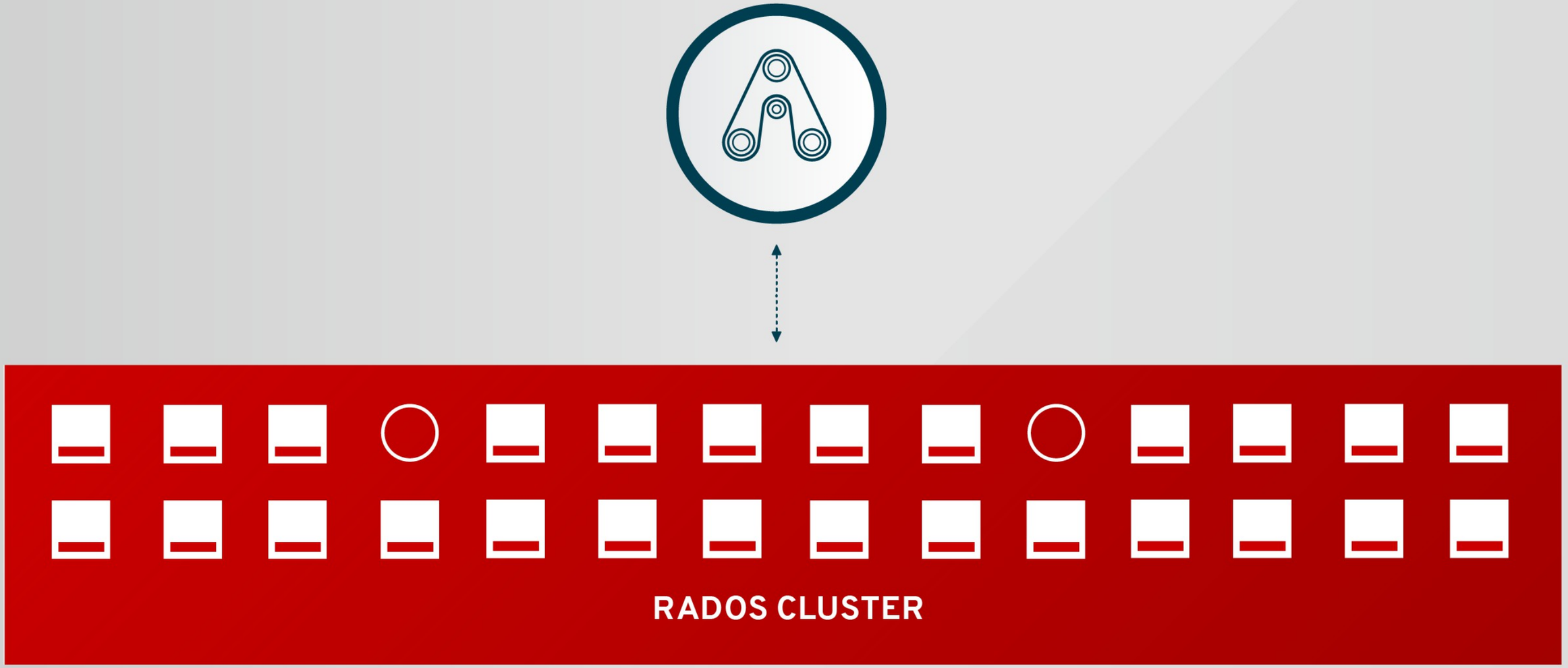
## LIBRADOS

A library allowing apps to directly access RADOS (C, C++, Java, Python, Ruby

## RADOS

A software-based reliable, autonomous, distributed object store comprised of self-healing, self-managing, intelligent storage nodes and lightweight monitors

# RADOS CLUSTER



# RADOS COMPONENTS

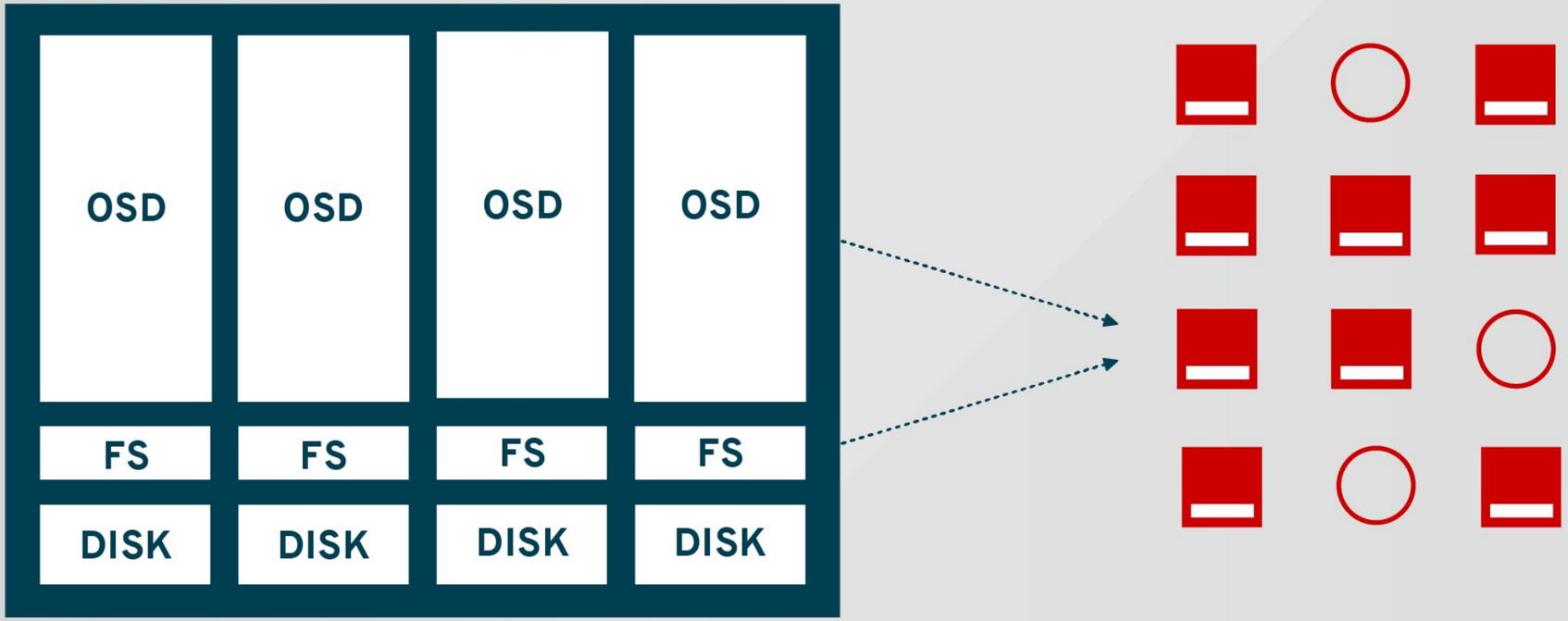
## OSDs

- 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

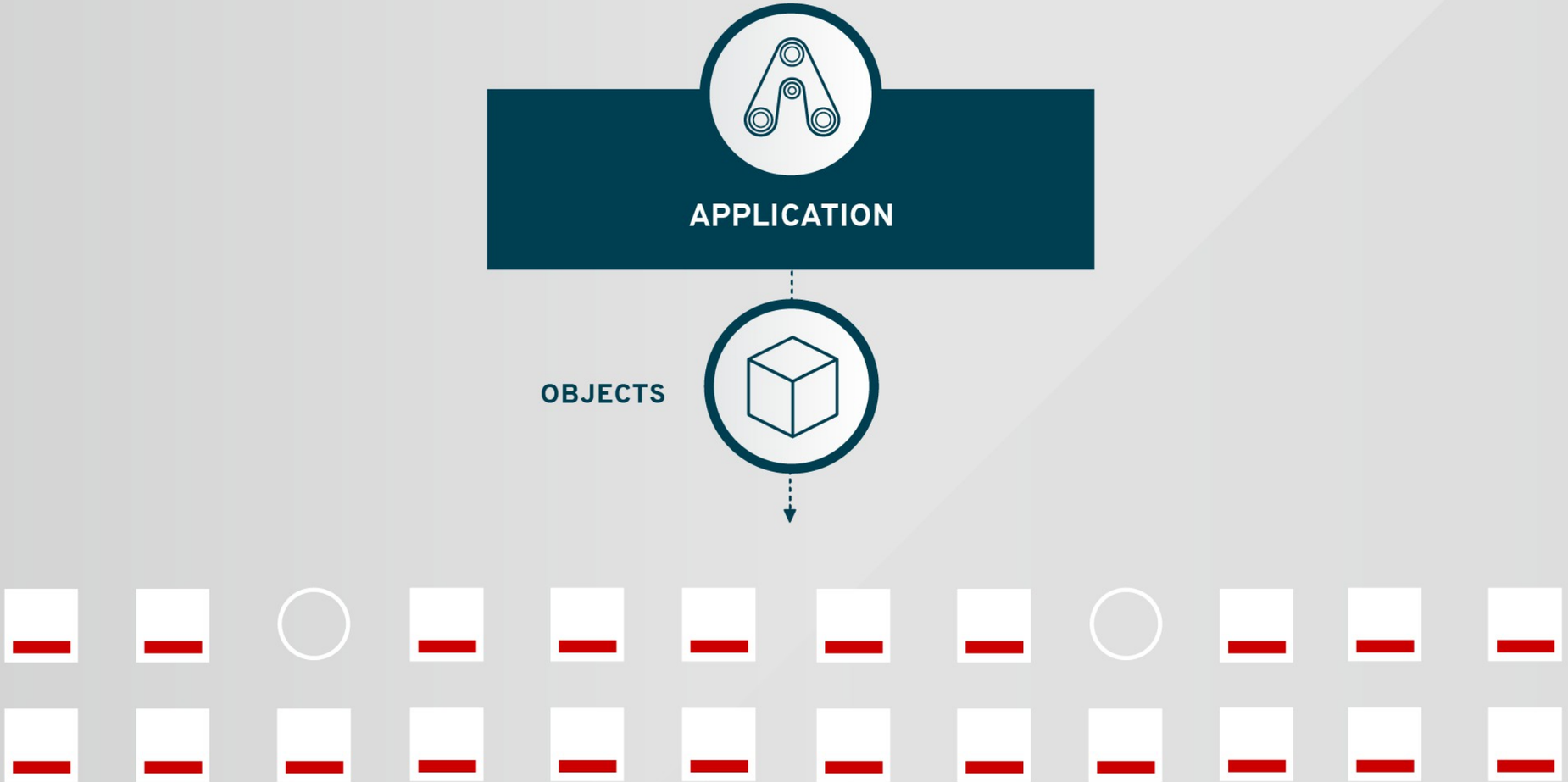
## Monitors

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

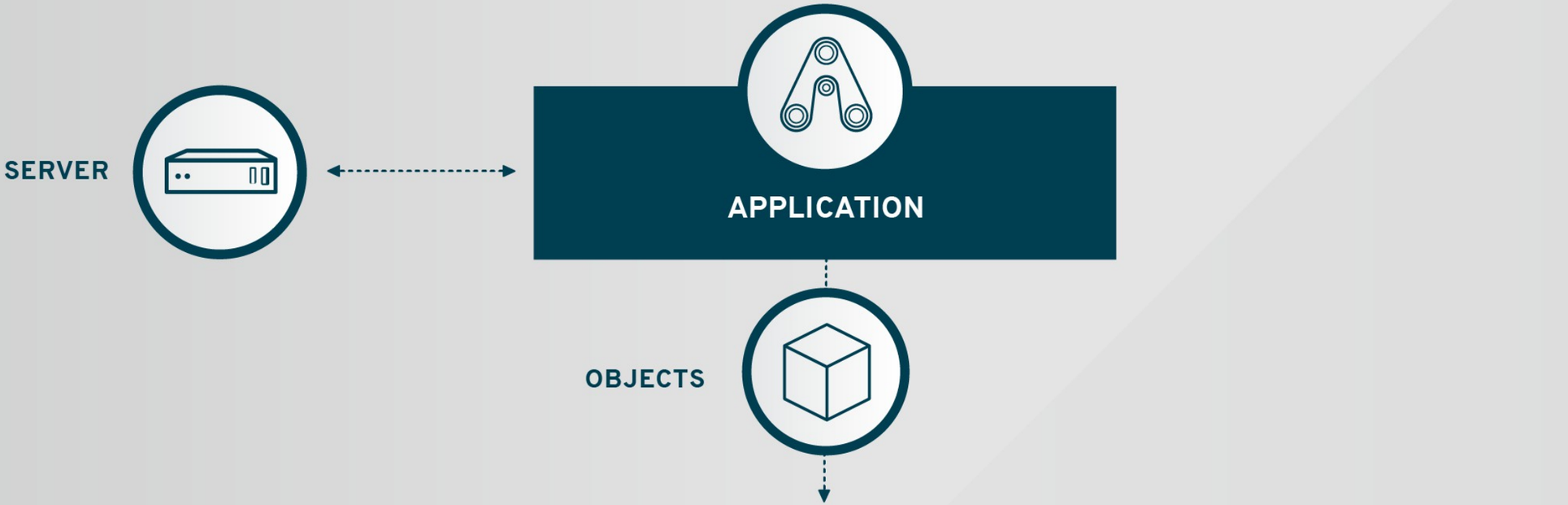
# OBJECT STORAGE DAEMONS



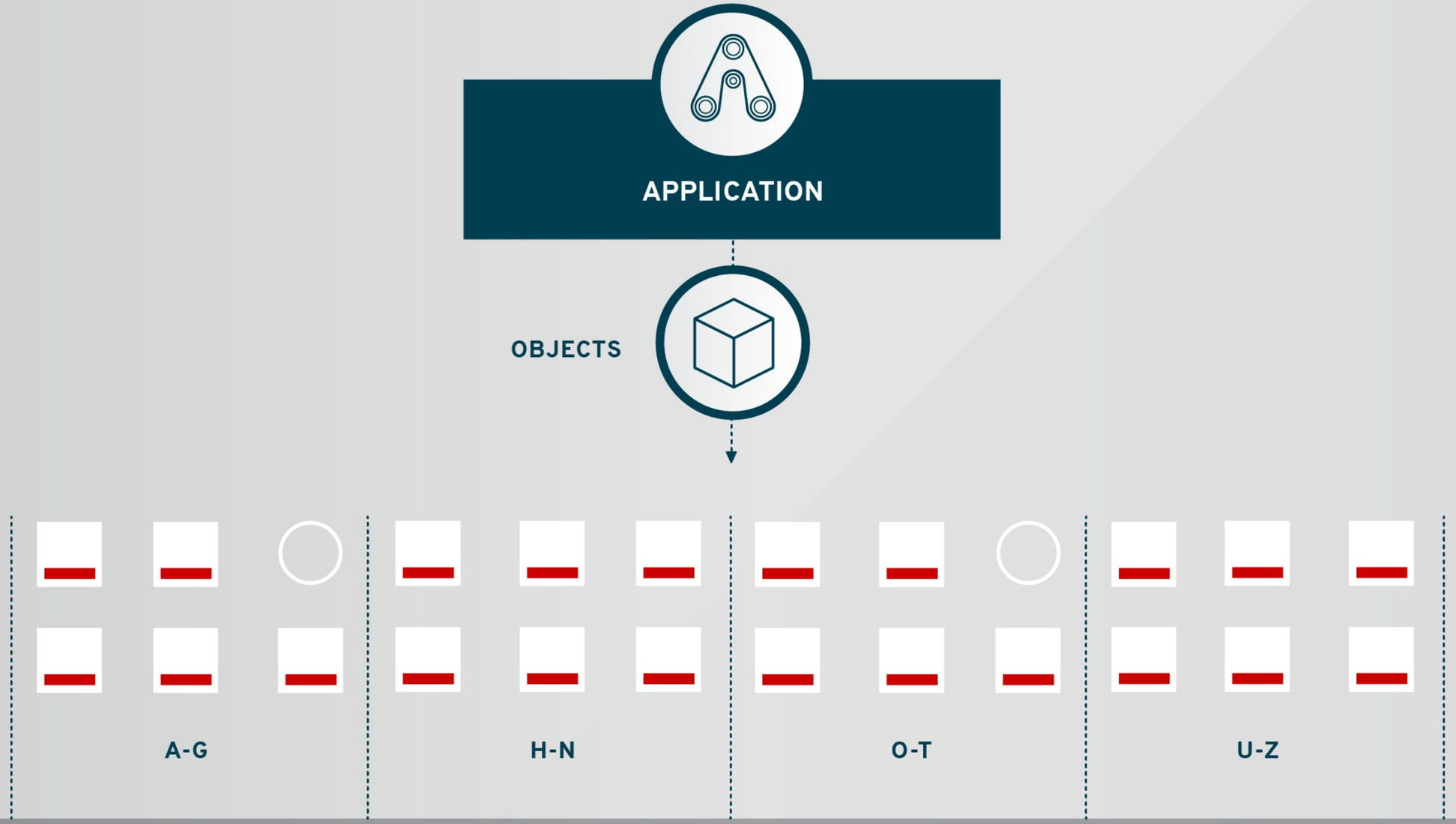
# WHERE DO OBJECTS LIVE?



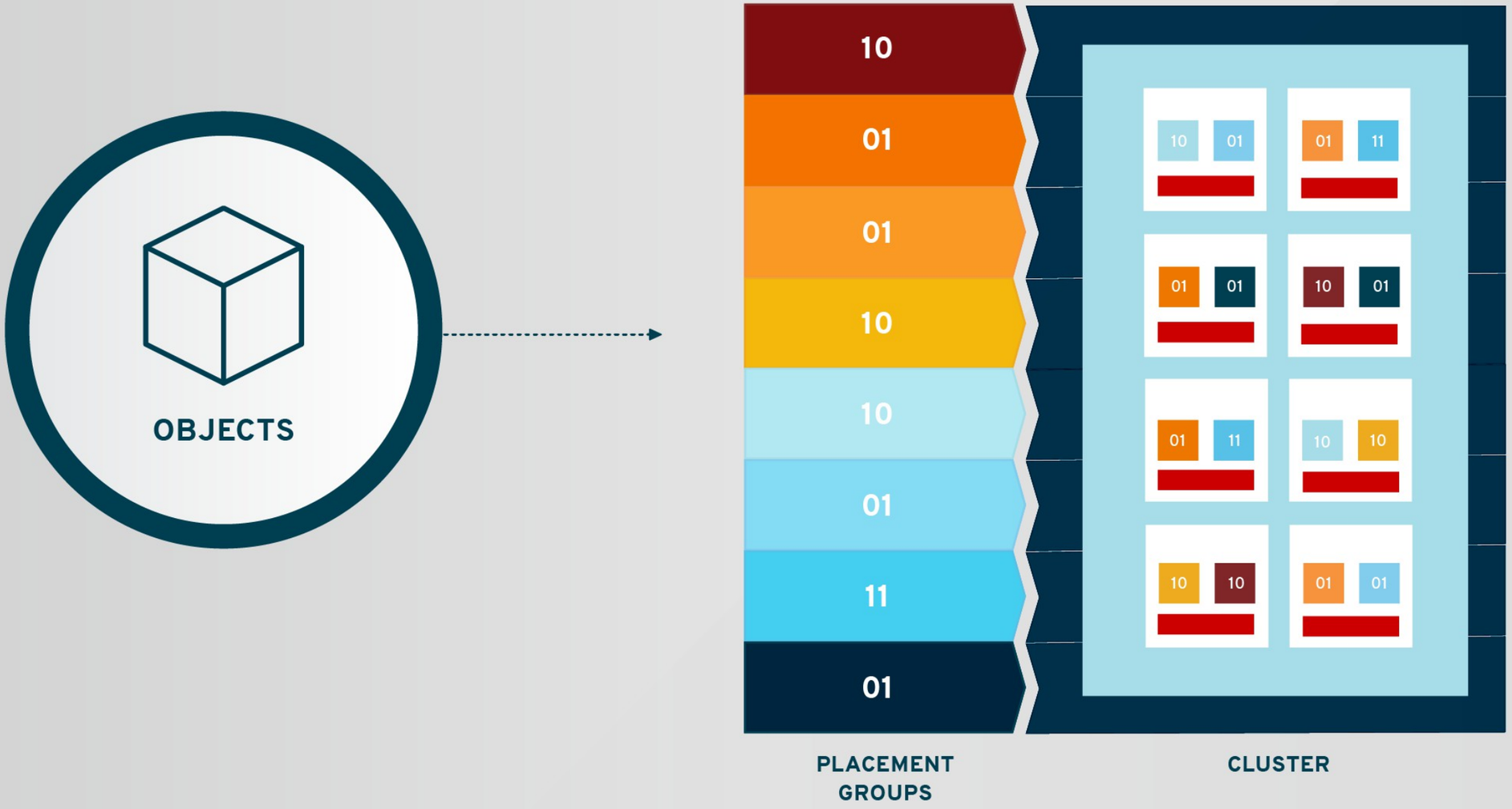
# A METADATA SERVER?



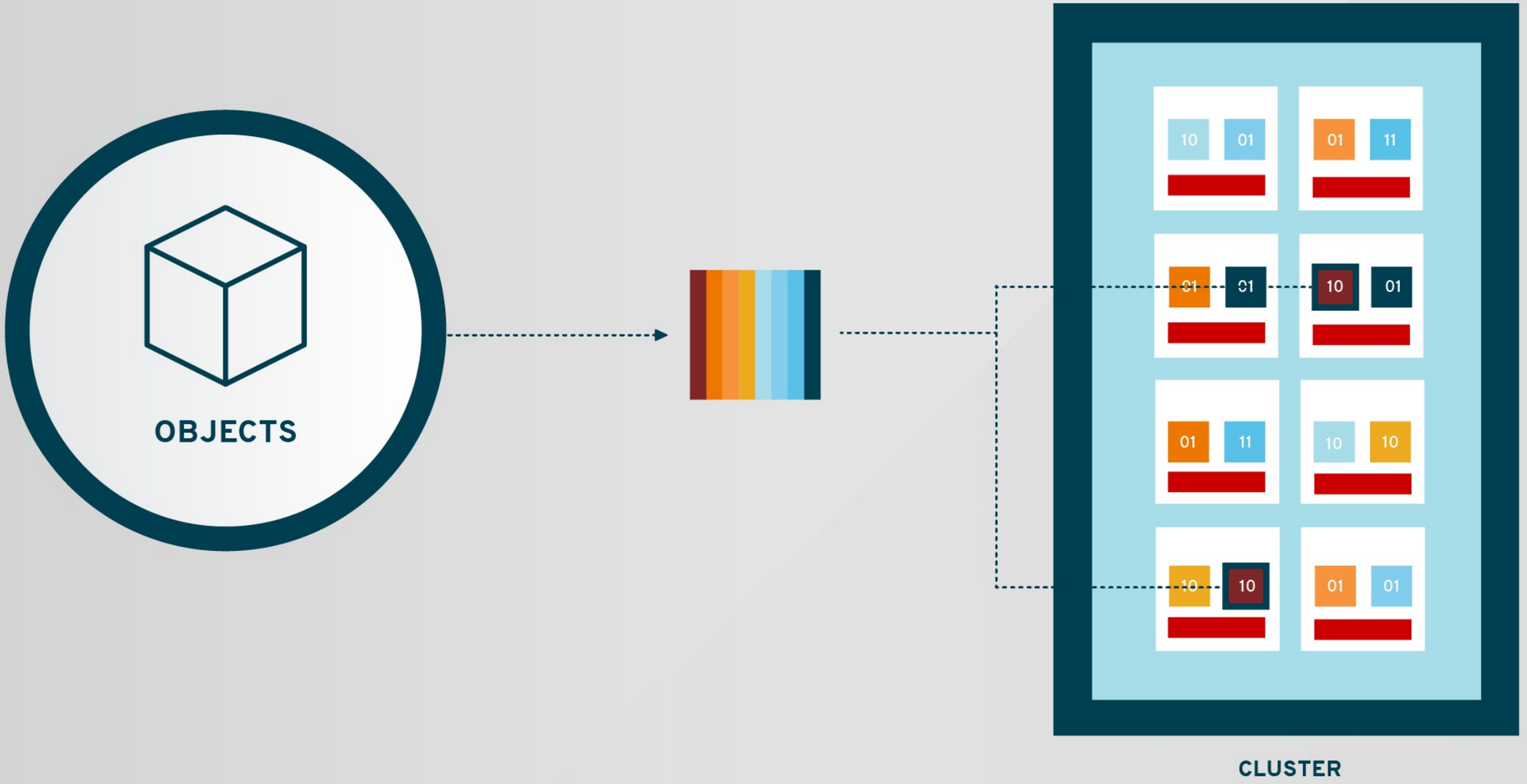
# CALCULATED PLACEMENT



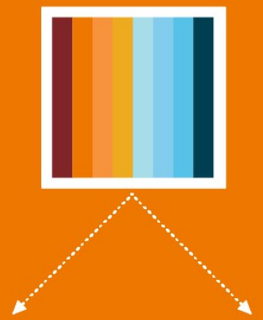
# EVEN BETTER - CRUSH!



# CRUSH IS A QUICK CALCULATION



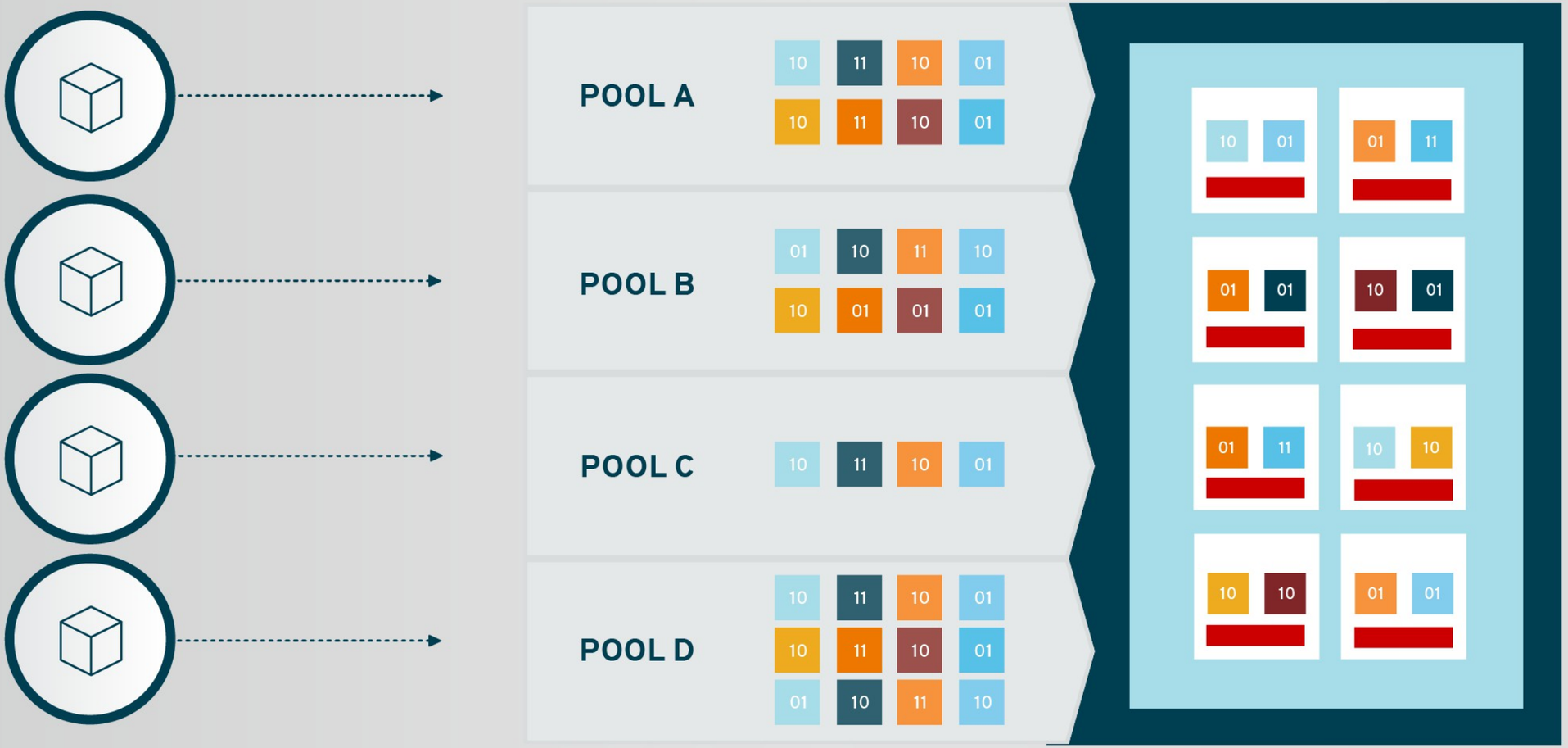
# CRUSH - DYNAMIC DATA PLACEMENT



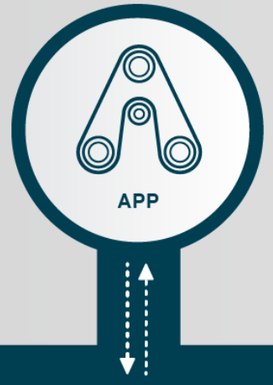
## CRUSH

- Psuedo-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

# DATA IS ORGANIZED INTO POOLS

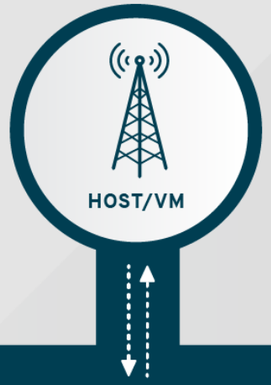


# ARCHITECTUAL COMPONENTS




## RGW

A web services gateway for object storage, compatible with S3 and Swift




## RBD

A reliable, fully distributed block device with cloud platform integration

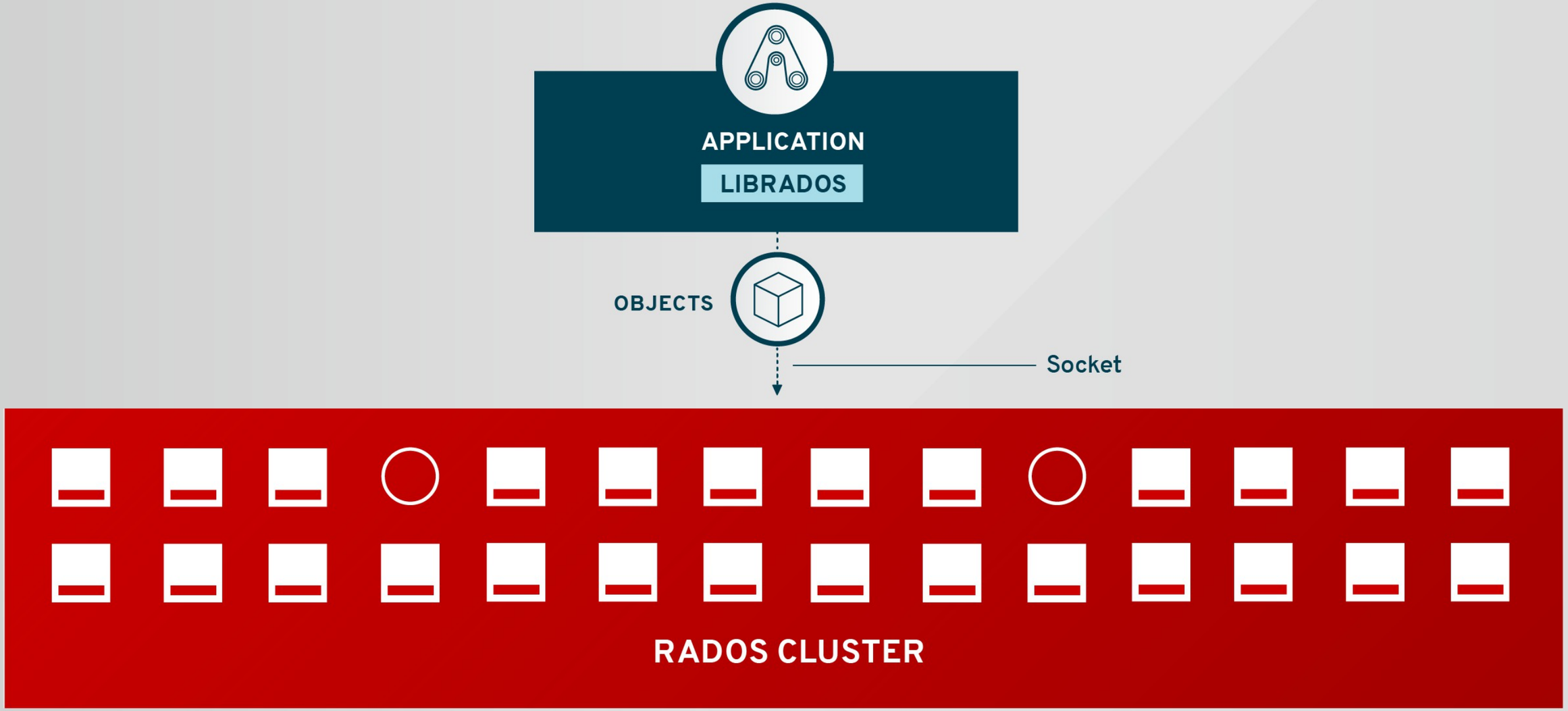
## LIBRADOS

A library allowing apps to directly access RADOS (C, C++, Java, Python, Ruby

## RADOS

A software-based reliable, autonomous, distributed object store comprised of self-healing, self-managing, intelligent storage nodes and lightweight monitors

# ACCESSING A RADOS CLUSTER

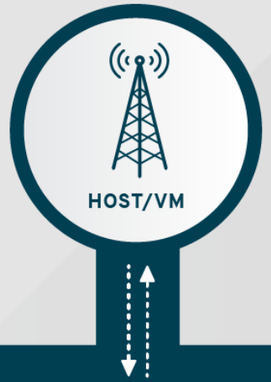


# LIBRADOS - RADOS ACCESS FOR APPS

## LIBRADOS

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

# ARCHITECTURAL COMPONENTS


## RGW

A web services gateway for object storage, compatible with S3 and Swift

## RBD

A reliable, fully distributed block device with cloud platform integration

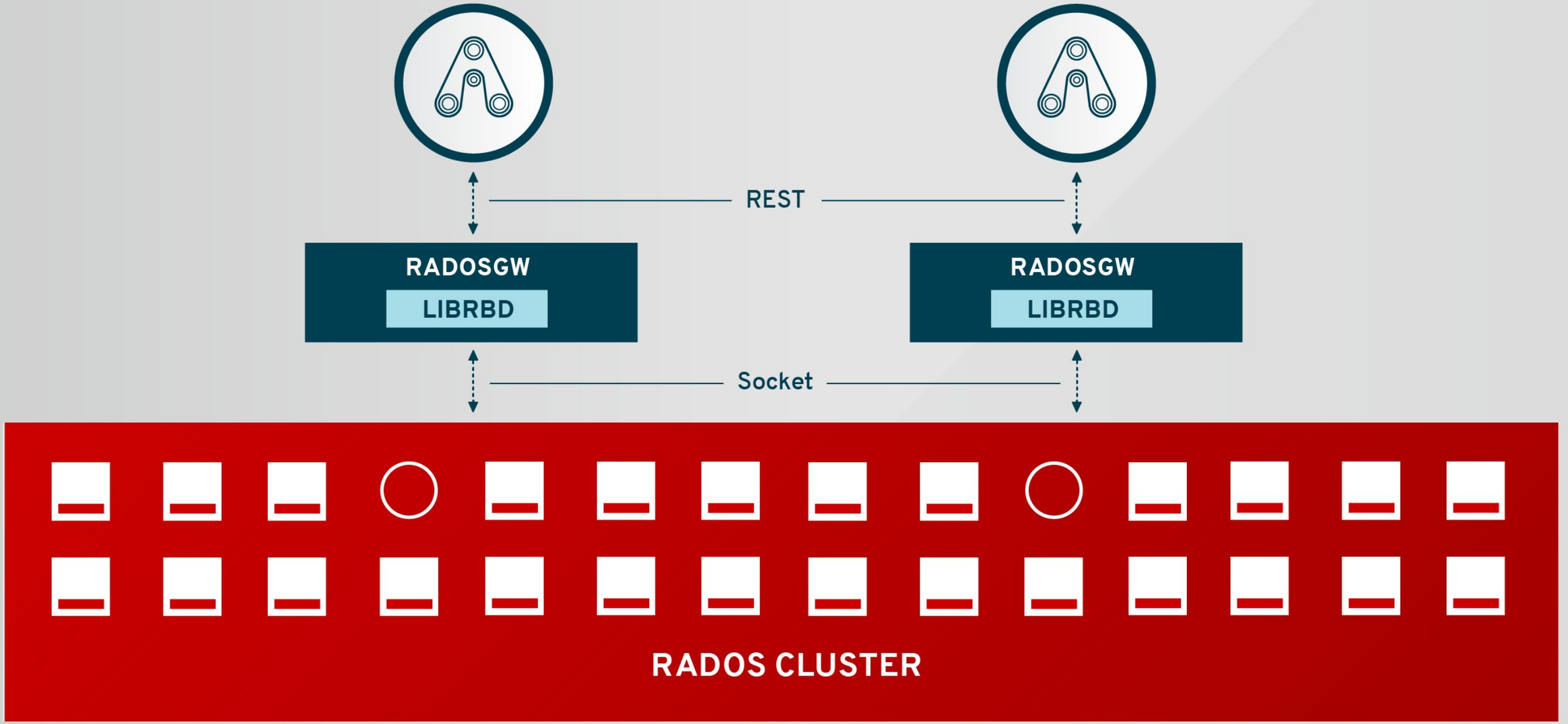
## LIBRADOS

A library allowing apps to directly access RADOS (C, C++, Java, Python, Ruby

## RADOS

A software-based reliable, autonomous, distributed object store comprised of self-healing, self-managing, intelligent storage nodes and lightweight monitors

# THE RADOS GATEWAY

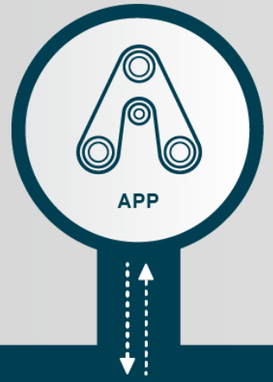


# RADOSGW MAKES RADOS WEBBY

## RADOSGW

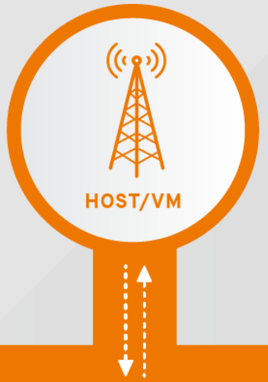
- REST-based object storage proxy
- Uses RADOS to store objects
- API supports buckets, accounts
- Usage accounting for billing
- Compatible with S3 and Swift applications

# ARCHITECTURAL COMPONENTS




## RGW

A web services gateway for object storage, compatible with S3 and Swift




## RBD

A reliable, fully distributed block device with cloud platform integration

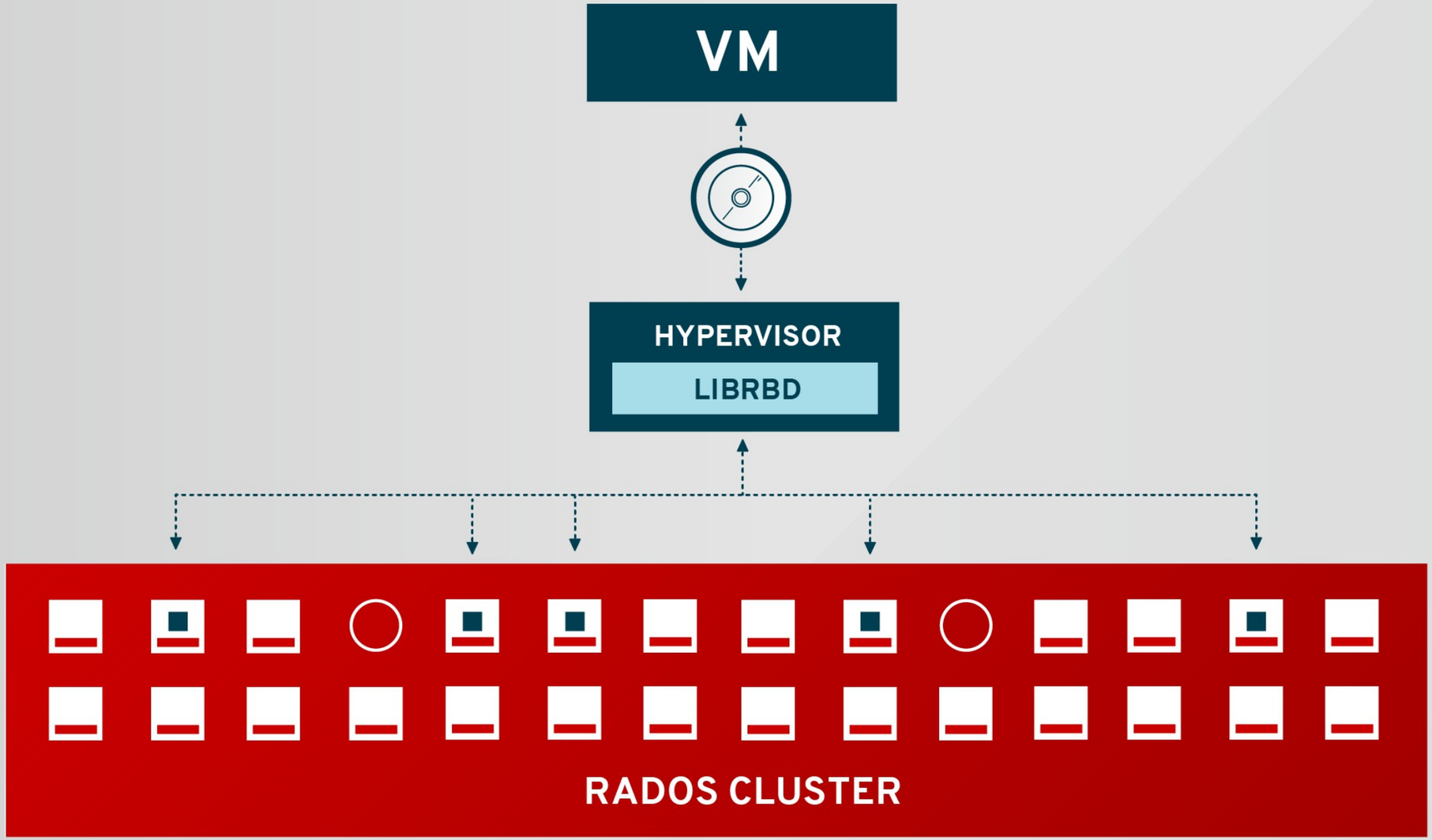
## LIBRADOS

A library allowing apps to directly access RADOS (C, C++, Java, Python, Ruby

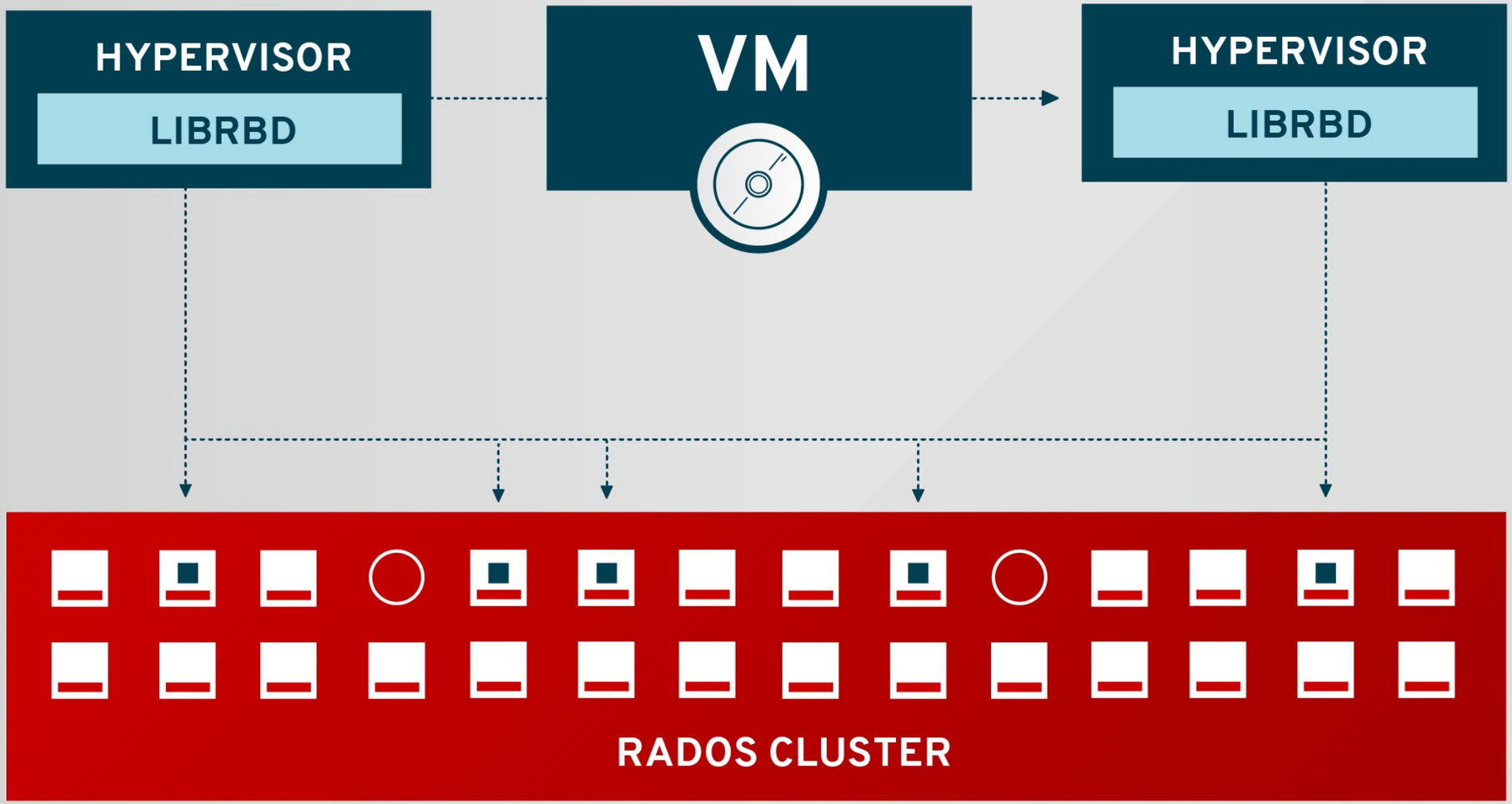
## RADOS

A software-based reliable, autonomous, distributed object store comprised of self-healing, self-managing, intelligent storage nodes and lightweight monitors

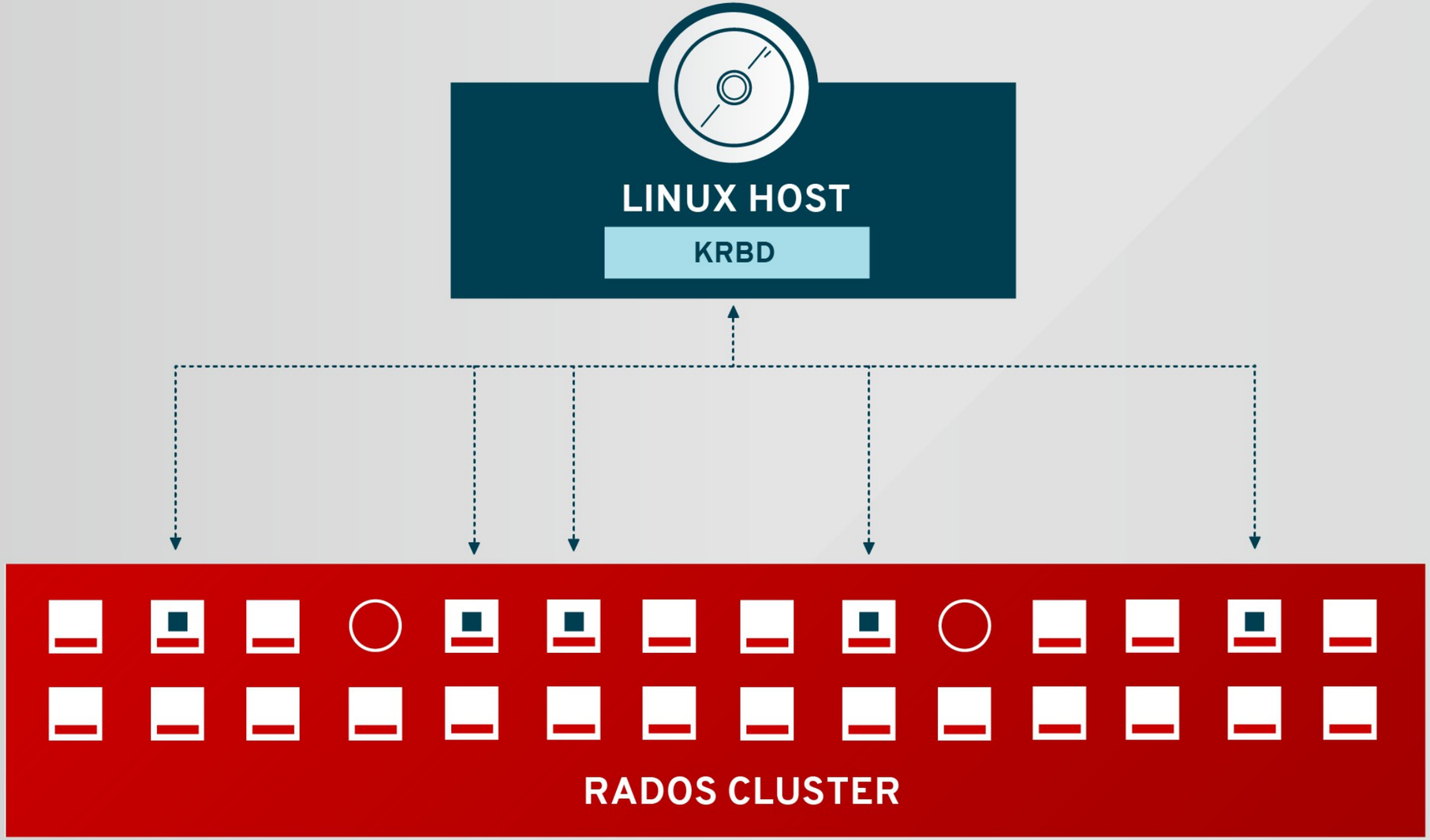
# STORING VIRTUAL DISKS



# SEPARATE COMPUTE FROM STORAGE



# KERNEL MODULE FOR MAX FLEXIBLE!

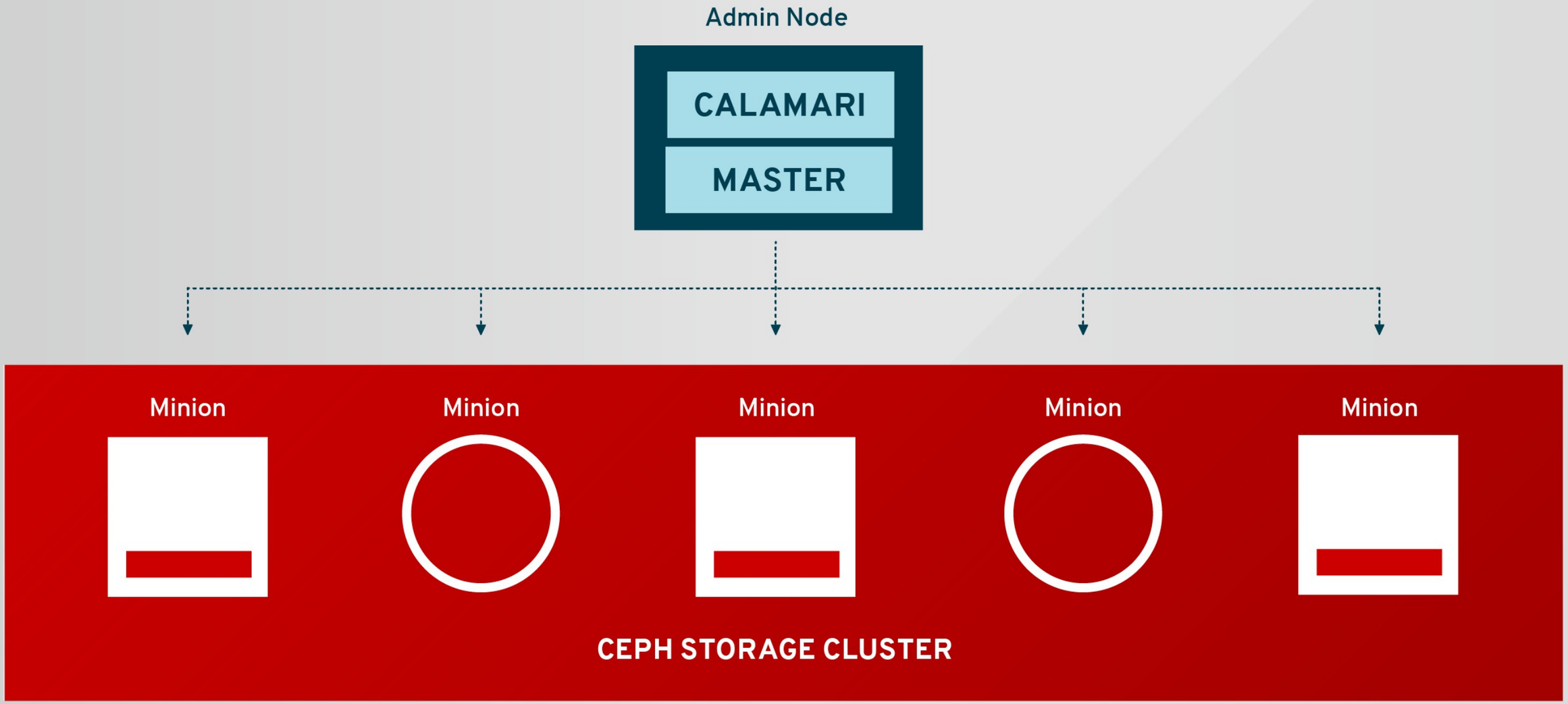


# RBD STORES VIRTUAL DISKS

## RADOS BLOCK DEVICE

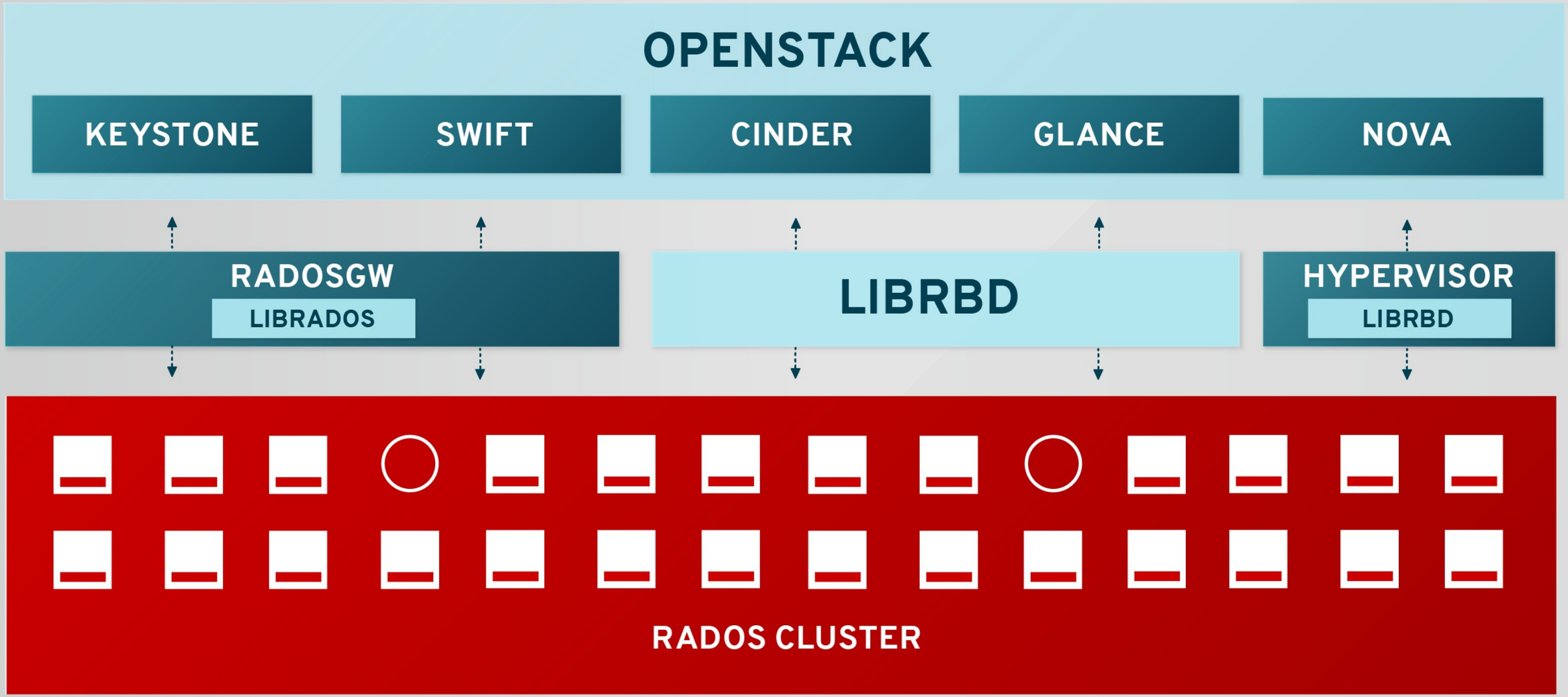
- Storage of disk images in RADOS
- Decouples VMs from host
- Images are striped across the cluster (pool)
- Snapshots
- Copy-on-write clones
- Support in:
  - Mainline Linux Kernel (2.6.39+)
  - Qemu/KVM
  - OpenStack

# CALAMARI ARCHITECTURE

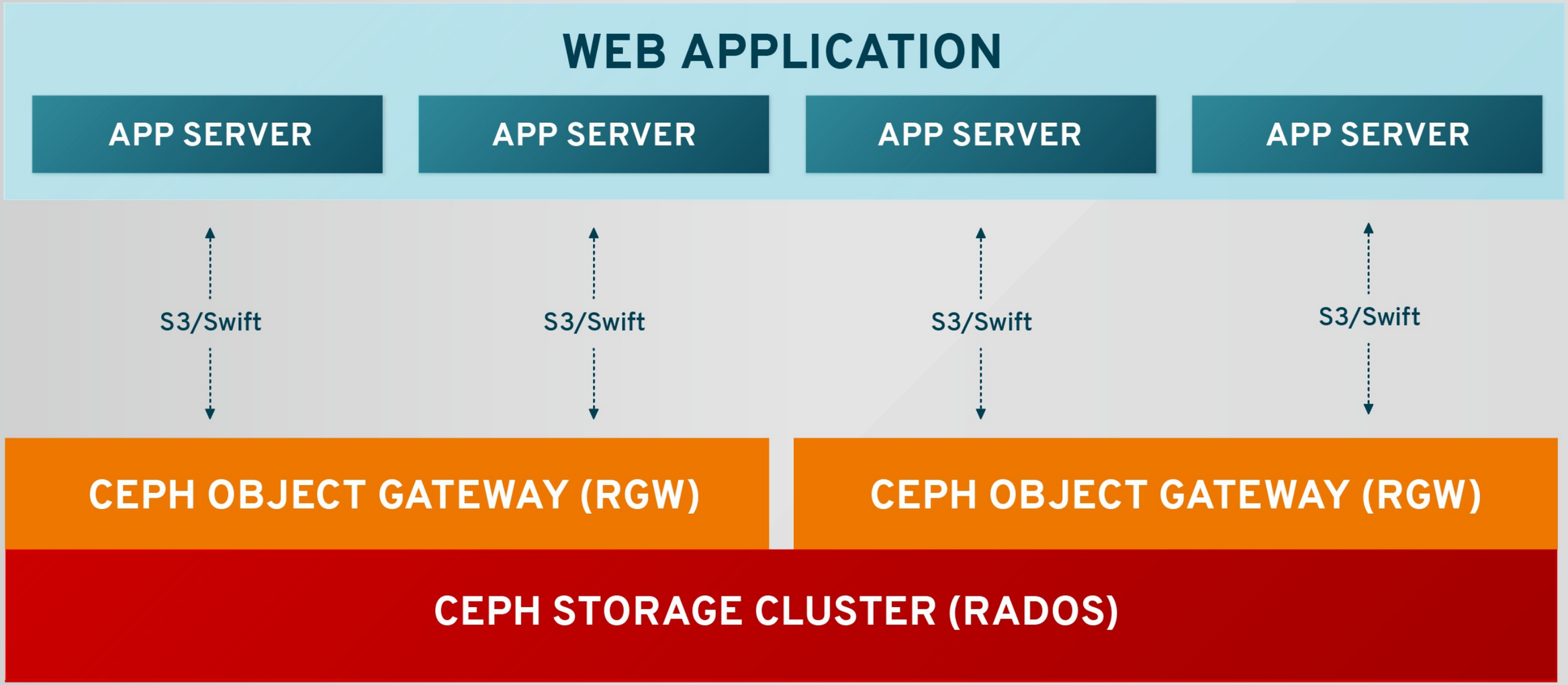



USE CASES

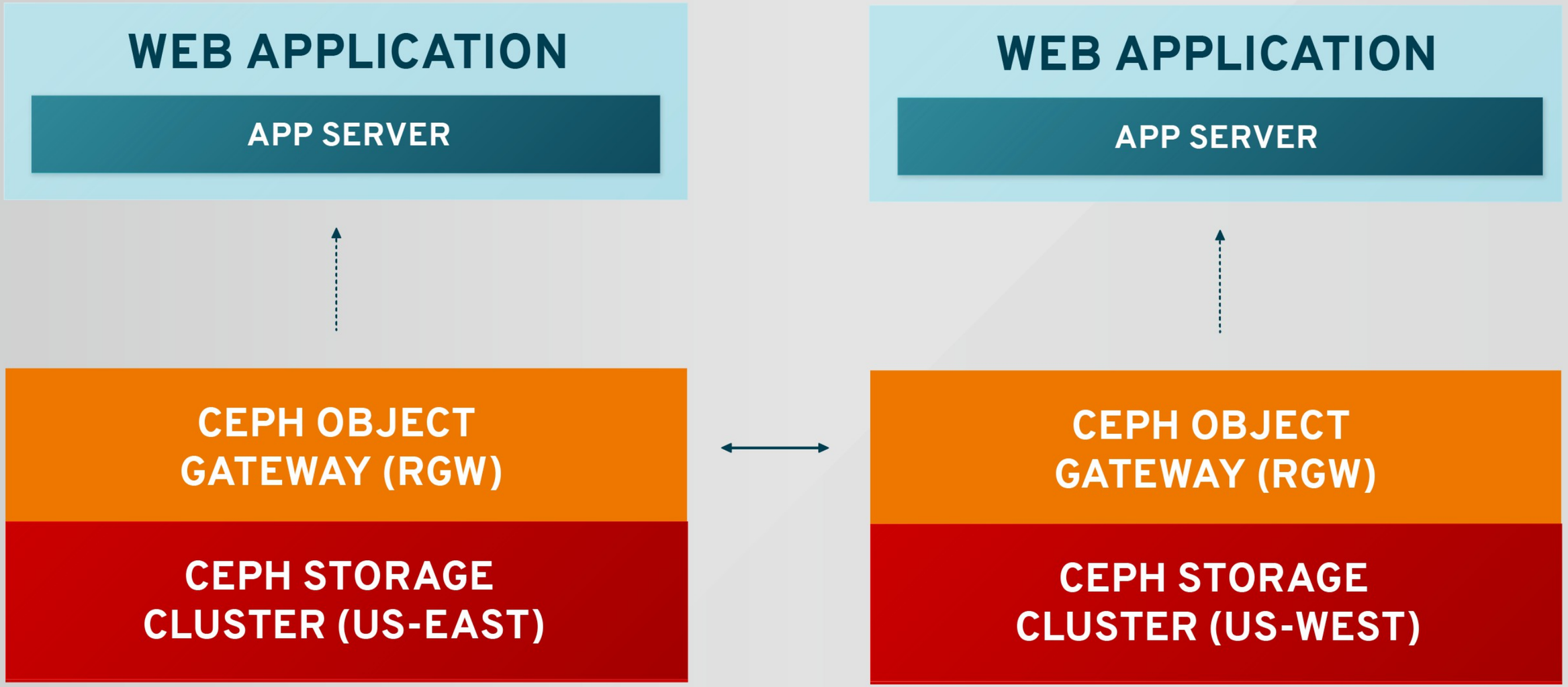
# CEPH AND OPENSTACK



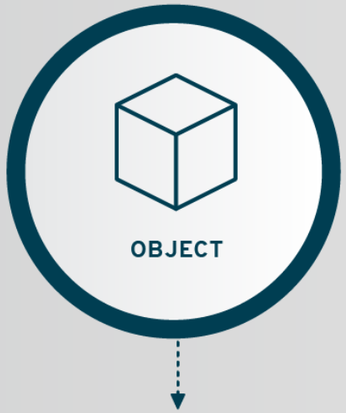
# WEB APPLICATION STORAGE

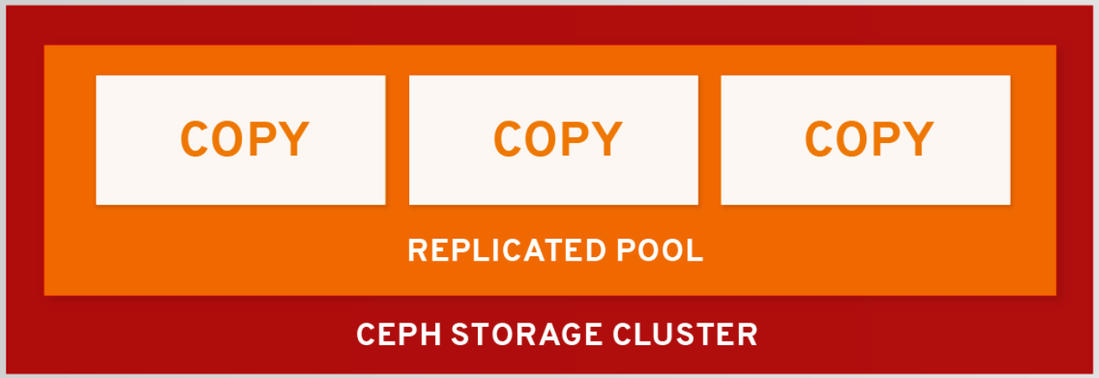


# MULTI-SITE OBJECT STORAGE



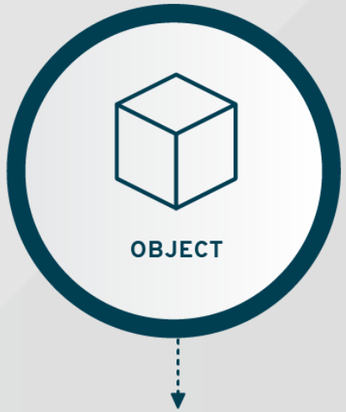
# ERASURE CODING




**FULL COPIES OF STORED OBJECTS**

- Very high durability
- Quicker recovery

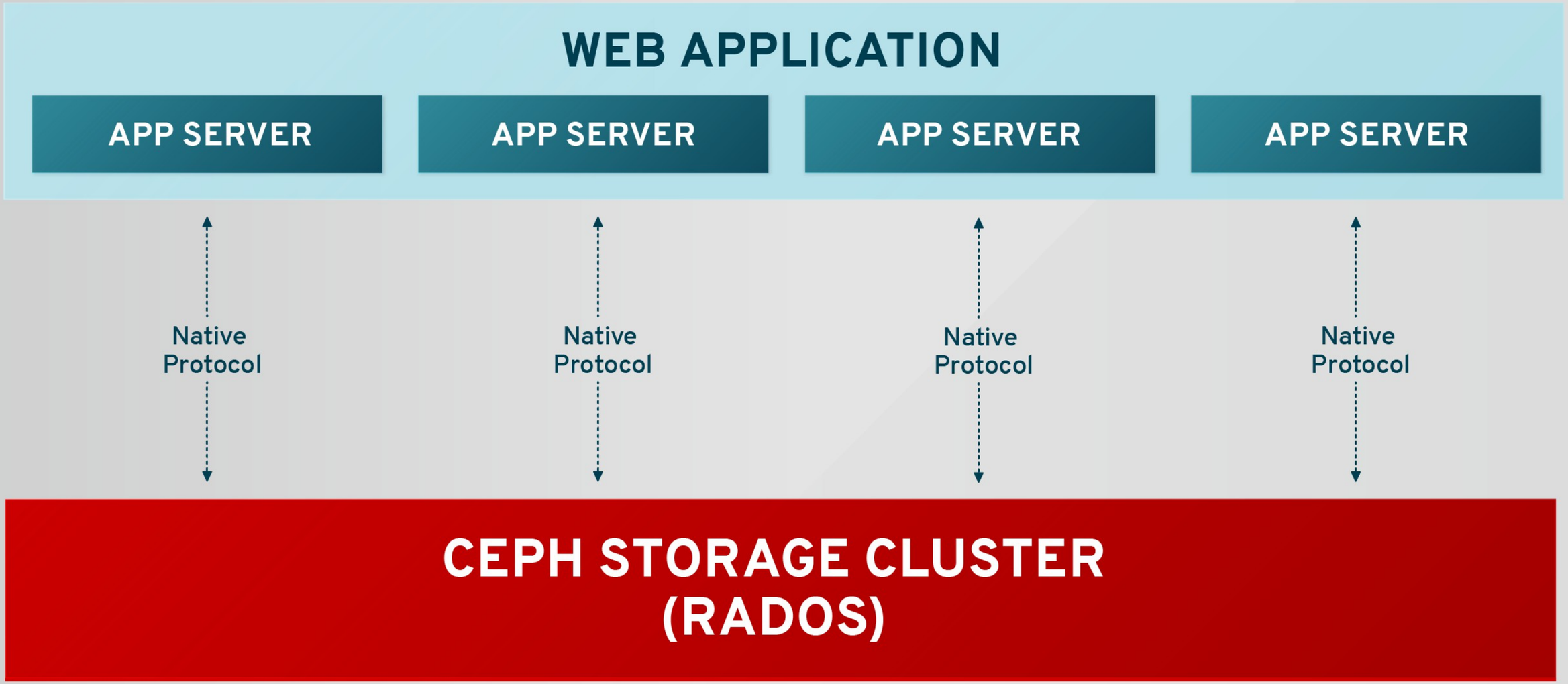




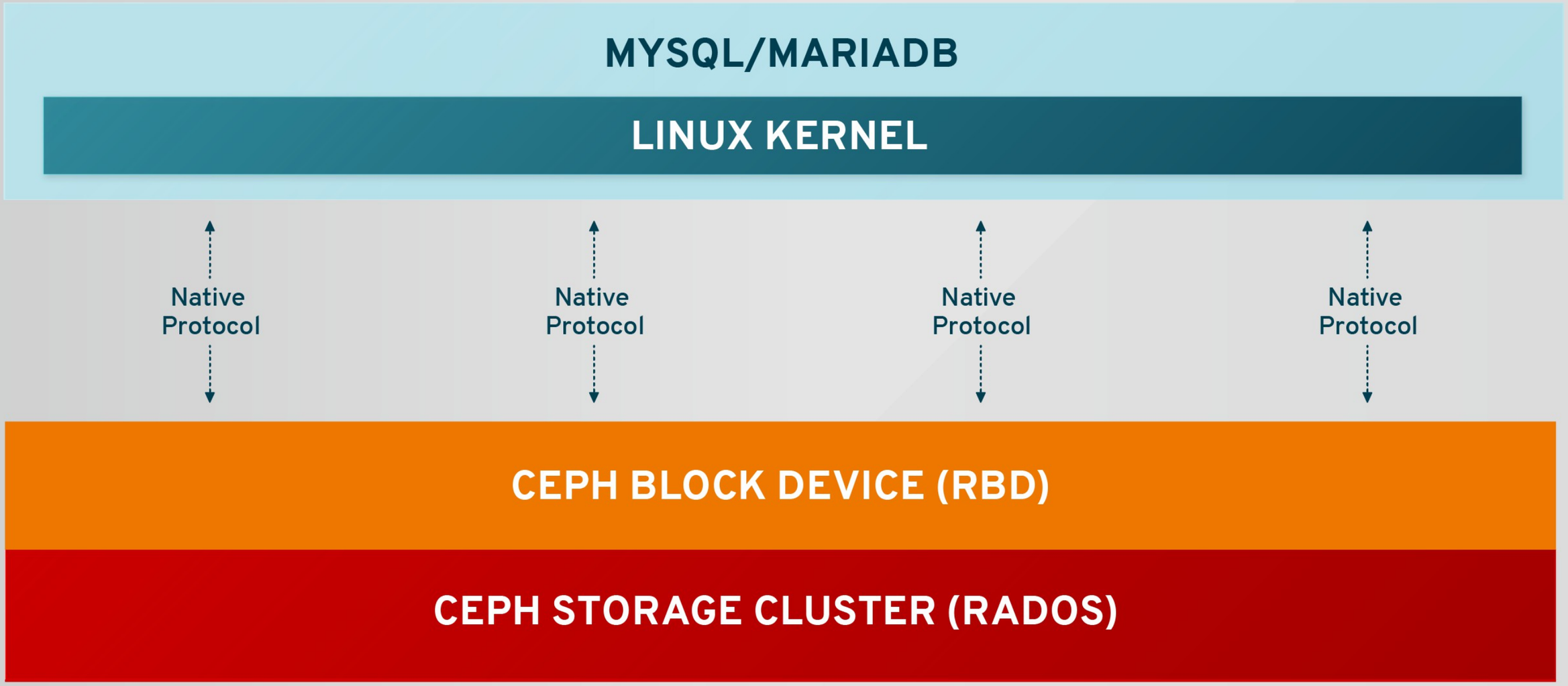
**ONE COPY PLUS PARITY**

- Cost-effective durability
- Expensive recovery

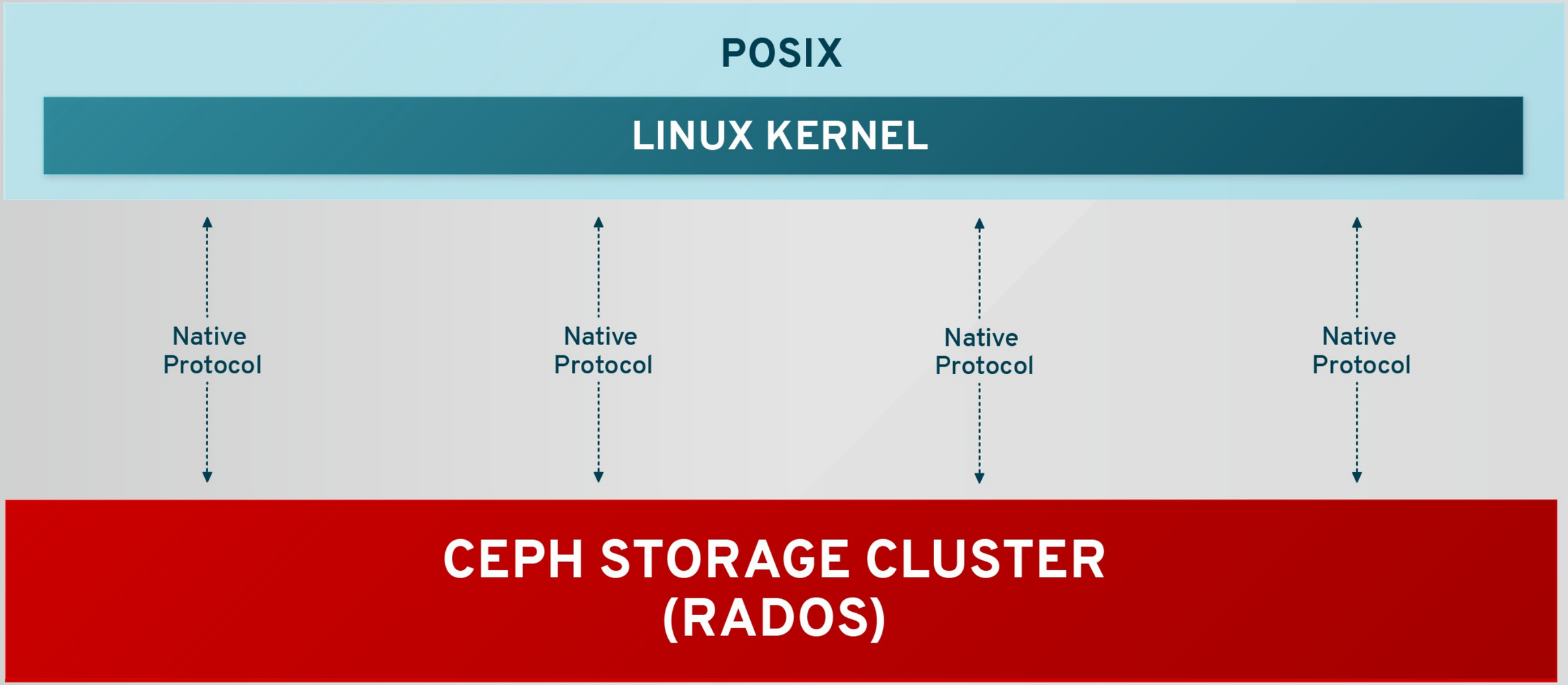
# WEBSCALE APPLICATIONS



# DATABASES



# KERNEL MODULE FLEXIBILITY




FEATURES

# FEATURE OVERVIEW

## SCALABILITY

**Scale-out Architecture**

- Grow a cluster from one node to thousands using the powerful CRUSH algorithm

**Automatic Rebalancing**

- Cope with outages automatically; grow and shrink capacity at will
- Make forklift upgrades and data migration obsolete

**Hot/Phased Software**

- Upgrade clusters in phases with no downtime

# FEATURE OVERVIEW

## APIs & EXTENSIBILITY

### Interoperable

- Amazon S3 and OpenStack Object Storage (Swift) API support
- Full integration with OpenStack block storage (via Cinder/Glance)

### Native Language Bindings

- Native bindings in a variety of languages provide the richest application experience

### Complete Management API

- Manage all cluster and object storage functions with a RESTful API

# FEATURE OVERVIEW

## SECURITY

### Access Control Lists

- Exert granular control over object storage user and bucket-level permissions.

### Quotas

- Set pool or object storage limits.

# FEATURE OVERVIEW

## RELIABILITY & OVERVIEW

### Striping, Erasure Coding, or Replication Across Nodes

- Enjoy data durability, high availability, and high performance

### Dynamic Block Resizing

- Expand or shrink block devices with no or minimal downtime

### Data Placement

- No need for lookups - every client can calculate where data is stored
- Policy-defined SLAs, performance tiers, and failure domains

### Automatic Failover

- Prevent failures from impacting integrity or availability

# FEATURE OVERVIEW

## MULTI-SITE & DISASTER RECOVERY

### Zones & Region Support

- Deploy topologies similar to S3 and others with a global namespace (object only).

### Data Center Synchronization

- Back-up full or partial sets of data between regions (object only).

### Read Affinity

- Always serve local copies of data to local users (object only).

# FEATURE OVERVIEW

## MULTI-SITE & DISASTER RECOVERY

### Export Snapshots to Geographically Dispersed Data Centers

- Institute disaster recovery (block only).

### Export Incremental Snapshots

- Minimize network bandwidth by only sending changes (block only).

### Stretched RADOS

- RADOS can be optimized for clusters operating across two geographic regions (if network latency permits)

# FEATURE OVERVIEW

## PERFORMANCE

### Copy-on-write Cloning

- Provision virtual machine (VM) block devices quickly (block only)

### In-memory Client Side Caching

- Accomadate both kernel and hypervisor (block only)

### Massive Parallelism for Data I/O

- Leverage a “client/cluster” model, not a “client/server” one

# FEATURE OVERVIEW

## COST EFFECTIVENESS

### Thin Provisioning

- Allows over-provisioning (block only).

### Commodity Hardware

- Tailor the price/performance mix to the workload.

### Heterogeneous Hardware

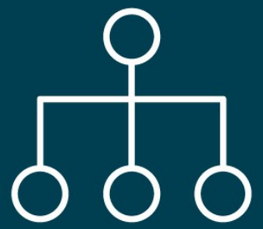
- Avoid having to replace older nodes as newer ones are added.

### Erasure Coding

- Enjoy the most cost-effective data-durability option.

# FEATURE OVERVIEW

## WEB-BASED MANAGEMENT



### Calamari Dashboard

- Get quick status on your cluster and take action with this integrated dashboard & management platform.

### Per-disk/Pool Performance Stats

- Measure IOPS over time, and spot bottlenecks quickly and easily.

### Diagnostics Workbench

- Survey cluster-wide component status with this speedy tool.

### Management

- Monitor disk usage and adjust cluster, pool, and OSD settings.


redhat