storage devices, to increase throughput and performance. Ceph clients can use data striping to increase performance when writing data to the cluster.

Data Distribution and Organization in Ceph

This section describes the mechanisms that Ceph uses to distribute and organize data across the various storage devices in a cluster.

Partitioning Storage with Pools

Ceph OSDs protect and constantly check the integrity of the data that is stored in the cluster. *Pools* are logical partitions of the Ceph storage cluster that are used to store objects under a common name tag. Ceph assigns each pool a specific number of hash buckets, called *Placement Groups (PGs)*, to group objects for storage.

Each pool has the following adjustable properties:

- Immutable ID
- Name
- Number of PGs to distribute the objects across the OSDs
- CRUSH rule to determine the mapping of the PGs for this pool
- Protection type (replicated or erasure coding)
- Parameters that are associated with the protection type
- Various flags to influence the cluster behavior

Configure the number of placement groups that are assigned to each pool independently to fit the type of data and the required access for the pool.

The CRUSH algorithm determines the OSDs that host the data for a pool. Each pool has a single CRUSH rule that is assigned as its placement strategy. The CRUSH rule determines which OSDs store the data for all the pools that are assigned that rule.

Placement Groups

A *Placement Group (PG)* aggregates a series of objects into a hash bucket, or group. Ceph maps each PG to a set of OSDs. An object belongs to a single PG, and all objects that belong to the same PG return the same hash result.

The CRUSH algorithm maps an object to its PG based on the hashing of the object's name. The placement strategy is also called the CRUSH placement rule. The placement rule identifies the failure domain to choose within the CRUSH topology to receive each replica or erasure code chunk.

When a client writes an object to a pool, it uses the pool's CRUSH placement rule to determine the object's placement group. The client then uses its copy of the cluster map, the placement group, and the CRUSH placement rule to calculate which OSDs to write a copy of the object to (or its erasure-coded chunks).

The layer of indirection that the placement group provides is important when new OSDs become available to the Ceph cluster. When OSDs are added to or removed from a cluster, placement groups are automatically rebalanced between operational OSDs.

Mapping an Object to Its Associated OSDs

A Ceph client gets the latest copy of the cluster map from a MON. The cluster map provides information to the client about all the MONs, OSDs, and MDSs in the cluster. It does not provide