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

heinlein

Ceph
Erasure Coding
und weitere neue Features

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 21 Mitarbeitern
- Eigener Betrieb eines ISPs seit 1992
- Täglich tiefe Einblicke in die Herzen der IT aller Unternehmensgrößen

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

Skalierbarkeit

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

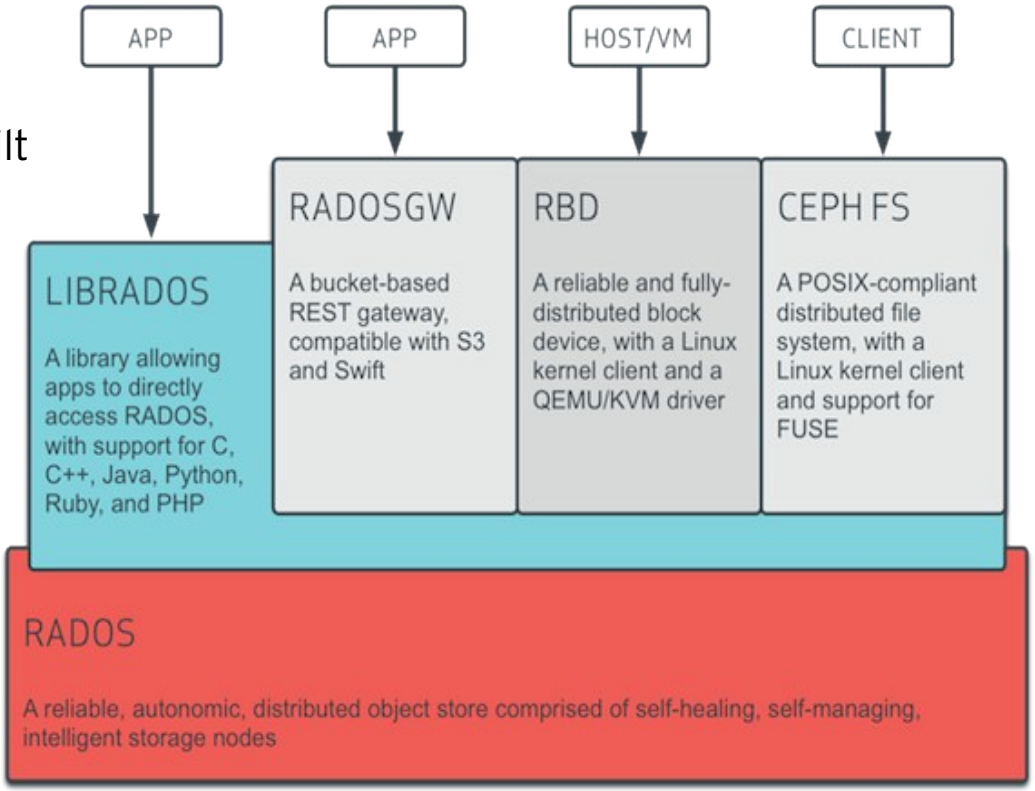

Was ist Ceph?

Ceph
Object Store

- RADOS: Reliable Autonomic Distributed Object Store
  - 2007 Doktorarbeit von Sage Weil
  - Jetzt ist Inktank Teil von RedHat
- 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

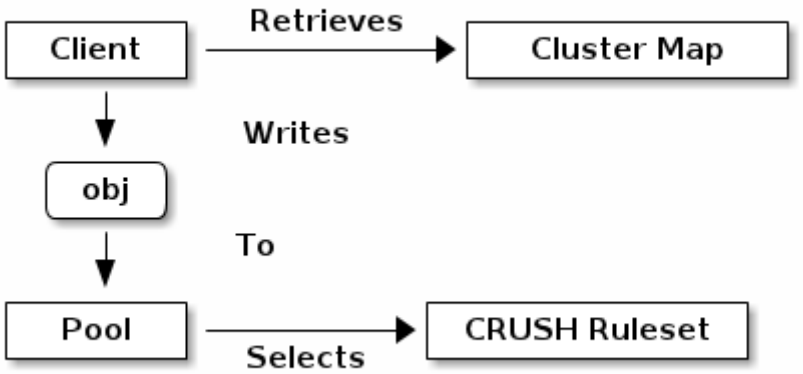
Ceph
Zugriff auf Daten

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



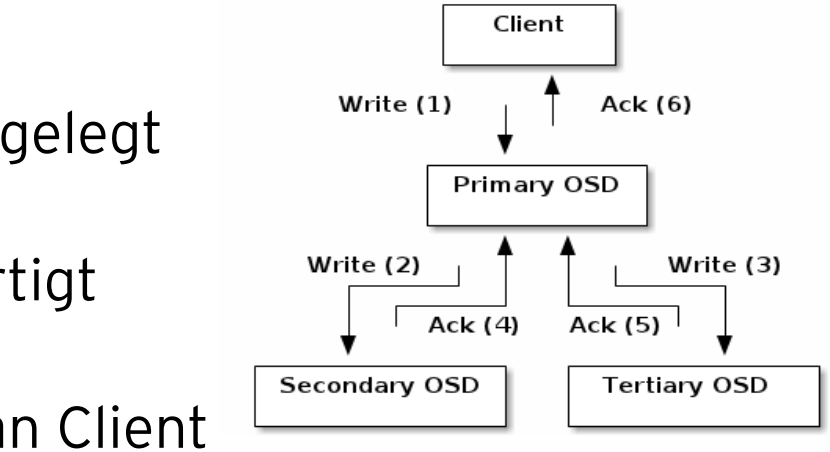
Ceph Clusterzustand

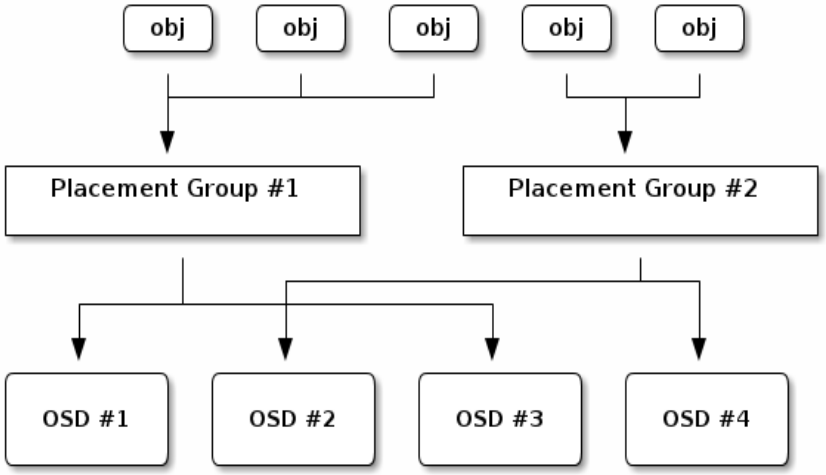
- Monitore
  - eigenen Prozesse
  - redundant
  - mit Quorum (also immer ungerade Anzahl)
  - günstig im Netzwerk verteilen
- CRUSH Map
  - Welches OSD auf welchem Knoten
  - Welches OSD aktiv
  - Pools
  - Redundanzen / Replikationen
  - Wo sind Ausfallzonen für Pools definiert
  - Datenplatzierung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



Ceph
Redundanz / Replikation

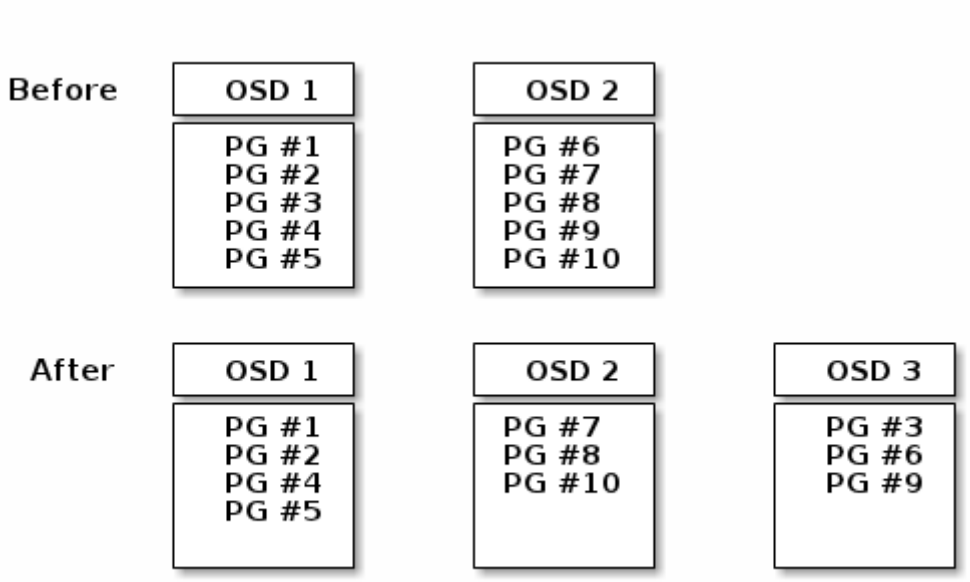
- Objekte werden in mehreren Kopien abgelegt
- Kopie wird vom „primary OSD“ angefertigt
- Erst nach Schreiben aller Kopien ACK an Client
- Lokalisierung durch CRUSH
  - Damit kennt der Client die Orte der Kopien
  - Fällt primary OSD aus, wird von Kopie gelesen
  - Gleichzeitig balanciert der Cluster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

Erasure Coding

- „RAID 5 over IP“
- braucht weniger Platz
- komplexer und langsamer
- Kennt nur einen Teil der Operationen (kein partial write)
- Einsatzzweck:
  - „Cold Storage“ von Archivdaten

Erasure Coding

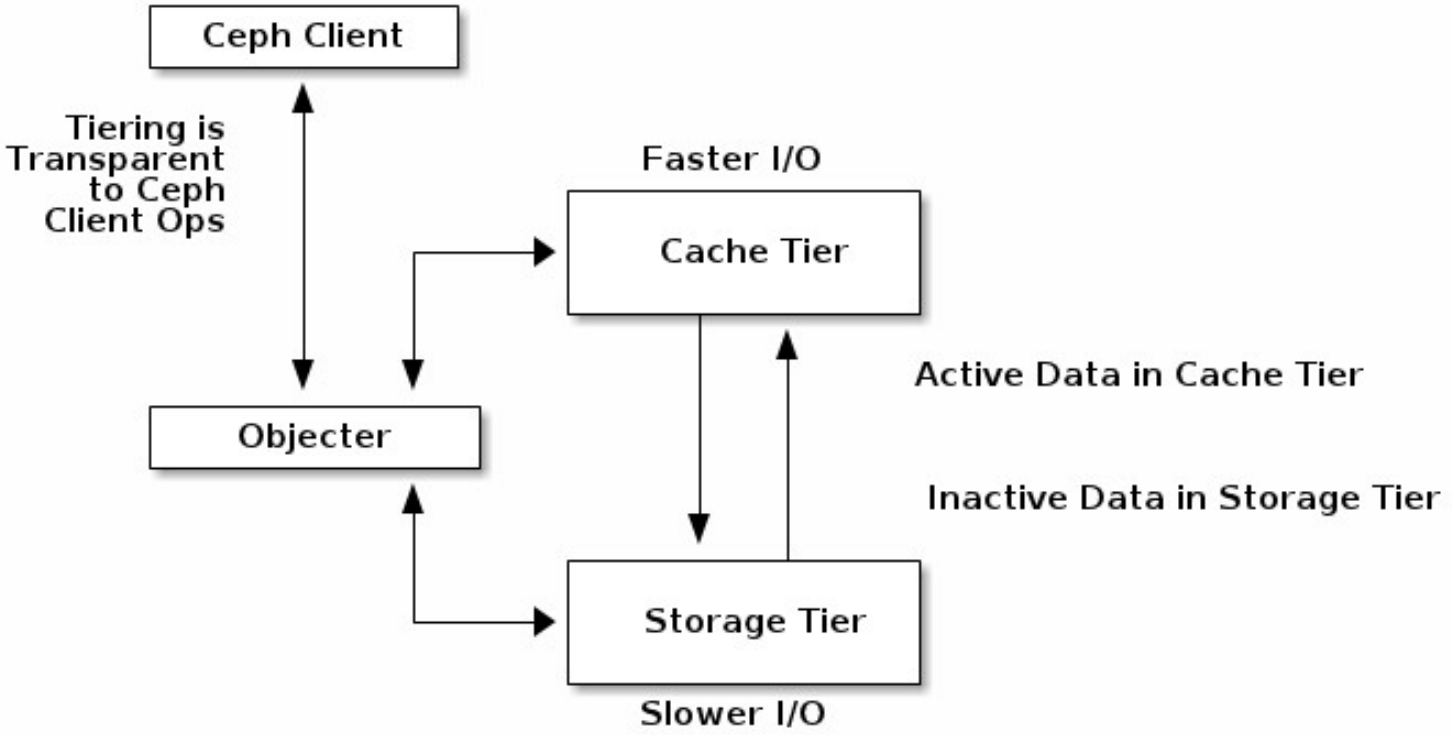
- Parameter in Profilen abgelegt
- k=data-chunks
  - Jedes Object wird in k Teile geteilt
  - Jedes Teil auf einem anderen OSD gespeichert
  - „Anzahl der RAID-Daten-Festplatten“
- m=coding-chunks
  - m „Coding Chunks“ werden für jedes Objekt berechnet
  - Jeder „Coding Chunk“ wird auf einem anderen OSD gespeichert
  - „Anzahl der RAID-Parity-Festplatten“
  - m ist die Anzahl der OSDs, die ohne Datenverlust down sein dürfen

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

Erasure Coding

- k=2, m=1
  - 1 OSD kann ausfallen
  - 4MB Objekt wird in 2 2MB Teile + 1 2MB Coding Chunk aufgeteilt = **6**MB brutto
  - Bei gleicher Replikation wären das 2 4MB Objekte = **8**MB brutto
  - Ersparnis: 25%
- k=10, m=4
  - 4 OSDs dürfen ausfallen
  - 4MB Objekt in 10 410KB Teile + 4 410KB Coding Chunks aufgeteilt = **5,6**MB brutto
  - Bei gleicher Replikation wären das 5 4MB Objekte = **20**MB brutto
  - Ersparnis: 72%
- Aber
  - Mehr CPU-Last
  - Keine RBDs direkt möglich, nur über Cache Tier

Cache Pool Tiering



Cache Pool Tiering

- Writeback
  - Schreib- und Lesezugriffe laufen über den Cache-Tier
  - Der Cache-Tier migriert ungenutzte Daten in den (langsamen) Storage-Tier
  - „Heiße“ Daten bleiben im schnellen Cache-Tier (z.B. auf SSDs)
  - Ideal für veränderliche Daten
- Read-Only
  - Schreibzugriffe direkt auf den Storage-Tier
  - Lesezugriffe über den Cache-Tier
  - Ideal für Write Once, Read Many Daten (Archive, Bildergalerien)

Cache Pool Tiering

- Zwei Pools notwendig
  - Hot-Storage: Cache Tier
  - Cold-Storage: Storage Tier
- Cold-Storage kann Erasure Coded sein
  - Platzersparnis
- CRUSH-Regeln separieren die Pools auf eigene OSDs
- `ceph osd tier add cold-storage hot-storage`
- `ceph osd tier cache-mode hot-storage writeback`
- `ceph osd tier set-overlay cold-storage hot-storage`
- `ceph osd pool set hot-storage hit_set_type bloom`

Weiteres

- Primary affinity
  - Das OSD mit der Primärkopie kann beeinflusst werden
  - Beschleunigt Lesezugriffe
- Rados-Gateway
  - Standalone Modus ohne Fast-CGI
- CephFS
  - Seit 0.80.9: Locking mit flock/fcntl


Was ist neu in Giant?
0.87: 29.10.2014
0.87.1: 27.2.2015

Verbesserungen

- Performance
  - libRADOS Code (OSD + Clients)
  - Cache Tiering
  - Monitore
- Recovery
  - Erasure Coding verbessert durch zusätzliche Datenblöcke
  - Werkzeuge für Debugging und Reparatur
- CephFS
  - Performance und Recovery verbessert
  - Aber immer noch nicht „production ready“

Upgrade

- Zuerst den Cluster auf Firefly (0.80) heben
- Dann die bewährte Reihenfolge:
  1. Monitore
  2. OSDs
  3. MDSs & Rados-Gateway
- RBD client-side caching default an
- Neue Statistiken (df & perf counter)
- CephFS Inodes mit mehreren Hardlinks vor dem Upgrade anfassen
  - `find /mnt/cephfs -type f -links +1 -exec touch \{\} \;`
- Bestimmte Cache Tier Modus-Änderungen nicht mehr erlaubt

Ceph – Erasure Coding und weitere neue Features
[Chemnitzer Linux-Tage 2015]
Robert Sander <r.sander@heinlein-support.de>

Verbesserungen

- Neuer CRUSH Bucket-Typ straw2
- RBD: Copy on Read
- Neuer Network-Messaging Stack
- CephFS: Quota auf Unterverzeichnisse
- Clustermap Checksummen für bessere Konsistenz

- Jede Menge Bugfixes

- Performance

Ausblick: Was steht auf der Roadmap?

- Rados-Gateway Sync Agent
- Samba VFS Modul
- Calamari Dashboard
- QA Testing
- Performance

- Ich suche: **Junior Consultants** für mein Team

- Natürlich und gerne stehe ich Ihnen jederzeit mit Rat und Tat zur Verfügung und freue mich auf neue Kontakte.
  - Robert Sander
  - Mail: r.sander@heinlein-support.de
  - Telefon: 030/40 50 51 – 43

- Wenn's brennt:
  - Heinlein Support 24/7 Notfall-Hotline: 030/40 505 - 110

Soweit, so gut.

Gleich sind Sie am Zug:
Fragen und Diskussionen!

heinlein

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.