

Storage mit Ceph

Scale-Out ganz einfach



Martin Gerhard Loschwitz
hastexo!

Wer?















20. April 2013 FH Joanneum



hastexo!

Skalierbares Storage

2 Arten von Skalierbarkeit

Scale-Up

Scale-Up vorher:



Scale-Up danach:



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

Scale-Out

Scale-Out vorher:



Scale-Out danach:



Scale-Out ist hip

Webserver

Datenbanken

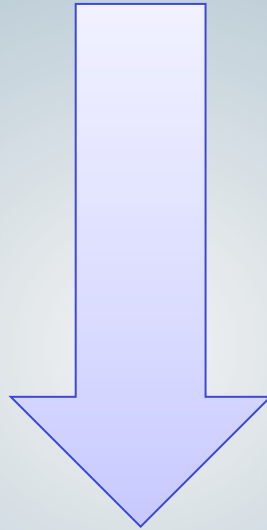
Sogar ich!

Storage? Meh.

Das Block-Problem

```
graph TD;
    A([User Space]) --> B[FS];
    B --> C[HDD];
```

User Space



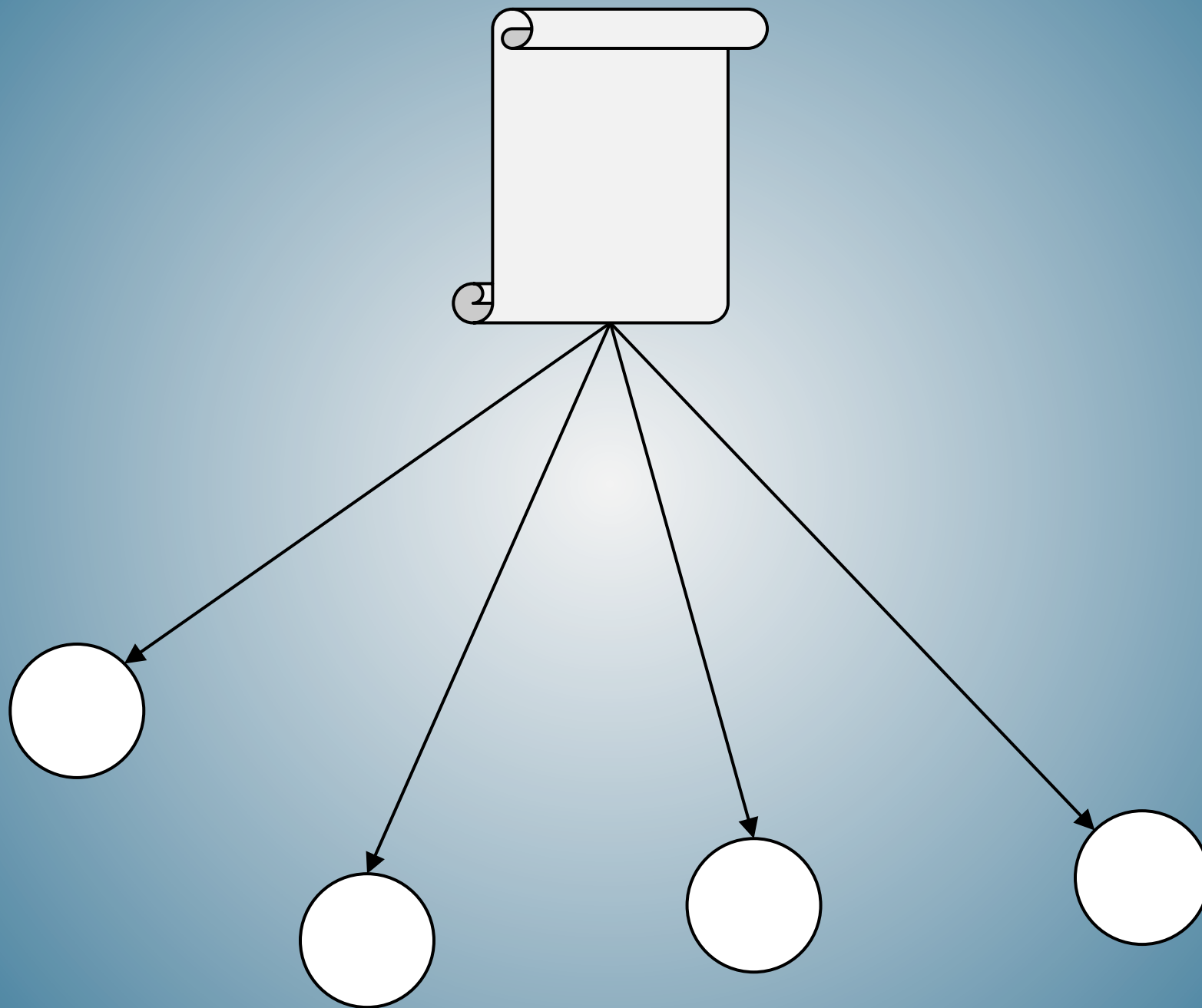
FS

HDD

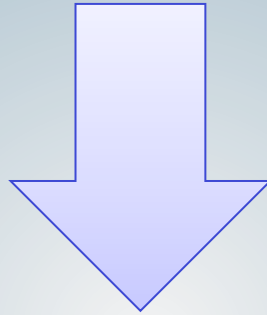
Blöcke für
verteilte Systeme?

Ich glaube nicht, Tim!

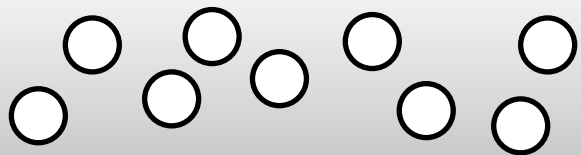
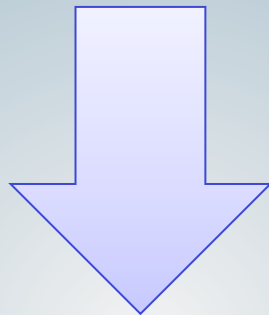
Object Stores



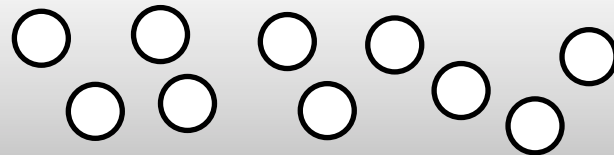
User Space



User Space



Objects



FS

FS

FS

FS

FS

FS

FS

HDD

HDD

HDD

HDD

HDD

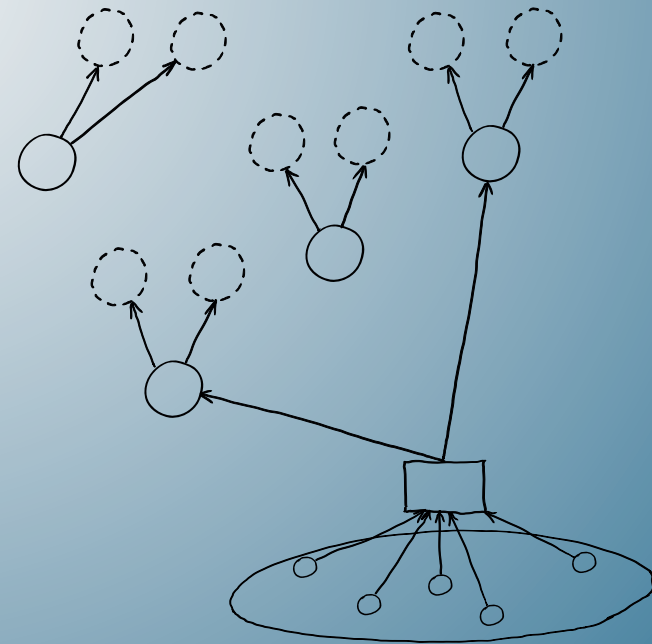
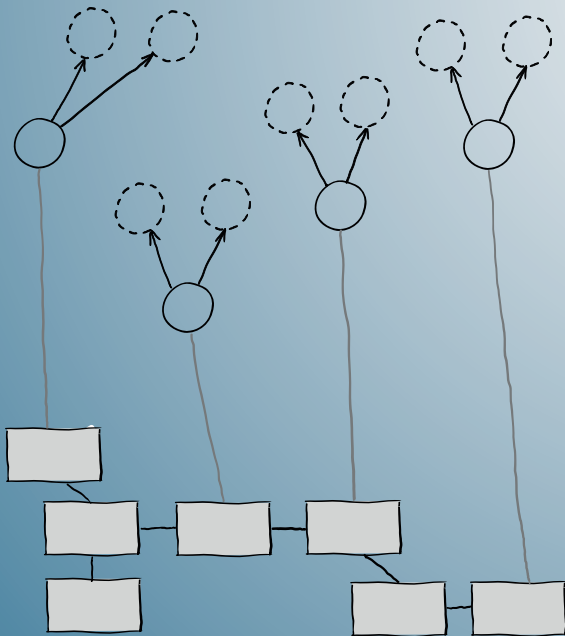
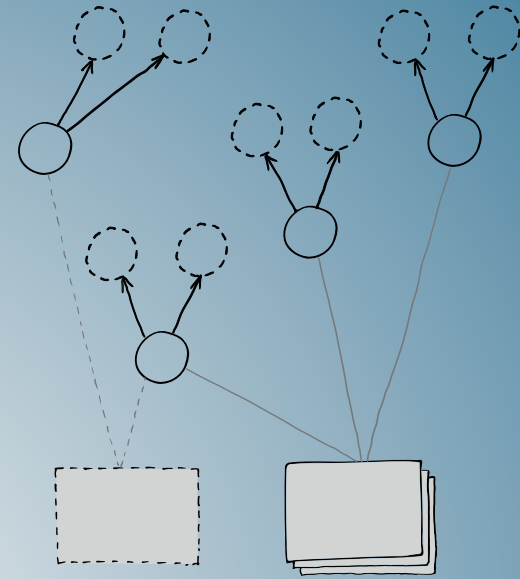
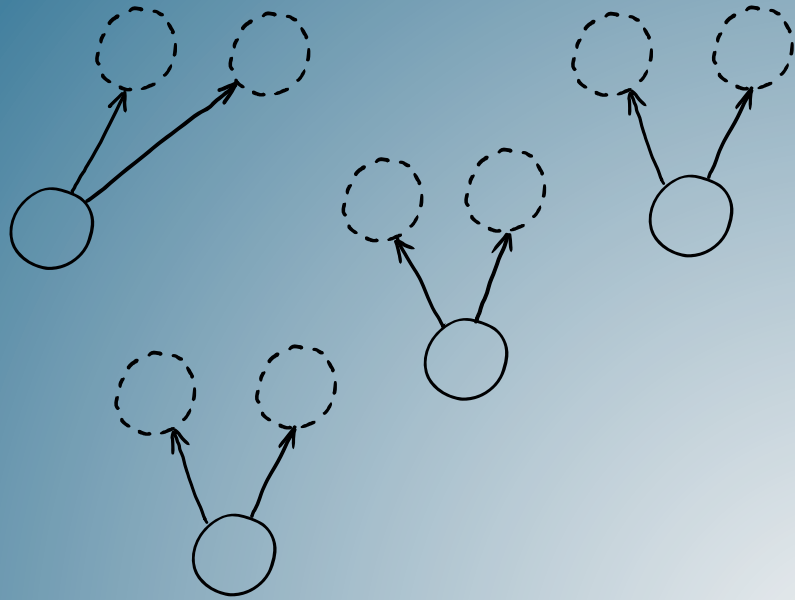
HDD

HDD

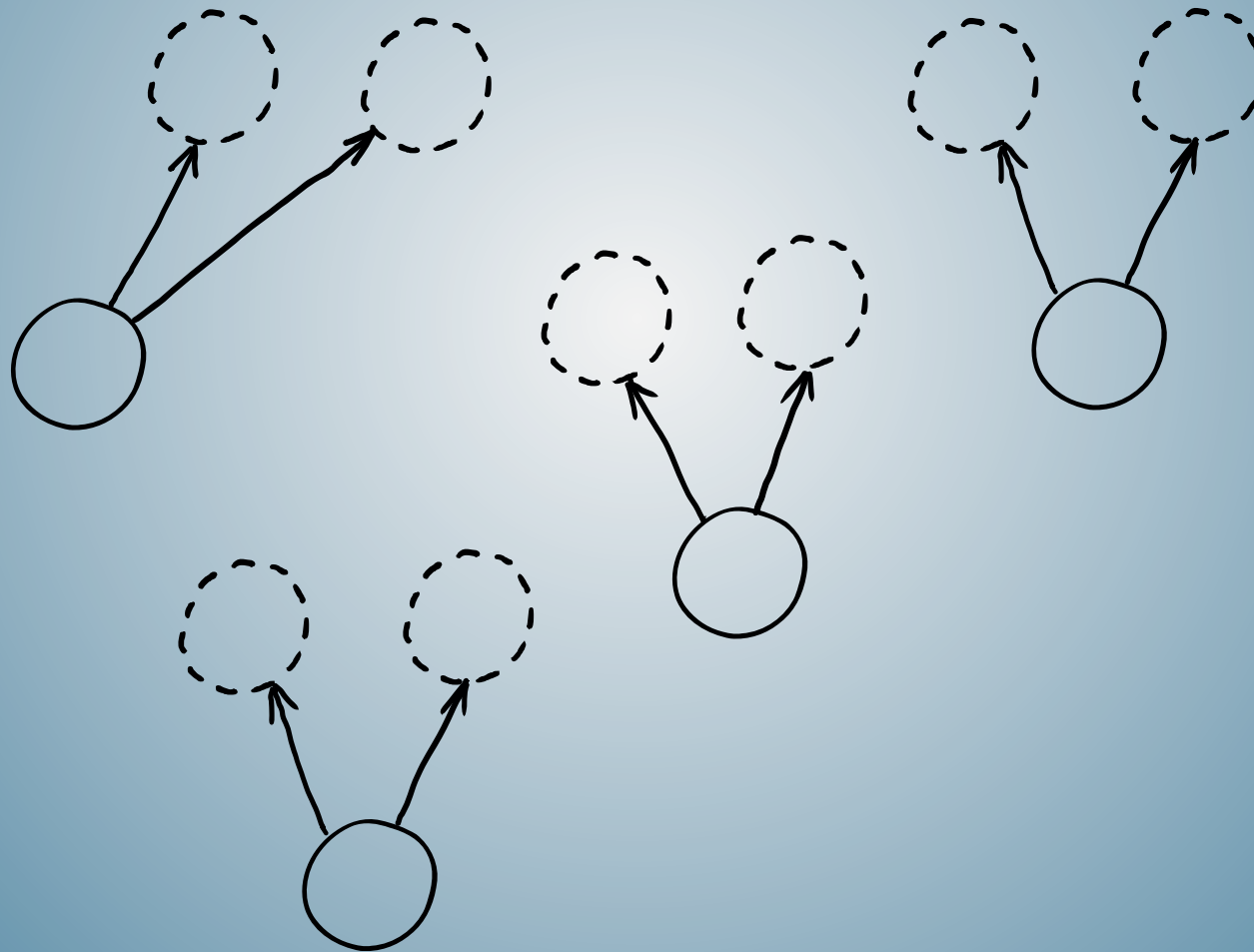


ceph

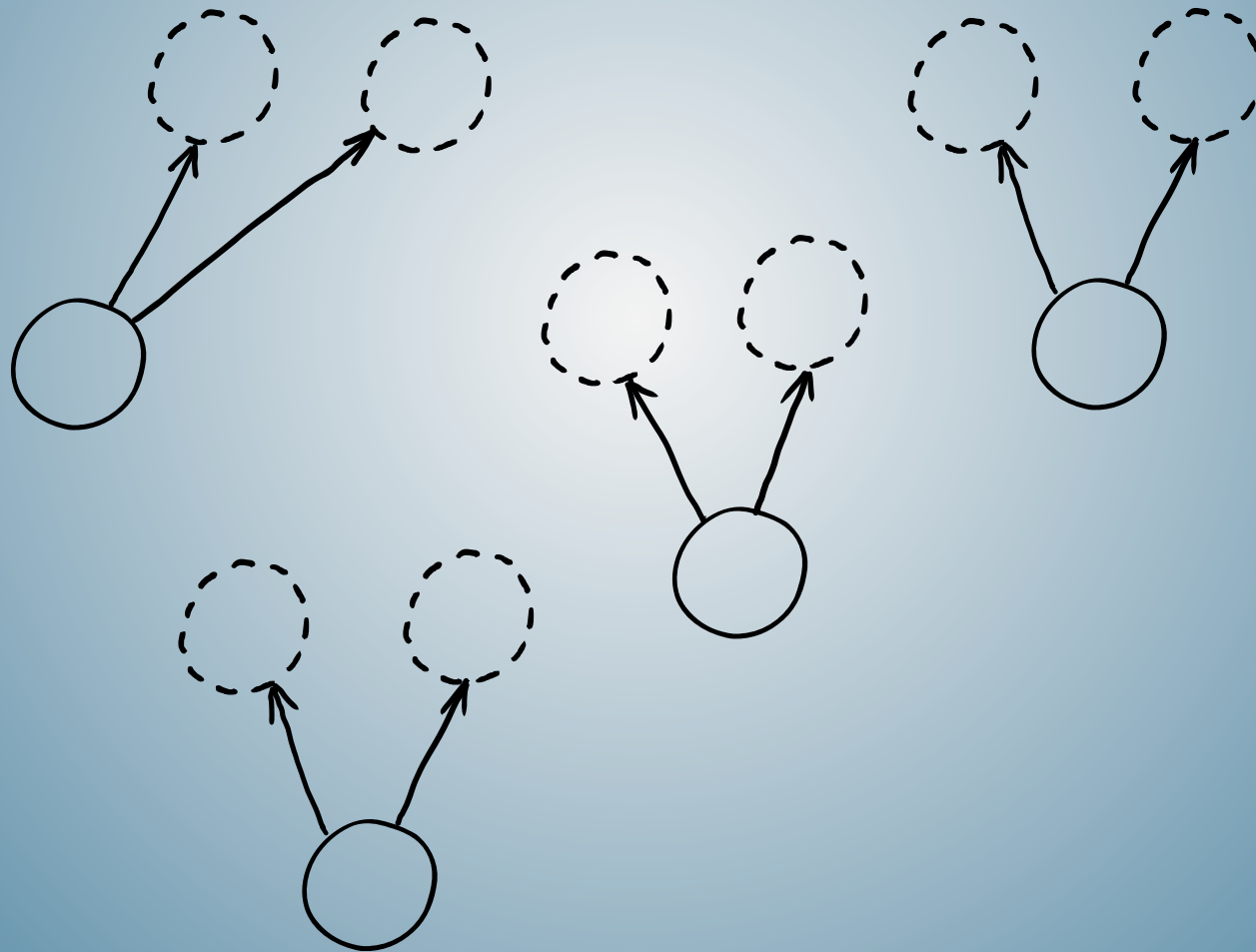
Ursprünglich eine Dissertation



Object Store



RADOS



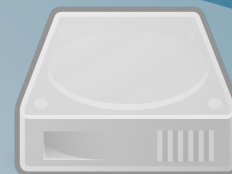
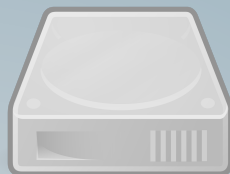
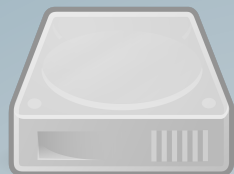
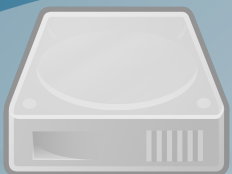
Redundant **A**utonomic **D**istributed **O**bject **S**ore

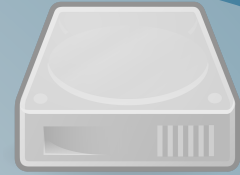
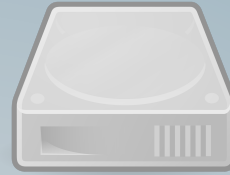
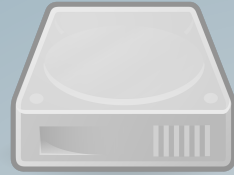
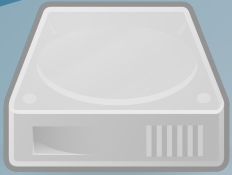
2 Komponenten

OSDs

MONs

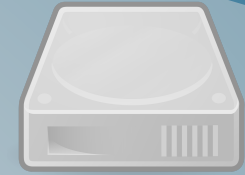
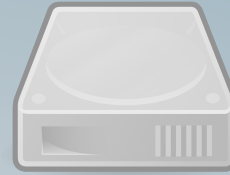
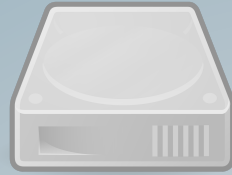
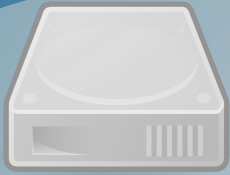
Data Placement



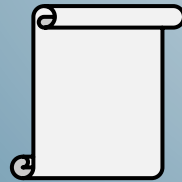


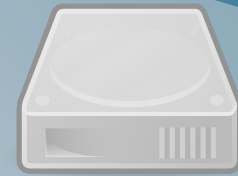
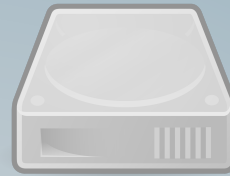
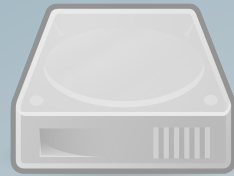
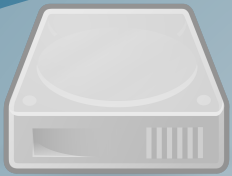
MONs



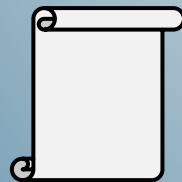


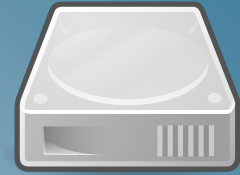
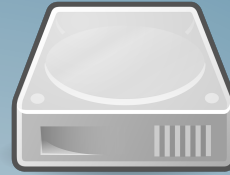
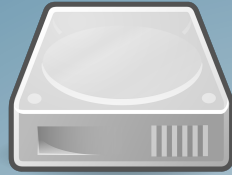
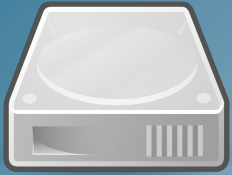
MONs



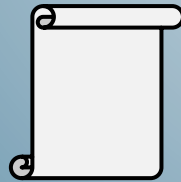


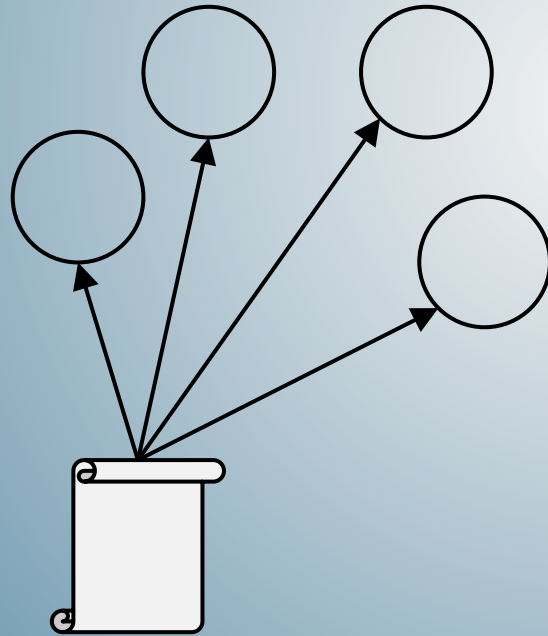
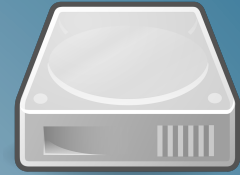
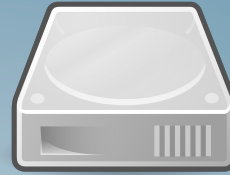
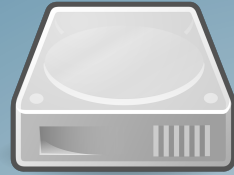
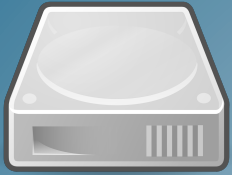
MONs



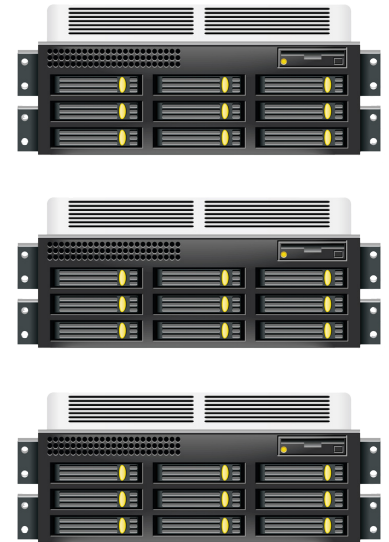


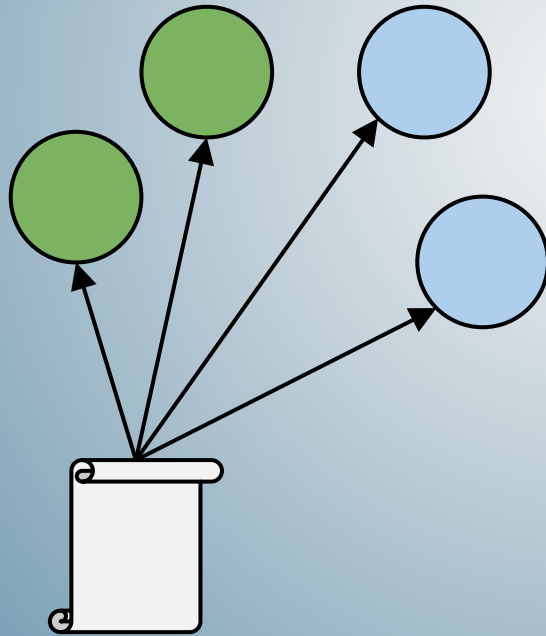
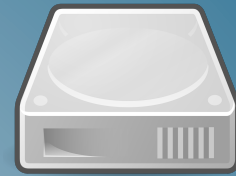
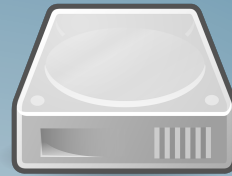
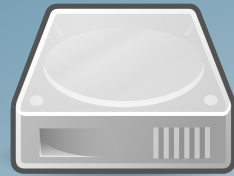
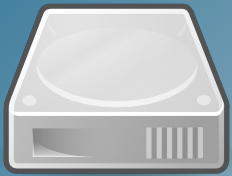
MONs





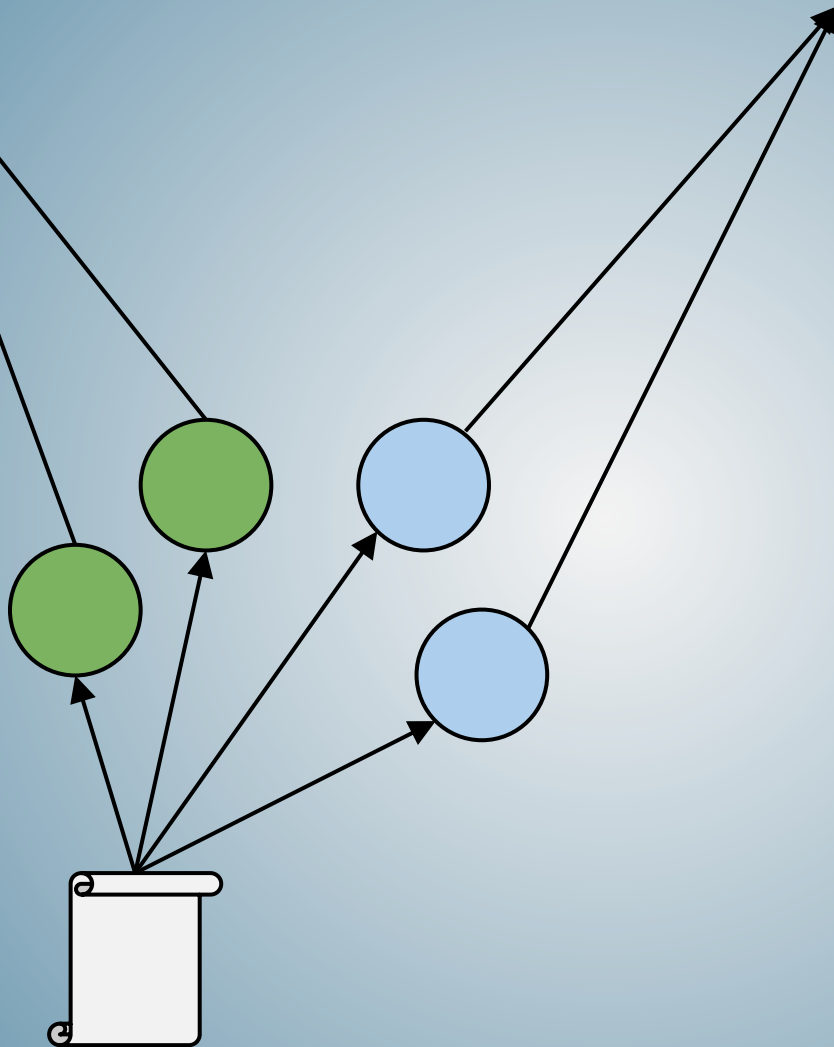
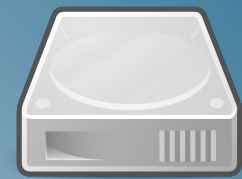
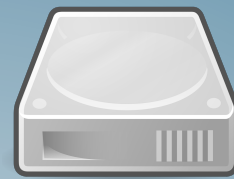
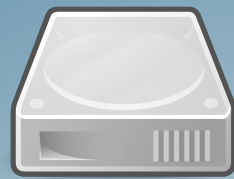
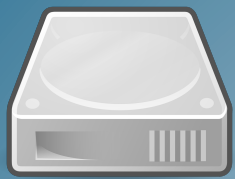
MONs





MONs

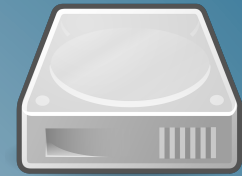
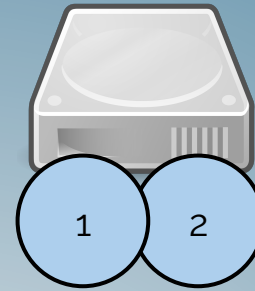
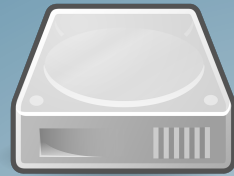
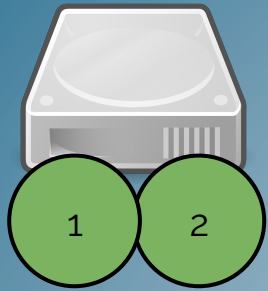




MONs

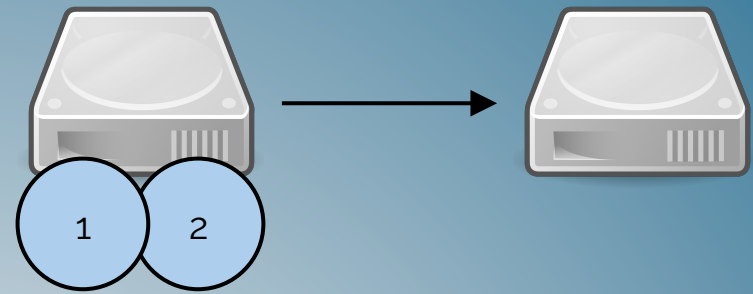
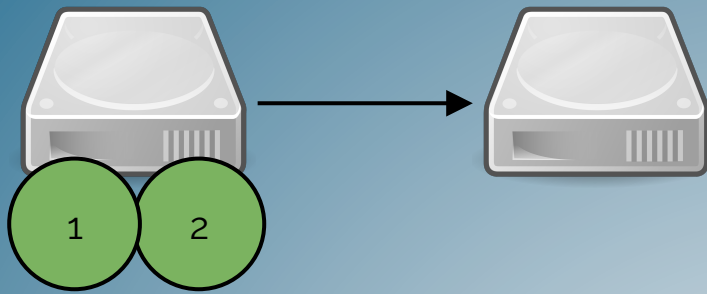


Parallelisierung

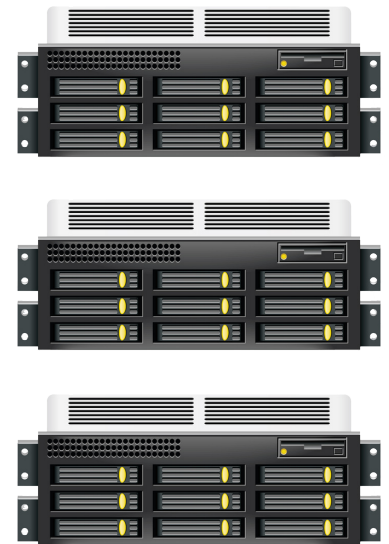


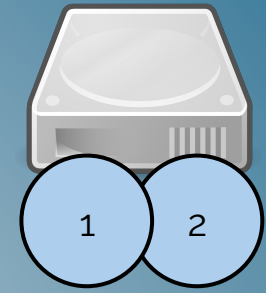
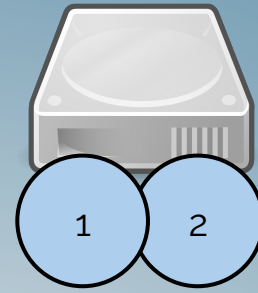
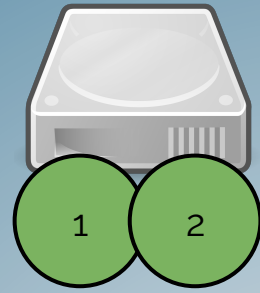
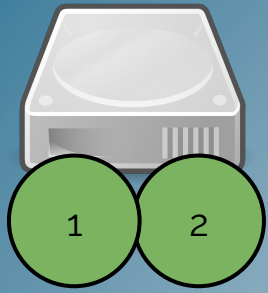
MONs





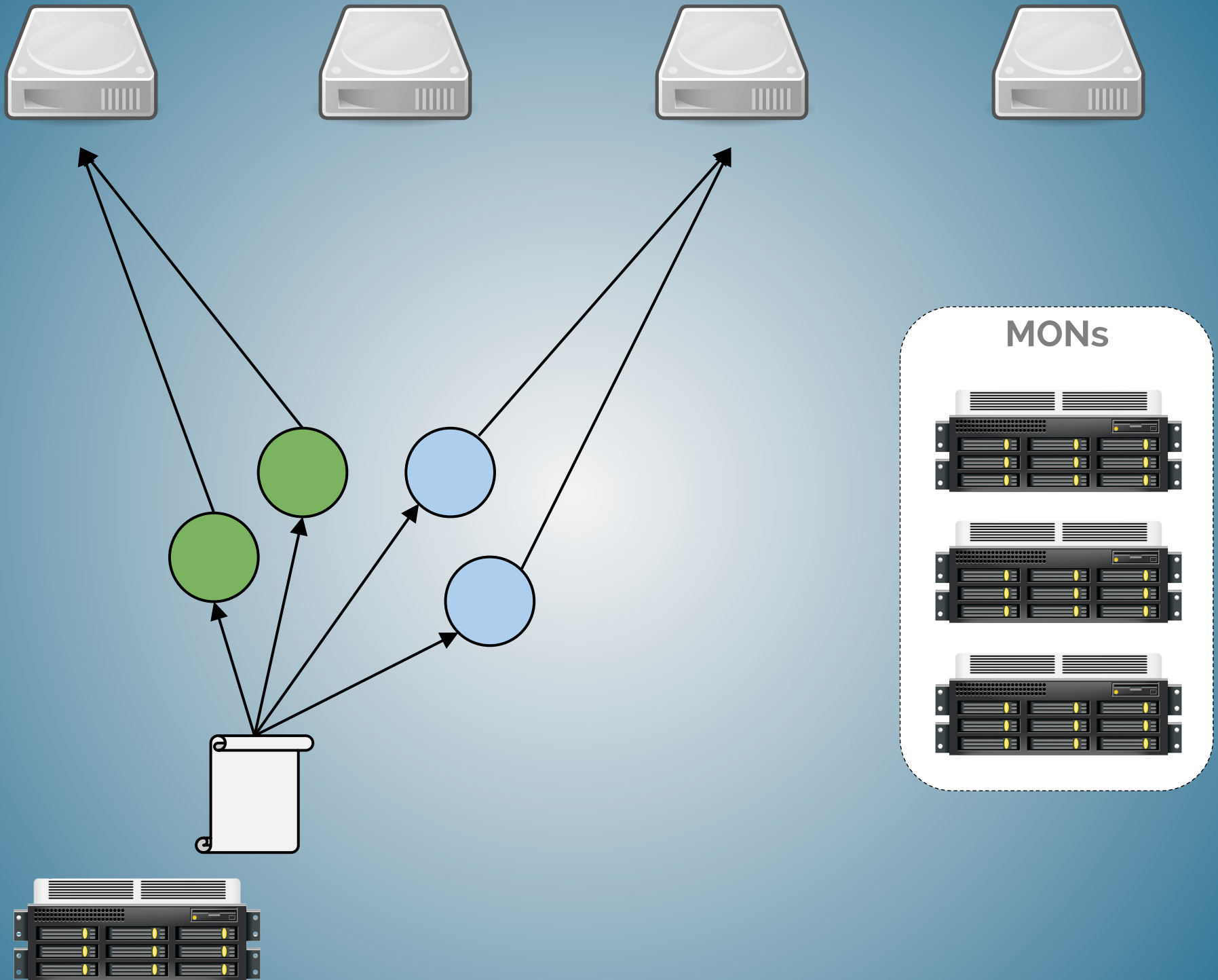
MONs





MONs





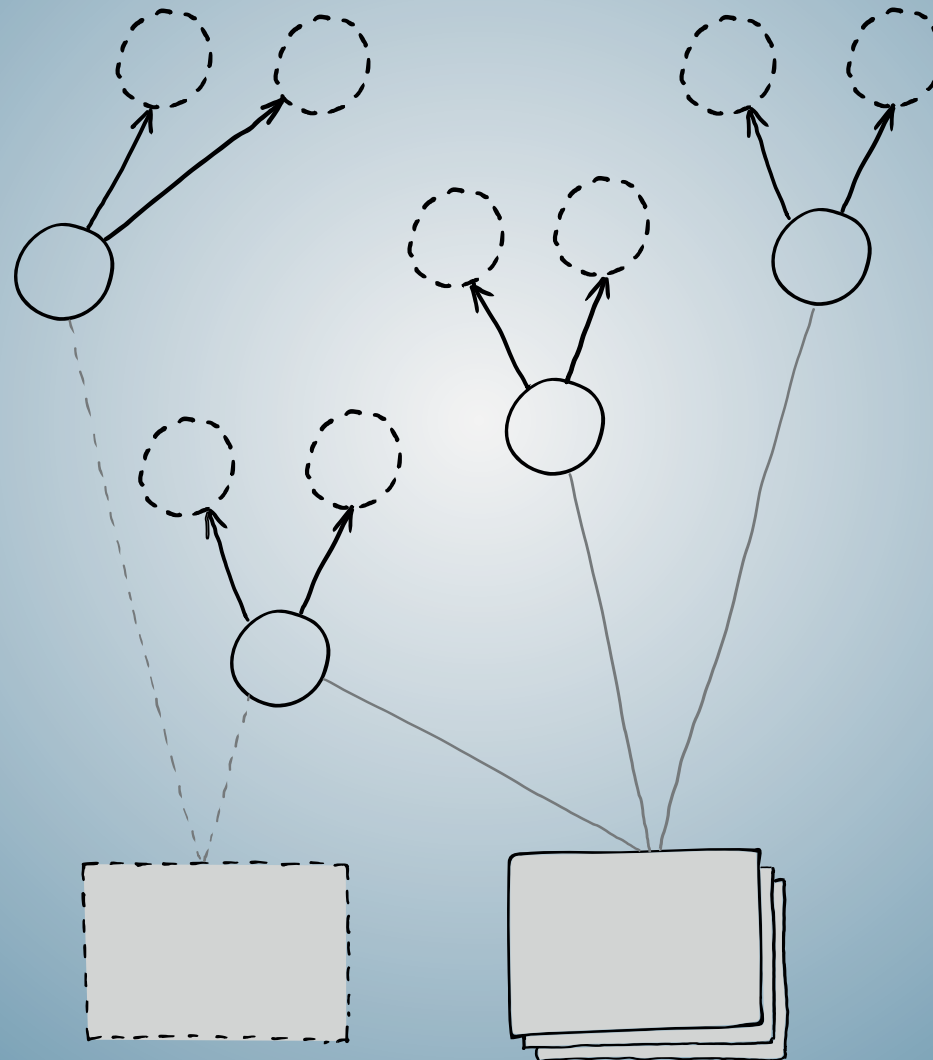
CRUSH

Controlled Replication Under Scalable Hashing

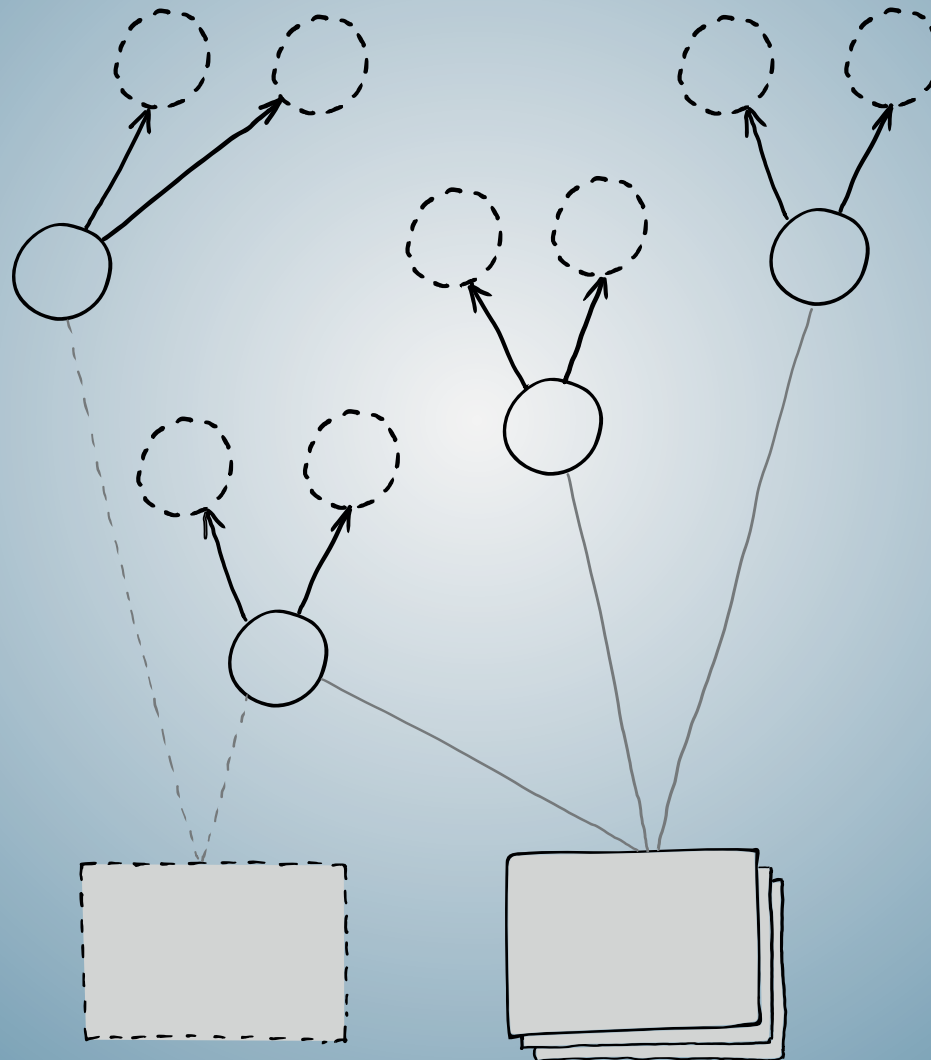
Rack aware

Clients?

Block Device Driver



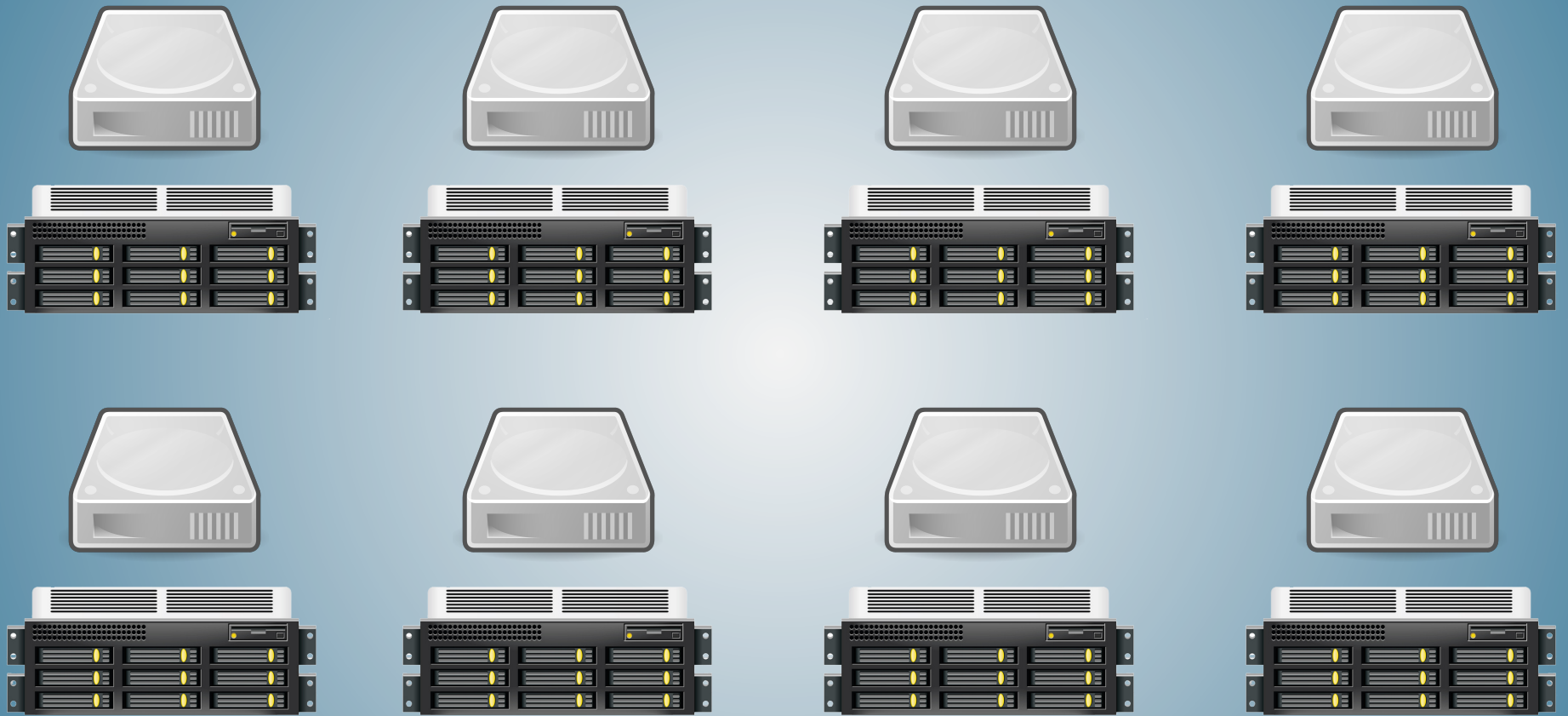
RBD (RADOS Block Device)



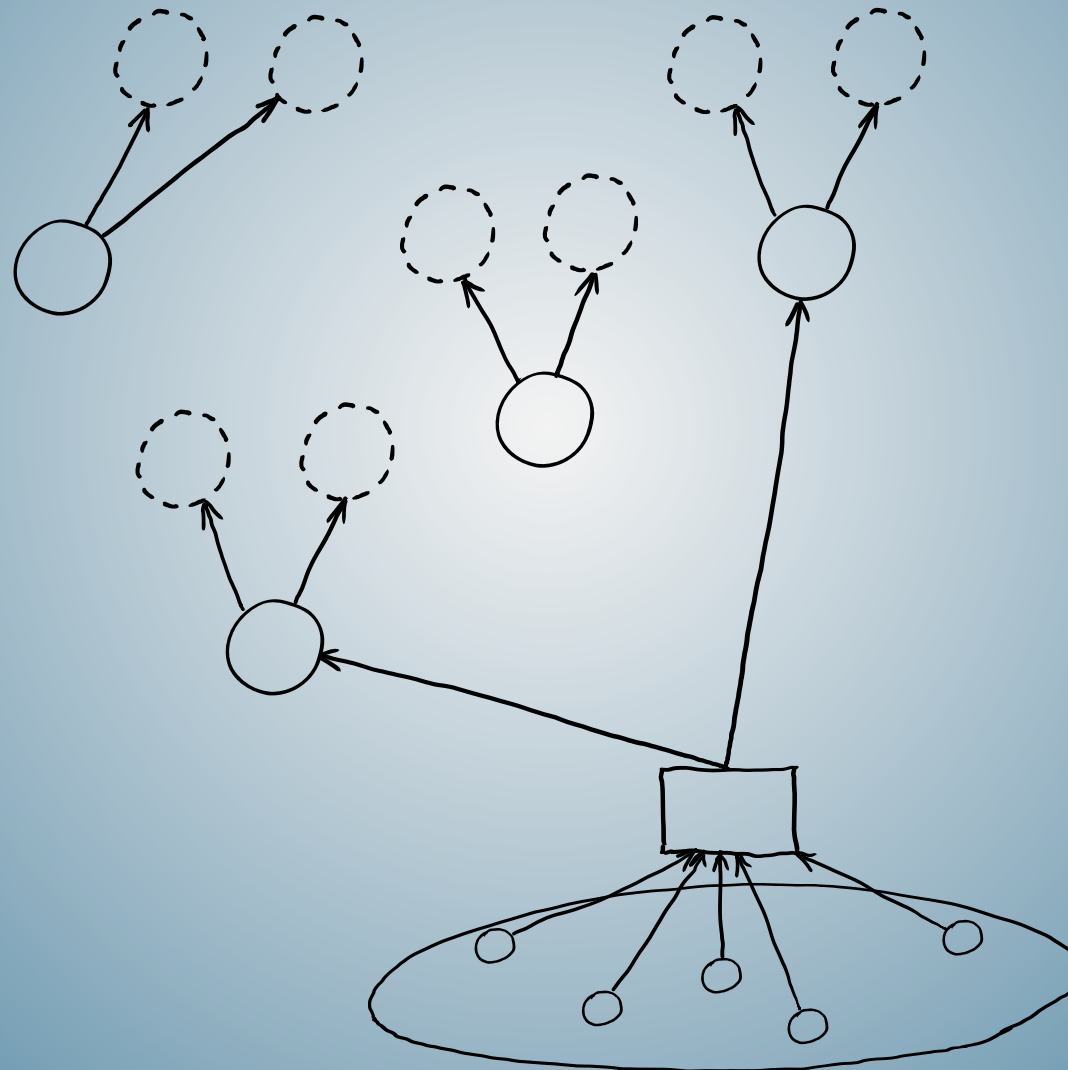
rbd

Qemu-RBD

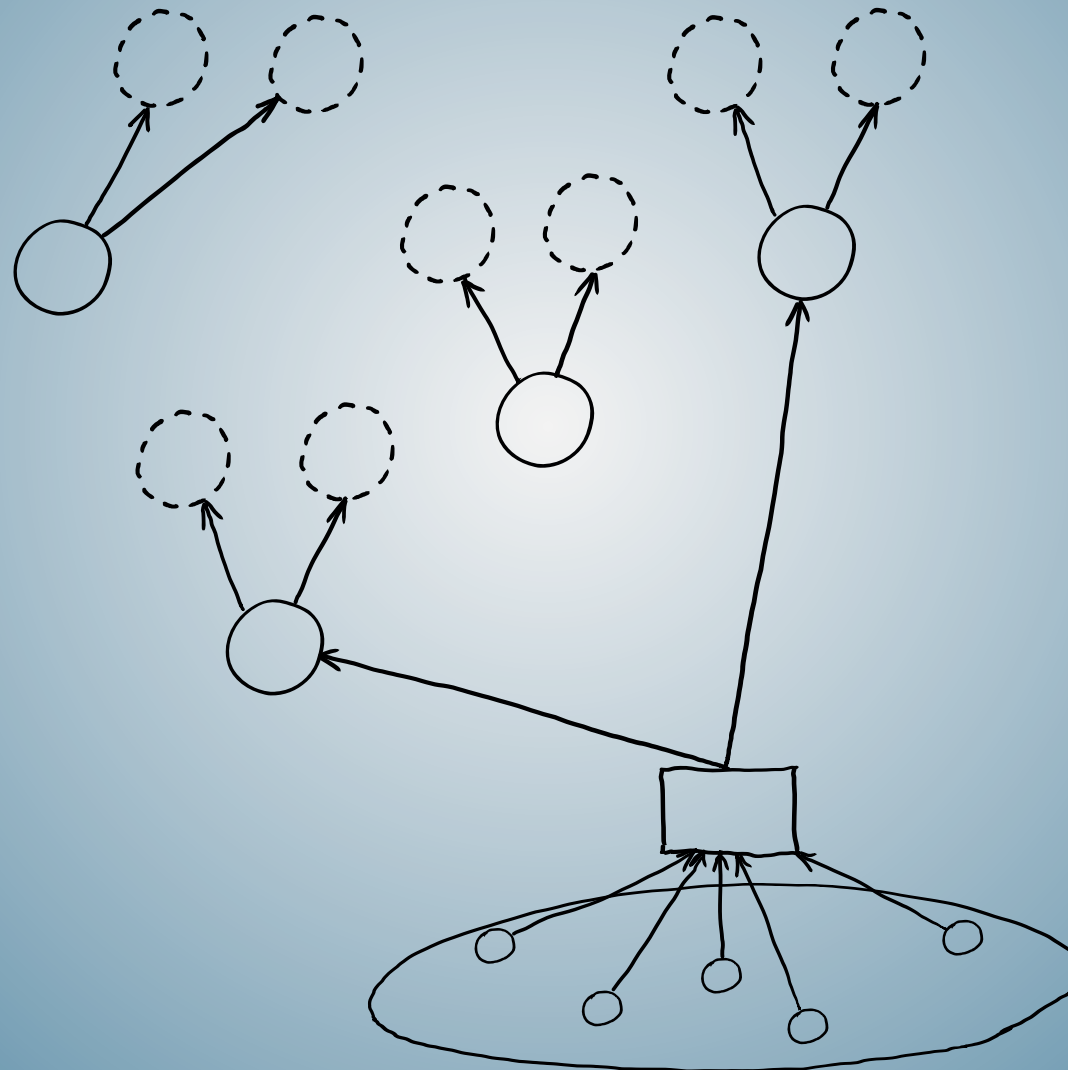
```
<disk type='network' device='disk'>
  <driver name='qemu' type='raw' cache='writeback' />
  <auth username='libvirt'>
    <secret type='ceph' usage='client.libvirt secret' />
  </auth>
  <source protocol='rbd' name='libvirt/ubuntu-amd64-alice'>
    <host name='192.168.133.111' port='6789' />
    <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>
```



ReSTful API

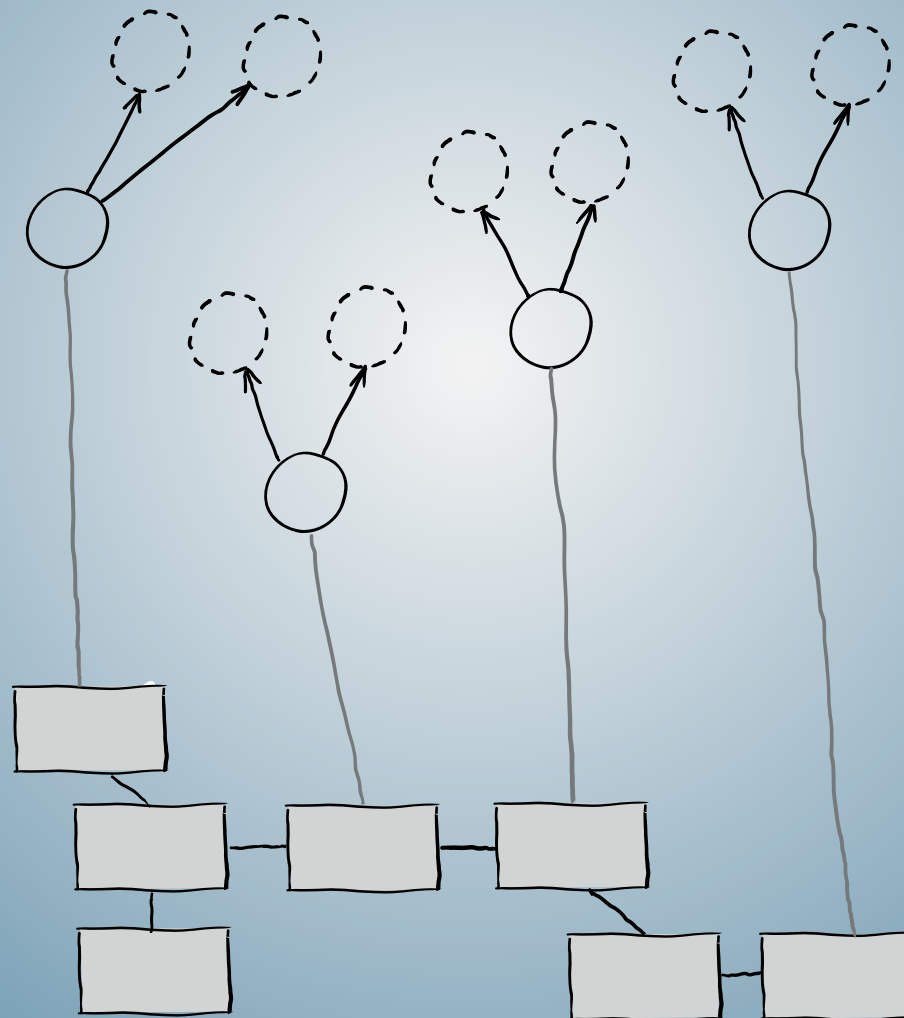


radosgw

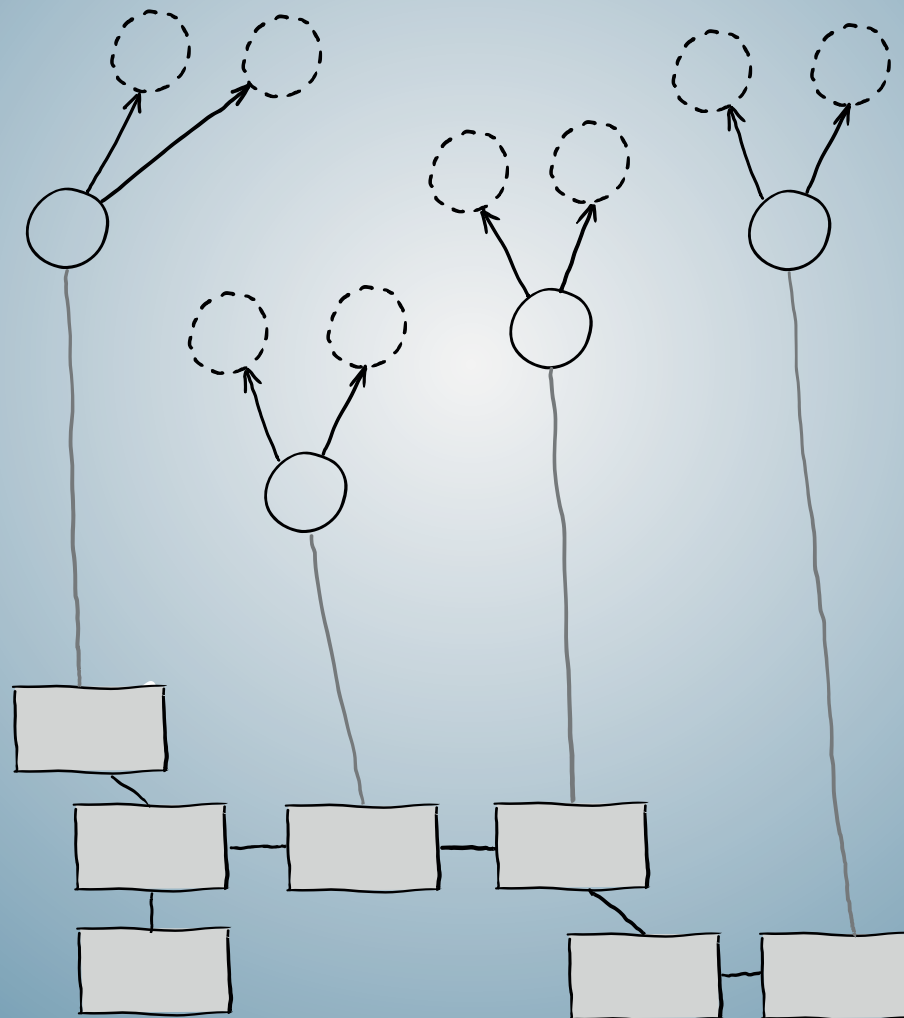


Kompatibel mit S3 und Swift

POSIX file system



CephFS



CephFS: Leider
noch Beta!

Keinen passenden
Client gefunden?

librados

Einsatzszenarien

Gigantic Storage

40TB, 3 Replikas = € 21.000

Virtualisierung

iSCSI Storage

Cloud

OpenStack

Live-Demo

24. Mai 2013

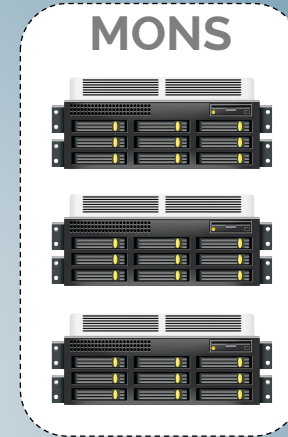
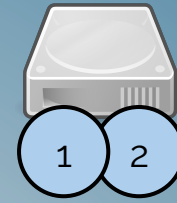
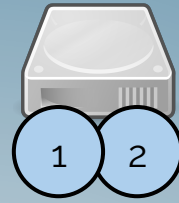
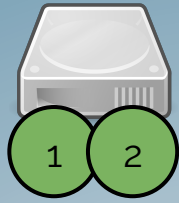
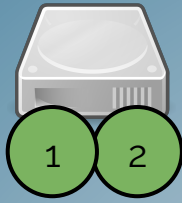
OpenStack DACH Tag 2013

<http://openstackdach2013.eventbrite.com>

Mit besonderem Dank an:

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

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



goo.gl/S1sYZ (Ich auf Google+)

goo.gl/LqWFB (Slides auf Slideshare)

twitter.com/hastexo

hastexo.com