PLACEMENT GROUP CONCEPTS

When you execute commands like `ceph -w`, `ceph osd dump`, and other commands related to placement groups, Ceph may return values using some of the following terms:

Peering

The process of bringing all of the OSDs that store a Placement Group (PG) into agreement about the state of all of the objects (and their metadata) in that PG. Note that agreeing on the state does not mean that they all have the latest contents.

Acting Set

The ordered list of OSDs who are (or were as of some epoch) responsible for a particular placement group.

Up Set

The ordered list of OSDs responsible for a particular placement group for a particular epoch according to CRUSH. Normally this is the same as the *Acting Set*, except when the *Acting Set* has been explicitly overridden via `pg_temp` in the OSD Map.

Current Interval or *Past Interval*

A sequence of OSD map epochs during which the *Acting Set* and *Up Set* for particular placement group do not change.

Primary

The member (and by convention first) of the *Acting Set*, that is responsible for coordination peering, and is the only OSD that will accept client-initiated writes to objects in a placement group.

Replica

A non-primary OSD in the *Acting Set* for a placement group (and who has been recognized as such and *activated* by the primary).

Stray

An OSD that is not a member of the current *Acting Set*, but has not yet been told that it can delete its copies of a particular placement group.

Recovery

Ensuring that copies of all of the objects in a placement group are on all of the OSDs in the *Acting Set*. Once *Peering* has been performed, the *Primary* can start accepting write operations, and *Recovery* can proceed in the background.

PG Info

Basic metadata about the placement group's creation epoch, the version for the most recent write to the placement group, *last epoch started*, *last epoch clean*, and the beginning of the *current interval*. Any inter-OSD communication about placement groups includes the *PG Info*, such that any OSD that knows a placement group exists (or once existed) also has a lower bound on *last epoch clean* or *last epoch started*.

PG Log

A list of recent updates made to objects in a placement group. Note that these logs can be truncated after all OSDs in the *Acting Set* have acknowledged up to a certain point.

Missing Set

Each OSD notes update log entries and if they imply updates to the contents of an object, adds that object to a list of needed updates. This list is called the *Missing Set* for that `<OSD,PG>`.

Authoritative History

A complete, and fully ordered set of operations that, if performed, would bring an OSD's copy of a placement group up to date.

Epoch

A (monotonically increasing) OSD map version number

Last Epoch Start

The last epoch at which all nodes in the *Acting Set* for a particular placement group agreed on an *Authoritative History*. At this point, *Peering* is deemed to have been successful.

up_thru

Before a *Primary* can successfully complete the *Peering* process, it must inform a monitor that is alive through the current OSD map *Epoch* by having the monitor set its *up_thru* in the osd map. This helps *Peering* ignore previous *Acting Sets* for which *Peering* never completed after certain sequences of failures, such as the second interval below:

- *acting set* = [A,B]
- *acting set* = [A]
- *acting set* = [] very shortly after (e.g., simultaneous failure, but staggered detection)
- *acting set* = [B] (B restarts, A does not)

Last Epoch Clean

The last *Epoch* at which all nodes in the *Acting set* for a particular placement group were completely up to date (both placement group logs and object contents). At this point, *recovery* is deemed to have been completed.