CEPHALOPODS AND SAMBA

IRA COOPER - SambaXP 2016.05.12

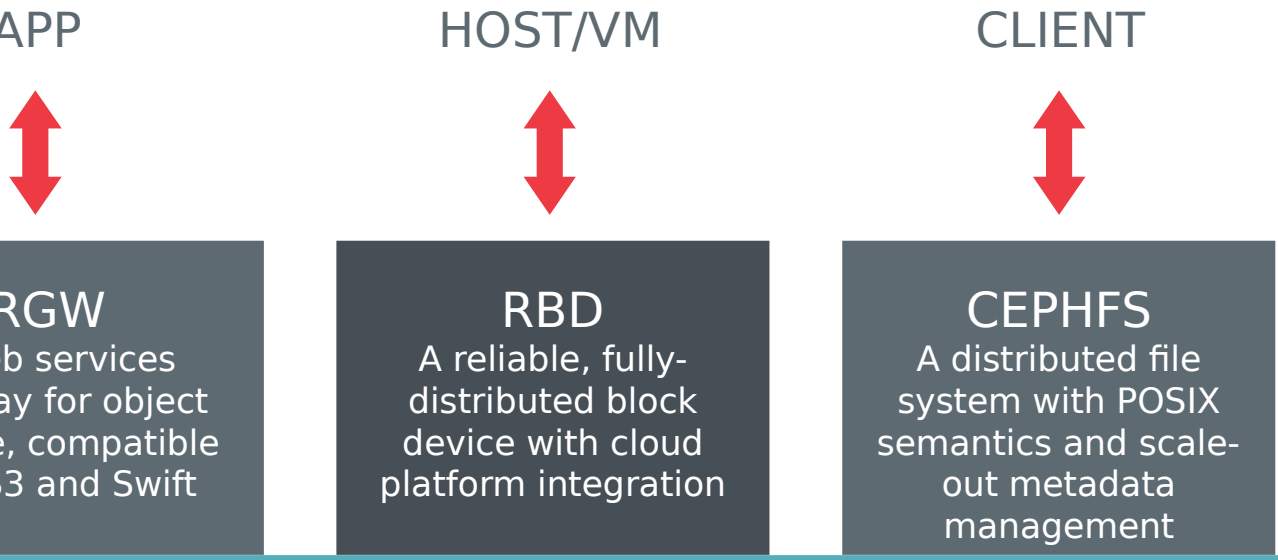
AGENDA

- 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



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

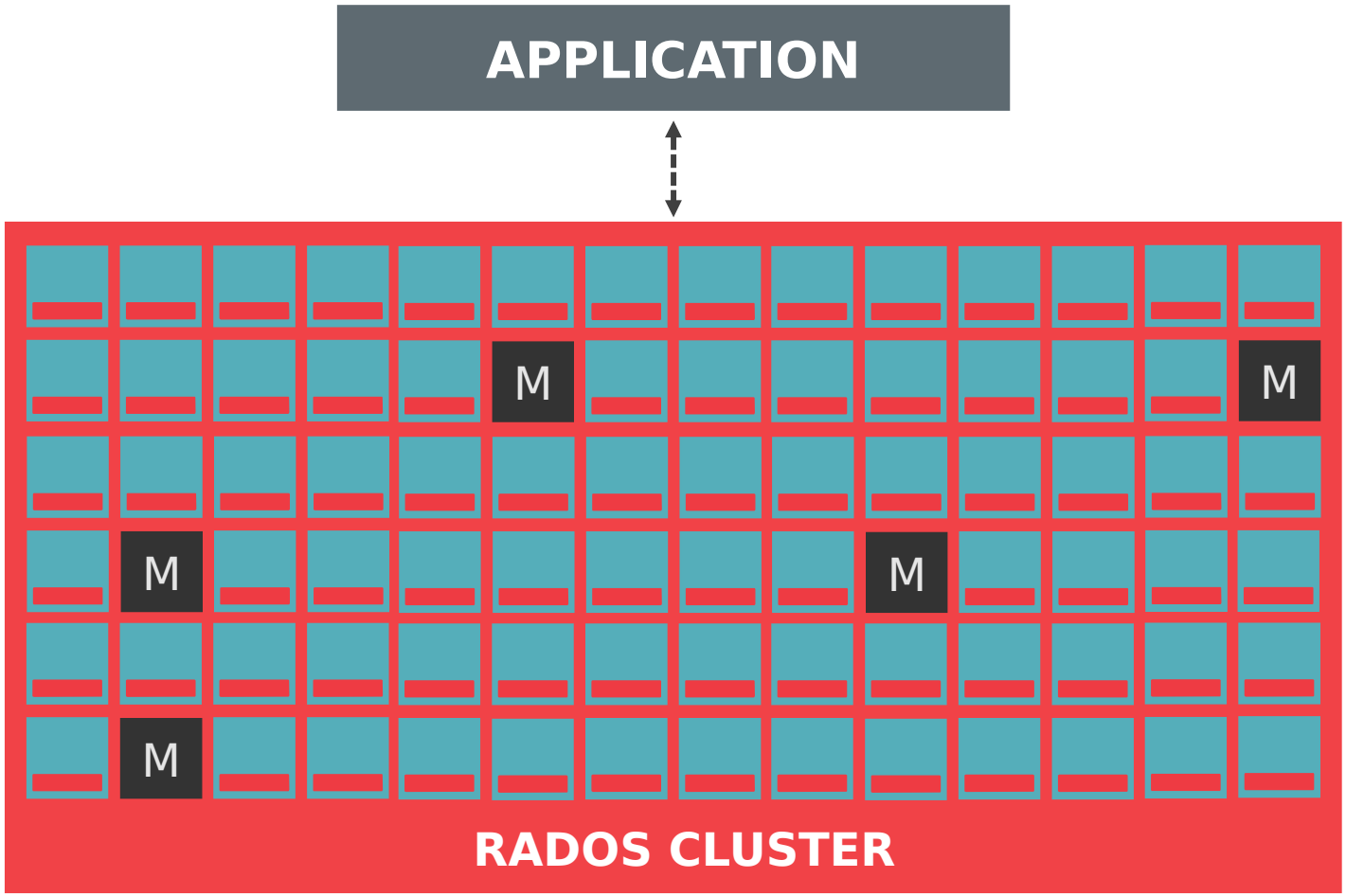
LIBRADOS
A library allowing apps to directly access RADOS (C, C++, Java, Python, Ruby, PHP)

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

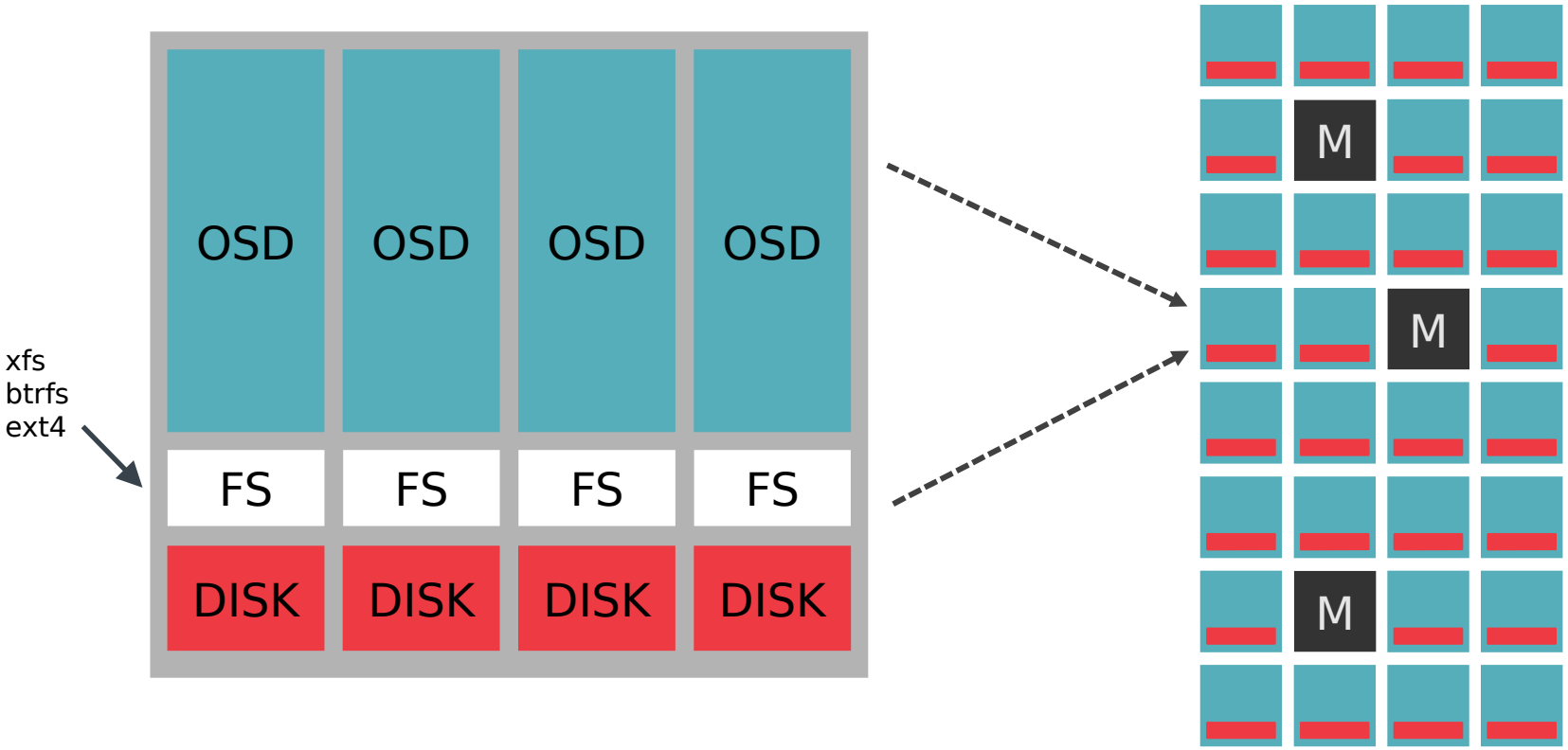
RADOS

- 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

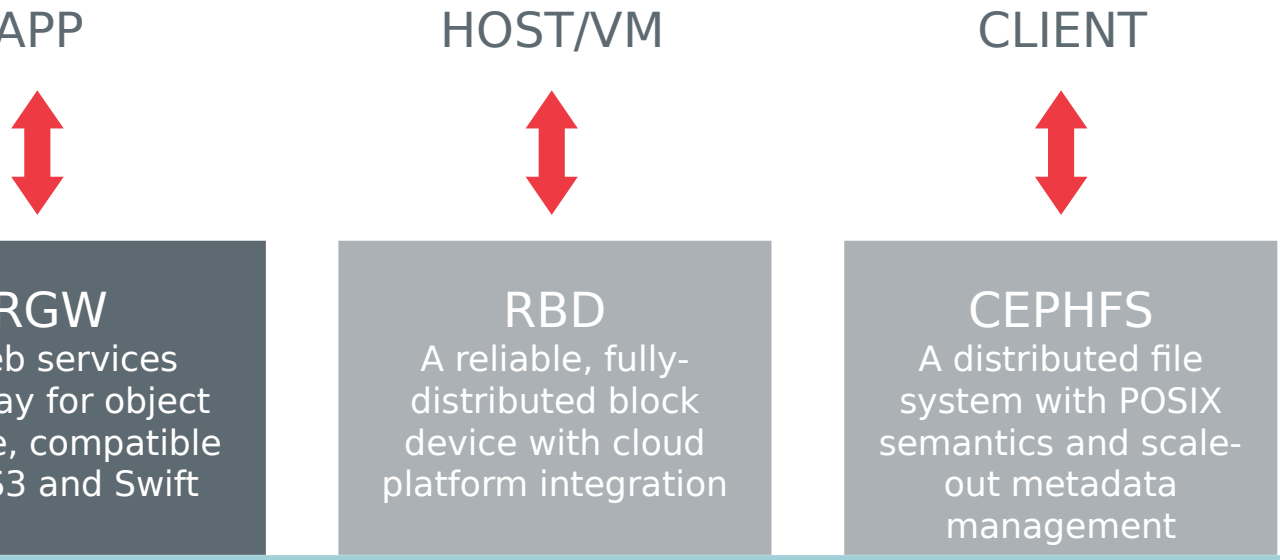
RADOS CLUSTER



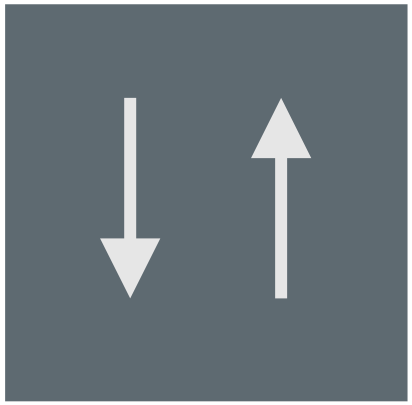
OBJECT STORAGE DAEMONS



ARCHITECTURAL COMPONENTS



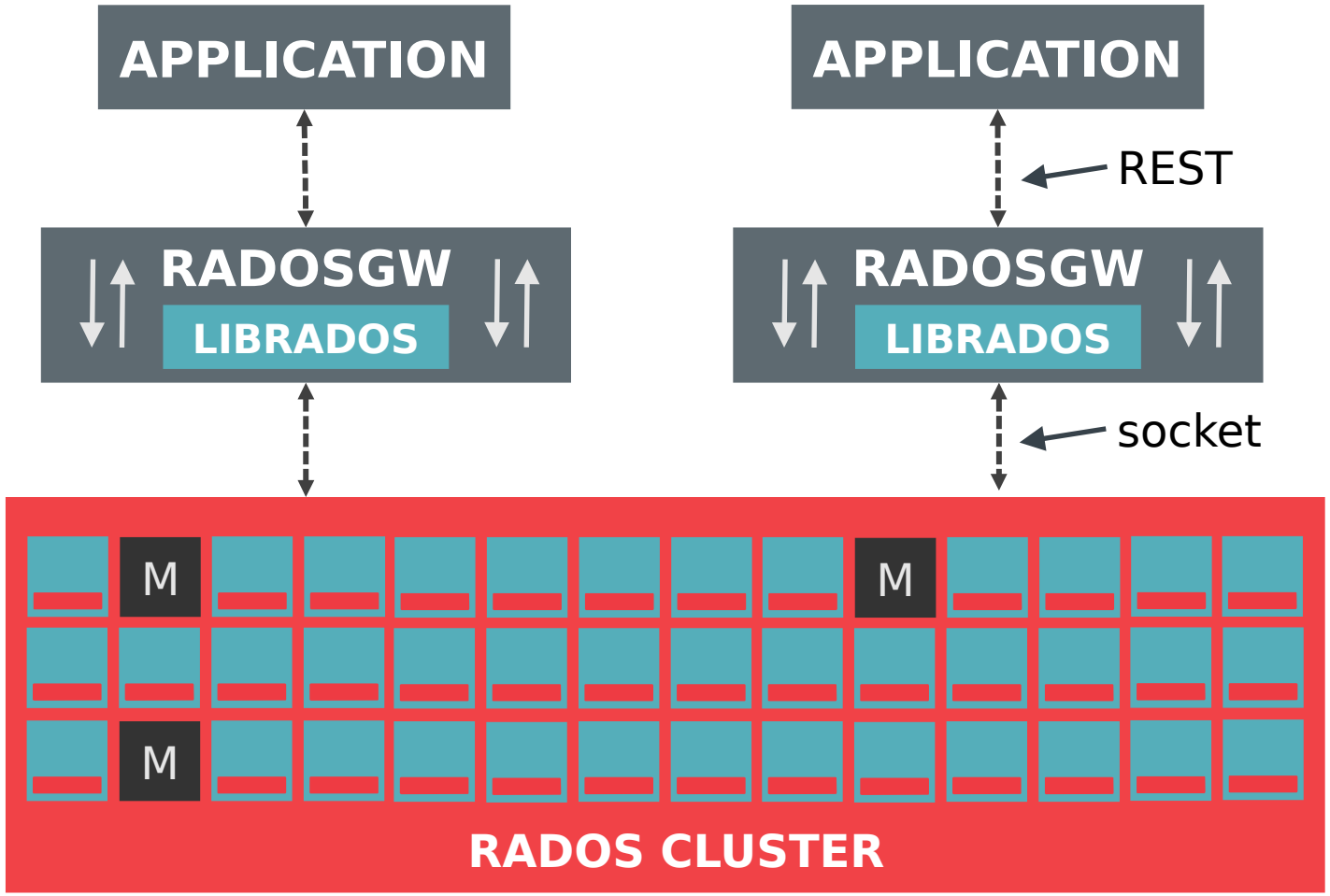
RADOSGW MAKES RADOS WEBBY



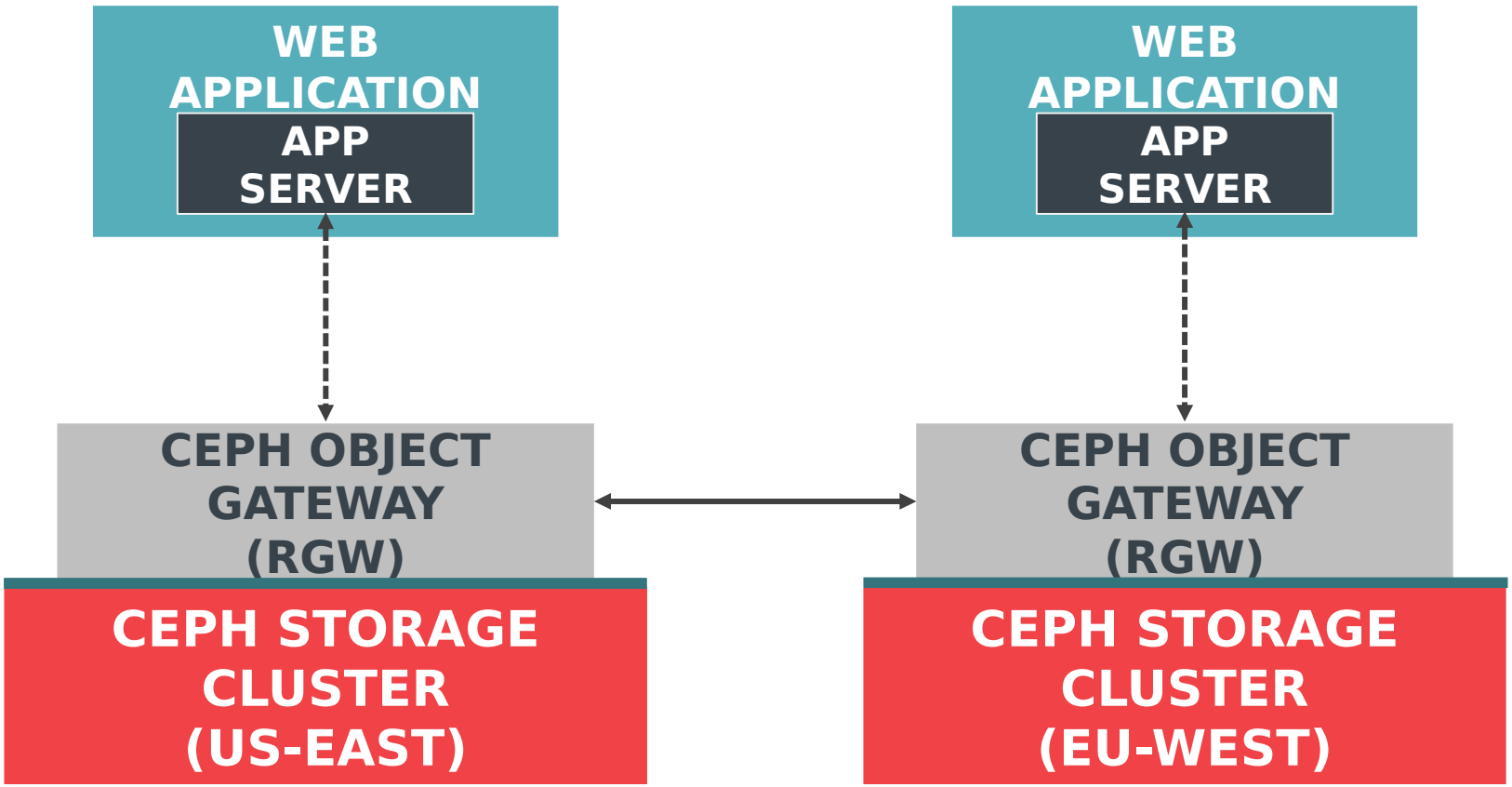
RADOSGW:

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

THE RADOS GATEWAY



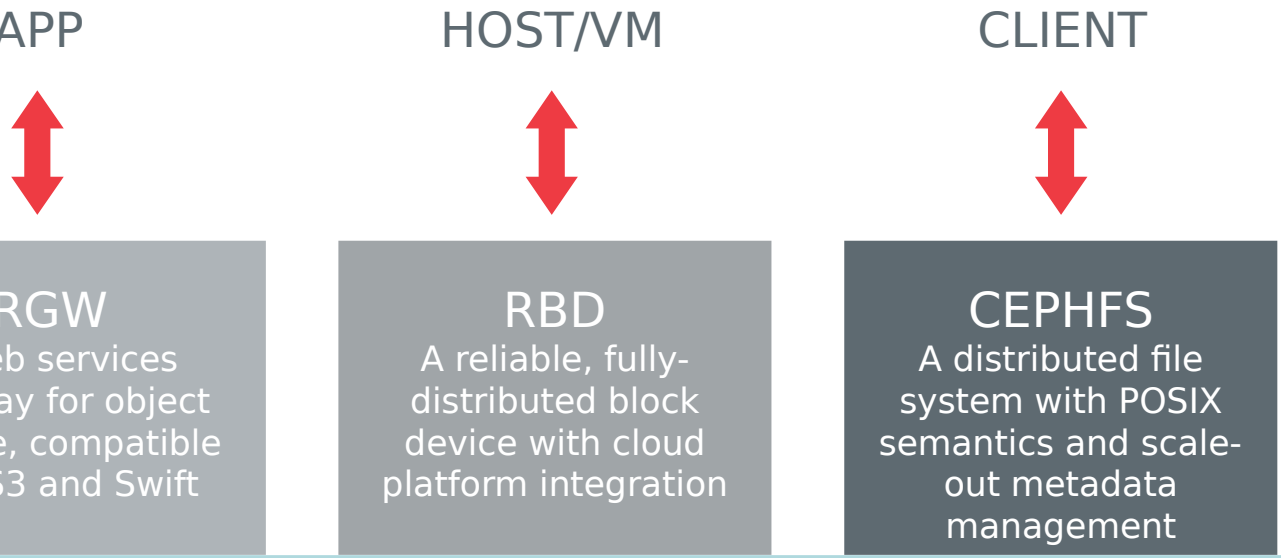
MULTI-SITE OBJECT STORAGE



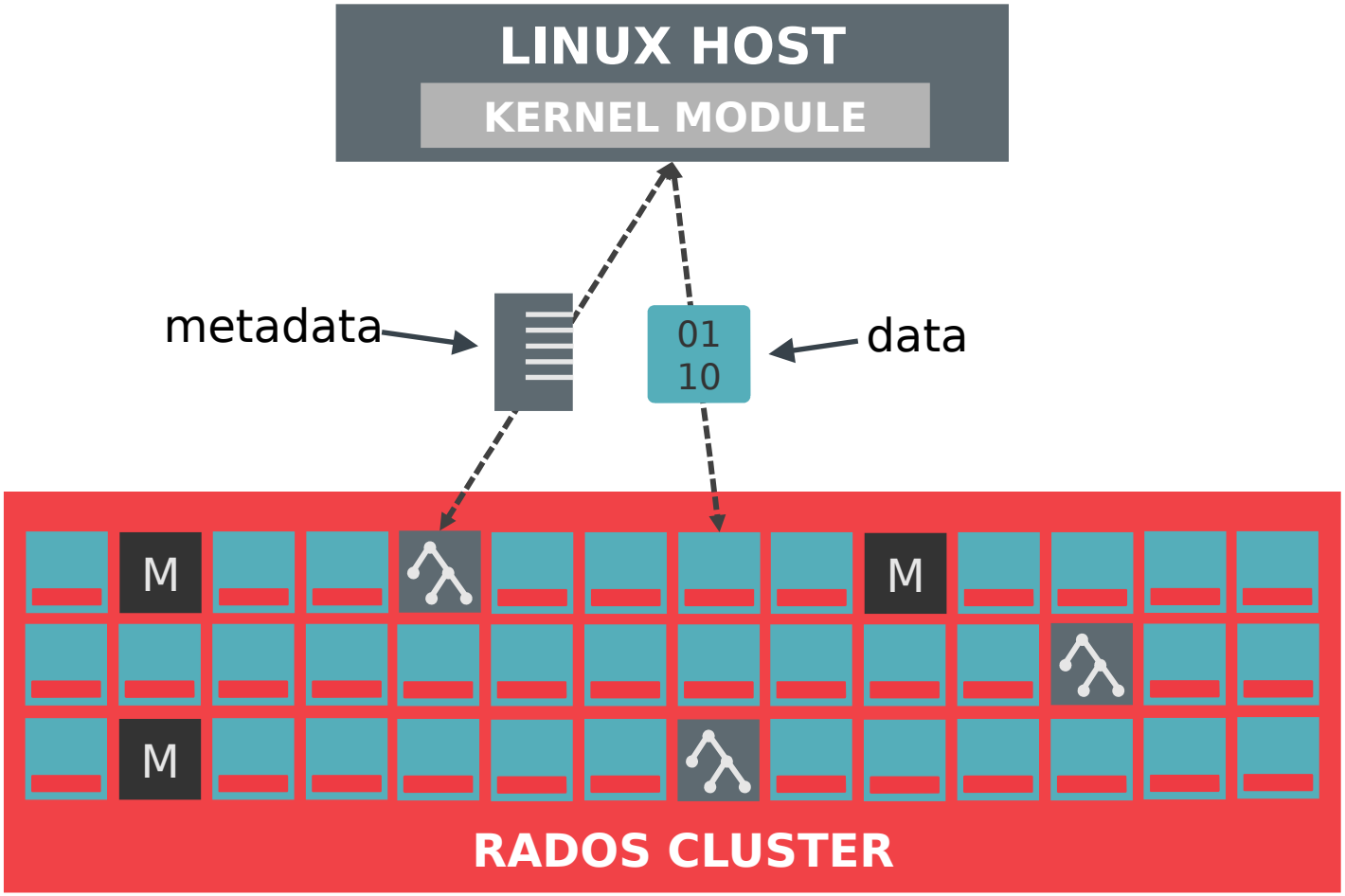
FEDERATED RGW

- 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

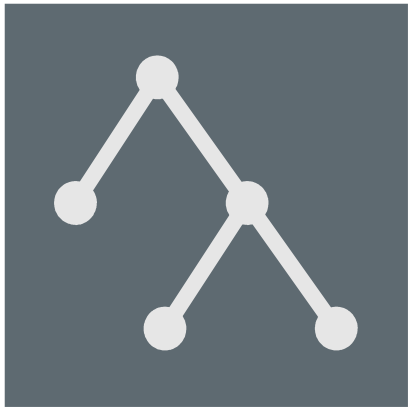
ARCHITECTURAL COMPONENTS



SEPARATE METADATA SERVER



SCALABLE METADATA SERVERS

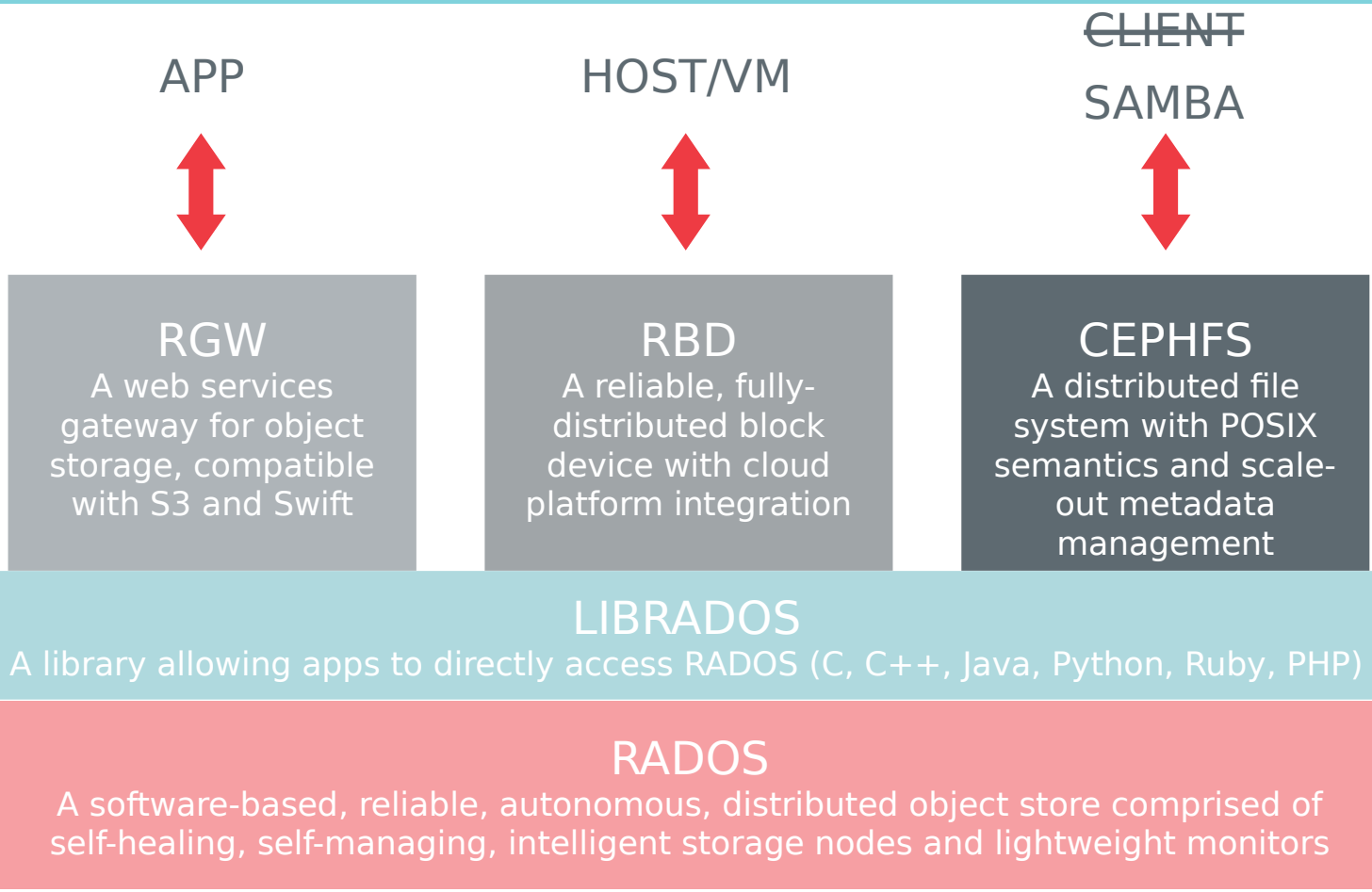


METADATA SERVER

- 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

SAMBA - TODAY

ARCHITECTURAL COMPONENTS



SAMBA INTEGRATION

- vfs_ceph
  - Since 2013.
  - Used as the outline for vfs_glusterfs
  - Been in testing in teuthology for a while now.
    - But not clustered :(.

- ACL Integration?
  - Patchset from Zheng Yan, still needs more work.
  - 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
  - Ceph team recommends ceph_fuse for now.
  - That's what the demo uses...

DEMO

FUTURE DIRECTIONS

- CTDB “fcntl lock” dependency removal.
  - etcd
    - Battle tested.
    - Push other config info into etcd?
      - nodes
      - public_addresses
    - I've already started on this.
      - Expect more info at SDC!
  - Zookeeper much the same as etcd.
    - Not working on it now.
- S3 style object stores.

FUTURE DIRECTIONS

- RGW
  - Export object data as files.
  - Export files as object data?
    - Not today in ceph.
  - Integrate where?
    - S3
    - RADOS

- RBD
  - With SMB Direct, who knows?

QUESTIONS?

THANK YOU!

Ira Cooper
SAMBA TEAM


ira@wakeful.net