Ceph CRUSH & device classes

Der Algorithmus (Controlled Replication Under Scalable Hashing) bildet die Grundlage von Ceph.

CRUSH berechnet, wo Daten gespeichert und woher sie abgerufen werden. Dies hat den Vorteil, dass kein zentraler Indexierungsdienst erforderlich ist. CRUSH arbeitet mit einer Karte von OSDs, Buckets (Gerätestandorten) und Regelsätzen (Datenreplikation) für Pools.

Diese Zuordnung kann geändert werden, um verschiedene Replikationshierarchien widerzuspiegeln. Die Objektreplikate können getrennt werden (z. B. Fehlerdomänen), während die gewünschte Verteilung beibehalten wird.

Eine gängige Konfiguration besteht darin, verschiedene Festplattenklassen für verschiedene Ceph-Pools zu verwenden. Aus diesem Grund hat Ceph Geräteklassen mit Luminous eingeführt, um dem Bedarf an einfacher Regelsatzgenerierung gerecht zu werden.

Die Geräteklassen sind in der Ceph OSD-Baumausgabe zu sehen. Diese Klassen stellen ihren eigenen Root-Bucket dar, der mit dem folgenden Befehl angezeigt werden kann.

```
ceph osd crush tree --show-shadow
```

Eine beispielhafte Ausgabe vom obrigen Kommando wäre:

```
ID  CLASS  WEIGHT   TYPE NAME
-2    hdd  0.29306  root default~hdd
-6    hdd  0.09769      host proxmox-a-0~hdd
 1    hdd  0.09769          osd.1
-8    hdd  0.09769      host proxmox-b-0~hdd
 2    hdd  0.09769          osd.2
-4    hdd  0.09769      host proxmox-c-0~hdd
 0    hdd  0.09769          osd.0
-1         0.29306  root default
-5         0.09769      host proxmox-a-0
 1    hdd  0.09769          osd.1
-7         0.09769      host proxmox-b-0
 2    hdd  0.09769          osd.2
-3         0.09769      host proxmox-c-0
 0    hdd  0.09769          osd.0
```

Um einen Pool anzuweisen, Objekte nur auf einer bestimmten Geräteklasse zu verteilen, müßt Du zunächst einen Regelsatz für die Geräteklasse erstellen:

```
ceph osd crush rule create-replicated <rule-name> <root> <failure-domain>
<class>

  <rule-name> - Name der Regel, um eine Verbindung mit einem Pool
herzustellen
  <root> - zu welcher Crush-Root es gehören soll (standardmäßiger Ceph-Root
„default“)
```

```
  <failure-domain> - an welche Fehlerdomäne die Objekte verteilt werden
sollen (normalerweise Host)
  <class> - welche Art von OSD-Backing-Store verwendet werden soll (z. B.
NVMe, SSD, HDD)
```

Sobald sich die Regel in der CRUSH-Map befindet, kannst Du einem Pool anweisen, den Regelsatz zu verwenden.

```
ceph osd pool set <pool-name> crush_rule <rule-name>
```

Wenn der Pool bereits Objekte enthält, müssen diese entsprechend verschoben werden. Abhängig von Ihrem Setup kann dies erhebliche Leistungseinbußen für Ihren Cluster mit sich bringen. Alternativ kannst Du einen neuen Pool erstellen und die Datenträger einzeln verschieben.