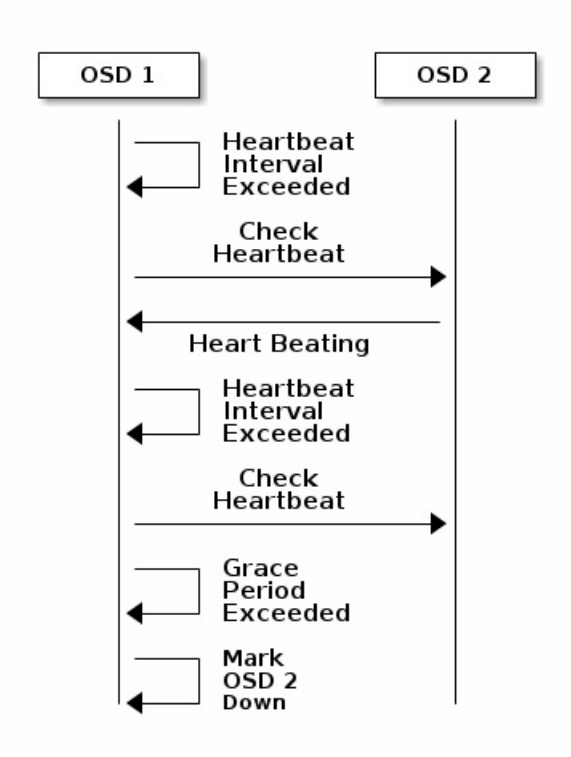
CONFIGURING MONITOR/OSD INTERACTION

After you have completed your initial Ceph configuration, you may deploy and run Ceph. When you execute a command such as `ceph health` or `ceph -s`, the Ceph Monitor reports on the current state of the Ceph Storage Cluster. The Ceph Monitor knows about the Ceph Storage Cluster by requiring reports from each Ceph OSD Daemon, and by receiving reports from Ceph OSD Daemons about the status of their neighboring Ceph OSD Daemons. If the Ceph Monitor doesn't receive reports, or if it receives reports of changes in the Ceph Storage Cluster, the Ceph Monitor updates the status of the Ceph Cluster Map.

Ceph provides reasonable default settings for Ceph Monitor/Ceph OSD Daemon interaction. However, you may override the defaults. The following sections describe how Ceph Monitors and Ceph OSD Daemons interact for the purposes of monitoring the Ceph Storage Cluster.

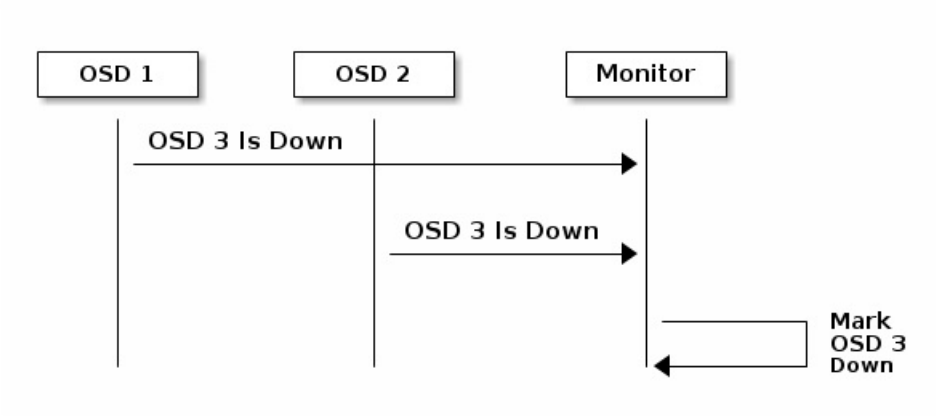
OSDS CHECK HEARTBEATS

Each Ceph OSD Daemon checks the heartbeat of other Ceph OSD Daemons every 6 seconds. You can change the heartbeat interval by adding an `osd heartbeat interval` setting under the `[osd]` section of your Ceph configuration file, or by setting the value at runtime. If a neighboring Ceph OSD Daemon doesn't show a heartbeat within a 20 second grace period, the Ceph OSD Daemon may consider the neighboring Ceph OSD Daemon down and report it back to a Ceph Monitor, which will update the Ceph Cluster Map. You may change this grace period by adding an `osd heartbeat grace` setting under the `[mon]` and `[osd]` or `[global]` section of your Ceph configuration file, or by setting the value at runtime.



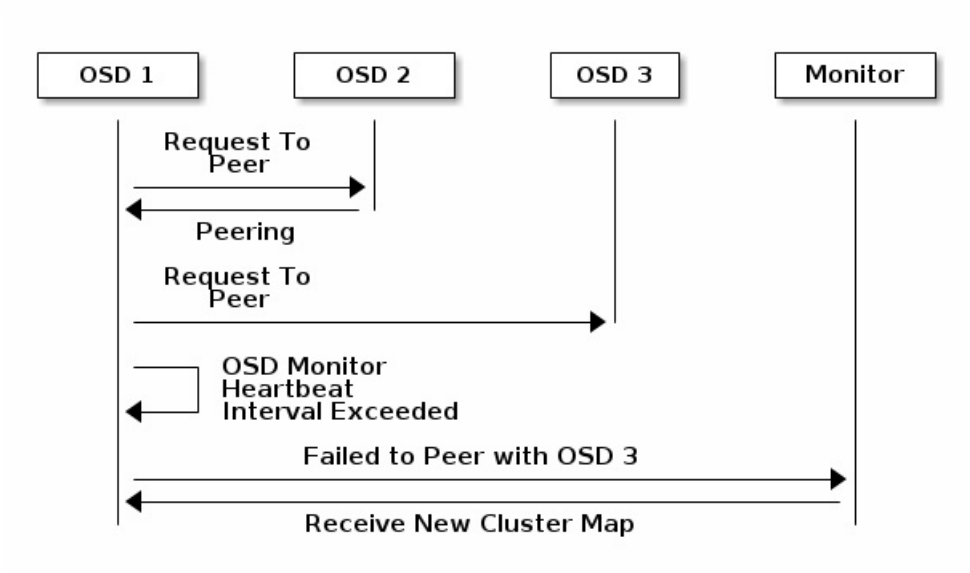
OSDS REPORT DOWN OSDS

By default, two Ceph OSD Daemons from different hosts must report to the Ceph Monitors that another Ceph OSD Daemon is down before the Ceph Monitors acknowledge that the reported Ceph OSD Daemon is down. But there is chance that all the OSDs reporting the failure are hosted in a rack with a bad switch which has trouble connecting to another OSD. To avoid this sort of false alarm, we consider the peers reporting a failure a proxy for a potential "subcluster" over the overall cluster that is similarly laggy. This is clearly not true in all cases, but will sometimes help us localize the grace correction to a subset of the system that is unhappy. `mon osd reporter subtree level` is used to group the peers into the "subcluster" by their common ancestor type in CRUSH map. By default, only two reports from different subtree are required to report another Ceph OSD Daemon down. You can change the number of reporters from unique subtrees and the common ancestor type required to report a Ceph OSD Daemon down to a Ceph Monitor by adding an `mon osd min down reporters` and `mon osd reporter subtree level` settings under the `[mon]` section of your Ceph configuration file, or by setting the value at runtime.



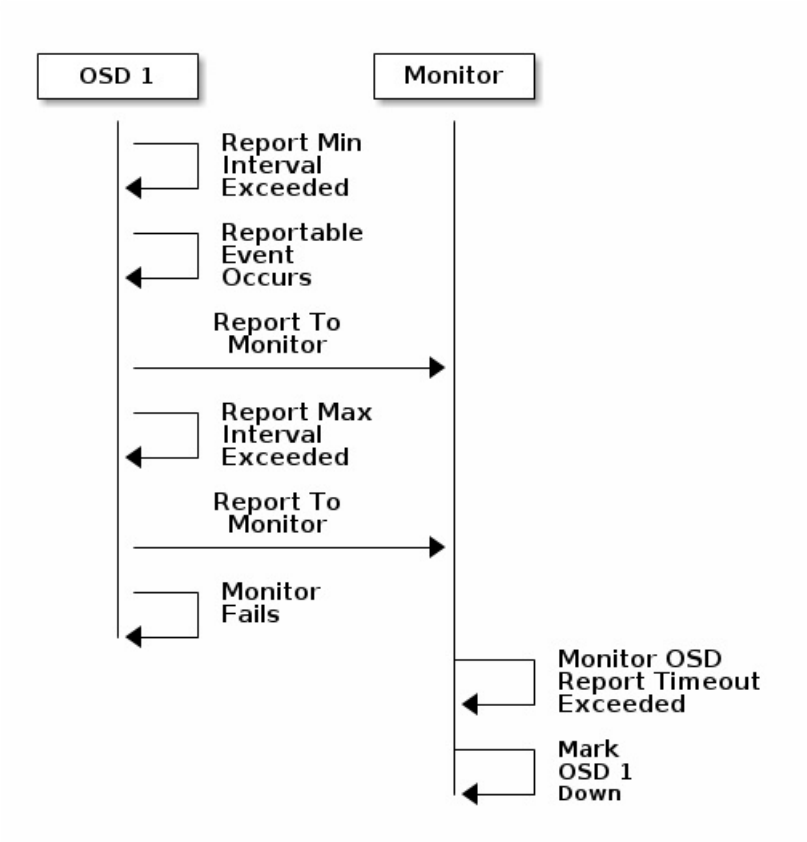
OSDS REPORT PEERING FAILURE

If a Ceph OSD Daemon cannot peer with any of the Ceph OSD Daemons defined in its Ceph configuration file (or the cluster map), it will ping a Ceph Monitor for the most recent copy of the cluster map every 30 seconds. You can change the Ceph Monitor heartbeat interval by adding an `osd mon heartbeat interval` setting under the `[osd]` section of your Ceph configuration file, or by setting the value at runtime.



OSDS REPORT THEIR STATUS

If an Ceph OSD Daemon doesn't report to a Ceph Monitor, the Ceph Monitor will consider the Ceph OSD Daemon down after the `mon osd report timeout` elapses. A Ceph OSD Daemon sends a report to a Ceph Monitor when a reportable event such as a failure, a change in placement group stats, a change in `up_thru` or when it boots within 5 seconds. You can change the Ceph OSD Daemon minimum report interval by adding an `osd mon report interval min` setting under the `[osd]` section of your Ceph configuration file, or by setting the value at runtime. A Ceph OSD Daemon sends a report to a Ceph Monitor every 120 seconds irrespective of whether any notable changes occur. You can change the Ceph Monitor report interval by adding an `osd mon report interval max` setting under the `[osd]` section of your Ceph configuration file, or by setting the value at runtime.



CONFIGURATION SETTINGS

When modifying heartbeat settings, you should include them in the `[global]` section of your configuration file.

MONITOR SETTINGS

`mon osd min up ratio`

Description: The minimum ratio of up Ceph OSD Daemons before Ceph will mark Ceph OSD Daemons down.
Type: Double
Default: `.3`

`mon osd min in ratio`

Description: The minimum ratio of in Ceph OSD Daemons before Ceph will mark Ceph OSD Daemons out.
Type: Double
Default: `.75`

`mon osd laggy halflife`

Description: The number of seconds laggy estimates will decay.
Type: Integer
Default: `60*60`

`mon osd laggy weight`

Description: The weight for new samples in laggy estimation decay.
Type: Double
Default: `0.3`

`mon osd laggy max interval`

Description: Maximum value of `laggy_interval` in laggy estimations (in seconds). Monitor uses an adaptive approach to evaluate the `laggy_interval` of a certain OSD. This value will be used to calculate the grace time for that OSD.
Type: Integer
Default: 300

`mon osd adjust heartbeat grace`

Description: If set to `true`, Ceph will scale based on laggy estimations.
Type: Boolean
Default: `true`

`mon osd adjust down out interval`

Description: If set to `true`, Ceph will scaled based on laggy estimations.
Type: Boolean
Default: `true`

`mon osd auto mark in`

Description: Ceph will mark any booting Ceph OSD Daemons as `in` the Ceph Storage Cluster.
Type: Boolean
Default: `false`

`mon osd auto mark auto out in`

Description: Ceph will mark booting Ceph OSD Daemons auto marked `out` of the Ceph Storage Cluster as `in` the cluster.
Type: Boolean
Default: `true`

`mon osd auto mark new in`

Description: Ceph will mark booting new Ceph OSD Daemons as `in` the Ceph Storage Cluster.
Type: Boolean
Default: `true`

`mon osd down out interval`

Description: The number of seconds Ceph waits before marking a Ceph OSD Daemon `down` and `out` if it doesn't respond.
Type: 32-bit Integer
Default: `600`

`mon osd down out subtree limit`

Description: The smallest CRUSH unit type that Ceph will **not** automatically mark out. For instance, if set to `host` and if all OSDs of a host are down, Ceph will not automatically mark out these OSDs.
Type: String
Default: `rack`

`mon osd report timeout`

Description: The grace period in seconds before declaring unresponsive Ceph OSD Daemons `down`.
Type: 32-bit Integer
Default: `900`

`mon osd min down reporters`

Description: The minimum number of Ceph OSD Daemons required to report a `down` Ceph OSD Daemon.
Type: 32-bit Integer
Default: 2

`mon osd reporter subtree level`

Description: In which level of parent bucket the reporters are counted. The OSDs send failure reports to monitor if they find its peer is not responsive. And monitor mark the reported OSD out and then down after a grace period.
Type: String
Default: `host`

OSD SETTINGS

`osd heartbeat address`

- **Description:** An Ceph OSD Daemon's network address for heartbeats.
- **Type:** Address
- **Default:** The host address.

`osd heartbeat interval`

- **Description:** How often an Ceph OSD Daemon pings its peers (in seconds).
- **Type:** 32-bit Integer
- **Default:** `6`

`osd heartbeat grace`

- **Description:** The elapsed time when a Ceph OSD Daemon hasn't shown a heartbeat that the Ceph Storage Cluster considers it `down`. This setting has to be set in both the [mon] and [osd] or [global] section so that it is read by both the MON and OSD daemons.
- **Type:** 32-bit Integer
- **Default:** `20`

`osd mon heartbeat interval`

- **Description:** How often the Ceph OSD Daemon pings a Ceph Monitor if it has no Ceph OSD Daemon peers.
- **Type:** 32-bit Integer
- **Default:** `30`

`osd mon report interval max`

- **Description:** The maximum time in seconds that a Ceph OSD Daemon can wait before it must report to a Ceph Monitor.
- **Type:** 32-bit Integer
- **Default:** `120`

`osd mon report interval min`

- **Description:** The minimum number of seconds a Ceph OSD Daemon may wait from startup or another reportable event before reporting to a Ceph Monitor.
- **Type:** 32-bit Integer
- **Default:** `5`
- **Valid Range:** Should be less than `osd mon report interval max`

`osd mon ack timeout`

- **Description:** The number of seconds to wait for a Ceph Monitor to acknowledge a request for statistics.
- **Type:** 32-bit Integer
- **Default:** `30`