Ceph Rados Gateway

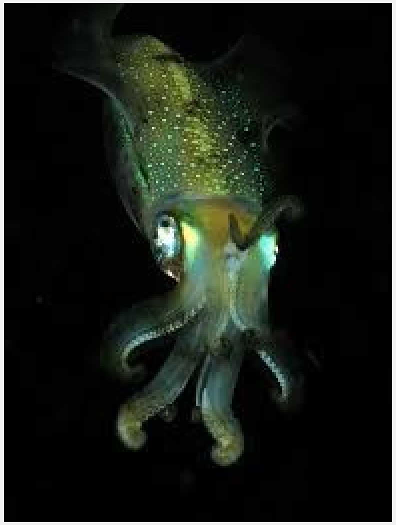
Orit Wasserman
owasserm@redhat.com
Fosdem 2016

AGENDA

- Short Ceph overview
- Rados Gateway architecture
- What's next
- questions

Ceph architecture

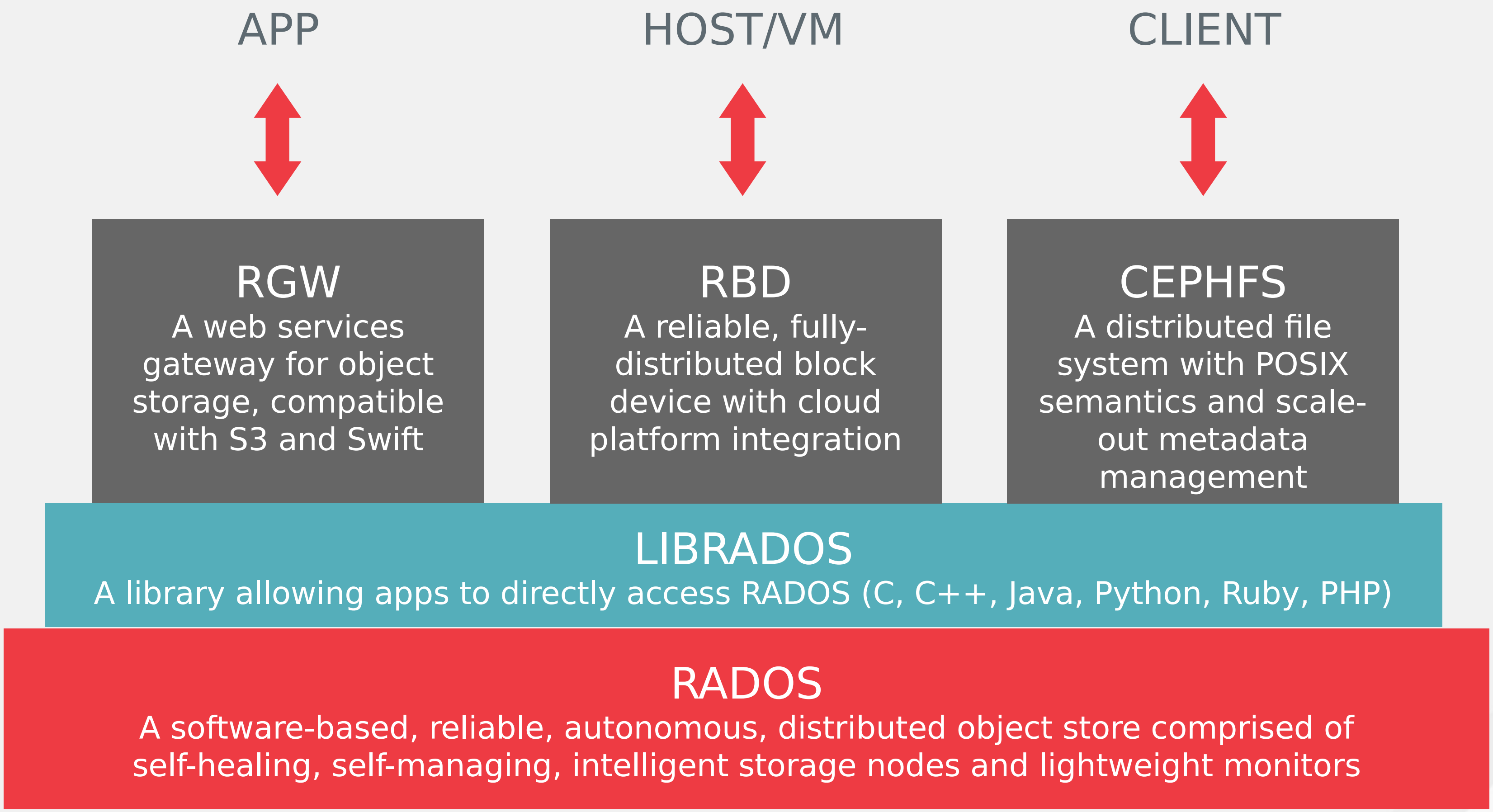
Cephalopod



Ceph

- Open source
- Software defined storage
- Distributed
- No single point of failure
- Massively scalable
- Self healing
- Unified storage: object, block and file

Ceph architecture



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

Rados

- Reliable Distributed Object Storage
- Replication
- 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 node

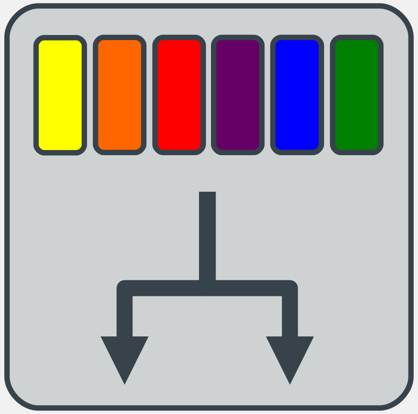
- 10s to 10000s in a cluster
- One per disk (or one per SSD, RAID group…)
- Serve stored objects to clients
- Intelligently peer for replication & recovery

Monitor node

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

Crush

- Pseudo-random placement algorithm
- Fast calculation, no lookup
- Ensures even distribution
- Repeatable, deterministic
- Rule-based configuration
  - specifiable replication
  - infrastructure topology aware
  - allows weighting

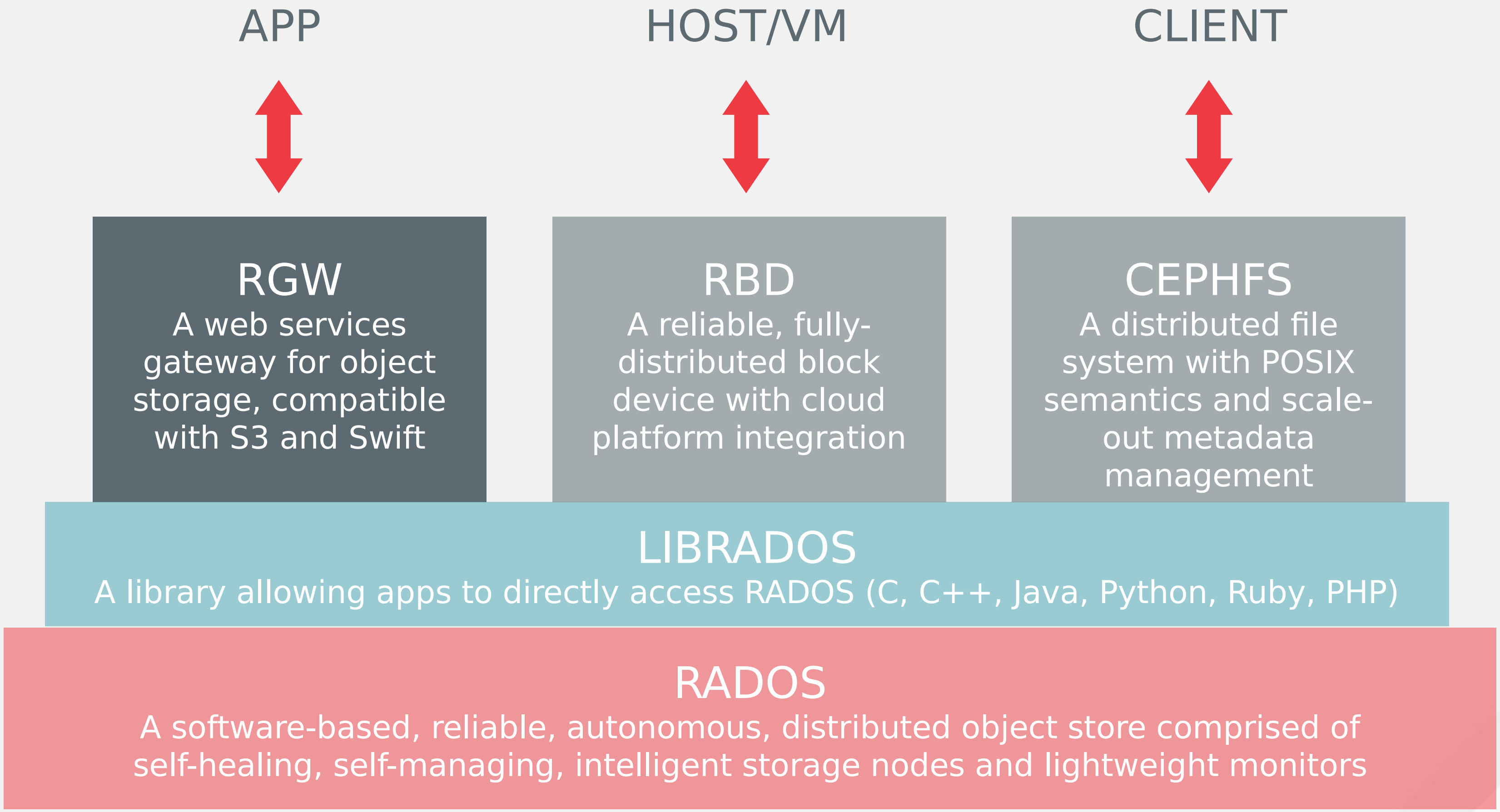


Librados API

- 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

Rados Gateway

Rados Gateway



APP

HOST/VM

CLIENT

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

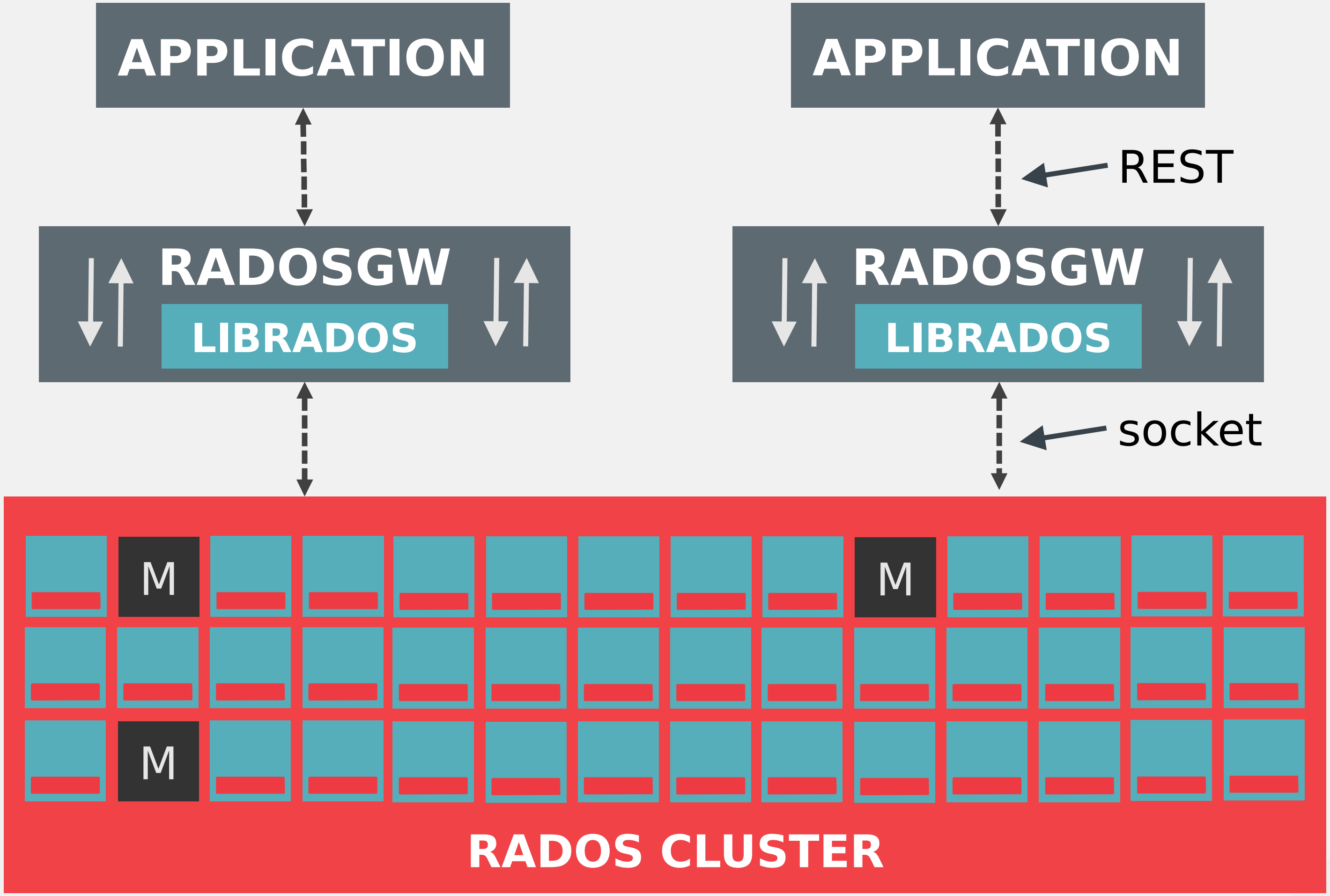
RBD
A reliable, fully-distributed block device with cloud platform integration

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

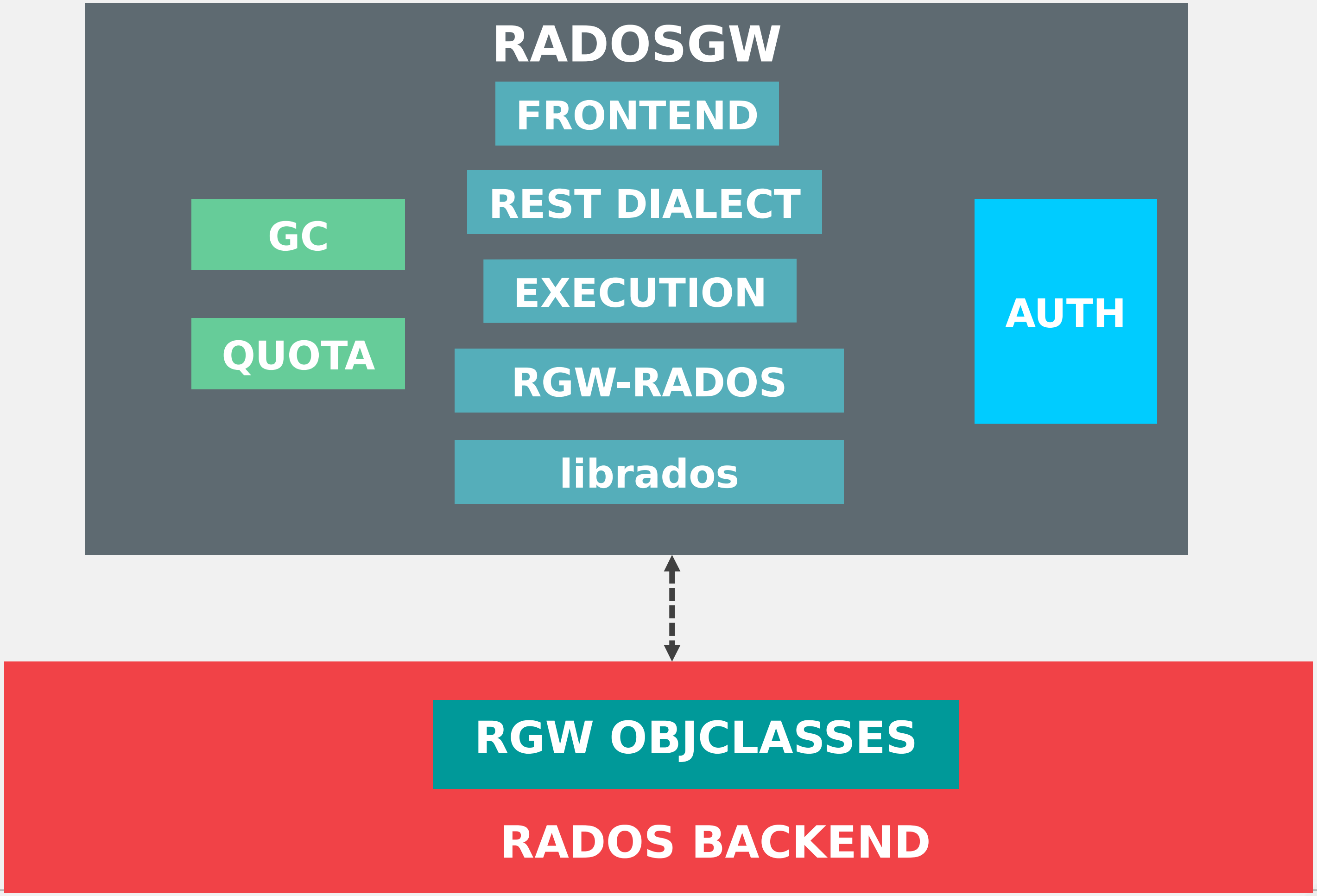
LIBRADOS
A library allowing apps to directly access RADOS (C, C++, Java, Python, Ruby, PHP)

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

Rados Gateway



RGW



RADOSGW

FRONTEND

REST DIALECT

EXECUTION

RGW-RADOS

librados

GC

QUOTA

AUTH

RGW OBJCLASSES

RADOS BACKEND

RGW Components

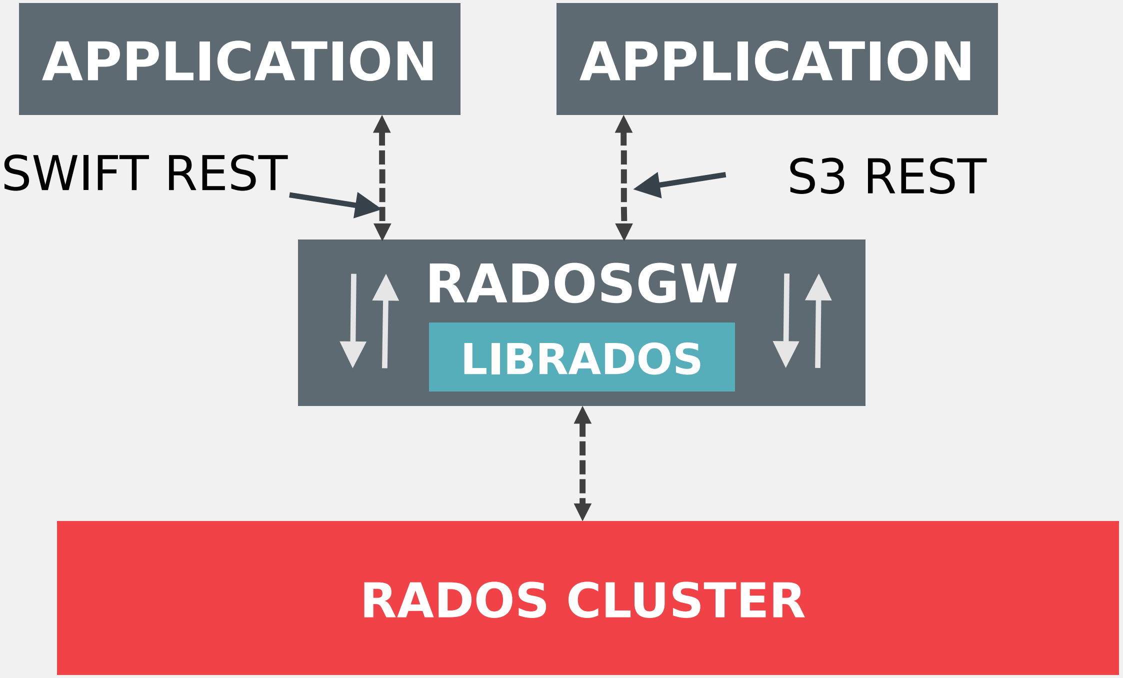
- Frontend
  - FastCGI - external web servers
  - Civetweb – embedded web server
- Rest Dialect
  - S3
  - Swift
  - Other API
- Execution layer – common layer for all dialects

RGW Components

- RGW Rados – manages RGW data by using rados
  - object striping
  - atomic overwrites
  - bucket index handling
  - Object classes that run on the OSDs
- Quota - handles user or bucket quotas.
- Authentication - handle users authentication
- GC - Garbage collection mechanism that runs in the background.

RESTful OBJECT STORAGE

- Data
  - Users
  - Buckets
  - Objects
  - ACLs
- Authentication
- APIs
  - S3
  - Swift



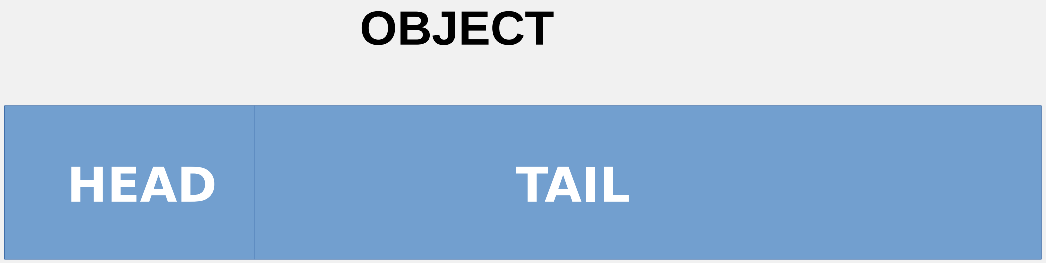
RGW vs RADOS object

- RADOS
  - Limited object sizes
  - Mutable objects
  - Not indexed
  - No per-object ACLs
- RGW
  - Large objects (Up to a few TB per object)
  - Immutable objects
  - Sorted bucket listing
  - Permissions

RGW objects

- Large objects
- Fast small object access
- Fast access to object attributes
- Buckets can consist of a very large number of objects

RGW objects



OBJECT

HEAD

TAIL

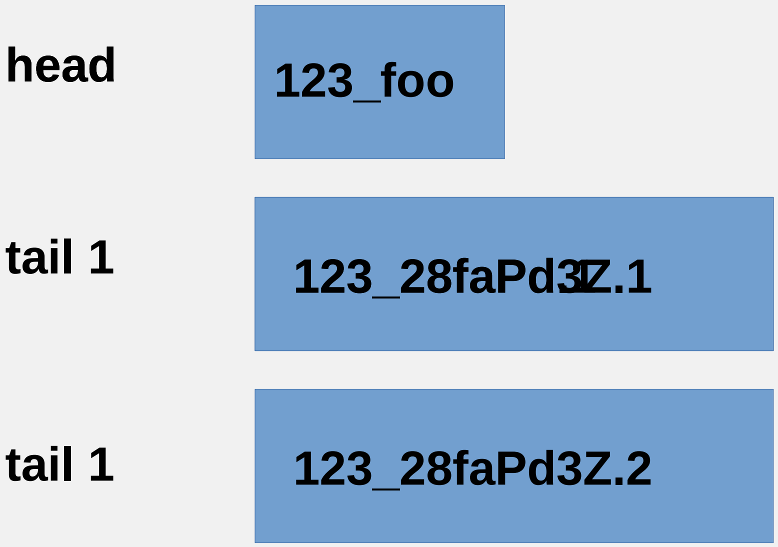
- Head
  - Single rados object
  - Object metadata (acls, user attributes, manifest)
  - Optional start of data
- Tail
  - Striped data
  - 0 or more rados objects

RGW Objects

OBJECT: foo

BUCKET: boo

BUCKET ID: 123



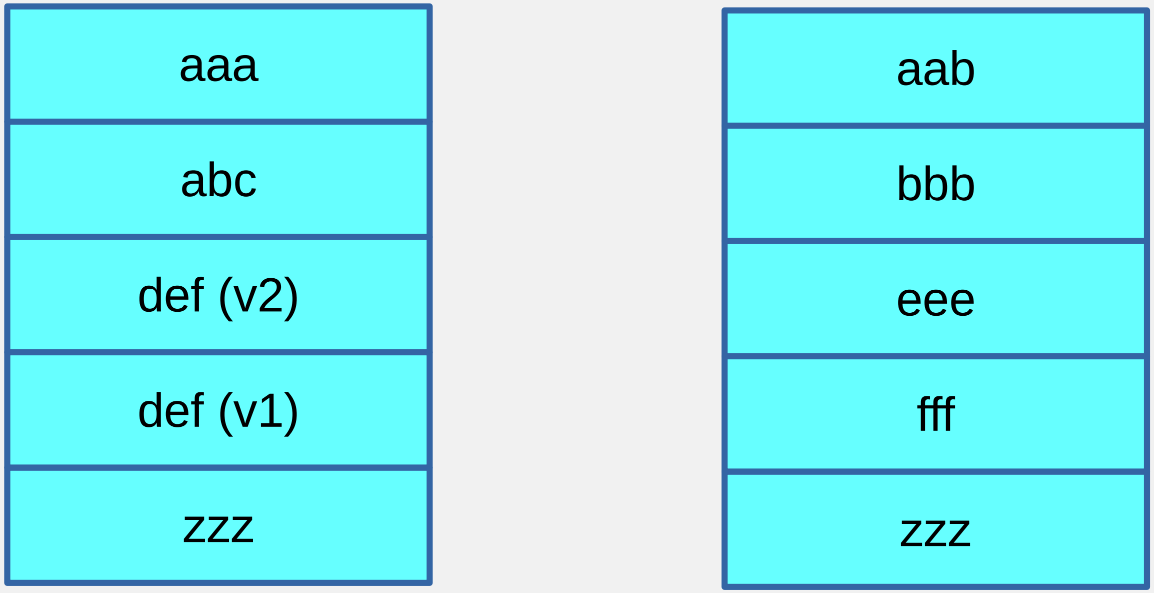
head 123_foo

tail 1 123_28faPd3Z.1

tail 1 123_28faPd3Z.2

RGW bucket index

BUCKET INDEX

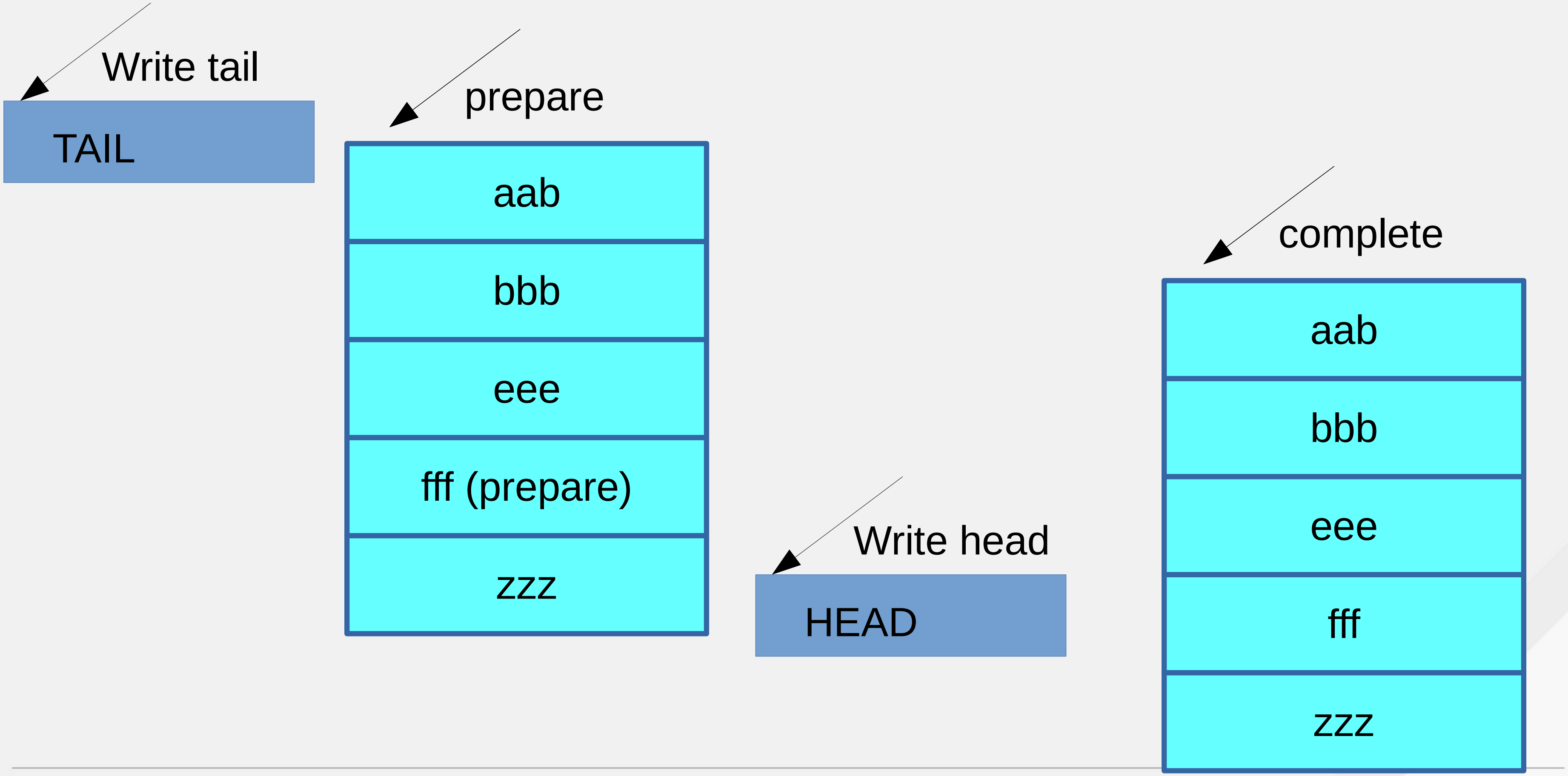


Shard 1	Shard 2
aaa	aab
abc	bbb
def (v2)	eee
def (v1)	fff
zzz	zzz

RGW object creation

- Update bucket index
- Create head object
- Create tail objects
- All those operations need to be consist

RGW object creation


Write tail

TAIL

prepare

aab
bbb
eee
fff (prepare)
zzz

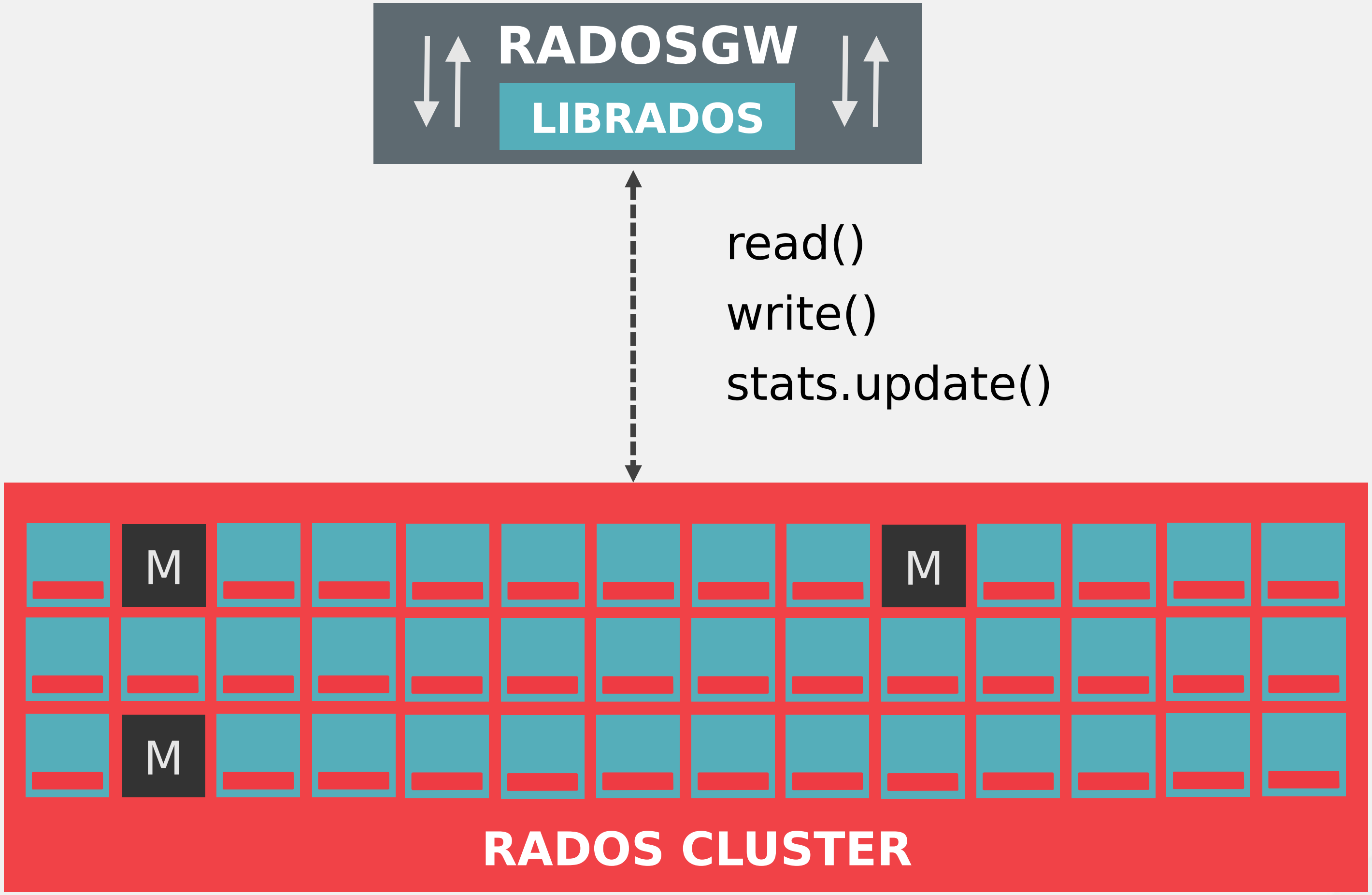
Write head

HEAD

complete

aab
bbb
eee
fff
zzz

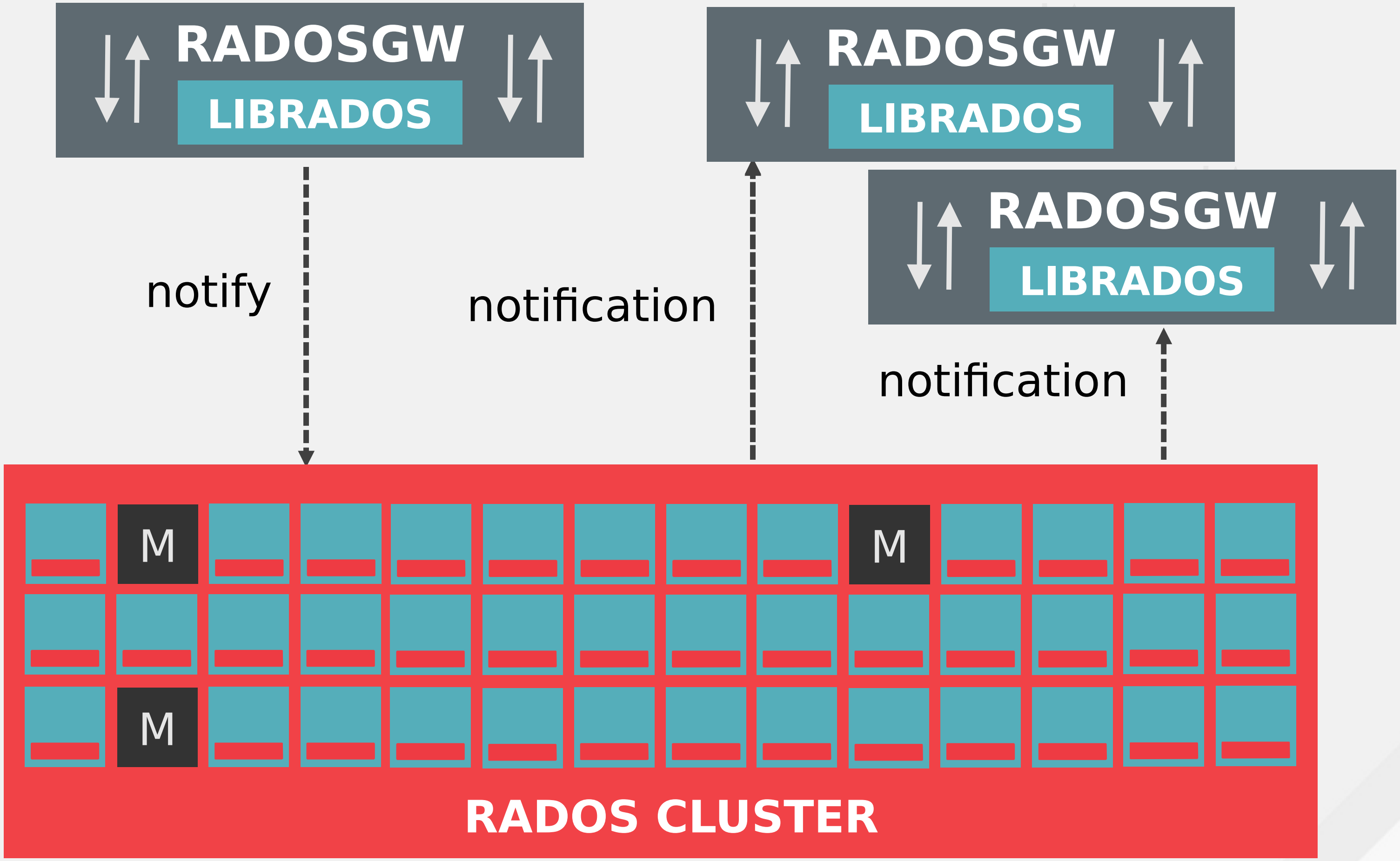
RGW quota


RADOSGW
LIBRADOS
read()
write()
stats.update()
M
M
M
RADOS CLUSTER

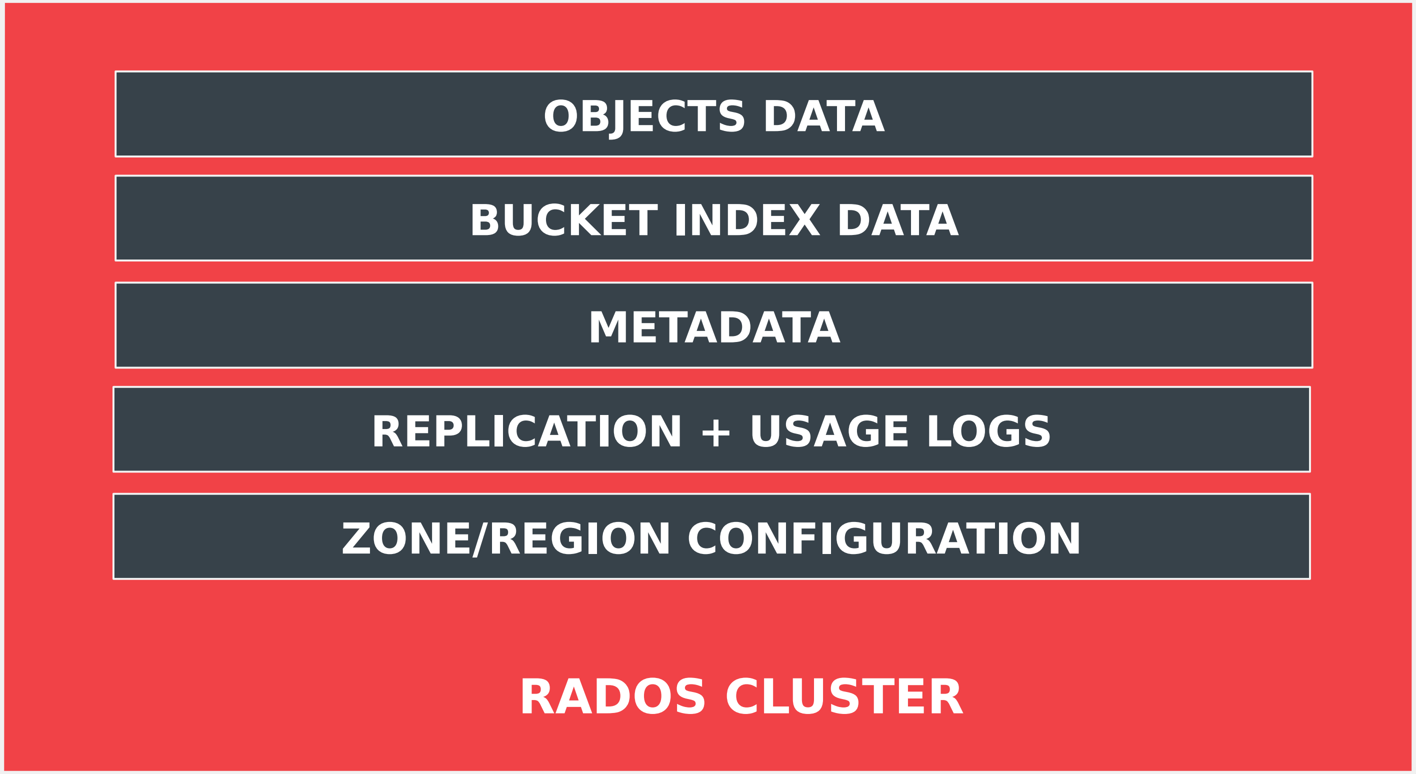
RGW metadata cache

- Metadata needed for each request:
  - User Info
  - Bucket Entry Point
  - Buck Instance Info

RGW metadata cache



RGW rados data



OBJECTS DATA

BUCKET INDEX DATA

METADATA

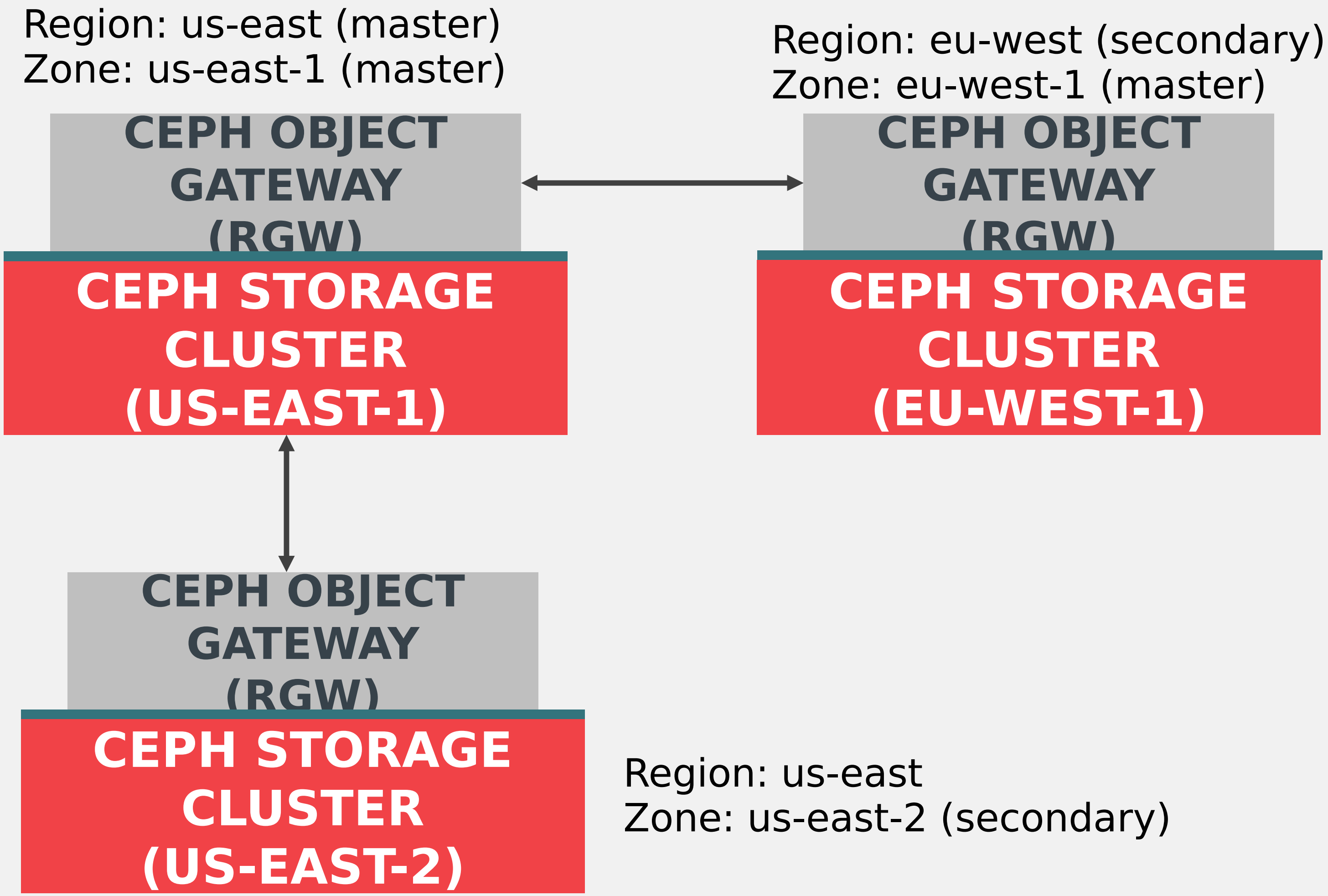
REPLICATION + USAGE LOGS

ZONE/REGION CONFIGURATION

RADOS CLUSTER

Multisite

Regions and zones



Region: us-east (master)
Zone: us-east-1 (master)

CEPH OBJECT GATEWAY (RGW)

CEPH STORAGE CLUSTER (US-EAST-1)

Region: eu-west (secondary)
Zone: eu-west-1 (master)

CEPH OBJECT GATEWAY (RGW)

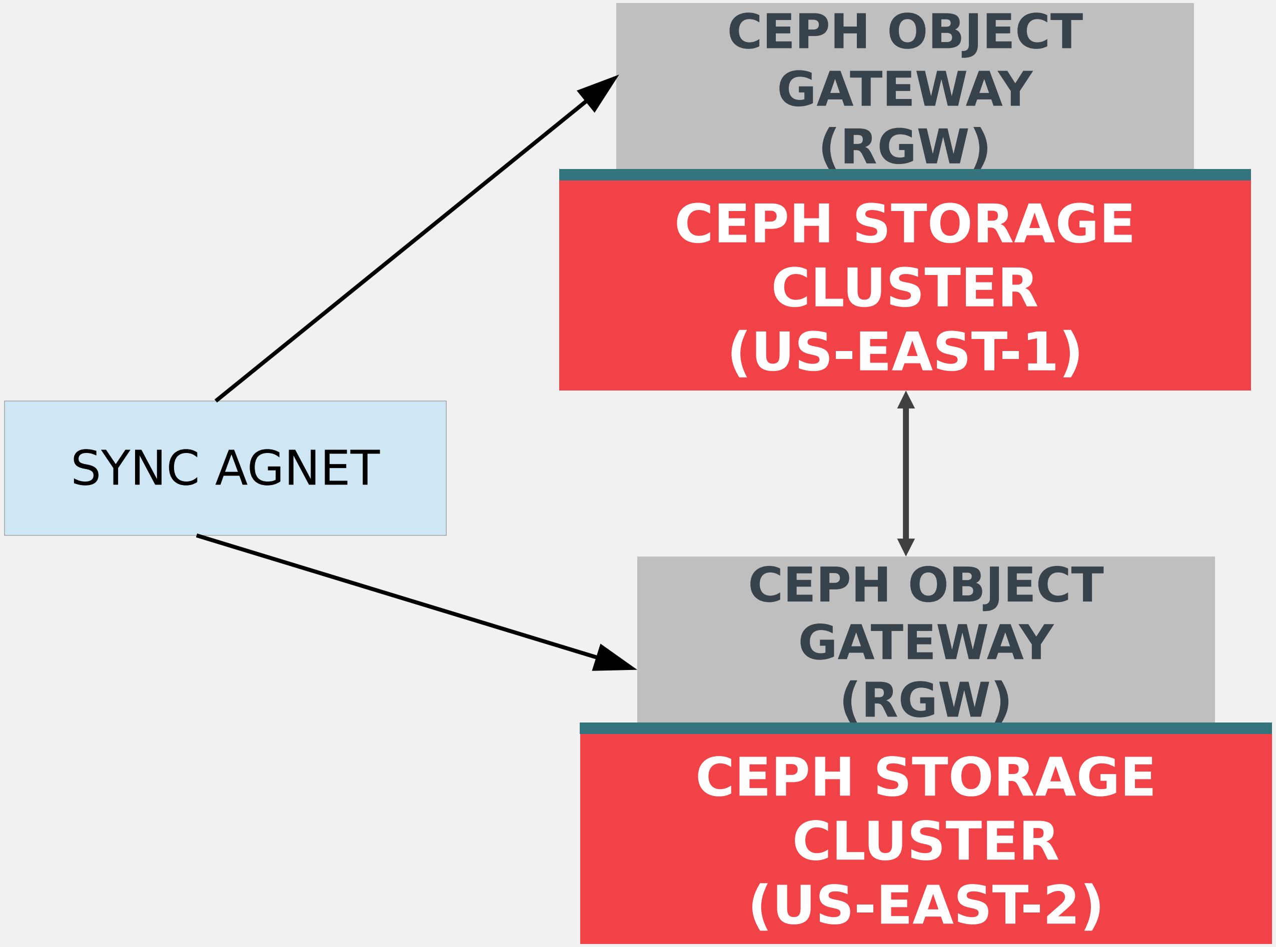
CEPH STORAGE CLUSTER (EU-WEST-1)

CEPH OBJECT GATEWAY (RGW)

CEPH STORAGE CLUSTER (US-EAST-2)

Region: us-east
Zone: us-east-2 (secondary)

RGW sync agent



SYNC AGNET

CEPH OBJECT GATEWAY (RGW)

CEPH STORAGE CLUSTER (US-EAST-1)

CEPH OBJECT GATEWAY (RGW)

CEPH STORAGE CLUSTER (US-EAST-2)

Problems

- No active/active
- External utility
- Confusing configuration semantics

What's next

WHAT'S NEXT

- Multi tenancy
  - different users on the same tenant can share data
  - Buckets names are not unique across tenants
- Object expiration
- AWS4
- NFS
  - For migration from NFS to RGW
  - Based on NFS Ganesha
- LibRGW – API to RGW, used by Ganesha
- Static website – root domain support
- Keystone v3
- Swift Large Object

New multisite

- New implementation as part of RGW
- Namespaces
- Simpler configuration
- Active/active support

THANK YOU

owasserm@redhat.com

ceph-users@ceph.com

ceph-devel@ceph.com

OFTC #ceph