Red Hat Ceph Storage 7.0

Release Notes

Release notes for Red Hat Ceph Storage 7.0

Last Updated: 2024-03-12

Red Hat Ceph Storage 7.0 Release Notes

Release notes for Red Hat Ceph Storage 7.0

Legal Notice

Abstract

The release notes describes the major features, enhancements, known issues, and bug fixes implemented for the Red Hat Ceph Storage 7.0 product release.

Table of Contents

MAKING OPEN SOURCE MORE INCLUSIVE

Red Hat is committed to replacing problematic language in our code, documentation, and web properties. We are beginning with these four terms: master, slave, blacklist, and whitelist. Because of the enormity of this endeavor, these changes will be implemented gradually over several upcoming releases. For more details, see our CTO Chris Wright's message.

PROVIDING FEEDBACK ON RED HAT CEPH STORAGE DOCUMENTATION

We appreciate your input on our documentation. Please let us know how we could make it better. To do so, create a Bugzilla ticket:

1. Go to the *Bugzilla* website.
2. In the Component drop-down, select **Documentation**.
3. In the Sub-Component drop-down, select the appropriate sub-component.
4. Select the appropriate version of the document.
5. Fill in the **Summary** and **Description** field with your suggestion for improvement. Include a link to the relevant part(s) of documentation.
6. Optional: Add an attachment, if any.
7. Click **Submit Bug**.

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-us/red_hat_ceph_storage/7.

CHAPTER 2. ACKNOWLEDGMENTS

Red Hat Ceph Storage version 7.0 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 ®
- Croit ™
- Clyso ™
- Cloudbase solutions ™

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. THE CEPHADM UTILITY

Add cmount_path option and generate unique user ID

With this enhancement, you can add the optional **cmount_path** option and generate a unique user ID for each Ceph File System to allow sharing CephFS clients across multiple Ganesha exports thereby reducing the memory usage for a single CephFS client.

Bugzilla:2246077

TLS is enabled across all monitoring components, enhancing security for Prometheus

With this enhancement, to safeguard data integrity, confidentiality, and alignment with the security best practices, TLS is enabled across the monitoring stack. The enhanced security feature for Prometheus, Alert manager, and Node exporter adds an additional layer of protection by using secure communication across the monitoring stack.

Bugzilla:1994251

Enhanced security of the monitoring stack

With this enhancement, to safeguard data integrity, confidentiality, and to align with the security best practices, the authentication feature for Prometheus, Alert manager, and the Node exporter is implemented. This enhances the security of the whole monitoring stack by enabling TLS in all the monitoring components and requiring users to provide valid credentials before accessing Prometheus and Alert Manager data. By using this new feature, an additional layer of protection is provided by using secure communication and preventing the unauthorized access to sensitive metrics and monitoring data. With TLS enabled in all the monitoring stack components and authentication in place, users must authenticate before accessing monitoring and metrics data, enhancing overall security and control over data access.

Bugzilla:2028335

Users can now put a host into maintenance mode

Previously, users could not put a host into maintenance mode as stopping all the daemons on that host would cause data unavailability.

With this enhancement, a stronger force flag for the **ceph orch host maintenance enter** command called **--yes-i-really-mean-it** is added. Users can now put a host into maintenance mode, even when Cephadm warns against it.

Bugzilla:2107339

Users can now drain a host of daemons without draining the client conf or keyring files

With this enhancement, users can drain a host of daemons, without also draining the client **conf** or **keyring** files deployed on the host by passing the **--keep-conf-keyring** flag to the **ceph orch host drain** command. Users can now mark a host to have all daemons drained or not placed there while still having Cephadm manage **conf** or **keyring** files on the host.

Bugzilla:2153827

Cephadm services can now be marked as managed or unmanaged

Previously, the only way to mark the Cephadm services as **managed** or **unmanaged** was to edit and re-apply the service specification for the service. This was inconvenient for scenarios, such as temporarily stop making OSDs on devices that match an OSD specification in Cephadm.

With this enhancement, commands are added to set the Cephadm services to be marked as **managed** or **unmanaged**. For example, one can run **ceph orch set-unmanaged mon** command or **ceph orch set-managed mon** command. These new commands allow toggling the states without having to edit and re-apply the service specification.

Bugzilla:2155625

Users can now apply client IP restrictions on the NFS deployment using the HAProxy protocol mode

Previously, users could not apply client IP restrictions, while still using HAProxy between the client and NFS. This is because only the HAProxy IP would be recognized by NFS, making proper client IP restriction impossible.

With this enhancement, it is possible to deploy an NFS service in HAProxy protocol mode by passing **--ingress-mode=haproxy-protocol** argument in the **ceph nfs cluster create** command or by setting **enable_haproxy_protocol: true** in both the NFS service specification and the corresponding ingress specification. Users can now apply proper client IP restriction on their NFS deployment using the new HAProxy protocol mode in their NFS deployment.

Bugzilla:2176300

3.2. CEPH FILE SYSTEM

quota.max_bytes is now set in more understandable size values

Previously, the **quota.max_bytes** value was set in bytes, resulting in often very large size values, which was hard to set or changed.

With this enhancement, the **quota.max_bytes** values can now be set with human-friendly values, such as M/Mi, G/Gi, or T/Ti. For example, 10GiB or 100K.

Bugzilla:2091154

Laggy clients are now evicted only if there are no laggy OSDs

Previously, monitoring performance dumps from the MDS would sometimes show that the OSDs were laggy, **objecter.op_laggy** and **objecter.osd_laggy**, causing laggy clients and the dirty data for cap revokes would not be flushed.

With this enhancement, if the **defer_client_eviction_on_laggy_osds** parameter is set to true and a client gets laggy because of a laggy OSD then client eviction does not take place until OSDs are no longer laggy.

Bugzilla:2228066

3.3. CEPH DASHBOARD

Improved overview dashboard utilization panel

With this enhancement, graphs and graph legends are improved for better usability. In addition, search queries are improved for giving better results.

Bugzilla:2238472

Upgrade the cluster from the dashboard

Previously, a Red Hat Ceph Storage cluster could only be upgraded through the command-line interface.

With this enhancement, you can easily view and upgrade the available versions of the storage cluster and also track the upgrade progress from the Ceph dashboard.

Bugzilla:2232295

The Ceph Dashboard has an Overview page that displays the Ceph Object Gateway information

With this enhancement, an Overview page is added in the Object Gateway section of the Ceph dashboard. From this Overview page, users can now access a dedicated object gateway overview in the Ceph dashboard. This feature enhances the user experience by providing insights into the Object Gateway performance, storage usage, and configuration details. When multi-site is configured in the cluster, the user can also see the multi-site sync status directly on this Overview page.

Bugzilla:2233356

The Ceph dashboard displays capacity usage information for Block Device images

With this enhancement, a new capacity usage progress bar is visible within the Block > Images table on the Ceph dashboard. The bar provides visible usage information, along with a percentage.

This progress bar is only available for block images with **fast-diff** enabled and no snapshot mirroring.

Bugzilla:2151171

Manage Ceph File System volumes on the dashboard

Previously, Ceph File System (CephFS) volumes could only be managed through the command-line interface.

With this enhancement, CephFS volumes can now be listed, created, edited, and removed through the Ceph dashboard.

Bugzilla:2232297

Manage Ceph File System subvolumes and subvolume groups on the dashboard

Previously, Ceph File System (CephFS) subvolumes and subvolumes groups could only be managed through the command-line interface.

With this enhancement, CephFS subvolumes and subvolume groups can now be listed, created, edited, and removed through the Ceph dashboard.

Bugzilla:2232298

Users can now specify the FQDN for Ceph Object Gateway host through the CLI and the dashboard

Previously, short hostnames were picked to resolve the Ceph Object gateway hostname which would cause issues.

With this enhancement, in the Ceph dashboard, the **rgw_dns_name** configuration option of the Ceph Object Gateway is used to resolve the hostname if it is provided. If you wants to specify the FQDN for the Ceph Object Gateway host, use the **rgw_dns_name** configuration option in the CLI and the dashboard then picks it up and the Ceph Object Gateway requests are made against it.

Bugzilla:2236109

Configure Ceph Object Gateway multi-site on the dashboard

With this enhancement, you can configure multi-site Ceph Object Gateway not only through the command-line interface, but also through the Ceph Dashboard. The dashboard now supports creating, updating, and deleting Object Gateway entities such as realms, zonegroups, and zones. In addition, this feature allows users to configure multi-site between two remote clusters, providing options for data replication and synchronization.

Bugzilla:2190355

3.4. CEPH OBJECT GATEWAY

The radosgw-admin bucket command prints bucket versioning

With this enhancement, the ` radosgw-admin bucket stats ` command prints the versioning status for buckets as **enabled** or **off** since versioning can be enabled or disabled after creation.

Bugzilla:2068491

S3 WORM certification with an external entity

With this enhancement, S3 WORM feature is certified with an external entity. This enables data retention in compliance with FSI regulations and a secured object storage deployment that guarantees data retrieval even if the object/buckets in the production zones have been lost or compromised.

For more information, see Introduction to WORM.

Multi-site sync instances now dynamically spread workloads among themselves

Previously, the Ceph Object Gateway multi-site throughput was generally limited to the available bandwidth of a single instance.

With this enhancement, Ceph Object Gateway multisite sync instances dynamically parallelize workload among themselves, using a work sharing algorithm. As a result, the scalability is now significantly improved and sync workloads are divided evenly throughout the sync.

Bugzilla:1740782

Enhanced bucket granular multi-site sync policies

IBM Storage Ceph now supports bucket granular multi-site sync policies.

Bucket granular multi-sync sync replication was previously available as limited release. This enhancement provides full availability for new and existing customers in production environments.

This feature allows enabling and disabling multi-site async replication on a per bucket level. This enhancement also increases the total RAW space required and increases the amount of synchronization traffic between sites.

For more information, see Bucket granular sync policies .

Reduced object storage query times for data analytics with added JSON, Parquet, and CSV-format object support

Ceph Object Gateway now supports operating on JSON-format, Parquet-format, and CSV-format objects, expanding potential application of S3 select to widely deployed analytics frameworks, for example, Apache Spark and Trino. This added support helps reduce query times.

For more information, see S3 select content from an object .

Data can now be transitioned to the Azure cloud service, using the multi-cloud gateway (MCG)

Previously, transitioning data to the Azure cloud service was a Technology Preview feature. With this enhancement, the feature is ready to be used in a production environment.

This feature enables data transition in a Ceph Object Gateway storage environment, into the Azure cloud, for archiving purposes.

For more information, see Transitioning data to Azure cloud service.

Enhanced S3 select feature for more efficient integration with Trino

Object storage now has enhanced integration with Trino for S3 select operations. This improves the query times for semi-structured and structured datasets stored in Ceph Object Gateway.

For more information, see Integrating Ceph Object Gateway with Trino .

3.5. RADOS

New performance counters introduced for messenger v2

Previously, there were no dedicated performance counters for accounting encrypted traffic in messenger v2.

With this enhancement, **msgr_recv_encrypted_bytes** and **msgr_send_encrypted_bytes**, are introduced to account for receiving and sending bytes respectively that facilitate rough validation of the encryption status.

Bugzilla:1846154

New reports available for sub-events for delayed operations

Previously, slow operations were marked as delayed but without a detailed description.

With this enhancement, you can view the detailed descriptions of delayed sub-events for operations.

Bugzilla:2240832

3.6. NFS GANESHA

Support HAProxy's PROXY protocol

With this enhancement, HAProxy uses load balancing servers. This allows load balancing and also enables client restrictions on access.

Bugzilla:2097490

CHAPTER 4. 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.

4.1. CEPH MANAGER PLUG INS

Python tasks no longer wait for the GIL

Previously, the Ceph manager daemon held the Python global interpreter lock (GIL) during some RPCs with the Ceph MDS, due to which, other Python tasks are starved waiting for the GIL.

With this fix, the GIL is released during all **libcephfs**/**librbd** calls and other Python tasks may acquire the GIL normally.

Bugzilla:2219093

4.2. THE CEPHADM UTILITY

cephadm can differentiate between a duplicated hostname and no longer adds the same host to a cluster

Previously, **cephadm** would consider a host with a shortname and a host with its FQDN as two separate hosts, causing the same host to be added twice to a cluster.

With this fix, **cephadm** now recognizes the difference between a host shortname and the FQDN, and does not add the host again to the system.

Bugzilla:2049445

cephadm no longer reports that a non-existing label is removed from the host

Previously, in **cephadm**, there was no check to verify if a label existed before removing it from a host. Due to this, the **ceph orch host label rm** command would report that a label was removed from the host, even when the label was non-existent. For example, a misspelled label.

With this fix, the command now provides clear feedback whether the label specified was successfully removed or not to the user.

Bugzilla:2113901

The keepalive daemons communicate and enter the main/primary state

Previously, keepalive configurations were populated with IPs that matched the host IP reported from the **ceph orch host ls** command. As a result, if the VIP was configured on a different subnet than the host IP listed, the keepalive daemons were not able to communicate, resulting in the keepalive daemons to enter a primary state.

With this fix, the IPs of keepalive peers in the keepalive configuration are now chosen to match the subnet of the VIP. The keepalive daemons can now communicate even if the VIP is in a different subnet than the host IP from **ceph orch host ls** command. In this case, only one keepalive daemon enters primary state.

Bugzilla:2222010

Stopped crash daemons now have the correct state

Previously, when a crash daemon stopped, the return code gave an **error** state, rather than the expected **stopped** state, causing **systemd** to think that the service had failed.

With this fix, the return code gives the expected **stopped** state.

Bugzilla:2126465

HA proxy now binds to the frontend port on the VIP

Previously, in Cephadm, multiple ingress services could not be deployed on the same host with the same frontend port as the port binding occurred across all host networks.

With this fix, multiple ingress services can now be present on the same host with the same frontend port as long as the services use different VIPs and different monitoring ports are set for the ingress service in the specification.

Bugzilla:2231452

4.3. CEPH FILE SYSTEM

User-space Ceph File System (CephFS) work as expected post upgrade

Previously, the user-space CephFS client would sometimes crash during a cluster upgrade. This would occur due to stale feature bits on the MDS side that were held on the user-space side.

With this fix, ensure that the user-space CephFS client has updated MDS feature bits that allows the clients to work as expected after a cluster upgrade.

Bugzilla:2247174

Blocklist and evict client for large session metadata

Previously, large client metadata buildup in the MDS would sometimes cause the MDS to switch to read-only mode.

With this fix, the client that is causing the buildup is blocklisted and evicted, allowing the MDS to work as expected.

Bugzilla:2238663

Deadlocks no longer occur between the unlink and reintegration requests

Previously, when fixing async dirop bug, a regression was introduced by previous commits, causing deadlocks between the unlink and reintegration request.

With this fix, the old commits are reverted and there is no longer a deadlock between unlink and reintegration requests.

Bugzilla:2228635

Client always sends a caps revocation acknowledgement to the MDS daemon

Previously, whenever an MDS daemon sent a caps revocation request to a client and during this time, if the client released the caps and removed the inode, then the client would drop the request directly, but the MDS daemon would need to wait for a caps revoking acknowledgement from the client. Due to this, even when there was no need for caps revocation, the MDS daemon would continue waiting for an acknowledgement from the client, causing a warning in MDS Daemon health status.

With this fix, the client always sends a caps revocation acknowledgement to the MDS Daemon, even when there is no inode existing and the MDS Daemon no longer stays stuck.

Bugzilla:2228000

MDS locks are obtained in the correct order

Previously, MDS would acquire metadata tree locks in the wrong order, resulting in a **create** and **getattr** RPC request to deadlock.

With this fix, locks are obtained in the correct order in MDS and the requests no longer deadlock.

Bugzilla:2235338

Sending split_realms information is skipped from CephFS MDS

Previously, the **split_realms** information would be incorrectly sent from the CephFS MDS which could not be correctly decoded by **kclient**. Due to this, the clients would not care about the **split_realms** and treat it as a corrupted snaptrace.

With this fix, **split_realms** are not sent to **kclient** and no crashes take place.

Bugzilla:2228003

Snapshot data is no longer lost after setting writing flags

Previously, in clients, if the **writing** flag was set to '1' when the **Fb** caps were used, it would be skipped in case of any dirty caps and reuse the existing capsnap, which is incorrect. Due to this, two consecutive snapshots would be overwritten and lose data.

With this fix, the **writing** flags are correctly set and no snapshot data is lost.

Bugzilla:2224241

Thread renaming no longer fails

Previously, in a few rare cases, during renaming, if another thread tried to lookup the dst dentry, there were chances for it to get inconsistent result, wherein both the src dentry and dst dentry would link to the same inode simultaneously. Due to this,the rename request would fail as two different dentries were being linked to the same inode.

With this fix, the thread waits for the renaming action to finish and everything works as expected.

Bugzilla:2227987

Revocation requests no longer get stuck

Previously, before the revoke request was sent out, which would increase the 'seq', if the clients released the corresponding caps and sent out the cap update request with the old **seq**, the MDS would miss checking the **seq**(s) and cap calculation. Due to this, the revocation requests would be stuck infinitely and would throw warnings about the revocation requests not responding from clients.

With this fix, an acknowledgement is always sent for revocation requests and they no longer get stuck.

Bugzilla:2227992

Errors are handled gracefully in MDLog::_recovery_thread

Previously, a write would fail if the MDS was already blocklisted due to the **fs fail** issued by the QA tests. For instance, the QA test **test_rebuild_moved_file** (tasks/data-scan) would fail due to this reason.

With this fix, the write failures are gracefully handled in **MDLog::_recovery_thread**.

Bugzilla:2228358

Ceph client now verifies the cause of lagging before sending out an alarm

Previously, Ceph would sometimes send out false alerts warning of laggy OSDs. For example, **X client(s) laggy due to laggy OSDs**. These alerts were sent out without verifying that the lagginess was actually due to the OSD, and not due to some other cause.

With this fix, the **X client(s) laggy due to laggy OSDs** message is only sent out if some clients and an OSD is laggy.

Bugzilla:2247187

4.4. CEPH DASHBOARD

Grafana panels for performance of daemons in the Ceph Dashboard now show correct data

Previously, the labels exporter were not compatible with the queries used in the Grafana dashboard. Due to this, the Grafana panels were empty for Ceph daemons performance in the Ceph Dashboard.

With this fix, the label names are made compatible with the Grafana dashboard queries and the Grafana panels for performance of daemons show correct data.

Bugzilla:2241309

Edit layering and deep-flatten features disabled on the Dashboard

Previously, in the Ceph dashboard, it was possible to allow editing the layering & deep-flatten features, which are immutable, resulting in an error - **rbd: failed to update image features: (22) Invalid argument**.

With this fix, editing the layering & deep-flatten features are disabled and everything works as expected.

Bugzilla:2166708

ceph_daemon label is added to the labeled performance counters in Ceph exporter

Previously, in Ceph exporter, adding the **ceph_daemon** label to the labeled performance counters was missed.

With this fix, **ceph_daemon** label is added to the labeled performance counters in Ceph exporter. **ceph daemon** label is now present on all Ceph daemons performance metrics and **instance_id** label for Ceph Object Gateway performance metrics.

Bugzilla:2240972

Protecting snapshot is enabled only if layering for its parent image is enabled

Previously, protecting snapshot was enabled even if layering was disabled for its parent image. This caused errors when trying to protect the snapshot of an image for which layering was disabled.

With this fix, protecting snapshot is disabled if layering for an image is disabled. Protecting snapshot is enabled only if layering for its parent image is enabled.

Bugzilla:2166705

Newly added host details are now visible on the cluster expansion review page

Previously, users could not see the information about the hosts that were added in the previous step.

With this fix, hosts that were added in the previous step are now visible on the cluster expansion review page.

Bugzilla:2232567

Ceph Object Gateway page now loads properly on the Ceph dashboard.

Previously, an incorrect regex matching caused the dashboard to break when trying to load the Ceph Object Gateway page. The Ceph Object Gateway page would not load with specific configurations like **rgw_frontends like beast port=80 ssl_port=443**.

With this fix, the regex matching in the codebase is updated and the Ceph Object Gateway page loads without any issues.

Bugzilla:2238470

4.5. CEPH OBJECT GATEWAY

Ceph Object Gateway daemon no longer crashes where phoneNumbers.addr is NULL

Previously, due to a syntax error, the query for **select * from s3object[*].phonenumbers where phoneNumbers.addr is NULL;** would cause the Ceph Object Gateway daemon to crash.

With this fix the wrong syntax is identified and reported, no longer causing the daemon to crash.

Bugzilla:2230234

Ceph Object Gateway daemon no longer crashes with cast(trim) queries

Previously, due to the trim skip type checking within the query for **select cast(trim(leading 132140533849470.72 from _3) as float) from s3object;**, the Ceph Object Gateway daemon would crash.

With this fix the type is checked and is identified if wrong and reported, no longer causing the daemon to crash.

Bugzilla:2248866

Ceph Object Gateway daemon no longer crashes with “where” clause in an s3select JSON query.

Previously, due to a syntax error, an **s3select** JSON query with a “where” clause would cause the the Ceph Object Gateway daemon to crash.

With this fix the wrong syntax is identified and reported, no longer causing the daemon to crash.

Bugzilla:2225434

Ceph Object Gateway daemon no longer crashes with s3 select phonenumbers.type query

Previously, due to a syntax error, the query for **select phonenumbers.type from s3object[*].phonenumbers;** would cause the Ceph Object Gateway daemon to crash.

With this fix the wrong syntax is identified and reported, no longer causing the daemon to crash.

Bugzilla:2230230

Ceph Object Gateway daemon validates arguments and no longer crashes

Previously, due to an operator with missing arguments, the daemon would crash when trying to access the nonexistent arguments.

With this fix the daemon validates the number of arguments per operator and the daemon no longer crashes.

Bugzilla:2230233

Ceph Object Gateway daemon no longer crashes with the trim command

Previously, due to the trim skip type checking within the query for **select trim(LEADING '1' from '111abcdef111') from s3object;**, the Ceph Object Gateway daemon would crash.

With this fix, the type is checked and is identified if wrong and reported, no longer causing the daemon to crash.

Bugzilla:2248862

Ceph Object Gateway daemon no longer crashes if a big value is entered

Previously, due to too large of a value entry, the query for **select DATE_DIFF(SECOND, utcnow(),date_add(year,1111111111111111111, utcnow())) from s3object;** would cause the Ceph Object Gateway daemon to crash.

With this fix, the crash is identified and an error is reported.

Bugzilla:2245145

Ceph Object Gateway now parses the CSV objects without processing failures

Previously, Ceph Object Gateway failed to properly parse CSV objects. When the process failed, the requests would stop without a proper error message.

With this fix, the CSV parser works as expected and processes the CSV objects with no failures.

Bugzilla:2241907

Object version instance IDs beginning with a hyphen are restored

Previously, when restoring the index on a versioned bucket, object versions with an instance ID beginning with a hyphen would not be properly restored into the bucket index.

With this fix, instance IDs beginning with a hyphen are now recognized and restored into the bucket index, as expected.

Bugzilla:2247138

Multi-delete function notifications work as expected

Previously, due to internal errors, such as a race condition in the code, the Ceph Object Gateway would crash or react unexpectedly when multi-delete functions were performed and the notifications were set for bucket deletions.

With this fix, notifications for multi-delete function work as expected.

Bugzilla:2239173

RADOS object multipart upload workflows complete properly

Previously, in some cases, a RADOS object that was part of a multipart upload workflow objects that were created on a previous upload would cause certain parts to not complete or stop in the middle of the upload.

With this fix, all parts upload correctly, once the multipart upload workflow is complete.

Bugzilla:2008835

Users belonging to a different tenant than the bucket owner can now manage notifications

Previously, a user that belonged to a different tenant than the bucket owner was not able to manage notifications. For example, modify, get, or delete.

With this fix, any user with the correct permissions can manage the notifications for the buckets.

Bugzilla:2180415

Ability to perform NFS setattr on buckets is removed

Previously, changing the attributes stored on a bucket via export as an NFS directory triggered an inconsistency in the Ceph Object gateway bucket information cache. Due to this, subsequent accesses to the bucket via NFS failed.

With this fix, the ability to perform NFS **setattr** on buckets is removed and attempts to perform NFS **setattr** on a bucket, for example, **chown** on the directory, have no effect.

NOTE

This might change in future releases.

Bugzilla:2241145

Testing for reshardable bucket layouts is added to prevent crashes

Previously, with the added bucket layout code to enable dynamic bucket resharding with multi-site, there was no check to verify if the bucket layout supported resharding during dynamic, immediate, or rescheduled resharding. Due to this, the Ceph Object gateway daemon would crash in case of dynamic bucket resharding and the **radosgw-admin** command would crash in case of immediate or scheduled resharding.

With this fix, a test for reshardable bucket layouts is added and the crashes no longer occur. When immediate and scheduled resharding occurs, an error message is displayed. When dynamic bucket resharding occurs, the bucket is skipped.

Bugzilla:2242987

The user modify -placement-id command can now be used with an empty --storage-class argument

Previously, if the **--storage-class** argument was not used when running the 'user modify --placement-id' command, the command would fail.

With this fix, the **--storage-class** argument can be left empty without causing the command to fail.

Bugzilla:2228157

Initialization now only unregisters watches that were previously registered

Previously, in some cases, an error in initialization could cause an attempt to unregister a watch that was never registered. This would result in some command line tools crashing unpredictably.

With this fix, only previously registered watches are unregistered.

Bugzilla:2224078

Multi-site replication now maintains consistent states between zones and prevents overwriting deleted objects

Previously, a race condition in multi-site replication would allow objects that should be deleted to be copied back from another site, resulting in an inconsistent state between zones. As a result, the zone which is receiving the workload ends up with some objects which should be deleted still present.

With this fix, a custom header is added to pass the destination zone's trace string and is then checked against the object's replication trace. If there is a match, a 304 response is returned, preventing the full sync from overwriting a deleted object.

Bugzilla:2219427

The memory footprint of Ceph Object Gateway has significantly been reduced

Previously, in some cases, a memory leak associated with Lua scripting integration caused excessive RGW memory growth.

With this fix, the leak is fixed and the memory footprint for Ceph Object Gateway is significantly reduced.

Bugzilla:2032001

Bucket index performance no longer impacted during versioned object operations

Previously, in some cases, space leaks would occur and reduce bucket index performance. This was caused by a race condition related to updates of object logical head (OLH), which relates to versioned bucket current version calculations during updates.

With this fix, logic errors in OLH update operations are fixed and space is no longer being leaked during versioned object operations.

Bugzilla:2219467

Delete markers are working correctly with the LC rule

Previously, optimization was attempted to reuse a sal object handle. Due to this, delete markers were not being generated as expected.

With this fix, the change to re-use sal object handle for get-object-attributes is reverted and delete markers are created correctly.

Bugzilla:2248116

SQL engine no longer causes Ceph Object Gateway crash with illegal calculations

Previously, in some cases, the SQL engine would throw an exception that was not handled, causing a Ceph Object Gateway crash. This was caused due to an illegal SQL calculation of a date-time operation.

With this fix, the exception is handled with an emitted error message, instead of crashing.

Bugzilla:2246150

The **select trim (LEADING '1' from '111abcdef111') from s3object;** query now works when capitals are used in query

Previously, if **LEADING** or **TRAILING** were written in all capitals, the string would not properly read, causing a float type to be referred to as a string type, thus leading to a wrong output.

With this fix, type checking is introduced before completing the query, and **LEADING** and **TRAILING** work written either capitalized or in lower case.

Bugzilla:2245575

JSON parsing now works for **select _1.authors.name from s3object[*] limit 1** query

Previously, an anonymous array given in the **select _1.authors.name from s3object[*] limit 1** would give the wrong value output.

With this fix, JSON parsing works, even if an anonymous array is provided to the query.

Bugzilla:2236462

4.6. MULTI-SITE CEPH OBJECT GATEWAY

Client no longer resets the connection for an incorrect **Content-Length** header field value

Previously, when returning an error page to the client, for example, a 404 or 403 condition, the **</body>** and **</html>** closing tags were missing, although their presence was accounted for in the request's **Content-Length** header field value. Due to this, depending on the client, the TCP connection between the client and the Rados Gateway would be closed by an RST packet from the client on account of incorrect Content-Length header field value, instead of a FIN packet under normal circumstances.

With this fix, send the **</body>** and **</html>** closing tags to the client under all the required conditions. The value of the **Content-Length** header field correctly represents the length of data sent to the client, and the client no longer resets the connection for an incorrect Content-Length reason.

Bugzilla:2189412

Sync notification are sent with the correct object size

Previously, when an object was synced between zones, and sync notifications were configured, the notification was sent with zero as the size of the object.

With this fix, sync notifications are sent with the correct object size.

Bugzilla:2238921

Multi-site sync properly filters and checks according to allowed zones and filters

Previously, when using the multi-site sync policy, certain commands, such as **radosgw-admin sync status**, would not filter restricted zones or empty sync group names. The lack of filter caused the output of these commands to be misleading.

With this fix, restricted zones are no longer checked or reported and empty sync group names are filtered out of the status results.

Bugzilla:2159966

4.7. RADOS

The ceph version command no longer returns the empty version list

Previously, if the MDS daemon was not deployed in the cluster then the **ceph version** command returned an empty version list for MDS daemons that represented version inconsistency. This should not be shown if the daemon is not deployed in the cluster.

With this fix, the daemon version information is skipped if the daemon version map is empty and the **ceph version** command returns the version information only for the Ceph daemons which are deployed in the cluster.

Bugzilla:2110933

ms_osd_compression_algorithm now displays the correct value

Previously, an incorrect value in **ms_osd_compression_algorithm** displayed a list of algorithms instead of the default value, causing a discrepancy by listing a set of algorithms instead of one.

With this fix, only the default value is displayed when using the **ms_osd_compression_algorithm** command.

Bugzilla:2155380

MGR no longer disconnects from the cluster without retries

Previously, during network issues, clusters would disconnect with MGR without retries and the authentication of **monclient** would fail.

With this fix, retries are added in scenarios where hunting and connection would both fail.

Bugzilla:2106031

Increased timeout retry value for client_mount_timeout

Previously, due to the mishandling of the **client_mount_timeout** configurable, the timeout for authenticating a client to monitors could reach up to 10 retries disregarding its high default value of 5 minutes.

With this fix, the previous single-retry behavior of the configurable is restored and the authentication timeout works as expected.

Bugzilla:2233800

4.8. RBD MIRRORING

Demoted mirror snapshot is removed following the promotion of the image

Previously, due to an implementation defect, the demoted mirror snapshots would not be removed following the promotion of the image, whether on the secondary image or on the primary image. Due to this, demoted mirror snapshots would pile up and consume storage space.

With this fix, the implementation defect is fixed and the appropriate demoted mirror snapshot is removed following the promotion of the image.

Bugzilla:2237304

Non-primary images are now deleted when the primary image is deleted

Previously, a race condition in the rbd-mirror daemon image replayer prevented a non-primary image from being deleted when the primary was deleted. Due to this, the non-primary image would not be deleted and the storage space was used.

With this fix, the rbd-mirror image replayer is modified to eliminate the race condition. Non-primary images are now deleted when the primary image is deleted.

Bugzilla:2230056

The librbd client correctly propagates the block-listing error to the caller

Previously, when the **rbd_support** module's RADOS client was block-listed, the module's **mirror_snapshot_schedule** handler would not always shut down correctly. The handler's **librbd** client would not propagate the block-list error, thereby stalling the handler's shutdown. This lead to the failures of the **mirror_snapshot_schedule** handler and the **rbd_support** module to automatically recover from repeated client block-listing. The **rbd_support** module stopped scheduling mirror snapshots after its client was repeatedly block-listed.

With this fix, the race in the **librbd** client between its exclusive lock acquisition and handling of block-listing is fixed. This allows the **librbd** client to propagate the block-listing error correctly to the caller, for example, the **mirror_snapshot_schedule** handler, while waiting to acquire an exclusive lock. The **mirror_snapshot_schedule** handler and the **rbd_support_module** automatically recovers from repeated client block-listing.

Bugzilla:2237303

CHAPTER 5. TECHNOLOGY PREVIEWS

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.

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 using them for production. These features provide early access to upcoming product features, enabling customers to test functionality and provide feedback during the development process. See the support scope for Red Hat Technology Preview features for more details.

5.1. CRIMSON OSD

Newly implemented Crimson-OSD of the core Ceph object storage daemon (OSD) component replaces ceph-osd

With this enhancement, the next generation ceph-osd is implemented for multi-core scalability and to improve performance with fast network and storage devices, employing state-of-the-art technologies that includes DPDK and SPDK. Crimson aims to be compatible with an earlier version of OSD daemon with the class ceph-osd.

For more information, see Crimson (Technology Preview).

5.2. CEPH OBJECT GATEWAY

Object storage archive zone in Red Hat Ceph Storage

With this enhancement, the archive zone receives all objects from the production zones and keeps every version for every object, providing the user with an object catalogue that contains the full history of the object. This provides a secured object storage deployment that guarantees data retrieval even if the object/buckets in the production zones have been lost or compromised.

For more information, see Configuring the archive zone (Technology Preview).

Protect object storage data outside of a production cluster using per-bucket enable and disable sync to an archive zone

As an administrator, you can now recover any version of any object that has existed on the primary site from the archive zone. In the case of data loss or a ransomware attack, valid versions of all objects are accessible, if needed.

For more information, see Configuring the archive zone (Technology Preview).

5.3. RADOS

Balancing Red Hat Ceph Storage cluster using read balancer

With this release, to ensure that each device gets its fair share of primary OSDs so that read requests get distributed across OSDs in the cluster, evenly, read balancer is implemented. Read balancing is cheap and the operation is fast as there is no data movement involved. Read balancing supports replicated pools only. Erasure coded pools are not supported.

For more information, see Balancing Red Hat Ceph Storage cluster using read balancer (Technology Preview) and Ceph rebalancing and recovery.

CHAPTER 6. KNOWN ISSUES

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

6.1. THE CEPHADM UTILITY

Some NFS daemons may not produce logs

Currently, in a cephadm-deployed NFS service with multiple NFS daemons, only one would be serving IO, causing the user to come across NFS daemons that produce no logs. There is no workaround for this as of now.

Bugzilla:2251653

Prevent Mutex from failing when unlocking.

Previously, when a Mutex that was not locked was attempted to be unlocked, the Mutex crashed.

As a workaround, verify if the Mutex is locked in the first place before unlocking.

Bugzilla:2216442

ceph orch ps command does not display a version for monitoring stack daemons

In **cephadm**, due to the version grabbing code currently being incompatible with the downstream monitoring stack containers, version grabbing fails for monitoring stack daemons, such as **node-exporter**, **prometheus**, and **alertmanager**.

Encryption of multipart uploads requires special handling around the part boundaries because each part is uploaded and encrypted separately. In multi-site, objects are encrypted, and multipart uploads are replicated as a single part. As a result, the replicated copy loses its knowledge about the original part boundaries required to decrypt the data correctly, which causes this corruption.

As a workaround, if the user needs to find the version, the daemons' container names include the version.

Bugzilla:2125382

6.2. CEPH DASHBOARD

Ceph Object Gateway page does not load after a multi-site configuration

The Ceph Object Gateway page does not load because the dashboard cannot find the correct access key and secret key for the new realm during multi-site configuration.

As a workaround, use the **ceph dashboard set-rgw-credentials** command to manually update the keys.

Bugzilla:2231072

6.3. CEPH OBJECT GATEWAY

JSON select count() from S3Object[]; queries are lagging and cause high CPU usage

When running the **select count() from S3Object[];** query, the time lapse and radosgw CPU utilization are very high compared to CSV object queries.

As a workaround, when running the JSON query, use **count()** instead of **count(*)** query.

Bugzilla:2240974

s3select and Trino fail when processing JSON object using s3select

When Trino processes a JSON object using the s3select request, the request fails causing Trino to fail too. This emits the **wrong json dataType should use DOCUMENT** error message in the Ceph Object Gateway logs.

As a workaround, it is sometimes possible to use the s3select request directly, not using Trino.

Bugzilla:249756

CHAPTER 7. ASYNCHRONOUS ERRATA UPDATES

This section describes the bug fixes, known issues, and enhancements of the z-stream releases.

7.1. RED HAT CEPH STORAGE 7.0Z1

Red Hat Ceph Storage release 7.0z1 is now available. The bug fixes that are included in the update are listed in the RHBA-2024:1214 and RHBA-2024:1215 advisories.

7.1.1. Enhancements

7.1.1.1. Ceph Block Devices

Improved rbd_diff_iterate2() API performance

Previously, RBD diff-iterate was not guaranteed to execute locally if exclusive lock was available when diffing against the beginning of time (**fromsnapname == NULL**) in fast-diff mode (**whole_object == true** with **fast-diff** image feature enabled and valid).

With this enhancement, **rbd_diff_iterate2()** API performance is improved, thereby increasing the performance for QEMU live disk synchronization and backup use cases, where the **fast-diff** image feature is enabled.

Bugzilla:2259052

7.1.1.2. Ceph Object Gateway

rgw-restore-bucket-index tool can now restore the bucket indices for versioned buckets

With this enhancement, the **rgw-restore-bucket-index** tool now works as broadly as possible, with the ability to restore the bucket indices for un-versioned as well as for versioned buckets.

Bugzilla:2240992

7.1.2. Known issues

7.1.2.1. Multi-site Ceph Object Gateway

Some Ceph Object Gateway applications using S3 client SDKs can experience unexpected errors

Presently, some applications using S3 client SDKs could experience an unexpected 403 error when uploading a zero-length object, if an external checksum is requested.

As a workaround, use Ceph Object Gateway services with SSL.

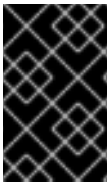
Bugzilla:2256969

7.1.3. Removed functionality

7.1.3.1. Ceph Object Gateway

Prometheus metrics are no longer used

This release introduces new feature-rich labeled perf counters, replacing the Object Gateway-related Prometheus metrics previously used. The new metrics are being introduced before complete removal to allow overlapping usage.



IMPORTANT

The Prometheus metrics are currently still available for use simultaneously with the newer metrics during this transition. However, the Prometheus metrics will be completely removed in the Red Hat Ceph Storage 8.0 release.

Use the following table for knowing the replaced metrics in 7.0z1 and later.

Table 7.1. Replacement metrics

Deprecated Prometheus metric	New metric in 7.0z1
ceph_rgw_get	**ceph_rgw_op_global_get_obj_ops**
ceph_rgw_get_b	**ceph_rgw_op_global_get_obj_bytes**
ceph_rgw_get_initial_lat_sum	**ceph_rgw_op_global_get_obj_lat_sum**
ceph_rgw_get_initial_lat_count	**ceph_rgw_op_global_get_obj_lat_count**
ceph_rgw_put	**ceph_rgw_op_global_put_obj_ops**
cdeph_rgw_put_b	**ceph_rgw_op_global_put_obj_bytes**
ceph_rgw_put_initial_lat_sum	**ceph_rgw_op_global_put_obj_lat_sum**
ceph_rgw_put_initial_lat_count	**ceph_rgw_op_global_put_obj_lat_count**

CHAPTER 8. SOURCES

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

- For Red Hat Enterprise Linux 9:
  https://ftp.redhat.com/redhat/linux/enterprise/9Base/en/RHCEPH/SRPMS/