Red Hat Ceph Storage 4.1

Release Notes

Release notes for Red Hat Ceph Storage 4.1z3

Last Updated: 2023-10-16

Red Hat Ceph Storage 4.1 Release Notes

Release notes for Red Hat Ceph Storage 4.1z3

Legal Notice

Abstract

The Release Notes describes the major features, enhancements, known issues, and bug fixes implemented in Red Hat Ceph Storage product, which includes previous notes of the Red Hat Ceph Storage 4.1 releases up to the current release.

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.

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 4.1 contains many contributions from the Red Hat Ceph Storage team. In addition, 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
- SUSE

CHAPTER 3. NEW FEATURES

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

The main features added by this release are:

- Vault support
- Support for co-locating the Grafana container alongside the OSD container
- Ability to clone subvolumes from snapshots in CephFS

Vault support

With the Red Hat Ceph Storage 4.1 release, the Ceph object storage gateway (RGW) can now interoperate with the Hashicorp Vault secure key management service. As a storage administrator, you can securely store keys, passwords and certificates in the HashiCorp Vault for use with the Ceph Object Gateway. The HashiCorp Vault provides a secure key management service for server-side encryption used by the Ceph Object Gateway.

For more information, see the *The HashiCorp Vault* section in the *Red Hat Ceph Storage Object Gateway Configuration and Administration Guide*.

Support for co-locating the Grafana container

The Red Hat Ceph Storage 4.1 release supports the co-location of the Grafana container with Ceph OSD and an additional scale-out daemon, cardinality 2. Cardinality 2 was previously only supported with the Ceph Object Gateway in the Red Hat Ceph Storage 4.0 release.

For more information, see the *Red Hat Ceph Supported Configurations* article.

Cloning subvolumes from snapshots in CephFS

Subvolumes can be created by cloning subvolume snapshots. It is an asynchronous operation involving copying data from a snapshot to a subvolume.

For information about cloning subvolumes from snapshots in CephFS, see the *Red Hat Ceph Storage File System Guide*.

3.1. THE CEPH ANSIBLE UTILITY

ceph-ansible now supports multisite deployments with multiple realms

Previously, **ceph-ansible** multisite deployments supported a single RGW realm. With this update, **ceph-ansible** now supports multiple realms with their associated zones, zonegroups, and endpoints.

For more information, see *Configuring multisite Ceph Object Gateways* in the *Red Hat Ceph Storage Installation Guide*.

The dedicated journal devices retain their configuration when migrating from Filestore OSD to Bluestore

Previously, dedicated journal devices for Filestore OSD could not be reused when migrating to Bluestore OSD DB. An example of a dedicated device configuration is using a HDD for data and an SSD for journaling.

With this update, dedicated journal devices retain their configuration during the migration, so that they can be reused with the Bluestore OSD DB.

For more information, see How to Migrate the Object Store from FileStore to BlueStore in the Administration Guide.

purge-container-cluster.yml playbook now supports clusters with three-digit IDs

Previously, **purge-container-cluster** only supported Red Hat Ceph Storage clusters with up to 99 OSDs. This is because the playbook supported ceph-osd services with two-digit indices. With this update, you can properly purge clusters with three-digit IDs.

OpenStack users can deploy Ceph Dashboard with a default admin account with read-only privileges

Previously, changes made from Ceph Dashboard by OpenStack users with full admin privileges could override cluster settings or status. With this feature, Ceph Dashboard admin account can only monitor Ceph cluster status and retrieve information and settings.

Added support for logrotate in containerized Red Hat Ceph Storage deployments

With this release, containerized Ceph logs are rotated using the **logrotate** program. This can help prevent the file system from filling up with log data.

3.2. CEPH FILE SYSTEM

Independent life-cycle operations for subvolumes and subvolume snapshots

Because the CSI protocol treats snapshots as first class objects, this requires source subvolumes and subvolume snapshots to operate independently of each other. Since the Kubernetes storage interface uses the CSI protocol, subvolume removal with a snapshot retention option (**--retain-snapshots**) has been implemented. This allows other life-cycle operations on a retained snapshot to proceed appropriately.

3.3. THE CEPH VOLUME UTILITY

Ceph OSD encryption support when using ceph-volume in raw mode

With this release, the **ceph-volume** command can prepare Ceph OSDs for encryption using **raw** mode.

3.4. DISTRIBUTION

An S3 client is included in the Ceph tools repository

Starting with this release, an S3 command-line client, **s3cmd** is included in the Red Hat Ceph Storage 4 Tools software repository. To install the S3 command-line client package, enable the **rhceph-4-tools-for-rhel-8-x86_64-rpms** repository first.

3.5. CEPH OBJECT GATEWAY

Support for Amazon S3 resources in Ceph Object Gateway

AWS provides the Secure Token Service (STS) to allow secure federation with existing OpenID Connect/ OAuth2.0 compliant identity services such as Keycloak. STS is a standalone REST service that provides temporary tokens for an application or user to access a Simple Storage Service (S3) endpoint after the user authenticates against an identity provider (IDP).

Previously, users without permanent Amazon Web Services (AWS) credentials could not access S3 resources through Ceph Object Gateway. With this update, Ceph Object Gateway supports STS AssumeRoleWithWebIdentity. This service allows web application users who have been authenticated with an OpenID Connect/OAuth 2.0 compliant IDP to access S3 resources through Ceph Object Gateway.

For more information, see Secure Token Service in the Developer Guide.

AWS S3 ListObjects v2 operation provides an improved mechanism to list the objects in the S3 bucket without additional configuration

Previously, S3 protocol clients, like S3A and the awscli command-line tool, had to be configured with the older ListObjects method. With this feature, AWS S3 ListObjects v2 operation is implemented, that provides an improved mechanism to list objects in an S3 bucket.

Ceph Object Gateway's default bucket-index shards increased to 11

The default number of bucket-index shards for new buckets has been increased, from 1 to 11. This increases the amount of write throughput for small buckets and delays the onset of dynamic resharding. This change only affects new deployments and zones.

For existing deployments, you can change this default value using the **radosgw-admin zonegroup modify --bucket-index-max-shards=11** command. If the zonegroup is part of a realm, the change must be committed with **radosgw-admin period update --commit** command. If a commit is not done, then the change will take effect until after the Ceph Object Gateways are restarted.

NOTE

After an upgrade, the new default value has to be increased manually, but on new Red Hat Ceph Storage 4.1 deployments the new default value is set automatically.

The Ceph Object Gateway log includes **access log** for Beast

With this release, Beast, the front-end web server, now includes an Apache-style **access log** line in the Ceph Object Gateway log. This update to the log helps diagnose connection and client network issues.

The minimum value of a session token's expiration is configurable

The **rgw_sts_min_session_duration** option can now have a value lower than the default value of 900 seconds.

Listing the contents of large buckets

With this release, the ordered listing when delimited, and possible prefix, are specified more efficiently by skipping over object in lower pseudo-directories. This allows fewer interactions between the Ceph client and the Ceph Object Gateway, and also between the Ceph Object Gateway and the Ceph OSDs. This enhancement, along with configuration changes to HA proxy allows the listing of contents in large buckets.

The Ceph Object Gateway log includes the access log for Beast

With this release, Beast, the front-end web server, now includes an Apache-style access log line in the Ceph Object Gateway log. This update to the log helps diagnose connection and client network issues.

3.6. RADOS

Update to use ping times to track network performance

Previously, when network problems occur, it was difficult to distinguish from other performance issues. With this release, a heath warning is generated if the average Red Hat Ceph Storage OSD heartbeat exceeds a configurable threshold for any computed intervals. The Red Hat Ceph Storage OSD computes 1 minute,5 minute and 15 minute intervals with the average, minimum and maximum values.

BlueStore compression stats added to the dashboard

With this release, compression related performance metrics for BlueStore OSDs will now be visible in the dashboard.

For more information about the dashboard, see the Dashboard Guide.

The storage cluster status changes when a Ceph OSD encounters an I/O error

With this release, the Ceph Monitor now has a **mon_osd_warn_num_repaired** option, which is set to **10** by default. If any Ceph OSD has repaired more than this many I/O errors in stored data, a **OSD_TOO_MANY_REPAIRS** health warning status is generated. To clear this warning, the new **clear_shards_repaired** option has been added to the **ceph tell** command. For example:

```
ceph tell osd.NUMBER clear_shards_repaired [COUNT]
```

By default, the **clear_shards_repaired** option sets the repair count to **0**. To be warned again if additional Ceph OSD repairs are performed, you can specify the value of the **mon_osd_warn_num_repaired** option.

Update to the heartbeat grace period

Previously, when there were no Ceph OSD failures for more than 48 hours, there was no mechanism to reset the grace timer back to the default value. With this release, the heartbeat grace timer is reset to the default value of 20 seconds, if there have been no failures on a Ceph OSD for 48 hours. When the failure interval between the last failure and the latest failure exceeds 48 hours, the grace timer is reset to the default value of 20 seconds.

The grace time is the interval in which a Ceph storage cluster considers a Ceph OSD as down by the absence of a heartbeat. The grace time is scaled based on lag estimations or on how frequently a Ceph ODS is experiencing failures.

The osd_client_message_cap option has been added back

Previously, the **osd_client_message_cap** option was removed, but with this release, the **osd_client_message_cap** option has been re-introduced. This option helps control the maximum number of in-flight client requests by throttling those requests. Doing this can be helpful when a Ceph OSD flaps due to overwhelming amount of client-based traffic.

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/.

4.1. BLOCK DEVICES (RBD)

Mapping RBD images to NBD images

The **rbd-nbd** utility maps RADOS Block Device (RBD) images to Network Block Devices (NBD) and enables Ceph clients to access volumes and images in Kubernetes environments. To use **rbd-nbd**, install the **rbd-nbd** package. For details, see the **rbd-nbd(7)** manual page.

4.2. OBJECT GATEWAY

Object Gateway archive site

With this release an archive site is supported as a Technology Preview. The archive site allows you to have a history of versions of S3 objects that can only be eliminated through the gateways associated with the archive zone. Including an archive zone in a multizone configuration allows you to have the flexibility of an S3 object history in only one zone while saving the space that the replicas of the versions S3 objects would consume in the rest of the zones.

CHAPTER 5. 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.

Ubuntu is no longer supported

Installing a Red Hat Ceph Storage 4 cluster on Ubuntu is no longer supported. Use Red Hat Enterprise Linux as the underlying operating system.

Configuring iSCSI gateway using **ceph-ansible** is no longer supported

Configuring the Ceph iSCSI gateway by using the **ceph-ansible** utility is no longer supported. Use **ceph-ansible** to install the gateway and then use the **gwcli** utility to configure the Ceph iSCSI gateway. For details, see the *The Ceph iSCSI Gateway* chapter in the *Red Hat Ceph Storage Block Device Guide*.

ceph-disk is deprecated

With this release, the **ceph-disk** utility is no longer supported. The **ceph-volume** utility is used instead. For details, see the *Why does **ceph-volume** replace `ceph-disk`* section in the *Administration Guide* for Red Hat Ceph Storage 4.

FileStore is no longer supported in production

The FileStore OSD back end is now deprecated because the new BlueStore back end is now fully supported in production. For details, see the *How to migrate the object store from FileStore to BlueStore* section in the *Red Hat Ceph Storage Installation Guide*.

Ceph configuration file is now deprecated

The Ceph configuration file (**ceph.conf**) is now deprecated in favor of new centralized configuration stored in Ceph Monitors. For details, see the *The Ceph configuration database* section in the *Red Hat Ceph Storage Configuration Guide*.

CHAPTER 6. BUG FIXES

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

6.1. THE CEPH ANSIBLE UTILITY

The size of the replication pool can now be modified after the Ceph cluster deployment

Previously, increasing the size of the replication pool failed after the Ceph cluster was deployed using director. This occurred because an issue with the task in charge of customizing the pool size prevented it from executing when the playbook was rerun. With this update, you can now modify pool size after cluster deployment.

(BZ#1743242)

Ceph Ansible supports multiple **grafana** instances during a Ceph dashboard deployment

Previously, in a multi-node environment, **ceph-ansible** was not able to configure multiple **grafana** instances as only one node was supported, leaving the remaining nodes unconfigured. With this update, **ceph-ansible** supports multiple instances and injects Ceph-specific layouts on all the Ceph Monitor nodes during the deployment of the Ceph Dashboard.

(BZ#1784011)

Running the Ansible **purge-cluster.yml** playbook no longer fails when the dashboard feature is disabled

Previously, using the **purge-cluster-yml** playbook to purge clusters failed when the dashboard feature was disabled with the following error message:

```
registry.redhat.io/openshift4/ose-prometheus-node-exporter:v4.1
  msg: '[Errno 2] No such file or directory'
```

This occurred because the **dashboard_enabled** variable was ignored. With this update, the **dashboard_enabled** variable is correctly handled, and **purge-cluster.yml** runs successfully.

(BZ#1785736)

Red Hat Ceph Storage installation on Red Hat OpenStack Platform no longer fails

Previously, the **ceph-ansible** utility became unresponsive when attempting to install Red Hat Ceph Storage with the Red Hat OpenStack Platform 16, and it returns an error similar to the following:

```
'Error: unable to exec into ceph-mon-dcn1-computehci1-2: no container with name or ID ceph-mon-dcn1-computehci1-2 found: no such container'
```

This occurred because **ceph-ansible** reads the value of the fact **container_exec_cmd** from the wrong node in handler_osds.yml

With this update, **ceph-ansible** reads the value of **container_exec_cmd** from the correct node, and the installation proceeds successfully.

(BZ#1792320)

Ansible unsets the **norebalance** flag after it completes

Previously, Ansible did not unset the **norebalance** flag and it had to be unset manually. With this update, the **rolling-update.yml** Ansible playbook unsets the **norebalance** flag automatically after it completes and a manual unset is not required.

(BZ#1793564)

Ansible upgrades a multisite Ceph Object Gateway when the Dashboard is enabled

Previously, when Red Hat Ceph Storage Dashboard is enabled and an attempt to use Ansible to upgrade to a later version of Red Hat Ceph Storage is made, the upgrade to the secondary Ceph Object Gateway site in a multisite setup failed. With this update to Red Hat Ceph Storage, upgrade of the secondary site works as expected.

(BZ#1794351)

Ceph Ansible works with Ansible 2.9

Previously, **ceph-ansible** versions 4.0 and above did not work with Ansible version 2.9. This occurred because the **ceph-validate** role did not allow **ceph-ansible** to be run against Ansible 2.9. With this update, **ceph-ansible** works with Ansible 2.9.

(BZ#1807085)

Ceph installations with custom software repositories no longer fail

Previously, using custom repositories to install Ceph were not allowed. This occurred because the **redhat_custom_repository.yml** file was removed. With this update, the **redhat_custom_repository.yml** file is included, and custom repositories can be used to install Red Hat Ceph Storage.

> NOTE
>
> Only Red Hat-signed packages can use custom software repositories to install Ceph. Custom third-party software repositories are not supported.

(BZ#1808362)

The **ceph-ansible** purge playbook does not fail if dashboard was not installed

Previously, when the dashboard was not deployed, the purge playbook failed when purging the cluster because it tried to remove dashboard related resources that did not exist. Consequently, the purge playbook stated that the dashboard is deployed, and purge failed. With this update, **ceph-ansible** does not purge dashboard related resources if not part of the deployment, and purge completes successfully.

(BZ#1808988)

Using a standalone **nfs-ganesha** daemon with an external Ceph storage cluster no longer fails to copy the keyring during deployment

Previously, in configurations consisting of a standalone **nfs-ganesha** daemon and an external Ceph storage cluster, the Ceph keyring was not copied to /etc/ceph during a Ceph Ansible deployment.

With this update, the Ceph keyring is copied to /etc/ceph/ directory.

(BZ#1814942)

Ceph Ansible updates the privileges of the dashboard admin user after initial install

Previously, **ceph-ansible** could only set the privileges of the dashboard user when it was first created. Running the playbooks after changing **dashboard_admin_user_ro: false** from its original setting during install would not update the privileges of the user. In Red Hat Ceph Storage 4.1z1 **ceph-ansible** has been updated to support changing the dashboard user privileges on successive runs of the playbooks.

(BZ#1826002)

The docker-to-podman.yml playbook now migrates dashboard containers

Previously, running the **docker-to-podman.yml** playbook migrated all the daemons from **docker** to **podman**, except for **grafana-server** and the dashboard containers. With this release, running **docker-to-podman.yml** successfully migrates all of the daemons.

(BZ#1829389)

Storage directories from old containers are removed

Previously, storage directories for old containers were not removed. This could cause high disk usage. This could be seen if you installed Red Hat Ceph Storage, purged it, and reinstalled it. In Red Hat Ceph Storage 4.1z1, storage directories for containers that are no longer being used are removed and excessive disk usage does not occur.

(BZ#1834974)

Upgrading a containerized cluster from 4.0 to 4.1 on Red Hat Enterprise Linux 8.1 no longer fails

Previously, when upgrading a Red Hat Ceph Storage cluster from 4.0 to 4.1 the upgrade could fail with an error on **set_fact ceph_osd_image_repodigest_before_pulling**. Due to an issue with how the container image tag was updated, **ceph-ansible** could fail. In Red Hat Ceph Storage 4.1z1 **ceph-ansible** has been updated so it no longer fails and upgrading works as expected.

(BZ#1844496)

Enabling the Ceph Dashboard fails on an existing OpenStack environment

In an existing OpenStack environment, when configuring the Ceph Dashboard's IP address and port after the Ceph Manager dashboard module was enabled, was causing a conflict with the HAProxy configuration. To avoid this conflict, configure the Ceph Dashboard's IP address and port before enabling the Ceph Manager dashboard module.

(BZ#1851455)

Red Hat Ceph Storage Dashboard fails when deploying a Ceph Object Gateway secondary site

Previously, the Red Hat Ceph Storage Dashboard would fail to deploy the secondary site in a Ceph Object Gateway multi-site deployment, because when Ceph Ansible ran the **radosgw-admin user create** command, the command would return an error. With this release, the Ceph Ansible task in the deployment process has been split into two different tasks. Doing this allows the Red Hat Ceph Storage Dashboard to deploy a Ceph Object Gateway secondary site successfully.

(BZ#1851764)

The Ceph File System Metadata Server installation fails when running a playbook with the --limit option

Some facts were not getting set on the first Ceph Monitor, but those facts were getting set on all respective Ceph Monitor nodes. When running a playbook with the **--limit** option, these facts were not set on the Ceph Monitor, if the Ceph Monitor was not part of the batch. This would cause the playbook to fail when these facts where used in a task for the Ceph Monitor. With this release, these facts are set on the Ceph Monitor whether the playbook uses the **--limit** option or not.

(BZ#1852796)

Adding a new Ceph Ojbect Gateway instance when upgrading fails

The **radosgw_frontend_port** option did not consider more than one Ceph Object Gateway instance, and configured port **8080** to all instances. With this release, the **radosgw_frontend_port** option is increased for each Ceph Object Gateway instance, allowing you to use more than one Ceph Object Gateway instance.

(BZ#1859872)

Ceph Ansible's **shrink-osd.yml** playbook fails when using FileStore in a containerized environment

A default value was missing in Ceph Ansible's **shrink-osd.yml** playbook, which was causing a failure when shrinking a FileStore-backed Ceph OSD in a containerized environment. A previously prepared Ceph OSD using **ceph-disk** and **dmcrypt**, was leaving the **encrypted** key undefined in the corresponding Ceph OSD file. With this release, a default value was added so the Ceph Ansible **shrink-osd.yml** playbook can ran on Ceph OSD that have been prepared using **dmcrypt** in containerized environments.

(BZ#1862416)

Using HTTPS breaks access to Prometheus and the alert manager

Setting the **dashboard_protocol** option to **https** was causing the Red Hat Ceph Storage Dashboard to try and access the Prometheus API, which does not support TLS natively. With this release, Prometheus and the alert manager are force to use the HTTP protocol, when setting the **dashboard_protocol** option to **https**.

(BZ#1866006)

The Ceph Ansible **shrink-osd.yml** playbook does not clean the Ceph OSD properly

The **zap** action done by the **ceph_volume** module does not handle the **osd_fsid** parameter. This caused the Ceph OSD to be improperly zapped by leaving logical volumes on the underlying devices. With this release, the **zap** action properly handles the **osd_fsid** parameter, and the Ceph OSD can be cleaned properly after shrinking.

(BZ#1873010)

The Red Hat Ceph Storage rolling update fails when multiple storage clusters exist

Running the Ceph Ansible **rolling_update.yml** playbook when multiple storage clusters are configured, would cause the rolling update to fail because a storage cluster name could not be specified. With this release, the **rolling_update.yml** playbook uses the **--cluster** option to allow for a specific storage cluster name.

(BZ#1876447)

The **hosts** field has an invalid value when doing a rolling update

A Red Hat Ceph Storage rolling update fails because the syntax changed in the evaluation of the **hosts** value in the Ceph Ansible **rolling_update.yml** playbook. With this release, a fix to the code updates the syntax properly when the **hosts** field is specified in the playbook.

(BZ#1876803)

Running the **rolling_update.yml** playbook does not retrieve the storage cluster **fsid**

When running the **rolling_update.yml** playbook, and the Ceph Ansible inventory does not have Ceph Monitor nodes defined, for example, in an external scenario, the storage cluster **fsid** is not retrieved. This causes the **rolling_update.yml** playbook to fail. With this release, the **fsid** retrieval is skipped when there are no Ceph Monitors defined in the inventory, allowing the **rolling_update.yml** playbook to execute when no Ceph Monitors are present.

(BZ#1877426)

6.2. THE COCKPIT CEPH INSTALLER

Cockpit Ceph Installer no longer deploys Civetweb instead of Beast for RADOS Gateway

Previously, the Cockpit Ceph Installer configured RADOS Gateway (RGW) to use the deprecated Civetweb frontend instead of the currently supported Beast front end. With this update to Red Hat Ceph Storage, the Cockpit Ceph Installer deploys the Beast frontend with RGW as expected.

(BZ#1806791)

The **ansible-runner-service.sh** script no longer fails due to a missing repository

Previously, the Cockpit Ceph Installer startup script could fail due to a missing repository in **/etc/containers/registries.conf**. The missing repository was **registry.redhat.io**. In Red Hat Ceph Storage 4.1z1, the **ansible-runner-service.sh** script has been updated to explicitly state the registry name so the repository does not have to be included in **/etc/containers/registries.conf**.

(BZ#1809003)

Cockpit Ceph Installer no longer fails on physical network devices with bridges

Previously, the Cockpit Ceph Installer failed if physical network devices were used in a Linux software bridge. This was due to a logic error in the code. In Red Hat Ceph Storage 4.1z1, the code has been fixed and you can use Cockpit Ceph Installer to deploy on nodes with bridges on the physical network interfaces.

(BZ#1816478)

Cluster installation no longer fails due to **cockpit-ceph-installer** not setting admin passwords for dashboard and grafana

Previously, **cockpit-ceph-installer** did not allow you to set the admin passwords for dashboard and Grafana. This caused storage cluster configuration to fail because **ceph-ansible** requires the default passwords to be changed.

With this update, **cockpit-ceph-installer** allows you to set the admin passwords in Cockpit so the storage cluster configuration can complete successfully.

(BZ#1839149)

Cockpit Ceph Installer allows RPM installation type on Red Hat Enterprise Linux 8

Previously, on Red Hat Enterprise Linux 8, the Cockpit Ceph Installer would not allow you to select RPM for Installation type, you could only install containerized. In Red Hat Ceph Storage 4.1z1, you can select RPM to install Ceph on bare-metal.

(BZ#1850814)

6.3. CEPH FILE SYSTEM

Improved Ceph File System performance as the number of subvolume snapshots increases

Previously, creating more than 400 subvolume snapshots was degrading the Ceph File System performance by slowing down file system operations. With this release, you can configure subvolumes to only support subvolume snapshots at the subvolume root directory, and you can prevent cross-subvolume links and renames. Doing this allows for the creation of higher numbers of subvolume snapshots, and does not degrade the Ceph File System performance.

(BZ#1848503)

Big-endian systems failed to decode metadata for Ceph MDS

Previously, decoding the Ceph MDS metadata on big-endian systems would fail. This was caused by Ceph MDS ignoring the endianness when decode structures from RADOS. The Ceph MDS metadata routines were fixed to correct this issue, resulting in Ceph MDS decoding the structure correctly.

(BZ#1896555)

6.4. CEPH MANAGER PLUGINS

Ceph Manager crashes when setting the alerts interval

There was a code bug in the alerts module for Ceph Manager, which was causing the Ceph Manager to crash. With this release, this code bug was fixed, and you can set the alerts interval without the Ceph Manager crashing.

(BZ#1849894)

6.5. THE CEPH VOLUME UTILITY

The ceph-volume lvm batch command fails with mixed device types

The **ceph-volume** command did not return the expected return code when devices are filtered using the **lvm batch** sub-command, and when the Ceph OSD strategy changed. This was causing **ceph-ansible** tasks to fail. With this release, the **ceph-volume** command returns the correct status code when the Ceph OSD strategy changes, allowing **ceph-ansible** to properly check if new Ceph OSDs can be added or not.

(BZ#1825113)

The ceph-volume command is treating a logical volume as a raw device

The **ceph-volume** command was treating a logical volume as a raw device, which was causing the **add-osds.yml** playbook to fail. This was not allowing additional Ceph OSD to be added to the storage cluster. With this release, a code bug was fixed in **ceph-volume** so it handles logical volumes properly, and the **add-osds.yml** playbook can be used to add Ceph OSDs to the storage cluster.

(BZ#1850955)

6.6. CONTAINERS

The **nfs-ganesha** daemon starts normally

Previously, a configuration using **nfs-ganesha** with the RADOS backend would not start because the **nfs-ganesha-rados-urls** library was missing. This occurred because the **nfs-ganesha** library package for the RADOS backend was moved to a dedicated package. With this update, the **nfs-ganesha-rados-urls** package is added to the Ceph container image, so the **nfs-ganesha** daemon starts successfully.

(BZ#1797075)

6.7. CEPH OBJECT GATEWAY

Ceph Object Gateway properly applies AWS request signing

Previously, the Ceph Object Gateway did not properly apply an AWS request for signing headers, and was generating the following error message:

```
SignatureDoesNotMatch
```

With this release, the Ceph Object Gateway code was fixed to properly sign headers. This results in the signing request to succeed when requested.

(BZ#1665683)

The **radosgw-admin bucket check** command no longer displays incomplete multipart uploads

Previously, running the **radosgw-admin bucket check** command displayed incomplete multipart uploads. This could cause confusion for a site admin because the output might have appeared as though the bucket index were damaged. With this update, the command displays only errors and orphaned objects, and the incomplete uploads are filtered out.

(BZ#1687971)

Uneven distribution of omap keys with bucket shard objects

In versioned buckets, occasionally delete object operations were unable to fully complete. In this state, the bucket index entries for these objects had their name and instance strings zeroed out. When there was a subsequent reshard, the empty name and instance strings caused the entry to be resharded to shard 0. Entries that did not belong on shard 0 ended up there. This put a disproportionate number of entries on shard 0 and was larger than each other shard. With this release, the name and instance strings are no longer cleared during this portion of the delete operation. If a reshard takes place, the entries that were not fully deleted nonetheless end up on the correct shard and are not forced to shard 0.

(BZ#1749090)

Increase in overall throughput of Object Gateway lifecycle processing performance

Previously, Object Gateway lifecycle processing performance was constrained by the lack of parallelism due to the increasing workload of objects or buckets with many buckets or containers in the given environment. With this update, parallelism is in two dimensions, a single object gateway instance can have several lifecycle processing threads, and each thread has multiple work-pool threads executing the lifecycle work. Additionally, this update improved the allocation of **shards** to workers, thereby increasing overall throughput.

(BZ#1794715)

Bucket tenanting status is interpreted correctly when **rgw_parse_bucket_key** is called

Previously, some callers of **rgw_parse_bucket_key** like **radosgw-admin bucket stats**, which processed keys in a loop, could incorrectly interpret untenanted buckets as tenanted, if some tenanted buckets were listed. If **rgw_parse_bucket_key** was called with a non-empty rgw bucket argument, it would not correctly assign an empty value for bucket::tenant when no tenant was present in the key. In Red Hat Ceph Storage 4.1z1 the bucket tenant member is now cleared if no tenant applies and bucket tenanting status is interpreted correctly.

(BZ#1830330)

The Ceph Object Gateway tries to cache and access anonymous user information

Previously, the Ceph Object Gateway tried to fetch anonymous user information for each request that has not been authenticated. This unauthenticated access was causing high load on a single Ceph OSD in the storage cluster. With this release, the Ceph Object Gateway will try not to fetch anonymous user information, resulting in a decrease in latency and load on a single Ceph OSD.

(BZ#1831865)

Lifecycle expiration is reported correctly for objects

Previously, incorrect lifecycle expiration could be reported for some objects, due to the presence of a prefix rule. This was caused because the optional prefix restriction in lifecycle expiration rules was ignored when generating expiration headers used in S3 HEAD and GET requests. In Red Hat Ceph Storage 4.1z1, the rule prefix is now part of the expiration header rule matching and lifecycle expiration for objects is reported correctly.

(BZ#1833309)

A high number of objects in the **rgw.none** bucket stats

The code that calculates stats failed to check, in some cases, whether a bucket index entry referenced an object that already existed. This was causing the bucket stats to be incorrect. With this release, code was added to check for existence, fixing the bucket stats.

(BZ#1846035)

A call to an ordered bucket listing gets stuck

A code bug in the bucket ordered list operation could cause, under specific circumstances, this operation to get stuck in a loop and never complete. With this release, this code bug was fixed, and as a result the call to an ordered bucket listing completes as expected.

(BZ#1853052)

Life-cycle processing ignores **NoncurrentDays** in **NoncurrentVersionExpiration**

A variable which is suppose to contain the modification time of objects during parallel life-cycle processing was incorrectly initialized. This caused non-current versions of objects in buckets with a non-current expiration rule to expire before their intended expiration time. With this release, the modificaction time (**mtime**) is correctly initialized and propagates to the life-cycle's processing queue. This results in the non-current expiration to happen after the correct time period.

(BZ#1875305)

Parts of some objects were erroneously added to garbage collection

When reading objects using the Ceph Object Gateway, if parts of those objects took more than half of the value, as defined by the **rgw_gc_obj_min_wait** option, then their tail object was added to the garbage collection list. Those tail objects in the garbage collection list were deleted, resulting in data loss. With this release, the garbage collection feature meant to delay garbage collection for deleted objects was disabled. As a result, reading objects using the Ceph Object Gateway that are taking a long time are not added to the garbage collection list.

(BZ#1892644)

6.8. MULTI-SITE CEPH OBJECT GATEWAY

The RGW daemon no longer crashes on shutdown

Previously, the RGW process would abort in certain circumstances due to a race condition during radosgw shutdown. One situation this was issue was seen was when deleting objects when using multisite. This was caused by dereferencing unsafe memory. In Red Hat Ceph Storage 4.1z1 unsafe memory is no longer dereferenced and the RGW daemon no longer crashes.

(BZ#1840858)

6.9. RADOS

A health warning status is reported when no Ceph Managers or OSDs are in the storage cluster

In previous Red Hat Ceph Storage releases, the storage cluster health status was **HEALTH_OK** even though there were no Ceph Managers or OSDs in the storage cluster. With this release, this health status has changed, and will report a health warning if a storage cluster is not set up with Ceph Managers, or if all the Ceph Managers go down. Because Red Hat Ceph Storage heavily relies on the Ceph Manager to deliver key features, it is not advisable to run a Ceph storage cluster without Ceph Managers or OSDs.

(BZ#1761474)

The **ceph config show** command displays the correct **fsid**

Previously, the **ceph config show** command only displayed the configuration keys present in the Ceph Monitor's database, and because the **fsid** is a **NO_MON_UPDATE** configuration value, the **fsid** was not displaying correctly. With this release, the **ceph config show** command displays the correct **fsid** value.

(BZ#1772310)

Small objects and files in RADOS no longer use more space than required

The Ceph Object Gateway and the Ceph file system (CephFS) stores small objects and files as individual objects in RADOS. Previously, objects smaller than BlueStore's default minimum allocation size (**min_alloc_size**) of 16 KB used more space than required. This happened because the earlier default value of BlueStore's **min_alloc_size** was 16 KB for solid state devices (SSDs). Currently, the default value of **min_alloc_size** for SSDs is 4 KB. This enables better use of space with no impact on performance.

(BZ#1788347)

Slow ops not being logged in cluster logs

Previously, slow ops were not being logged in cluster logs. They were logged in the **osd** or **mon** logs, but lacked the expected level of detail. With this release, slow ops are being logged in cluster logs, at a level of detail that makes the logs useful for debugging.

(BZ#1807184)

Backfills are no longer delayed during placement group merging

Previously, in Red Hat Ceph Storage placement group merges could take longer than expected if the acting set for the source and target placement groups did not match before merging. Backfills done when there is a mismatch can appear to stall. In Red Hat Ceph Storage 4.1z1 the code has been updated to only merge placement groups whose acting sets match. This change allows merges to complete without delay.

(BZ#1810949)

Ceph Monitors can grow beyond the memory target

Auto-tuning the memory target was only done on the Ceph Monitor leader and not the Ceph Monitors following the leader. This was causing the Ceph Monitor followers to exceed the set memory target, resulting in the Ceph Monitors crashing once its memory was exhausted. With this release, the auto-tuning process applies the memory target for the Ceph Monitor leader and its followers so memory is not exhausted on the system.

(BZ#1827856)

Disk space usage does not increase when OSDs are down for a long time

Previously, when an OSD was down for a long time, a large number of osdmaps were stored and not trimmed. This led to excessive disk usage. In Red Hat Ceph Storage 4.1z1, osdmaps are trimmed regardless of whether or not there are down OSDs and disk space is not overused.

(BZ#1829646)

Health metrics are correctly reported when **smartctl** exits with a non-zero error code

Previously, the **ceph device get-health-metrics** command could fail to report metrics if **smartctl** exited with a non-zero error code even though running **smartctl** directly reported the correct information. In this case a JSON error was reported instead. In Red Hat Ceph Storage 4.1z1, the **ceph device get-health-metrics** command reports metrics even if **smartctl** exits with a non-zero error code as long as **smartctl** itself reports correct information.

(BZ#1837645)

Crashing Ceph Monitors caused by a negative time span

Previously, Ceph Monitors could crash when triggered by a monotonic clock going back in time. These crashes caused a negative monotonic time span and triggered an assertion into the Ceph Monitor leading them to crash. The Ceph Monitor code was updated to tolerate this assertion and interprets it as a zero-length interval and not a negative value. As a result, the Ceph Monitor does not crash when this assertion is made.

(BZ#1847685)

Improvements to the encoding and decoding of messages on storage clusters

When deploying a Red Hat Ceph Storage cluster containing a heterogeneous architecture, such as x86_64 and s390, could cause system crashes. Also, under certain workloads for CephFS, Ceph Monitors on s390x nodes could crash unexpectedly. With this release, properly decoding **entity_addrvec_t** with a marker of **1**, properly decoding the **enum** types on big-endian systems by using an intermediate integer variable type, and fixed encoding and decoding **float** types on big-endian systems. As a result, heterogeneous storage clusters, and Ceph Monitors on s390x nodes no longer crash.

(BZ#1895040)

6.10. RADOS BLOCK DEVICES (RBD)

Multiple **rbd unmap** commands can be issued concurrently and the corresponding RBD block devices are unmapped successfully

Previously, issuing concurrent **rbd unmap** commands could result in udev-related event race conditions. The commands would sporadically fail, and the corresponding RBD block devices might remain mapped to their node. With this update, the udev-related event race conditions have been fixed, and the commands no longer fail.

(BZ#1784895)

CHAPTER 7. KNOWN ISSUES

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

7.1. THE CEPH ANSIBLE UTILITY

Deploying the placement group autoscaler does not work as expected on CephFS related pools only

To work around this issue, the placement group autoscaler can be manually enabled on CephFS related pools after the playbook has run.

(BZ#1836431)

The filestore-to-bluestore playbook does not support the`osd_auto_discovery` scenario

Red Hat Ceph Storage 4 deployments based on **osd_auto_recovery** scenario can't use the **filestore-to-bluestore** playbook to ease the **BlueStore** migration.

To work around this issue, use **shrink-osd** playbook and redeploy the shrinked OSD with **osd_objectstore: bluestore**.

(BZ#1881523)

7.2. CEPH MANAGEMENT DASHBOARD

The Dashboard does not provide correct Ceph iSCSI error messages

If the Ceph iSCSI returns an error, for example the HTTP "400" code when trying to delete an iSCSI target while a user is logged in, the Red Hat Ceph Storage Dashboard does not forward that error code and message to the Dashboard user using the pop-up notifications, but displays a generic "500 Internal Server Error". Consequently, the message that the Dashboard provides is not informative and even misleading; an expected behavior ("users cannot delete a busy resource") is perceived as an operational failure ("internal server error"). To work around this issue, see the Dashboard logs.

(BZ#1786457)

7.3. THE CEPH VOLUME UTILITY

Ceph OSD fails to start because udev resets the permissions for BlueStore DB and WAL devices

When specifying the BlueStore DB and WAL partitions for an OSD using the **ceph-volume lvm create** command or specifying the partitions, using the **lvm_volume** option with Ceph Ansible can cause those devices to fail on startup. The **udev** subsystem resets the partition permissions back to **root:disk**.

To work around this issue, manually start the systemd **ceph-volume** service. For example, to start the OSD with an ID of 8, run the following: **systemctl start 'ceph-volume@lvm-8-*'**. You can also use the **service** command, for example: **service ceph-volume@lvm-8-4c6ddc44-9037-477d-903c-63b5a789ade5 start**. Manually starting the OSD results in the partition having the correct permission, **ceph:ceph**.

(BZ#1822134)

7.4. CEPH OBJECT GATEWAY

Deleting buckets or objects in the Ceph Object Gateway causes orphan RADOS objects

Deleting buckets or objects after the Ceph Object Gateway garbage collection (GC) has processed the GC queue causes large quantities of orphan RADOS objects. These RADOS objects are "leaked" data that belonged to the deleted buckets.

Over time, the number of orphan RADOS objects can fill the data pool and degrade the performance of the storage cluster.

To reclaim the space from these orphan RADOS objects, refer to the *Finding orphan and leaky objects* section of the *Red Hat Ceph Storage Object Gateway Configuration and Administration Guide* .

(BZ#1844720)

7.5. MULTI-SITE CEPH OBJECT GATEWAY

The **radosgw-admin** commands that create and modify users are not allowed in secondary zones for multi-site Ceph Obejct Gateway environments

Using the **radosgw-admin** commands to create or modify users and subusers on the secondary zone does not propagate those changes to the master zone, even if the **--yes-i-really-mean-it** option was used.

To workaround this issue, use the REST APIs instead of the **radosgw-admin** commands. The REST APIs enable you to create and modify users in secondary zone, and then propagate those changes to the master zone.

(BZ#1553202)

7.6. PACKAGES

Current version of Grafana causes certain bugs in the Dashboard

Red Hat Ceph Storage 4 uses the Grafana version 5.2.4. This version causes the following bugs in the Red Hat Ceph Storage Dashboard:

- When navigating to **Pools** > **Overall Performance**, Grafana returns the following error:

  ```
  TypeError: l.c[t.type] is undefined
  true
  ```

- When viewing a pool's performance details (**Pools** > select a pool from the list > **Performance Details**) the Grafana bar is displayed along with other graphs and values, but it should not be there.

These bugs will be fixed after rebasing to a newer Grafana version in a future release of Red Hat Ceph Storage.

(BZ#1786107)

7.7. RADOS

The **ceph device** command does not work when querying MegaRaid devices

Currently, the **ceph device query-daemon-health-metrics** command does not support querying the health metrics of disks attached to MegaRaid devices. This command displays an error similar to the following:

```
smartctl returned invalid JSON
```

The disk failure prediction module for MegaRaid devices is unusable at this time. Currently, there is no workaround for this issue.

See the *Red Hat Ceph Storage Hardware Guide* for more information on using RAID solutions with Red Hat Ceph Storage.

(BZ#1810396)

CHAPTER 8. SOURCES

The updated Red Hat Ceph Storage source code packages are available at the following location:

- For Red Hat Enterprise Linux 7: http://ftp.redhat.com/redhat/linux/enterprise/7Server/en/RHCEPH/SRPMS/
- For Red Hat Enterprise Linux 8: http://ftp.redhat.com/redhat/linux/enterprise/8Base/en/RHCEPH/SRPMS/