Managing and Customizing the CRUSH Map

Objectives

After completing this section, you should be able to administer and update the cluster CRUSH map used by the Ceph cluster.

CRUSH and Object Placement Strategies

Ceph calculates which OSDs should hold which objects by using a placement algorithm called CRUSH (Controlled Replication Under Scalable Hashing). Objects are assigned to placement groups (PGs) and CRUSH determines which OSDs those placement groups should use to store their objects.

The CRUSH Algorithm

The CRUSH algorithm enables Ceph clients to directly communicate with OSDs; this avoids a centralized service bottleneck. Ceph clients and OSDs use the CRUSH algorithm to efficiently compute information about object locations, instead of having to depend on a central lookup table. Ceph clients retrieve the cluster maps and use the CRUSH map to algorithmically determine how to store and retrieve data. This enables massive scalability for the Ceph cluster by avoiding a single point of failure and a performance bottleneck.

The CRUSH algorithm works to uniformly distribute the data in the object store, manage replication, and respond to system growth and hardware failures. When new OSDs are added or an existing OSD or OSD host fails, Ceph uses CRUSH to rebalance the objects in the cluster among the active OSDs.

CRUSH Map Components

Conceptually, a CRUSH map contains two major components:

A CRUSH hierarchy
: This lists all available OSDs and organizes them into a treelike structure of *buckets*.

  The CRUSH hierarchy is often used to represent where OSDs are located. By default, there is a root bucket representing the whole hierarchy, which contains a host bucket for each OSD host.

  The OSDs are the leaves of the tree, and by default all OSDs on the same OSD host are placed in that host's bucket. You can customize the tree structure to rearrange it, add more levels, and group OSD hosts into buckets representing their location in different server racks or data centers.

At least one CRUSH rule
: *CRUSH rules* determine how placement groups are assigned OSDs from those buckets. This determines where objects for those placement groups are stored. Different pools might use different CRUSH rules from the CRUSH map.