Storage with Ceph

Scale-Out made easy

Martin Gerhard Loschwitz

hastexo!

© 2013 hastexo Professional Services GmbH. All rights reserved.

h!

Wer?

h!

h!

h!

h!


LINUX
MAGAZIN


ADMIN
www.admin-magazin.de


Pacemaker


SLAC
Secure Linux
Administration
Conference 2013

h!


XTREME BOWL


WIEN

hastexo!

Skalierbares Storage

2 Arten der Skalierbarkeit

Scale-Up

Scale-Up vorher:

Scale-Up nachher:

Scale-Up stößt schnell
an seine **Grenzen**

Scale-Out

Scale-Out vorher:

Scale-Out nachher:

Scale-Out ist hip

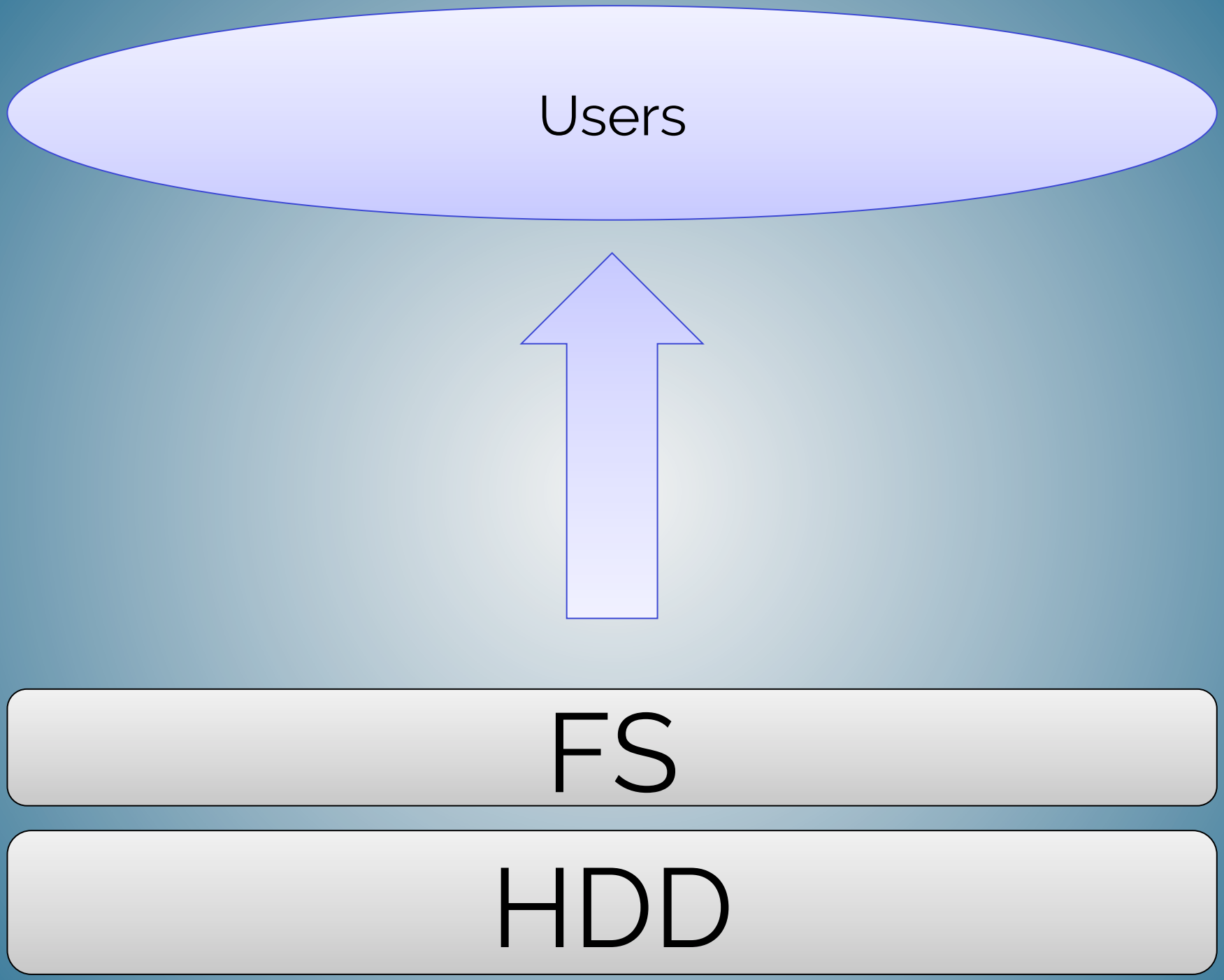
Webserver

h!

Datenbanken

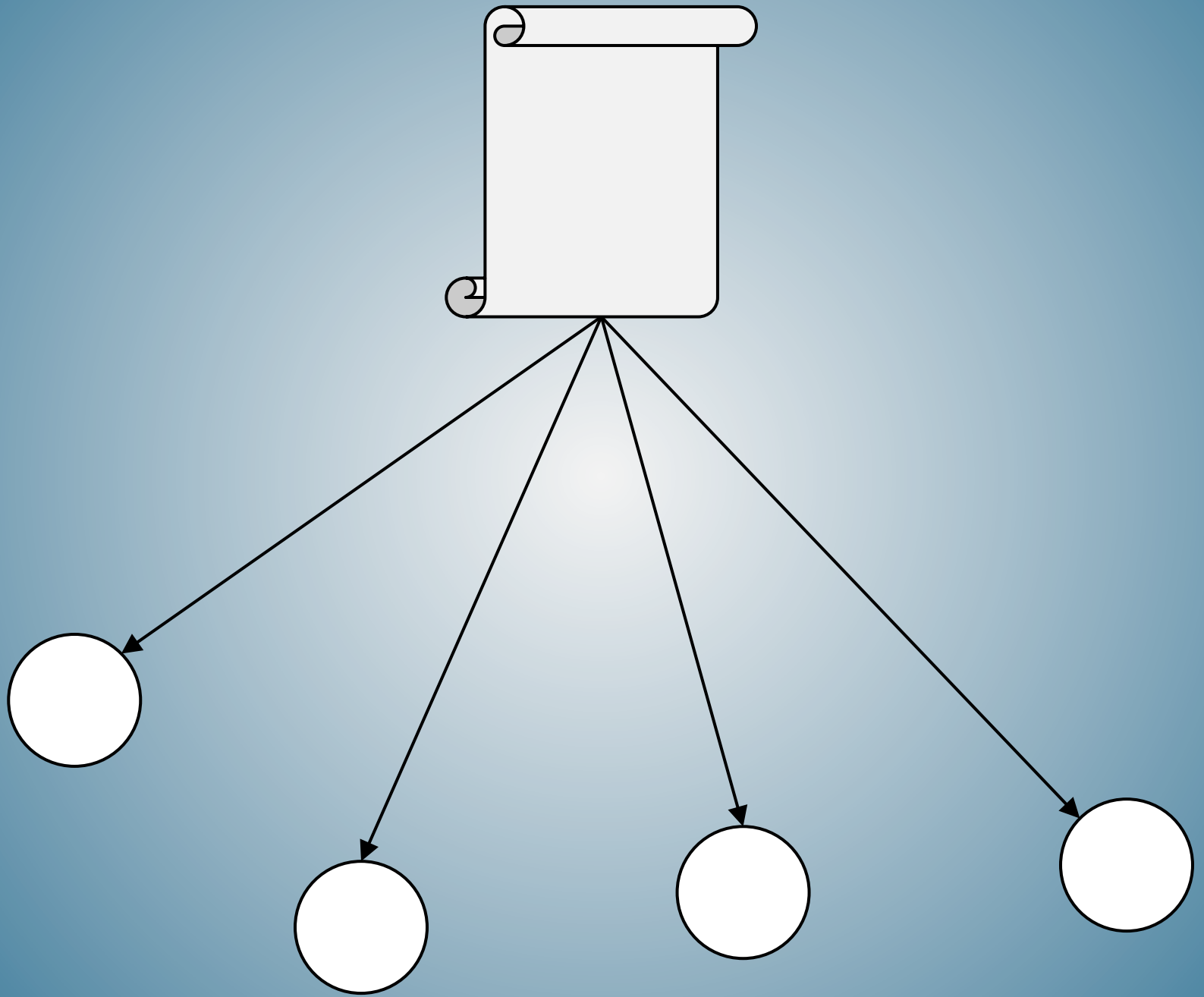
Storage? Meh.

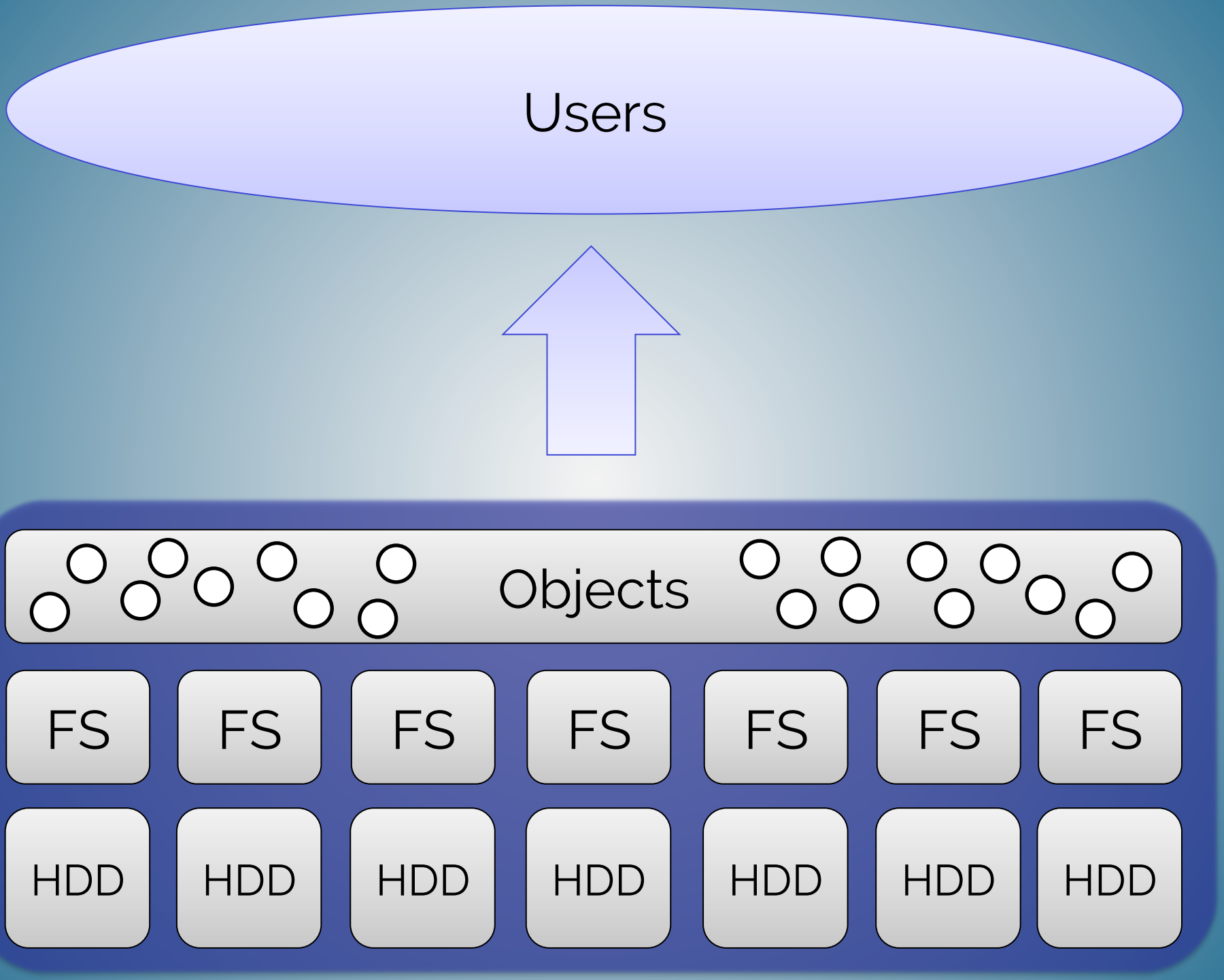
Das Block-Problem


Users
FS
HDD

Mit Blöcken geht
Scale-Out
nicht vernünftig

Object Stores



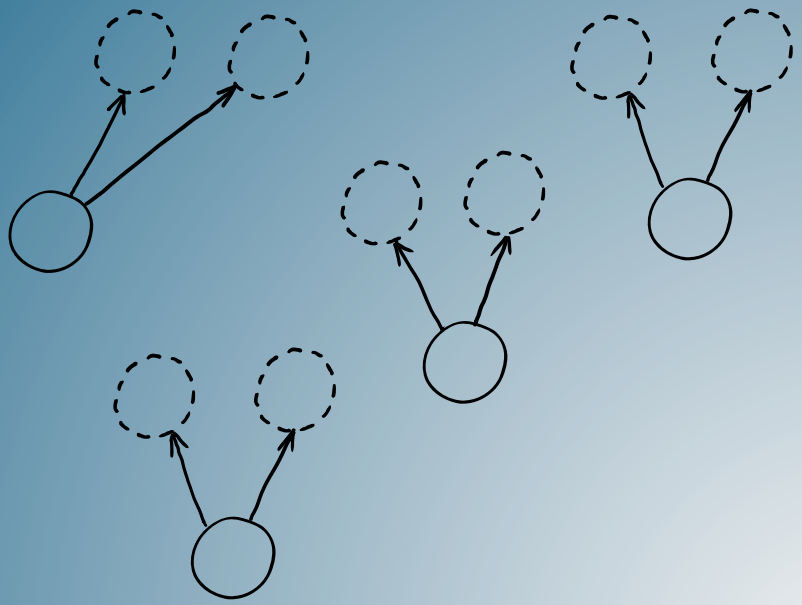

Users
Objects
FS
FS
FS
FS
FS
FS
FS
HDD
HDD
HDD
HDD
HDD
HDD
HDD

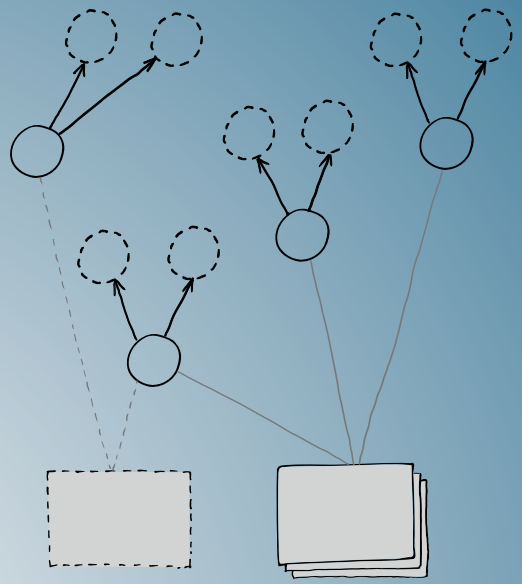
ceph

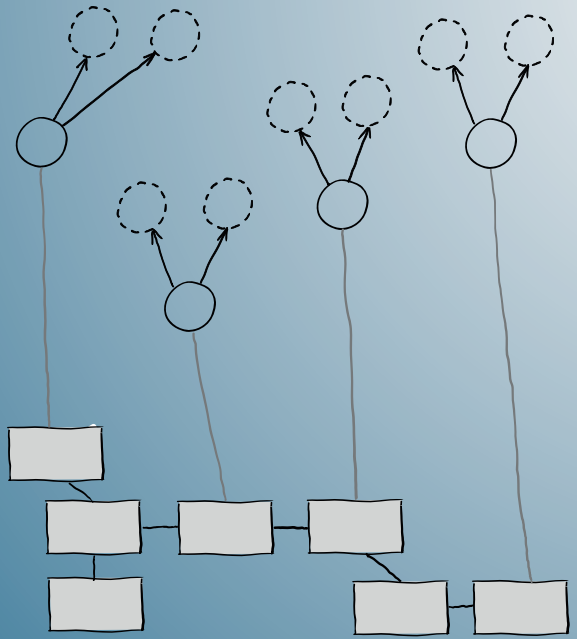
h!

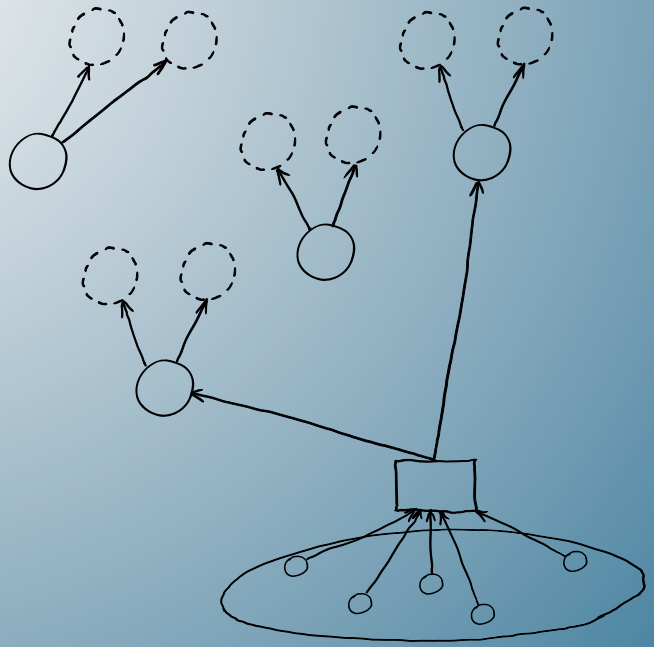
Cephalopod (Wikipedia, user Nhobgood)

Doktorarbeit von Sage Weil

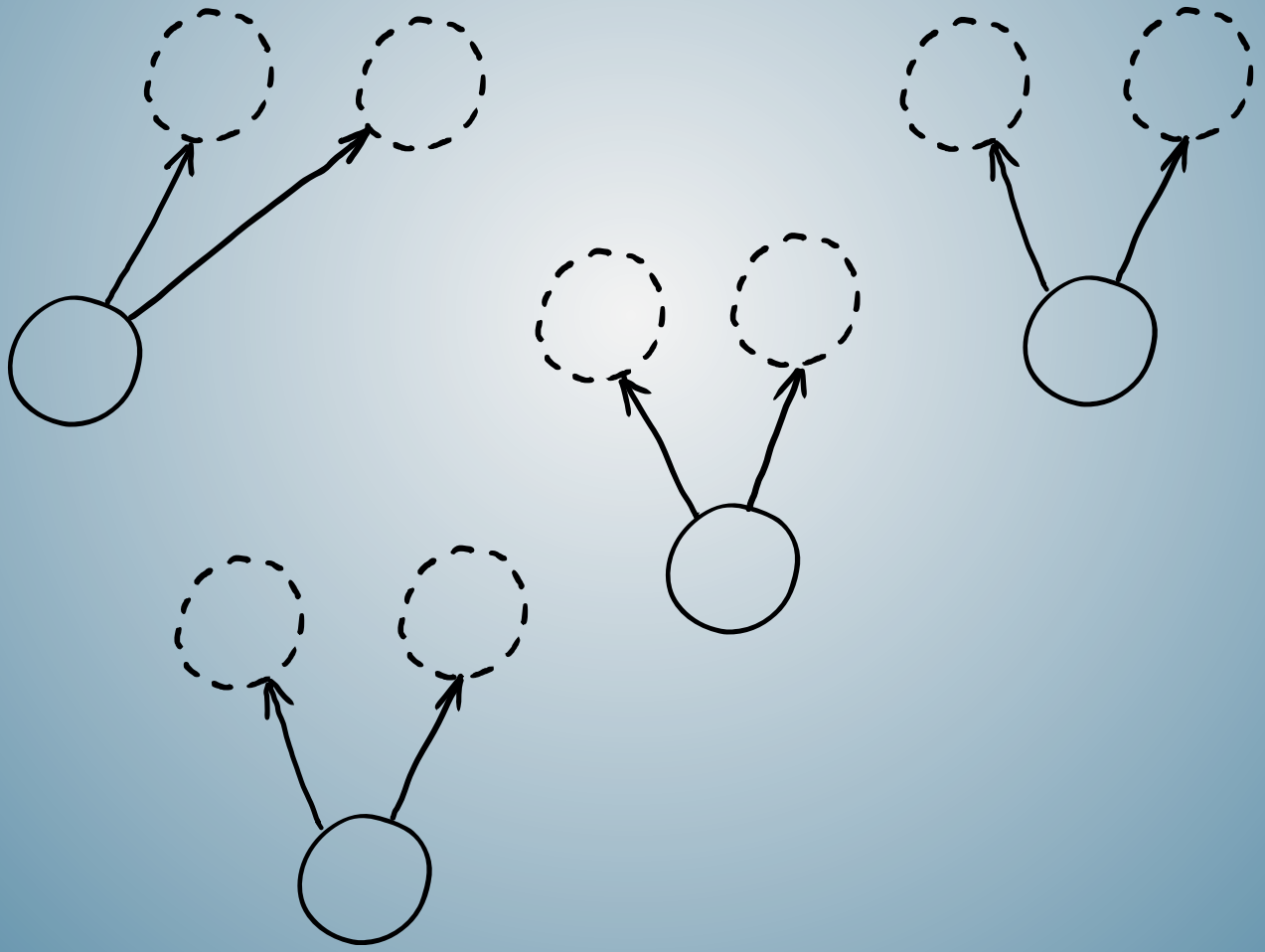








RADOS



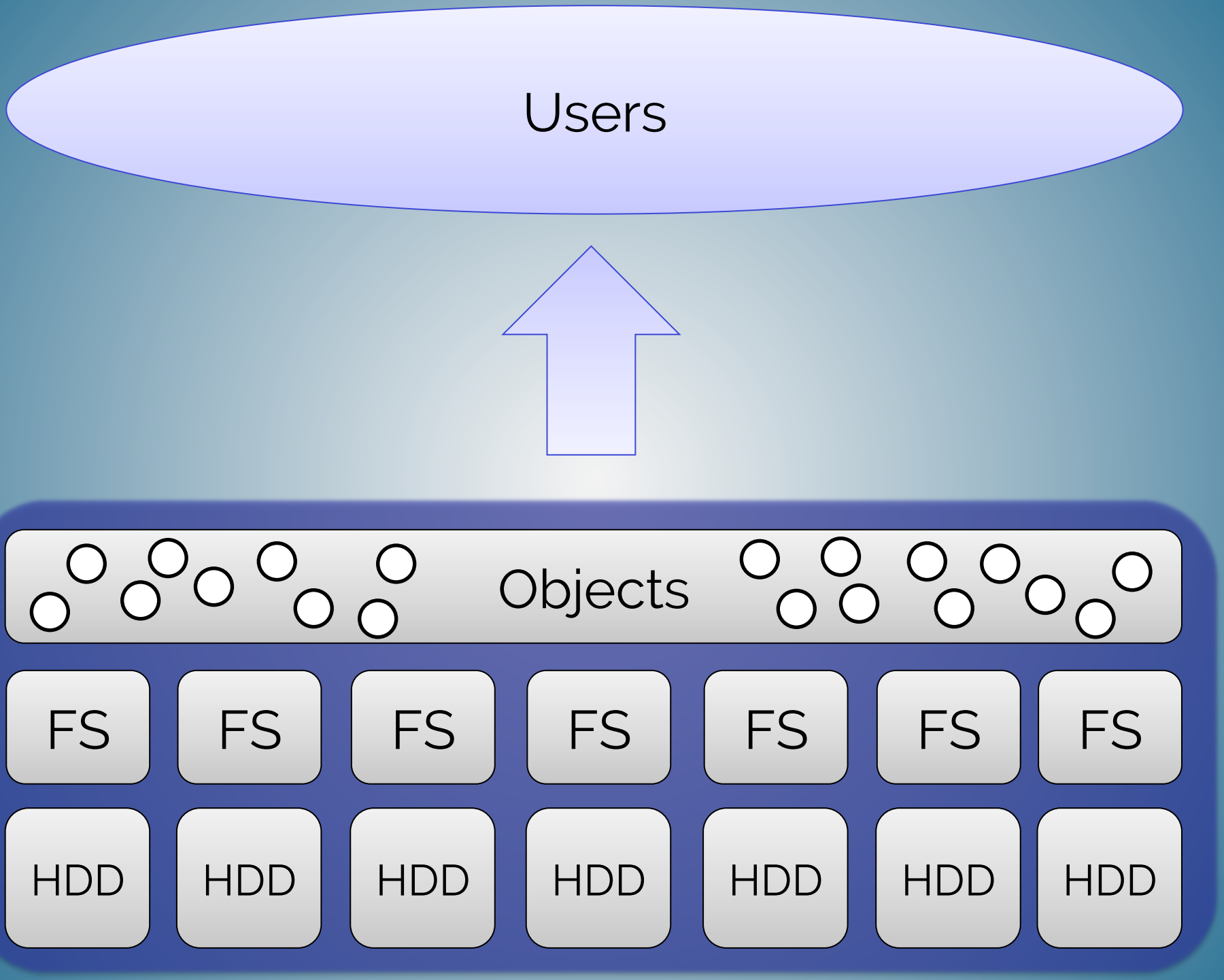
h!

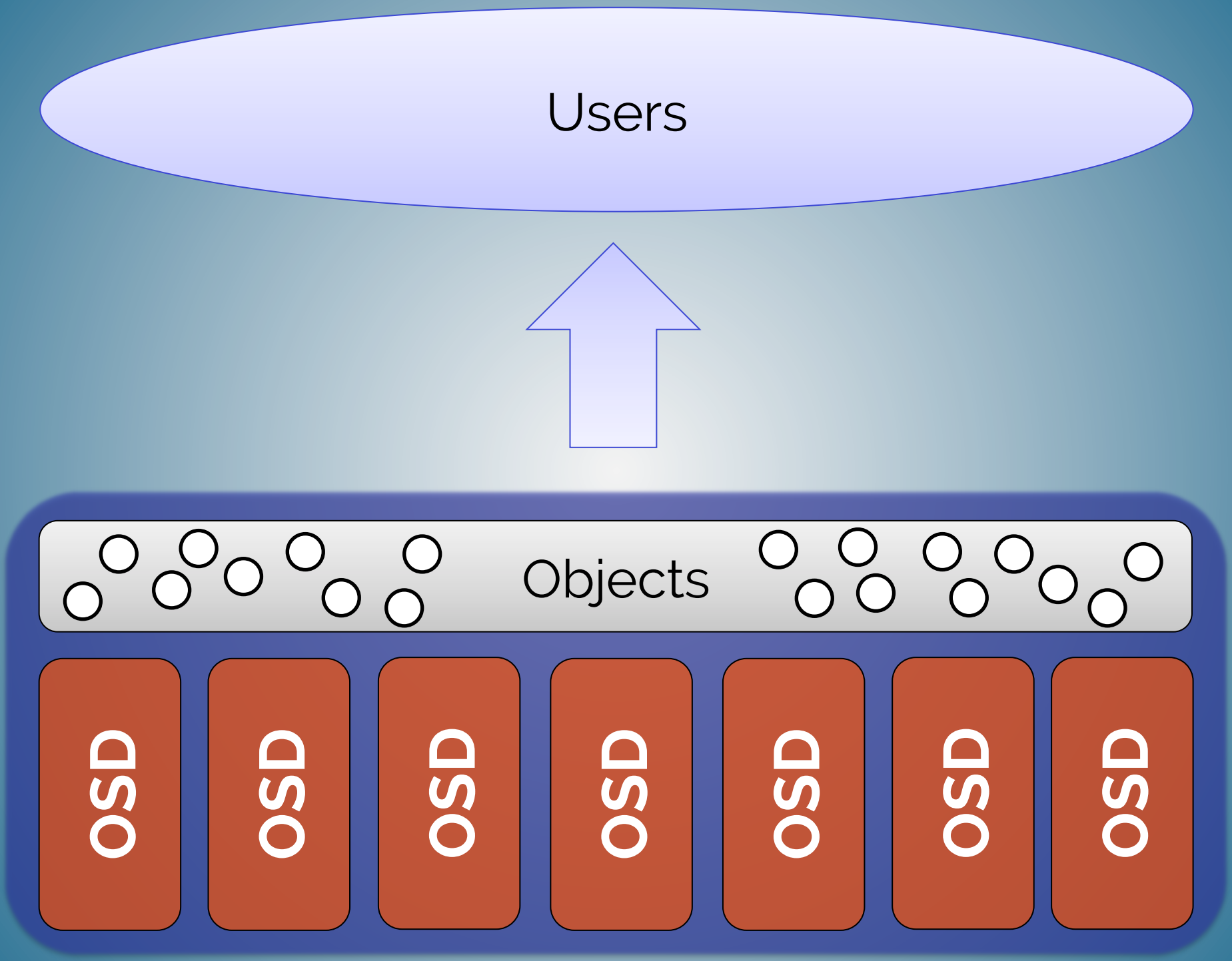
Redundant
Autonomic
Distributed
Object
Store

2 Komponenten

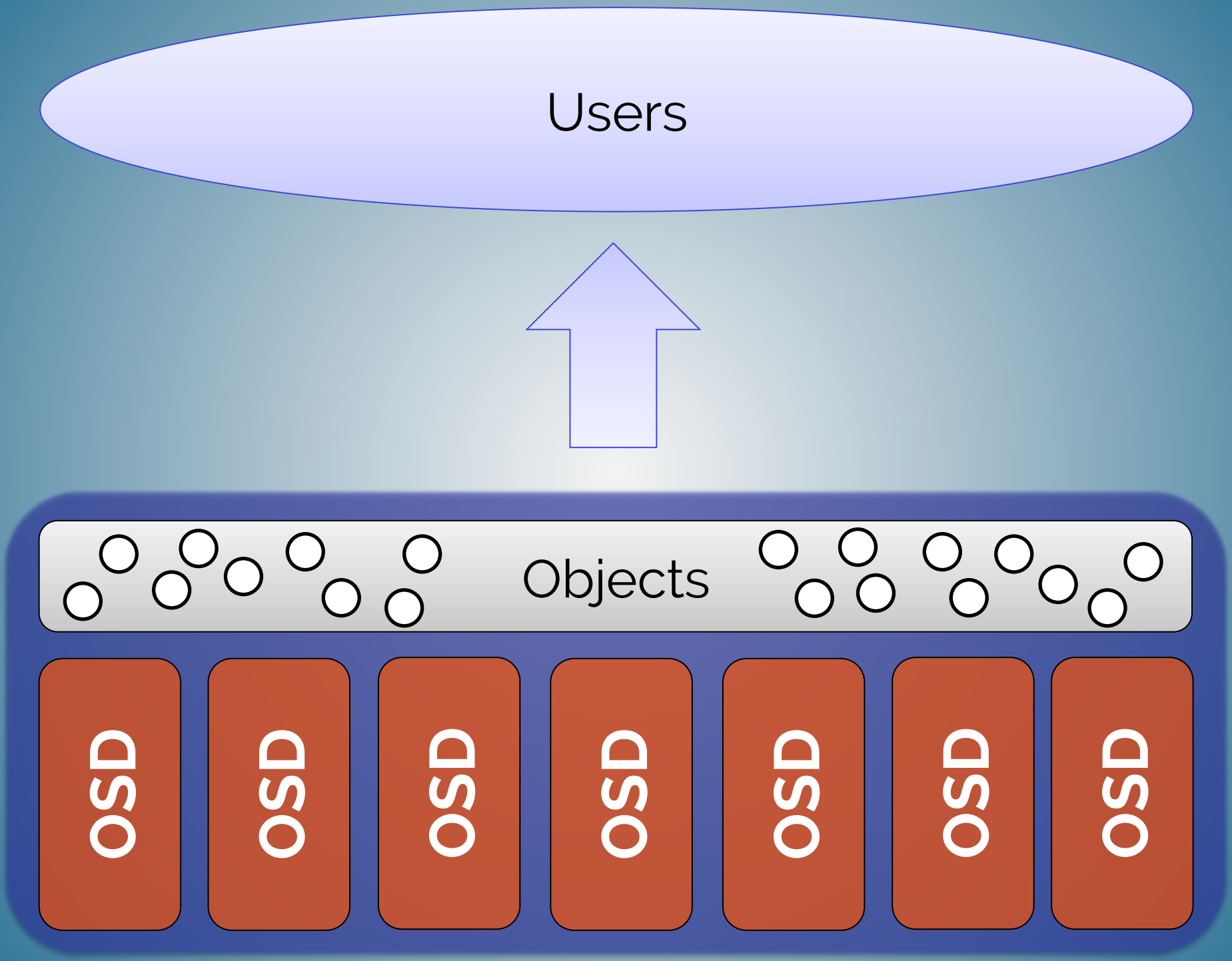
h!

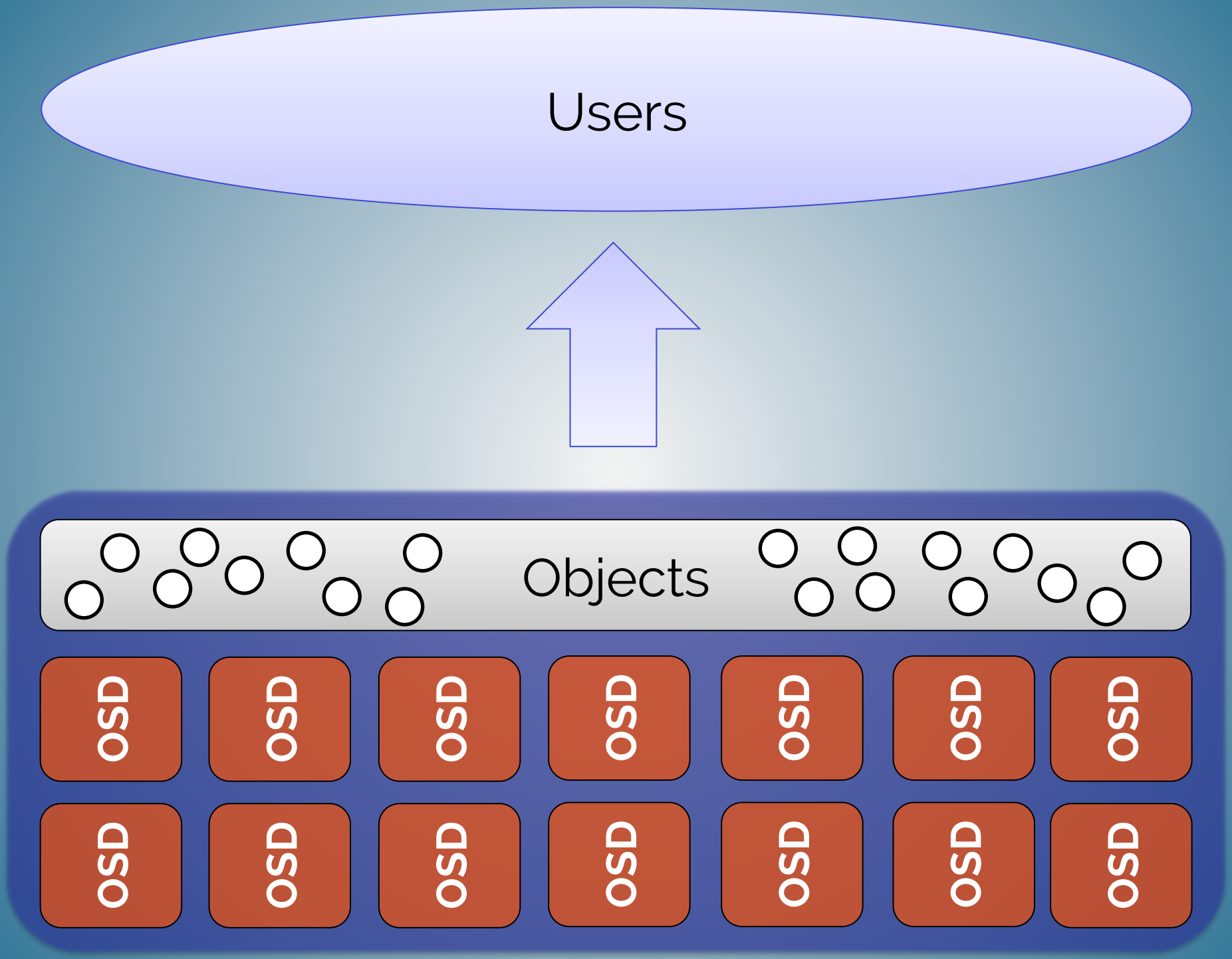
OSDs

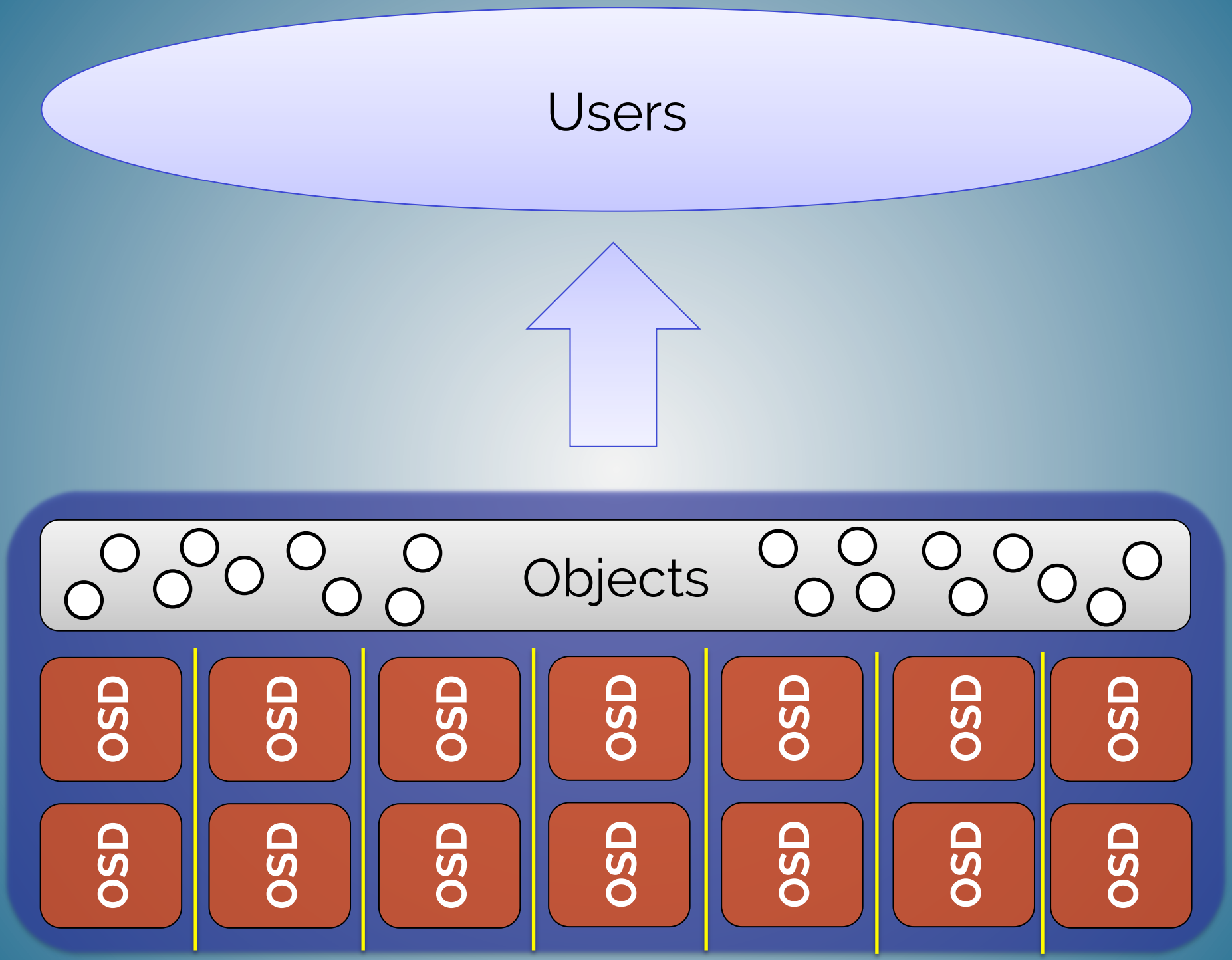

Users
Objects
FS
FS
FS
FS
FS
FS
FS
HDD
HDD
HDD
HDD
HDD
HDD
HDD

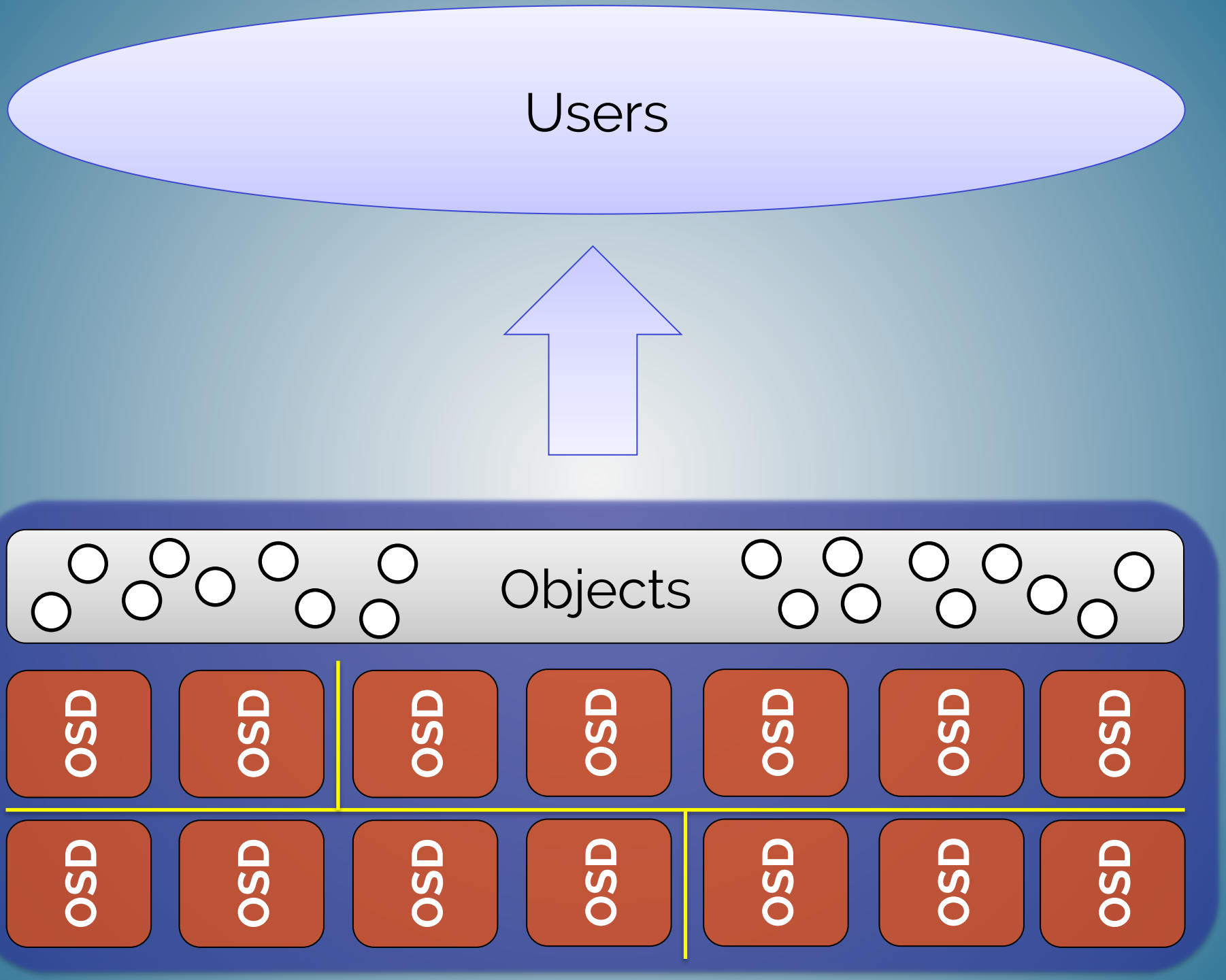

Users
Objects
OSD
OSD
OSD
OSD
OSD
OSD
OSD

Unified Storage

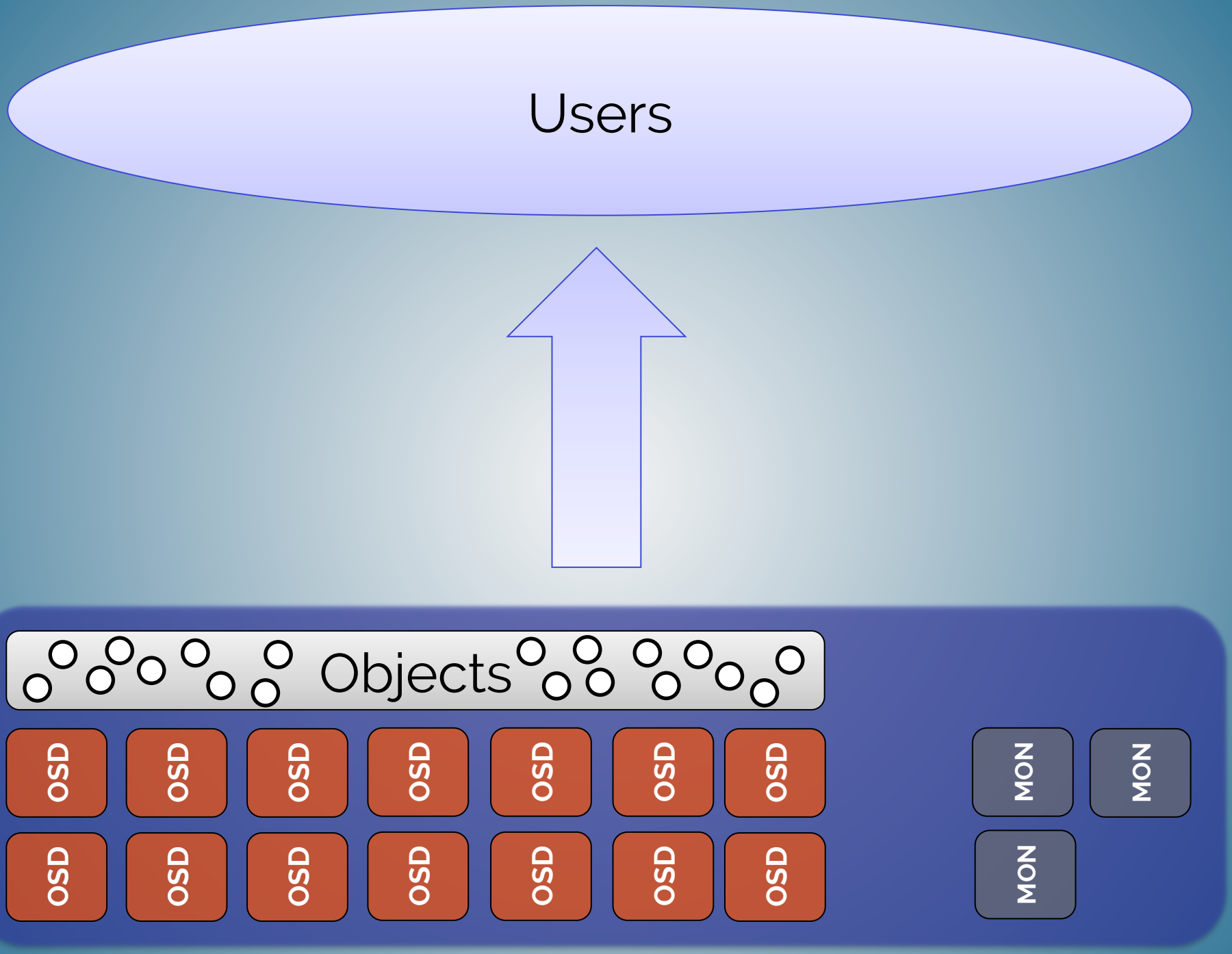

Users
Objects
OSD
OSD
OSD
OSD
OSD
OSD
OSD


Users
Objects
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD

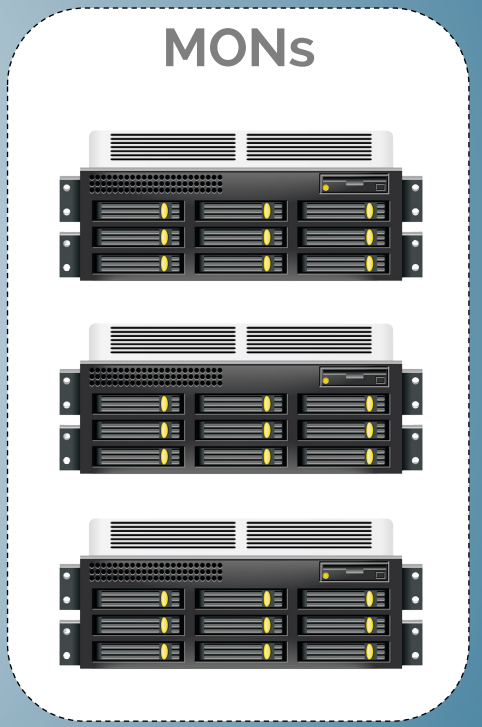

Users
Objects
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD


Users
Objects
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD

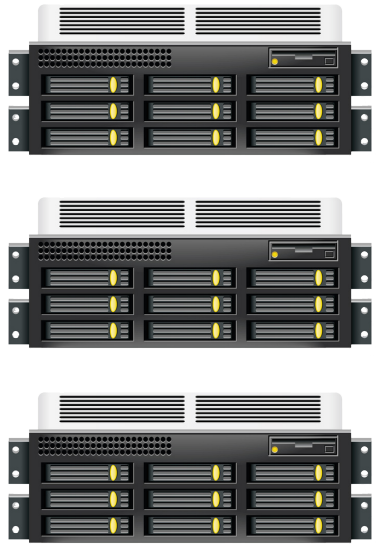
MONs

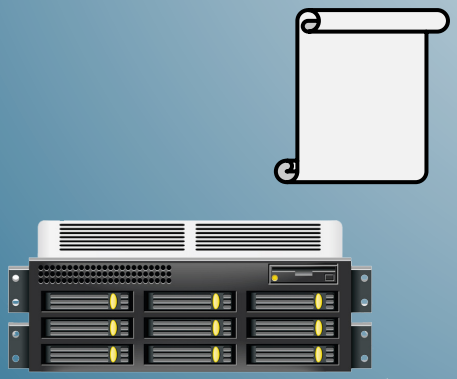

Users
Objects
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD
MON
MON
MON

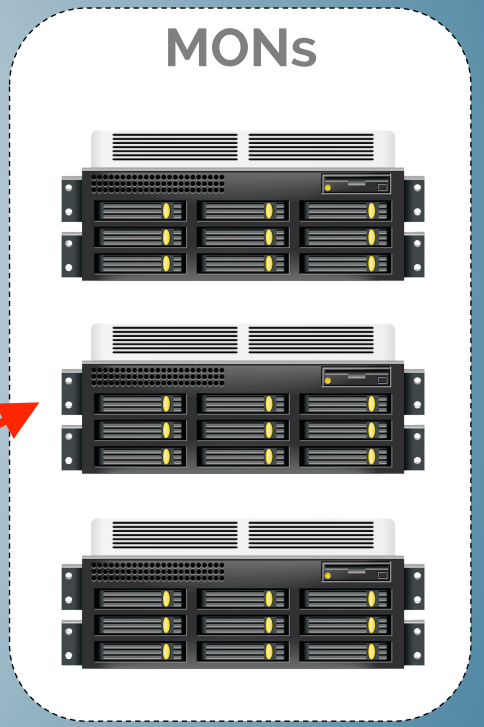
Data Placement

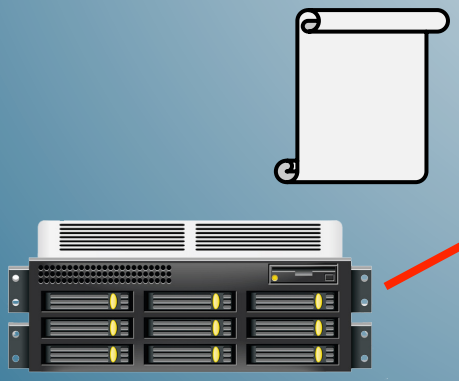

MONs

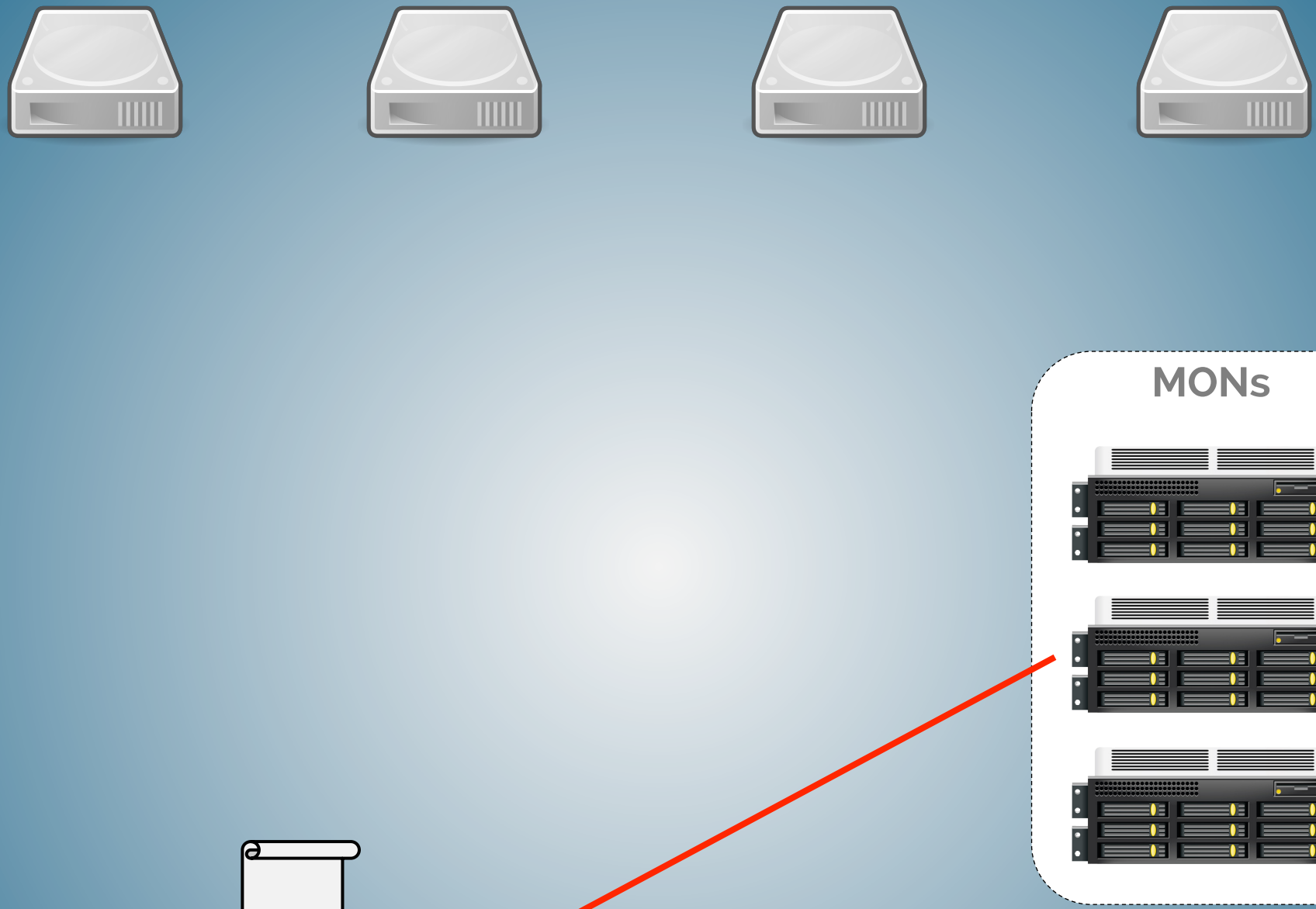
MONs

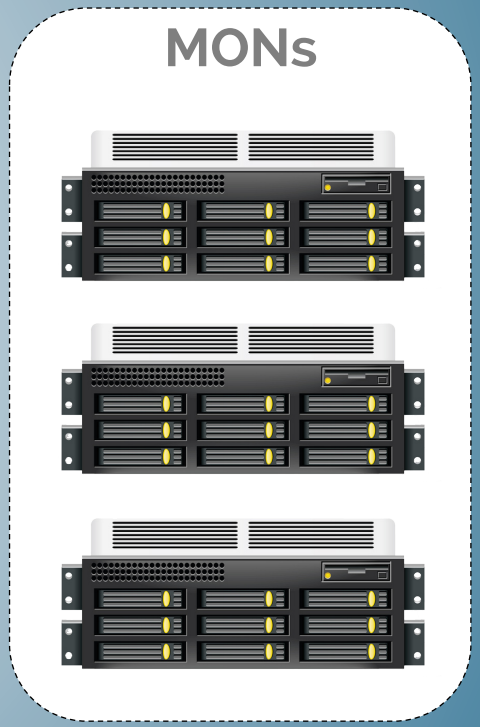


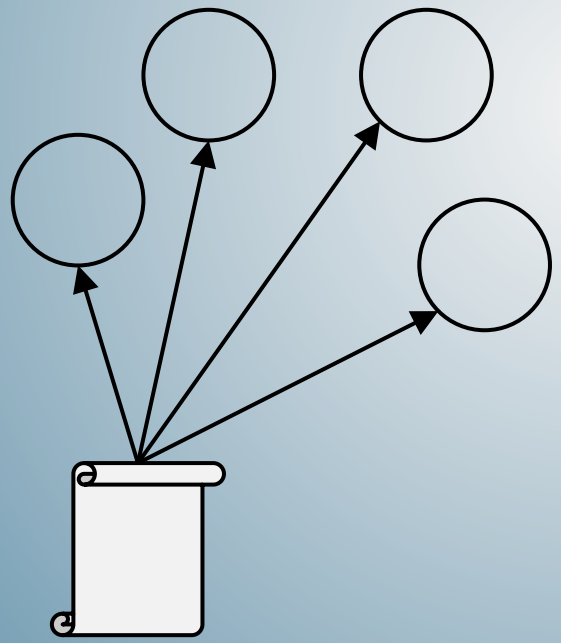


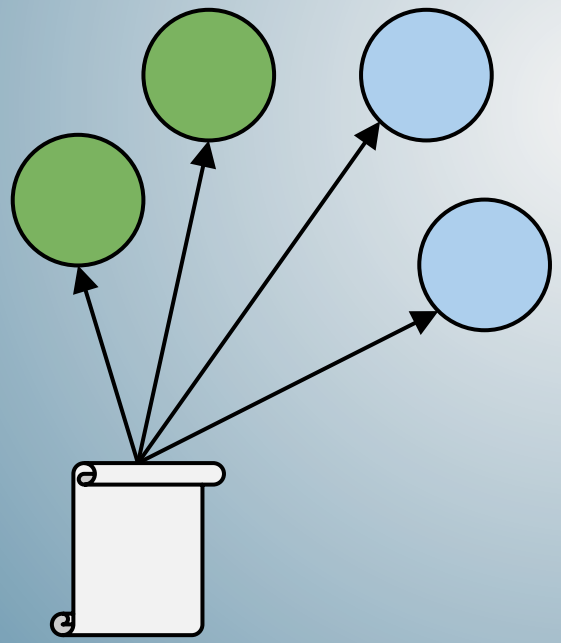

MONs

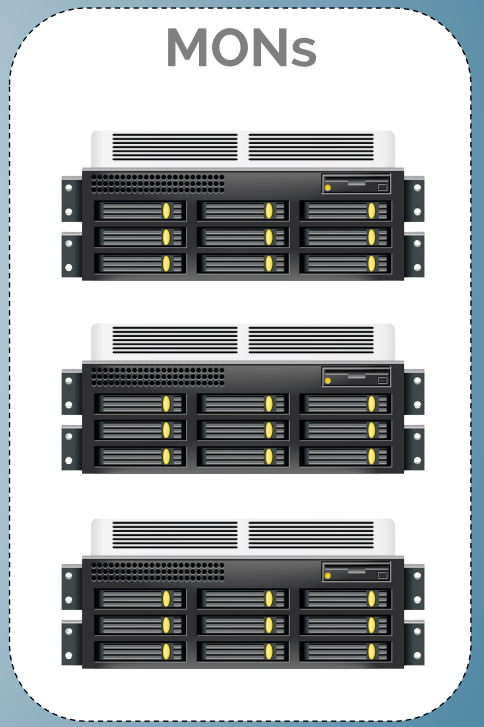


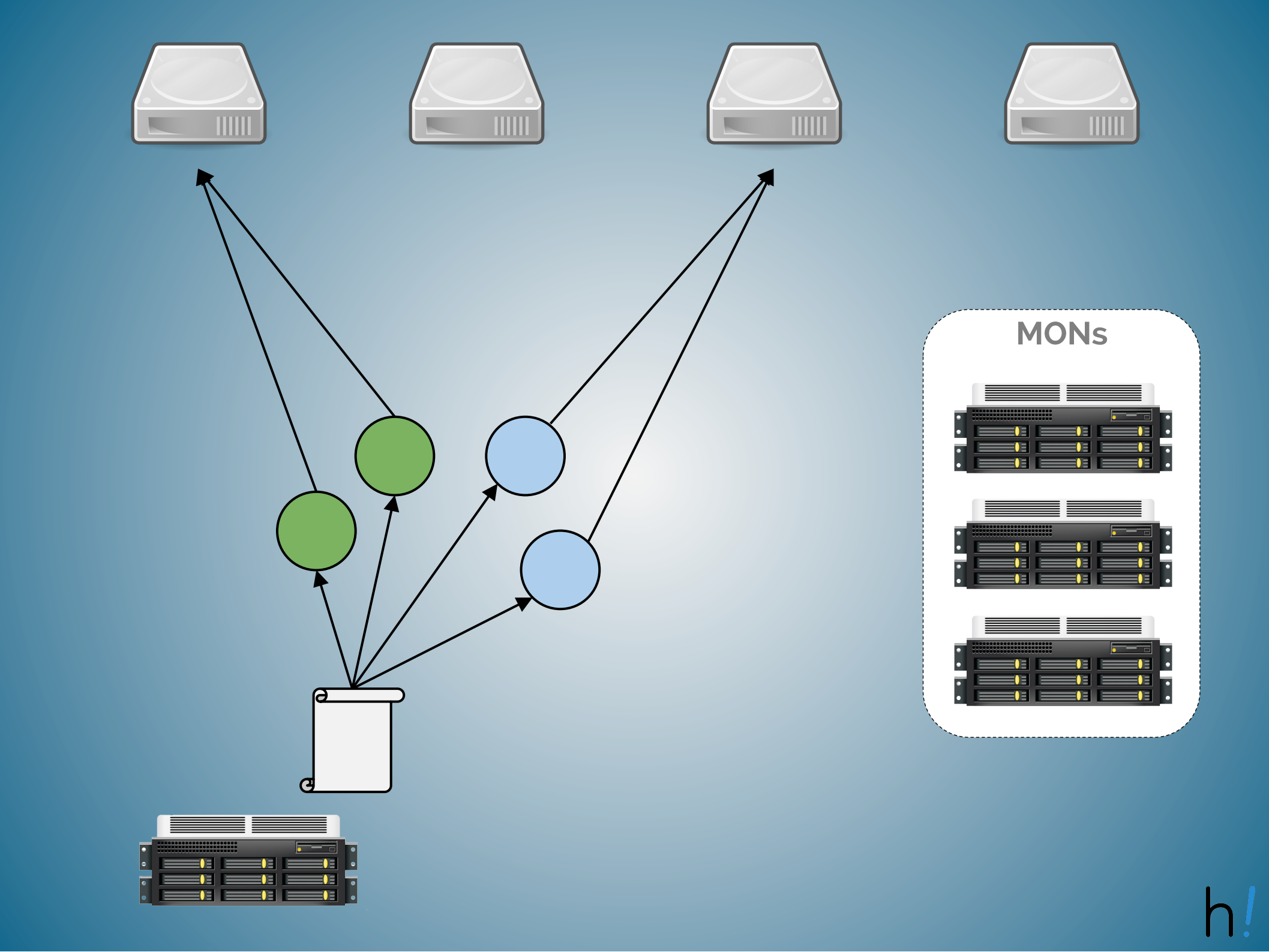

MONs


MONs



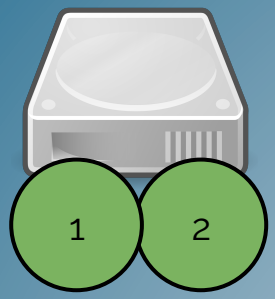


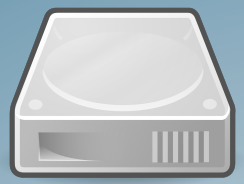

MONs

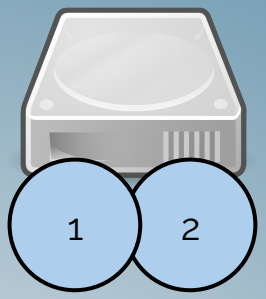

MONs

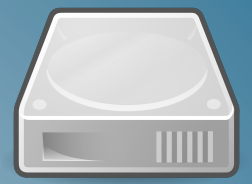
Parallelisierung

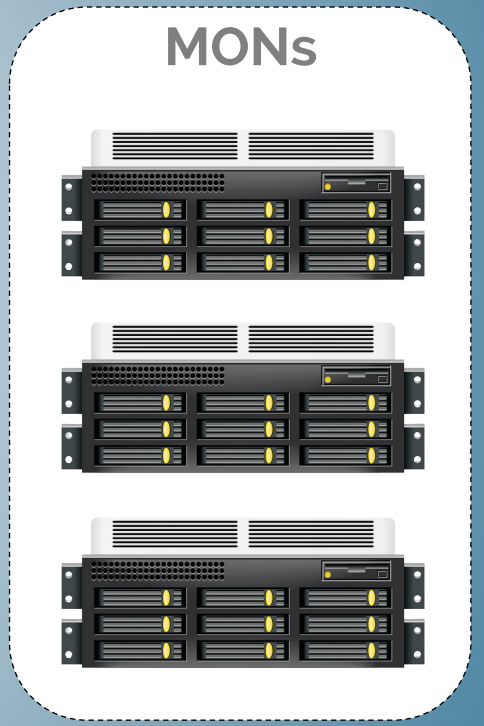
h!

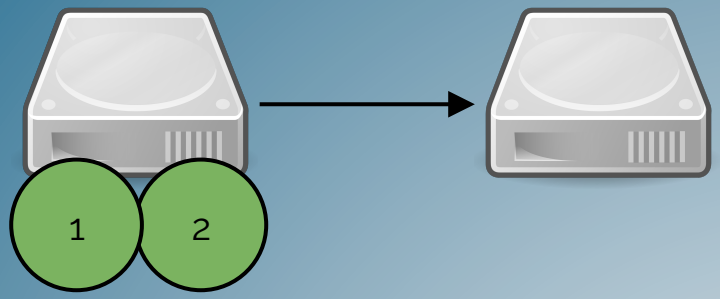

1
2

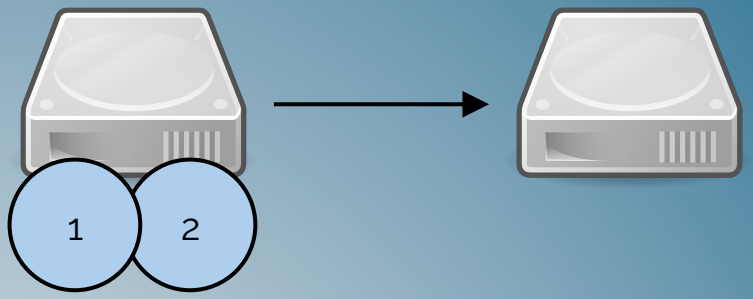


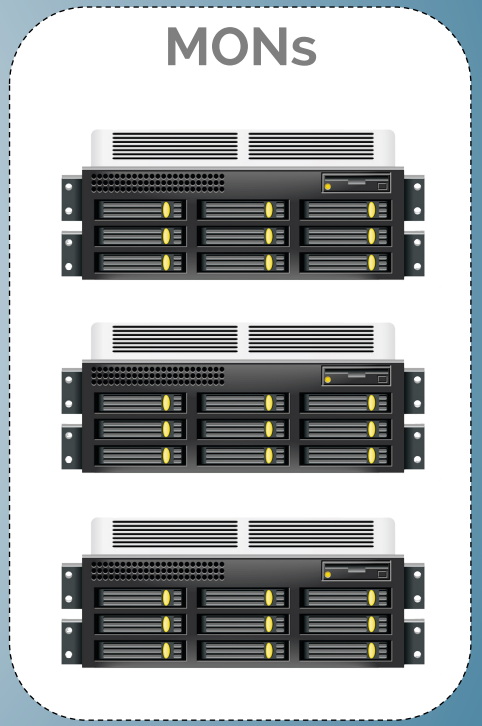

1
2

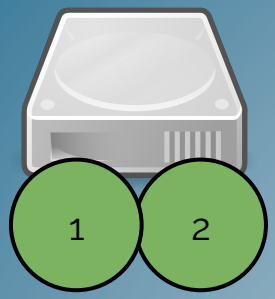


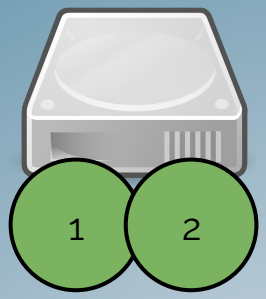

MONs

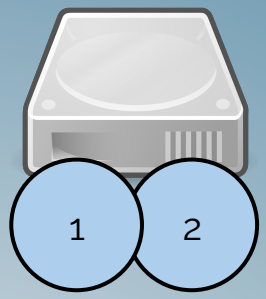

1
2

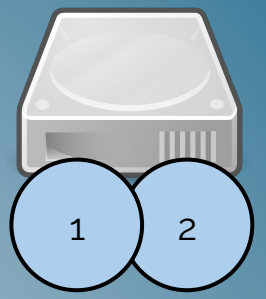

1
2

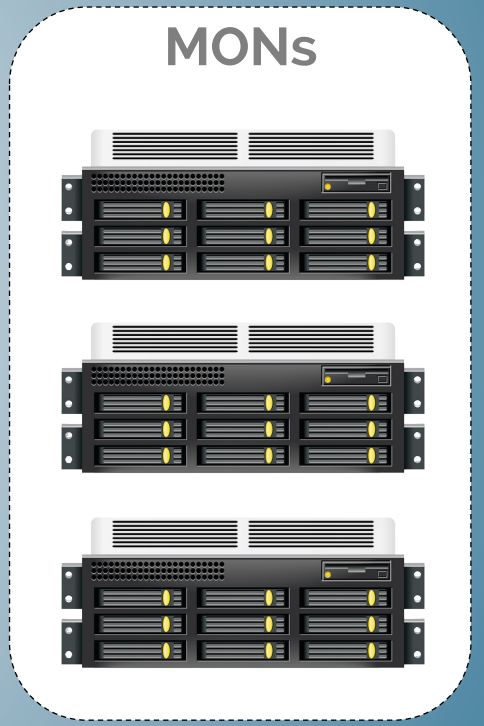

MONs

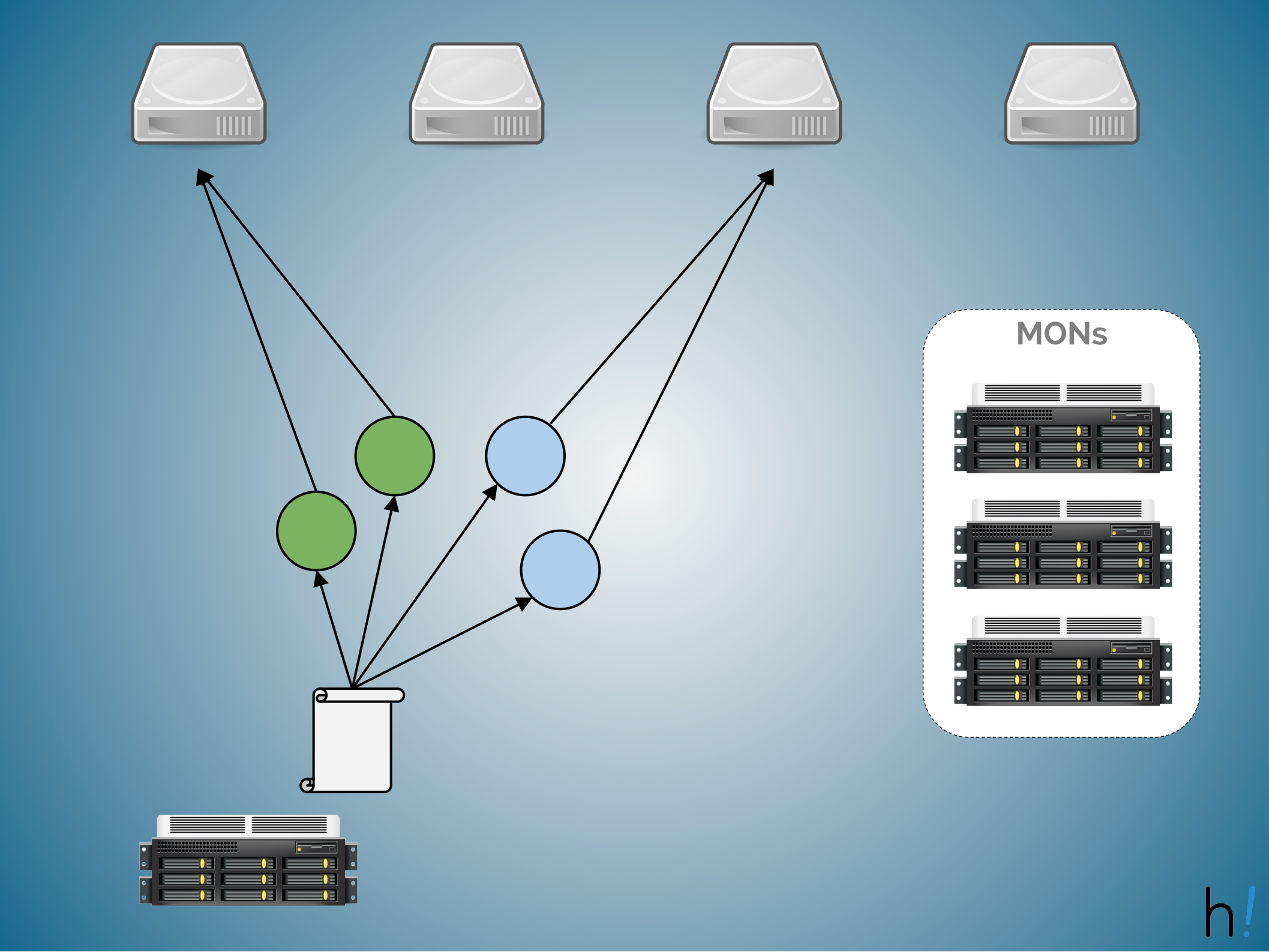

1
2


1
2


1
2


1
2


MONs

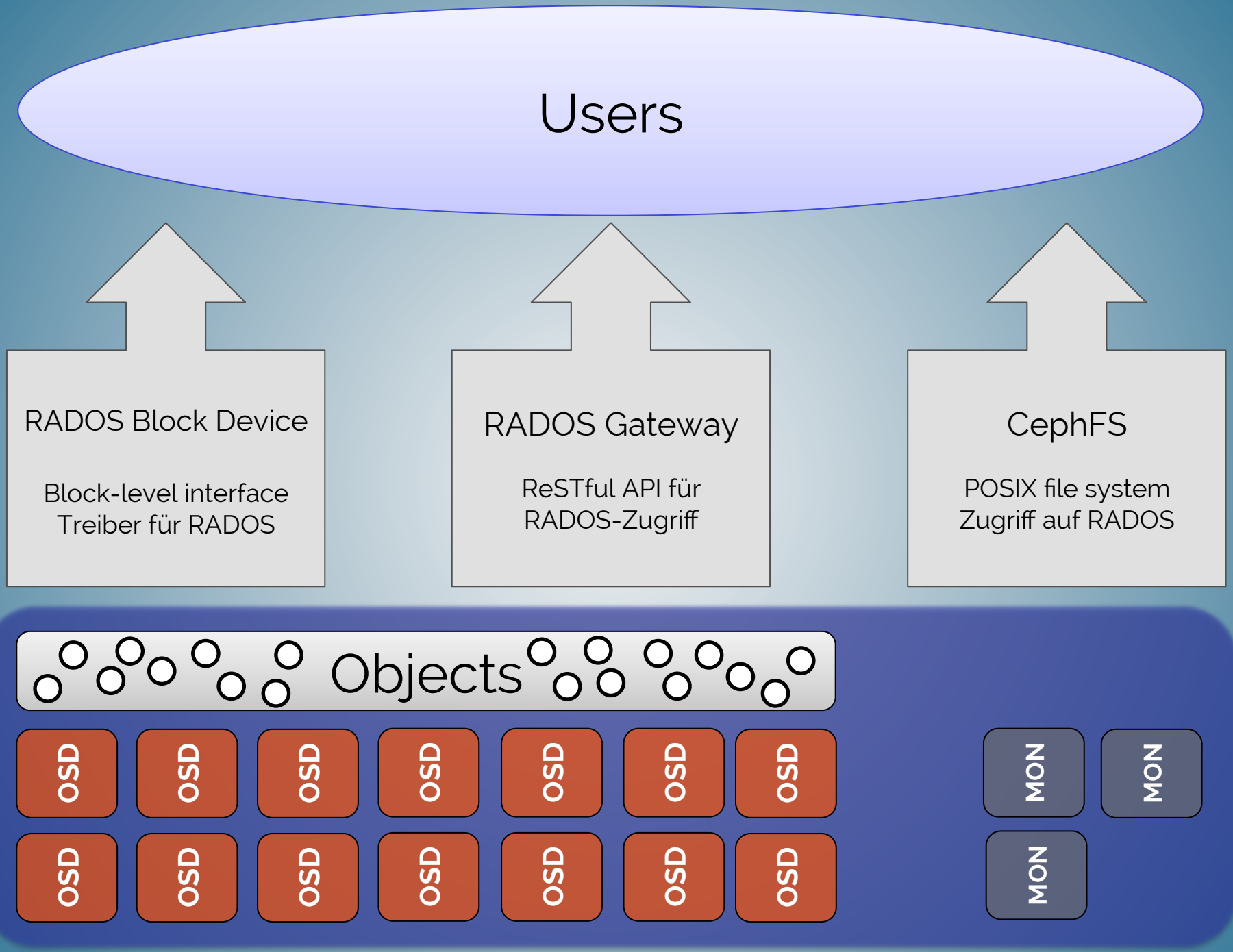

MONs

CRUSH

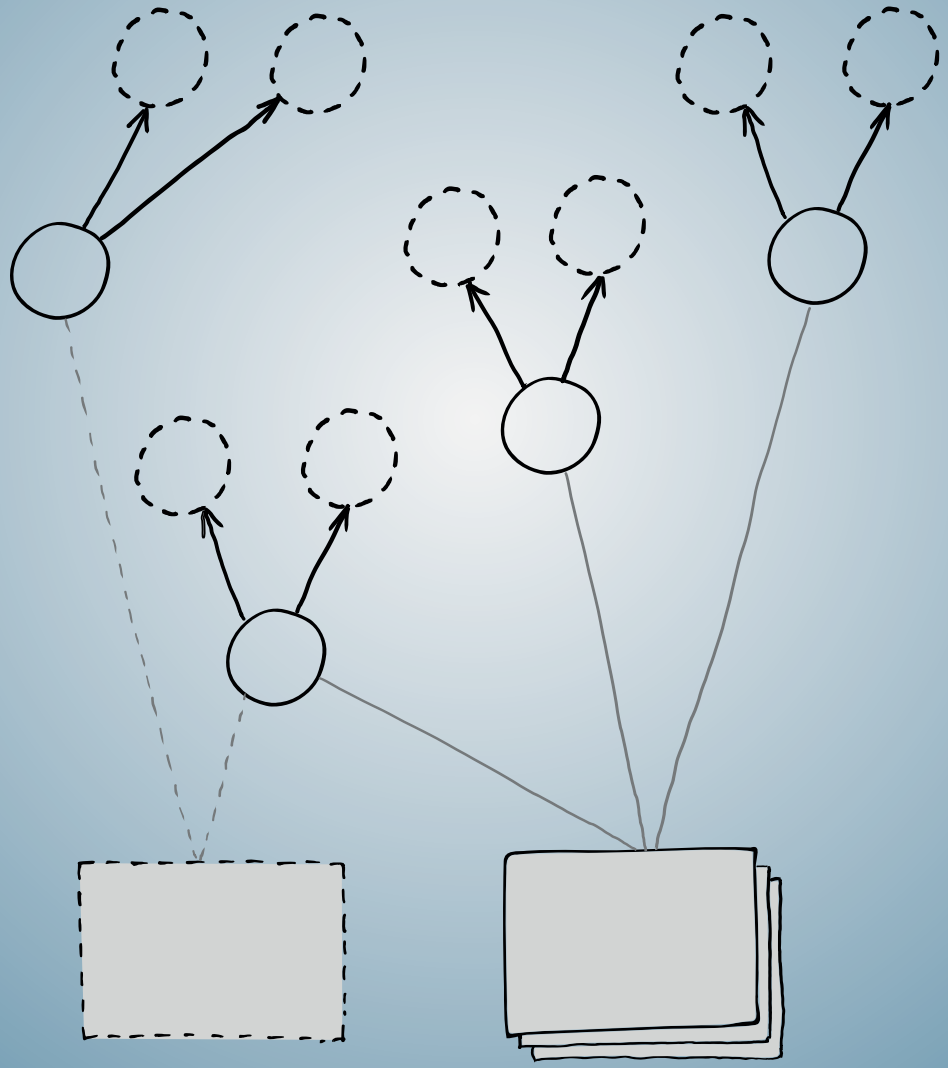
Controlled
Replication
Under
Scalable
Hashing

Rack aware

Clients?


Users
RADOS Block Device
Block-level interface
Treiber für RADOS
RADOS Gateway
ReSTful API für
RADOS-Zugriff
CephFS
POSIX file system
Zugriff auf RADOS
Objects
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD
MON
MON
MON

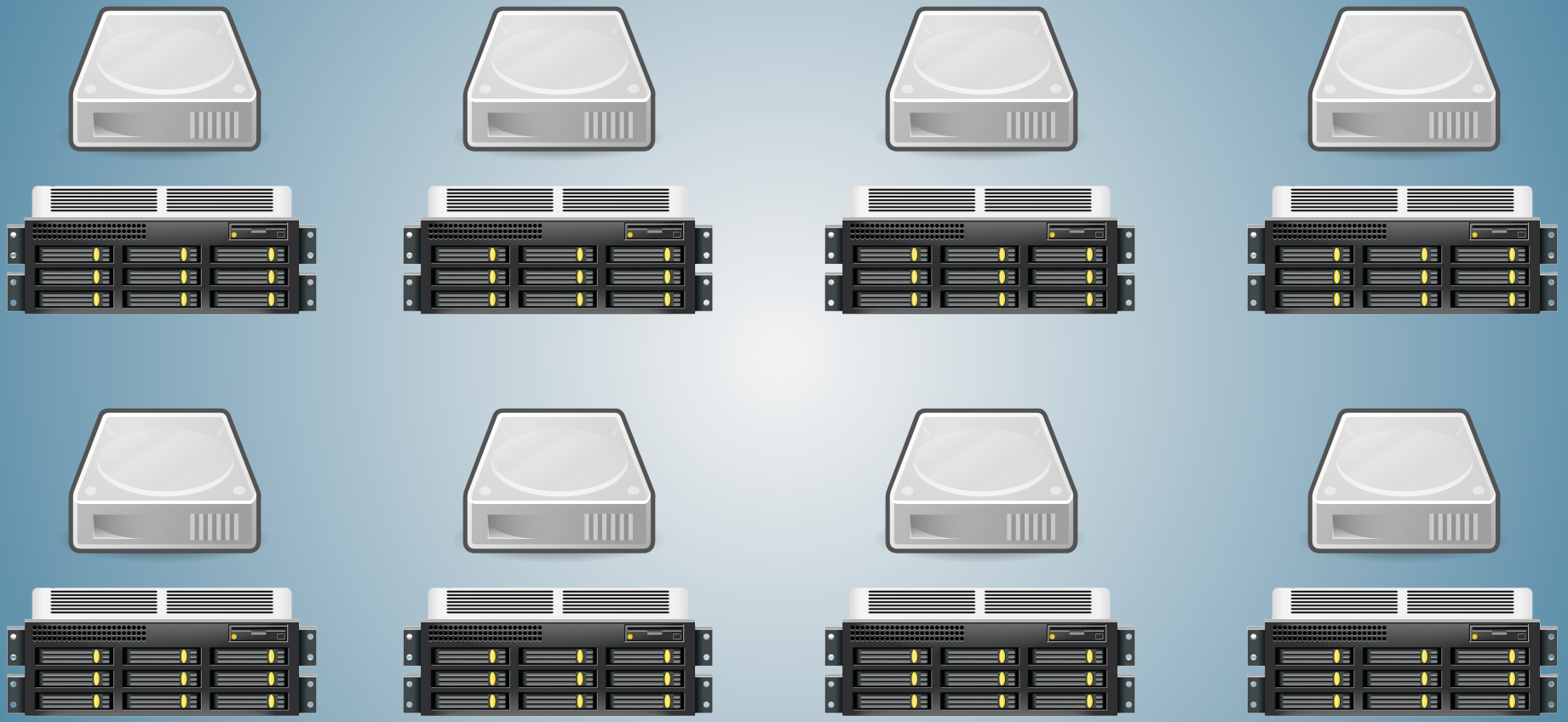
RBD (RADOS Block Device)



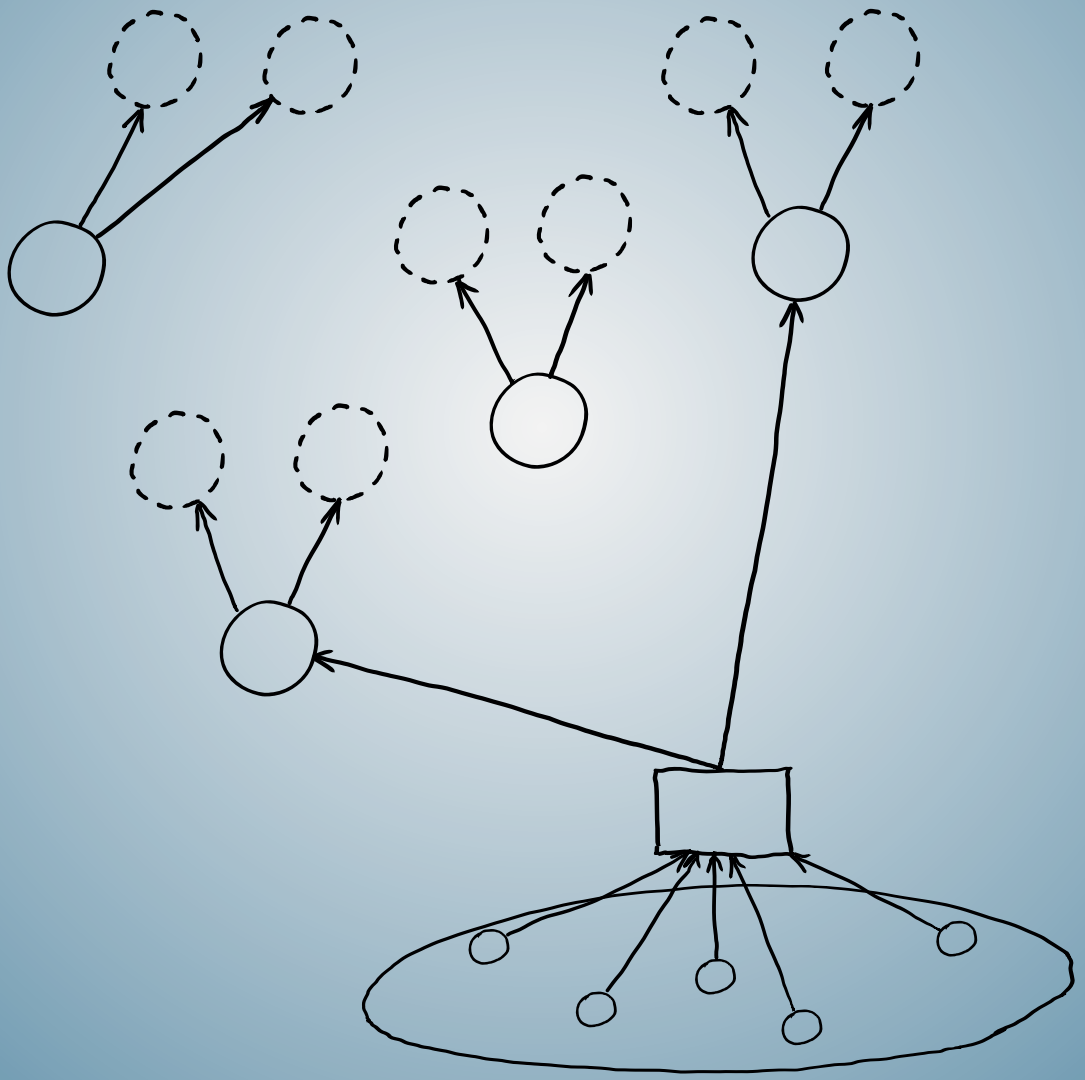
rbd

Qemu-RBD

```
<disk type='network' device='disk'>
  <driver name='qemu' type='raw' cache='writeback'/>
  <auth username='libvirt'>
    <secret type='ceph' usage='client.libvirt secret'/>
  </auth>
  <source protocol='rbd' name='libvirt/ubuntu-amd64-alice'>
    <host name='192.168.133.111' port='6789'/>
    <host name='192.168.133.112' port='6789'/>
    <host name='192.168.133.113' port='6789'/>
  </source>
  <target dev='vda' bus='virtio'/>
  <address type='pci' domain='0x0000' bus='0x00' slot='0x05' function='0x0'/>
</disk>
```



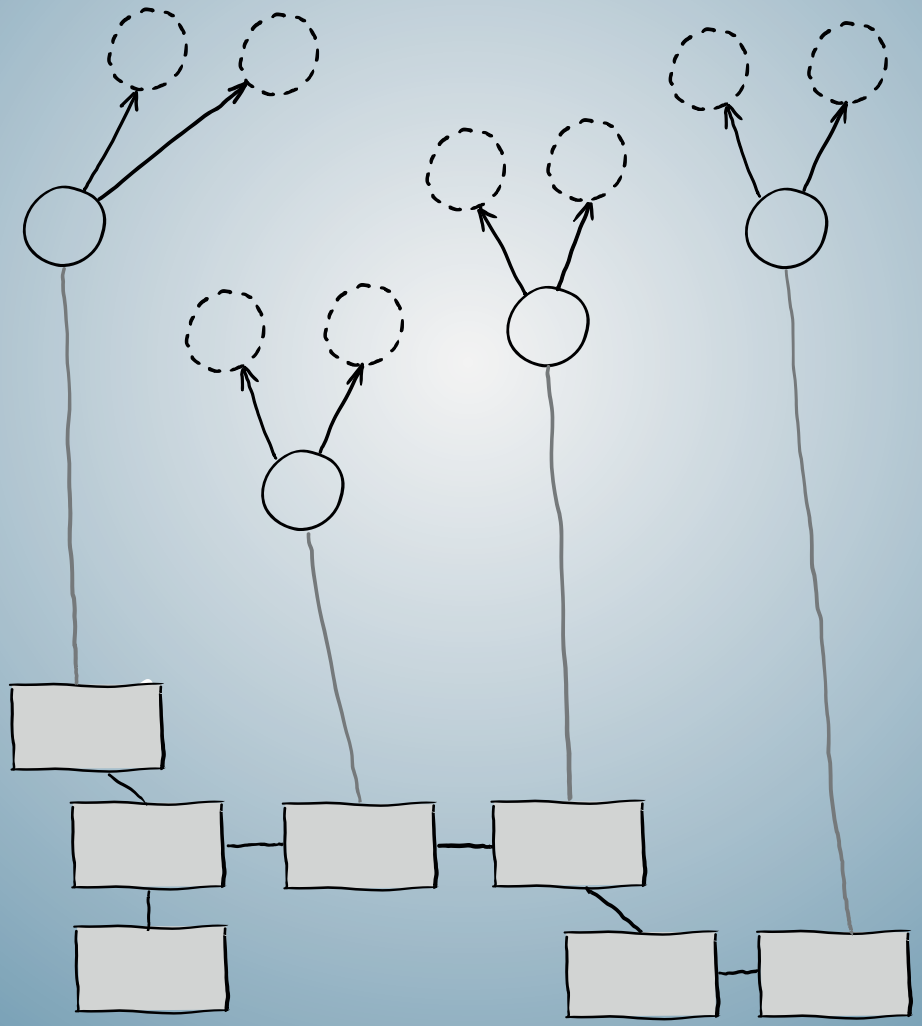
radosgw



Kompatibel mit S3 und Swift

h!

CephFS



CephFS: Leider noch beta

Nicht genug? Der **eigene Client** mit **librados**!

Einsatzszenarien

Gigantic Storage

h!

40TB, 3 Replikas= € 24.000

Virtualisierung

iSCSI Storage

OpenStack!

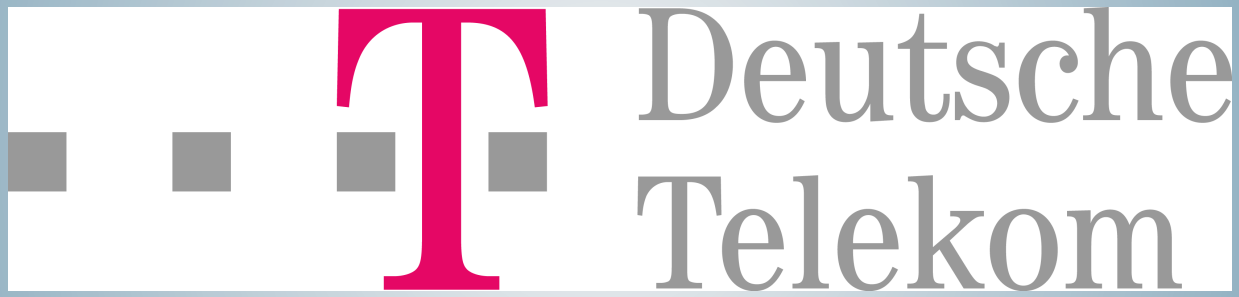
Nahtlose Integration

Cinder & RADOS: Geht!

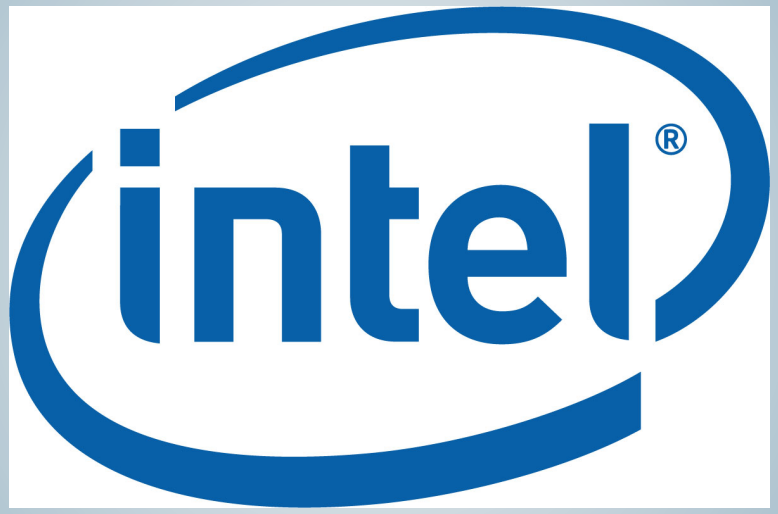
h!

Glance & RADOS: Geht!

Referenzen?


Deutsche
Telekom


Bloomberg



h!

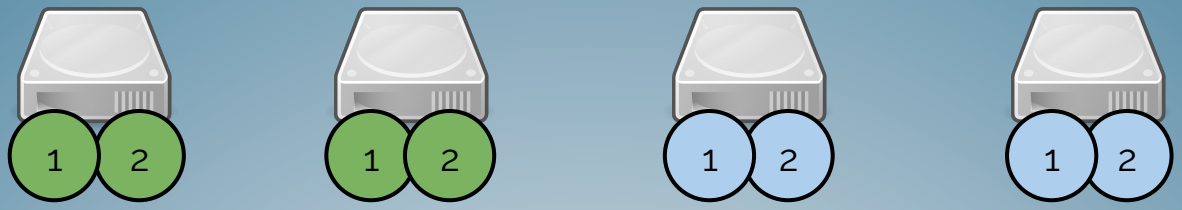
Live demo

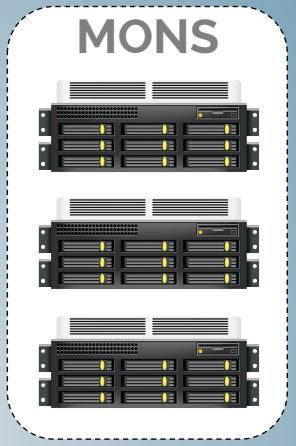
h!

Mit besonderem Dank an:

Sage Weil (Twitter: @liewegas)
& Crew für Ceph

Inktank (Twitter: @inktank)
für das Ceph-Logo





ceph

goo.gl/S1sYZ (me on Google+)

goo.gl/LqWFB (Slides on Slideshare)

twitter.com/hastexo

hastexo.com

h!