Storage Cluster mit Ceph

CeBIT 2015

20. März 2015

Michel Rode
Linux/Unix Consultant & Trainer
B1 Systems GmbH
rode@b1-systems.de

Vorstellung B1 Systems

- gegründet 2004
- primär Linux/Open Source-Themen
- national & international tätig
- über 60 Mitarbeiter
- unabhängig von Soft- und Hardware-Herstellern
- Leistungsangebot:
  - Beratung & Consulting
  - Support
  - Entwicklung
  - Training
  - Betrieb
  - Lösungen
- dezentrale Strukturen

Schwerpunkte

- Virtualisierung (XEN, KVM & RHEV)
- Systemmanagement (Spacewalk, Red Hat Satellite, SUSE Manager)
- Konfigurationsmanagement (Puppet & Chef)
- Monitoring (Nagios & Icinga)
- IaaS Cloud (OpenStack & SUSE Cloud & RDO)
- Hochverfügbarkeit (Pacemaker)
- Shared Storage (GPFS, OCFS2, DRBD & CEPH)
- Dateiaustausch (ownCloud)
- Paketierung (Open Build Service)
- Administratoren oder Entwickler zur Unterstützung des Teams vor Ort

Storage Cluster

Was sind Storage Cluster?

- hochverfügbare Systeme
- verteilte Standorte
- skalierbar (mehr oder weniger)
- Problem: Häufig Vendor-Lock-In

Beispiele 1/2

- Dell PowerVault
- IBM SVC
- Netapp Metro Cluster
- Netapp Clustered Ontap
- ...

Beispiele 2/2

- AWS S3
- Rackspace Files
- Google Cloud Storage
- Microsoft Azure

Alternativen

- DRBD
- CEPH
- ...

Was ist Ceph?

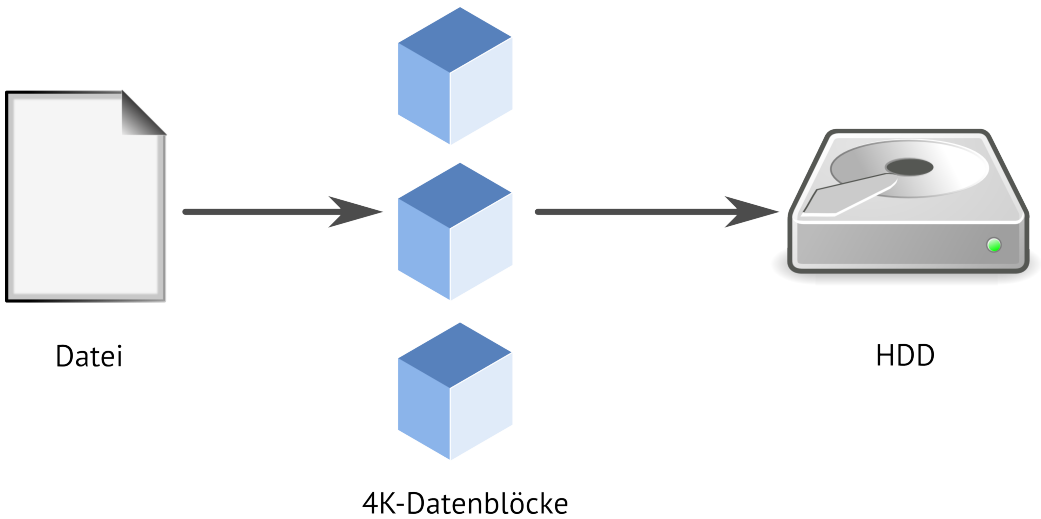
Was ist Ceph?

- Storage Cluster (Distributed Object Store)
- Open Source (LGPL)
- Object/Block/File Storage

Ziele bei der Entwicklung von Ceph

- kein SPOF (*Single Point of Failure*)
- hohe Skalierbarkeit
- gute Parallelisierung

Block Storage



- Block Storage selber:
  - Files werden gesplittet → Blocks
  - jeweils eigene Adresse
  - keine Metadata

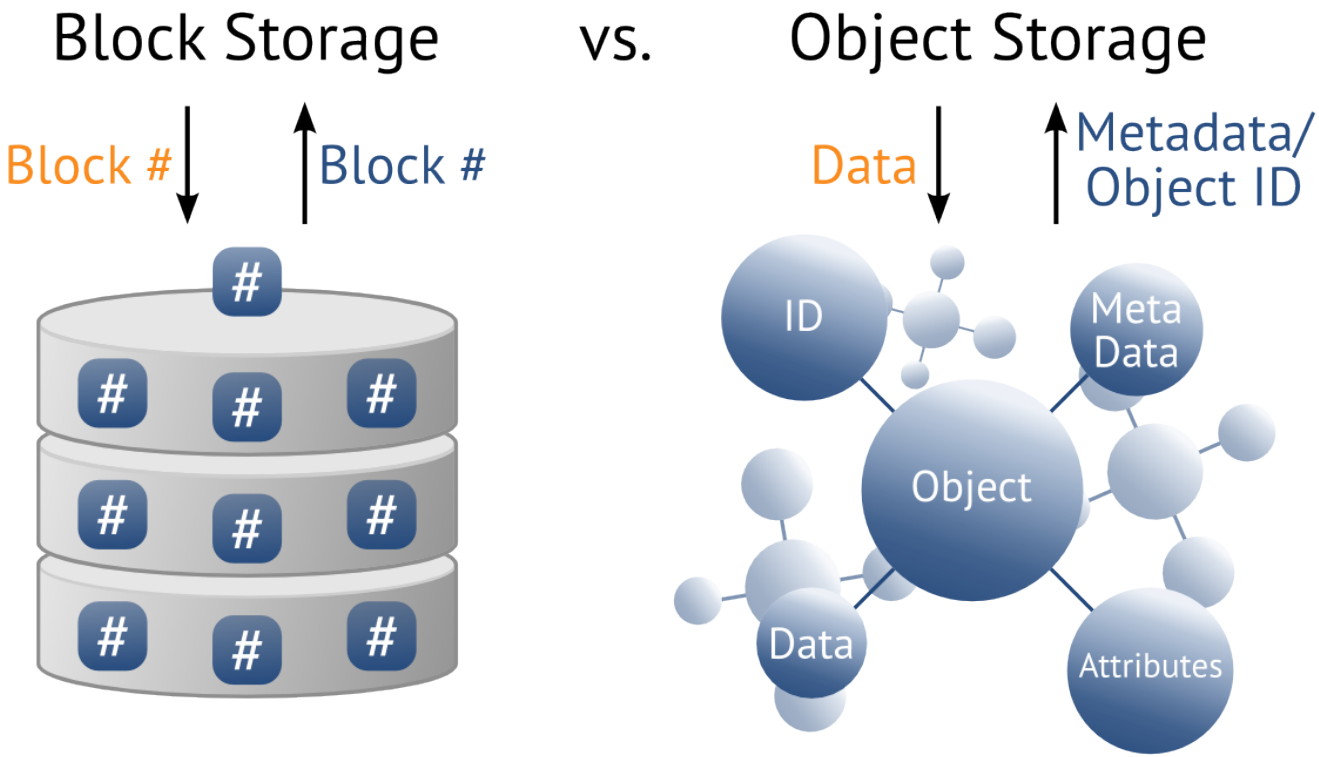
Block Storage

- RADOS Block Device/RBD
- Integration in KVM
  - OpenStack
  - SUSE OpenStack Cloud
  - Proxmox
- resizeable images
- read-only snapshots
- revert to snapshots

Object Storage

- Data – Bilder bis Manuals bis Videos
- Metadata – Kontextinformationen für die Daten
- Index/Identifier – natürlich unique!

Object vs Block



Quelle: `http://www.druva.com/wp-content/uploads/Screen-Shot-2014-08-18-at-11.02.02-AM-500x276.png`

File Storage

- „Stronger data safety for mission-critical applications“
- POSIX-konform
- automatisches Verteilen – bessere Performance!
- CephFS

Gateway/RGW

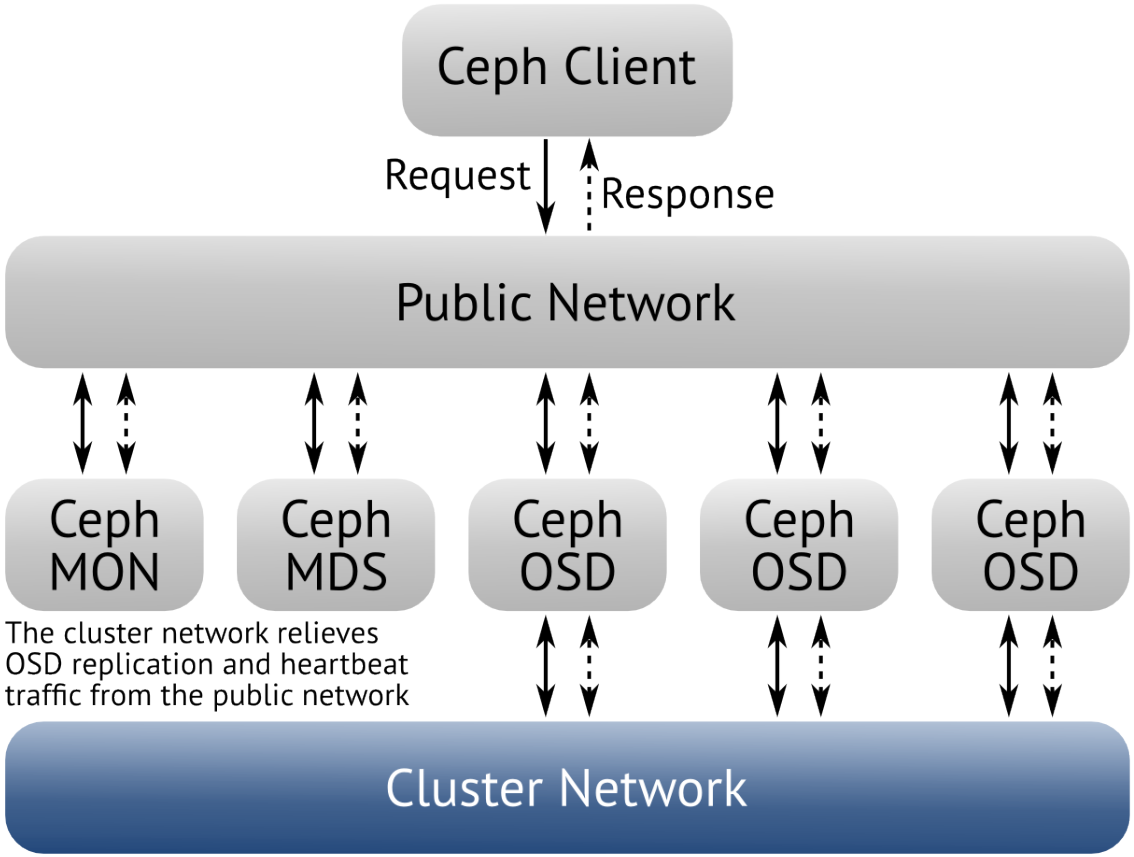
- RESTful API
- Interface für
  - OpenStack Swift
  - Amazon S3

Aufbau von Ceph

Aufbau von Ceph

- Object Storage Device – OSD
- Monitor – MON
- Metadata Server – MDS

Aufbau

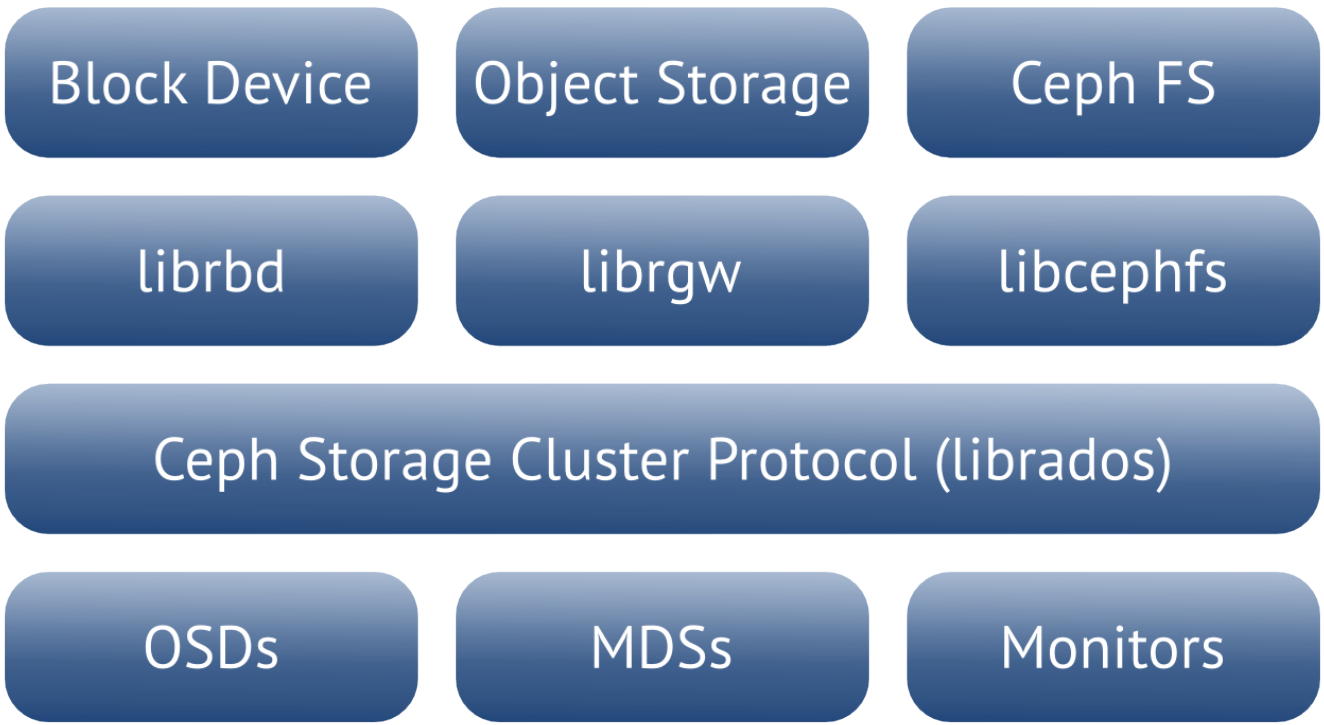

Ceph Client
Request
Response
Public Network
Ceph MON
Ceph MDS
Ceph OSD
Ceph OSD
Ceph OSD
The cluster network relieves OSD replication and heartbeat traffic from the public network
Cluster Network

Funktionsweise von Ceph

Funktionsweise von Ceph

- automatisches Verteilen und Replizieren der Daten
- RAID-0
- CRUSH Map
- Client kommuniziert direkt mit allen Systemen im Cluster

Funktionsweise von Ceph



ceph-mon – Ceph Monitor Daemon

- Map – aktive/inaktive Nodes
- mindestens 1
- hochverfügbar!
- mit Paxos zum Quorum (2/3, 3/5)

ceph-osd – Ceph Object Storage Daemon 1/4

- kann und darf ausfallen
- mindestens drei Knoten
- paralleler Zugriff
- CRUSH-Map

ceph-osd – Ceph Object Storage Daemon 2/4

- Object → File → Disk

Tabelle

ID	Binary	Metadata
1234	100101	name1 value1
4321	010010	name2 value2

- Semantik liegt beim Client
- ID ist eindeutig

ceph-osd – Ceph Object Storage Daemon 3/4

Dateisystem:

- Test-Umgebungen:
  - BTRFS
  - ZFS
- Produktiv-Systeme:
  - ext3 (kleine Umgebung)
  - XFS (Enterprise-Umgebung)

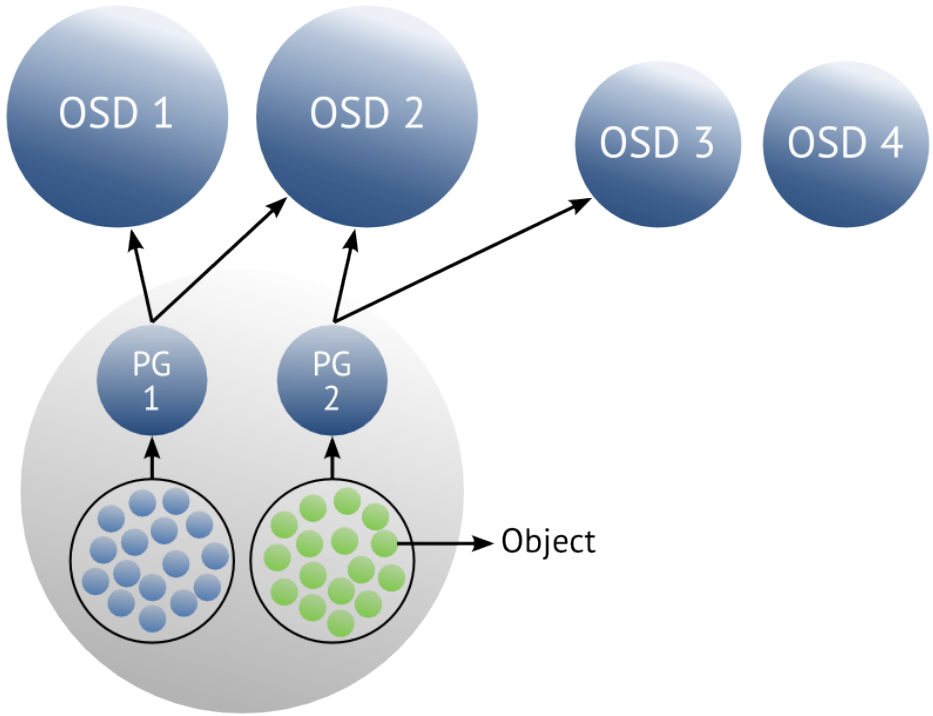
ceph-osd – Ceph Object Storage Daemon 4/4

- Daten werden erst in Journal geschrieben
- Tipp: 4 OSD pro SSD

`ceph-mds` – Ceph Metadata Server Daemon

- speichert Inodes und Directories
- erforderlich für CephFS
- kein separater Speicher

CRUSH Maps



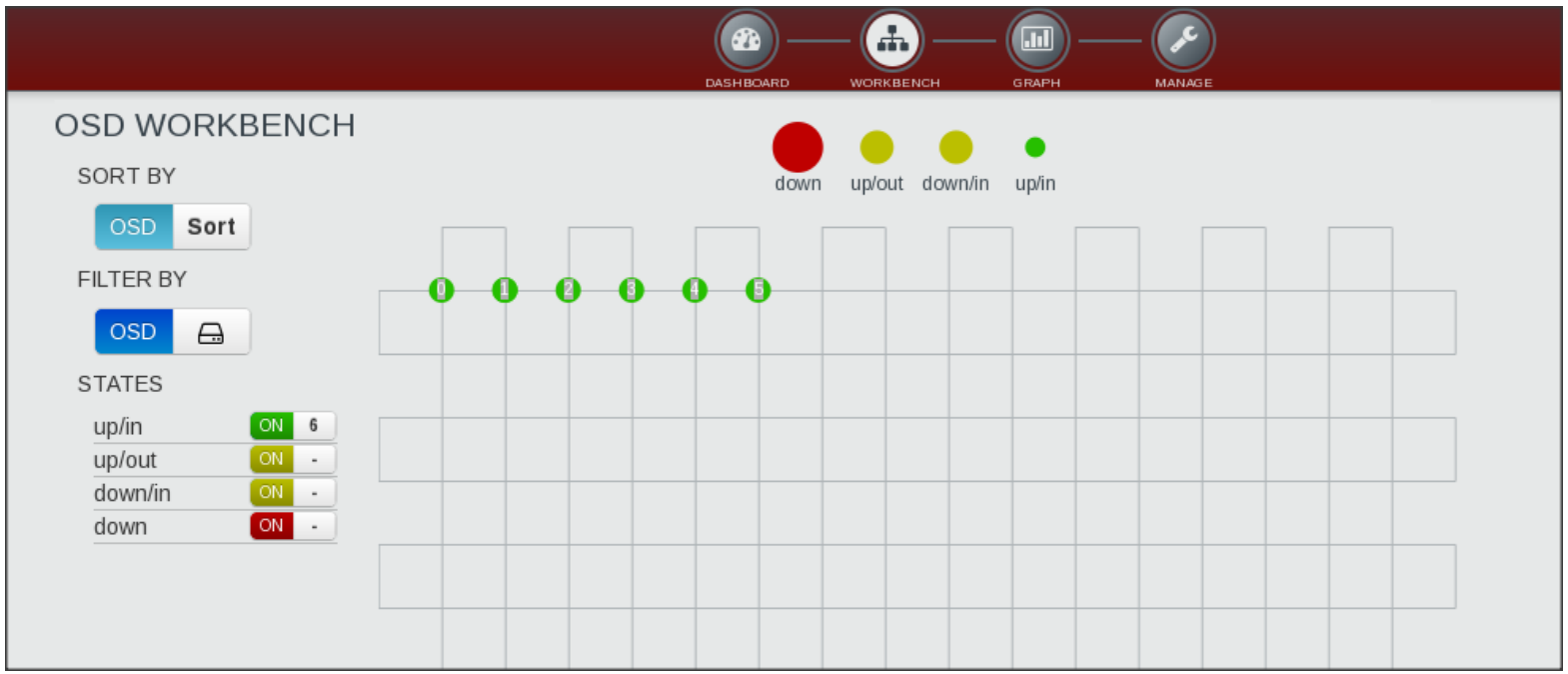
- CRUSH – *Controlled Replication Under Scalable Hashing*
- Datei (oid) → Objekt (pgid) → PGs → CRUSH (pgid) → osd1,osd2
- Jeder mit Jedem!
- Platzierungsregeln

Quelle: `http://www.sebastien-han.fr/images/ceph-data-placement.jpg`

War das alles?

- Erasure Coding
- Tiering
- Federation
- Chef
- Calamari

Calamari


DASHBOARD
WORKBENCH
GRAPH
MANAGE
OSD WORKBENCH
SORT BY
OSD
Sort
FILTER BY
OSD
STATES
up/in ON 6
up/out ON -
down/in ON -
down ON -
down
up/out
down/in
up/in
0
1
2
3
4
5

Getting Started

Getting Started

- Der Weg zum Ceph-Cluster
  - `ceph-deploy`
  - Sandbox ok
  - Produktion nogo
- oder händisch
- Was noch?
  - OSD tree
  - Pools

Release VS Client

- Fehlermeldung: Feature mismatch

```
ceph osd crush tunables legacy
```

Mini-Howto

Schritt 1 – Initiale Konfiguration erstellen

```
$ ceph-deploy new <mons>
[ceph_deploy.new][DEBUG ] Creating new cluster named ceph
[ceph_deploy.new][DEBUG ] Resolving host ceph01
[ceph_deploy.new][DEBUG ] Monitor ceph01 at 192.168.122.191
[ceph_deploy.new][INFO  ] making sure passwordless SSH succeeds
[ceph_deploy.new][DEBUG ] Monitor initial members are ['ceph01']
[ceph_deploy.new][DEBUG ] Monitor addrs are ['192.168.122.191']
[ceph_deploy.new][DEBUG ] Creating a random mon key...
[ceph_deploy.new][DEBUG ] Writing initial config to ceph.conf...
[ceph_deploy.new][DEBUG ] Writing monitor keyring to ceph.mon.keyring...
```

Schritt 2 – Pakete installieren

```
$ ceph-deploy install <nodes>
[ceph_deploy.install][INFO  ] Distro info: Fedora 20 Heisenbug
[ceph01][INFO  ] installing ceph on ceph01
[ceph01][INFO  ] Running command: rpm --import \
  https://ceph.com/git/?p=ceph.git;a=blob_plain;f=keys/release.asc
[ceph01][INFO  ] Running command: rpm -Uvh --replacepkgs --force --quiet \
  http://ceph.com/rpm-firefly/fc20/noarch/ceph-release-1-0.fc20.noarch.rpm
[...]
[ceph01][INFO  ] Running command: yum -y -q install ceph
[ceph01][INFO  ] Running command: ceph --version
[ceph01][DEBUG ] ceph version 0.81 (8de9501df275a5fe29f2c64cb44f195130e4a8fc)
[ceph_deploy.install][DEBUG ] Detecting platform for host ceph02 ...
```

Schritt 3 – Monitor(e) erstellen 1/3

```
$ ceph-deploy mon create-initial
[ceph_deploy.mon][DEBUG ] Deploying mon, cluster ceph hosts ceph01
[ceph_deploy.mon][DEBUG ] detecting platform for host ceph01 ...
[...]
[ceph_deploy.mon][INFO  ] distro info: Fedora 20 Heisenbug
[ceph01][DEBUG ] determining if provided host has same hostname in remote
[ceph01][DEBUG ] write cluster configuration to /etc/ceph/{cluster}.conf
[ceph01][DEBUG ] create the mon path if it does not exist
[ceph01][DEBUG ] checking for done path: /var/lib/ceph/mon/ceph-ceph01/done
[ceph01][DEBUG ] create a done file to avoid re-doing the mon deployment
[ceph01][DEBUG ] create the init path if it does not exist
[ceph01][DEBUG ] locating the ‘service‘ executable...
[...]
```

Schritt 3 – Monitor(e) erstellen 2/3

```
$ ceph-deploy mon create-initial
[...]
[ceph01][INFO  ] Running command: /usr/sbin/service ceph
  -c /etc/ceph/ceph.conf start mon.ceph01
[ceph01][DEBUG ] === mon.ceph01 ===
[ceph01][DEBUG ] Starting Ceph mon.ceph01 on ceph01...
[ceph01][DEBUG ] Starting ceph-create-keys on ceph01...
[ceph01][INFO  ] Running command: ceph --cluster=ceph
  --admin-daemon /var/run/ceph/ceph-mon.ceph01.asok mon_status
[ceph01][DEBUG ] ********************************************
[ceph01][DEBUG ] status for monitor: mon.ceph01
[...]
```

Schritt 3 – Monitor(e) erstellen 3/3

```
$ ceph-deploy mon create-initial
[...]
[ceph01][DEBUG ]     "mons": [
[ceph01][DEBUG ]       {
[ceph01][DEBUG ]         "addr": "192.168.122.191:6789/0",
[ceph01][DEBUG ]         "name": "ceph01",
[ceph01][DEBUG ]         "rank": 0
[ceph01][DEBUG ]       }
[ceph01][DEBUG ]     ]
[ceph01][DEBUG ]   },
[...]
[ceph_deploy.mon][INFO  ] mon.ceph01 monitor has reached quorum!
[ceph_deploy.mon][INFO  ] Running gatherkeys...
```

Schritt 4 – Verzeichnisse anlegen

```
$ mkdir /var/local/osdX
```

Schritt 5 – OSD prepare

```
$ ceph-deploy osd prepare ceph01:/var/loca/ceph01 ceph02...
[ceph_deploy.conf][DEBUG ] found configuration file at:
  /root/.cephdeploy.conf
[ceph_deploy.cli][INFO  ] Invoked (1.5.11): /usr/bin/ceph-deploy
  osd prepare ceph01:/var/local/osd1 ceph02:/var/local/osd2
  ceph03:/var/local/osd3
[ceph_deploy.osd][DEBUG ] Preparing cluster ceph disks
  ceph01:/var/local/osd1: ceph02:/var/local/osd2: ceph03:/var/local/osd3:
[ceph01][DEBUG ] write cluster configuration to /etc/ceph/{cluster}.conf
[ceph01][INFO  ] Running command: udevadm trigger --subsystem-match=block
  --action=add
[ceph_deploy.osd][DEBUG ] Preparing host ceph01 disk /var/local/osd1
  journal None activate False
[ceph01][INFO  ] Running command: ceph-disk -v prepare --fs-type xfs
  --cluster ceph -- /var/local/osd1
[ceph01][WARNIN] DEBUG:ceph-disk:Preparing osd data dir /var/local/osd1
[ceph01][INFO  ] checking OSD status...
[ceph01][INFO  ] Running command: ceph --cluster=ceph osd stat
  --format=json
[ceph_deploy.osd][DEBUG ] Host ceph01 is now ready for osd use.
```

Schritt 6 – OSD activate

```
$ ceph-deploy osd activate ceph01:/var/loca/ceph01 ceph02...
[ceph_deploy.osd][DEBUG ] Activating cluster ceph disks
  ceph01:/var/local/osd1: ceph02:/var/local/osd2: ceph03:/var/local/osd3:
[ceph01][INFO  ] Running command: ceph-disk -v activate --mark-init sysvinit
  --mount /var/local/osd1
[ceph01][DEBUG ] === osd.0 ===
[ceph01][DEBUG ] Starting Ceph osd.0 on ceph01...
[ceph01][DEBUG ] starting osd.0 at :/0 osd_data /var/lib/ceph/osd/ceph-0
  /var/lib/ceph/osd/ceph-0/journal
[ceph01][WARNIN] DEBUG:ceph-disk:Cluster uuid is 1d6a5501-5b8f-4a3a-8c92-...
[ceph01][WARNIN] DEBUG:ceph-disk:Cluster name is ceph
[ceph01][WARNIN] DEBUG:ceph-disk:OSD uuid is 3e05a33e-785d-41d3-8d4b-...
[ceph01][WARNIN] DEBUG:ceph-disk:OSD id is 0
[ceph01][WARNIN] DEBUG:ceph-disk:Marking with init system sysvinit
[ceph01][WARNIN] DEBUG:ceph-disk:ceph osd.0 data dir is ready at
  /var/local/osd1
[ceph01][WARNIN] DEBUG:ceph-disk:Creating symlink /var/lib/ceph/osd/ceph-0
  -> /var/local/osd1
[ceph01][WARNIN] DEBUG:ceph-disk:Starting ceph osd.0...
[ceph01][WARNIN] create-or-move updating item name 'osd.0' weight 0.01
  at location {host=ceph01,root=default} to crush map
```

Schritt 7 – Deploy Adminkeys

```
$ ceph-deploy admin <nodes>
```

Schritt 8 – Check Health

```
$ ceph -s
cluster 8ae29b47-245a-4ef6-a5cc-d5d5fb7417bd
health HEALTH_OK
monmap e1: 1 mons at {ceph01=192.168.122.191:6789/0}, election epoch 1,
  quorum 0 ceph01
osdmap e22: 3 osds: 3 up, 3 in
pgmap v40: 192 pgs, 3 pools, 0 bytes data, 0 objects
19478 MB used, 5314 MB / 26191 MB avail
192 active+clean
```

Vielen Dank für Ihre Aufmerksamkeit!

Bei weiteren Fragen wenden Sie sich bitte an info@b1-systems.de
oder +49 (0)8457 - 931096.

Besuchen Sie uns auch hier auf der CeBIT,
Halle 6, H16/312.