Red Hat Ceph Storage 2.4

Release Notes

Release notes for Red Hat Ceph Storage 2.4

Last Updated: 2018-01-23

Red Hat Ceph Storage 2.4 Release Notes

Release notes for Red Hat Ceph Storage 2.4

Legal Notice

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

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 2.4 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 new structure for detecting duplicate operations

Under certain circumstances, it is better to recover by using the backfill process instead of log-based recovery. The most direct way to force backfilling is to reduce the size of the placement group log. This setting, however, undermines the detection of duplicate operations. This update introduces a separate structure to detect duplicate operations beyond the placement group log entries. As a result, duplicate operations are detected as expected.

`RocksDB` is enabled as an option to replace`levelDB`

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.

Scrubbing is blocked for any PG if the primary or any replica OSDs are recovering

The **`osd_scrub_during_recovery`** parameter now defaults to **`false`**, so that when an OSD is recovering, the scrubbing process is not initialized on it. Previously, **`osd_scrub_during_recovery`** was set to **`true`** by default allowing scrubbing and recovery to run simultaneously. In addition, in previous releases if the user set **`osd_scrub_during_recovery`** to **`false`**, only the primary OSD was checked for recovery activity.

A new `compact` command

With this update, the OSD administration socket supports the **`compact`** command. A large number of **`omap`** create and delete operations can cause the normal compaction of the **`levelDB`** database during those operations to be too slow to keep up with the workload. As a result, **`levelDB`** can grow very large and inhibit performance. The **`compact`** command compacts the **`omap`** database (**`levelDB`** or **`RocksDB`**) to a smaller size to provide more consistent performance.

Improved delete handling

With this update, delete handling has been improved and it is enabled by default in new clusters. To enable this improvement in previously deployed clusters, use the **`ceph osd set recovery_deletes`** command after upgrading to version 2.4.

A new Compatibility Guide

A new Compatibility Guide is now available. The guide provides a matrix that lists which versions of various products are compatible with this version of Red Hat Ceph Storage. See the Compatibility Guide for details.

CHAPTER 4. NOTABLE BUG FIXES

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

Ceph now handles delete operations during recovery instead of the peering process

Previously, bringing an OSD that was down or out for longer than 15 minutes back to the cluster caused placement group peering times to be elongated. The peering process took a long time to complete because delete operations were processed inline while merging the placement group log as part of peering. As a consequence, operations to the placement group that were in the peering state were blocked. With this update, Ceph handles delete operations during normal recovery instead of the peering process. As a result, the peering process completes faster and operations are no longer blocked.

(BZ#1452780)

Several AWS version 4 signature bugs are fixed

This update fixes several Amazon Web Service (AWS) version 4 signature bugs.

(BZ#1456060)

Repairing bucket indexes works as expected

Previously, the **`cls`** method of the Ceph Object Gateway that is used for repairing bucket indexes failed when its output result was too large. Consequently, affected bucket index objects could not be repaired using the **`bucket check --fix`** command, and the command failed with the "(90) Message too long" error. This update introduces a paging mechanism that ensures that bucket indexes can be repaired as expected.

(BZ#1463969)

Fixed incorrect handling of source headers containing the slash character

Incorrect handling of source headers that contained slash ("/") characters caused the unexpected authentication failure of an Amazon Web Services (AWS) version 4 signature. This error prevented specific operations, such as copying Hadoop Amazon Simple Storage Services (S3A) multipart objects, from completing. With this update, handling of slash characters in source headers has been improved, and the affected operations can be performed as expected.

(BZ#1470301)

Fixed incorrect handling of headers containing the plus character

Incorrect handling of the plus character ("+") in Amazon Web Services (AWS) version 4 canonical headers caused unexpected authentication failures when operating on such objects. As a consequence, some operations, such as Hadoop Amazon Simple Storage Services (S3A) distributed copy (DistCp), failed unexpectedly. This update ensures that the plus character is escaped as required, and affected operations no longer fail.

(BZ#1470836)

CRUSH calculations for removed OSDs match on kernel clients and the cluster

When an OSD was removed with the **`ceph osd rm`** command, but was still present in the CRUSH map, the CRUSH calculations for that OSD on kernel clients and the cluster did not match. Consequently, kernel clients returned I/O errors. The mismatch between client and server behavior has been fixed and kernel clients do not return the I/O errors anymore in this situation.

(BZ#1471939)

OSDs now wait up to three hours for other OSD to complete its initialization sequence

At boot time, an OSD daemon could fail to start when it took more than five minutes to wait for other OSD to complete its initialization sequence. As a consequence, such OSDs had to be started manually. With this update, OSDs wait up to three hours. As a result, OSDs no longer fail to start when the initialization sequence of other OSDs takes too long.

(BZ#1472409)

The garbage collection now properly handles parts of resent multipart objects

Previously, when parts of multipart uploads were resent, they were mistakenly made eligible for garbage collection. As a consequence, attempts to read such multipart objects failed with the "404 Not Found" error. With this update, the garbage collection has been fixed to properly handle this case. As a result, such multipart objects can be read as expected.

(BZ#1476865)

The multi-site synchronization works as expected

Due to an object lifetime defect in the Ceph Object Gateway multi-site synchronization code path, a failure could occur during incremental sync. The underlying source code has been modified, and the multi-site synchronization works as expected.

(BZ#1476888)

A new serialization mechanism for upload completions is supported

A race condition in completion of multipart upload operations could fail if a client retried its complete operation while the original completion was still in progress. As a consequence, a multipart upload failed, especially, when it was slow to complete. This update introduces a new serialization mechanism for upload completions, and the multipart upload failures no longer occur.

(BZ#1477754)

Encrypted OSDs no longer fail after upgrading to 2.3

Since version 2.3, a test has been added that checks if the **`ceph_fsid`** file exists inside the **`lockbox`** directory. If the file does not exist, an attempt to start encrypted OSDs fails. Because previous versions did not include this test, after upgrading to 2.3, the encrypted OSDs failed to start after rebooting. This bug has been fixed, and encrypted OSDs no longer fail after upgrading to version 2.3 or later.

(BZ#1477775)

Fixing bucket indexes no longer damages them

Previously, a bug in the Ceph Object Gateway namespacing could cause the bucket index repair process to incorrectly delete object entries. As a consequence, an attempt to fix a bucket index could damage the index. The bug has been fixed, and fixing bucket indexes no longer damages them.

(BZ#1479949)

Encrypted containerized OSDs starts as expected after a reboot

Encrypted containerized OSD daemons failed to start after a reboot. In addition, the following log message was added to the OSD log file:

```
filestore(/var/lib/ceph/osd/bb-1) mount failed to open journal
/var/lib/ceph/osd/bb-1/journal: (2) No such file or directory
```

This bug has been fixed, and such OSDs start as expected in this situation.

(BZ#1488149)

`ceph-disk` retries up to ten times to find files that represents newly created OSD partitions

When deploying a new OSD with the **`ceph-ansible`** playbook, the file under the **`/sys/`** directory that represents a newly created OSD partition failed to show up right after the **`partprobe`** utility returned it. Consequently, the **`ceph-disk`** utility failed to activate the OSD, and **`ceph-ansible`** could not deploy the OSD successfully. With this update, if **`ceph-disk`** cannot find the file, it retries up to ten times to find it before it terminates. As a result, **`ceph-disk`** can activate the newly prepared OSD as expected.

(BZ#1491780)

Bugs in the Ceph Object Gateway quota have been fixed

An integer underflow in cached quota values in the Ceph Object Gateway server could allow users to exceed quota. In addition, a double counting error in the quota check for multipart uploads caused early enforcement for that operation when it was performed near the quota limit. This update fixes these two errors.

(BZ#1498280)

Multi-site synchronization no longer terminates unexpectedly with a segmentation fault

In a multi-site configuration of the Ceph Object Gateway, when data synchronization started and the data sync status was **`status=Init`**, the synchronization process reinitialized the sync status but set the number of shards incorrectly to 0. Consequently, the synchronization terminated unexpectedly with a segmentation fault. This bug has been fixed by updating the number of sync log shards, and synchronization works as expected.

(BZ#1500206)

CHAPTER 5. 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/