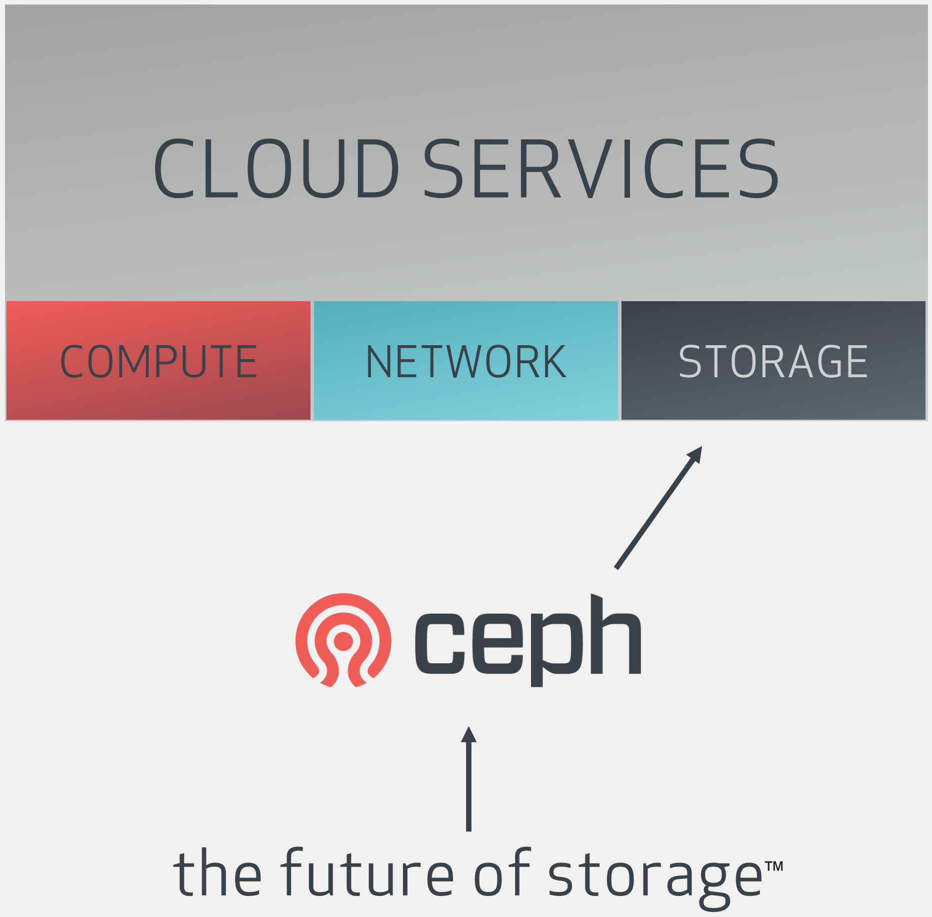
Ceph Intro & Architectural Overview

Federico Lucifredi
Product Management Director, Ceph Storage
Vancouver & Guadalajara, May 18th, 2015



CLOUD SERVICES

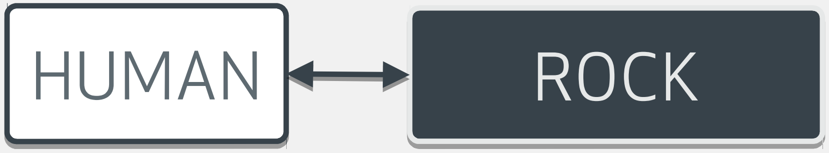
COMPUTE

NETWORK

STORAGE

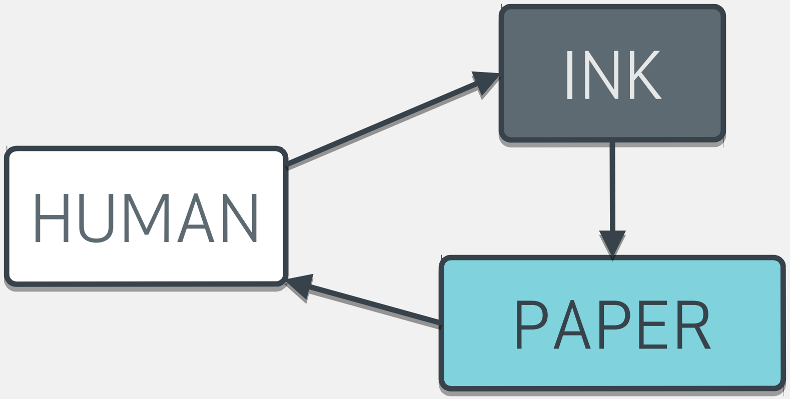
ceph

the future of storage™



HUMAN

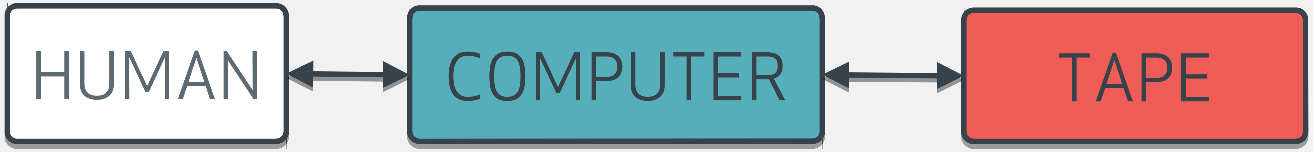
ROCK



INK

HUMAN

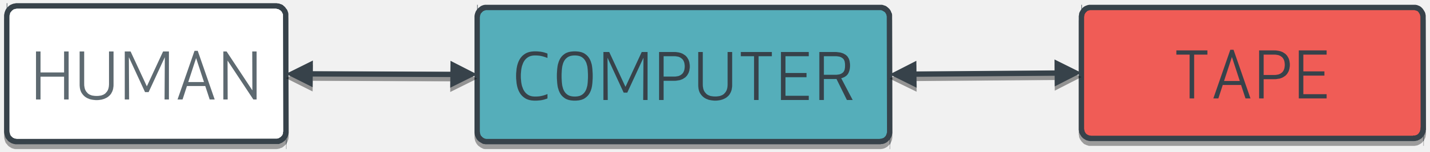
PAPER



HUMAN

COMPUTER

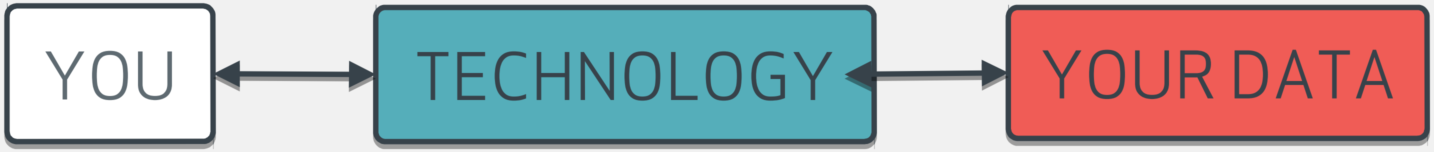
TAPE



HUMAN

COMPUTER

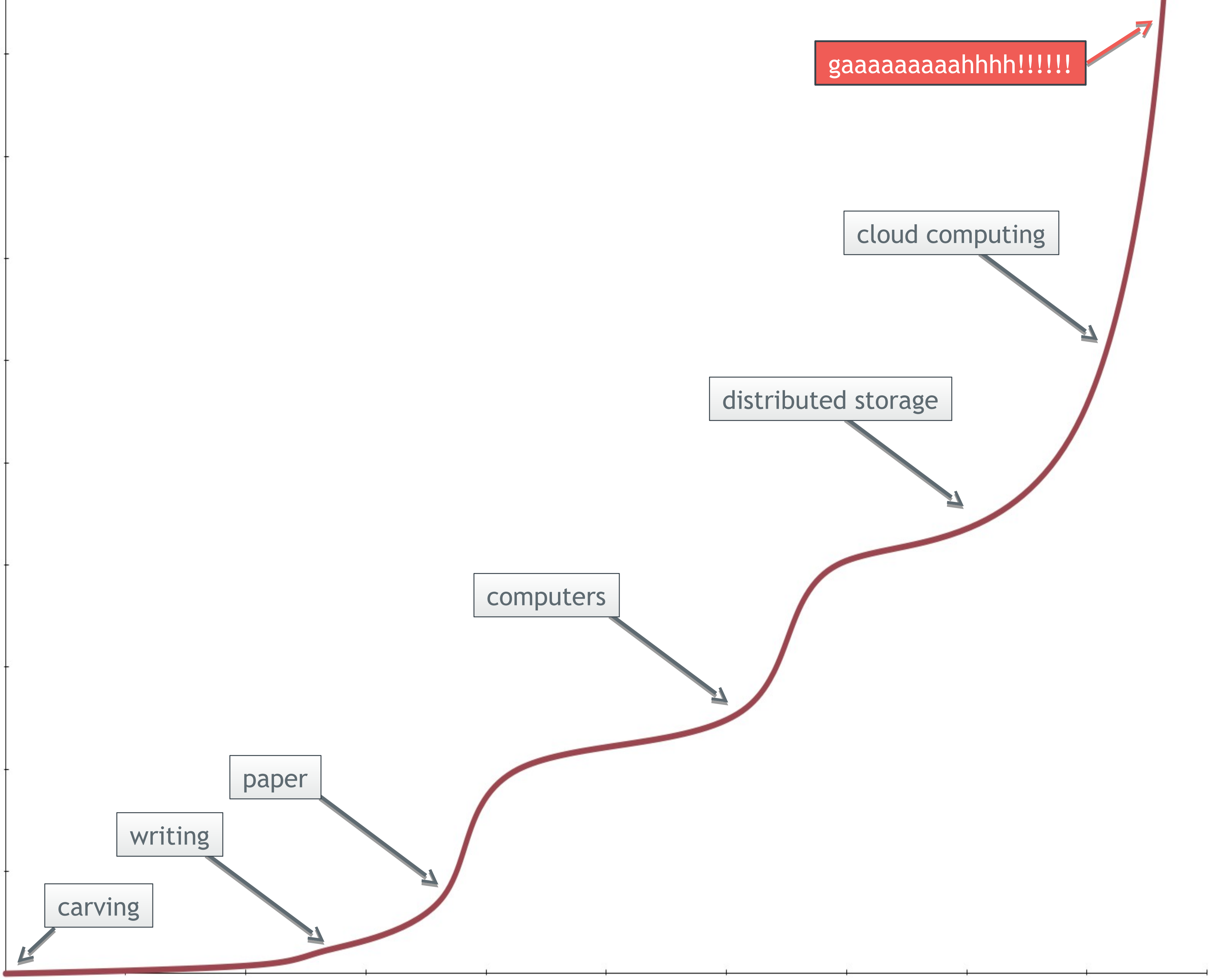
TAPE

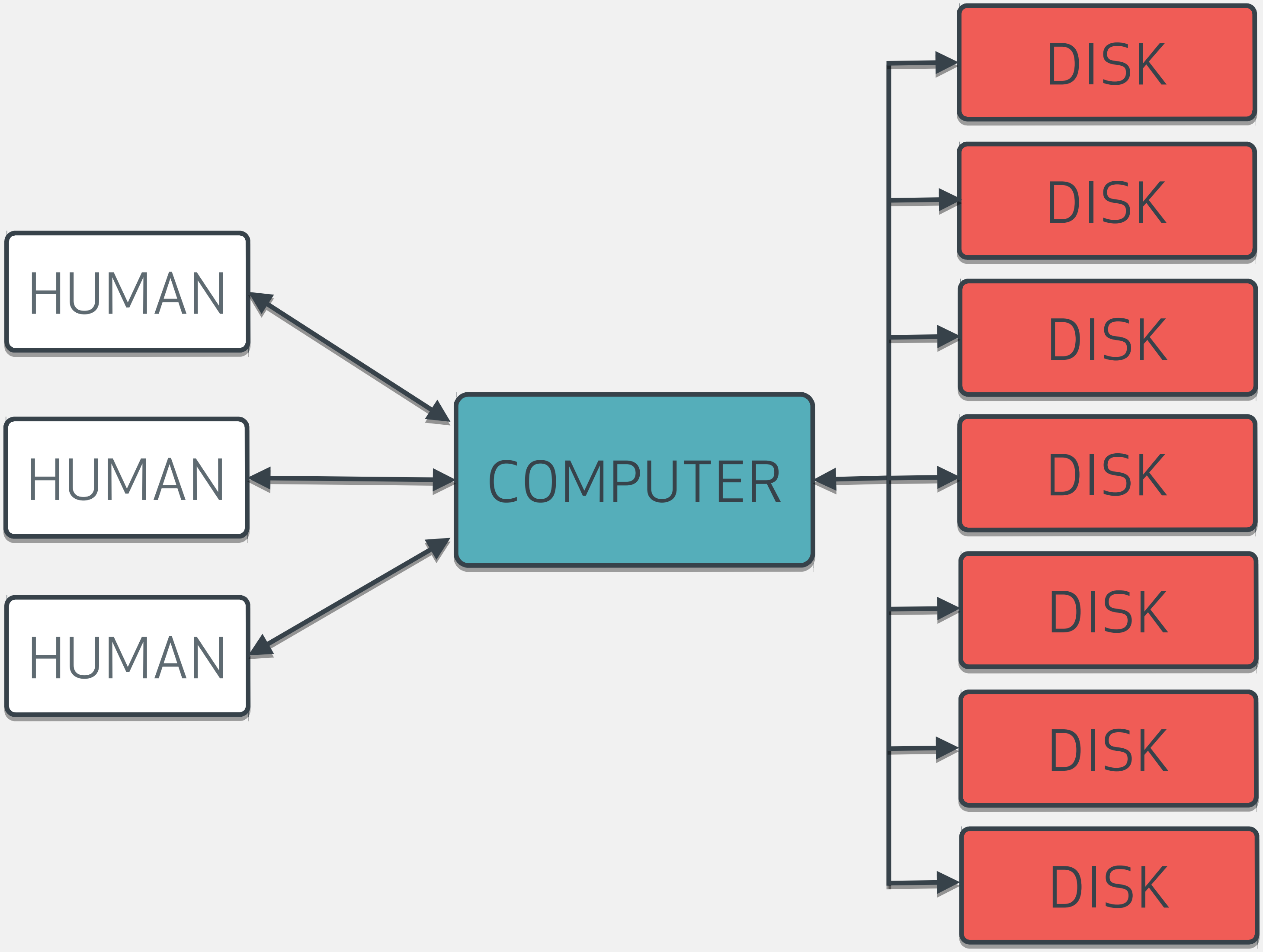


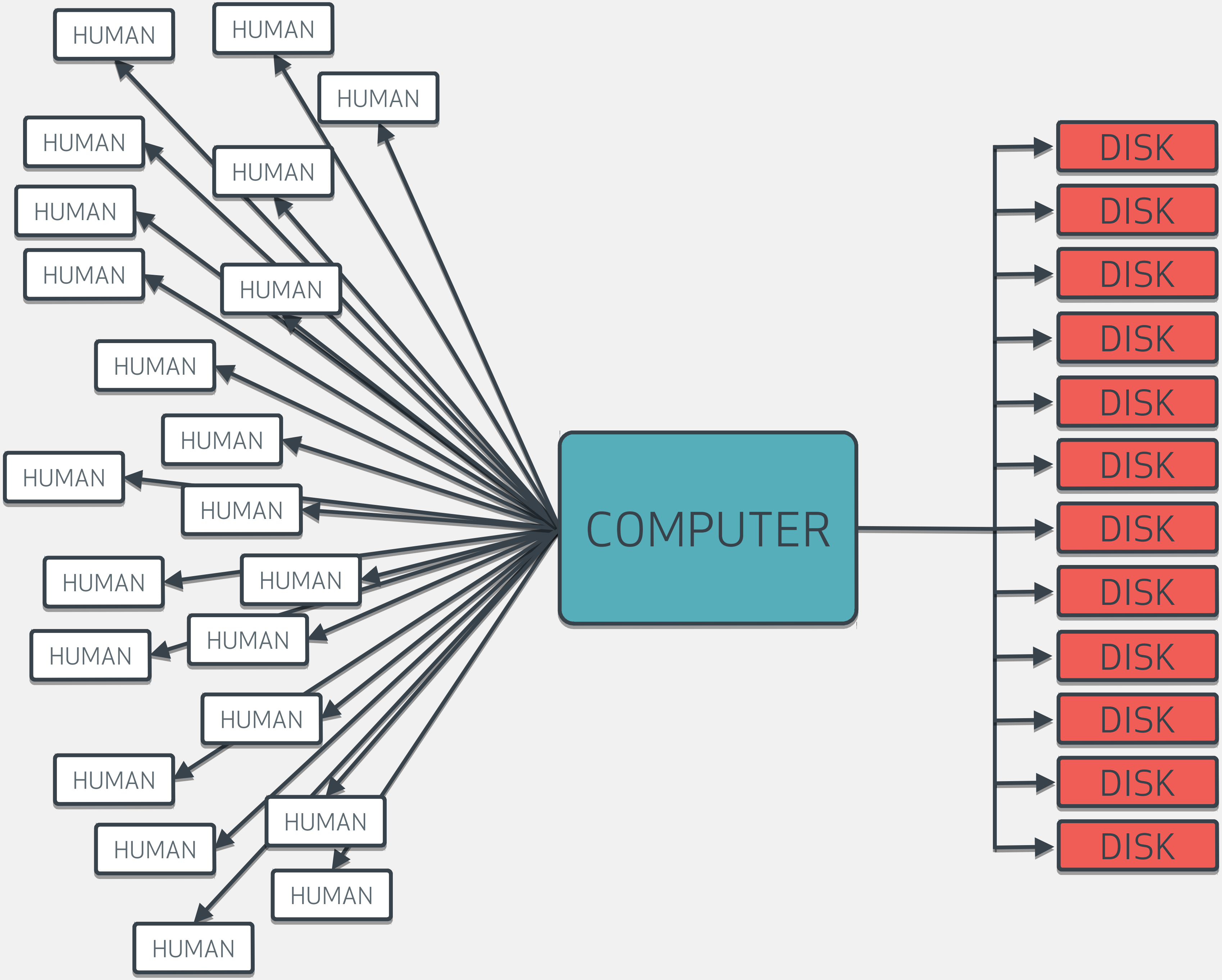
YOU

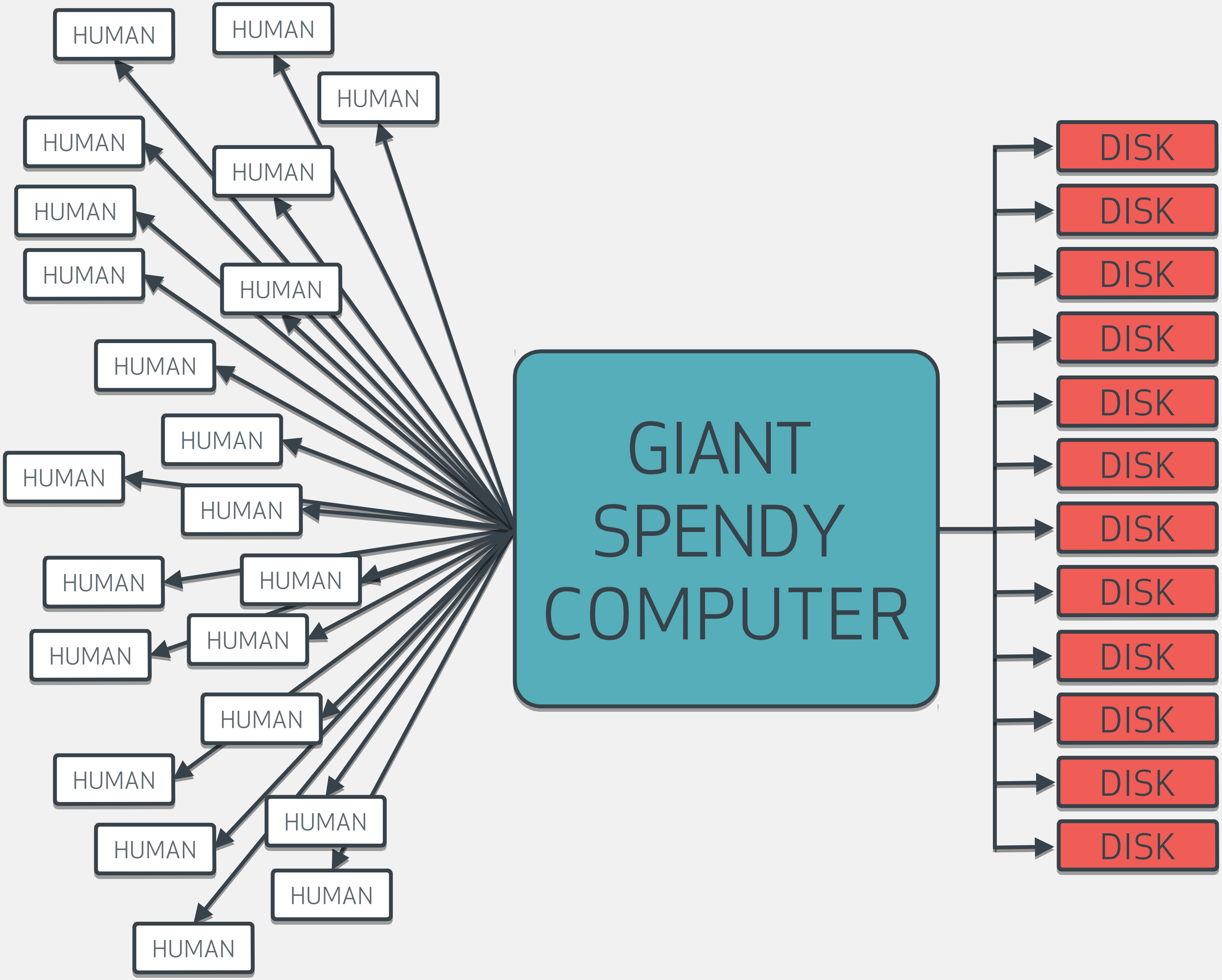
TECHNOLOGY

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
HUMAN
COMPUTER
DISK
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

HUMAN

HUMAN

GIANT SPENDY COMPUTER

DISK

DISK

DISK

DISK

DISK

DISK

DISK

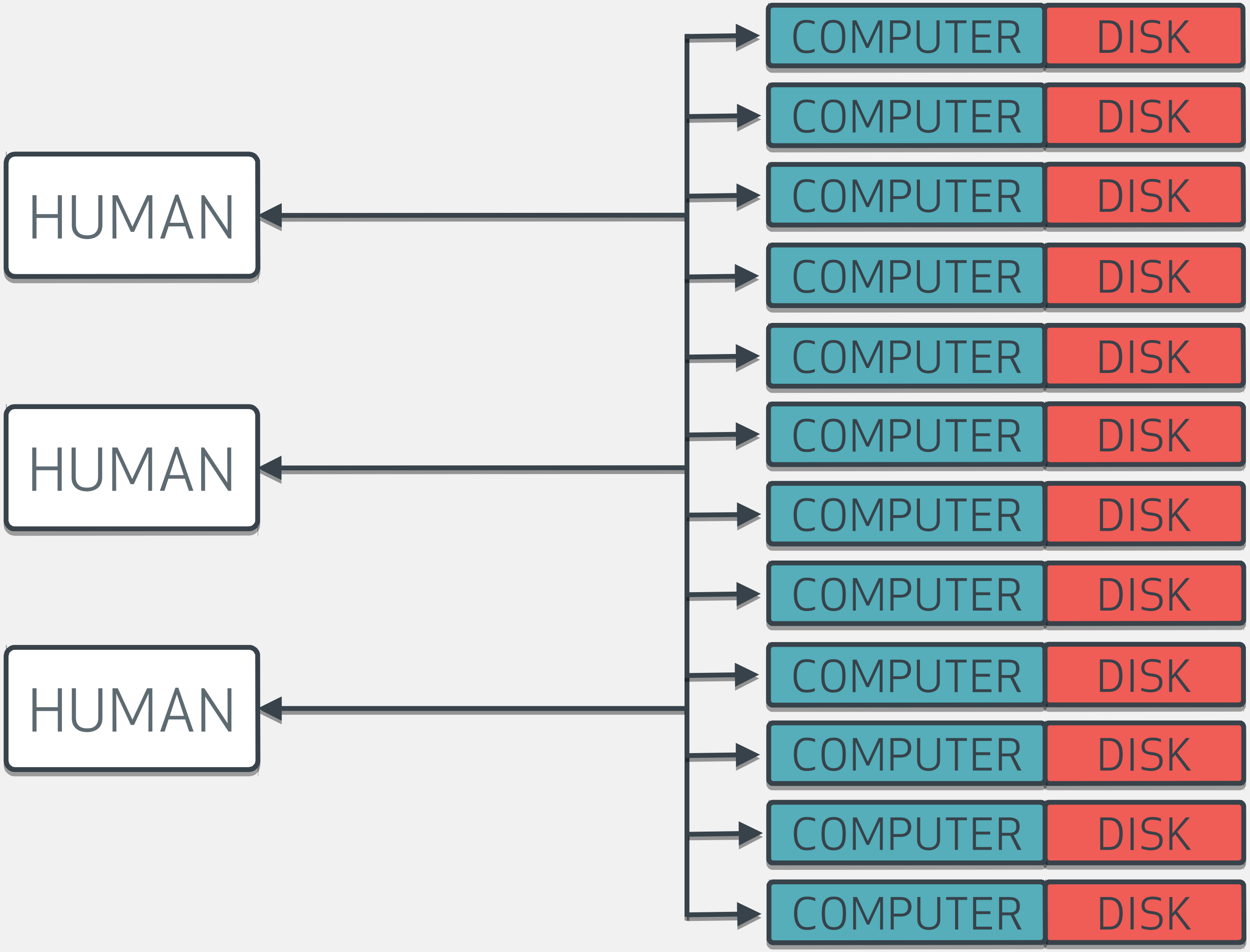
DISK

DISK

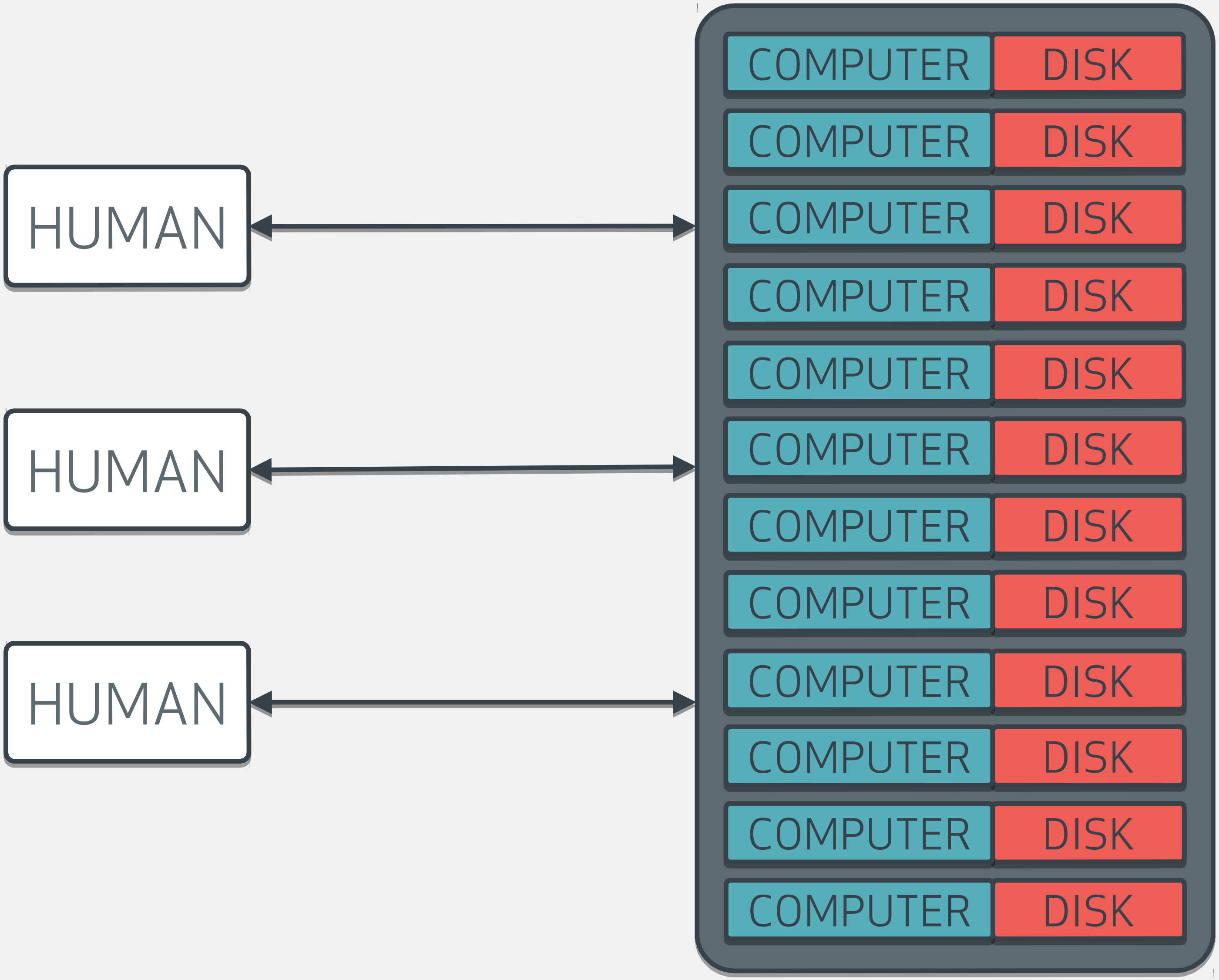
DISK

DISK

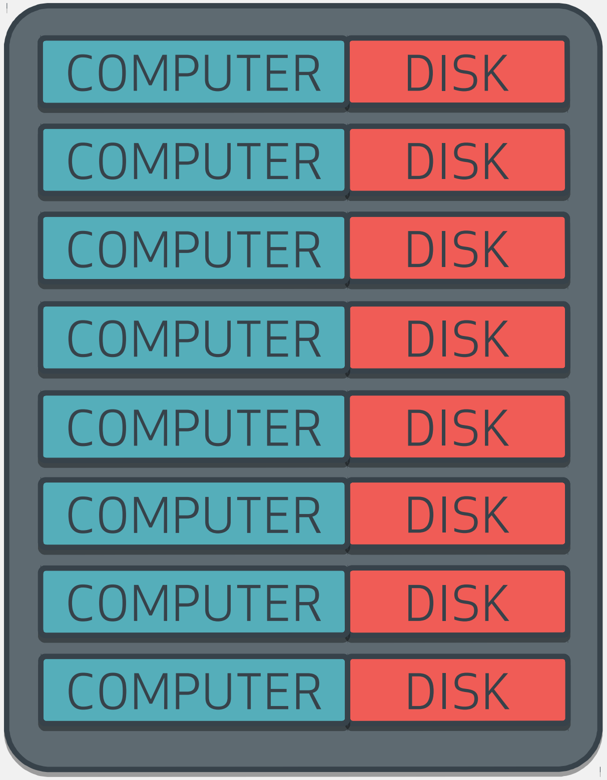
DISK



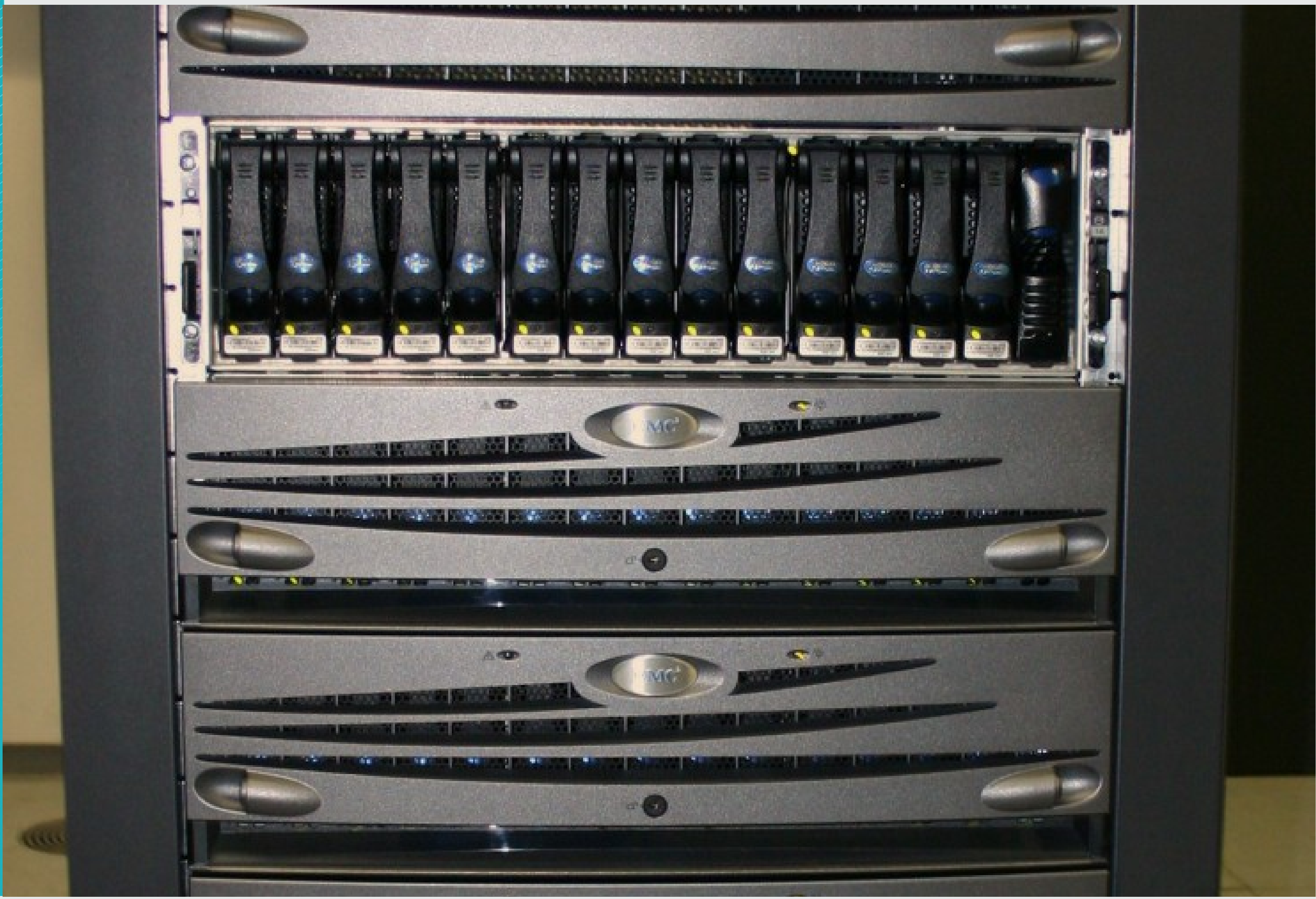
HUMAN
HUMAN
HUMAN
COMPUTER DISK
COMPUTER DISK
COMPUTER DISK
COMPUTER DISK
COMPUTER DISK
COMPUTER DISK
COMPUTER DISK
COMPUTER DISK
COMPUTER DISK
COMPUTER DISK
COMPUTER DISK
COMPUTER DISK


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

“STORAGE APPLIANCE”



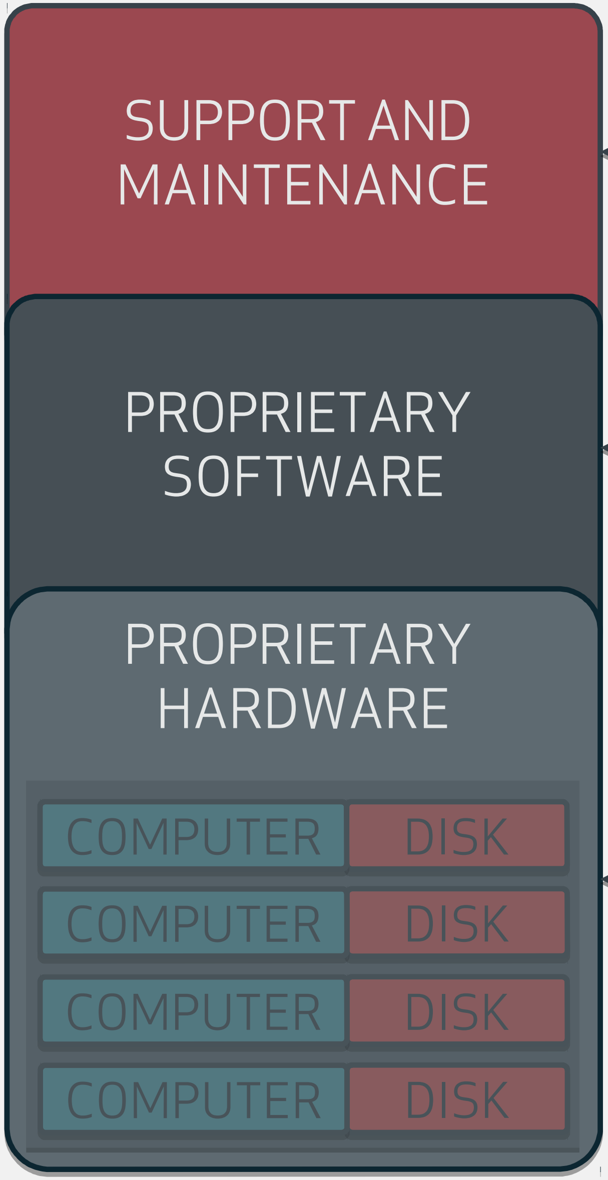
COMPUTER	DISK
COMPUTER	DISK
COMPUTER	DISK
COMPUTER	DISK
COMPUTER	DISK
COMPUTER	DISK
COMPUTER	DISK
COMPUTER	DISK



Storage Appliance

Michael Moll, Wikipedia / CC BY-SA 2.0

$NYSE:EMC, FY2014 10K

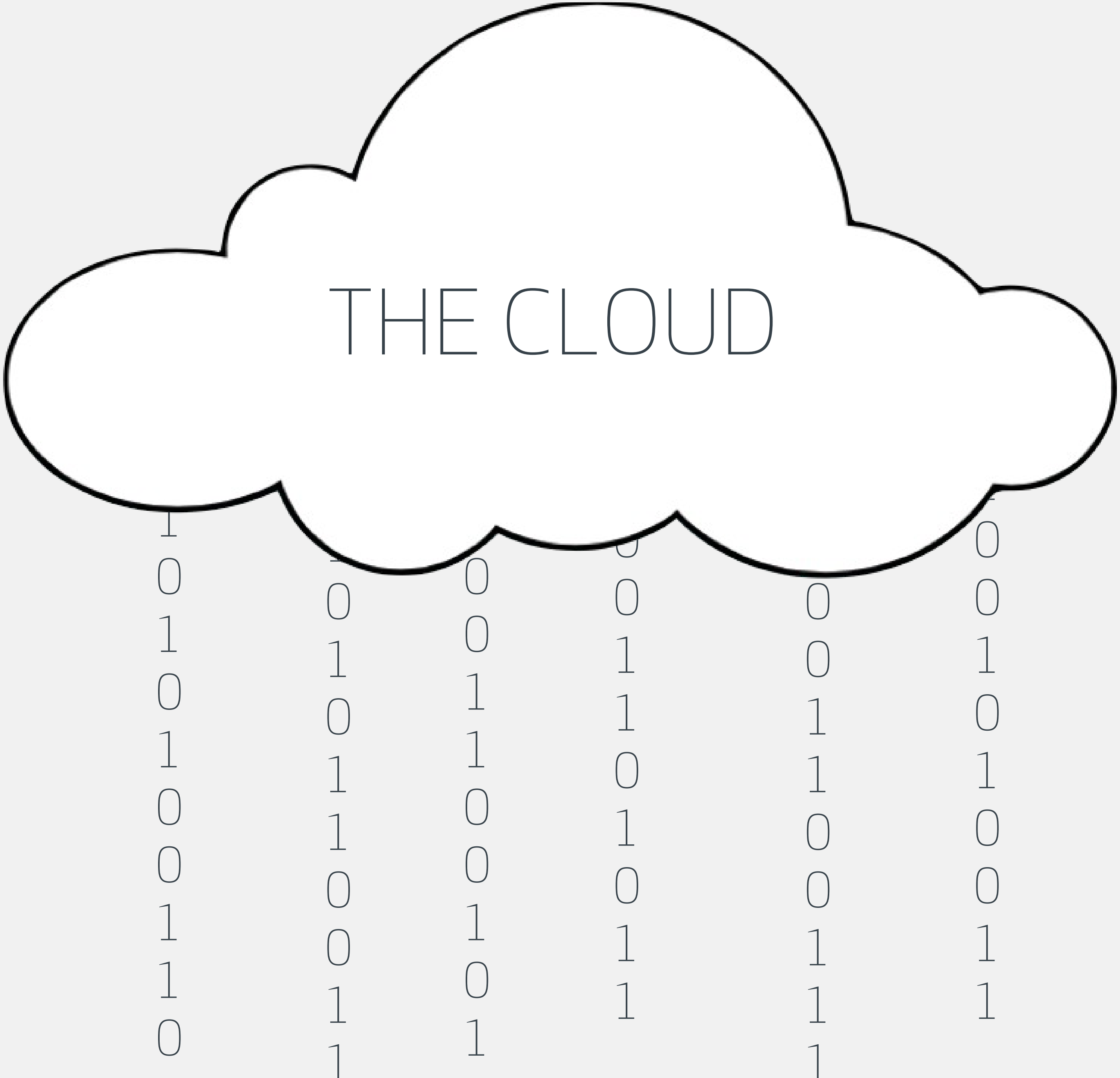


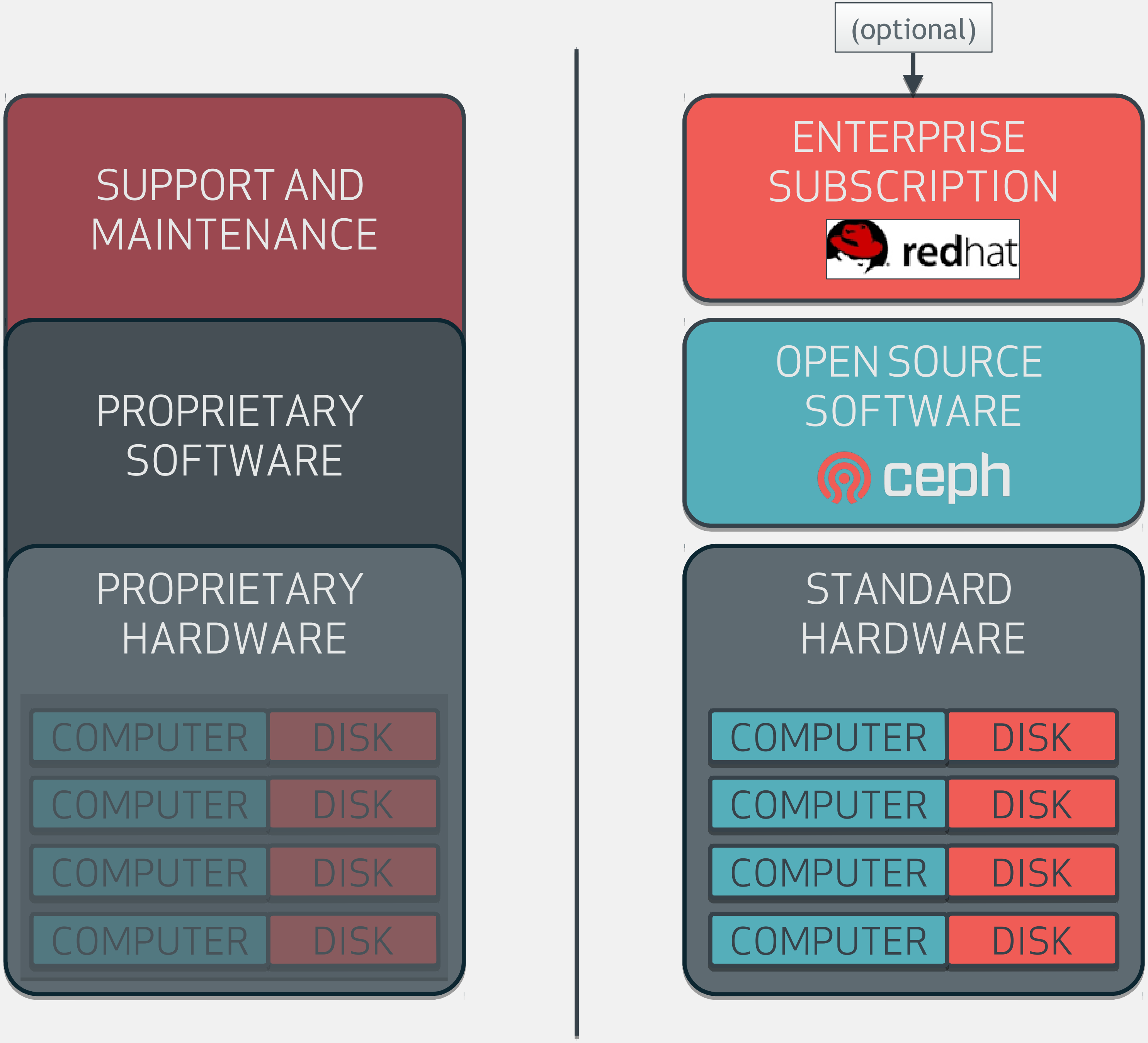
34% of revenue
(5.7 billion dollars)

1.3 billion in R&D
Spent in a year

1.6+ million square feet
of manufacturing space

THE CLOUD


1010100110
010110011
001100101
001101011
001100111
001010011



SUPPORT AND MAINTENANCE

PROPRIETARY SOFTWARE

PROPRIETARY HARDWARE

COMPUTER DISK

COMPUTER DISK

COMPUTER DISK

COMPUTER DISK

(optional)

ENTERPRISE SUBSCRIPTION

redhat

OPEN SOURCE SOFTWARE

ceph

STANDARD HARDWARE

COMPUTER DISK

COMPUTER DISK

COMPUTER DISK

COMPUTER DISK

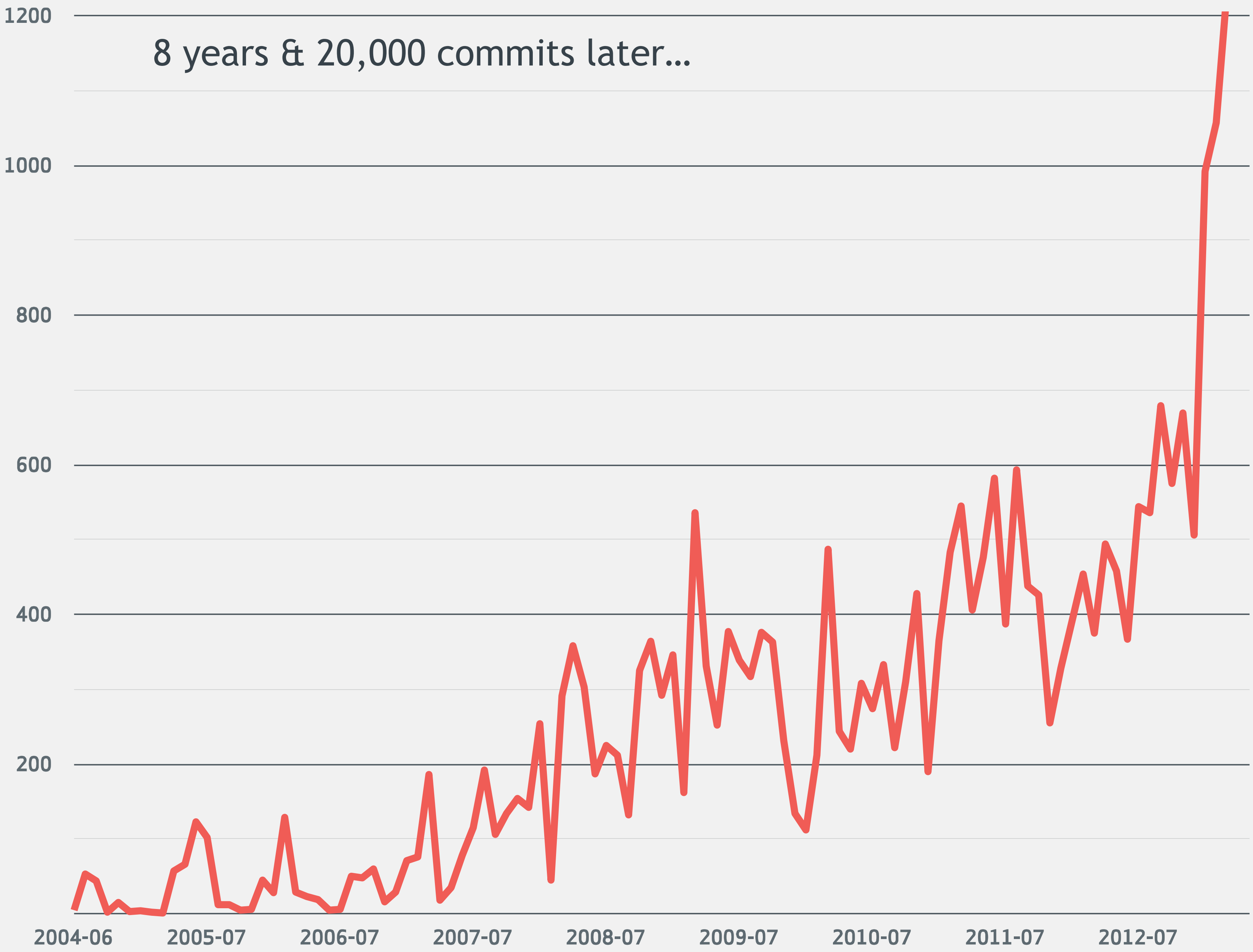
ceph

philosophy

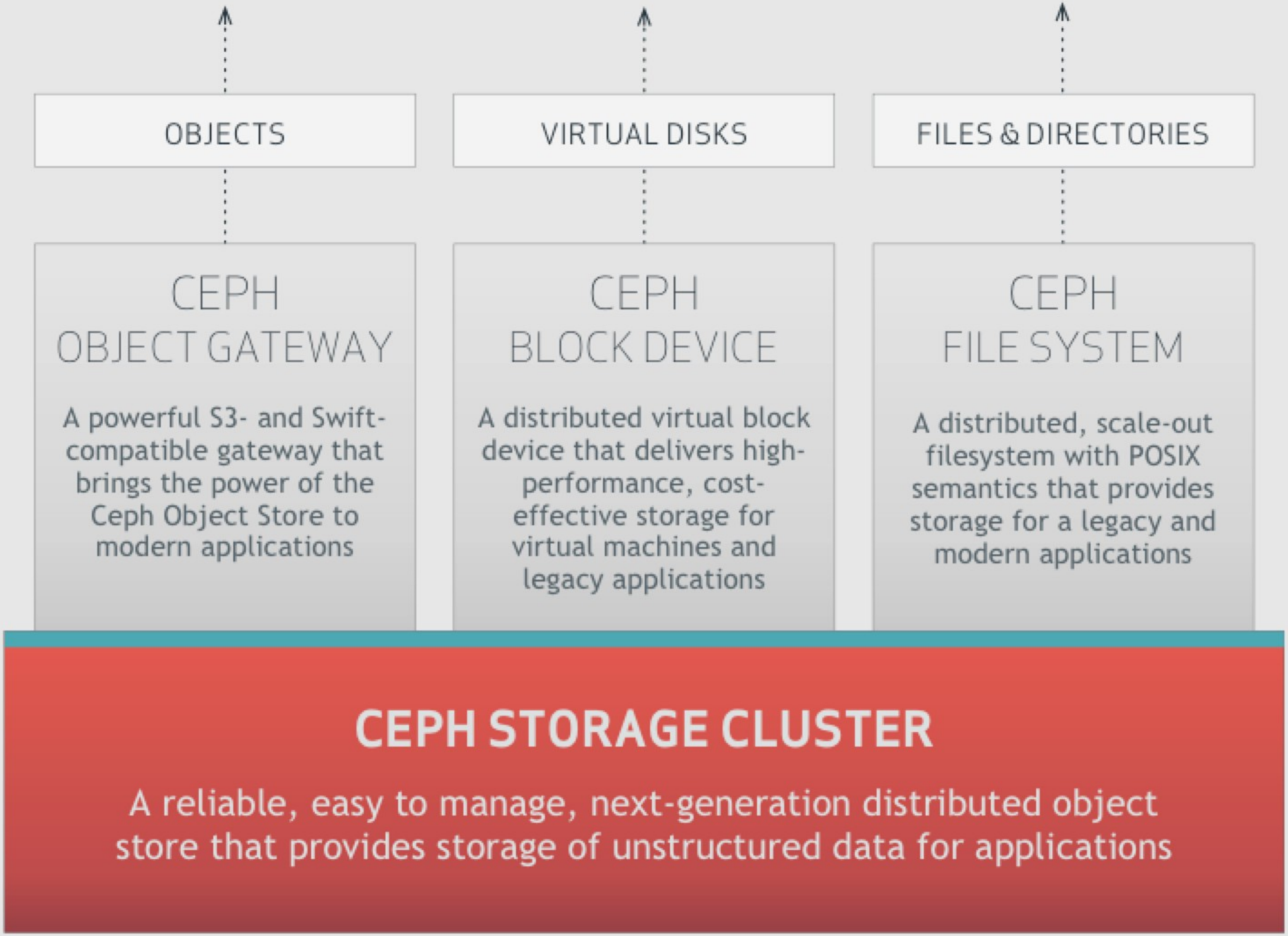
- OPEN SOURCE
- COMMUNITY-FOCUSED

design

- SCALABLE
- NO SINGLE POINT OF FAILURE
- SOFTWARE BASED
- SELF-MANAGING



8 years & 20,000 commits later…



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

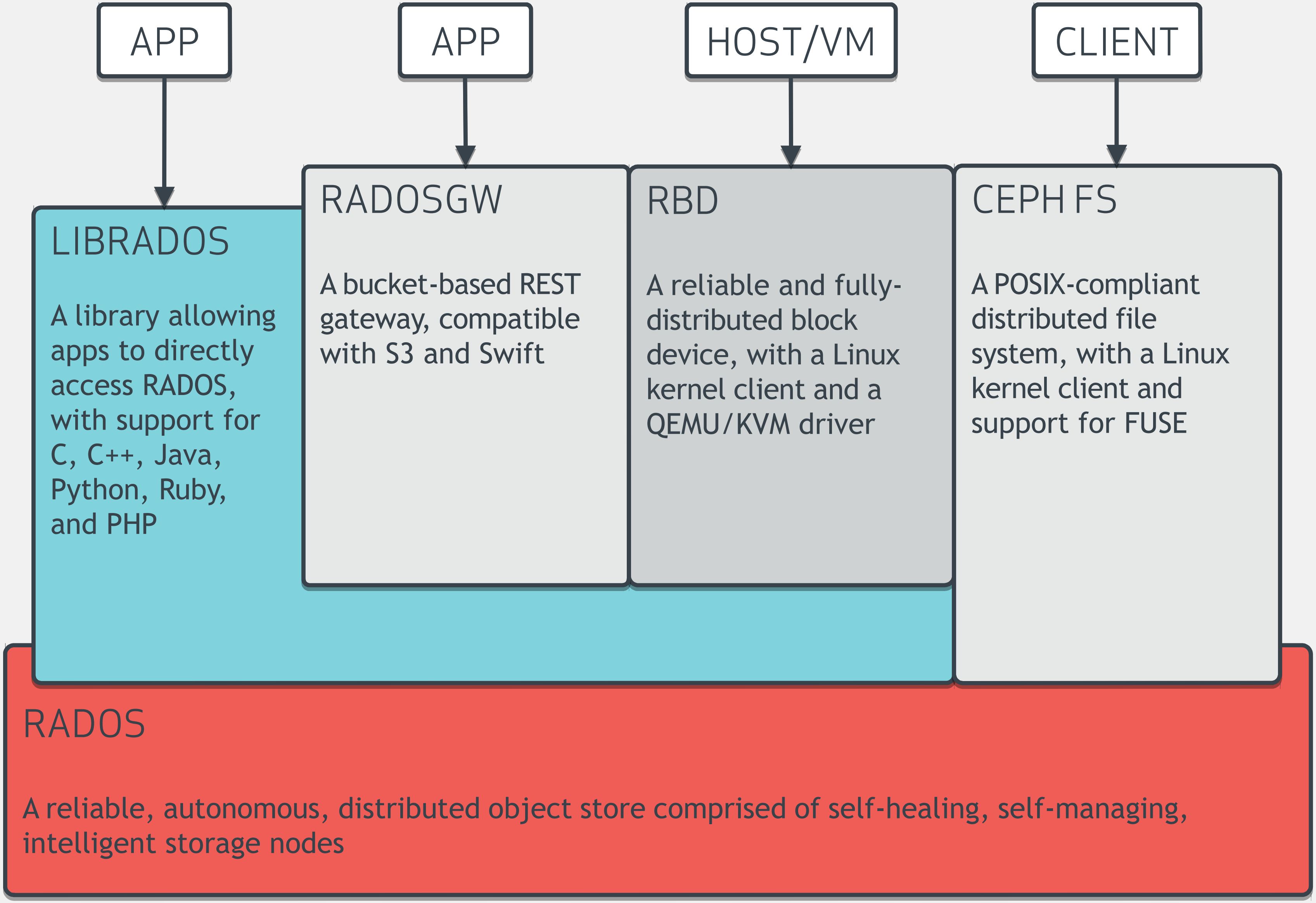
A distributed virtual block device that delivers high-performance, cost-effective storage for virtual machines and legacy applications

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

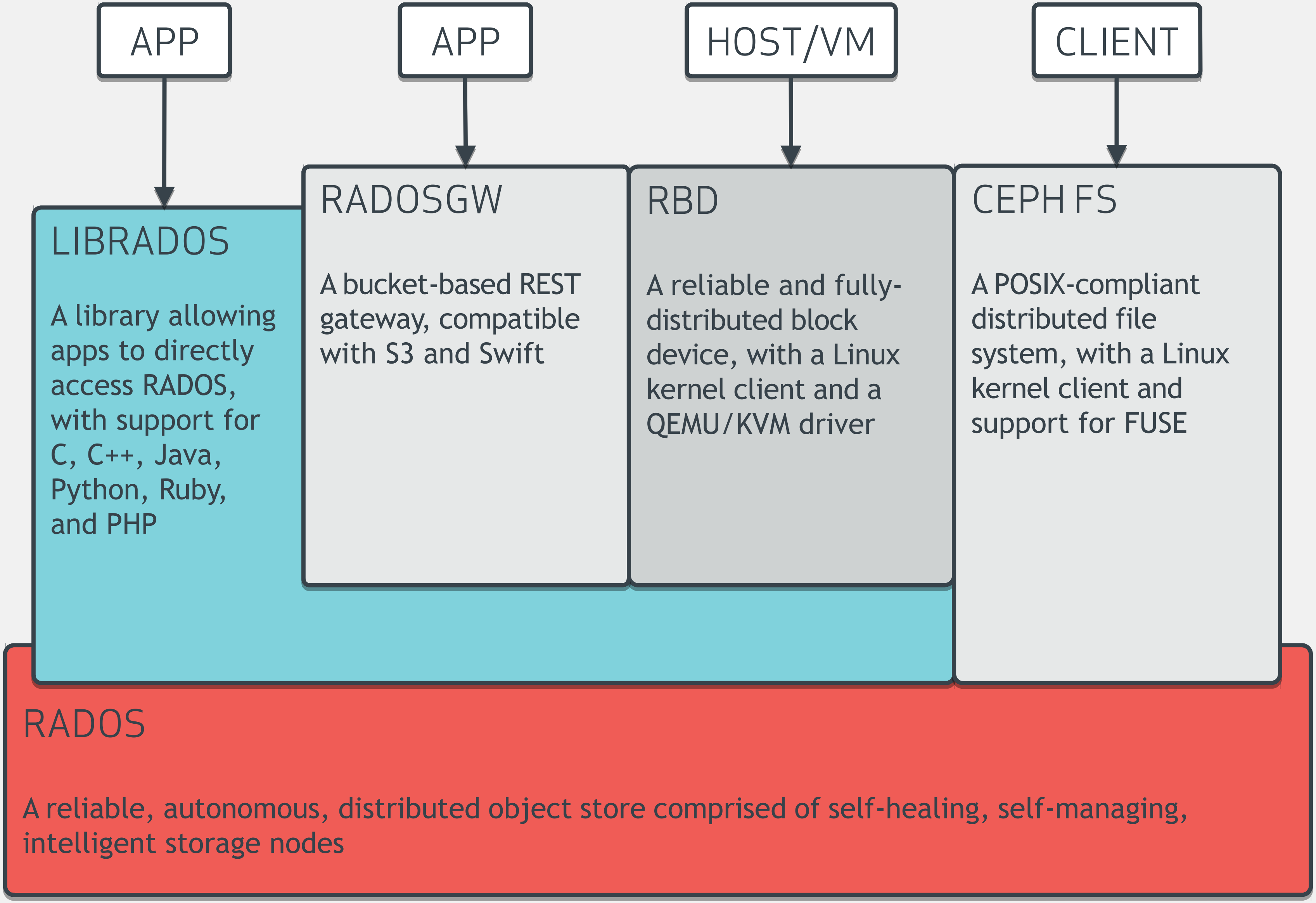
A reliable and fully-distributed block device, with a Linux kernel client and a QEMU/KVM driver

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

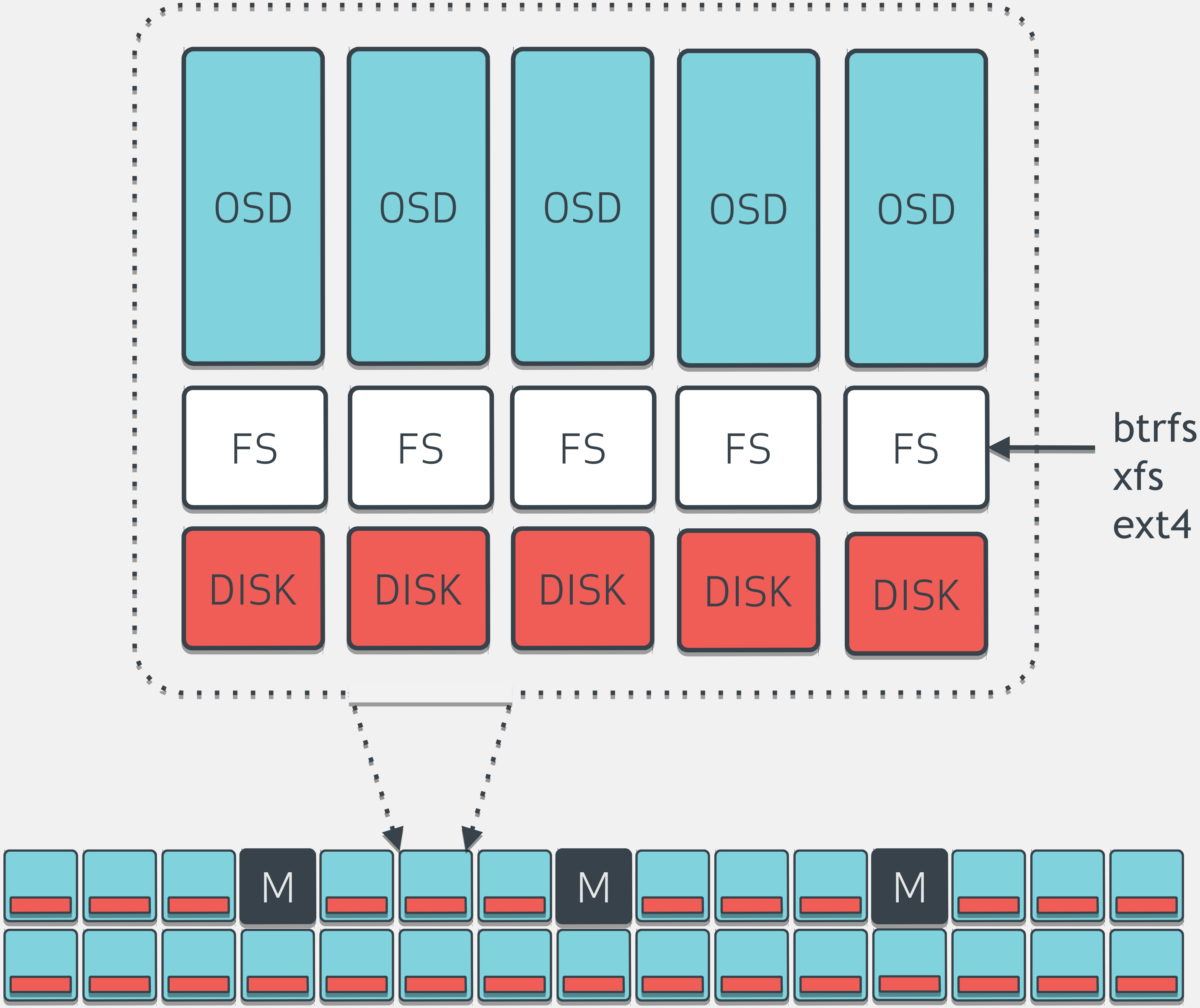
A reliable and fully-distributed block device, with a Linux kernel client and a QEMU/KVM driver

CEPH FS

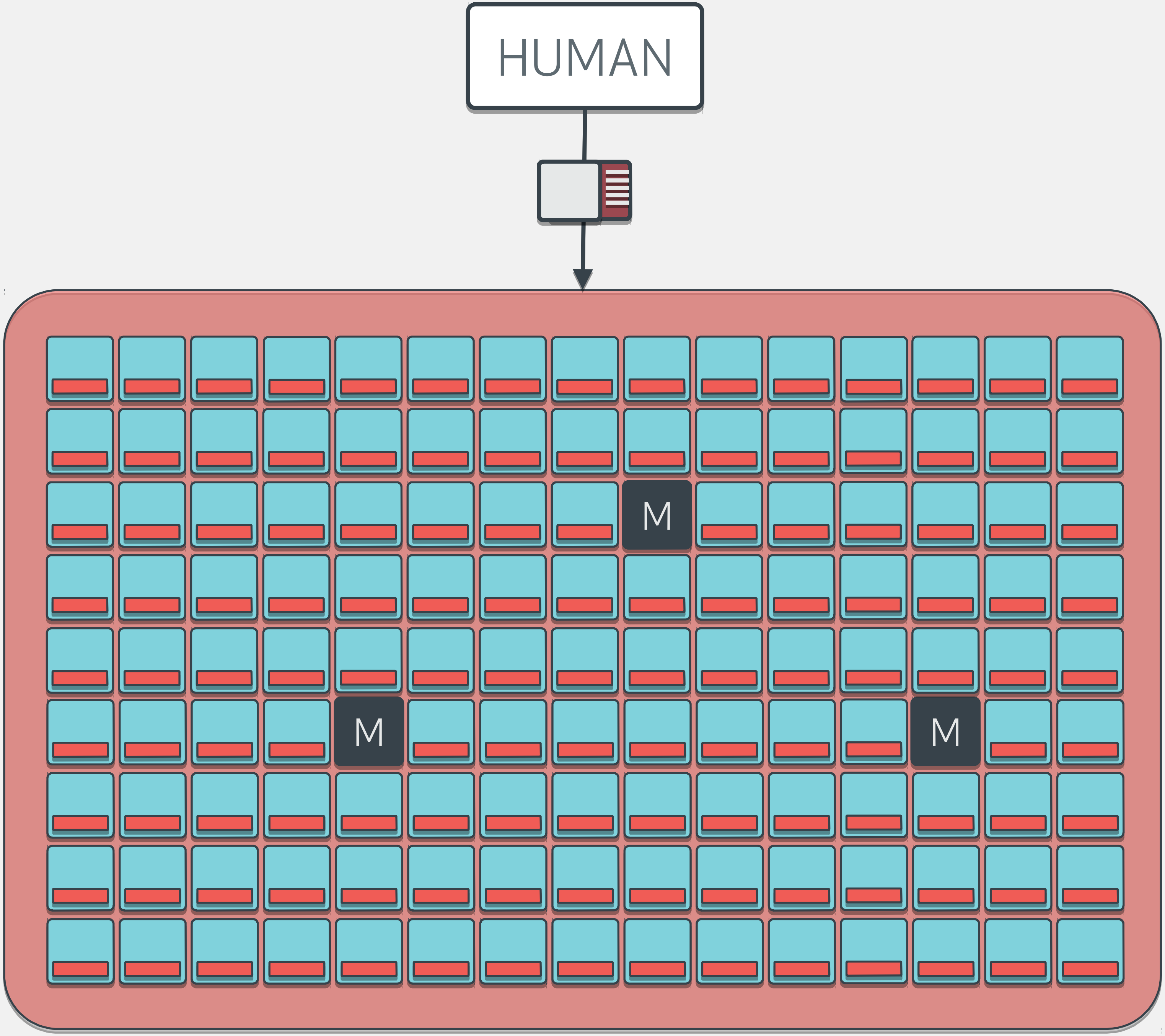
A POSIX-compliant distributed file system, with a Linux kernel client and support for FUSE

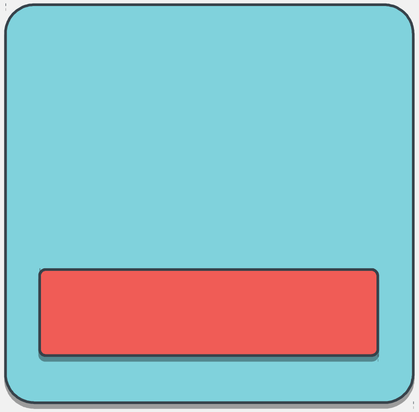
RADOS

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



btrfs
xfs
ext4


HUMAN
M
M
M

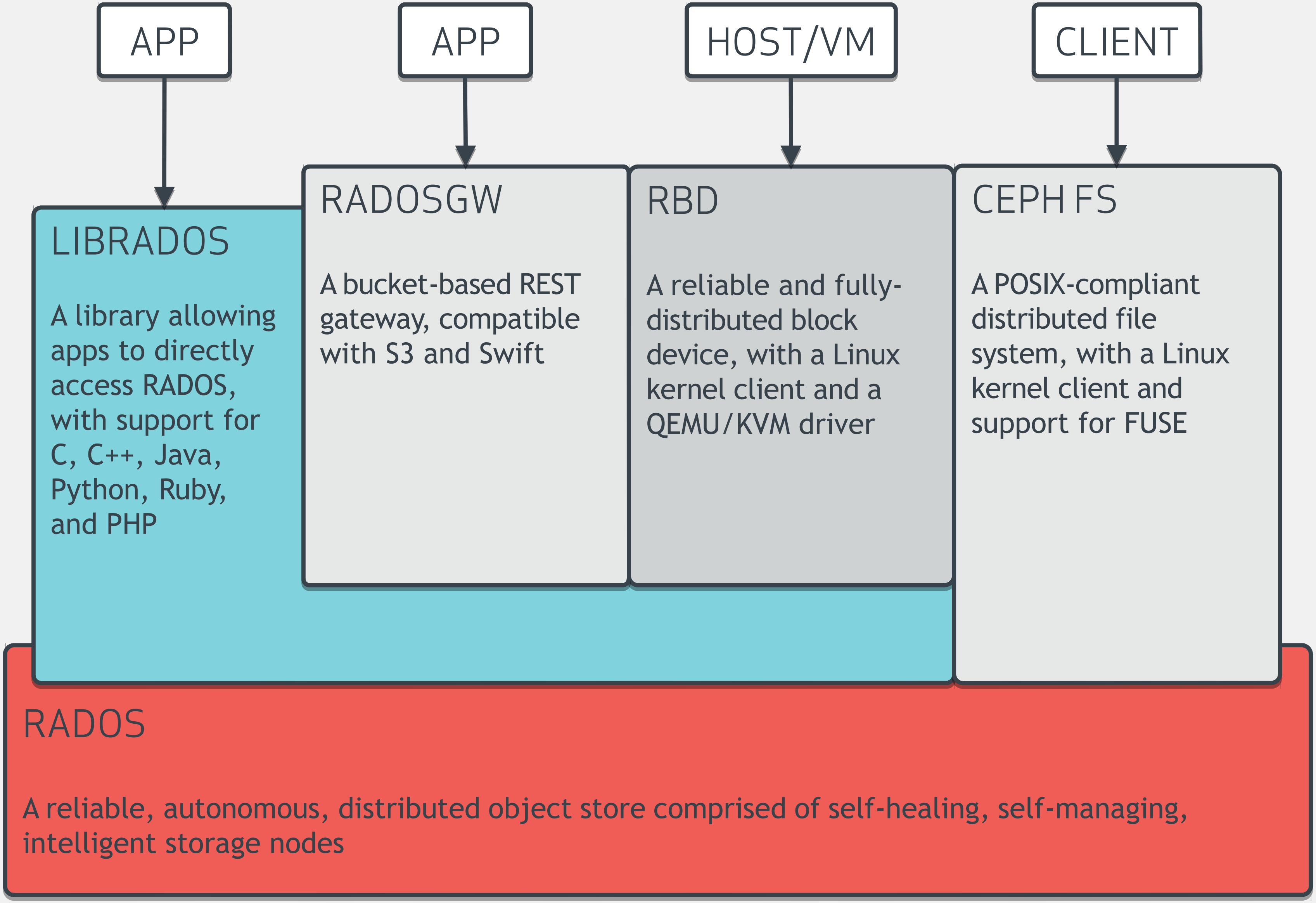


OSDs:

- 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

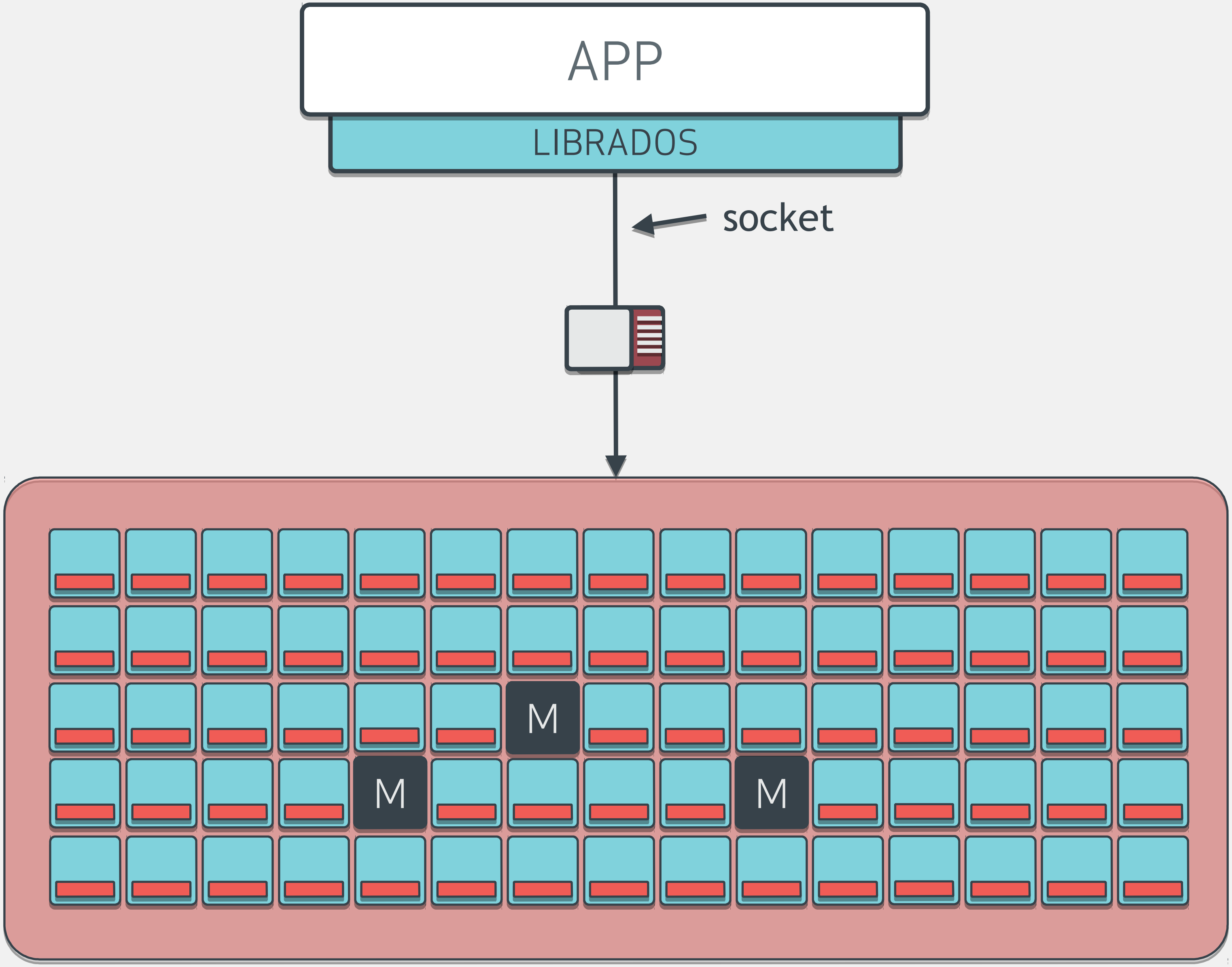
A reliable and fully-distributed block device, with a Linux kernel client and a QEMU/KVM driver

CEPH FS

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

RADOS

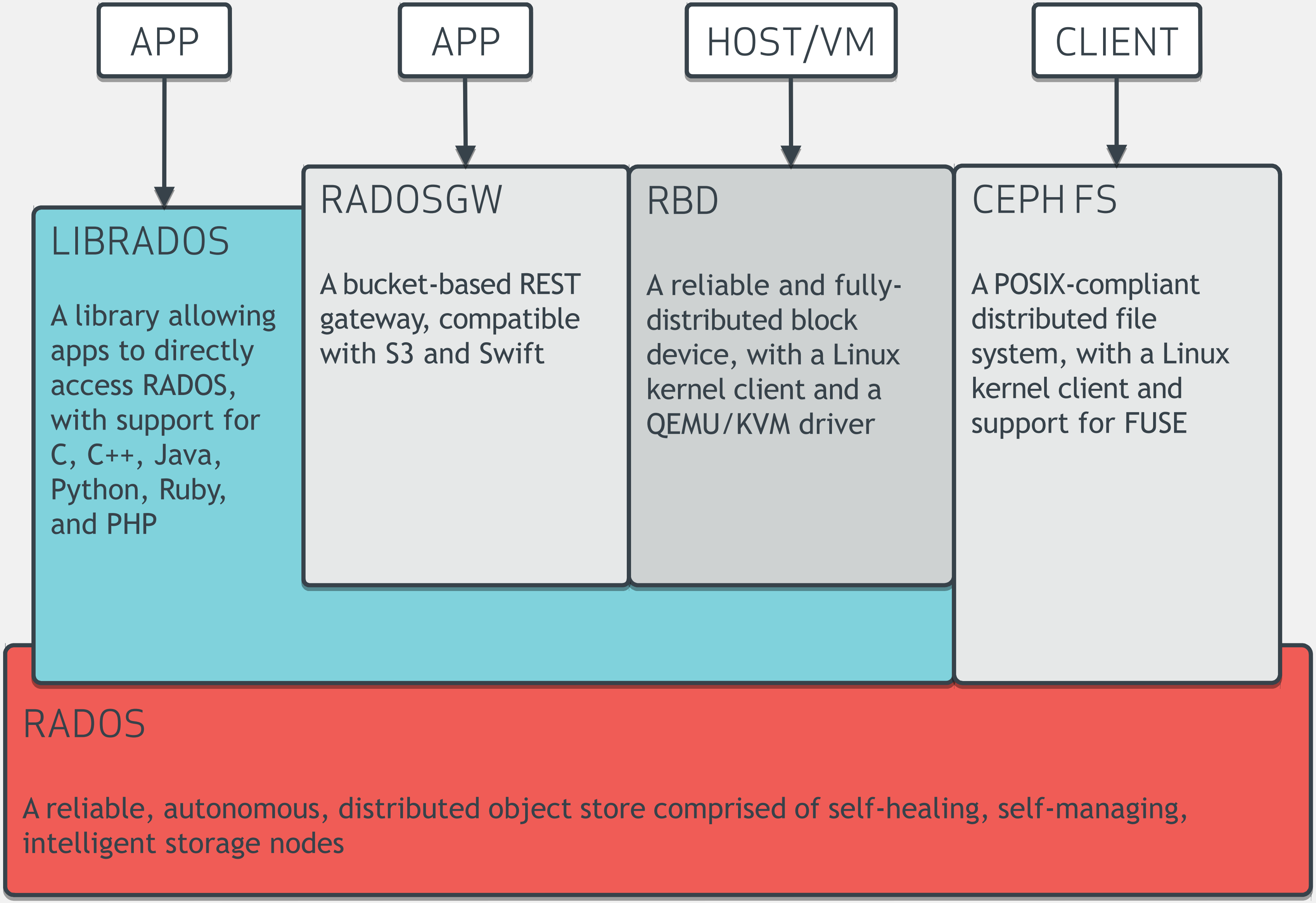
A reliable, autonomous, distributed object store comprised of self-healing, self-managing, intelligent storage nodes


APP
LIBRADOS
socket
M
M
M

L

LIBRADOS

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



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

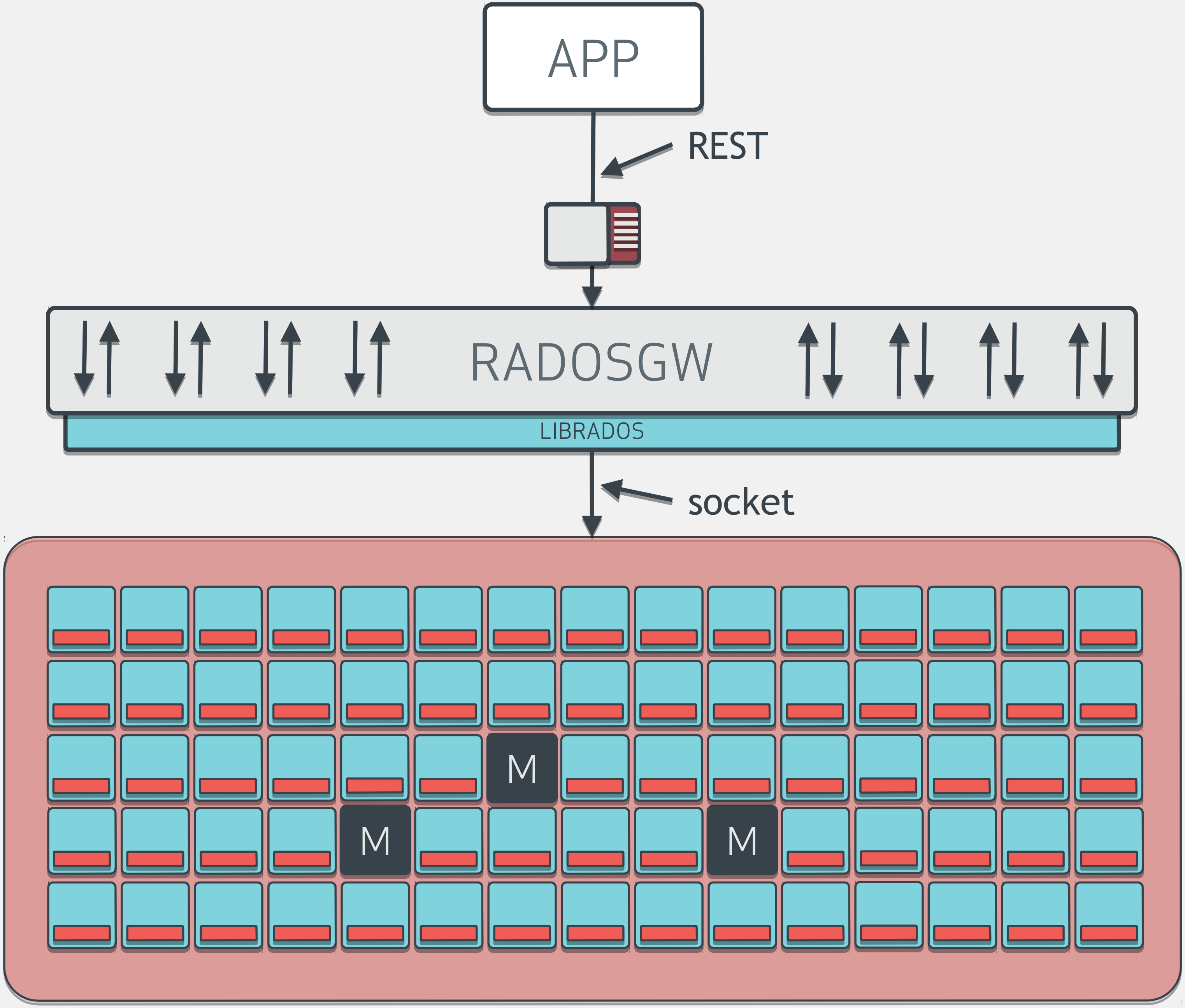
A reliable and fully-distributed block device, with a Linux kernel client and a QEMU/KVM driver

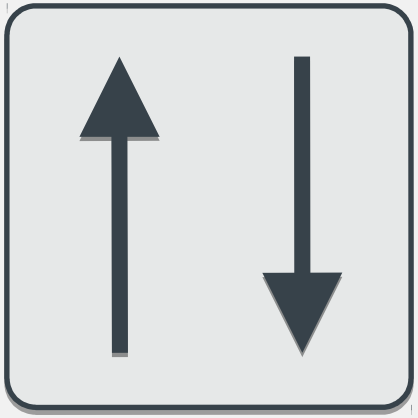
CEPH FS

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

RADOS

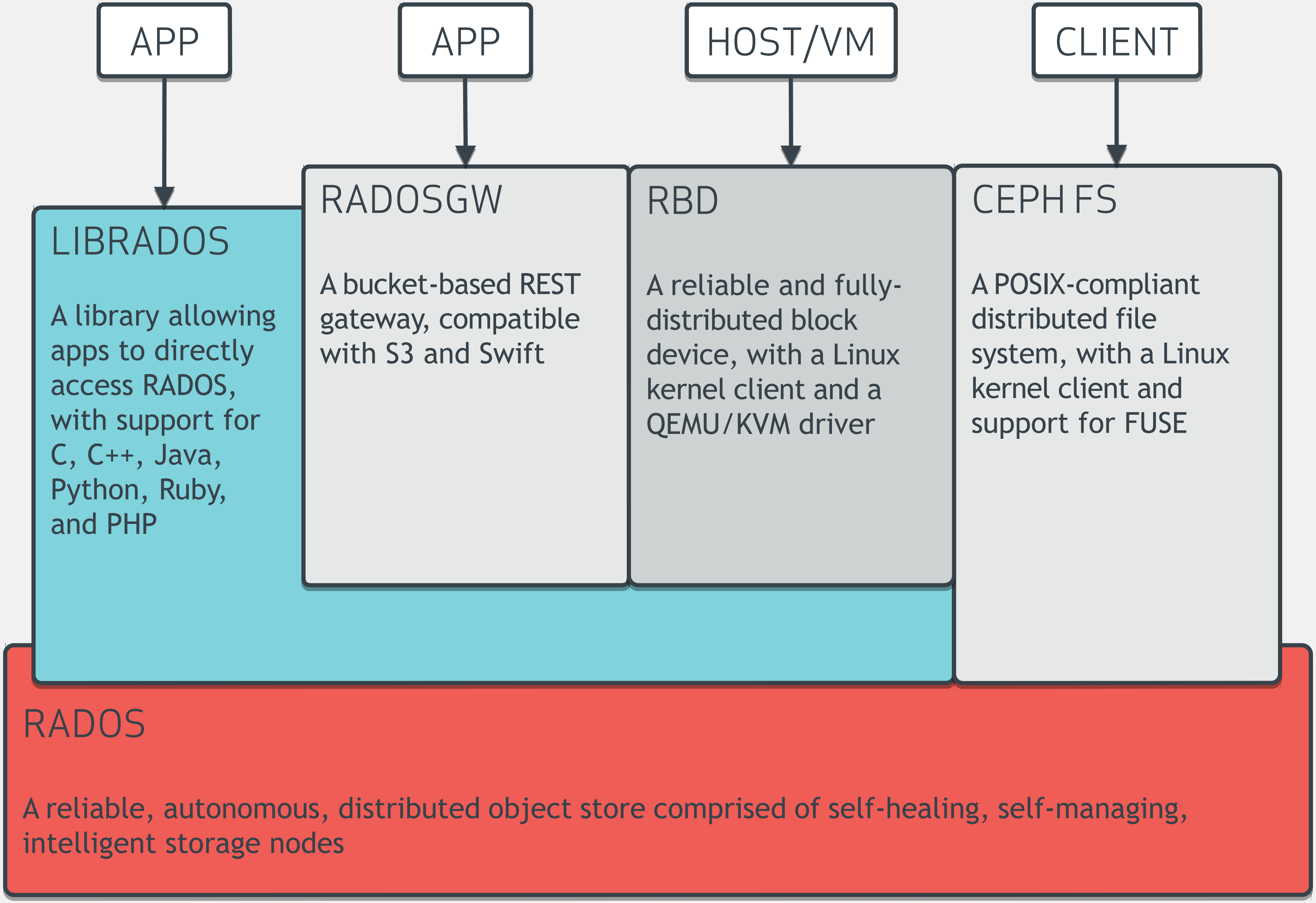
A reliable, autonomous, distributed object store comprised of self-healing, self-managing, intelligent storage nodes


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

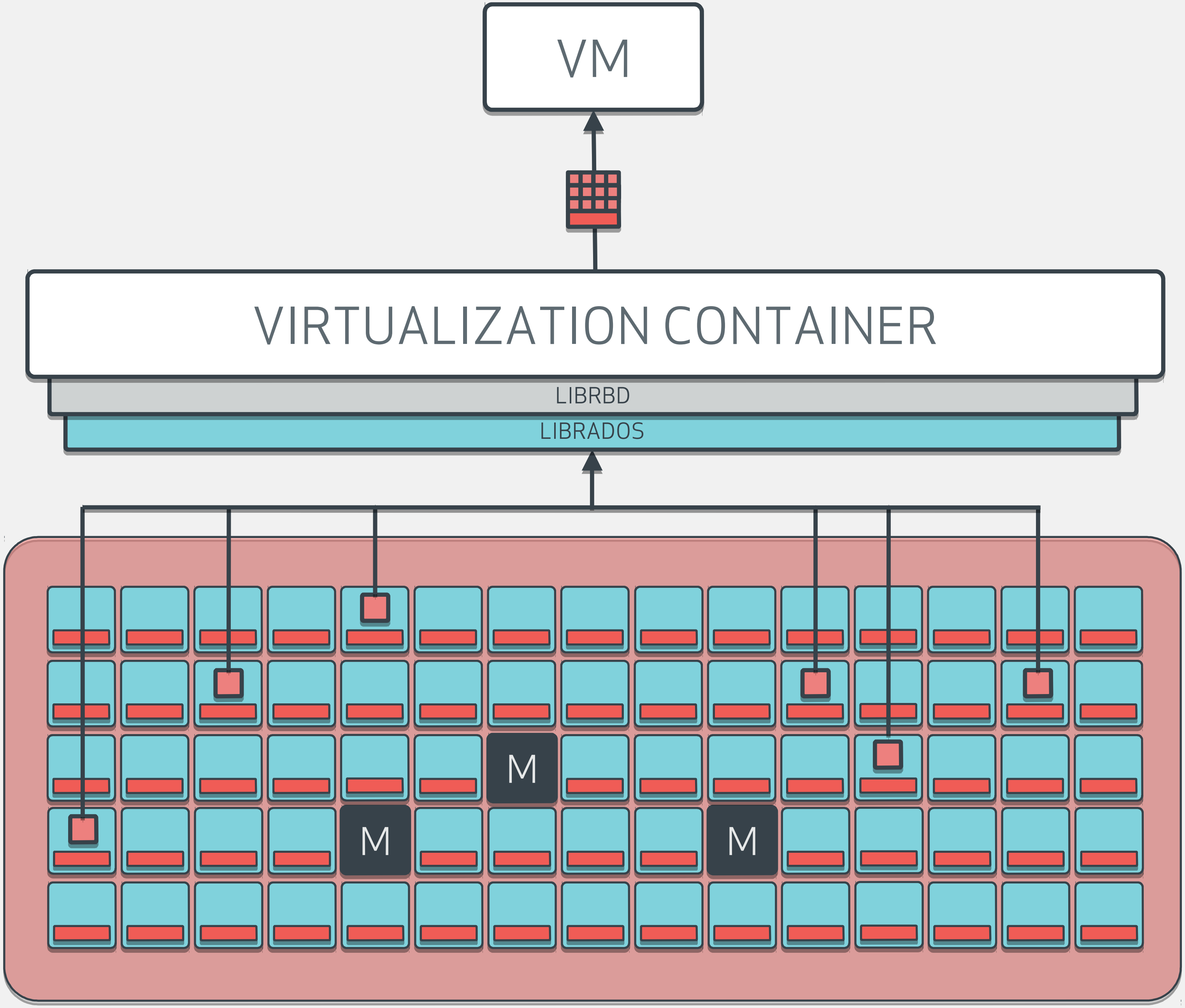
A reliable and fully-distributed block device, with a Linux kernel client and a QEMU/KVM driver

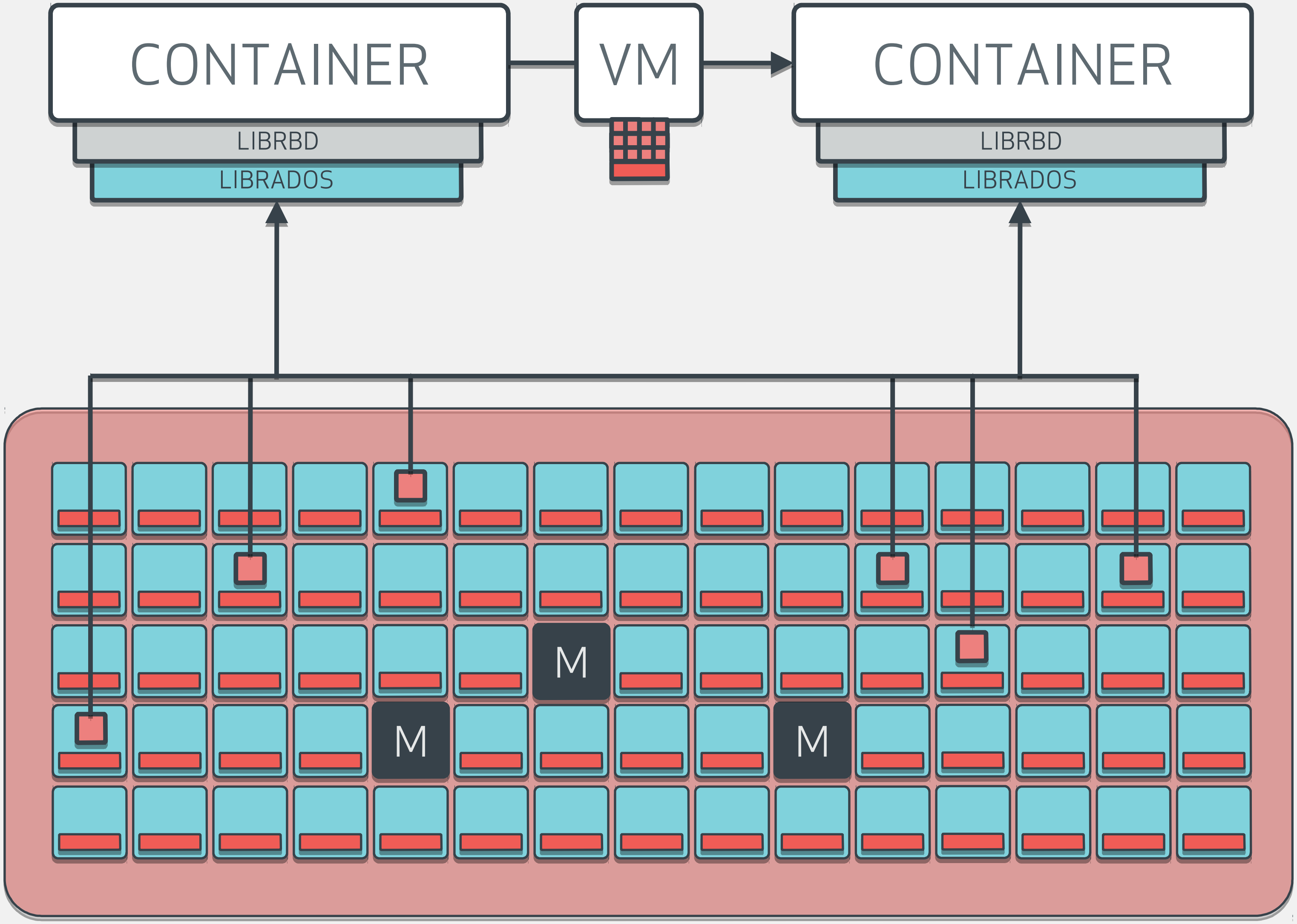
CEPH FS

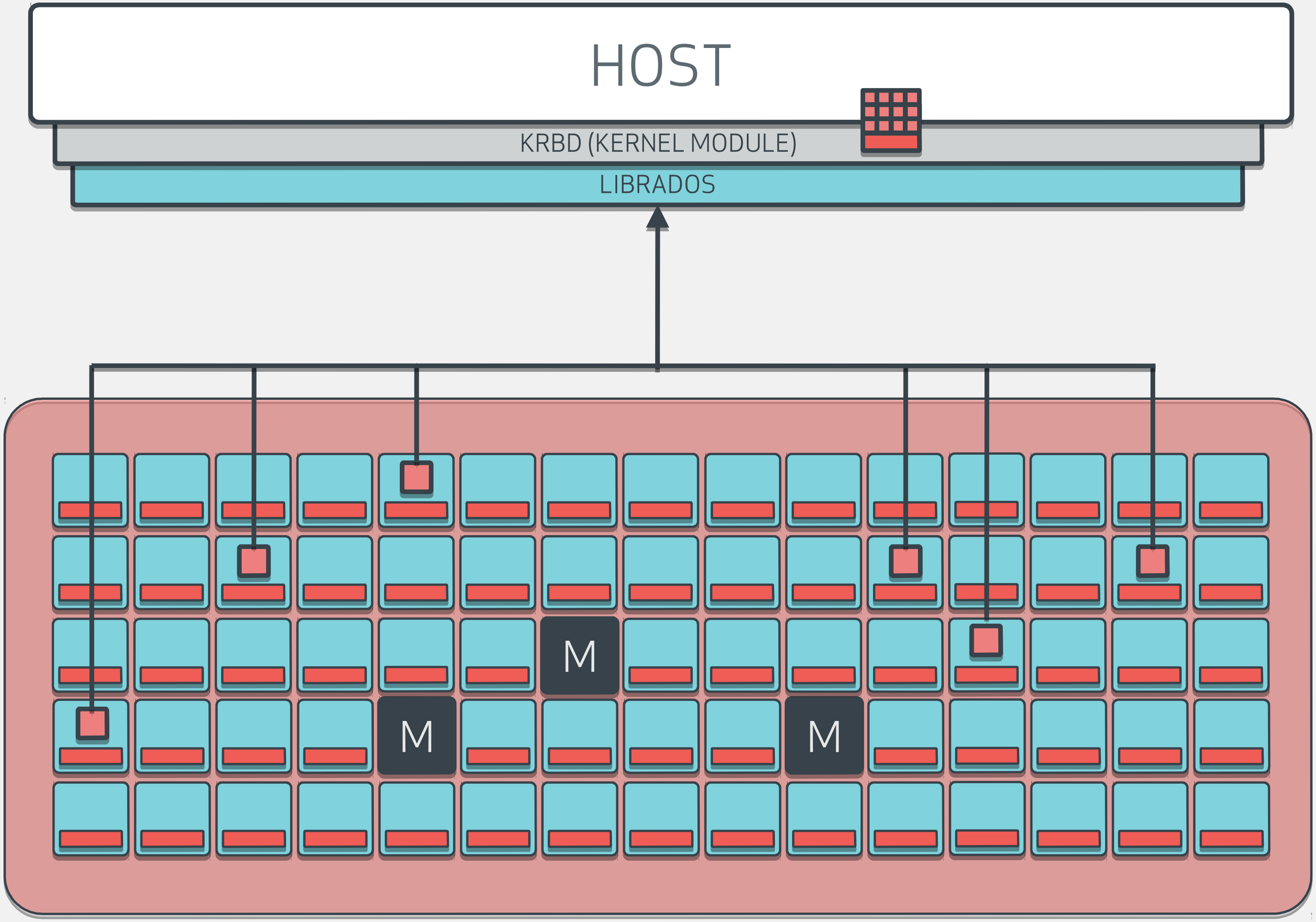
A POSIX-compliant distributed file system, with a Linux kernel client and support for FUSE

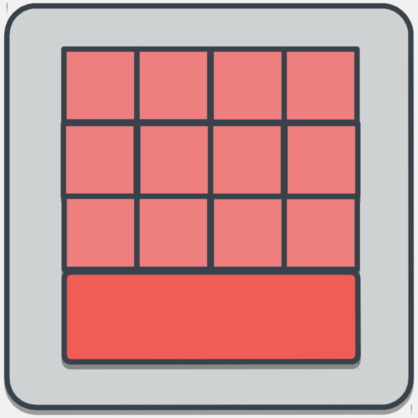
RADOS

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


VM
VIRTUALIZATION CONTAINER
LIBRBD
LIBRADOS
M
M
M

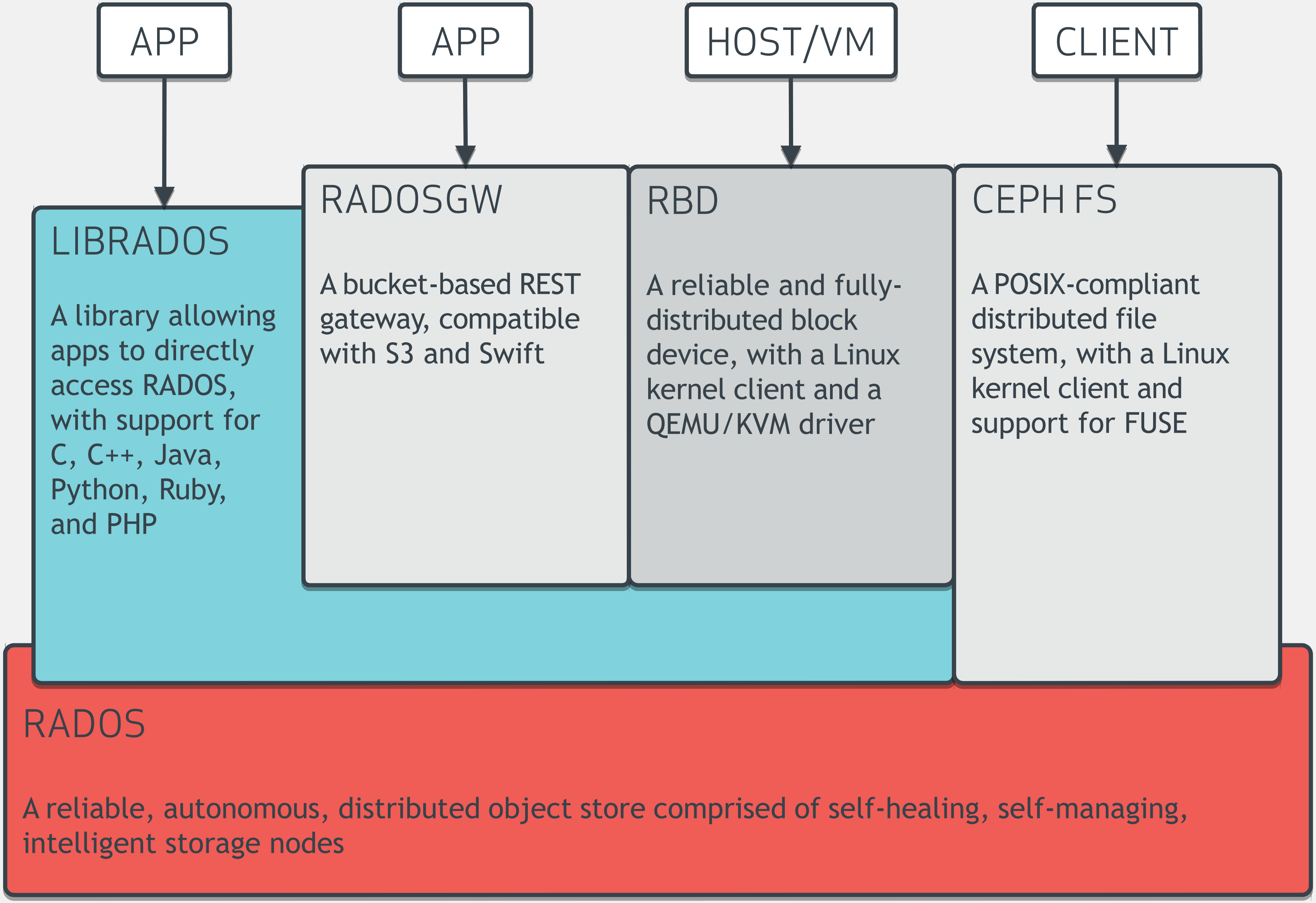

CONTAINER
VM
CONTAINER
LIBRBD
LIBRADOS
LIBRBD
LIBRADOS
M
M
M


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
  - OpenStack, CloudStack



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

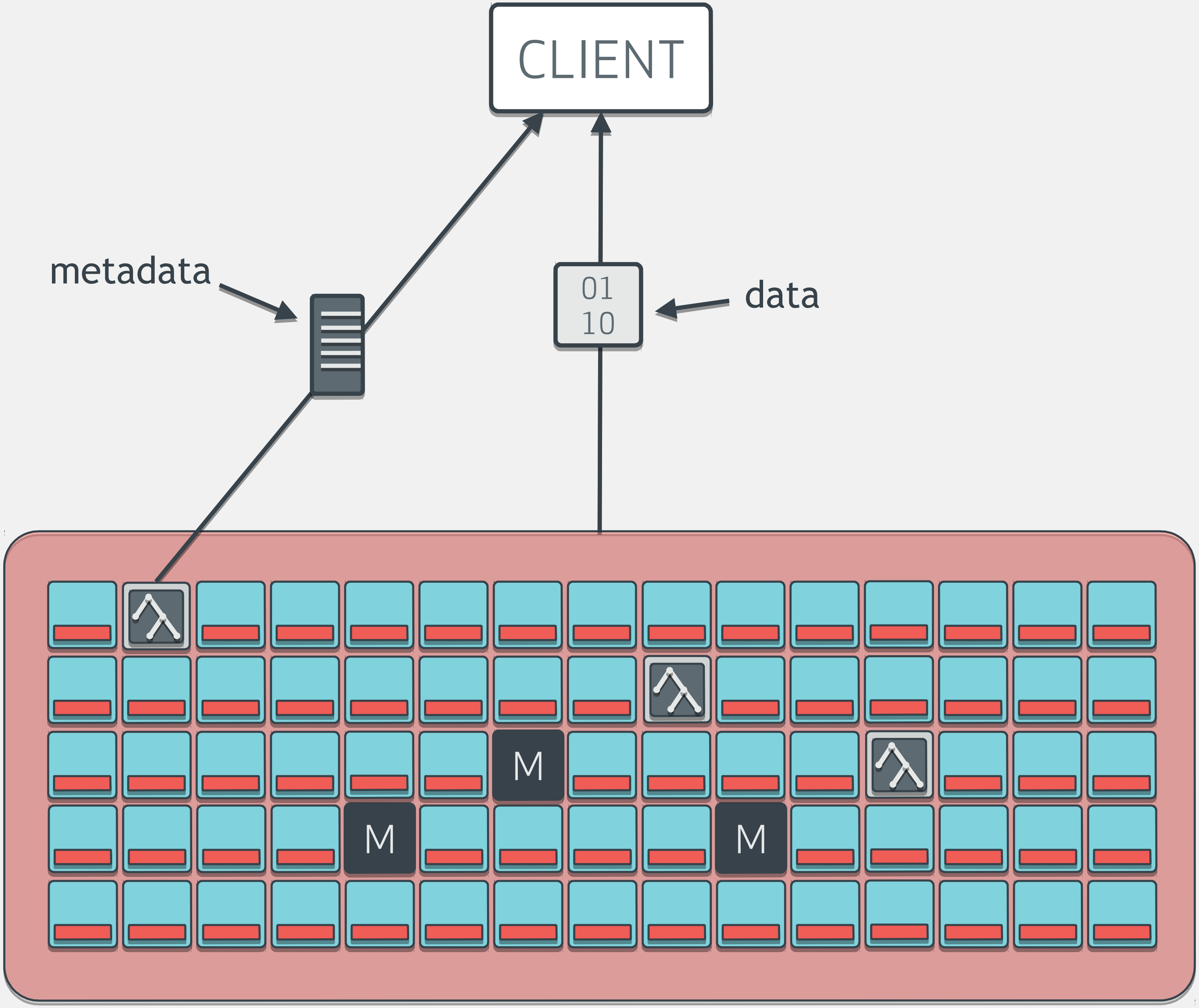
A reliable and fully-distributed block device, with a Linux kernel client and a QEMU/KVM driver

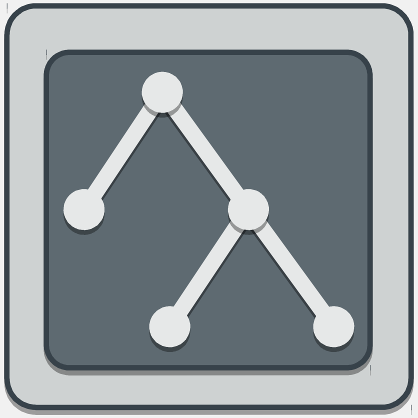
CEPH FS

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

Questions?

Federico Lucifredi
PM Director, Ceph

federico@redhat.com
@0xF2

redhat.com | ceph.com