

During the cluster life cycle, the number of PGs must be adjusted as the cluster layout changes. CRUSH attempts to ensure a uniform distribution of objects among OSDs in the pool, but there are scenarios where the PGs become unbalanced. The placement group autoscaler can be used to optimize PG distribution, and is on by default. You can also manually set the number of PGs per pool, if required.

Objects are typically distributed uniformly, provided that there are one or two orders of magnitude (factors of ten) more placement groups than OSDs in the pool. If there are not enough PGs, then objects might be distributed unevenly. If there is a small number of very large objects stored in the pool, then object distribution might become unbalanced.



Note

PGs should be configured so that there are enough to evenly distribute objects across the cluster. If the number of PGs is set too high, then it increases CPU and memory use significantly. Red Hat recommends approximately 100 to 200 placement groups per OSD to balance these factors.

Calculating the Number of Placement Groups

For a cluster with a single pool, you can use the following formula, with 100 placement groups per OSD:

$$\text{Total PGs} = (\text{OSDs} * 100) / \text{Number of replicas}$$

Red Hat recommends the use of the Ceph Placement Groups per Pool Calculator, <https://access.redhat.com/labs/cephpgc/>, from the Red Hat Customer Portal Labs.

Mapping PGs Manually

Use the `ceph osd pg-upmap-items` command to manually map PGs to specific OSDs. Because older Ceph clients do not support it, you must configure the `ceph osd set-require-min-compat-client` setting to enable the `pg-upmap` command.

```
[ceph: root@node /]# ceph osd set-require-min-compat-client luminous
set require_min_compat_client to luminous
```

The following example remaps the PG 3.25 from OSDs 2 and 0 to 1 and 0:

```
[ceph: root@node /]# ceph pg map 3.25
osdmap e384 pg 3.25 (3.25) -> up [2,0] acting [2,0]
[ceph: root@node /]# ceph osd pg-upmap-items 3.25 2 1
set 3.25 pg_upmap_items mapping to [2->1]
[ceph: root@node /]# ceph pg map 3.25
osdmap e387 pg 3.25 (3.25) -> up [1,0] acting [1,0]
```

Remapping hundreds of PGs this way is not practical. The `osdmapprool` command is useful here. It takes the actual map for a pool, analyses it, and generates the `ceph osd pg-upmap-items` commands to run for an optimal distribution:

1. Export the map to a file. The following command saves the map to the `./om` file: