

# Ceph Intro & Architectural Overview

Ross Turk

VP Community, Inktank

# ME ME ME ME ME ME.

I made a slide today. It's all about me.

Ross Turk

VP Community, Inktank

ross@inktank.com

@rossturk

inktank.com | ceph.com



# CLOUD SERVICES

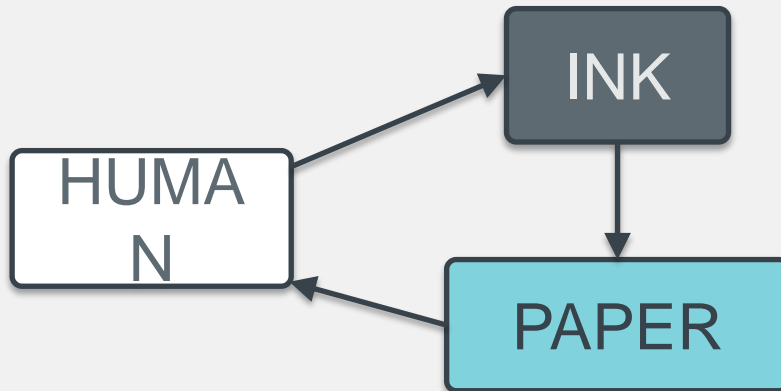
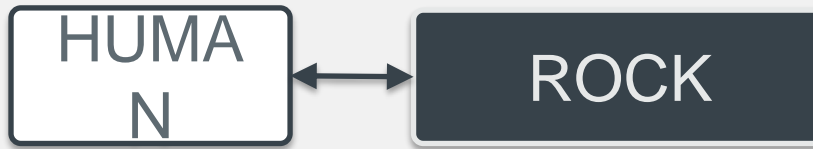
COMPUTE

NETWORK

STORAGE

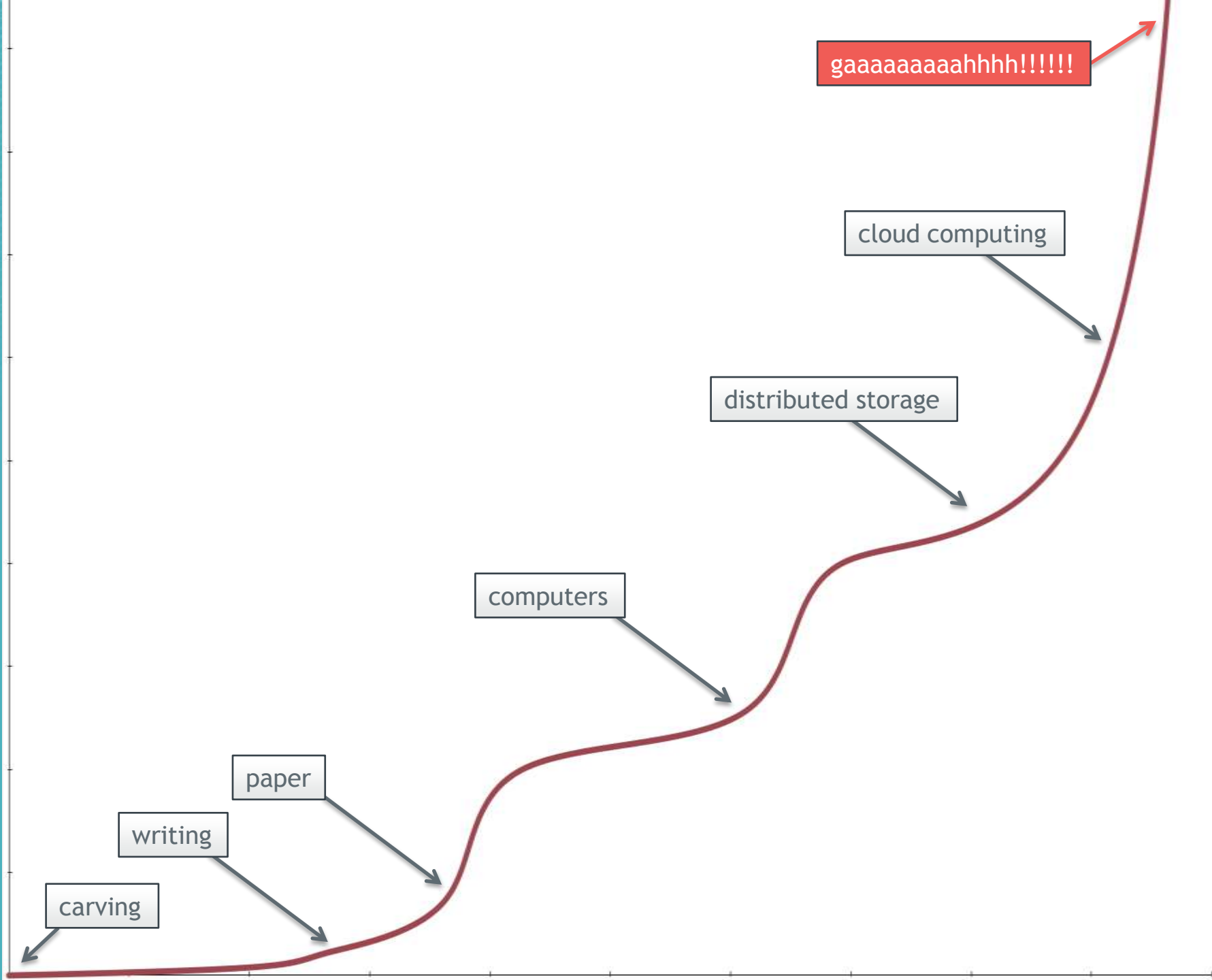


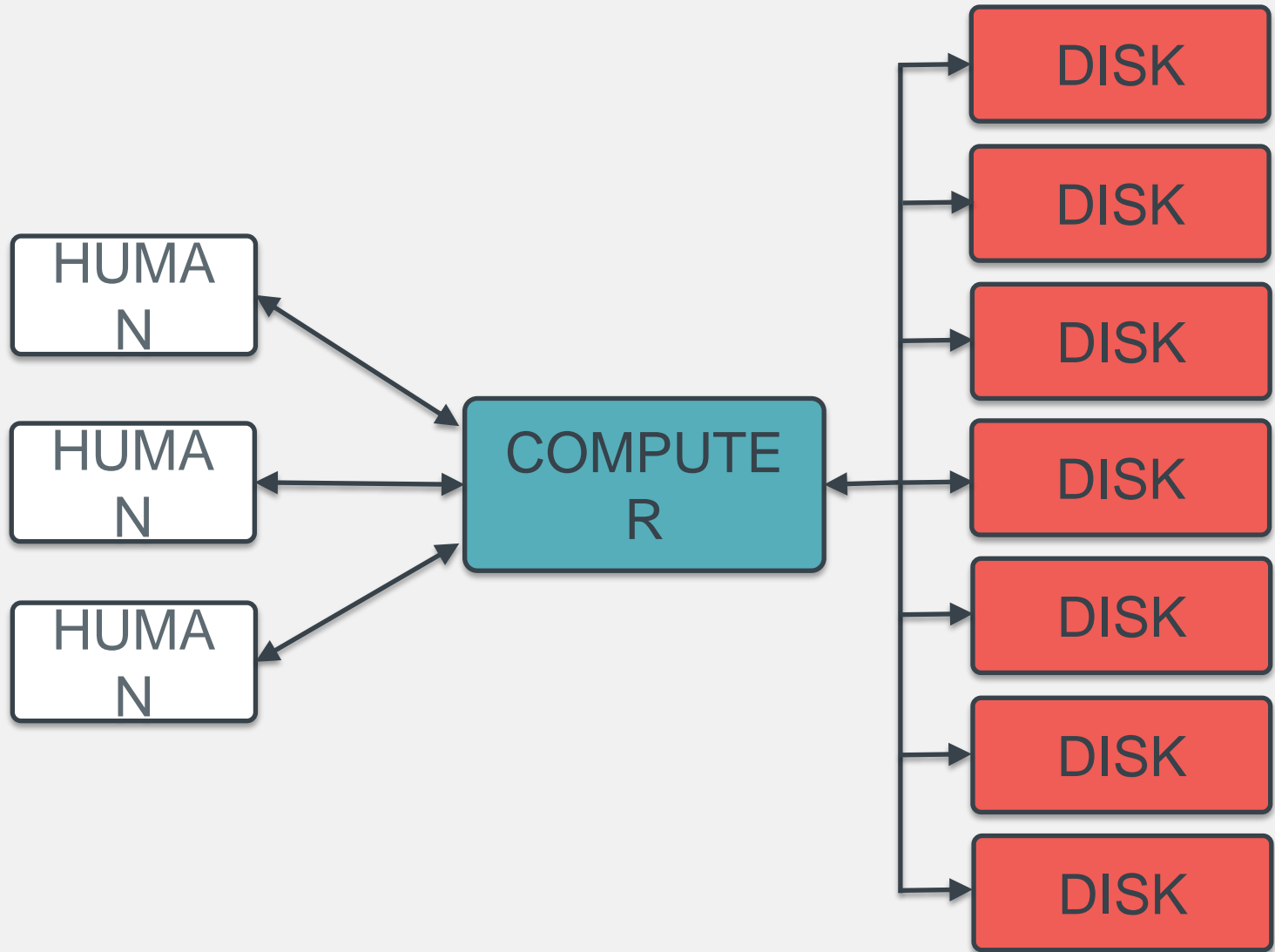
the future of storage™



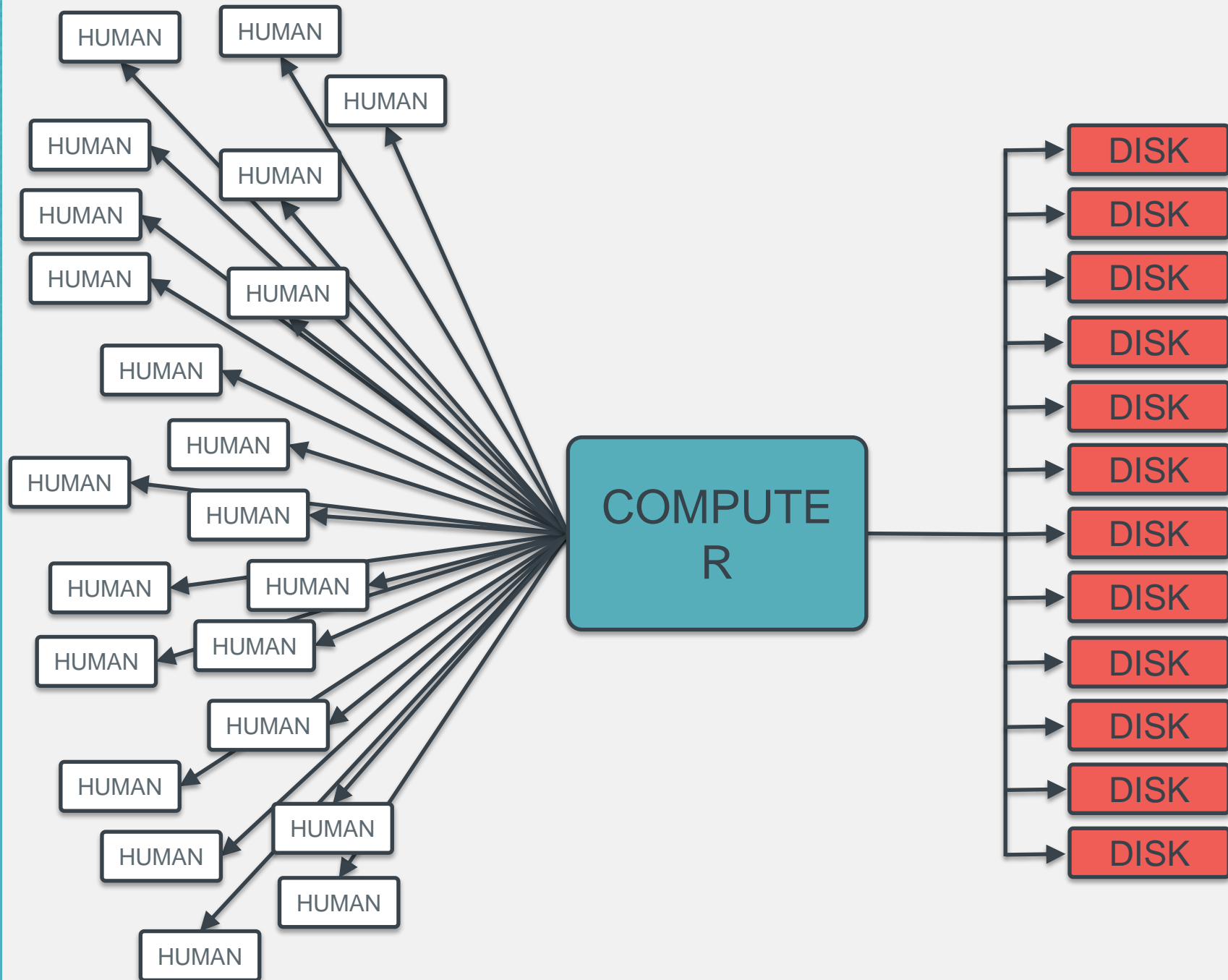




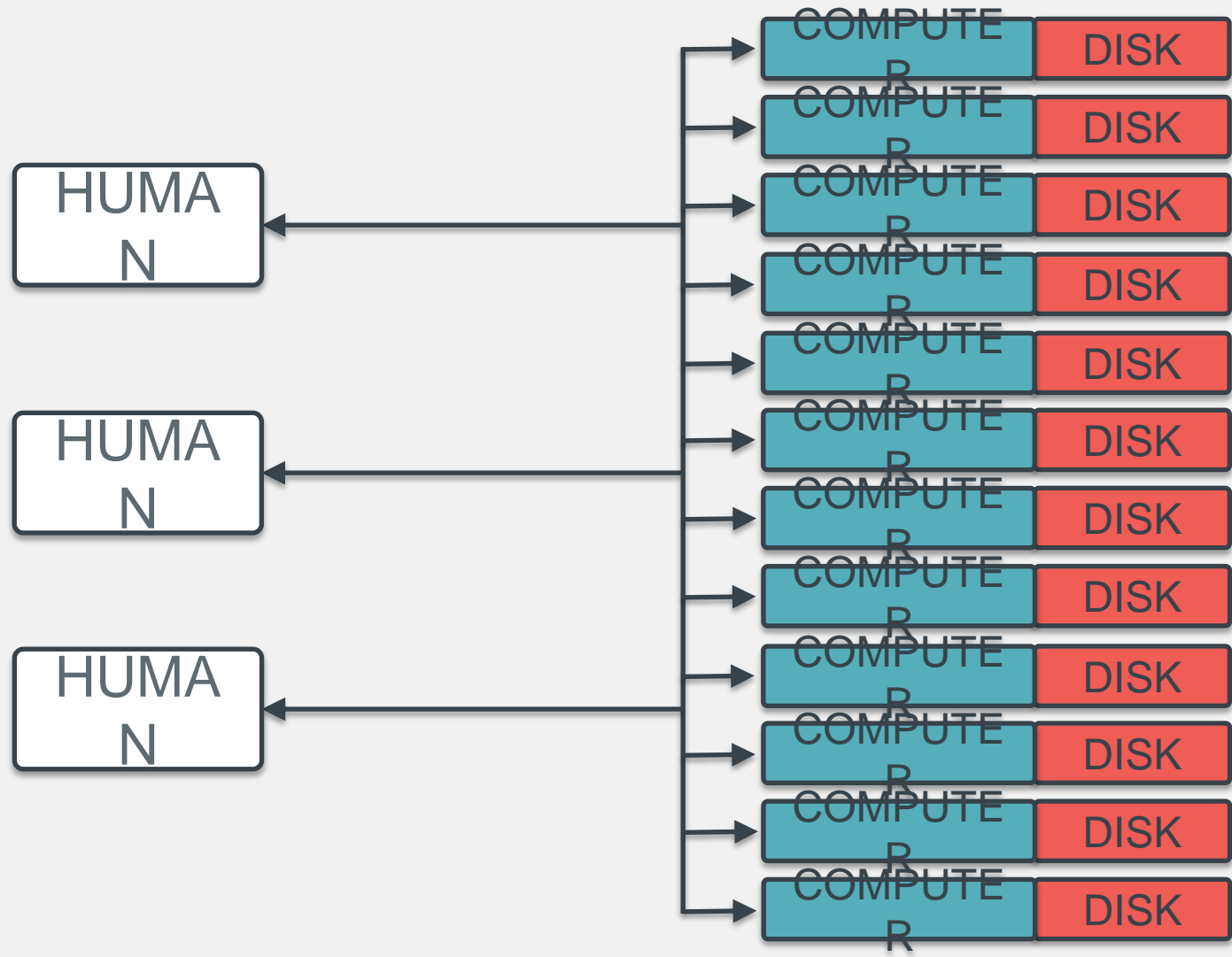


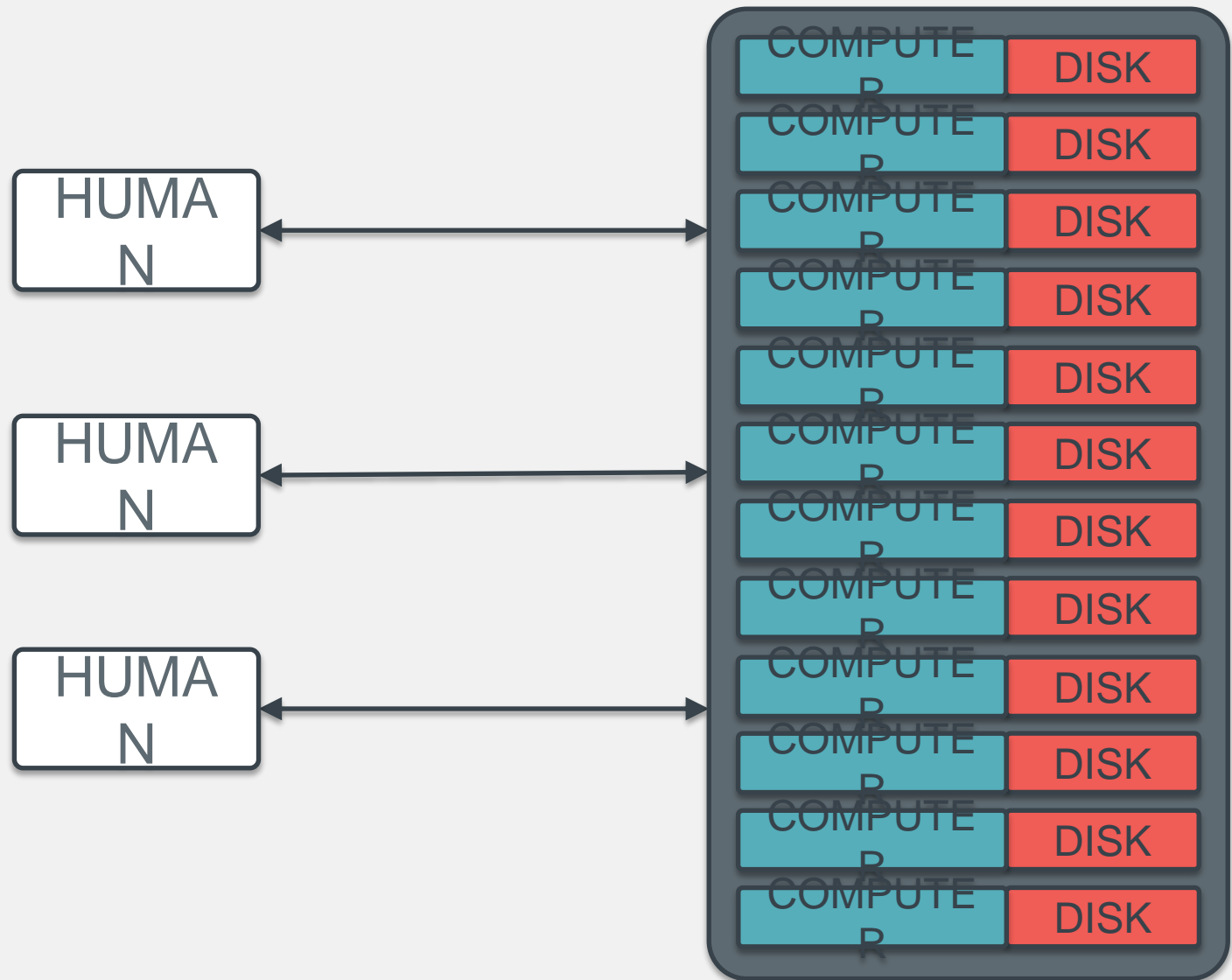






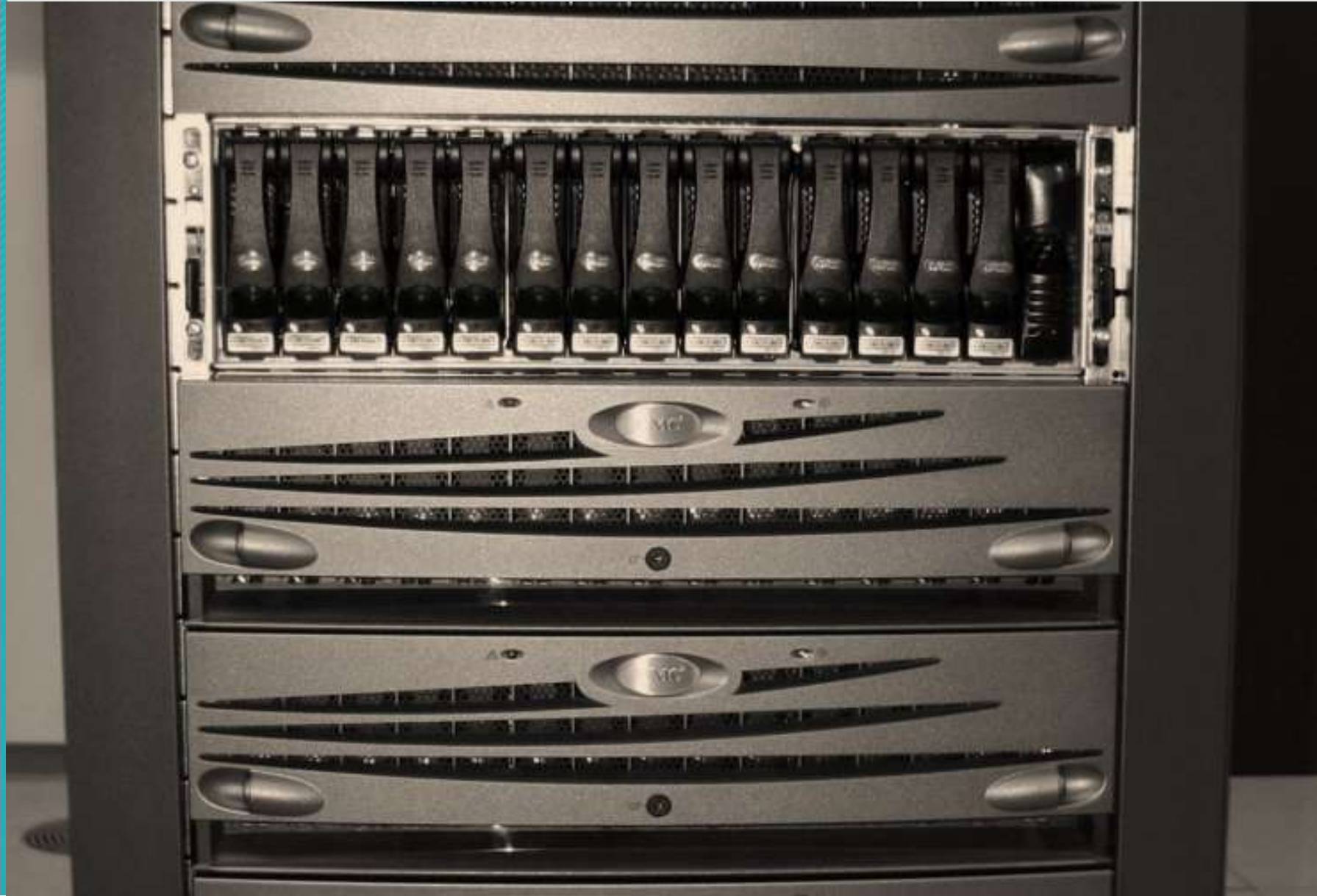






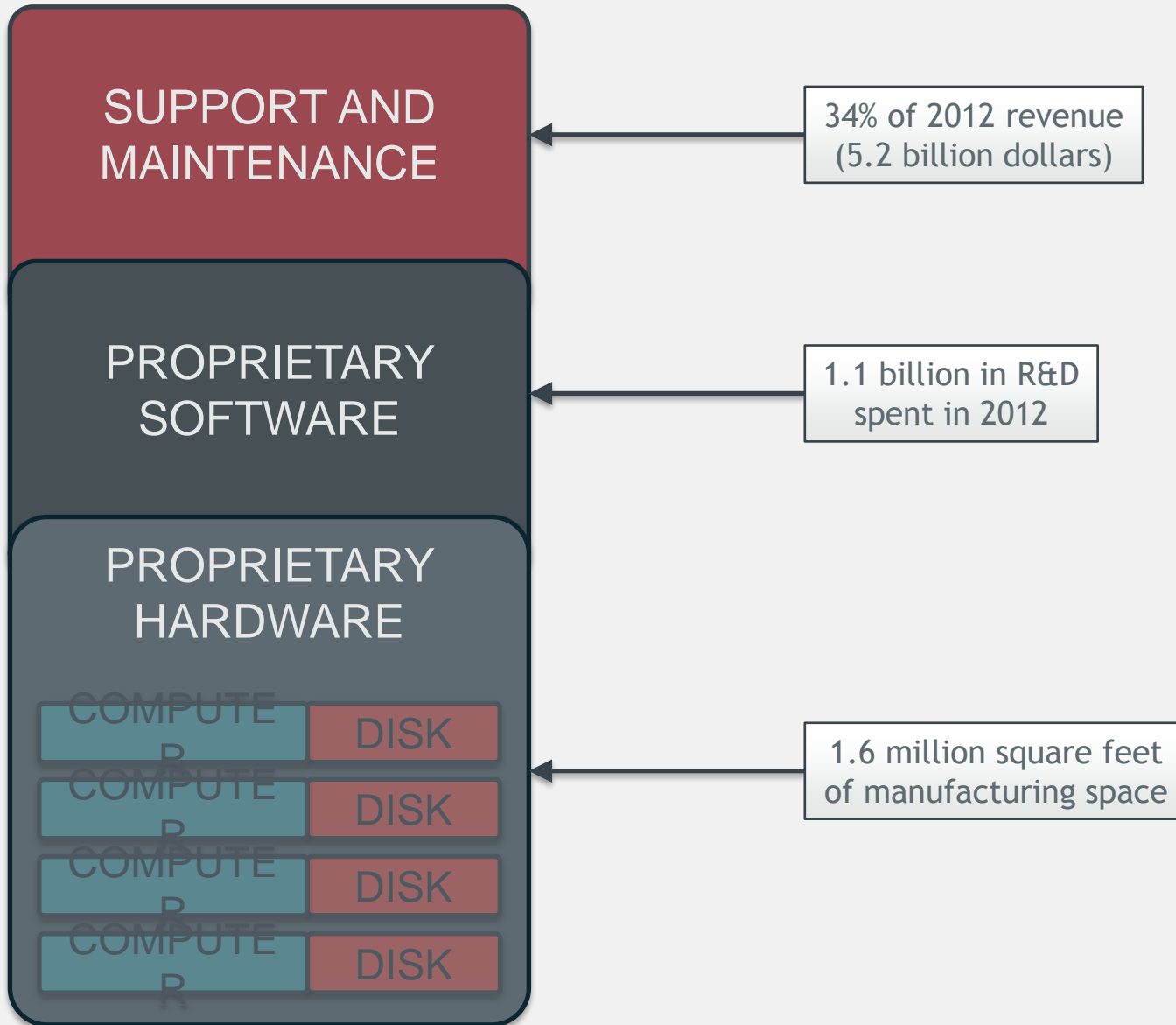
# “STORAGE APPLIANCE”





## Storage Appliance

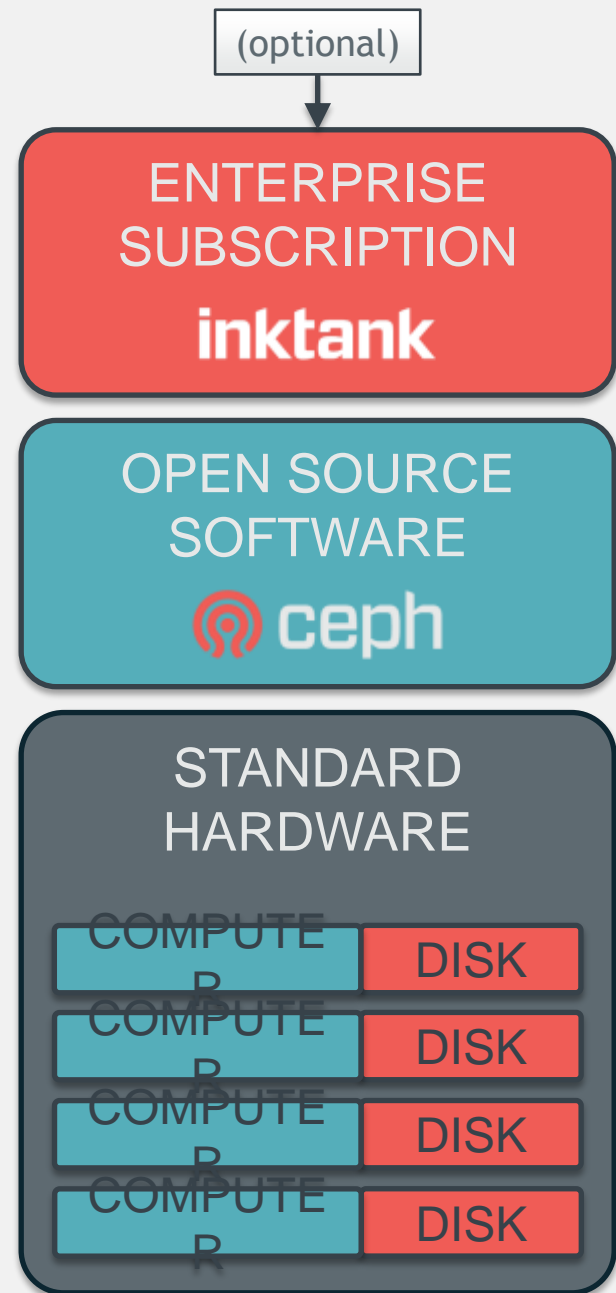
Michael Moll, Wikipedia / CC BY-SA 2.0



# THE CLOUD

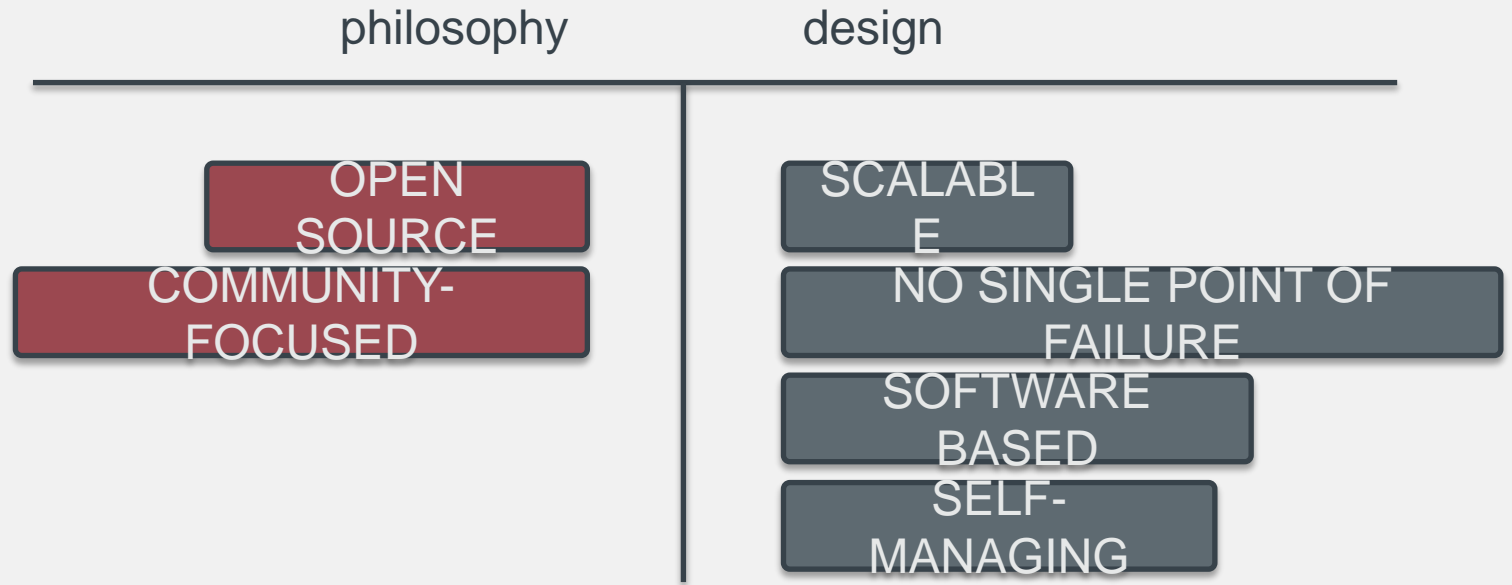
0	0	0	0	0	0
1	1	0	1	0	0
0	0	1	1	1	1
1	1	1	0	1	0
0	1	0	1	0	1
0	0	0	0	0	0
1	0	1	1	1	1
1	1	0	1	1	1
0	1	1	1	1	1



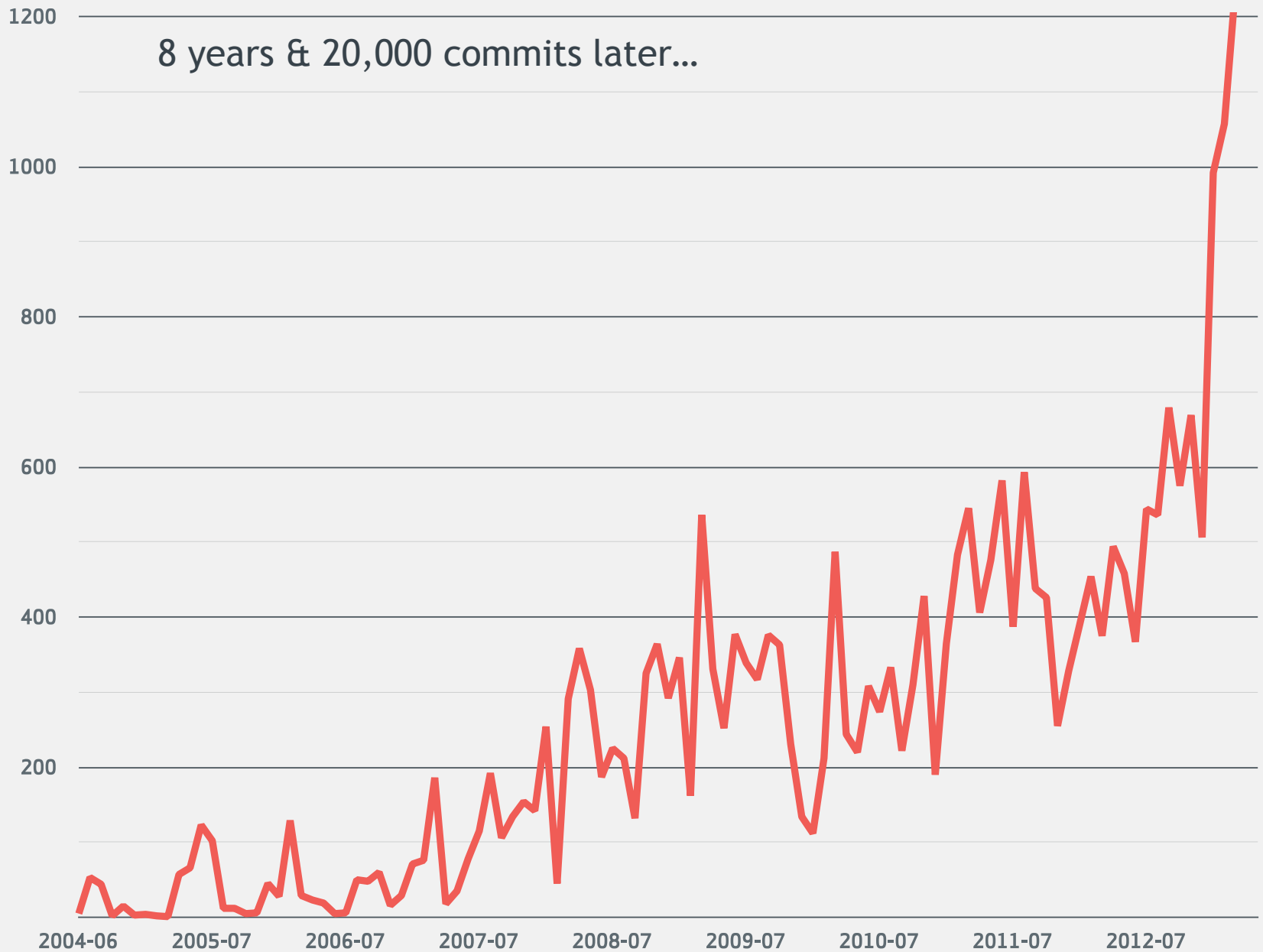


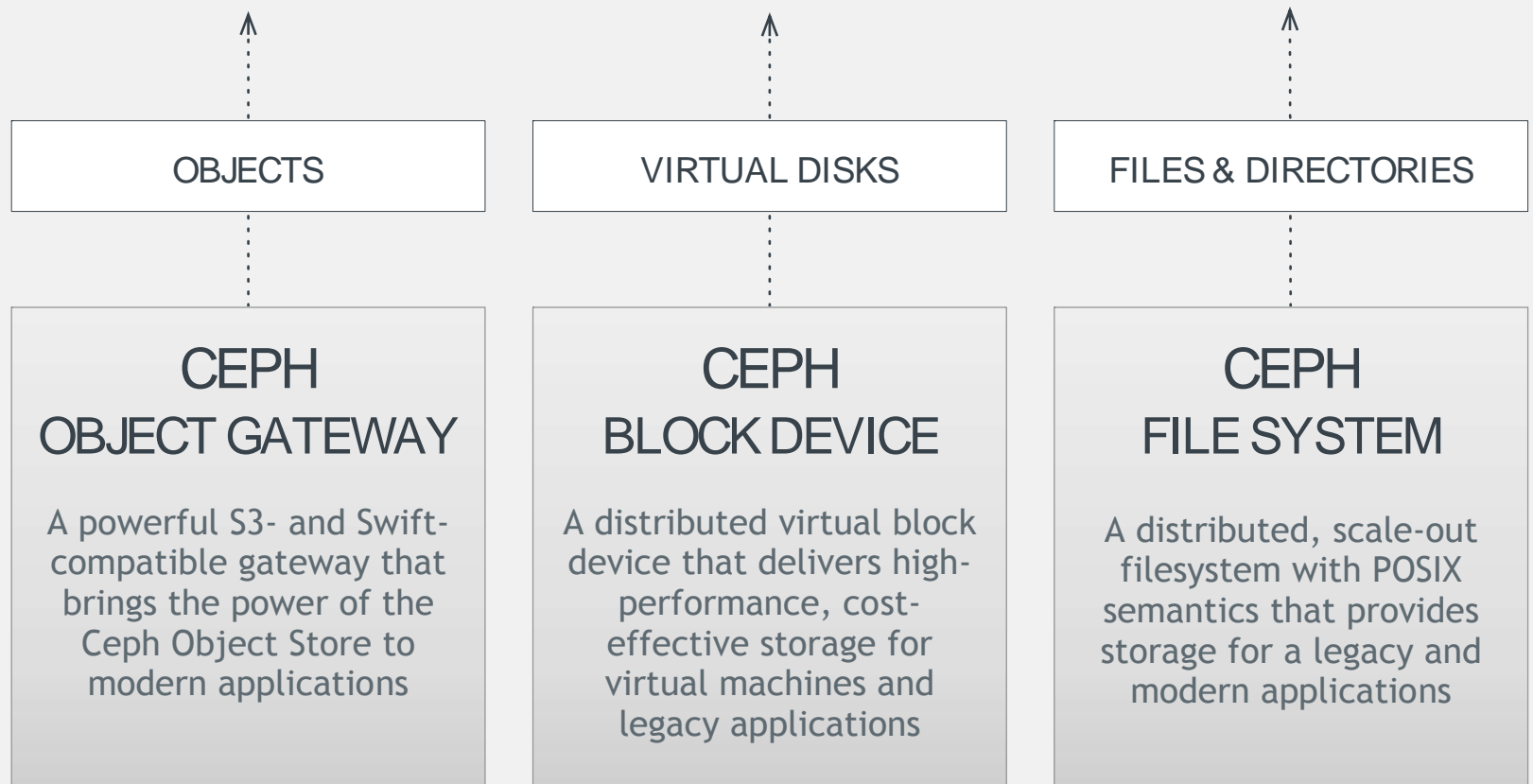


ceph



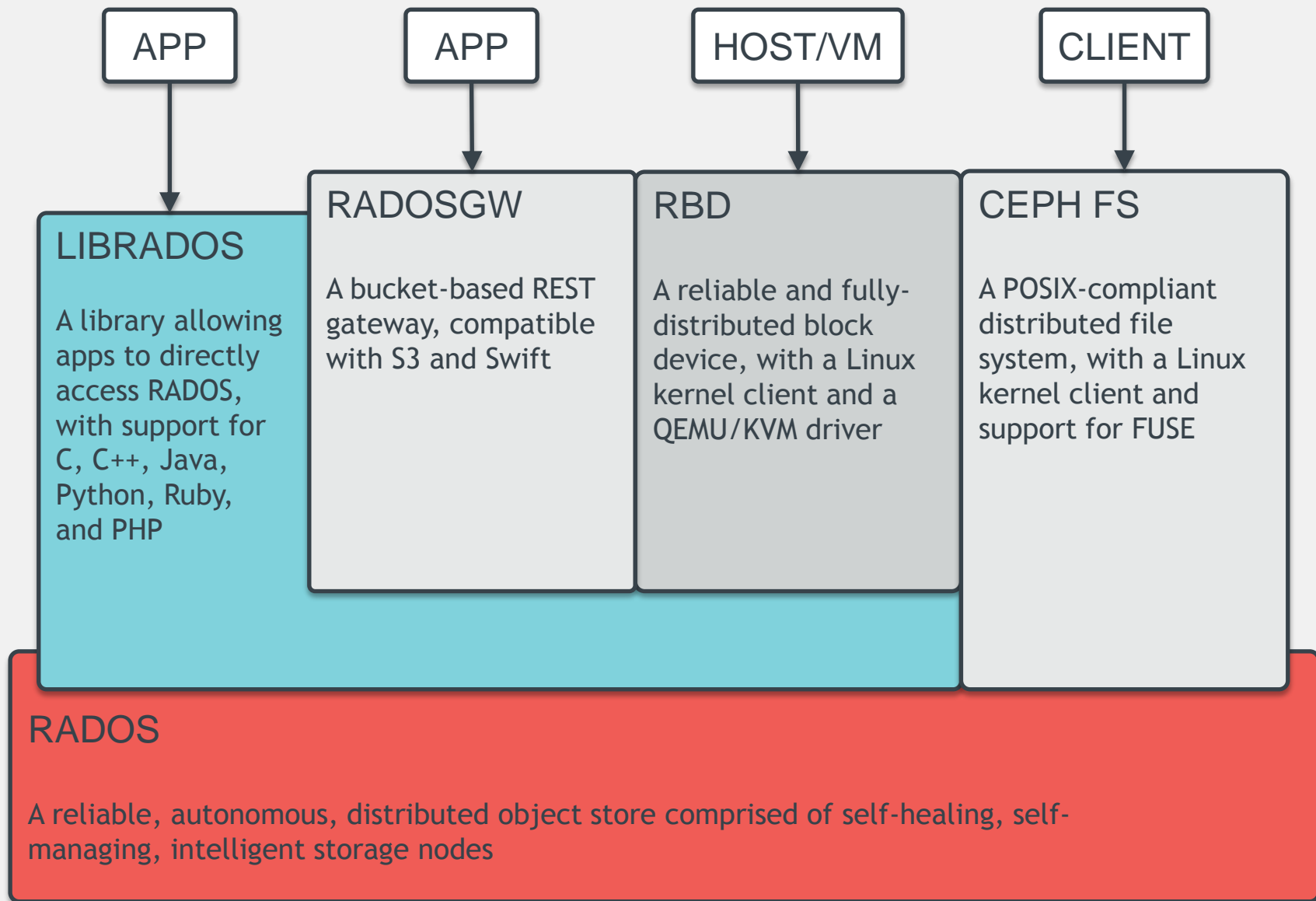
8 years & 20,000 commits later...

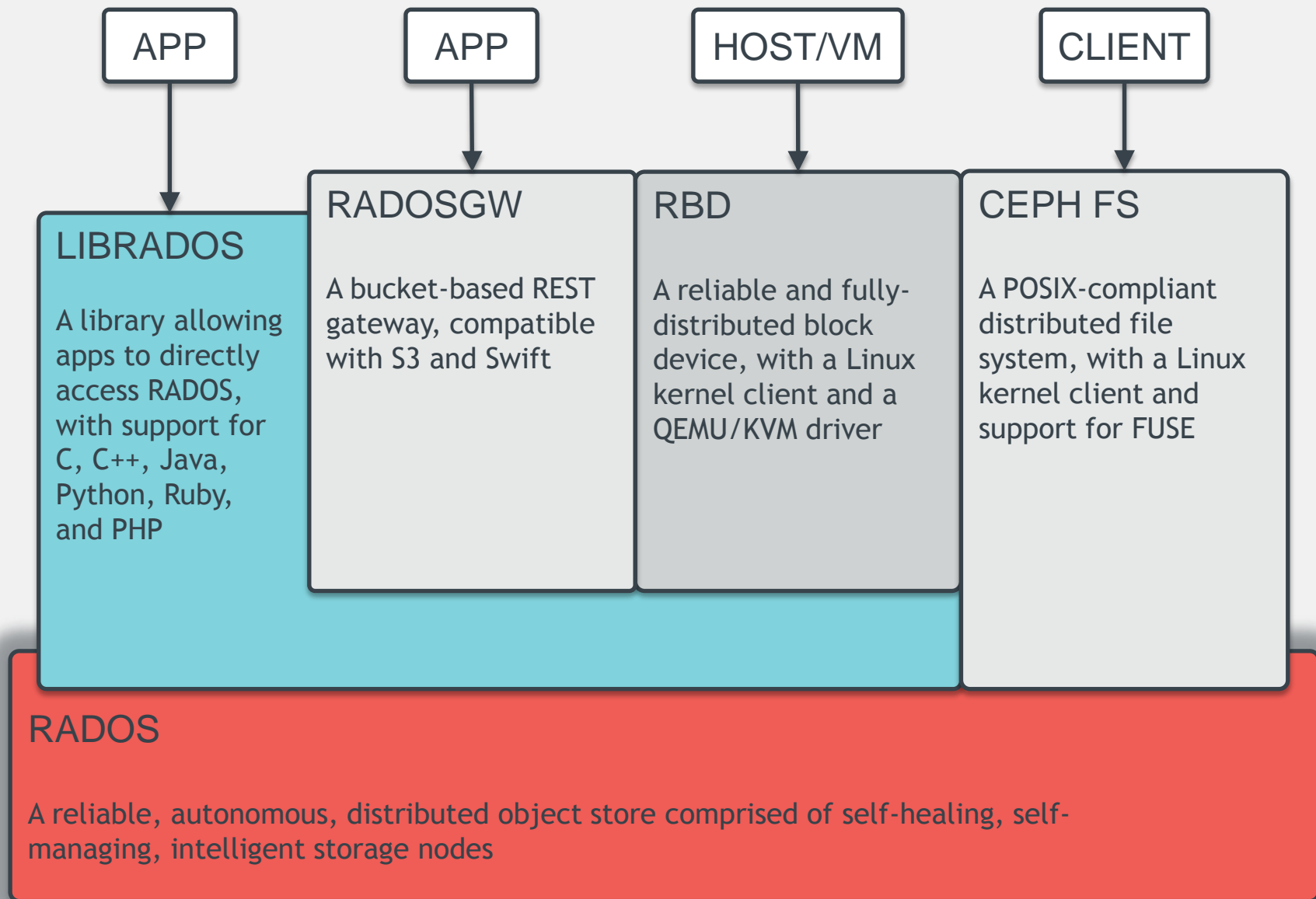


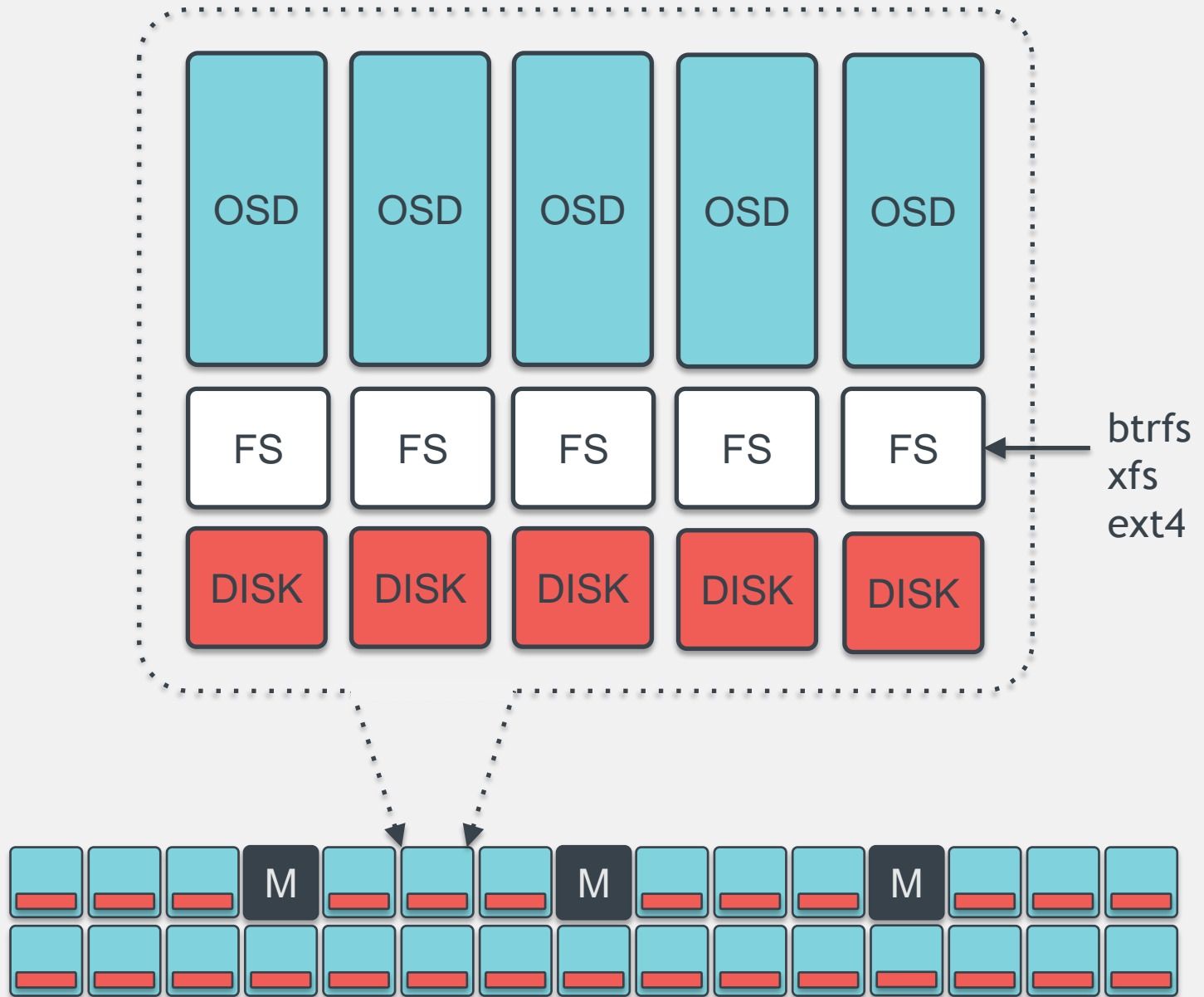


## CEPH STORAGE CLUSTER

A reliable, easy to manage, next-generation distributed object store that provides storage of unstructured data for applications

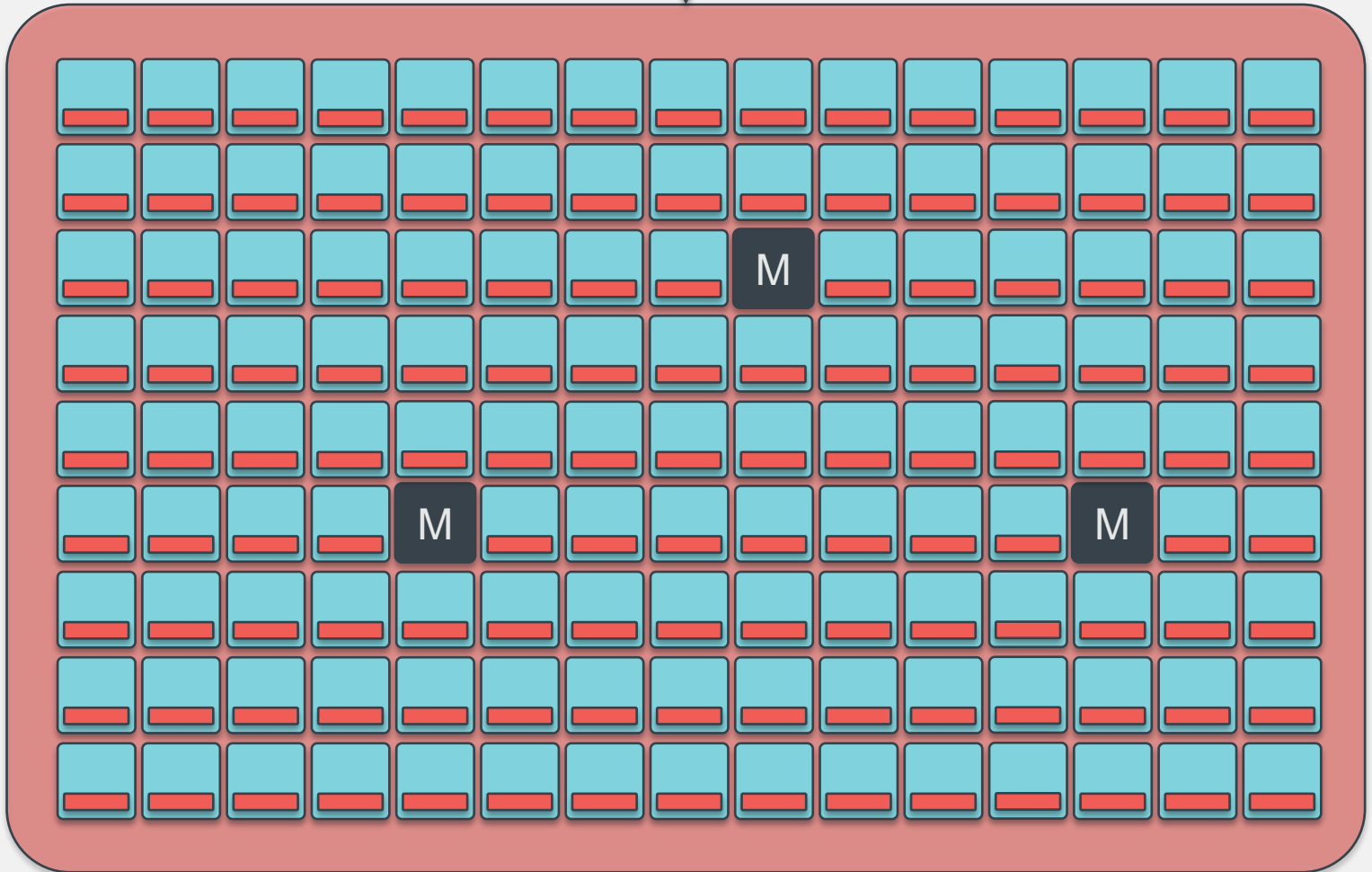


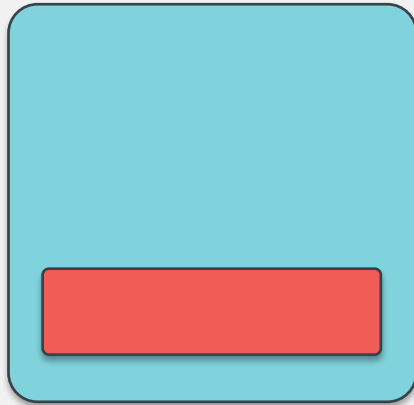






HUMAN





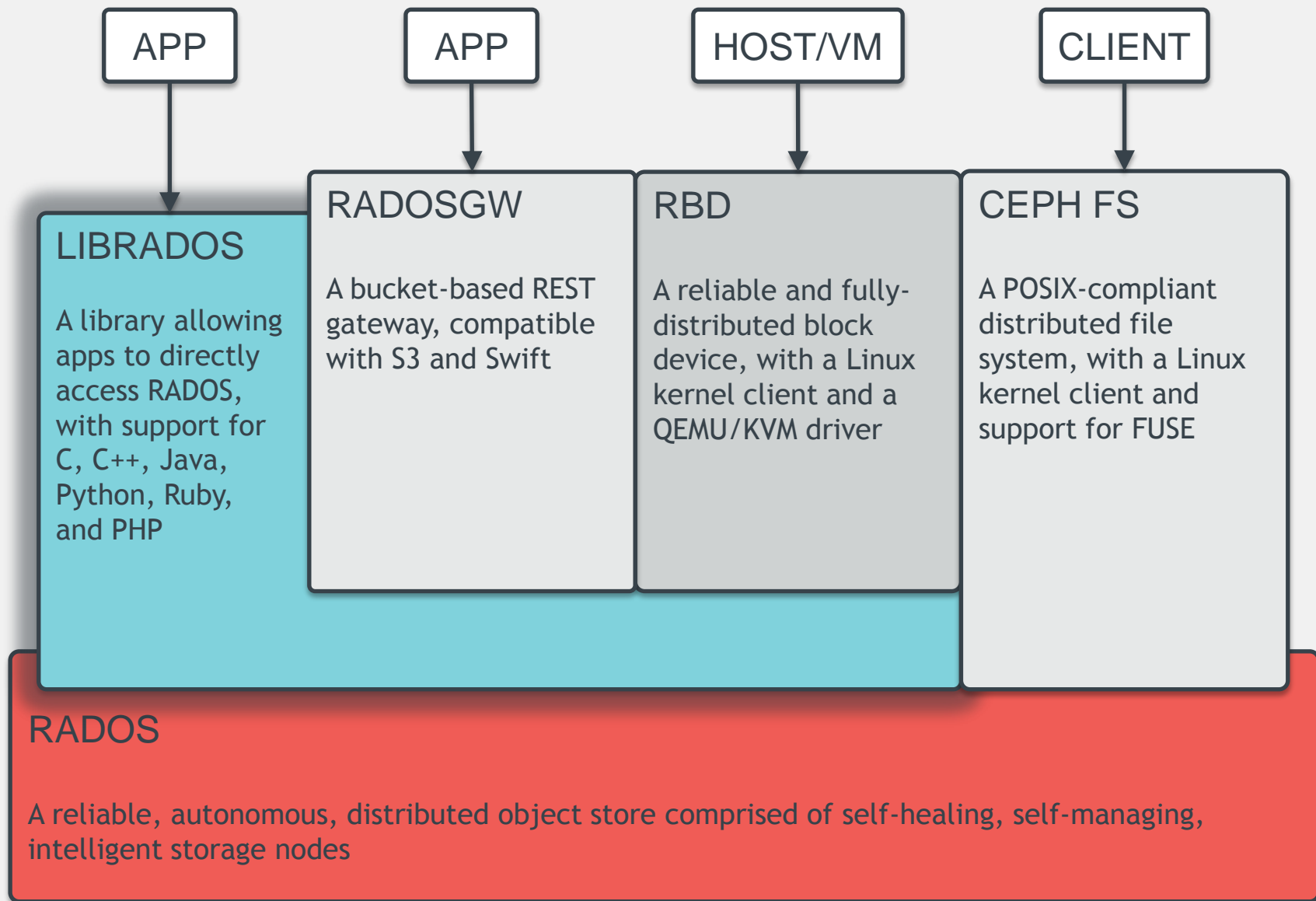
### OSDs:

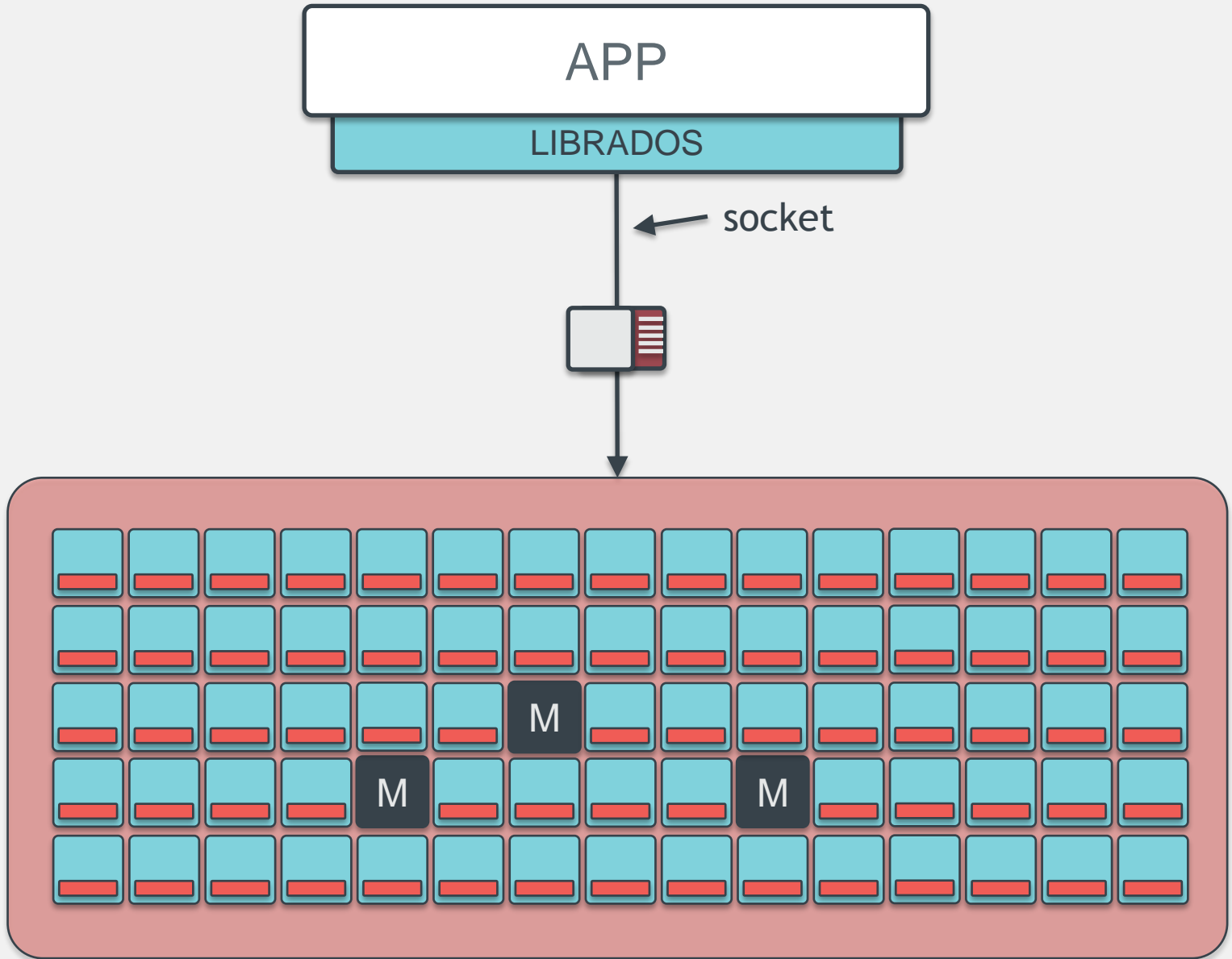
- 10s to 10000s in a cluster
- One per disk
  - (or one per SSD, RAID group...)
- Serve stored objects to clients
- Intelligently peer to perform replication and recovery tasks

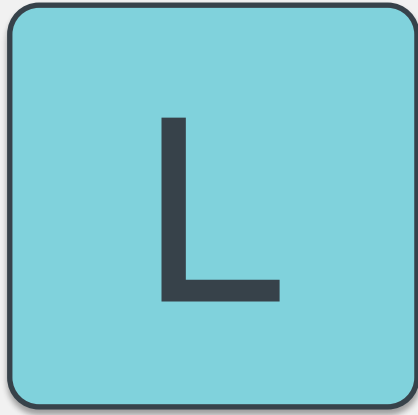


### Monitors:

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

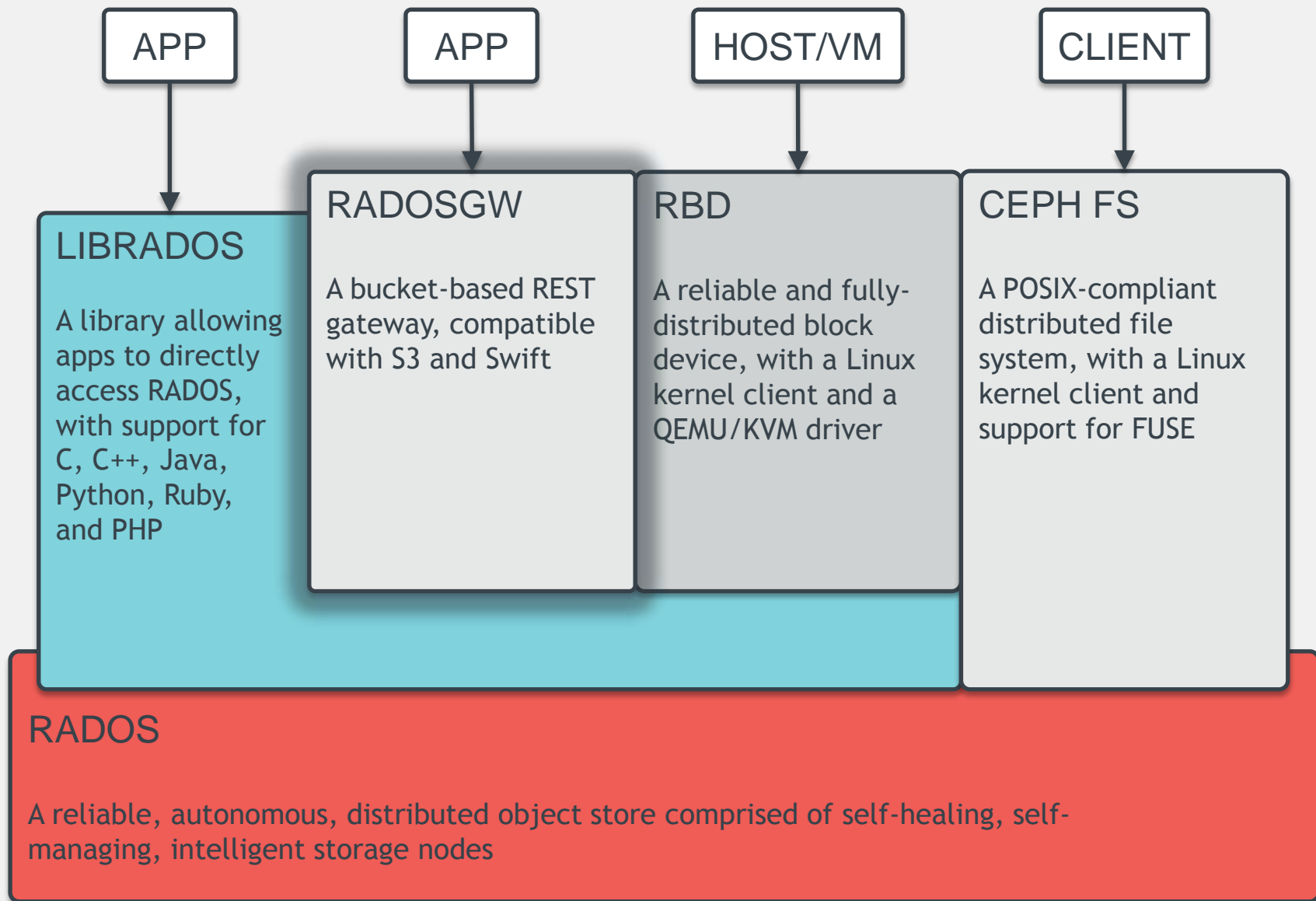


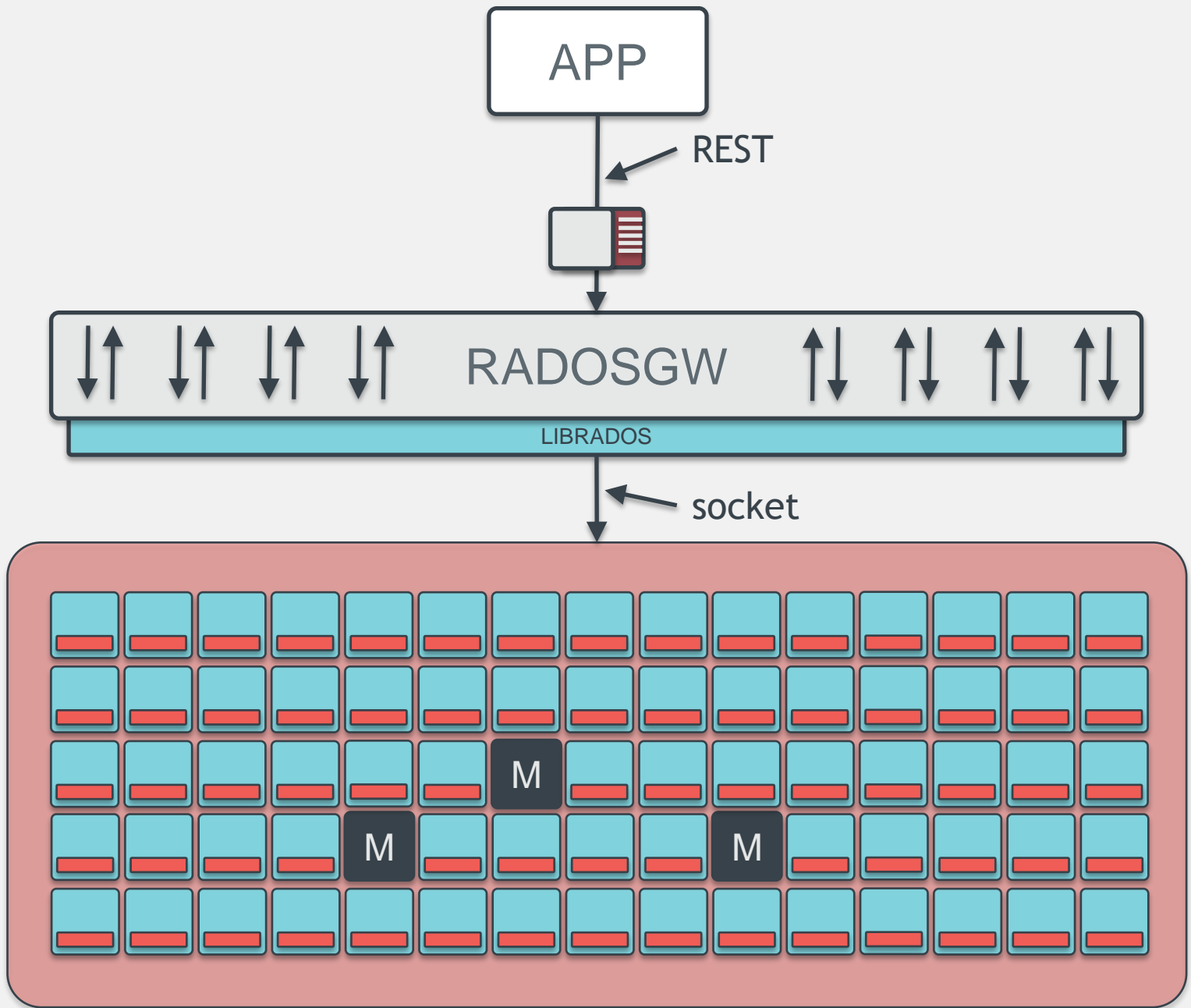


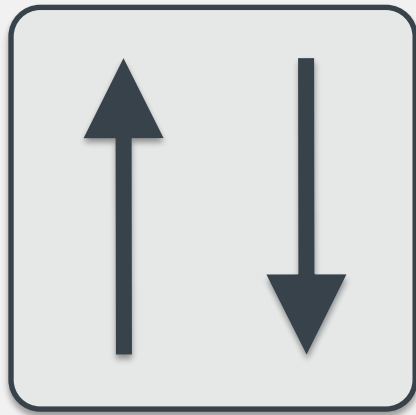


## LIBRADOS

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



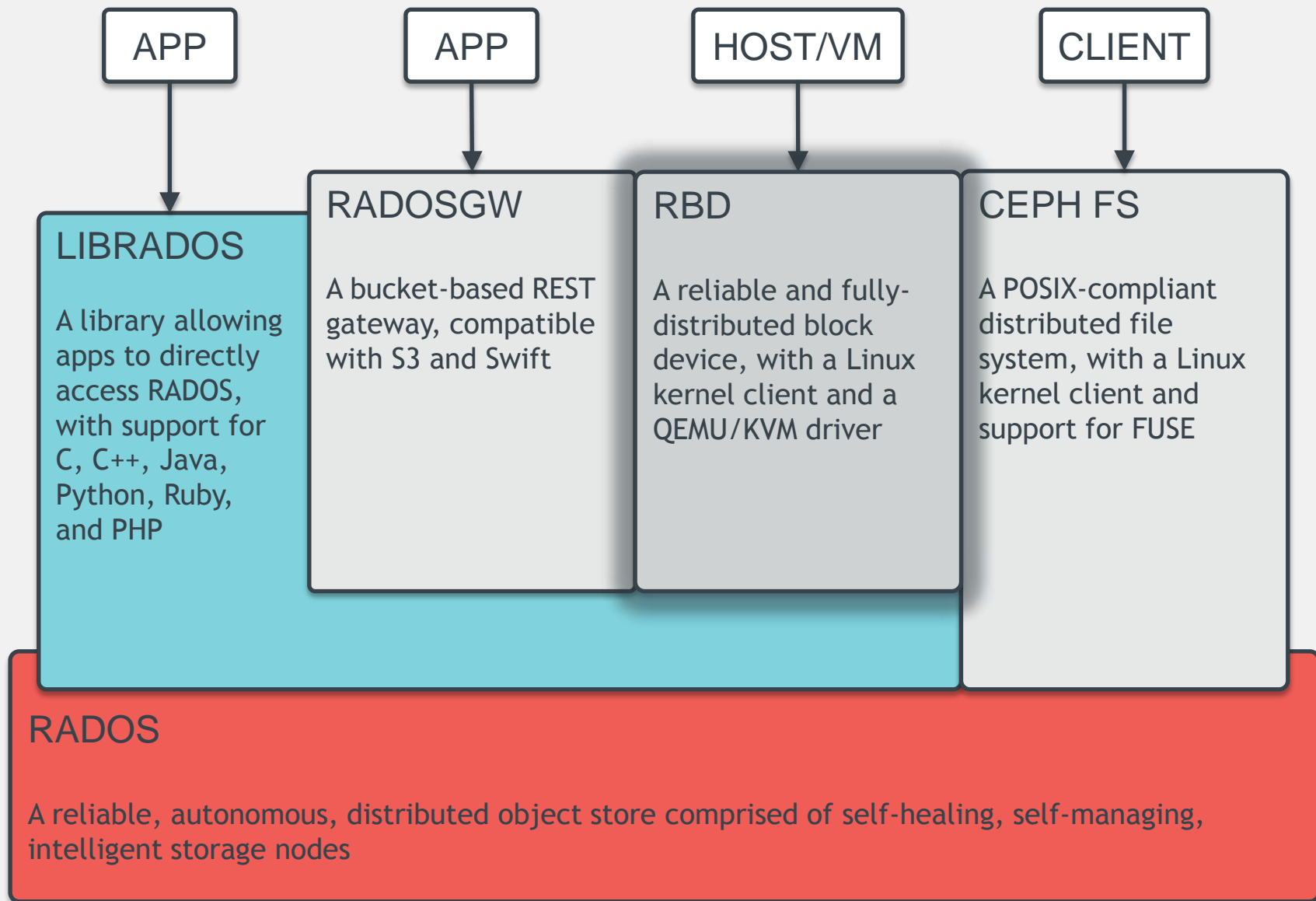


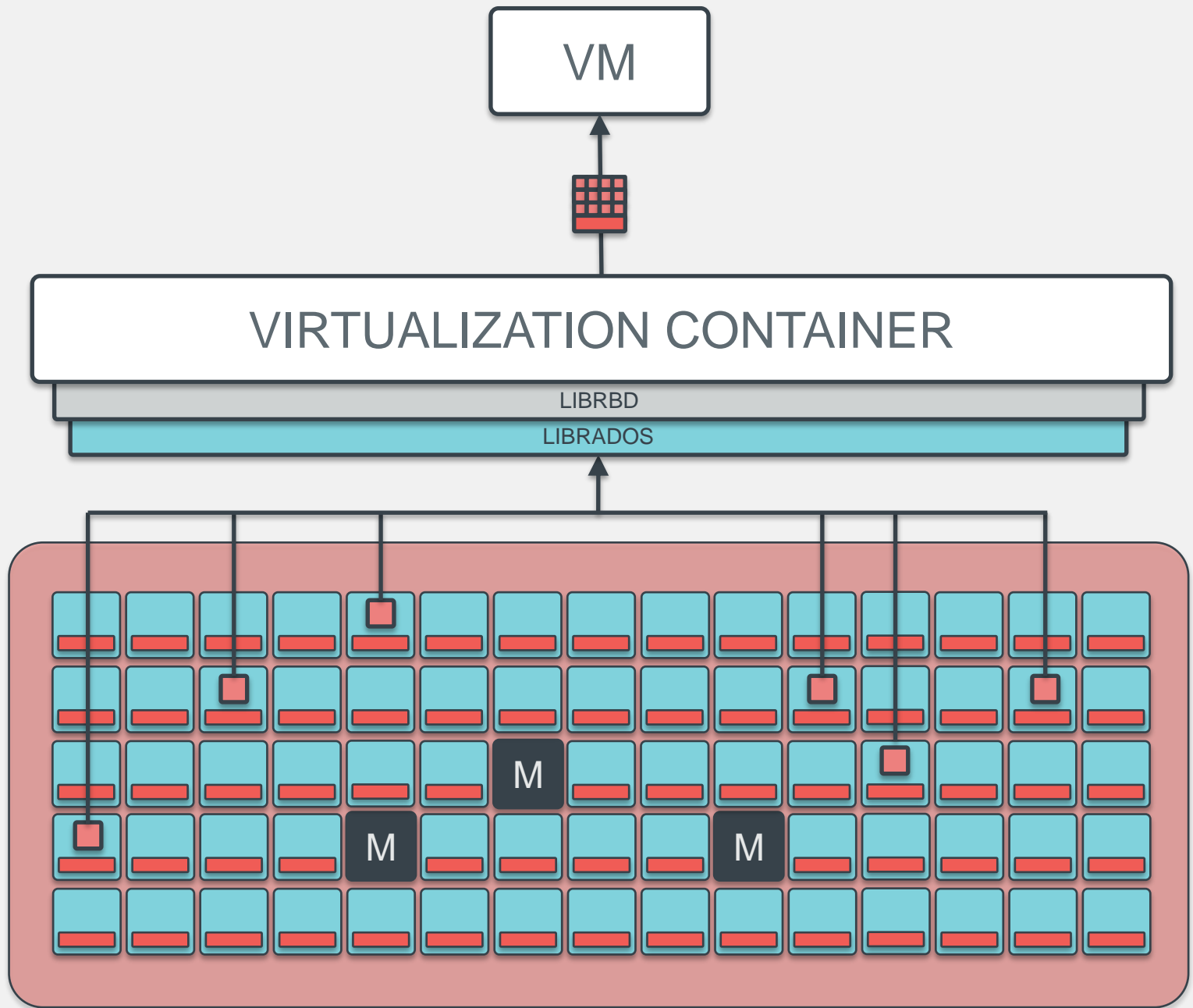


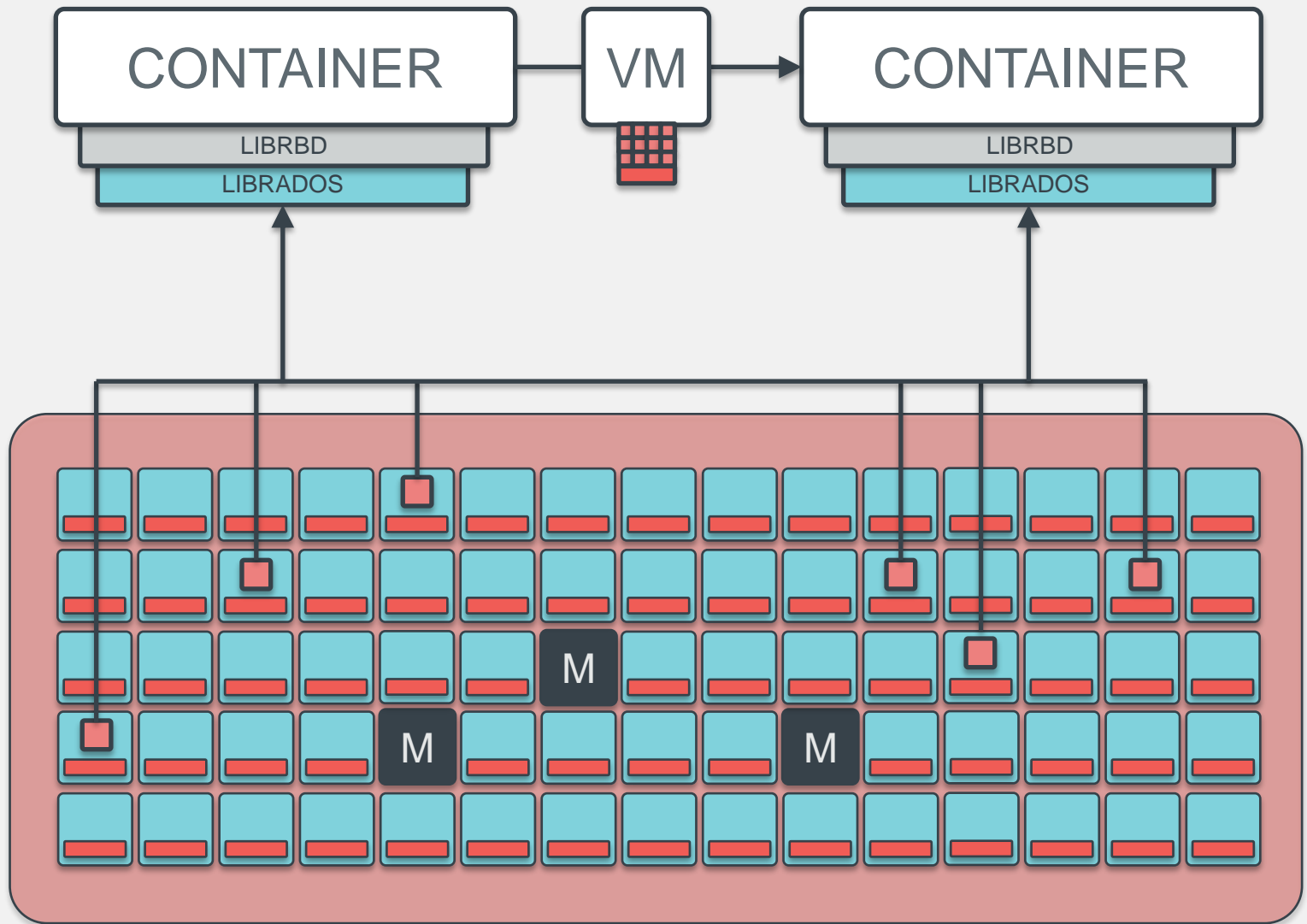
## RADOS Gateway:

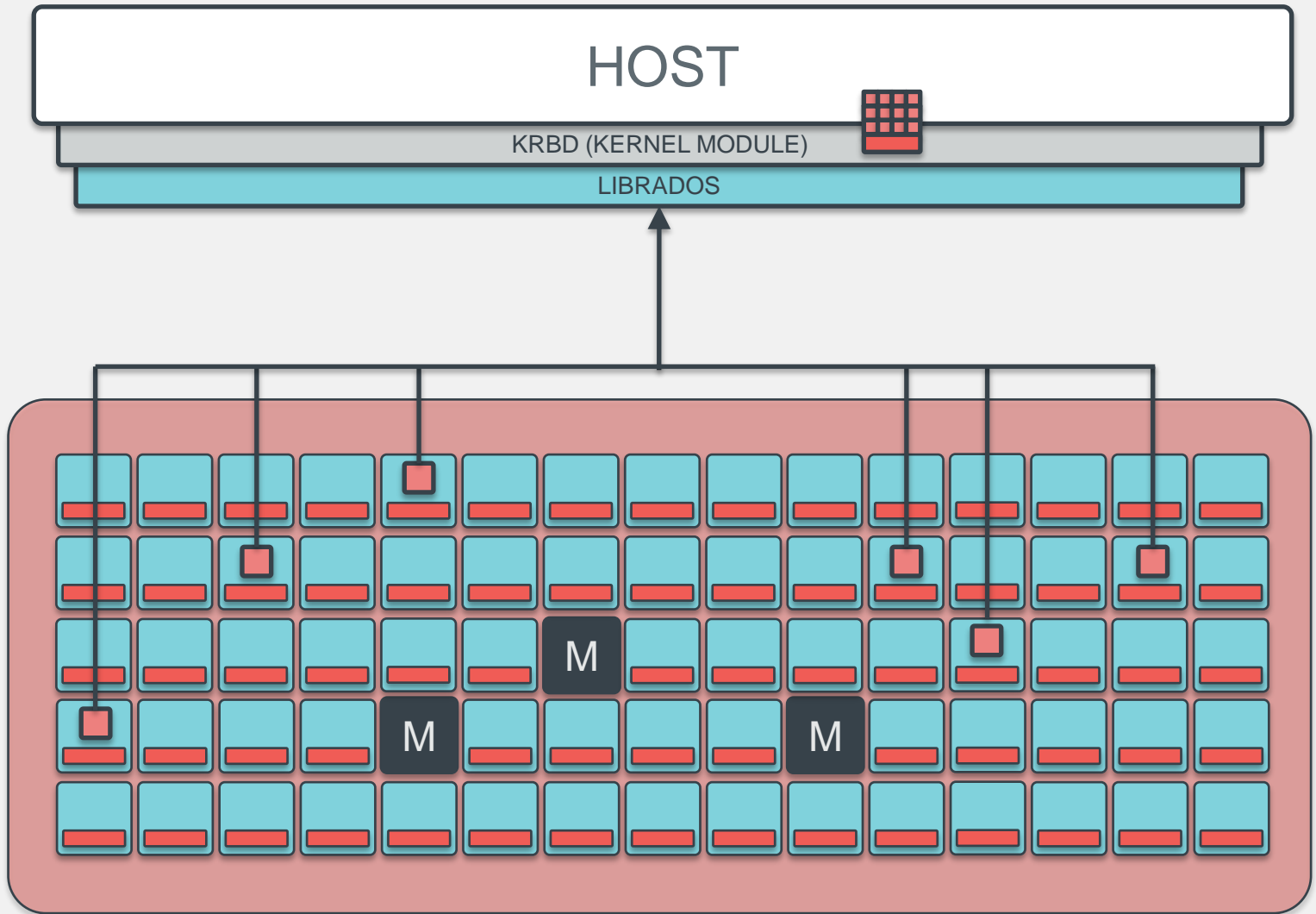
- REST-based object storage proxy
- Uses RADOS to store objects
- API supports buckets, accounts
- Usage accounting for billing
- Compatible with S3 and Swift applications

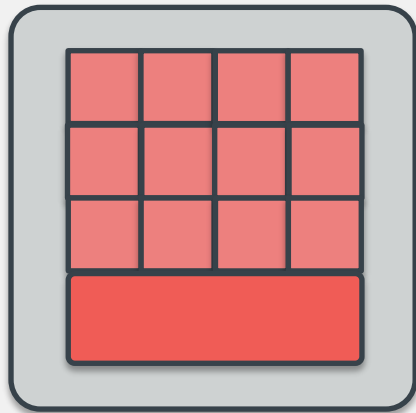






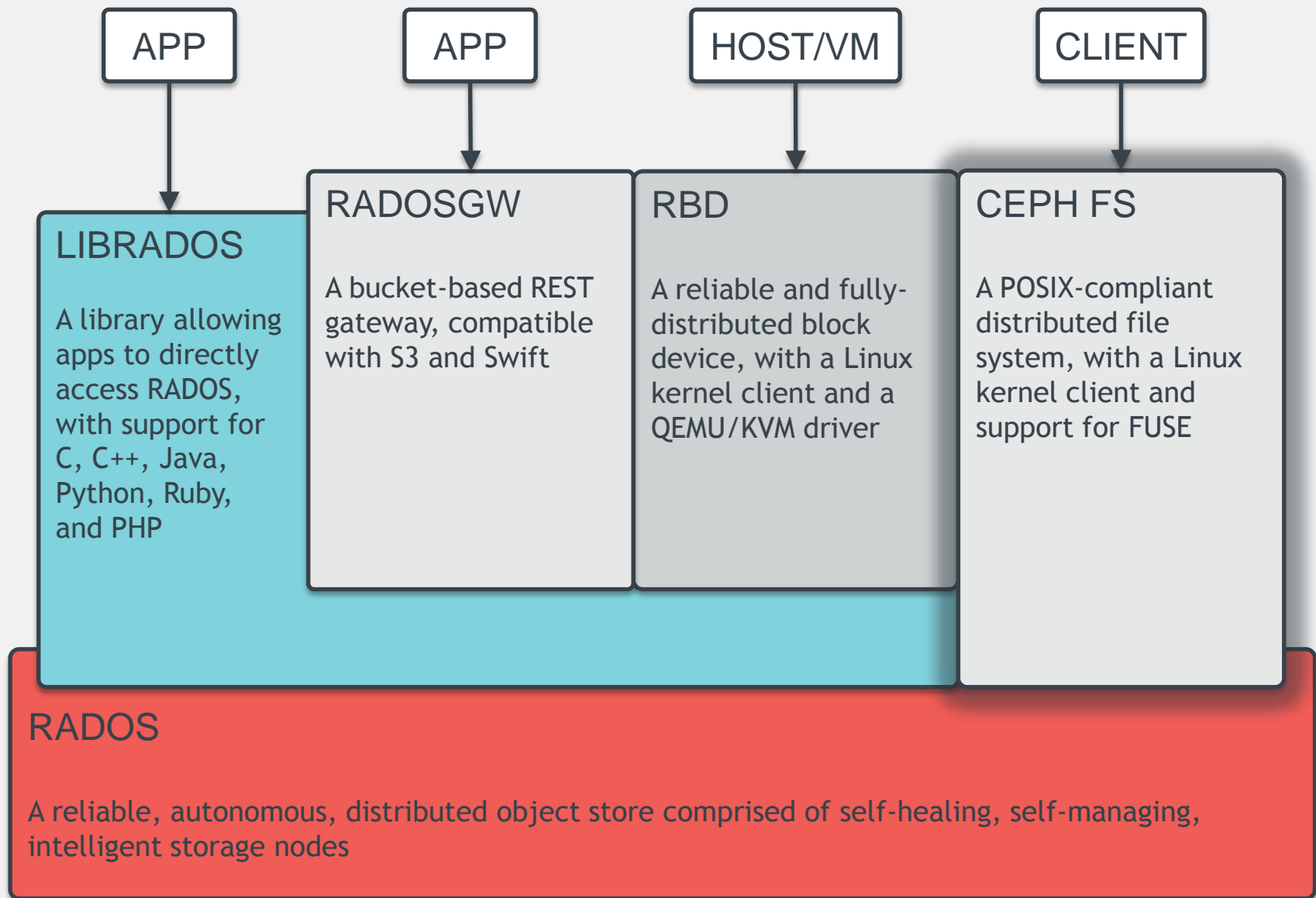


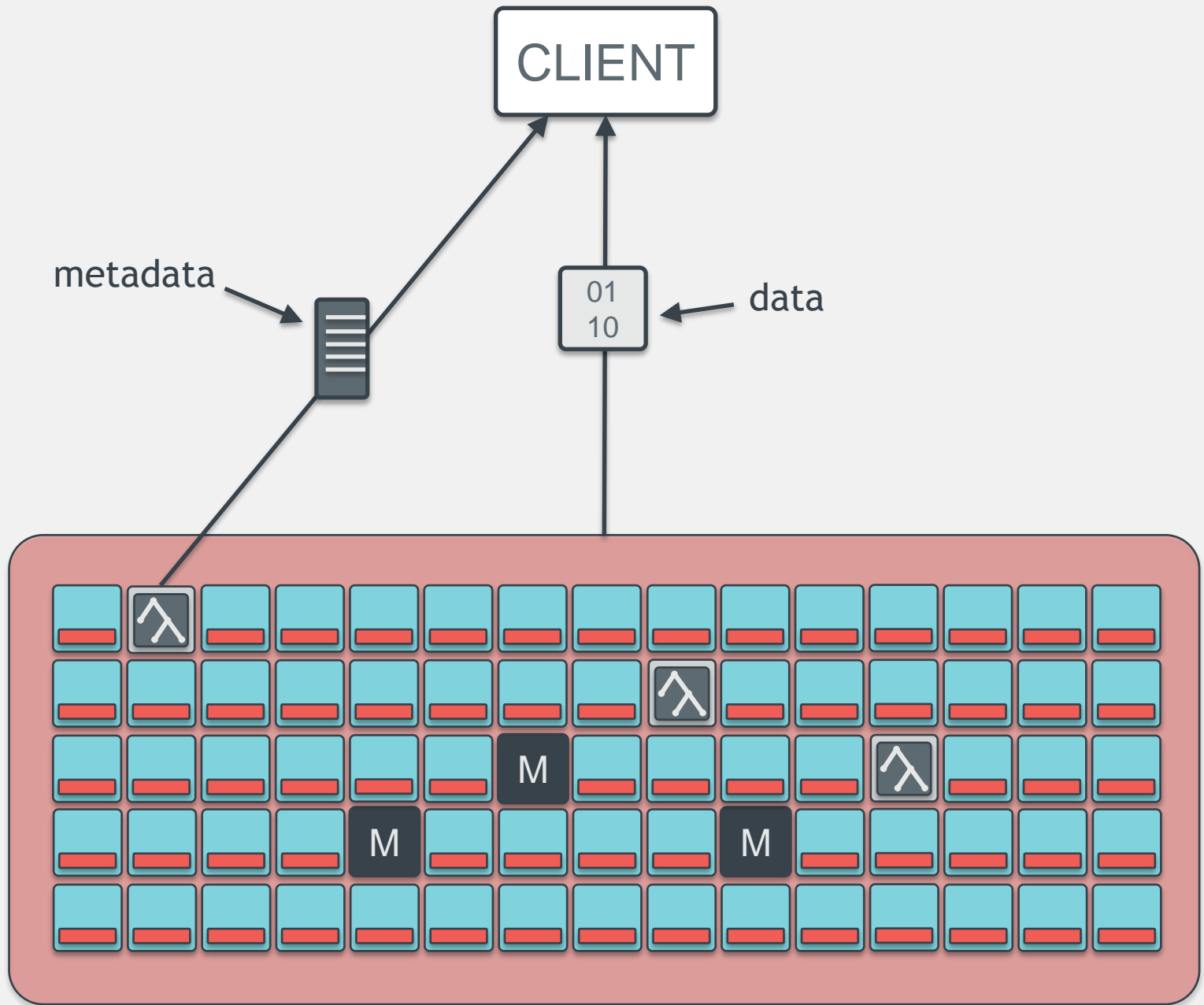


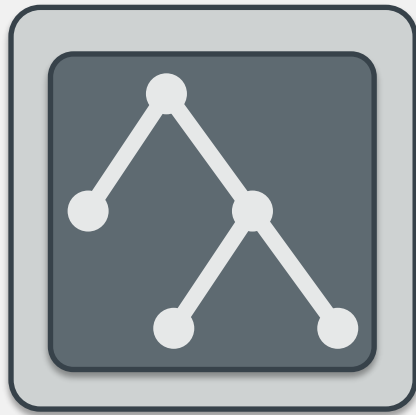


## RADOS Block Device:

- 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







## Metadata Server

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



# What Makes Ceph Unique?

Part one: CRUSH

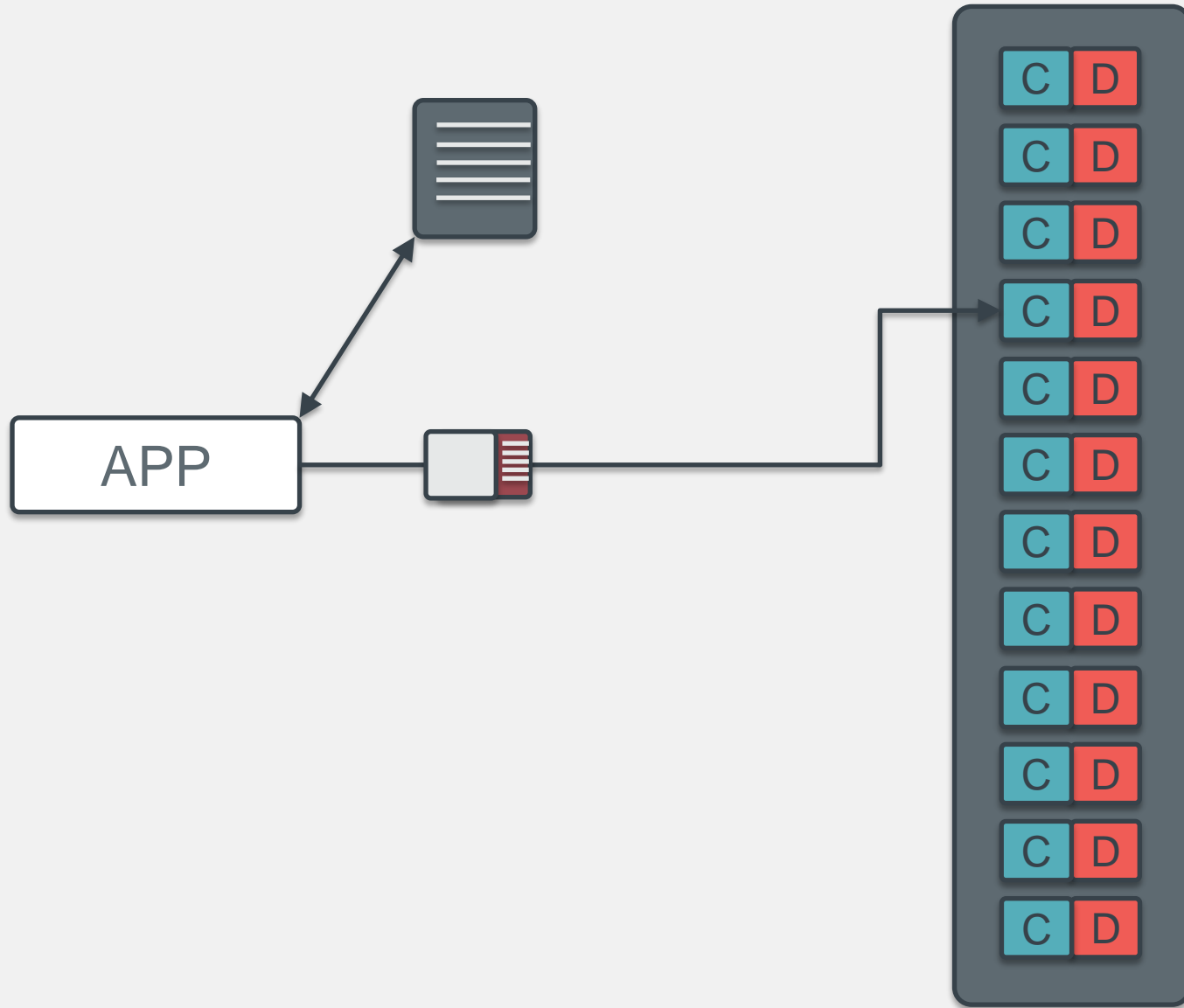






How Long Did It Take You To Find Your Keys This Morning?

azmeen, Flickr / CC BY 2.0



Jan 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 well evident to my mind that a stupid dunce or a wrong-doer cannot possibly be happy under he personally keeps a hypocritical mask unwearied to bow it head & receive the easy yoke of Christ. Man may, out of the perverseness of his heart, refuse the pursuit of God's written revelation & thus endeavor to sever his soul 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 Law of God, the great Jehovah, to whom be glory, praise, Dominion & power, now & forever. Amen.

Jan 26 of 1870. I have been very busy & have not had time to write. I have been very busy & have not had time to write. I have been very busy & have not had time to write.

Jan 27. The New Year of 1870. And the Old Year is gone at last gone with all its joys & fears - gone with all its happiness & sorrow. No 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 for ever into the bosom of Eternity.

Jan 28. 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 - of my untutored 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. Thank thee only 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 aim at, be my Lord's, my Country's & Truth's.

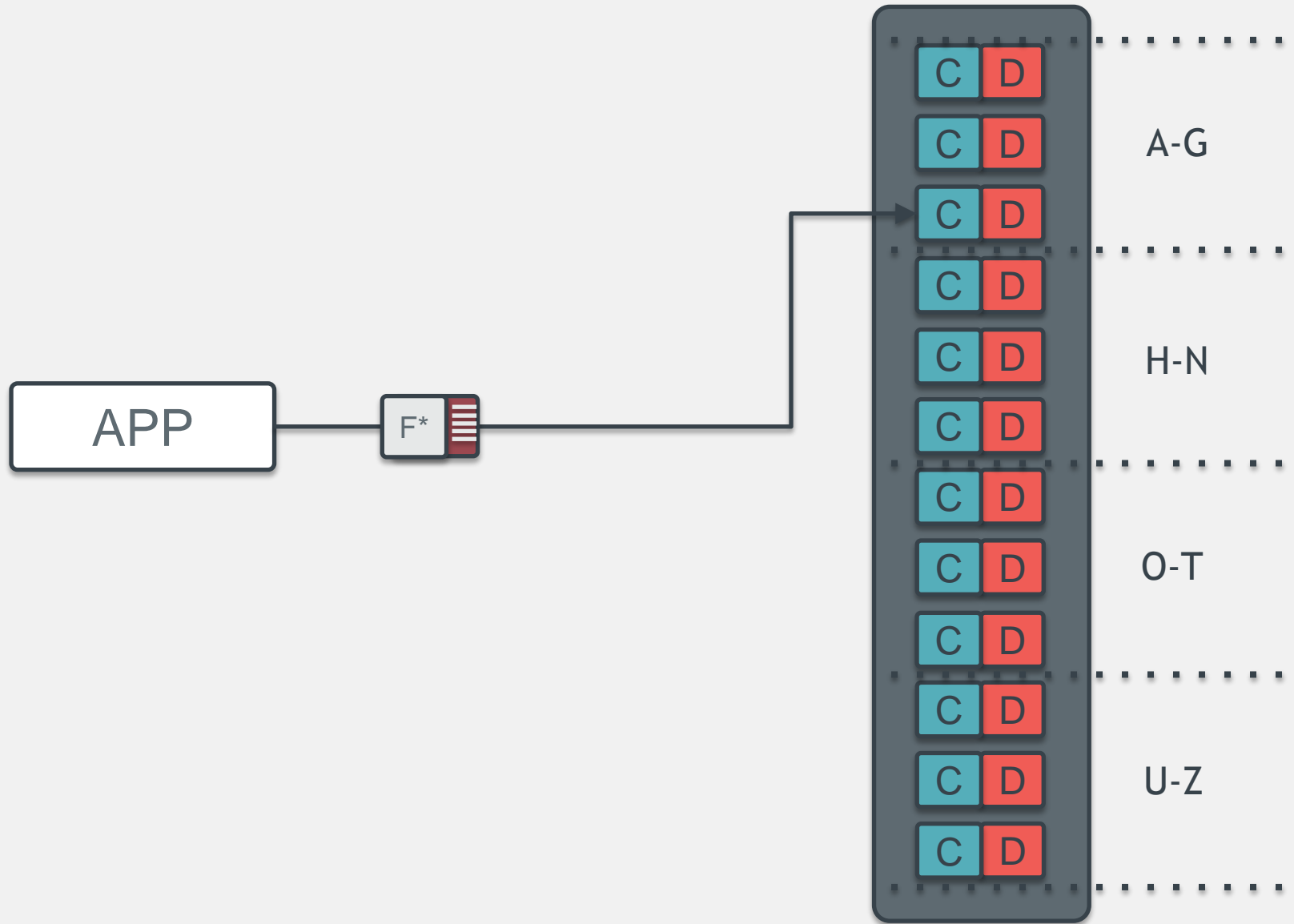
Thine Obedient Son,  
Montgomery County, Maryland

July 8. I have been very busy & have not had time to write. I have been very busy & have not had time to write. I have been very busy & have not had time to write.

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

Barnaby, Flickr / CC BY 2.0







I Always Put My Keys on the Hook By the Door

vitamindave, Flickr / CC BY 2.0



HOW DO YOU  
FIND YOUR KEYS  
WHEN YOUR HOUSE  
IS  
**INFINITELY BIG**  
AND  
**ALWAYS CHANGING?**





**The Answer: CRUSH!!!!**

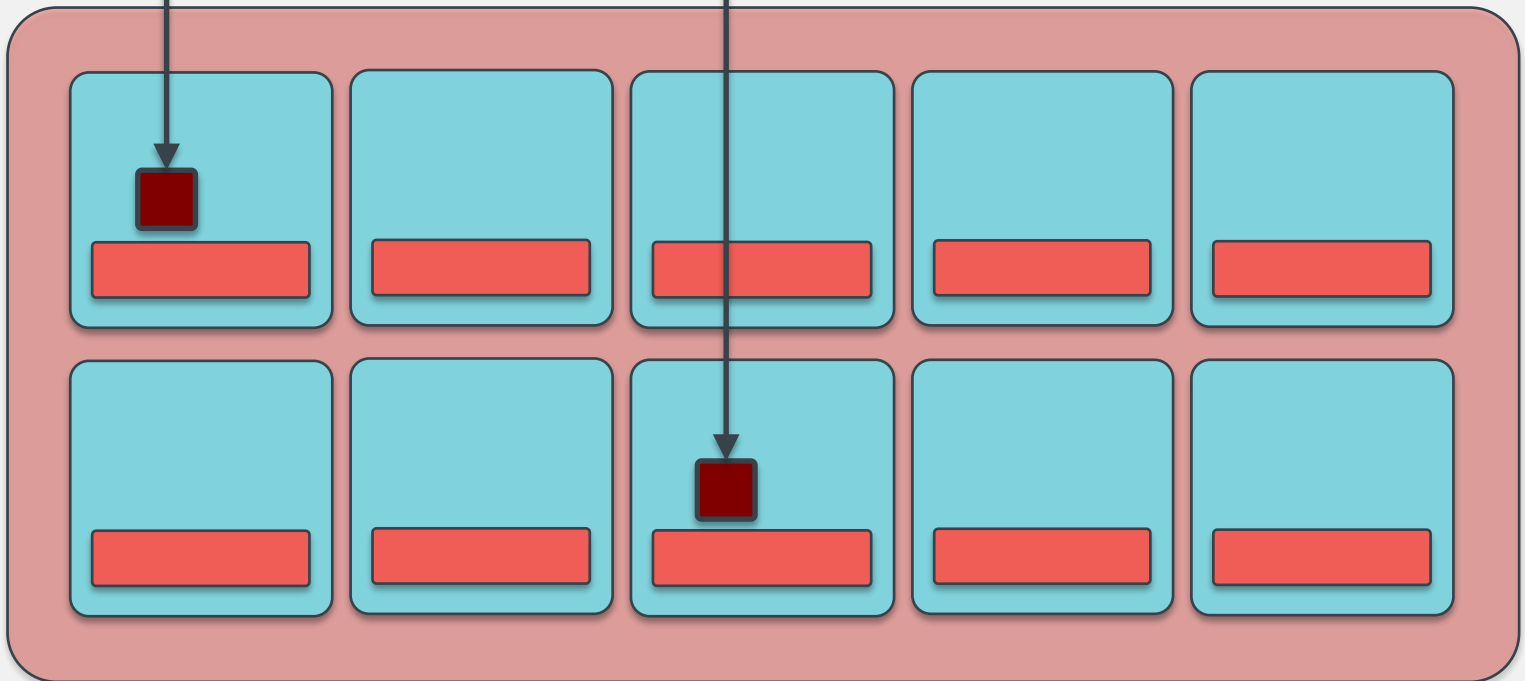
pasukaru76, Flickr / CC SA 2.0

10 10 01 01 10 10 01 11 01 10

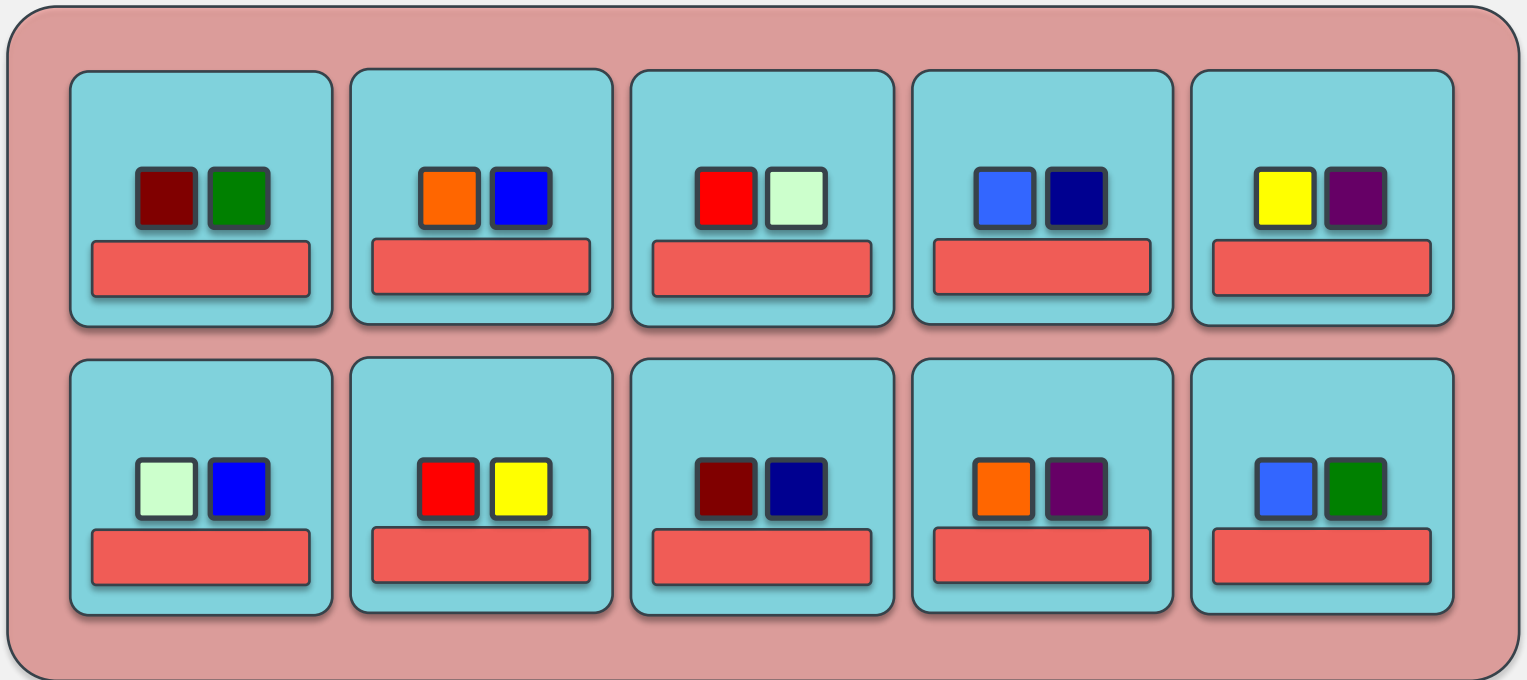
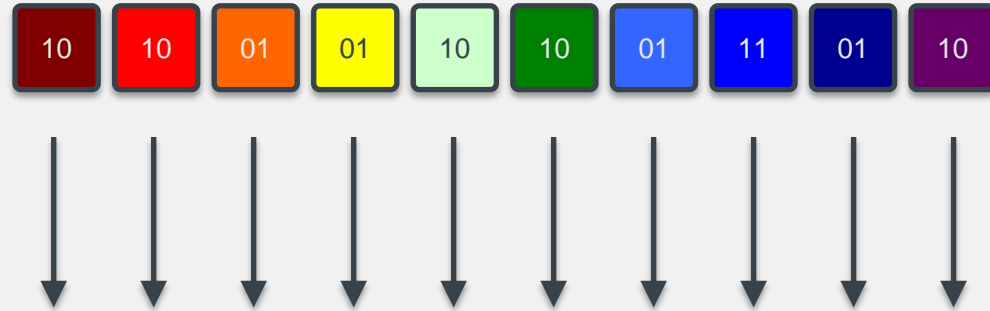
$\text{hash}(\text{object name}) \% \text{num pg}$

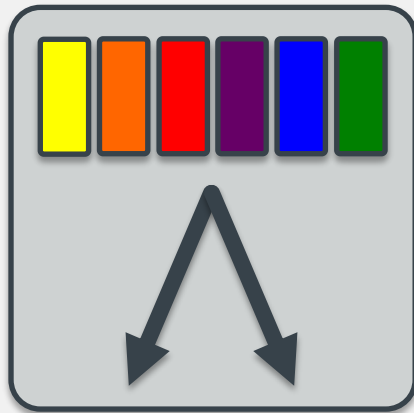


$\text{CRUSH}(\text{pg}, \text{cluster state}, \text{rule set})$



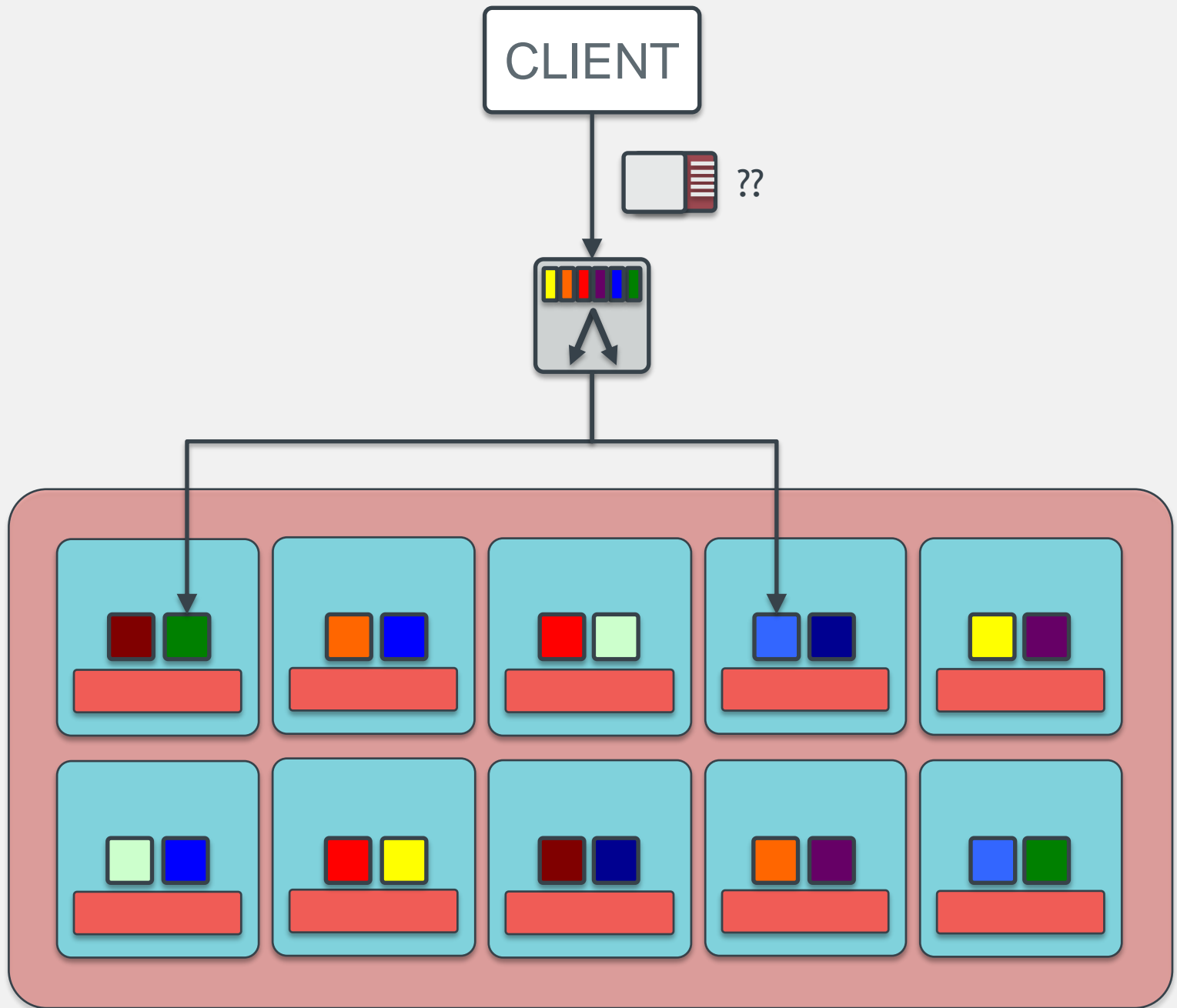
10 10 01 01 10 10 01 11 01 10

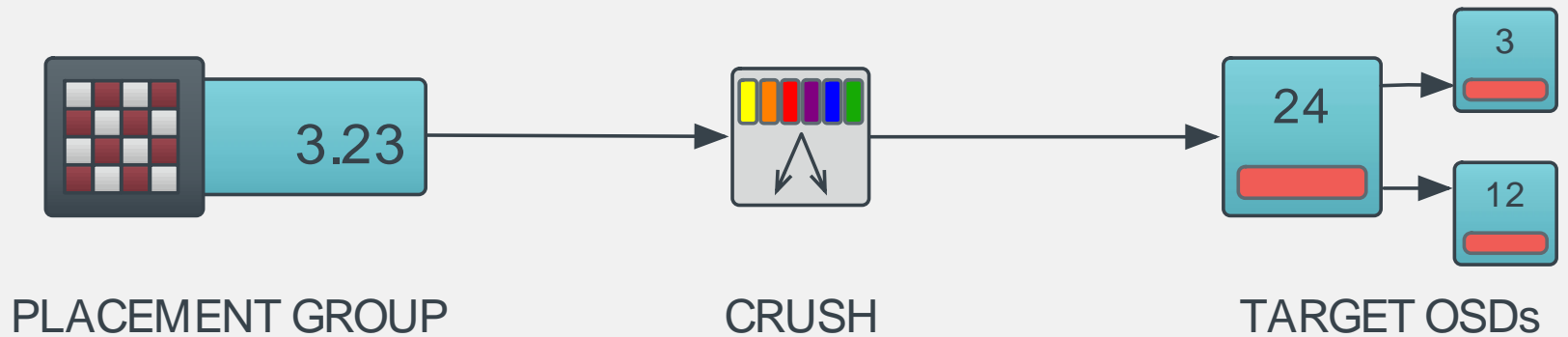
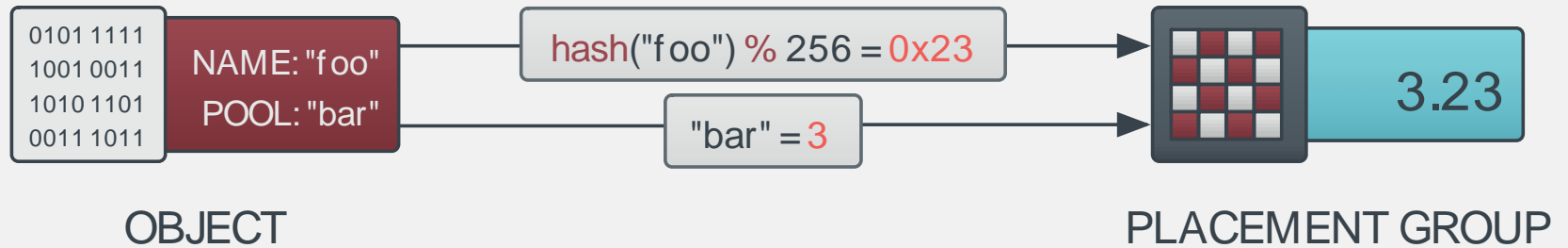


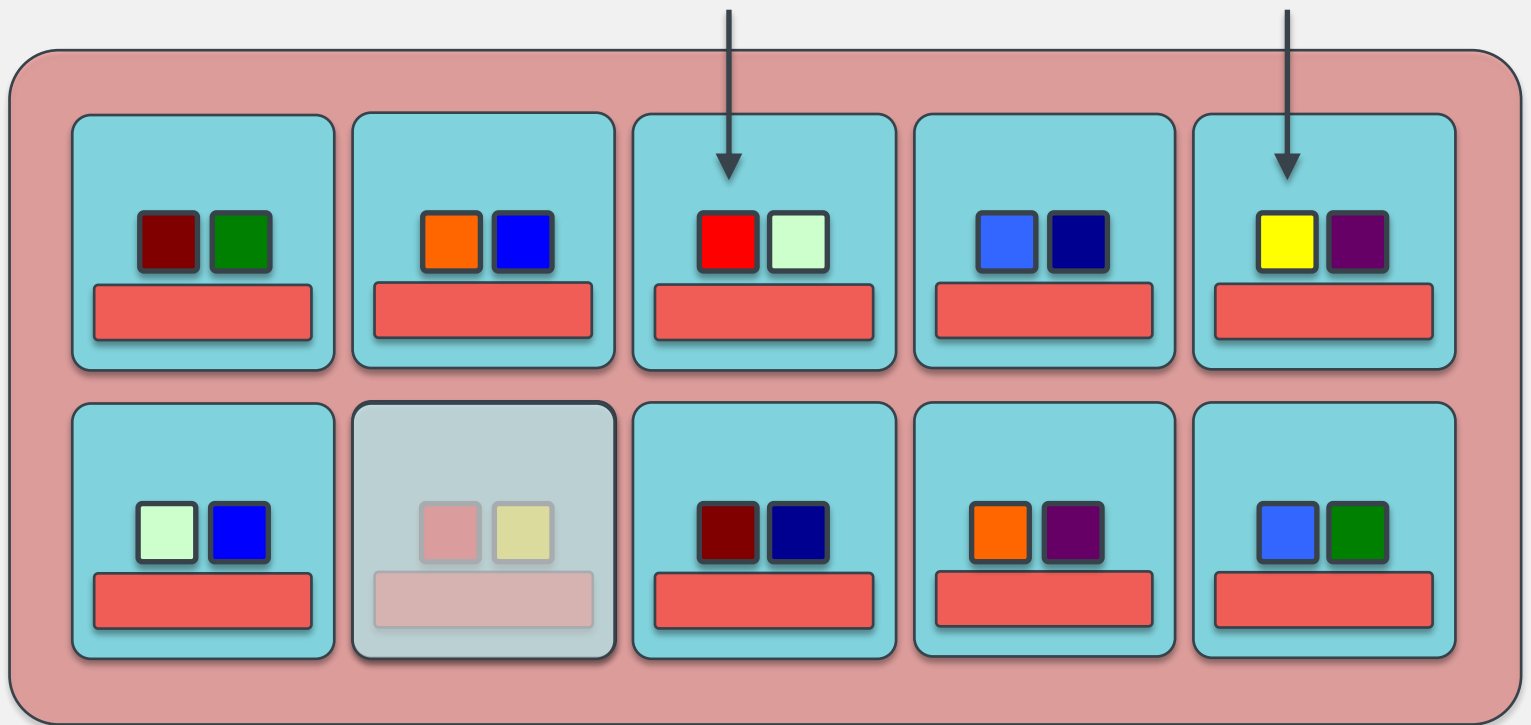


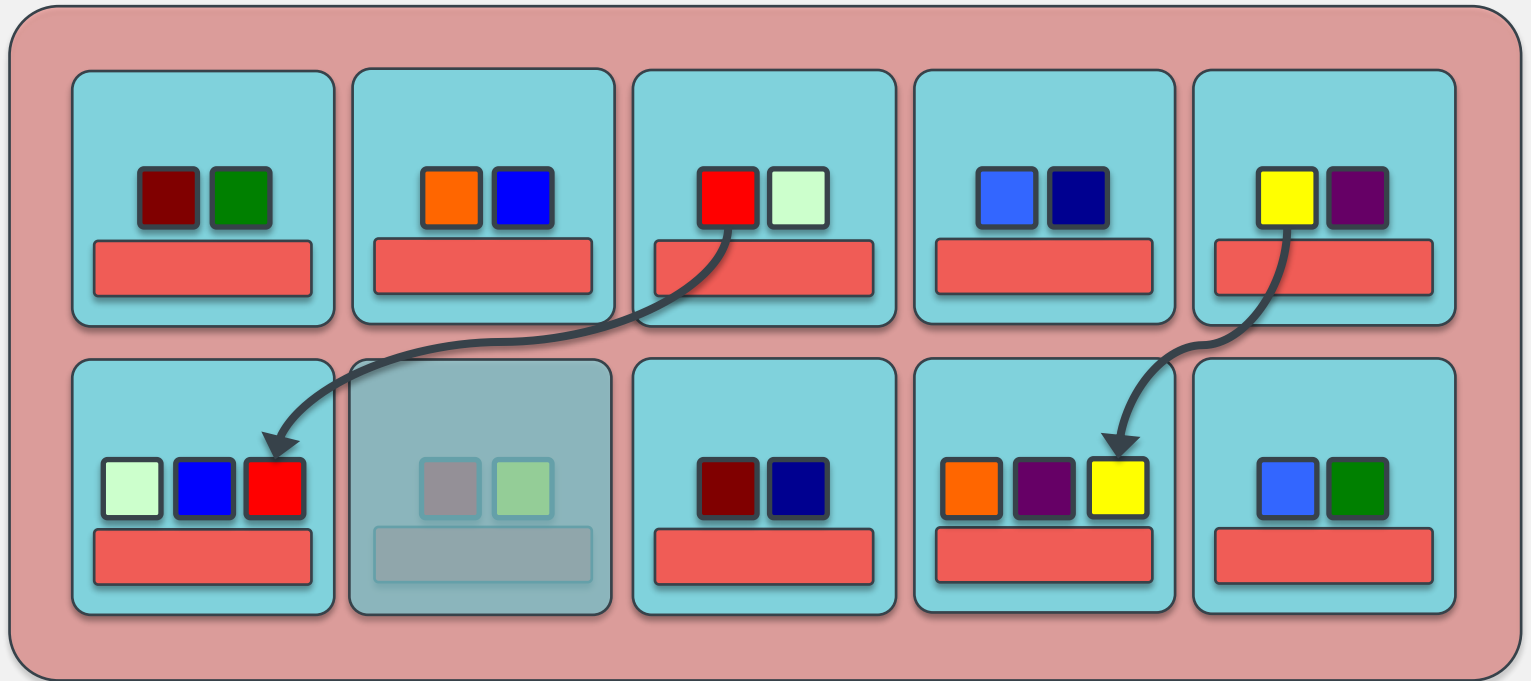
## CRUSH

- 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

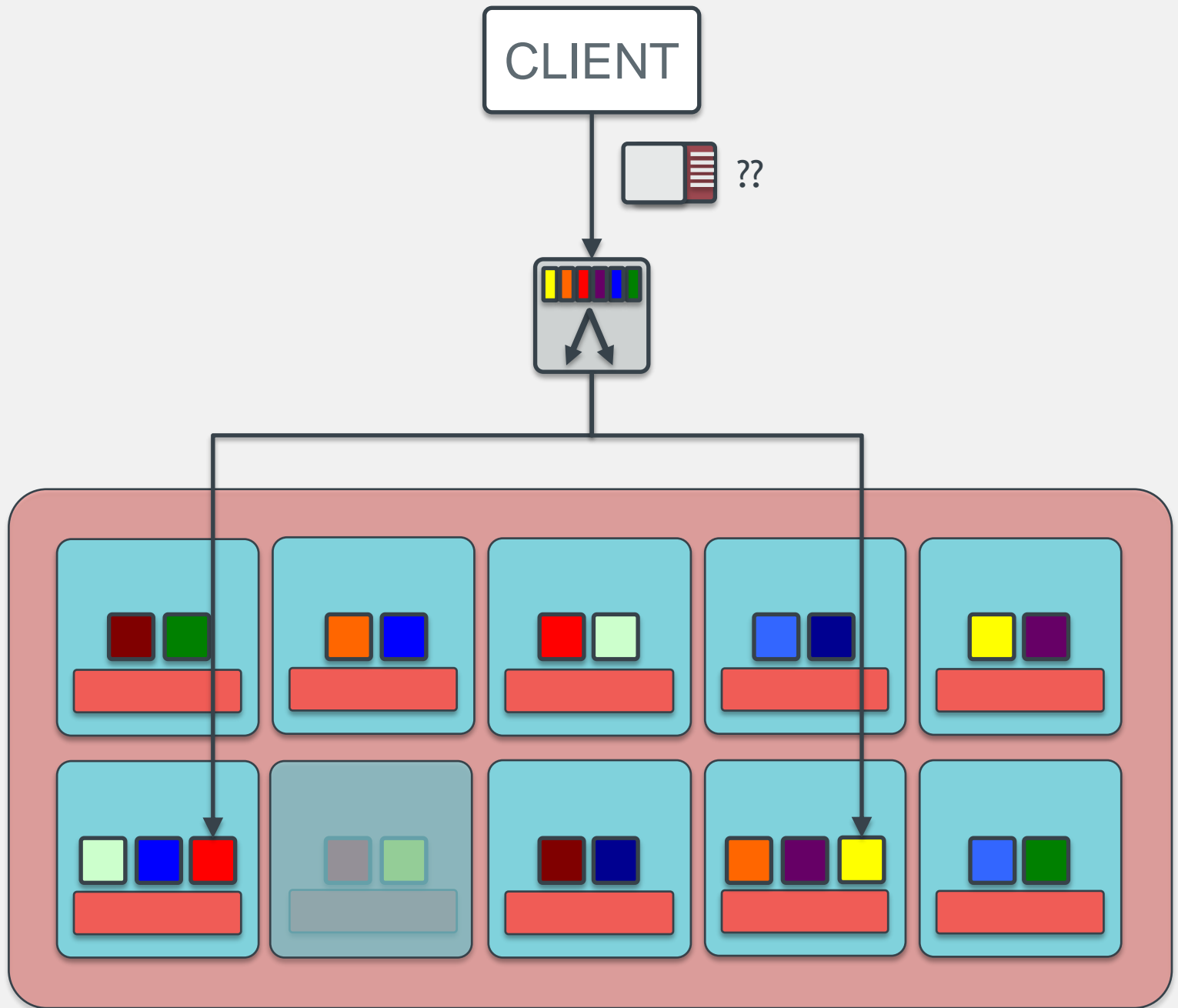








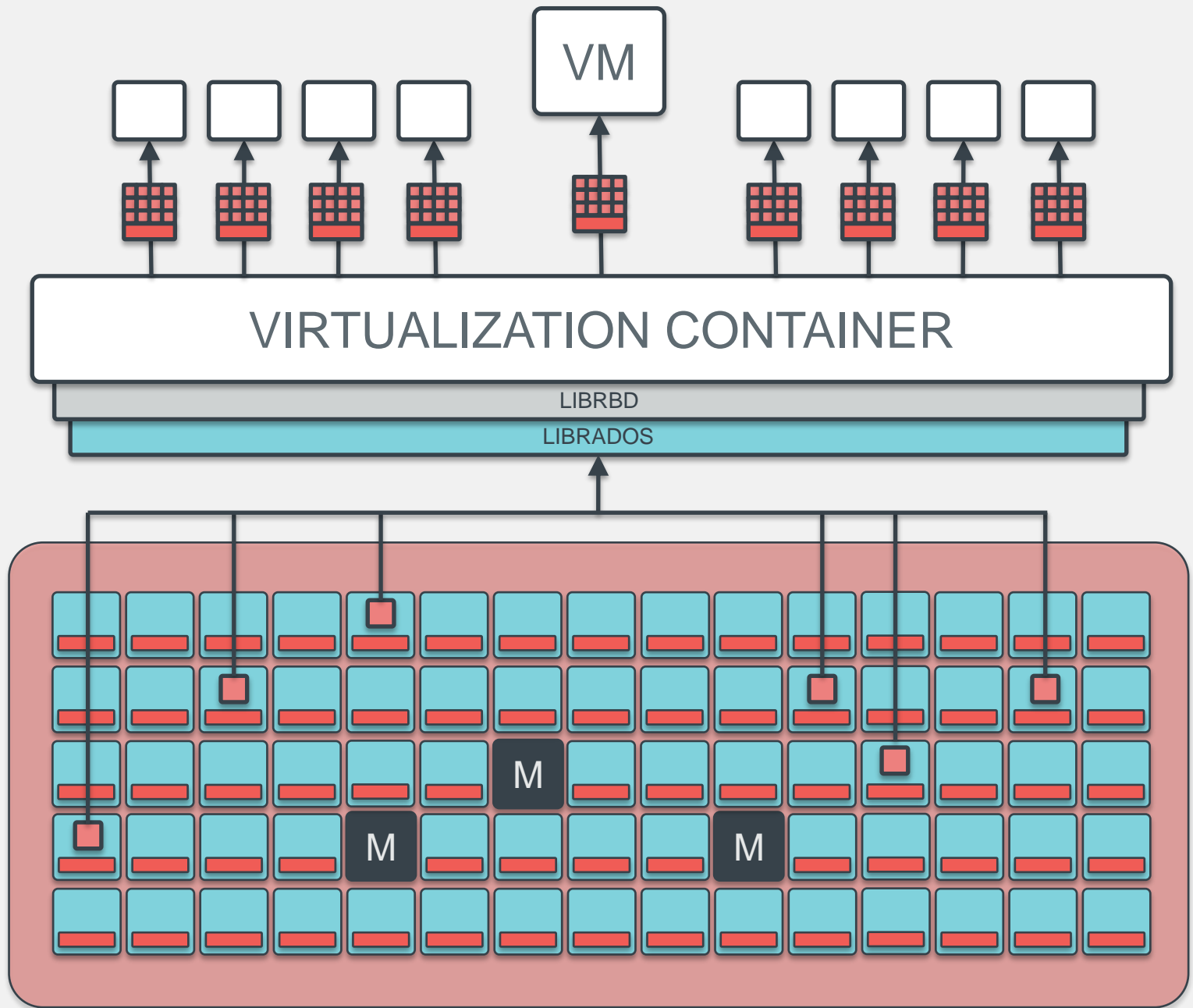





# What Makes Ceph Unique

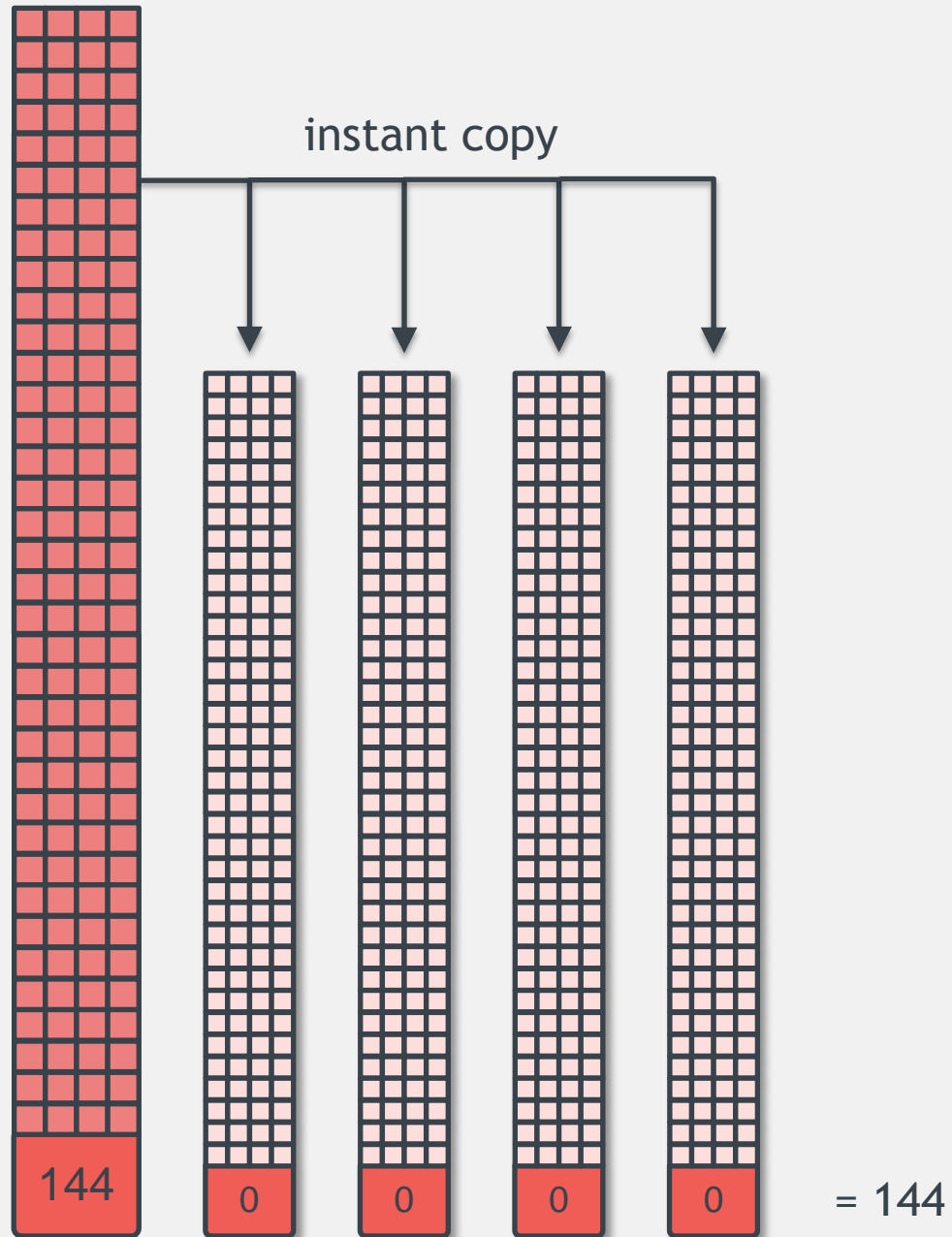
Part two: thin provisioning

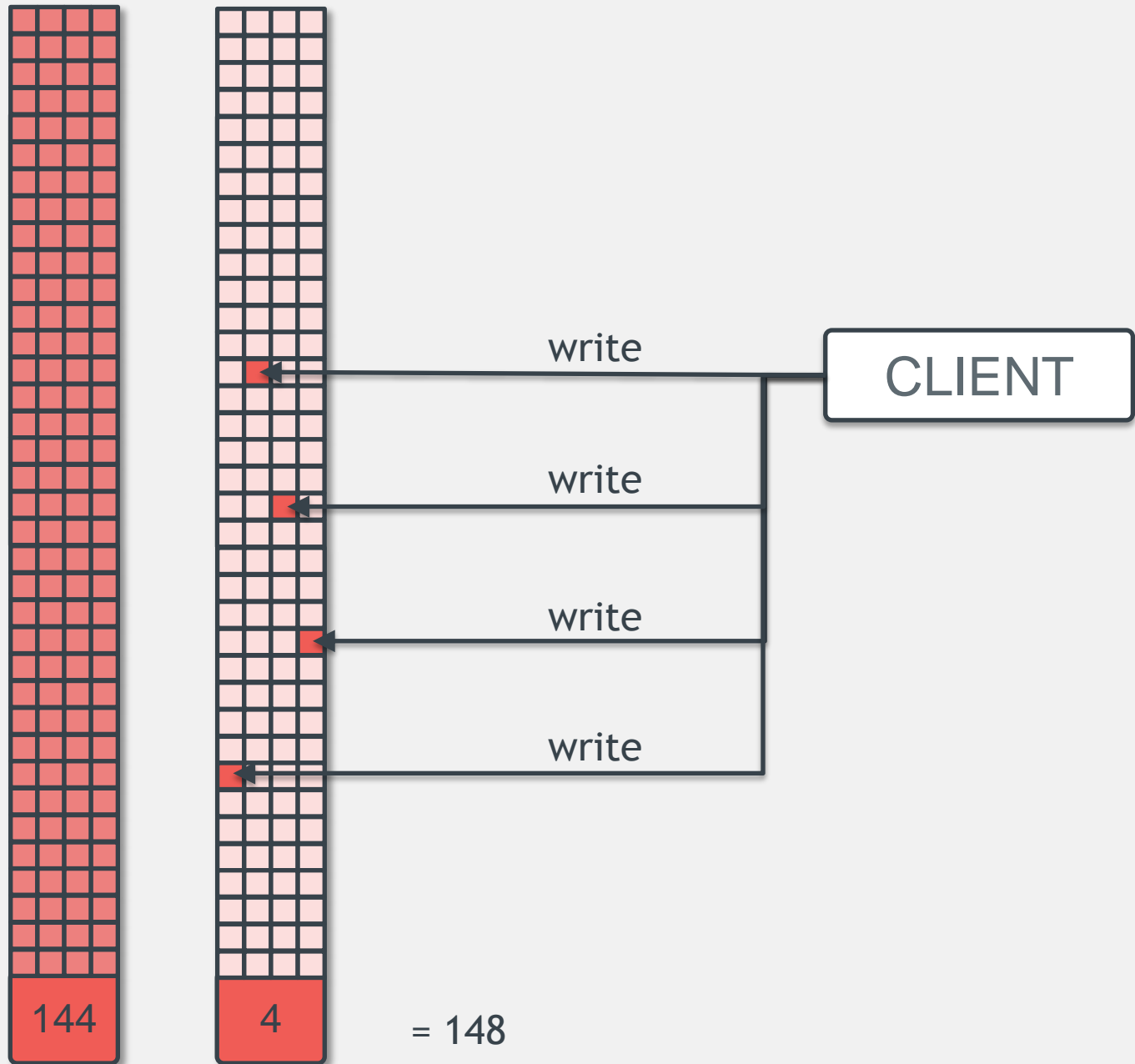


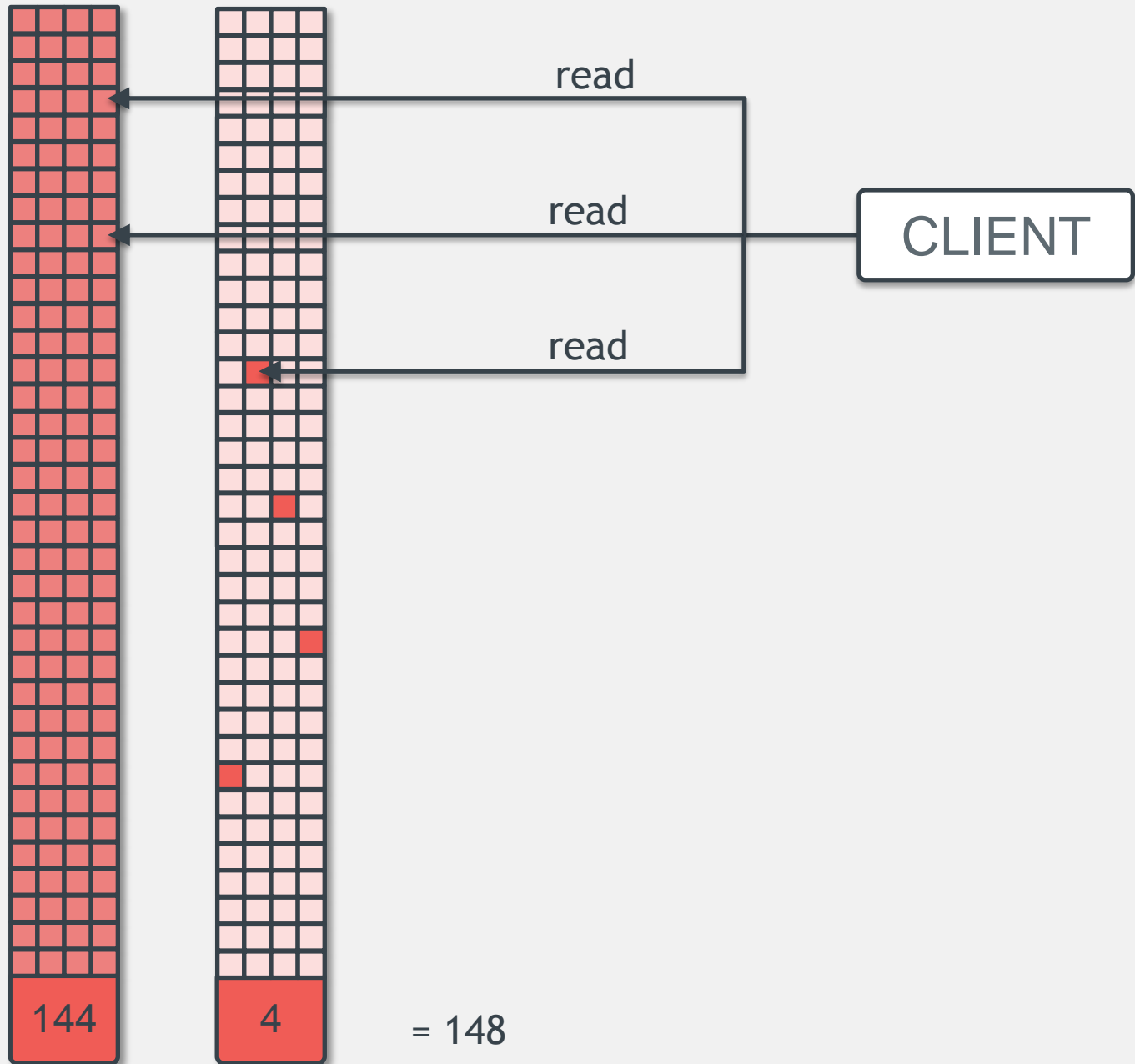




HOW DO YOU  
SPIN UP  
THOUSANDS OF VMs  
**INSTANTLY**  
AND  
**EFFICIENTLY?**







# What Makes Ceph Unique?

Part three: clustered metadata

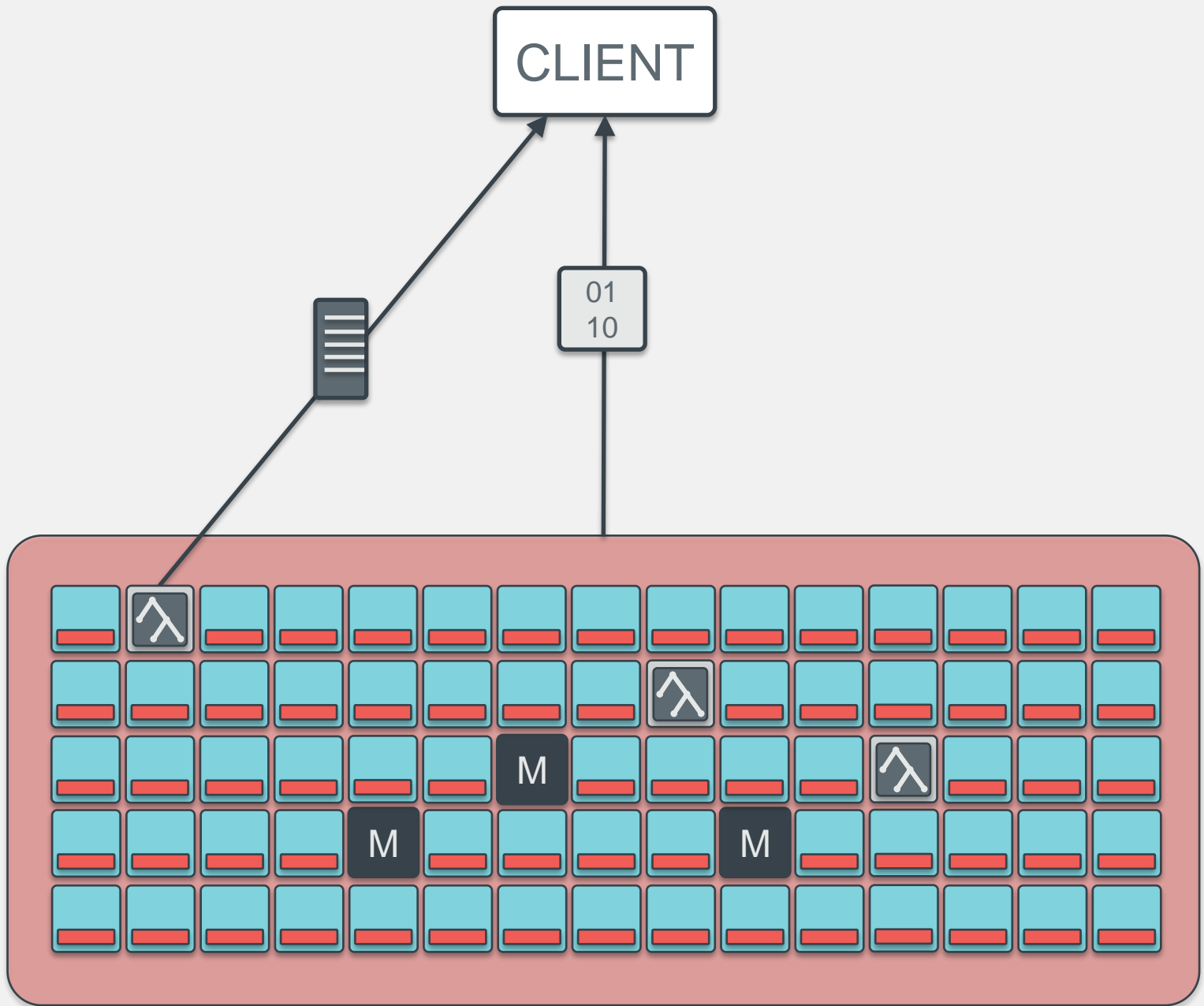


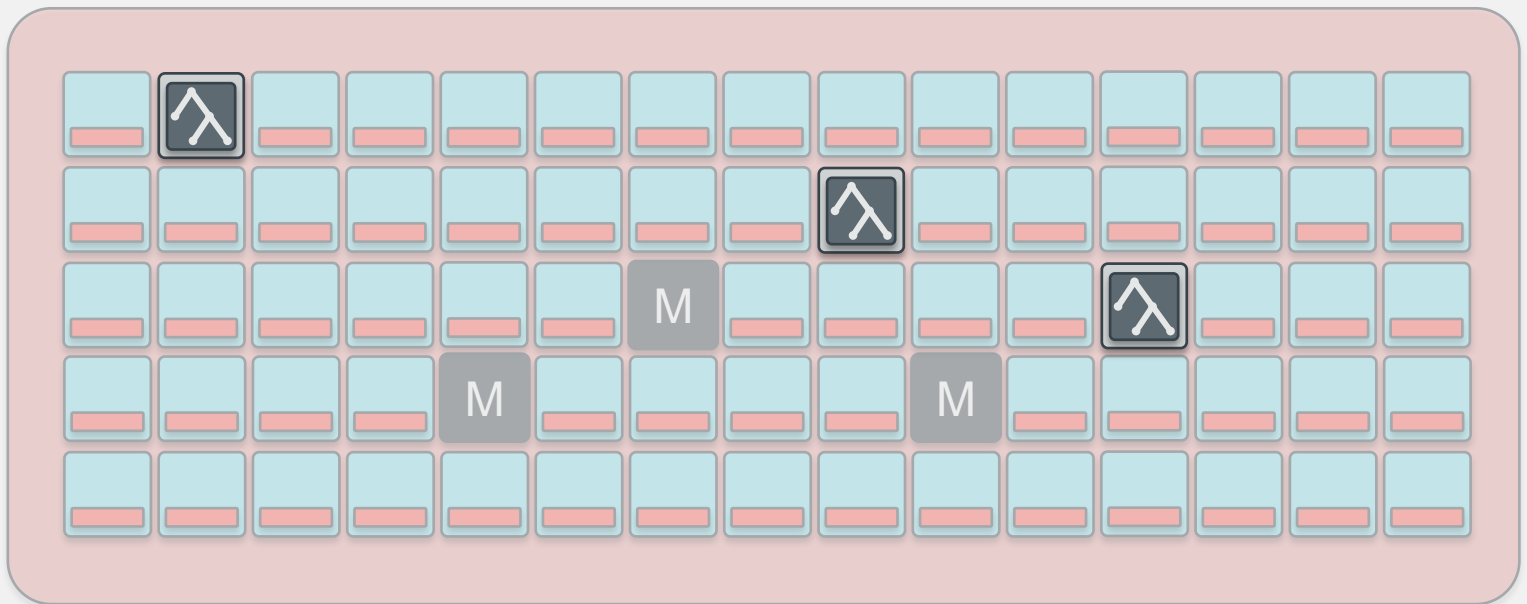


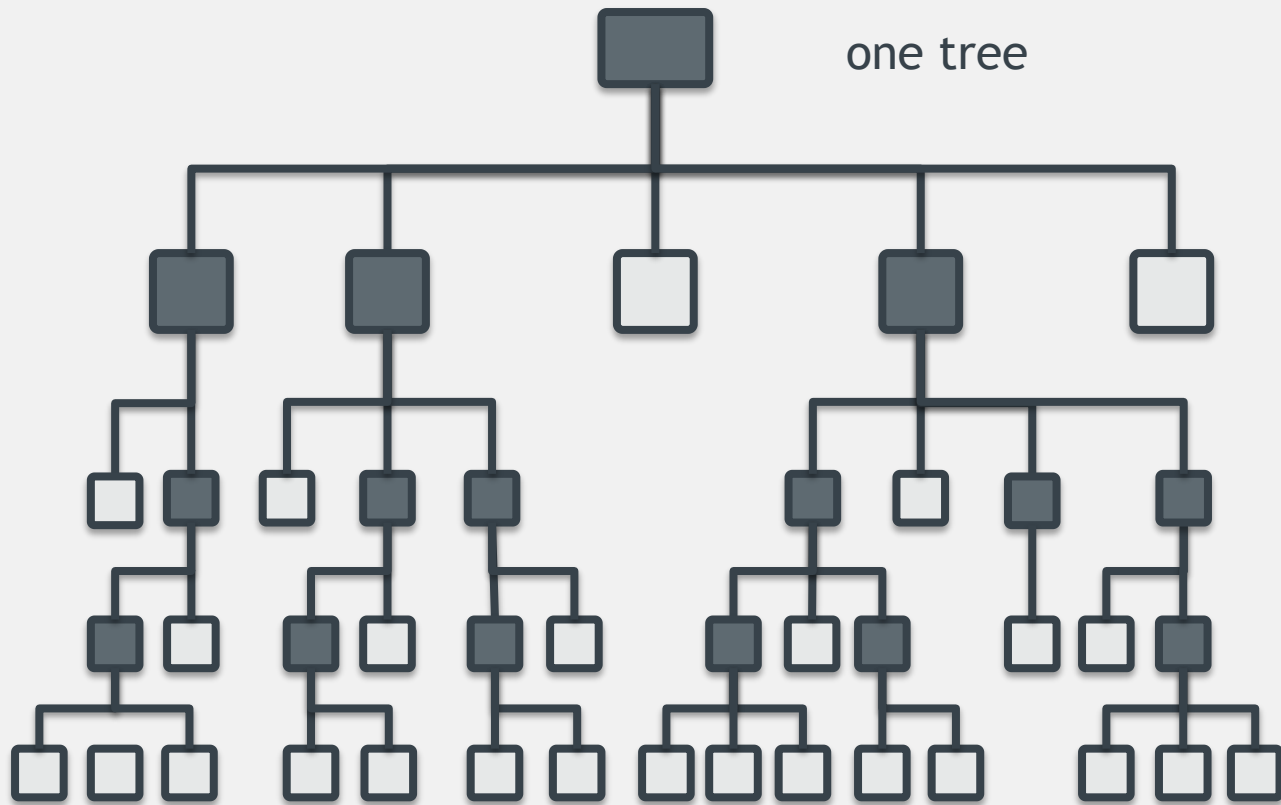
lrwxrwxrwx	1	root	root	26	Apr	26	17:54	libgssapi_krb5.so	-> mit-krb5/libgssapi_krb5.so.2.2
lrwxrwxrwx	1	root	root	21	Apr	26	17:55	libgssapi_krb5.so.2	-> libgssapi_krb5.so.2.2
-rw-r--r--	50	root	root	216824	Jul	31	2012	libgssapi_krb5.so.2.2	
lrwxrwxrwx	1	root	root	13	Apr	26	17:54	libgs.so.8	-> libgs.so.8.71
-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
lrwxrwxrwx	1	root	root	26	Apr	26	17:55	libgthread-2.0.so.0	-> libgthread-2.0.so.0.2400
-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
lrwxrwxrwx	1	root	root	26	Apr	26	17:55	libgtk-x11-2.0.so.0	-> libgtk-x11-2.0.so.0.2000
-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

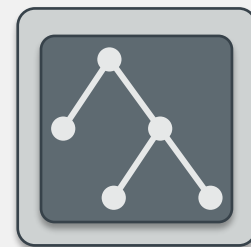
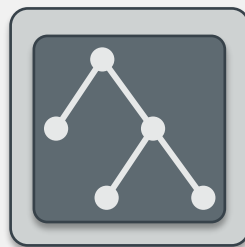
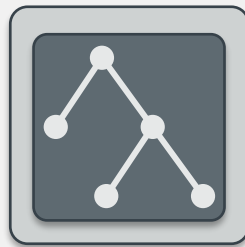
Barnaby, Flickr / CC BY 2.0



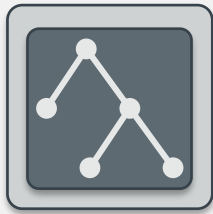


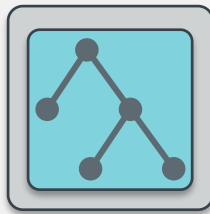
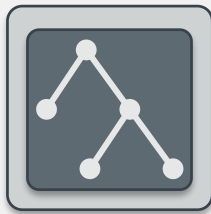
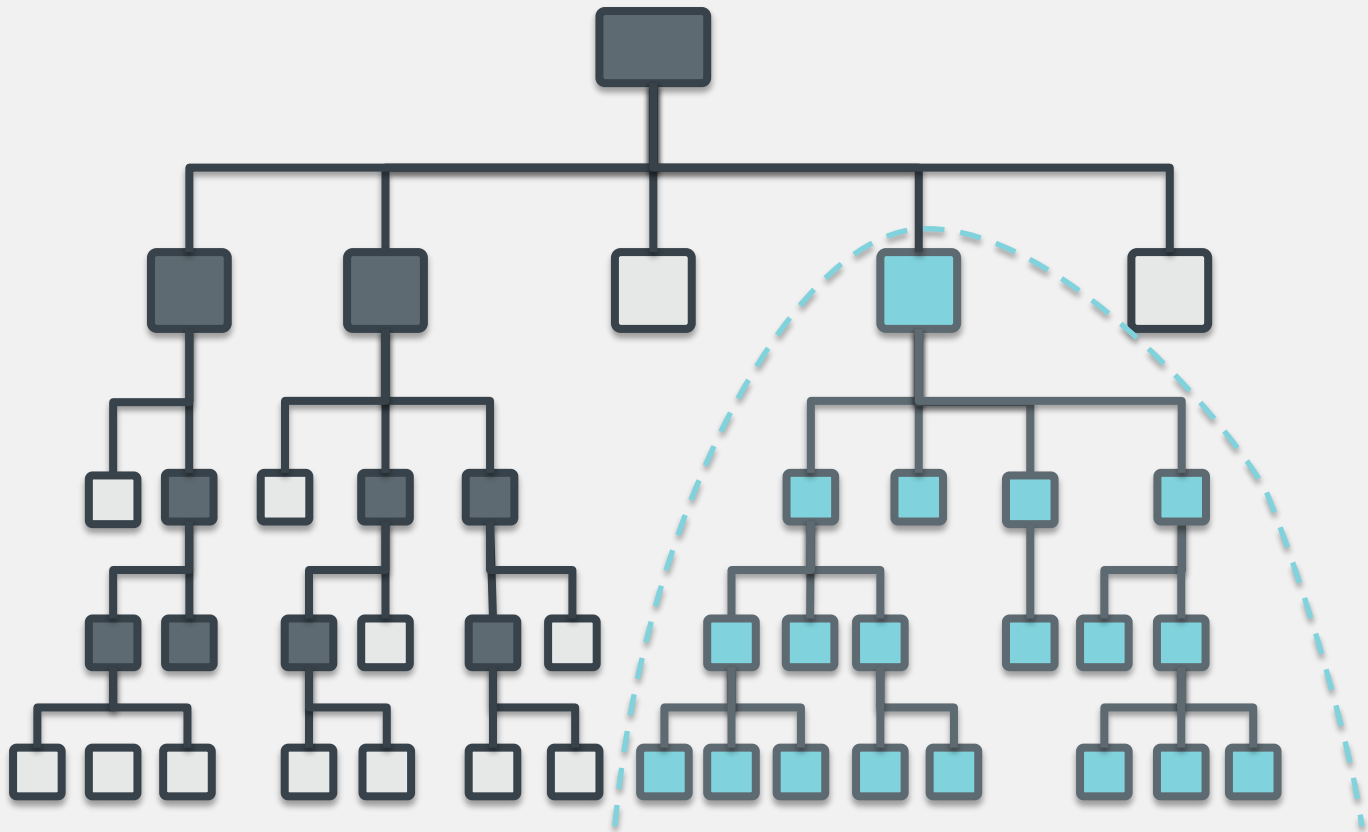


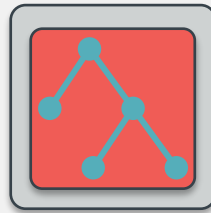
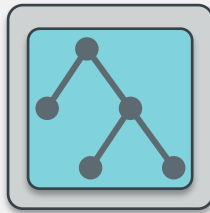
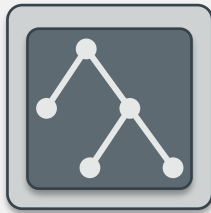
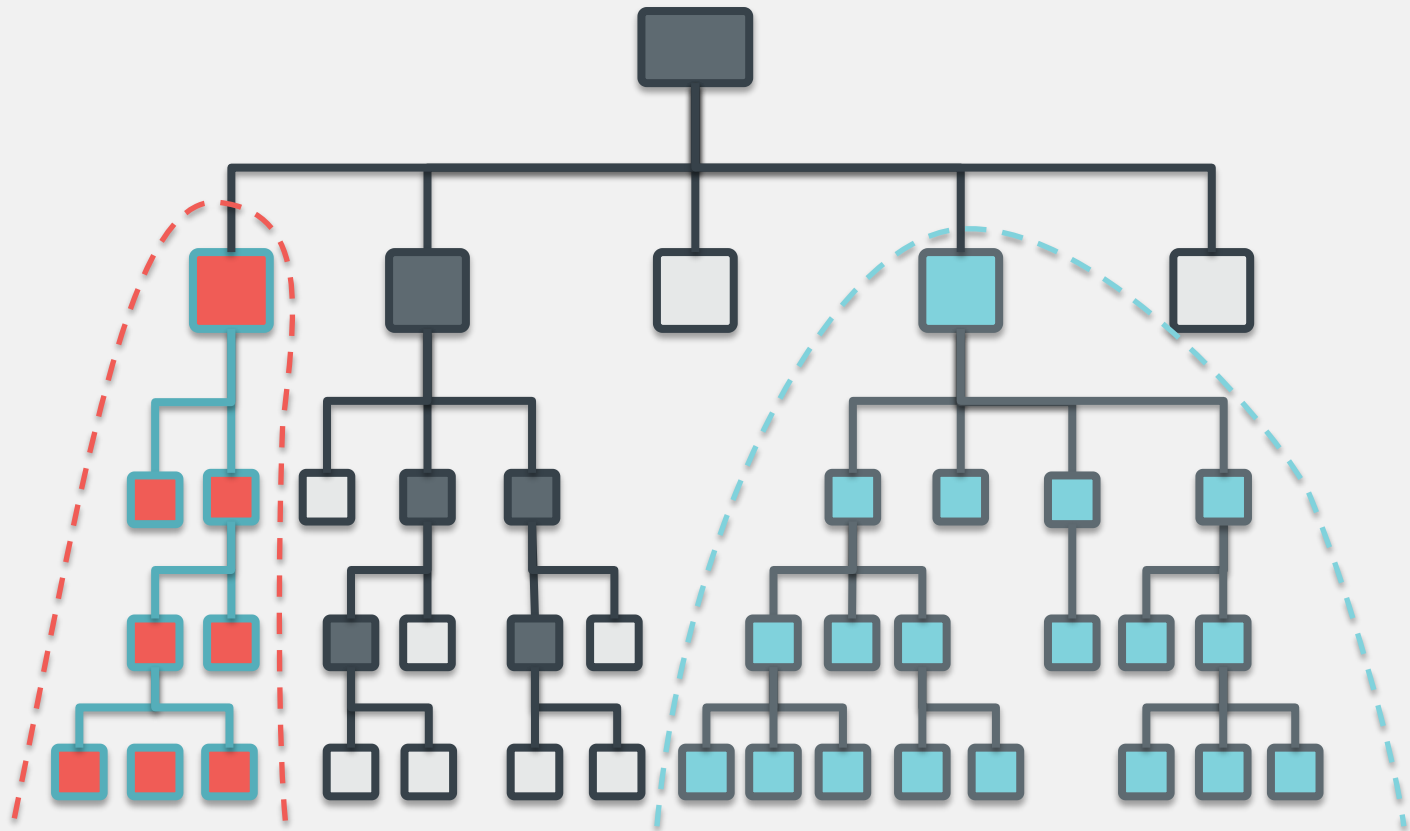
three metadata servers

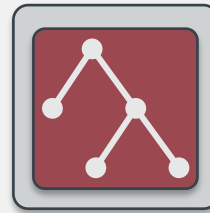
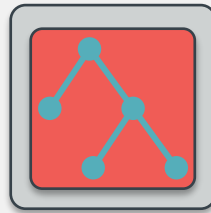
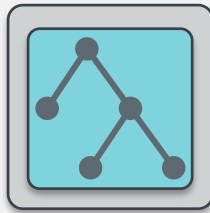
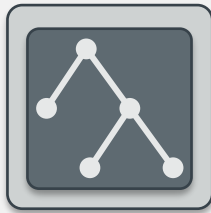
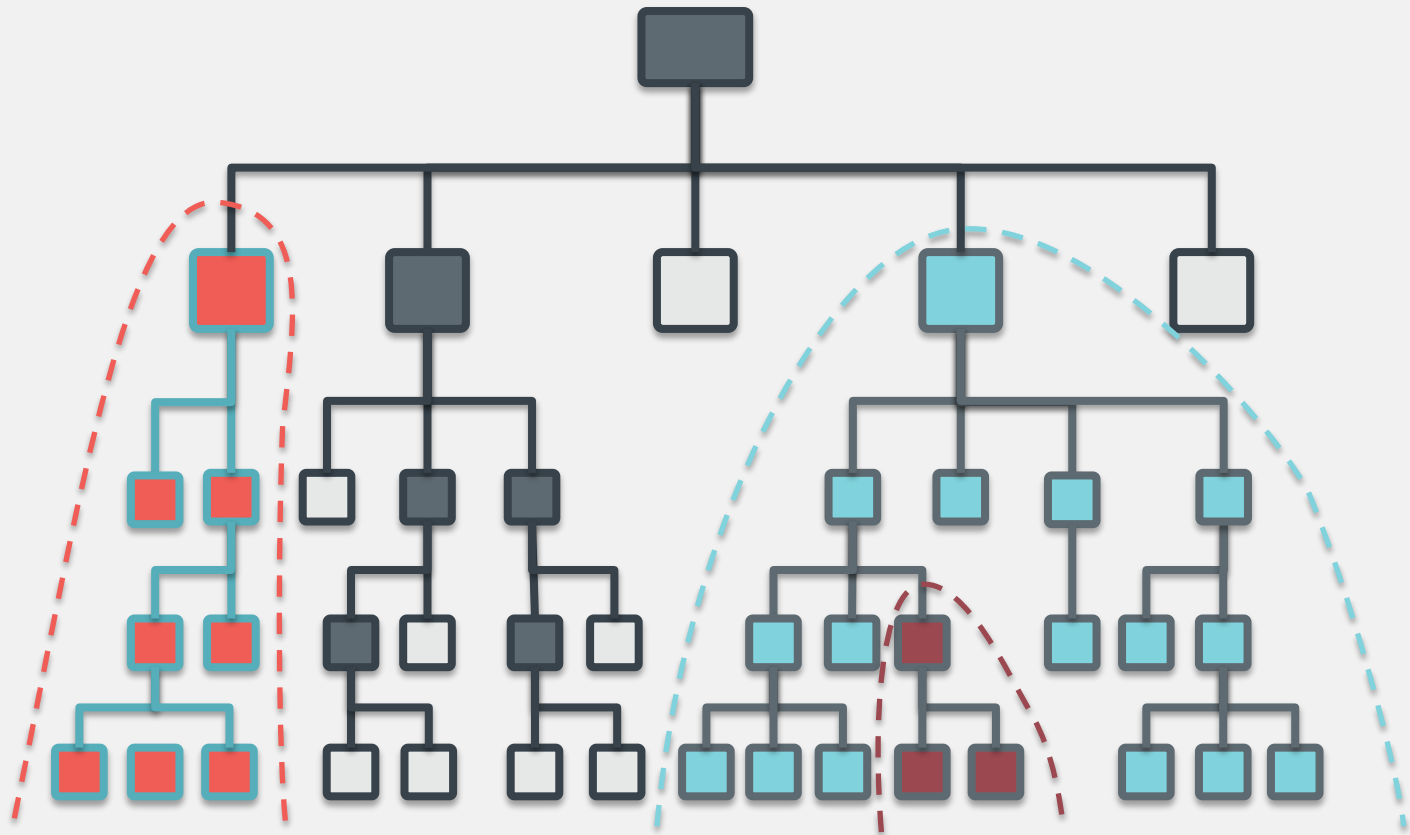


??

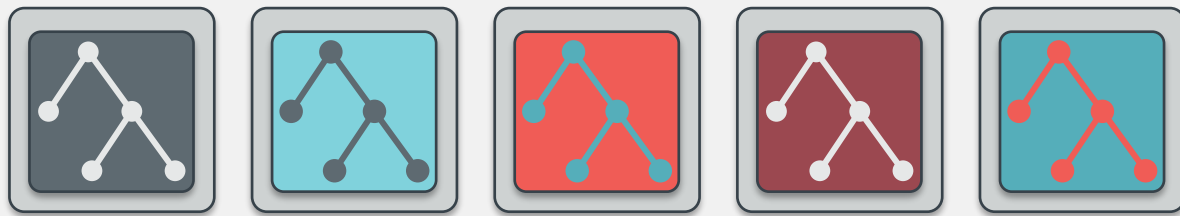
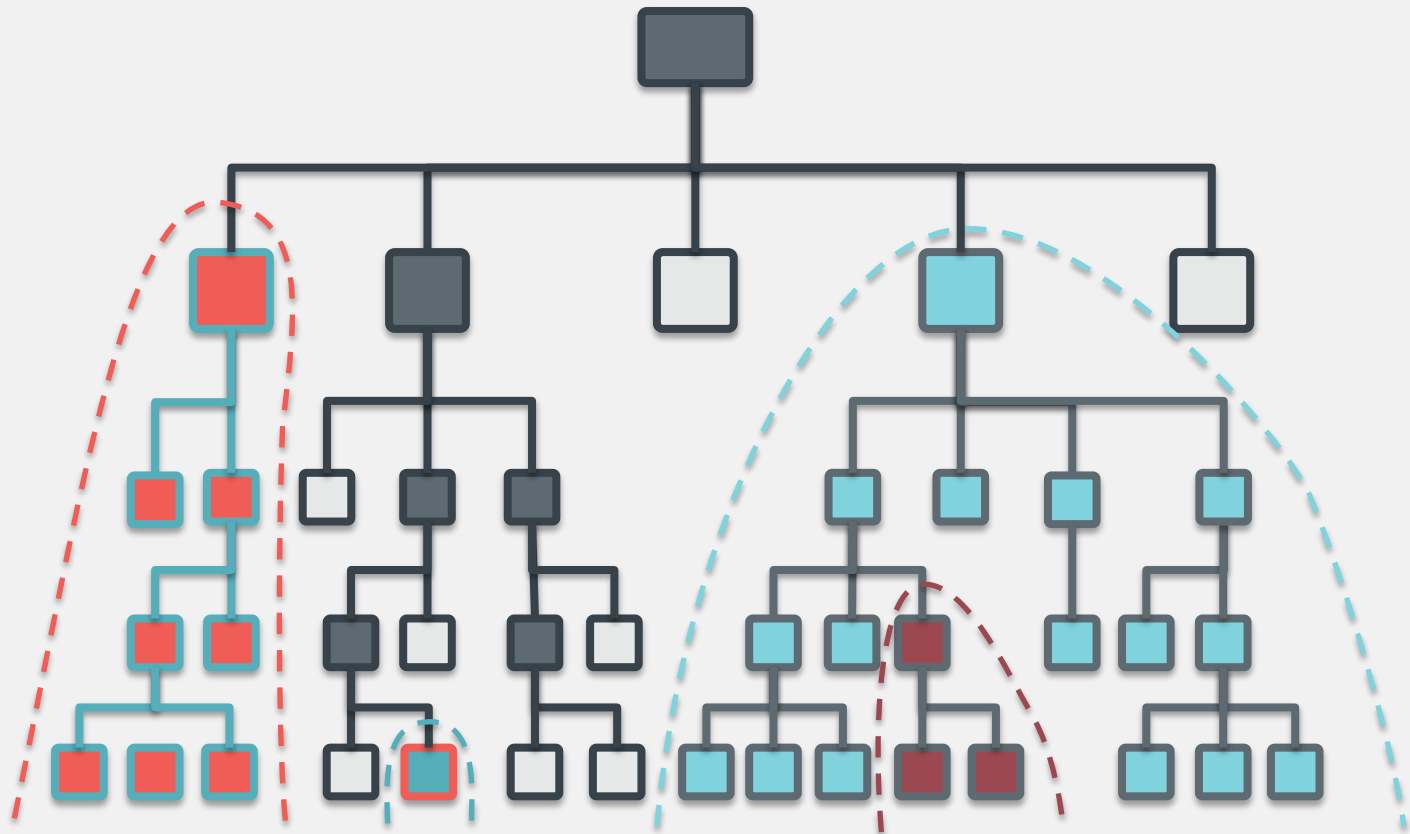












DYNAMIC SUBTREE PARTITIONING

# 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>

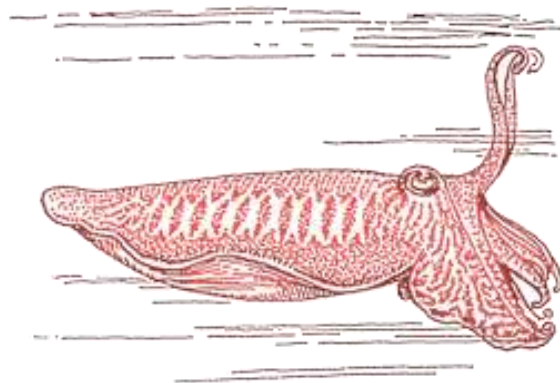
Doc updates and suggestions are always welcome.

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

# Ceph Cuttlefish (v0.61.x)

Best Ceph ever.

1. New ceph-deploy provisioning tool
2. New Chef cookbooks
3. Fully-tested packages for RHEL (in EPEL)
4. RGW authentication management API
5. RADOS pool quotas
6. New ceph df
7. RBD incremental snapshots



CUTTLEFISH  
CUTTLEFISH  
CUTTLEFISH

# Questions?

Ross Turk  
VP Community, Inktank

ross@inktank.com  
@rossturk

inktank.com | ceph.com

