Interpreting OSD Status

An OSD daemon can be in one of four states, based on the combination of these two flags:

- **down** or **up** - indicating whether the daemon is running and communicating with the MONs.
- **out** or **in** - indicating whether the OSD is participating in cluster data placement.

The state of an OSD in normal operation is up and in.

If an OSD fails and the daemon goes offline, the cluster might report it as down and in for a short period of time. This is intended to give the OSD a chance to recover on its own and rejoin the cluster, avoiding unnecessary recovery traffic.

For example, a brief network interruption might cause the OSD to lose communication with the cluster and be temporarily reported as down. After a short interval controlled by the mon_osd_down_out_interval configuration option (five minutes by default), the cluster reports the OSD as down and out. At this point, the placement groups assigned to the failed OSD are migrated to other OSDs.

If the failed OSD then returns to the up and in states, the cluster reassigns placement groups based on the new set of OSDs and by rebalancing the objects in the cluster.

Note

Use the ceph osd set noout and ceph osd unset noout commands to enable or disable the noout flag on the cluster. However, the ceph osd out *osdid* command tells the Ceph cluster to ignore an OSD for data placement and marks the OSD with the out state.

OSDs verify each other's status at regular time intervals (six seconds by default). They report their status to the MONs every 120 seconds, by default. If an OSD is down, the other OSDs or the MONs do not receive heartbeat responses from that down OSD.

The following configuration settings manage OSD heartbeats:

Configuration option	Description
osd_heartbeat_interval	Number of seconds between OSD peer checks.
osd_heartbeat_grace	Number of seconds before an unresponsive OSD moves to the down state.
mon_osd_min_down_reporters	Number of peers reporting that an OSD is down before a MON considers it to be down.
mon_osd_min_down_reports	Number of times an OSD is reported to be down before a MON considers it to be down.
mon_osd_down_out_subtree_limit	Prevents a CRUSH unit type (such as a host) from being automatically marked as out when it fails.
osd_mon_report_interval_min	A newly booted OSD has to report to a MON within this number of seconds.
osd_mon_report_interval_max	Maximum number of seconds between reports from an OSD to a MON.