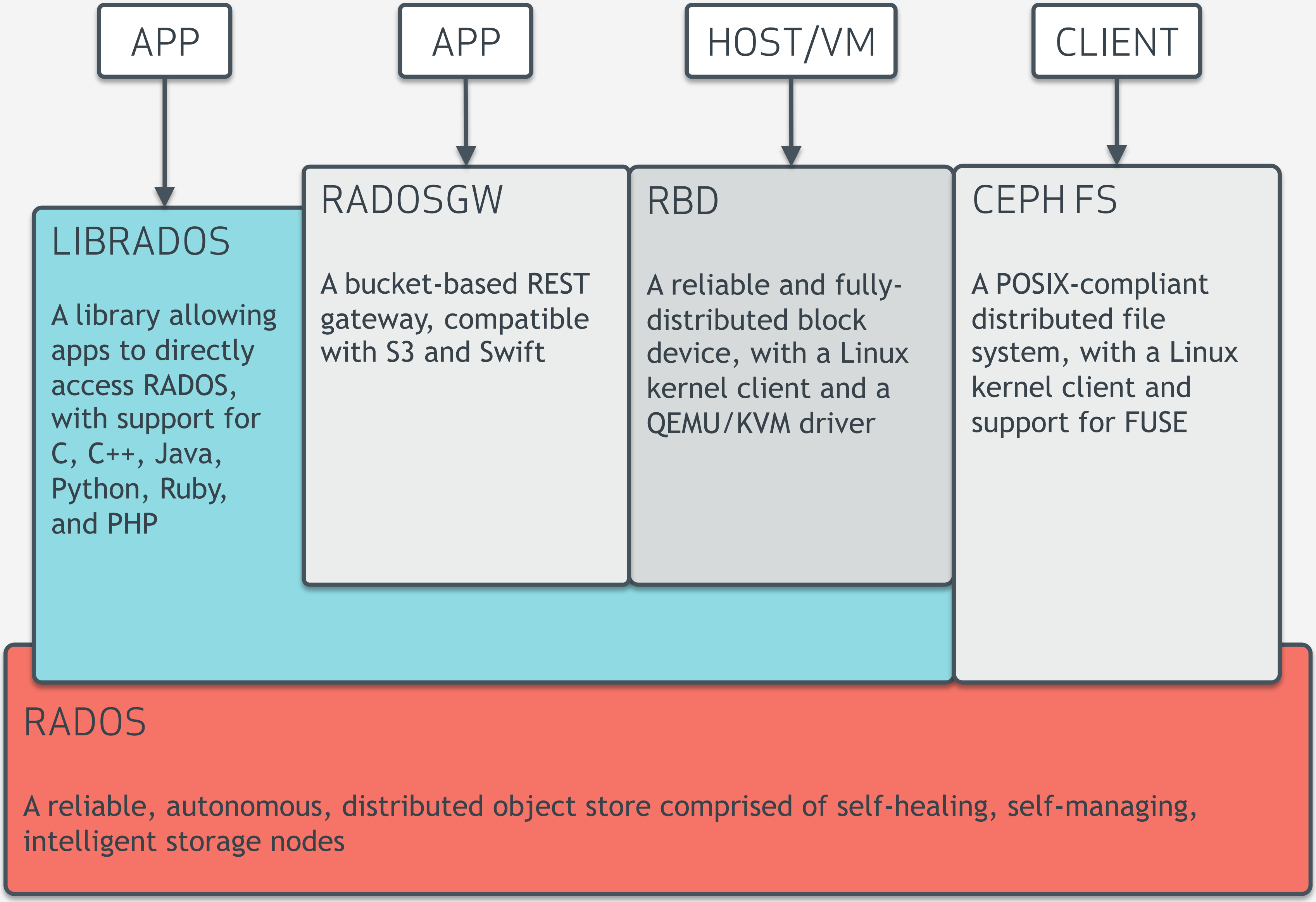
inktank

New Features for Ceph with Cinder and Beyond

ceph

Why Ceph?

- Low cost
- Flexible
- Scalable
- Open source



APP

APP

HOST/VM

CLIENT

LIBRADOS

A library allowing apps to directly access RADOS, with support for C, C++, Java, Python, Ruby, and PHP

RADOSGW

A bucket-based REST gateway, compatible with S3 and Swift

RBD

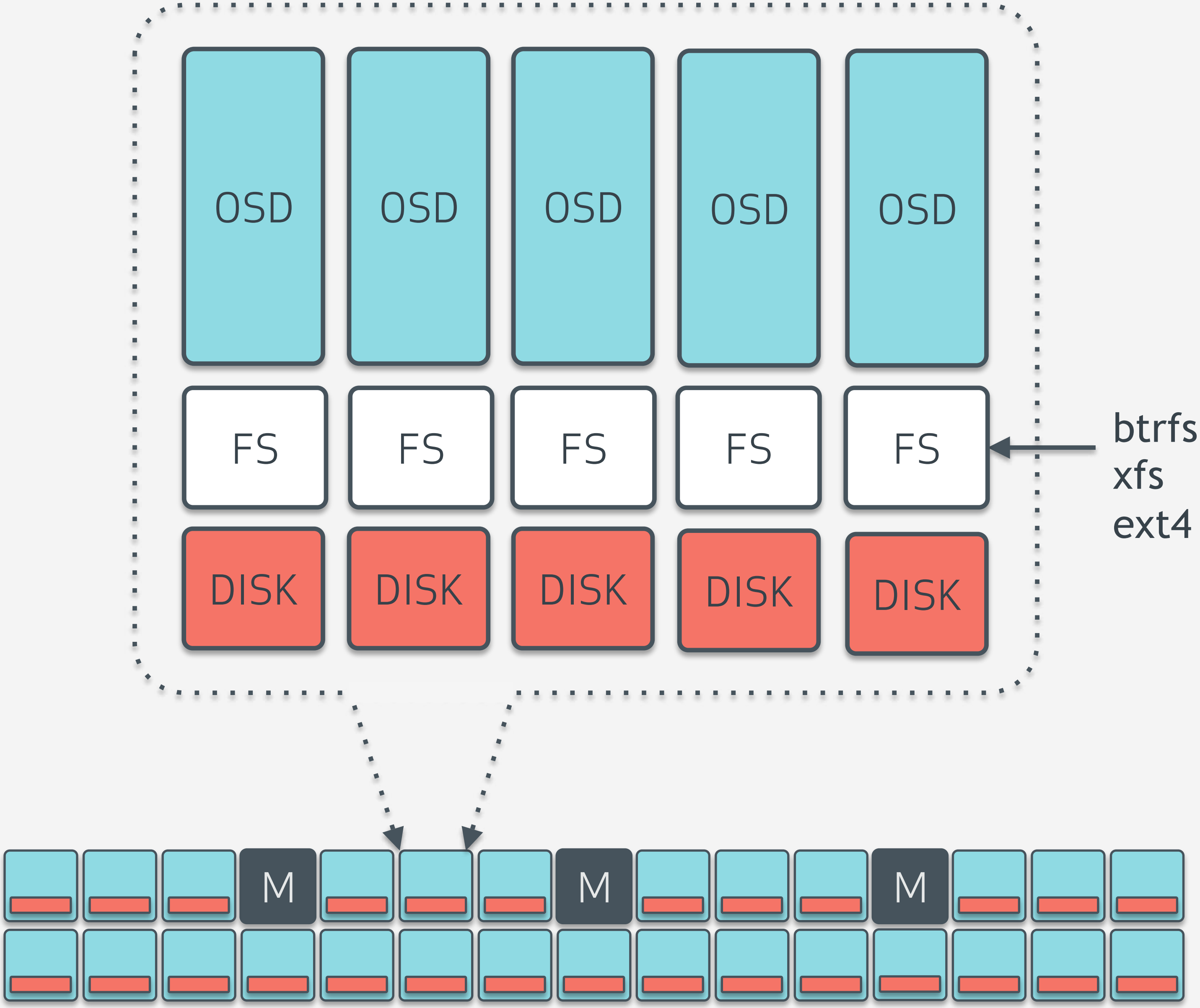
A reliable and fully-distributed block device, with a Linux kernel client and a QEMU/KVM driver

CEPH FS

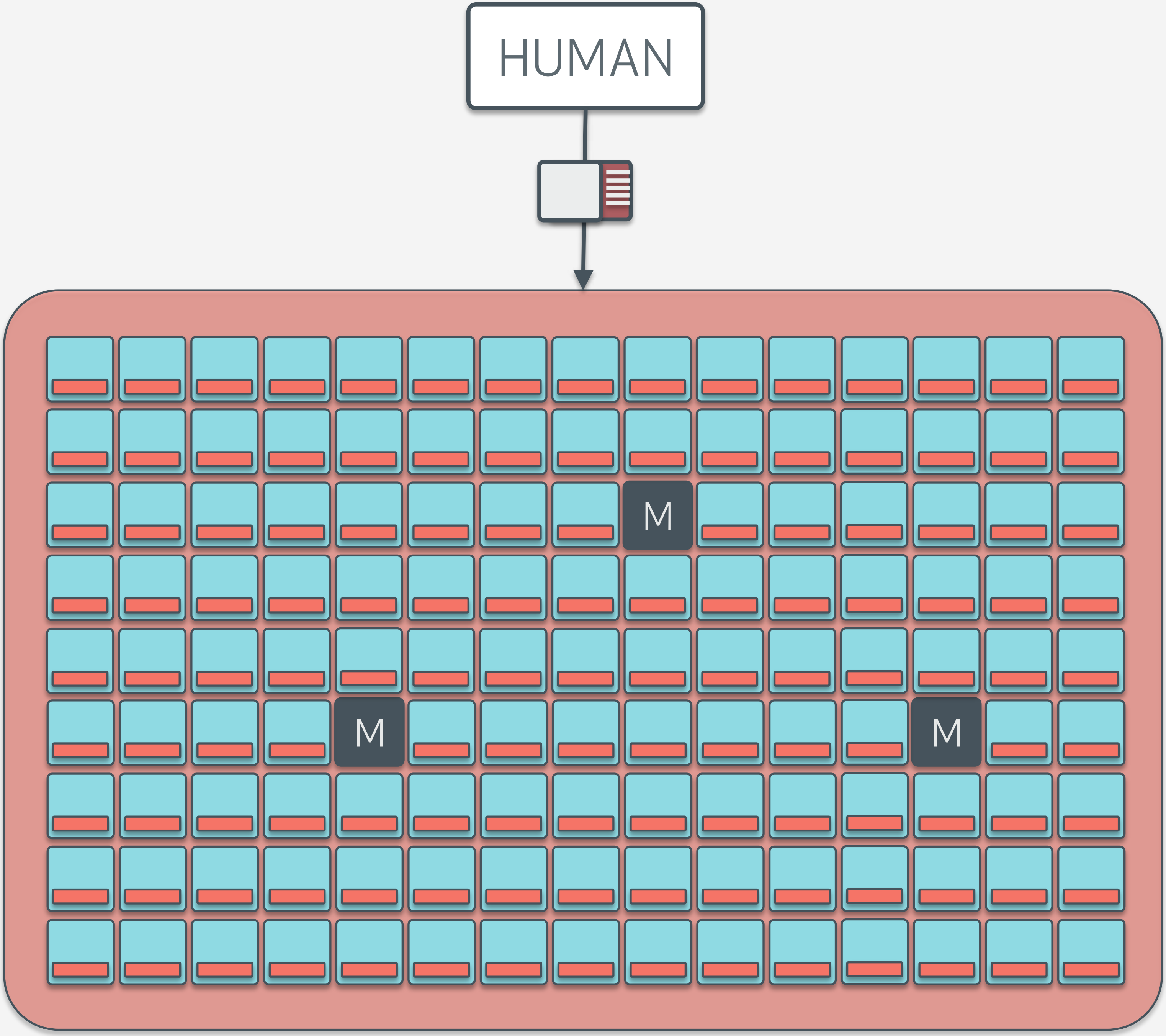
A POSIX-compliant distributed file system, with a Linux kernel client and support for FUSE

RADOS

A reliable, autonomous, distributed object store comprised of self-healing, self-managing, intelligent storage nodes



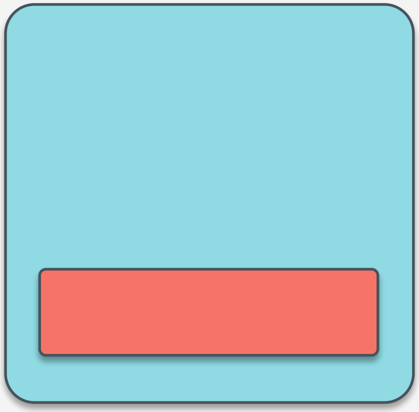
btrfs
xfs
ext4



HUMAN
M
M
M

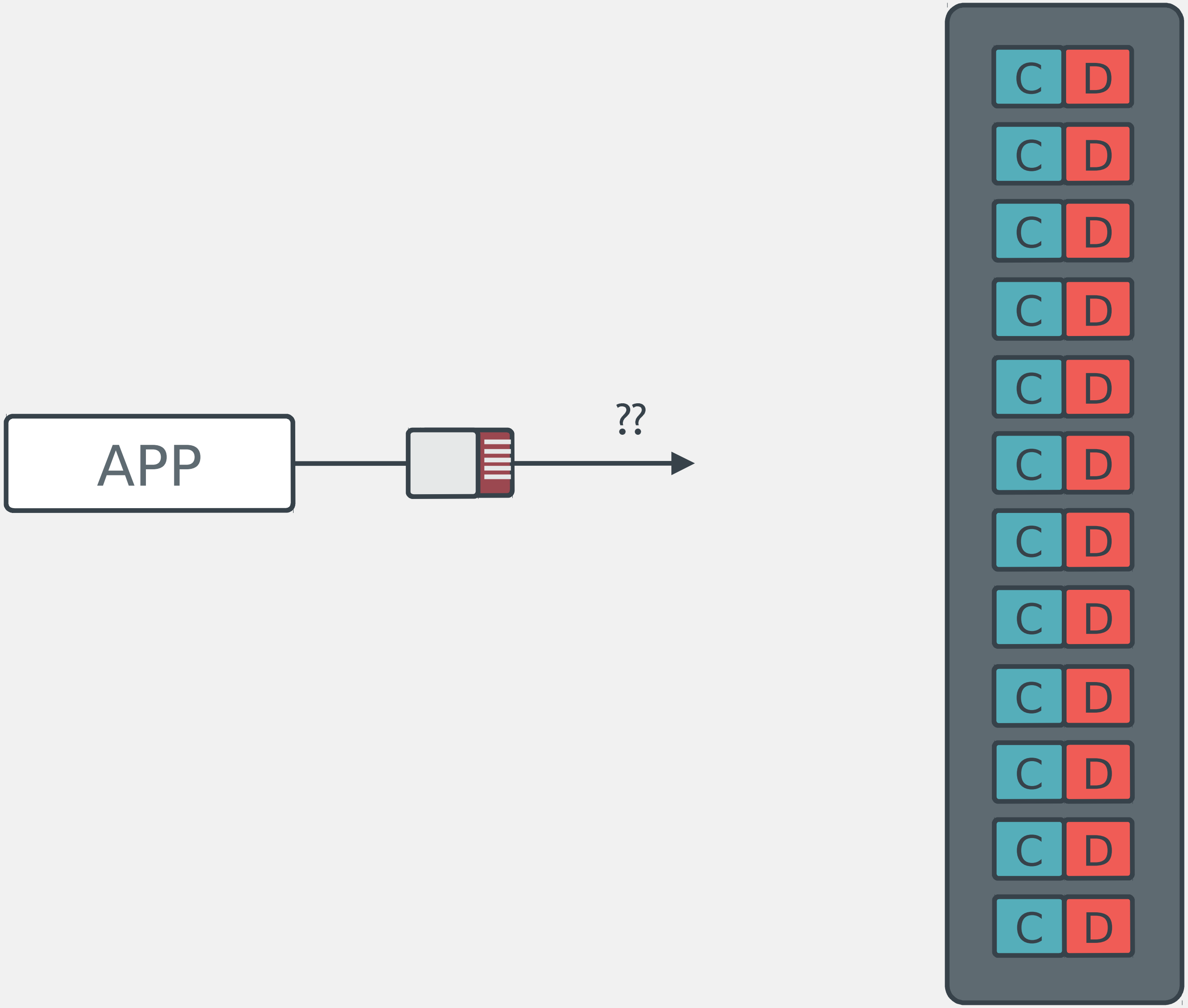
Monitors:

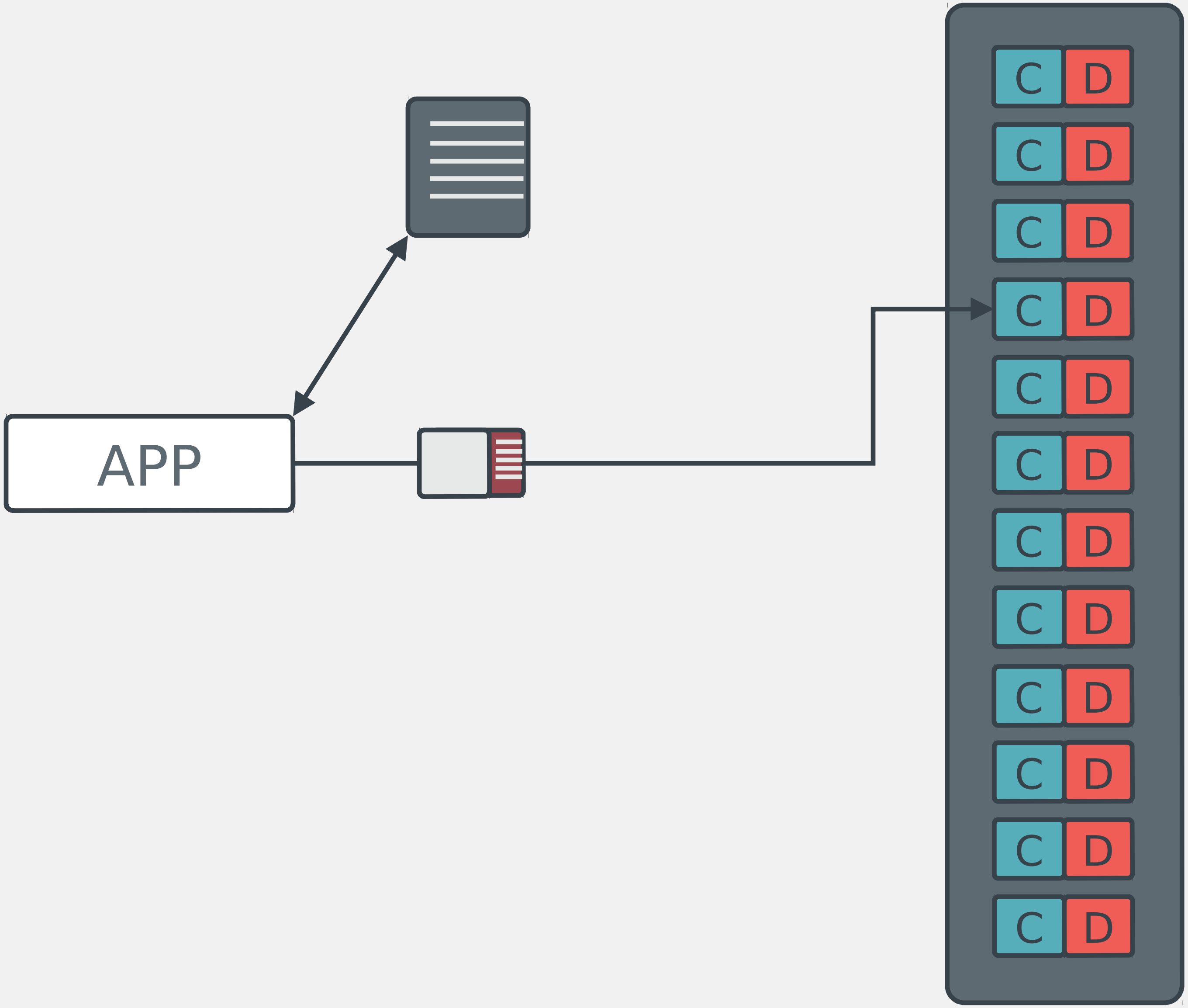
- Maintain cluster map
- Provide consensus for distributed decision-making
- Must have an odd number
- These do **not** serve stored objects to clients

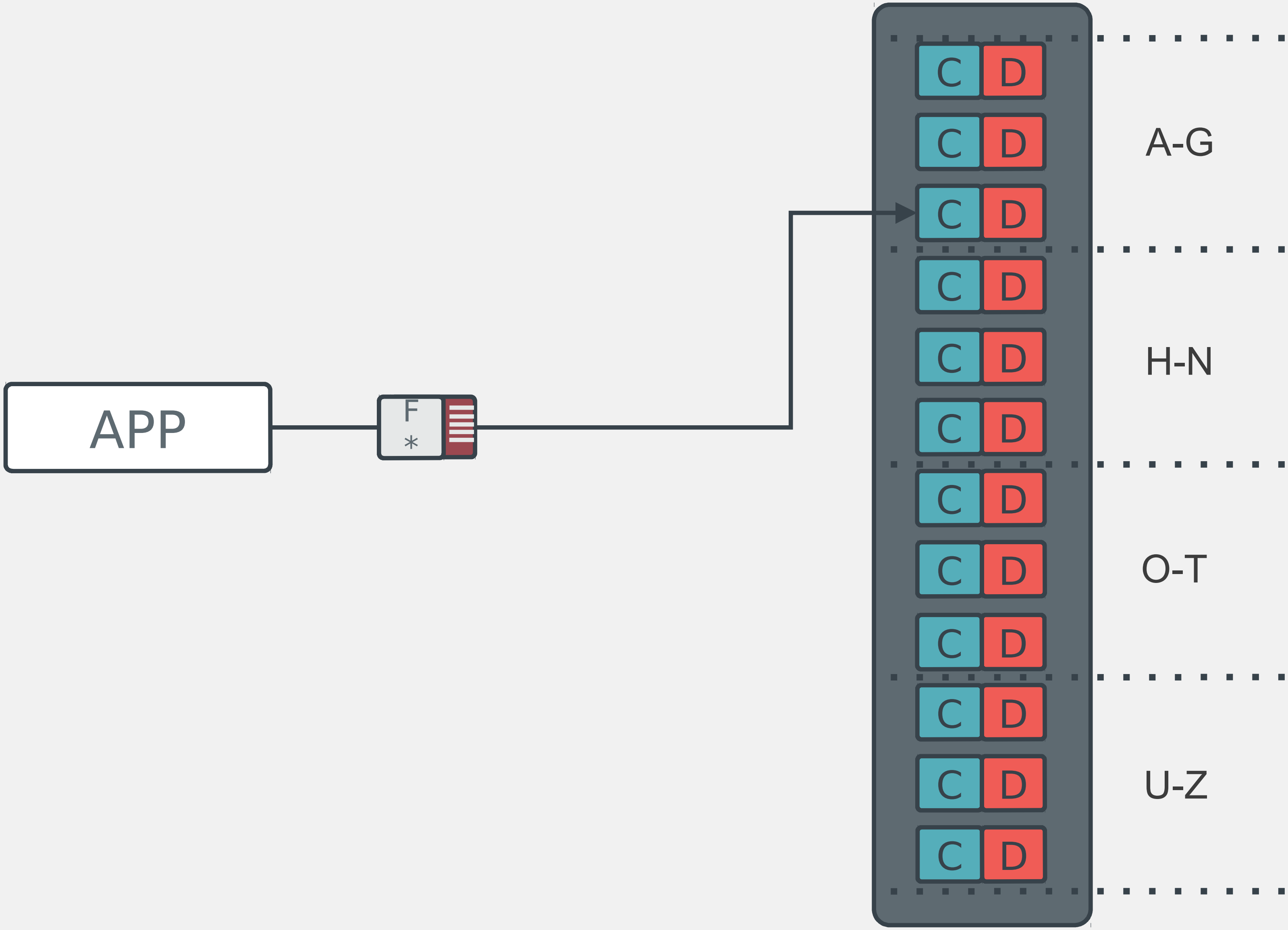


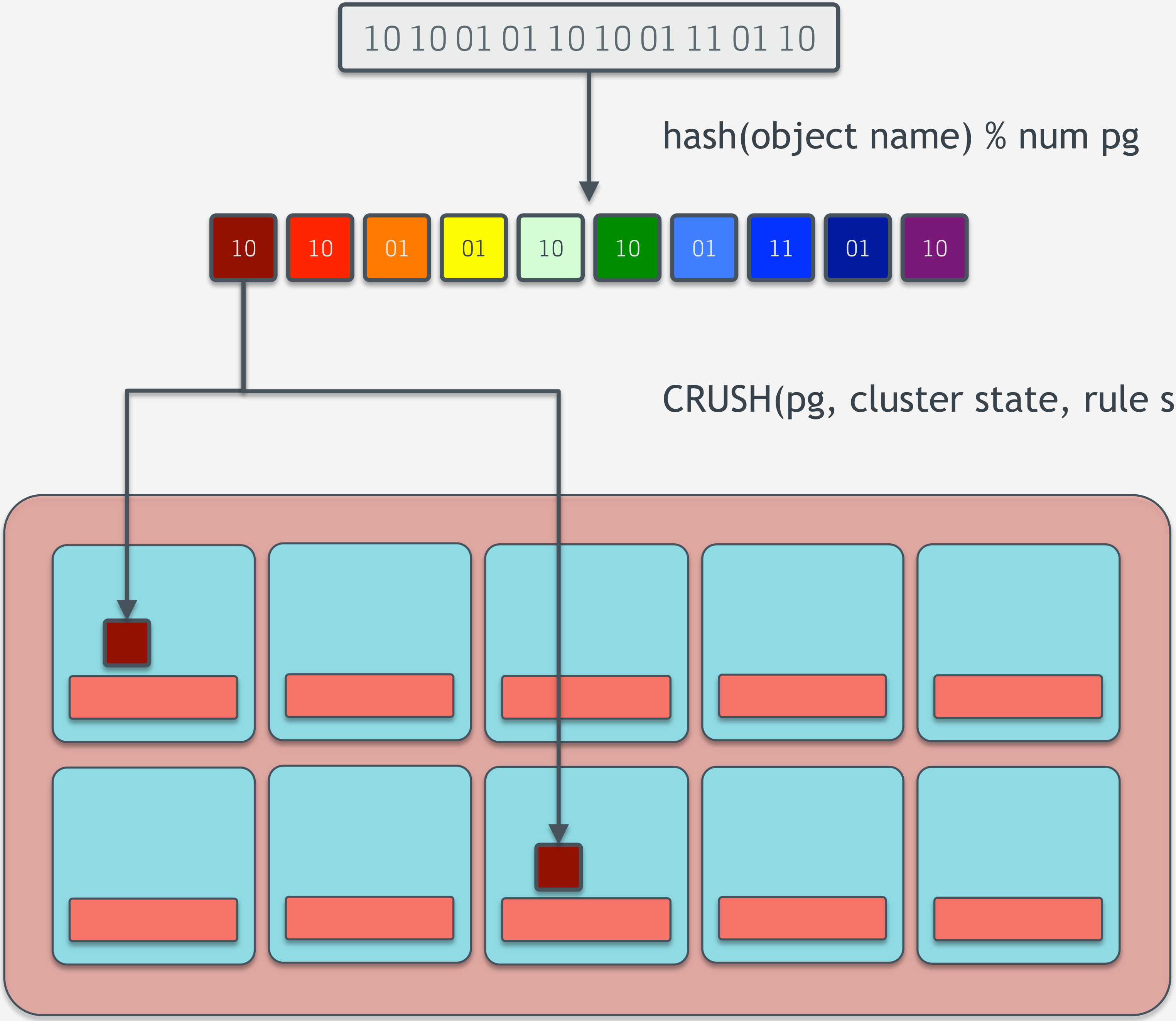
OSDs:

- One per disk (recommended)
- At least three in a cluster
- Serve stored objects to clients
- Intelligently peer to perform replication tasks
- Supports object classes


APP
??
C D
C D
C D
C D
C D
C D
C D
C D
C D
C D
C D
C D

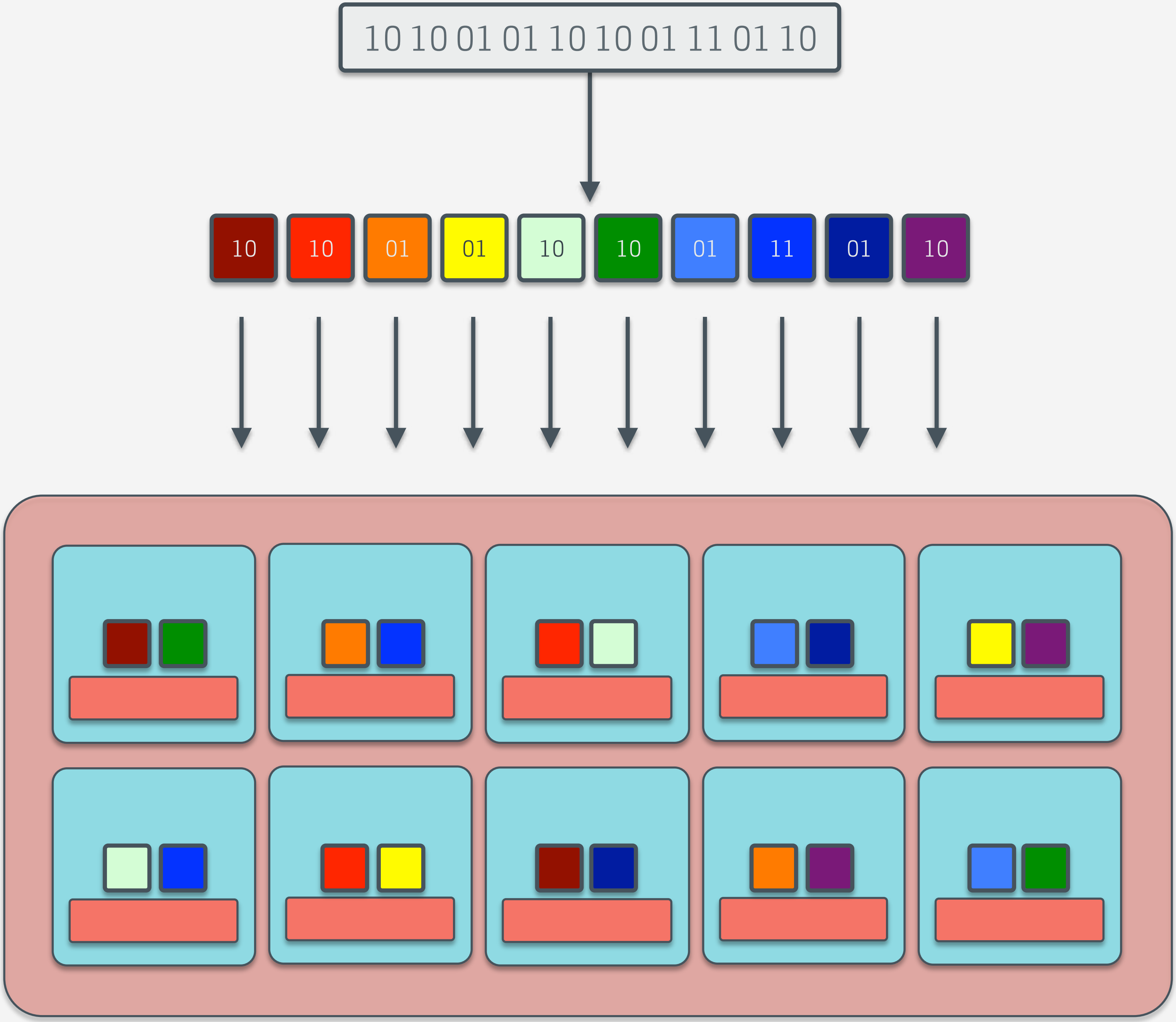

APP
C D
C D
C D
C D
C D
C D
C D
C D
C D
C D
C D
C D

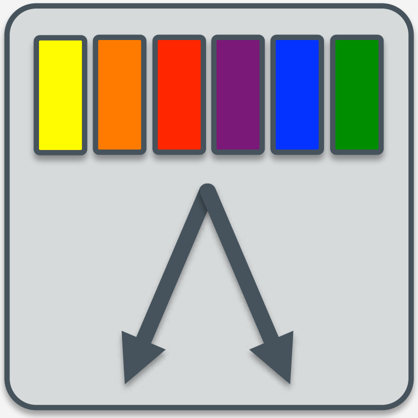

APP
F
*
C
D
A-G
H-N
O-T
U-Z



hash(object name) % num pg

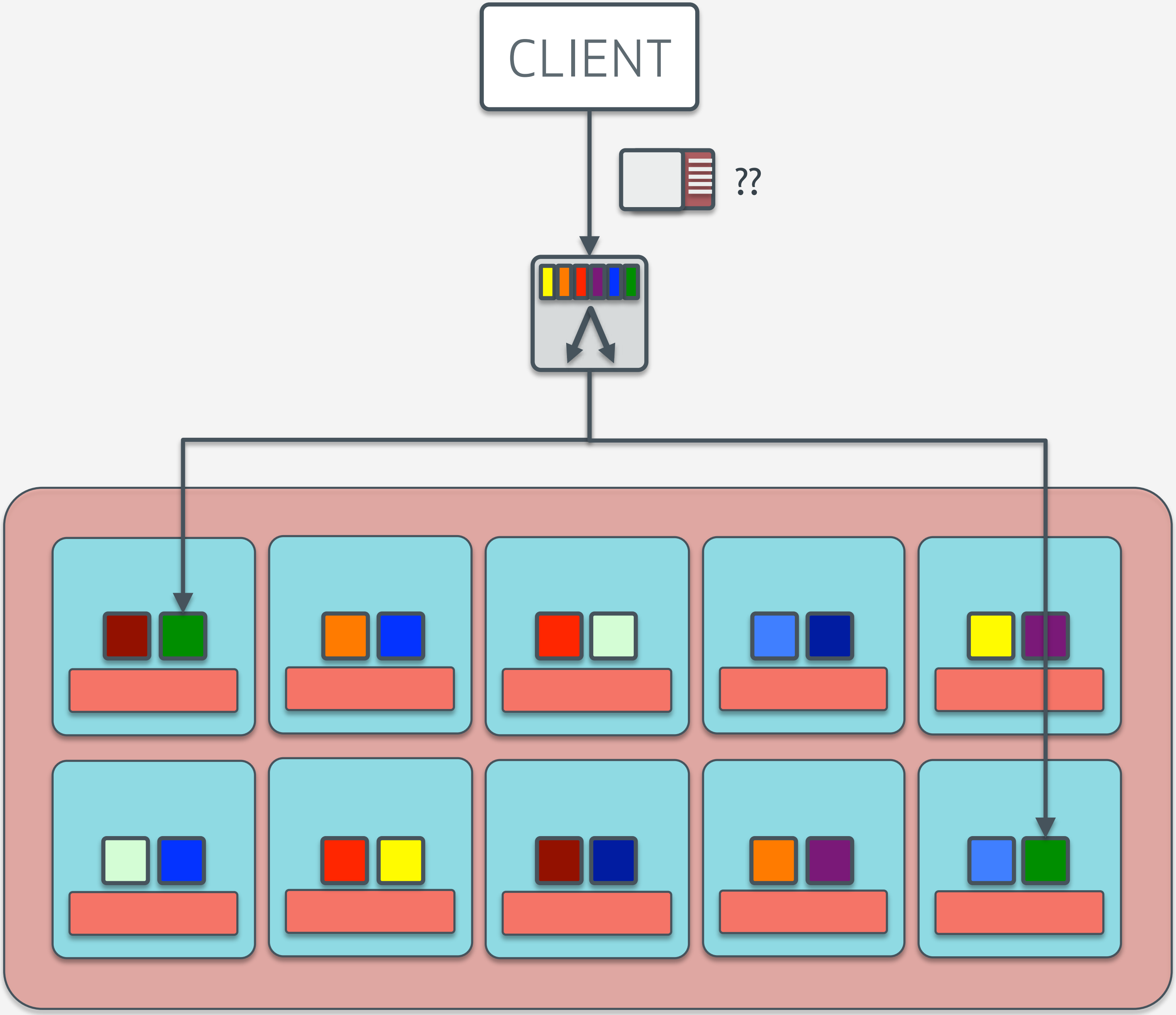
CRUSH(pg, cluster state, rule set)

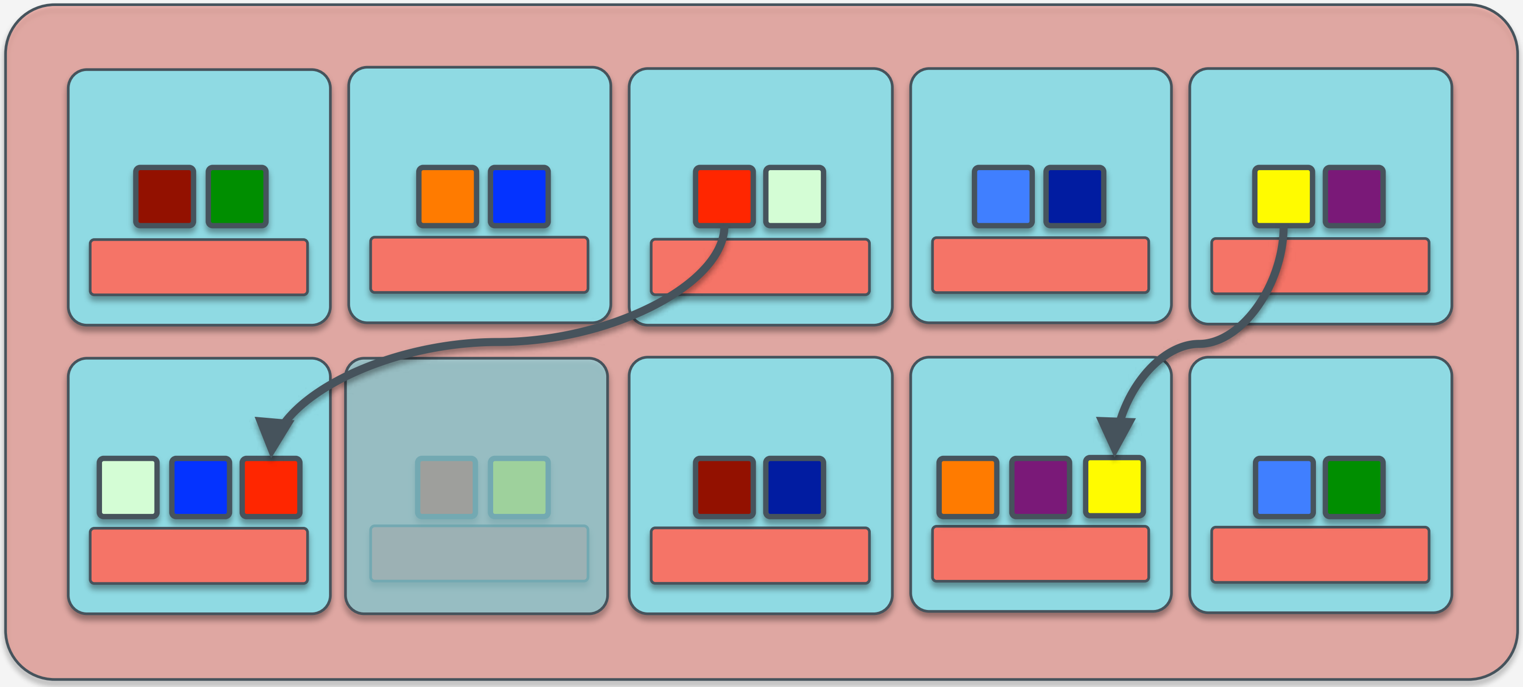

10 10 01 01 10 10 01 11 01 10
10
10
01
01
10
10
01
11
01
10

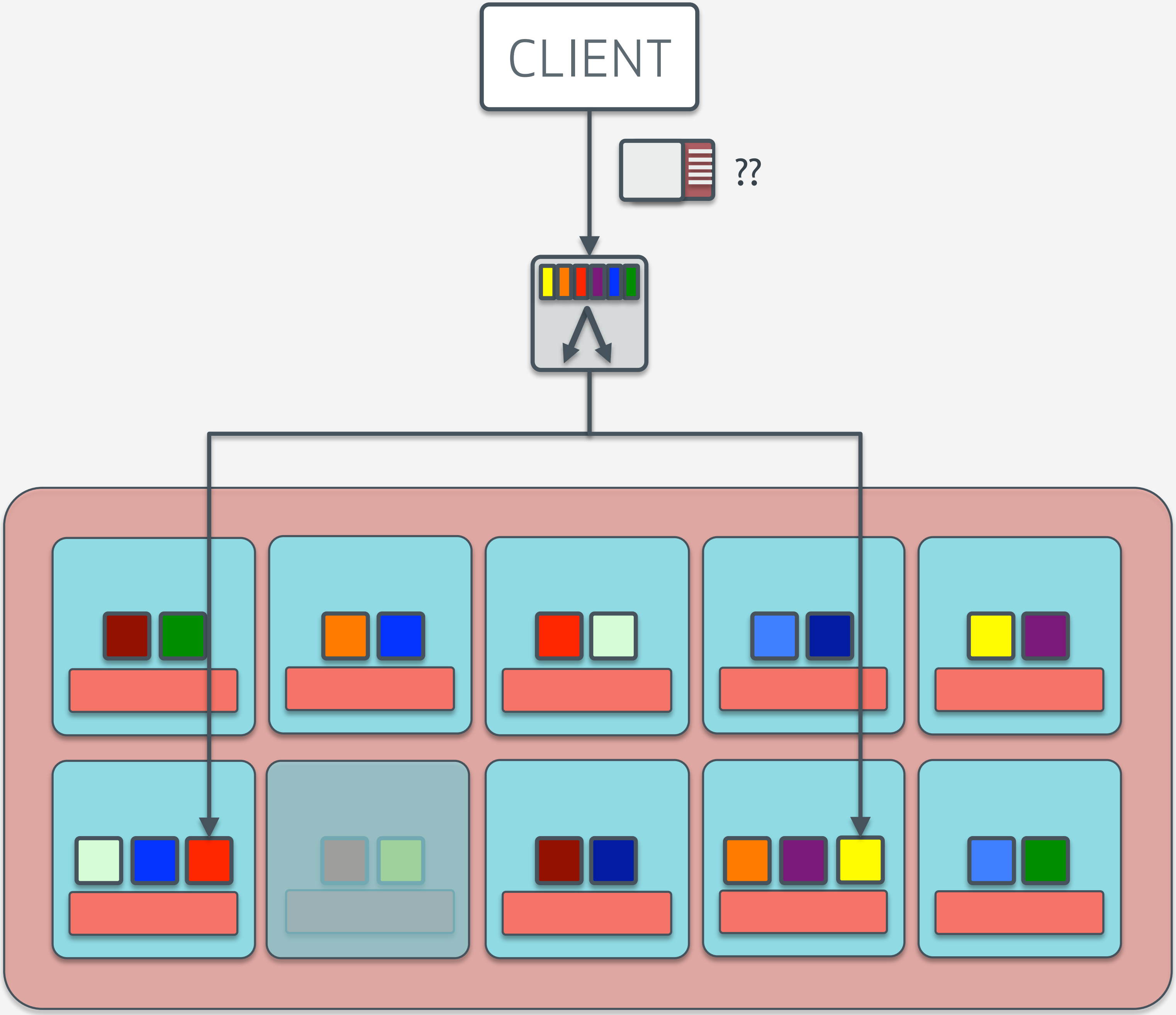


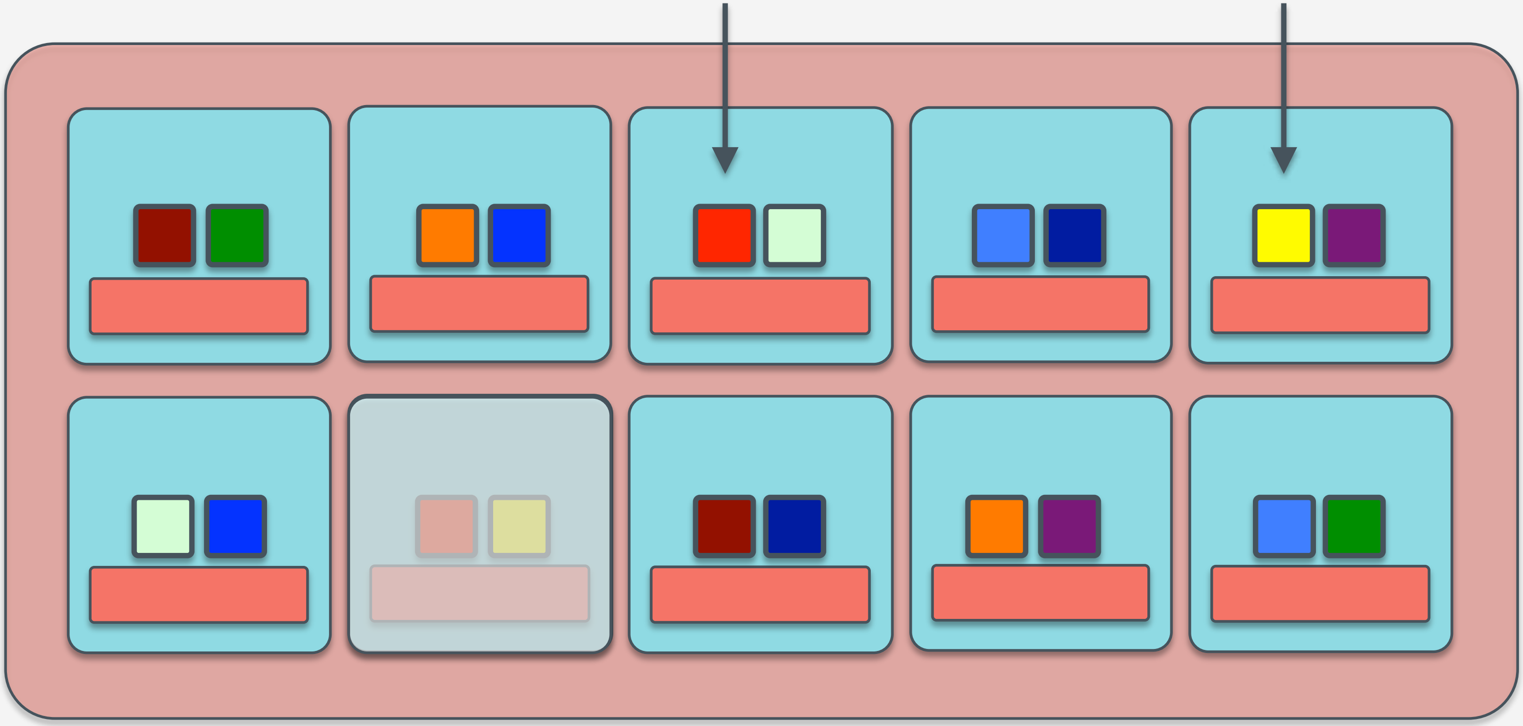
CRUSH

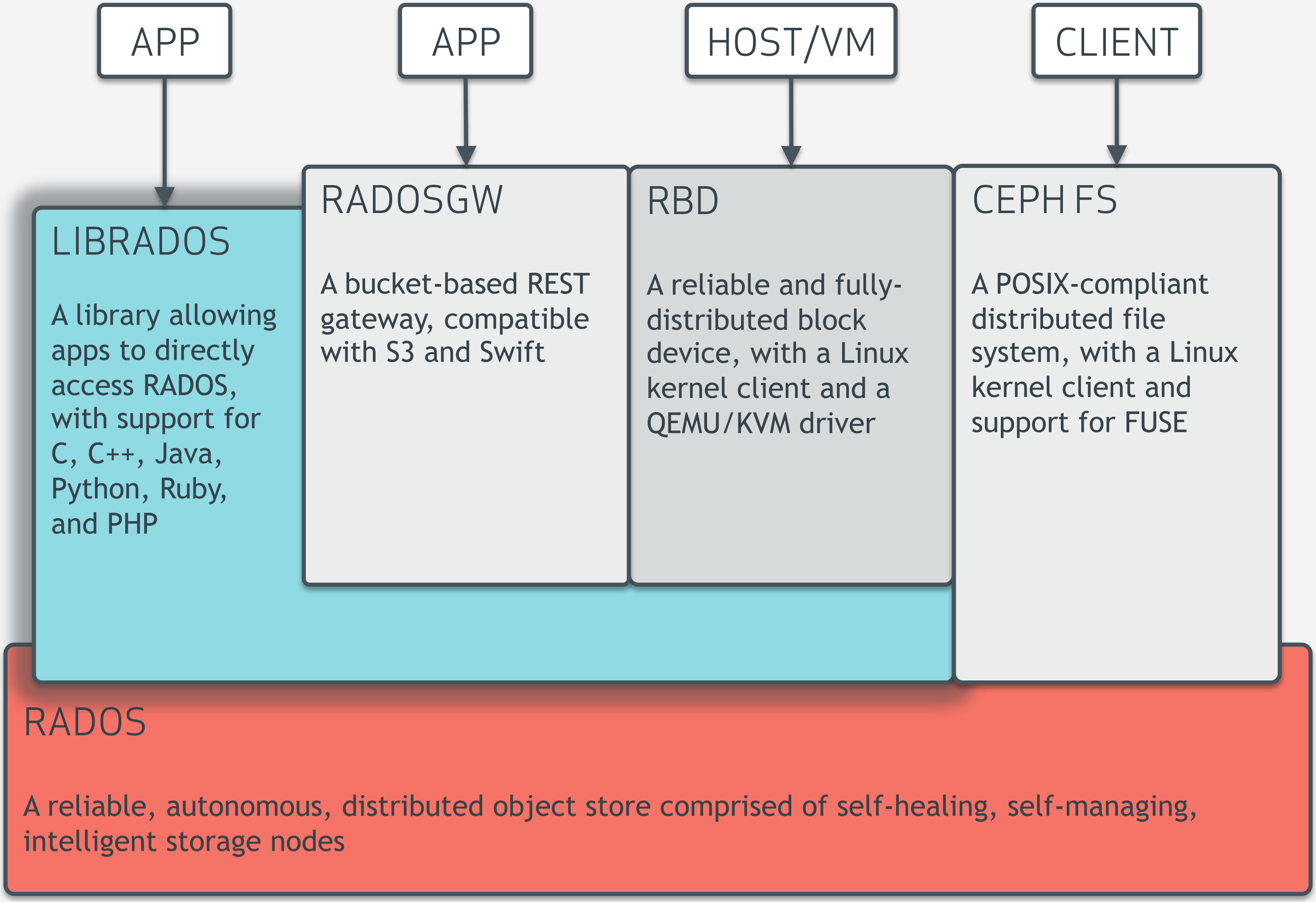
- Pseudo-random placement algorithm
- Ensures even distribution
- Repeatable, deterministic
- Rule-based configuration
  - Replica count
  - Infrastructure topology
  - Weighting


CLIENT
??




CLIENT
??





APP

APP

HOST/VM

CLIENT

LIBRADOS

A library allowing apps to directly access RADOS, with support for C, C++, Java, Python, Ruby, and PHP

RADOSGW

A bucket-based REST gateway, compatible with S3 and Swift

RBD

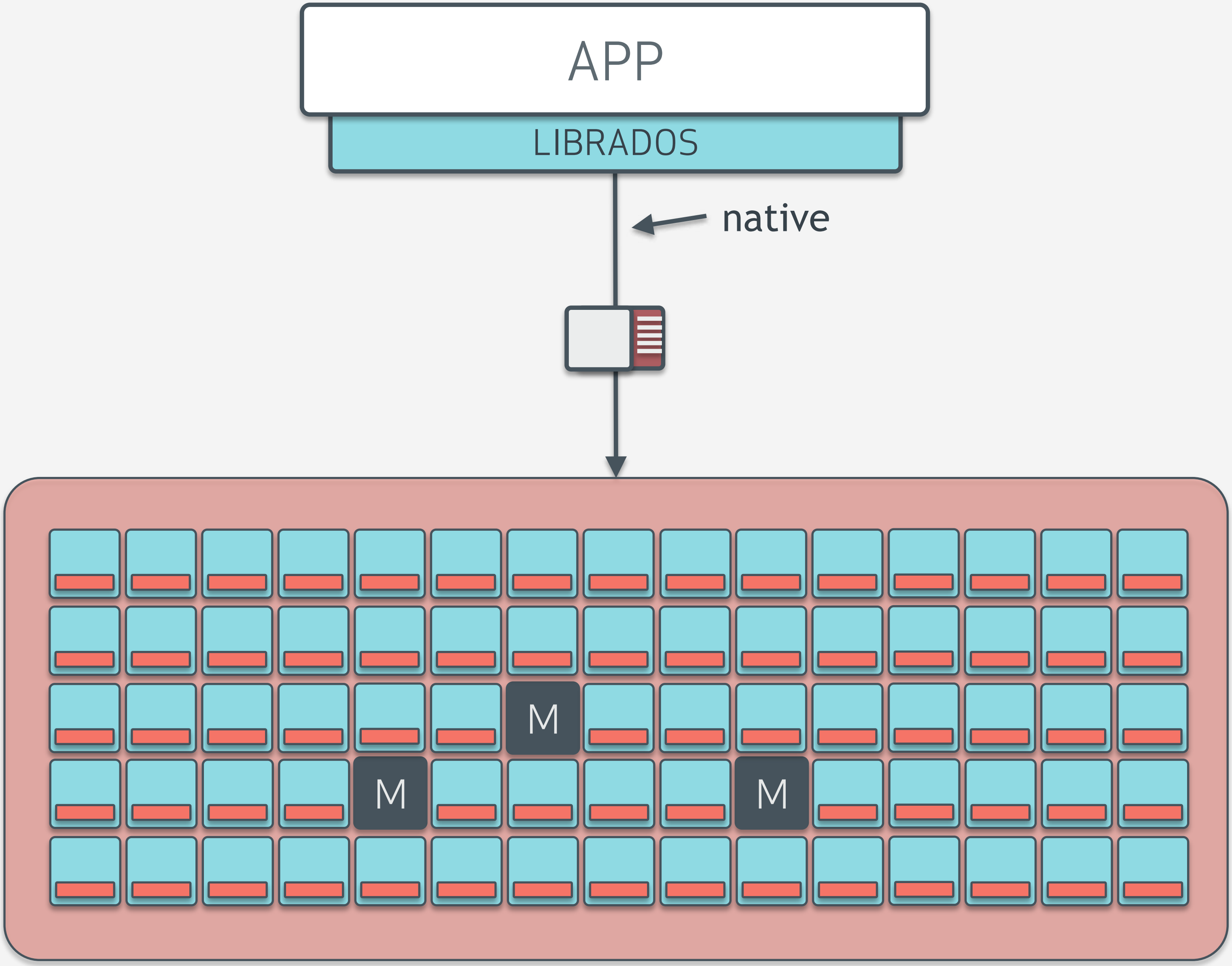
A reliable and fully-distributed block device, with a Linux kernel client and a QEMU/KVM driver

CEPH FS

A POSIX-compliant distributed file system, with a Linux kernel client and support for FUSE

RADOS

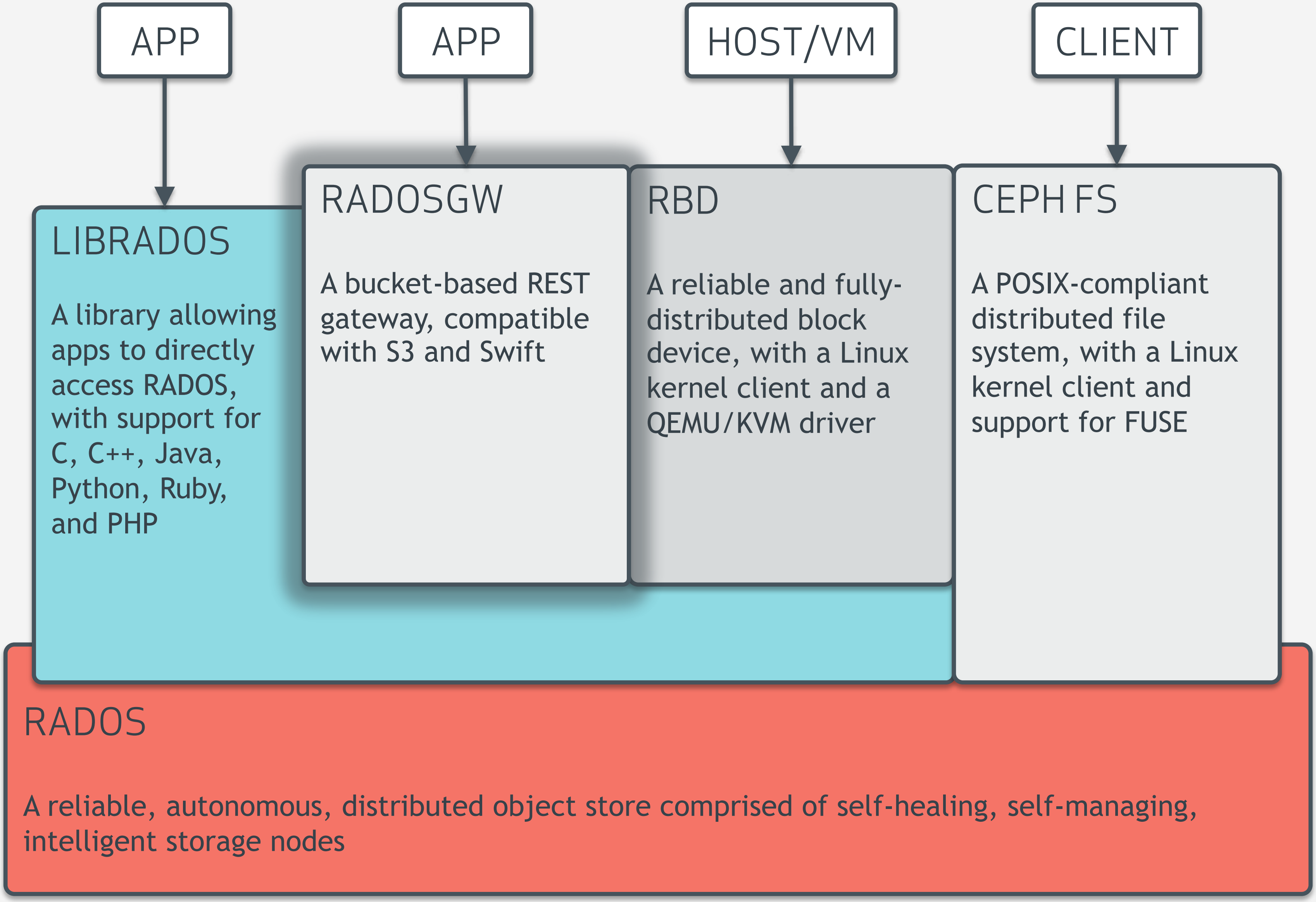
A reliable, autonomous, distributed object store comprised of self-healing, self-managing, intelligent storage nodes


APP
LIBRADOS
native
M
M
M

L

LIBRADOS

- Provides direct access to RADOS for applications
- C, C++, Python, PHP, Java
- No HTTP overhead



APP

APP

HOST/VM

CLIENT

LIBRADOS

A library allowing apps to directly access RADOS, with support for C, C++, Java, Python, Ruby, and PHP

RADOSGW

A bucket-based REST gateway, compatible with S3 and Swift

RBD

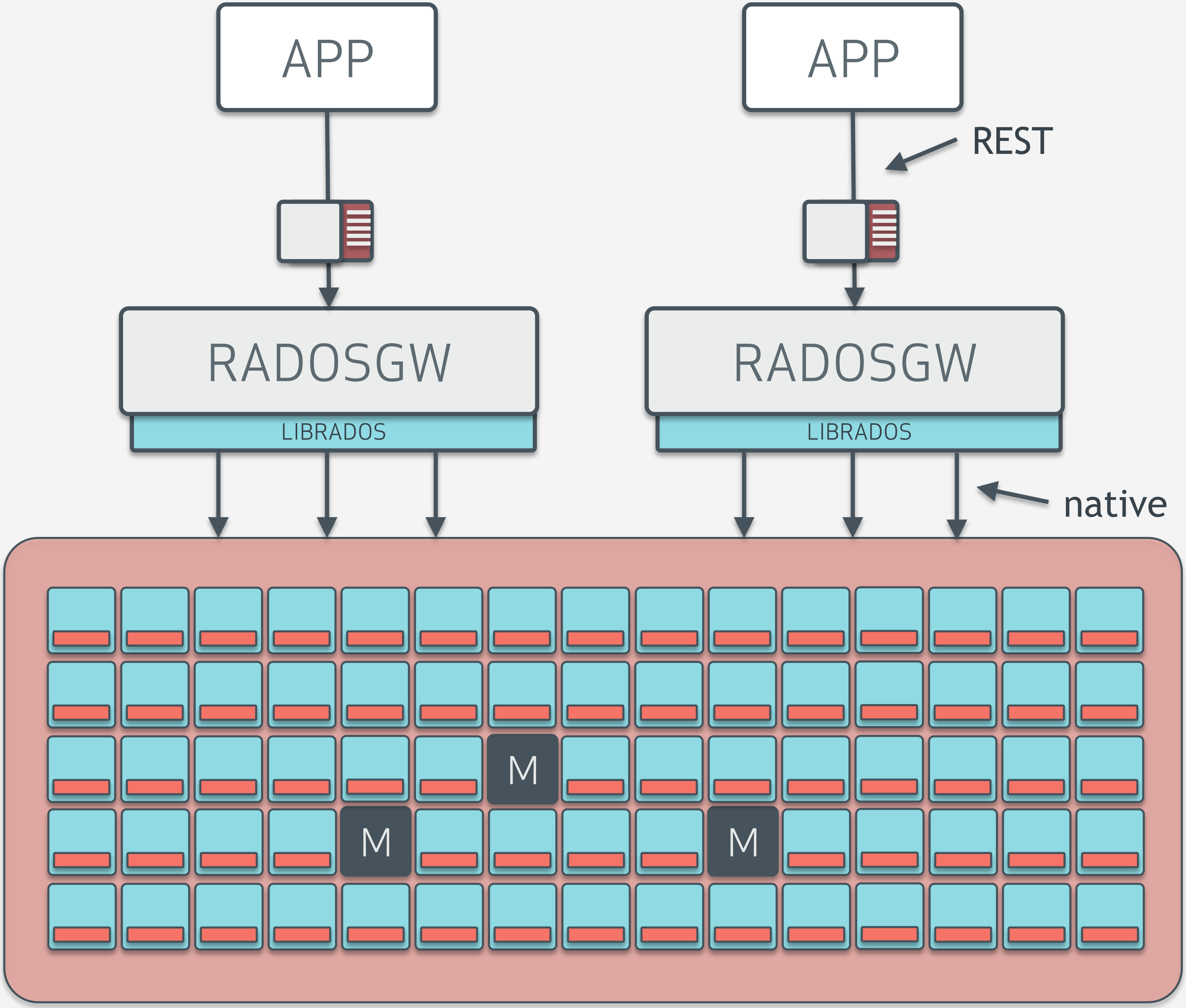
A reliable and fully-distributed block device, with a Linux kernel client and a QEMU/KVM driver

CEPH FS

A POSIX-compliant distributed file system, with a Linux kernel client and support for FUSE

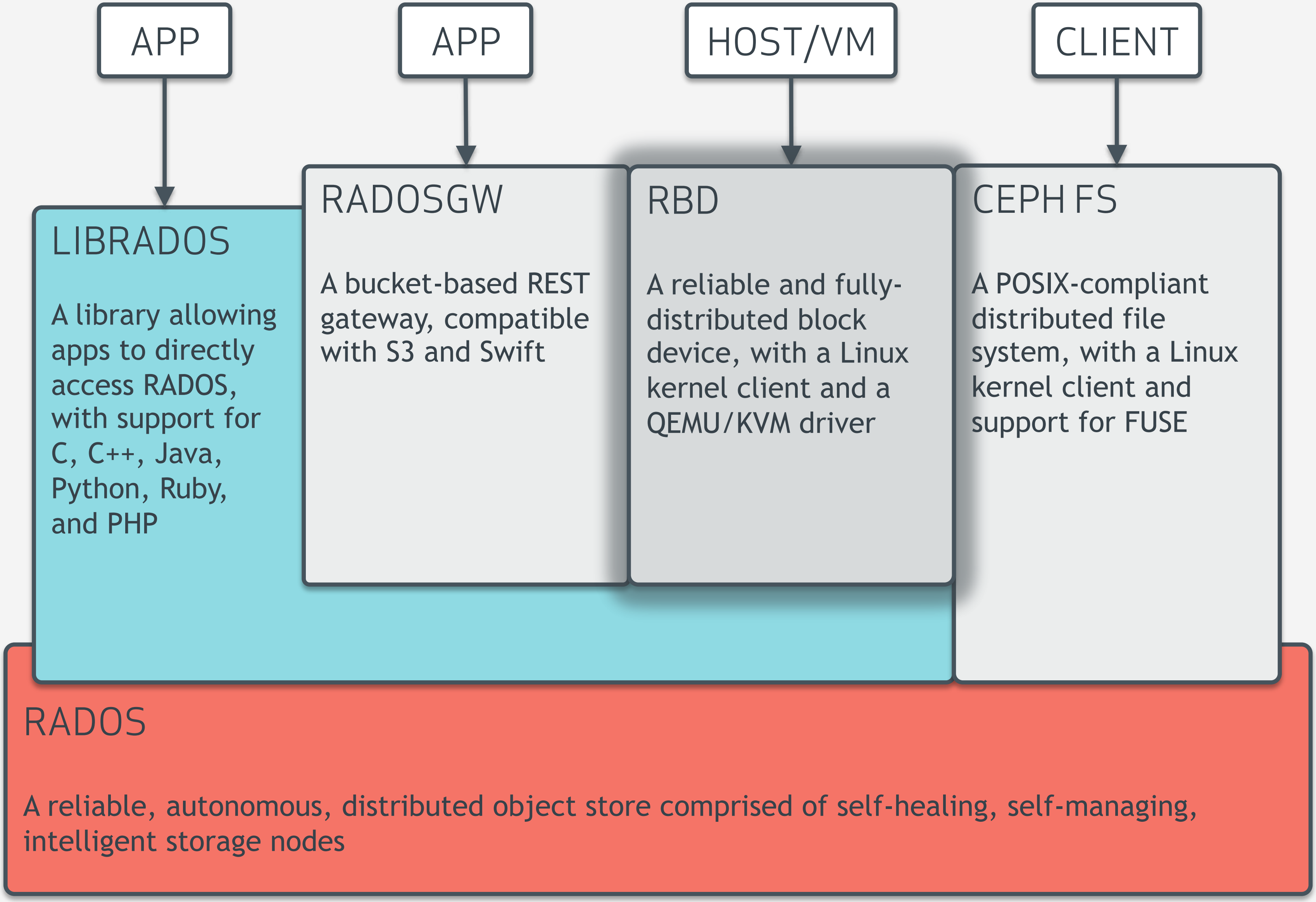
RADOS

A reliable, autonomous, distributed object store comprised of self-healing, self-managing, intelligent storage nodes


APP
APP
REST
RADOSGW
LIBRADOS
RADOSGW
LIBRADOS
native
M
M
M

RADOS Gateway:

- REST-based interface to RADOS
- Supports buckets, accounting
- Compatible with S3 and Swift applications



APP

APP

HOST/VM

CLIENT

LIBRADOS

A library allowing apps to directly access RADOS, with support for C, C++, Java, Python, Ruby, and PHP

RADOSGW

A bucket-based REST gateway, compatible with S3 and Swift

RBD

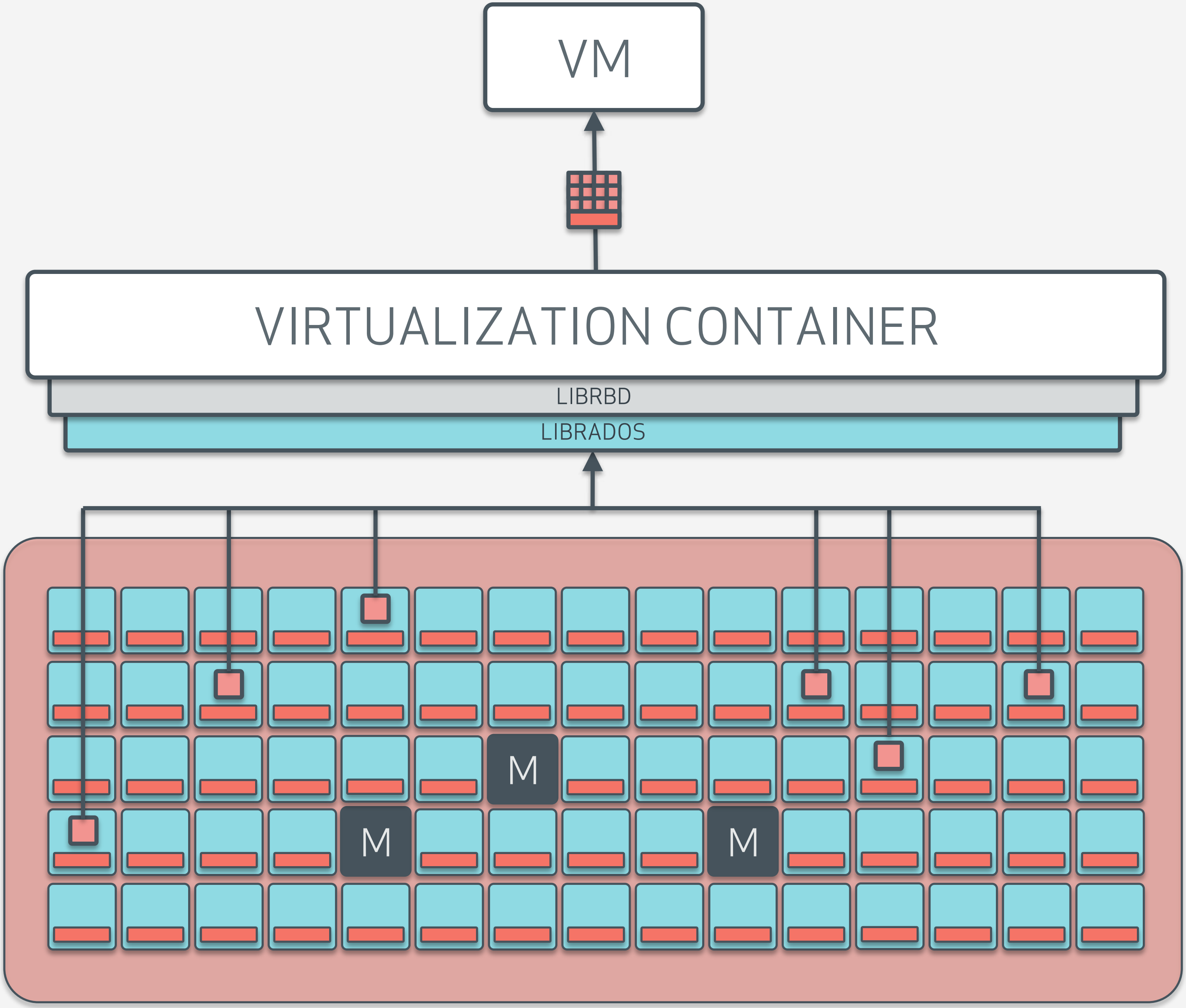
A reliable and fully-distributed block device, with a Linux kernel client and a QEMU/KVM driver

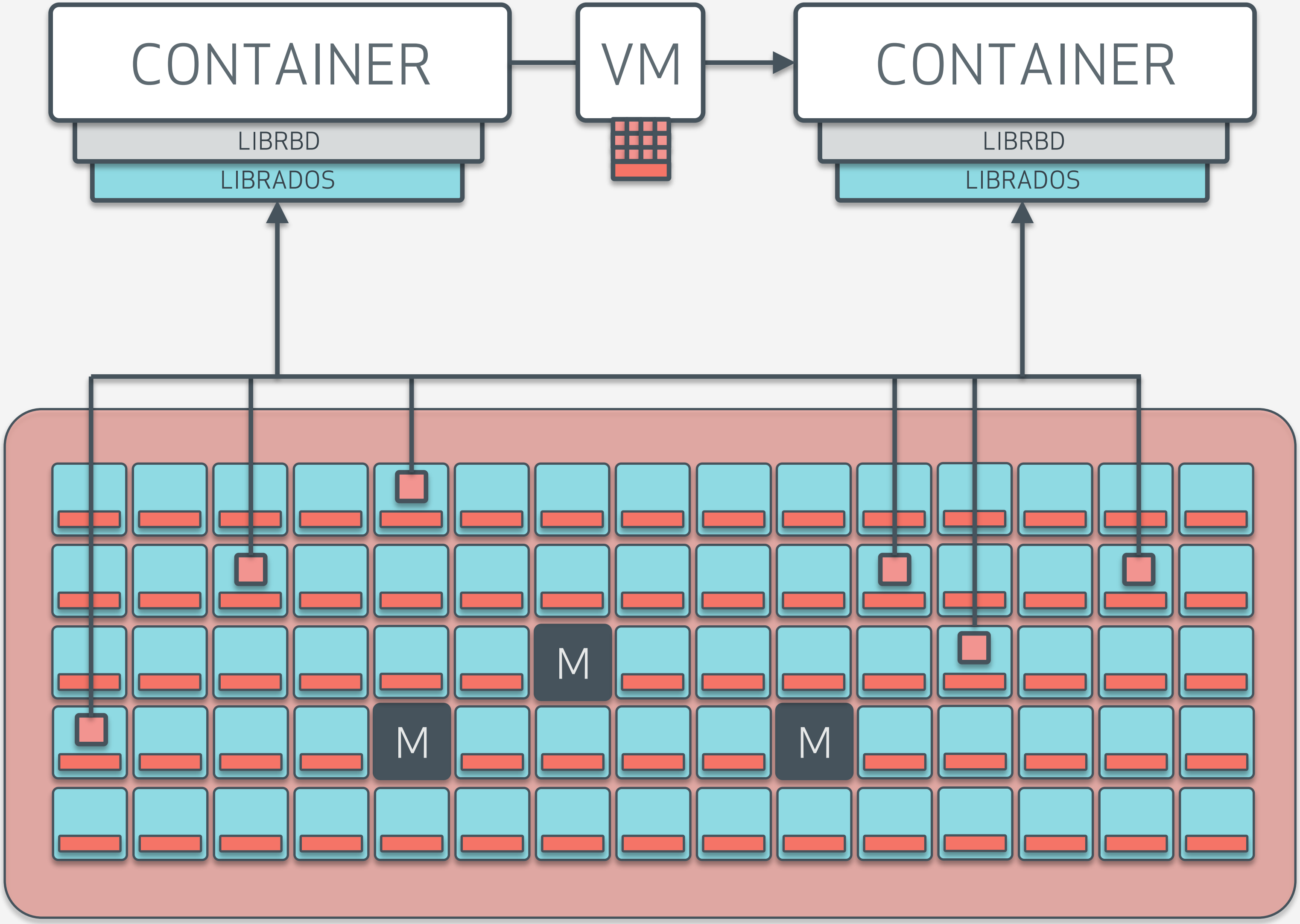
CEPH FS

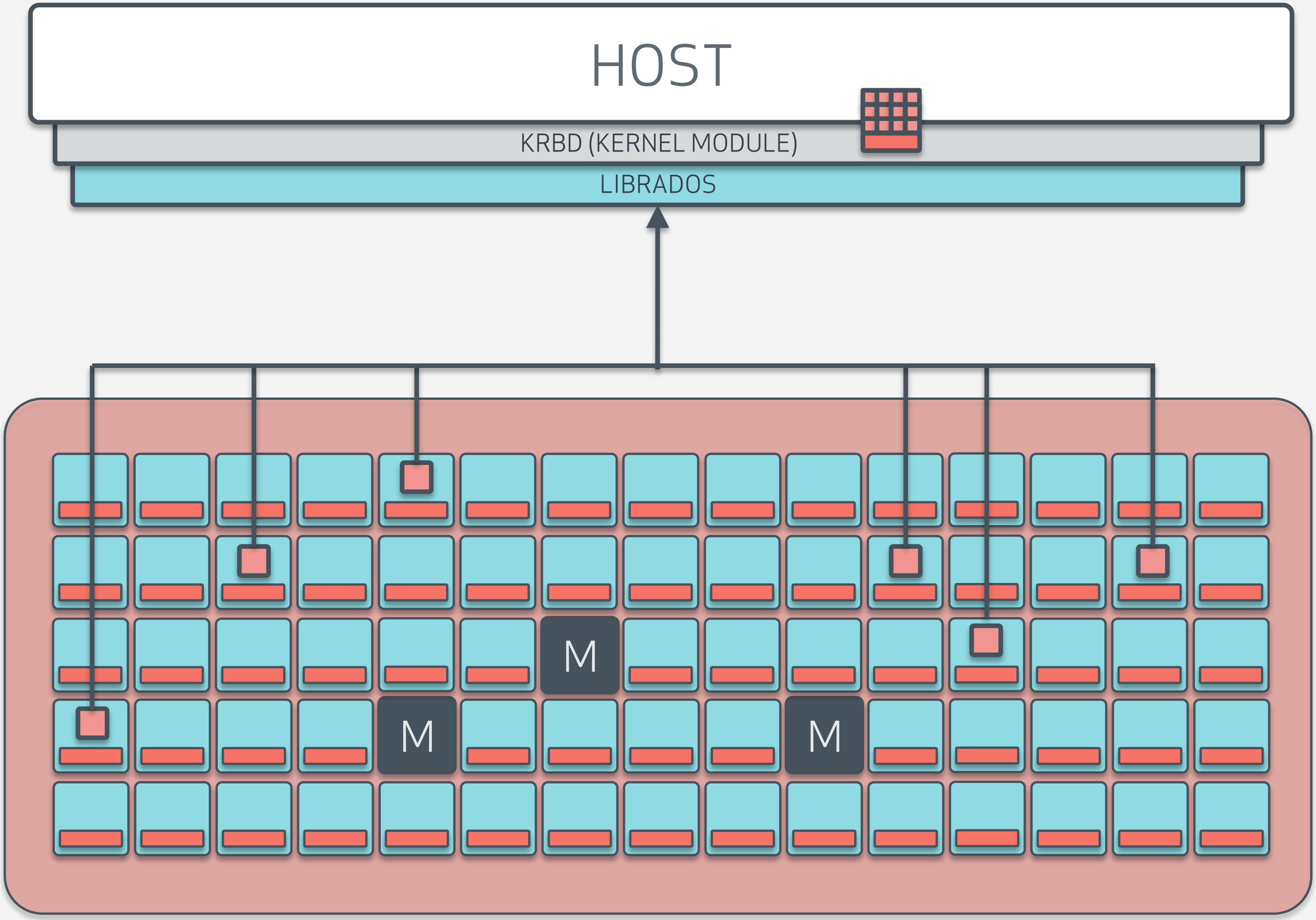
A POSIX-compliant distributed file system, with a Linux kernel client and support for FUSE

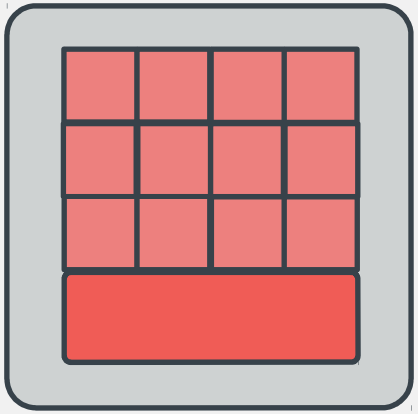
RADOS

A reliable, autonomous, distributed object store comprised of self-healing, self-managing, intelligent storage nodes


VM
VIRTUALIZATION CONTAINER
LIBRBD
LIBRADOS
M
M
M

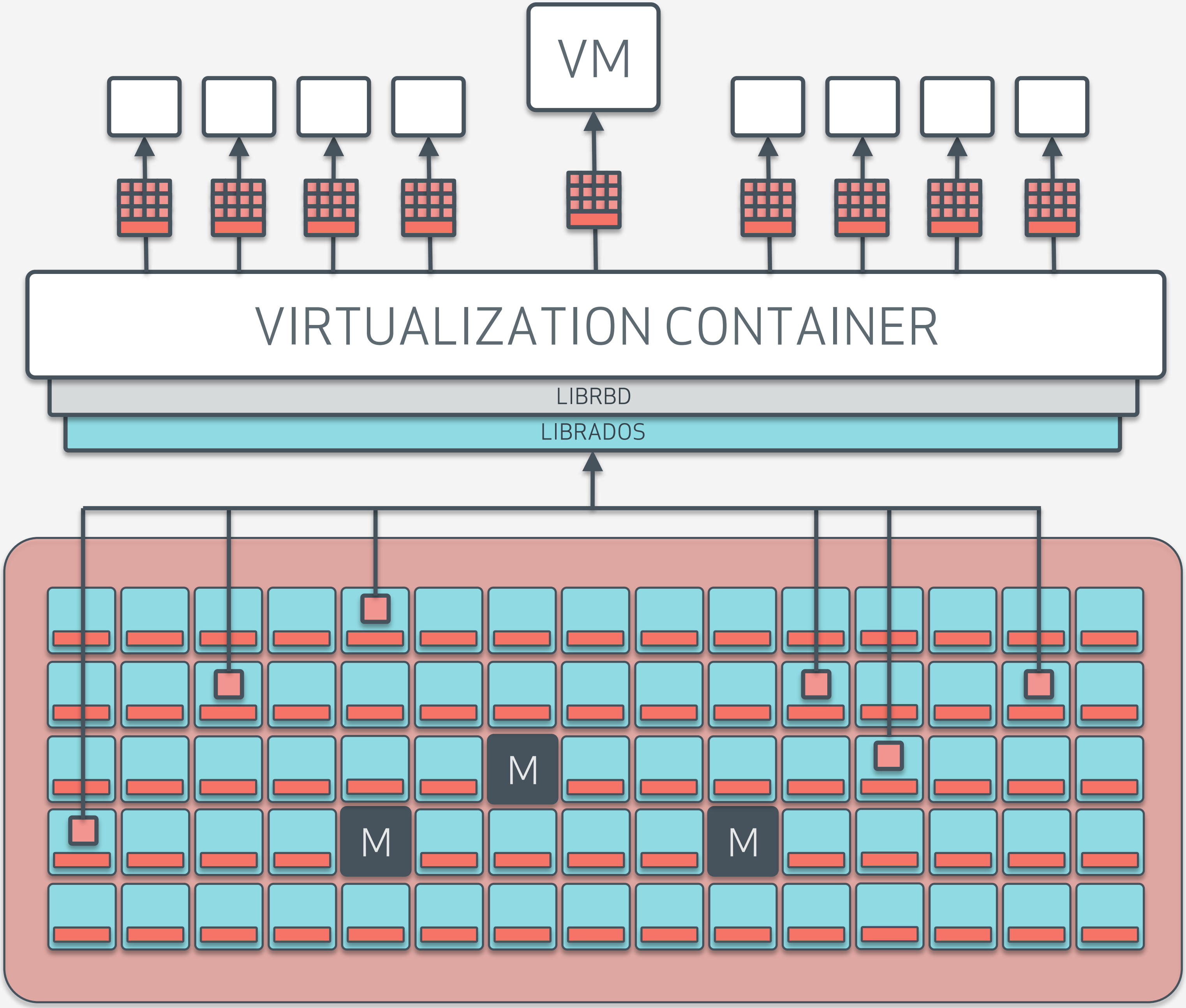

CONTAINER
VM
CONTAINER
LIBRBD
LIBRADOS
LIBRBD
LIBRADOS
M
M
M


HOST
KRBD (KERNEL MODULE)
LIBRADOS
M
M
M



RADOS Block Device:

- Storage of virtual disks in RADOS
- Allows decoupling of VMs and containers
- Live migration!
- Images are striped across the cluster
- Thin-provisioning
- Snapshots and cloning


VM
VIRTUALIZATION CONTAINER
LIBRBD
LIBRADOS
M
M
M

HOW DO YOU

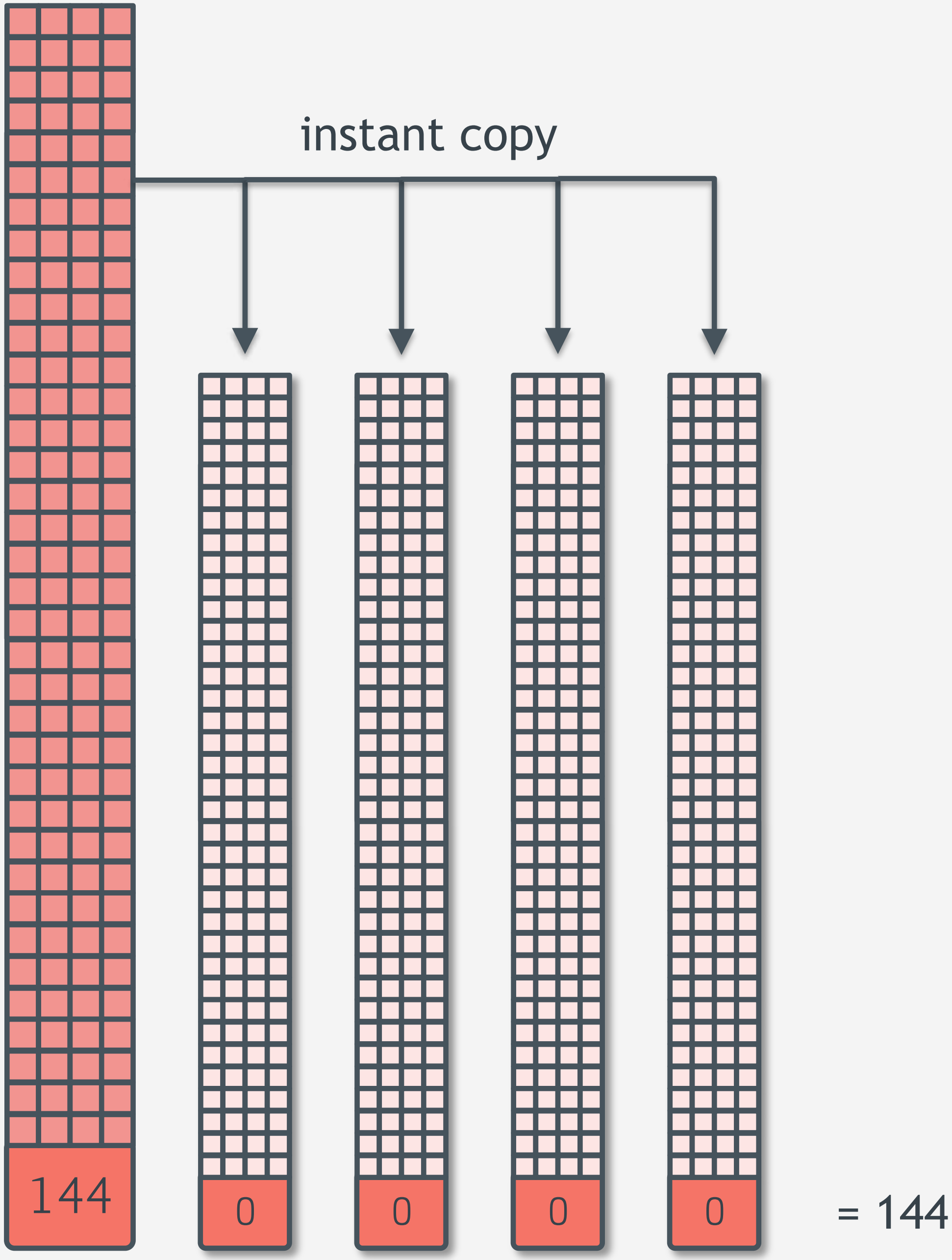
SPIN UP

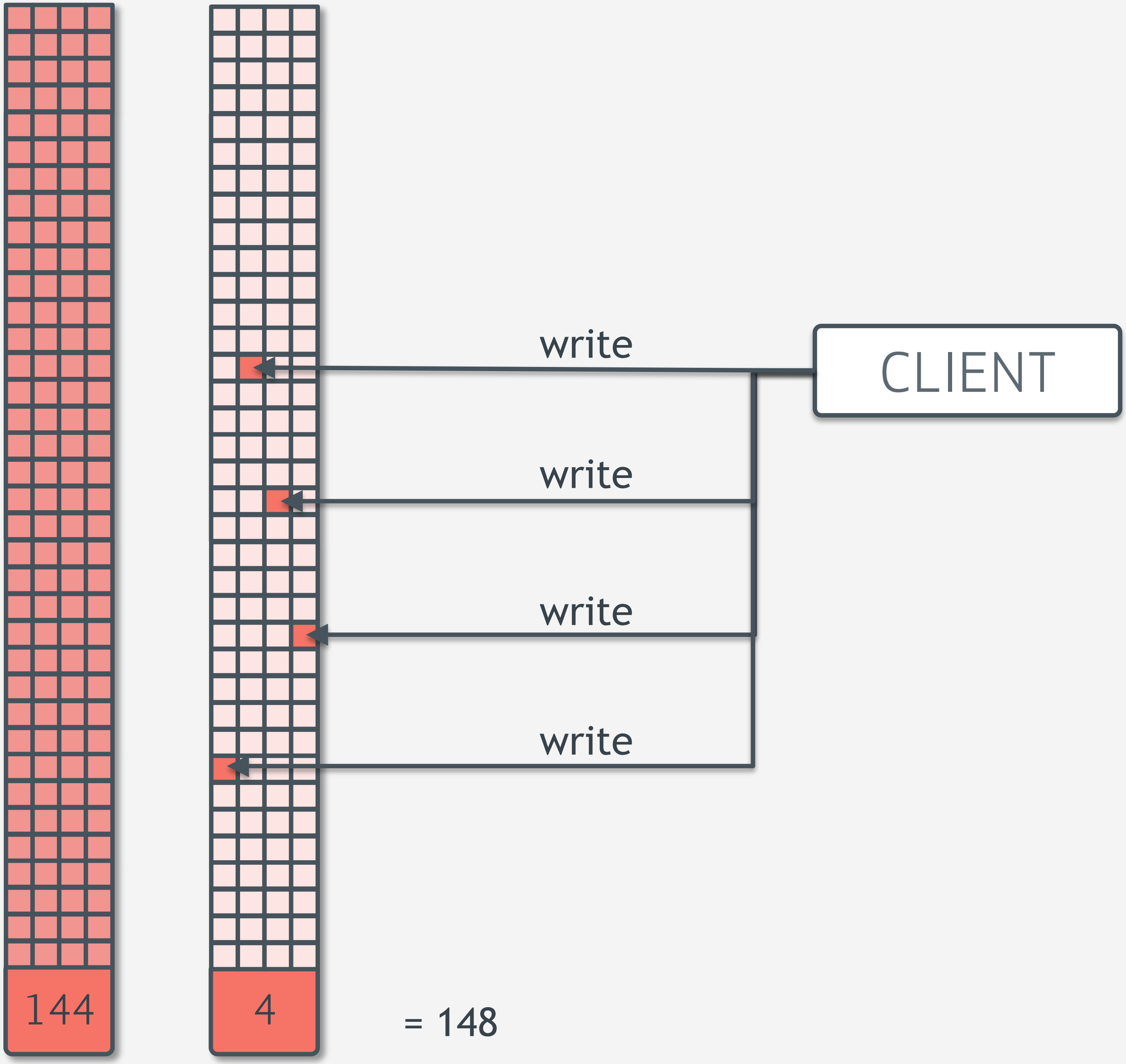
THOUSANDS OF VMs

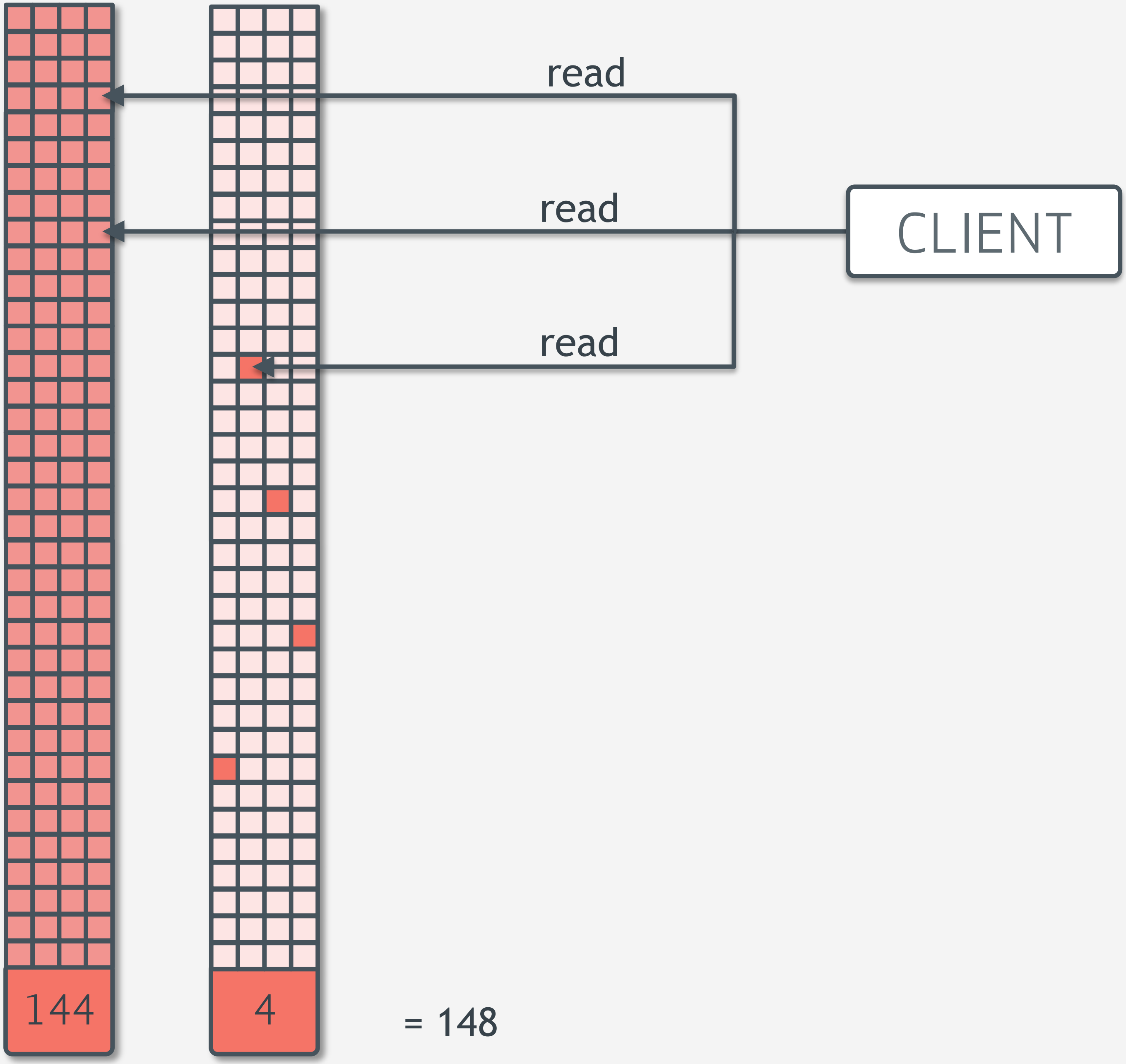
INSTANTLY

AND

EFFICIENTLY?

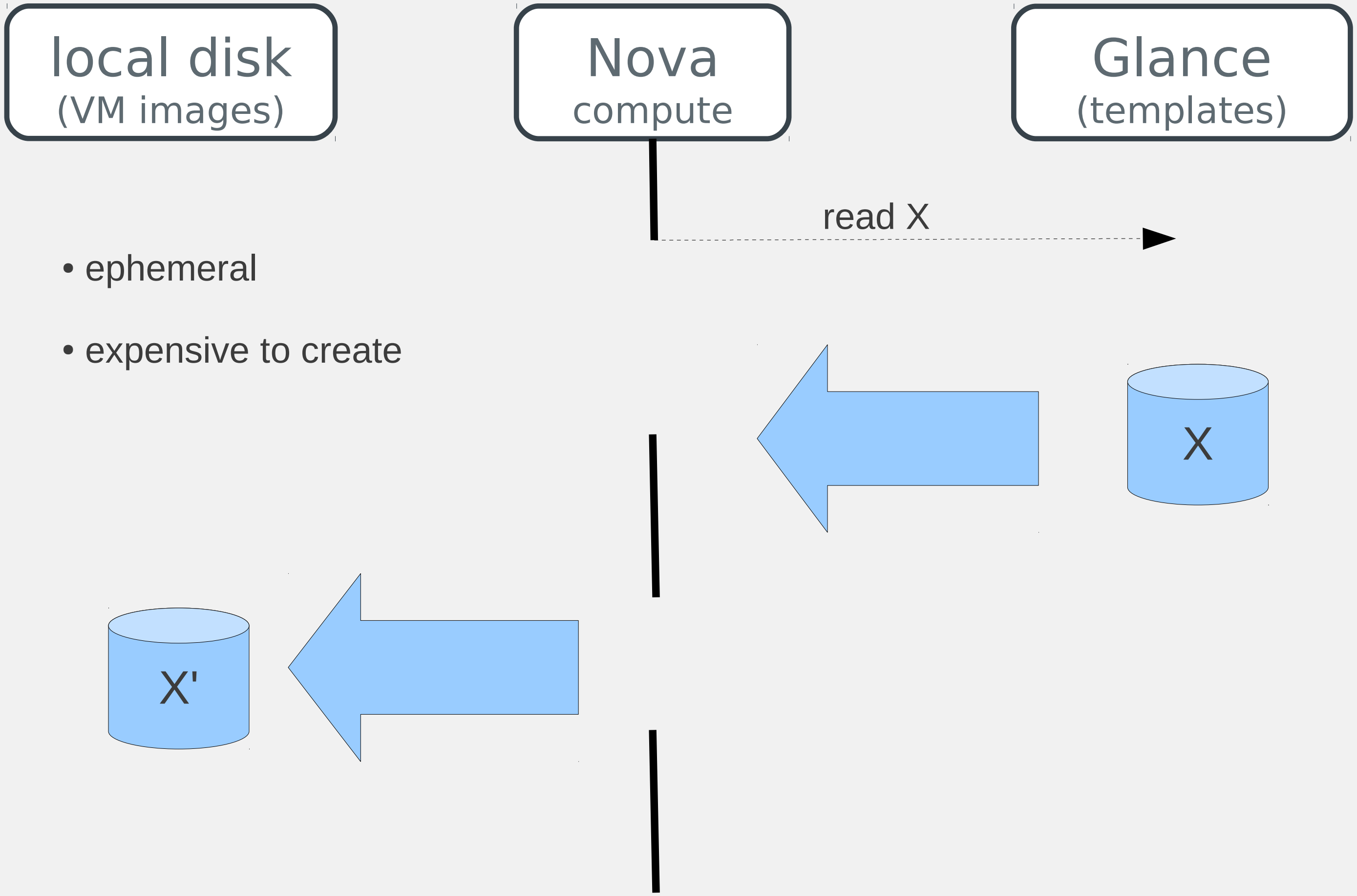

instant copy
144
0
0
0
0
= 144


write
CLIENT
write
write
write
144
4
= 148



= 148

old-style VM image creation

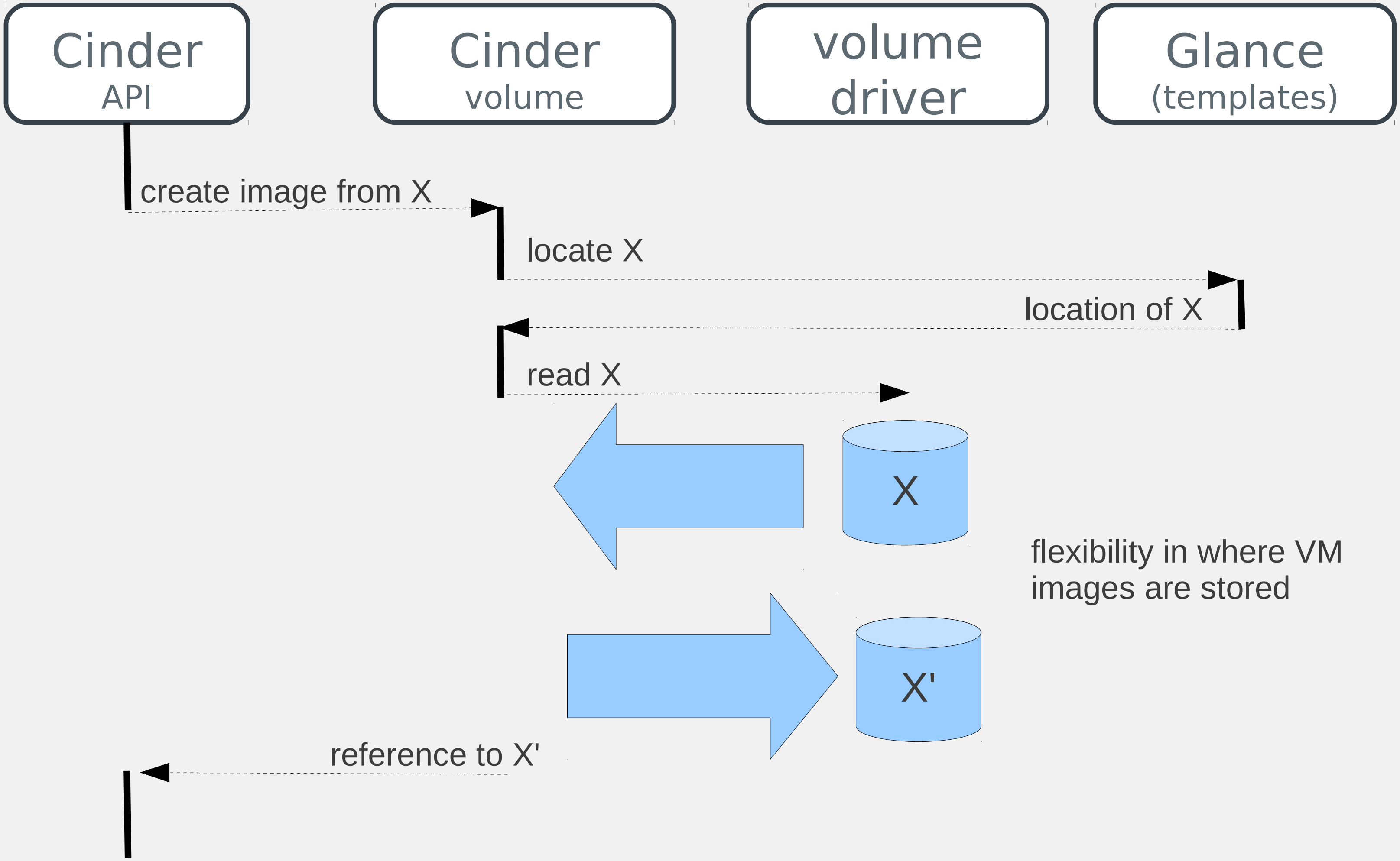


- ephemeral
- expensive to create

Why use block storage?

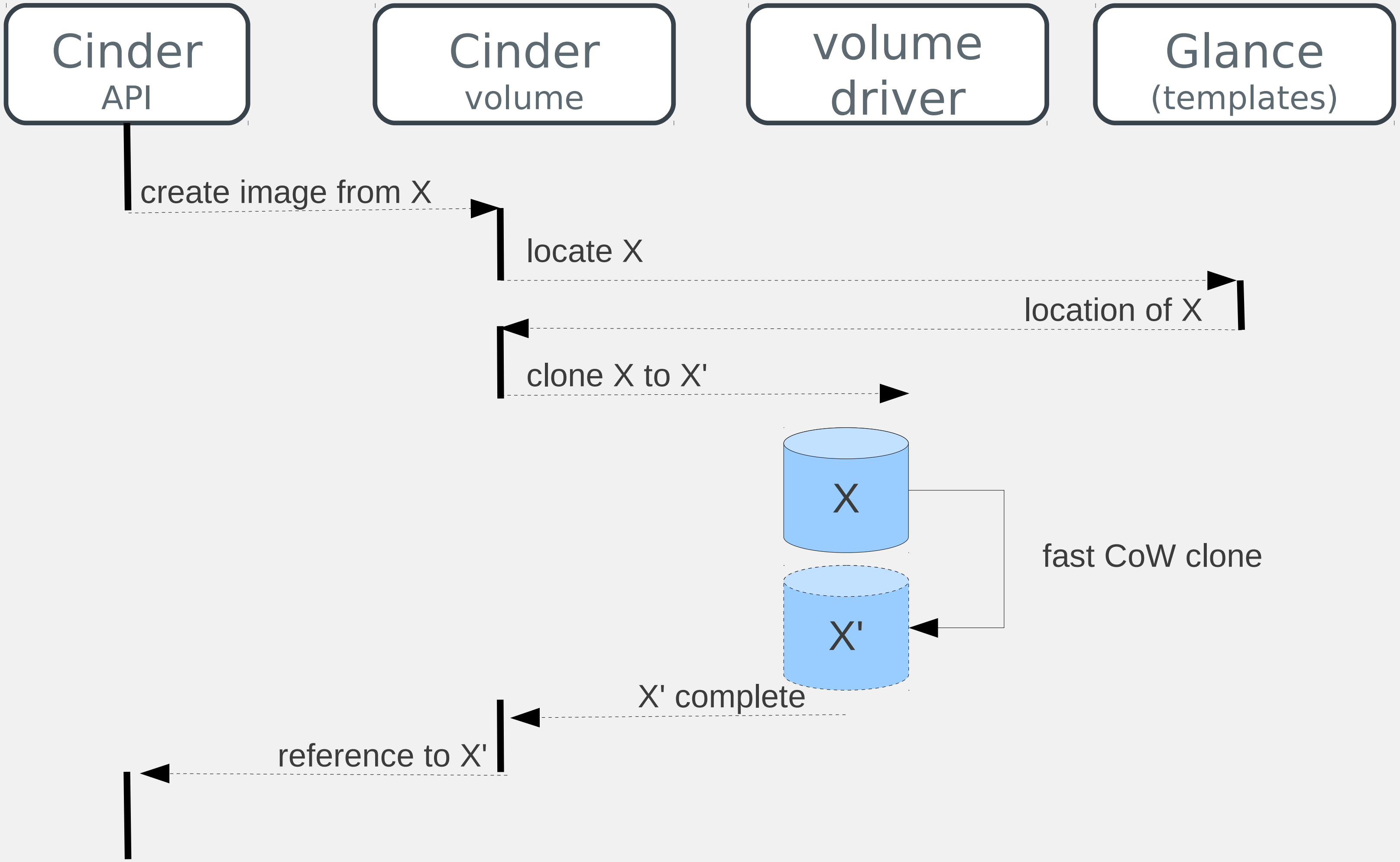
- Persistent
  - More familiar to users
- Not tied to a single host
  - Decouples compute and storage
  - Enables Live migration
- Extra capabilities of storage system
  - Efficient snapshots
  - Different types of storage available
  - Cloning for fast restore or scaling

Cinder volume creation



flexibility in where VM images are stored

Efficient volume creation



What's new in Bobtail: Improved OSD threading

- Filesystem and journal related-locks are now more fine-grained
- Boosted single disk IOPS from 6k to 22k
- Restructured how map updates are handled, letting each placement group process them independently

What's new in Bobtail: Recovery QoS

- Message priority system reworked to prevent starvation
- Recovery operations can be lower priority than client I/O without starving
- Requests to access an object can increase recovery priority for that object

What's new in Bobtail: Block Device Cloning

- Instantly create new volumes based on templates (snapshots)
- Integrated with Cinder in Folsom
- Grizzly adds the ability to copy (not clone) non-raw images to RBD

What's new in Bobtail: Keystone Integration

- RADOS gateway can talk to keystone to authenticate swift api requests
- Let keystone manage your users
- Supported by the Ceph juju charm

What's next: Cuttlefish

- Incremental backup for block devices
- On-disk encryption
- REST management API for RADOS gateway
- More performance improvements (especially for small I/O)
- More! (http://www.inktank.com/about-inktank/roadmap/)

What's next: Dumpling

- Geo-replication for RADOS gateway
- REST management API for Ceph cluster
- ...

(virtual) Ceph Developer Summit May 6

Questions?

Josh Durgin
josh.durgin@inktank.com
jdurgin on freenode

inktank.com | ceph.com

inktank