Red Hat Ceph Storage 2.3 Release Notes

Release notes for Red Hat Ceph Storage 2.3

Red Hat Ceph Storage Documentation Team

Red Hat Ceph Storage 2.3 Release Notes

Release notes for Red Hat Ceph Storage 2.3

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.

Important

For customers using the Ceph Object Gateway multi-site feature, Red Hat recommends to not upgrade to Red Hat Ceph Storage 2.3 due to the known issues identified in the release. You can find details about these issues in the Known Issues section. Red Hat is investigating the fixes for these bugs and will make them available in the first update for Red Hat Ceph Storage 2.3. That will be the build Ceph Object Gateway multi-site customers can upgrade to.

CHAPTER 2. ACKNOWLEDGMENTS

Red Hat Ceph Storage version 2.3 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
- UbuntuKylin
- 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.

Scrub processes can now be disabled during recovery

A new option **`osd_scrub_during_recovery`** has been added with this release. Setting this option to **`false`** in the Ceph configuration file disables starting new scrub processes during recovery. As a result, the speed of the recovery is enhanced.

The radosgw-admin utility supports a new bucket limitcheck command

The **`radosgw-admin`** utility has a new command **`bucket limitcheck`** to warn the administrator when a bucket needs resharding. Previously, buckets with more objects than is recommended could be unnoticed and cause performance issues. This new command reports on bucket status with respect to the configured bucket sharding recommendations ensuring that administrators can detect overloaded buckets easily.

Red Hat Ceph Storage now ships with a companion ISO that contains the debuginfo packages

Previously, it was difficult for users to consume the **`debuginfo`** packages in restricted environments where the Red Hat CDN was not directly accessible. Red Hat Ceph Storage now ships with a companion ISO for Red Hat Enterprise Linux that contains the **`debuginfo`** packages for the product. User can now use this ISO to obtain the **`debuginfo`** packages.

The process of enabling SELinux on a Ceph Storage Cluster has been improved

A new subcommand has been added to the ceph-disk utility that can help make the process of enabling SELinux on a Ceph Storage Cluster faster. Previously, the standard way of SELinux labeling did not take into account the fact that OSDs usually reside on different disks. This caused the labelling process to be slow. This new subcommand is designed to speed up the process by labeling the Ceph files in parallel per OSD.

Subscription Manager now reports on the raw disk capacity available per OSD

With this release, Red Hat Subscription Manager can report on the raw disk capacity available per OSD. To do so:

```
# subscription-manager facts
```

The Ceph Object Gateway data logs are trimmed automatically

Previously, after data logs in the Ceph Object Gateway were processed by data syncs and were no longer needed, they remained in the Ceph Object Gateway host taking up space. Ceph now automatically removes these logs.

Improved error messaging in the Ceph Object Gateway

Previously, when an invalid placement group configuration prevented the Ceph Object Gateway from creating any of its internal pools, the error message was insufficient, making it difficult to deduce the root cause of failure.

The error message now suggests there might be an issue with the configuration such as an insufficient number of placement groups or inconsistent values set for the **pg_num** and **pgp_num** parameters, making it easier for the administrator to solve the problem.

Use CEPH_ARGS to ensure all commands work for clusters with unique names

In Red Hat Ceph Storage, the **cluster** variable **in group_vars/all** determines the name of the cluster. Changing the default value to something else means that all the command line calls need to be changed as well. For example, if the cluster name is **foo**, then **ceph health** becomes **ceph --cluster foo health**.

An easier way to handle this is to use the environment variable **CEPH_ARGS**. In this case, run **export CEPH_ARGS="--cluster foo"**. With that, you can run all command line calls normally.

Improvements to the snapshot trimmer

This release improves control and throttling of the snapshot trimmer in the underlying Reliable Autonomic Distributed Object Store (RADOS).

A new **osd_max_trimming_pgs** option has been introduced, which limits how many placement groups on an OSD can be trimming snapshots at any given time. The default setting for this option is 2.

This release also restores the safe use of the **osd_snap_trim_sleep** option. This option adds the given number of seconds in delay between every dispatch of snapshot trim operations to the underlying system. By default, this option is set to 0.

The new version of the Ceph container is fully supported

The new version of the Ceph container image is based on the Red Hat Ceph Storage 2.3 and Red Hat Enterprise Linux 7.3. This version is now fully supported.

For details, see the Deploying Red Hat Ceph Storage 2 as a Container Image Red Hat Knowledgebase article.

Exporting namespaces to NFS-Ganesha

NFS-Ganesha is an NFS interface for the Ceph Object Gateway that presents buckets and objects as directories and files. With this update, NFS-Ganesha fully supports the ability to export Amazon S3 object namespaces by using NFS 4.1.

For details, see the Exporting the Namespace to NFS-Ganesha section in the Red Hat Ceph Storage 2 Object Gateway for Red Hat Enterprise Linux.

In addition, NFSv3 is newly added as a technology preview. For details, see the Technology Previews section.

The Troubleshooting Guide is available

With this release, the Troubleshooting Guide is available. The guide contains information about fixing the most common errors you can encounter with the Ceph Storage Cluster.

CHAPTER 4. TECHNOLOGY PREVIEWS

This section provides an overview of Technology Preview features introduced or updated in this release of Red Hat Ceph Storage.

Important

Technology Preview features are not supported with Red Hat production service level agreements (SLAs), might not be functionally complete, and Red Hat does not recommend to use them for production. These features provide early access to upcoming product features, enabling customers to test functionality and provide feedback during the development process.

For more information on Red Hat Technology Preview features support scope, see https://access.redhat.com/support/offerings/techpreview/.

The Ceph Object Gateway now supports NFSv3 protocol as a technology preview

Support has been added for accessing Ceph objects by way of the NFSv3 protocol, alongside the existing NFSv4. The NFSv3 protocol has been superseded by NFSv4, but remains widely supported and used.

Object upload semantics are slightly different than when using NFSv4 due to the lack of OPEN and CLOSE operations in NFSv3. For NFSv3 clients, the Gateway emulates OPEN operations, and when objects are uploaded, commits the update transaction when no data has been seen for a short time period, instead of on CLOSE.

Additionally, NFS-Ganesha now fully supports the ability to export Amazon S3 object namespaces by using NFS 4.1. For details, see the Major Updates section.

CHAPTER 5. KNOWN ISSUES

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

Adding an MDS to an existing cluster fails

Adding a Ceph Metadata Server (MDS) to an existing cluster fails with the error:

```
osd_pool_default_pg_num is undefined\n\nThe error appears to have been
in '/usr/share/ceph-ansible/roles/ceph-
mon/tasks/create_mds_filesystems.yml
```

As a consequence, an attempt to create an MDS pool fails.

To work around this issue, add the **`osd_pool_default_pg_num`** parameter to **`ceph_conf_overrides`** in the **`/usr/share/ceph-ansible/group_vars/all.yml`** file, for example:

```
ceph_conf_overrides:
  global:
  osd_pool_default_pg_num: 64
```

(BZ#1461367)

OSD activation fails when running the osd_disk_activate.sh script in the Ceph container image when a cluster name contains numbers

In the Ceph container image, the **`osd_disk_activate.sh`** script considers all numbers included in a cluster name as an OSD ID. As a consequence, OSD activation fails when running the script because the script is seeking a keyring on a path based on an OSD ID that does not exist.

To workaround this issue, do not use cluster names that contain numbers. (BZ#1458512)

Multi-site configuration of the Ceph Object Gateway sometimes fails when options are changed at runtime

When the **`rgw md log max shards`** and **`rgw data log num shards`** options are changed at runtime in multi-site configuration of the Ceph Object Gateway, the **`radosgw`** process terminates unexpectedly with a segmentation fault.

To avoid this issue, do not change the aforementioned options at runtime, but set them during the initial configuration of the Ceph Object Gateway. (BZ#1330952)

Simultaneous upload operations to the same file cause I/O errors

Simultaneous upload operations to the same file location by different NFS clients cause I/O errors on both clients. Consequently, no data is updated in the Ceph Object Gateway cluster; if an object already existed in the cluster in the same location, it is unchanged.

To work around this problem, do not simultaneously upload to the same file location. (BZ#1420328)

Old zone group name is sometimes displayed alongside with the new one

In a multi-site configuration when a zone group is renamed, other zones can in some cases continue to display the old zone group name in the output of the **`radosgw-admin zonegroup list`** command.

To work around this issue:

1. Verify that the new zone group name is present on each cluster.
2. Remove the old zone group name:

```
$ rados -p .rgw.root rm zonegroups_names.<old-name>
```

(BZ#1423402)

Some OSDs fail to come up after reboot

On a machine with more than five OSDs, some OSDs fail to come up after a reboot because the systemd unit for the **`ceph-disk`** utility times out after 120 seconds.

To work around this problem, edit the **`/usr/lib/systemd/system/ceph-disk\@.service`** file and replace 120 with 7200. (BZ#1458007)

The GNU tar utility currently cannot extract archives directly into the Ceph Object Gateway NFS mounted file systems

The current version of the GNU tar utility makes overlapping write operations when extracting files. This behavior breaks the strict sequential write restriction in the current version of the Ceph Object Gateway NFS. In addition, GNU tar reports these errors in the usual way, but it also by default continues extracting the files after reporting the errors. As a result, the extracted files can contain incorrect data.

To work around this problem, use alternate programs to copy file hierarchies into the Ceph Object Gateway NFS. Recursive copying by using the **`cp -r`** command works correctly. Non-GNU archive utilities might be able to correctly extract the tar archives, but none have been verified. (BZ#1418606)

Updating a Ceph cluster deployed as a container rolling_update.yml fails

After updating a Ceph cluster deployed as a container image by using the **`rolling_update.yml`** playbook, the **`ceph-mon`** daemons are not restarted. As a consequence, they are unable to join the quorum after the upgrade.

To work around this issue, follow the steps described in the Updating Red Hat Ceph Storage deployed as a Container Image Knowledgebase article on the Red Hat Customer Portal instead of using **`rolling_update.yml`**. (BZ#1458024)

The --inconsistent-index option of the radosgw-admin bucket rm should never be used

Using the **`--inconsistent-index`** option with **`radosgw-admin bucket rm`** can cause corruption of the bucket index if the command fails or is stopped. Do not use this option. (BZ#1464554)

Failover and failback cause data sync issues in multi-site environments

In environments using the Ceph Object Gateway multi-site feature, failover and failback cause data sync to stall. This is because the **`radosgw-admin sync status`** command reports that **`data sync is behind`** for an extended period of time.

To workaround this issue, run **`radosgw-admin data sync init`** and restart gateways. (BZ#1459967)

The container image has incorrect owner and group IDs

In the Red Hat Ceph Storage container image, the owner and group IDs for some processes are incorrect. The group ID of the **`ceph-osd`** process is **`disk`** when it is supposed to be **`ceph`**. The owner and group IDs for the files **`/etc/ceph/`** **`root:root`** when they it is supposed to be **`ceph:ceph`**. (BZ#1451349)

Using IPv6 addressing is not supported with containerized Ceph clusters

An attempt to deploy a Ceph cluster as a container image fails if IPv6 addressing is used. To work around this issue, use IPv4 addressing only. (BZ#1451786)

Ceph Object Gateway multi-site replication does not work

In the Ceph Object Gateway there is an option to set the hostname in the zonegroup for each gateway. When using multi-site replication, the Ceph Object Gateways responsible for replication (the gateways that are part of primary and secondary site zones as endpoints) should not use this option. This causes the multi-site replication feature to fail.

To workaround this issue, please use the default NULL. Comment out the **`rgw dns name`** option for the respective Ceph Object Gateways and restart them. (BZ#1464268)

Ceph Object Gateway crashes with Swift DLO operations

The Ceph Object Gateway crashes when a system user attempts Swift DLO operations. (BZ#1469355)

CHAPTER 6. NOTABLE BUG FIXES

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

The output of the `radosgw-admin realm rename` command now alerts the administrator to run the command separately on each of the realm's clusters

In a multi-site configuration, the name of a realm is only stored locally and is not shared as part of the period. As a consequence, when it is changed on one cluster, the name is not updated on the other cluster. Previously, users could easily miss this step, which could lead to confusion. With this update, the output of the **`radosgw-admin realm rename`** command contains instructions to rename the realm on other clusters as well. (BZ#1423886)

In the Ceph Object Gateway multi-site configuration, when the data log for replication was larger than 1000 entries, queries to list the data log entered to an infinite loop and used all memory. As a consequence, objects were not replicated from the primary to the secondary Ceph Object Gateway. With this update, the queries no longer loop over the entries, which prevents them to enter to infinite loops. As a result, the objects are replicated as expected. (BZ#1465446)

"radosgw-admin zone create" no longer creates an incorrect zone ID

Previously, the **`radosgw-admin zone create`** command with a specified zone ID created a zone with a different zone ID. This bug has been fixed, and the command now creates a zone with the specified zone ID. (BZ#1418235)

The radosgw-admin utility no longer logs an unnecessary message

Previously, the **`radosgw-admin`** utility logged the following message every time even if the message was not relevant:

```
`2017-02-11 00:09:56.704029 7f9011d259c0 0 System already converted`
```

The log level of this message has been changed from 0 to 20. As a result, the **`radosgw-admin`** command logs the aforementioned message only when appropriate. (BZ#1421819)

Results from deep scrubbing are no longer overwritten by shallow scrubbing

Previously, when performing shallow scrubbing after deep scrubbing, results from deep scrubbing were overwritten by results from shallow scrubbing. As a consequence, the deep scrubbing results were lost. Now, unless the **`nodeep_scrub`** flag is set, no shallow scrubbing is performed regularly, so the information from deep scrubbing is regenerated. (BZ#1330023)

An OSD failure no longer causes significant delay

Previously, when an OSD and a Monitor were colocated on the same node and the node failed, Ceph waited some time before sending the note to the Monitor so the Monitor could decide if it wanted to mark the OSD as down. This could lead to a significant delay. With this update, when an OSD is known to be down, the cluster becomes aware immediately after the failure report, and Ceph sends the note to the Monitor right away. (BZ#1425115)

The Ceph Object Gateway provides valid time stamps for newly created objects

Previously, the Ceph Object Gateway was storing 0 in the **x-timestamp** fields for all objects. This bug has been fixed, and newly created objects have the correct time stamps. Note that old objects will still retain the 0 time stamps. (BZ#1439917)

Swift SLOs can now be read from any other zones

Previously, the Ceph Object Gateway failed to fetch manifest files of Swift Static Large Objects (SLO). As a consequence, an attempt to read those objects from any other zone than the zone where the object was originally uploaded failed. This bug has been fixed, and the objects are read from all zones as expected. (BZ#1423858)

Ansible and "ceph-disk" no longer fail to create encrypted OSDs if the cluster name is different than "ceph"

Previously, the **ceph-disk** utility did not support configuring the **dmcrypt** utility if the cluster name was different than "ceph". Consequently, it was not possible to use the **ceph-ansible** utility to create encrypted OSDs if you use a custom cluster name.

This bug has been fixed, and custom cluster names can now be used. (BZ#1391920)

Two new parameters have been introduced to cope with the errors caused by modern Keystone token types

The token revocation API that the Ceph Object Gateway uses no longer works with modern token types in OpenStack and Keystone. This causes errors in the Ceph log and Python backtraces in Keystone.

To cope with these errors, two new parameters **rgw_keystone_token_cache_size** and **rgw_keystone_revocation_interval** have been introduced. Setting the **rgw_keystone_toke_cache_size** parameter to 0 in the Ceph configuration file removes the errors. Setting the **rgw_keystone_revocation_interval** parameter to 0 improves performance, but removes the ability to revoke tokens. (BZ#1438965)

ceph-radosgw starts as expected after upgrading from 1.3 to 2 when a non-default value is used for rgw_region_root_pool and rgw_zone_root_pool

Previously, the **ceph-radosgw** service did not start after upgrading the Ceph Object Gateway from 1.3 to 2, when the Gateway used non-default values for the **rgw_region_root_pool** and **rgw_zone_root_pool** parameters. This bug has been fixed and the **ceph-radosgw** now starts as expected. (BZ#1396956)

bi-list operations now perform as expected

Previously, the addition of new bucket index key ranges for multi-site replication induced an unintended bucket index entry decoding problem in the **bi-list operation**, which is now used during bucket resharding. Consequently, bucket resharding failed when multi-site replication was used.

The logic has been changed in the **bi-list** operation to resolve this bug, and **bi-list** operations can be performed as expected when multi-site replication is used. (BZ#1446665)

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/