Ceph – Software Defined Storage für die Cloud

CeBIT 2016

15. 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 70 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
- 80%+ basieren auf FC

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?

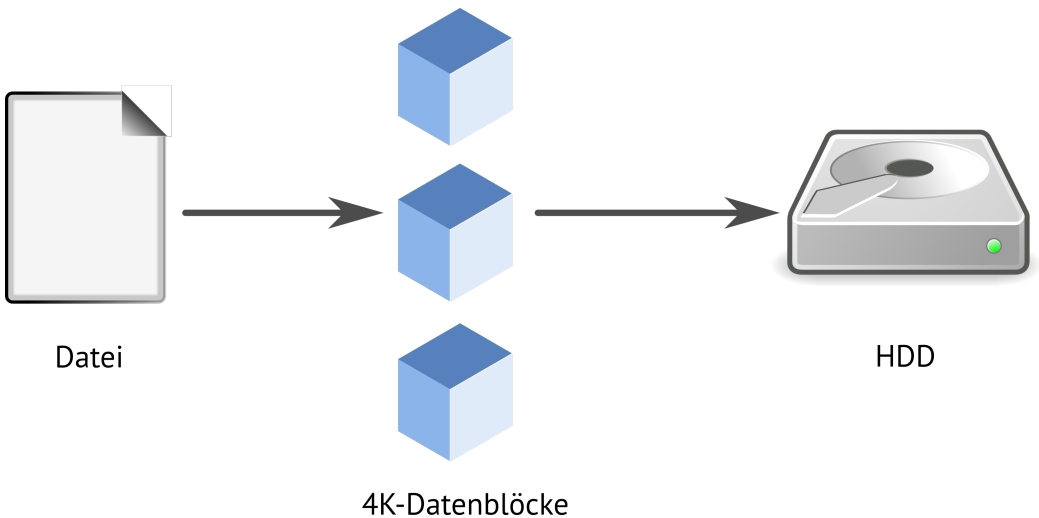
Was ist Ceph?

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

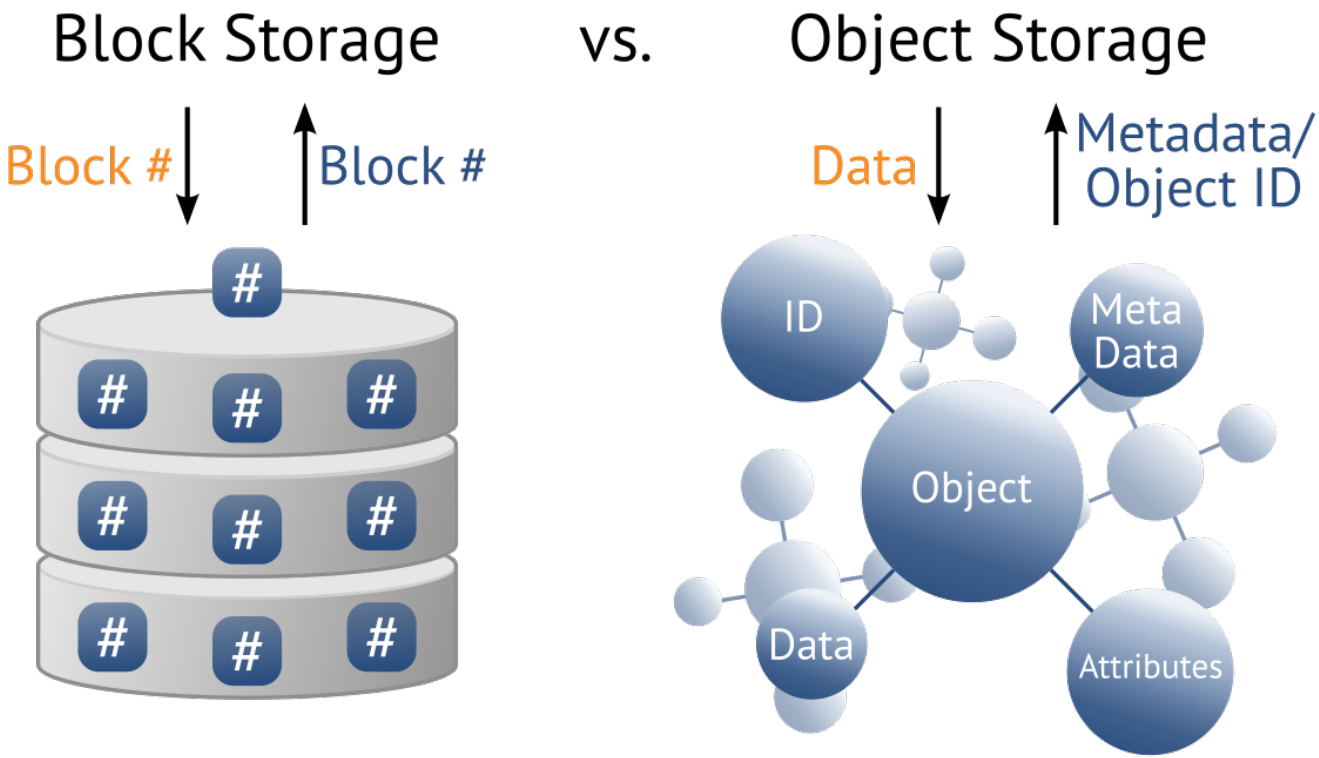
Block Storage

- 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

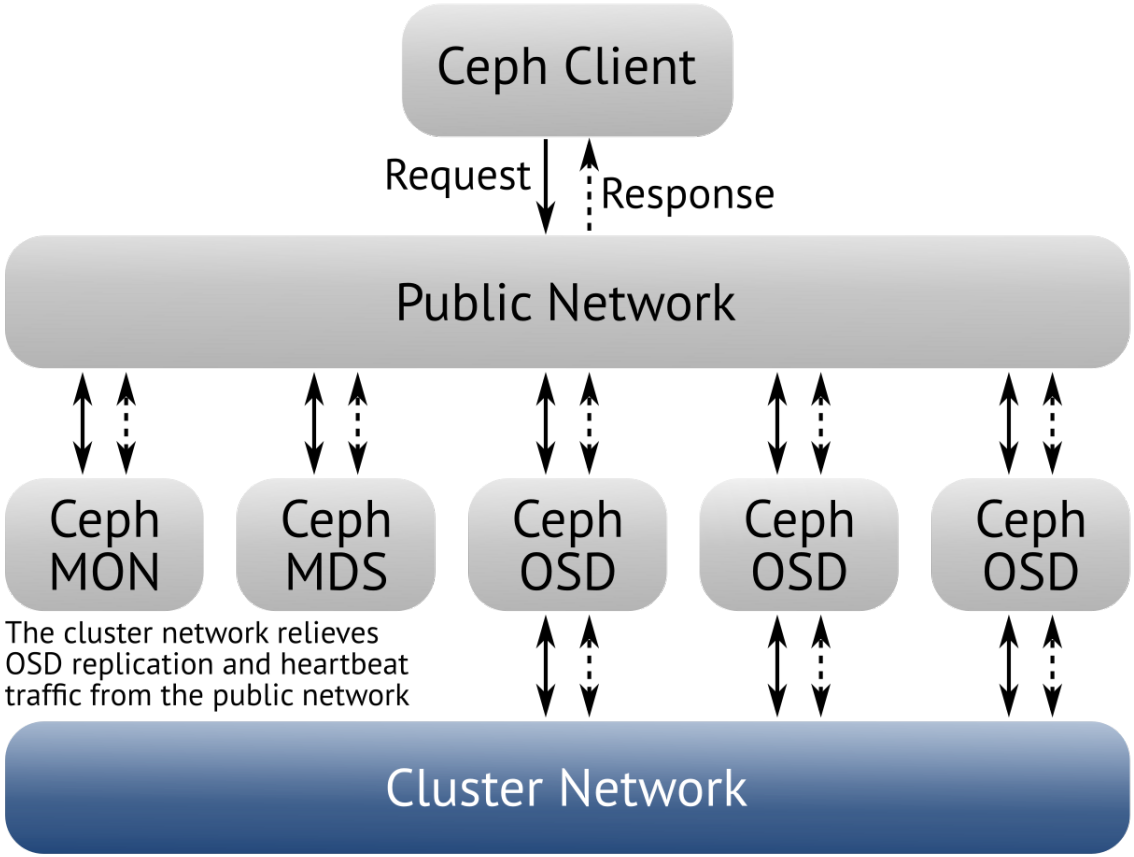
- RESTful API
- Interface für
  - OpenStack Swift
  - Amazon S3

Aufbau von Ceph

Aufbau von Ceph

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

Aufbau

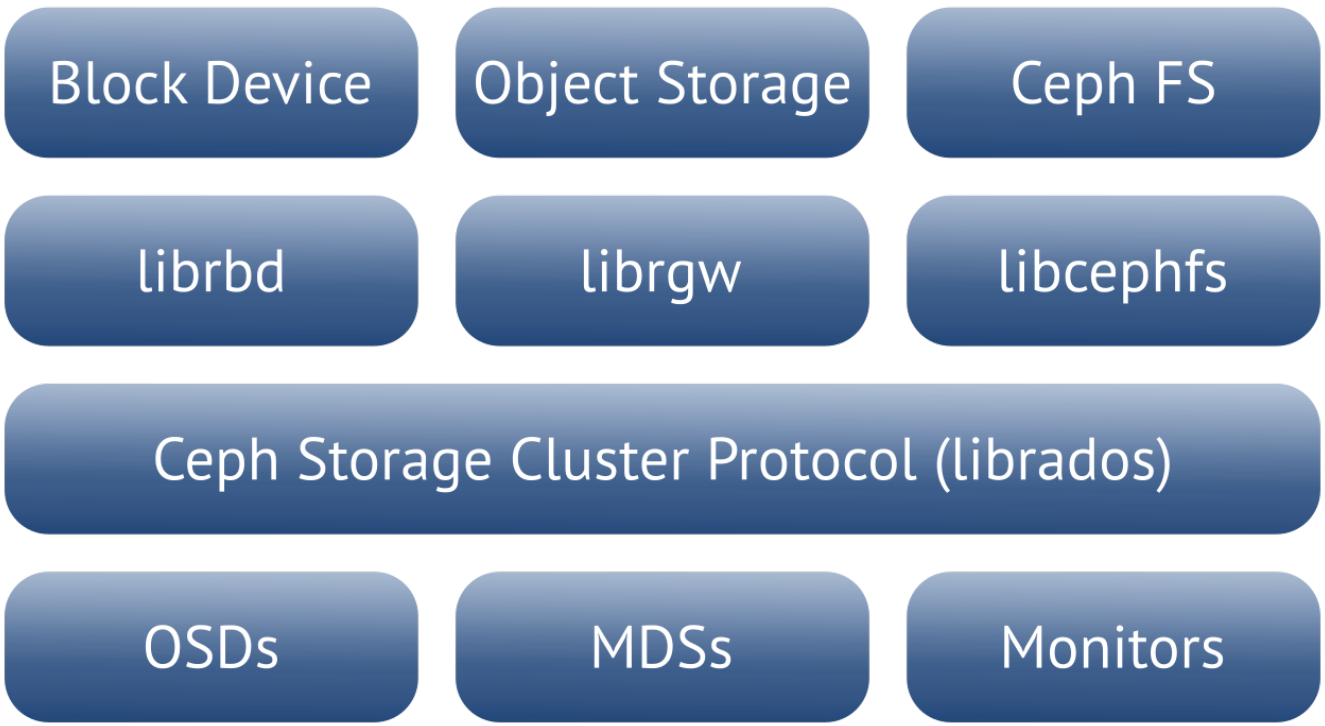

Ceph Client
Request
Response
Public Network
Ceph MON
Ceph MDS
Ceph OSD
Ceph OSD
Ceph OSD
The cluster network relieves OSD replication and heartbeat traffic from the public network
Cluster Network

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)

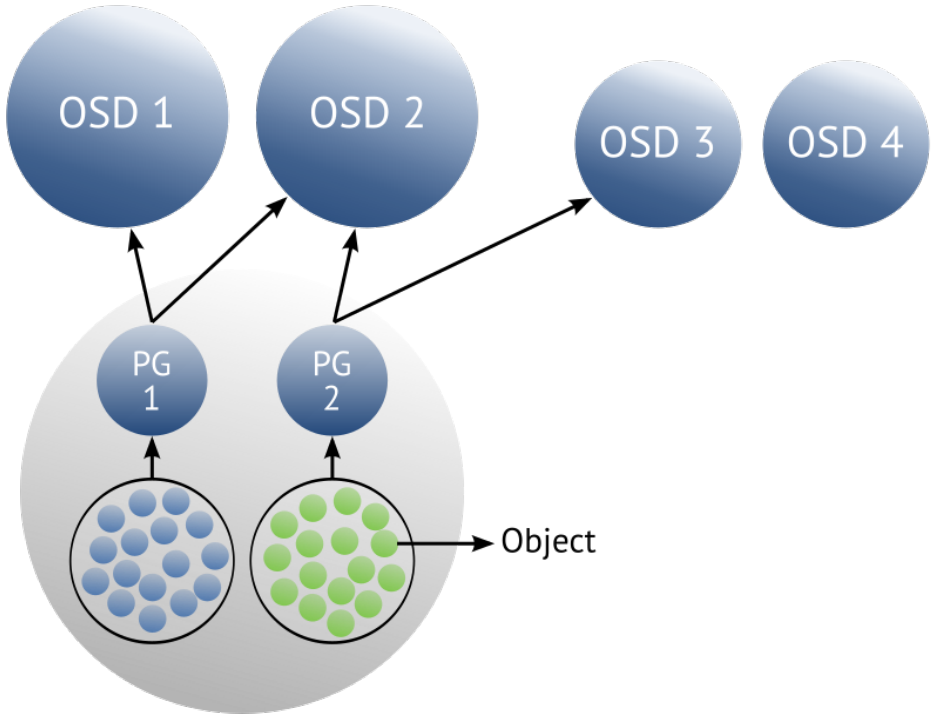
ceph-osd – Ceph Object Storage Daemon 4/4

- 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



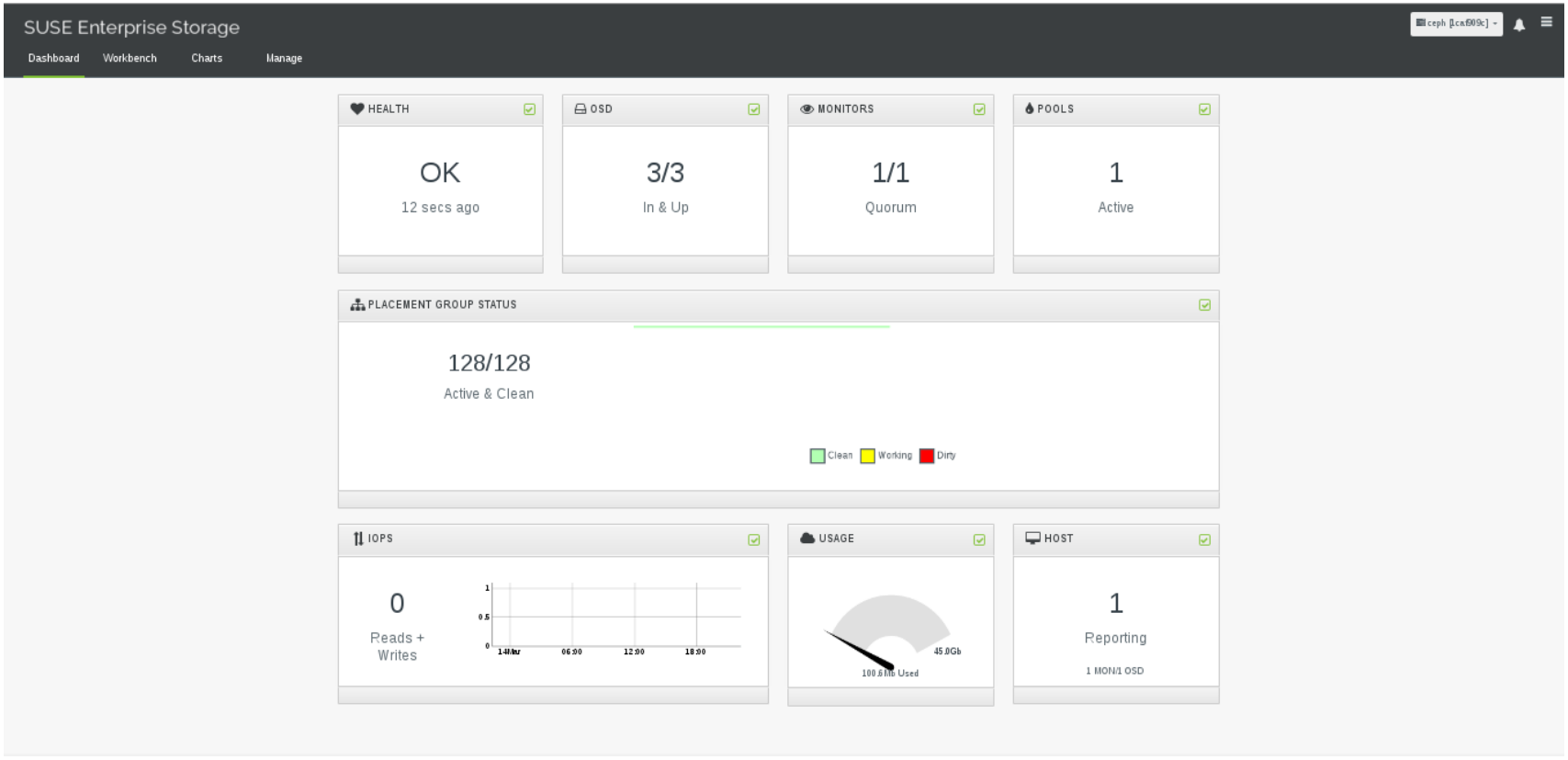
- CRUSH – *Controlled Replication Under Scalable Hashing*
- Datei (oid) → Objekt (pgid) → PGs → CRUSH (pgid) → osd1,osd2
- Jeder mit Jedem!
- Platzierungsregeln

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

War das alles?

- Pools
  - Replicated
  - Erasure Coding
- Tiering
- Federation
- Chef
- Calamari
- Backend for LIO (lrbd)

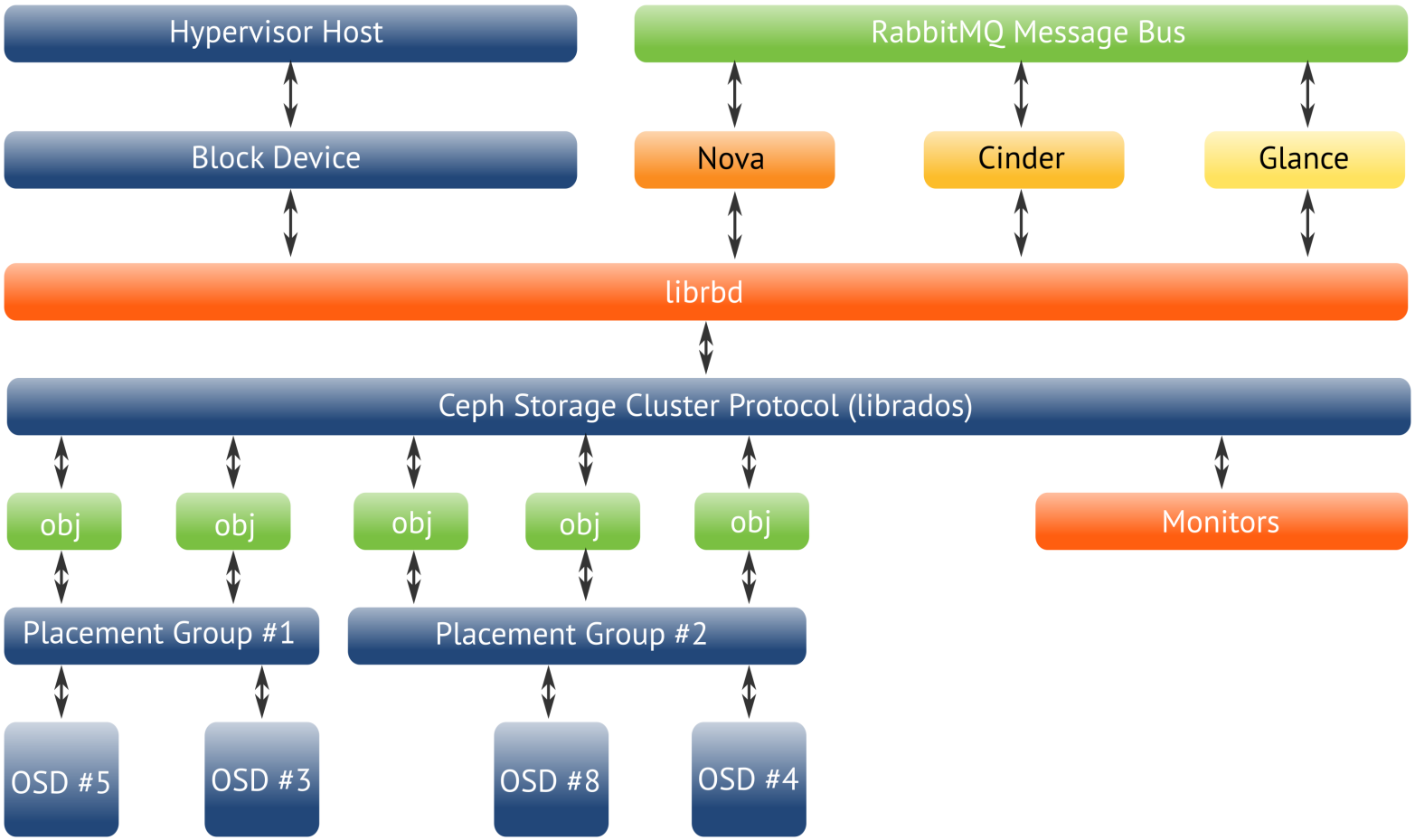
Calamari 1/2


SUSE Enterprise Storage
Dashboard
Workbench
Charts
Manage
HEALTH
OK
12 secs ago
OSD
3/3
In & Up
MONITORS
1/1
Quorum
POOLS
1
Active
PLACEMENT GROUP STATUS
128/128
Active & Clean
Clean
Working
Dirty
IOPS
0
Reads + Writes
USAGE
45.0Gb
HOST
1
Reporting
1 MON/1 OSD

Calamari 2/2


SUSE Enterprise Storage
ceph [1caf909c]
Dashboard
Workbench
Charts
Manage
Host ceph01 IOPS Per Device
System
Disk
Network
Host - ceph01
Summary
All CPUs
Bytes
Await
IOPS
Capacity
Inodes
Bytes
Packets
vda IOPS
IOPS
120
100
80
60
40
20
0
14Mar
06:00
12:00
18:00
Time (Last 24 hours)
Time Axis
24 Hours
1w
3d
24h
12h
1h
Legend
IOPS

Openstack & Ceph 1/2



Openstack & Ceph 2/2

- Glance
  - Upload, Download, Status, Snapshots, ...
- Cinder
  - Volumes, Boot Volume, Resizing, ...
- Nova
  - Live-Migration, Ephemeral, ...

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 3, D36/410.