

Red Hat Ceph Storage 2.5

Release Notes

Release notes for Red Hat Ceph Storage 2.5

Last Updated: 2018-07-24

Release notes for Red Hat Ceph Storage 2.5

Legal Notice

Copyright © 2018 Red Hat, Inc.

The text of and illustrations in this document are licensed by Red Hat under a Creative Commons Attribution–Share Alike 3.0 Unported license ("CC-BY-SA"). An explanation of CC-BY-SA is available at

http://creativecommons.org/licenses/by-sa/3.0/

. In accordance with CC-BY-SA, if you distribute this document or an adaptation of it, you must provide the URL for the original version.

Red Hat, as the licensor of this document, waives the right to enforce, and agrees not to assert, Section 4d of CC-BY-SA to the fullest extent permitted by applicable law.

Red Hat, Red Hat Enterprise Linux, the Shadowman logo, JBoss, OpenShift, Fedora, the Infinity logo, and RHCE are trademarks of Red Hat, Inc., registered in the United States and other countries.

Linux ® is the registered trademark of Linus Torvalds in the United States and other countries.

Java ® is a registered trademark of Oracle and/or its affiliates.

XFS ® is a trademark of Silicon Graphics International Corp. or its subsidiaries in the United States and/or other countries.

MySQL ® is a registered trademark of MySQL AB in the United States, the European Union and other countries.

Node.js ® is an official trademark of Joyent. Red Hat Software Collections is not formally related to or endorsed by the official Joyent Node.js open source or commercial project.

The OpenStack ® Word Mark and OpenStack logo are either registered trademarks/service marks or trademarks/service marks of the OpenStack Foundation, in the United States and other countries and are used with the OpenStack Foundation's permission. We are not affiliated with, endorsed or sponsored by the OpenStack Foundation, or the OpenStack community.

All other trademarks are the property of their respective owners.

Abstract

The Release Notes document describes the major features and enhancements implemented in Red Hat Ceph Storage in a particular release. The document also includes known issues and bug fixes.

Table of Contents

CHAPTER 1. INTRODUCTION	3
CHAPTER 2. ACKNOWLEDGMENTS	4
CHAPTER 3. MAJOR UPDATES	5
CHAPTER 4. DEPRECATED FUNCTIONALITY	8
CHAPTER 5. KNOWN ISSUES	9
CHAPTER 6. NOTABLE BUG FIXES	11
CHAPTER 7. SOURCES	17

CHAPTER 1. INTRODUCTION

Red Hat Ceph Storage is a massively scalable, open, software-defined storage platform that combines the most stable version of the Ceph storage system with a Ceph management platform, deployment utilities, and support services.

CHAPTER 2. ACKNOWLEDGMENTS

Red Hat Ceph Storage version 2.0 contains many contributions from the Red Hat Ceph Storage team. Additionally, the Ceph project is seeing amazing growth in the quality and quantity of contributions from individuals and organizations in the Ceph community. We would like to thank all members of the Red Hat Ceph Storage team, all of the individual contributors in the Ceph community, and additionally (but not limited to) the contributions from organizations such as:

- Intel
- Fujitsu
- UnitedStack
- Yahoo
- Ubuntu Kylin
- Mellanox
- CERN
- Deutsche Telekom
- Mirantis
- SanDisk

CHAPTER 3. MAJOR UPDATES

This section lists all major updates, enhancements, and new features introduced in this release of Red Hat Ceph Storage.

A heartbeat message for Jumbo frames has been added

Previously, if a network included jumbo frames and the maximum transmission unit (MTU) was not configured properly on all network parts, a lot of problems, such as slow requests, and stuck peering and backfilling processes occurred. In addition, the OSD logs did not include any heartbeat timeout messages because the heartbeat message packet size is below 1500 bytes. This update adds a heartbeat message for Jumbo frames.

The osd_hack_prune_past_interval option is now supported

The **osd_hack_prune_past_interval** option helps to reduce memory usage for the past intervals entries, which can help with recovery of unhealthy clusters.

WARNING

This option can cause data loss, therefore, use it only when instructed by the Red Hat Support Engineers.

The default value for themin_in_ratio option has been increased to 0.75

The min_in_ratio option prevents Monitors from marking OSDs as **out** when the amount of **out** OSDs drops below certain fraction of all OSDs in the cluster. In previous releases, the default value of **min_in_ratio** was set to **0.3**. With this update, the value has been increased to **0.75**.

RocksDB is enabled as an option to replacelevelDB

This update enables an option to use the **RocksDB** back end for the **omap** database as opposed to **levelDB**. **RocksDB** uses the multi-threading mechanism in compaction so that it better handles the situation when the **omap** directories become very large (more than 40 G). **LevelDB** compaction takes a lot of time in such a situation and causes OSD daemons to time out.

For details about conversion from **levelDB** to **RocksDB**, see the Ceph - Steps to convert OSD omap backend from leveldb to rocksdb solution on Red Hat Customer Portal.

The software repository containing the ceph-ansible package has changed

Earlier versions of Red Hat Ceph Storage relied on the **ceph-ansible** package in the **rhel-7server-rhscon-2-installer-rpms** repository. The Red Hat Storage Console 2 product is nearing end-of-life. Therefore, use the **ceph-ansible** package provided by the **rhel-7-server-rhceph-2tools-rpms** repository instead.

Changing the default compression behavior of rocksdb

Disabling compression reduces the size of the I/O operations, but not the I/O operations themselves.

The old default values:

```
filestore_rocksdb_options =
"max_background_compactions=8;compaction_readahead_size=2097152"
```

The new default values:

```
filestore_rocksdb_options =
"max_background_compactions=8;compaction_readahead_size=2097152;compressio
n=kNoCompression"
```

Also, this does not effect any existing OSD, only those OSDs that have been manually converted or newly provisioned OSDs.

The RocksDB cache size can now be larger than 2 GB

Previously, you could not set values larger than 2 GB. Now, the value for **rocksdb_cache_size** parameter, can be set to a larger size, such as 4 GB.

Support for the Red Hat Ceph Storage Dashboard

The Red Hat Ceph Storage Dashboard provides a monitoring dashboard for Ceph clusters to visualize the 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 network.

For details, see the Monitoring Ceph Clusters with Red Hat Ceph Storage Dashboard section in the Administration Guide for Red Hat Ceph Storage 2.

Split threshold is now randomized

Previously, the split threshold was not randomized, so that many OSDs reached it at the same time. As a consequence, such OSDs incurred high latency because they all split directories at once. With this update, the split threshold is randomized which ensures that OSDs split directories over a large period of time.

Logging the time out of disk operations

Ceph OSDs now log when they shutdown due to disk operations timing out by default.

The --yes-i-really-mean-it override option is mandatory for executing theradosgwadmin orphans find command

The **radosgw-admin orphans find** command can inadvertently remove data objects still in use, if followed by another operation, such as, a **rados rm** command. Users are now warned before attempting to produce lists of potentially orphaned objects.

Perform offline compaction on an OSD

The **ceph-osdomap-tool** now has a **compact** command to perform offline compaction on an OSD's **omap** directory.

For S3 and Swift protocols, an option to list buckets/containers in natural (partial) order has been added

Listing containers in sorted order is canonical in both protocols, but is costly, and not required by some client applications. The performance and workload cost of S3 and Swift bucket/container listings is reduced for sharded buckets/containers when the **allow_unordered** extension is used.

Asynchronous Garbage Collection

An asynchronous mechanism for executing the Ceph Object Gateway garbage collection using the **librados** APIs has been introduced. The original garbage collection mechanism serialized all processing, and lagged behind applications in specific workloads. Garbage collection performance has been significantly improved, and can be tuned to specific site requirements.

Deploying Ceph using ceph-ansible on Ubuntu

Previously, Red Hat did not provide the **ceph-ansible** package for Ubuntu. With this release, you can use the Ansible automation application to deploy a Ceph Storage Cluster from an Ubuntu node. See Chapter 3 in the Red Hat Ceph Storage Installation Guide for Ubuntu for more details.

CHAPTER 4. DEPRECATED FUNCTIONALITY

This section provides an overview of functionality that has been deprecated in all minor releases up to this release of Red Hat Ceph Storage.

The Red Hat Storage Console

The Red Hat Storage Console has reached the end of its life cycle, and is deprecated. Use the Ansible automation application with the **ceph-ansible** playbooks to install a Red Hat Storage Ceph cluster. For details, see the Installation Guide for Red Hat Enterprise Linux or Ubuntu.

For cluster monitoring, you can use the Red Hat Ceph Storage Dashboard that provides a monitoring dashboard to visualize the state of a cluster. For details, see the Monitoring Ceph Clusters with Red Hat Ceph Storage Dashboard section in the Administration Guide.

The salt packages have been deprecated and removed

The installation of the Calamari web interface no longer requires the **salt** packages, because of several security vulnerabilities. Remove the **salt** packages manually to mitigate those security vulnerabilities, if they are installed.

CHAPTER 5. KNOWN ISSUES

This section documents known issues found in this release of Red Hat Ceph Storage.

The nfs-ganesha-rgw utility cannot write to the NFS mount, if SELinux is enabled

Currently, **nfs-ganesha-rgw** utility does not run in an unconfined SELinux domain. When SELinux is enabled, write operations to the NFS mount fails. To work around this issue, use SELinux in permissive mode.

(BZ#1535906)

The Ceph-pools Dashboard displays previously deleted pools

The *Ceph-pools* Dashboard continues to reflect the pools, which have been deleted from the Ceph Storage Cluster.

(BZ#1537035)

Installing the Red Hat Ceph Storage Dashboard with a non-default password fails

Currently, you must use the default password when installing the Red Hat Ceph Storage Dashboard. After the installation finishes, you can change the default password. For details see the *Changing the default Red Hat Ceph Storage Dashboard password* section in the *Administration Guide* for Red Hat Ceph Storage 2.

(BZ#1537390)

In the Red Hat Ceph Storage Dashboard, on reboot of an OSD node the OSD IDs are repeated

In the Ceph OSD Information dashboard, the OSD IDs are repeated on a reboot of an OSD node.

(BZ#1537505)

The Red Hat Ceph Storage Dashboard does not display certain graphs correctly

The graphs *Top 5 Pools by Capacity Used*, and *Pool Capacity - Past 7 days* are not displayed on the *Ceph Pools* and *Ceph cluster* dashboards. Also, on the *Ceph OSD Information* dashboard, the graph *BlueStore IO Summary - all OSD's @ 95%ile* displays an *Internal Error*.

(BZ#1538124)

The Ceph version is not reflected in the Red Hat Ceph Storage Dashboard after the Monitor node is rebooted

In the *Ceph Version Configuration* section of the *Ceph Cluster* Dashboard, the Ceph Monitor version is reflects as "-" on a reboot of Ceph Monitor node.

(BZ#1538319)

When adding an OSD to the Red Hat Ceph Storage Dashboard, some details are not displayed

On adding an OSD node to a Ceph storage cluster, the details such as *hostname*, *Disk/OSD Summary*, and *OSD versions* of the newly added OSD are not displayed in the *Ceph Backend Storage*, *Ceph Cluster*, and *Ceph OSD Information* dashboards.

(BZ#1538331)

Installing and upgrading containerized Ceph fails

Using Full Qualified Domain Names (FQDN) in the **/etc/hostname** file for containerized Ceph deployments will fail when installing and upgrading Ceph. When using the **ceph-ansible** playbook to install Ceph, the installation will fail with the following error message:

"msg": "The task includes an option with an undefined variable. The error was: 'osd_pool_default_pg_num' is undefined

To work around the installation failure, change the FQDN in the **/etc/hostname** file to the short host name on all nodes in the storage cluster. Next, rerun the **ceph-ansible** playbook to install Ceph. When upgrading Ceph with the **rolling_update** playbook, the upgrade will fail with the following error message:

 $"\ensuremath{\mathsf{FAILED}}$ - <code>RETRYING: container</code> | waiting for the containerized monitor to join the quorum"

To work around the upgrade failure, change the FQDN in the **/etc/hostname** file to the short host name on all nodes in the storage cluster. Next, restart the corresponding Ceph daemons running on each node in the storage cluster, then rerun the **rolling_update** playbook to upgrade Ceph.

(BZ#1546127)

The copy_admin_key option, when set to true, does not copy the keyring to other nodes

When using **ceph-ansible** with the **copy_admin_key** option set to **true**, the administrator's keyring does not copy to the other nodes in the Ceph Storage Cluster. To work around this issue, you must manually copy the administrator's keyring file to the other nodes. For example, as the **root** user on the Ceph Monitor node:

#scp /etc/ceph/<cluster_name>.client.admin.keyring
<destination_node_name>:/etc/ceph/

(BZ#1546175)

CHAPTER 6. NOTABLE BUG FIXES

This section describes bugs fixed in this release of Red Hat Ceph Storage that have significant impact on users.

OSD operations are no longer blocked when theosd_scrub_sleep option is set

Previously, the scrubbing operation was moved into the unified operations queue, but the **osd_scrub_sleep** option did not change. Consequently, setting the **osd_scrub_sleep** option was blocking OSD operations. With this update, setting the **osd_scrub_sleep** option no longer blocks OSD operations by doing the scrub sleep operation asynchronously.

(BZ#1444139)

Some OSDs fail to come up after reboot

Previously, on a machine with more than five OSDs, some OSDs failed to come up after a reboot because the **systemd** unit for the **ceph-disk** utility timed out after 120 seconds. With this release, the **ceph-disk** code no longer fails, if the OSD **udev** rule triggers prior to the mounting of the **/var/** directory.

(BZ#1458007)

Swift POST operations no longer generate random 500 errors

Previously, when making changes to the same bucket through multiple Ceph Object Gateways, under certain circumstances under heavy load, the Ceph Object Gateway returned a 500 error. With this release, the chances are reduced to cause a race condition.

(BZ#1491739)

In containerized deployment, ceph-disk now retries when it cannot find a partition

Due to a possible race condition between kernel and the **udev** utility, the device file for a partition was not created when the **ceph-disk** utility searched for it. Consequently, the **ceph-disk** utility could not create the partition. In this release, when **ceph-disk** cannot find the partition, it retries the search again, thus working around this possible race condition.

(BZ#1496509)

Stale bucket index entries are no longer left over after object deletions

Previously, under certain circumstances, deleted objects were incorrectly interpreted as incomplete delete transactions because of an incorrect time. As a consequence, the delete operations were reported successful in the Ceph Object Gateway logs, but the deleted objects were not correctly removed from bucket indexes. The incorrect time comparison has been fixed, and deleting objects works correctly.

(BZ#1500904)

The Ceph Object Gateway no longer refuses an S3 upload when the Content-Type field is missing

When doing a Simple Storage Service (S3) upload, if the Content-Type field was missing from the policy part of the upload, then Ceph Object Gateway refused the upload with a 403 error:

Policy missing condition: Content-Type

With this update, the S3 POST policy does not require the Content-Type field.

(BZ#1502780)

The OSD no longer times out when backfilling

Previously, backfilling objects with hundreds of thousands of **omap** entries could cause the OSD to time out. With this release, the backfilling process now reads a maximum of 8096 **omap** entries at a time, allowing the OSDs to continue running.

(BZ#1505561)

Swift buckets can now be accessed anonymously

Previously, it was not possible anonymously access a Swift bucket, even if the permissions allowed it. The logic to anonymously access Swift buckets was missing. With this release, that logic has been added. As a result, Swift buckets can now be accessed anonymously, if the permissions allow the access.

(BZ#1507120)

Unreconstructable object errors are not handled properly

During a backfill or recovery operation of an erasure coded pool, the unreconstructable object errors were not handled properly. As a consequence, the OSDs terminated unexpectedly with the following message:

osd/ReplicatedPG.cc: recover_replicas: object added to missing set for backfill, but is not in recovering, error!

With this update, the error handling has been corrected and two new placement group (PG) states have been added: **recover_unfound** and **backfill_unfound**. As a result, the OSD does not terminate unexpectedly, and the PG state indicates that the PG contains unfound objects.

(BZ#1508935)

The Ceph Object Gateway successfully starts after upgrading Red Hat OpenStack Platform 11 to 12

Previously, when upgrading Red Hat OpenStack Platform 11 to 12, the Ceph Object Gateway would fail to start because port 8080 was already in use by **haproxy**. With this release, you can specify the IP address and port bindings for the Ceph Object Gateway. As a result, the Ceph Object Gateway will start properly.

(BZ#1509584)

Objects eligible for expiration are no longer infrequently passed over

Previously, due to an off-by-one error in expiration processing in the Ceph Object Gateway, objects eligible for expiration could infrequently be passed over, and consequently were not removed. The underlying source code has been modified, and the objects are no longer passed over.

(BZ#1514210)

Folders starting with an underscore (_) are not in the bucket index

Previously, a server-side copy mishandled object names starting with an underscore. This led to objects being created with two leading underscores. The Ceph Object Gateway code has been fixed to properly handle leading underscores. As a result, objects names with leading underscores behave correctly.

(BZ#1515275)

Fixed a memory leak with the Ceph Object Gateway

A buffer used to transfer incoming PUT data was incorrectly sized at the maximum chunk value of 4 MB. This was leading to a space leak of unused buffer space when PUTs of smaller objects were processed. The Ceph Object Gateway can leak space when processing large numbers of PUT requests less than 4M in size. As a result, the incorrect buffer sizing logic was fixed.

(BZ#1522881)

ceph-ansible now disables the Ceph Object Gateway service as expected when upgrading the OpenStack container

When upgrading the OpenStack container from version 11 to 12, the **ceph-ansible** utility did not properly disable the Ceph Object Gateway service provided by the overcloud image. Consequently, the containerized Ceph Object Gateway service entered a failed state because the port it used was bound. The **ceph-ansible** utility has been updated to properly disable the system Ceph Object Gateway service. As a result, the containerized Ceph Object Gateway service Gateway service starts as expected after upgrading the OpenStack container from version 11 to 12.

(BZ#1525209)

The ceph-ansible utility no longer fails when upgrading from Red Hat OpenStack Platform 11 to 12

Ceph containers set their socket name by either the short host name or by the fully qualified domain name (FQDN). Previously, the **ceph-ansible** utility could not predict which host name type would be used. With this release, **ceph-ansible** checks for both socket names.

(BZ#1541303)

Relocated some OSD options the rolling-update.yml Ceph Ansible playbook

Previously, when doing a minor Ceph upgrade, for example, upgrading version 10.2.9 to 10.2.10, the **noout**, **noscrub** and **nodeep-scrub** OSD options did not get applied. Since a daemon does not exist for these versions, the **mgr** section in the **rolling-update.yml** file was skipped. With this release, the OSD options are set properly after all the Ceph Monitors have been upgraded.

(BZ#1548071)

Slow OSD startup after upgrading to Red Hat Ceph Storage 2.5

Ceph Storage Clusters that have large **omap** databases experience slow OSD startup due to scanning and repairing during the upgrade from Red Hat Ceph Storage 2.4 to 2.5. The rolling update may take longer than the specified time out of 5 minutes. Before running the Ansible **rolling_update.yml** playbook, set the **handler_health_osd_check_delay** option to 180 in the **group_vars/all.yml** file.

(BZ#1548481)

An inconsistent PG state can reappear long after the PG was repaired

In rare circumstances after a PG has been repaired and the primary changes, the inconsistent state can falsely reappear, even without a scrub being performed. This bug fix cleans up stray scrub error counts to prevent this.

(BZ#1550892)

The expected_num_objects option was not working as expected

Previously, when using the **ceph osd pool create** command with **expected_num_objects** option, placement group (PG) directories were not pre-created at pool creation time as expected, resulting in performance drops when filestore splitting occurred. With this update, the **expected_num_objects** parameter is now passed through to filestore correctly, and PG directories for the expected number of objects are pre-created at pool creation time.

(BZ#1554963)

The Ceph Object Gateway handles requests for negative byte-range objects correctly

The Ceph Object Gateway was treating negative byte-range object requests as invalid, whereas such requests succeed and return the whole object in AWS S3. This caused applications which expected the AWS behavior for negative or other invalid range requests saw unexpected errors and possible failure. In this update, a new option **rgw_ignore_get_invalid_range** was added to Ceph Object Gateway. When the **rgw_ignore_get_invalid_range** option is **true** (non-default), the Ceph Object Gateway behavior for invalid range requests is backward compatible with AWS.

(BZ#1576487)

Fixed the Ceph Object Gateway multiple site sync for versioned buckets and objects

Previously, internal Ceph Object Gateway multi-site sync logic behaved incorrectly in some scenarios when attempting to sync containers with S3 object versioning enabled. In particular, when a new object upload was followed immediately by an attribute or ACL setting operation. Objects in versioning-enabled containers would fail to sync in some scenarios. For example, when using the **s3cmd sync** command to mirror a filesystem directory. With this update, the Ceph Object Gateway multi-site replication logic has been corrected for the known failure cases.

(BZ#1578401)(BZ#1584763)

Improving performance of the Ceph Object Gateway when using SSL

The Ceph Object Gateway's use of **libcurl** was slower and had a bigger memory footprint. With this update, the Ceph Object Gateway reuses the **libcurl** data structures, which improves SSL efficiency when authenticating using OpenStack Keystone with SSL.

(BZ#1578670)

Update to the ceph-disk Unit Files

The transition to containerized Ceph left some **ceph-disk** unit files. The files were harmless, but appeared as failing, which could be distressing to the operator. With this update, executing the **switch-from-non-containerized-to-containerized-ceph-daemons.yml** playbook disables the **ceph-disk** unit files too.

(BZ#1581579)

The listing of versioned-bucket objects was returning an additional entry

When listing large versioning-enabled buckets with a marker, that marker was also listed. Listings would include an extra duplicate entry for each 1000 objects in the bucket. With this update, the marker is no longer included in bucket listings, and the correct number of entries are returned.

(BZ#1584218)

Cache entries were not refreshing as expected

The new time-based metadata cache entry expiration logic did not include logic to update the expiration time on already-cached entries being updated in place. Cache entries became permanently stale after expiration, leading to a performance regression as metadata objects were effectively not cached and always read from the cluster. Logic was added to update the expiration time of cached entries when updated.

(BZ#1584829)

The Ceph Object Gateway was using large amounts of the CPU

Because of the linkage between the Ceph Object Gateway and **tcmalloc**, a bug was found causing high CPU utilization when **tcmalloc** tries to reclaim space for specific workloads. This bug seems to be **tcmalloc** version specific. The linkage between the Ceph Object Gateway and **tcmalloc** has been reverted for Red Hat Ceph Storage 2.

(BZ#1591455)

The Ceph Object Gateway retry logic can cause high-CPU utilization

A bug in the Ceph Object Gateway retry logic can cause a non-terminating condition in operations processing, which can lead to high-CPU utilization and less noticeable side effects. This high-CPU utilization triggered a "busy-loop" in various workloads. In this update, the error-handling logic has been fixed to handle this condition.

(BZ#1595386)

Reduce OSD memory usage for Ceph Object Gateway workloads

The OSD memory usage was tuned to reduce unnecessary usage, especially for Ceph Object Gateway workloads.

(BZ#1599507)

Ceph Monitor does not mark an OSD down when the cluster network is down

In some cases, when a new OSD was added into an existed storage cluster, the OSD heartbeat peers were not updated. This happened if the new OSD did not get any PGs mapped to it. With this fix, the existing OSDs will refresh their heartbeat peers when a new OSD joins the storage cluster. As a result, the Ceph Monitor marks the OSD daemon as down if the OSD node's cluster network is down.

(BZ#1488389)

Better recovery from cache inconsistency for the Ceph Object Gateway nodes

When the Ceph Object Gateway nodes were experiencing a heavy load, cache transmission updates sent through the watch or notify requests were being lost. Consequently, certain Ceph Object Gateway nodes were left with stale cache data, which causes several problems, notably that the bucket index contained old data. With this update, secondary cache coherency mechanisms have been added:

• Bucket operations now look for errors that indicate cache inconsistency.

- Ability to forcibly refresh has been added.
- A timeout to bound the age of any cache entry, forcing an eventual refresh at some point has been added.

As a result, the Ceph Object Gateway nodes recover from cache inconsistency rather than entering a persistent, user-visible error state.

(BZ#1491723)

CHAPTER 7. SOURCES

The updated Red Hat Ceph Storage packages are available at the following locations:

- For Red Hat Enterprise Linux: http://ftp.redhat.com/redhat/linux/enterprise/7Server/en/RHCEPH/SRPMS/
- For Ubuntu: https://rhcs.download.redhat.com/ubuntu/