heinlein

Storage Cluster Systeme
[Chemnitzer Linuxtage 2014]
Robert Sander <r.sander@heinlein-support.de>

Clustered Storage

Ceph und Gluster im Vergleich

Robert Sander <r.sander@heinlein-support.de>

Wer sind wir?

- wir bieten seit 20 Jahren Wissen und Erfahrung rund um Linux-Server und E-Mails
- IT-Consulting und 24/7 Linux-Support mit > 20 Mitarbeitern
- Eigener Betrieb eines ISPs seit 1992
  - jpberlin.de
  - mailbox.org
- Täglich tiefe Einblicke in die Herzen der IT aller Unternehmensgrößen

heinlein

Robert Sander <r.sander@heinlein-support.de>

Software defined Storage

Abstraktion von Hardware

- Hardware ist „egal“
- Fing eigentlich schon mit LVM an
- Beschränkt sich nicht nur auf eine Maschine
- Redundanz nicht über RAID-Controller
- Jede Hardware kann ausfallen
  - Software natürlich auch

Robert Sander <r.sander@heinlein-support.de>

Skalierbarkeit

- Beliebig in die Breite skalieren
- Keine „teure“ vertikale Skalierung notwendig
- günstigere Commodity Hardware einsetzbar
- Trotzdem: Blick auf Performance wichtig

Storage Cluster Systeme
[Chemnitzer Linuxtage 2014]
Robert Sander <r.sander@heinlein-support.de>

Ceph

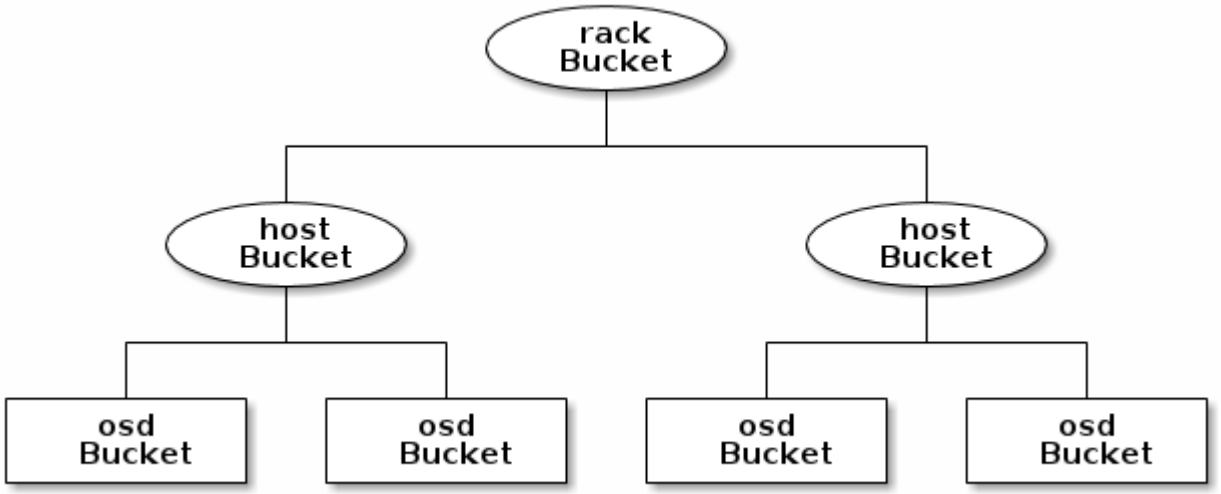
Robert Sander <r.sander@heinlein-support.de>

Ceph
Object Store

- RADOS: Reliable Autonomic Distributed Object Store
  - 2007 Doktorarbeit von Sage Weil
- Ein Object hat einen Namen in einem flachen Namensraum
  - Metadaten / Attribute
  - Daten / Payload
- Placement Groups
- Object Storage Devices
- Verteilung durch Algorithmus
  - keine Zentrale, keine verteilte Tabelle o.ä.
  - CRUSH: Controlled Replication Under Scalable Hashing

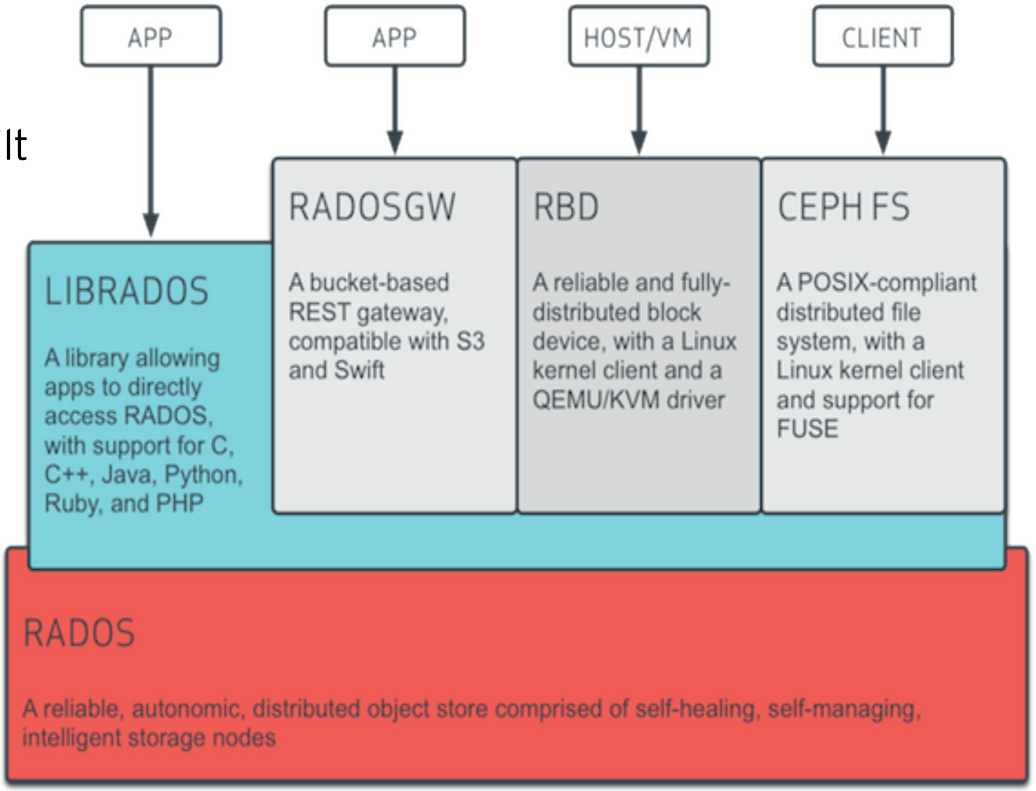
Ceph Ausfallzonen

- Alles kann ausfallen
  - Festplatten
  - Controller
  - CPUs
  - NICs
  - Software
  - Switches
  - Strom
- CRUSH Map
  - Wo möchte ich meine Redundanz platzieren?



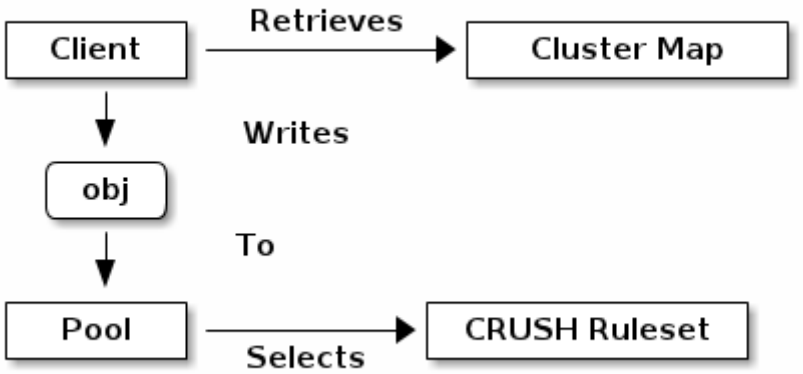
Ceph
Zugriff auf Daten

- RADOS Block Device
  - thin provisioned
  - Daten über mehrere Objekte verteilt
  - Snapshots
  - Cloning
  - Als Kernel-Device oder qemu-rbd
- REST API: radosgw
  - FastCGI
  - Amazon S3 & OpenStack Swift
- CephFS
  - POSIX-Dateisystem



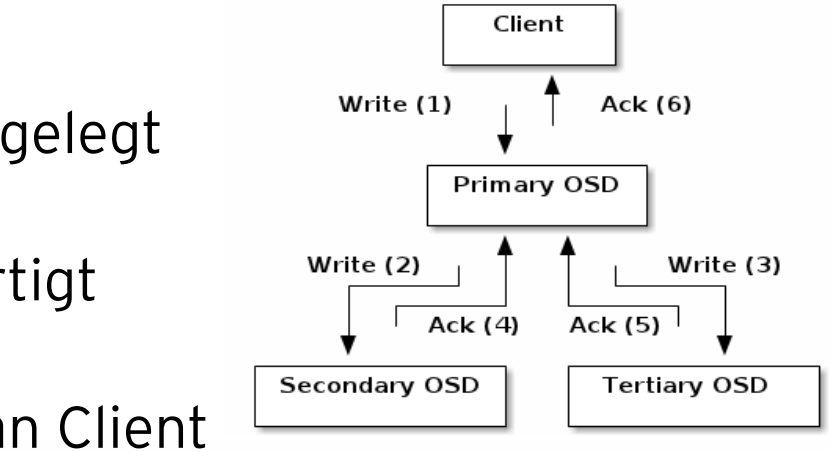
Ceph Clusterzustand

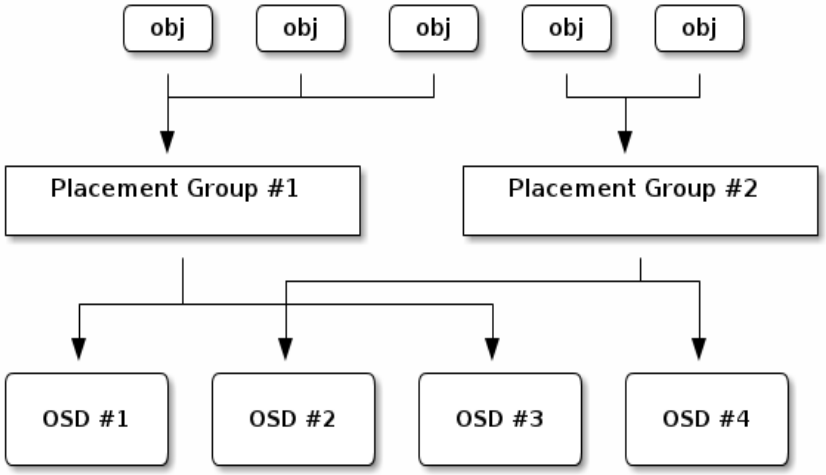
- Monitore
  - eigenen Prozesse
  - redundant
  - mit Quorum (also immer ungerade Anzahl)
  - günstig im Netzwerk verteilen
- CRUSH Map
  - Welches OSD auf welchem Knoten
  - Welches OSD aktiv
  - Pools
  - Redundanzen / Replikationen
  - Wo sind Ausfallzonen für Pools definiert
  - Datenplatzierung dann über CRUSH Algorithmus



Ceph
Redundanz / Replikation

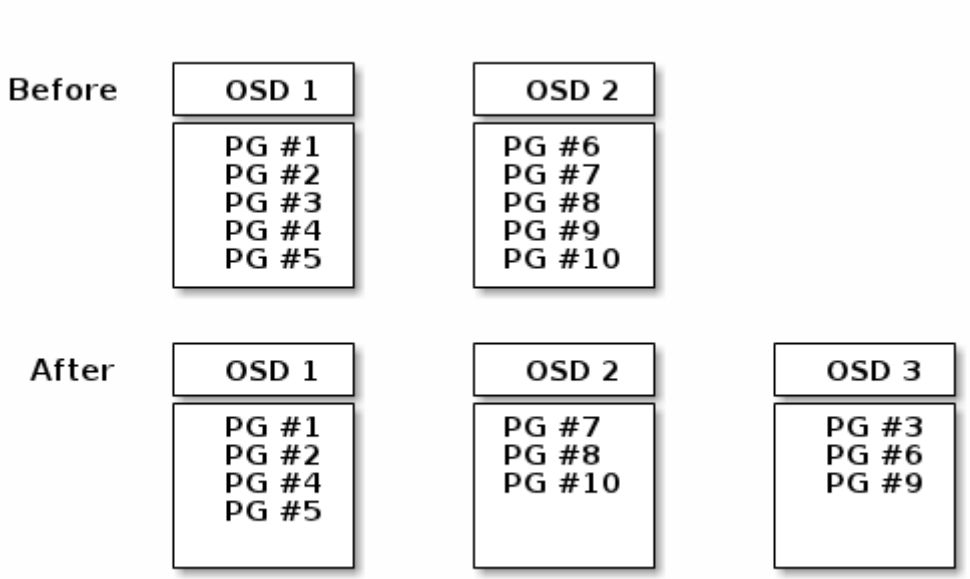
- Objekte werden in mehreren Kopien abgelegt
- Kopie wird vom „primary OSD“ angefertigt
- Erst nach Schreiben aller Kopien ACK an Client
- Lokalisierung durch CRUSH
  - Damit kennt der Client die Orte der Kopien
  - Fällt primary OSD aus, wird von Kopie gelesen
  - Gleichzeitig balanciert der Cluster die Daten neu





Ceph Skalierung

- Ausbalancierung der Placement Groups durch CRUSH
- Komplett Online
- Reduzierung auch Online
  - mit passender Replikation
- Wartung einzelner Knoten



Ceph Performance

- Durch parallelen Zugriff auf OSDs Saturierung des Netzwerks
- Schreiben kostet
  - Inter-OSD Clusternetzwerk tunen
  - 10 GB/s empfohlen
  - 1 GB/s bonding möglich
  - Journaling auf SSD
  - Auf HDD-Controller achten
- http://ceph.com/docs/master/start/hardware-recommendations/

Robert Sander <r.sander@heinlein-support.de>

Ceph Roadmap

- Erasure Coding
  - RAID5
  - Mehr Platz
  - Weniger Performance
  - Für „kalte“ Daten
- Ceph Enterprise von Inktank

Robert Sander <r.sander@heinlein-support.de>

Gluster

Robert Sander <r.sander@heinlein-support.de>

Gluster
Verteiltes Dateisystem

- Einfache Server
  - kein zentraler Metadatenserver
- Zustandslose Clients
  - aber mit konsistenter Konfiguration
- POSIX kompatibel („ähnlich“)
- erstes Release 2006
- seit 2012: RedHat Storage Server

Gluster Architektur

- Brick
  - ein Filesystem-Mountpoint auf dem Knoten (Server)
- Volume
  - wird mit Translatoren aus Bricks zusammengesetzt
- Translator
  - kombiniert Bricks zu Subvolumes und Subvolumes zu Volumes
  - Am Ende entsteht ein Graph zwischen Brick und Volume

  - Distribute (RAID0 auf Dateiebene)
  - Replicate (RAID1 auf Dateiebene)
  - Stripe (RAID0 auf Blockebene)
    - empfohlen, wenn Dateigröße > Filesystemgröße der Bricks

Gluster Ausfallzonen

- Beim Anlegen des Volumes abzubilden
- Keine einfache Rekonfiguration
- Kein dynamisches Rebalancing

Robert Sander <r.sander@heinlein-support.de>

Gluster
Zugriff auf Daten

- GlusterFS als FUSE-Dateisystem
  - libgfapi für eigene Applikationen
- NFS serverseitig eingebaut
  - als glusterfsd-Client im glusterd
  - mit Samba + ctdb auch SMB-Cluster
- REST API
  - S3 & SWIFT
  - Account → Volume
  - Container → Verzeichis
  - Objekt → Datei

Robert Sander <r.sander@heinlein-support.de>

Gluster
Redundanz / Replikation

- Distribute: Elastic Hash Algorithmus
  - 32 Bit großer Davis-Mayers-Hashraum, aufgeteilt in N Bereiche (N Subvolumes)
  - Verzeichnisse auf den Bricks erhalten jeweils disjunkten Bereich zugewiesen
  - Hash auf Dateiname
    - Bestimmt dann über den Bereich den Brick, auf dem die Datei abgelegt wird
- Replicate
  - Der Client schreibt auf alle beteilgten Subvolumes
  - Transaktionsbasiert, dadurch konsistent
  - Beliebige Anzahl Kopien möglich
- Kombination ergibt „RAID10“

Gluster Skalierung

- Dynamisch Bricks hinzufügen
  - Bei „Replicate“ immer Multiplikat der Kopienanzahl
- Ausbalancierung der Datenverteilung im Volume
- Ersatz eines Bricks

Robert Sander <r.sander@heinlein-support.de>

Gluster Performance

- Native Clients sprechen mit jedem Brickserver (glusterfsd)
  - Clients für Kopien zuständig
  - dort performantes Netzwerk notwendig
- NFS / SMB Clients werden mit Round-Robin-DNS verteilt
  - sprechen dann nur mit einem Knoten

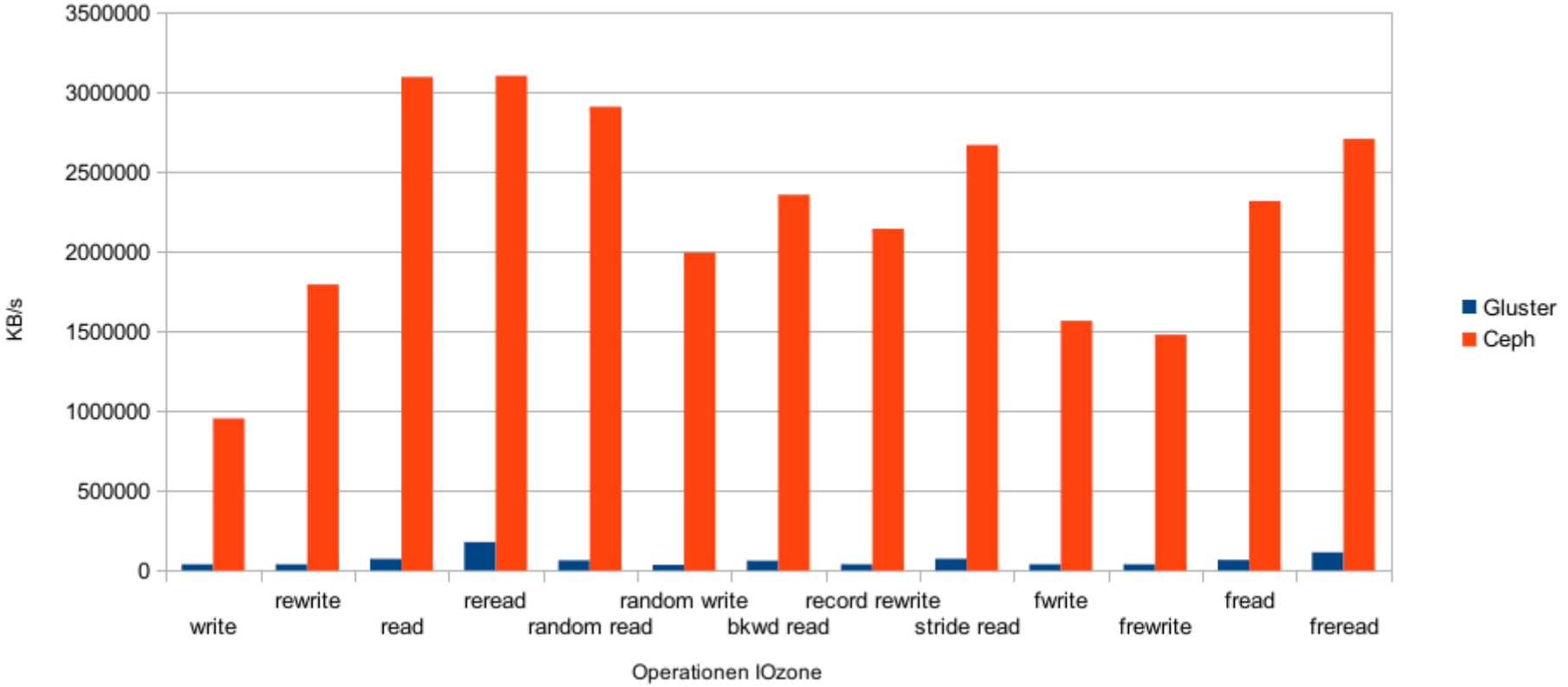
Robert Sander <r.sander@heinlein-support.de>

Gluster Roadmap

- Verteilte Geo-Replikation
- Dateisnapshots
- Kompression
- Quota serverseitig
- Volumesnapshots
- pNFS
- HSM

Performance Vergleich

→ Nicht repräsentativ



Fazit

- Ceph ist für blockbasierte Anwendungen (KVM, ...)
- Gluster ist für dateibasierte Anwendungen (NFS, SMB)
- Object Store können sie beide

- Hohe Flexibilität durch Skalierung in die Breite
- Enorme Kostenreduktion ggü. klassischen Storage-Systemen

Soweit, so gut.

Gleich sind Sie am Zug:
Fragen und Diskussionen!

Wir suchen:
Admins, Consultants, Trainer!

Wir bieten:
Spannende Projekte, Kundenlob, eigenständige Arbeit, keine Überstunden, Teamarbeit

...und natürlich: Linux, Linux, Linux...

http://www.heinlein-support.de/jobs

Und nun...

- Vielen Dank für's Zuhören...
- Schönen Tag noch...
- Und viel Erfolg an der Tastatur...

Bis bald.

heinlein

Robert Sander <r.sander@heinlein-support.de>

Heinlein Support hilft bei allen Fragen rund um Linux-Server

HEINLEIN AKADEMIE

Von Profis für Profis: Wir vermitteln die oberen 10% Wissen: geballtes Wissen und umfangreiche Praxiserfahrung.

HEINLEIN HOSTING

Individuelles Business-Hosting mit perfekter Maintenance durch unsere Profis. Sicherheit und Verfügbarkeit stehen an erster Stelle.

HEINLEIN CONSULTING

Das Backup für Ihre Linux-Administration: LPIC-2-Profis lösen im CompetenceCall Notfälle, auch in SLAs mit 24/7-Verfügbarkeit.

HEINLEIN ELEMENTS

Hard- und Software-Appliances und speziell für den Serverbetrieb konzipierte Software rund ums Thema eMail.