CEPHALOPODS AND SAMBA

IRA COOPER – SNIA SDC 2016.09.18

DISCLAIMER

- These opinions are my opinions.

- They do not represent promises from:
  - Red Hat Inc.
  - Samba Team
  - Me
  - My Mom

AGENDA

- CEPH Architecture.
  - Why CEPH?
  - RADOS
  - RGW
  - CEPHFS

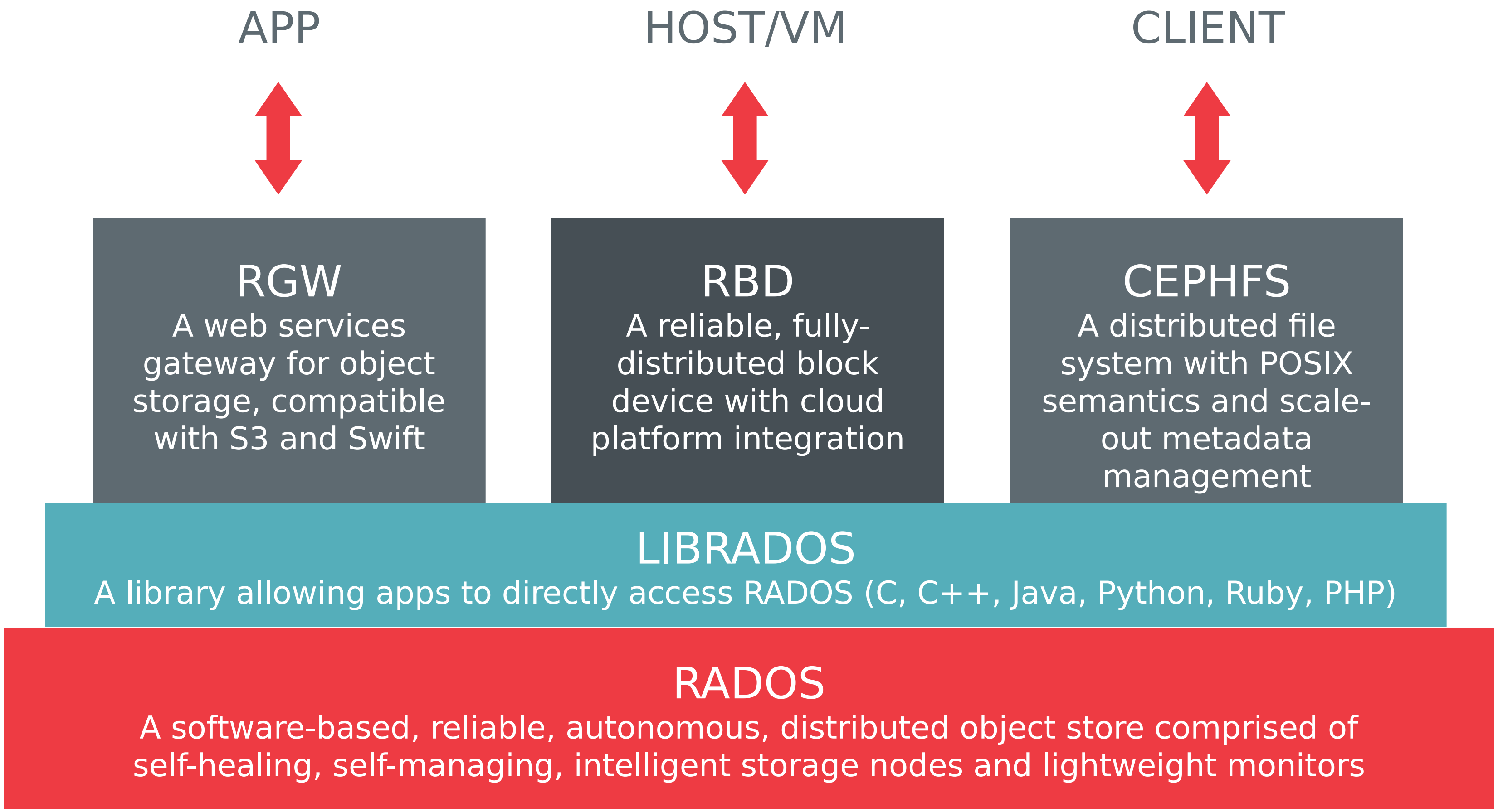
- Current Samba integration with CEPH.
- Future directions.

- Maybe a demo?

CEPH MOTIVATING PRINCIPLES

- All components must scale horizontally.
- There can be no single point of failure.
- The solution must be hardware agnostic.
- Should use commodity hardware.
- Self-manage whenever possible.
- Open source.

ARCHITECTURAL COMPONENTS



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

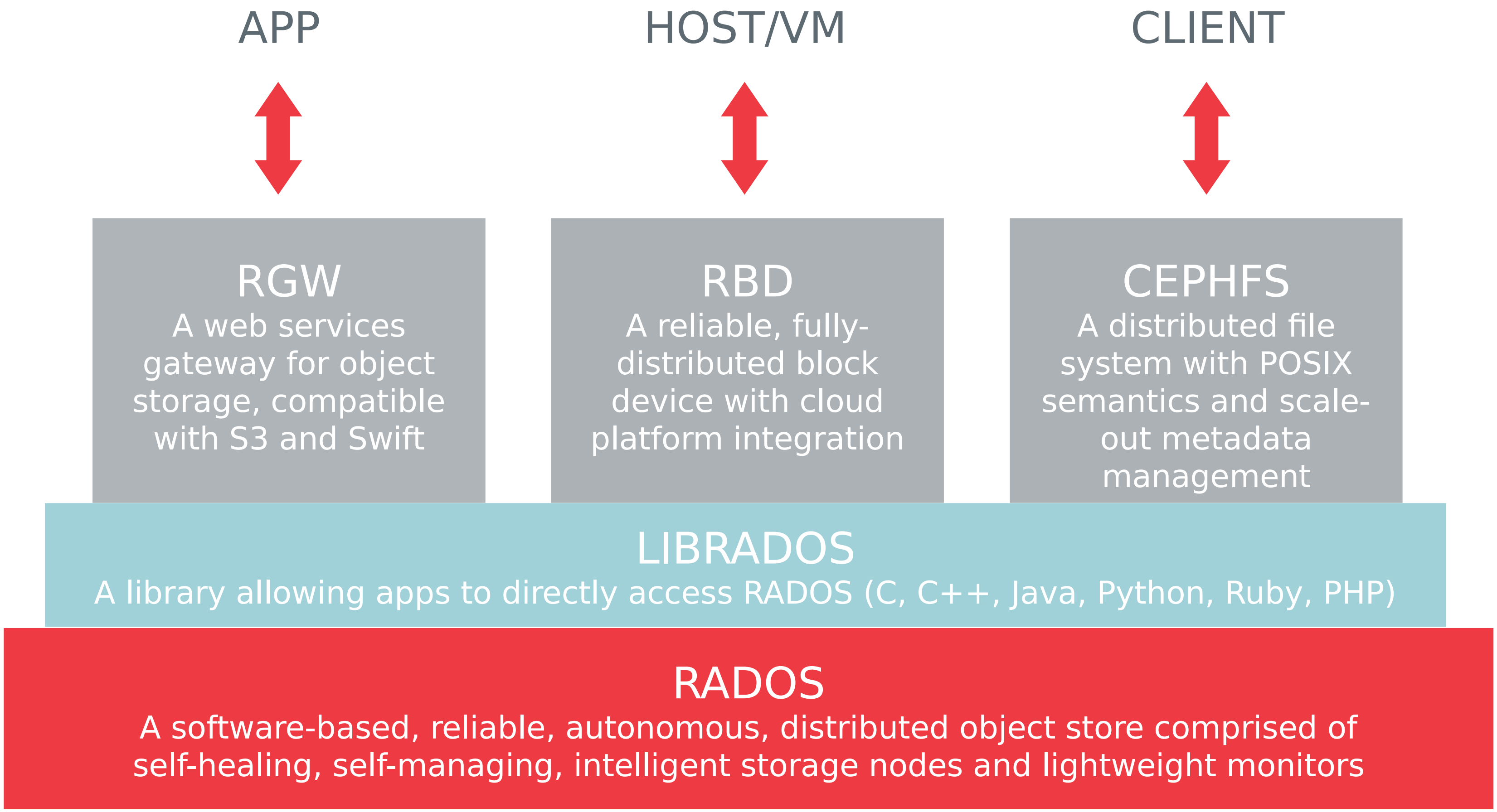
LIBRADOS

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

ARCHITECTURAL COMPONENTS



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)

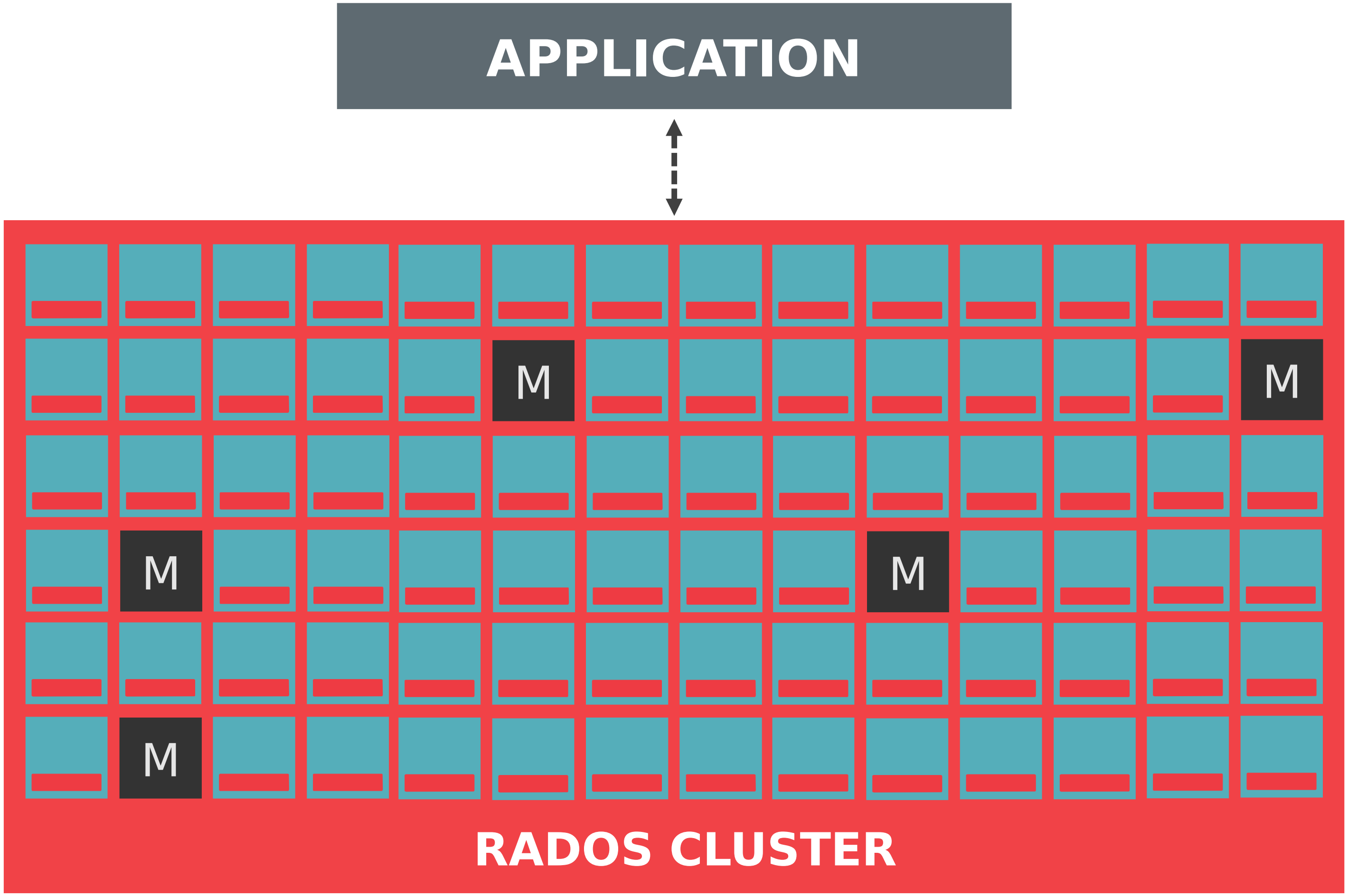
RADOS

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

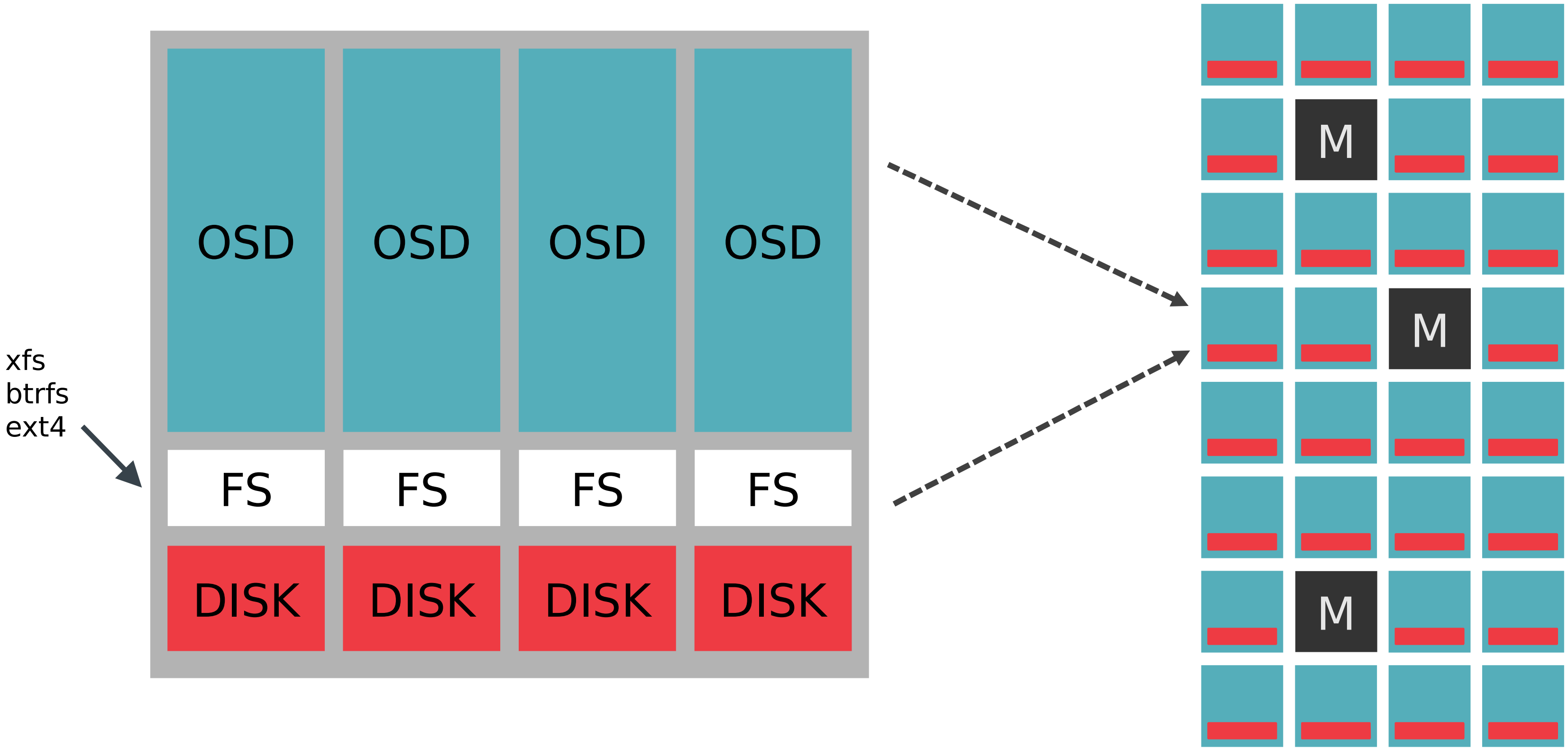
RADOS

- Flat object namespace within each pool
- Rich object API (librados)
  - Bytes, attributes, key/value data
  - Partial overwrite of existing data
  - Single-object compound operations
  - RADOS classes (stored procedures)
- Strong consistency (CP system)
- Infrastructure aware, dynamic topology
- Hash-based placement (CRUSH)
- Direct client to server data path

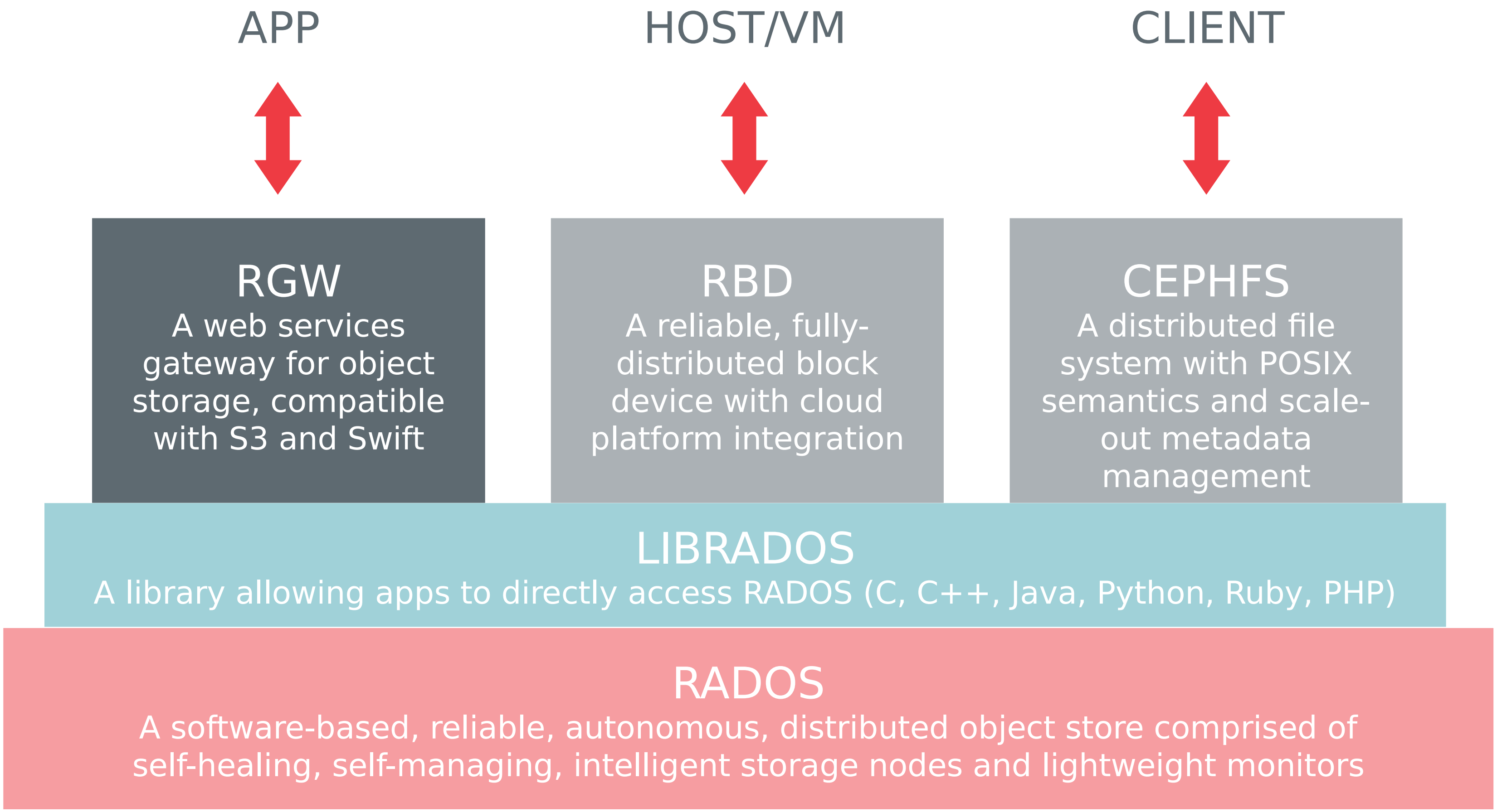
RADOS CLUSTER



OBJECT STORAGE DAEMONS



ARCHITECTURAL COMPONENTS



APP

HOST/VM

CLIENT

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

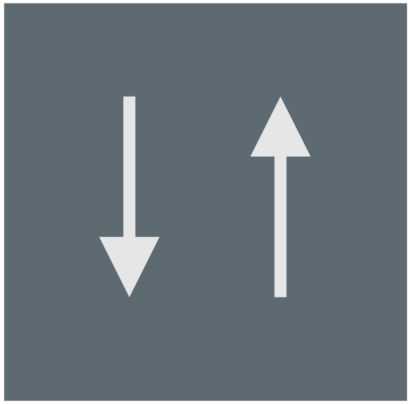
RBD
A reliable, fully-distributed block device with cloud platform integration

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

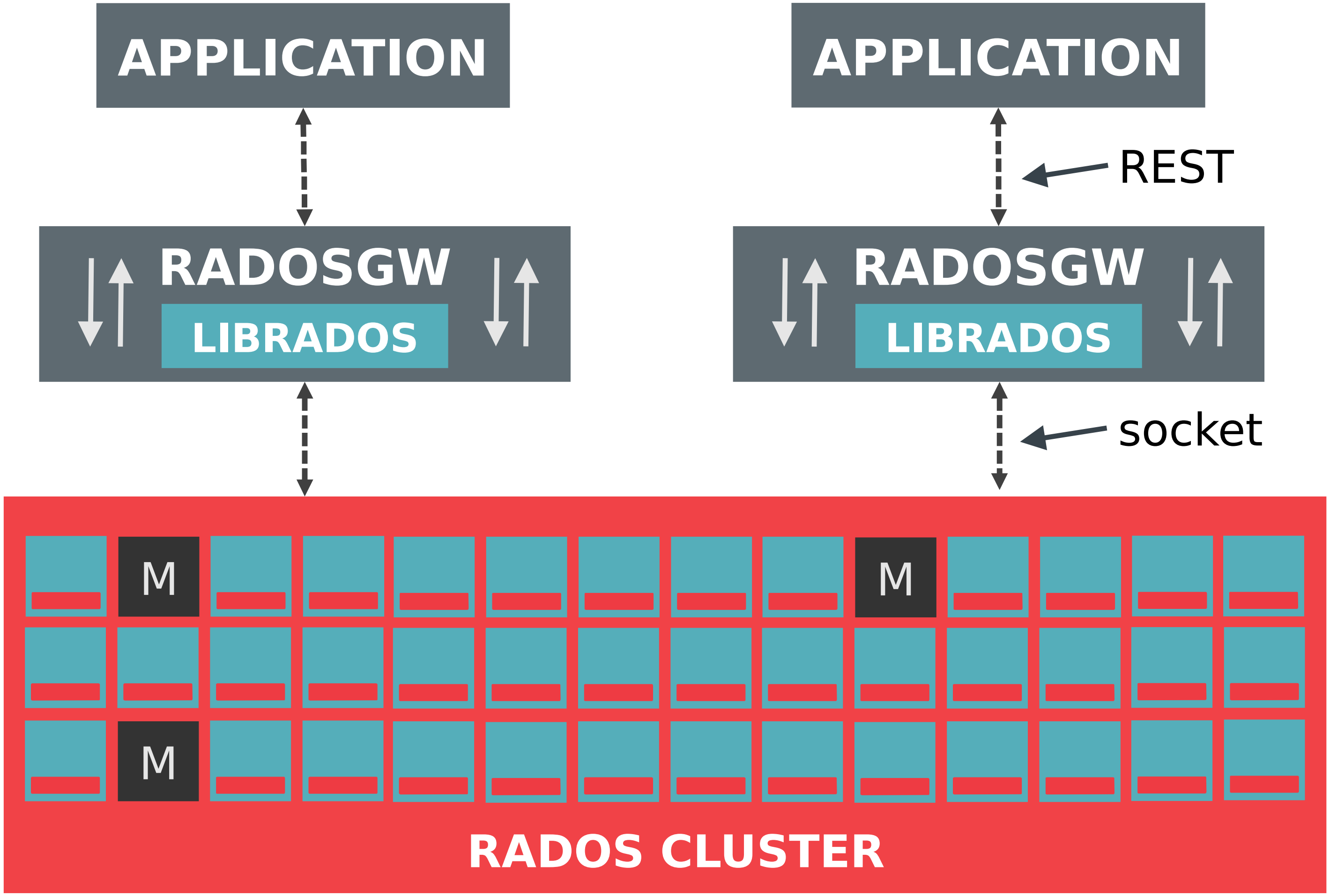
RADOSGW MAKES RADOS WEBBY



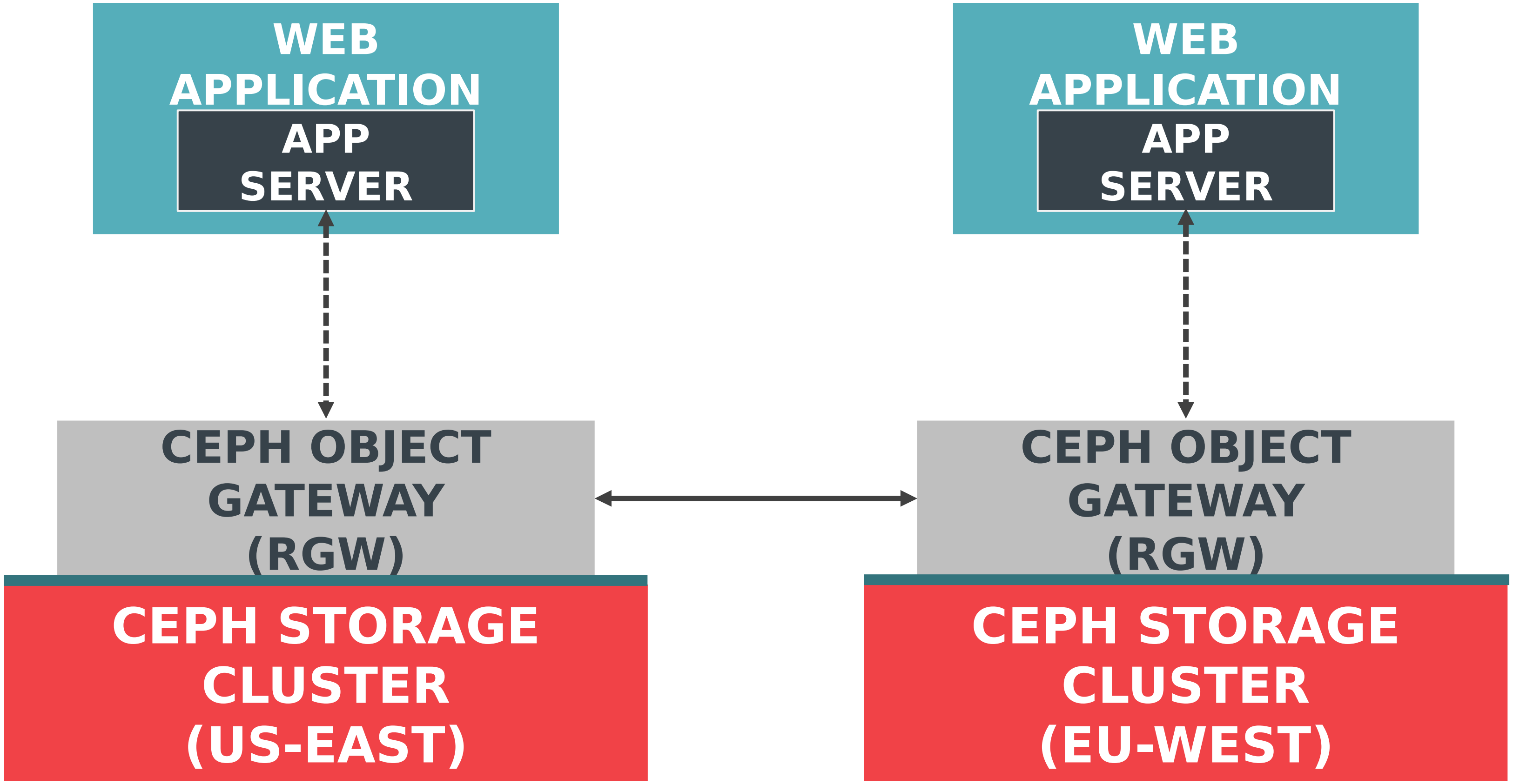
RADOSGW:

- REST-based object storage proxy
- Uses RADOS to store objects
    - Stripes large RESTful objects across many RADOS objects
    - Space efficient for small objects
- API supports buckets, accounts
- Usage accounting for billing
- Compatible with S3 and Swift applications

THE RADOS GATEWAY



MULTI-SITE OBJECT STORAGE



WEB APPLICATION

APP SERVER

CEPH OBJECT GATEWAY (RGW)

CEPH STORAGE CLUSTER (US-EAST)

WEB APPLICATION

APP SERVER

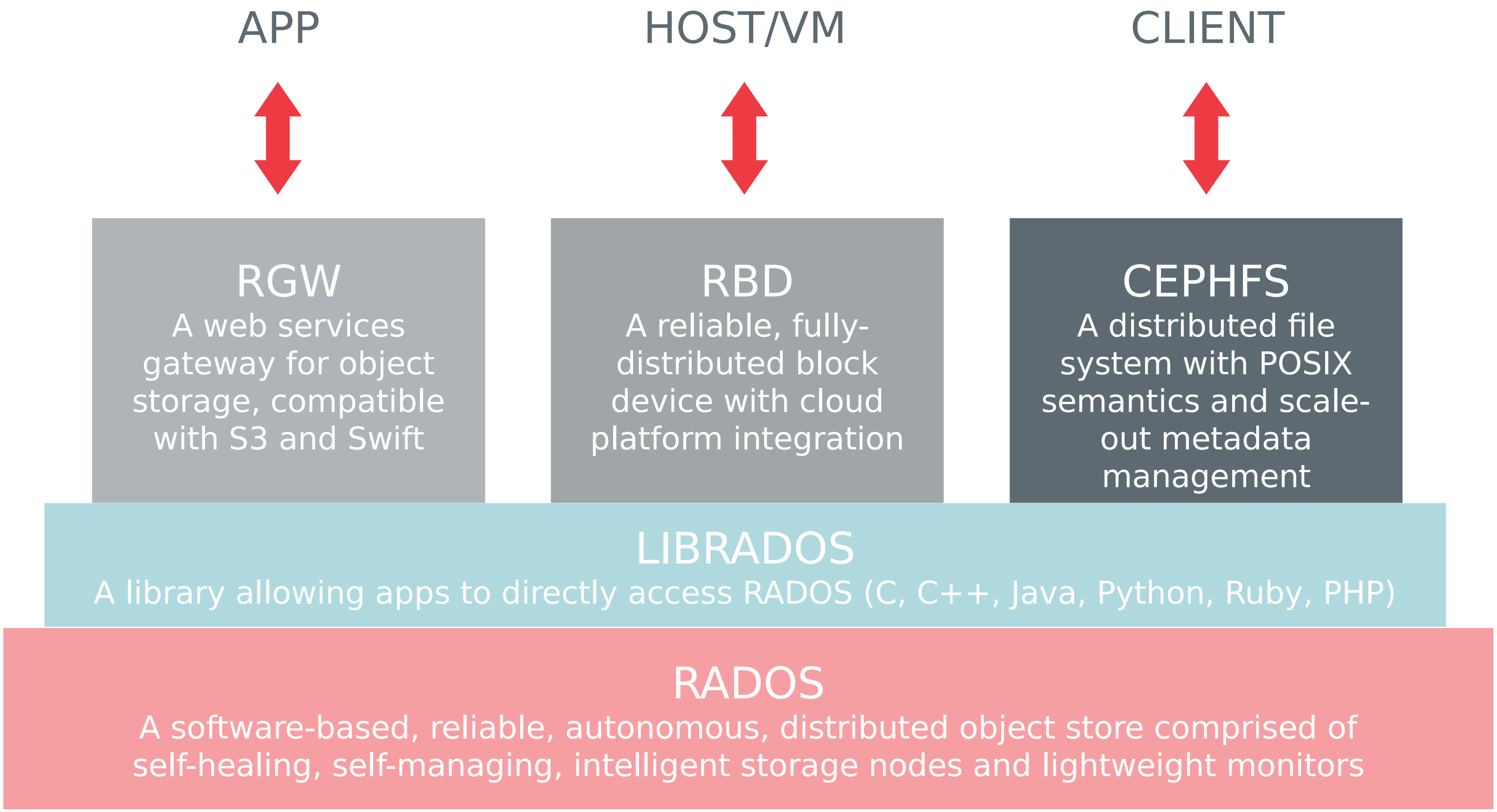
CEPH OBJECT GATEWAY (RGW)

CEPH STORAGE CLUSTER (EU-WEST)

FEDERATED RGW

- Zones and regions
  - Topologies similar to S3 and others
  - Global bucket and user/account namespace
- Cross data center synchronization
  - Asynchronously replicate buckets between regions
- Read affinity
  - Serve local data from local DC
  - Dynamic DNS to send clients to closest DC

ARCHITECTURAL COMPONENTS



APP

HOST/VM

CLIENT

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

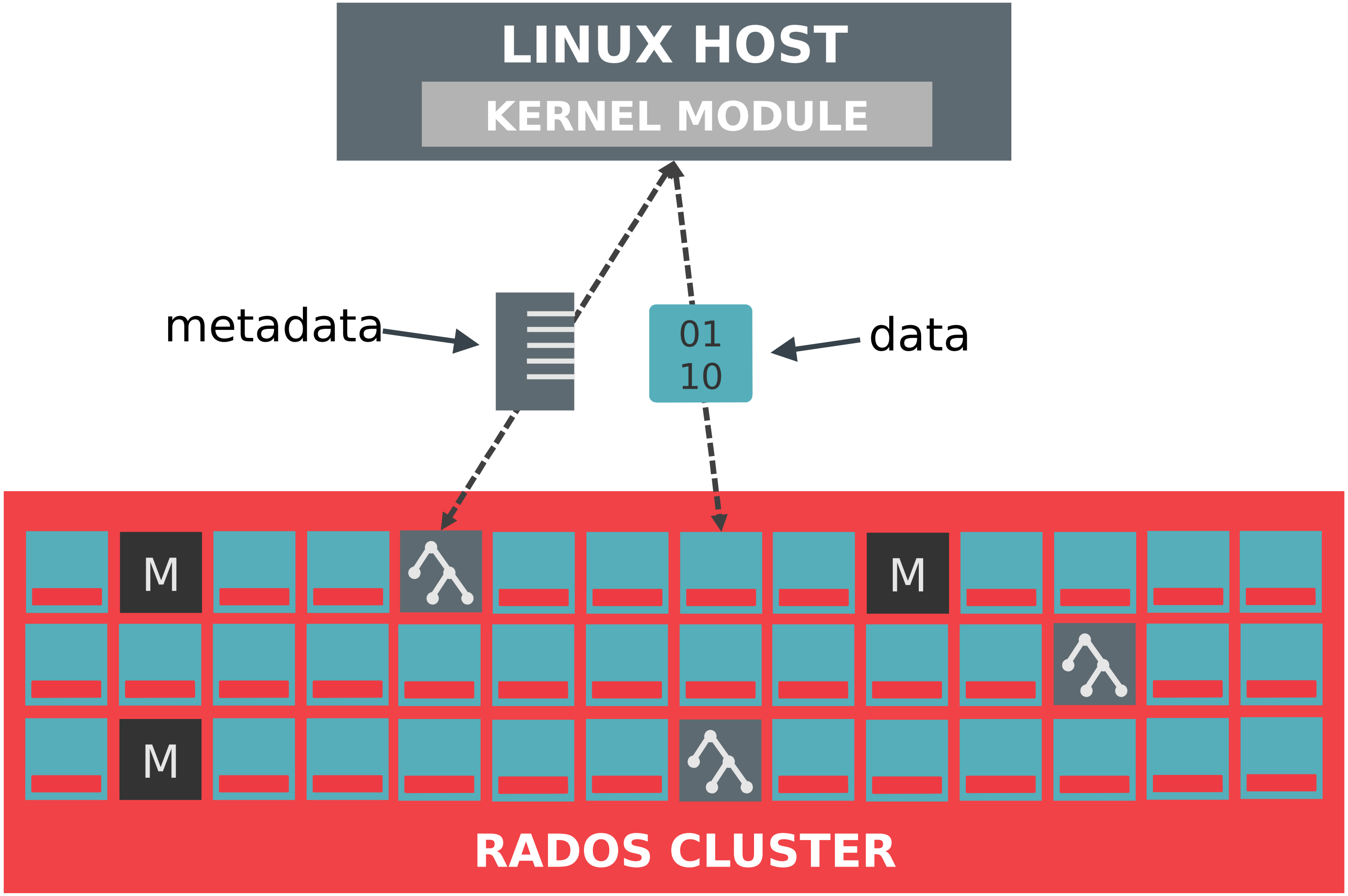
RBD
A reliable, fully-distributed block device with cloud platform integration

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

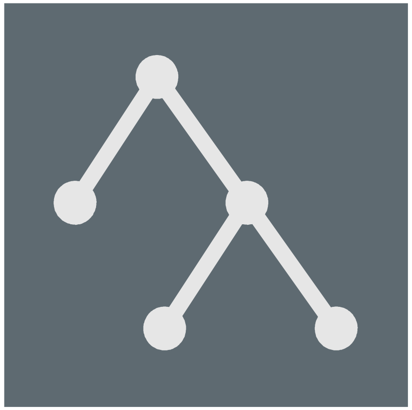
LIBRADOS
A library allowing apps to directly access RADOS (C, C++, Java, Python, Ruby, PHP)

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

SEPARATE METADATA SERVER



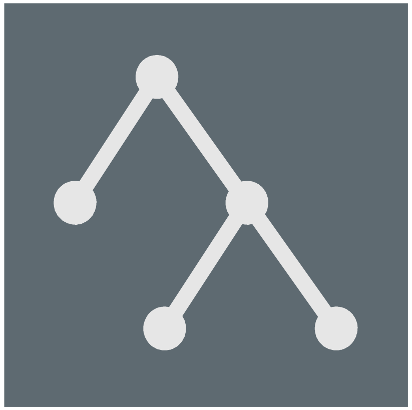
SCALABLE METADATA SERVERS



METADATA SERVER

- Manages metadata for a POSIX-compliant shared filesystem
    - Directory hierarchy
    - File metadata (owner, timestamps, mode, etc.)
- Clients stripe file data in RADOS
    - MDS not in data path
- MDS stores metadata in RADOS
    - Key/value objects
- Dynamic cluster scales to 10s or 100s
- Only required for shared filesystem

METADATA SERVERS – FUTURE

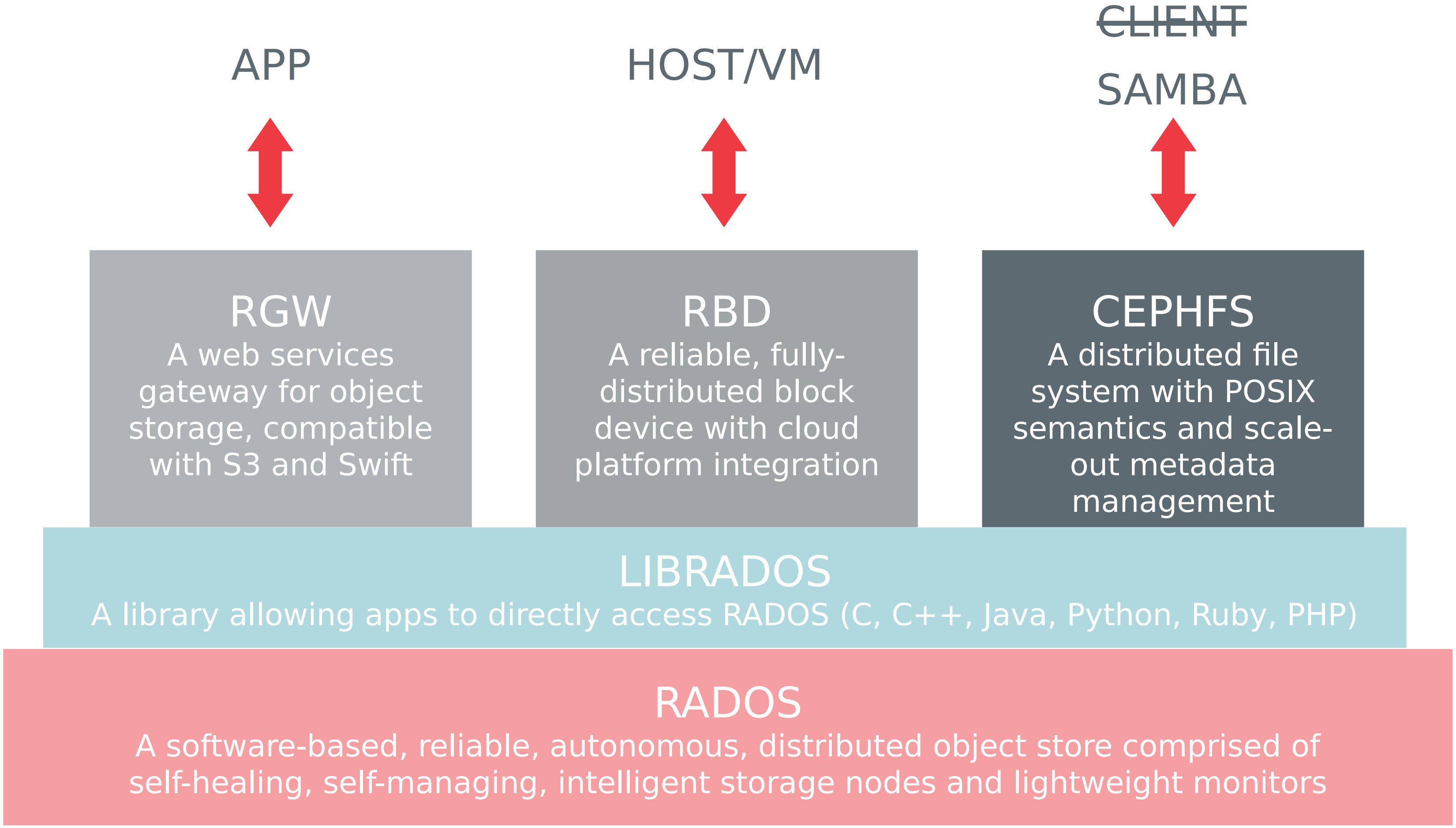


METADATA SERVER

- Sharding of the MDS (MetaData Server)
  - More scalable performance.
- Active – Passive Failover
  - Allowing for better availability
- Both features are in the codebase
  - In active development
  - Not production ready

SAMBA - TODAY

ARCHITECTURAL COMPONENTS



APP

HOST/VM

~~CLIENT~~

SAMBA

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

SAMBA INTEGRATION

- vfs_ceph
  - Since 2013.
  - Used as the outline for vfs_glusterfs
  - Been in testing in teuthology for a while now.
  - Patches up to be used as a testbed for statx.

- ACL Integration?
  - Patchset for POSIX ACLs committed for Samba 4.5
    - Thank you to Zheng Yan
  - Work on RichACLs is on going.

CTDB INTEGRATION

- fcntl locks
  - Does any filesystem get this right at the start.
  - 0/2 so far.
  - Ceph's have been fixed, they work for CTDB.
    - If you tweak the time outs.
      - But these tweaks aren't production ready!

- Both kernel and FUSE clients have been tested
  - CephFS team recommends ceph_fuse.
  - That's what our initial integration used.

CTDB INTEGRATION

- CTDB “fcntl lock” dependency removal.
  - etcd
    - Battle tested.
    - Push other config info into etcd?
      - nodes
      - public_addresses
    - The demo will show basic etcd integration.
      - Thank you to Jose Rivera for his work here.
  - Zookeeper
    - Much the same as etcd for this use.
    - Not working on it now.

DEMO

FUTURE DIRECTIONS – OBJECT

- RGW
  - Export object data as files.
  - Export files as object data?
    - Not today in ceph.
  - Integrate where?
    - S3
    - RADOS
    - Librgw
    - CephFS / vfs_ceph
- S3
  - Not being worked on at this time.
- Non file system based locking makes all this possible.

QUESTIONS?

THANK YOU!

Ira Cooper

SAMBA TEAM

✉ ira@wakeful.net