TERMINOLOGY

A Ceph cluster may have zero or more CephFS *filesystems*. CephFS filesystems have a human readable name (set in `fs new`) and an integer ID. The ID is called the filesystem cluster ID, or *FSCID*.

Each CephFS filesystem has a number of *ranks*, one by default, which start at zero. A rank may be thought of as a metadata shard. Controlling the number of ranks in a filesystem is described in Configuring multiple active MDS daemons

Each CephFS ceph-mds process (a *daemon*) initially starts up without a rank. It may be assigned one by the monitor cluster. A daemon may only hold one rank at a time. Daemons only give up a rank when the ceph-mds process stops.

If a rank is not associated with a daemon, the rank is considered *failed*. Once a rank is assigned to a daemon, the rank is considered *up*.

A daemon has a *name* that is set statically by the administrator when the daemon is first configured. Typical configurations use the hostname where the daemon runs as the daemon name.

Each time a daemon starts up, it is also assigned a *GID*, which is unique to this particular process lifetime of the daemon. The GID is an integer.

REFERRING TO MDS DAEMONS

Most of the administrative commands that refer to an MDS daemon accept a flexible argument format that may contain a rank, a GID or a name.

Where a rank is used, this may optionally be qualified with a leading filesystem name or ID. If a daemon is a standby (i.e. it is not currently assigned a rank), then it may only be referred to by GID or name.

For example, if we had an MDS daemon which was called 'myhost', had GID 5446, and was assigned rank 0 in the filesystem 'myfs' which had FSCID 3, then any of the following would be suitable forms of the 'fail' command:

```
ceph mds fail 5446     # GID
ceph mds fail myhost   # Daemon name
ceph mds fail 0        # Unqualified rank
ceph mds fail 3:0      # FSCID and rank
ceph mds fail myfs:0   # Filesystem name and rank
```

MANAGING FAILOVER

If an MDS daemon stops communicating with the monitor, the monitor will wait `mds_beacon_grace` seconds (default 15 seconds) before marking the daemon as *laggy*.

Each file system may specify a number of standby daemons to be considered healthy. This number includes daemons in standby-replay waiting for a rank to fail (remember that a standby-replay daemon will not be assigned to take over a failure for another rank or a failure in a another CephFS file system). The pool of standby daemons not in replay count towards any file system count. Each file system may set the number of standby daemons wanted using:

```
ceph fs set <fs name> standby_count_wanted <count>
```

Setting `count` to 0 will disable the health check.

CONFIGURING STANDBY DAEMONS

There are four configuration settings that control how a daemon will behave while in standby:

```
mds_standby_for_name
mds_standby_for_rank
mds_standby_for_fscid
```

```
mds_standby_replay
```

These may be set in the ceph.conf on the host where the MDS daemon runs (as opposed to on the monitor). The daemon loads these settings when it starts, and sends them to the monitor.

By default, if none of these settings are used, all MDS daemons which do not hold a rank will be used as standbys for any rank.

The settings which associate a standby daemon with a particular name or rank do not guarantee that the daemon will *only* be used for that rank. They mean that when several standbys are available, the associated standby daemon will be used. If a rank is failed, and a standby is available, it will be used even if it is associated with a different rank or named daemon.

MDS_STANDBY_REPLAY

If this is set to true, then the standby daemon will continuously read the metadata journal of an up rank. This will give it a warm metadata cache, and speed up the process of failing over if the daemon serving the rank fails.

An up rank may only have one standby replay daemon assigned to it, if two daemons are both set to be standby replay then one of them will arbitrarily win, and the other will become a normal non-replay standby.

Once a daemon has entered the standby replay state, it will only be used as a standby for the rank that it is following. If another rank fails, this standby replay daemon will not be used as a replacement, even if no other standbys are available.

Historical note: In Ceph prior to v10.2.1, this setting (when `false`) is always true when `mds_standby_for_*` is also set.

MDS_STANDBY_FOR_NAME

Set this to make the standby daemon only take over a failed rank if the last daemon to hold it matches this name.

MDS_STANDBY_FOR_RANK

Set this to make the standby daemon only take over the specified rank. If another rank fails, this daemon will not be used to replace it.

Use in conjunction with `mds_standby_for_fscid` to be specific about which filesystem's rank you are targeting, if you have multiple filesystems.

MDS_STANDBY_FOR_FSCID

If `mds_standby_for_rank` is set, this is simply a qualifier to say which filesystem's rank is referred to.

If `mds_standby_for_rank` is not set, then setting FSCID will cause this daemon to target any rank in the specified FSCID. Use this if you have a daemon that you want to use for any rank, but only within a particular filesystem.

MON_FORCE_STANDBY_ACTIVE

This setting is used on monitor hosts. It defaults to true.

If it is false, then daemons configured with standby_replay=true will **only** become active if the rank/name that they have been configured to follow fails. On the other hand, if this setting is true, then a daemon configured with standby_replay=true may be assigned some other rank.

EXAMPLES

These are example ceph.conf snippets. In practice you can either copy a ceph.conf with all daemons' configuration to all your servers, or you can have a different file on each server that contains just that server's daemons' configuration.

SIMPLE PAIR

Two MDS daemons 'a' and 'b' acting as a pair, where whichever one is not currently assigned a rank will be the standby replay

follower of the other.

```
[mds.a]
mds standby replay = true
mds standby for rank = 0

[mds.b]
mds standby replay = true
mds standby for rank = 0
```

FLOATING STANDBY

Three MDS daemons ‘a’, ‘b’ and ‘c’, in a filesystem that has `max_mds` set to 2.

```
# No explicit configuration required: whichever daemon is
# not assigned a rank will go into 'standby' and take over
# for whichever other daemon fails.
```

TWO MDS CLUSTERS

With two filesystems, I have four MDS daemons, and I want two to act as a pair for one filesystem and two to act as a pair for the other filesystem.

```
[mds.a]
mds standby for fscid = 1

[mds.b]
mds standby for fscid = 1

[mds.c]
mds standby for fscid = 2

[mds.d]
mds standby for fscid = 2
```