RECOVERY RESERVATION

Recovery reservation extends and subsumes backfill reservation. The reservation system from backfill recovery is used for local and remote reservations.

When a PG goes active, first it determines what type of recovery is necessary, if any. It may need log-based recovery, backfill recovery, both, or neither.

In log-based recovery, the primary first acquires a local reservation from the OSDService's local_reserver. Then a MRemoteReservationRequest message is sent to each replica in order of OSD number. These requests will always be granted (i.e., cannot be rejected), but they may take some time to be granted if the remotes have already granted all their remote reservation slots.

After all reservations are acquired, log-based recovery proceeds as it would without the reservation system.

After log-based recovery completes, the primary releases all remote reservations. The local reservation remains held. The primary then determines whether backfill is necessary. If it is not necessary, the primary releases its local reservation and waits in the Recovered state for all OSDs to indicate that they are clean.

If backfill recovery occurs after log-based recovery, the local reservation does not need to be reacquired since it is still held from before. If it occurs immediately after activation (log-based recovery not possible/necessary), the local reservation is acquired according to the typical process.

Once the primary has its local reservation, it requests a remote reservation from the backfill target. This reservation CAN be rejected, for instance if the OSD is too full (backfillfull_ratio osd setting). If the reservation is rejected, the primary drops its local reservation, waits (osd_backfill_retry_interval), and then retries. It will retry indefinitely.

Once the primary has the local and remote reservations, backfill proceeds as usual. After backfill completes the remote reservation is dropped.

Finally, after backfill (or log-based recovery if backfill was not necessary), the primary drops the local reservation and enters the Recovered state. Once all the PGs have reported they are clean, the primary enters the Clean state and marks itself active+clean.

THINGS TO NOTE

We always grab the local reservation first, to prevent a circular dependency. We grab remote reservations in order of OSD number for the same reason.

The recovery reservation state chart controls the PG state as reported to the monitor. The state chart can set:

- recovery_wait: waiting for local/remote reservations
- recovering: recovering
- recovery_toofull: recovery stopped, OSD(s) above full ratio
- backfill_wait: waiting for remote backfill reservations
- backfilling: backfilling
- backfill_toofull: backfill stopped, OSD(s) above backfillfull ratio

SEE ALSO

The Active substate of the automatically generated OSD state diagram.