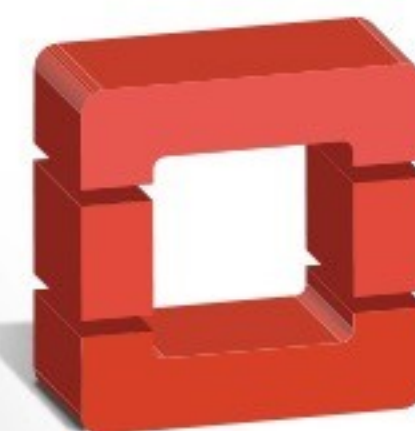# CEPH
storage cluster to go exabyte and beyond

Alvaro Soto
OpenStack & Ceph engineer

KIO Networks

OpenStack Days
MÉXICO

whoami

OS Lover

OpenStack Days
MÉXICO

whoami

- Software developer
- Full Linux sysadmin stack
- Cepher / Stacker

Agenda

- *Storage Background (*** I'm not a storage guy ***)*
- *Ceph Intro & architecture*
- *Myth or fact*
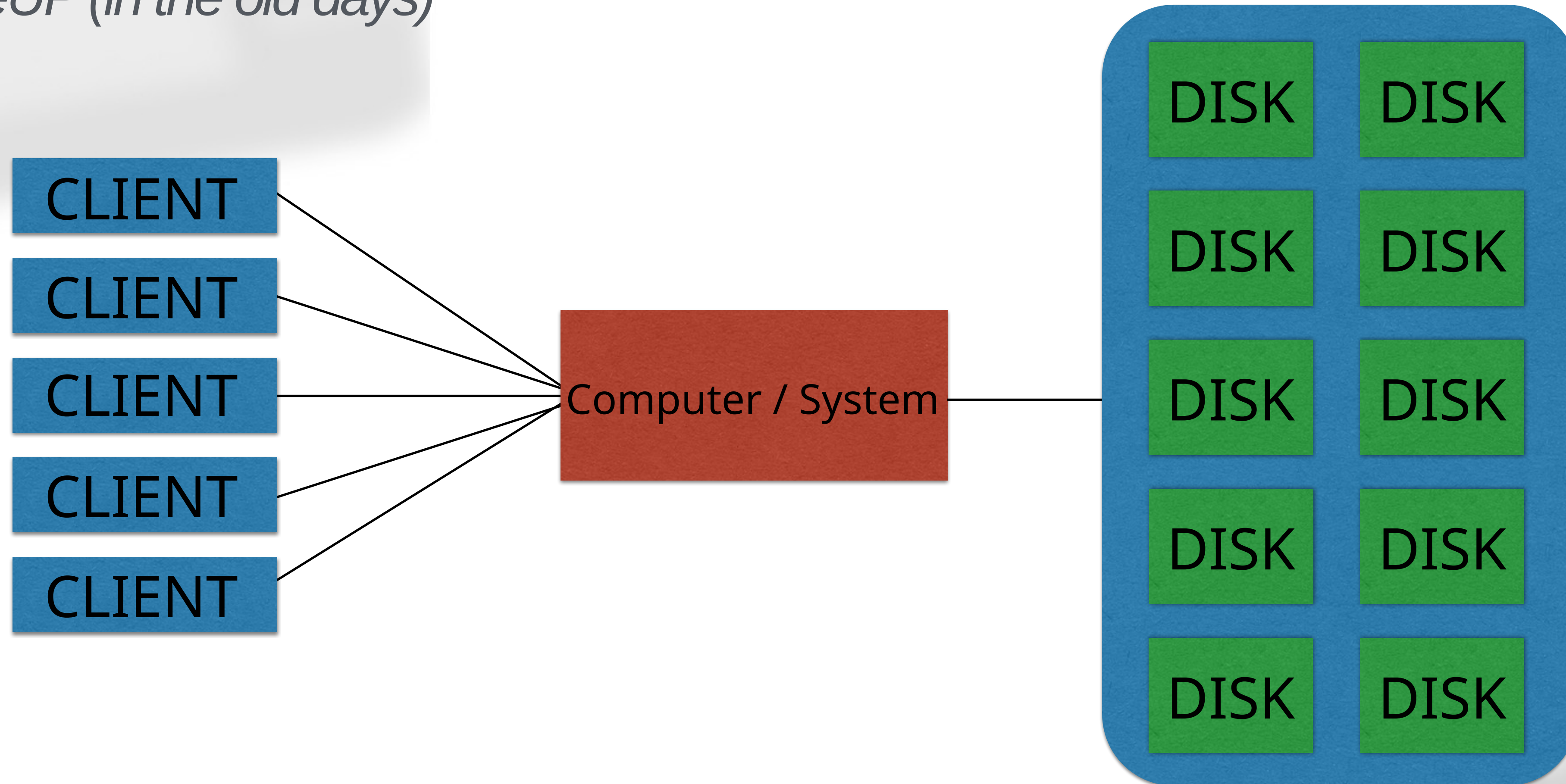- *Ceph & OpenStack*

Storage Background
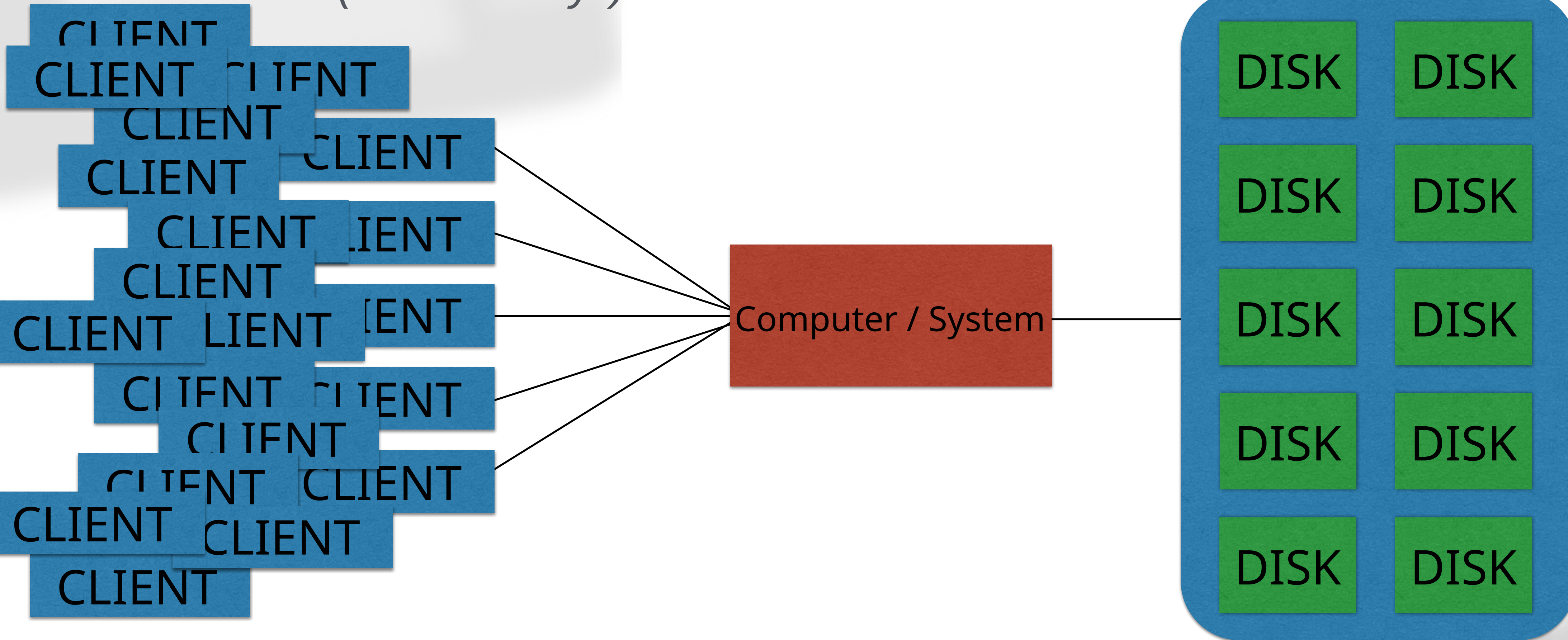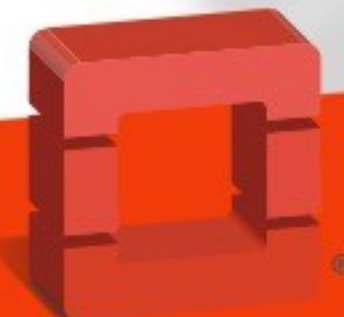*** *I'm not a storage guy* ***

OpenStack Days
MÉXICO

Storage Background

ScaleUP (in the old days)

Storage Background

ScaleUP (in the old days)

CLIENT (×many)

Computer / System

DISK DISK
DISK DISK
DISK DISK
DISK DISK
DISK DISK

OpenStack Days
MÉXICO

Storage Background
ScaleUP (in the old days)

CLIENT
CLIENT CLIENT
CLIENT
CLIENT
CLIENT
CLIENT
CLIENT
CLIENT
CLIENT
CLIENT CLIENT
CLIENT
CLIENT
CLIENT CLIENT
CLIENT
CLIENT CLIENT
CLIENT CLIENT
CLIENT

Computer / System

DISK	DISK	DISK	DISK
DISK	DISK	DISK	DISK
DISK	DISK	DISK	DISK
DISK	DISK	DISK	DISK
DISK	DISK	DISK	DISK

Storage Background
ScaleUP (in the old days)

CLIENT
CLIENT CLIENT
CLIENT
CLIENT CLIENT
CLIENT
CLIENT CLIENT
CLIENT
CLIENT CLIENT CLIENT
CLIENT
CLIENT CLIENT
CLIENT CLIENT
CLIENT CLIENT
CLIENT CLIENT
CLIENT

DISK DISK DISK DISK
DISK DISK DISK DISK
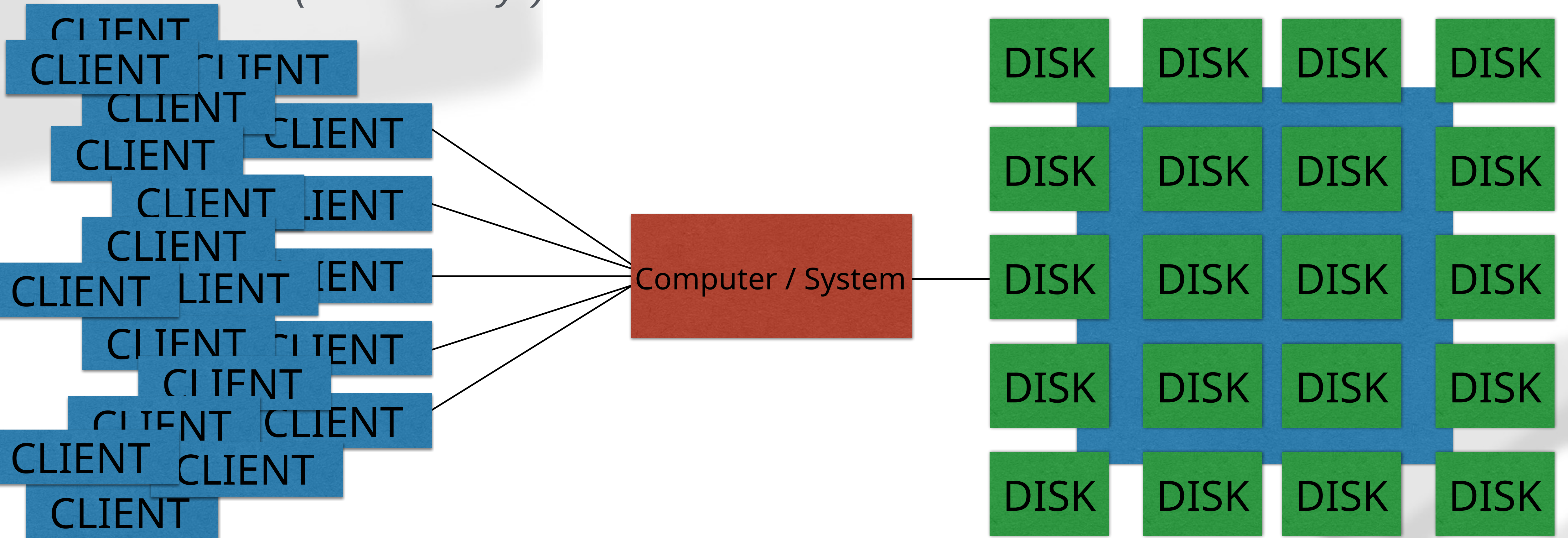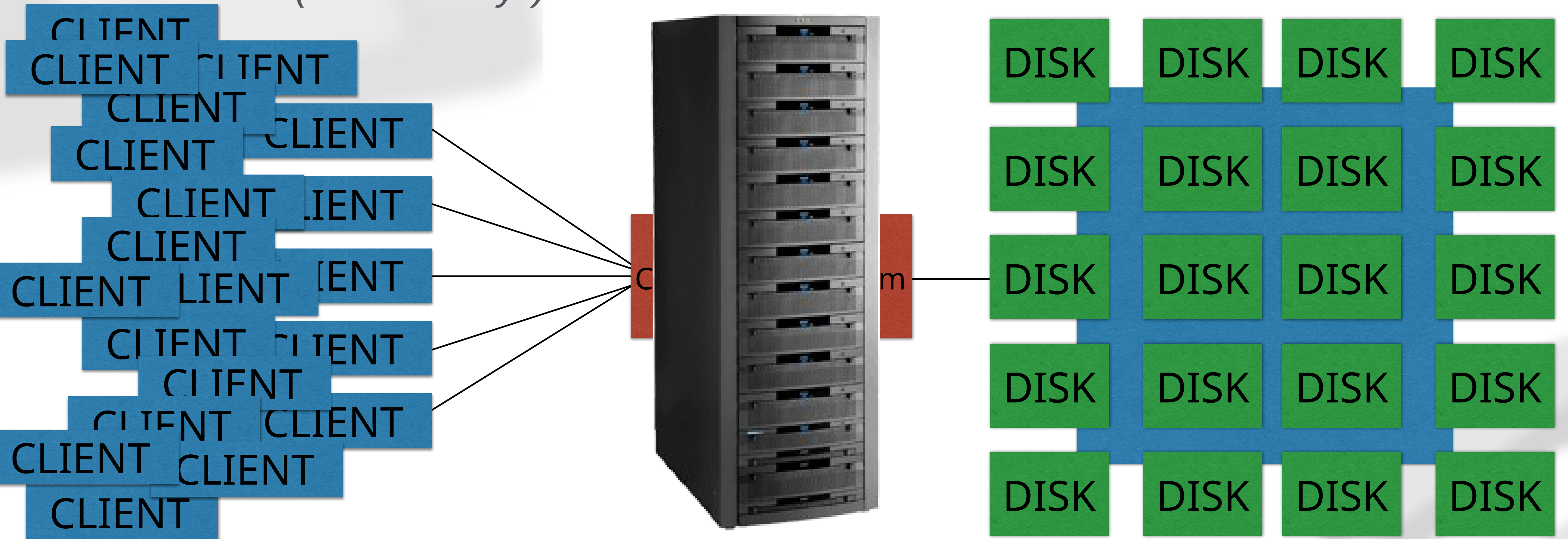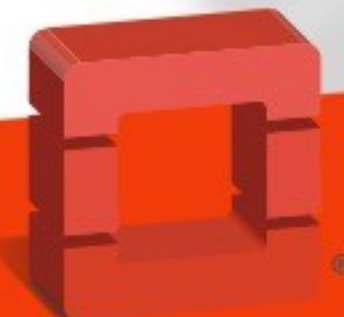DISK DISK DISK DISK
DISK DISK DISK DISK
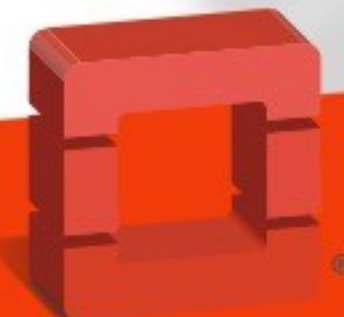DISK DISK DISK DISK

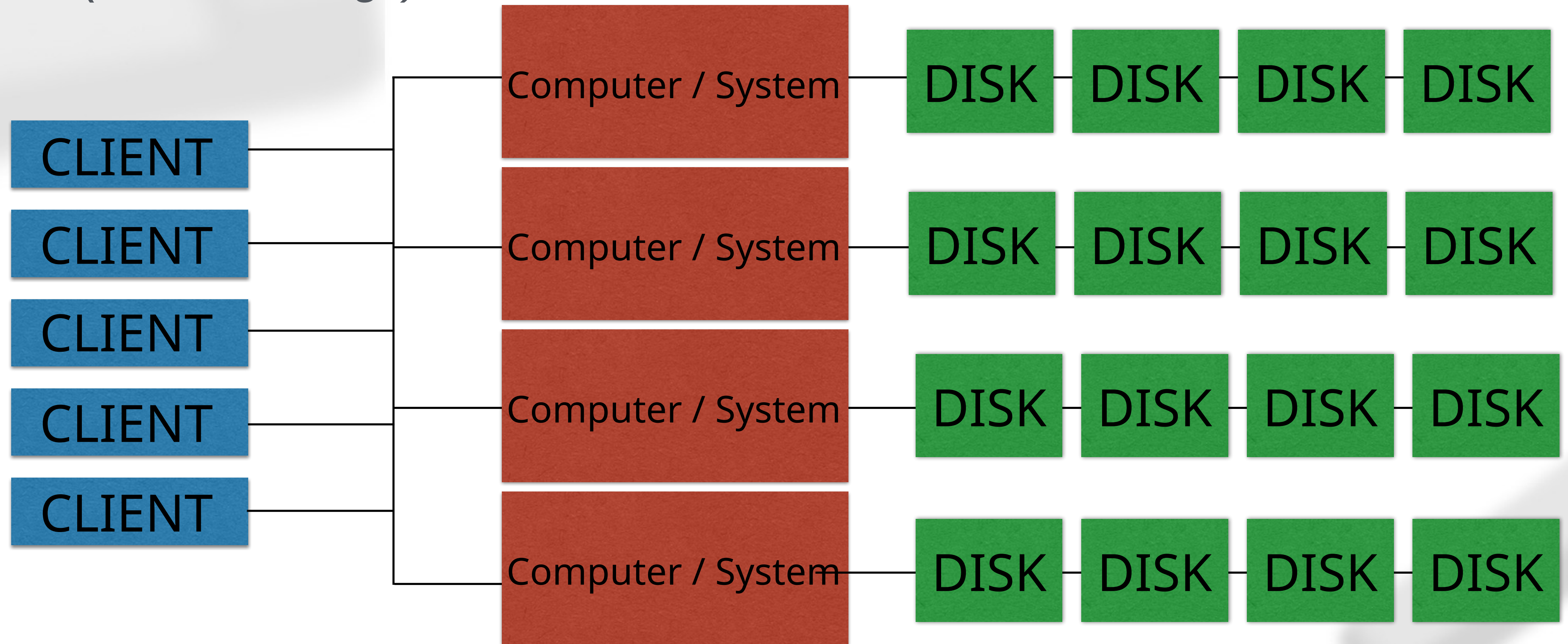# Storage Background
ScaleUP (in the old days)

Storage Background
ScaleOUT (in the cloud age)
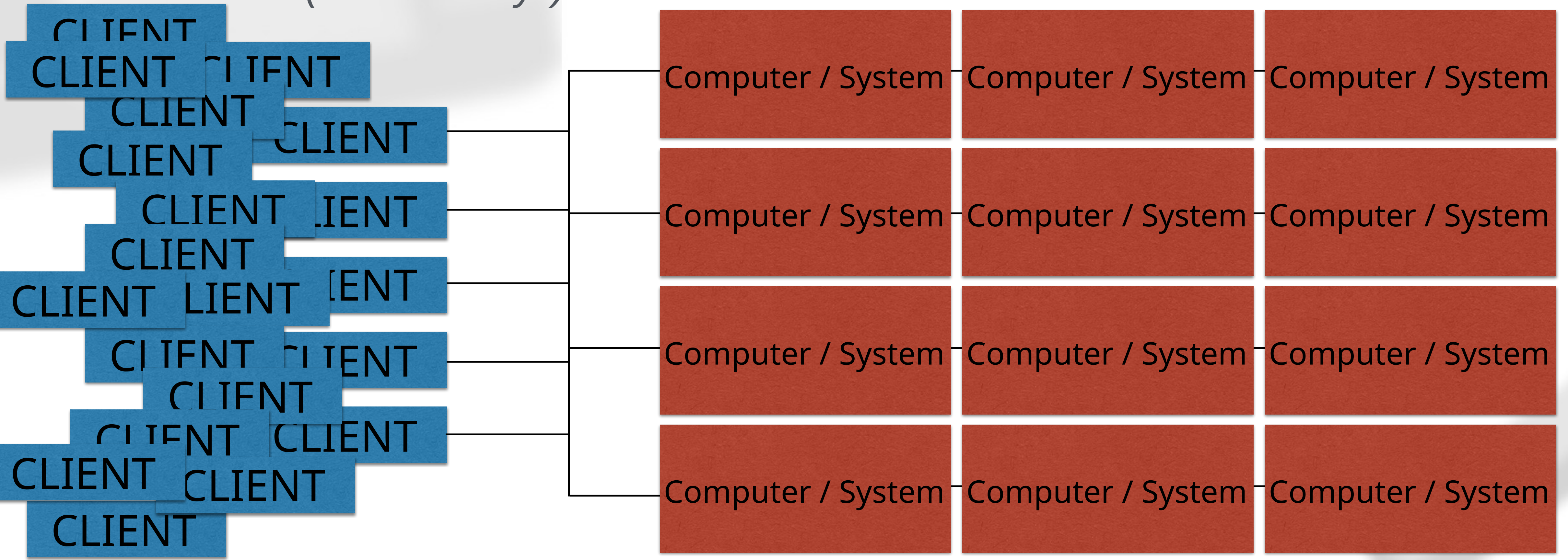
Storage Background
ScaleUP (in the old days)

CLIENT
CLIENT CLIENT
CLIENT
CLIENT CLIENT
CLIENT
CLIENT CLIENT
CLIENT
CLIENT CLIENT
CLIENT LIENT
CLIENT CLIENT
CLIENT
CLIENT CLIENT
CLIENT CLIENT
CLIENT CLIENT
CLIENT

| Computer / System	Computer / System	Computer / System
Computer / System	Computer / System	Computer / System
Computer / System	Computer / System	Computer / System
Computer / System	Computer / System	Computer / System

OpenStack Days
MÉXICO

Ceph
introduction & architecture

- Ceph was initially created by Sage Weil (DreamHost), 2007
- Linus Torvalds merged the CephFS client into Linux kernel, 2010
- Weil created Inktank Storage, 2012
- First stable release, code name Argonaut, 2012
- Minimun two release per year (10)
- Red Hat purchased Inktank, 2014
- Last stable release, code name Jewel, 2016
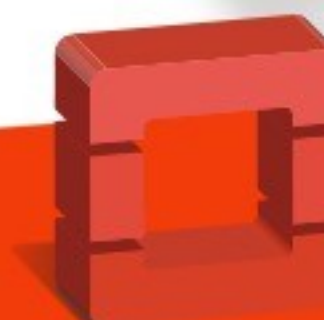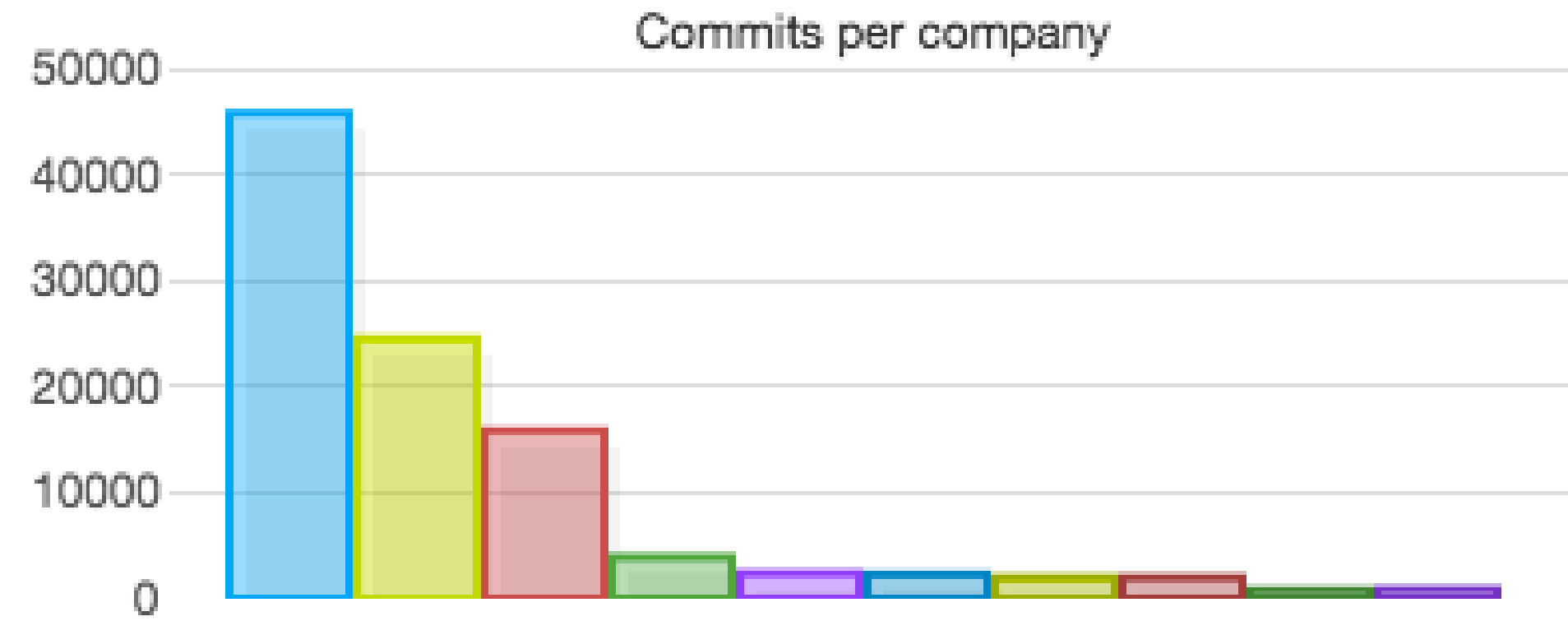
OpenStack Days
MÉXICO

Ceph
introduction & architecture

Commits per company

- Red Hat
- Inktank
- DreamHost
- The Linux Box
- CohortFS, LLC
- SUSE
- Intel
- Deutsche Telekom
- Cloudwatt
- UnitedStack

- University of California, Santa Cruz
- Mirantis
- ZTE Corporation
- Github
- Reliance Jio Infocomm, Ltd.
- 42on
- eBay
- Flipkart
- Gentoo Linux
- Ubuntu Kylin

OpenStack Days
MÉXICO

Ceph
introduction & architecture

Commits per company

- Red Hat
- Inktank
- DreamHost
- The Linux Box
- CohortFS, LLC
- SUSE
- Intel
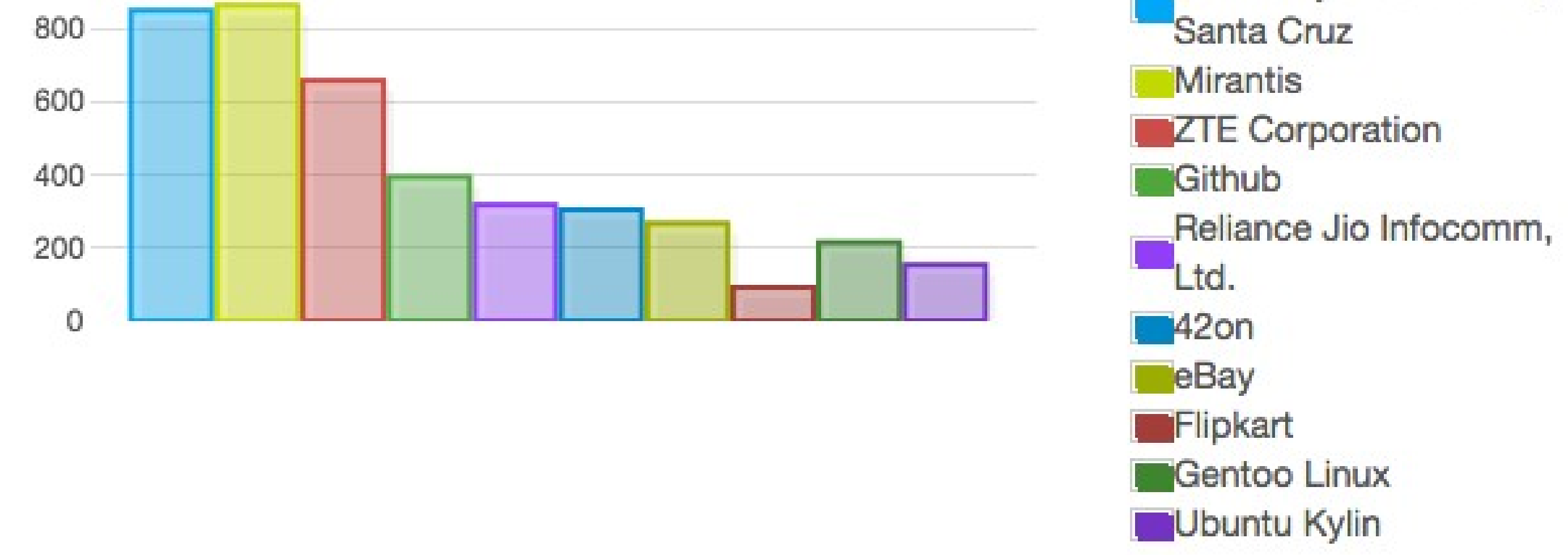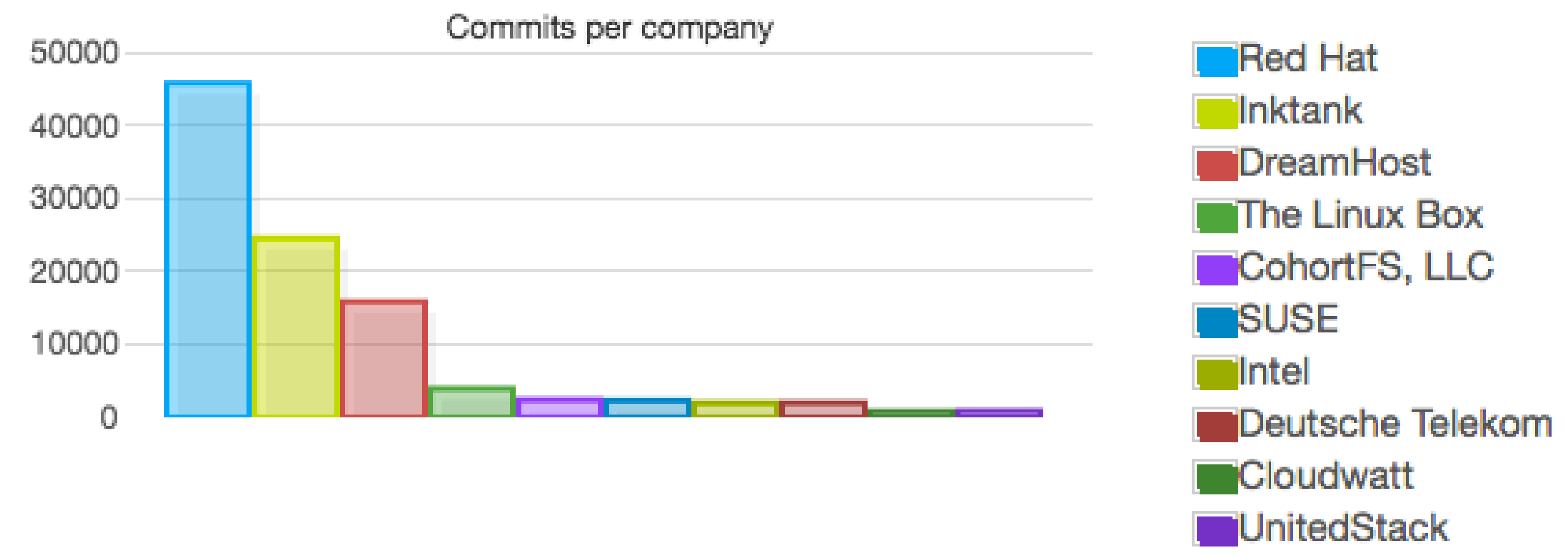- Deutsche Telekom
- Cloudwatt
- UnitedStack
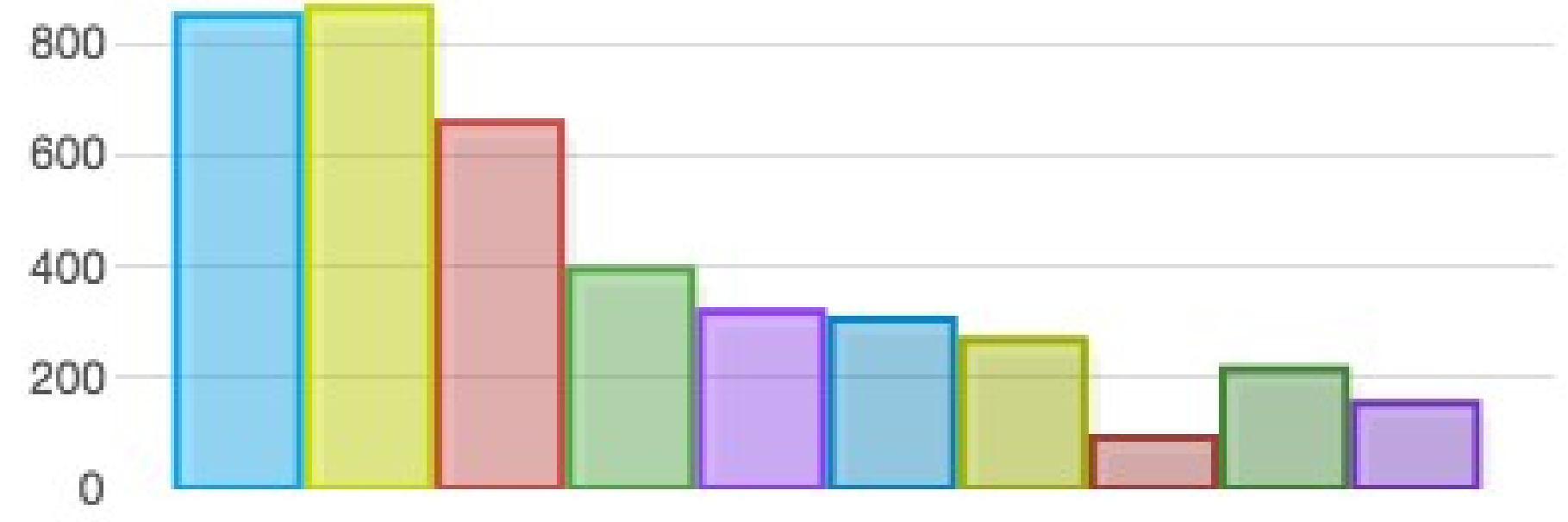
- University of California, Santa Cruz
- Mirantis
- ZTE Corporation
- Github
- Reliance Jio Infocomm, Ltd.
- 42on
- eBay
- Flipkart
- Gentoo Linux
- Ubuntu Kylin

Community focused!!!!!

OpenStack Days
MÉXICO

enterprise support**

opensource

commodity hardware
or standard hardware

OpenStack Days
MÉXICO

Ceph
introduction & architecture

enterprise support**

commodity hardware
or standard hardware

OpenSource
Community focused
No single point of failure

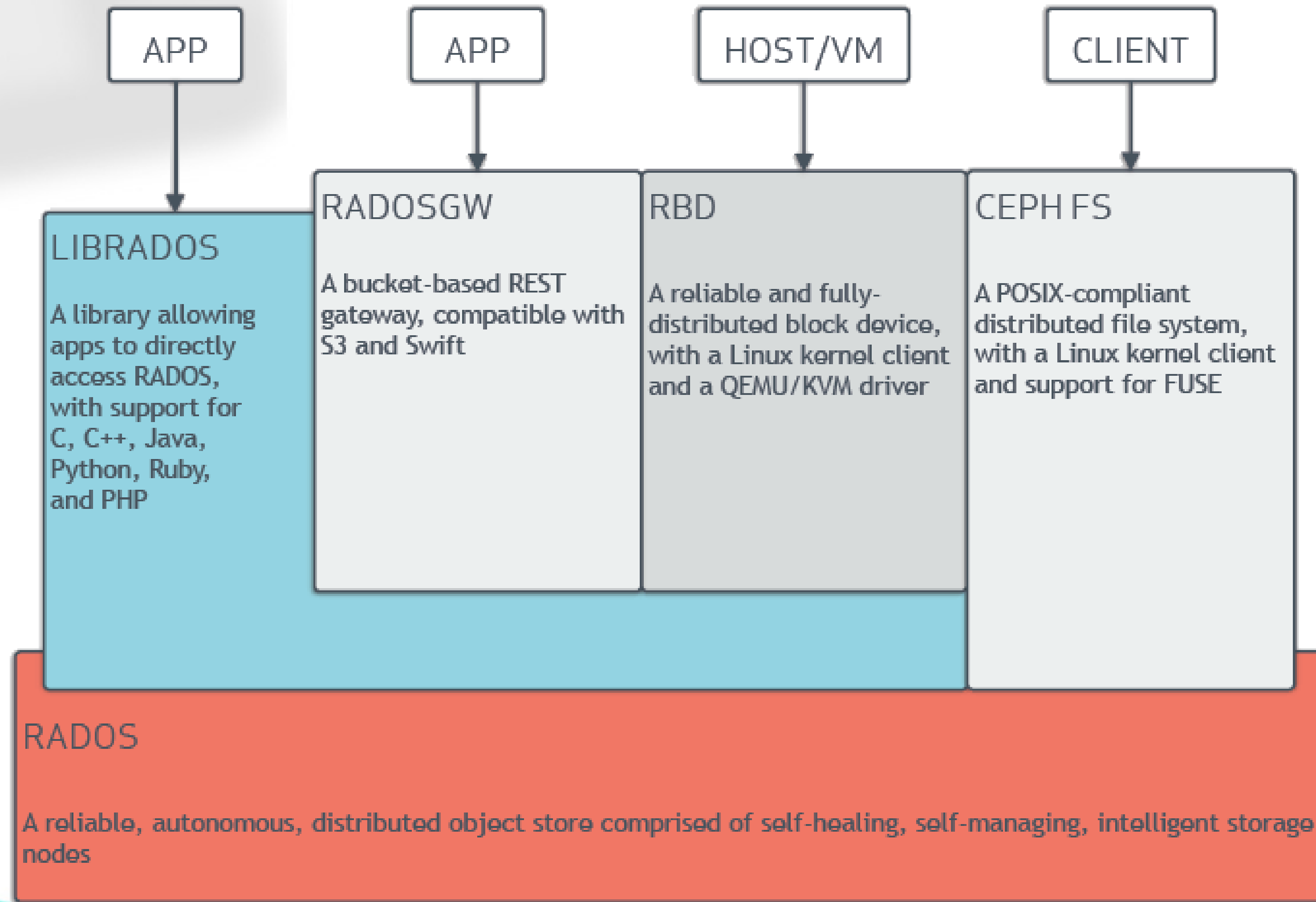**philosophy**

design

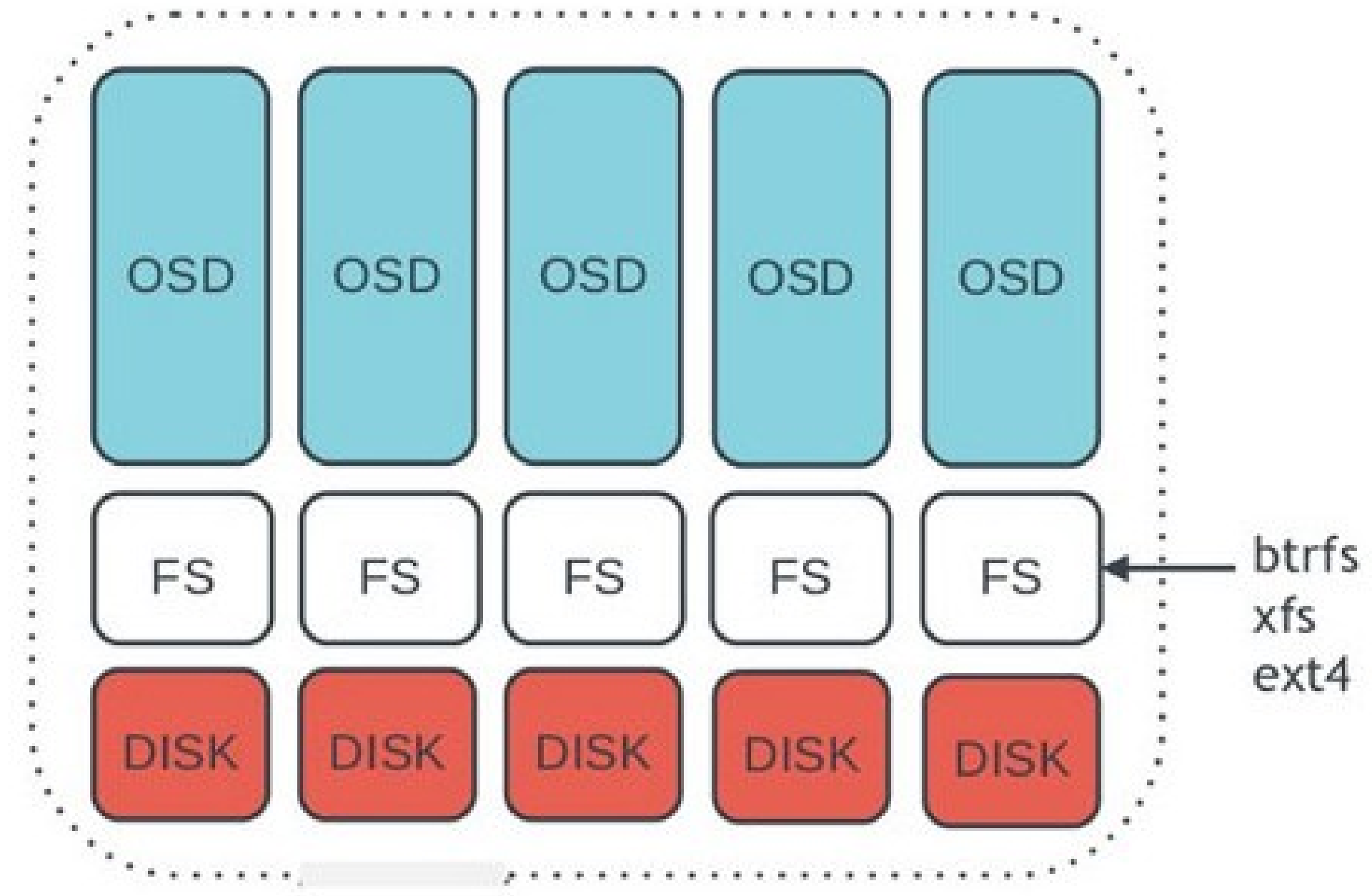Scalable

Software base

Self managing / healing

OpenStack Days
MÉXICO

Ceph
introduction & architecture

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

Ceph
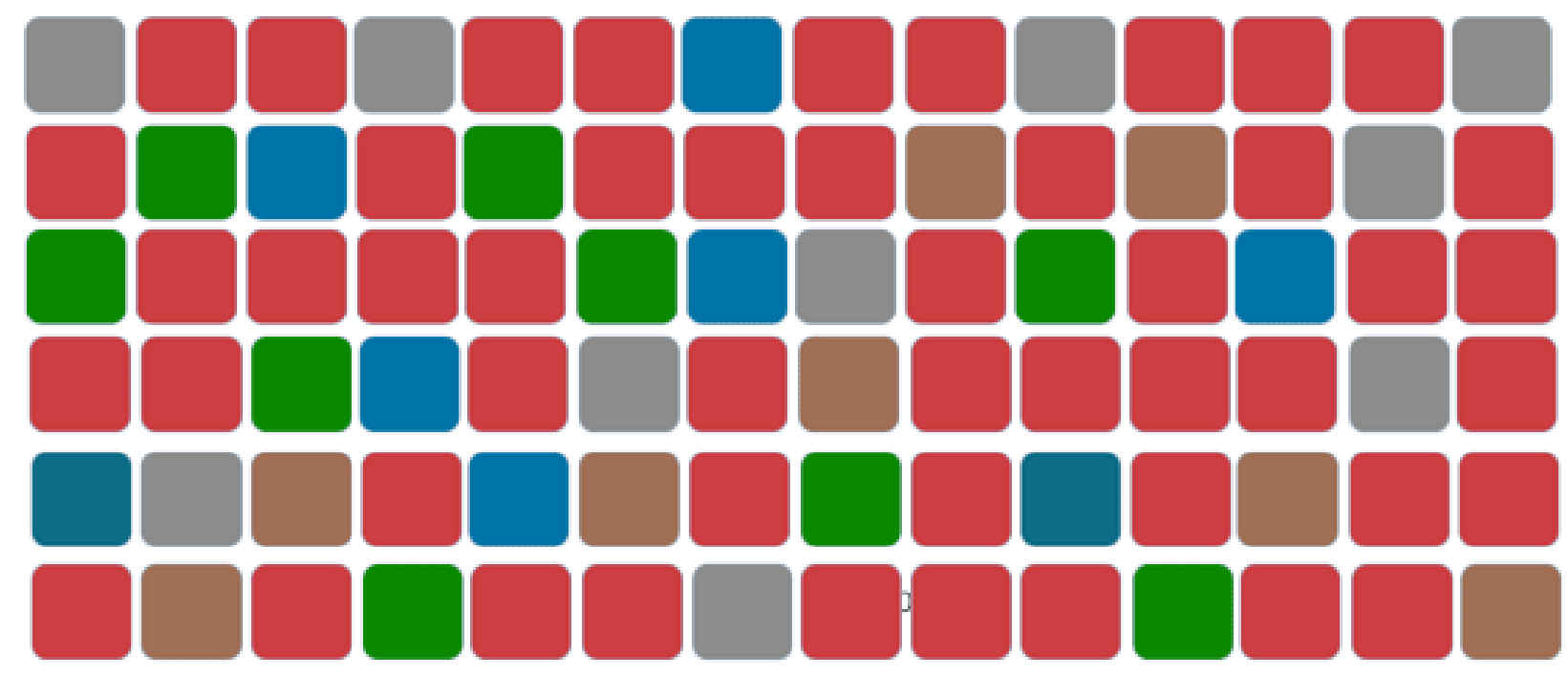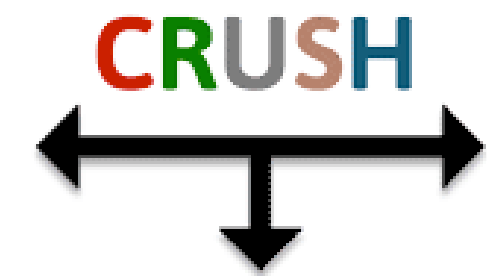*introduction & architecture*

btrfs
xfs
ext4

Data placement with CRUSH

Pseudo-random placement algorithm

Rule based configuration

CRUSH

Controlled Replication Under Scalable Hashing

Myth or fact

Myth or fact
Performance???

Ceph is a distributed object store and file system designed to provide excellent performance, reliability and scalability.

http://www.ceph.com

1.5.2 Optimizations for Computer Architectures

The rapid evolution of computer architectures has also led to an insatiable demand for new compiler technology. Almost all high-performance systems take advantage of the same two basic techniques: *parallelism* and *memory hierarchies*. Parallelism can be found at several levels: at the *instruction level*, where multiple operations are executed simultaneously and at the *processor level*, where different threads of the same application are run on different processors. Memory hierarchies are a response to the basic limitation that we can build very fast storage or very large storage, but not storage that is both fast and large.

**Compilers: Principles, Techniques, and Tools
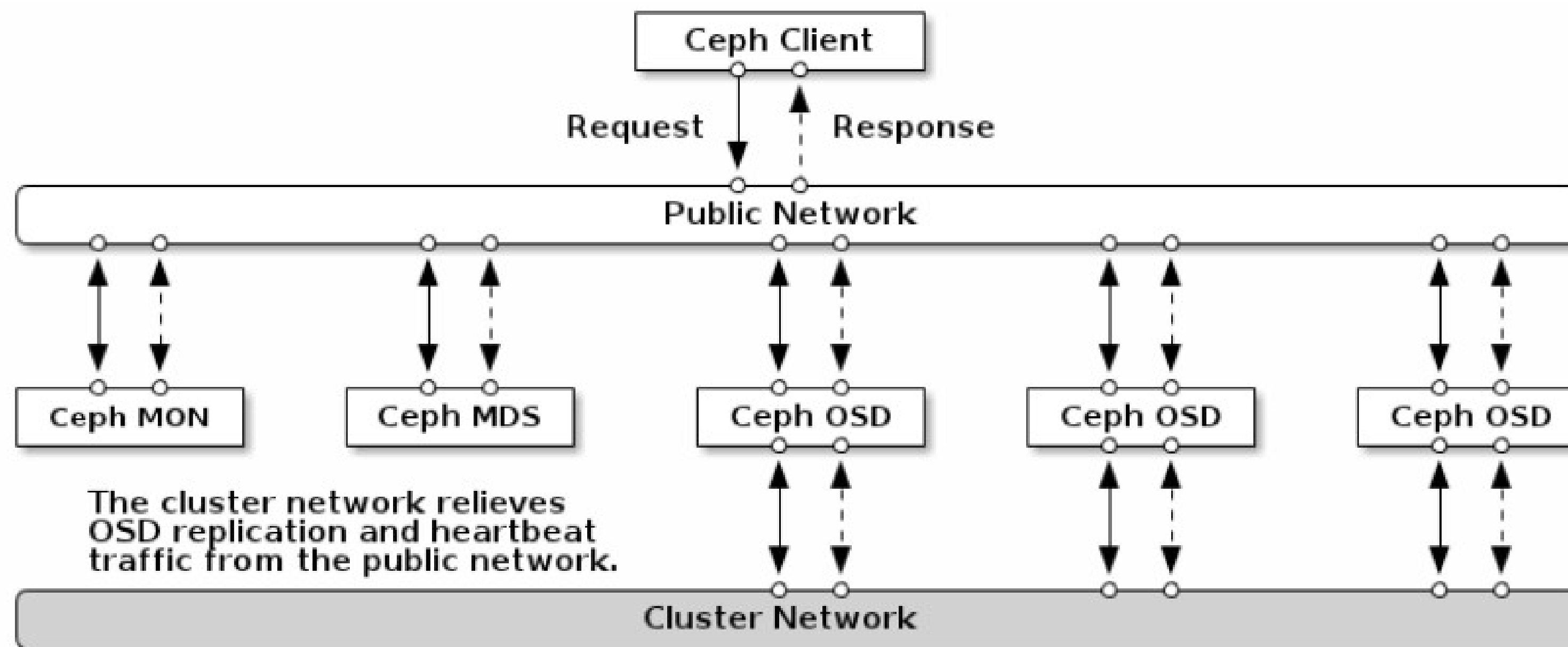by Alfred Aho, Jeffrey Ullman, Monica S. Lam, and Ravi Sethi**

OpenStack Days
MÉXICO

Myth or fact
Reliability???

Ceph is a distributed object store and file system designed to provide excellent performance, reliability and scalability.

http://www.ceph.com

OpenStack Days
MÉXICO

Myth or fact
Scalability???

Ceph is a distributed object store and file system designed to provide excellent performance, reliability and scalability.

http://www.ceph.com

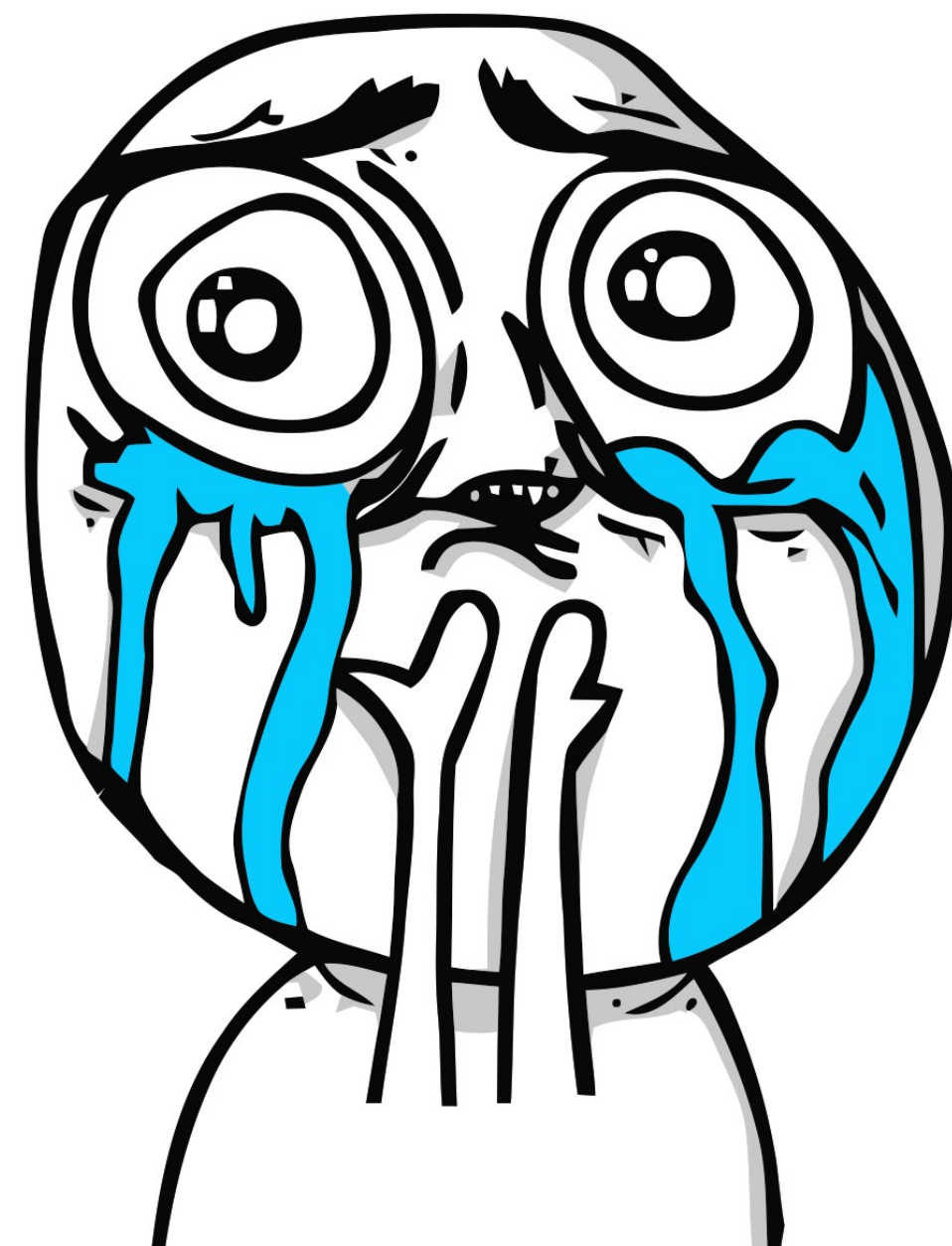RADOS: A Scalable, Reliable Storage Service for Petabyte-scale Storage Clusters

OpenStack Days
MÉXICO

Myth or fact
Scalability???

Ceph is a distributed object store and file system designed to provide excellent performance, reliability and scalability.

http://www.ceph.com

Myth or fact
Scalability???

Ceph is a distributed object store and file system designed to provide excellent performance, reliability and scalability.
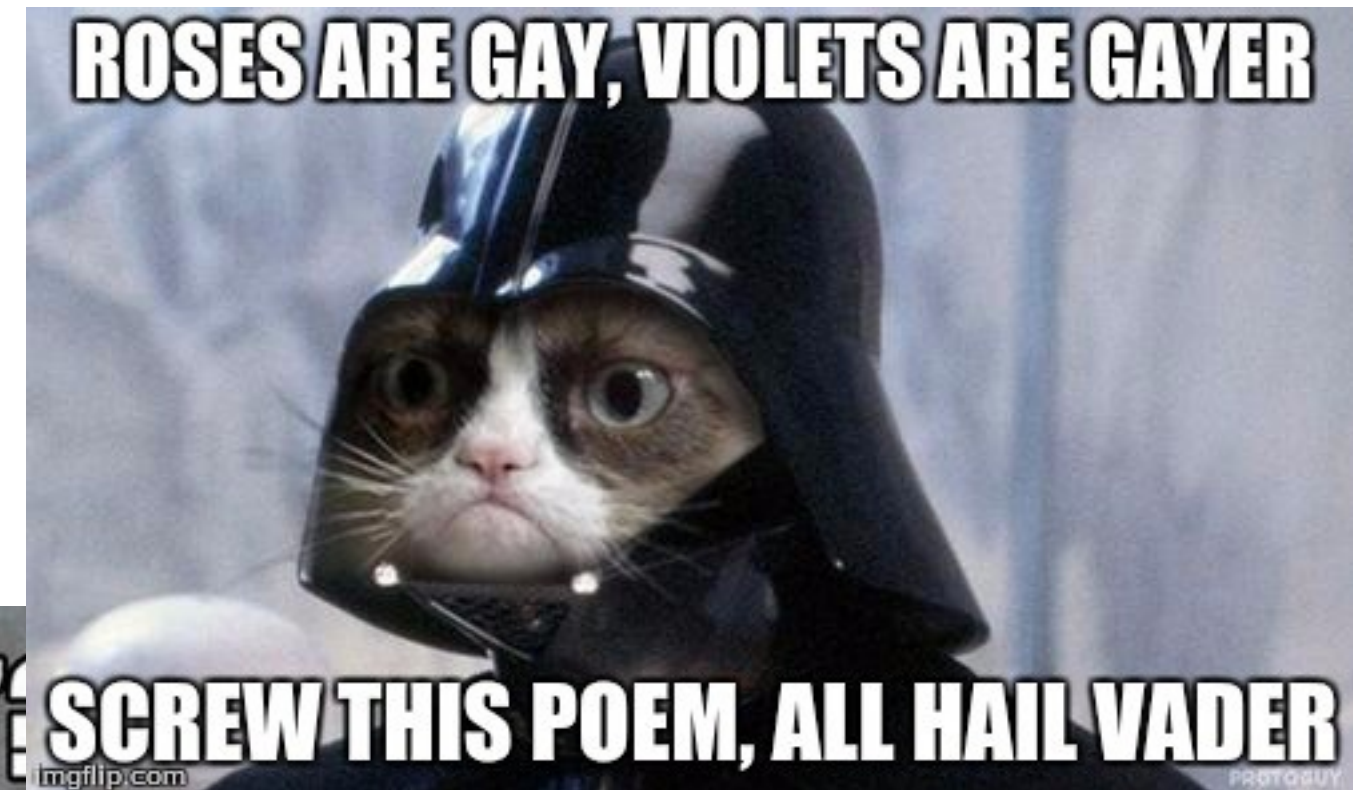
http://www.ceph.com

ROSES ARE GAY, VIOLETS ARE GAYER

SCREW THIS POEM, ALL HAIL VADER

"Good Morning":

No such thing.

RAI...   ...torage Service for Petabyte-scale
                          ...e Clusters

OpenStack Days
MÉXICO

Myth or fact
Scalabi...

Ceph i... ...and file
syste... ...t
perfo...

http://www.cep...

ROSES ARE GAY, VIOLETS ARE GAYER

SCREW THIS POEM, ALL HAIL VADER

...Morning"...

LIKE A

BOSS

No such thing.

RAI... ...e-scale

OpenStack Days
MÉXICO

Myth or fact
Scalability???

Ceph is a distributed object store and file system designed to provide excellent performance,

INSIDE THE CEPH EXASCALE STORAGE AT YAHOO

April 16, 2015   Timothy Prickett Morgan

The hyperscale giants can't wait for the IT technologies that they need for their own when the time is right an open source alt etabyte-scale

Storage Clusters

OpenStack Days
MÉXICO

Myth or fact
Scalability???

Ceph is a distributed object store and file system designed to provide excellent performance,

http://www.ceph.com

INSIDE THE CEPH EXASCALE STORAGE AT YAHOO

April 16, 2015   Timothy Prickett Morgan

The hyperscale giants can't wait for the IT technologies that they need for their own when the time is right an open source alt etabyte-scale

Storage Clusters

The architecture is inherently scalable without any theoretical boundaries

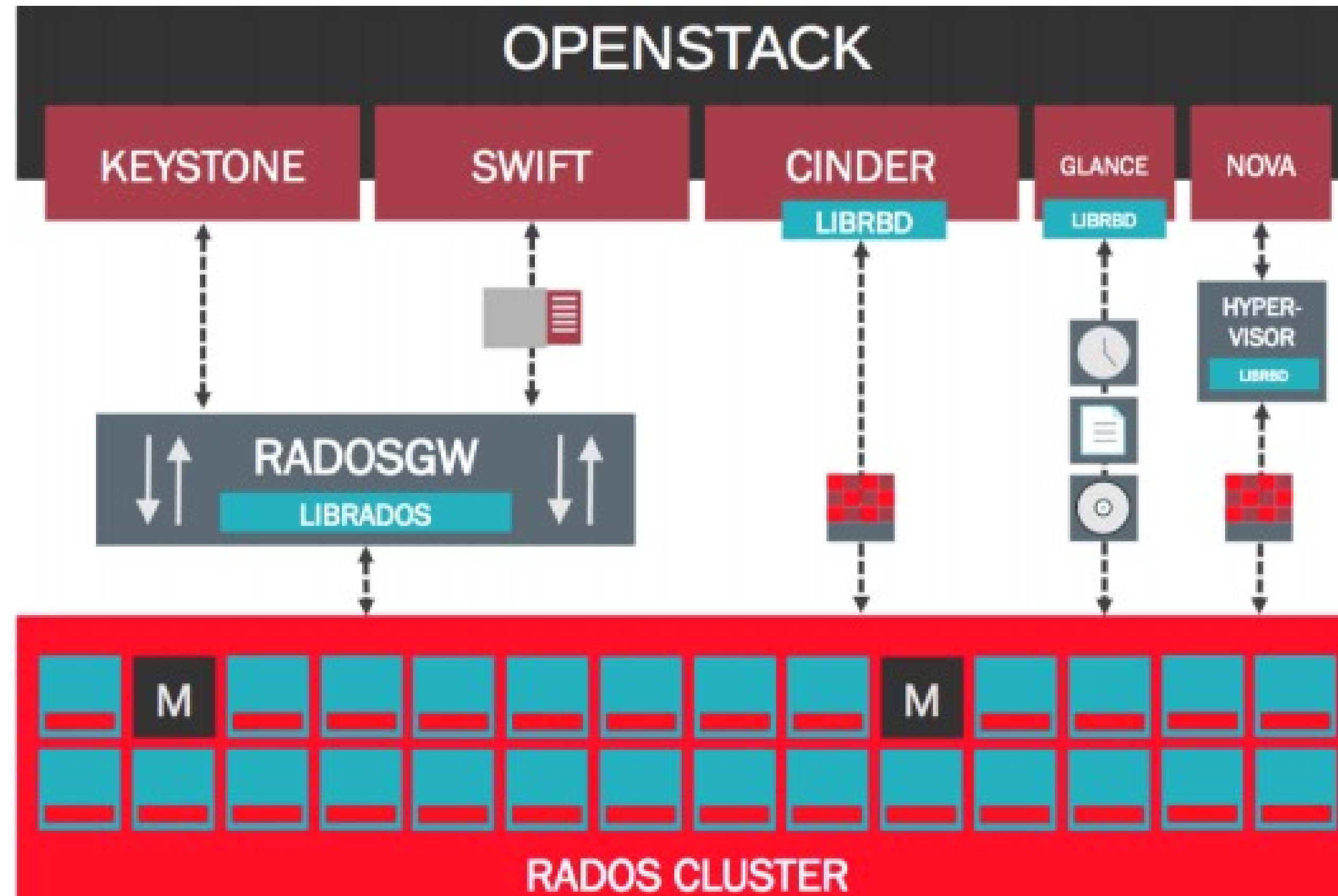OpenStack Days
MÉXICO

Ceph & OpenStack
Integration

Image by RedHat

Ceph & OpenStack
Features (Some cool)

Copy-on-write snapshots (RBD)
KRBD for BARE (RBD)
Tiering (Ceph Pools)
Leaf configuration
Ceilometer integration for RGW
Multi-attach for RBD (Cinder)

Ceph & OpenStack
Features (Some cool)

Import/export snapshots and RBD (Cinder)
Differential backup orchestration (Cinder)
Deep flatten (RBD Snap)
RBD mirroring integration (Jewel)
CephFS with Ganesha NFS -> Manila (Jewel)
DevStack Ceph (From Kilo)

Q & A

OpenStack Days
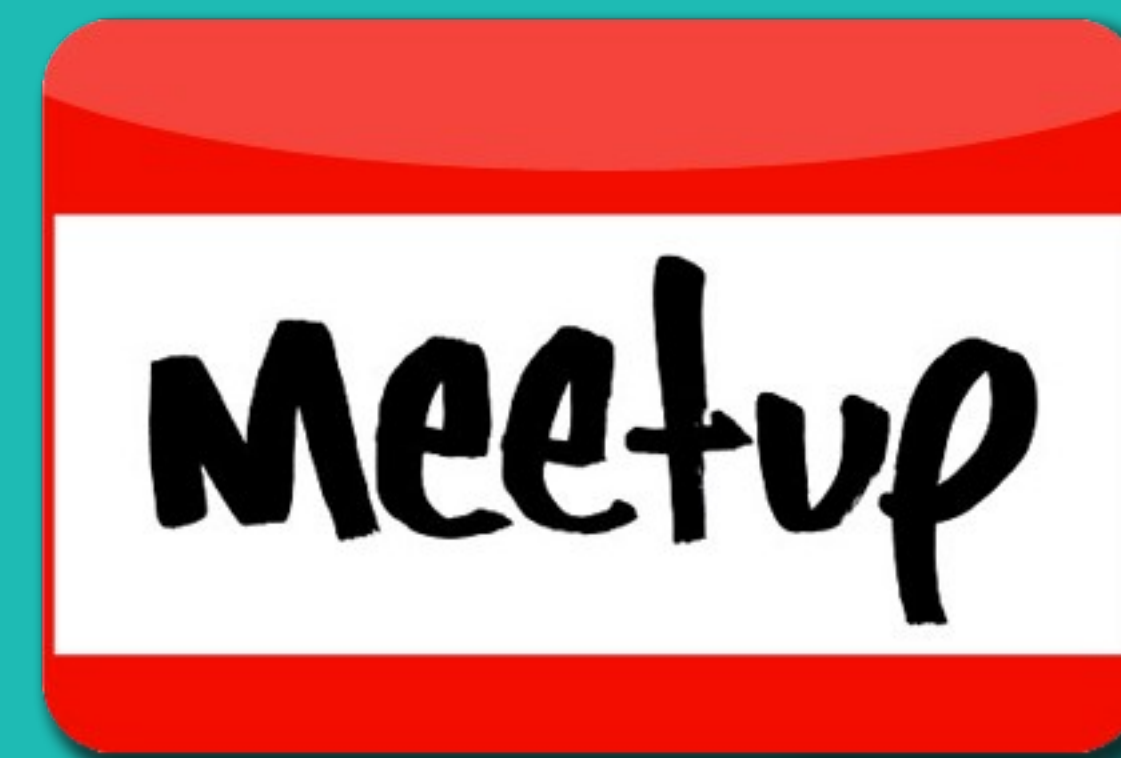**MÉXICO**

THANK YOU

locate

http://headup.ws

#irc khyron

@alsotoes

alsotoes@gmail.com

asoto@kionetworks.com

Meetup

OpenStack Days
MÉXICO

CephersMX

StackersMX

Myth or fact
Storage backend de facto???

Block Storage Drivers: Production

| Driver	May 2015	November 2014
Ceph RBD	44%	37%
LVM (default)	16%	23%
Net App	11%	10%
NFS	5%	7%
GlusterFS	9%	10%
EMC	4%	4%
SolidFire	5%	3%
VMWare VMDK	1%	2%
Dell EqualLogic	2%	1%
HP3PAR	2%	1%

OpenStack Days
MÉXICO
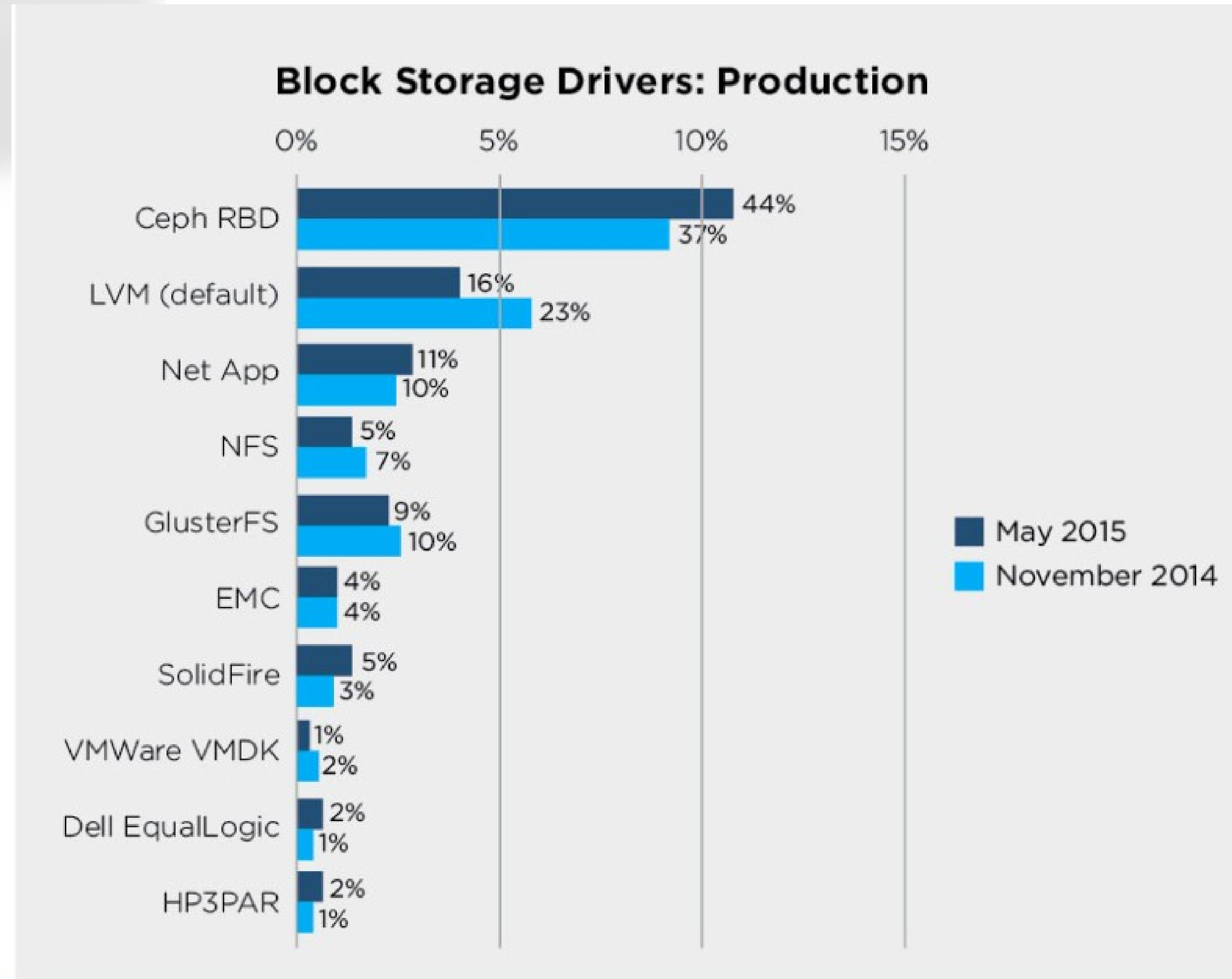
Ceph & OpenStack
Features (Some cool)

```
ceph-deploy disk zap vm04:vdb
ceph-deploy osd create --dmcrypt vm04:vdb

ceph osd getcrushmap -o crushmap.compiled
crushtool -d crushmap.compiled -o crushmap.decompiled
```
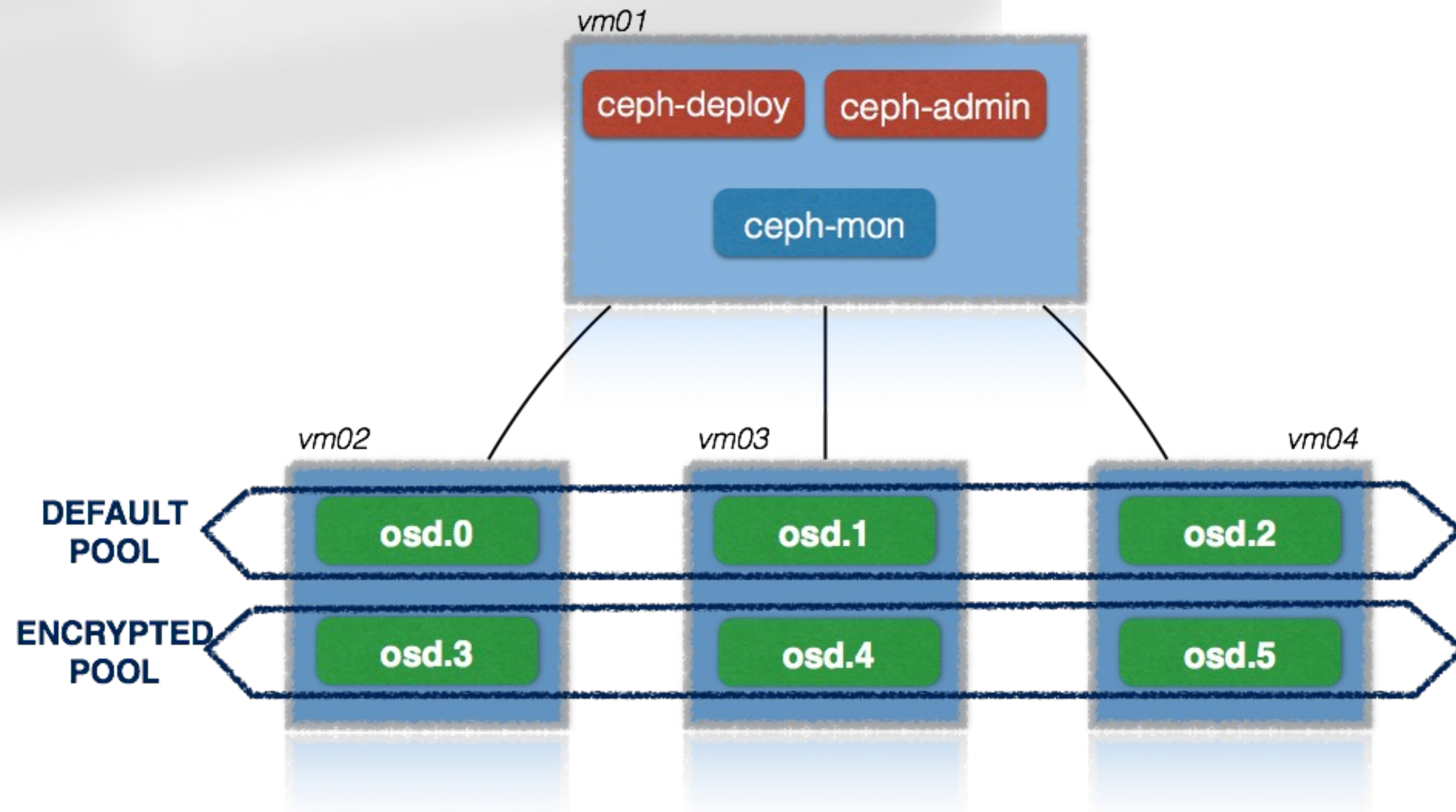


```
host vm04-encr {
    id -7        # do not change unnecessarily
    # weight 0.080
    alg straw
    hash 0  # rjenkins1
    item osd.5 weight 0.040
}

root encrypted {
    id -8        # do not change unnecessarily
    # weight 0.120
    alg straw
    hash 0  # rjenkins1
    item vm02-encr weight 0.040
    item vm03-encr weight 0.040
    item vm04-encr weight 0.040
}
```

```
rule encrypted_ruleset {
    ruleset 1
    type replicated
    min_size 1
    max_size 10
    step take encrypted
    step chooseleaf firstn 0 type host
    step emit
}
```

```
ceph osd pool create encrypted 128
ceph osd pool set encrypted crush_ruleset 1
```