Ceph Object Storage Devices

Ceph Object Storage Devices (OSDs) are the building blocks of a Ceph storage cluster. OSDs connect a storage device (such as a hard disk or other block device) to the Ceph storage cluster. An individual storage server can run multiple OSD daemons and provide multiple OSDs to the cluster. Red Hat Ceph Storage 5 supports a feature called BlueStore to store data within RADOS. BlueStore uses the local storage devices in raw mode and is designed for high performance.

One design goal for OSD operation is to bring computing power as close as possible to the physical data so that the cluster can perform at peak efficiency. Both Ceph clients and OSD daemons use the Controlled Replication Under Scalable Hashing (CRUSH) algorithm to efficiently compute information about object location, instead of depending on a central lookup table.

CRUSH Map

CRUSH assigns every object to a Placement Group (PG), which is a single hash bucket. PGs are an abstraction layer between the objects (application layer) and the OSDs (physical layer). CRUSH uses a pseudo-random placement algorithm to distribute the objects across the PGs and uses rules to determine the mapping of the PGs to the OSDs. In the event of failure, Ceph remaps the PGs to different physical devices (OSDs) and synchronizes their content to match the configured data protection rules.

Primary and Secondary OSDs

One OSD is the *primary OSD* for the object's placement group, and Ceph clients always contact the primary OSD in the acting set when it reads or writes data. Other OSDs are *secondary OSDs* and play an important role in ensuring the resilience of data in the event of failures in the cluster.

Primary OSD functions:

- Serve all I/O requests
- Replicate and protect the data
- Check data coherence
- Rebalance the data
- Recover the data

Secondary OSD functions:

- Act always under control of the primary OSD
- Can become the primary OSD

Warning

A host that runs OSDs must not mount Ceph RBD images or CephFS file systems by using the kernel-based client. Mounted resources can become unresponsive due to memory deadlocks or blocked I/O that is pending on stale sessions.

Ceph Managers

Ceph Managers (MGRs) provide for the collection of cluster statistics.

If no MGRs are available in a cluster, client I/O operations are not negatively affected, but attempts to query cluster statistics fail. To avoid this scenario, Red Hat recommends that you deploy at least two Ceph MGRs for each cluster, each running in a separate failure domain.

The MGR daemon centralizes access to all data that is collected from the cluster and provides a simple web dashboard to storage administrators. The MGR daemon can also export status information to an external monitoring server, such as Zabbix.