https://gregorjs.github.io/ceph_at_uibk

CEPH

an der

UNIVERSITÄT INNSBRUCK

gregor.schwab@uibk.ac.at

Was ist Ceph?

Ceph ist ein

Software defined Storage (SDS)

System

Ceph wurde von

Sage Weil

in der Firma

Inktank

entwickelt und 2014 von

Redhat

gekauft

Warum Ceph

- distributed
- scalable
- free as in speech

Warum nicht Ceph?

- Performance
- Stabilität
- Benutzerfreundlichkeit

Kann Ceph produktiv eingesetzt werden?

Dies hängt von der Betrachtung ab

RADOS object store, RBD, und RadosGW

gelten als ausreichend stabil grosse Organisationen verwenden es

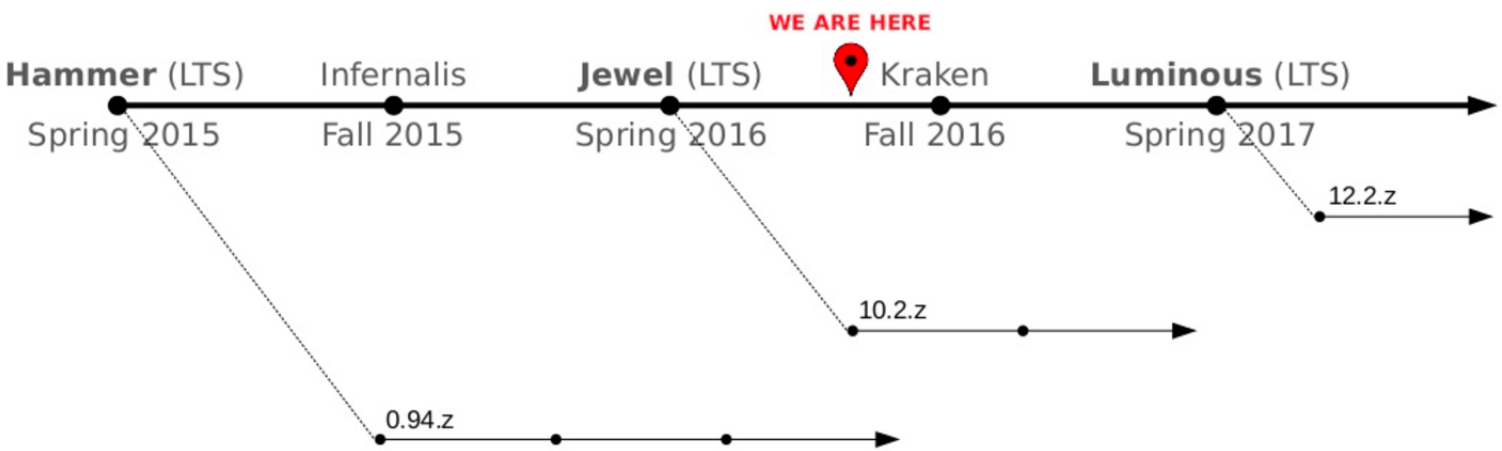
CephFS is stabil seit Jewel

(ohne Snapshots)

tendentiell stabiler als so manch kommerzielles System

Ceph Release Cycle

RELEASE CADENCE



Ceph Developer Community

GROWING DEVELOPMENT COMMUNITY

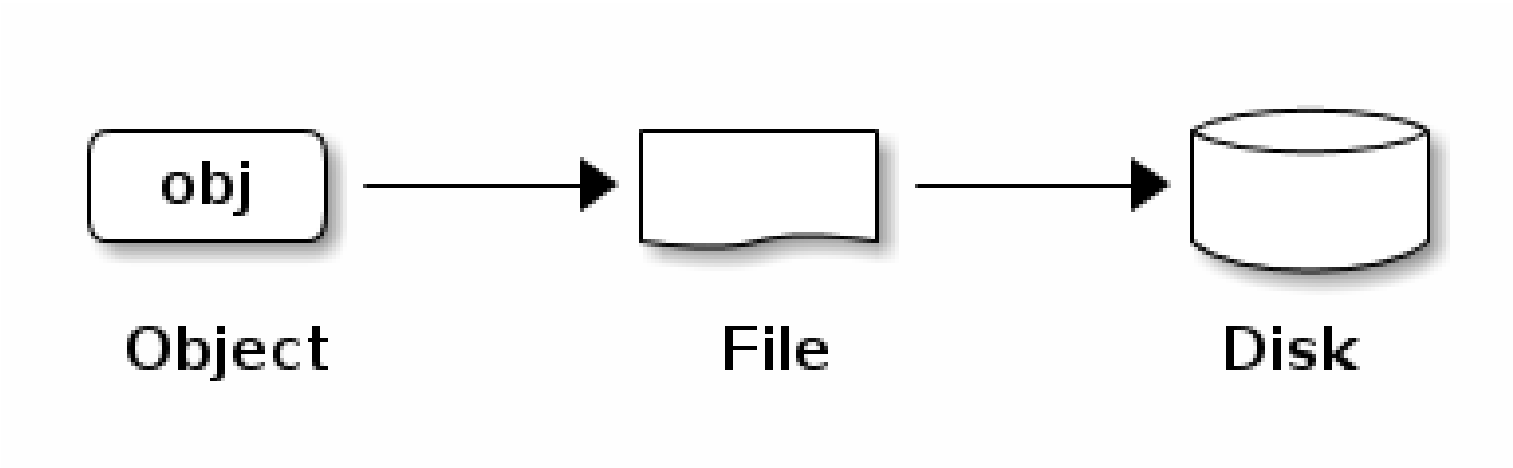
- **Red Hat**
- **Mirantis**
- **SUSE**
- SanDisk
- XSKY
- ZTE
- LETV
- Quantum
- **EasyStack**
- H3C
- **UnitedStack**
- Digiware
- Mellanox
- Intel
- Walmart Labs
- DreamHost
- Tencent
- Deutsche Telekom
- Igalia
- Fujitsu
- DigitalOcean
- University of Toronto

Ceph besteht aus

- OSDs
- Monitors
- MDS

Ceph ist eigentlich ein

OBJECT STORE



Was ist Ceph nicht?

CEPH ≠ RAID

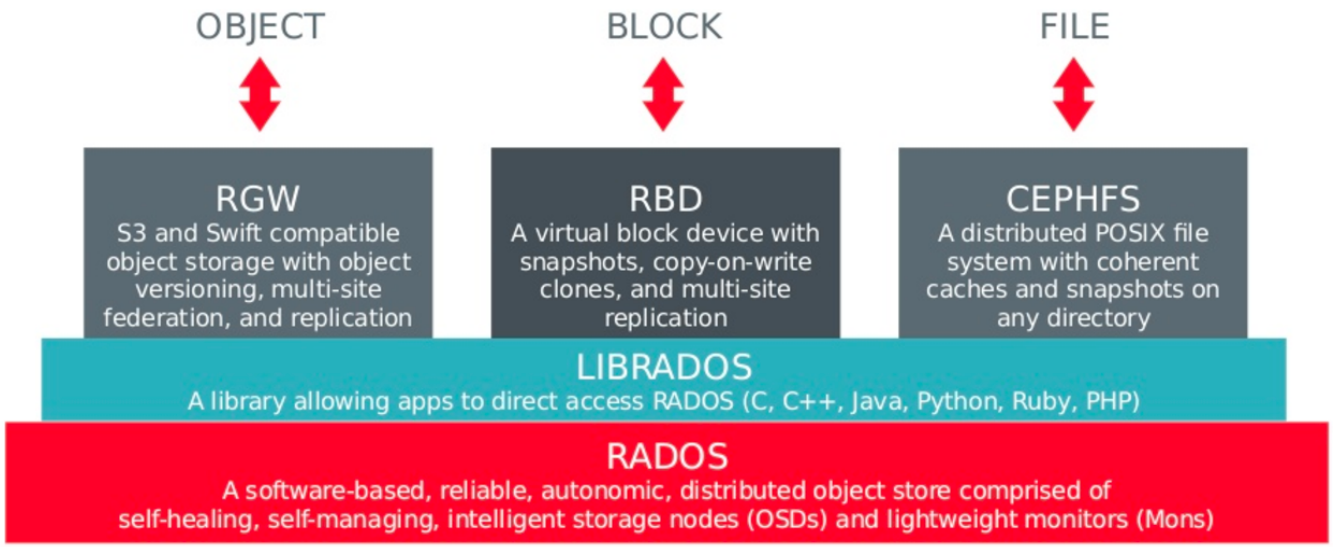
Ceph beherrscht aber

- Striping
- Replicas
- Erasure Coding
- Cache Pools

Ceph bietet mehrere

Storage Frontends

- RBD
- RADOS Gateway
- CephFS


OBJECT
BLOCK
FILE
RGW
S3 and Swift compatible object storage with object versioning, multi-site federation, and replication
RBD
A virtual block device with snapshots, copy-on-write clones, and multi-site replication
CEPHFS
A distributed POSIX file system with coherent caches and snapshots on any directory
LIBRADOS
A library allowing apps to direct access RADOS (C, C++, Java, Python, Ruby, PHP)
RADOS
A software-based, reliable, autonomic, distributed object store comprised of self-healing, self-managing, intelligent storage nodes (OSDs) and lightweight monitors (Mons)

RBD

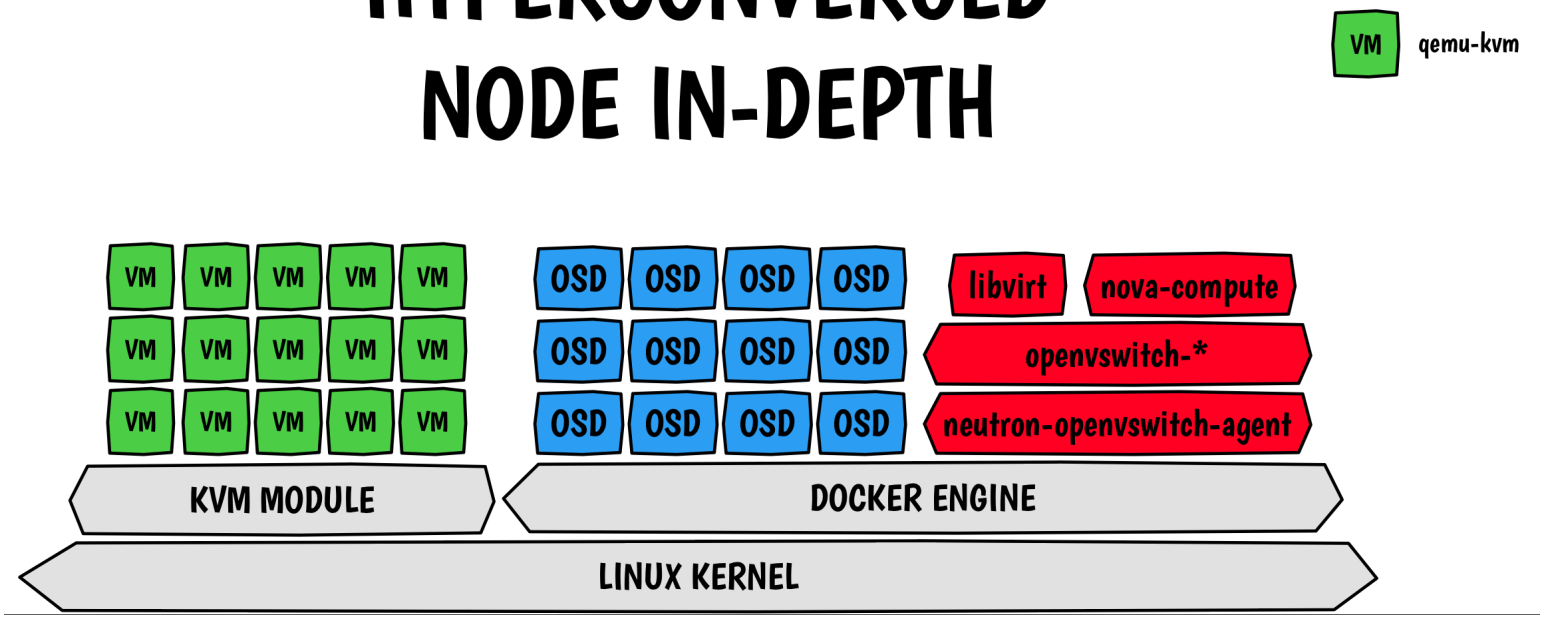
über

qemu-librbd

(caching)

CEPH UND HYPERCONVERGED ?

HYPERCONVERGED NODE IN-DEPTH

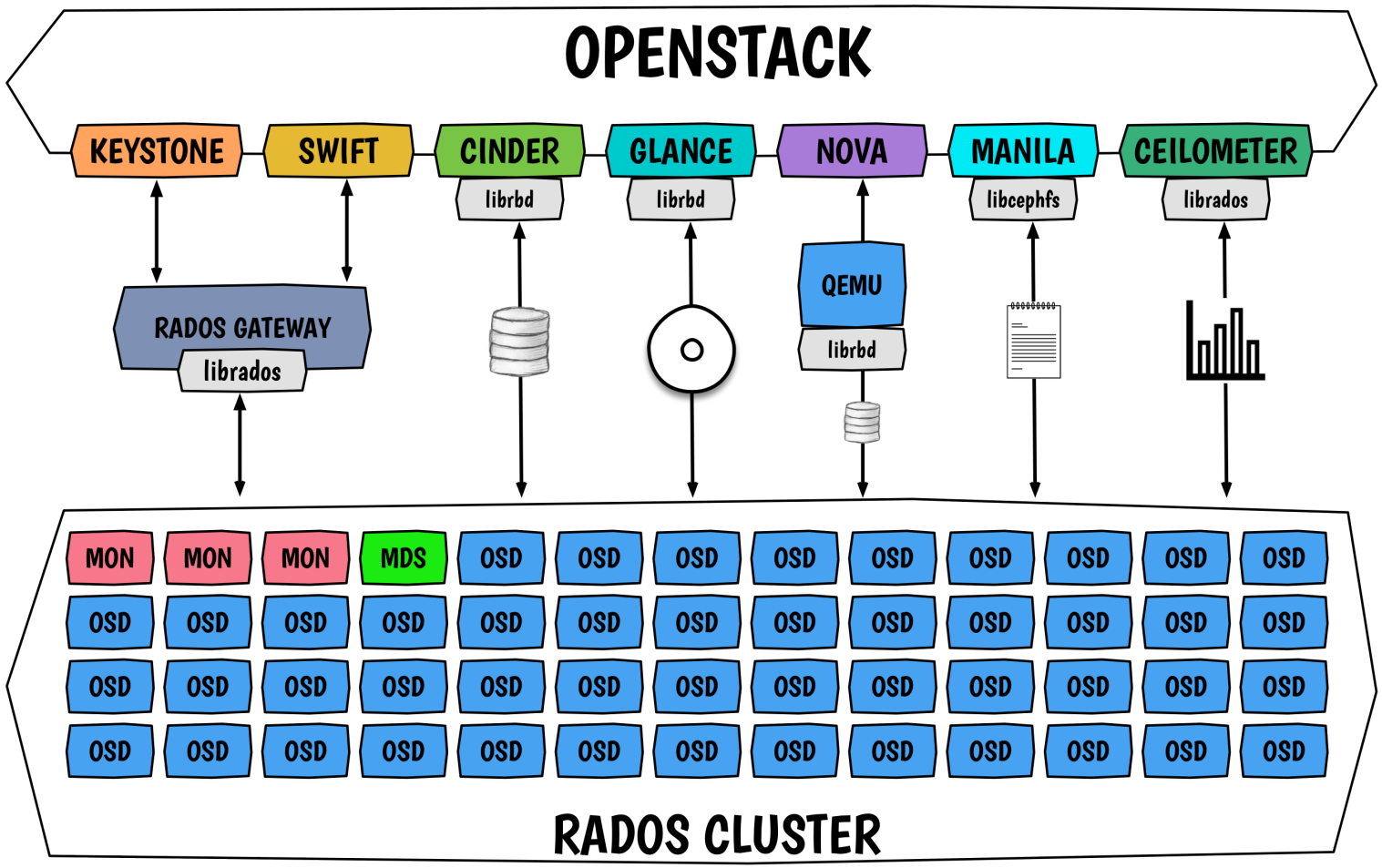


Unterstützung für

CEPH

in

OPENSTACK?


OPENSTACK
KEYSTONE
SWIFT
CINDER
GLANCE
NOVA
MANILA
CEILOMETER
librbd
librbd
libcephfs
librados
RADOS GATEWAY
librados
QEMU
librbd
MON
MON
MON
MDS
OSD
RADOS CLUSTER

Unterstützung für

CEPH

in

RHEV?

- seit RHEV 3.6

CEPH @ UIBK

Einige Fakten:

- Einsatz seit 2014
- RBD produktiv seit 2015
- RADOSGW seit 2015/16
- Vanilla Variante

Warum Vanilla?

- kein Vendor-lockin
- kein Hardware lockin
- weniger Gesamtkosten
- Transparenz
- Lernkurve für unified storage

2 Ceph Cluster

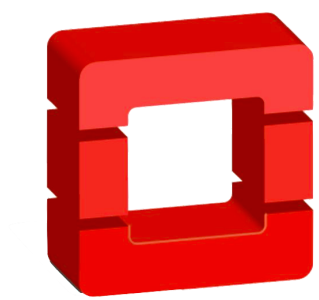
- Ceph Produktiv Cluster
  - 3 Mons, 4 Osd Nodes, Hammer
  - SSD journals
- Ceph Backup Cluster
  - 3 Mons, 2 Osd Nodes, Jewel
  - SSD journals

Anbindung über

Storage VLAN

Storage Backend über VLAN

OPENSTACK @ UIBK



Einige Fakten:

- Erstinistallation 2014
- Produktiv seit 2015
- Vanilla Variante

einige Produktivtenants

- Rados Gateway
- CI Runner
- Puppet / Foreman Infrastructure

Vorteil:

flexibler

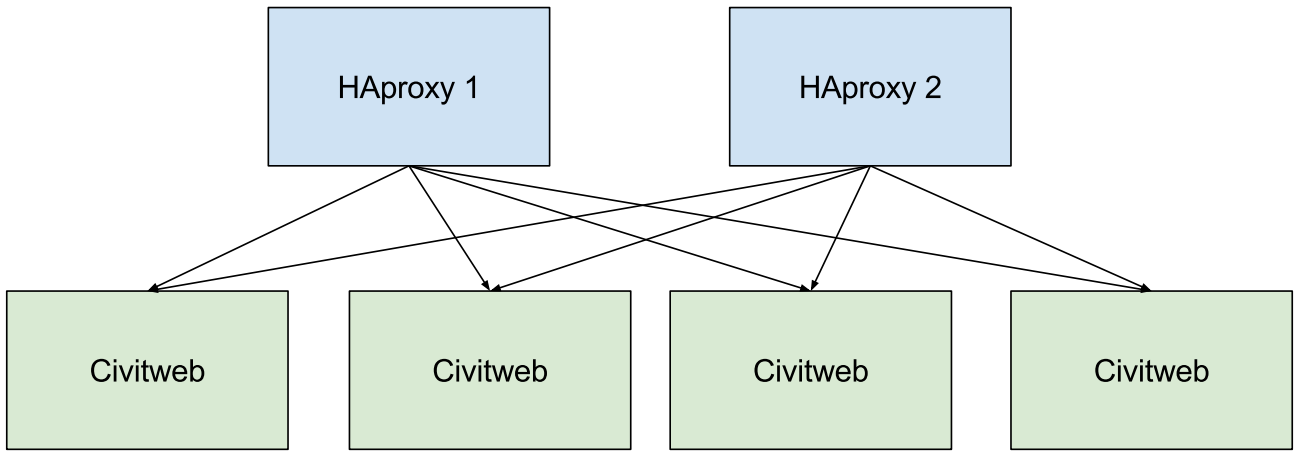
als andere Virtualisierungen

Nachteil: immer noch viele

Entwicklungen

RadosGW über 2 HAProxys

RadosGW Architektur



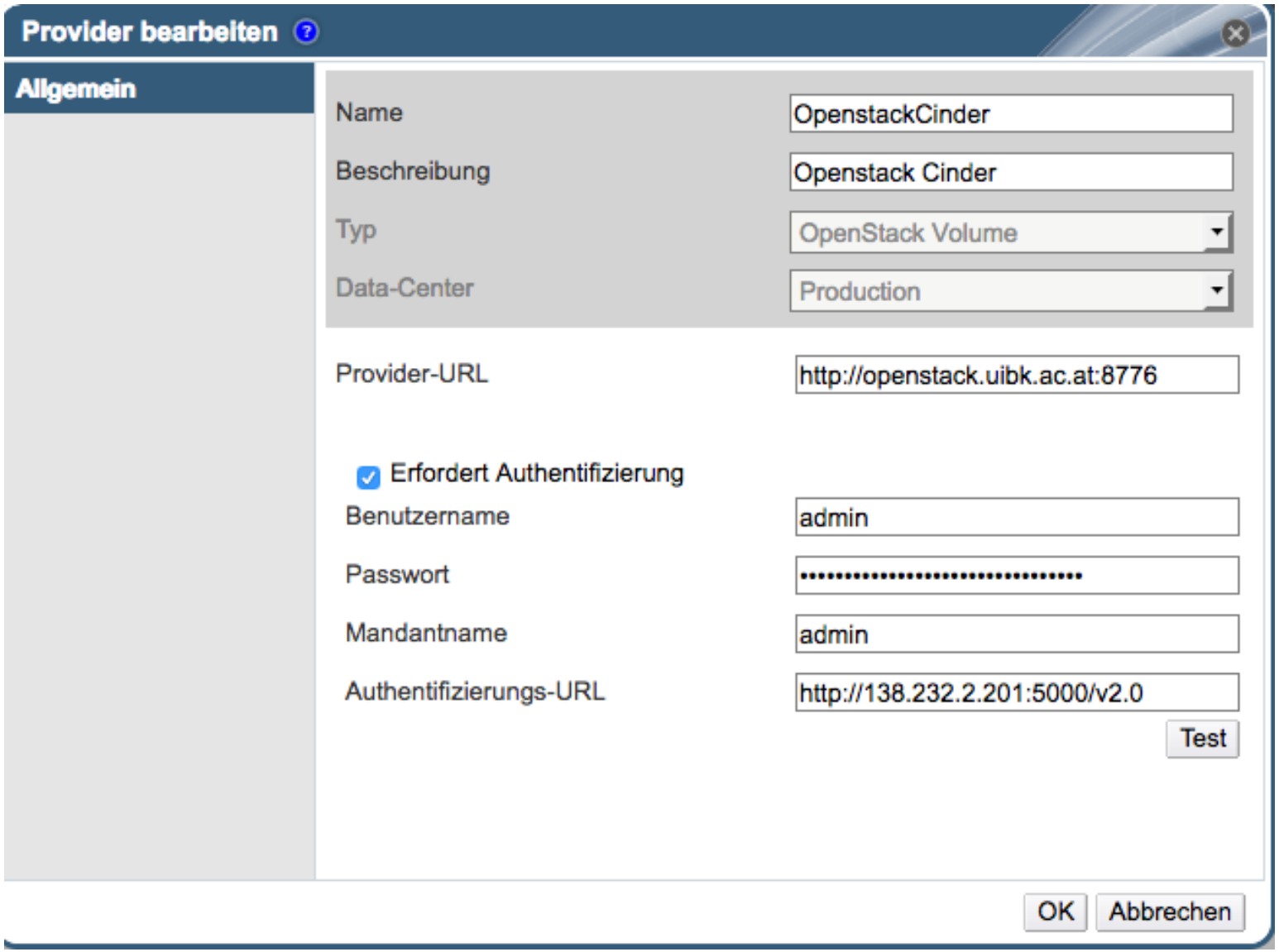
RBD

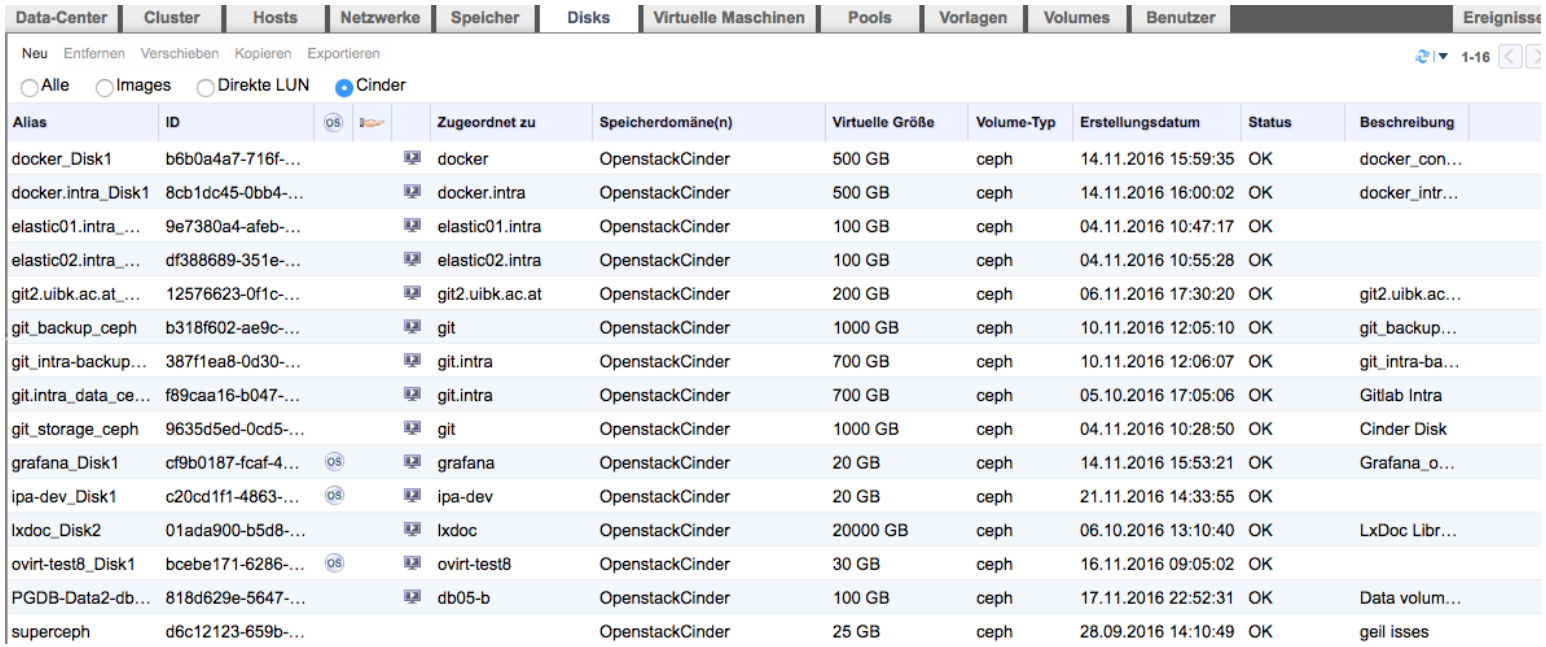
über

Cinder

in

- Openstack
- RHEV


Provider bearbeiten
Allgemein
Name
OpenstackCinder
Beschreibung
Openstack Cinder
Typ
OpenStack Volume
Data-Center
Production
Provider-URL
http://openstack.uibk.ac.at:8776
Erfordert Authentifizierung
Benutzername
admin
Passwort
Mandantname
admin
Authentifizierungs-URL
http://138.232.2.201:5000/v2.0
Test
OK
Abbrechen


Data-Center
Cluster
Hosts
Netzwerke
Speicher
Disks
Virtuelle Maschinen
Pools
Vorlagen
Volumes
Benutzer
Ereignisse
Neu
Entfernen
Verschieben
Kopieren
Exportieren
1-16
Alle
Images
Direkte LUN
Cinder
Alias
ID
Zugeordnet zu
Speicherdomäne(n)
Virtuelle Größe
Volume-Typ
Erstellungsdatum
Status
Beschreibung
docker_Disk1 b6b0a4a7-716f-… docker OpenstackCinder 500 GB ceph 14.11.2016 15:59:35 OK docker_con…
docker.intra_Disk1 8cb1dc45-0bb4-… docker.intra OpenstackCinder 500 GB ceph 14.11.2016 16:00:02 OK docker_intr…
elastic01.intra_… 9e7380a4-afeb-… elastic01.intra OpenstackCinder 100 GB ceph 04.11.2016 10:47:17 OK
elastic02.intra_… df388689-351e-… elastic02.intra OpenstackCinder 100 GB ceph 04.11.2016 10:55:28 OK
git2.uibk.ac.at_… 12576623-0f1c-… git2.uibk.ac.at OpenstackCinder 200 GB ceph 06.11.2016 17:30:20 OK git2.uibk.ac…
git_backup_ceph b318f602-ae9c-… git OpenstackCinder 1000 GB ceph 10.11.2016 12:05:10 OK git_backup…
git_intra-backup… 387f1ea8-0d30-… git.intra OpenstackCinder 700 GB ceph 10.11.2016 12:06:07 OK git_intra-ba…
git.intra_data_ce… f89caa16-b047-… git.intra OpenstackCinder 700 GB ceph 05.10.2016 17:05:06 OK Gitlab Intra
git_storage_ceph 9635d5ed-0cd5-… git OpenstackCinder 1000 GB ceph 04.11.2016 10:28:50 OK Cinder Disk
grafana_Disk1 cf9b0187-fcaf-4… grafana OpenstackCinder 20 GB ceph 14.11.2016 15:53:21 OK Grafana_o…
ipa-dev_Disk1 c20cd1f1-4863-… ipa-dev OpenstackCinder 20 GB ceph 21.11.2016 14:33:55 OK
lxdoc_Disk2 01ada900-b5d8-… lxdoc OpenstackCinder 20000 GB ceph 06.10.2016 13:10:40 OK LxDoc Libr…
ovirt-test8_Disk1 bcebe171-6286-… ovirt-test8 OpenstackCinder 30 GB ceph 16.11.2016 09:05:02 OK
PGDB-Data2-db… 818d629e-5647-… db05-b OpenstackCinder 100 GB ceph 17.11.2016 22:52:31 OK Data volum…
superceph d6c12123-659b-… OpenstackCinder 25 GB ceph 28.09.2016 14:10:49 OK geil isses

RBD Backups

über

Incremental Snapshots

RADOS Backups

über

RGW Multisite

(noch s3sync, ersetzt federated gateways)

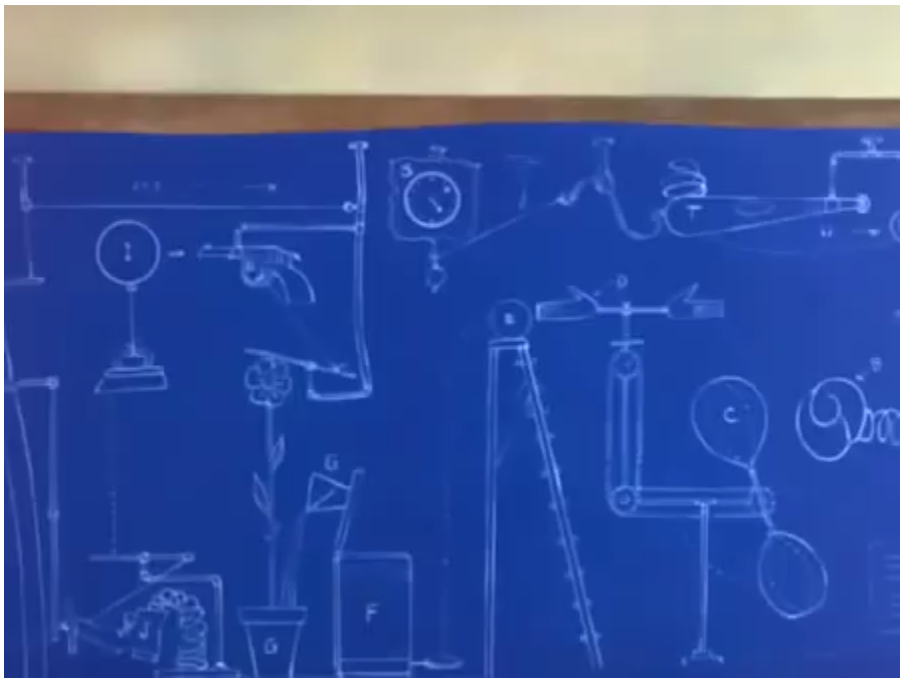
INSTALLATION:

keine

manuellen

Installationen

Automatisieren!



Installationsarten:

- Puppet (am Anfang)
- **Ansible**
- **Docker?**

positive Erfahrungen:

- sehr stabil
- läuft ohne Unterbruch
- Rolling Upgrades (manuell)
- ansible_ceph beste Codequalität

weniger schöne Erfahrungen:

- RPM Dependency Hell in Hammer
- RBD Kernelmodul
- richtigen Controller wählen!
- **parted-3.2" package BUG**

Performance

mit

caching

sehr gut

Operations:

Auskonfigurieren einer kaputten OSD

```
#!/bin/bash
ceph osd out $1
sudo service ceph-osd@$1 stop
ceph osd crush remove osd.$1
ceph auth del osd.$1
ceph osd rm $1
sudo umount /var/lib/ceph/osd/ceph-$1
sudo rmdir /var/lib/ceph/osd/ceph-$1
```

Einhängen einer neuen OSD

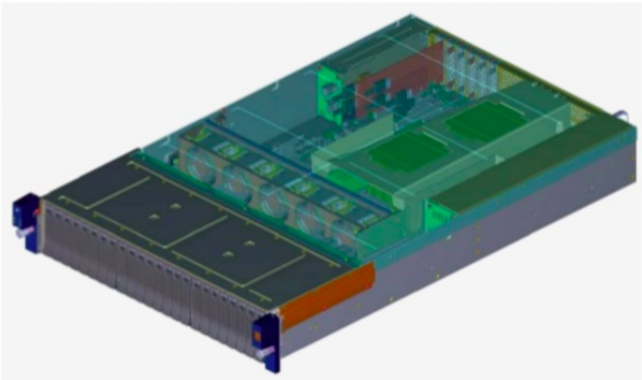
```
#!/bin/bash
CLUSTER_UUID="YOUR UUID"
disk=/dev/sdx
journal=/dev/sda2
ceph-disk prepare --cluster ceph --cluster-uuid $CLUSTER_UUID --fs-type x
ceph-disk activate ${disk}
```

Ausblick Ceph:

- Hybrid Cluster (SSD, HDD)
- Ceph on Flash
- Bluestore (Kraken Release)
- DPDK Stack (Kraken Release)
- Erasure code overwrites, append only (Kraken Release)
- ceph-mgr for metrics
- RBD ordered writeback cache, low latency

CEPH ON FLASH

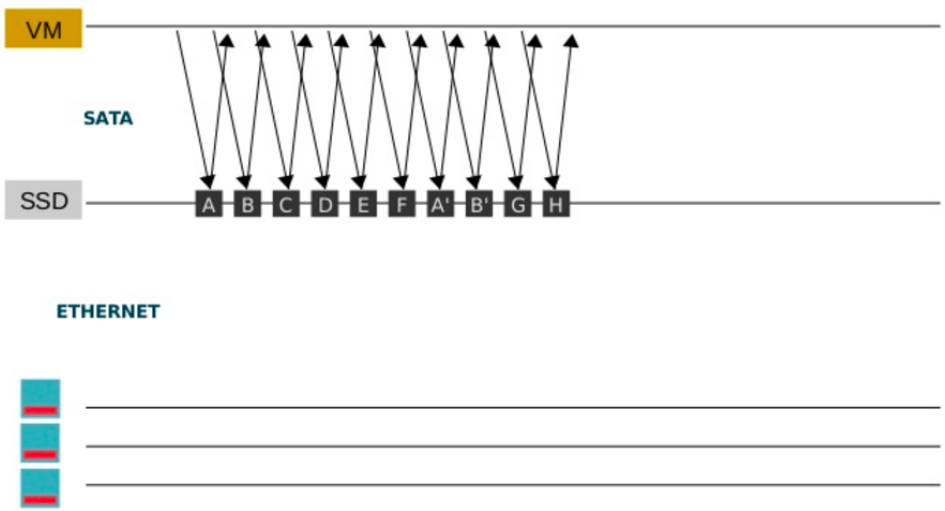
- Samsung
  - up to 153 TB in 2u
  - 700K IOPS, 30 GB/s



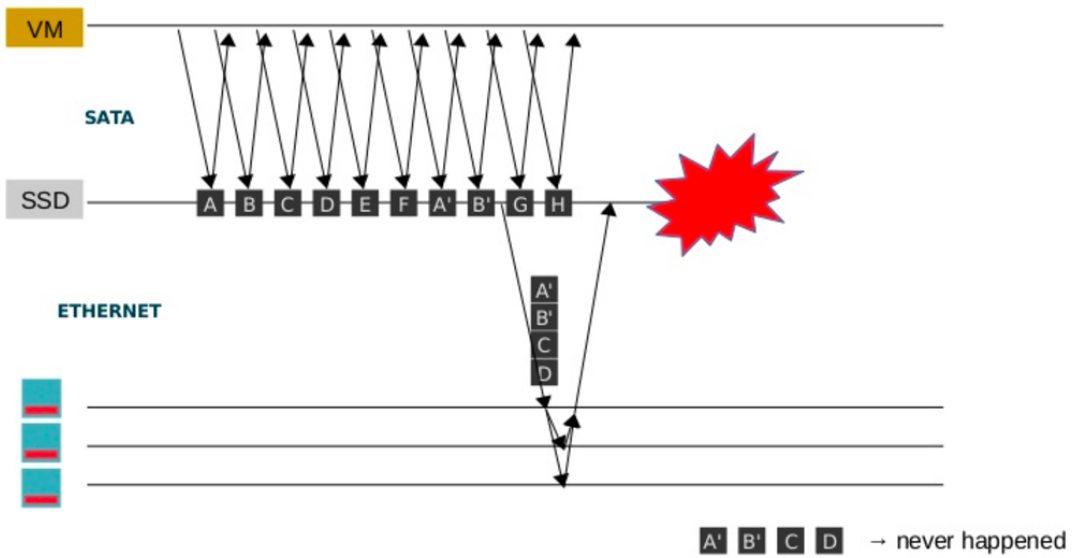
- SanDisk
  - up to 512 TB in 3u
  - 780K IOPS, 7 GB/s

Ordered writeback cache

LOCAL SSD → LOW LATENCY

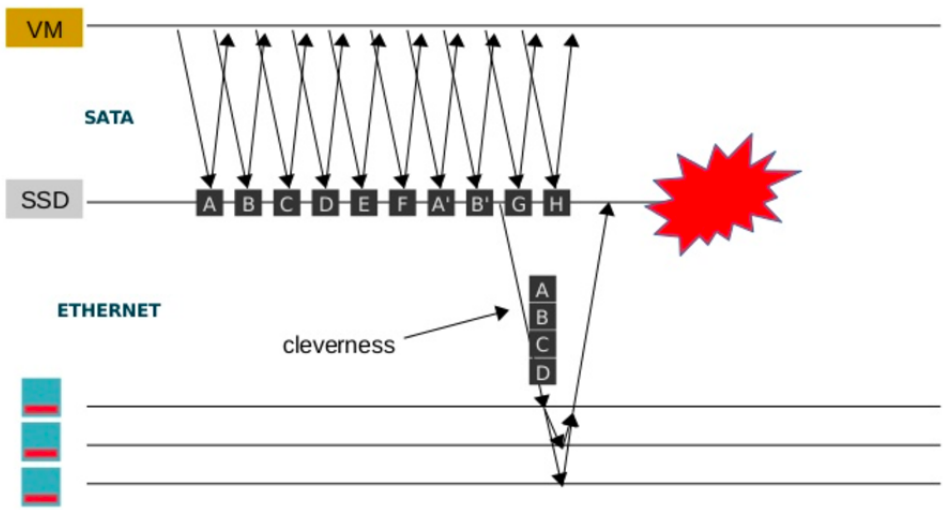


WRITEBACK IS UNORDERED



A' B' C D → never happened

RBD ORDERED WRITEBACK CACHE



A B C D → CRASH CONSISTENT!

Ausblick Openstack:

- Newton Release Upgrade (in 2 Wochen)
- Docker Integration

Kurze Demo

Die Präsentation enthält hier eine kleine Live Demo.

DANKE!