Ceph Intro & Architectural Overview

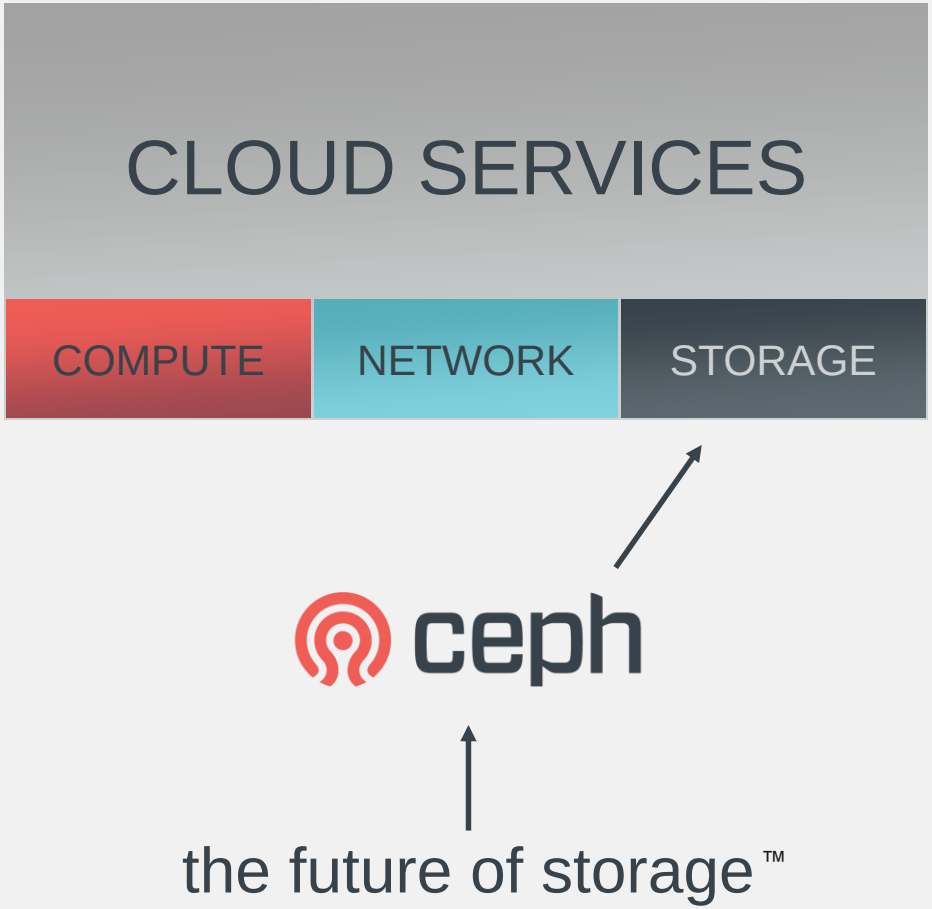
Abbas Bangash
Intercloud Systems

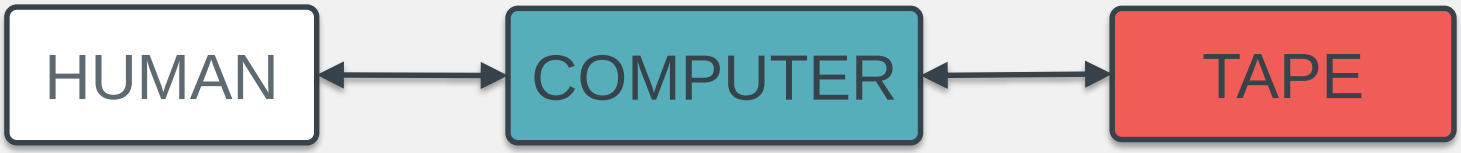
About Me

Abbas Bangash
Systems Team Lead,
Intercloud Systems

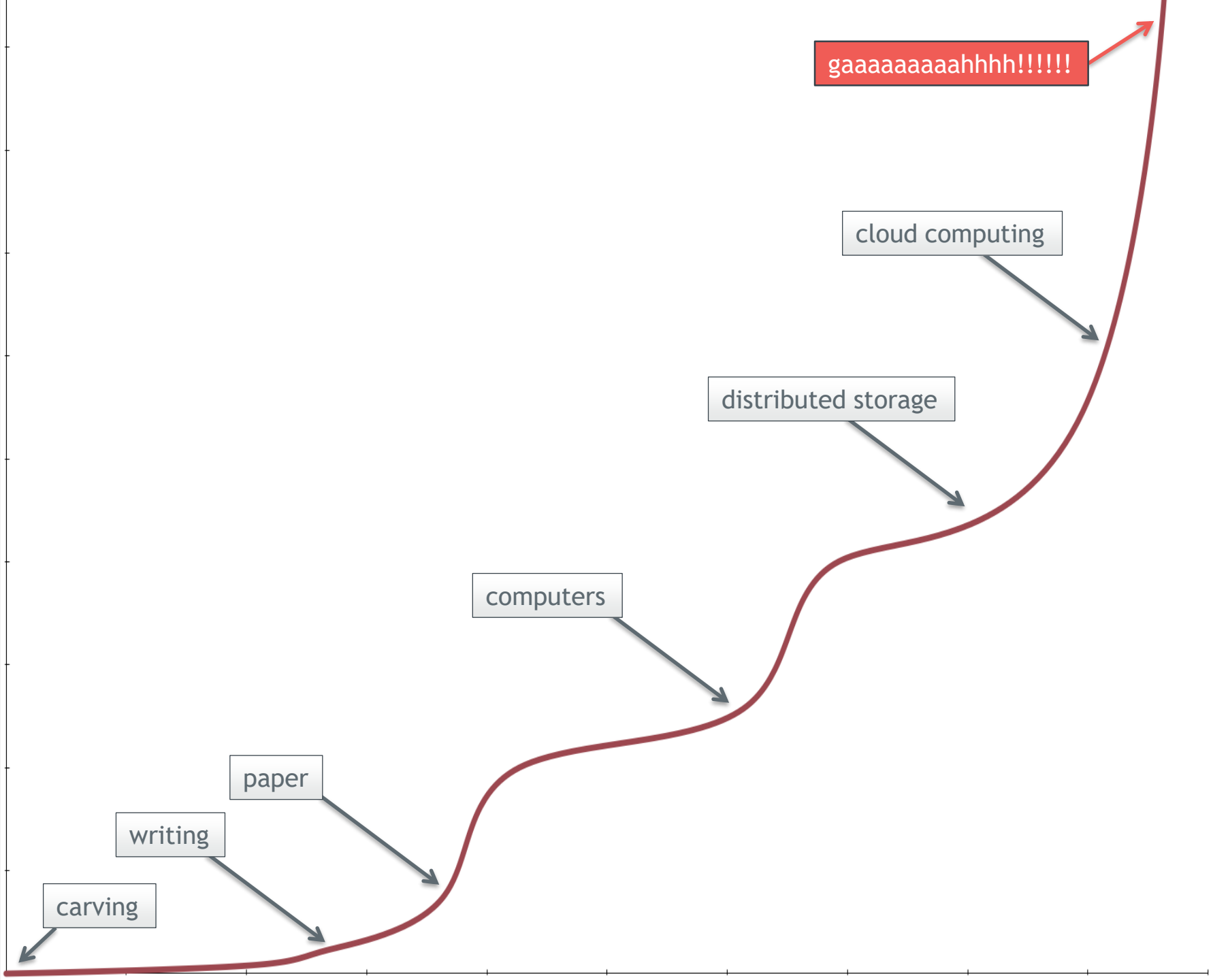
abangash@intercloudsys.com

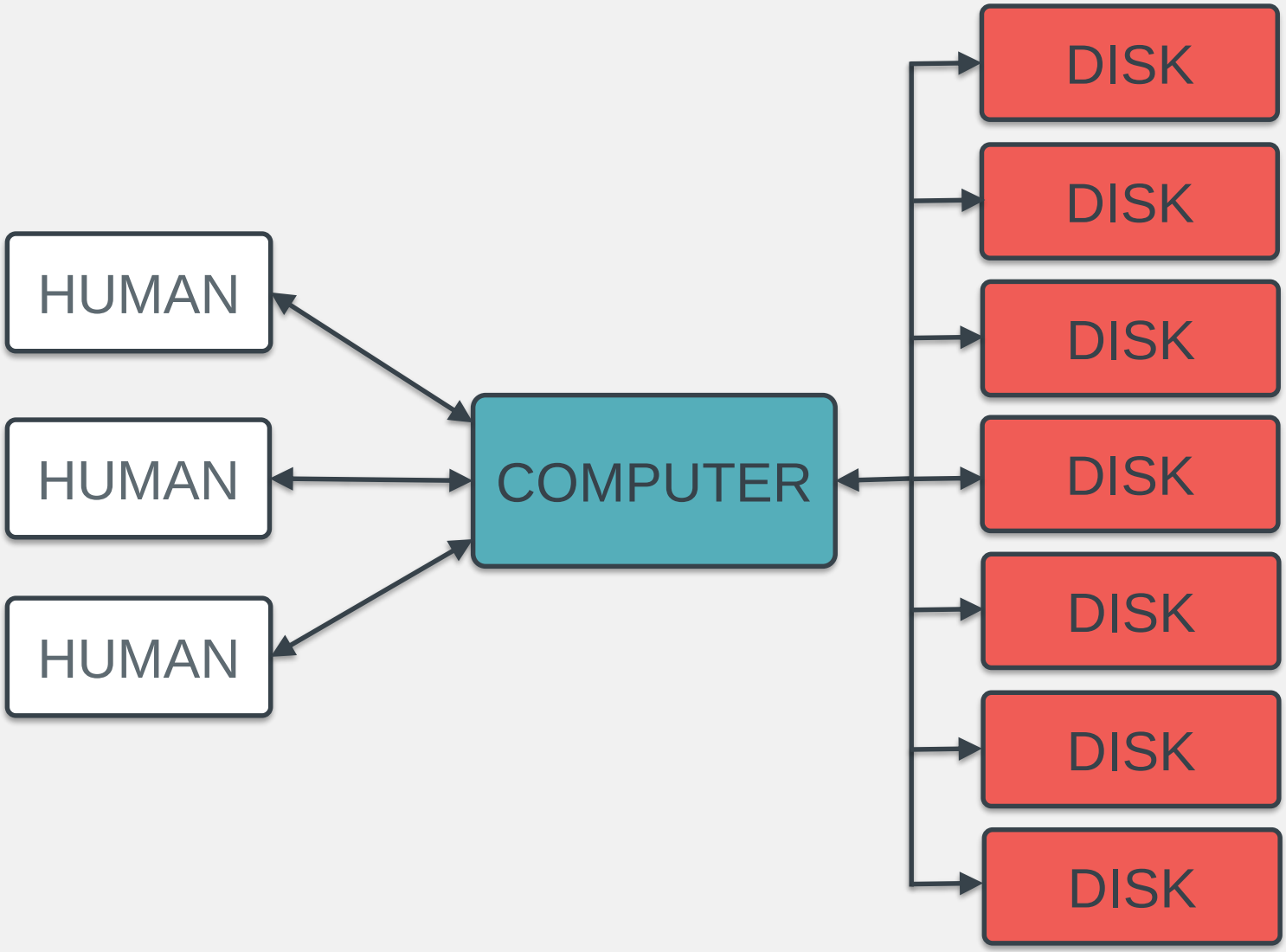
intercloudsys.com

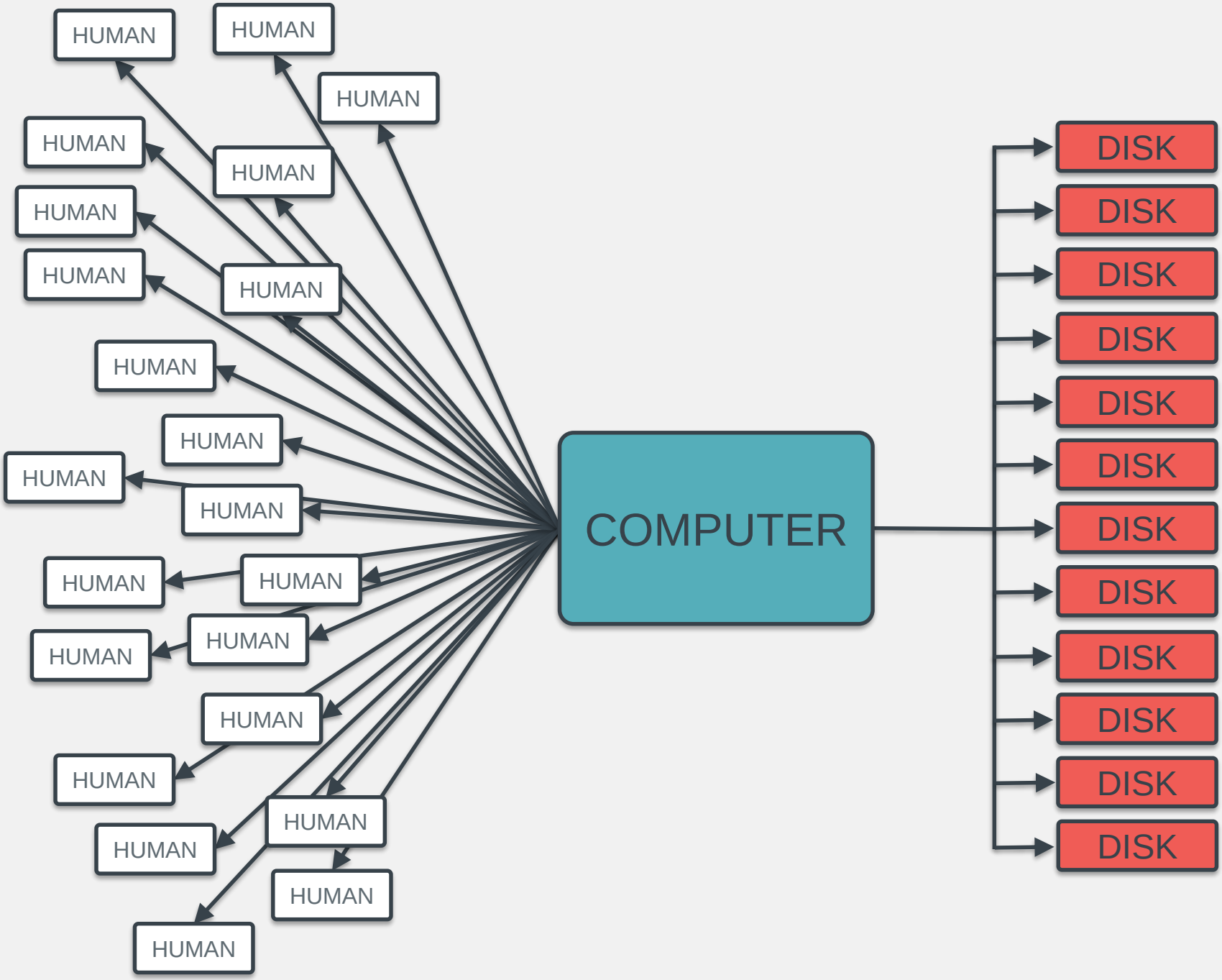

CLOUD SERVICES
COMPUTE
NETWORK
STORAGE
ceph
the future of storage™

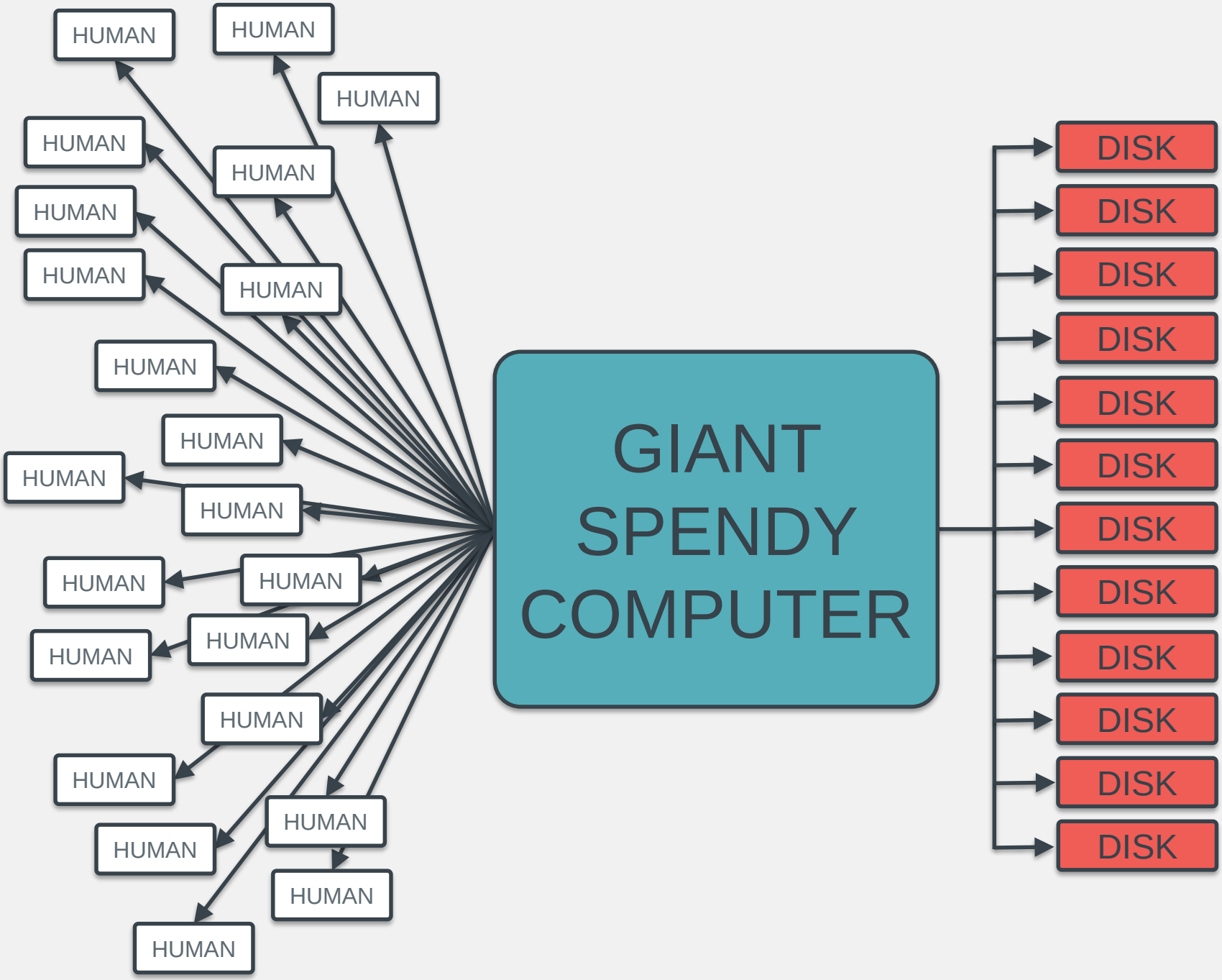

HUMAN
COMPUTER
TAPE

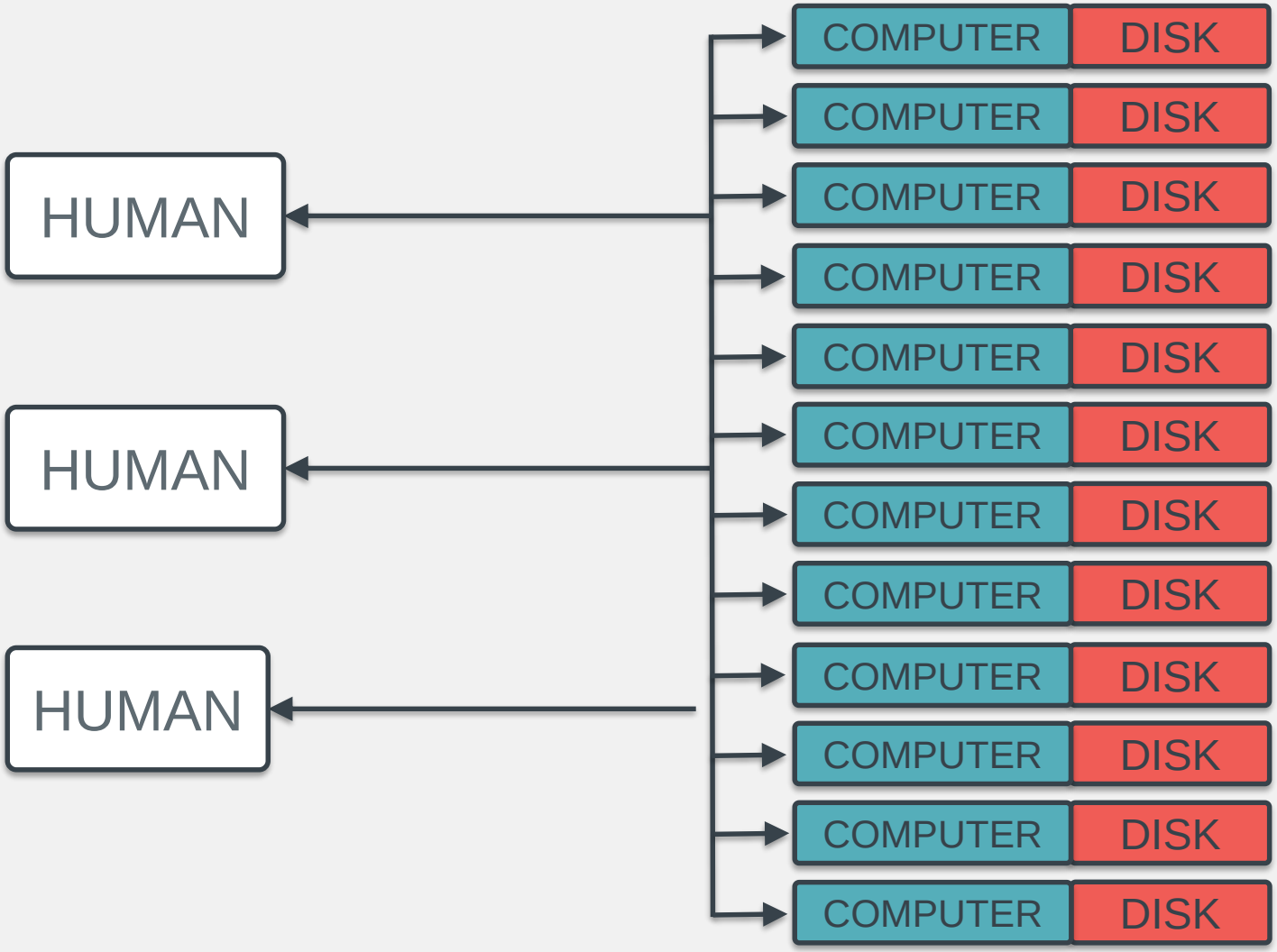

YOU
TECHNOLOGY
YOUR DATA

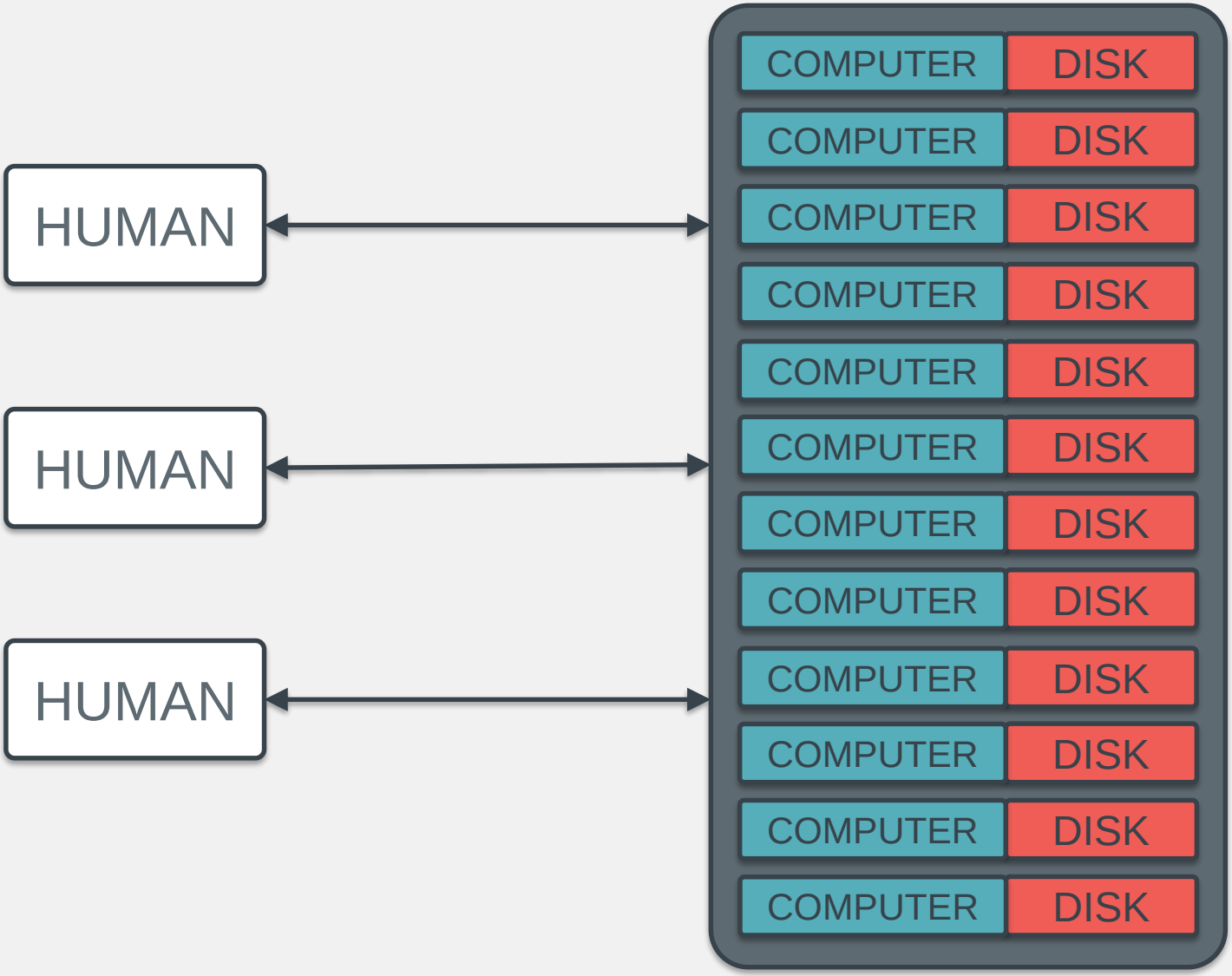

gaaaaaaaaahhhh!!!!!!
cloud computing
distributed storage
computers
paper
writing
carving


HUMAN
HUMAN
HUMAN
COMPUTER
DISK
DISK
DISK
DISK
DISK
DISK
DISK

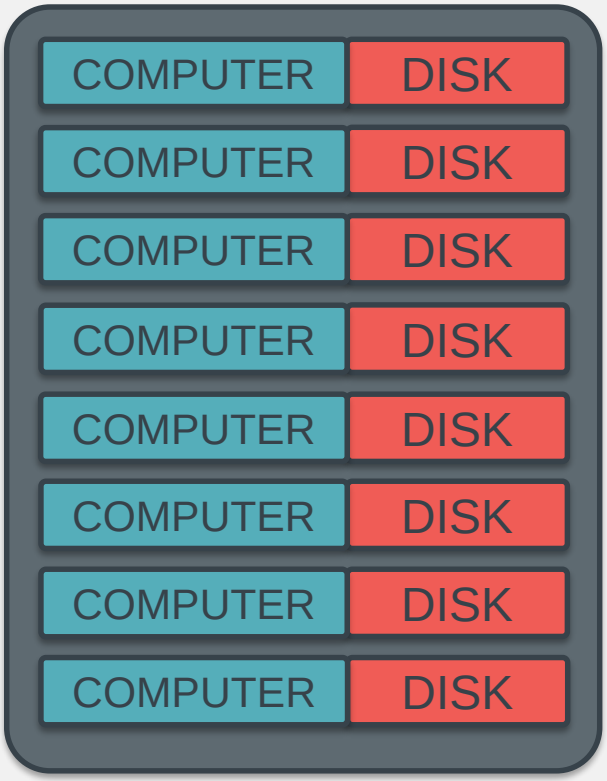

HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
COMPUTER
DISK
DISK
DISK
DISK
DISK
DISK
DISK
DISK
DISK
DISK
DISK


HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
HUMAN
GIANT SPENDY COMPUTER
DISK
DISK
DISK
DISK
DISK
DISK
DISK
DISK
DISK
DISK
DISK

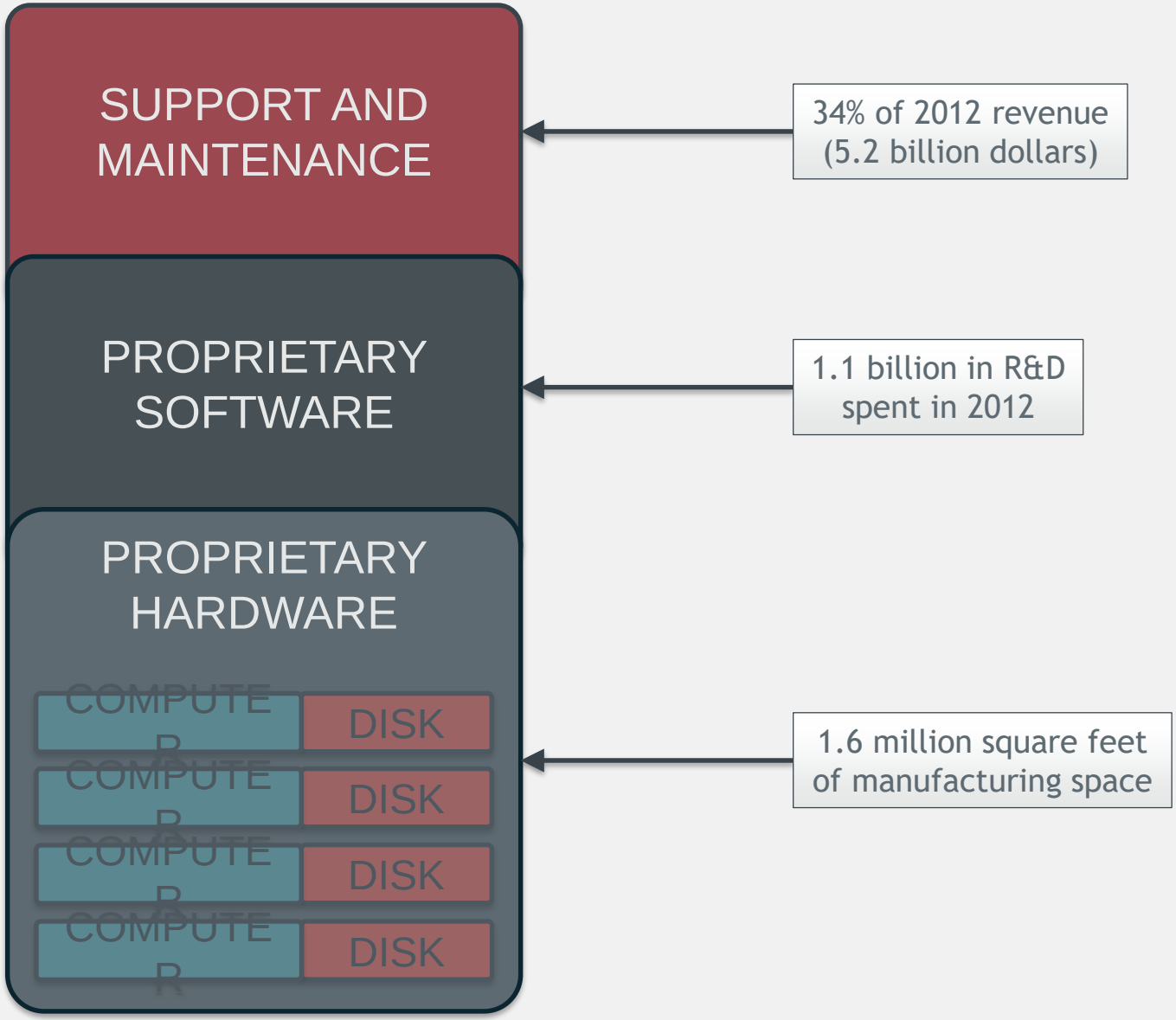

HUMAN
HUMAN
HUMAN
COMPUTER
DISK
COMPUTER
DISK
COMPUTER
DISK
COMPUTER
DISK
COMPUTER
DISK
COMPUTER
DISK
COMPUTER
DISK
COMPUTER
DISK
COMPUTER
DISK
COMPUTER
DISK


HUMAN
HUMAN
HUMAN
COMPUTER
DISK
COMPUTER
DISK
COMPUTER
DISK
COMPUTER
DISK
COMPUTER
DISK
COMPUTER
DISK
COMPUTER
DISK
COMPUTER
DISK
COMPUTER
DISK
COMPUTER
DISK

“STORAGE APPLIANCE”

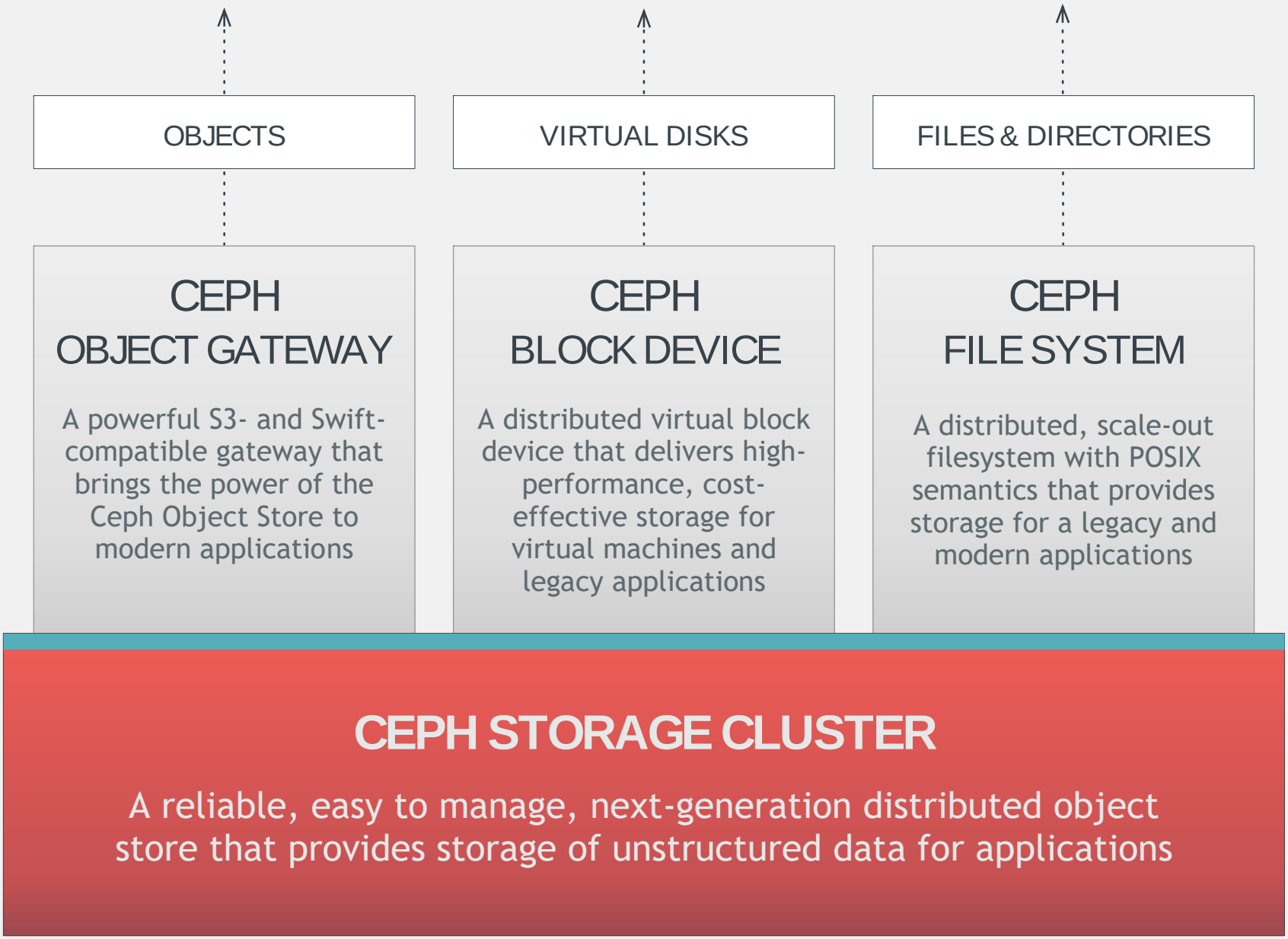


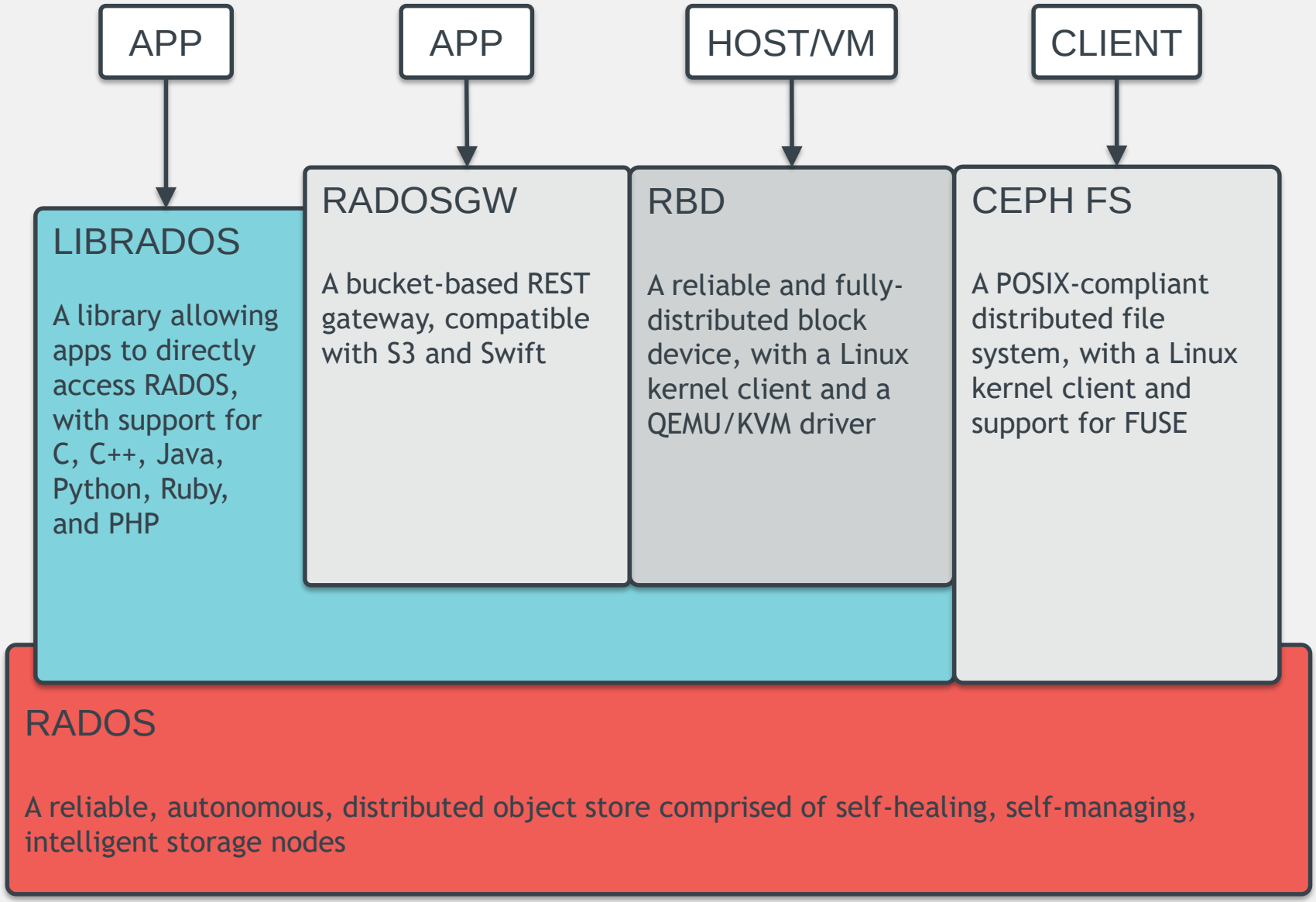
Storage Appliance

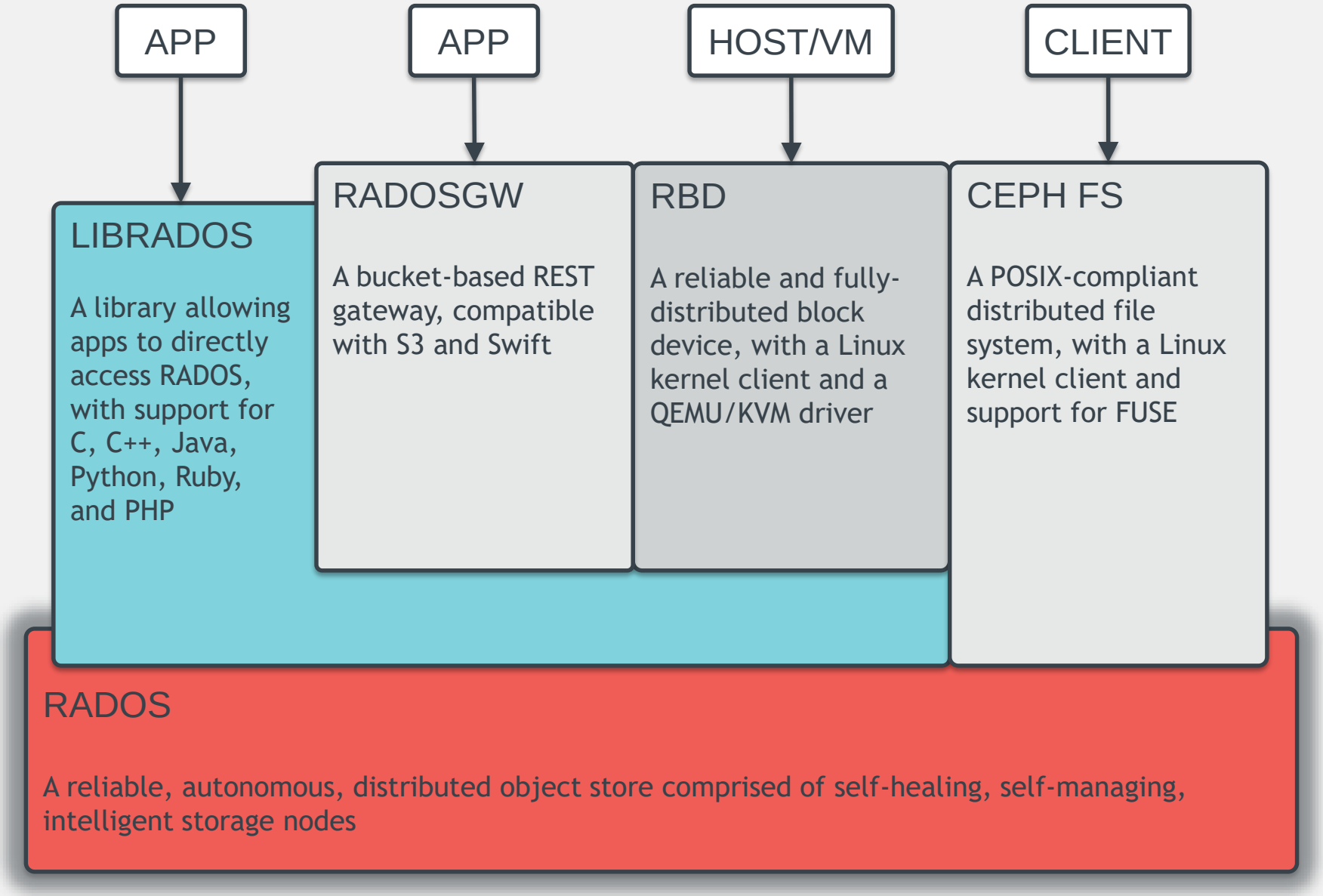

SUPPORT AND MAINTENANCE
PROPRIETARY SOFTWARE
PROPRIETARY HARDWARE
COMPUTER
DISK
COMPUTER
DISK
COMPUTER
DISK
COMPUTER
DISK
34% of 2012 revenue (5.2 billion dollars)
1.1 billion in R&D spent in 2012
1.6 million square feet of manufacturing space

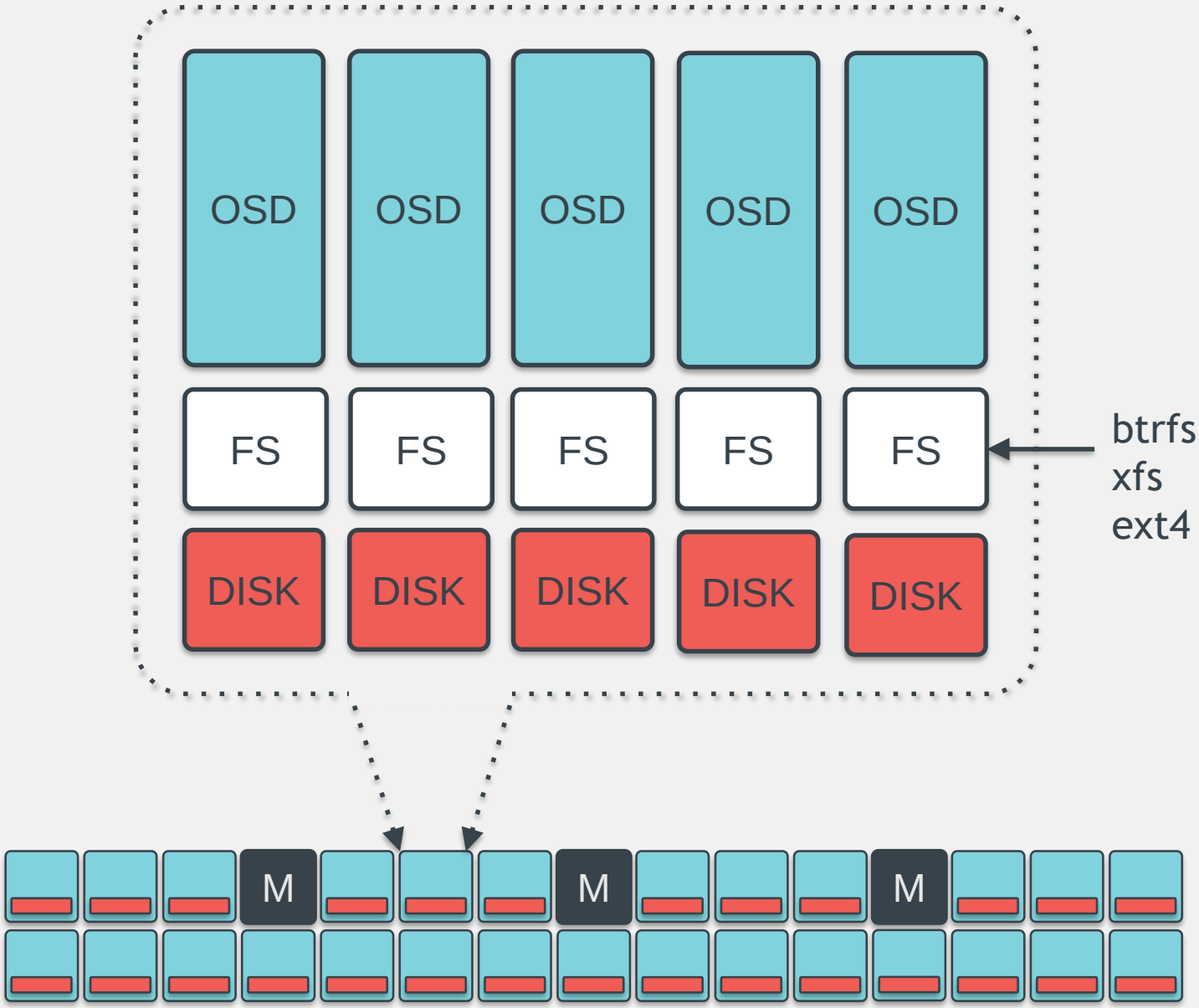

ceph

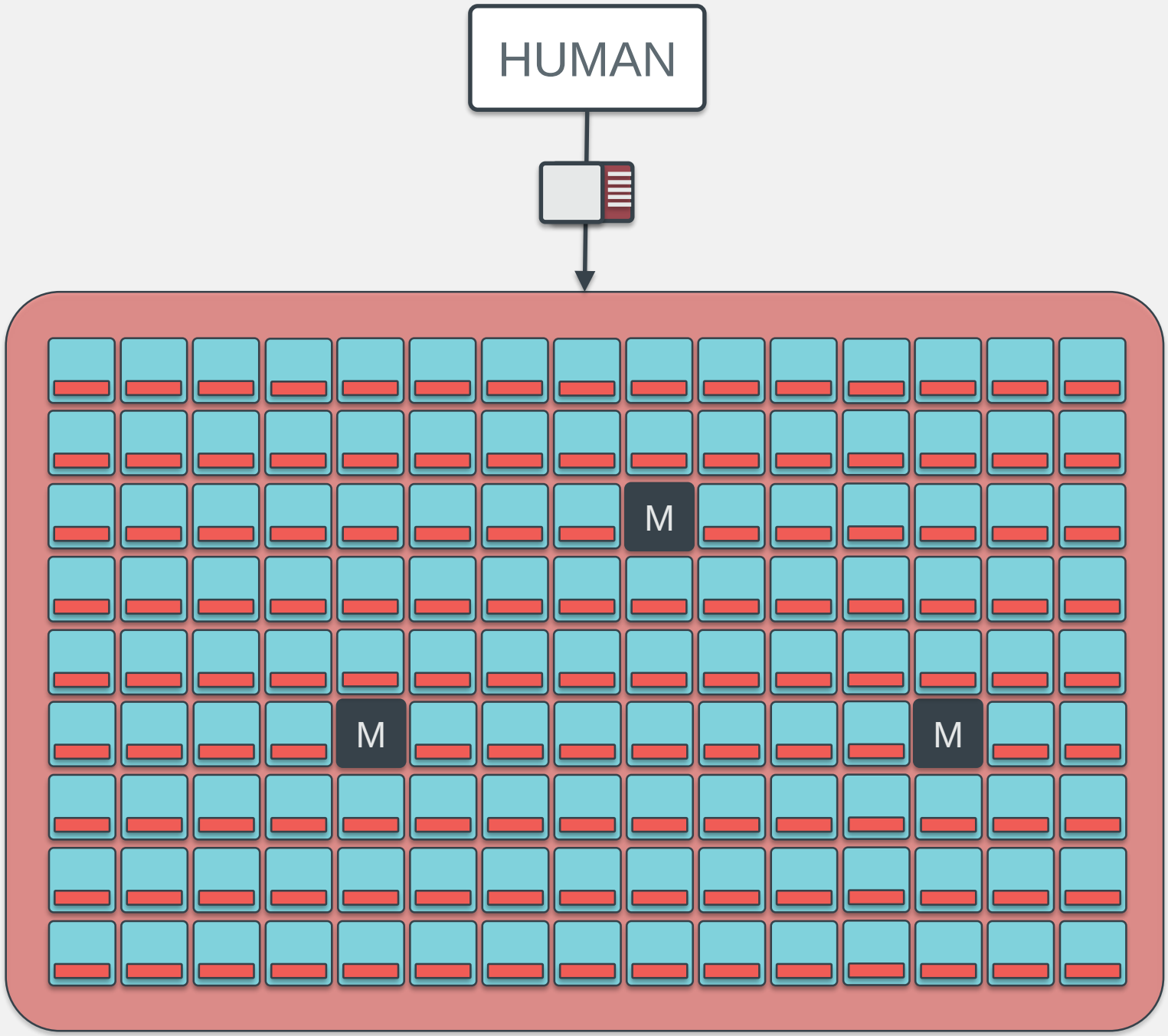
philosophy	design
OPEN SOURCE	SCALABLE
COMMUNITY-FOCUSED	NO SINGLE POINT OF FAILURE
	SOFTWARE BASED
	SELF-MANAGING

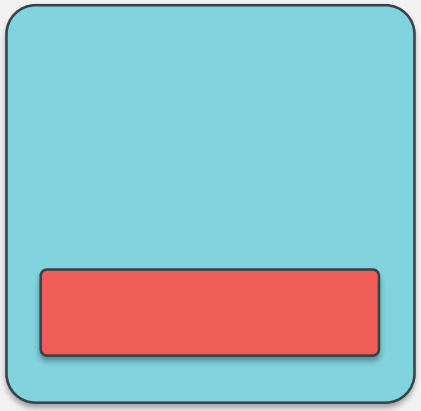

OBJECTS
VIRTUAL DISKS
FILES & DIRECTORIES
CEPH
OBJECT GATEWAY
A powerful S3- and Swift-compatible gateway that brings the power of the Ceph Object Store to modern applications
CEPH
BLOCK DEVICE
A distributed virtual block device that delivers high-performance, cost-effective storage for virtual machines and legacy applications
CEPH
FILE SYSTEM
A distributed, scale-out filesystem with POSIX semantics that provides storage for a legacy and modern applications
CEPH STORAGE CLUSTER
A reliable, easy to manage, next-generation distributed object store that provides storage of unstructured data for 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


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


OSD
OSD
OSD
OSD
OSD
FS
FS
FS
FS
FS
DISK
DISK
DISK
DISK
DISK
btrfs
xfs
ext4
M
M
M


HUMAN
M
M
M

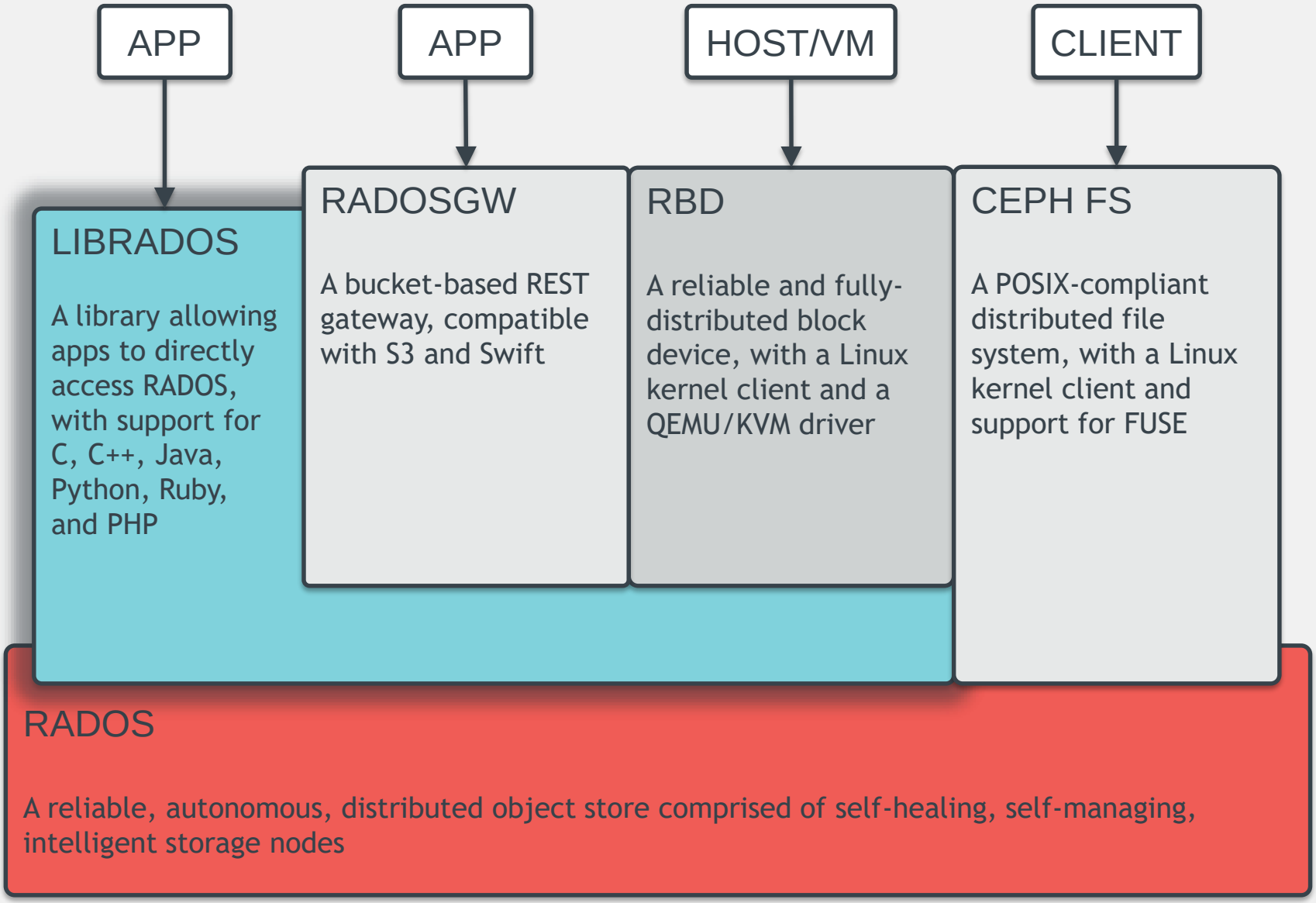


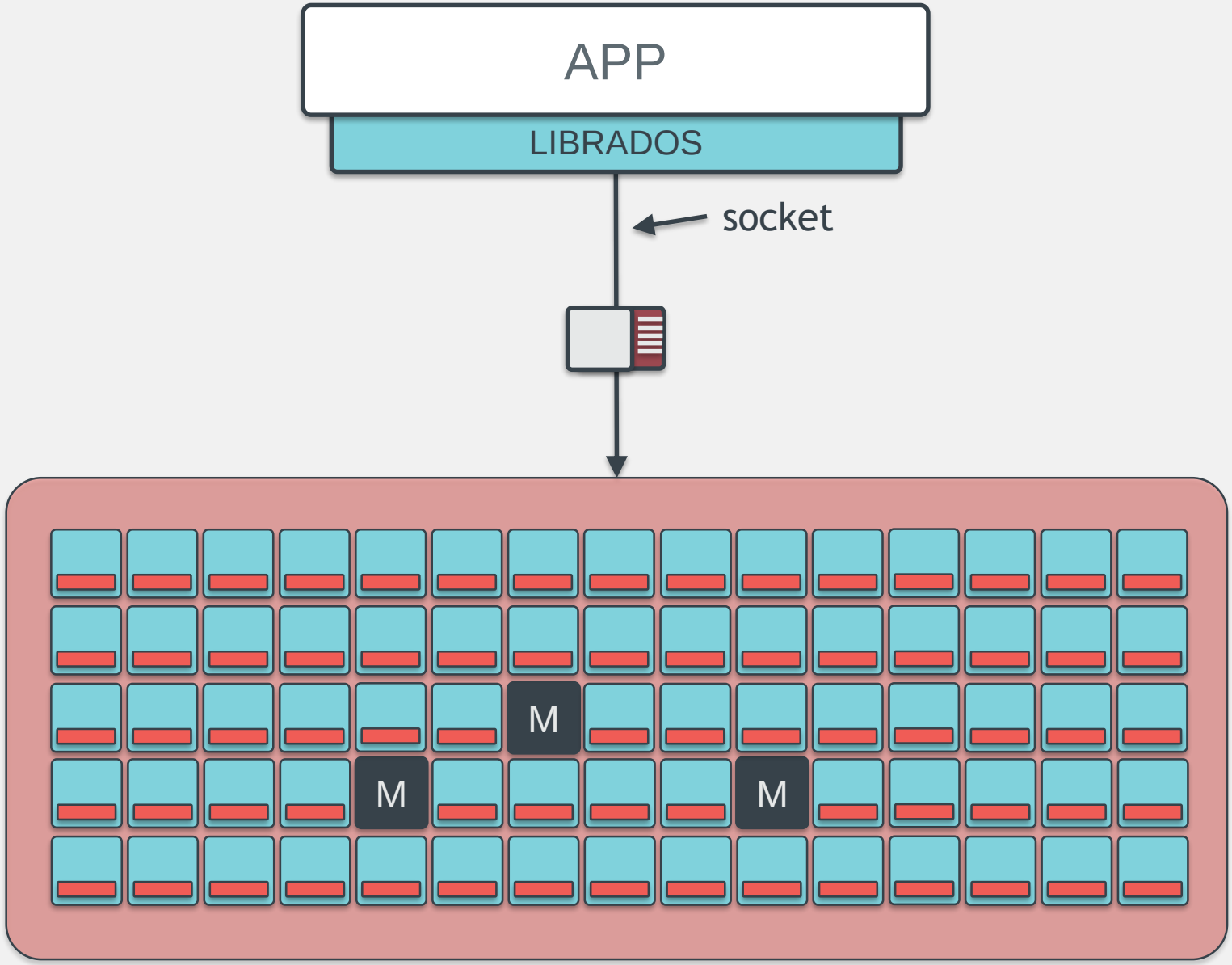
OSDs:

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

Monitors:

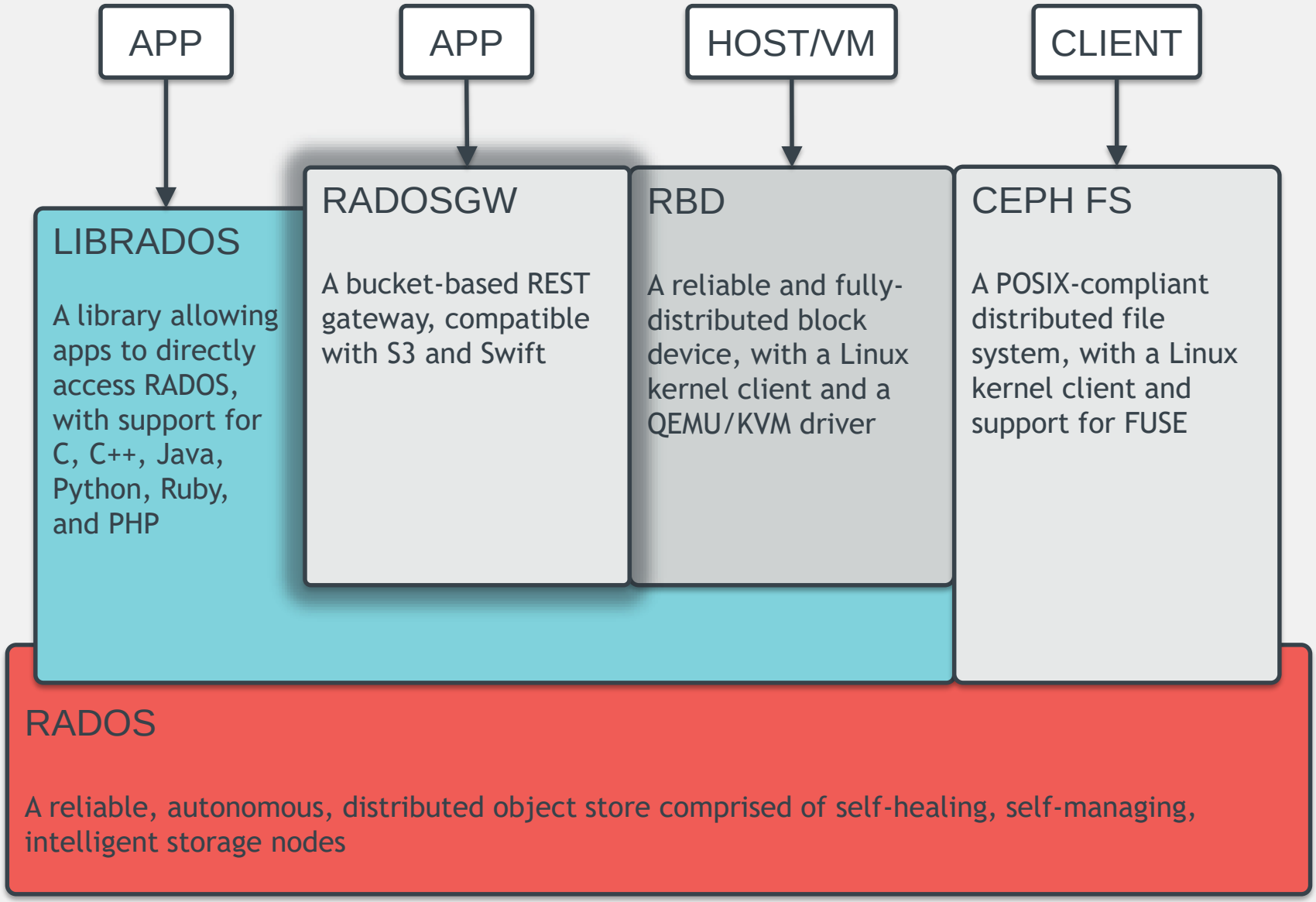
- Maintain cluster membership and state
- Provide consensus for distributed decision-making
- Small, odd number
- These do **not** serve stored objects to clients

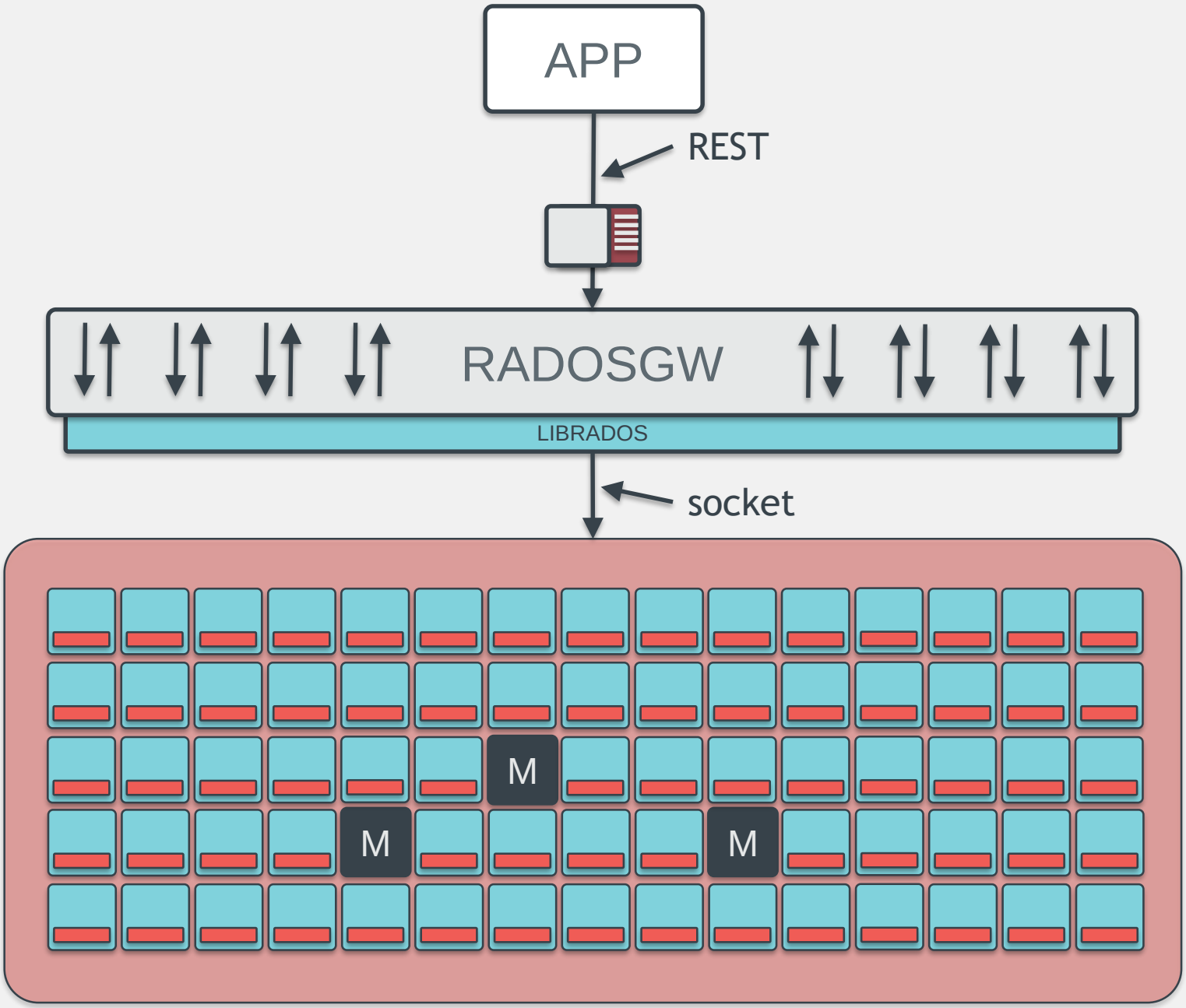

APP
APP
HOST/VM
CLIENT
LIBRADOS
A library allowing apps to directly access RADOS, with support for C, C++, Java, Python, Ruby, and PHP
RADOSGW
A bucket-based REST gateway, compatible with S3 and Swift
RBD
A reliable and fully-distributed block device, with a Linux kernel client and a QEMU/KVM driver
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

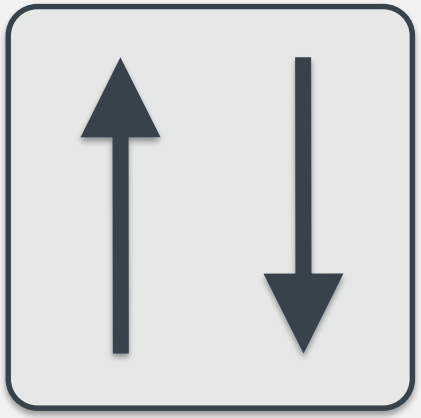

APP
LIBRADOS
socket
M
M
M

LIBRADOS

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

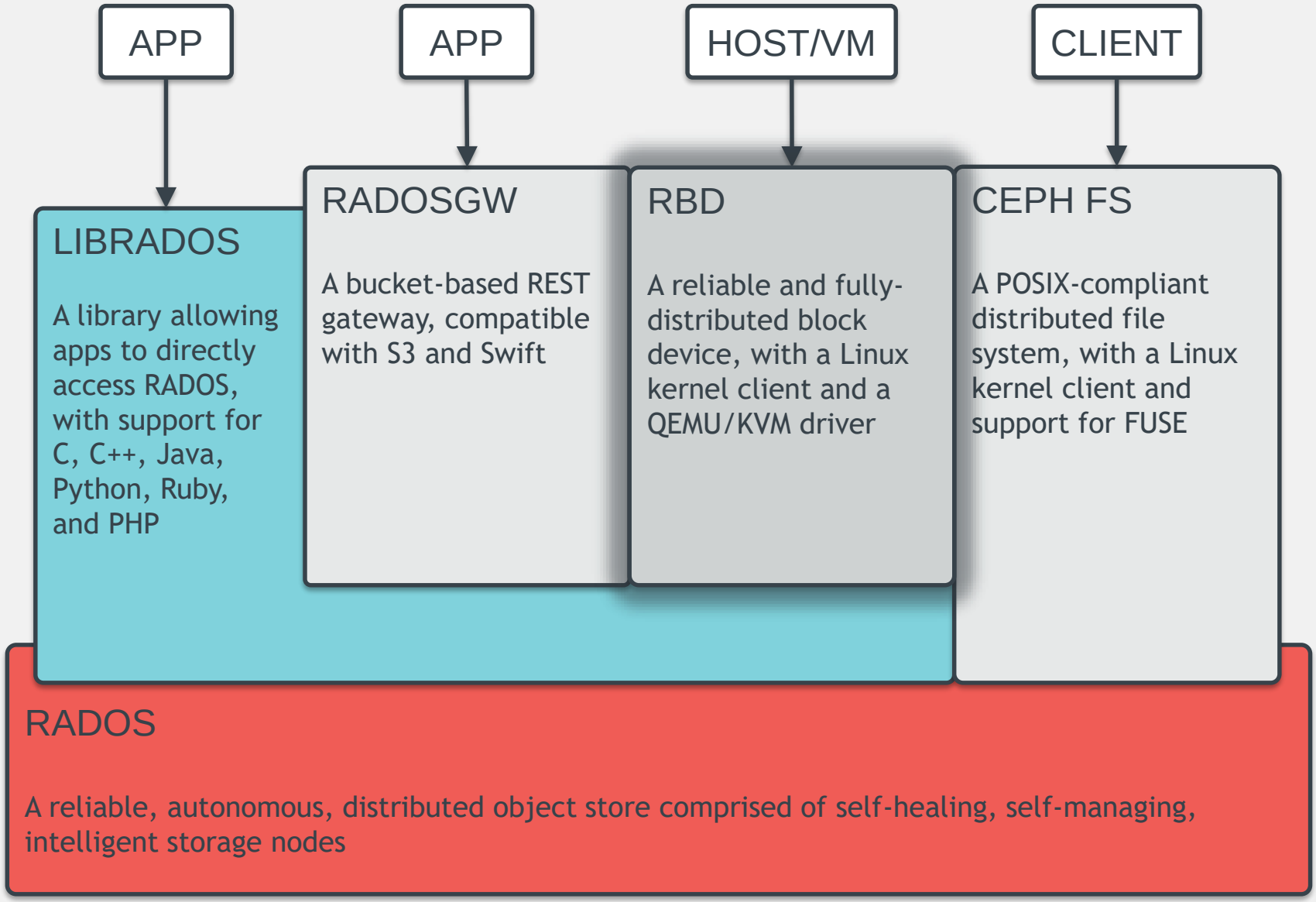

APP
APP
HOST/VM
CLIENT
LIBRADOS
A library allowing apps to directly access RADOS, with support for C, C++, Java, Python, Ruby, and PHP
RADOSGW
A bucket-based REST gateway, compatible with S3 and Swift
RBD
A reliable and fully-distributed block device, with a Linux kernel client and a QEMU/KVM driver
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

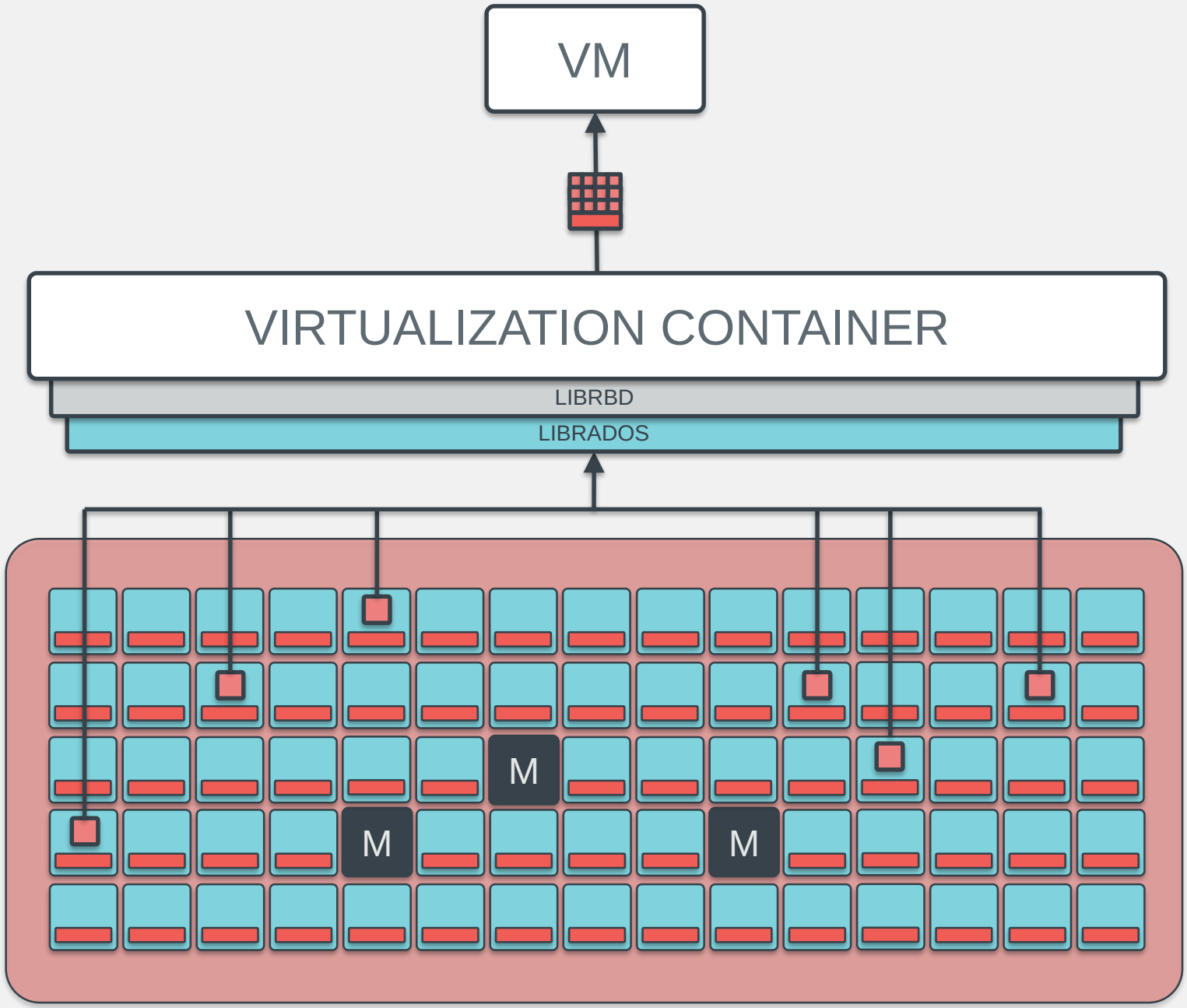

APP
REST
RADOSGW
LIBRADOS
socket
M
M
M

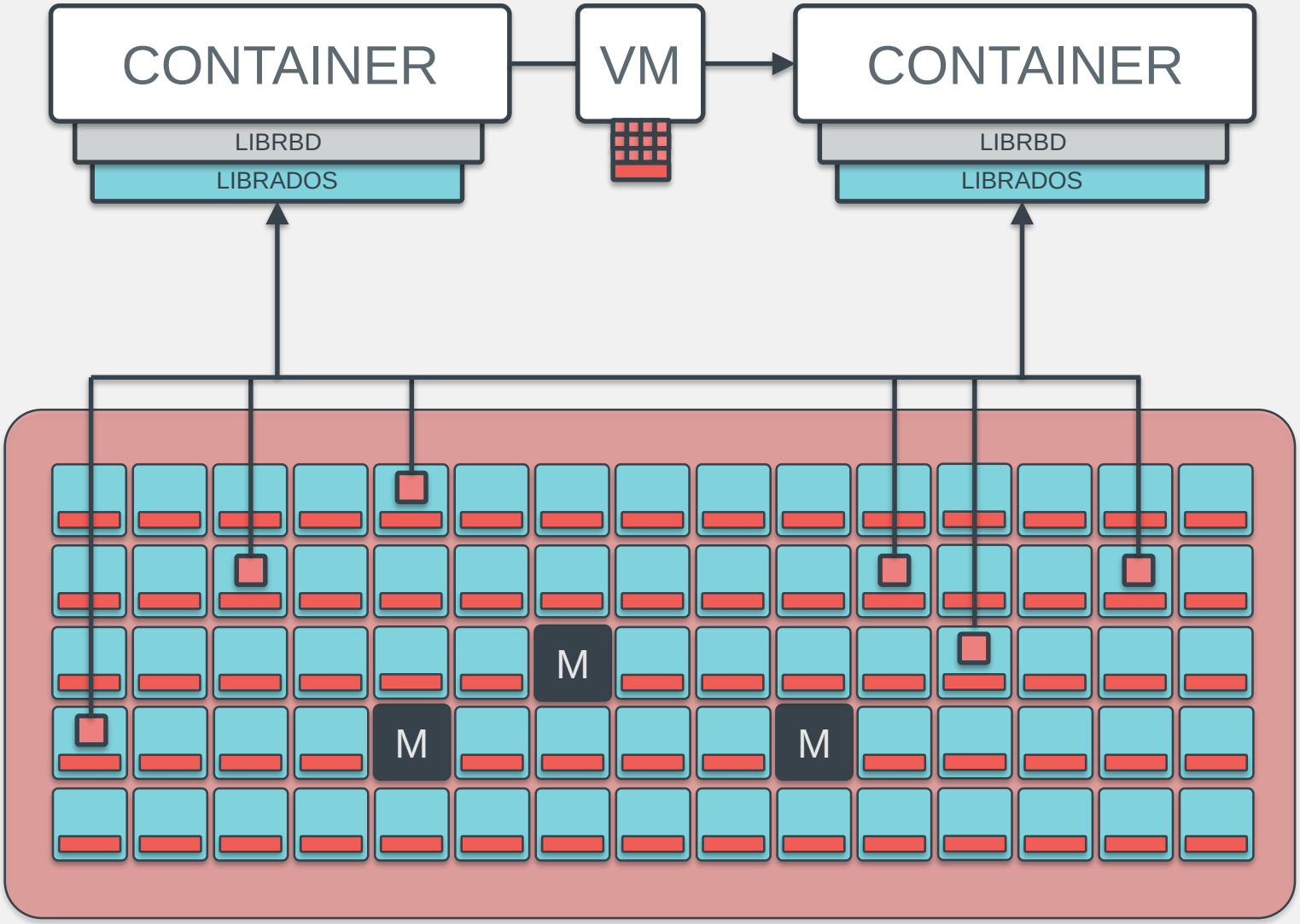


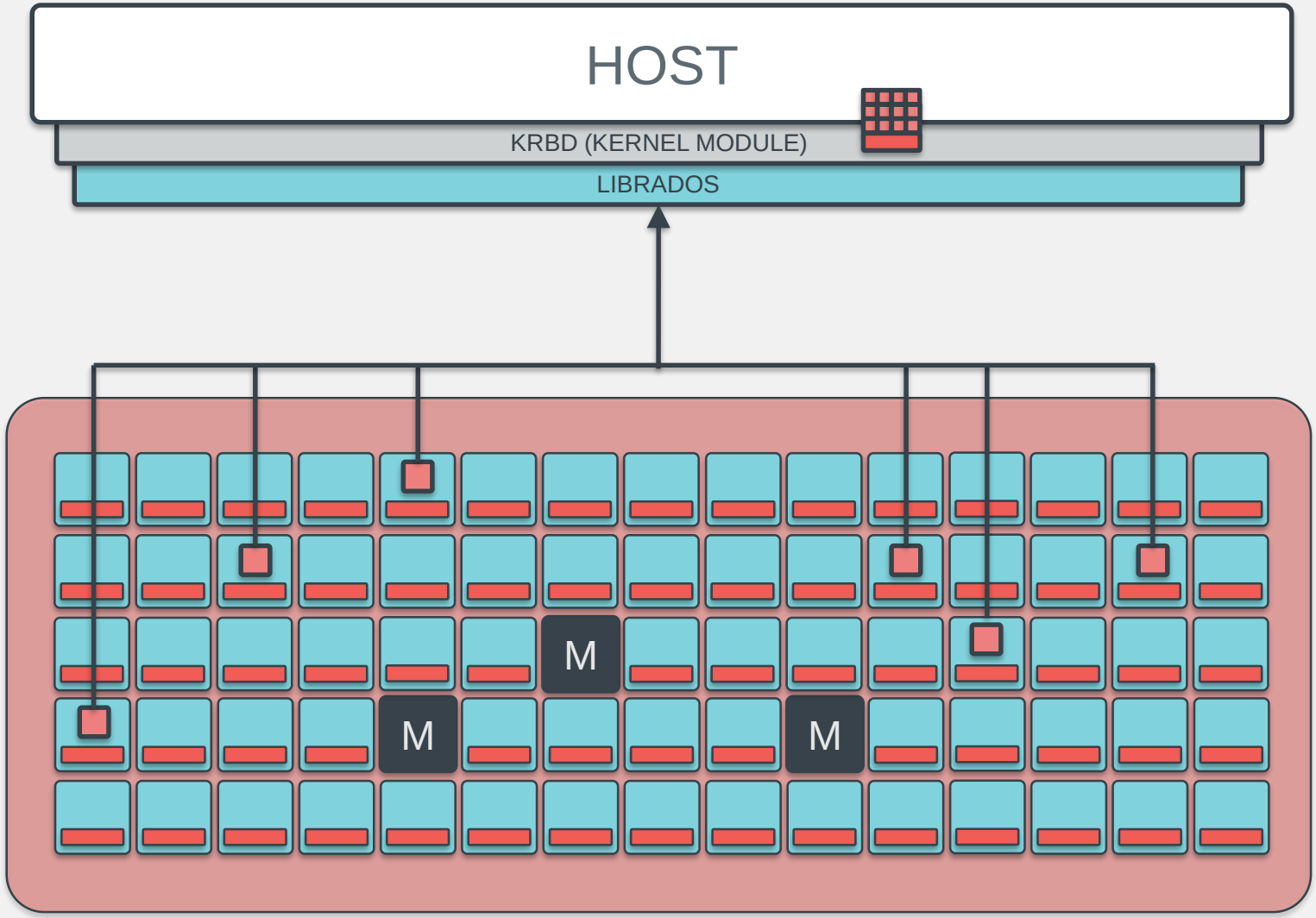
RADOS Gateway:

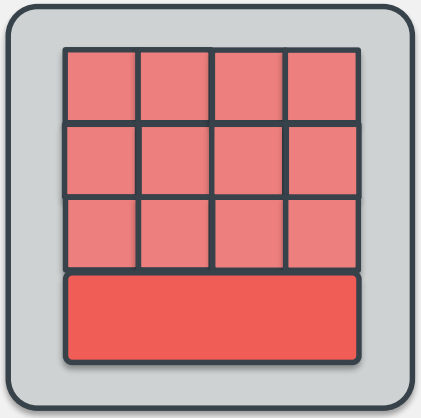
- REST-based object storage proxy
- Uses RADOS to store objects
- API supports buckets, accounts
- Usage accounting for billing
- Compatible with S3 and Swift applications


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

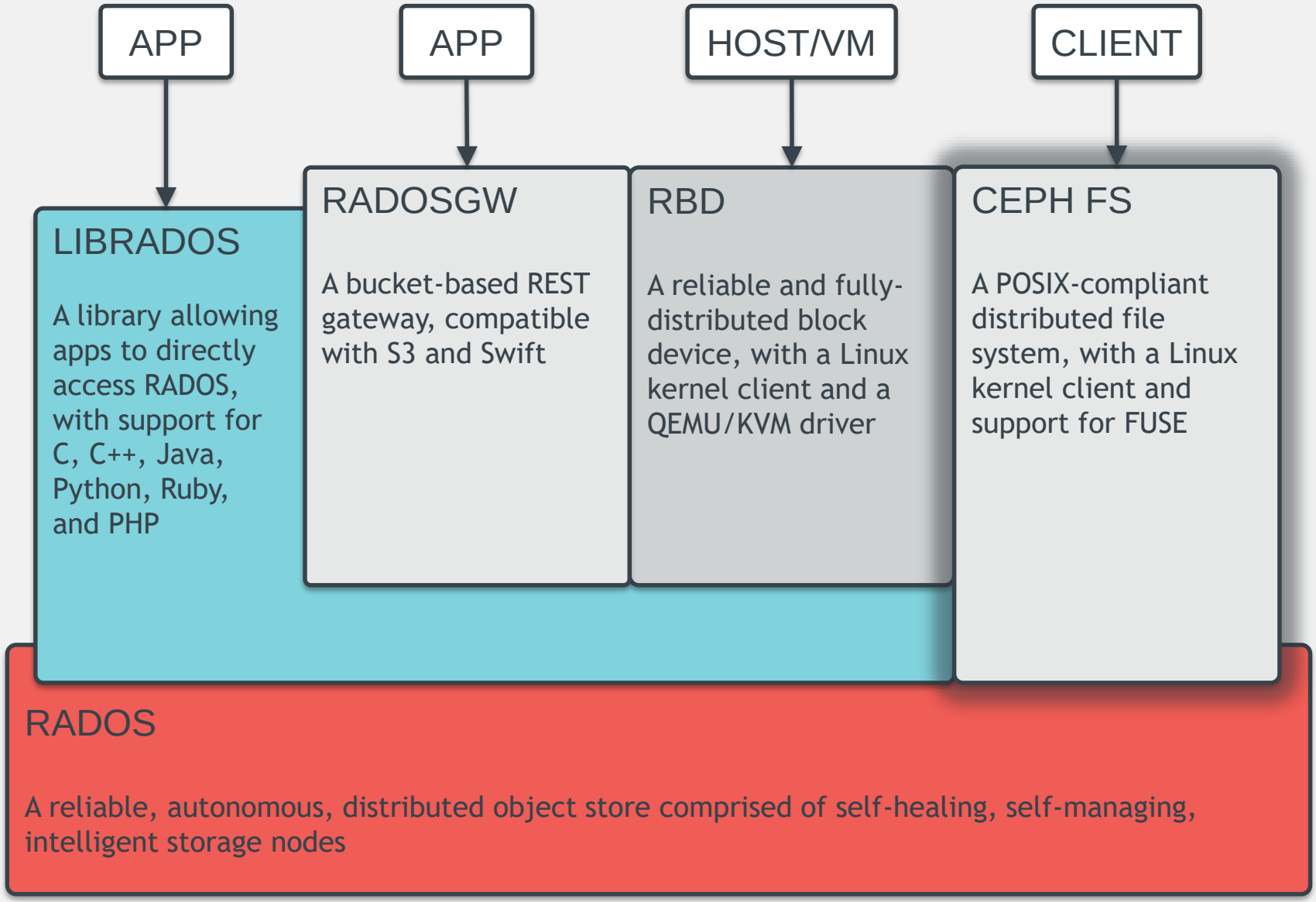

CONTAINER
LIBRBD
LIBRADOS
VM
CONTAINER
LIBRBD
LIBRADOS
M
M
M

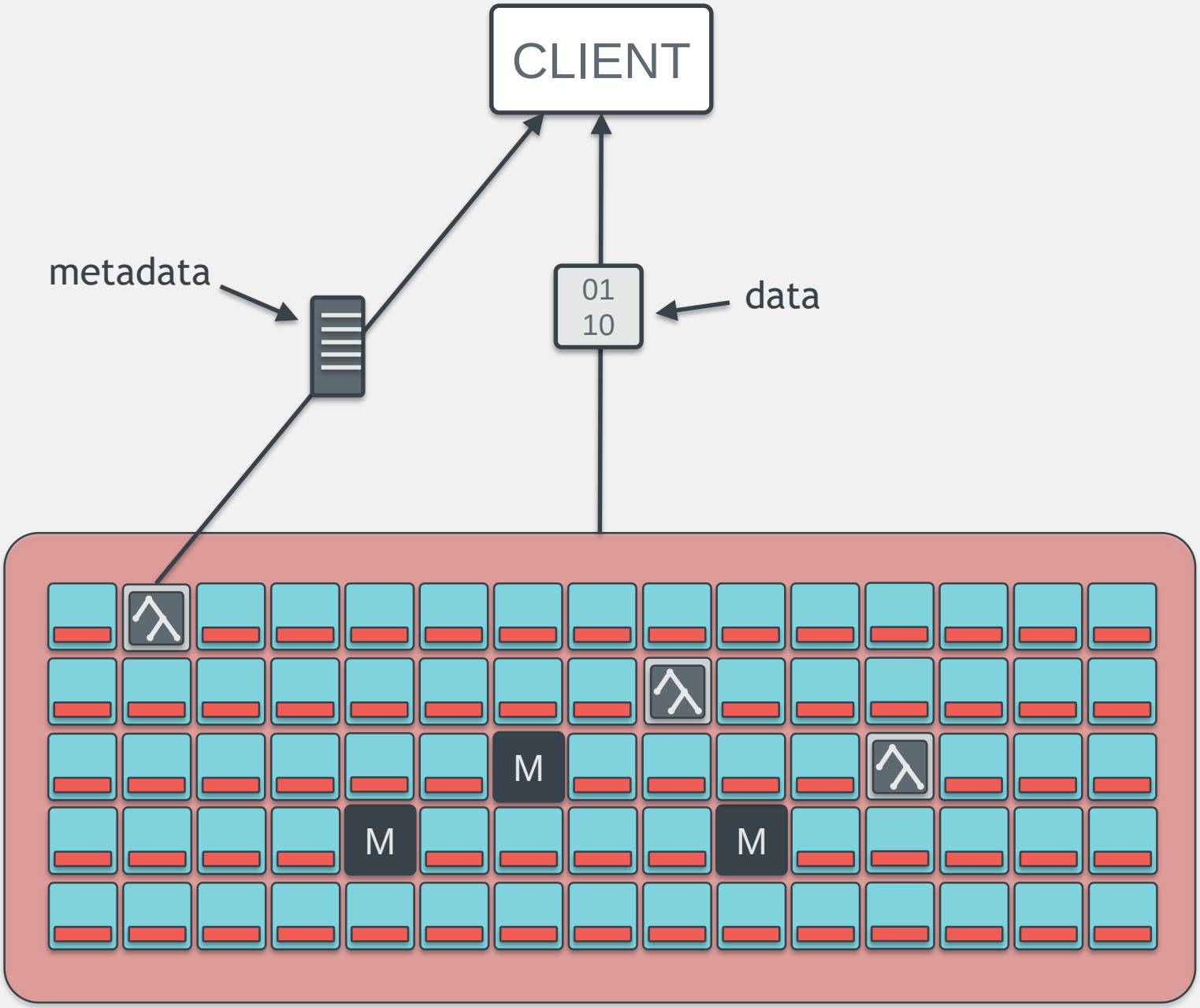

HOST
KRBD (KERNEL MODULE)
LIBRADOS
M
M
M

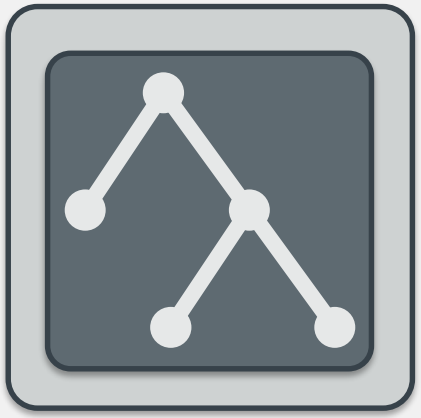


RADOS Block Device:

- 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, native Xen coming soon
  - OpenStack, CloudStack, Nebula, Proxmox


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


CLIENT
metadata
01
10
data
M
M
M

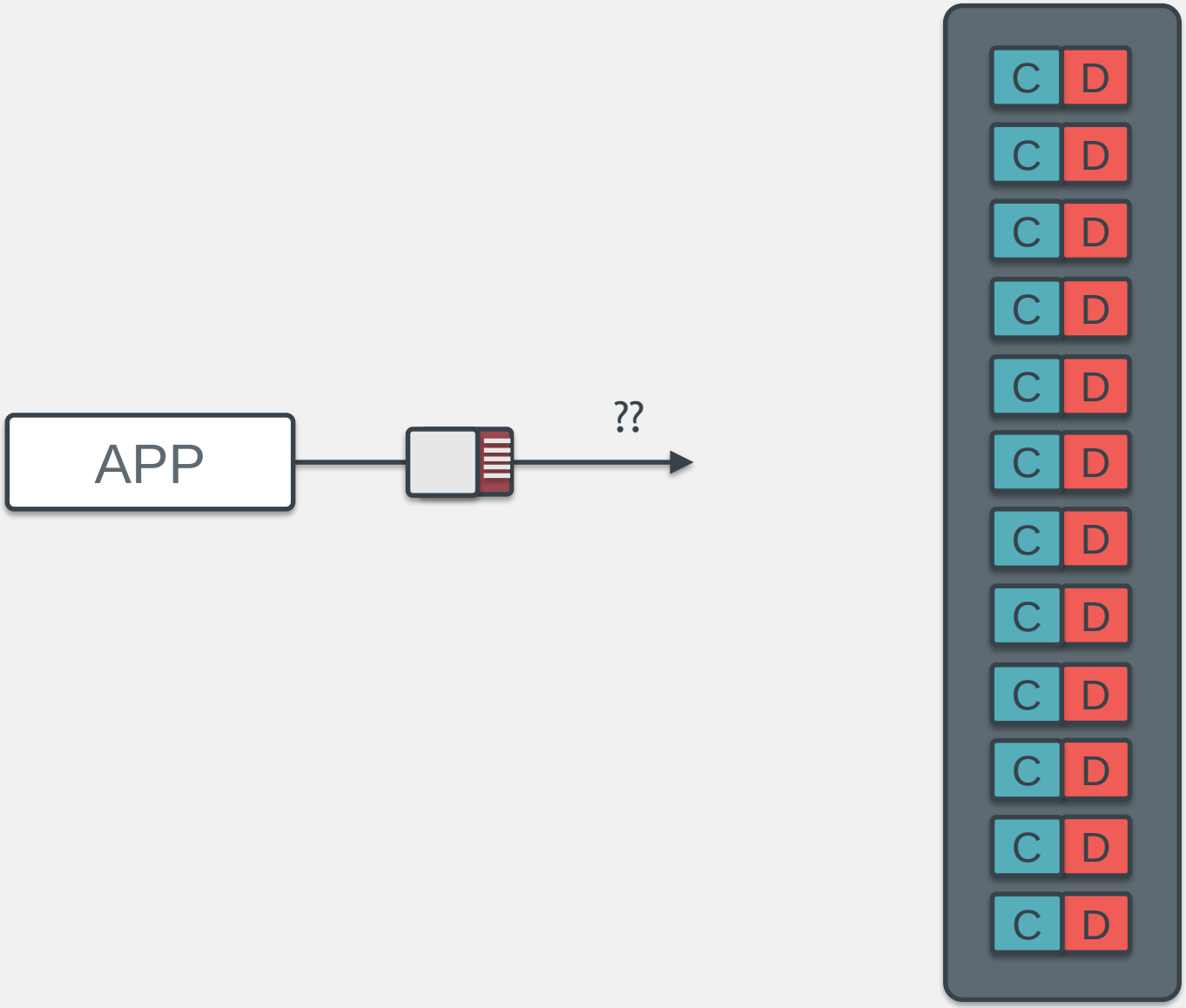


Metadata Server

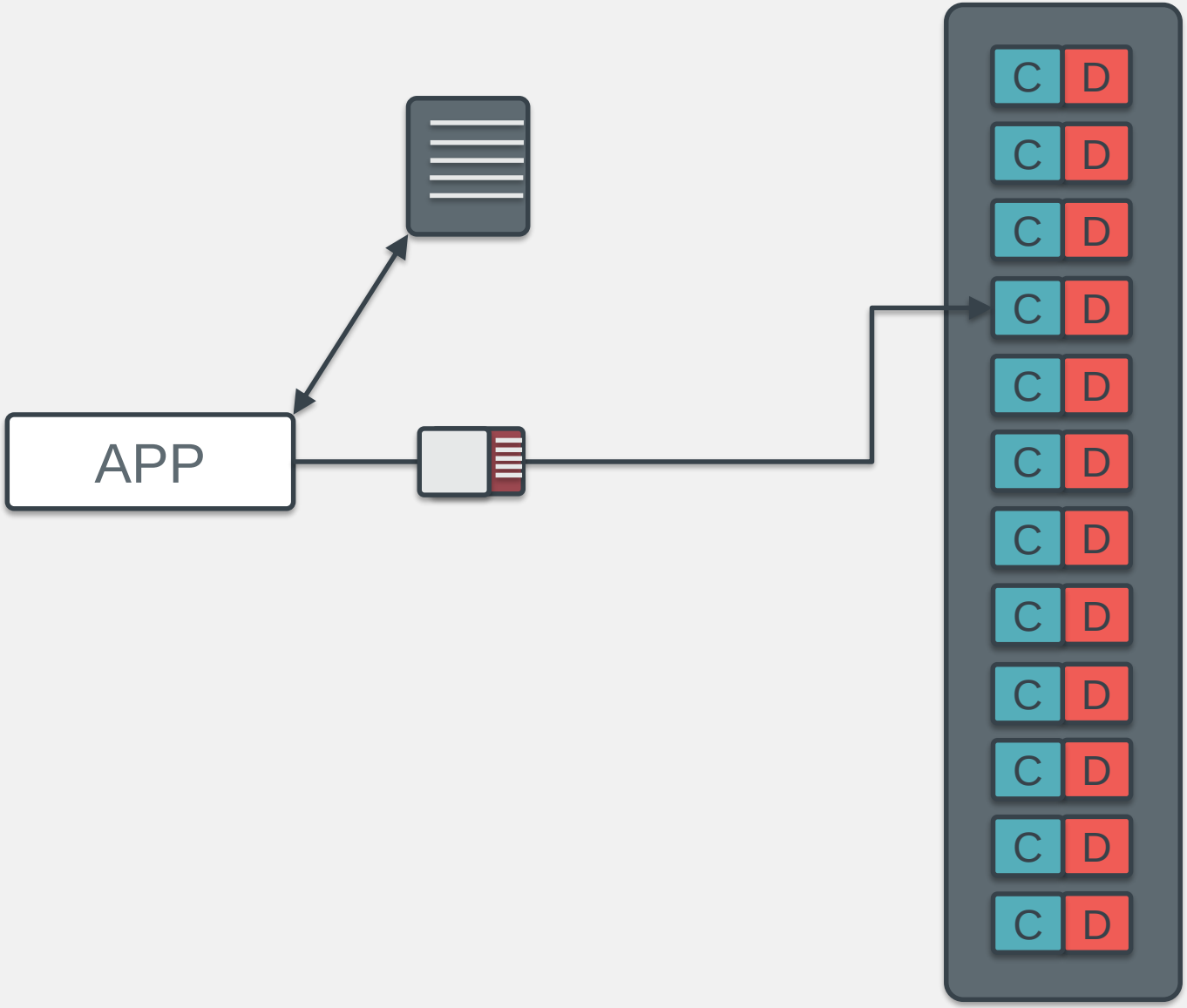
- Manages metadata for a POSIX-compliant shared filesystem
  - Directory hierarchy
  - File metadata (owner, timestamps, mode, etc.)
- Stores metadata in RADOS
- Does not serve file data to clients
- Only required for shared filesystem

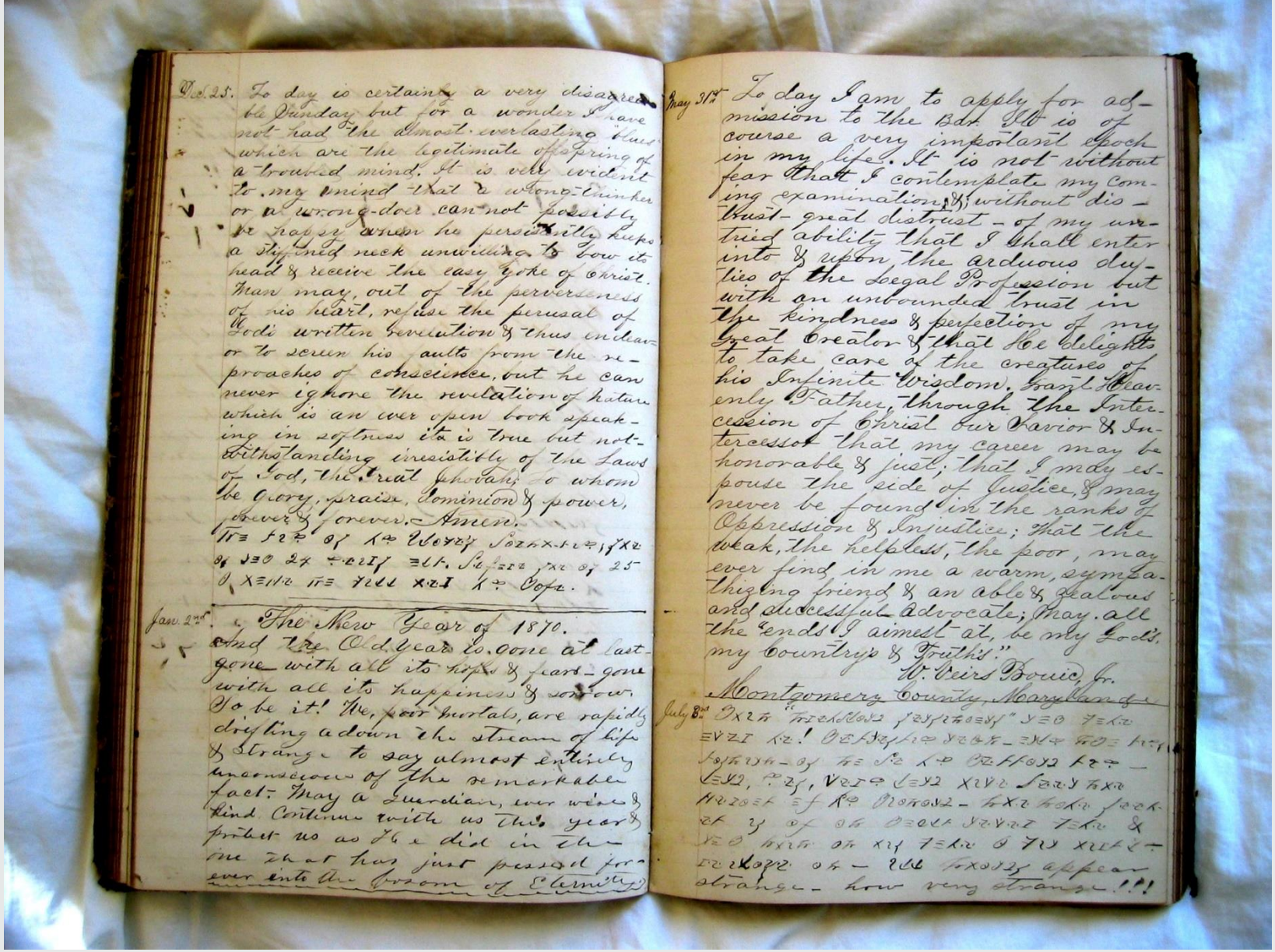
What Makes Ceph Unique?

Part one: CRUSH


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

How Long Did It Take You To Find Your Keys This Morning?


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



Dec. 25. To day is certainly a very disagreeable Sunday but for a wonder I have not had the almost everlasting blues which are the legitimate offspring of a troubled mind. It is very evident to my mind that a wrong-thinker or a wrong-doer can not possibly be happy when he persistently keeps a stiffened neck unwilling to bow its head & receive the easy yoke of Christ. Man may, out of the perverseness of his heart, refuse the perusal of God's written revelation & thus endeavor to screen his faults from the reproaches of conscience, but he can never ignore the revelation of Nature which is an ever open book speaking in softness it is true but notwithstanding irresistibly of the Laws of God, the Great Jehovah; to whom be Glory, praise, Dominion & power, forever & forever. Amen.

[illegible] 25

Jan. 2nd. The New Year of 1870. And the Old Year is gone at last — gone with all its hopes & fears — gone with all its happiness & sorrow. So be it! We, poor mortals, are rapidly drifting adown the stream of life & strange to say almost entirely unconscious of the remarkable fact. May a Guardian, ever wise & kind, continue with us this year & protect us as He did in the one that has just passed forever into the bosom of Eternity.

May 31st. To day I am to apply for admission to the Bar. It is of course a very important epoch in my life. It is not without fear that I contemplate my coming examination & without distrust — great distrust — of my untried ability that I shall enter into & upon the arduous duties of the Legal Profession but with an unbounded trust in the kindness & perfection of my Great Creator & that He delights to take care of the creatures of his Infinite Wisdom. Grant Heavenly Father, through the Intercession of Christ our Savior & Intercessor that my career may be honorable & just; that I may espouse the side of Justice, & may never be found in the ranks of Oppression & Injustice; that the weak, the helpless, the poor, may ever find in me a warm, sympathizing friend & an able & zealous and successful advocate; May all the Ends I aimest at, be my God's, my Country's & Truth's."

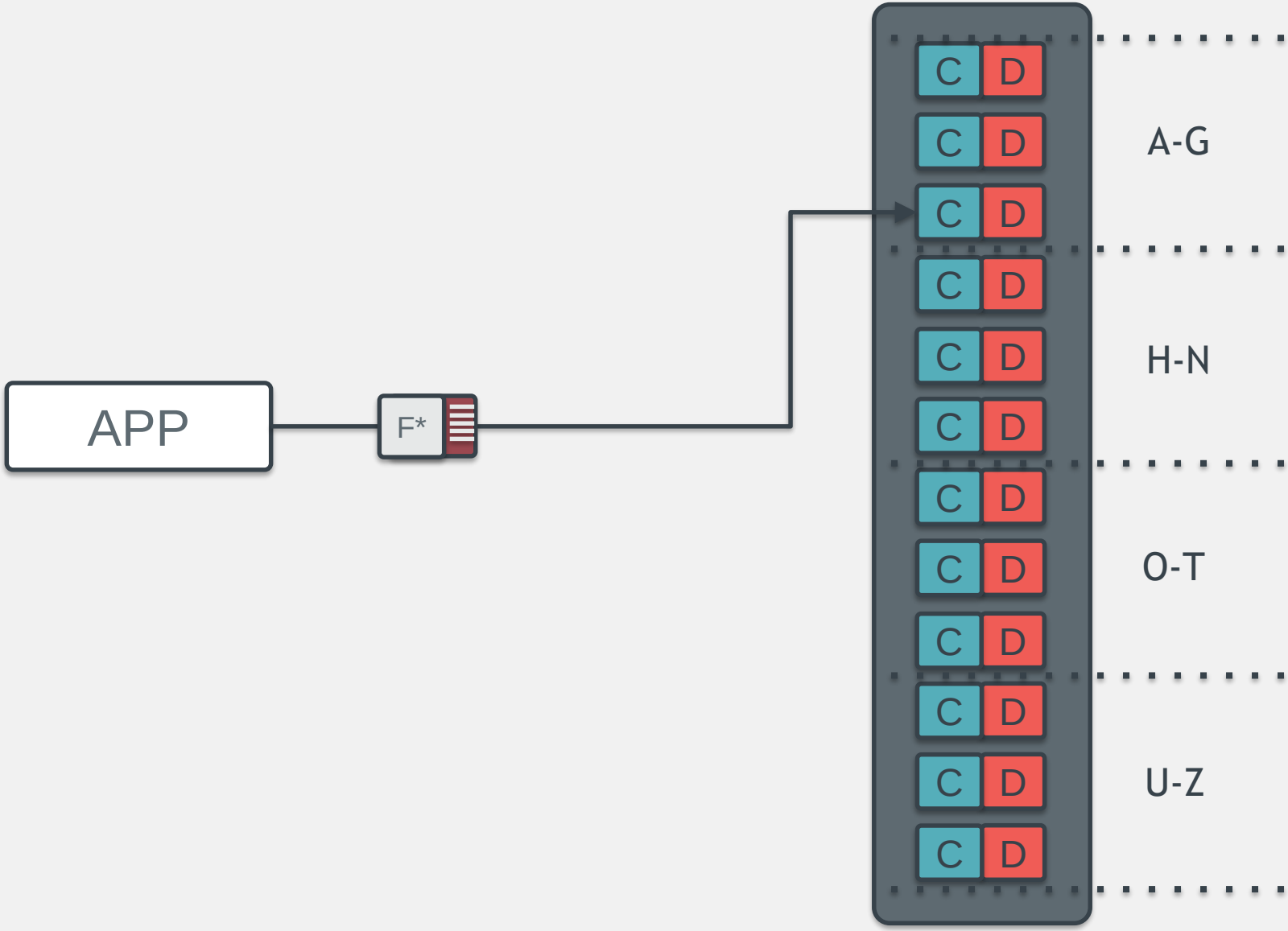
W. Peirs Bouic, Jr.

Montgomery County, Maryland

July 8th [illegible]

Source: Barnaby, Flickr / CC BY 2.0

Dear Diary: Today I Put My Keys on the Kitchen Counter


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

HOW DO YOU

FIND YOUR KEYS

WHEN YOUR HOUSE

IS

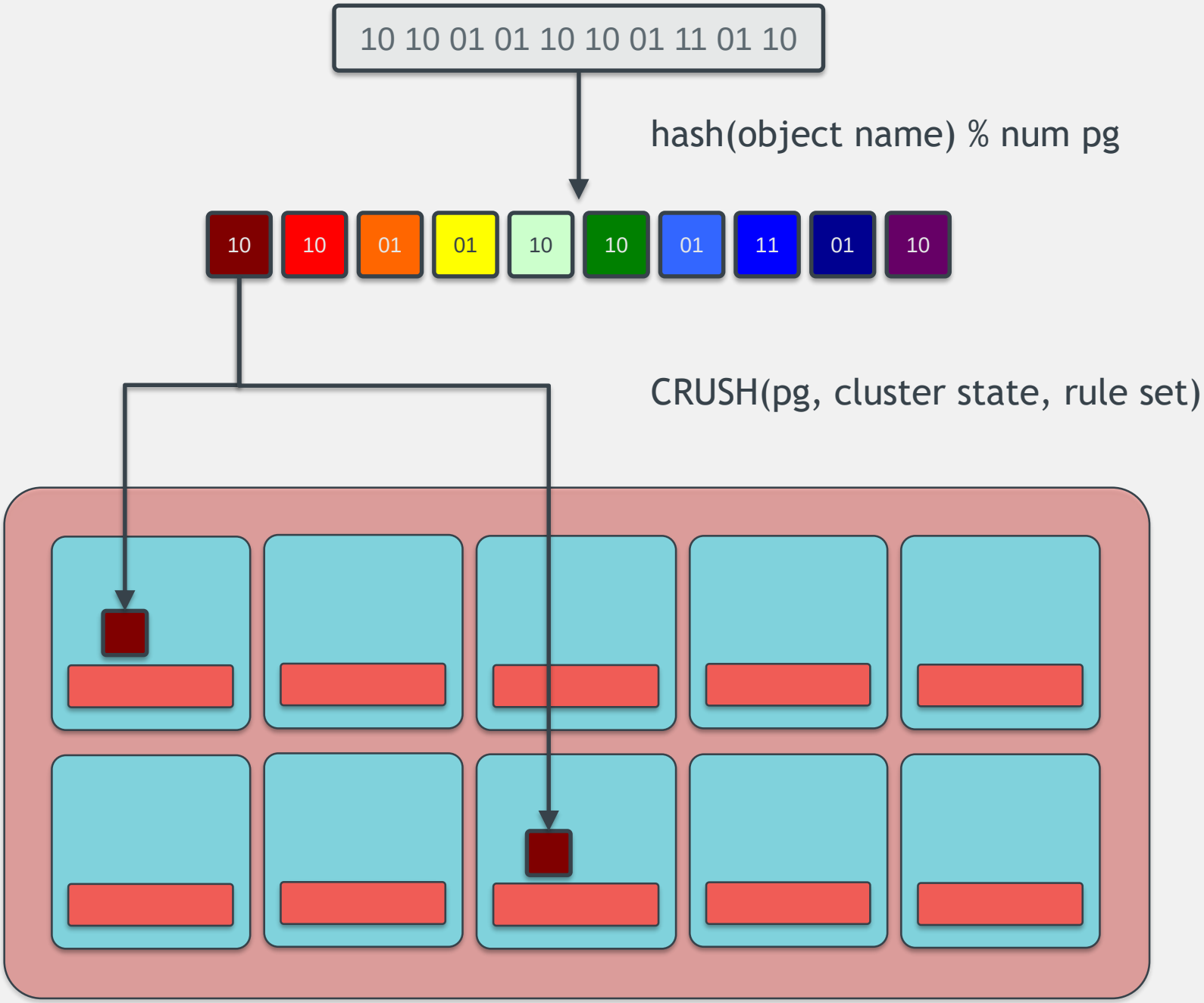
INFINITELY BIG

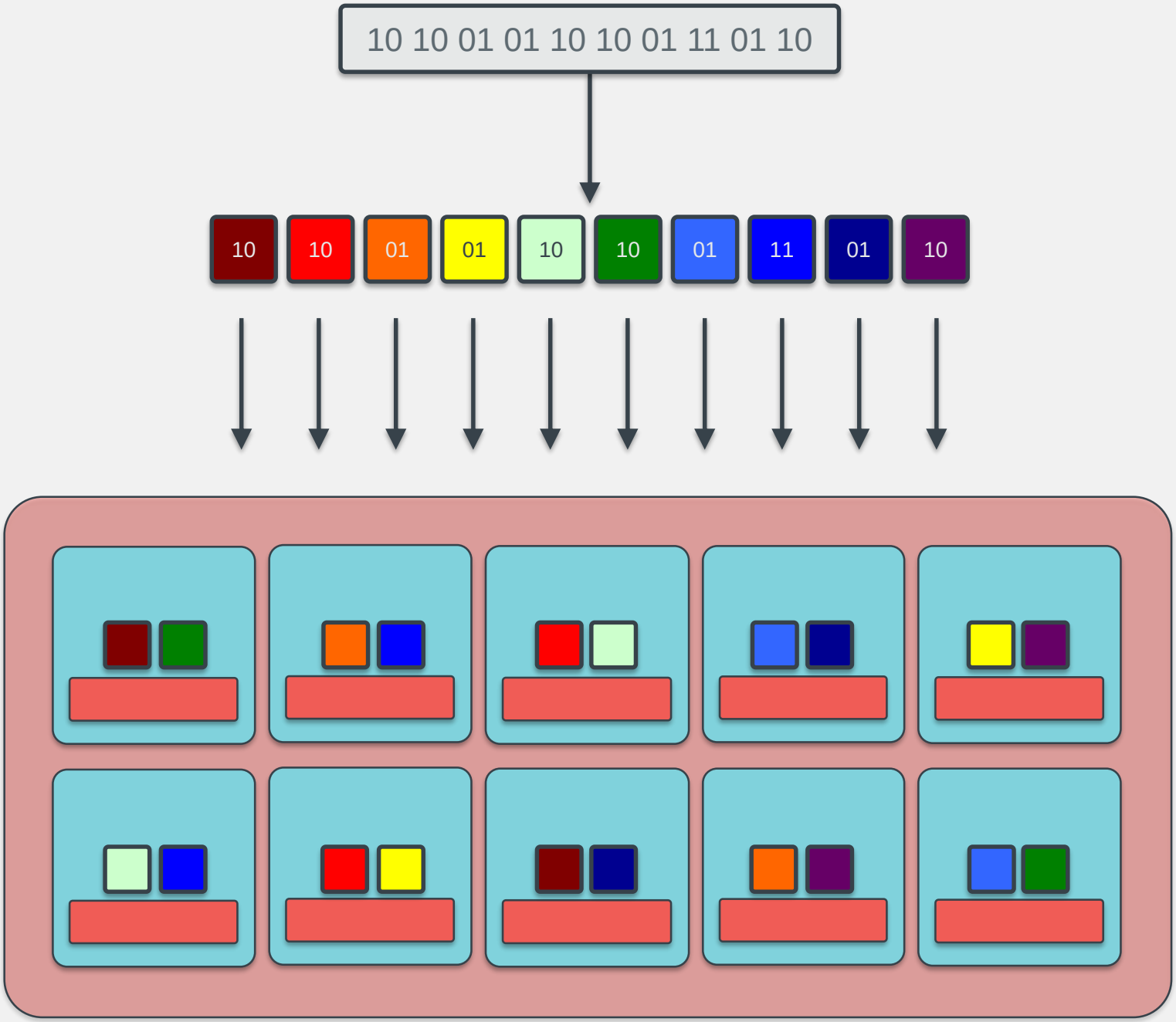
AND

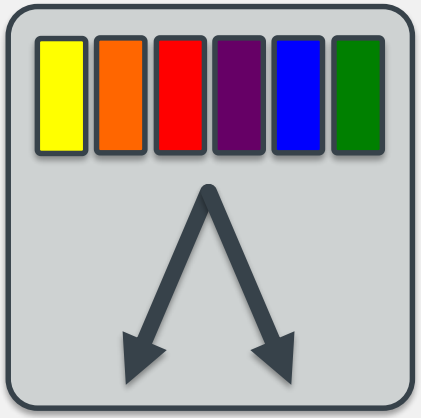
ALWAYS CHANGING?

The Answer: CRUSH!!!!!

pasukaru76, Flickr / CC SA 2.0

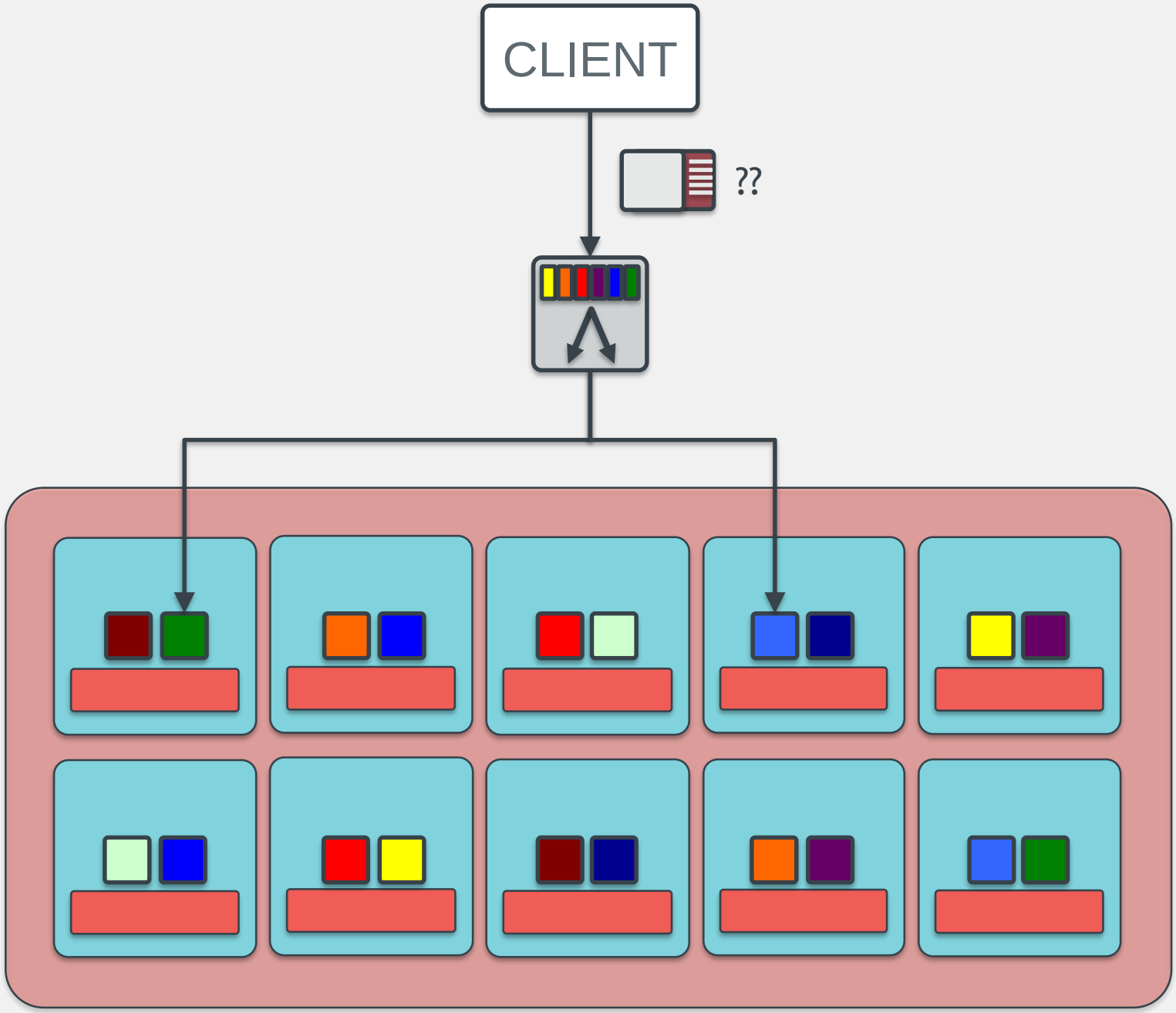

10 10 01 01 10 10 01 11 01 10
hash(object name) % num pg
10
10
01
01
10
10
01
11
01
10
CRUSH(pg, cluster state, rule set)

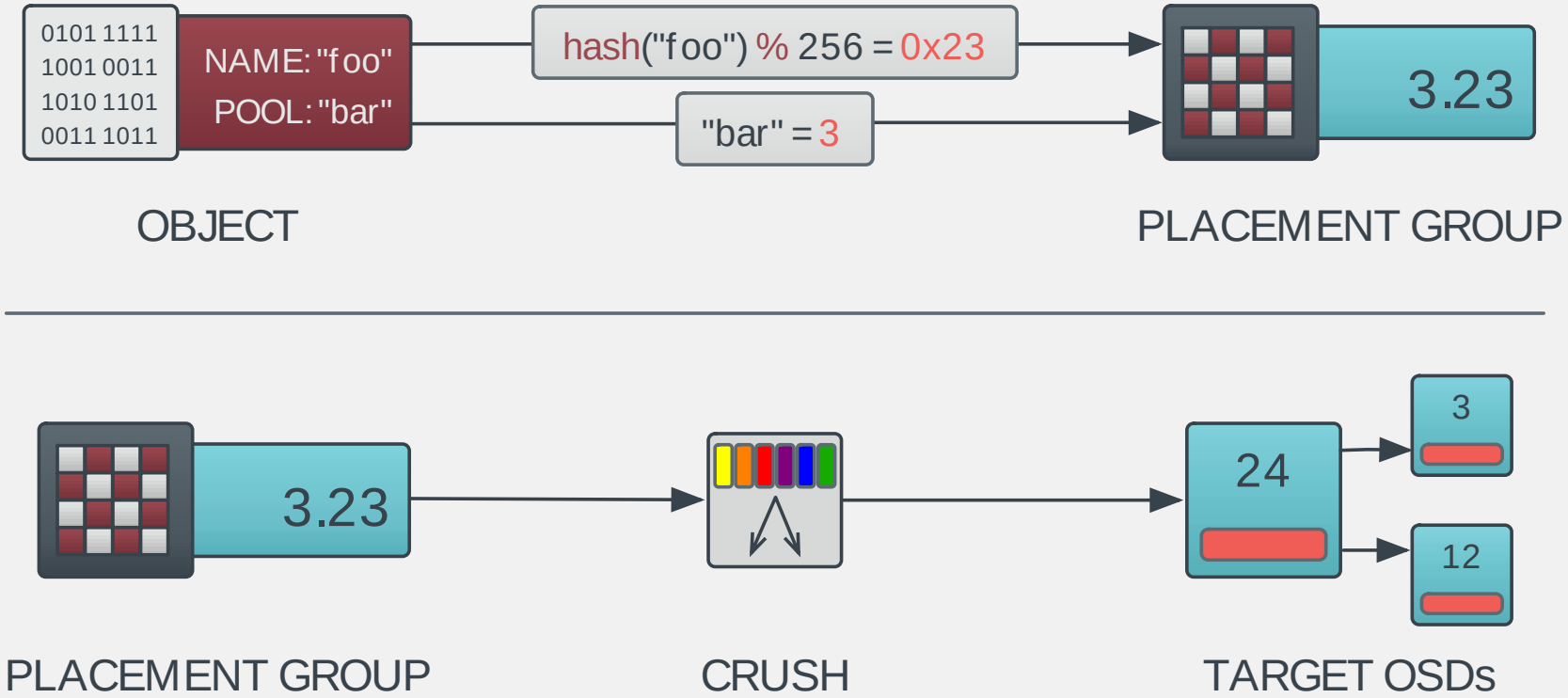

10 10 01 01 10 10 01 11 01 10
10
10
01
01
10
10
01
11
01
10

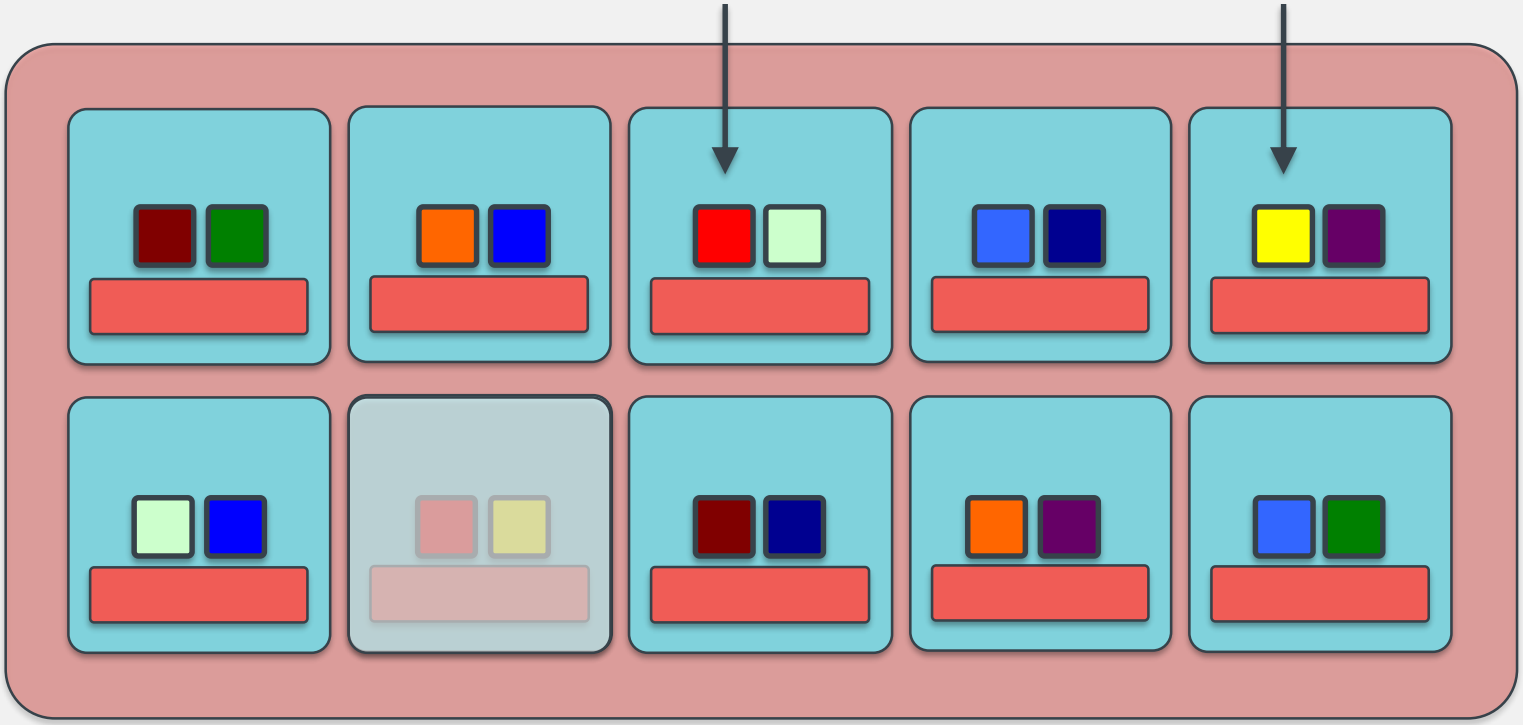


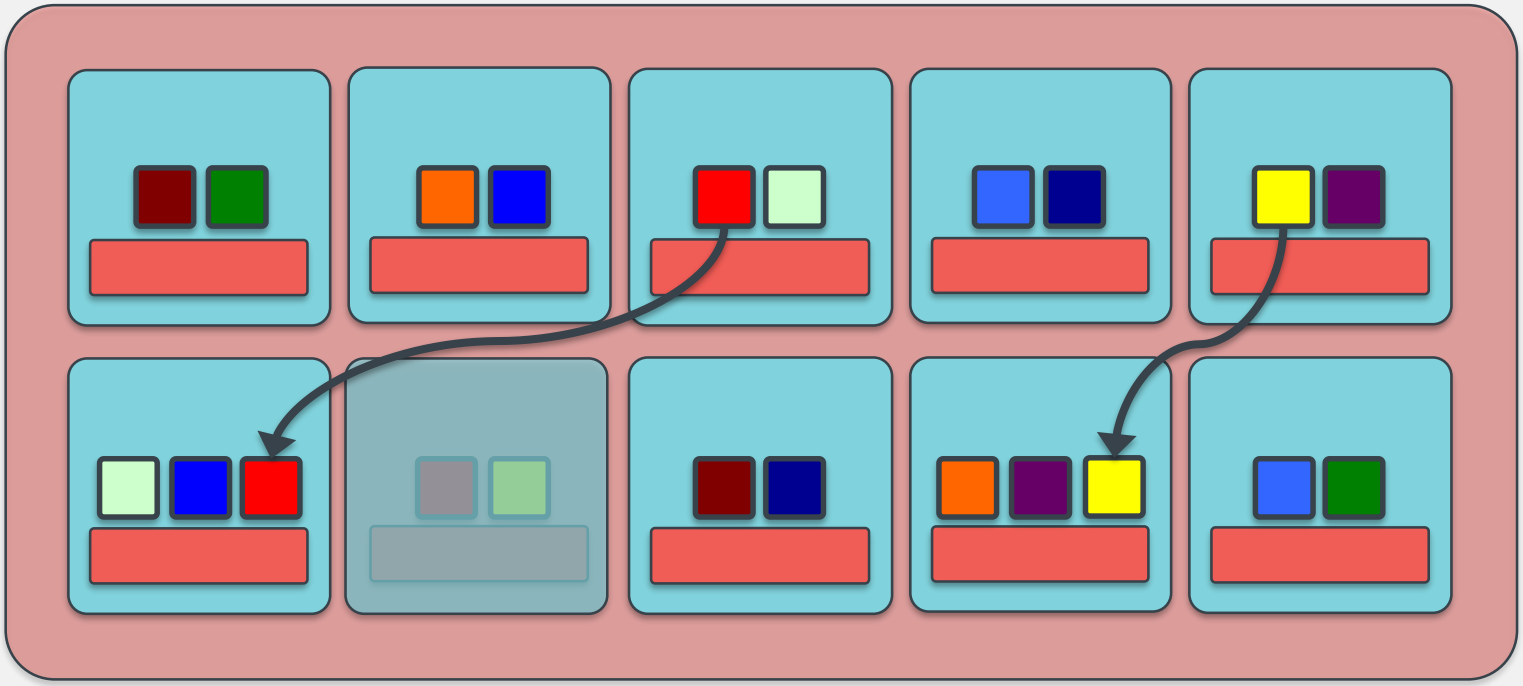
CRUSH

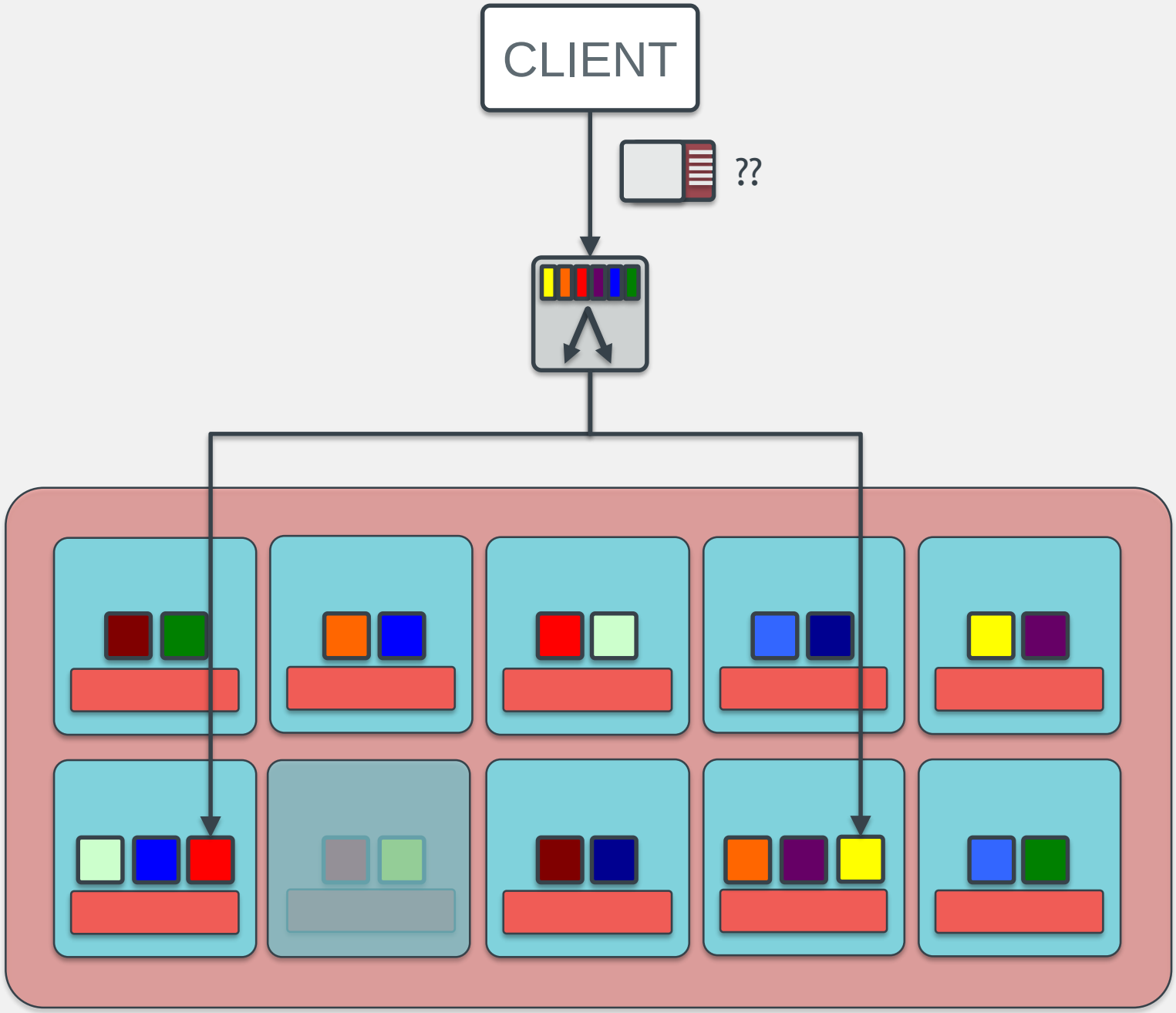
- Pseudo-random placement algorithm
  - Fast calculation, **no lookup**
  - Repeatable, deterministic
- Statistically uniform distribution
- Stable mapping
  - Limited data migration on change
- Rule-based configuration
  - Infrastructure topology aware
  - Adjustable replication
  - Weighting


CLIENT
??


0101 1111
1001 0011
1010 1101
0011 1011
NAME: "foo"
POOL: "bar"
hash("foo") % 256 = 0x23
"bar" = 3
3.23
OBJECT
PLACEMENT GROUP
3.23
24
3
12
PLACEMENT GROUP
CRUSH
TARGET OSDs

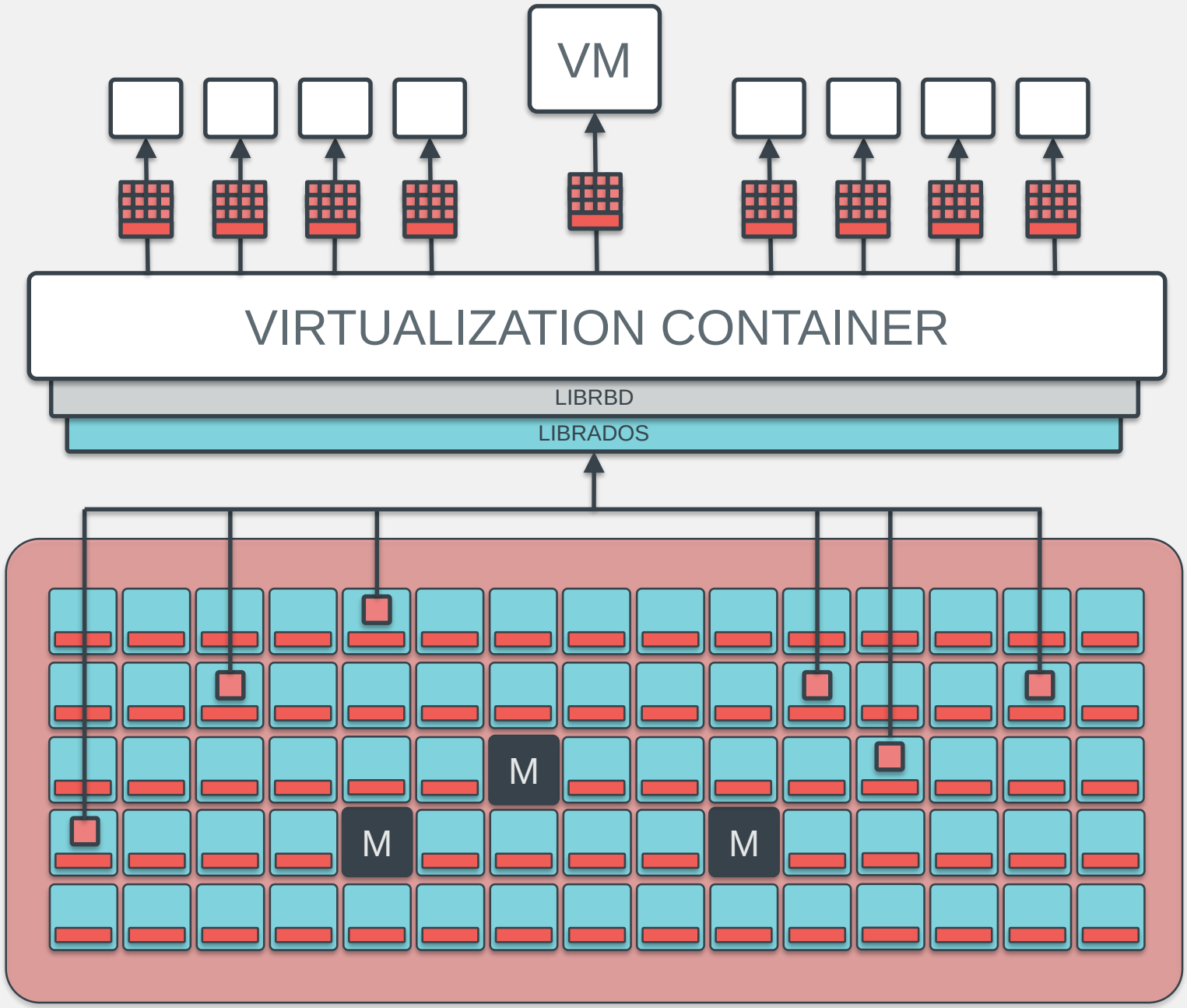





CLIENT
??

What Makes Ceph Unique

Part two: thin provisioning


VM
VIRTUALIZATION CONTAINER
LIBRBD
LIBRADOS
M
M
M

HOW DO YOU

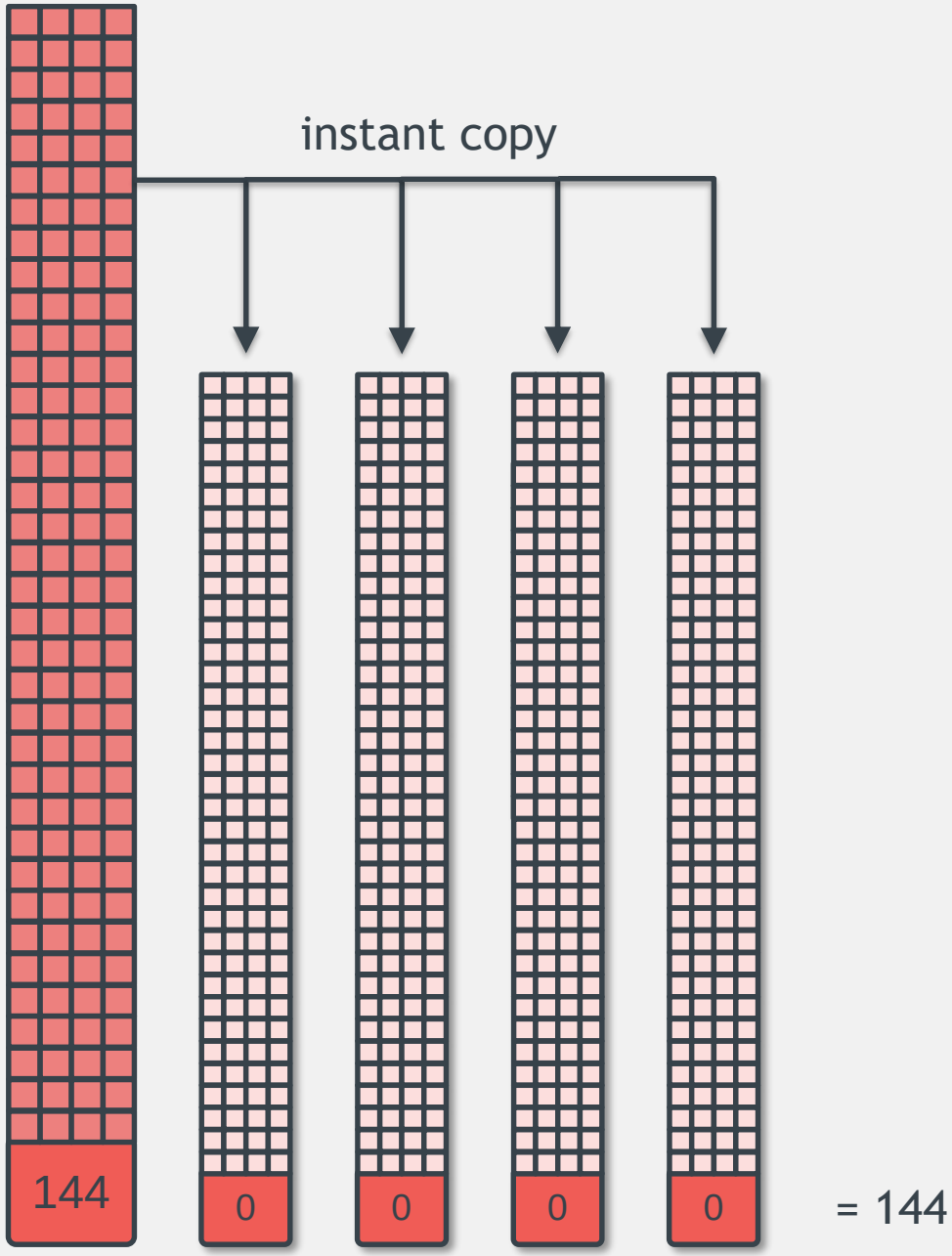
SPIN UP

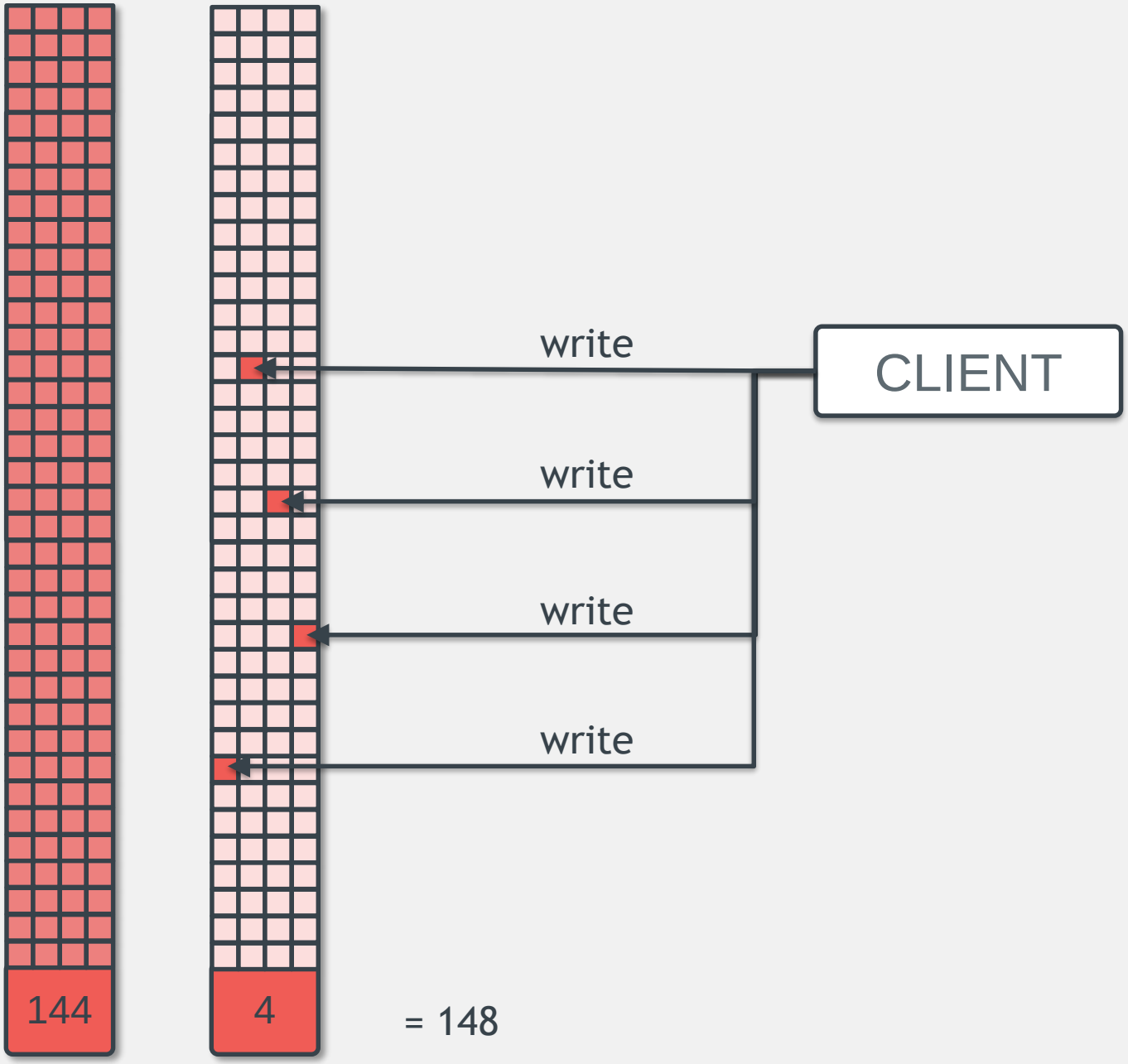
THOUSANDS OF VMs

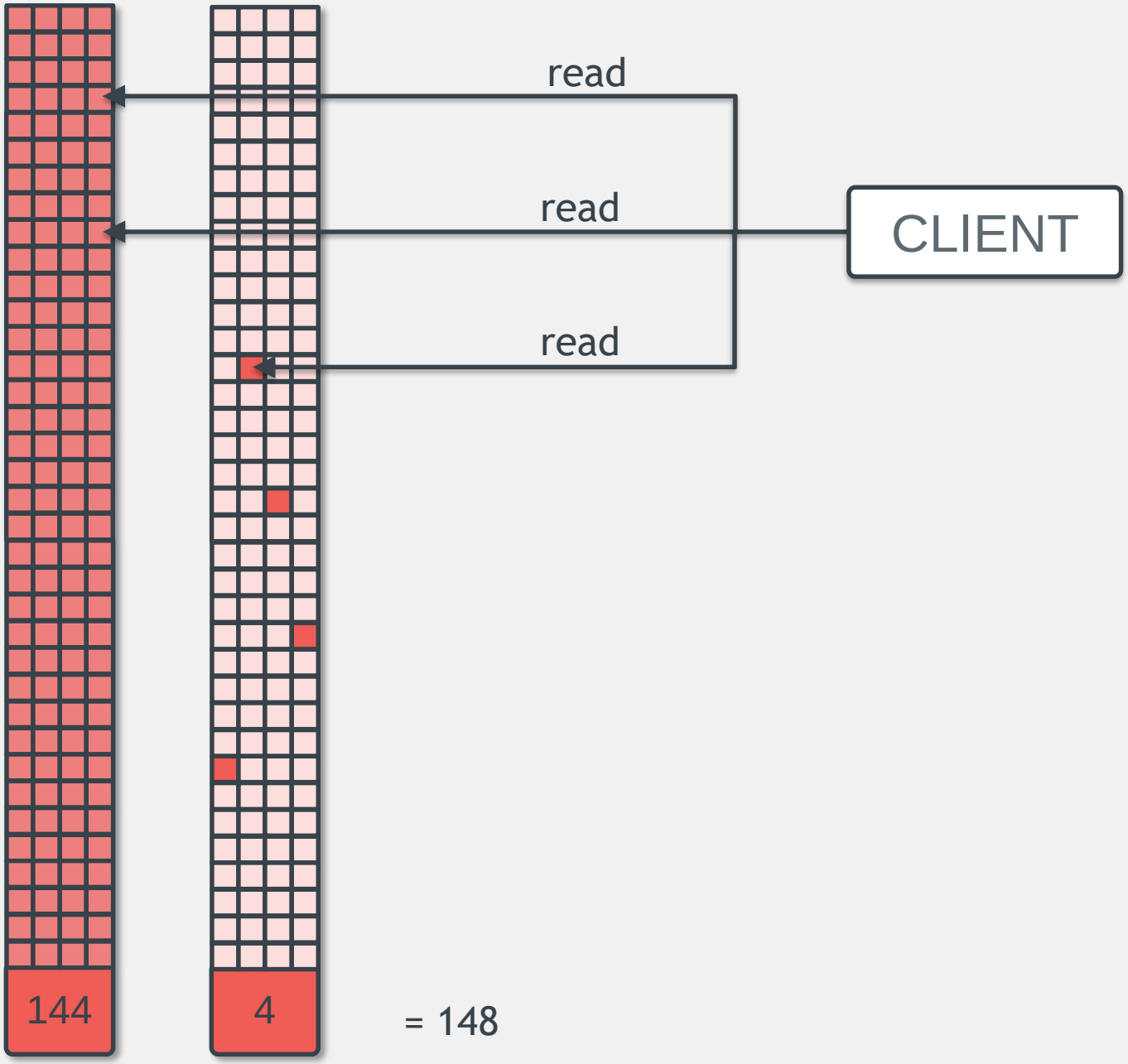
INSTANTLY

AND

EFFICIENTLY?


instant copy
144
0
0
0
0
= 144


write
CLIENT
write
write
write
144
4
= 148

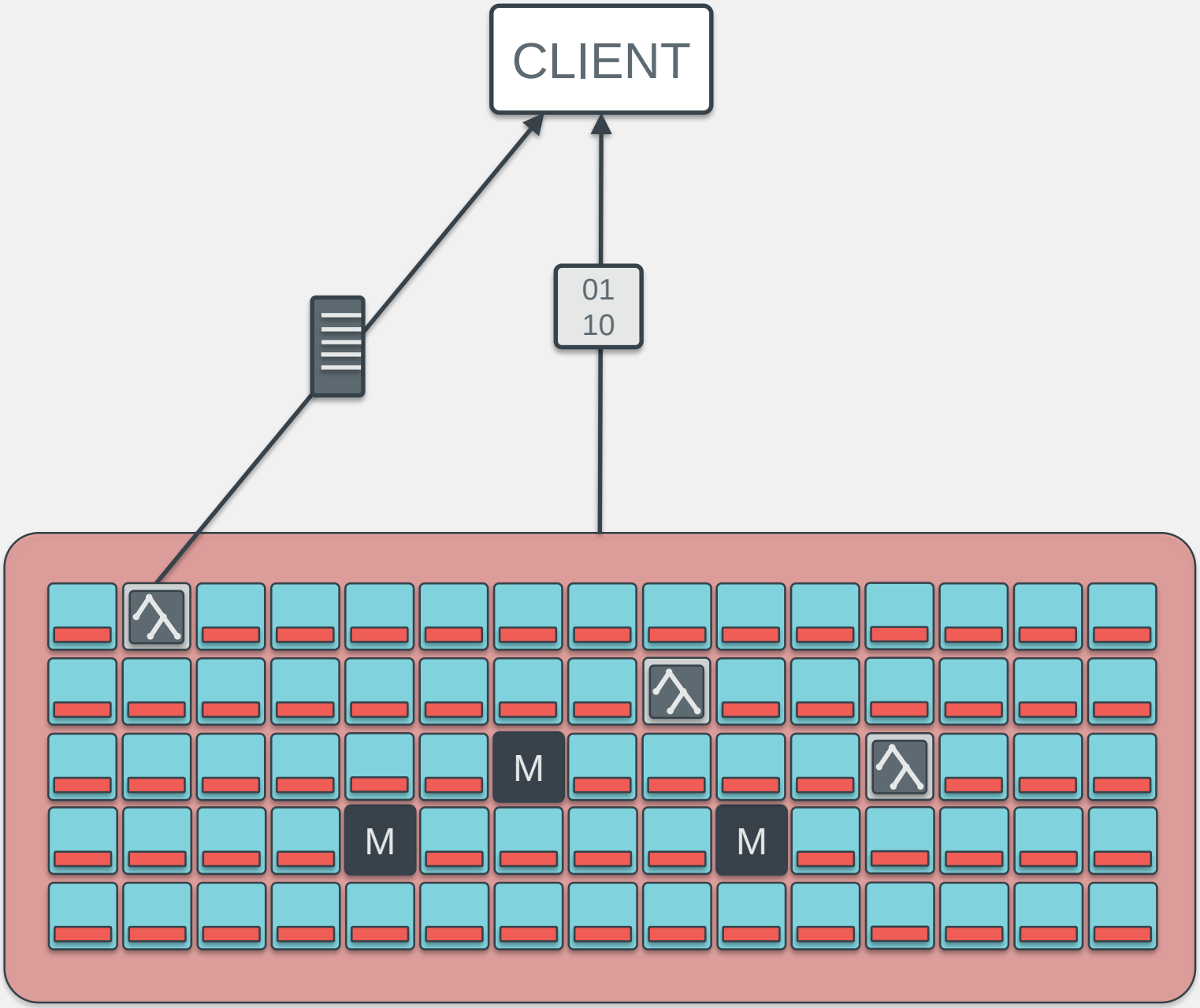

read
read
read
CLIENT
144
4
= 148

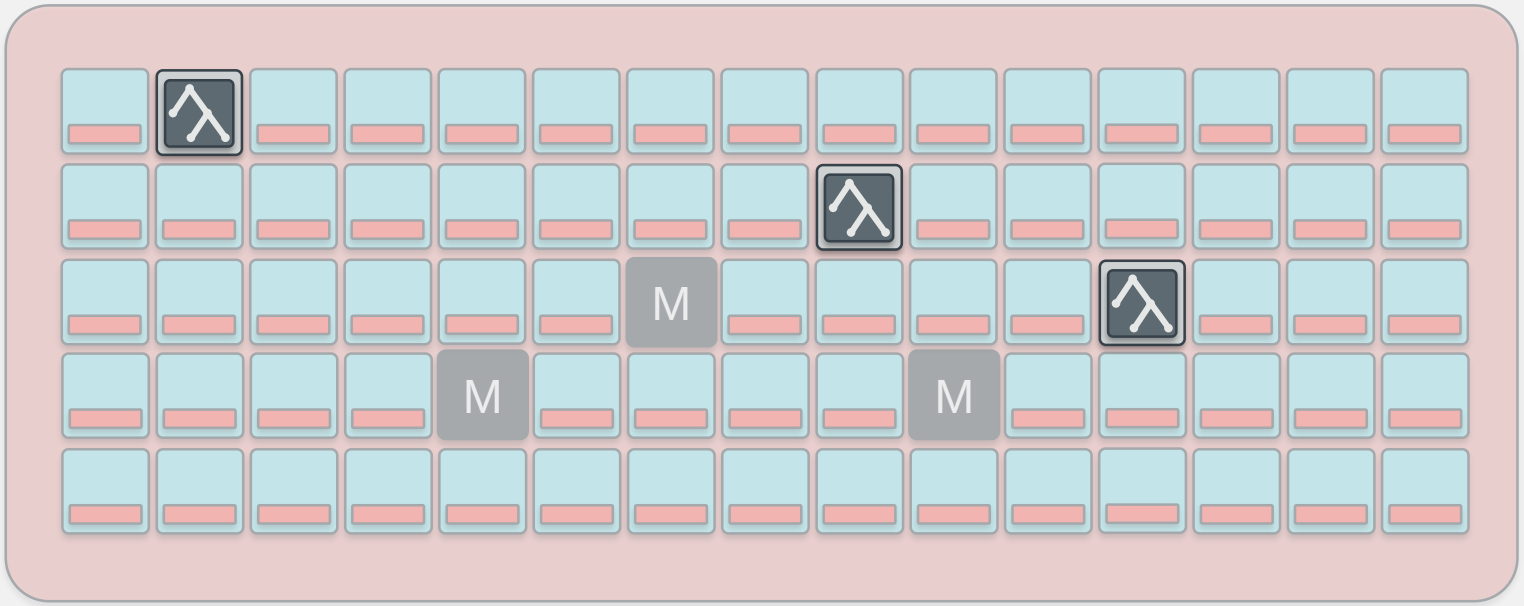
What Makes Ceph Unique?

Part three: clustered metadata

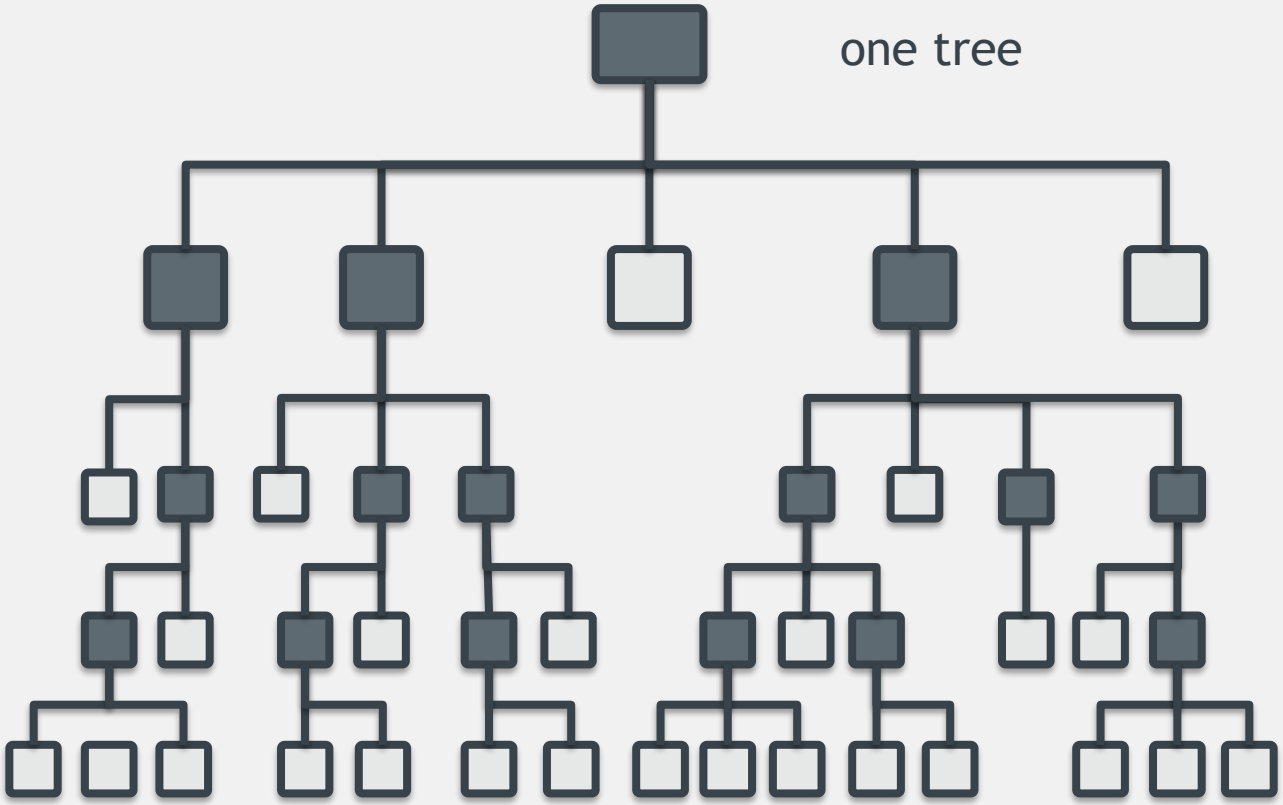
```
lrwxrwxrwx  1 root root       26 Apr 26 17:54 libgssapi_krb5.so -> mit-krb5/libgssapi_krb5.so
lrwxrwxrwx  1 root root       21 Apr 26 17:55 libgssapi_krb5.so.2 -> libgssapi_krb5.so.2.2
-rw-r--r-- 50 root root   216824 Jul 31  2012 libgssapi_krb5.so.2.2
lrwxrwxrwx  1 root root       13 Apr 26 17:54 libgs.so.8 -> libgs.so.8.71
-rw-r--r-- 17 root root  9478048 Jan 25  2011 libgs.so.8.71
lrwxrwxrwx  1 root root       21 Apr 26 17:55 libgssrpc.so -> mit-krb5/libgssrpc.so
lrwxrwxrwx  1 root root       16 Apr 26 17:55 libgssrpc.so.4 -> libgssrpc.so.4.1
-rw-r--r-- 50 root root   115352 Jul 31  2012 libgssrpc.so.4.1
-rw-r--r-- 50 root root    21832 Sep  8  2010 libgthread-2.0.a
-rw-r--r-- 50 root root      972 Sep  8  2010 libgthread-2.0.la
lrwxrwxrwx  1 root root       26 Apr 26 17:55 libgthread-2.0.so -> libgthread-2.0.so.0.2400.2
lrwxrwxrwx  1 root root       26 Apr 26 17:55 libgthread-2.0.so.0 -> libgthread-2.0.so.0.2400.2
-rw-r--r-- 50 root root    17704 Sep  8  2010 libgthread-2.0.so.0.2400.2
drwxr-xr-x  2 root root     4096 Apr 26 18:00 libgtk2.0-0
-rw-r--r-- 49 root root  9275282 Oct 14  2010 libgtk-x11-2.0.a
-rw-r--r-- 49 root root      981 Oct 14  2010 libgtk-x11-2.0.la
lrwxrwxrwx  1 root root       26 Apr 26 17:55 libgtk-x11-2.0.so -> libgtk-x11-2.0.so.0.2000.1
lrwxrwxrwx  1 root root       26 Apr 26 17:55 libgtk-x11-2.0.so.0 -> libgtk-x11-2.0.so.0.2000.1
-rw-r--r-- 49 root root  4319784 Oct 14  2010 libgtk-x11-2.0.so.0.2000.1
lrwxrwxrwx  1 root root       15 Apr 26 17:55 libgvc.so -> libgvc.so.5.0.0
lrwxrwxrwx  1 root root       15 Apr 26 17:55 libgvc.so.5 -> libgvc.so.5.0.0
-rw-r--r-- 49 root root   504424 Jul  5  2010 libgvc.so.5.0.0
lrwxrwxrwx  1 root root       16 Apr 26 17:55 libgvpr.so -> libgvpr.so.1.0.0
lrwxrwxrwx  1 root root       16 Apr 26 17:55 libgvpr.so.1 -> libgvpr.so.1.0.0
-rw-r--r-- 50 root root   482856 Jul  5  2010 libgvpr.so.1.0.0
-rw-r--r-- 50 root root   267948 Apr 13  2009 libHalf.a
lrwxrwxrwx  1 root root       16 Apr 26 17:55 libHalf.so -> libHalf.so.6.0.0
lrwxrwxrwx  1 root root       16 Apr 26 17:55 libHalf.so.6 -> libHalf.so.6.0.0
-rw-r--r-- 50 root root   269992 Apr 13  2009 libHalf.so.6.0.0
-rw-r--r-- 50 root root    52850 Nov  1  2009 libhistory.a
```

POSIX Filesystem Metadata

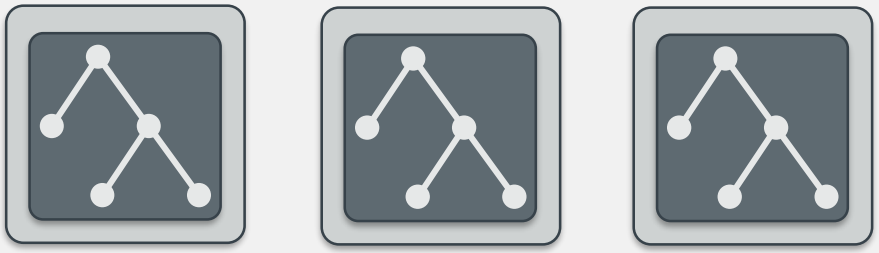

CLIENT
01
10
M
M
M


M
M
M

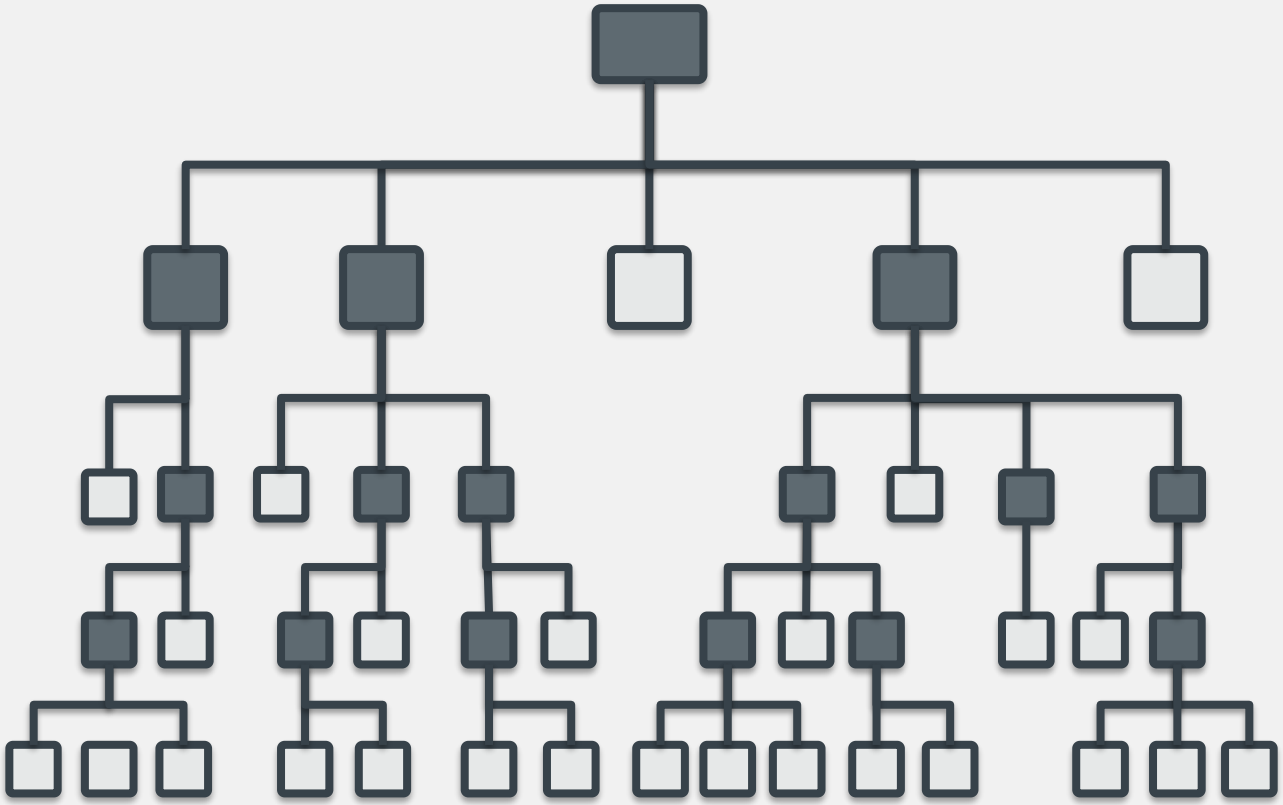
one tree

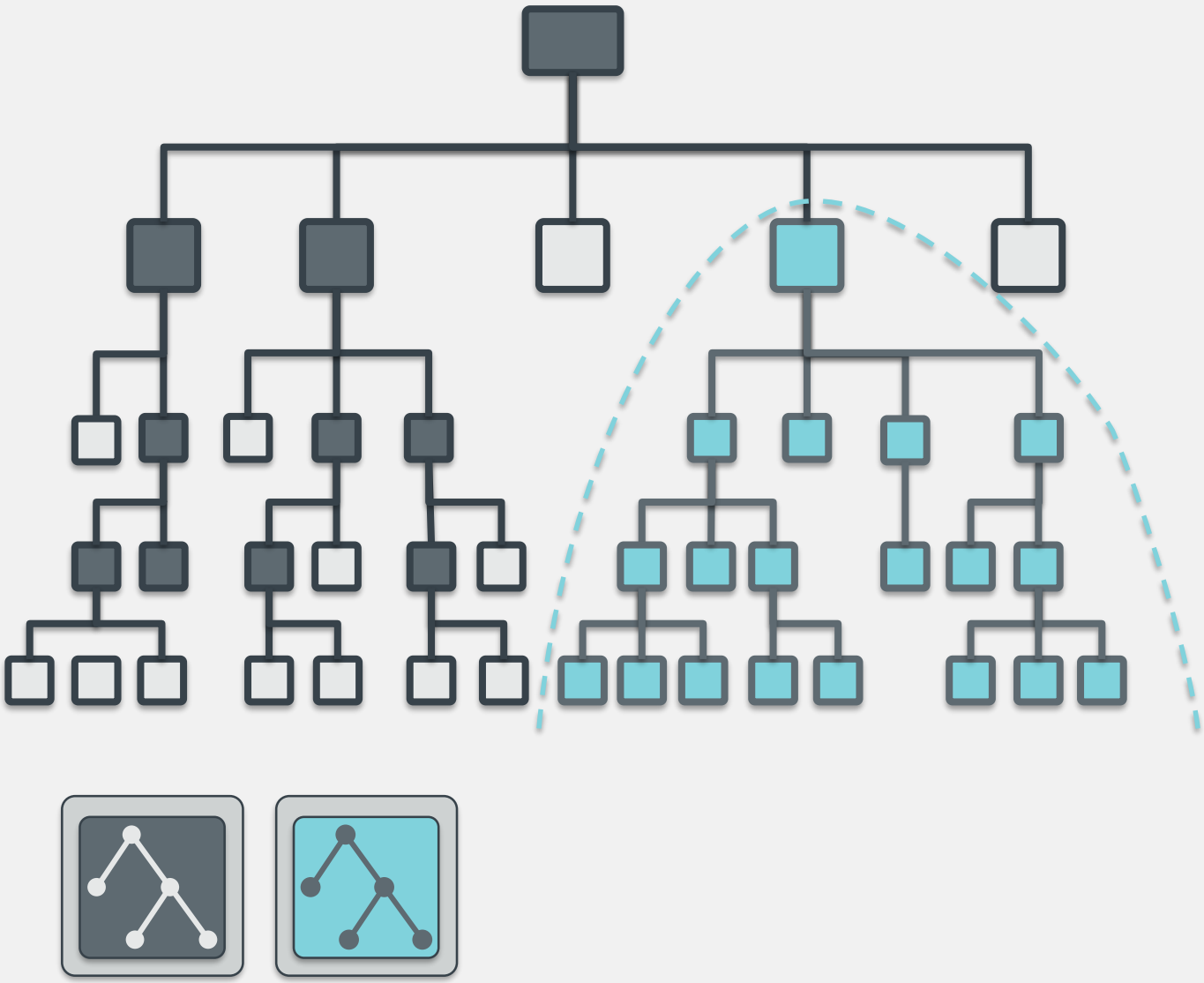


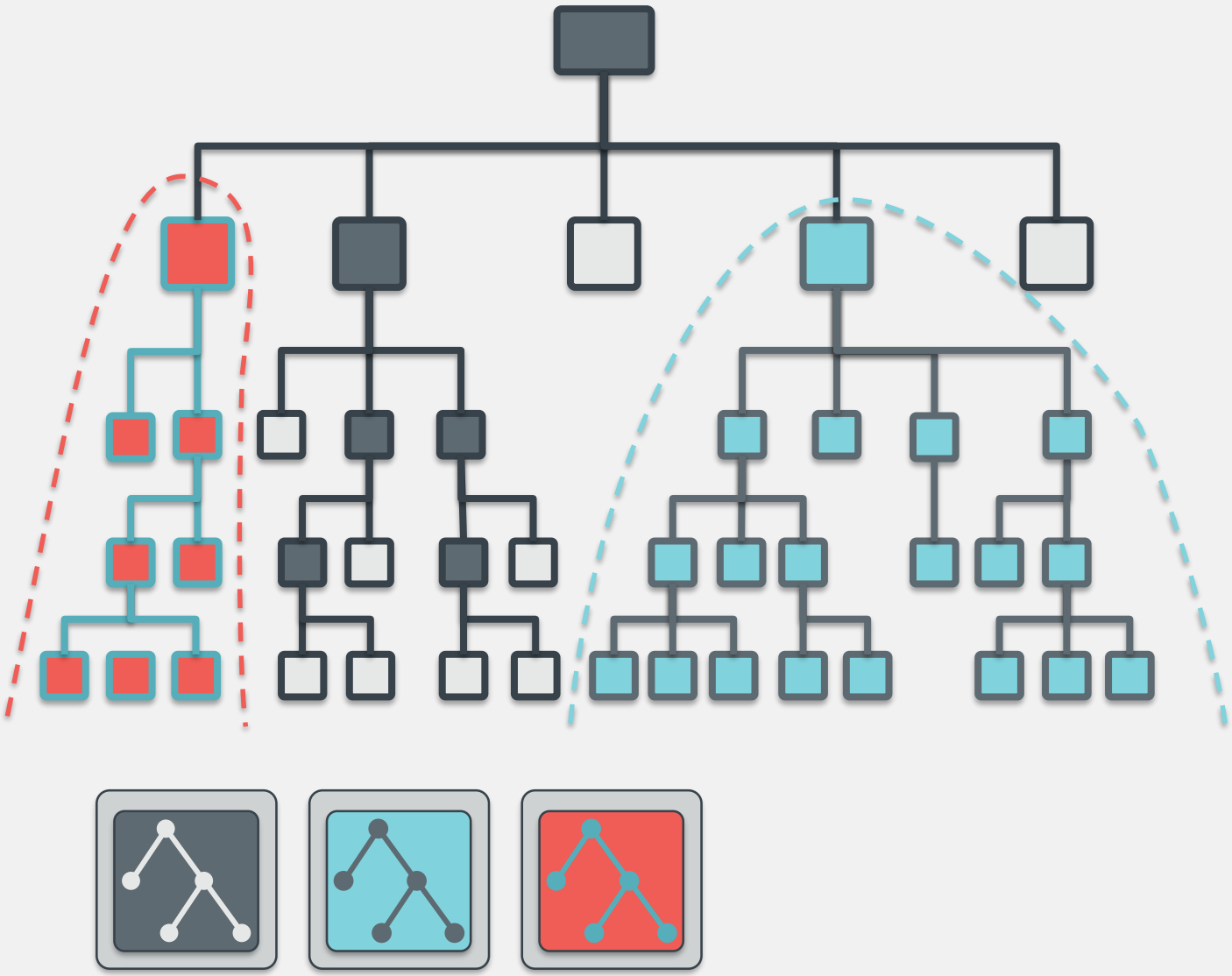
three metadata servers

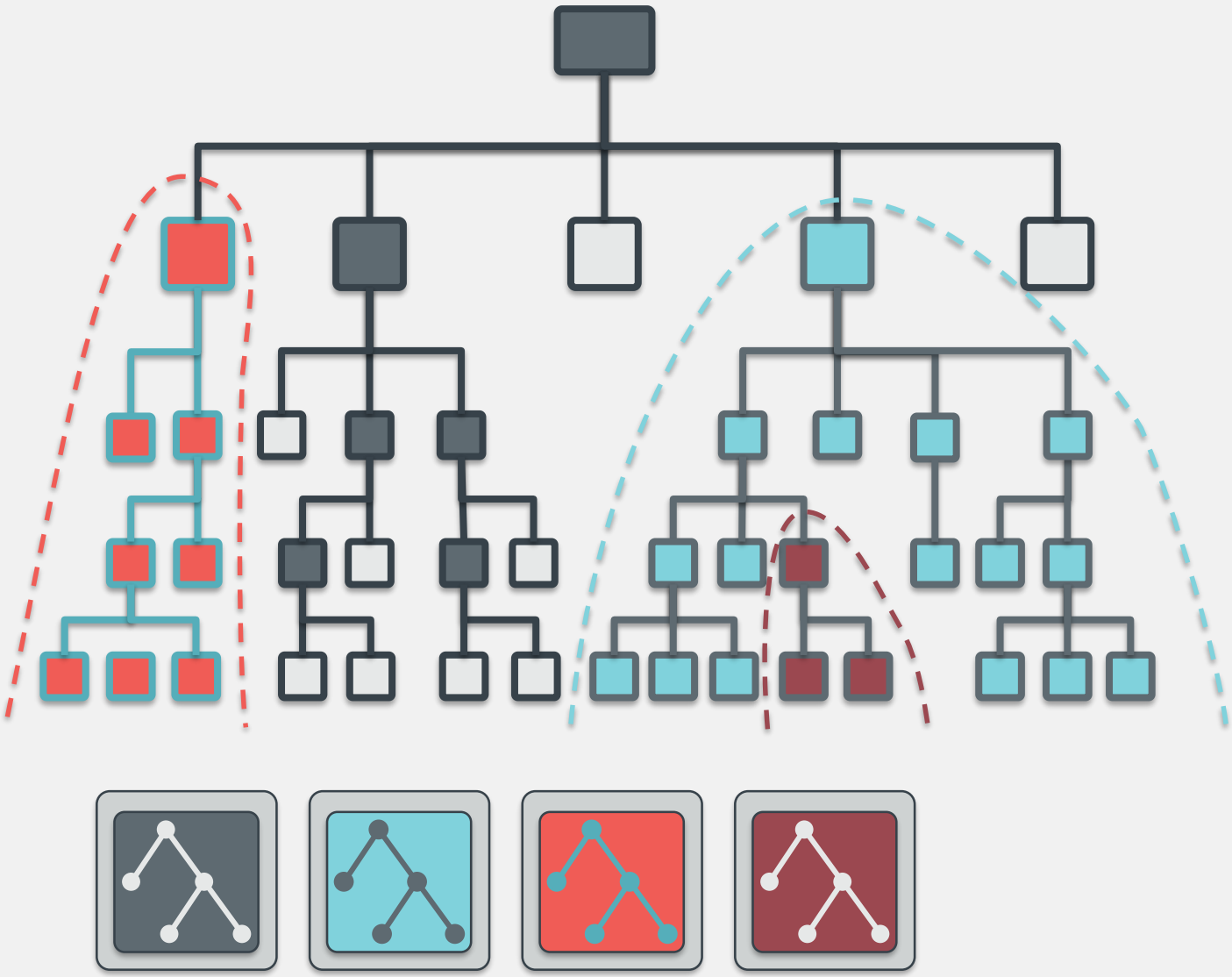


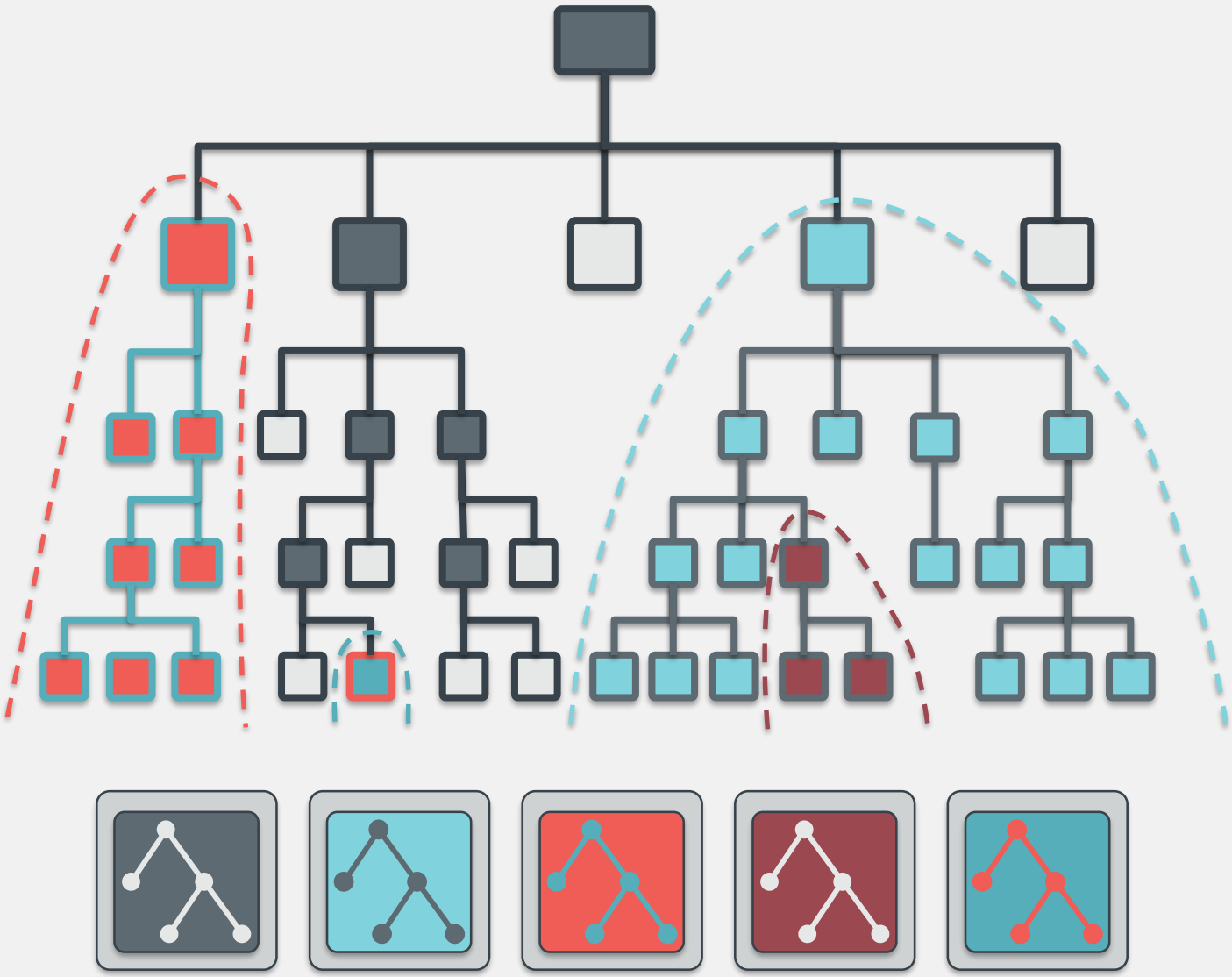
??











DYNAMIC SUBTREE PARTITIONING

OpenStack + Ceph

Used for *Glance* Images, *Cinder* Volumes and *Nova* ephemeral disk (coming soon)

Ceph + OpenStack offers compelling features:

- CoW clones, layered volumes, snapshots, boot from volume, live migration
- Cost effective with Thin Provisioning
  - ~110TB “used”, ~45TB * replicas on disk

Ceph is the most popular network block storage backend for OpenStack

Deployment

Automated deployment using Cephdeploy

Automated machine commissioning and maintenance

- Add a server to the hostgroup (osd, mon, radosgw)
- OSD disks are detected, formatted, prepared, auth'd
  - Also after disk replacement
- Auto-generated ceph.conf
- Last step is manual/controlled: service ceph start

Cephdeploy for bulk operations on the servers

- Ceph rpm upgrades
- daemon restarts

Getting Started With Ceph

Have a working cluster up quickly.

Read about the latest version of Ceph.

- The latest stuff is always at http://ceph.com/get

Deploy a test cluster using ceph-deploy.

- Read the quick-start guide at http://ceph.com/qsg

Deploy a test cluster on the AWS free-tier using Juju.

- Read the guide at http://ceph.com/juju

Read the rest of the docs!

- Find docs for the latest release at http://ceph.com/docs

Getting Involved With Ceph

Help build the best storage system around!

Most project discussion happens on the mailing list.

- Join or view archives at http://ceph.com/list

IRC is a great place to get help (or help others!)

- Find details and historical logs at http://ceph.com/irc

The tracker manages our bugs and feature requests.

- Register and start looking around at http://ceph.com/tracker

Doc updates and suggestions are always welcome.

- Learn how to contribute docs at http://ceph.com/docwriting