Hyperkonvergente Ceph-Cluster

Ceph ist ein verteilter Objektspeicher und Dateisystem, das für hervorragende Leistung, Zuverlässigkeit und Skalierbarkeit konzipiert ist.

- Selbstheilung
- Skalierbar auf Exabyte-Ebene
- Bietet Block-, Dateisystem- und Objektspeicher
- Einrichten von Pools mit unterschiedlichen Leistungs- und Redundanzmerkmalen
- Daten werden repliziert, wodurch sie fehlertolerant sind
- Läuft auf Standardhardware
- Open Source

Terminology

Ceph besteht aus mehreren DAEMONS zur Verwendung als RBD-Speicher:

- Ceph Monitor (ceph-mon, oder MON)
- Ceph Manager (ceph-mgr, oder MGS)
- Ceph Metadata Service (ceph-mds, oder MDS)
- Ceph Object Storage Daemon (ceph-osd, oder OSD)

Empfehlungen für einen gesunden Ceph-Cluster

Um einen hyperkonvergenten Ceph-Cluster aufzubauen, müßt Du mindestens drei (vorzugsweise) identische Server für das Setup verwenden.

CPU

Ceph-Dienste können in zwei Kategorien eingeteilt werden:

- Intensive CPU-Auslastung, die von hohen CPU-Basisfrequenzen und mehreren Kernen profitiert. Mitglieder dieser Kategorie sind:
  - **Object Storage Daemon (OSD)-Dienste**
  - Meta Data Service (MDS), der für CephFS verwendet wird
- Moderate CPU-Auslastung, die nicht mehrere CPU-Kerne benötigt. Dies sind:
  - Monitor (MON)-Dienste
  - Manager (MGR)-Dienste

Als einfache Faustregel solltest Du jedem Ceph-Dienst mindestens einen CPU-Kern (oder Thread) zuweisen, um die Mindestressourcen bereitzustellen, die für eine stabile und dauerhafte Ceph-Leistung erforderlich sind.

Wenn Du beispielsweise planst, einen Ceph-Monitor, einen Ceph-Manager und 6 Ceph OSD-Dienste auf einem Knoten auszuführen, solltest Du 8 CPU-Kerne ausschließlich für Ceph reservieren, wenn Du

eine grundlegende und stabile Leistung anstrebst.

Beachte, dass die CPU-Auslastung von OSDs hauptsächlich von der Festplattenleistung abhängt. Je höher die möglichen IOPS (IO Operations per Second) einer Festplatte sind, desto mehr CPU kann von einem OSD-Dienst genutzt werden. Bei modernen Enterprise-SSD-Festplatten, wie NVMe, die dauerhaft eine hohe IOPS-Last von über 100.000 mit einer Latenz von unter einer Millisekunde aufrechterhalten können, kann jedes OSD mehrere CPU-Threads verwenden. Beispielsweise sind bei Festplatten mit sehr hoher Leistung wahrscheinlich vier bis sechs CPU-Threads pro NVMe-gestütztem OSD erforderlich.

Hauptpeicher (RAM)

Insbesondere in einer hyperkonvergenten Konfiguration muss der Speicherverbrauch sorgfältig geplant und überwacht werden.

Als Faustregel gilt: Für etwa **1 TiB Daten wird 1 GiB Speicher** von einem OSD verwendet. Während die Nutzung unter normalen Bedingungen geringer sein kann, wird sie bei kritischen Vorgängen wie recovery, re-balancing oder backfilling (Auffüllen) am meisten genutzt. Das bedeutet, dass Du vermeiden solltest, Deinen verfügbaren Speicher bereits im Normalbetrieb voll auszuschöpfen, sondern etwas Spielraum lassen solltest, um mit Ausfällen fertig zu werden.

Der OSD-Dienst selbst wird zusätzlichen Speicher verwenden. Das Ceph BlueStore-Backend des Daemons benötigt standardmäßig 3-5 GiB Speicher (anpassbar).

Netzwerk

Network

Ich empfehle eine Netzwerkbandbreite von mindestens 10 Gbit/s oder mehr, die ausschließlich für Ceph-Verkehr verwendet wird. Ein 1-Mesh-Netzwerk-Setup ist auch eine Option für Cluster mit drei bis fünf Knoten, wenn keine Switches mit 10+ Gbit/s verfügbar sind. Um Deinen Bandbreitenbedarf abzuschätzen, müßt Du die Leistung Deiner Festplatten berücksichtigen. Während eine einzelne Festplatte eine 1-GB-Verbindung möglicherweise nicht auslastet, können mehrere HDD-OSDs pro Knoten bereits 10 Gbit/s auslasten. Wenn moderne NVMe-SSDs verwendet werden, kann eine einzelne bereits 10 Gbit/s Bandbreite oder mehr auslasten. Für solche Hochleistungs-Setups empfehle ich mindestens 25 Gbit/s, während sogar 40 Gbit/s oder 100+ Gbit/s erforderlich sein können, um das volle Leistungspotenzial der zugrunde liegenden Festplatten auszuschöpfen.

Wenn Du Dir nicht sicher bist, empfehle ich die Verwendung von zwei (physischen) separaten Netzwerken für Hochleistungs-Setups:

- ein Netzwerk mit sehr hoher Bandbreite (25+ Gbit/s) für den (internen) Ceph-Cluster-Datenverkehr.
- ein Netzwerk mit hoher Bandbreite (10+ Gbit/s) für den (public) Ceph-Datenverkehr zwischen dem Ceph-Server und dem Ceph-Client-Storage Datenverkehr.

Disks

Bei der Planung der Größe Deines Ceph-Clusters ist es wichtig, die Wiederherstellungszeit zu berücksichtigen.

Besonders bei kleinen Clustern kann die Wiederherstellung lange dauern. Es wird empfohlen, in kleinen Setups SSDs anstelle von HDDs zu verwenden, um die Wiederherstellungszeit zu verkürzen und die Wahrscheinlichkeit eines nachfolgenden Fehlerereignisses während der Wiederherstellung zu minimieren.

Im Allgemeinen bieten SSDs mehr IOPS als rotierende Festplatten. In Anbetracht dessen kann es neben den höheren Kosten sinnvoll sein, eine klassenbasierte Trennung der Pools zu implementieren. Eine andere Möglichkeit, OSDs zu beschleunigen, besteht darin, eine schnellere Festplatte als Journal- oder DB/Write-Ahead-Log-Gerät zu verwenden. Wenn eine schnellere Festplatte für mehrere OSDs verwendet wird, muss ein angemessenes Gleichgewicht zwischen OSD- und WAL-/DB- (oder Journal-)Festplatte ausgewählt werden, da sonst die schnellere Festplatte zum Engpass für alle verknüpften OSDs wird.

Abgesehen vom Festplattentyp funktioniert Ceph am besten mit einer gleichmäßig großen und gleichmäßig verteilten Anzahl von Festplatten pro Knoten. Beispielsweise sind 4 x 500 GB-Festplatten in jedem Knoten besser als eine gemischte Konfiguration mit einer einzelnen 1 TB- und drei 250 GB-Festplatten.

Du musst auch die Anzahl der OSDs und die Kapazität einzelner OSDs ausbalancieren. Mehr Kapazität ermöglicht Dir eine Erhöhung der Speicherdichte, bedeutet aber auch, dass ein einzelner OSD-Ausfall Ceph dazu zwingt, mehr Daten auf einmal wiederherzustellen.

Vermeide RAID

Da Ceph Datenobjektredundanz und mehrere parallele Schreibvorgänge auf Festplatten (OSDs) selbst handhabt, verbessert die Verwendung eines RAID-Controllers normalerweise weder die Leistung noch die Verfügbarkeit. Im Gegenteil, Ceph ist darauf ausgelegt, ganze Festplatten allein und ohne Abstraktion dazwischen zu handhaben. RAID-Controller sind nicht für die Ceph-Arbeitslast ausgelegt und können die Dinge verkomplizieren und manchmal sogar die Leistung verringern, da ihre Schreib- und Caching-Algorithmen mit denen von Ceph in Konflikt geraten können.