TRALIOS
Linux Server Management

Intern: Ceph

Kurzeinführung in die verteile Storage-Lösung

Dominik Vallendor ▪ 29.05.2017

Motivation

TRALIOS
Linux Server Management

- Lokale Speicher sind unflexibel, selbst mit Redundanzlösungen (bsp. DRBD)
- Storages meist nicht redundant, teuer oder nicht erweiterbar
- iSCSI komplex zu handhaben
- Trend zu Cloud-Lösungen benötigt technisch soliden Unterbau

Was ist Ceph?

TRALIOS
Linux Server Management

- Verteilte Storage-Lösung
- Unterstützt beliebige Anzahl von Servern
- Stark in die Breite skalierbar (Scale-Out)
- Objektspeicher mit S3-kompatibler Schnittstelle
- Virtuelles Blockdevice
- CephFS

Vorteile

TRALIOS
Linux Server Management

- Blockdevice sehr einfach einzubinden (rbd)
- Native Unterstützung des Blockdevices durch KVM/Libvirt
- Selbstheilend
- Kein Single Point of Failure
- Einfach erweiterbar mit neuen Festplatten und/oder Servern im laufenden Betrieb
- Redundanz-Anzahl frei wählbar und im Betrieb änderbar (meist 3)
- Upgrade-Pfade für Versions-Upgrade
- Updates von Nodes im laufenden Betrieb
- Snapshots und COW (ähnlich LVM) möglich

Komponenten

TRALIOS
Linux Server Management

- Object Storage Nodes
- 1 bis x Object Storage Devices (OSD) pro Node
- Monitor Nodes
- Admin Node (optional)
- Separates Netz empfohlen

In der Regel:

mind. drei Server mit einem oder mehreren OSDs und jeweils einem Monitor

Verwaltung

TRALIOS
Linux Server Management

- *cephadmin*-Benutzer auf jedem Server
- Verwaltung vom Admin-Node aus, Zugriff per SSH (public key auth, Sudo)
- Verwaltung erfolgt meist durch *ceph-deploy*
- *ceph-deploy* ruft betriebssystem-spezifische Kommandos auf
- Software-Installation, starten von Diensten, etc. alles automatisch

Object Storage Device (OSD)

TRALIOS
Linux Server Management

- liegen jeweils auf einer Partition
- *ceph-deploy* geht von direktem Festplatten-Zugriff aus
- LVM unerwünscht
- RAID unerwünscht und macht nur wenig Sinn
- XFS-Dateisystem auf der Partition
- Dateisystem wird im Storage Node lokal gemountet
- Verwaltung durch *ceph-osd* Daemon

Monitoring Node

TRALIOS
Linux Server Management

- Überwacht den Cluster
- Wird durch *ceph-mon* Daemon bereit gestellt
- Quorum erfordert mind. drei Server
- Läuft i.d.R. mit auf dem Storage Node (Doppelfunktion)

Arbeitsweise

TRALIOS
Linux Server Management

- RADOS (reliable autonomic distributed object store)
- Storage Pools
- Pool ID und Object ID
- Cluster-Map
- Hashing der Object ID und gleichmäßige Verteilung auf die Placement Groups
- CRUSH Algorithmus (Controlled Replication Under Scalable Hashing)

Betrieb

TRALIOS
Linux Server Management

- Funktionsweise testen: *ceph health*
- Detaillierte Informationen: *ceph status*
- OSD-Status abfragen: *ceph osd tree*
- Umgang mit OSDs: *ceph osd out, stop, crush, del, rm, create*

Performance

TRALIOS
Linux Server Management

- Messergebnisse nocht nicht repräsentativ
- Testsystem mit langsamer Hardware, Partitionen schlecht verteilt
- Performance scheint insgesamt aber gut zu sein
- Zugriffsgeschwindigkeit ca. 30% bis über 100% einer einzelnen Festplatte
- Geschwindigkeit ist auch beim Hintergrund-Rebuild immer noch gut
- Detaillierte Übertragungs/Zugriffsdaten live einsehbar
- Performance angeblich stark abhängig vom Netzwerk. 10 Gbit/s empfohlen

Fehlerszenarien

TRALIOS
Linux Server Management

Storage Node Ausfall:

- Server wird aus dem Cluster entfernt
- Reaktion sehr schnell

Festplatten-Ausfall:

- Ausfall wird vom Storage Node erkannt
- OSD wird erst beim Zugriff entfernt
- Verzeichnis muss manuell geunmounted werden, etc.
- Prinzipiell aber kein Problem

OSD Replacement, Erweiterung

TRALIOS
Linux Server Management

- Austausch eines defekten OSDs nicht vorgesehen
- Austausch = Entfernen + Erweitern von/mit OSDs
- Defekte OSDs müssen manuell deaktivert/entfernt werden
- Cluster organisiert sich nach Änderung von OSDs automatisch neu

Ausblick

TRALIOS
Linux Server Management

- Weitere Performance-Tests notwenig
- SSD-only vs. SSD-Journal
- Verschiedene Netzwerksetups und -bandbreiten testen
- Wie viele OSDs pro Storage Node?
- CRUSH-Map verstehen/bearbeiten
- KVM-Livemigration