Administration of Red Hat Ceph Storage
Abstract

This document describes how to manage processes, monitor cluster states, manage users, and add and remove daemons for Red Hat Ceph Storage.
# Table of Contents

CHAPTER 1. CEPH ADMINISTRATION OVERVIEW ............................................. 5.
  Additional Resources ............................................................... 5

CHAPTER 2. UNDERSTANDING PROCESS MANAGEMENT FOR CEPH ................ 6.
  2.1. PREREQUISITES ............................................................... 6
  2.2. AN OVERVIEW OF PROCESS MANAGEMENT FOR CEPH ................. 6
     Additional Resources ......................................................... 6
  2.3. STARTING, STOPPING, AND RESTARTING ALL THE CEPH DAEMONS ........ 6
     Prerequisites ................................................................. 6
     Procedure ................................................................. 6
  2.4. STARTING, STOPPING, AND RESTARTING THE CEPH DAEMONS BY TYPE .... 7
     Prerequisites ................................................................. 7
     Procedure ................................................................. 7
  2.5. STARTING, STOPPING, AND RESTARTING A CEPH DAEMONS BY INSTANCE ... 8
     Prerequisites ................................................................. 8
     Procedure ................................................................. 8
  2.6. ADDITIONAL RESOURCES ................................................... 9

CHAPTER 3. MONITORING ................................................................. 10.
  3.1. HIGH-LEVEL MONITORING .................................................... 10
     3.1.1. Interactive Mode ....................................................... 10
     3.1.2. Checking Cluster Health ........................................... 10
     3.1.3. Watching a Cluster .................................................. 10
     3.1.4. Checking a Cluster’s Usage Statistics ................................ 12
     3.1.5. Checking a Cluster’s Status ........................................ 12
     3.1.6. Checking Monitor Status ............................................ 13
     3.1.7. Using the Administration Socket .................................... 14
     3.1.8. Checking OSD Status ................................................ 15
  3.2. LOW-LEVEL MONITORING .................................................... 16
     3.2.1. Placement Group Sets ............................................... 16
     3.2.2. Peering ............................................................... 17
     3.2.3. Monitoring Placement Group States ................................ 17
        3.2.3.1. Creating ......................................................... 21
        3.2.3.2. Peering ......................................................... 21
        3.2.3.3. Active .......................................................... 22
        3.2.3.4. Clean .......................................................... 22
        3.2.3.5. Degraded ....................................................... 22
        3.2.3.6. Recovering .................................................... 22
        3.2.3.7. Backfilling .................................................... 23
        3.2.3.8. Remapped ...................................................... 23
        3.2.3.9. Stale ............................................................ 23
        3.2.3.10. Misplaced ..................................................... 24
        3.2.3.11. Incomplete .................................................... 24
     3.2.4. Identifying Troubled Placement Groups ................................ 24
     3.2.5. Finding an Object Location .......................................... 25
  3.3. MONITORING A CEPH STORAGE CLUSTER WITH THE RED HAT CEPH STORAGE DASHBOARD 25
     Prerequisites ................................................................. 26
     3.3.1. The Red Hat Ceph Storage Dashboard .................................. 26
     3.3.2. Installing the Red Hat Ceph Storage Dashboard ..................... 28
     3.3.3. Accessing the Red Hat Ceph Storage Dashboard .................... 30
        Prerequisites ............................................................. 30
        Procedure .............................................................. 30
3.3.4. Changing the default Red Hat Ceph Storage dashboard password 31
3.3.5. The Prometheus plugin for Red Hat Ceph Storage 32
3.3.5.1. Prerequisites 32
3.3.5.2. The Prometheus plugin 32
3.3.5.3. Managing the Prometheus environment 32
3.3.5.4. Using the Prometheus expression browser 34
3.3.5.5. Working with the Prometheus data and queries 34
3.3.5.6. Additional Resources 36
3.3.6. The Red Hat Ceph Storage Dashboard alerts 36
3.3.6.1. Prerequisites 36
3.3.6.2. About Alerts 36
3.3.6.3. Accessing the Alert Status dashboard 37
   Procedure 37
3.3.6.4. Configuring the Notification Target 37
   Procedure 37
3.3.6.5. Changing the Default Alerts and Adding New Ones 37
   Procedure 37
Additional Resources 38

CHAPTER 4. OVERRIDES .............................................................. 39
4.1. SETTING AND UNSETTING OVERRIDES 39
4.2. USE CASES 40

CHAPTER 5. USER MANAGEMENT ............................................. 41
5.1. BACKGROUND 41
   5.1.1. User 41
   5.1.2. Authorization (Capabilities) 42
   5.1.3. Pool 43
   5.1.4. Namespace 43
5.2. MANAGING USERS 44
   5.2.1. List Users 44
   5.2.2. Get a User 45
   5.2.3. Add a User 45
   5.2.4. Modify User Capabilities 46
   5.2.5. Delete a User 46
   5.2.6. Print a User’s Key 47
   5.2.7. Import a User 47
5.3. KEYRING MANAGEMENT 47
   5.3.1. Create a Keyring 48
   5.3.2. Add a User to a Keyring 48
   5.3.3. Create a User 49
   5.3.4. Modify a User 49
5.4. COMMAND LINE USAGE 49
5.5. LIMITATIONS 50

CHAPTER 6. USING THE CEPH-VOLUME UTILITY TO DEPLOY OSDS .............................................. 52
6.1. USING THE CEPH-VOLUME LVM PLUGIN 52
   6.1.1. Preparing OSDs 52
   6.1.2. Activating OSDs 54
   6.1.3. Creating OSDs 54
   6.1.4. Using batch mode 55

CHAPTER 7. BENCHMARKING PERFORMANCE ...................................... 56
CHAPTER 8. PERFORMANCE COUNTERS ................................................................. 62
  8.1. ACCESS ........................................................................................................ 62
  8.2. SCHEMA ....................................................................................................... 62
  8.3. DUMP ............................................................................................................ 64
      8.3.1. Average Count and Sum ...................................................................... 65
      8.3.2. Monitor Metrics Description Tables .................................................... 65
      8.3.3. OSD Metrics Description Tables ......................................................... 70
      8.3.4. The Ceph Object Gateway Metrics Tables ......................................... 82

CHAPTER 9. BLUESTORE ...................................................................................... 87
  9.1. ABOUT BLUESTORE ................................................................................. 87
  9.2. BLUESTORE DEVICES ............................................................................ 88
  9.3. SIZING CONSIDERATIONS FOR BLUESTORE .................................. 88
  9.4. ADDING OSDS THAT USE BLUESTORE ............................................ 89
      Prerequisites .................................................................................................. 89
      Procedure ..................................................................................................... 89
      Additional Resources .................................................................................... 93
A Red Hat Ceph Storage cluster is the foundation for all Ceph deployments. After deploying a Red Hat Ceph Storage cluster, there are administrative operations for keeping a Red Hat Ceph Storage cluster healthy and performing optimally.

The Red Hat Ceph Storage Administration Guide helps storage administrators to perform such tasks as:

- How do I check the health of my Red Hat Ceph Storage cluster?
- How do I start and stop the Red Hat Ceph Storage cluster services?
- How do I add or remove an OSD from a running Red Hat Ceph Storage cluster?
- How do I manage user authentication and access controls to the objects stored in a Red Hat Ceph Storage cluster?
- I want to understand how to use overrides with a Red Hat Ceph Storage cluster.
- I want to monitor the performance of the Red Hat Ceph Storage cluster.

A basic Ceph Storage Cluster consist of two types of daemons:

- A Ceph Object Storage Device (OSD) stores data as objects within placement groups assigned to the OSD
- A Ceph Monitor maintains a master copy of the cluster map

A production system will have three or more Ceph Monitors for high availability and typically a minimum of 50 OSDs for acceptable load balancing, data re-balancing and data recovery.

**Additional Resources**

- Red Hat Ceph Storage Installation Guide for:
  - Red Hat Enterprise Linux
  - Ubuntu
CHAPTER 2. UNDERSTANDING PROCESS MANAGEMENT FOR CEPH

As a storage administrator, you can manipulate the Ceph daemons in various ways. Manipulating these daemons allows you to start, stop and restart all of the Ceph services as needed.

- Starting, Stopping, and Restarting All Ceph Daemons
- Starting, Stopping, Restarting a Ceph Daemon by Type
- Starting, Stopping, Restarting a Ceph Daemon by Instance

2.1. PREREQUISITES

- A running Red Hat Ceph Storage cluster.

2.2. AN OVERVIEW OF PROCESS MANAGEMENT FOR CEPH

In Red Hat Ceph Storage 3, all process management is done through the Systemd service. Each time you want to start, restart, and stop the Ceph daemons, you must specify the daemon type or the daemon instance.

Additional Resources

- For more information about using Systemd, see Chapter 9 in the Red Hat Enterprise Linux System Administrator’s Guide.

2.3. STARTING, STOPPING, AND RESTARTING ALL THE CEPH DAEMONS

To start, stop, or restart all the running Ceph daemons on a node, follow these procedures.

Prerequisites

- Having root access to the node.

Procedure

- Starting all Ceph daemons:

  [root@admin ~]# systemctl start ceph.target

- Stopping all Ceph daemons:

  [root@admin ~]# systemctl stop ceph.target

- Restarting all Ceph daemons:

  [root@admin ~]# systemctl restart ceph.target
2.4. STARTING, STOPPING, AND RESTARTING THE CEPH DAEMONS BY TYPE

To start, stop, or restart all Ceph daemons of a particular type, follow these procedures on the node running the Ceph daemons.

Prerequisites
- Having *root* access to the node.

Procedure
- **On Ceph Monitor nodes:**
  - Starting
    ```
    [root@mon ~]# systemctl start ceph-mon.target
    ```
  - Stopping
    ```
    [root@mon ~]# systemctl stop ceph-mon.target
    ```
  - Restarting
    ```
    [root@mon ~]# systemctl restart ceph-mon.target
    ```
- **On Ceph OSD nodes:**
  - Starting
    ```
    [root@osd ~]# systemctl start ceph-osd.target
    ```
  - Stopping
    ```
    [root@osd ~]# systemctl stop ceph-osd.target
    ```
  - Restarting
    ```
    [root@osd ~]# systemctl restart ceph-osd.target
    ```
- **On Ceph Object Gateway nodes:**
  - Starting
    ```
    [root@rgw ~]# systemctl start ceph-radosgw.target
    ```
  - Stopping
    ```
    [root@rgw ~]# systemctl stop ceph-radosgw.target
    ```
  - Restarting
    ```
    [root@rgw ~]# systemctl restart ceph-radosgw.target
    ```
2.5. STARTING, STOPPING, AND RESTARTING A CEPH DAEMONS BY INSTANCE

To start, stop, or restart a Ceph daemon by instance, follow these procedures on the node running the Ceph daemons.

Prerequisites

- Having root access to the node.

Procedure

- On a Ceph Monitor node:
  
  Starting
  
  [root@mon ~]# systemctl start ceph-mon@$MONITOR_HOST_NAME

  Stopping
  
  [root@mon ~]# systemctl stop ceph-mon@$MONITOR_HOST_NAME

  Restarting
  
  [root@mon ~]# systemctl restart ceph-mon@$MONITOR_HOST_NAME

  Replace
  
  - $MONITOR_HOST_NAME with the name of the Ceph Monitor node.

- On a Ceph OSD node:
  
  Starting
  
  [root@osd ~]# systemctl start ceph-osd@$OSD_NUMBER

  Stopping
  
  [root@osd ~]# systemctl stop ceph-osd@$OSD_NUMBER

  Restarting
  
  [root@osd ~]# systemctl restart ceph-osd@$OSD_NUMBER

  Replace
  
  - $OSD_NUMBER with the ID number of the Ceph OSD.
    
    For example, when looking at the ceph osd tree command output, osd.0 has an ID of 0.
● On a Ceph Object Gateway node:

Starting

```
[root@rgw ~]# systemctl start ceph-radosgw@rgw.$OBJ_GATEWAY_HOST_NAME
```

Stopping

```
[root@rgw ~]# systemctl stop ceph-radosgw@rgw.$OBJ_GATEWAY_HOST_NAME
```

Restarting

```
[root@rgw ~]# systemctl restart ceph-radosgw@rgw.$OBJ_GATEWAY_HOST_NAME
```

Replace

- Replace $OBJ_GATEWAY_HOST_NAME with the name of the Ceph Object Gateway node.

2.6. ADDITIONAL RESOURCES

● For more information about installing Red Hat Ceph Storage, see:

  - Installation Guide for Red Hat Enterprise Linux
  - Installation Guide for Ubuntu
CHAPTER 3. MONITORING

Once you have a running cluster, you may begin monitoring the storage cluster to ensure that the Ceph Monitor and OSD daemons are running, at a high-level. Ceph storage cluster clients must connect to a Ceph monitor and receive the latest version of the Ceph cluster map before they can read and write data to the Ceph pools of the storage cluster. So the monitor cluster must have agreement on the state of the cluster before Ceph clients can read and write data.

Ceph OSDs must peer the placement groups on the primary OSD with the copies of the placement groups on secondary OSDs. If faults arise, peering will reflect something other than the active + clean state.

3.1. HIGH-LEVEL MONITORING

High-level monitoring of a storage cluster typically involves checking the status of Ceph OSD and Monitor daemons to ensure that they are up and running. High level monitoring also involves checking the storage cluster capacity to ensure that the cluster doesn’t exceed its full ratio. The Calamari instance on the Ansible Tower or Red Hat Storage Console node is the most common way to conduct high-level monitoring. However, you may also use the command line, the admin socket or the Ceph API to monitor the storage cluster.

3.1.1. Interactive Mode

To run the ceph utility in interactive mode, type ceph at the command line with no arguments, for example:

```
# ceph
ceph> health
ceph> status
ceph> quorum_status
ceph> mon_status
```

3.1.2. Checking Cluster Health

After you start the Ceph storage cluster, and before you start reading and/or writing data, check the storage cluster’s health first. You can check on the health of the Ceph storage cluster with the following:

```
# ceph health
```

If you specified non-default locations for the configuration or keyring, you may specify their locations:

```
# ceph -c /path/to/conf -k /path/to/keyring health
```

Upon starting the Ceph cluster, you will likely encounter a health warning such as HEALTH_WARN XXX num placement groups stale. Wait a few moments and check it again. When the storage cluster is ready, ceph health should return a message such as HEALTH_OK. At that point, it is okay to begin using the cluster.

3.1.3. Watching a Cluster

To watch the cluster’s ongoing events on the command line, open a new terminal. Then, enter:
Ceph will print each event. For example, a tiny Ceph cluster consisting of one monitor and two OSDs may print the following:

```
# ceph -w

Ceph will print each event. For example, a tiny Ceph cluster consisting of one monitor and two OSDs may print the following:

```

The output provides:

- Cluster ID
- Cluster health status
- The monitor map epoch and the status of the monitor quorum
- The OSD map epoch and the status of OSDs
- The placement group map version
- The number of placement groups and pools
- The *notional* amount of data stored and the number of objects stored
- The total amount of data stored

**How Ceph Calculates Data Usage**

The *used* value reflects the *actual* amount of raw storage used. The **xxx GB / xxx GB** value means the amount available, the lesser of the two numbers, of the overall storage capacity of the cluster. The
notional number reflects the size of the stored data before it is replicated, cloned or snapshotted. Therefore, the amount of data actually stored typically exceeds the notional amount stored, because Ceph creates replicas of the data and may also use storage capacity for cloning and snapshotting.

### 3.1.4. Checking a Cluster’s Usage Statistics

To check a cluster’s data usage and data distribution among pools, you can use the `df` option. It is similar to Linux `df`. Execute the following:

```
# ceph df
```

The **GLOBAL** section of the output provides an overview of the amount of storage the storage cluster uses for data.

- **SIZE**: The overall storage capacity of the storage cluster.
- **AVAIL**: The amount of free space available in the storage cluster.
- **RAW USED**: The amount of raw storage used.
- **% RAW USED**: The percentage of raw storage used. Use this number in conjunction with the full ratio and near full ratio to ensure that you are not reaching the storage cluster’s capacity.

The **POOLS** section of the output provides a list of pools and the notional usage of each pool. The output from this section **DOES NOT** reflect replicas, clones or snapshots. For example, if you store an object with 1MB of data, the notional usage will be 1MB, but the actual usage may be 3MB or more depending on the number of replicas (e.g., `size = 3`, clones and snapshots).

- **NAME**: The name of the pool.
- **ID**: The pool ID.
- **USED**: The notional amount of data stored in kilobytes, unless the number appends `M` for megabytes or `G` for gigabytes.
- **%USED**: The notional percentage of storage used per pool.
- **Objects**: The notional number of objects stored per pool.

**NOTE**

The numbers in the **POOLS** section are notional. They are not inclusive of the number of replicas, snapshots or clones. As a result, the sum of the **USED** and **%USED** amounts will not add up to the **RAW USED** and **%RAW USED** amounts in the **GLOBAL** section of the output. See [How Ceph Calculates Data Usage](#) for details.

### 3.1.5. Checking a Cluster’s Status

To check a cluster’s status, execute the following:

```
# ceph status
```

Or:

```
```
In interactive mode, type `status` and press Enter:

```
ceph> status
```

Ceph will print the cluster status. For example, a tiny Ceph cluster consisting of one monitor, and two
OSDs may print the following:

```
cluster b370a29d-9287-4ca3-ab57-3d824f65e339
health HEALTH_OK
monmap e1: 1 mons at {ceph1=10.0.0.8:6789/0}, election epoch 2, quorum 0
ceph1
osdmap e63: 2 osds: 2 up, 2 in
  pgmap v41332: 952 pgs, 20 pools, 17130 MB data, 2199 objects
    115 GB used, 167 GB / 297 GB avail
      1 active+clean+scrubbing+deep
    951 active+clean
```

### 3.1.6. Checking Monitor Status

If the storage cluster has multiple monitors, which is required for high availability for production Ceph
storage clusters. You should check the Ceph Monitor quorum status after you start the Ceph storage
cluster before reading and/or writing data. A quorum must be present when multiple monitors are
running. You should also check Ceph Monitor status periodically to ensure that they are running. If there
is a problem with the Monitor, that prevents an agreement on the state of the storage cluster, the fault
may prevent Ceph clients from reading and writing data.

- To display the monitor map, execute the following:

  ```
  # ceph mon stat
  or
  # ceph mon dump
  ```

- To check the quorum status for the storage cluster, execute the following:

  ```
  # ceph quorum_status -f json-pretty
  ```

Ceph will return the quorum status. For example, a Ceph storage cluster consisting of three
monitors may return the following:

```
{
  "election_epoch": 10,
  "quorum": [
    0,
    1,
    2],
  "monmap": {
    "epoch": 1,
    "fsid": "444b489c-4f16-4b75-83f0-cb8097468898",
    "modified": "2011-12-12 13:28:27.505520",
    "created": "2011-12-12 13:28:27.505520",
    "mons": [
```
3.1.7. Using the Administration Socket

Use the administration socket to interact with a given daemon directly by using a UNIX socket file. For example, the socket enables you to:

- List the Ceph configuration at runtime
- Set configuration values at runtime directly without relaying on Monitors. This is useful when Monitors are **down**.
- Dump historic operations
- Dump the operation priority queue state
- Dump operations without rebooting
- Dump performance counters

In addition, using the socket is helpful when troubleshooting problems related to Monitors or OSDs. For details, see the Troubleshooting Guide for Red Hat Ceph Storage 3.

To use the socket:

```
ceph daemon <type>.<id> <command>
```

Replace:

- `<type>` with the type of the Ceph daemon (**mon**, **osd**, **mds**).
- `<id>` with the daemon ID
- `<command>` with the command to run. Use `help` to list the available commands for a given daemon.

For example, to view a Monitor status of a Monitor named **mon.0**:

```
# ceph daemon mon.0 mon_status
```

Alternatively, specify the daemon by using its socket file.

```
ceph daemon /var/run/ceph/<socket-file> <command>
```
For example, to view the status of an OSD named osd.2:

```
# ceph daemon /var/run/ceph/ceph-osd.2.asok status
```

To list all socket files for the Ceph processes:

```
$ ls /var/run/ceph
```

### 3.1.8. Checking OSD Status

An OSD’s status is either in the cluster, **in**, or out of the cluster, **out**; and, it is either up and running, **up**, or it is down and not running, or **down**. If an OSD is **up**, it may be either **in** the storage cluster, which data can be read and written, or it is **out** of the storage cluster. If it was **in** the cluster and recently moved **out** of the cluster, Ceph will migrate placement groups to other OSDs. If an OSD is **out** of the cluster, CRUSH will not assign placement groups to the OSD. If an OSD is **down**, it should also be **out**.

**NOTE**

If an OSD is **down** and **in**, there is a problem and the cluster will not be in a healthy state.

If you execute a command such as `ceph health`, `ceph -s` or `ceph -w`, you may notice that the cluster does not always echo back **HEALTH OK**. Don’t panic. With respect to OSDs, you should expect that the cluster will **NOT** echo **HEALTH OK** in a few expected circumstances:

- You haven’t started the cluster yet, it won’t respond.
- You have just started or restarted the cluster and it’s not ready yet, because the placement groups are getting created and the OSDs are in the process of peering.
- You just added or removed an OSD.
- You just have modified the cluster map.

An important aspect of monitoring OSDs is to ensure that when the cluster is up and running that all OSDs that are **in** the cluster are **up** and running, too. To see if all OSDs are running, execute:

```
# ceph osd stat
```

or

```
# ceph osd dump
```
The result should tell you the map epoch, eNNNN, the total number of OSDs, x, how many, y, are up, and how many, z, are in:

```
eNNNN: x osds: y up, z in
```

If the number of OSDs that are in the cluster is more than the number of OSDs that are up, execute the following command to identify the ceph-osd daemons that aren’t running:

```
# ceph osd tree
```

Example output:

```
# id  weight type name   up/down reweight
-1  3   pool default
-3  3       rack mainrack
-2  3           host osd-host
 0  1               osd.0   up  1
 1  1               osd.1   up  1
 2  1               osd.2   up  1
```

**TIP**

The ability to search through a well-designed CRUSH hierarchy may help you troubleshoot the storage cluster by identifying the physical locations faster.

If an OSD is down, connect to the node and start it. You can use Red Hat Storage Console to restart the OSD node, or you can use the command line, for example:

```
# systemctl start ceph-osd@<osd_id>
```

### 3.2. LOW-LEVEL MONITORING

Lower-level monitoring typically involves ensuring that OSDs are peering. When faults occur, placement groups operate in a degraded state. This can be due to many things such as failed hardware, hung or crashed daemon, network latency or outage among other things.

#### 3.2.1. Placement Group Sets

When CRUSH assigns placement groups to OSDs, it looks at the number of replicas for the pool and assigns the placement group to OSDs such that each replica of the placement group gets assigned to a different OSD. For example, if the pool requires three replicas of a placement group, CRUSH may assign them to osd.1, osd.2 and osd.3 respectively. CRUSH actually seeks a pseudo-random placement that will take into account failure domains you set in the CRUSH map, so you will rarely see placement groups assigned to nearest neighbor OSDs in a large cluster. We refer to the set of OSDs that should contain the replicas of a particular placement group as the **Acting Set**. In some cases, an OSD in the Acting Set is down or otherwise not able to service requests for objects in the placement group. When these situations arise, don’t panic. Common examples include:

- You added or removed an OSD. Then, CRUSH reassigned the placement group to other OSDs —thereby changing the composition of the Acting Set and spawning the migration of data with a "backfill" process.
- An OSD was **down**, was restarted and is now **recovering**.

- An OSD in the Acting Set is **down** or unable to service requests, and another OSD has temporarily assumed its duties.

Ceph processes a client request using the Up Set, which is the set of OSDs that will actually handle the requests. In most cases, the Up Set and the Acting Set are virtually identical. When they are not, it may indicate that Ceph is migrating data, an OSD is recovering, or that there is a problem, that is, Ceph usually echoes a **HEALTH WARN** state with a “stuck stale” message in such scenarios.

- To retrieve a list of placement groups:
  ```
  # ceph pg dump
  ```

- To view which OSDs are in the Acting Set or in the Up Set for a given placement group:
  ```
  # ceph pg map <pg-num>
  ```

  The result should tell you the osdmap epoch, \texttt{eNNN}, the placement group number, \texttt{<pg-num>}, the OSDs in the Up Set \texttt{up[]} and the OSDs in the acting set, \texttt{acting[]}:

  ```
  osdmap eNNN pg <pg-num> -> up [0,1,2] acting [0,1,2]
  ```

  **NOTE**

  If the Up Set and Acting Set do not match, this may be an indicator that the cluster rebalancing itself or of a potential problem with the cluster.

### 3.2.2. Peering

Before you can write data to a placement group, it must be in an **active** state, and it **should** be in a **clean** state. For Ceph to determine the current state of a placement group, the primary OSD of the placement group (i.e., the first OSD in the acting set), peers with the secondary and tertiary OSDs to establish agreement on the current state of the placement group (assuming a pool with 3 replicas of the PG).

![Peering Diagram](CEPH_459704_1077)

### 3.2.3. Monitoring Placement Group States

If you execute a command such as `ceph health`, `ceph -s` or `ceph -w`, you may notice that the cluster does not always echo back **HEALTH OK**. After you check to see if the OSDs are running, you should also check placement group states. You should expect that the cluster will **NOT** echo **HEALTH OK**
in a number of placement group peering-related circumstances:

- You have just created a pool and placement groups haven’t peered yet.
- The placement groups are recovering.
- You have just added an OSD to or removed an OSD from the cluster.
- You have just modified the CRUSH map and the placement groups are migrating.
- There is inconsistent data in different replicas of a placement group.
- Ceph is scrubbing a placement group’s replicas.
- Ceph doesn’t have enough storage capacity to complete backfilling operations.

If one of the foregoing circumstances causes Ceph to echo **HEALTH WARN**, don’t panic. In many cases, the cluster will recover on its own. In some cases, you may need to take action. An important aspect of monitoring placement groups is to ensure that when the cluster is up and running that all placement groups are **active**, and preferably in the **clean** state. To see the status of all placement groups, execute:

```bash
# ceph pg stat
```

The result should tell you the placement group map version, **vNNNNNN**, the total number of placement groups, **x**, and how many placement groups, **y**, are in a particular state such as **active+clean**:

```bash
vNNNNNN: x pgs: y active+clean; z bytes data, aa MB used, bb GB / cc GB avail
```

**NOTE**

It is common for Ceph to report multiple states for placement groups.

### Snapshot Trimming PG States

When snapshots exist, two additional PG states will be reported.

- **snaptrim**: The PGs are currently being trimmed
- **snaptrim_wait**: The PGs are waiting to be trimmed

Example Output:

```bash
244 active+clean+snaptrim_wait
32 active+clean+snaptrim
```

**NOTE**

See the miscellaneous OSD settings in the Red Hat Ceph Storage 3 **Configuration Guide** for more details on the snapshot trimming settings.

In addition to the placement group states, Ceph will also echo back the amount of data used, **aa**, the amount of storage capacity remaining, **bb**, and the total storage capacity for the placement group. These
numbers can be important in a few cases:

- You are reaching the near full ratio or full ratio.
- Your data isn’t getting distributed across the cluster due to an error in the CRUSH configuration.

### Placement Group IDs

Placement group IDs consist of the pool number, and not the pool name, followed by a period (.) and the placement group ID—a hexadecimal number. You can view pool numbers and their names from the output of `ceph osd lspools`. The default pool names data, metadata and rbd correspond to pool numbers 0, 1 and 2 respectively. A fully qualified placement group ID has the following form:

```plaintext
<pool_num>.<pg_id>
```

Example output:

```plaintext
0.1f
```

- To retrieve a list of placement groups:
  ```bash
  # ceph pg dump
  ```

- To format the output in JSON format and save it to a file:
  ```bash
  # ceph pg dump -o <file_name> --format=json
  ```

- To query a particular placement group:
  ```bash
  # ceph pg <pool_num>.<pg_id> query
  ```

Example output in JSON format:

```json
{
  "state": "active+clean",
  "up": [1, 0],
  "acting": [1, 0],
  "info": {
    "pgid": "1.e",
    "last_update": "4'1",
    "last_complete": "4'1",
    "log_tail": "0'0",
    "last_backfill": "MAX",
    "purged_snaps": "[]",
    "history": {
      "epoch_created": 1,
      "last_epoch_started": 537,
      "last_epoch_clean": 537,
    }
  }
}
```
"last_epoch_split": 534,
"same_up_since": 536,
"same_interval_since": 536,
"same_primary_since": 536,
"last_scrub": "4'1",
"last_scrub_stamp": "2013-01-25 10:12:23.828174"
},
"stats": {
  "version": "4'1",
  "reported": "536'782",
  "state": "active+clean",
  "last_fresh": "2013-01-25 10:12:23.828271",
  "last_change": "2013-01-25 10:12:23.828271",
  "last_active": "2013-01-25 10:12:23.828271",
  "last_clean": "2013-01-25 10:12:23.828271",
  "last_unstale": "2013-01-25 10:12:23.828271",
  "mapping_epoch": 535,
  "log_start": "0'0",
  "ondisk_log_start": "0'0",
  "created": 1,
  "last_epoch_clean": 1,
  "parent": "0.0",
  "parent_split_bits": 0,
  "last_scrub": "4'1",
  "last_scrub_stamp": "2013-01-25 10:12:23.828174",
  "log_size": 128,
  "ondisk_log_size": 128,
  "stat_sum": {
    "num_bytes": 205,
    "num_objects": 1,
    "num_object_clones": 0,
    "num_object_copies": 0,
    "num_objects_missing_on_primary": 0,
    "num_objects_degraded": 0,
    "num_objects_unfound": 0,
    "num_read": 1,
    "num_read_kb": 0,
    "num_write": 3,
    "num_write_kb": 1
  },
  "stat_cat_sum": {
  },
  "up": [
    1,
    0
  ],
  "acting": [
    1,
    0
  ],
  "empty": 0,
  "dne": 0,
  "incomplete": 0
},
"recovery_state": [
{
"name": "Started/Primary/Active",
"enter_time": "2013-01-23 09:35:37.594691",
"might_have_unfound": [
],
"scrub": {
"scrub_epoch_start": "536",
"scrub_active": 0,
"scrub_blockWrites": 0,
"finalizing_scrub": 0,
"scrub_waiting_on": 0,
"scrub_waiting_on_whom": [
]
},
"ended": 1,
"ended_time": "2013-01-23 09:35:38.392865"
},
{
"name": "Started",
"enter_time": "2013-01-23 09:35:31.581160"
}
]
}

The following subsections describe common states in greater detail.

3.2.3.1. Creating

When you create a pool, it will create the number of placement groups you specified. Ceph will echo creating when it is creating one or more placement groups. Once they are created, the OSDs that are part of a placement group’s Acting Set will peer. Once peering is complete, the placement group status should be active+clean, which means a Ceph client can begin writing to the placement group.

![Monitoring Placement Group States]

3.2.3.2. Peering

When Ceph is Peering a placement group, Ceph is bringing the OSDs that store the replicas of the placement group into agreement about the state of the objects and metadata in the placement group. When Ceph completes peering, this means that the OSDs that store the placement group agree about the current state of the placement group. However, completion of the peering process does NOT mean that each replica has the latest contents.

Authoritative History
Ceph will **NOT** acknowledge a write operation to a client, until all OSDs of the acting set persist the write operation. This practice ensures that at least one member of the acting set will have a record of every acknowledged write operation since the last successful peering operation.

With an accurate record of each acknowledged write operation, Ceph can construct and disseminate a new authoritative history of the placement group—a complete, and fully ordered set of operations that, if performed, would bring an OSD’s copy of a placement group up to date.

### 3.2.3.3. Active

Once Ceph completes the peering process, a placement group may become **active**. The **active** state means that the data in the placement group is generally available in the primary placement group and the replicas for read and write operations.

### 3.2.3.4. Clean

When a placement group is in the **clean** state, the primary OSD and the replica OSDs have successfully peered and there are no stray replicas for the placement group. Ceph replicated all objects in the placement group the correct number of times.

### 3.2.3.5. Degraded

When a client writes an object to the primary OSD, the primary OSD is responsible for writing the replicas to the replica OSDs. After the primary OSD writes the object to storage, the placement group will remain in a **degraded** state until the primary OSD has received an acknowledgement from the replica OSDs that Ceph created the replica objects successfully.

The reason a placement group can be **active+degraded** is that an OSD may be **active** even though it doesn’t hold all of the objects yet. If an OSD goes **down**, Ceph marks each placement group assigned to the OSD as **degraded**. The OSDs must peer again when the OSD comes back online. However, a client can still write a new object to a **degraded** placement group if it is **active**.

If an OSD is **down** and the **degraded** condition persists, Ceph may mark the **down** OSD as **out** of the cluster and remap the data from the **down** OSD to another OSD. The time between being marked **down** and being marked **out** is controlled by **mon osd down out interval**, which is set to **300** seconds by default.

A placement group can also be **degraded**, because Ceph cannot find one or more objects that Ceph thinks should be in the placement group. While you cannot read or write to unfound objects, you can still access all of the other objects in the **degraded** placement group.

Let’s say there are 9 OSDs with three copies of an object. If OSD number 9 goes down, the PGs assigned to OSD 9 go in a degraded state. If OSD 9 doesn’t recover, it goes out of the cluster and the cluster rebalances. In that scenario, the PGs are degraded and then recover to an active state.

### 3.2.3.6. Recovering

Ceph was designed for fault-tolerance at a scale where hardware and software problems are ongoing. When an OSD goes **down**, its contents may fall behind the current state of other replicas in the placement groups. When the OSD is back **up**, the contents of the placement groups must be updated to reflect the current state. During that time period, the OSD may reflect a **recovering** state.

Recovery isn’t always trivial, because a hardware failure might cause a cascading failure of multiple OSDs. For example, a network switch for a rack or cabinet may fail, which can cause the OSDs of a
number of host machines to fall behind the current state of the cluster. Each one of the OSDs must recover once the fault is resolved.

Ceph provides a number of settings to balance the resource contention between new service requests and the need to recover data objects and restore the placement groups to the current state. The osd recovery delay start setting allows an OSD to restart, re-peer and even process some replay requests before starting the recovery process. The osd recovery threads setting limits the number of threads for the recovery process, by default one thread. The osd recovery thread timeout sets a thread timeout, because multiple OSDs may fail, restart and re-peer at staggered rates. The osd recovery max active setting limits the number of recovery requests an OSD will entertain simultaneously to prevent the OSD from failing to serve. The osd recovery max chunk setting limits the size of the recovered data chunks to prevent network congestion.

3.2.3.7. Backfilling

When a new OSD joins the cluster, CRUSH will reassign placement groups from OSDs in the cluster to the newly added OSD. Forcing the new OSD to accept the reassigned placement groups immediately can put excessive load on the new OSD. Backfilling the OSD with the placement groups allows this process to begin in the background. Once backfilling is complete, the new OSD will begin serving requests when it is ready.

During the backfill operations, you may see one of several states: backfill_wait indicates that a backfill operation is pending, but isn’t underway yet; backfill indicates that a backfill operation is underway; and, backfill_too_full indicates that a backfill operation was requested, but couldn’t be completed due to insufficient storage capacity. When a placement group cannot be backfilled, it may be considered incomplete.

Ceph provides a number of settings to manage the load spike associated with reassigning placement groups to an OSD, especially a new OSD. By default, osd_max_backfills sets the maximum number of concurrent backfills to or from an OSD to 10. The osd backfill full ratio enables an OSD to refuse a backfill request if the OSD is approaching its full ratio, by default 85%. If an OSD refuses a backfill request, the osd backfill retry interval enables an OSD to retry the request, by default after 10 seconds. OSDs can also set osd backfill scan min and osd backfill scan max to manage scan intervals, by default 64 and 512.

For some workloads, it is beneficial to avoid regular recovery entirely and use backfill instead. Since backfilling occurs in the background, this allows I/O to proceed on the objects in the OSD. To force backfill rather than recovery, set osd_min_pg_log_entries to 1, and set osd_max_pg_log_entries to 2. Contact your Red Hat Support account team for details on when this situation is appropriate for your workload.

3.2.3.8. Remapped

When the Acting Set that services a placement group changes, the data migrates from the old acting set to the new acting set. It may take some time for a new primary OSD to service requests. So it may ask the old primary to continue to service requests until the placement group migration is complete. Once data migration completes, the mapping uses the primary OSD of the new acting set.

3.2.3.9. Stale

While Ceph uses heartbeats to ensure that hosts and daemons are running, the ceph-osd daemons may also get into a stuck state where they aren’t reporting statistics in a timely manner, for example, a temporary network fault. By default, OSD daemons report their placement group, up thru, boot and
failure statistics every half second, that is, 0.5, which is more frequent than the heartbeat thresholds. If the Primary OSD of a placement group’s acting set fails to report to the monitor or if other OSDs have reported the primary OSD down, the monitors will mark the placement group stale.

When you start the storage cluster, it is common to see the stale state until the peering process completes. After the storage cluster has been running for awhile, seeing placement groups in the stale state indicates that the primary OSD for those placement groups is down or not reporting placement group statistics to the monitor.

3.2.3.10. Misplaced

There are some temporary backfilling scenarios where a PG gets mapped temporarily to an OSD. When that temporary situation should no longer be the case, the PGs might still reside in the temporary location and not in the proper location. In which case, they are said to be misplaced. That’s because the correct number of extra copies actually exist, but one or more copies is in the wrong place.

Lets say there are 3 OSDs: 0,1,2 and all PGs map to some permutation of those three. If you add another OSD (OSD 3), some PGs will now map to OSD 3 instead of one of the others. However, until OSD 3 is backfilled, the PG will have a temporary mapping allowing it to continue to serve I/O from the old mapping. During that time, the PG is misplaced, because it has a temporary mapping, but not degraded, since there are 3 copies.

Example

```plaintext
pg 1.5: up=acting: [0,1,2]
<add osd 3>
pg 1.5: up: [0,3,1] acting: [0,1,2]
```

Here, [0,1,2] is a temporary mapping, so the up set is not equal to the acting set and the PG is misplaced but not degraded since [0,1,2] is still three copies.

Example

```plaintext
pg 1.5: up=acting: [0,3,1]
```

OSD 3 is now backfilled and the temporary mapping is removed, not degraded and not misplaced.

3.2.3.11. Incomplete

A PG goes into an incomplete state when there is incomplete content and peering fails, that is, when there are no complete OSDs which are current enough to perform recovery.

Lets say OSD 1, 2, and 3 are the acting OSD set and it switches to OSD 1, 4, and 3, then osd.1 will request a temporary acting set of OSD 1, 2, and 3 while backfilling 4. During this time, if OSD 1, 2, and 3 all go down, osd.4 will be the only one left which might not have fully backfilled all the data. At this time, the PG will go incomplete indicating that there are no complete OSDs which are current enough to perform recovery.

Alternately, if osd.4 is not involved and the acting set is simply OSD 1, 2, and 3 when OSD 1, 2, and 3 go down, the PG would likely go stale indicating that the mons have not heard anything on that PG since the acting set changed. The reason being there are no OSDs left to notify the new OSDs.

3.2.4. Identifying Troubled Placement Groups
As previously noted, a placement group isn’t necessarily problematic just because its state isn’t **active+clean**. Generally, Ceph’s ability to self repair may not be working when placement groups get stuck. The stuck states include:

- **Unclean**: Placement groups contain objects that are not replicated the desired number of times. They should be recovering.

- **Inactive**: Placement groups cannot process reads or writes because they are waiting for an OSD with the most up-to-date data to come back up.

- **Stale**: Placement groups are in an unknown state, because the OSDs that host them have not reported to the monitor cluster in a while, and can be configured with the `mon osd report timeout` setting.

To identify stuck placement groups, execute the following:

```
# ceph pg dump_stuck [unclean|inactive|stale|undersized|degraded]
```

### 3.2.5. Finding an Object Location

To store object data in the Ceph Object Store, a Ceph client must:

1. Set an object name
2. Specify a pool

The Ceph client retrieves the latest cluster map and the CRUSH algorithm calculates how to map the object to a placement group, and then calculates how to assign the placement group to an OSD dynamically. To find the object location, all you need is the object name and the pool name. For example:

```
# ceph osd map <pool_name> <object_name>
```

### 3.3. MONITORING A CEPH STORAGE CLUSTER WITH THE RED HAT CEPH STORAGE DASHBOARD

The Red Hat Ceph Storage Dashboard provides a monitoring dashboard to visualize the state of a Ceph Storage Cluster. Also, the Red Hat Ceph Storage Dashboard architecture provides a framework for additional modules to add functionality to the storage cluster.

- To learn about the Dashboard, see Section 3.3.1, “The Red Hat Ceph Storage Dashboard”.

- To install the Dashboard, see Section 3.3.2, “Installing the Red Hat Ceph Storage Dashboard”.

- To access the Dashboard, see Section 3.3.3, “Accessing the Red Hat Ceph Storage Dashboard”.

- To change the default password after installing the Dashboard, see Section 3.3.4, “Changing the default Red Hat Ceph Storage dashboard password”.

- To learn about the Prometheus plugin, see Section 3.3.5, “The Prometheus plugin for Red Hat Ceph Storage”.

---

25
To learn about the Red Hat Ceph Storage Dashboard alerts and how to configure them, see Section 3.3.6, "The Red Hat Ceph Storage Dashboard alerts".

Prerequisites

- A running Red Hat Ceph Storage cluster

3.3.1. The Red Hat Ceph Storage Dashboard

The Red Hat Ceph Storage Dashboard provides a monitoring dashboard for Ceph clusters to visualize the storage cluster state. The dashboard is accessible from a web browser and provides a number of metrics and graphs about the state of the cluster, Monitors, OSDs, Pools, or the network.

With the previous releases of Red Hat Ceph Storage, monitoring data was sourced through a [collectd](http://collectd.org) plugin, which sent the data to an instance of the Graphite monitoring utility. Starting with Red Hat Ceph Storage 3.2, monitoring data is sourced directly from the [ceph-mgr](http://ceph-mgr.io) daemon, using the [ceph-mgr](http://ceph-mgr.io) Prometheus plugin.

The introduction of Prometheus as the monitoring data source simplifies deployment and operational management of the Red Hat Ceph Storage Dashboard solution, along with reducing the overall hardware requirements. By sourcing the Ceph monitoring data directly, the Red Hat Ceph Storage Dashboard solution is better able to support Ceph clusters deployed in containers.

**NOTE**

With this change in architecture, there is no migration path for monitoring data from Red Hat Ceph Storage 2.x and 3.0 to Red Hat Ceph Storage 3.2.

The Red Hat Ceph Storage Dashboard uses the following utilities:

- The Ansible automation application for deployment.
- The embedded Prometheus [ceph-mgr](http://ceph-mgr.io) plugin.
- The Prometheus [node-exporter](http://node-exporter.io) daemon, running on each node of the storage cluster.
- The Grafana platform to provide a user interface and alerting.

The Red Hat Ceph Storage Dashboard supports the following features:

**General Features**

- Support for Red Hat Ceph Storage 3.1 and higher
- SELinux support
- Support for FileStore and BlueStore OSD back ends
- Support for encrypted and non-encrypted OSDs
- Support for Monitor, OSD, the Ceph Object Gateway, and iSCSI roles
- Initial support for the Metadata Servers (MDS)
- Drill down and dashboard links
• 15 second granularity

• Support for Hard Disk Drives (HDD), Solid-state Drives (SSD), Non-volatile Memory Express (NVMe) interface, and Intel® Cache Acceleration Software (Intel® CAS)

Node Metrics

• CPU and RAM usage

• Network load

Configurable Alerts

• Out-of-Band (OOB) alerts and triggers

• Notification channel is automatically defined during the installation

• The Ceph Health Summary dashboard created by default

See the Red Hat Ceph Storage Dashboard Alerts section for details.

Cluster Summary

• OSD configuration summary

• OSD FileStore and BlueStore summary

• Cluster versions breakdown by role

• Disk size summary

• Host size by capacity and disk count

• Placement Groups (PGs) status breakdown

• Pool counts

• Device class summary, HDD vs. SSD

Cluster Details

• Cluster flags status (noout, nodown, and others)

• OSD or Ceph Object Gateway hosts up and down status

• Per pool capacity usage

• Raw capacity utilization

• Indicators for active scrub and recovery processes

• Growth tracking and forecast (raw capacity)

• Information about OSDs that are down or near full, including the OSD host and disk

• Distribution of PGs per OSD

• OSDs by PG counts, highlighting the over or under utilized OSDs
OSD Performance

- Information about I/O operations per second (IOPS) and throughput by pool
- OSD performance indicators
- Disk statistics per OSD
- Cluster wide disk throughput
- Read/write ratio (client IOPS)
- Disk utilization heat map
- Network load by Ceph role

The Ceph Object Gateway Details

- Aggregated load view
- Per host latency and throughput
- Workload breakdown by HTTP operations

The Ceph iSCSI Gateway Details

- Aggregated views
- Configuration
- Performance
- Per Gateway resource utilization
- Per client load and configuration
- Per Ceph Block Device image performance

3.3.2. Installing the Red Hat Ceph Storage Dashboard

The Red Hat Ceph Storage Dashboard provides a visual dashboard to monitor various metrics in a running Ceph Storage Cluster.

NOTE

For information on upgrading the Red Hat Ceph Storage Dashboard see Upgrading Red Hat Ceph Storage Dashboard in the Installation Guide for Red Hat Enterprise Linux.

Prerequisites

- The storage cluster nodes use Red Hat Enterprise Linux 7.
- A separate node, the Red Hat Ceph Storage Dashboard node, for receiving data from the cluster nodes and providing the Red Hat Ceph Storage Dashboard.
Prepare the Red Hat Ceph Storage Dashboard node:

- Enable the Tools repository.
- If using a firewall, then ensure that the following TCP ports are open:

<table>
<thead>
<tr>
<th>Port</th>
<th>Use</th>
<th>Where?</th>
</tr>
</thead>
<tbody>
<tr>
<td>3000</td>
<td>Grafana</td>
<td>The Red Hat Ceph Storage Dashboard node.</td>
</tr>
<tr>
<td>9090</td>
<td>Basic Prometheus graphs</td>
<td>The Red Hat Ceph Storage Dashboard node.</td>
</tr>
<tr>
<td>9100</td>
<td>Prometheus' <code>node-exporter</code> daemon</td>
<td>All storage cluster nodes.</td>
</tr>
<tr>
<td>9283</td>
<td>Gathering Ceph data</td>
<td>All <code>ceph-mgr</code> nodes.</td>
</tr>
<tr>
<td>9287</td>
<td>Ceph iSCSI gateway data</td>
<td>All Ceph iSCSI gateway nodes.</td>
</tr>
</tbody>
</table>

For more details see the *Using Firewalls* chapter in the *Security Guide* for Red Hat Enterprise Linux 7.

**Procedure**

Use the following commands on the Ansible administration node as the `root` user.

1. Install the `cephmetrics-ansible` package.

   ```
   [root@admin ~]# yum install cephmetrics-ansible
   ```

2. Using the Ceph Ansible inventory as a base, add the Red Hat Ceph Storage Dashboard node under the `[ceph-grafana]` section of the Ansible inventory file, by default located at `/etc/ansible/hosts`.

   ```
   [ceph-grafana] $HOST_NAME
   
   Replace:
   ```

   - `$HOST_NAME` with the name of the Red Hat Ceph Storage Dashboard node

   For example:

   ```
   [ceph-grafana]
   node0
   ```

3. Change to the `/usr/share/cephmetrics-ansible/` directory.

   ```
   [root@admin ~]# cd /usr/share/cephmetrics-ansible
   ```
4. Use the Ansible playbook.

```
[root@admin cephmetrics-ansible]# ansible-playbook -v playbook.yml
```

**IMPORTANT**

Every time you update the cluster configuration, for example, you add or remove a MON or OSD node, you must re-run the `cephmetrics` Ansible playbook.

**NOTE**

The `cephmetrics` Ansible playbook does the following actions:

- Updates the `ceph-mgr` instance to enable the prometheus plugin and opens TCP port 9283.
- Deploys the Prometheus `node-exporter` daemon to each node in the storage cluster.
  - Opens TCP port 9100.
  - Starts the `node-exporter` daemon.
- Deploys Grafana and Prometheus containers under Docker/systemd on the Red Hat Ceph Storage Dashboard node.
  - Prometheus is configured to gather data from the ceph-mgr nodes and the node-exporters running on each ceph host
  - Opens TCP port 3000.
  - The dashboards, themes and user accounts are all created in Grafana.
- Outputs the URL of Grafana for the administrator.

### 3.3.3. Accessing the Red Hat Ceph Storage Dashboard

Accessing the Red Hat Ceph Storage Dashboard gives you access to the web-based management tool for administrating Red Hat Ceph Storage clusters.

**Prerequisites**

- **Install the Red Hat Ceph Storage Dashboard.**
- Ensure that NTP is synchronizing clocks properly because a time lag can occur among the Ceph Storage Dashboard node, cluster nodes, and a browser when the nodes are not properly synced. See the *Configuring the Network Time Protocol for Red Hat Ceph Storage* section in the Red Hat Ceph Storage 3 *Installation Guide for Red Hat Enterprise Linux* or *Ubuntu.*

**Procedure**

1. Enter the following URL to a web browser:

   ```
   http://$HOST_NAME:3000
   ```
Replace:

- **$HOST_NAME** with the name of the Red Hat Ceph Storage Dashboard node

For example:

```
http://cephmetrics:3000
```

2. Enter the password for the **admin** user. If you did not set the password during the installation, use **admin**, which is the default password.

Once logged in, you are automatically placed on the *Ceph At a Glance* dashboard. The *Ceph At a Glance* dashboard provides a high-level overviews of capacity, performance, and node-level performance information.

**Example**

![Ceph At a Glance dashboard](image)

**Additional Resources**

- See the [Changing the Default Red Hat Ceph Storage Dashboard Password](#) section in the Red Hat Ceph Storage Administration Guide.

### 3.3.4. Changing the default Red Hat Ceph Storage dashboard password

The default user name and password for accessing the Red Hat Ceph Storage Dashboard is set to **admin** and **admin**. For security reasons, you might want to change the password after the installation.

**NOTE**

If you redeploy the Red Hat Ceph Storage dashboard using Ceph Ansible, then the password will be reset to the default value. Update the Ceph Ansible inventory file (`/etc/ansible/hosts`) with the custom password to prevent the password from resetting to the default value.

**Prerequisites**
● Install the Red Hat Ceph Storage Dashboard.

● Log in to the Red Hat Ceph Storage Dashboard.

**Procedure**

1. Click the Grafana icon in the upper-left corner.

2. Hover over the user name you want to modify the password for. In this case **admin**.

3. Click **Profile**.

4. Click **Change Password**.

5. Enter the new password twice and click **Change Password**.

**Additional Resource**

● If you forgot the password, follow the **Reset admin password** procedure on the Grafana web pages.

### 3.3.5. The Prometheus plugin for Red Hat Ceph Storage

As a storage administrator, you can gather performance data, export that data using the Prometheus plugin module for the Red Hat Ceph Storage Dashboard, and then perform queries on this data. The Prometheus module allows **ceph-mgr** to expose Ceph related state and performance data to a Prometheus server.

#### 3.3.5.1. Prerequisites

● Running Red Hat Ceph Storage 3.1 or higher.

● Installation of the Red Hat Ceph Storage Dashboard.

#### 3.3.5.2. The Prometheus plugin

The Prometheus plugin provides an exporter to pass on Ceph performance counters from the collection point in **ceph-mgr**. The Red Hat Ceph Storage Dashboard receives **MMgrReport** messages from all **MgrClient** processes, such as Ceph Monitors and OSDs. A circular buffer of the last number of samples contains the performance counter schema data and the actual counter data. This plugin creates an HTTP endpoint and retrieves the latest sample of every counter when polled. The HTTP path and query parameters are ignored; all extant counters for all reporting entities are returned in a text exposition format.

**Additional Resources**

● See the [Prometheus documentation](#) for more details on the text exposition format.

#### 3.3.5.3. Managing the Prometheus environment

To monitor a Ceph storage cluster with Prometheus you can configure and enable the Prometheus exporter so the metadata information about the Ceph storage cluster can be collected.
A running Red Hat Ceph Storage 3.1 cluster

Installation of the Red Hat Ceph Storage Dashboard

Procedure

1. Open and edit the `/etc/prometheus/prometheus.yml` file.
   
   a. Under the `global` section, set the `scrape_interval` and `evaluation_interval` options to 15 seconds.

   **Example**

   ```yaml
   global:
     scrape_interval: 15s
     evaluation_interval: 15s
   ```

   b. Under the `scrape_configs` section, add the `honor_labels: true` option, and edit the `targets`, and `instance` options for each of the `ceph-mgr` nodes.

   **Example**

   ```yaml
   scrape_configs:
     - job_name: 'node'
       honor_labels: true
       static_configs:
         - targets: ['node1.example.com:9100']
           labels:
             instance: "node1.example.com"
         - targets: ['node2.example.com:9100']
           labels:
             instance: "node2.example.com"
   ```

   **NOTE**

   Using the `honor_labels` option enables Ceph to output properly-labelled data relating to any node in the Ceph storage cluster. This allows Ceph to export the proper `instance` label without Prometheus overwriting it.

   c. To add a new node, simply add the `targets`, and `instance` options in the following format:

   **Example**

   ```yaml
   - targets: ['new-node.example.com:9100']
     labels:
       instance: "new-node"
   ```

   **NOTE**

   The `instance` label has to match what appears in Ceph’s OSD metadata `instance` field, which is the short host name of the node. This helps to correlate Ceph stats with the node’s stats.
2. Add Ceph targets to the `/etc/prometheus/ceph_targets.yml` file in the following format.

   **Example**

   ```yaml
   [ 
     { 
       "targets": [ "cephnode1.example.com:9283" ],
       "labels": {} 
     }
   ]
   ```

3. Enable the Prometheus module:

   ```bash
   # ceph mgr module enable prometheus
   ```

### 3.3.5.4. Using the Prometheus expression browser

Use the builtin Prometheus expression browser to run queries against the collected data.

**Prerequisites**

- A running Red Hat Ceph Storage 3.1 cluster
- Installation of the Red Hat Ceph Storage Dashboard

**Procedure**

1. Enter the URL for the Prometheus web browser:

   ```bash
   http://${DASHBOARD_SERVER_NAME}:9090/graph
   ```

   Replace...

   - `${DASHBOARD_SERVER_NAME}` with the name of the Red Hat Ceph Storage Dashboard server.

2. Click on *Graph*, then type in or paste the query into the query window and press the *Execute* button.

   a. View the results in the console window.

3. Click on *Graph* to view the rendered data.

**Additional Resources**

- See the [Prometheus expression browser](#) documentation on the Prometheus web site for more information.

### 3.3.5.5. Working with the Prometheus data and queries

The statistic names are exactly as Ceph names them, with illegal characters translated to underscores, and `ceph_` prefixed to all names. All Ceph daemon statistics have a `ceph_daemon` label that identifies the type and ID of the daemon they come from, for example: `osd.123`. Some statistics can come from different types of daemons, so when querying you will want to filter on Ceph daemons starting with `osd` to avoid mixing in the Ceph Monitor and RocksDB stats. The global Ceph storage cluster statistics have
labels appropriate to what they report on. For example, metrics relating to pools have a \texttt{pool_id} label. The long running averages that represent the histograms from core Ceph are represented by a pair of sum and count performance metrics.

The following example queries can be used in the Prometheus expression browser:

**Show the physical disk utilization of an OSD**

\[
\frac{\text{irate(node_disk_io_time_ms}[1m])}{10} \text{ and on(device,instance)} \text{ ceph_disk_occupation}{\text{ceph_daemon}=\text{osd.1}}
\]

**Show the physical IOPS of an OSD as seen from the operating system**

\[
\text{irate(node_disk_reads_completed}[1m]) + \text{irate(node_disk_writes_completed}[1m]) \text{ and on (device, instance)} \text{ ceph_disk_occupation}{\text{ceph_daemon}=\text{osd.1}}
\]

**Pool and OSD metadata series**

Special data series are output to enable the displaying and the querying on certain metadata fields. Pools have a \texttt{ceph_pool_metadata} field, for example:

\[
\text{ceph_pool_metadata}{pool_id=\text{2},name=\text{cephfs_metadata_a}} \text{ 1.0}
\]

OSDs have a \texttt{ceph_osd_metadata} field, for example:

\[
\text{ceph_osd_metadata}{cluster_addr=172.21.9.34:6802/19096,device_class=\text{ssd},ceph_daemon=\text{osd.0},public_addr=172.21.9.34:6801/19096,weight=\text{1.0}} \text{ 1.0}
\]

**Correlating drive statistics with \texttt{node_exporter}**

The Prometheus output from Ceph is designed to be used in conjunction with the generic node monitoring from the Prometheus node exporter. Correlation of Ceph OSD statistics with the generic node monitoring drive statistics, special data series are output, for example:

\[
\text{ceph_disk_occupation}{\text{ceph_daemon}=\text{osd.0},device=\text{sdd},exported_instance=\text{node1}}
\]

To get disk statistics by an OSD ID, use either the \texttt{and} operator or the asterisk (*) operator in the Prometheus query. All metadata metrics have the value of 1 so they act neutral with asterisk operator. Using asterisk operator allows to use \texttt{group_left} and \texttt{group_right} grouping modifiers, so that the resulting metric has additional labels from one side of the query. For example:

\[
\text{rate(node_disk_bytes_written}[30s]) \text{ and on (device,instance)} \text{ ceph_disk_occupation}{\text{ceph_daemon}=\text{osd.0}}
\]

**Using \texttt{label_replace}**

The \texttt{label_replace} function can add a label to, or alter a label of, a metric within a query. To correlate an OSD and its disks write rate, the following query can be used:
label_replace(rate(node_disk_bytes_written[30s]), "exported_instance", "$1", "instance", "\(.*\):.*\)\) and on (device,exported_instance) ceph_disk_occupation{ceph_daemon="osd.0"}
- OSD Host Less - Free Capacity Check
- OSD's With High Response Times
- Network Errors
- Pool Capacity High
- Monitors Down
- Overall Cluster Capacity Low
- OSDs With High PG Count

3.3.6.3. Accessing the Alert Status dashboard

Certain Red Hat Ceph Storage Dashboard alerts are configured by default in the Alert Status dashboard. This section shows two ways to access it.

**Procedure**

To access the dashboard:

- In the main At the Glance dashboard, click the Active Alerts panel in the upper-right corner.

Or...

- Click the dashboard menu from in the upper-left corner next to the Grafana icon. Select Alert Status.

3.3.6.4. Configuring the Notification Target

A notification channel called cephmetrics is automatically created during installation. All preconfigured alerts reference the cephmetrics channel but before you can receive the alerts, complete the notification channel definition by selecting the desired notification type. The Grafana platform supports a number of different notification types including email, Slack, and PagerDuty.

**Procedure**

- To configure the notification channel, follow the instructions in the Alert Notifications section on the Grafana web page.

3.3.6.5. Changing the Default Alerts and Adding New Ones

This section explains how to change the trigger value on already configured alerts and how to add new alerts to the Alert Status dashboard.

**Procedure**

- To change the trigger value on alerts or to add new alerts, follow the Alerting Engine & Rules Guide on the Grafana web pages.

**IMPORTANT**

To prevent overriding custom alerts, the Alert Status dashboard will not be updated when upgrading the Red Hat Ceph Storage Dashboard packages when you change the trigger values or add new alerts.
Additional Resources

- The Grafana web page
CHAPTER 4. OVERRIDES

By default, Ceph will reflect the current status of OSDs and perform normal operations such as rebalancing, recovering, and scrubbing. From time to time, it may be advantageous to override Ceph’s default behavior.

4.1. SETTING AND UNSETTING OVERRIDES

To override Ceph’s default behavior, use the `ceph osd set` command and the behavior you wish to override. For example:

```bash
# ceph osd set <flag>
```

Once you set the behavior, `ceph health` will reflect the override(s) that you have set for the cluster.

To cease overriding Ceph’s default behavior, use the `ceph osd unset` command and the override you wish to cease. For example:

```bash
# ceph osd unset <flag>
```

<table>
<thead>
<tr>
<th>Flag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>noin</td>
<td>Prevents OSDs from being treated as in the cluster.</td>
</tr>
<tr>
<td>noout</td>
<td>Prevents OSDs from being treated as out of the cluster.</td>
</tr>
<tr>
<td>noun</td>
<td>Prevents OSDs from being treated as up and running.</td>
</tr>
<tr>
<td>nodown</td>
<td>Prevents OSDs from being treated as down.</td>
</tr>
<tr>
<td>full</td>
<td>Makes a cluster appear to have reached its full ratio, and thereby prevents write operations.</td>
</tr>
<tr>
<td>pause</td>
<td>Ceph will stop processing read and write operations, but will not affect OSD in, out, up or down statuses.</td>
</tr>
<tr>
<td>nobackfill</td>
<td>Ceph will prevent new backfill operations.</td>
</tr>
<tr>
<td>norebalance</td>
<td>Ceph will prevent new rebalancing operations.</td>
</tr>
<tr>
<td>norecover</td>
<td>Ceph will prevent new recovery operations.</td>
</tr>
<tr>
<td>noscrub</td>
<td>Ceph will prevent new scrubbing operations.</td>
</tr>
<tr>
<td>nodeep-scrub</td>
<td>Ceph will prevent new deep scrubbing operations.</td>
</tr>
<tr>
<td>notieragent</td>
<td>Ceph will disable the process that is looking for cold/dirty objects to flush and evict.</td>
</tr>
</tbody>
</table>
4.2. USE CASES

- **noin**: Commonly used with **noout** to address flapping OSDs.

- **noout**: If the **mon osd report timeout** is exceeded and an OSD has not reported to the monitor, the OSD will get marked **out**. If this happens erroneously, you can set **noout** to prevent the OSD(s) from getting marked **out** while you troubleshoot the issue.

- **noup**: Commonly used with **nodown** to address flapping OSDs.

- **nodown**: Networking issues may interrupt Ceph 'heartbeat' processes, and an OSD may be **up** but still get marked down. You can set **nodown** to prevent OSDs from getting marked down while troubleshooting the issue.

- **full**: If a cluster is reaching its **full_ratio**, you can pre-emptively set the cluster to **full** and expand capacity. NOTE: Setting the cluster to **full** will prevent write operations.

- **pause**: If you need to troubleshoot a running Ceph cluster without clients reading and writing data, you can set the cluster to **pause** to prevent client operations.

- **nobackfill**: If you need to take an OSD or node **down** temporarily, (e.g., upgrading daemons), you can set **nobackfill** so that Ceph will not backfill while the OSD(s) is **down**.

- **norecover**: If you need to replace an OSD disk and don’t want the PGs to recover to another OSD while you are hotswapping disks, you can set **norecover** to prevent the other OSDs from copying a new set of PGs to other OSDs.

- **noscrub** and **nodeep-scrub**: If you want to prevent scrubbing (e.g., to reduce overhead during high loads, recovery, backfilling, rebalancing, etc.), you can set **noscrub** and/or **nodeep-scrub** to prevent the cluster from scrubbing OSDs.

- **notieragent**: If you want to stop the tier agent process from finding cold objects to flush to the backing storage tier, you may set **notieragent**.
CHAPTER 5. USER MANAGEMENT

This section describes Ceph client users, and their authentication and authorization with the Red Hat Ceph Storage cluster. Users are either individuals or system actors such as applications, which use Ceph clients to interact with the Red Hat Ceph Storage cluster daemons.

When Ceph runs with authentication and authorization enabled (enabled by default), you must specify a user name and a keyring containing the secret key of the specified user (usually by using the command line). If you do not specify a user name, Ceph will use the `client.admin` administrative user as the default user name. If you do not specify a keyring, Ceph will look for a keyring by using the `keyring` setting in the Ceph configuration. For example, if you execute the `ceph health` command without specifying a user or keyring:

```
# ceph health
```

Ceph interprets the command like this:

```
# ceph -n client.admin --keyring=/etc/ceph/ceph.client.admin.keyring health
```

Alternatively, you may use the `CEPH_ARGS` environment variable to avoid re-entry of the user name and secret.

For details on configuring the Red Hat Ceph Storage cluster to use authentication, see Configuration Guide for Red Hat Ceph Storage 3.

5.1. BACKGROUND

Irrespective of the type of Ceph client, for example, block device, object store, file system, native API, or the Ceph command line, Ceph stores all data as objects within pools. Ceph users must have access to pools in order to read and write data. Additionally, administrative Ceph users must have permissions to execute Ceph’s administrative commands. The following concepts will help you understand Ceph user management.

5.1.1. User

A user of the Red Hat Ceph Storage cluster is either an individual or a system actor such as an application. Creating users allows you to control who (or what) can access the storage cluster, its pools, and the data within pools.

Ceph has the notion of a **type** of user. For the purposes of user management, the type will always be `client`. Ceph identifies users in period (.) delimited form consisting of the user type and the user ID: for example, `TYPE.ID, client.admin, or client.user1`. The reason for user typing is that Ceph
monitors, and OSDs also use the Cephx protocol, but they are not clients. Distinguishing the user type helps to distinguish between client users and other users—streamlining access control, user monitoring and traceability.

Sometimes Ceph’s user type may seem confusing, because the Ceph command line allows you to specify a user with or without the type, depending upon the command line usage. If you specify --user or --id, you can omit the type. So client.user1 can be entered simply as user1. If you specify --name or -n, you must specify the type and name, such as client.user1. We recommend using the type and name as a best practice wherever possible.

NOTE

A Red Hat Ceph Storage cluster user is not the same as a Ceph Object Storage user. The object gateway uses a Red Hat Ceph Storage cluster user to communicate between the gateway daemon and the storage cluster, but the gateway has its own user management functionality for its end users.

5.1.2. Authorization (Capabilities)

Ceph uses the term "capabilities" (caps) to describe authorizing an authenticated user to exercise the functionality of the monitors and OSDs. Capabilities can also restrict access to data within a pool or a namespace within a pool. A Ceph administrative user sets a user’s capabilities when creating or updating a user.

Capability syntax follows the form:

```
<daemon_type> 'allow <capability>' [<daemon_type> 'allow <capability>']
```

- **Monitor Caps:** Monitor capabilities include r, w, x, allow profile <cap>, and profile rbd. For example:

  ```
  mon 'allow rwx`
  mon 'allow profile osd'
  ```

- **OSD Caps:** OSD capabilities include r, w, x, class-read, class-write, profile osd, profile rbd, and profile rbd-read-only. Additionally, OSD capabilities also allow for pool and namespace settings. :

  ```
  osd 'allow <capability>' [pool=<pool_name>] [namespace=<namespace_name>]
  ```

NOTE

The Ceph Object Gateway daemon (radosgw) is a client of the Ceph Storage Cluster, so it isn’t represented as a Ceph Storage Cluster daemon type.

The following entries describe each capability.

<table>
<thead>
<tr>
<th>allow</th>
<th>Precedes access settings for a daemon.</th>
</tr>
</thead>
<tbody>
<tr>
<td>r</td>
<td>Gives the user read access. Required with monitors to retrieve the CRUSH map.</td>
</tr>
</tbody>
</table>
### 5.1.3. Pool

A pool defines a storage strategy for Ceph clients, and acts as a logical partition for that strategy.

In Ceph deployments, it is common to create a pool to support different types of use cases, for example, cloud volumes/images, object storage, hot storage, cold storage, and so on. When deploying Ceph as a back end for OpenStack, a typical deployment would have pools for volumes, images, backups and virtual machines, and users such as `client.glance`, `client.cinder`, and so on.

### 5.1.4. Namespace

Objects within a pool can be associated to a namespace—a logical group of objects within the pool. A user’s access to a pool can be associated with a namespace such that reads and writes by the user take place only within the namespace. Objects written to a namespace within the pool can only be accessed by users who have access to the namespace.

**NOTE**

Currently, namespaces are only useful for applications written on top of `librados`. Ceph clients such as block device and object storage do not currently support this feature.

The rationale for namespaces is that pools can be a computationally expensive method of segregating...
data by use case, because each pool creates a set of placement groups that get mapped to OSDs. If multiple pools use the same CRUSH hierarchy and ruleset, OSD performance may degrade as load increases.

For example, a pool should have approximately 100 placement groups per OSD. So an exemplary cluster with 1000 OSDs would have 100,000 placement groups for one pool. Each pool mapped to the same CRUSH hierarchy and ruleset would create another 100,000 placement groups in the exemplary cluster. By contrast, writing an object to a namespace simply associates the namespace to the object name with out the computational overhead of a separate pool. Rather than creating a separate pool for a user or set of users, you may use a namespace.

**NOTE**

Only available using **librados** at this time.

### 5.2. MANAGING USERS

User management functionality provides system administrators with the ability to create, update and delete Red Hat Ceph Storage cluster users.

When you create or delete users in a Red Hat Ceph Storage cluster, you may need to distribute keys to clients so that they can be added to keyrings. See Keyring Management for details.

#### 5.2.1. List Users

To list the users in the storage cluster, execute the following:

```
# ceph auth list
```

Ceph will list out all users in the storage cluster. For example, in a two-node exemplary storage cluster, `ceph auth list` will output something that looks like this:

```
nstalled auth entries:
osd.0
  key: AQCvCbtToC6MDhAATtuT70S1+DymPCfDSsyV4w==
caps: [mon] allow profile osd
caps: [osd] allow *
osd.1
  key: AQC4CbtTCFJBChAAVq5spj0ff4eHZICxI0VZeA==
caps: [mon] allow profile osd
caps: [osd] allow *
client.admin
  key: AQBHCbtT6APDHhAA5W00cBchwkQjh3dkKsyPjw==
caps: [mds] allow
  caps: [mon] allow *
caps: [osd] allow *
client.bootstrap-mds
  key: AQBICbtT0K9uGBAAde5zcIGHZL3T/u2g6EBw==
caps: [mon] allow profile bootstrap-mds
client.bootstrap-osd
  key: AQBHCbtT4Gxq0RAADE5u7RkpCN/0o4e5W0uBtw==
caps: [mon] allow profile bootstrap-osd
```
Note that the **TYPE.ID** notation for users applies such that **osd.0** is a user of type **osd** and its ID is **0**, **client.admin** is a user of type **client** and its ID is **admin**, that is, the default **client.admin** user. Note also that each entry has a **key: <value>** entry, and one or more **caps: entries**.

You may use the `-o <file_name>` option with `ceph auth list` to save the output to a file.

### 5.2.2. Get a User

To retrieve a specific user, key and capabilities, execute the following:

**Syntax**

```
# ceph auth get <TYPE.ID>
```

**Example**

```
# ceph auth get client.admin
```

You may also use the `-o <file_name>` option with `ceph auth get` to save the output to a file. Developers may also execute the following:

**Syntax**

```
# ceph auth export <TYPE.ID>
```

**Example**

```
# ceph auth export client.admin
```

The **auth export** command is identical to **auth get**, but also prints out the internal **auid**, which isn’t relevant to end users.

### 5.2.3. Add a User

Adding a user creates a username, that is, **TYPE.ID**, a secret key and any capabilities included in the command you use to create the user.

A user’s key enables the user to authenticate with the Ceph Storage Cluster. The user’s capabilities authorize the user to read, write, or execute on Ceph monitors (**mon**), Ceph OSDs (**osd**) or Ceph Metadata Servers (**mds**).

There are a few ways to add a user:

- **ceph auth add**: This command is the canonical way to add a user. It will create the user, generate a key and add any specified capabilities.

- **ceph auth get-or-create**: This command is often the most convenient way to create a user, because it returns a keyfile format with the user name (in brackets) and the key. If the user already exists, this command simply returns the user name and key in the keyfile format. You may use the `-o <file_name>` option to save the output to a file.

- **ceph auth get-or-create-key**: This command is a convenient way to create a user and return the user’s key only. This is useful for clients that need the key only, for example,
libvirt. If the user already exists, this command simply returns the key. You may use the -o <file_name> option to save the output to a file.

When creating client users, you may create a user with no capabilities. A user with no capabilities is useless beyond mere authentication, because the client cannot retrieve the cluster map from the monitor. However, you can create a user with no capabilities if you wish to defer adding capabilities later using the ceph auth caps command.

A typical user has at least read capabilities on the Ceph monitor and read and write capability on Ceph OSDs. Additionally, a user’s OSD permissions are often restricted to accessing a particular pool:

```
# ceph auth add client.john mon 'allow r' osd 'allow rw pool=liverpool'
# ceph auth get-or-create client.paul mon 'allow r' osd 'allow rw pool=liverpool'
# ceph auth get-or-create client.george mon 'allow r' osd 'allow rw pool=liverpool' -o george.keyring
# ceph auth get-or-create-key client.ringo mon 'allow r' osd 'allow rw pool=liverpool' -o ringo.key
```

### IMPORTANT

If you provide a user with capabilities to OSDs, but you DO NOT restrict access to particular pools, the user will have access to ALL pools in the cluster!

### 5.2.4. Modify User Capabilities

The ceph auth caps command allows you to specify a user and change the user’s capabilities. To add capabilities, use the form:

**Syntax**

```
# ceph auth caps <USERTYPE.USERID> <daemon> 'allow [r|w|x|*|...] [pool= <pool_name>] [namespace=<namespace_name>]'
```

**Example**

```
# ceph auth caps client.john mon 'allow r' osd 'allow rw pool=liverpool'
# ceph auth caps client.paul mon 'allow rw' osd 'allow rwx pool=liverpool'
# ceph auth caps client.brian-manager mon 'allow *' osd 'allow *'
```

To remove a capability, you may reset the capability. If you want the user to have no access to a particular daemon that was previously set, specify an empty string. For example:

```
# ceph auth caps client.ringo mon ' ' osd ' '
```

See Section 5.1.2, “Authorization (Capabilities)” for additional details on capabilities.

### 5.2.5. Delete a User

To delete a user, use ceph auth del:

```
# ceph auth del {TYPE}.{ID}
```
Where \{\text{TYPE}\} is one of \text{client}, \text{osd}, \text{mon}, \text{or mds}, and \{\text{ID}\} is the user name or ID of the daemon.

### 5.2.6. Print a User’s Key

To print a user’s authentication key to standard output, execute the following:

```bash
# ceph auth print-key \{\text{TYPE}\}.\{\text{ID}\}
```

Where \{\text{TYPE}\} is one of \text{client}, \text{osd}, \text{mon}, \text{or mds}, and \{\text{ID}\} is the user name or ID of the daemon.

Printing a user’s key is useful when you need to populate client software with a user’s key, for example, \text{libvirt}.

```bash
# mount -t ceph \langle \text{hostname} \rangle:/\langle \text{mount\_point} \rangle -o name=client.user,secret=`ceph auth print-key client.user`
```

### 5.2.7. Import a User

To import one or more users, use \texttt{ceph auth import} and specify a keyring:

**Syntax**

```bash
# ceph auth import -i </path/to/keyring>
```

**Example**

```bash
# ceph auth import -i /etc/ceph/ceph.keyring
```

**NOTE**

The ceph storage cluster will add new users, their keys and their capabilities and will update existing users, their keys and their capabilities.

### 5.3. KEYRING MANAGEMENT

When you access Ceph by using a Ceph client, the Ceph client will look for a local keyring. Ceph presets the \texttt{keyring} setting with the following four keyring names by default so you don’t have to set them in the Ceph configuration file unless you want to override the defaults, which is not recommended:

- \texttt{/etc/ceph/\$cluster.$name.keyring}
- \texttt{/etc/ceph/\$cluster.keyring}
- \texttt{/etc/ceph/keyring}
- \texttt{/etc/ceph/keyring.bin}

The \texttt{\$cluster} metavariable is the Ceph storage cluster name as defined by the name of the Ceph configuration file, that is, \texttt{ceph.conf} means the cluster name is \texttt{ceph}; thus, \texttt{ceph.keyring}. The \texttt{\$name} metavariable is the user type and user ID, for example, \texttt{client.admin}; thus, \texttt{ceph.client.admin.keyring}. 


NOTE

When executing commands that read or write to `/etc/ceph`, you may need to use `sudo` to execute the command as `root`.

After you create a user, for example, `client.ringo`, you must get the key and add it to a keyring on a Ceph client so that the user can access the Ceph Storage Cluster.

See Chapter 5, User Management for details on how to list, get, add, modify and delete users directly in the Ceph Storage Cluster. However, Ceph also provides the `ceph-authtool` utility to allow you to manage keyrings from a Ceph client.

### 5.3.1. Create a Keyring

When you use the procedures in the Managing Users_ section to create users, you need to provide user keys to the Ceph client(s) so that the Ceph client can retrieve the key for the specified user and authenticate with the Ceph Storage Cluster. Ceph Clients access keyrings to lookup a user name and retrieve the user's key.

The `ceph-authtool` utility allows you to create a keyring. To create an empty keyring, use `--create-keyring` or `-C`. For example:

```
# ceph-authtool --create-keyring /path/to/keyring
```

When creating a keyring with multiple users, we recommend using the cluster name, for example, `$cluster.keyring` for the keyring file name and saving it in the `/etc/ceph/` directory so that the `keyring` configuration default setting will pick up the filename without requiring you to specify it in the local copy of the Ceph configuration file. For example, create `ceph.keyring` by executing the following:

```
# ceph-authtool -C /etc/ceph/ceph.keyring
```

When creating a keyring with a single user, we recommend using the cluster name, the user type and the user name and saving it in the `/etc/ceph/` directory. For example, `ceph.client.admin.keyring` for the `client.admin` user.

To create a keyring in `/etc/ceph/`, you must do so as `root`. This means the file will have `rw` permissions for the `root` user only, which is appropriate when the keyring contains administrator keys. However, if you intend to use the keyring for a particular user or group of users, ensure that you execute `chown` or `chmod` to establish appropriate keyring ownership and access.

### 5.3.2. Add a User to a Keyring

When you add a user to the Ceph storage cluster, you can use the `get` procedure to retrieve a user, key and capabilities, then save the user to a keyring file.

When you only want to use one user per keyring, the Get a User_ procedure with the `-o` option will save the output in the keyring file format. For example, to create a keyring for the `client.admin` user, execute the following:

```
# ceph auth get client.admin -o /etc/ceph/ceph.client.admin.keyring
```

Notice that we use the recommended file format for an individual user.
When you want to import users to a keyring, you can use `ceph-authtool` to specify the destination keyring and the source keyring. For example:

```
# ceph-authtool /etc/ceph/ceph.keyring --import-keyring
/etc/ceph/ceph.client.admin.keyring
```

### 5.3.3. Create a User

Ceph provides the Add a User function to create a user directly in the Ceph Storage Cluster. However, you can also create a user, keys and capabilities directly on a Ceph client keyring. Then, you can import the user to the Ceph Storage Cluster. For example:

```
# ceph-authtool -n client.ringo --cap osd 'allow rwx' --cap mon 'allow rwx' /etc/ceph/ceph.keyring
```

See [Section 5.1.2, “Authorization (Capabilities)”](#) for additional details on capabilities.

You can also create a keyring and add a new user to the keyring simultaneously. For example:

```
# ceph-authtool -C /etc/ceph/ceph.keyring -n client.ringo --cap osd 'allow rwx' --cap mon 'allow rwx' --gen-key
```

In the foregoing scenarios, the new user `client.ringo` is only in the keyring. To add the new user to the Ceph Storage Cluster, you must still add the new user to the Ceph Storage Cluster:

```
# ceph auth add client.ringo -i /etc/ceph/ceph.keyring
```

### 5.3.4. Modify a User

To modify the capabilities of a user record in a keyring, specify the keyring, and the user followed by the capabilities, for example:

```
# ceph-authtool /etc/ceph/ceph.keyring -n client.ringo --cap osd 'allow rwx' --cap mon 'allow rwx'
```

To update the user to the Ceph storage cluster, you must update the user in the keyring to the user entry in the the Ceph storage cluster:

```
# ceph auth import -i /etc/ceph/ceph.keyring
```

See [Section 5.2.7, “Import a User”](#) for details on updating a Ceph Storage Cluster user from a keyring.

You may also modify user capabilities directly in the storage cluster, store the results to a keyring file; then, import the keyring into the main `ceph.keyring` file.

### 5.4. COMMAND LINE USAGE

Ceph supports the following usage for user name and secret:

```
--id | --user
```

**Description**
Ceph identifies users with a type and an ID (e.g., **TYPE.ID** or **client.admin**, **client.user1**). The **id**, **name** and **-n** options enable you to specify the ID portion of the user name (e.g., **admin**, **user1**, **foo**, etc.). You can specify the user with the **--id** and omit the type. For example, to specify user **client.foo** enter the following:

```bash
# ceph --id foo --keyring /path/to/keyring health
# ceph --user foo --keyring /path/to/keyring health
```

**--name | -n**

**Description**

Ceph identifies users with a type and an ID (e.g., **TYPE.ID** or **client.admin**, **client.user1**). The **--name** and **-n** options enables you to specify the fully qualified user name. You must specify the user type (typically **client**) with the user ID. For example:

```bash
# ceph --name client.foo --keyring /path/to/keyring health
# ceph -n client.foo --keyring /path/to/keyring health
```

**--keyring**

**Description**

The path to the keyring containing one or more user name and secret. The **--secret** option provides the same functionality, but it does not work with Ceph RADOS Gateway, which uses **--secret** for another purpose. You may retrieve a keyring with **ceph auth get-or-create** and store it locally. This is a preferred approach, because you can switch user names without switching the keyring path. For example:

```bash
# rbd map foo --pool rbd myimage --id client.foo --keyring
/path/to/keyring
```

### 5.5. LIMITATIONS

The **cephx** protocol authenticates Ceph clients and servers to each other. It is not intended to handle authentication of human users or application programs run on their behalf. If that effect is required to handle the access control needs, you must have another mechanism, which is likely to be specific to the front end used to access the Ceph object store. This other mechanism has the role of ensuring that only acceptable users and programs are able to run on the machine that Ceph will permit to access its object store.

The keys used to authenticate Ceph clients and servers are typically stored in a plain text file with appropriate permissions in a trusted host.

**IMPORTANT**

Storing keys in plaintext files has security shortcomings, but they are difficult to avoid, given the basic authentication methods Ceph uses in the background. Those setting up Ceph systems should be aware of these shortcomings.

In particular, arbitrary user machines, especially portable machines, should not be configured to interact directly with Ceph, since that mode of use would require the storage of a plaintext authentication key on an insecure machine. Anyone who stole that machine or obtained surreptitious access to it could obtain the key that will allow them to authenticate their own machines to Ceph.
Rather than permitting potentially insecure machines to access a Ceph object store directly, users should be required to sign in to a trusted machine in the environment using a method that provides sufficient security for the purposes. That trusted machine will store the plaintext Ceph keys for the human users. A future version of Ceph may address these particular authentication issues more fully.

At the moment, none of the Ceph authentication protocols provide secrecy for messages in transit. Thus, an eavesdropper on the wire can hear and understand all data sent between clients and servers in Ceph, even if he cannot create or alter them. Those storing sensitive data in Ceph should consider encrypting their data before providing it to the Ceph system.

For example, Ceph Object Gateway provides S3 API Server-side Encryption, which encrypts unencrypted data received from a Ceph Object Gateway client before storing it in the Ceph Storage cluster and similarly decrypts data retrieved from the Ceph Storage cluster before sending it back to the client. To ensure encryption in transit between the client and the Ceph Object Gateway, the Ceph Object Gateway should be configured to use SSL.
CHAPTER 6. USING THE CEPH-VOLUME UTILITY TO DEPLOY OSDS

The ceph-volume utility is a single purpose command-line tool to deploy logical volumes as OSDs. It uses a plugin-type framework to deploying OSDs with different device technologies. The ceph-volume utility follows a similar workflow of the ceph-disk utility for deploying OSDs, with a predictable, and robust way of preparing, activating, and starting OSDs. Currently, the ceph-volume utility only supports the lvm plugin, with the plan to support others technologies in the future.

**IMPORTANT**
The ceph-disk command is deprecated.

6.1. USING THE CEPH-VOLUME LVM PLUGIN

By making use of LVM tags, the lvm subcommand is able to store and re-discover by querying devices associated with OSDs so they can be activated. This includes support for lvm-based technologies like dm-cache as well.

When using ceph-volume, the use of dm-cache is transparent, and treats dm-cache like a logical volume. The performance gains and losses when using dm-cache will depend on the specific workload. Generally, random and sequential reads will see an increase in performance at smaller block sizes; while random and sequential writes will see a decrease in performance at larger block sizes.

To use the LVM plugin, add lvm as a subcommand to the ceph-volume command:

```
ceph-volume lvm
```

There are three subcommands to the lvm subcommand, as follows:

- prepare
- activate
- create
- batch

**NOTE**
Using the create subcommand combines the prepare and activate subcommands into one subcommand. See the create subcommand section for more details.

6.1.1. Preparing OSDs

The prepare subcommand prepares an OSD backend object store and consumes logical volumes for both the OSD data and journal. There is no default object storage type. The object storage type requires either the --filestore or --bluestore option to be set at preparation time. Starting with Red Hat Ceph Storage 3.2, support for the BlueStore object storage type is available. The prepare subcommand will not create or modify the logical volumes, except for adding some extra metadata using LVM tags.
LVM tags makes volumes easier to discover later, and help identify them as part of a Ceph system, and what role they have. The `ceph-volume lvm prepare` command adds the following list of LVM tags:

- `cluster_fsid`
- `data_device`
- `journal_device`
- `encrypted`
- `osd_fsid`
- `osd_id`
- `journal_uuid`

The `prepare` process is very strict, it requires two logical volumes that are ready for use, and requires the minimum size for an OSD data and journal. The journal device can be either a logical volume or a partition.

Here is the `prepare` workflow process:

1. Accept logical volumes for data and journal
2. Generate a UUID for the OSD
3. Ask the Ceph Monitor to get an OSD identifier reusing the generated UUID
4. OSD data directory is created and data volume mounted
5. Journal is symlinked from data volume to journal location
6. The `monmap` is fetched for activation
7. Device is mounted and the data directory is populated by `ceph-osd`
8. LVM tags are assigned to the OSD data and journal volumes

Do the following step on an OSD node, and as the `root` user, to prepare a simple OSD deployment using LVM:

```
ceph-volume lvm prepare --bluestore --data $VG_NAME/$LV_NAME
```

For example:

```
# ceph-volume lvm prepare --bluestore --data example_vg/data_lv
```

For BlueStore, you can also specify the `--block.db` and `--block.wal` options, if you want to use a separate device for RocksDB.

Here is an example of using FileStore with a partition as a journal device:

```
# ceph-volume lvm prepare --filestore --data example_vg/data_lv --journal /dev/sdc1
```
When using a partition, it must contain a **PARTUUID** discoverable by the `blkid` command, this way it can be identified correctly regardless of the device name or path.

**IMPORTANT**

The `ceph-volume` LVM plugin does not create partitions on a raw disk device. Creating this partition has to be done before using a partition for the OSD journal device.

### 6.1.2. Activating OSDs

Once the prepare process is done, the OSD is ready to go active. The activation process enables a Systemd unit at boot time which allows the correct OSD identifier and its UUID to be enabled and mounted.

Here is the `activate` workflow process:

1. Requires both OSD id and OSD uuid
2. Enable the systemd unit with matching OSD id and OSD uuid
3. The systemd unit will ensure all devices are ready and mounted
4. The matching `ceph-osd` systemd unit will get started

Do the following step on an OSD node, and as the `root` user, to activate an OSD:

```
ceph-volume lvm activate --filestore $OSD_ID $OSD_UUID
```

For example:

```
# ceph-volume lvm activate --filestore 0 0263644D-0BF1-4D6D-BC34-28BD98AE3BC8
```

**NOTE**

There are no side-effects when running this command multiple times.

### 6.1.3. Creating OSDs

The `create` subcommand wraps the two-step process to deploy a new OSD by calling the `prepare` subcommand and then calling the `activate` subcommand into a single subcommand. The reason to use `prepare` and then `activate` separately is to gradually introduce new OSDs into a storage cluster, and avoiding large amounts of data being rebalanced. There is nothing different to the process except the OSD will become *up* and *in* immediately after completion.

Do the following step on an OSD node, and as the `root` user:

```
ceph-volume lvm create --filestore --data $VG_NAME/$LV_NAME --journal $JOURNAL_DEVICE
```

For example:
# ceph-volume lvm create --filestore --data example_vg/data_lv --journal example_vg/journal_lv

6.1.4. Using batch mode

The `batch` subcommand automates the creation of multiple OSDs when single devices are provided. The `ceph-volume` command decides the best method in creating the OSDs based on drive type. This best method is dependant on the object store format, BlueStore or FileStore.

BlueStore is the default object store type for OSDs. When using BlueStore, OSD optimization depends on three different scenarios based on the devices being used. If all devices are traditional hard drives, then one OSD per device is created. If all devices are solid state drives, then two OSDs per device are created. If there is a mix of traditional hard drives and solid state drives, then data is put on the traditional hard drives, and the journal (`block.db`) is created as large as possible on the solid state drive.

**NOTE**

The `batch` subcommand does not support the creating of a separate logical volume for the write-ahead-log (`block.wal`) device.

**BlusStore example**

```
# ceph-volume lvm batch --bluestore /dev/sda /dev/sdb /dev/nvme0n1
```

When using FileStore, OSD optimization depends on two different scenarios based on the devices being used. If all devices are traditional hard drives or are solid state drives, then one OSD per device is created, collocating the journal on the same device. If there is a mix of traditional hard drives and solid state drives, then data is put on the traditional hard drives, and the journal is created on the solid state drive using the sizing parameters specified in the Ceph configuration file, by default `ceph.conf`, with a default journal size of 5 GB.

**FileStore example**

```
# ceph-volume lvm batch --filestore /dev/sda /dev/sdb
```
CHAPTER 7. BENCHMARKING PERFORMANCE

The purpose of this section is to give Ceph administrators a basic understanding of Ceph’s native benchmarking tools. These tools will provide some insight into how the Ceph storage cluster is performing. This is not the definitive guide to Ceph performance benchmarking, nor is it a guide on how to tune Ceph accordingly.

7.1. PERFORMANCE BASELINE

The OSD (including the journal) disks and the network throughput should each have a performance baseline to compare against. You can identify potential tuning opportunities by comparing the baseline performance data with the data from Ceph’s native tools. Red Hat Enterprise Linux has many built-in tools, along with a plethora of open source community tools, available to help accomplish these tasks. For more details about some of the available tools, see this Knowledgebase article.

7.2. STORAGE CLUSTER

Ceph includes the rados bench command to do performance benchmarking on a RADOS storage cluster. The command will execute a write test and two types of read tests. The --no-cleanup option is important to use when testing both read and write performance. By default the rados bench command will delete the objects it has written to the storage pool. Leaving behind these objects allows the two read tests to measure sequential and random read performance.

NOTE

Before running these performance tests, drop all the file system caches by running the following:

```
# echo 3 | sudo tee /proc/sys/vm/drop_caches && sudo sync
```

1. Create a new storage pool:

```
# ceph osd pool create testbench 100 100
```

2. Execute a write test for 10 seconds to the newly created storage pool:

```
# rados bench -p testbench 10 write --no-cleanup
```

Example Output

```
Maintaining 16 concurrent writes of 4194304 bytes for up to 10 seconds or 0 objects
Object prefix: benchmark_data_cephn1.home.network_10510

sec Cur ops started finished avg MB/s cur MB/s last lat
avg lat
0 0 0 0 0 0 0 -
0 1 16 16 0 0 0 -
0 2 16 16 0 0 0 -
0 3 16 16 0 0 0 -
```
3. Execute a sequential read test for 10 seconds to the storage pool:

```
# rados bench -p testbench 10 seq
```

**Example Output**

```
sec  Cur ops  started  finished  avg MB/s  cur MB/s  last lat  avg lat
 0     0       0        0         0         0        0        0       -
```
Total time run: 0.804869
Total reads made: 28
Read size: 4194304
Bandwidth (MB/sec): 139.153
Average Latency: 0.420841
Max latency: 0.706133
Min latency: 0.0816332

4. Execute a random read test for 10 seconds to the storage pool:

```bash
# rados bench -p testbench 10 rand
```

**Example Output**

<table>
<thead>
<tr>
<th>sec</th>
<th>Cur ops</th>
<th>started</th>
<th>finished</th>
<th>avg MB/s</th>
<th>cur MB/s</th>
<th>last lat</th>
<th>avg lat</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>0.388125</td>
<td>16</td>
<td>46</td>
<td>30</td>
<td>119.801</td>
<td>120</td>
<td>0.440184</td>
<td>0.388125</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>81</td>
<td>65</td>
<td>129.408</td>
<td>140</td>
<td>0.577359</td>
<td>0.417461</td>
</tr>
<tr>
<td>0.409318</td>
<td>3</td>
<td>16</td>
<td>120</td>
<td>138.175</td>
<td>156</td>
<td>0.597435</td>
<td>0.409318</td>
</tr>
<tr>
<td>1419964</td>
<td>4</td>
<td>15</td>
<td>157</td>
<td>141.485</td>
<td>152</td>
<td>0.683111</td>
<td>0.419964</td>
</tr>
<tr>
<td>0.408343</td>
<td>5</td>
<td>16</td>
<td>206</td>
<td>151.553</td>
<td>192</td>
<td>0.310578</td>
<td>0.408343</td>
</tr>
<tr>
<td>0.387207</td>
<td>6</td>
<td>16</td>
<td>253</td>
<td>157.608</td>
<td>188</td>
<td>0.0745175</td>
<td>0.387207</td>
</tr>
<tr>
<td>0.39043</td>
<td>7</td>
<td>16</td>
<td>287</td>
<td>154.412</td>
<td>136</td>
<td>0.792774</td>
<td>0.39043</td>
</tr>
<tr>
<td>0.39876</td>
<td>8</td>
<td>16</td>
<td>325</td>
<td>154.044</td>
<td>152</td>
<td>0.314254</td>
<td>0.39876</td>
</tr>
<tr>
<td>0.406032</td>
<td>9</td>
<td>16</td>
<td>362</td>
<td>153.245</td>
<td>148</td>
<td>0.355576</td>
<td>0.406032</td>
</tr>
<tr>
<td>0.398372</td>
<td>10</td>
<td>16</td>
<td>405</td>
<td>155.092</td>
<td>172</td>
<td>0.64734</td>
<td>0.398372</td>
</tr>
</tbody>
</table>

Total time run: 10.302229
Total reads made: 405
Read size: 4194304
Bandwidth (MB/sec): 157.248
Average Latency: 0.405976
Max latency: 1.00869
Min latency: 0.0378431

5. To increase the number of concurrent reads and writes, use the `-t` option, which the default is 16 threads. Also, the `-b` parameter can adjust the size of the object being written. The default object size is 4MB. A safe maximum object size is 16MB. Red Hat recommends running multiple copies of these benchmark tests to different pools. Doing this shows the changes in performance from multiple clients.

Add the `--run-name <label>` option to control the names of the objects that get written.
during the benchmark test. Multiple `rados bench` commands may be ran simultaneously by changing the `--run-name` label for each running command instance. This prevents potential I/O errors that can occur when multiple clients are trying to access the same object and allows for different clients to access different objects. The `--run-name` option is also useful when trying to simulate a real world workload. For example:

```sh
# rados bench -p testbench 10 write -t 4 --run-name client1
```

Example Output

```
Maintaining 4 concurrent writes of 4194304 bytes for up to 10 seconds or 0 objects
Object prefix: benchmark_data_node1_12631

<table>
<thead>
<tr>
<th>sec</th>
<th>Cur ops</th>
<th>started</th>
<th>finished</th>
<th>avg MB/s</th>
<th>cur MB/s</th>
<th>last lat avg lat</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>3.99099</td>
<td>4</td>
<td>1.94755</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>8</td>
<td>4</td>
<td>5.32498</td>
<td>8</td>
<td>2.978</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>8</td>
<td>4</td>
<td>3.99504</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>10</td>
<td>6</td>
<td>4.79504</td>
<td>4</td>
<td>2.92419</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>10</td>
<td>7</td>
<td>4.64471</td>
<td>4</td>
<td>3.02498</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>12</td>
<td>8</td>
<td>4.55287</td>
<td>4</td>
<td>3.12204</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>14</td>
<td>10</td>
<td>4.9821</td>
<td>8</td>
<td>2.55901</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>16</td>
<td>12</td>
<td>5.31621</td>
<td>8</td>
<td>2.68769</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>17</td>
<td>13</td>
<td>5.18488</td>
<td>4</td>
<td>2.11937</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>17</td>
<td>13</td>
<td>4.71431</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>18</td>
<td>14</td>
<td>4.65486</td>
<td>2</td>
<td>2.4836</td>
</tr>
<tr>
<td>13</td>
<td>4</td>
<td>18</td>
<td>14</td>
<td>4.29757</td>
<td>0</td>
<td>-</td>
</tr>
</tbody>
</table>

Total time run:         13.123548
Total writes made:      18
Write size:             4194304
Bandwidth (MB/sec):     5.486
Stddev Bandwidth:       3.0991
Max bandwidth (MB/sec): 8
Min bandwidth (MB/sec): 0
Average Latency:        2.91578
Stddev Latency:         0.956993
Max latency:            5.72685
Min latency:            1.91967
```
6. Remove the data created by the `rados bench` command:

```
# rados -p testbench cleanup
```

### 7.3. BLOCK DEVICE

Ceph includes the `rbd bench-write` command to test sequential writes to the block device measuring throughput and latency. The default byte size is 4096, the default number of I/O threads is 16, and the default total number of bytes to write is 1 GB. These defaults can be modified by the `--io-size`, `--io-threads` and `--io-total` options respectively. For more information on the `rbd` command, see the Block Device Commands section in the Ceph Block Device Guide for Red Hat Ceph Storage 3.

**Creating a Ceph Block Device**

1. As `root`, load the `rbd` kernel module, if not already loaded:

```
# modprobe rbd
```

2. As `root`, create a 1 GB `rbd` image file in the `testbench` pool:

```
# rbd create image01 --size 1024 --pool testbench
```

**NOTE**

When creating a block device image these features are enabled by default: `layering`, `object-map`, `deep-flatten`, `journaling`, `exclusive-lock`, and `fast-diff`.

On Red Hat Enterprise Linux 7.2 and Ubuntu 16.04, users utilizing the kernel RBD client will not be able to map the block device image. You must first disable all these features, except, `layering`.

**Syntax**

```
# rbd feature disable <image_name> <feature_name>
```

**Example**

```
# rbd feature disable image1 journaling deep-flatten exclusive-lock fast-diff object-map
```

Using the `--image-feature layering` option on the `rbd create` command will only enable `layering` on newly created block device images.

This is a known issue, see the Red Hat Ceph Storage 3.2 Release Notes for more details.

All these features will work for users utilizing the user-space RBD client to access the block device images.

3. As `root`, map the image file to a device file:
4. As root, create an **ext4** file system on the block device:

```
# mkfs -t ext4 -m0 /dev/rbd/testbench/image01
```

5. As root, create a new directory:

```
# mkdir /mnt/ceph-block-device
```

6. As root, mount the block device under `/mnt/ceph-block-device/`

```
# mount /dev/rbd/testbench/image01 /mnt/ceph-block-device
```

**Execute the write performance test against the block device**

```
# rbd bench-write image01 --pool=testbench
```

**Example**

```
bench-write  io_size 4096 io_threads 16 bytes 1073741824 pattern seq
SEC       OPS   OPS/SEC   BYTES/SEC
2     11127   5479.59  22444382.79
3     11692   3901.91  15982220.33
4     12372   2953.34  12096895.42
5     12580   2300.05  9421008.60
6     13141   2101.80  8608975.15
7     13195    356.07  1458459.94
8     13820    390.35  1598876.60
9     14124    325.46  1333066.62
```
CHAPTER 8. PERFORMANCE COUNTERS

The Ceph performance counters are a collection of internal infrastructure metrics. The collection, aggregation, and graphing of this metric data can be done by an assortment of tools and can be useful for performance analytics.

8.1. ACCESS

The performance counters are available through a socket interface for the Ceph Monitors and the OSDs. The socket file for each respective daemon is located under `/var/run/ceph`, by default. The performance counters are grouped together into collection names. These collection names represent a subsystem or an instance of a subsystem.

Here is the full list of the Monitor and the OSD collection name categories with a brief description for each:

**Monitor Collection Name Categories**

- Cluster Metrics - Displays information about the storage cluster: Monitors, OSDs, Pools, and PGs
- Level Database Metrics - Displays information about the back-end `KeyValueStore` database
- Monitor Metrics - Displays general monitor information
- Paxos Metrics - Displays information on cluster quorum management
- Throttle Metrics - Displays the statistics on how the monitor is throttling

**OSD Collection Name Categories**

- Write Back Throttle Metrics - Displays the statistics on how the write back throttle is tracking unflushed IO
- Filestore Metrics - Displays information on all related filestore statistics
- Level Database Metrics - Displays information about the back-end `KeyValueStore` database
- Objecter Metrics - Displays information on various object-based operations
- Read and Write Operations Metrics - Displays information on various read and write operations
- Recovery State Metrics - Displays latencies on various recovery states
- OSD Throttle Metrics - Display the statistics on how the OSD is throttling

**RADOS Gateway Collection Name Categories**

- Object Gateway Client Metrics - Displays statistics on GET and PUT requests
- Objecter Metrics - Displays information on various object-based operations
- Object Gateway Throttle Metrics - Display the statistics on how the OSD is throttling

8.2. SCHEMA
The `ceph daemon .. perf schema` command outputs the available metrics. Each metric has an associated bit field value type.

To view the metric's schema:

```
# ceph daemon <daemon_name> perf schema
```

**NOTE**

You must run the `ceph daemon` command from the node running the daemon.

Executing `ceph daemon .. perf schema` command from the Monitor node:

```
# ceph daemon mon.`hostname -s` perf schema
```

Example

```
{
  "cluster": {
    "num_mon": {
      "type": 2
    },
    "num_mon_quorum": {
      "type": 2
    },
    "num_osd": {
      "type": 2
    },
    "num_osd_up": {
      "type": 2
    },
    "num_osd_in": {
      "type": 2
    },
    ...
  }
```

Executing the `ceph daemon .. perf schema` command from the OSD node:

```
# ceph daemon osd.0 perf schema
```

Example

```
...
"filestore": {
  "journal_queue_max_ops": {
    "type": 2
  },
  "journal_queue_ops": {
    "type": 2
  },
  "journal_ops": {
    "type": 10
  }
```
Table 8.1. The bit field value definitions

<table>
<thead>
<tr>
<th>Bit</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Floating point value</td>
</tr>
<tr>
<td>2</td>
<td>Unsigned 64-bit integer value</td>
</tr>
<tr>
<td>4</td>
<td>Average (Sum + Count)</td>
</tr>
<tr>
<td>8</td>
<td>Counter</td>
</tr>
</tbody>
</table>

Each value will have bit 1 or 2 set to indicate the type, either a floating point or an integer value. When bit 4 is set, there will be two values to read, a sum and a count. See Section 8.3.1, “Average Count and Sum” for more details. When bit 8 is set, the average for the previous interval would be the sum delta, since the previous read, divided by the count delta. Alternatively, dividing the values outright would provide the lifetime average value. Typically these are used to measure latencies, the number of requests and a sum of request latencies. Some bit values are combined, for example 5, 6 and 10. A bit value of 5 is a combination of bit 1 and bit 4. This means the average will be a floating point value. A bit value of 6 is a combination of bit 2 and bit 4. This means the average value will be an integer. A bit value of 10 is a combination of bit 2 and bit 8. This means the counter value will be an integer value.

8.3. DUMP

The ceph daemon .. perf dump command outputs the current values and groups the metrics under the collection name for each subsystem.

To view the current metric data:

```
# ceph daemon {daemon_name} perf dump
```

**NOTE**

You must run the ceph daemon command from the node running the daemon.

Executing ceph daemon .. perf dump command from the Monitor node:
# ceph daemon mon.`hostname -s` perf dump

**Example**

```
{
    "cluster": {
        "num_mon": 1,
        "num_mon_quorum": 1,
        "num_osd": 2,
        "num_osd_up": 2,
        "num_osd_in": 2,
        ...
```

To view a short description of each Monitor metric available, please see the table below.

Executing the `ceph daemon .. perf dump` command from the OSD node:

```bash
# ceph daemon osd.0 perf dump
```

**Example**

```
...
"filestore": {
    "journal_queue_max_ops": 300,
    "journal_queue_ops": 0,
    "journal_ops": 992,
    "journal_queue_max_bytes": 33554432,
    "journal_queue_bytes": 0,
    "journal_bytes": 934537,
    "journal_latency": {
        "avgcount": 992,
        "sum": 254.975925772
    }
},
```

8.3.1. Average Count and Sum

All latency numbers have a bit field value of 5. This field contains floating point values for the average count and sum. The `avgcount` is the number of operations within this range and the `sum` is the total latency in seconds. When dividing the `sum` by the `avgcount` this will provide you with an idea of the latency per operation.

To view a short description of each OSD metric available, please see the `OSD table` below.

8.3.2. Monitor Metrics Description Tables

- Cluster Metrics Table
- Level Database Metrics Table
- General Monitor Metrics Table
- Paxos Metrics Table
### Table 8.2. Cluster Metrics Table

<table>
<thead>
<tr>
<th>Collection Name</th>
<th>Metric Name</th>
<th>Bit Field Value</th>
<th>Short Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>cluster</td>
<td>num_mon</td>
<td>2</td>
<td>Number of monitors</td>
</tr>
<tr>
<td></td>
<td>num_mon_quorum</td>
<td>2</td>
<td>Number of monitors in quorum</td>
</tr>
<tr>
<td></td>
<td>num_osd</td>
<td>2</td>
<td>Total number of OSD</td>
</tr>
<tr>
<td></td>
<td>num_osd_up</td>
<td>2</td>
<td>Number of OSDs that are up</td>
</tr>
<tr>
<td></td>
<td>num_osd_in</td>
<td>2</td>
<td>Number of OSDs that are in cluster</td>
</tr>
<tr>
<td></td>
<td>osd_epoch</td>
<td>2</td>
<td>Current epoch of OSD map</td>
</tr>
<tr>
<td></td>
<td>osd_bytes</td>
<td>2</td>
<td>Total capacity of cluster in bytes</td>
</tr>
<tr>
<td></td>
<td>osd_bytes_used</td>
<td>2</td>
<td>Number of used bytes on cluster</td>
</tr>
<tr>
<td></td>
<td>osd_bytes_avail</td>
<td>2</td>
<td>Number of available bytes on cluster</td>
</tr>
<tr>
<td></td>
<td>num_pool</td>
<td>2</td>
<td>Number of pools</td>
</tr>
<tr>
<td></td>
<td>num_pg</td>
<td>2</td>
<td>Total number of placement groups</td>
</tr>
<tr>
<td></td>
<td>num_pg_active_clean</td>
<td>2</td>
<td>Number of placement groups in active+clean state</td>
</tr>
<tr>
<td></td>
<td>num_pg_active</td>
<td>2</td>
<td>Number of placement groups in active state</td>
</tr>
<tr>
<td></td>
<td>num_pg_peering</td>
<td>2</td>
<td>Number of placement groups in peering state</td>
</tr>
<tr>
<td></td>
<td>num_object</td>
<td>2</td>
<td>Total number of objects on cluster</td>
</tr>
<tr>
<td></td>
<td>num_object_degraded</td>
<td>2</td>
<td>Number of degraded (missing replicas) objects</td>
</tr>
</tbody>
</table>
### Table 8.3. Level Database Metrics Table

<table>
<thead>
<tr>
<th>Collection Name</th>
<th>Metric Name</th>
<th>Bit Field Value</th>
<th>Short Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>leveldb</td>
<td>leveldb_get</td>
<td>10</td>
<td>Gets</td>
</tr>
<tr>
<td></td>
<td>leveldb_transaction</td>
<td>10</td>
<td>Transactions</td>
</tr>
<tr>
<td></td>
<td>leveldb_compact</td>
<td>10</td>
<td>Compactions</td>
</tr>
<tr>
<td></td>
<td>leveldb_compact_range</td>
<td>10</td>
<td>Compactions by range</td>
</tr>
<tr>
<td></td>
<td>leveldb_compact_queue_merge</td>
<td>10</td>
<td>Mergings of ranges in compaction queue</td>
</tr>
<tr>
<td></td>
<td>leveldb_compact_queue_len</td>
<td>2</td>
<td>Length of compaction queue</td>
</tr>
</tbody>
</table>

### Table 8.4. General Monitor Metrics Table

<table>
<thead>
<tr>
<th>Collection Name</th>
<th>Metric Name</th>
<th>Bit Field Value</th>
<th>Short Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>mon</td>
<td>num_sessions</td>
<td>2</td>
<td>Current number of opened monitor sessions</td>
</tr>
</tbody>
</table>
### Table 8.5. Paxos Metrics Table

<table>
<thead>
<tr>
<th>Collection Name</th>
<th>Metric Name</th>
<th>Bit Field Value</th>
<th>Short Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>paxos</td>
<td>start_leader</td>
<td>10</td>
<td>Starts in leader role</td>
</tr>
<tr>
<td></td>
<td>start_peon</td>
<td>10</td>
<td>Starts in peon role</td>
</tr>
<tr>
<td></td>
<td>restart</td>
<td>10</td>
<td>Restarts</td>
</tr>
<tr>
<td></td>
<td>refresh</td>
<td>10</td>
<td>Refreshes</td>
</tr>
<tr>
<td></td>
<td>refresh_latency</td>
<td>5</td>
<td>Refresh latency</td>
</tr>
<tr>
<td></td>
<td>begin</td>
<td>10</td>
<td>Started and handled begins</td>
</tr>
<tr>
<td></td>
<td>begin_keys</td>
<td>6</td>
<td>Keys in transaction on begin</td>
</tr>
<tr>
<td></td>
<td>begin_bytes</td>
<td>6</td>
<td>Data in transaction on begin</td>
</tr>
<tr>
<td></td>
<td>begin_latency</td>
<td>5</td>
<td>Latency of begin operation</td>
</tr>
<tr>
<td></td>
<td>commit</td>
<td>10</td>
<td>Commits</td>
</tr>
<tr>
<td>Collection Name</td>
<td>Metric Name</td>
<td>Bit Field Value</td>
<td>Short Description</td>
</tr>
<tr>
<td>----------------------</td>
<td>----------------------</td>
<td>-----------------</td>
<td>--------------------------------------------------------</td>
</tr>
<tr>
<td>commit_keys</td>
<td></td>
<td>6</td>
<td>Keys in transaction on commit</td>
</tr>
<tr>
<td>commit_bytes</td>
<td></td>
<td>6</td>
<td>Data in transaction on commit</td>
</tr>
<tr>
<td>commit_latency</td>
<td></td>
<td>5</td>
<td>Commit latency</td>
</tr>
<tr>
<td>collect</td>
<td></td>
<td>10</td>
<td>Peon collects</td>
</tr>
<tr>
<td>collect_keys</td>
<td></td>
<td>6</td>
<td>Keys in transaction on peon collect</td>
</tr>
<tr>
<td>collect_bytes</td>
<td></td>
<td>6</td>
<td>Data in transaction on peon collect</td>
</tr>
<tr>
<td>collect_latency</td>
<td></td>
<td>5</td>
<td>Peon collect latency</td>
</tr>
<tr>
<td>collect_uncommitted</td>
<td></td>
<td>10</td>
<td>Uncommitted values in started and handled collects</td>
</tr>
<tr>
<td>collect_timeout</td>
<td></td>
<td>10</td>
<td>Collect timeouts</td>
</tr>
<tr>
<td>accept_timeout</td>
<td></td>
<td>10</td>
<td>Accept timeouts</td>
</tr>
<tr>
<td>lease_ack_timeout</td>
<td></td>
<td>10</td>
<td>Lease acknowledgement timeouts</td>
</tr>
<tr>
<td>lease_timeout</td>
<td></td>
<td>10</td>
<td>Lease timeouts</td>
</tr>
<tr>
<td>store_state</td>
<td></td>
<td>10</td>
<td>Store a shared state on disk</td>
</tr>
<tr>
<td>store_state_keys</td>
<td></td>
<td>6</td>
<td>Keys in transaction in stored state</td>
</tr>
<tr>
<td>store_state_bytes</td>
<td></td>
<td>6</td>
<td>Data in transaction in stored state</td>
</tr>
<tr>
<td>store_state_latency</td>
<td></td>
<td>5</td>
<td>Storing state latency</td>
</tr>
<tr>
<td>share_state</td>
<td></td>
<td>10</td>
<td>Sharings of state</td>
</tr>
<tr>
<td>share_state_keys</td>
<td></td>
<td>6</td>
<td>Keys in shared state</td>
</tr>
<tr>
<td>share_state_bytes</td>
<td></td>
<td>6</td>
<td>Data in shared state</td>
</tr>
</tbody>
</table>
### Collection Name | Metric Name | Bit Field Value | Short Description
---|---|---|---
new_pn | 10 | New proposal number queries
new_pn_latency | 5 | New proposal number getting latency

#### Table 8.6. Throttle Metrics Table

<table>
<thead>
<tr>
<th>Collection Name</th>
<th>Metric Name</th>
<th>Bit Field Value</th>
<th>Short Description</th>
</tr>
</thead>
</table>
| throttle-* | val | 10 | Currently available throttle
| | max | 10 | Max value for throttle
| | get | 10 | Gets
| | get_sum | 10 | Got data
| | get_or_fail_fail | 10 | Get blocked during get_or_fail
| | get_or_fail_success | 10 | Successful get during get_or_fail
| | take | 10 | Takes
| | take_sum | 10 | Taken data
| | put | 10 | Puts
| | put_sum | 10 | Put data
| | wait | 5 | Waiting latency

### 8.3.3. OSD Metrics Description Tables

- Write Back Throttle Metrics Table
- Filestore Metrics Table
- Level Database Metrics Table
- Objecter Metrics Table
- Read and Write Operations Metrics Table
- Recovery State Metrics Table
### OSD Throttle Metrics Table

Table 8.7. Write Back Throttle Metrics Table

<table>
<thead>
<tr>
<th>Collection Name</th>
<th>Metric Name</th>
<th>Bit Field Value</th>
<th>Short Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>WBThrottle</td>
<td>bytes_dirtied</td>
<td>2</td>
<td>Dirty data</td>
</tr>
<tr>
<td></td>
<td>bytes_wb</td>
<td>2</td>
<td>Written data</td>
</tr>
<tr>
<td></td>
<td>ios_dirtied</td>
<td>2</td>
<td>Dirty operations</td>
</tr>
<tr>
<td></td>
<td>ios_wb</td>
<td>2</td>
<td>Written operations</td>
</tr>
<tr>
<td></td>
<td>inodes_dirtied</td>
<td>2</td>
<td>Entries waiting for write</td>
</tr>
<tr>
<td></td>
<td>inodes_wb</td>
<td>2</td>
<td>Written entries</td>
</tr>
</tbody>
</table>

### Filestore Metrics Table

Table 8.8. Filestore Metrics Table

<table>
<thead>
<tr>
<th>Collection Name</th>
<th>Metric Name</th>
<th>Bit Field Value</th>
<th>Short Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>filestore</td>
<td>journal_queue_max_ops</td>
<td>2</td>
<td>Max operations in journal queue</td>
</tr>
<tr>
<td></td>
<td>journal_queue_ops</td>
<td>2</td>
<td>Operations in journal queue</td>
</tr>
<tr>
<td></td>
<td>journal_ops</td>
<td>10</td>
<td>Total journal entries written</td>
</tr>
<tr>
<td></td>
<td>journal_queue_max_bytes</td>
<td>2</td>
<td>Max data in journal queue</td>
</tr>
<tr>
<td></td>
<td>journal_queue_bytes</td>
<td>2</td>
<td>Size of journal queue</td>
</tr>
<tr>
<td></td>
<td>journal_bytes</td>
<td>10</td>
<td>Total operations size in journal</td>
</tr>
<tr>
<td></td>
<td>journal_latency</td>
<td>5</td>
<td>Average journal queue completing latency</td>
</tr>
<tr>
<td></td>
<td>journal_wr</td>
<td>10</td>
<td>Journal write IOs</td>
</tr>
<tr>
<td></td>
<td>journal_wr_bytes</td>
<td>6</td>
<td>Journal data written</td>
</tr>
<tr>
<td></td>
<td>journal_full</td>
<td>10</td>
<td>Journal writes while full</td>
</tr>
<tr>
<td></td>
<td>committing</td>
<td>2</td>
<td>Is currently committing</td>
</tr>
</tbody>
</table>
### Collection Name | Metric Name | Bit Field Value | Short Description
--- | --- | --- | ---
commitcycle | | 10 | Commit cycles
commitcycle_interval | | 5 | Average interval between commits
commitcycle_latency | | 5 | Average latency of commit
op_queue_max_ops | | 2 | Max operations count in queue
op_queue_ops | | 2 | Operations count in queue
ops | | 10 | Operations
op_queue_max_bytes | | 2 | Max size of queue
op_queue_bytes | | 2 | Size of queue
bytes | | 10 | Data written to store
apply_latency | | 5 | Apply latency
queue_transaction_latency_avg | | 5 | Store operation queue latency

### Collection Name | Metric Name | Bit Field Value | Short Description
--- | --- | --- | ---
leveldb | leveldb_get | 10 | Gets
leveldb_transaction | | 10 | Transactions
leveldb_compact | | 10 | Compactions
leveldb_compact_range | | 10 | Compactions by range
leveldb_compact_queue_merge | | 10 | Mergings of ranges in compaction queue
leveldb_compact_queue_len | | 2 | Length of compaction queue
Table 8.10. Objecter Metrics Table

<table>
<thead>
<tr>
<th>Collection Name</th>
<th>Metric Name</th>
<th>Bit Field Value</th>
<th>Short Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>objecter</td>
<td>op_active</td>
<td>2</td>
<td>Active operations</td>
</tr>
<tr>
<td></td>
<td>op_laggy</td>
<td>2</td>
<td>Laggy operations</td>
</tr>
<tr>
<td></td>
<td>op_send</td>
<td>10</td>
<td>Sent operations</td>
</tr>
<tr>
<td></td>
<td>op_send_bytes</td>
<td>10</td>
<td>Sent data</td>
</tr>
<tr>
<td></td>
<td>op_resend</td>
<td>10</td>
<td>Resent operations</td>
</tr>
<tr>
<td></td>
<td>op_ack</td>
<td>10</td>
<td>Commit callbacks</td>
</tr>
<tr>
<td></td>
<td>op_commit</td>
<td>10</td>
<td>Operation commits</td>
</tr>
<tr>
<td></td>
<td>op</td>
<td>10</td>
<td>Operation</td>
</tr>
<tr>
<td></td>
<td>op_r</td>
<td>10</td>
<td>Read operations</td>
</tr>
<tr>
<td></td>
<td>op_w</td>
<td>10</td>
<td>Write operations</td>
</tr>
<tr>
<td></td>
<td>op_rmw</td>
<td>10</td>
<td>Read-modify-write operations</td>
</tr>
<tr>
<td></td>
<td>op_pg</td>
<td>10</td>
<td>PG operation</td>
</tr>
<tr>
<td></td>
<td>osdop_stat</td>
<td>10</td>
<td>Stat operations</td>
</tr>
<tr>
<td></td>
<td>osdop_create</td>
<td>10</td>
<td>Create object operations</td>
</tr>
<tr>
<td></td>
<td>osdop_read</td>
<td>10</td>
<td>Read operations</td>
</tr>
<tr>
<td></td>
<td>osdop_write</td>
<td>10</td>
<td>Write operations</td>
</tr>
<tr>
<td></td>
<td>osdop_writefull</td>
<td>10</td>
<td>Write full object operations</td>
</tr>
<tr>
<td></td>
<td>osdop_append</td>
<td>10</td>
<td>Append operation</td>
</tr>
<tr>
<td></td>
<td>osdop_zero</td>
<td>10</td>
<td>Set object to zero operations</td>
</tr>
<tr>
<td></td>
<td>osdop_truncate</td>
<td>10</td>
<td>Truncate object operations</td>
</tr>
<tr>
<td></td>
<td>osdop_delete</td>
<td>10</td>
<td>Delete object operations</td>
</tr>
<tr>
<td></td>
<td>osdop_mapext</td>
<td>10</td>
<td>Map extent operations</td>
</tr>
<tr>
<td>Collection Name</td>
<td>Metric Name</td>
<td>Bit Field Value</td>
<td>Short Description</td>
</tr>
<tr>
<td>--------------------</td>
<td>------------------------</td>
<td>-----------------</td>
<td>--------------------------------------------------------</td>
</tr>
<tr>
<td>osdop_sparse_read</td>
<td></td>
<td>10</td>
<td>Sparse read operations</td>
</tr>
<tr>
<td>osdop_clonerange</td>
<td></td>
<td>10</td>
<td>Clone range operations</td>
</tr>
<tr>
<td>osdop_getxattr</td>
<td></td>
<td>10</td>
<td>Get xattr operations</td>
</tr>
<tr>
<td>osdop_setxattr</td>
<td></td>
<td>10</td>
<td>Set xattr operations</td>
</tr>
<tr>
<td>osdop_cmpxattr</td>
<td></td>
<td>10</td>
<td>Xattr comparison operations</td>
</tr>
<tr>
<td>osdop_rmxattr</td>
<td></td>
<td>10</td>
<td>Remove xattr operations</td>
</tr>
<tr>
<td>osdop_resetxattr</td>
<td></td>
<td>10</td>
<td>Reset xattr operations</td>
</tr>
<tr>
<td>osdop_tmap_up</td>
<td></td>
<td>10</td>
<td>TMAP update operations</td>
</tr>
<tr>
<td>osdop_tmap_put</td>
<td></td>
<td>10</td>
<td>TMAP put operations</td>
</tr>
<tr>
<td>osdop_tmap_get</td>
<td></td>
<td>10</td>
<td>TMAP get operations</td>
</tr>
<tr>
<td>osdop_call</td>
<td></td>
<td>10</td>
<td>Call (execute) operations</td>
</tr>
<tr>
<td>osdop_watch</td>
<td></td>
<td>10</td>
<td>Watch by object operations</td>
</tr>
<tr>
<td>osdop_notify</td>
<td></td>
<td>10</td>
<td>Notify about object operations</td>
</tr>
<tr>
<td>osdop_src_cmpxattr</td>
<td></td>
<td>10</td>
<td>Extended attribute comparison in multi operations</td>
</tr>
<tr>
<td>osdop_other</td>
<td></td>
<td>10</td>
<td>Other operations</td>
</tr>
<tr>
<td>linger_active</td>
<td></td>
<td>2</td>
<td>Active lingering operations</td>
</tr>
<tr>
<td>linger_send</td>
<td></td>
<td>10</td>
<td>Sent lingering operations</td>
</tr>
<tr>
<td>linger_resend</td>
<td></td>
<td>10</td>
<td>Resent lingering operations</td>
</tr>
<tr>
<td>linger_ping</td>
<td></td>
<td>10</td>
<td>Sent pings to lingering operations</td>
</tr>
<tr>
<td>poolop_active</td>
<td></td>
<td>2</td>
<td>Active pool operations</td>
</tr>
<tr>
<td>poolop_send</td>
<td></td>
<td>10</td>
<td>Sent pool operations</td>
</tr>
<tr>
<td>Collection Name</td>
<td>Metric Name</td>
<td>Bit Field Value</td>
<td>Short Description</td>
</tr>
<tr>
<td>------------------</td>
<td>----------------</td>
<td>-----------------</td>
<td>--------------------------------------------------------</td>
</tr>
<tr>
<td>poolop_resend</td>
<td></td>
<td>10</td>
<td>Resent pool operations</td>
</tr>
<tr>
<td>poolstat_active</td>
<td></td>
<td>2</td>
<td>Active get pool stat operations</td>
</tr>
<tr>
<td>poolstat_send</td>
<td></td>
<td>10</td>
<td>Pool stat operations sent</td>
</tr>
<tr>
<td>poolstat_resend</td>
<td></td>
<td>10</td>
<td>Resent pool stats</td>
</tr>
<tr>
<td>statfs_active</td>
<td></td>
<td>2</td>
<td>Statfs operations</td>
</tr>
<tr>
<td>statfs_send</td>
<td></td>
<td>10</td>
<td>Sent FS stats</td>
</tr>
<tr>
<td>statfs_resend</td>
<td></td>
<td>10</td>
<td>Resent FS stats</td>
</tr>
<tr>
<td>command_active</td>
<td></td>
<td>2</td>
<td>Active commands</td>
</tr>
<tr>
<td>command_send</td>
<td></td>
<td>10</td>
<td>Sent commands</td>
</tr>
<tr>
<td>command_resend</td>
<td></td>
<td>10</td>
<td>Resent commands</td>
</tr>
<tr>
<td>map_epoch</td>
<td></td>
<td>2</td>
<td>OSD map epoch</td>
</tr>
<tr>
<td>map_full</td>
<td></td>
<td>10</td>
<td>Full OSD maps received</td>
</tr>
<tr>
<td>map_inc</td>
<td></td>
<td>10</td>
<td>Incremental OSD maps received</td>
</tr>
<tr>
<td>osd_sessions</td>
<td></td>
<td>2</td>
<td>Open sessions</td>
</tr>
<tr>
<td>osd_session_open</td>
<td></td>
<td>10</td>
<td>Sessions opened</td>
</tr>
<tr>
<td>osd_session_close</td>
<td></td>
<td>10</td>
<td>Sessions closed</td>
</tr>
<tr>
<td>osd_laggy</td>
<td></td>
<td>2</td>
<td>Laggy OSD sessions</td>
</tr>
</tbody>
</table>

Table 8.11. Read and Write Operations Metrics Table

<table>
<thead>
<tr>
<th>Collection Name</th>
<th>Metric Name</th>
<th>Bit Field Value</th>
<th>Short Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>osd</td>
<td>op_wip</td>
<td>2</td>
<td>Replication operations currently being processed (primary)</td>
</tr>
<tr>
<td></td>
<td>op_in_bytes</td>
<td>10</td>
<td>Client operations total write size</td>
</tr>
<tr>
<td>Collection Name</td>
<td>Metric Name</td>
<td>Bit Field Value</td>
<td>Short Description</td>
</tr>
<tr>
<td>-----------------</td>
<td>-------------</td>
<td>-----------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>op_out_bytes</td>
<td></td>
<td>10</td>
<td>Client operations total read size</td>
</tr>
<tr>
<td>op_latency</td>
<td></td>
<td>5</td>
<td>Latency of client operations (including queue time)</td>
</tr>
<tr>
<td>op_process_latency</td>
<td></td>
<td>5</td>
<td>Latency of client operations (excluding queue time)</td>
</tr>
<tr>
<td>op_r</td>
<td></td>
<td>10</td>
<td>Client read operations</td>
</tr>
<tr>
<td>op_r_out_bytes</td>
<td></td>
<td>10</td>
<td>Client data read</td>
</tr>
<tr>
<td>op_r_latency</td>
<td></td>
<td>5</td>
<td>Latency of read operation (including queue time)</td>
</tr>
<tr>
<td>op_r_process_latency</td>
<td></td>
<td>5</td>
<td>Latency of read operation (excluding queue time)</td>
</tr>
<tr>
<td>op_w</td>
<td></td>
<td>10</td>
<td>Client write operations</td>
</tr>
<tr>
<td>op_w_in_bytes</td>
<td></td>
<td>10</td>
<td>Client data written</td>
</tr>
<tr>
<td>op_w_rlat</td>
<td></td>
<td>5</td>
<td>Client write operation readable/applied latency</td>
</tr>
<tr>
<td>op_w_latency</td>
<td></td>
<td>5</td>
<td>Latency of write operation (including queue time)</td>
</tr>
<tr>
<td>op_w_process_latency</td>
<td></td>
<td>5</td>
<td>Latency of write operation (excluding queue time)</td>
</tr>
<tr>
<td>op_rw</td>
<td></td>
<td>10</td>
<td>Client read-modify-write operations</td>
</tr>
<tr>
<td>op_rw_in_bytes</td>
<td></td>
<td>10</td>
<td>Client read-modify-write operations write in</td>
</tr>
<tr>
<td>op_rw_out_bytes</td>
<td></td>
<td>10</td>
<td>Client read-modify-write operations read out</td>
</tr>
<tr>
<td>op_rw_rlat</td>
<td></td>
<td>5</td>
<td>Client read-modify-write operation readable/applied latency</td>
</tr>
<tr>
<td>Collection Name</td>
<td>Metric Name</td>
<td>Bit Field Value</td>
<td>Short Description</td>
</tr>
<tr>
<td>------------------------</td>
<td>-------------------</td>
<td>-----------------</td>
<td>-----------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>op_rw_latency</td>
<td></td>
<td>5</td>
<td>Latency of read-modify-write operation (including queue time)</td>
</tr>
<tr>
<td>op_rw_process_latency</td>
<td></td>
<td>5</td>
<td>Latency of read-modify-write operation (excluding queue time)</td>
</tr>
<tr>
<td>subop</td>
<td></td>
<td>10</td>
<td>Suboperations</td>
</tr>
<tr>
<td>subop_in_bytes</td>
<td></td>
<td>10</td>
<td>Suboperations total size</td>
</tr>
<tr>
<td>subop_latency</td>
<td></td>
<td>5</td>
<td>Suboperations latency</td>
</tr>
<tr>
<td>subop_w</td>
<td></td>
<td>10</td>
<td>Replicated writes</td>
</tr>
<tr>
<td>subop_w_in_bytes</td>
<td></td>
<td>10</td>
<td>Replicated written data size</td>
</tr>
<tr>
<td>subop_w_latency</td>
<td></td>
<td>5</td>
<td>Replicated writes latency</td>
</tr>
<tr>
<td>subop_pull</td>
<td></td>
<td>10</td>
<td>Suboperations pull requests</td>
</tr>
<tr>
<td>subop_pull_latency</td>
<td></td>
<td>5</td>
<td>Suboperations pull latency</td>
</tr>
<tr>
<td>subop_push</td>
<td></td>
<td>10</td>
<td>Suboperations push messages</td>
</tr>
<tr>
<td>subop_push_in_bytes</td>
<td></td>
<td>10</td>
<td>Suboperations pushed size</td>
</tr>
<tr>
<td>subop_push_latency</td>
<td></td>
<td>5</td>
<td>Suboperations push latency</td>
</tr>
<tr>
<td>pull</td>
<td></td>
<td>10</td>
<td>Pull requests sent</td>
</tr>
<tr>
<td>push</td>
<td></td>
<td>10</td>
<td>Push messages sent</td>
</tr>
<tr>
<td>push_out_bytes</td>
<td></td>
<td>10</td>
<td>Pushed size</td>
</tr>
<tr>
<td>push_in</td>
<td></td>
<td>10</td>
<td>Inbound push messages</td>
</tr>
<tr>
<td>push_in_bytes</td>
<td></td>
<td>10</td>
<td>Inbound pushed size</td>
</tr>
<tr>
<td>recovery_ops</td>
<td></td>
<td>10</td>
<td>Started recovery operations</td>
</tr>
<tr>
<td>Collection Name</td>
<td>Metric Name</td>
<td>Bit Field Value</td>
<td>Short Description</td>
</tr>
<tr>
<td>-----------------</td>
<td>----------------------</td>
<td>-----------------</td>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td></td>
<td>loadavg</td>
<td>2</td>
<td>CPU load</td>
</tr>
<tr>
<td></td>
<td>buffer_bytes</td>
<td>2</td>
<td>Total allocated buffer size</td>
</tr>
<tr>
<td></td>
<td>numpg</td>
<td>2</td>
<td>Placement groups</td>
</tr>
<tr>
<td></td>
<td>numpg_primary</td>
<td>2</td>
<td>Placement groups for which this osd is primary</td>
</tr>
<tr>
<td></td>
<td>numpg_replica</td>
<td>2</td>
<td>Placement groups for which this osd is replica</td>
</tr>
<tr>
<td></td>
<td>numpg_stray</td>
<td>2</td>
<td>Placement groups ready to be deleted from this osd</td>
</tr>
<tr>
<td></td>
<td>heartbeat_to_peers</td>
<td>2</td>
<td>Heartbeat (ping) peers we send to</td>
</tr>
<tr>
<td></td>
<td>heartbeat_from_peers</td>
<td>2</td>
<td>Heartbeat (ping) peers we recv from</td>
</tr>
<tr>
<td></td>
<td>map_messages</td>
<td>10</td>
<td>OSD map messages</td>
</tr>
<tr>
<td></td>
<td>map_message_epochs</td>
<td>10</td>
<td>OSD map epochs</td>
</tr>
<tr>
<td></td>
<td>map_message_epochs_duplicates</td>
<td>10</td>
<td>OSD map duplicates</td>
</tr>
<tr>
<td></td>
<td>stat_bytes</td>
<td>2</td>
<td>OSD size</td>
</tr>
<tr>
<td></td>
<td>stat_bytes_used</td>
<td>2</td>
<td>Used space</td>
</tr>
<tr>
<td></td>
<td>stat_bytes_avail</td>
<td>2</td>
<td>Available space</td>
</tr>
<tr>
<td></td>
<td>copyfrom</td>
<td>10</td>
<td>Rados 'copy-from' operations</td>
</tr>
<tr>
<td></td>
<td>tier_promote</td>
<td>10</td>
<td>Tier promotions</td>
</tr>
<tr>
<td></td>
<td>tier_flush</td>
<td>10</td>
<td>Tier flushes</td>
</tr>
<tr>
<td></td>
<td>tier_flush_fail</td>
<td>10</td>
<td>Failed tier flushes</td>
</tr>
<tr>
<td></td>
<td>tier_try_flush</td>
<td>10</td>
<td>Tier flush attempts</td>
</tr>
<tr>
<td>Collection Name</td>
<td>Metric Name</td>
<td>Bit Field Value</td>
<td>Short Description</td>
</tr>
<tr>
<td>----------------------</td>
<td>--------------------</td>
<td>-----------------</td>
<td>------------------------------------------</td>
</tr>
<tr>
<td>tier_try_flush_fail</td>
<td>10</td>
<td>Failed tier flush attempts</td>
<td></td>
</tr>
<tr>
<td>tier_evict</td>
<td>10</td>
<td>Tier evictions</td>
<td></td>
</tr>
<tr>
<td>tier_whiteout</td>
<td>10</td>
<td>Tier whiteouts</td>
<td></td>
</tr>
<tr>
<td>tier_dirty</td>
<td>10</td>
<td>Dirty tier flag set</td>
<td></td>
</tr>
<tr>
<td>tier_clean</td>
<td>10</td>
<td>Dirty tier flag cleaned</td>
<td></td>
</tr>
<tr>
<td>tier_delay</td>
<td>10</td>
<td>Tier delays (agent waiting)</td>
<td></td>
</tr>
<tr>
<td>tier_proxy_read</td>
<td>10</td>
<td>Tier proxy reads</td>
<td></td>
</tr>
<tr>
<td>agent_wake</td>
<td>10</td>
<td>Tiering agent wake up</td>
<td></td>
</tr>
<tr>
<td>agent_skip</td>
<td>10</td>
<td>Objects skipped by agent</td>
<td></td>
</tr>
<tr>
<td>agent_flush</td>
<td>10</td>
<td>Tiering agent flushes</td>
<td></td>
</tr>
<tr>
<td>agent_evict</td>
<td>10</td>
<td>Tiering agent evictions</td>
<td></td>
</tr>
<tr>
<td>object_ctx_cache_hit</td>
<td>10</td>
<td>Object context cache hits</td>
<td></td>
</tr>
<tr>
<td>object_ctx_cache_total</td>
<td>10</td>
<td>Object context cache lookups</td>
<td></td>
</tr>
</tbody>
</table>

Table 8.12. Recovery State Metrics Table

<table>
<thead>
<tr>
<th>Collection Name</th>
<th>Metric Name</th>
<th>Bit Field Value</th>
<th>Short Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>recoverystate_perf</td>
<td>initial_latency</td>
<td>5</td>
<td>Initial recovery state latency</td>
</tr>
<tr>
<td></td>
<td>started_latency</td>
<td>5</td>
<td>Started recovery state latency</td>
</tr>
<tr>
<td></td>
<td>reset_latency</td>
<td>5</td>
<td>Reset recovery state latency</td>
</tr>
<tr>
<td></td>
<td>start_latency</td>
<td>5</td>
<td>Start recovery state latency</td>
</tr>
<tr>
<td></td>
<td>primary_latency</td>
<td>5</td>
<td>Primary recovery state latency</td>
</tr>
<tr>
<td></td>
<td>peering_latency</td>
<td>5</td>
<td>Peering recovery state latency</td>
</tr>
</tbody>
</table>

CHAPTER 8. PERFORMANCE COUNTERS
<table>
<thead>
<tr>
<th>Collection Name</th>
<th>Metric Name</th>
<th>Bit Field Value</th>
<th>Short Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>backfilling_latency</td>
<td>5</td>
<td>Backfilling recovery state latency</td>
<td></td>
</tr>
<tr>
<td>waitremotebackfillreserved_latency</td>
<td>5</td>
<td>Wait remote backfill reserved recovery state latency</td>
<td></td>
</tr>
<tr>
<td>waitlocalbackfillreserved_latency</td>
<td>5</td>
<td>Wait local backfill reserved recovery state latency</td>
<td></td>
</tr>
<tr>
<td>notbackfilling_latency</td>
<td>5</td>
<td>Notbackfilling recovery state latency</td>
<td></td>
</tr>
<tr>
<td>repnotrecovering_latency</td>
<td>5</td>
<td>Repnotrecovering recovery state latency</td>
<td></td>
</tr>
<tr>
<td>repwaitrecoveryreserved_latency</td>
<td>5</td>
<td>Rep wait recovery reserved recovery state latency</td>
<td></td>
</tr>
<tr>
<td>repwaitbackfillreserved_latency</td>
<td>5</td>
<td>Rep wait backfill reserved recovery state latency</td>
<td></td>
</tr>
<tr>
<td>RepRecovering_latency</td>
<td>5</td>
<td>RepRecovering recovery state latency</td>
<td></td>
</tr>
<tr>
<td>activating_latency</td>
<td>5</td>
<td>Activating recovery state latency</td>
<td></td>
</tr>
<tr>
<td>waitlocalrecoveryreserved_latency</td>
<td>5</td>
<td>Wait local recovery reserved recovery state latency</td>
<td></td>
</tr>
<tr>
<td>waitremoterecoveryreserved_latency</td>
<td>5</td>
<td>Wait remote recovery reserved recovery state latency</td>
<td></td>
</tr>
<tr>
<td>recovering_latency</td>
<td>5</td>
<td>Recovering recovery state latency</td>
<td></td>
</tr>
<tr>
<td>recovered_latency</td>
<td>5</td>
<td>Recovered recovery state latency</td>
<td></td>
</tr>
<tr>
<td>clean latency</td>
<td>5</td>
<td>Clean recovery state latency</td>
<td></td>
</tr>
<tr>
<td>Collection Name</td>
<td>Metric Name</td>
<td>Bit Field Value</td>
<td>Short Description</td>
</tr>
<tr>
<td>-------------------</td>
<td>------------------------</td>
<td>-----------------</td>
<td>------------------------------------------</td>
</tr>
<tr>
<td>active_latency</td>
<td></td>
<td>5</td>
<td>Active recovery state latency</td>
</tr>
<tr>
<td>replicaactive_latency</td>
<td></td>
<td>5</td>
<td>Replicaactive recovery state latency</td>
</tr>
<tr>
<td>stray_latency</td>
<td></td>
<td>5</td>
<td>Stray recovery state latency</td>
</tr>
<tr>
<td>getinfo_latency</td>
<td></td>
<td>5</td>
<td>Getinfo recovery state latency</td>
</tr>
<tr>
<td>getlog_latency</td>
<td></td>
<td>5</td>
<td>Getlog recovery state latency</td>
</tr>
<tr>
<td>waitactingchange_latency</td>
<td></td>
<td>5</td>
<td>Waitactingchange recovery state latency</td>
</tr>
<tr>
<td>incomplete_latency</td>
<td></td>
<td>5</td>
<td>Incomplete recovery state latency</td>
</tr>
<tr>
<td>getmissing_latency</td>
<td></td>
<td>5</td>
<td>Getmissing recovery state latency</td>
</tr>
<tr>
<td>waitupthru_latency</td>
<td></td>
<td>5</td>
<td>Waitupthru recovery state latency</td>
</tr>
</tbody>
</table>

Table 8.13. OSD Throttle Metrics Table

<table>
<thead>
<tr>
<th>Collection Name</th>
<th>Metric Name</th>
<th>Bit Field Value</th>
<th>Short Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>throttle-*</td>
<td>val</td>
<td>10</td>
<td>Currently available throttle</td>
</tr>
<tr>
<td></td>
<td>max</td>
<td>10</td>
<td>Max value for throttle</td>
</tr>
<tr>
<td></td>
<td>get</td>
<td>10</td>
<td>Gets</td>
</tr>
<tr>
<td></td>
<td>get_sum</td>
<td>10</td>
<td>Got data</td>
</tr>
<tr>
<td></td>
<td>get_or_fail_fail</td>
<td>10</td>
<td>Get blocked during get_or_fail</td>
</tr>
<tr>
<td></td>
<td>get_or_fail_success</td>
<td>10</td>
<td>Successful get during get_or_fail</td>
</tr>
<tr>
<td></td>
<td>take</td>
<td>10</td>
<td>Takes</td>
</tr>
<tr>
<td></td>
<td>take_sum</td>
<td>10</td>
<td>Taken data</td>
</tr>
<tr>
<td></td>
<td>put</td>
<td>10</td>
<td>Puts</td>
</tr>
<tr>
<td>Collection Name</td>
<td>Metric Name</td>
<td>Bit Field Value</td>
<td>Short Description</td>
</tr>
<tr>
<td>-----------------</td>
<td>-------------</td>
<td>-----------------</td>
<td>-------------------</td>
</tr>
<tr>
<td><strong>put_sum</strong></td>
<td></td>
<td>10</td>
<td>Put data</td>
</tr>
<tr>
<td><strong>wait</strong></td>
<td></td>
<td>5</td>
<td>Waiting latency</td>
</tr>
</tbody>
</table>

### 8.3.4. The Ceph Object Gateway Metrics Tables

- RADOS Gateway Client Table
- Objecter Metrics Table
- RADOS Gateway Throttle Metrics Table

#### Table 8.14. RADOS Client Metrics Table

<table>
<thead>
<tr>
<th>Collection Name</th>
<th>Metric Name</th>
<th>Bit Field Value</th>
<th>Short Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>client.rgw.&lt;rgw_node_name&gt;</td>
<td>req</td>
<td>10</td>
<td>Requests</td>
</tr>
<tr>
<td></td>
<td>failed_req</td>
<td>10</td>
<td>Aborted requests</td>
</tr>
<tr>
<td></td>
<td>get</td>
<td>10</td>
<td>Gets</td>
</tr>
<tr>
<td></td>
<td>get_b</td>
<td>10</td>
<td>Size of gets</td>
</tr>
<tr>
<td></td>
<td>get_initial_lat</td>
<td>5</td>
<td>Get latency</td>
</tr>
<tr>
<td></td>
<td>put</td>
<td>10</td>
<td>Puts</td>
</tr>
<tr>
<td></td>
<td>put_b</td>
<td>10</td>
<td>Size of puts</td>
</tr>
<tr>
<td></td>
<td>put_initial_lat</td>
<td>5</td>
<td>Put latency</td>
</tr>
<tr>
<td></td>
<td>qlen</td>
<td>2</td>
<td>Queue length</td>
</tr>
<tr>
<td></td>
<td>qactive</td>
<td>2</td>
<td>Active requests queue</td>
</tr>
<tr>
<td></td>
<td>cache_hit</td>
<td>10</td>
<td>Cache hits</td>
</tr>
<tr>
<td></td>
<td>cache_miss</td>
<td>10</td>
<td>Cache miss</td>
</tr>
<tr>
<td></td>
<td>keystone_token_cache_hit</td>
<td>10</td>
<td>Keystone token cache hits</td>
</tr>
</tbody>
</table>
Table 8.15. Objecter Metrics Table

<table>
<thead>
<tr>
<th>Collection Name</th>
<th>Metric Name</th>
<th>Bit Field Value</th>
<th>Short Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>objecter</td>
<td>op_active</td>
<td>2</td>
<td>Active operations</td>
</tr>
<tr>
<td></td>
<td>op_laggy</td>
<td>2</td>
<td>Laggy operations</td>
</tr>
<tr>
<td></td>
<td>op_send</td>
<td>10</td>
<td>Sent operations</td>
</tr>
<tr>
<td></td>
<td>op_send_bytes</td>
<td>10</td>
<td>Sent data</td>
</tr>
<tr>
<td></td>
<td>op_resend</td>
<td>10</td>
<td>Resent operations</td>
</tr>
<tr>
<td></td>
<td>op_ack</td>
<td>10</td>
<td>Commit callbacks</td>
</tr>
<tr>
<td></td>
<td>op_commit</td>
<td>10</td>
<td>Operation commits</td>
</tr>
<tr>
<td></td>
<td>op</td>
<td>10</td>
<td>Operation</td>
</tr>
<tr>
<td></td>
<td>op_r</td>
<td>10</td>
<td>Read operations</td>
</tr>
<tr>
<td></td>
<td>op_w</td>
<td>10</td>
<td>Write operations</td>
</tr>
<tr>
<td></td>
<td>op_rmw</td>
<td>10</td>
<td>Read-modify-write operations</td>
</tr>
<tr>
<td></td>
<td>op_pg</td>
<td>10</td>
<td>PG operation</td>
</tr>
<tr>
<td></td>
<td>osdop_stat</td>
<td>10</td>
<td>Stat operations</td>
</tr>
<tr>
<td></td>
<td>osdop_create</td>
<td>10</td>
<td>Create object operations</td>
</tr>
<tr>
<td></td>
<td>osdop_read</td>
<td>10</td>
<td>Read operations</td>
</tr>
<tr>
<td></td>
<td>osdop_write</td>
<td>10</td>
<td>Write operations</td>
</tr>
<tr>
<td></td>
<td>osdop_writefull</td>
<td>10</td>
<td>Write full object operations</td>
</tr>
<tr>
<td></td>
<td>osdop_append</td>
<td>10</td>
<td>Append operation</td>
</tr>
<tr>
<td></td>
<td>osdop_zero</td>
<td>10</td>
<td>Set object to zero operations</td>
</tr>
<tr>
<td>Collection Name</td>
<td>Metric Name</td>
<td>Bit Field Value</td>
<td>Short Description</td>
</tr>
<tr>
<td>-------------------------</td>
<td>---------------------</td>
<td>-----------------</td>
<td>-------------------------------------------------------</td>
</tr>
<tr>
<td>osdop_truncate</td>
<td>10</td>
<td></td>
<td>Truncate object operations</td>
</tr>
<tr>
<td>osdop_delete</td>
<td>10</td>
<td></td>
<td>Delete object operations</td>
</tr>
<tr>
<td>osdop_mapext</td>
<td>10</td>
<td></td>
<td>Map extent operations</td>
</tr>
<tr>
<td>osdop_sparse_read</td>
<td>10</td>
<td></td>
<td>Sparse read operations</td>
</tr>
<tr>
<td>osdop_clonerange</td>
<td>10</td>
<td></td>
<td>Clone range operations</td>
</tr>
<tr>
<td>osdop_getxattr</td>
<td>10</td>
<td></td>
<td>Get xattr operations</td>
</tr>
<tr>
<td>osdop_setxattr</td>
<td>10</td>
<td></td>
<td>Set xattr operations</td>
</tr>
<tr>
<td>osdop_cmpxattr</td>
<td>10</td>
<td></td>
<td>Xattr comparison operations</td>
</tr>
<tr>
<td>osdop_rmxattr</td>
<td>10</td>
<td></td>
<td>Remove xattr operations</td>
</tr>
<tr>
<td>osdop_resetxattr</td>
<td>10</td>
<td></td>
<td>Reset xattr operations</td>
</tr>
<tr>
<td>osdop_tmap_up</td>
<td>10</td>
<td></td>
<td>TMAP update operations</td>
</tr>
<tr>
<td>osdop_tmap_put</td>
<td>10</td>
<td></td>
<td>TMAP put operations</td>
</tr>
<tr>
<td>osdop_tmap_get</td>
<td>10</td>
<td></td>
<td>TMAP get operations</td>
</tr>
<tr>
<td>osdop_call</td>
<td>10</td>
<td></td>
<td>Call (execute) operations</td>
</tr>
<tr>
<td>osdop_watch</td>
<td>10</td>
<td></td>
<td>Watch by object operations</td>
</tr>
<tr>
<td>osdop_notify</td>
<td>10</td>
<td></td>
<td>Notify about object operations</td>
</tr>
<tr>
<td>osdop_src_cmpxattr</td>
<td>10</td>
<td></td>
<td>Extended attribute comparison in multi operations</td>
</tr>
<tr>
<td>osdop_other</td>
<td>10</td>
<td></td>
<td>Other operations</td>
</tr>
<tr>
<td>linger_active</td>
<td>2</td>
<td></td>
<td>Active lingering operations</td>
</tr>
<tr>
<td>linger_send</td>
<td>10</td>
<td></td>
<td>Sent lingering operations</td>
</tr>
<tr>
<td>linger_resend</td>
<td>10</td>
<td></td>
<td>Resent lingering operations</td>
</tr>
<tr>
<td>Collection Name</td>
<td>Metric Name</td>
<td>Bit Field Value</td>
<td>Short Description</td>
</tr>
<tr>
<td>-------------------</td>
<td>-------------------</td>
<td>-----------------</td>
<td>--------------------------------------------------------</td>
</tr>
<tr>
<td>linger_ping</td>
<td></td>
<td>10</td>
<td>Sent pings to lingering operations</td>
</tr>
<tr>
<td>poolop_active</td>
<td></td>
<td>2</td>
<td>Active pool operations</td>
</tr>
<tr>
<td>poolop_send</td>
<td></td>
<td>10</td>
<td>Sent pool operations</td>
</tr>
<tr>
<td>poolop_resend</td>
<td></td>
<td>10</td>
<td>Resent pool operations</td>
</tr>
<tr>
<td>poolstat_active</td>
<td></td>
<td>2</td>
<td>Active get pool stat operations</td>
</tr>
<tr>
<td>poolstat_send</td>
<td></td>
<td>10</td>
<td>Pool stat operations sent</td>
</tr>
<tr>
<td>poolstat_resend</td>
<td></td>
<td>10</td>
<td>Resent pool stats</td>
</tr>
<tr>
<td>statfs_active</td>
<td></td>
<td>2</td>
<td>Statfs operations</td>
</tr>
<tr>
<td>statfs_send</td>
<td></td>
<td>10</td>
<td>Sent FS stats</td>
</tr>
<tr>
<td>statfs_resend</td>
<td></td>
<td>10</td>
<td>Resent FS stats</td>
</tr>
<tr>
<td>command_active</td>
<td></td>
<td>2</td>
<td>Active commands</td>
</tr>
<tr>
<td>command_send</td>
<td></td>
<td>10</td>
<td>Sent commands</td>
</tr>
<tr>
<td>command_resend</td>
<td></td>
<td>10</td>
<td>Resent commands</td>
</tr>
<tr>
<td>map_epoch</td>
<td></td>
<td>2</td>
<td>OSD map epoch</td>
</tr>
<tr>
<td>map_full</td>
<td></td>
<td>10</td>
<td>Full OSD maps received</td>
</tr>
<tr>
<td>map_inc</td>
<td></td>
<td>10</td>
<td>Incremental OSD maps received</td>
</tr>
<tr>
<td>osd_sessions</td>
<td></td>
<td>2</td>
<td>Open sessions</td>
</tr>
<tr>
<td>osd_session_open</td>
<td></td>
<td>10</td>
<td>Sessions opened</td>
</tr>
<tr>
<td>osd_session_close</td>
<td></td>
<td>10</td>
<td>Sessions closed</td>
</tr>
<tr>
<td>osd_laggy</td>
<td></td>
<td>2</td>
<td>Laggy OSD sessions</td>
</tr>
</tbody>
</table>

Table 8.16. RADOS Gateway Throttle Metrics Table
<table>
<thead>
<tr>
<th>Collection Name</th>
<th>Metric Name</th>
<th>Bit Field Value</th>
<th>Short Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>throttle-*</td>
<td>val</td>
<td>10</td>
<td>Currently available throttle</td>
</tr>
<tr>
<td></td>
<td>max</td>
<td>10</td>
<td>Max value for throttle</td>
</tr>
<tr>
<td></td>
<td>get</td>
<td>10</td>
<td>Gets</td>
</tr>
<tr>
<td></td>
<td>get_sum</td>
<td>10</td>
<td>Got data</td>
</tr>
<tr>
<td></td>
<td>get_or_fail_fail</td>
<td>10</td>
<td>Get blocked during get_or_fail</td>
</tr>
<tr>
<td></td>
<td>get_or_fail_succeed</td>
<td>10</td>
<td>Successful get during get_or_fail</td>
</tr>
<tr>
<td></td>
<td>take</td>
<td>10</td>
<td>Takes</td>
</tr>
<tr>
<td></td>
<td>take_sum</td>
<td>10</td>
<td>Taken data</td>
</tr>
<tr>
<td></td>
<td>put</td>
<td>10</td>
<td>Puts</td>
</tr>
<tr>
<td></td>
<td>put_sum</td>
<td>10</td>
<td>Put data</td>
</tr>
<tr>
<td></td>
<td>wait</td>
<td>5</td>
<td>Waiting latency</td>
</tr>
</tbody>
</table>
CHAPTER 9. BLUESTORE

BlueStore is a new back end object store for the OSD daemons. The original object store, FileStore, requires a file system on top of raw block devices. Objects are then written to the file system. BlueStore does not require a file system. It stores objects directly on the block device.

IMPORTANT

BlueStore provides a high-performance backend for OSD daemons in a production environment. By default, BlueStore is configured to be self-tuning. If you determine that your environment performs better with BlueStore tuned manually, please contact Red Hat support and share the details of your configuration to help us improve the auto-tuning capability. Red Hat looks forward to your feedback and appreciates your recommendations.

9.1. ABOUT BLUESTORE

BlueStore is a new back end for the OSD daemons. Unlike the original FileStore back end, BlueStore stores objects directly on the block devices without any file system interface, which improves the performance of the cluster.

The following are some of the main features of using BlueStore:

Direct management of storage devices
BlueStore consumes raw block devices or partitions. This avoids any intervening layers of abstraction, such as local file systems like XFS, that might limit performance or add complexity.

Metadata management with RocksDB
BlueStore uses the RocksDB key-value database to manage internal metadata, such as the mapping from object names to block locations on a disk.

Full data and metadata checksumming
By default all data and metadata written to BlueStore is protected by one or more checksums. No data or metadata are read from disk or returned to the user without verification.

Inline compression
Data written can be optionally compressed before being written to disk.

Efficient copy-on-write
The Ceph Block Device and Ceph File System snapshots rely on a copy-on-write clone mechanism that is implemented efficiently in BlueStore. This results in efficient I/O both for regular snapshots and for erasure coded pools which rely on cloning to implement efficient two-phase commits.

No large double-writes
BlueStore first writes any new data to unallocated space on a block device, and then commits a RocksDB transaction that updates the object metadata to reference the new region of the disk. Only when the write operation is below a configurable size threshold, it falls back to a write-ahead journaling scheme, similar to what FileStore operates.

Multi-device support
BlueStore can use multiple block devices for storing different data, for example: Hard Disk Drive (HDD) for the data, Solid-state Drive (SSD) for metadata, Non-volatile Memory (NVM) or Non-volatile random-access memory (NVRAM) or persistent memory for the RocksDB write-ahead log (WAL). See Section 9.2, “BlueStore Devices” for details.
The `ceph-disk` utility does not yet provision multiple devices. To use multiple devices, OSDs must be set up manually.

Efficient block device usage
Because BlueStore does not use any file system, it minimizes the need to clear the storage device cache.

9.2. BLUESTORE DEVICES
This section explains what block devices the BlueStore back end uses.

BlueStore manages either one, two, or (in certain cases) three storage devices.

- primary
- WAL
- DB

In the simplest case, BlueStore consumes a single (primary) storage device. The storage device is partitioned into two parts that contain:

- **OSD metadata**: A small partition formatted with XFS that contains basic metadata for the OSD. This data directory includes information about the OSD, such as its identifier, which cluster it belongs to, and its private keyring.

- **Data**: A large partition occupying the rest of the device that is managed directly by BlueStore and that contains all of the OSD data. This primary device is identified by a block symbolic link in the data directory.

You can also use two additional devices:

- **A WAL (write-ahead-log) device**: A device that stores BlueStore internal journal or write-ahead log. It is identified by the `block.wal` symbolic link in the data directory. Consider using a WAL device only if the device is faster than the primary device, for example, when the WAL device uses an SSD disk and the primary devices uses an HDD disk.

- **A DB device**: A device that stores BlueStore internal metadata. The embedded RocksDB database puts as much metadata as it can on the DB device instead on the primary device to improve performance. If the DB device is full, it starts adding metadata to the primary device. Consider using a DB device only if the device is faster than the primary device.

If you have only a less than a gigabyte storage available on fast devices, Red Hat recommends using it as a WAL device. If you have more fast devices available, consider using it as a DB device. The BlueStore journal is always places on the fastest device, so using a DB device provides the same benefit that the WAL device while also allows for storing additional metadata.

9.3. SIZING CONSIDERATIONS FOR BLUESTORE
When mixing traditional and solid state drives using BlueStore OSDs, it is important to size the RocksDB logical volume (`block.db`) appropriately. Red Hat recommends that the RocksDB logical volume be no less than 4% of the block size. For example, if the block size is 1 TB, then at a minimum, create a 40 GB RocksDB logical volume.
In this scenario, BlueStore's cache memory is determined by the `bluestore_cache_size_ssd` and `bluestore_cache_size_hdd` options.

When not mixing drive types, there is no requirement to have a separate RocksDB logical volume. BlueStore will automatically manage the sizing of RocksDB.

In this scenario, BlueStore's cache memory is determined by the `bluestore_cache_size` option.

BlueStore’s cache memory is used for the key-value pair metadata for RocksDB, BlueStore metadata and object data.

NOTE

The BlueStore cache memory values are in addition to the memory footprint already being consumed by the OSD.

9.4. ADDING OSDS THAT USE BLUESTORE

This section describes how to install a new Ceph OSD node with the BlueStore back end.

**Prerequisites**

- A working Ceph cluster. See the Installation Guide for Red Hat Enterprise Linux or Ubuntu.

**Procedure**

Use the following commands on the Ansible administration node.

1. Add a new OSD node to the `[osds]` section in Ansible inventory file, by default located at `/etc/ansible/hosts`.

   ```
   [osds]
   node1
   node2
   node3
   <hostname>
   
   Replace:
   
   - `<hostname>` with the name of the OSD node
   
   For example:
   
   ```
   [osds]
   node1
   node2
   node3
   node4
   ```

2. Navigate to the `/usr/share/ceph-ansible/` directory.

   ```
   [user@admin ~]$ cd /usr/share/ceph-ansible
   ```

3. Create the `host_vars` directory.
4. Create the configuration file for the newly added OSD in `host_vars`.

```
[root@admin ceph-ansible] touch host_vars/<hostname>.yml
```

*Replace:*

- `<hostname>` with the host name of the newly added OSD

For example:

```
[root@admin ceph-ansible] touch host_vars/node4.yml
```

5. Add the following setting to the newly created file:

```
osd_objectstore: bluestore
```

**NOTE**
To use BlueStore for all OSDs, add `osd_objectstore:bluestore` to the `group_vars/all.yml` file.

6. Optional. If you want to store the `block.wal` and `block.db` partitions on dedicated devices, edit the `host_vars/<hostname>.yml` file as follows.

**a.** To use dedicated devices for `block.wal`:

```
osd_scenario: non-collocated
bluestore_wal_devices:
  - <device>
  - <device>
```

*Replace:*

- `<device>` with the path to the device

For example:

```
osd_scenario: non-collocated
bluestore_wal_devices:
  - /dev/sdf
  - /dev/sgd
```

**b.** To use dedicated devices for `block.db`:

```
osd_scenario: non-collocated
dedicated_devices:
```

```
```
- <device>
- <device>

Replace:

- <device> with the path to the device

For example:

```yaml
osd_scenario: non-collocated
dedicated_devices:
  - /dev/sdh
  - /dev/sdi
```

**NOTE**
If you use the `osd_scenario: collocated` parameter, the `block.wal` and `block.db` partitions will use the same device as specified with the `devices` parameter. For details, see the *Installing a Red Hat Ceph Storage Cluster* section in the Red Hat Ceph Storage 3 Installation Guide for Red Hat Enterprise Linux or Ubuntu.

**NOTE**
To use BlueStore for all OSDs, add the aforementioned parameters to the `group_vars/osds.yml` file.

7. To configure LVM based BlueStore OSDs, use `osd_scenario: lvm` in `host_vars/<hostname>.yml`:

```yaml
osd_scenario: lvm
lvm_volumes:
  - data: <datalv>
    data_vg: <datavg>
```

Replace:

- <datalv> with the data logical volume name
- <datavg> with the data logical volume group name

For example:

```yaml
osd_scenario: lvm
lvm_volumes:
  - data: data-lv1
    data_vg: vg1
```

8. Optional. If you want to store the `block.wal` and `block.db` on dedicated logical volumes, edit the `host_vars/<hostname>.yml` file as follows:

```yaml
osd_scenario: lvm
lvm_volumes:
  - data: data-lv1
    data_vg: vg1
```

CHAPTER 9. BLUESTORE
lvm_volumes:
  - data: <datalv>
  wal: <wallv>
  wal_vg: <vg>
  db: <dblv>
  db_vg: <vg>

Replace:

- <datalv> with the logical volume where the data should be contained
- <wallv> with the logical volume where the write-ahead-log should be contained
- <vg> with the volume group the WAL and/or DB device LVs are on
- <dblv> with the logical volume the BlueStore internal metadata should be contained

For example:

```yaml
osd_scenario: lvm
lvm_volumes:
  - data: data-lv3
  wal: wal-lv1
  wal_vg: vg3
  db: db-lv3
  db_vg: vg3
```

**NOTE**

When using `lvm_volumes:` with `osd_objectstore: bluestore` the `lvm_volumes` YAML dictionary must contain at least `data`. When defining `wal` or `db`, it must have both the LV name and VG name (`db` and `wal` are not required). This allows for four combinations: just data, data and `wal`, data and `wal` and `db`, or data and `db`. Data can be a raw device, lv or partition. The `wal` and `db` can be a lv or partition. When specifying a raw device or partition `ceph-volume` will put logical volumes on top of them.

**NOTE**

Currently, `ceph-ansible` does not create the volume groups or the logical volumes. This must be done before running the Anisble playbook.

9. Open and edit the `group_vars/all.yml` file, and uncomment the `osd_memory_target` option. Adjust the value on how much memory you want the OSD to consume.

**NOTE**

The default value for the `osd_memory_target` option is `4000000000`, which is 4 GB. This option pins the BlueStore cache in memory.

**IMPORTANT**

The `osd_memory_target` option only applies to BlueStore-backed OSDs.
10. Use the **ansible-playbook**:

```
[user@admin ceph-ansible]$ ansible-playbook site.yml
```

11. From a Monitor node, verify that the new OSD has been successfully added:

```
[root@monitor ~]# ceph osd tree
```

**Additional Resources**

- [Adding an OSD with Ansible](#)