Notice

This document is for a development version of Ceph.

Report a Documentation Bug

ADDING/REMOVING OSDS

When a cluster is up and running, it is possible to add or remove OSDs.

ADDING OSDS

OSDs can be added to a cluster in order to expand the cluster's capacity and resilience. Typically, an OSD is a Ceph `ceph-osd` daemon running on one storage drive within a host machine. But if your host machine has multiple storage drives, you may map one `ceph-osd` daemon for each drive on the machine.

It's a good idea to check the capacity of your cluster so that you know when it approaches its capacity limits. If your cluster has reached its `near full` ratio, then you should add OSDs to expand your cluster's capacity.

Warning

Do not let your cluster reach its `full ratio` before adding an OSD. OSD failures that occur after the cluster reaches its `near full ratio` might cause the cluster to exceed its `full ratio`.

DEPLOYING YOUR HARDWARE

If you are also adding a new host when adding a new OSD, see Hardware Recommendations for details on minimum recommendations for OSD hardware. To add an OSD host to your cluster, begin by making sure that an appropriate version of Linux has been installed on the host machine and that all initial preparations for your storage drives have been carrie

Next, add your OSD host to a rack in your cluster, connect the host to the network, and ensure that the host has network connectivity. For details, see Network Configuration Reference.

If your cluster has been manually deployed, you will need to install Ceph software packages manually. For details, see Installing Ceph (Manual). Configure SSH for the appropriate user to have both passwordless authentication and root permissions.

ADDING AN OSD (MANUAL)

The following procedure sets up a `ceph-osd` daemon, configures this OSD to use one drive, and configures the cluster to distribute data to the OSD. If your host machine has multiple drives, you may add an OSD for each drive on the host by repeating this procedure.

As the following procedure will demonstrate, adding an OSD involves creating a metadata directory for it, configuring a data storage drive, adding the OSD to the cluster, and then adding it to the CRUSH map.

When you add the OSD to the CRUSH map, you will need to consider the weight you assign to the new OSD. Since storage drive capacities increase over time, newer OSD hosts are likely to have larger hard drives than the older hosts in the cluster have and therefore might have greater weight as well.

Tip

Ceph works best with uniform hardware across pools. It is possible to add drives of dissimilar size and then adjust their weights accordingly. However, for best performance, consider a CRUSH hierarchy that has drives of the same type and size. It is better to add larger drives uniformly to existing hosts. This can be done incrementally, replacing smaller drives each time the new drives are added.

1. Create the new OSD by running a command of the following form. If you opt not to specify a UUID in this command, the UUID will be set automatically when the OSD starts up. The OSD number, which is needed for subsequent steps, is found in the command's output:

   ```
   $ ceph osd create [{uuid} [{id}]]
   ```

   If the optional parameter {id} is specified it will be used as the OSD ID. However, if the ID number is already in use, the command will fail.

   Warning

array, and any skipping of entries consumes extra memory. This memory consumption can become significant if there are large gaps or if clusters are large. By leaving the `{id}` parameter unspecified, we ensure that Ceph uses the smallest ID number available and that these problems are avoided.

2. Create the default directory for your new OSD by running commands of the following form:

```
$ ssh {new-osd-host}
$ sudo mkdir /var/lib/ceph/osd/ceph-{osd-number}
```

3. If the OSD will be created on a drive other than the OS drive, prepare it for use with Ceph. Run commands of the following form:

```
$ ssh {new-osd-host}
$ sudo mkfs -t {fstype} /dev/{drive}
$ sudo mount -o user_xattr /dev/{hdd} /var/lib/ceph/osd/ceph-{osd-number}
```

4. Initialize the OSD data directory by running commands of the following form:

```
$ ssh {new-osd-host}
$ ceph-osd -i {osd-num} --mkfs --mkkey
```

   Make sure that the directory is empty before running `ceph-osd`.

5. Register the OSD authentication key by running a command of the following form:

```
$ ceph auth add osd.{osd-num} osd 'allow *' mon 'allow rwx' -i /var/lib/ceph/osd/ceph-{osd-num}/keyring
```

   This presentation of the command has `ceph-{osd-num}` in the listed path because many clusters have the name `ceph`. However, if your cluster name is not `ceph`, then the string `ceph` in `ceph-{osd-num}` needs to be replaced with your cluster name. For example, if your cluster name is `cluster1`, then the path in the command should be `/var/lib/ceph/osd/cluster1-{osd-num}/keyring`.

6. Add the OSD to the CRUSH map by running the following command. This allows the OSD to begin receiving data. The `ceph osd crush add` command can add OSDs to the CRUSH hierarchy wherever you want. If you specify one or more buckets, the command places the OSD in the most specific of those buckets, and it moves that bucket underne

will attach the OSD directly to the root, but CRUSH rules expect OSDs to be inside of hosts. If the OSDs are not inside hosts, the OSDs will likely not receive any data.

```
$ ceph osd crush add {id-or-name} {weight}  [{bucket-type}={bucket-name} ...]
```

Note that there is another way to add a new OSD to the CRUSH map: decompile the CRUSH map, add the OSD to the device list, add the host as a bucket (if it is not already in the CRUSH map), add the device as an item in the host, assign the device a weight, recompile the CRUSH map, and set the CRUSH map. For details, see Adding/Moving an OSD. This is rarely necessary with recent releases (this sentence was written the month that Reef was released).

REPLACING AN OSD

Note

If the procedure in this section does not work for you, try the instructions in the `cephadm` documentation: Replacing an OSD.

Sometimes OSDs need to be replaced: for example, when a disk fails, or when an administrator wants to reprovision OSDs with a new back end (perhaps when switching from Filestore to BlueStore). Replacing an OSD differs from Removing the OSD in that the replaced OSD's ID and CRUSH map entry must be kept intact after the OSD is destroyed for replacement.

1. Make sure that it is safe to destroy the OSD:

```
$ while ! ceph osd safe-to-destroy osd.{id} ; do sleep 10 ; done
```

2. Destroy the OSD:

```
$ ceph osd destroy {id} --yes-i-really-mean-it
```

3. *Optional*: If the disk that you plan to use is not a new disk and has been used before for other purposes, zap the disk:

```
$ ceph-volume lvm zap /dev/sdX
```

4. Prepare the disk for replacement by using the ID of the OSD that was destro
steps:

5. Finally, activate the OSD:

```
$ ceph-volume lvm activate {id} {fsid}
```

Alternatively, instead of carrying out the final two steps (preparing the disk and activating the OSD), you can re-create the OSD by running a single command of the following form:

```
$ ceph-volume lvm create --osd-id {id} --data /dev/sdX
```

STARTING THE OSD

After an OSD is added to Ceph, the OSD is in the cluster. However, until it is started, the OSD is considered `down` and `in`. The OSD is not running and will be unable to receive data. To start an OSD, either run `service ceph` from your admin host or run a command of the following form to start the OSD from its host machine:

```
$ sudo systemctl start ceph-osd@{osd-num}
```

After the OSD is started, it is considered `up` and `in`.

OBSERVING THE DATA MIGRATION

After the new OSD has been added to the CRUSH map, Ceph begins rebalancing the cluster by migrating placement groups (PGs) to the new OSD. To observe this process by using the ceph tool, run the following command:

```
$ ceph -w
```

Or:

```
$ watch ceph status
```

return to `active+clean` when migration completes. When you are finished observing, press Ctrl-C to exit.

REMOVING OSDS (MANUAL)

It is possible to remove an OSD manually while the cluster is running: you might want to do this in order to reduce the size of the cluster or when replacing hardware. Typically, an OSD is a Ceph `ceph-osd` daemon running on one storage drive within a host machine. Alternatively, if your host machine has multiple storage drives, you might need to remove multiple `ceph-osd` daemons: one daemon for each drive on the machine.

Warning

Before you begin the process of removing an OSD, make sure that your cluster is not near its `full ratio`. Otherwise the act of removing OSDs might cause the cluster to reach or exceed its `full ratio`.

TAKING THE OSD `OUT` OF THE CLUSTER

OSDs are typically `up` and `in` before they are removed from the cluster. Before the OSD can be removed from the cluster, the OSD must be taken `out` of the cluster so that Ceph can begin rebalancing and copying its data to other OSDs. To take an OSD `out` of the cluster, run a command of the following form:

```
$ ceph osd out {osd-num}
```

OBSERVING THE DATA MIGRATION

After the OSD has been taken `out` of the cluster, Ceph begins rebalancing the cluster by migrating placement groups out of the OSD that was removed. To observe this process by using the ceph tool, run the following command:

```
$ ceph -w
```

return to `active+clean` when migration completes. When you are finished observing, press Ctrl-C to exit.

Note

Under certain conditions, the action of taking `out` an OSD might lead CRUSH to encounter a corner case in which some PGs remain stuck in the `active+remapped` state. This problem sometimes occurs in small clusters with few hosts (for example, in a small testing cluster). To address this problem, mark the OSD `in` by running a command of the following form:

```
$ ceph osd in {osd-num}
```

After the OSD has come back to its initial state, do not mark the OSD `out` again. Instead, set the OSD's weight to `0` by running a command of the following form:

```
$ ceph osd crush reweight osd.{osd-num} 0
```

After the OSD has been reweighted, observe the data migration and confirm that it has completed successfully. The difference between marking an OSD `out` and reweighting the OSD to `0` has to do with the bucket that contains the OSD. When an OSD is marked `out`, the weight of the bucket is not changed. But when an OSD is reweighted to `0`, the weight of the bucket is updated (namely, the weight of the OSD is subtracted from the overall weight of the bucket). When operating small clusters, it can sometimes be preferable to use the above reweight command.

STOPPING THE OSD

After you take an OSD `out` of the cluster, the OSD might still be running. In such a case, the OSD is `up` and `out`. Before it is removed from the cluster, the OSD must be stopped by running commands of the following form:

```
$ ssh {osd-host}
$ sudo systemctl stop ceph-osd@{osd-num}
```

After the OSD has been stopped, it is `down`.

The following procedure removes an OSD from the cluster map, removes the OSD's authentication key, removes the OSD from the OSD map, and removes the OSD from the `ceph.conf` file. If your host has multiple drives, it might be necessary to remove an OSD from each drive by repeating this procedure.

1. Begin by having the cluster forget the OSD. This step removes the OSD from the CRUSH map, removes the OSD's authentication key, and removes the OSD from the OSD map. (The purge subcommand was introduced in Luminous. For older releases, see the procedure linked here.):

```
$ ceph osd purge {id} --yes-i-really-mean-it
```

2. Navigate to the host where the master copy of the cluster's `ceph.conf` file is kept:

```
$ ssh {admin-host}
$ cd /etc/ceph
$ vim ceph.conf
```

3. Remove the OSD entry from your `ceph.conf` file (if such an entry exists):

```
[osd.1]
    host = {hostname}
```

4. Copy the updated `ceph.conf` file from the location on the host where the master copy of the cluster's `ceph.conf` is kept to the `/etc/ceph` directory of the other hosts in your cluster.

If your Ceph cluster is older than Luminous, you will be unable to use the `ceph osd purge` command. Instead, carry out the following procedure:

1. Remove the OSD from the CRUSH map so that it no longer receives data (for more details, see Removing an OSD):

```
$ ceph osd crush remove {name}
```

   Instead of removing the OSD from the CRUSH map, you might opt for one o
   alternatives: (1) decompile the CRUSH map, remove the OSD from the devic
   remove the device from the host bucket; (2) remove the host bucket from the CRUSH map

map, and set it:

2. Remove the OSD authentication key:

```
$ ceph auth del osd.{osd-num}
```

3. Remove the OSD:

```
$ ceph osd rm {osd-num}
```

For example:

```
$ ceph osd rm 1
```