OSZ IMT

ceph

Aufbau und Anwendung eines objektbasierten Speichersystems

guug Frühjahrsfachgespräch
25.02.2016

Christian Schubert

www.oszimt.de

Überblick

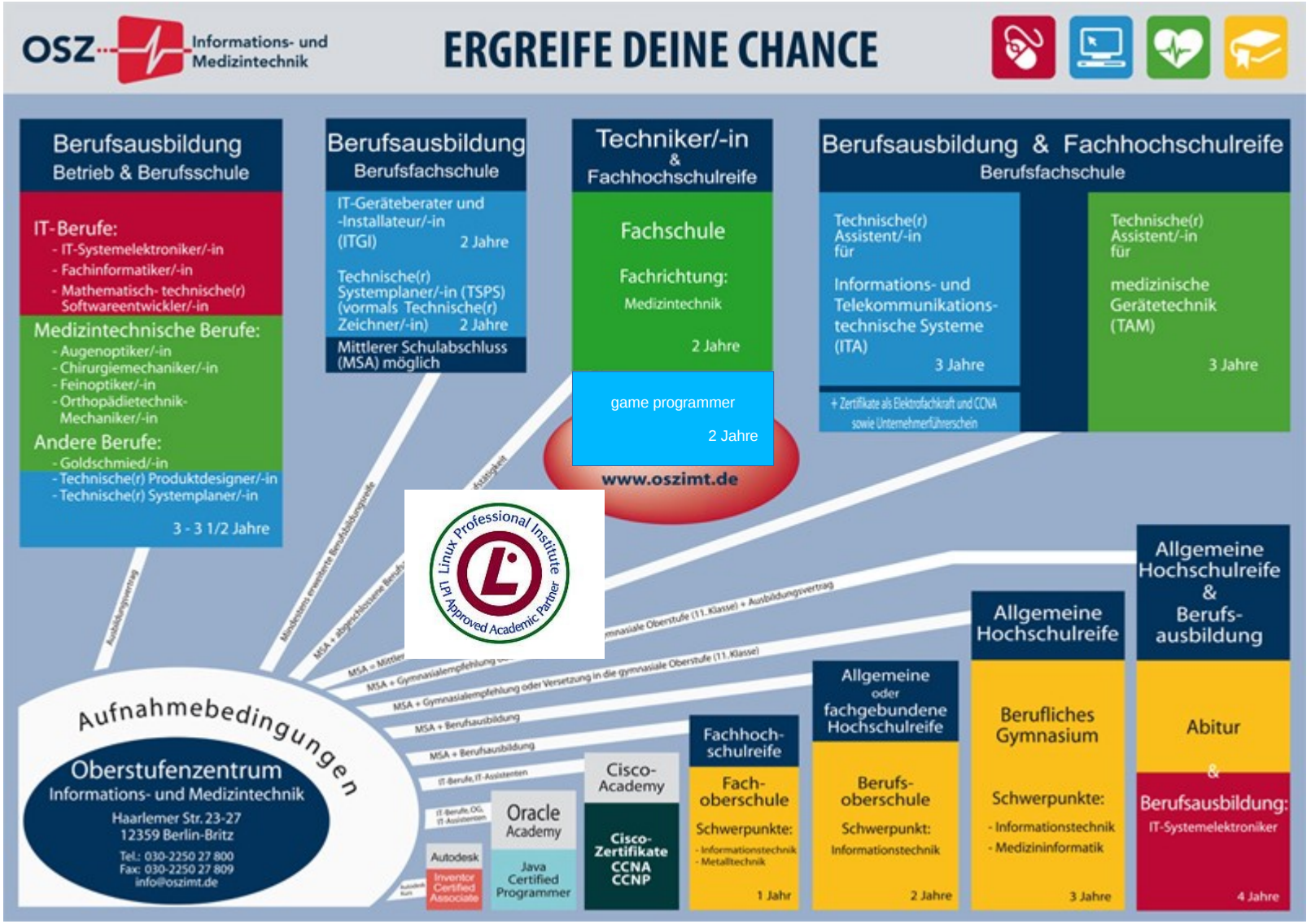
- OSZimt – Portfolio
- Motivation
- Ceph – Aufbau & Funktion
- Zwei Anwendungsbeispiele
  - Ceph im OSZimt
  - Ceph im Krankenhaus
- Hinweise & Tipps
- Fazit

Über mich

Christian Schubert
Lehrer
schubert@oszimt.de

Oberstufenzentrum Informations- und Medizintechnik
Haarlemer Straße 23-27
12359 Berlin
Fon +49 30 225027 800
Fax +49 30 225027 809

OSZimt – Portfolio



OSZimt – in Zahlen

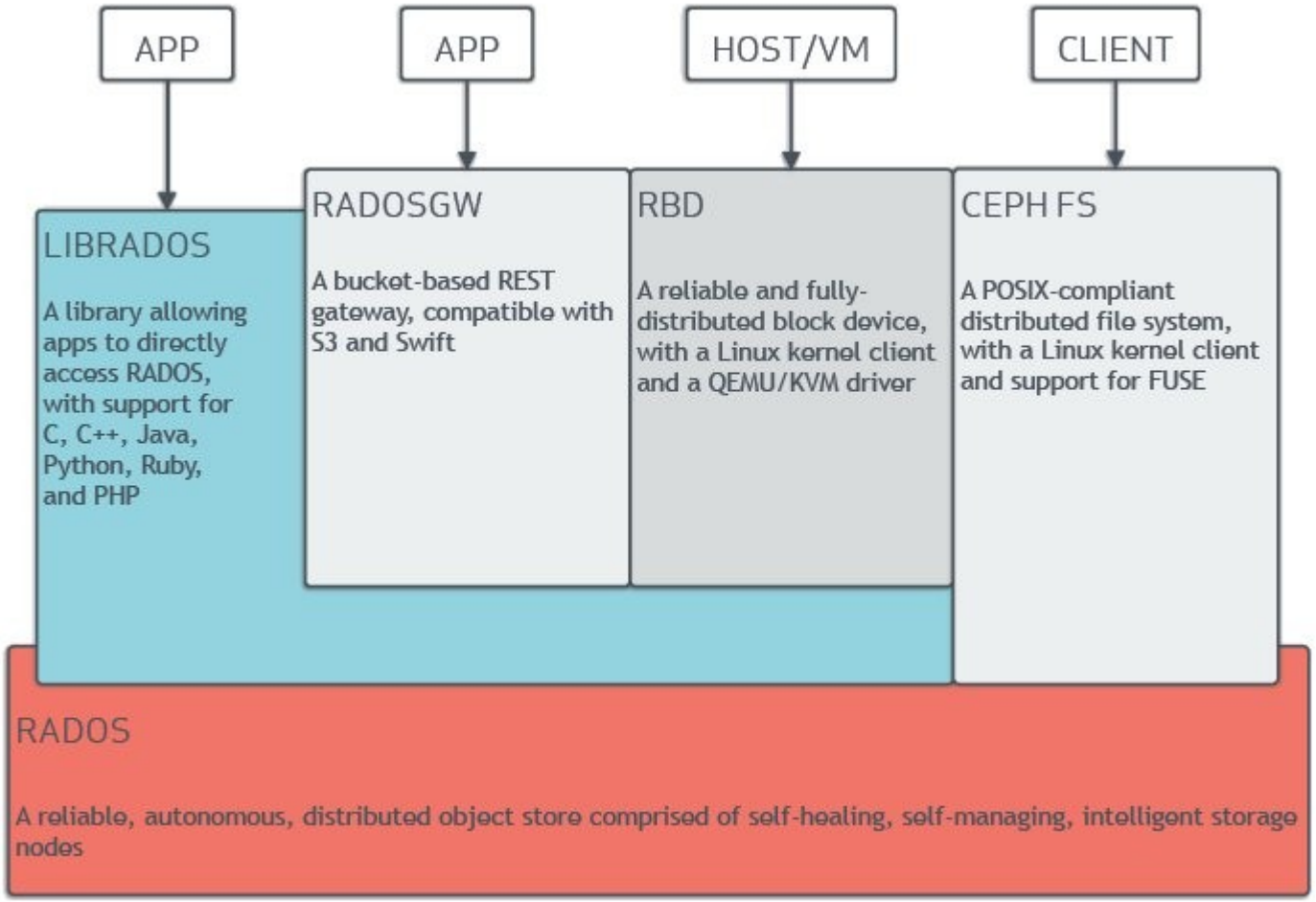
	Anzahl 2015/16
Schüler	2.471
Berufsoberschule	25
Berufsschule Auszubildende	1.590
Fachschule	47
Berufsfachschule mehrjährig	502
Fachoberschule	55
Berufliches Gymnasium	252
Mitarbeiter	136
Unterrichtsstunden	2.899,7 (96,5%)

Motivation

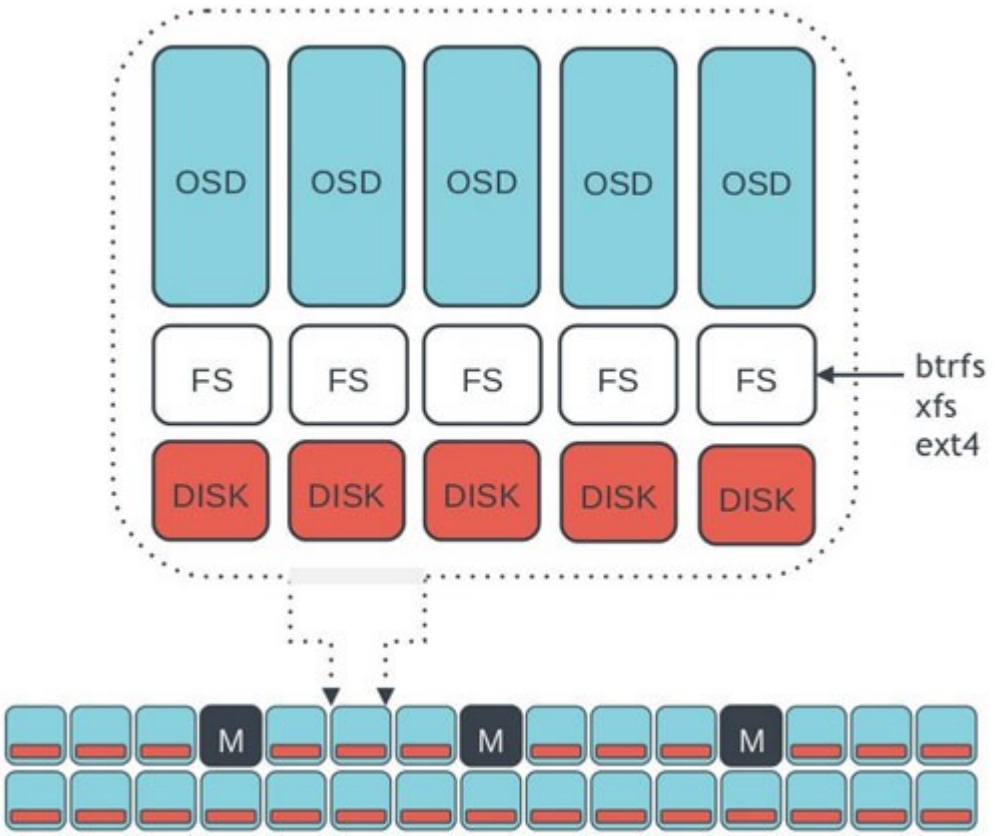
Warum RAID nicht immer Spaß macht

Ceph – Aufbau & Funktion

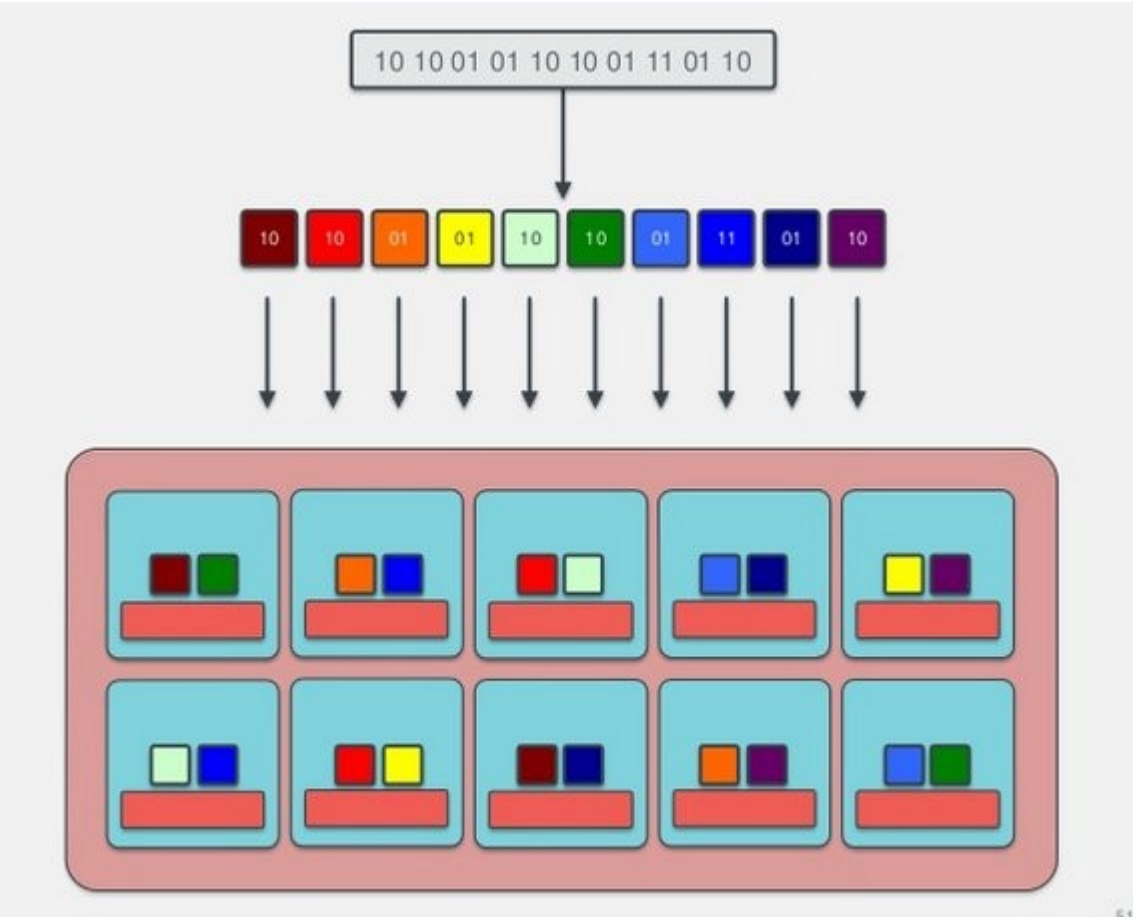
Ceph – Aufbau & Funktion



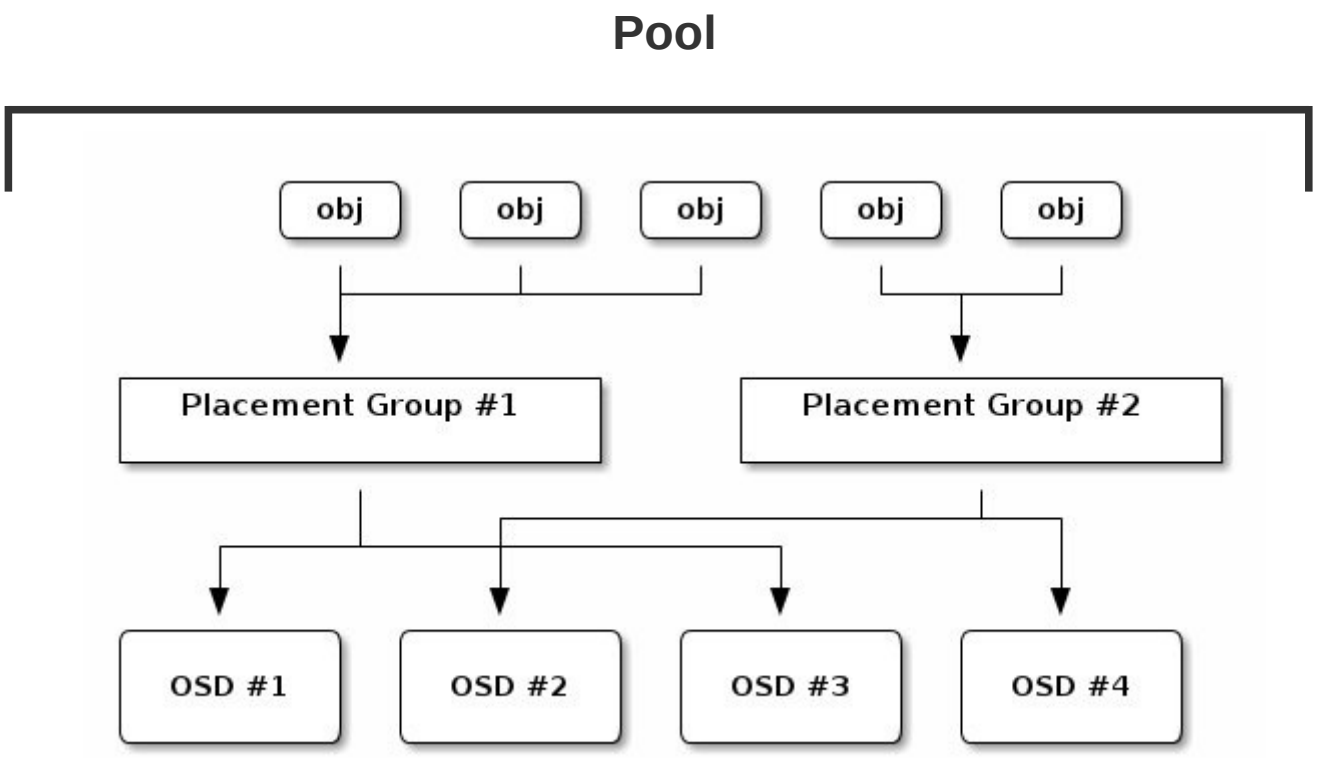
Ceph – Aufbau & Funktion



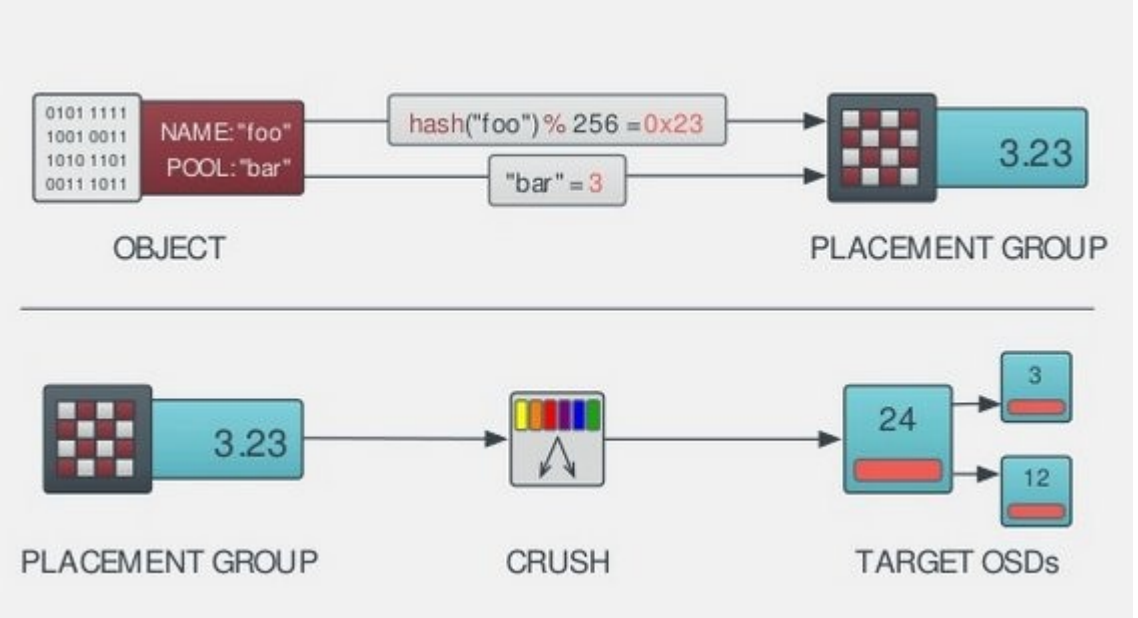
Ceph – Aufbau & Funktion



Ceph – Aufbau & Funktion



Ceph – Aufbau & Funktion

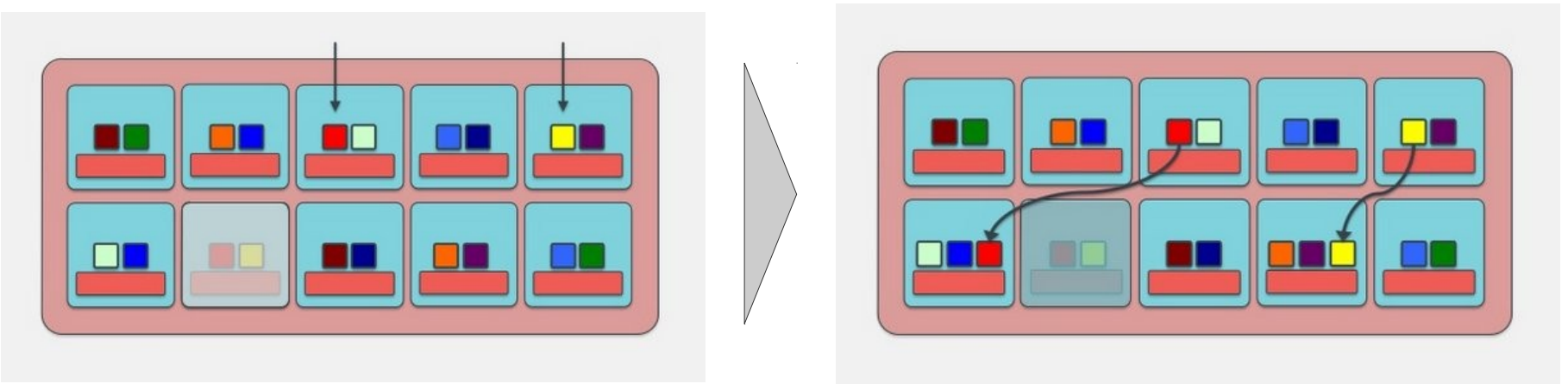


CRUSH(pg, cluster state, rule set)

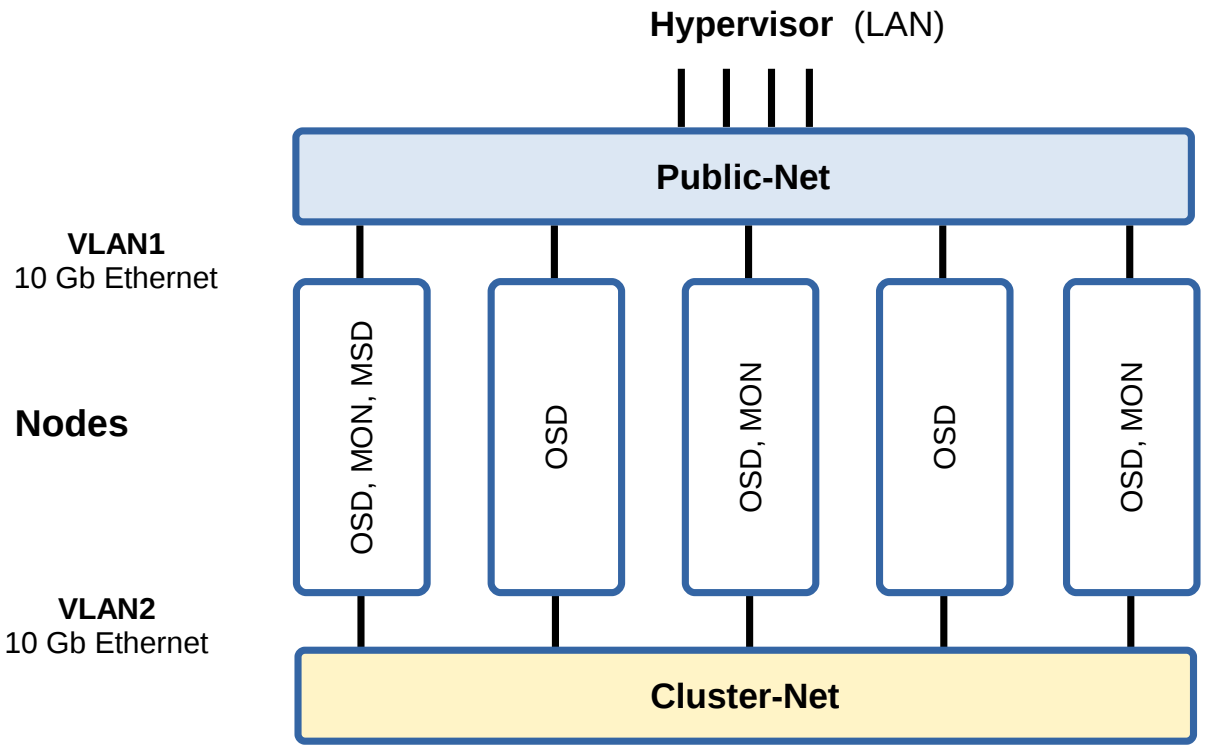
CRUSH: Controlled Replication Under Scalable Hashing

Ceph – Aufbau & Funktion

Rebalancing nach einem Ausfall einer OSD



Ceph – Aufbau & Funktion



Alternativ:
10 Gb **Ethernet**
MTU 9000 (Jumbo frames)

Alternativ:
Infiniband with IPoIB
DDR/QDR/FDR (10/40/56 Gbps)
MTU 65520 (connected mode)

Ceph – Aufbau & Funktion

- RADOS Gateway (RGW)

  Verbindung (API) zu anderen Speichersystemen (z.B. multi side)

- Erasure Coding (EC)

  'RAID' mit einstellbarer HDD-Anzahl (n = k + m) (r = k / n)

- Cache Tiering

  Hot-Cold-Pool in Kombination mit EC

- Ceph Client

  Mounten eines RADOS Block Devices (RBD)

- MDS

  CephFS und die Ausfallsicherheit

Ceph – Aufbau & Funktion

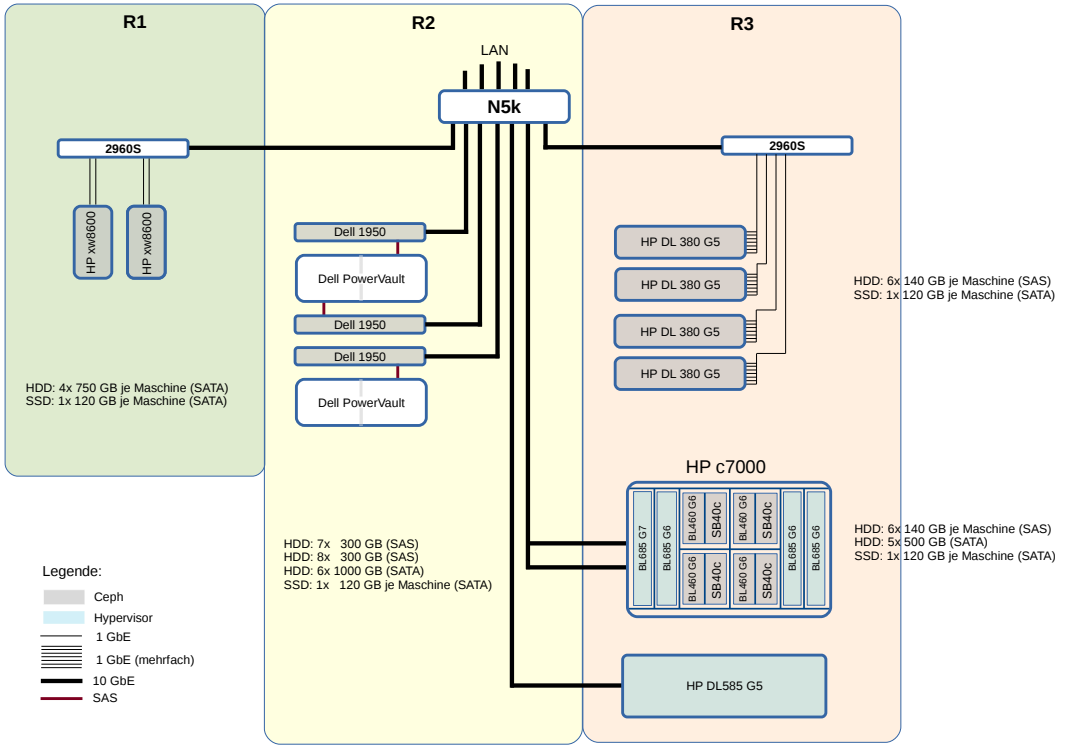
Process	Criteria	Minimum Recommended
ceph-osd	Processor	• 1x 64-bit AMD-64 • 1x 32-bit ARM dual-core or better • 1x i386 dual-core
	RAM	~1GB for 1TB of storage per daemon
	Volume Storage	1x storage drive per daemon
	Journal	1x SSD partition per daemon (optional)
	Network	2x 1GB Ethernet NICs
ceph-mon	Processor	• 1x 64-bit AMD-64/i386 • 1x 32-bit ARM dual-core or better • 1x i386 dual-core
	RAM	1 GB per daemon
	Disk Space	10 GB per daemon
	Network	2x 1GB Ethernet NICs
ceph-mds	Processor	• 1x 64-bit AMD-64 quad-core • 1x 32-bit ARM quad-core • 1x i386 quad-core
	RAM	1 GB minimum per daemon
	Disk Space	1 MB per daemon
	Network	2x 1GB Ethernet NICs

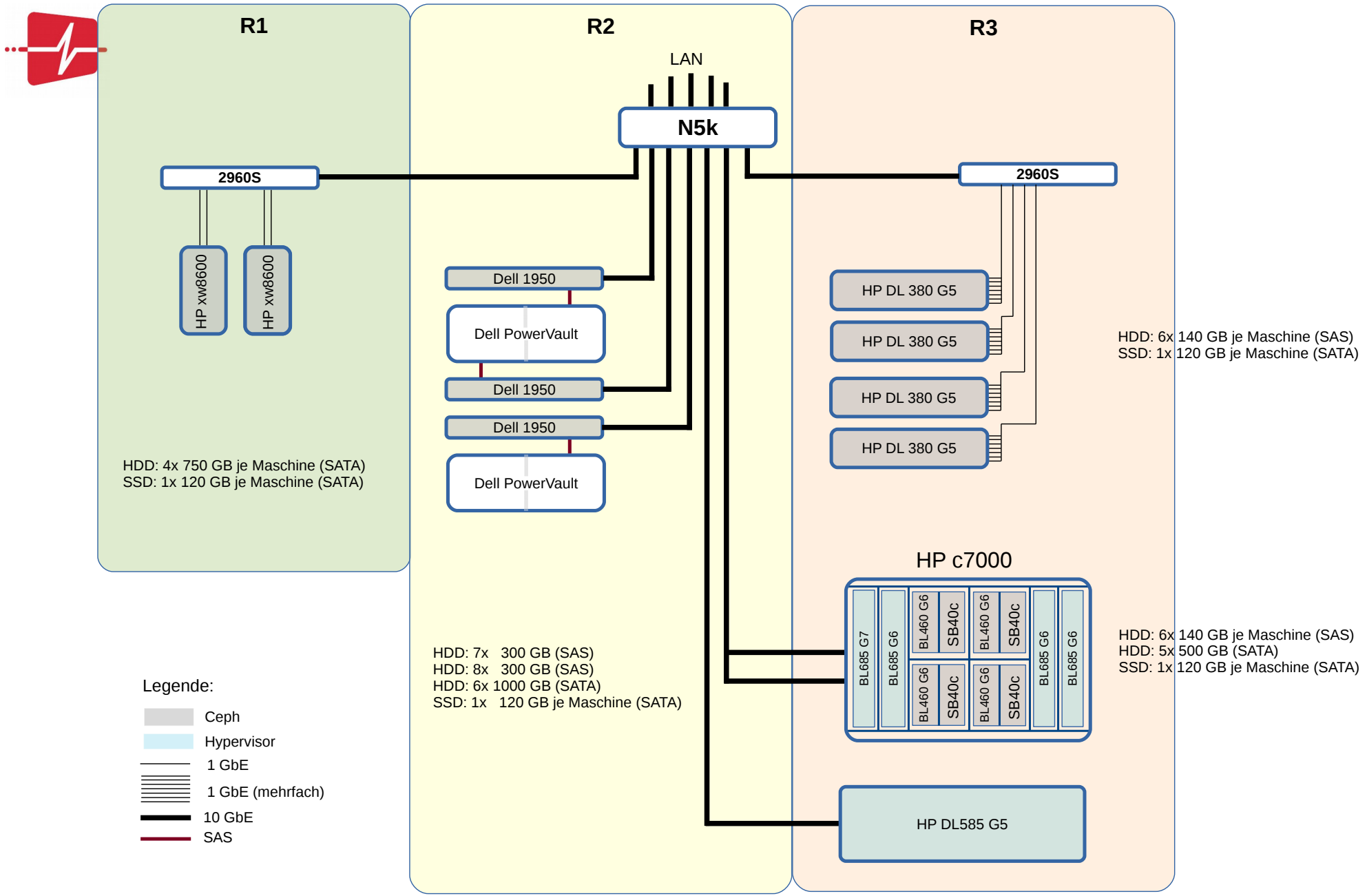
Anwendungsbeispiel – Ceph im OSZimt

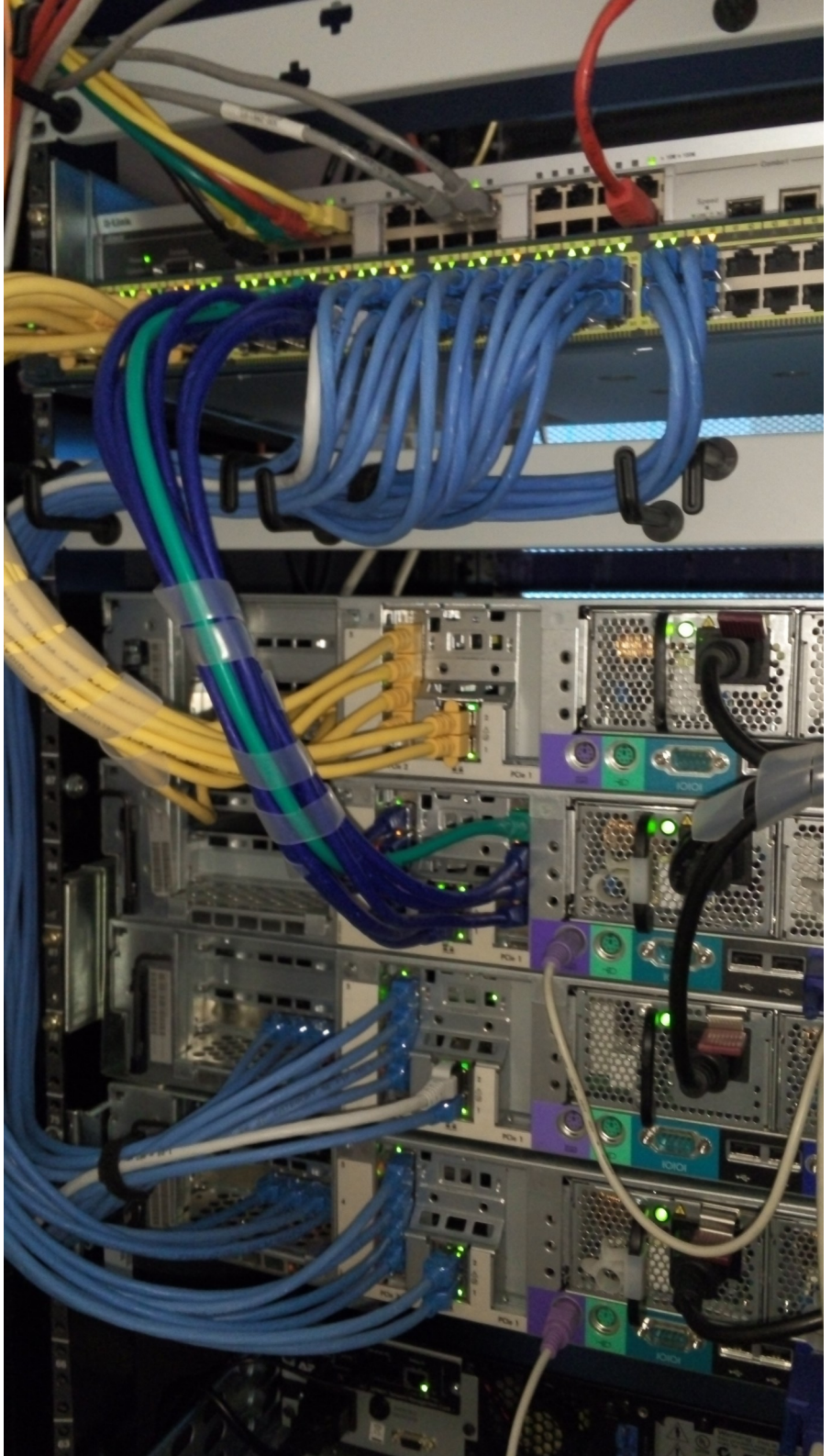
Ceph im OSZimt – Motivation

- OSZimt-Cloud = Proxmox + Ceph
- Cloud-Management
  Bestandteil der Ausbildung (FISI, ITA)
- VDI – alte Hardware mit neuster Software
- Service-VMs für den Fachunterricht (SQL, IDE)
- Storage für weitere Hypervisoren

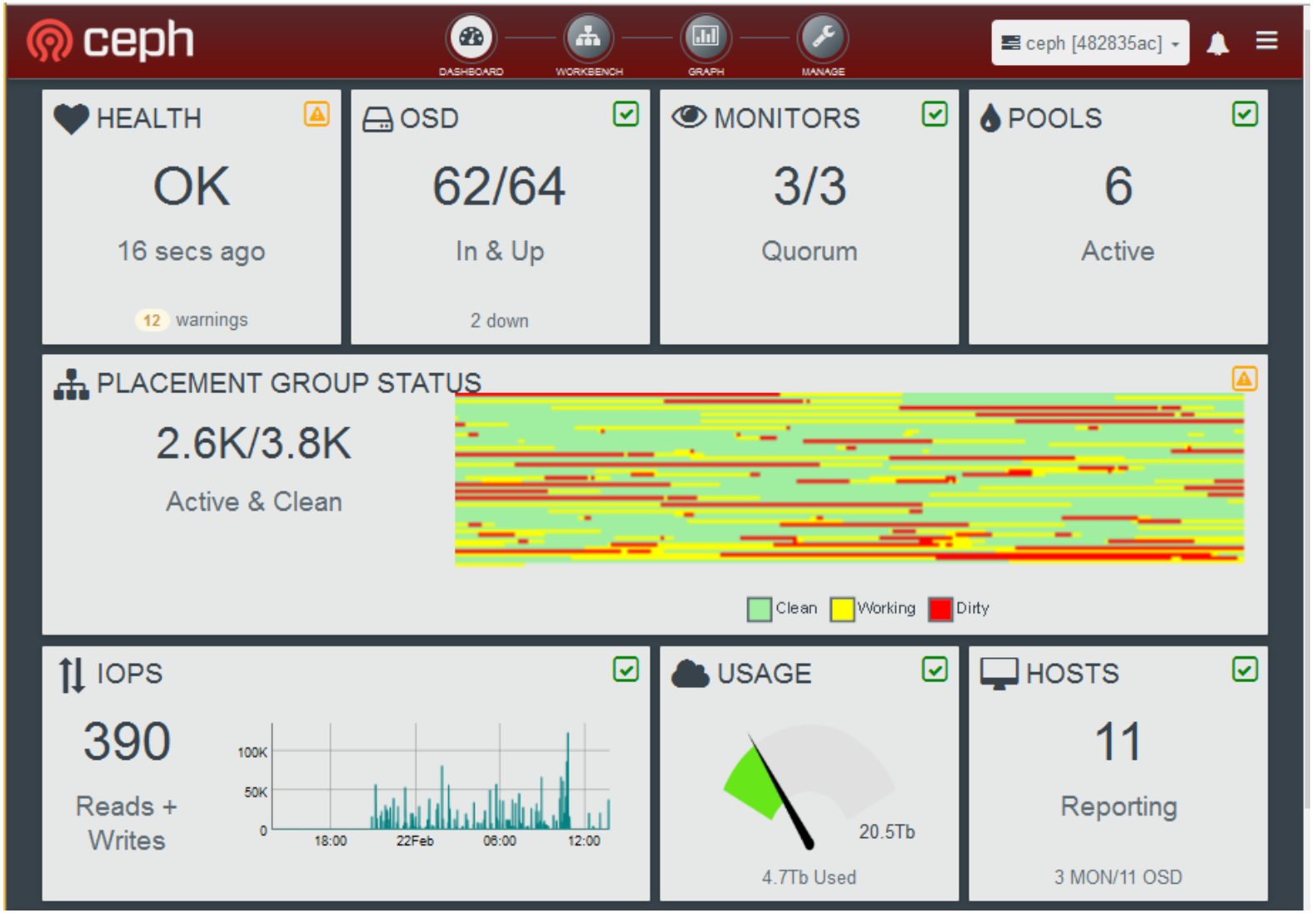
Ceph im OSZimt – Aufbau



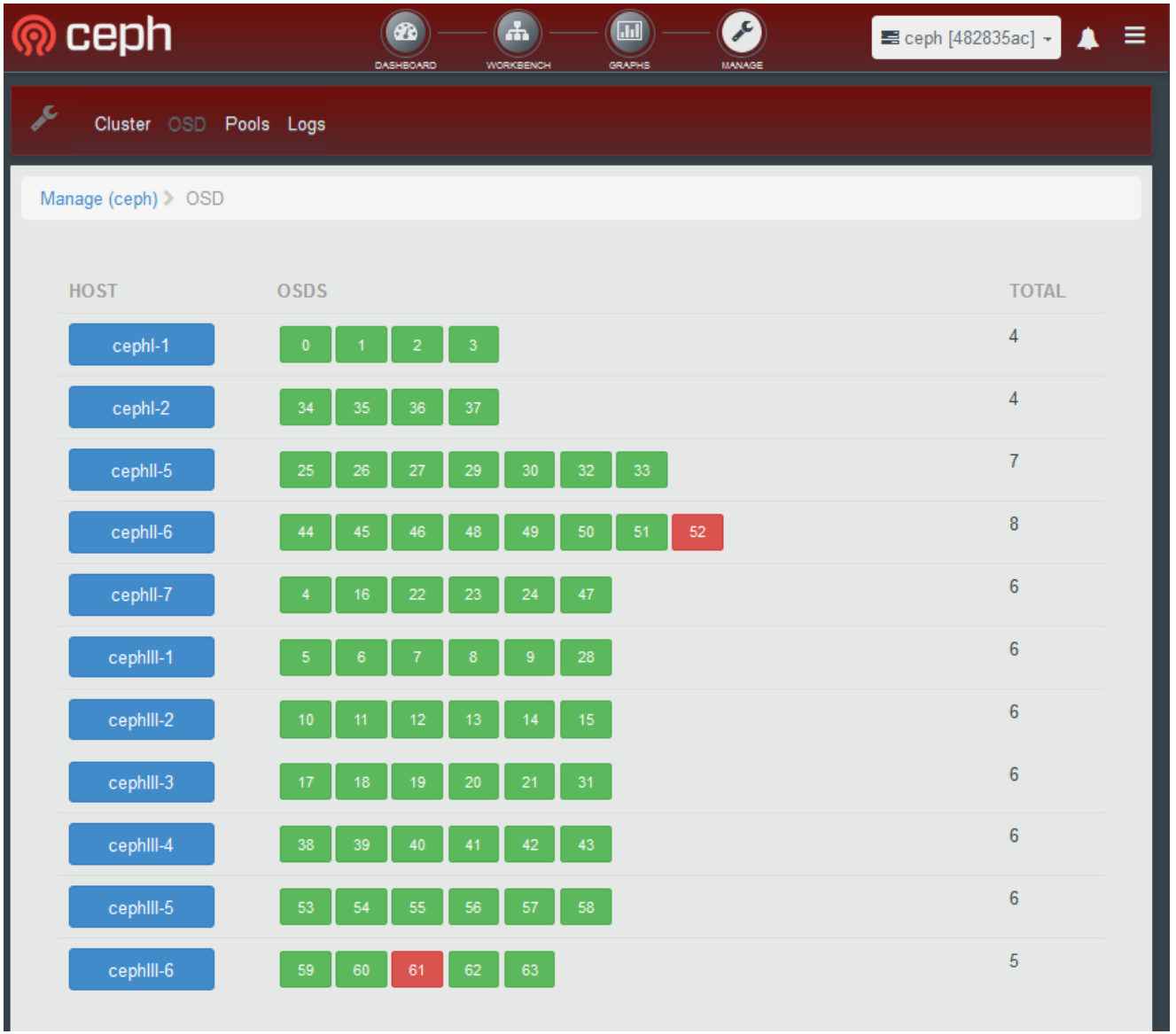

R1
R2
R3
LAN
N5k
2960S
2960S
HP xw8600
HP xw8600
Dell 1950
Dell PowerVault
Dell 1950
Dell 1950
Dell PowerVault
HP DL 380 G5
HP DL 380 G5
HP DL 380 G5
HP DL 380 G5
HDD: 6x 140 GB je Maschine (SAS)
SSD: 1x 120 GB je Maschine (SATA)
HDD: 4x 750 GB je Maschine (SATA)
SSD: 1x 120 GB je Maschine (SATA)
HP c7000
BL685 G7
BL685 G6
BL460 G6
SB40c
BL460 G6
SB40c
BL460 G6
SB40c
BL460 G6
SB40c
BL685 G6
BL685 G6
HDD: 6x 140 GB je Maschine (SAS)
HDD: 5x 500 GB (SATA)
SSD: 1x 120 GB je Maschine (SATA)
HDD: 7x 300 GB (SAS)
HDD: 8x 300 GB (SAS)
HDD: 6x 1000 GB (SATA)
SSD: 1x 120 GB je Maschine (SATA)
HP DL585 G5
Legende:
Ceph
Hypervisor
1 GbE
1 GbE (mehrfach)
10 GbE
SAS



Ceph im OSZimt – Calamari

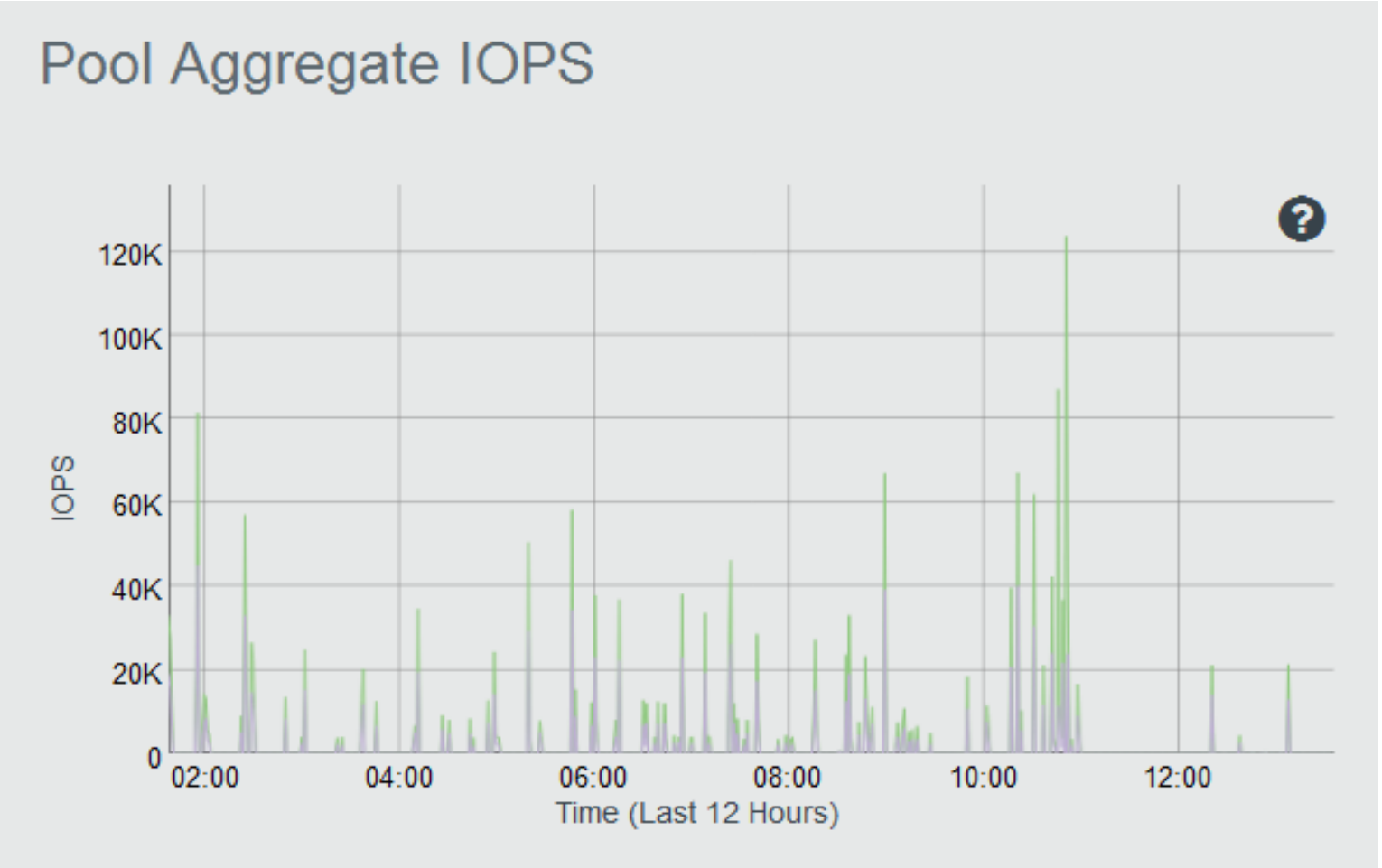


Ceph im OSZimt – Calamari

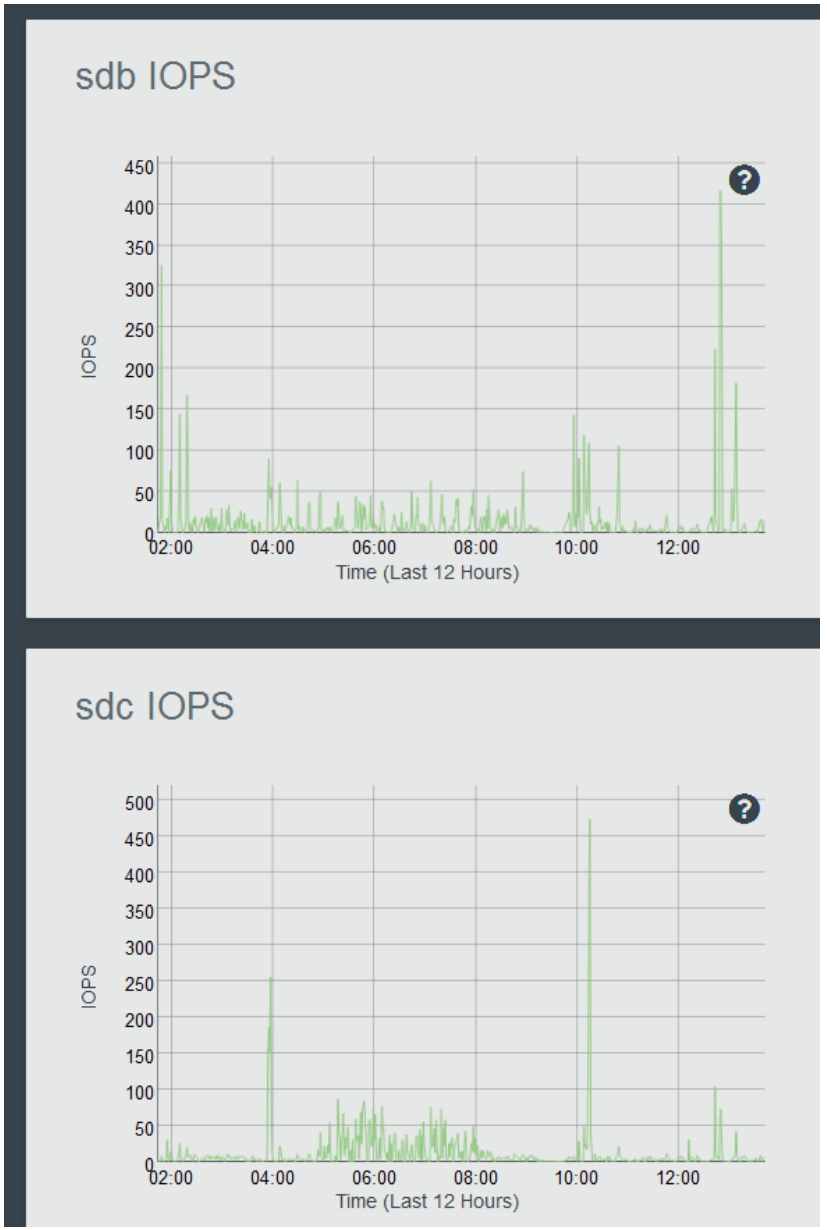


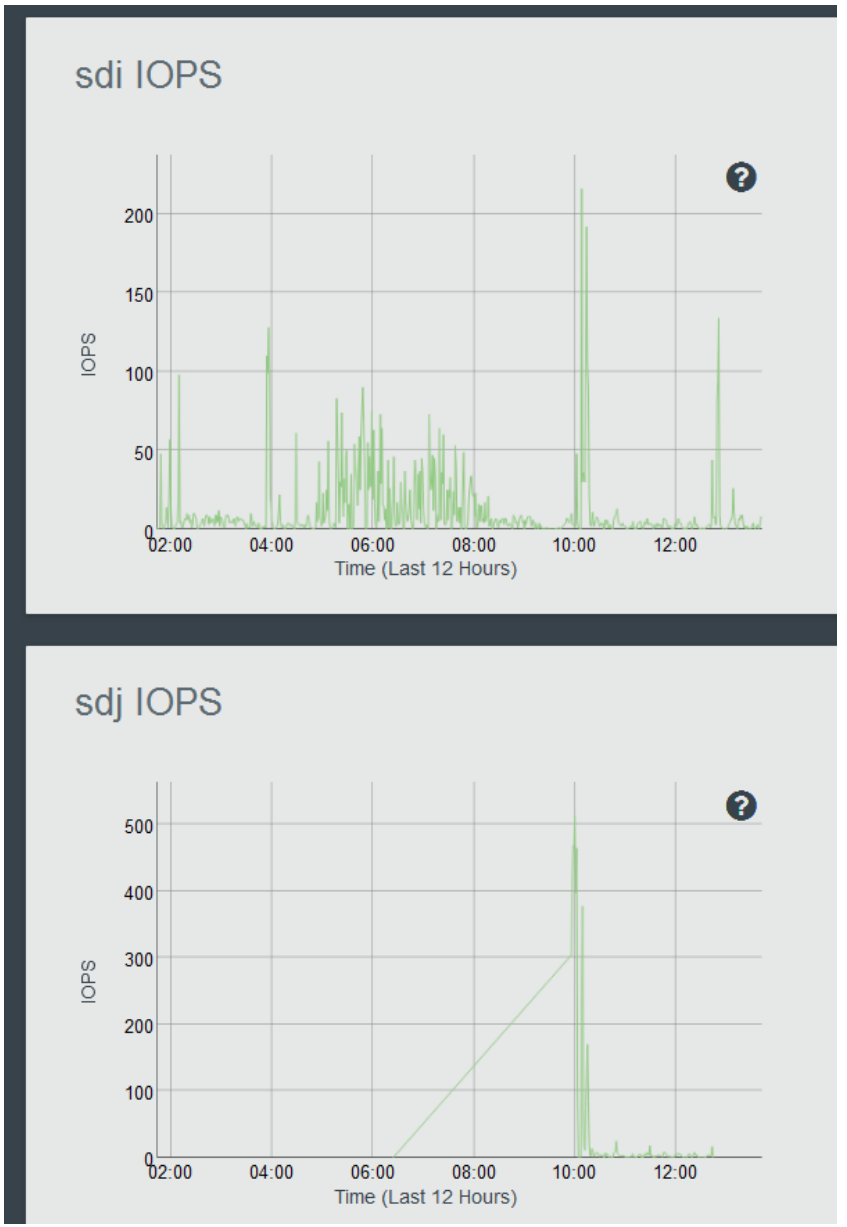
Ceph im OSZimt – Calamari

Ceph im OSZimt – Calamari



Ceph im OSZimt – Calamari

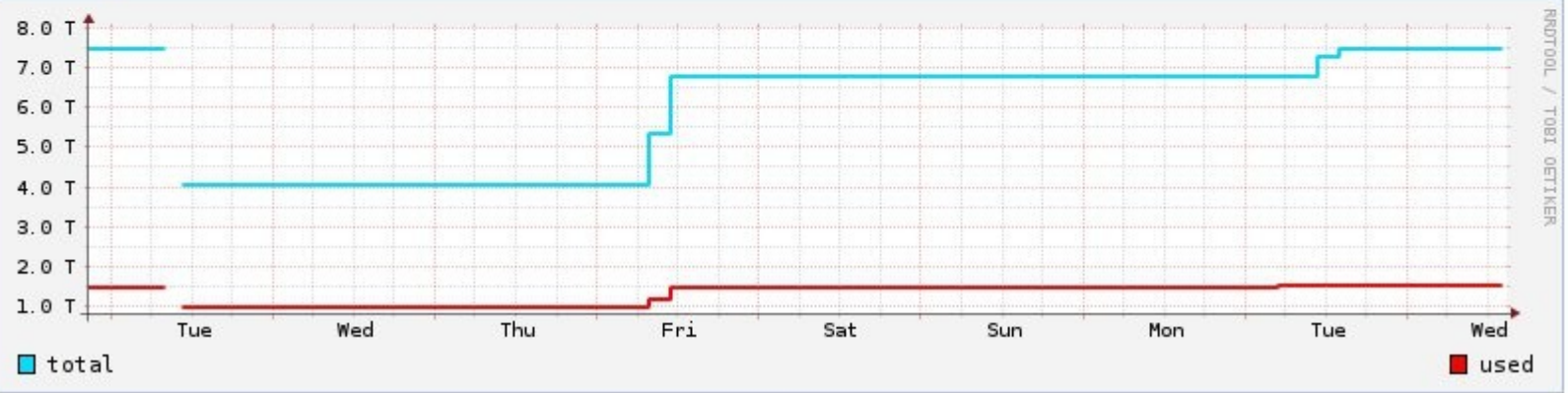




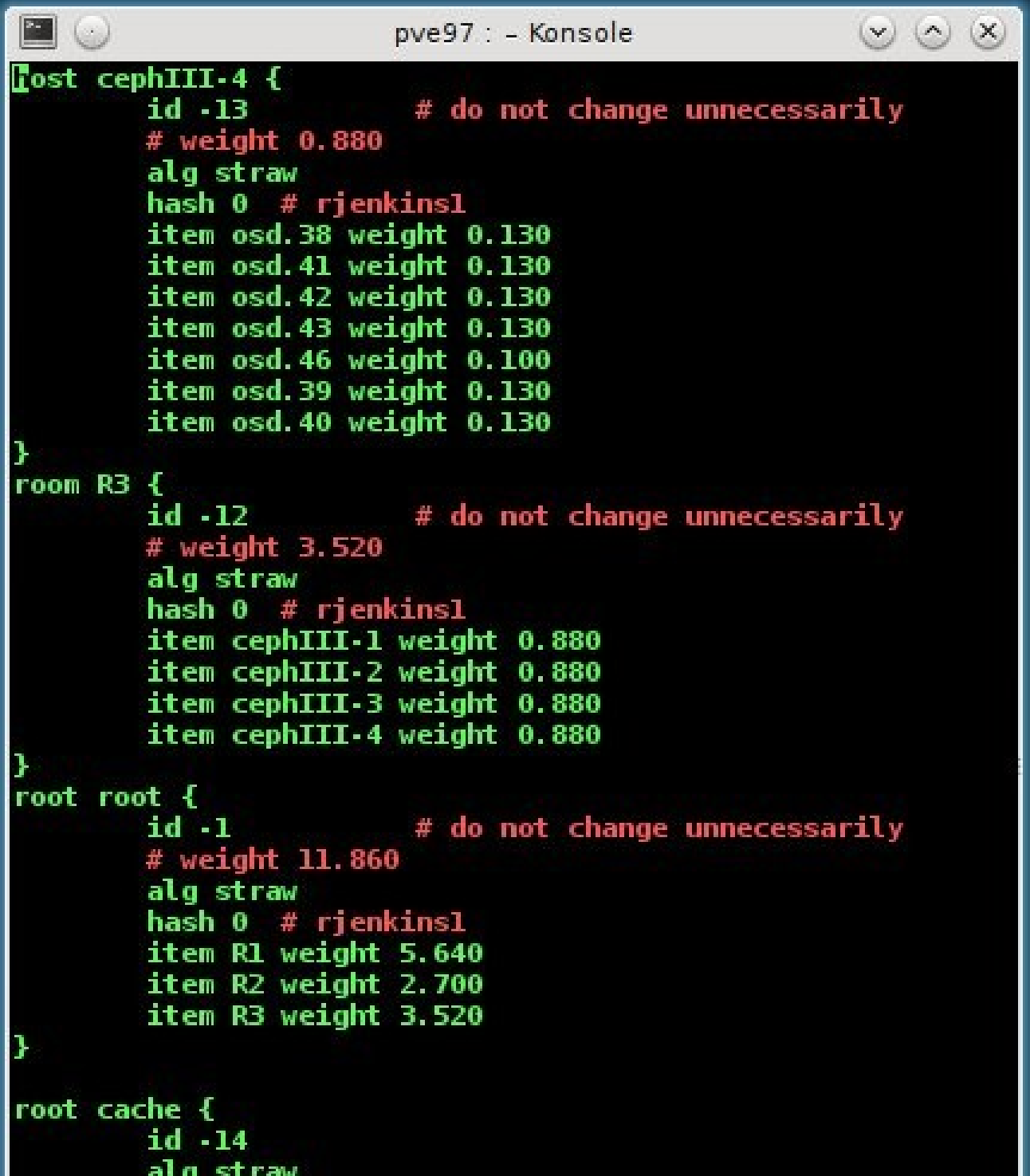
Ceph im OSZimt – Anwendung

Status	
Freigegeben	Ja
Aktiv	Ja
Inhalt	Disk-Image
Typ	RBD
Verteilt	Ja
Größe	6.80TB
Verwendet	1.41TB
Verfügbar	5.39TB

Auslastung



```
-14      0.5999  root cache
4        0.09999         osd.4   up      1
10       0.09999         osd.10  up      1
16       0.09999         osd.16  up      1
44       0.09999         osd.44  up      1
45       0.09999         osd.45  up      1
46       0.09999         osd.46  up      1
-1       12.13   root root
-10      5.64            room R1
-2       2.82                    host cephI-1
0        0.68                            osd.0   up      1
1        0.68                            osd.1   up      1
2        0.68                            osd.2   up      1
3        0.68                            osd.3   up      1
44       0.09999                         osd.44  up      1
-9       2.82                    host cephI-2
34       0.68                            osd.34  up      1
35       0.68                            osd.35  up      1
36       0.68                            osd.36  up      1
37       0.68                            osd.37  up      1
45       0.09999                         osd.45  up      1
-11      2.97            room R2
-6       1.08                    host cephII-1
24       0.27                            osd.24  up      1
25       0.27                            osd.25  up      1
22       0.27                            osd.22  up      1
23       0.27                            osd.23  up      1
-7       0.81                    host cephII-2
26       0.27                            osd.26  up      1
27       0.27                            osd.27  up      1
29       0.27                            osd.29  up      1
-8       1.08                    host cephII-3
32       0.27                            osd.32  up      1
33       0.27                            osd.33  up      1
48       0.27                            osd.48  up      1
30       0.27                            osd.30  down    0
-12      3.52            room R3
-3       0.88                    host cephIII-1
4        0.09999                         osd.4   up      1
5        0.13                            osd.5   up      1
```


pve97 : – Konsole
host cephIII-4 {
id -13 # do not change unnecessarily
weight 0.880
alg straw
hash 0 # rjenkins1
item osd.38 weight 0.130
item osd.41 weight 0.130
item osd.42 weight 0.130
item osd.43 weight 0.130
item osd.46 weight 0.100
item osd.39 weight 0.130
item osd.40 weight 0.130
}
room R3 {
id -12 # do not change unnecessarily
weight 3.520
alg straw
hash 0 # rjenkins1
item cephIII-1 weight 0.880
item cephIII-2 weight 0.880
item cephIII-3 weight 0.880
item cephIII-4 weight 0.880
}
root root {
id -1 # do not change unnecessarily
weight 11.860
alg straw
hash 0 # rjenkins1
item R1 weight 5.640
item R2 weight 2.700
item R3 weight 3.520
}
root cache {
id -14


# rules
rule data {
ruleset 0
type replicated
min_size 1
max_size 10
step take root
step chooseleaf firstn 0 type room
step emit
}
rule metadata {
ruleset 1
type replicated
min_size 1
max_size 10
step take root
step chooseleaf firstn 0 type room
step emit
}
rule rbd {
ruleset 2
type replicated
min_size 1
max_size 10
step take root
step chooseleaf firstn 0 type room
step emit
}
rule cache-pool {
ruleset 3
type replicated
min_size 1
max_size 10
step take cache
step chooseleaf firstn 0 type osd
step emit
}
# end crush map
251,9 Ende

Anwendungsbeispiel – Ceph im Krankenhaus

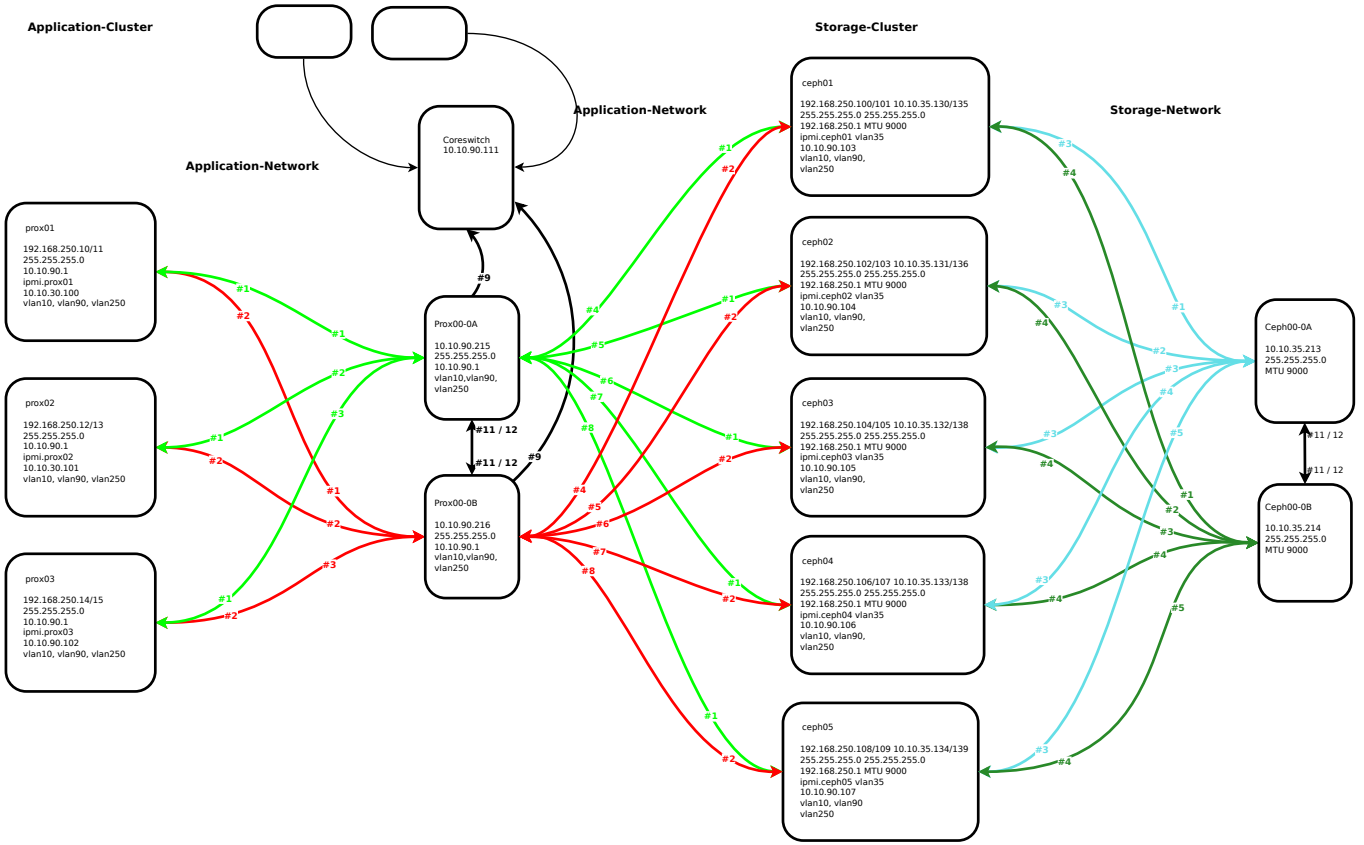
Diakoniewerk Bethel

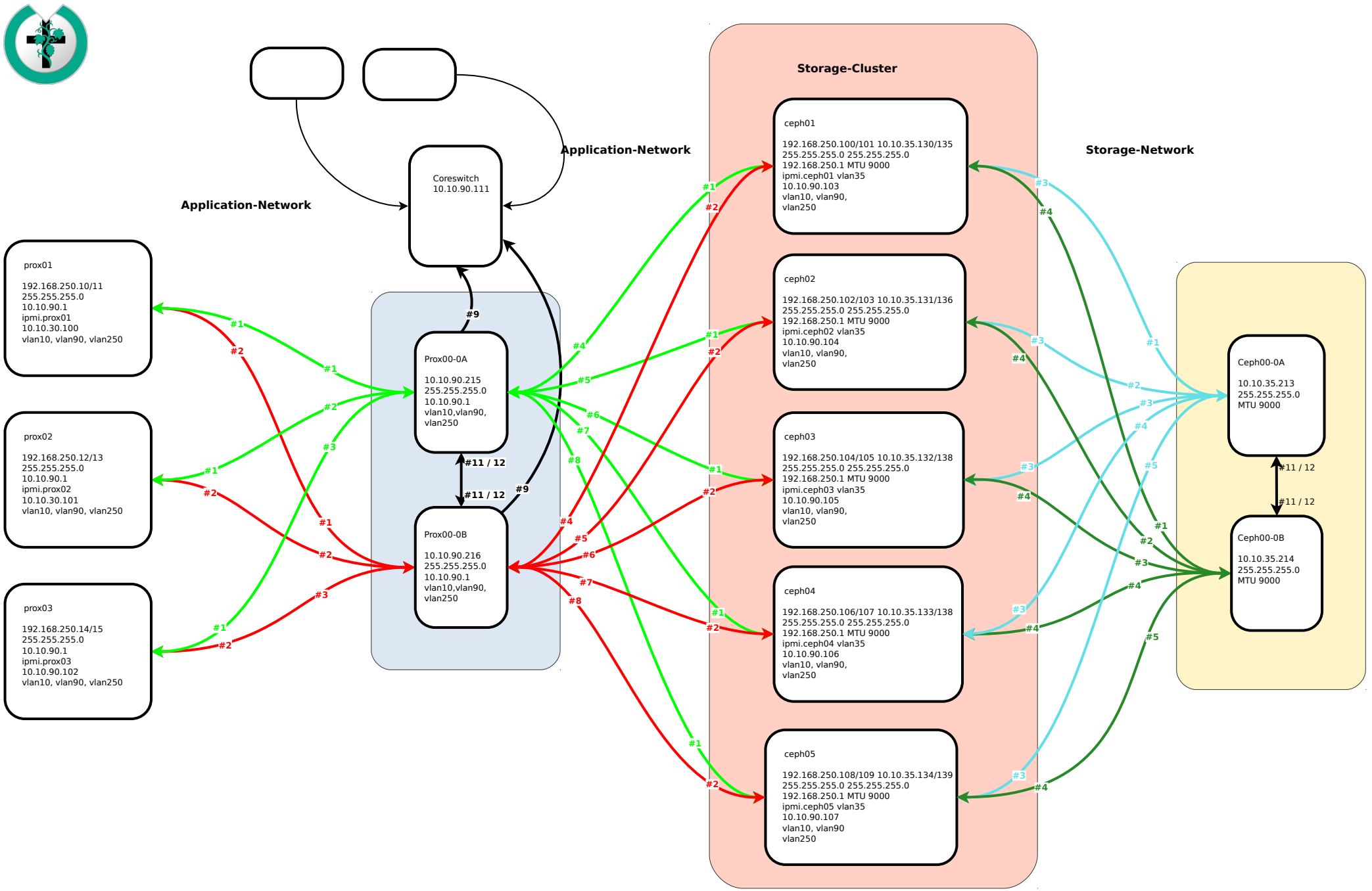
- *Krankenhaus* in Berlin (276 Betten), *geriatrischen Reha-Kliniken* in Welzheim und Trossingen
- *Seniorenwohnungen* und *stationäre Pflege* in Berlin, Bad Oeynhausen, Wiehl, Welzheim, Trossingen und München
- *Pflegedienste* in Berlin, Bad Oeynhausen, Welzheim

Ceph im Krankenhaus – Motivation

- Einführung eines neuen DMS
- HW musste ausgetauscht werden
- Angebot VMware inkl. Storage ~150 k€
- VMware mit hoher HW-Anforderung (1 SSD pro 6 HDDs)
- Ausfallsicherheit konnte nicht gewährleistet werden
- Ceph-Demo überzeugte,
  gleichzeitiger Ausfall mehrerer OSDs (Node)
- Testaufbau Hypervisor und Ceph bewies
  die Ausfallsicherheit der Anwendungen
- guter Erfahrungsaustausch mit dem OSZimt

Ceph im Krankenhaus – Aufbau




Storage-Cluster
Application-Network
Application-Network
Storage-Network
Coreswitch
10.10.90.111
prox01
192.168.250.10/11
255.255.255.0
10.10.90.1
ipmi.prox01
10.10.30.100
vlan10, vlan90, vlan250
prox02
192.168.250.12/13
255.255.255.0
10.10.90.1
ipmi.prox02
10.10.30.101
vlan10, vlan90, vlan250
prox03
192.168.250.14/15
255.255.255.0
10.10.90.1
ipmi.prox03
10.10.90.102
vlan10, vlan90, vlan250
Prox00-0A
10.10.90.215
255.255.255.0
10.10.90.1
vlan10,vlan90,
vlan250
Prox00-0B
10.10.90.216
255.255.255.0
10.10.90.1
vlan10,vlan90,
vlan250
#11 / 12
#11 / 12
#9
#9
ceph01
192.168.250.100/101 10.10.35.130/135
255.255.255.0 255.255.255.0
192.168.250.1 MTU 9000
ipmi.ceph01 vlan35
10.10.90.103
vlan10, vlan90,
vlan250
ceph02
192.168.250.102/103 10.10.35.131/136
255.255.255.0 255.255.255.0
192.168.250.1 MTU 9000
ipmi.ceph02 vlan35
10.10.90.104
vlan10, vlan90,
vlan250
ceph03
192.168.250.104/105 10.10.35.132/138
255.255.255.0 255.255.255.0
192.168.250.1 MTU 9000
ipmi.ceph03 vlan35
10.10.90.105
vlan10, vlan90,
vlan250
ceph04
192.168.250.106/107 10.10.35.133/138
255.255.255.0 255.255.255.0
192.168.250.1 MTU 9000
ipmi.ceph04 vlan35
10.10.90.106
vlan10, vlan90,
vlan250
ceph05
192.168.250.108/109 10.10.35.134/139
255.255.255.0 255.255.255.0
192.168.250.1 MTU 9000
ipmi.ceph05 vlan35
10.10.90.107
vlan10, vlan90
vlan250
Ceph00-0A
10.10.35.213
255.255.255.0
MTU 9000
Ceph00-0B
10.10.35.214
255.255.255.0
MTU 9000
#11 / 12
#11 / 12

Ceph im Krankenhaus – Ausstattung

- 5 x **Ceph-Node**
  - MoBo: Supermicro SM-X10DRC-T4+ (4 x 10 Gigabit LAN - onboard)
  - CPU: XE5-2623V3T - 3 GHz - 4-Core
  - RAM: 64 GB
  - HDDs: 6 x 900 GB SAS 10k (OSD)
6 x 2000 GB SATA 128 MB Cache (OSD)
  - SSD: 1 x Intel SSD DC P3600, 1.2TB, PCIe, NVMe 3.0 x4 (OSD)

- 3 x **Proxmox-Node**
  - MoBo: Supermicro SM-X10DRI-T4+ (4x 10GBase-T LAN - onboard)
  - CPU: XE5-2630V3 8-Core
  - RAM: 128 GB

Ceph im Krankenhaus – Ausstattung

Ceph im Krankenhaus – Anwendung

- Storage-Volumen 95 TB
- Replica 3
- 3 Pools (SSD, SAS, SATA)
- 10 Gbps Cu, Cluster-/Public-Netz getrennt
- DMS Testbetrieb seit 09/2015
  Fachanwendung (MediFox) produktiv seit 09/2015
- Performance liegt weit über den Erwartungen,
  Netzwerk saturiert (SSD-/SAS-Pool gleiche Performance)
- Projektkosten ca. 95 k€
  (Hypervisor, Storage, Consulting, Setup)

Ceph versus RAID

- Anzahl der möglichen, gleichzeitigen HD-Ausfälle (kann > 2 sein)
- Dauer eines Rebuilds/Rebalancing (⇘)
- Anzahl der möglichen Replikate (1-10)
- Scale out (jederzeit möglich)

→ es gehen keine Daten verloren

→ ich schlafe ruhiger ☺

→ Backup ist weiterhin erforderlich

Hinweise & Tipps

- SPoF vermeiden
- commodity HW ist OK, besser ist aktuelle HW
- Dienste auf separaten Nodes betreiben (z.B. Hypervisor, OSDs)
- 1 HDD = 1 OSD (OS auf extra HW), 1GB RAM pro 1TB OSD
- RAID-Cache nutzen (JBOD hat meist keinen)
- viele 'kleinere' HDDs sind sinnvoller, als wenige große
- je mehr Nodes desto besser (im Fehlerfall den Füllgrad beachten)

Hinweise & Tipps

- Public-Net vom Cluster-Net trennen
- max. Netzwerkgeschwindigkeit bei schnellen OSDs aktivieren
  ≥10 Gbps Ethernet, Infiniband, MTU (Jumbo frames, connected mode)
- Journale auf SSD schreiben, Partition alignment
- SSD-HDD-Verhältnis (Schreib-/Lesegeschwindigkeit)

Fazit

- Ceph ist nicht für alle oder alles gut
- fast 'unkaputtbar'
- unterschiedliche HW kann gleichzeitig genutzt werden
- mindest HW-Anforderung berücksichtigen, Dimensionierungsgrösse ist der worst case (Rebalancing)
- das eigene know-how kann mit dem Ceph-Cluster wachsen

In eigener Sache

Wer unser Haus und unsere Arbeit unterstützen möchte

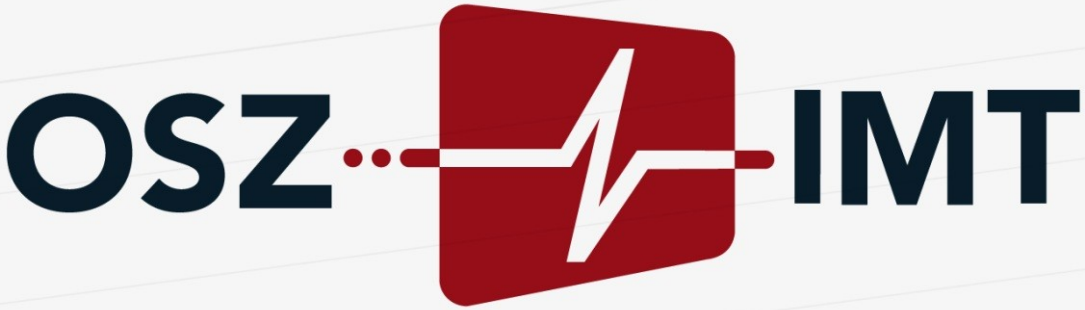
kann dies gerne über unseren Förderverein VeFF e.V. tun.

http://www.oszimt.de/de/ueber-uns/organisation/foerderverein.html

Hilfe (monitärer Art, Hardware, Hands on) ist immer willkommen.

Danke

VeFF e.V.:
Verein der Freunde und Förderer des Oberstufenzentrums Informations- und Medizintechnik Berlin e.V.



Christian Schubert

Schubert@OSZimt.de