Red Hat Ceph Storage 3.1

Release Notes

Release notes for Red Hat Ceph Storage 3.1

Last Updated: 2018-10-08

Red Hat Ceph Storage 3.1 Release Notes

Release notes for Red Hat Ceph Storage 3.1

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 **3**

CHAPTER 2. ACKNOWLEDGMENTS **4**

CHAPTER 3. NEW FEATURES **5**
3.1. CEPH ANSIBLE 5
3.2. CEPH DASHBOARD 5
3.3. CEPHFS 5
3.4. ISCSI GATEWAY 6
3.5. OBJECT GATEWAY 6
3.6. OBJECT GATEWAY MULTISITE 8
3.7. PACKAGES 8
3.8. RADOS 8

CHAPTER 4. BUG FIXES **9**
4.1. CEPH ANSIBLE 9
4.2. CEPH DASHBOARD 11
4.3. CEPH-DISK UTILITY 11
4.4. CEPHFS 12
4.5. CEPH MANAGER PLUGINS 12
4.6. CEPH-VOLUME UTILITY 12
4.7. CONTAINERS 12
4.8. ISCSI GATEWAY 13
4.9. OBJECT GATEWAY 13
4.10. OBJECT GATEWAY MULTISITE 15
4.11. RADOS 15

CHAPTER 5. TECHNOLOGY PREVIEWS **16**

CHAPTER 6. KNOWN ISSUES **17**
6.1. CEPH ANSIBLE 17
6.2. CEPH DASHBOARD 18
6.3. CEPH-VOLUME UTILITY 19
6.4. ISCSI GATEWAY 19
6.5. OBJECT GATEWAY 20
6.6. RADOS 20

CHAPTER 7. SOURCES **22**

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.

The Red Hat Ceph Storage documentation is available at https://access.redhat.com/documentation/en/red-hat-ceph-storage/.

CHAPTER 2. ACKNOWLEDGMENTS

Red Hat Ceph Storage version 3.1 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
- SUSE

CHAPTER 3. NEW FEATURES

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

3.1. CEPH ANSIBLE

Support for iSCSI gateway upgrades through rolling updates

Previously, when using a Ceph iSCSI gateway node, **`iscsi-gws`** could not be updated by **`ceph-ansible`** during a rolling upgrade. With this update to Red Hat Ceph Storage, **`ceph-ansible`** now supports upgrading **`iscsi-gws`** using the **`rolling_update.yml`** Ansible playbook.

Support NVMe based bucket index pools

Previously, configuring Ceph to optimize storage on high speed NVMe or SATA SSDs when using Object Gateway was a completely manual process which required complicated LVM configuration.

With this release, the **`ceph-ansible`** package provides two new Ansible playbooks that facilitate setting up SSD storage using LVM to optimize performance when using Object Gateway. See the Using NVMe with LVM Optimally chapter in the Red Hat Ceph Storage Object Gateway for Production Guide for more information.

3.2. CEPH DASHBOARD

The Prometheus plugin for the Red Hat Ceph Storage Dashboard

Previously, the Red Hat Ceph Storage Dashboard used **`collectd`** and Graphite for gathering and reporting on Ceph metrics. With this release, Prometheus is now used for data gathering and reporting, and provides querying capabilities. Also, Prometheus is much less resource intensive. See the Red Hat Ceph Storage Administration Guide for more details on the Prometheus plugin.

Installation of Red Hat Ceph Storage Dashboard using the `ansible` user

Previously, installing Red Hat Ceph Storage Dashboard (cephmetrics) with Ansible required root access. Traditionally, Ansible uses passwordless ssh and sudo with a regular user to install and make changes to systems. In this release, the Red Hat Ceph Storage Dashboard can be installed with **`ansible`** using a regular user. For more information on the Red Hat Ceph Storage Dashboard, see the Administration Guide.

The Red Hat Ceph Storage Dashboard displays the amount of used and available RAM on the storage cluster nodes

Previously, there was no way to view the actual memory usage on cluster nodes from the *Red Hat Ceph Storage Dashboard*. With this update to Red Hat Ceph Storage, a memory usage graph has been added to the *OSD Node Detail* dashboard.

The Red Hat Ceph Storage Dashboard supports OSDs provisioned by the `ceph-volume` utility

In this release, an update to the Red Hat Ceph Storage Dashboard adds support for displaying information on **`ceph-volume`** provisioned OSDs.

3.3. CEPHFS

More accurate CephFS free space information

The CephFS kernel client now reports the same, more accurate free space information as the fuse client via the **df** command.

3.4. ISCSI GATEWAY

The `max_data_area_mb` option is configurable per-LUN

Previously, the amount of memory the kernel used to pass SCSI command data to **tcmu-runner** was hard coded to 8MB. The hard coded limit was too small for many workloads and resulted in reduced throughput and/or TASK SET FULL errors filling initiator side logs. This can now be configured by setting the **max_data_area_mb** value with **gwcli**. Information on the new setting and command can be found in the Red Hat Ceph Storage Block Device Guide.

iSCSI gateway command-line utility (`gwcli`) supports snapshot create, delete, and rollback capabilities

Previously, to manage the snapshots of RBD-backed LUN images the **rbd** command line utility was utilized for this purpose. The **gwcli** utility now includes built-in support for managing LUN snapshots. With this release, all snapshot related operations can now be handled directly within the **gwcli** utility.

Disabling CHAP for iSCSI gateway authentication

Previously, CHAP authentication was required when using the Ceph iSCSI gateway. With this release, disabling CHAP authentication can be configured with the **gwcli** utility or with Ceph Ansible. However, mixing clients with CHAP enabled and disabled is not supported. All clients must either have CHAP enabled or disabled. If enabled, clients might have different CHAP credentials.

3.5. OBJECT GATEWAY

Improved Swift container ACL conformance has been added

Previously, Red Hat Ceph Storage did not support certain ACL use cases, including setting of container ACLs whose subject is a Keystone project/tenant.

With this update of Ceph, many Swift container ACLs which were previously unsupported are now supported.

Improvements to *radosgw-admin* sync status commands

With this update of Red Hat Ceph Storage a new **radosgw-admin bucket sync status** command has been added, as well as improvements to the existing **sync status** and **data sync status** commands.

These changes will make it easier to inspect the progress of multisite syncs.

Automated trimming of bucket index logs

When multisite sync is used, all changes are logged in the bucket index. These logs can grow excessively large. They also are no longer needed once they have been processed by all peer zones.

With this update of Red Hat Ceph Storage, the bucket index logs are automatically trimmed and do not grow beyond a reasonable size.

Admin socket command to invalidate cache

Two new admin socket commands to manipulate the cache were added to the **radosgw-admin** tool.

The **cache erase <objectname>** command flushes the given object from the cache.

The **cache zap** command erases the entire cache.

These commands can be used to help debug problems with the cache or provide a temporary workaround when an RGW node is holding stale information in the cache. Administrators can now flush any and all objects from the cache.

New administrative sockets added for the radosgw-admin command to view the Object Gateway cache

Two new administrative sockets were added to the **radosgw-admin** command to view the contents of the Ceph Object Gateway cache.

The **cache list [string]** sub-command lists all objects in the cache. If the optional **string** is provided, it only matches those objects containing the string.

The **cache inspect <objectname>** sub-command prints detailed information about the object.

These commands can be used to help debug caching problems on any Ceph Object Gateway node.

Implementation of partial order bucket/container listing

Previously, list bucket/container operations always returned elements in a sorted order. This has high overhead with sharded bucket indexes. Some protocols can tolerate receiving elements in arbitrary order so this is now allowed. An example **curl** command using this new feature:

```
curl GET http://server:8080/tb1?allow-unordered=True
```

With this update to Red Hat Ceph Storage, unordered listing via Swift and S3 is supported.

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.

Relaxed region constraint enforcement

In Red Hat Ceph Storage 3.x when using **s3cmd** and option **--region** with a zonegroup that does not exist an **InvalidLocationConstraint** error will be generated. This did not occur in Ceph 2.x because it did not have strict checking on the region. With this update Ceph 3.1 adds a new **rgw_relaxed_region_enforcement** boolean option to enable relaxed (non-enforcement of region constraint) behavior backward compatible with Ceph 2.x. The option defaults to False.

Default rgw_thread_pool_size value change to 512

The default **rgw_thread_pool_size** value changed from 100 to 512. This change accommodates larger workloads. Decrease this value for smaller workloads.

Increased the default value for the objecter_inflight_ops option

The default value for the **objecter_inflight_ops** option was changed from 1024 to 24576. The original default value was insufficient to support a typical Object Gateway workload. With this enhancement, larger workloads are supported by default.

3.6. OBJECT GATEWAY MULTISITE

Add option --trim-delay-ms in radosgw-admin sync error trim command - to limit the frequency of osd ops

A "trim delay" option has been added to the "radosgw-admin sync error trim" command in Ceph Object Gateway multisite. Previously, many OMAP keys could have been deleted by the full operation, leading to potential for impact on client workload. With the new option, trimming can be requested with low client workload impact.

3.7. PACKAGES

Rebase Ceph to version 12.2.5

Red Hat Ceph Storage 3.1 is now based on upstream Ceph Luminous 12.2.5.

3.8. RADOS

Warnings about objects with too many omap entries

With this update to Red Hat Ceph Storage warnings are displayed about pools which contain large omap objects. They can be seen in the output of **`ceph health detail`**. Information about the large objects in the pool are printed in the cluster logs. The settings which control when the warnings are printed are **`osd_deep_scrub_large_omap_object_key_threshold`** and **`osd_deep_scrub_large_omap_object_value_sum_threshold`**.

The `filestore_merge_threshold` option default has changed

Subdirectory merging has been disabled by default. The default value of the **`filestore_merge_threshold`** option has changed to -10 from 10. It has been observed to improve performance significantly on larger systems with a minimal performance impact to smaller systems. To take advantage of this performance increase set the **`expected-num-objects`** value when creating new data pools. See the Object Gateway for Production Guide for more information.

CHAPTER 4. BUG FIXES

This section describes bugs fixed in this release of Red Hat Ceph Storage that have significant impact on users. In addition, it includes descriptions of fixed known issues from previous versions.

4.1. CEPH ANSIBLE

Containerized OSDs start after reboot

Previously, in a containerized environment, after rebooting Ceph storage nodes some OSDs might not have started. This was due to a race condition. The race condition was resolved and now all OSD nodes start properly after a reboot.

(BZ#1486830)

Ceph Ansible no longer overwrites existing OSD partitions

On a OSD node reboot, it is possible that disk devices will get a different device path. For example, prior to restarting the OSD node, **`/dev/sda`** was an OSD, but after a reboot, the same OSD is now **`/dev/sdb`**. Previously, if no "ceph" partition was found on the disk, it was a valid OSD disk. With this release, if any partition is found on the disk, then the disk will not be used as an OSD.

(BZ#1498303)

Ansible no longer creates unused `systemd` unit files

Previously, when installing the Ceph Object Gateway by using the **`ceph-ansible`** utility, **`ceph-ansible`** created **`systemd`** unit files for the Ceph Object Gateway host corresponding to all Object Gateway instances located on other hosts. However, the only unit file that was active was the one that corresponded to the hostname of the Ceph Object Gateway. The others were not active and as such they did not cause problems. With this update of Ceph the other unit files are no longer created.

(BZ#1508460)

The OpenStack keys are copied to all Ceph Monitors

When Red Hat Ceph Storage was configured with **`run_once: true`** and **`inventory_hostname == groups.get(client_group_name) | first`** it can cause a bug when the only node being run is not the first node in the group. In a deployment with a single client node the keyrings will not be created since the task can be skipped. With this release this situation no longer occurs and all the OpenStack keys are copied to the monitor nodes.

(BZ#1588093)

The ceph-ansible utility removes the ceph-create-keys container from the same node where it was created.

Previously, the **`ceph-ansible`** utility did not always remove the **`ceph-create-keys`** container from the same node where it was created. Because of this, the deployment could fail with the message "Error response from daemon: No such container: ceph-create-keys." With this update to Red Hat Ceph Storage, **`ceph-ansible`** only tries to remove the container from the node where it was actually created, thus avoiding the error and not causing the deployment to fail.

(BZ#1590746)

Upgrading Red Hat Ceph Storage 2 to version 3 will set the`sortbitwise` option properly

Previously, a rolling upgrade from Red Hat Ceph Storage 2 to Red Hat Ceph Storage 3 would fail because the OSDs would never initialize. This is because **sortbitwise** was not properly set by Ceph Ansible. With this release, Ceph Ansible sets **sortbitwise** properly, so the ODSs can start.

(BZ#1600943)

Ceph ceph-ansible now installs the gwcli command during iscsi-gw install

Previously, when using Ansible playbooks from **ceph-ansible** to configure an iSCSI target, the **gwcli** command needed to verify the installation was not available. This was because the **ceph-iscsi-cli** package, which provides the **gwcli** command, was not included as a part of the install for the Ansible playbooks. With this update to Red Hat Ceph Storage, the Ansible playbooks now install the **ceph-iscsi-cli** package as a part of iSCSI target configuration.

(BZ#1602785)

Setting the `mon_use_fqdn` or the `mds_use_fqdn` options to `true` fails the Ceph Ansible playbook

Starting with Red Hat Ceph Storage 3.1, Red Hat no longer supports deployments with fully qualified domain names. If either the **mon_use_fqdn** or **mds_use_fqdn** options are set to **true**, then the Ceph Ansible playbook will fail. If the storage cluster is already configured with fully qualified domain names, then you must set the **use_fqdn_yes_i_am_sure** option to **true** in the **group_vars/all.yml** file.

(BZ#1613155)

Containerized OSDs for which `osd_auto_discovery` flag was set to `true` properly restart during a rolling update

Previously, when using the Ansible rolling update playbook in a containerized environment, OSDs for which **osd_auto_discovery** flag is set to **true** are not restarted and the OSD services run with old image. With this release, the OSDs are restarting as expected.

(BZ#1613626)

Ceph installation no longer fails when trying to deploy the Object Gateway

When deploying the Ceph Object Gateway using Ansible, the **rgw_hostname** variable was not being set on the Object Gateway node, but was incorrectly set on the Ceph Monitor node. In this release, the **rgw_hostname** variable is set properly and applied to the Ceph Object Gateway node.

(BZ#1618678)

Installing the Object Gateway no longer fails for container deployments

When installing the Object Gateway into a container the following error was observed:

```
fatal: [aio1_ceph-rgw_container-fc588f0a]: FAILED! => {"changed": false,
"cmd": "ceph --cluster ceph -s -f json", "msg": "[Errno 2] No such file
or directory"
```

An execution task failed because there was no **ceph-common** package installed. This Ansible task was delegated to a Ceph Monitor node, which allows the execution to happen in the correct order.

(BZ#1619098)

RADOS index object creation no longer assumes rados command available on the baremetal

Previously, the creation of the rados index object in **ceph-ansible** assumed the **rados** command was available on the bare metal node, but that is not always true when deploying in containers. This can cause the task which starts NFS to fail because the **rados** command is missing on the host. With this update to Red Hat Ceph Storage the Ansible playbook runs **rados** commands from the Ceph container instead during containerized deployment.

(BZ#1624417)

4.2. CEPH DASHBOARD

The *Ceph-pools* Dashboard no longer displays previously deleted pools

Previously in the *Red Hat Ceph Storage Dashboard*, the *Ceph-pools* Dashboard continued to reflect pools which were deleted from the Ceph Storage Cluster. With this update to Ceph they are no longer shown after being deleted.

(BZ#1537035)

Installation of Red Hat Ceph Storage Dashboard with a non-default password no longer fails

Previously, the Red Hat Storage Dashboard (cephmetrics) could only be deployed with the default password. To use a different password it had to be changed in the Web UI afterwards.

With this update to Red Hat Ceph Storage you can now set the Red Hat Ceph Storage Dashboard admin username and password using Ansible variables **grafana.admin_user** and **grafana.admin_password**.

For an example of how to set these variables, see the **group_vars/all.yml.sample** file.

(BZ#1537390)

OSD ids in 'Filestore OSD latencies' are no longer repeated

Previously, after rebooting OSDs, on the Red Hat Storage Dashboard page *Ceph OSD Information* the OSD IDs were repeated in the section *Filestore OSD Latencies*.

With this update to Red Hat Ceph Storage the OSD IDs are no longer repeated on reboot of an OSD node in the *Ceph OSD Information* dashboard. This was fixed as a part of a redesign of the underlying data reporting.

(BZ#1537505)

4.3. CEPH-DISK UTILITY

The ceph-disk utility defaults to BlueStore and when replacing an OSD, passing--filestore option is required

Previously, the **ceph-disk** utility used BlueStore as the default object store when creating OSDs. If the **--filestore** option was not used, then this caused problems in storage clusters using FileStore. In this release, the **ceph-disk** utility now defaults to FileStore as it had originally.

(BZ#1572722)

4.4. CEPHFS

Load on MDS daemons is not always balanced fairly or evenly in multiple active MDS configurations

Previously, in certain cases, the MDS balancers offloaded too much metadata to another active daemon, or none at all.

As of this update to Red Hat Ceph Storage this is no longer an issue as several balancer fixes and optimization have been made which address the issue.

(BZ#1494256)

MDS no longer asserts while in starting/resolve state

Previously, when increasing "max_mds" from "1" to "2", if the Metadata Server (MDS) daemon was in the starting/resolve state for a long period of time, then restarting the MDS daemon led to an assert. This caused the Ceph File System (CephFS) to enter a degraded state. With this update to Red Hat Ceph Storage, the underlying issue has been fixed, and increasing "max_mds" no longer causes CephFS to enter a degraded state.

(BZ#1578142)

Client I/O sometimes fails for CephFS FUSE clients

Client I/O sometimes failed for Ceph File System (CephFS) as a File System in User Space (FUSE) client with the error **`transport endpoint shutdown`** due to an assert in the FUSE service. With this update to Red Hat Ceph Storage, the issue is resolved.

(BZ#1585029)

4.5. CEPH MANAGER PLUGINS

The fixes for pg_num/pgp_num setting through the RESTful API

Previously, attempts to change **`pgp_num`** or **`pg_num`** via the RESTful API plugin failed. With this update to Red Hat Ceph Storage, the API is able to change the **`pgp_num`** and **`pg_num`** parameter successfully.

(BZ#1506102)

4.6. `CEPH-VOLUME` UTILITY

The SELinux context is set correctly when using `ceph-volume` for new filesystems

The **`ceph-volume`** utility was not labeling newly created filesystems, which was causing **`AVC`** denial messages in the **`/var/log/audit/audit.log`** file. In this release, the **`ceph-volume`** utility sets the proper SELinux context (**`ceph_var_lib_t`**), on the OSD filesystem.

(BZ#1609427)

4.7. CONTAINERS

The containerized Object Gateway daemon will read options from the Ceph configuration file now

When launching the Object Gateway daemon in a container, the daemon would override any

rgw_frontends options. This made it impossible to add extra options, such as, the **radosgw_civetweb_num_threads** option. In this release, the Object Gateway daemon will read options found in the Ceph configuration file, by default, **/etc/ceph/ceph.conf**.

(BZ#1582411)

A dmcrypt OSD comes up after upgrading a containerized Red Hat Ceph Storage cluster to 3.x

Previously, on FileStore, **ceph-disk** created the lockbox partition for **dmcrypt** on partition number 3. With the introduction of BlueStore, this partition is now on position number 5, but **ceph-disk** was trying to create the partition on position number 3 causing the OSD to fail. In this release, **ceph-disk** can now detect the correct partition to use for the lockbox partition.

(BZ#1609007)

4.8. ISCSI GATEWAY

LUN resize on target side Ceph is now reflected on clients

Previously, when using the iSCSI gateway, resized Logical Unit Numbers (LUNs) were not immediately visible to initiators. This required a work around of restarting the iSCSI gateway after resizing a LUN to expose it to the initiators.

With this update to Red Hat Ceph Storage, iSCSI initiators can now see a resized LUN immediately after rescan.

(BZ#1492342)

The iSCSI gateway supports custom cluster names

Previously, the Ceph iSCSI gateway only worked with the default storage cluster name (**ceph**). In this release, the **rbd-target-gw** now supports arbitrary Ceph configuration file locations, which allows the use of storage clusters not named **ceph**.

The Ceph iSCSI gateway can be deployed using Ceph Ansible or using the command-line interface with a custom cluster name.

(BZ#1502021)

Pools and images with hyphens ('-') are no longer rejected by the API

Previously, the iSCSI **gwcli** utility did not support hyphens in pool or image names. As such it was not possible to create a disk using a pool or image name that included hyphens ("-") by using the iSCSI **gwcli** utility.

With this update to Red Hat Ceph Storage, the iSCSI **gwcli** utility correctly handles hyphens. As such creating a disk using a pool or image name with hyphens is now supported.

(BZ#1508451)

4.9. OBJECT GATEWAY

Quota stats cache is no longer invalid

Previously in Red Hat Ceph Storage, quota values sometimes were not properly decremented. This could cause exceed errors when the quota was not actually exceeded.

With this update to Ceph, quota values are properly decremented and no incorrect errors are printed.

(BZ#1472868)

Object compression works properly

Previously, when using zlib compression with Object Gateway, objects were not being compressed properly. The actual size and used size were listed as the same despite log messages saying compression was in use. This was due to the usage of smaller buffers. With this update to Red Hat Ceph Storage, larger buffers are used and compression works as expected.

(BZ#1501380)

Marker objects no longer appear twice when listing objects

Previously, due to an error in processing, "marker" objects that were used to continue multi-segment listings were included incorrectly in the listing result. Consequently, such objects appeared twice in the listing output. With this update to Red Hat Ceph Storage, objects are only listed once, as expected.

(BZ#1504291)

Resharding a bucket that has ACLs set no longer alters the bucket ACL

Previously, in the Ceph Object Gateway (RGW), resharding a bucket with an access control list (ACL) set alters the bucket ACL. With this update to Red Hat Ceph Storage, ACLs on a bucket are preserved even if they are resharded.

(BZ#1536795)

Intermittent HTTP error code 409 no longer occurs with compression enabled

Previously, HTTP error codes could be encountered due to EEXIST being incorrectly handled in **`RGWPutObj::execute()`** in a special case. This caused the PUT operation to be incorrectly failed to the client, when it should have been retried. In this update to Red Hat Ceph Storage, the EEXIST condition handling has been corrected and this issue no longer occurs.

(BZ#1537737)

RGW no longer spikes to 100% CPU usage with no op traffic

Previously in certain situations an infinite loop could be encountered in **`rgw_get_system_obj()`**. This could cause spikes in CPU usage. With this update to Red Hat Ceph Storage this specific issue has been resolved.

(BZ#1560101)

Cache entries now refresh 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. To resolve this issue, in Red Hat Ceph Storage 3.1, logic has been added to update the expiration time of cached entries when updated.

(BZ#1585750)

Ceph is now able to delete/remove swift ACLs

Previously, the Swift CLI client could be used to set, but not to delete ACLs because the Swift header parsing logic could not detect ACL delete requests. With this update to Red Hat Ceph Storage, the header parsing logic has been fixed, and users can delete ACLs with the Swift client.

(BZ#1602882)

4.10. OBJECT GATEWAY MULTISITE

Some versioned objects do not sync when uploaded with 's3cmd sync'

Operations like **`PutACL`** that only modify object metadata do not generate a LINK_OLH entry in the bucket index log. When processed by multisite sync, these operations were skipped with the message **`versioned object will be synced on link_olh`**. Because of sync squashing, this caused the original LINK_OLH operation to be skipped as well, preventing the object version from syncing at all. With this update to Red Hat Ceph Storage this issue no longer occurs.

(BZ#1585239)

4.11. RADOS

The Ceph OSD daemon segfaults with `in thread 7f02ae07d700 thread_name:safe_timer`

Previously, a subtle race condition in the **`ceph-osd`** daemon can lead to the corruption of the **`osd_health_metrics`** data structure which results in corrupted data being sent to, and reported by, Ceph manager. This ultimately caused a segmentation fault. With this update to Red Hat Ceph Storage, a lock is now acquired before modifying the **`osd_health_metrics`** data structure.

(BZ#1580300)

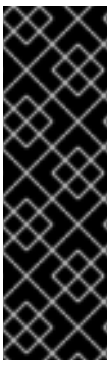
Reduced OSD memory usage

Buffers from client operations were not being rebuilt, which was leading to unnecessary memory growth by an OSD process. Rebuilding the buffers has reduced the memory footprint for OSDs in Object Gateway workloads.

(BZ#1599859)

CHAPTER 5. 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/.

OSD BlueStore

BlueStore is a new back end for the OSD daemons that allows for storing objects directly on the block devices. Because BlueStore does not need any file system interface, it improves performance of Ceph Storage Clusters.

To learn more about the BlueStore OSD back end, see the OSD BlueStore (Technology Preview) chapter in the Administration Guide.

Support for RBD mirroring to multiple secondary clusters

Mirroring RADOS Block Devices (RBD) from one primary cluster to multiple secondary clusters is now supported as a technology preview.

Erasure Coding for Ceph Block Devices

Erasure coding for Ceph Block Devices is now supported as a Technology Preview. For details, see the Erasure Coding with Overwrites (Technology Preview) section in the Storage Strategies Guide for Red Hat Ceph Storage 3.

CHAPTER 6. KNOWN ISSUES

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

6.1. CEPH ANSIBLE

The `shrink-osd.yml` playbook currently has no support for removing OSDs created by `ceph-volume`

The **`shrink-osd.yml`** playbook assumes all OSDs are created by **`ceph-disk`**. As a result, OSDs deployed using **`ceph-volume`** cannot be shrunk.

As a workaround, OSDs deployed using **`ceph-volume`** can be removed manually.

(BZ#1569413)

The container does not restart on option changes

When changing an option, for example, **`ceph_osd_docker_memory_limit`**, the change will not trigger a restart of the container.

To work around this issue restart the container manually.

(BZ#1596061)

Purging the cluster will try to unmount a partition from /var/lib/ceph

If you mount a partition to /var/lib/ceph, running the purge playbook will cause a failure when it tries to unmount it.

To work around this issue, do not mount a partition to /var/lib/ceph.

(BZ#1615872)

When putting a dedicated journal on an NVMe device installation can fail

If **`dedicated_devices`** contains an NVMe device and it has partitions or signatures on it Ansible installation might fail with an error like the following:

```
journal check: ondisk fsid 00000000-0000-0000-0000-000000000000 doesn't
match expected c325f439-6849-47ef-ac43-439d9909d391, invalid (someone
else's?) journal
```

To work around this issue ensure there are no partitions or signatures on the NVMe device.

(BZ#1619090)

Running the Ansible playbook, `purge-iscsi-gateways.yml` does not stop and disable the iSCSI gateway services

When purging the Ceph iSCSI gateways using Ceph Ansible, the iSCSI gateway services are still running. You must manually stop and disable these services by doing the following as **`root`**:

```
systemctl stop rbd-target-api
systemctl stop rbd-target-rbd
systemctl stop tcmu-runner
```

```
systemctl disable rbd-target-api
systemctl disable rbd-target-rbd
systemctl disable tcmu-runner
```

If you are using the **gwcli** command to manage the iSCSI gateways, then do not stop or disable these services.

(BZ#1621255)

6.2. CEPH DASHBOARD

The 'iSCSI Overview' page does not disply correctly

When using the Red Hat Ceph Storage Dashboard, the 'iSCSI Overview' page does not display any graphs or values as it is expected to.

(BZ#1595288)

Ceph OSD encryption summary is not displayed in the Red Hat Ceph Storage Dashboard

On the *Ceph OSD Information* dashboard, under the *OSD Summary* panel, the *OSD Encryption Summary* information is not displayed. Currently, there is no work around for this issue.

(BZ#1605241)

The Prometheus `node-exporter` service is not removed after doing a purge

When doing a purge of the Red Hat Ceph Storage Dashboard, the **node-exporter** service is not removed, and is still running. To work around this issue, you must manually stop and remove the **node-exporter** service.

Do the following as **root**:

```
# systemctl stop prometheus-node-exporter
# systemctl disable prometheus-node-exporter
# rpm -e prometheus-node-exporter
# reboot
```

For Ceph Monitor, OSD, Object Gateway, MDS, and Dashboard nodes, reboot these one at a time.

(BZ#1609713)

The OSD node details are not displayed in the *Host OSD Breakdown* panel

In the Red Hat Ceph Storage Dashboard, the *Host OSD Breakdown* information is not displayed on the *OSD Node Detail* panel under *All*.

(BZ#1610876)

Red Hat Ceph Storage Dashboard does not reflect correct OSDs

Currently, in the *Ceph Cluster* dashboard in some situations the *Cluster Configuration* tab can show the wrong number of OSDs. To work around this issue open the *Ceph OSD Information* dashboard and view the *OSD Summary* tab for the correct number of OSDs.

(BZ#1627725)

6.3. CEPH-VOLUME UTILITY

Using custom storage cluster names fails to start OSDs

When using a custom storage cluster name other than **ceph**, the OSDs might not start after a reboot.

To work around this issue, either do not use custom names when creating a new storage cluster, or create a symbolic link with the same name as the default configuration file name (**/etc/ceph/ceph.conf**) pointing to the custom named configuration file:

```
# mv /etc/ceph/ceph.conf /etc/ceph/ceph.conf.backup
# ln -s /etc/ceph/<custom-name>.conf /etc/ceph/ceph.conf
```

As a result, the OSDs will start properly.

(BZ#1621901)

6.4. ISCSI GATEWAY

Using Ceph Ansible to deploy the iSCSI gateway does not allow the user to adjust the max_data_area_mb option

Setting the **max_data_area_mb** option with Ceph Ansible will set a default value of 8 MB. To adjust this value, you must set it manually using the **gwcli** command. See the Red Hat Ceph Storage Block Device Guide for details on setting the **max_data_area_mb** option.

(BZ#1613826)

An iSCSI device is busy according to the systemd-udevd service

In Red Hat Enterprise Linux 7.5, the kernel's ALUA layer reduced the number of times an initiator retries the SCSI sense code **ALUA State Transition**. This code is returned from the target side by the **tcmu-runner** service when taking the RBD exclusive lock during a failover or failback scenario and when doing a device discovery. As a consequence, the maximum number of retries occurs before the discovery process has completed, and the SCSI layer will return a failure to the multipath IO layer. The multipath IO layer will try the next available path, and the same problem will occur. This causes a loop of path checking, resulting in failed IO, and management operations to the multipath device to fail. The logs on the initiator node will print messages about devices being removed and then re-added. To workaround this issued, downgrade the initiator's kernel to Red Hat Enterprise Linux 7.4.

(BZ#1623601)

Rebooting an iSCSI initiator with connected devices leads to an error

During device and path setup, the initiator will send commands to all paths at the same time. This will cause the Ceph iSCSI gateways to take the RBD lock from one device and set it on another device. In some cases the iSCSI gateway will interpret the lock being taken away in this manner, as a hard error and escalate its error handler by dropping the iSCSI connection, reopening the RBD devices to clear old states, and then enabling the iSCSI target port group to allow a new iSCSI connection. When disabling and enabling the iSCSI target port group this will cause a disruption to the device and path discovery. In turn, this will cause the multipath IO layer to continually disable and enable all paths and IO is suspended, or device and path discovery can fail and the device is not setup. Currently, there is no workaround for this issue.

(BZ#1623650)

6.5. OBJECT GATEWAY

The Ceph Object Gateway requires applications to write sequentially

The Ceph Object Gateway requires applications to write sequentially from offset 0 to the end of a file. Attempting to write out of order causes the upload operation to fail. To work around this issue, use utilities like **cp**, **cat**, or **rsync** when copying files into NFS space. Always mount with the **sync** option.

(BZ#1492589)

RGW garbage collection fails to keep pace during evenly balanced delete-write workloads

In testing during an evenly balanced delete-write (50% / 50%) workload the cluster fills completely in eleven hours. Object Gateway garbage collection fails to keep pace. This causes the cluster to fill completely and the status switches to HEALTH_ERR state. Aggressive settings for the new parallel/async garbage collection tunables did significantly delay the onset of cluster fill in testing, and can be helpful for many workloads. Typical real world cluster workloads are not likely to cause a cluster fill due primarily to garbage collection.

(BZ#1595833)

RGW garbage collection decreases client performance by up to 50% during mixed workload

In testing during a mixed workload of 60% reads, 16% writes, 14% deletes, and 10% lists, at 18 hours into the testing run, client throughput and bandwidth drop to half their earlier levels.

(BZ#1596401)

Large objects handled incorrectly on versioned swift containers

During uploads of large objects to versioned swift containers, please use the option **--leave-segments** in the upload using **python-swiftclient**. Not using this option will lead to an overwrite of the manifest file in which case an existing object is overwritten, leading to data loss.

(BZ#1601876)

6.6. RADOS

High object counts can degrade IO performance

The overhead with directory merging on FileStore can degrade the client's IO performance for pools with high object counts.

To work around this issue, use the 'expected_num_objects' option during pool creation. Creating pools is described in the Red Hat Ceph Storage Object Gateway for Production Guide.

(BZ#1592497)

When two or more RADOS Gateway daemons have the same name in a cluster Ceph Manager can crash

Currently, Ceph Manager can crash if some RADOS Gateway daemons have the same name. The following assert will be generated in this case:

```
DaemonPerfCounters::update(MMgrReport*)
```

To work around this issue, rename all the RADOS Gateway daemons that have the same name with new unique names.

(BZ#1634964)

CHAPTER 7. SOURCES

The updated Red Hat Ceph Storage source code 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/