Red Hat Ceph Storage 6.1

6.1 Release Notes

Release notes for Red Hat Ceph Storage 6.1

Last Updated: 2024-08-27

Red Hat Ceph Storage 6.1 6.1 Release Notes

Release notes for Red Hat Ceph Storage 6.1

Legal Notice

Abstract

The release notes describes the major features, enhancements, known issues, and bug fixes implemented for the Red Hat Ceph Storage 6.1 product release. 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.

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:

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

CHAPTER 2. ACKNOWLEDGMENTS

Red Hat Ceph Storage version 6.1 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 AND ENHANCEMENTS

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:

Compression on-wire with msgr2 protocol is now available

With this release, in addition to encryption on wire, compression on wire is also supported to secure network opertaions within the storage cluster.

See the *Encryption and key management* section in the *Red Hat Ceph Storage Data Security and Hardening Guide* for more details.

Python notifications are more efficient

Previously, there were some unused notifications that no modules needed at the moment. This caused inefficiency.

With this release, the **NotifyType** parameter is introduced. It is annotated, which events modules consume at the moment, for example **NotifyType.mon_map**, **NotifyType.osd_map**, and the like. As a consequence, only events that modules ask for are queued. The events that no modules consume are issued. Because of these changes, python notifications are now more efficient.

The changes to pg_num are limited

Previously, if drastic changes were made to **pg_num** that outpaced **pgp_num**, the user could hit the per-osd placement group limits and cause errors.

With this release, the changes to **pg_num** are limited to avoid the issue with per-osd placement group limits.

New pg_progress item is created to avoid dumping all placement group statistics for progress updates

Previously, the **pg_dump** item included unnecessary fields that wasted CPU if it was copied to **python-land**. This tended to lead to long **ClusterState::lock** hold times, leading to long **ms_dispatch** delays and generally slowing the processes.

With this release, a new **pg_progress** item is created to dump only the fields that **mgr tasks** or **progress** needs.

The mgr_ip is no longer re-fetched

Previously, the **mgr_ip** had to be re-fetched during the lifetime of an active Ceph manager module.

With this release, the **mgr_ip** does not change for the lifetime of an active Ceph manager module, thereby, there is no need to call back into Ceph Manager for re-fetching.

QoS in the Ceph OSD is based on the mClock algorithm, by default

Previously, the scheduler defaulted to the Weighted Priority Queue (WPQ). Quality of service (QoS) based on the mClock algorithm was in an experimental phase and was not yet recommended for production.

With this release, the mClock based operation queue enables QoS controls to be applied to Ceph OSD specific operations, such as client input and output (I/O) and recovery or backfill, as well as other background operations, such as **pg scrub**, **snap trim**, and **pg deletion**. The allocation of resources to

each of the services is based on the input and output operations per second (IOPS) capacity of each Ceph OSD and is achieved using built-in mClock profiles.

Also, this release includes the following enhancements:

- Hands-off automated baseline performance measurements for the OSDs determine Ceph OSD IOPS capacity with safeguards to fallback to default capacity when an unrealistic measurement is detected.
- Setting sleep throttles for background tasks is eliminated.
- Higher default values for recoveries and max backfills options with the ability to override them using an override flag.
- Configuration sets using mClock profiles hide complexity of tuning mClock and Ceph parameters.

See *The mClock OSD scheduler* in *Red Hat Ceph Storage Administration Guide* for details.

WORM compliance certification is now supported

Red Hat now supports WORM compliance certification.

See the *Enabling object lock for S3* for more details.

Set rate limits on users and buckets

With this release, you can set rate limits on users and buckets based on the operations in a Red Hat Ceph Storage cluster. See the *Rate limits for ingesting data* for more details.

librbd plugin named persistent write log cache to reduce latency

With this release, the new **librbd** plugin named Persistent Write Log Cache (PWL) provides a persistent, fault-tolerant write-back cache targeted with SSD devices. It greatly reduces latency and also improves performance at low **io_depths**. This cache uses a log-ordered write-back design which maintains checkpoints internally, so that writes that get flushed back to the cluster are always crash consistent. Even if the client cache is lost entirely, the disk image is still consistent; but the data will appear to be stale.

Ceph File System (CephFS) now supports high availability asynchronous replication for snapshots

Previously, only one **cephfs-mirror** daemon would be deployed per storage cluster, thereby a CephFS supported only asynchronous replication of snapshots directories.

With this release, multiple **cephfs-mirror** daemons can be deployed on two or more nodes to achieve concurrency in snapshot synchronization, thereby providing high availability.

See the *Ceph File System mirroring* section in the *Red Hat Ceph Storage File System Guide* for more details.

BlueStore is uprgaded to V3

With this release, BlueStore object store is upgraded to V3. Following are the two features:

- The allocation metadata is removed from RocksDB and now performas a full destage of the allocator object with the OSD allocation.

- With cache age binning, older onodes might be assigned a lower priority than the hot workload data. See the *Ceph BlueStore* for more details.

Use cephadm to manage operating system tuning profiles

With this release, you can use **cephadm** to create and manage operating susyem tuning profiles for better performance of the Red Hat Ceph Storage cluster. See the *Managing operating system tuning profiles with `cephadm`* for more details.

A direct upgrade from Red Hat Ceph Storage 5 to Red Hat Ceph Storage 7 will be available

For upgrade planning awareness, directly upgrading Red Hat Ceph Storage 5 to Red Hat Ceph Storage 7 (N=2) will be available.

The new cephfs-shell option is introduced to mount a filesystem by name

Previously, cephfs-shell could only mount the default filesystem.

With this release, a CLI option is added in cephfs-shell that allows the mounting of a different filesystem by name, that is, something analogous to the **mds_namespace=** or **fs= options** for **kclient** and **ceph-fuse**.

Day-2 tasks can now be performed through the Ceph Dashboard

With this release, in the Ceph Dashboard, a user can perform every day-2 tasks that require daily or weekly frequency of actions. This enhancement improves the Dashboard's assessment capabilities, customer experience, and strengthens its usability and maturity. In addition to this, new on-screen elements are also included to help and guide the user in retrieving additional information to complete a task.

3.1. THE CEPHADM UTILITY

Users can now rotate the authentication key for Ceph daemons

For security reasons, some users might desire to occasionally rotate the authentication key used for daemons in the storage cluster.

With this release, the ability to rotate the authentication key for ceph daemons using the **ceph orch daemon rotate-key *DAEMON_NAME*** command is introduced. For MDS, OSD, and MGR daemons, this does not require a daemon restart. However, for other daemons, such as Ceph Object Gateway daemons, the daemon might require restarting to switch to the new key.

Bugzilla:1783271

Bootstrap logs are now logged to STDOUT

With this release, to reduce potential errors, bootstrap logs are now logged to **STDOUT** instead of **STDERR** in successful bootstrap scenarios.

Bugzilla:1932764

Ceph Object Gateway zonegroup can now be specified in the specification used by the orchestrator

Previously, the orchestrator could handle setting the realm and zone for the Ceph Object Gateway. However, setting the zonegroup was not supported.

With this release, users can specify a **rgw_zonegroup** parameter in the specification that is used by the orchestrator. Cephadm sets the zonegroup for Ceph Object Gateway daemons deployed from the specification.

Bugzilla:2016288

ceph orch daemon add osd now reports if the hostname specified for deploying the OSD is unknown

Previously, since the **ceph orch daemon add osd** command gave no output, users would not notice if the hostname was incorrect. Due to this, Cephadm would discard the command.

With this release, the **ceph orch daemon add osd** command reports to the user if the hostname specified for deploying the OSD on is unknown.

Bugzilla:2016949

cephadm shell command now reports the image being used for the shell on startup

Previously, users would not always know which image was being used for the shell. This would affect the packages that were used for commands being run within the shell.

With this release, **cephadm shell** command reports the image used for the shell on startup. Users can now see the packages being used within the shell, as they can see the container image being used, and when that image was created as the shell starts up.

Bugzilla:2029714

Cluster logs under` /var/log/ceph` are now deleted

With this release, to better clean up the node as part of removing the Ceph cluster from that node, cluster logs under **/var/log/ceph** are deleted when **cephadm rm-cluster** command is run. The cluster logs are removed as long as **--keep-logs** is not passed to the **rm-cluster** command.

NOTE

If the **cephadm rm-cluster** command is run on a host that is part of a still existent cluster, the host is managed by Cephadm, and the Cephadm mgr module is still enabled and running, then Cephadm might immediately start deploying new daemons, and more logs could appear.

Bugzilla:2036063

Better error handling when daemon names are passed to ceph orch restart command

Previously, in cases where the daemon passed to **ceph orch restart** command was a haproxy or keepalived daemon, it would return a traceback. This made it unclear to users if they had made a mistake or Cephadm had failed in some other way.

With this release, better error handling is introduced to identify when the users pass a daemon name to **ceph orch restart** command instead of the expected service name. Upon encountering a daemon name, Cephadm reports and requests the user to check **ceph orch ls** for valid services to pass.

Bugzilla:2080926

Users can now create Ceph Object Gateway realm,zone, and zonegroup using ceph rgw realm bootstrap -i rgw_spec.yaml command

With this release, to streamline the process of setting up Ceph Object Gateway on a Red Hat Ceph Storage cluster, users can create a Ceph Object Gateway realm, zone, and zonegroup using **ceph rgw realm bootstrap -i rgw_spec.yaml** command. The specification file should be modeled similar to the one that is used to deploy Ceph Object Gateway daemons using the orchestrator. The command then creates the realm,zone, and, zonegroup, and passes the specification on to the orchestrator, which then deploys the Ceph Object Gateway daemons.

Example

```
rgw_realm: myrealm
rgw_zonegroup: myzonegroup
rgw_zone: myzone
placement:
  hosts:
   - rgw-host1
   - rgw-host2
spec:
  rgw_frontend_port: 5500
```

Bugzilla:2109224

crush_device_class and location fields are added to OSD specifications and host specifications respectively

With this release, the **crush_device_class** field is added to the OSD specifications, and the **location** field, referring to the initial crush location of the host, is added to host specifications. If a user sets the **location** field in a host specification, cephadm runs **ceph osd crush add-bucket** with the hostname, and given location to add it as a bucket in the crush map. For OSDs, they are set with the given **crush_device_class** in the crush map upon creation.

NOTE

This is only for OSDs that were created based on the specification with the field set. It does not affect the already deployed OSDs.

Bugzilla:2124441

Users can enable Ceph Object Gateway manager module

With this release, Ceph Object Gateway Manager module is now available, and can be turned on with **ceph mgr module enable rgw** command to enable users to gain access to the functionality of the Ceph Object Gateway manager module, such as **ceph rgw realm bootstrap**, and **ceph rgw realm tokens** commands.

Bugzilla:2133802

Users can enable additional metrics for node-exporter daemons

With this release, to enable users to have more customization of their node-exporter deployments, without requiring explicit support for each individual option, additional metrics are introduced that can now be enabled for **node-exporter** daemons deployed by Cephadm, using the **extra_entrypoint_args** field.

service_type: node-exporter service_name: node-exporter placement: label: "node-exporter" extra_entrypoint_args: - "--collector.textfile.directory=/var/lib/node_exporter/textfile_collector2" ---

Bugzilla:2142431

Users can set the crush location for a Ceph Monitor to replace tiebreaker monitors

With this release, users can set the crush location for a monitor deployed on a host. It should be assigned in the mon specification file.

Example

```
service_type: mon
service_name: mon
placement:
  hosts:
  - host1
  - host2
  - host3
spec:
  crush_locations:
    host1:
    - datacenter=a
    host2:
    - datacenter=b
    - rack=2
    host3:
    - datacenter=a
```

This is primarily added to make the replacing of a tiebreaker monitor daemon in stretch clusters deployed by Cephadm, more feasible. Without this change, users would have to manually edit the files written by Cephadm to deploy the tiebreaker monitor, as the tiebreaker monitor is not allowed to join without declaring its crush location.

Bugzilla:2149533

crush_device_class can now be specified per path in an OSD specification

With this release, to allow users more flexibility with **crush_device_class** settings when deploying OSDs through Cephadm, **crush_device_class**, you can specify per path inside an OSD specification. It is also supported to provide these per-path **crush_device_classes** along with a service-wide **crush_device_class** for the OSD service. In cases of service-wide **crush_device_class**, the setting is considered as default, and the path-specified settings take priority.

Example

```
service_type: osd
  service_id: osd_using_paths
  placement:
    hosts:
      - Node01
      - Node02
  crush_device_class: hdd
  spec:
    data_devices:
      paths:
      - path: /dev/sdb
        crush_device_class: ssd
      - path: /dev/sdc
```

```
    crush_device_class: nvme
  - /dev/sdd
db_devices:
  paths:
  - /dev/sde
wal_devices:
  paths:
  - /dev/sdf
```

Bugzilla:2151189

Cephadm now raises a specific health warning UPGRADE_OFFLINE_HOST when the host goes offline during upgrade

Previously, when upgrades failed due to a host going offline, a generic **UPGRADE_EXCEPTION** health warning would be raised that was too ambiguous for users to understand.

With this release, when an upgrade fails due to a host being offline, Cephadm raises a specific health warning - **UPGRADE_OFFLINE_HOST**, and the issue is now made transparent to the user.

Bugzilla:2152963

All the Cephadm logs are no longer logged into cephadm.log when --verbose is not passed

Previously, some Cephadm commands, such as **gather-facts**, would spam the log with massive amounts of command output every time they were run. In some cases, it was once per minute.

With this release, in Cephadm, all the logs are no longer logged into **cephadm.log** when **--verbose** is not passed. The **cephadm.log** is now easier to read since most of the spam previously written is no longer present.

Bugzilla:2180110

3.2. CEPH DASHBOARD

A new metric is added for OSD blocklist count

With this release, to configure a corresponding alert, a new metric **ceph_cluster_osd_blocklist_count** is added on the Ceph Dashboard.

Bugzilla:2067709

Introduction of ceph-exporter daemon

With this release, **ceph-exporter** daemon is introduced to collect and expose performance counters of all Ceph daemons as Prometheus metrics. It is deployed on each node of the cluster to be performant at large scale clusters.

Bugzilla:2076709

Support force promote for RBD mirroring through Dashboard

Previously, although RBD mirror promote/demote was implemented on the Ceph Dashboard, there was no option to force promote.

With this release, support for force promoting RBD mirroring through Ceph Dashboard is added. If the promotion fails on the Ceph Dashboard, the user is given the option to force the promotion.

Bugzilla:2133341

Support for collecting and exposing the labeled performance counters

With this release, support for collecting and exposing the labeled performance counters of Ceph daemons as Prometheus metrics with labels is introduced.

Bugzilla:2146544

3.3. CEPH FILE SYSTEM

cephfs-top limitation is increased for more client loading

Previously, due to a limitation in the **cephfs-top** utility, only less than 100 clients were loaded at a time, which could not be scrolled and also hung if more clients were loaded.

With this release, **cephfs-top** users can scroll vertically, as well as, horizontally. This enables **cephfs-top** to load nearly 10,000 clients. The users can scroll the loaded clients, and view them on the screen.

Bugzilla:2138793

Users now have the option to sort clients based on the fields of their choice in cephfs-top

With this release, users have the option to sort the clients based on the fields of their choice in **cephfs-top** and also, limit the number of clients to be displayed. This enables the user to analyze the metrics based on the order of fields as per requirement.

Bugzilla:2158689

Non-head omap entries are now included in the omap entries

Previously, a directory fragment would not split if non-head snapshotted entries were not taken into account when deciding to merge or split a fragment. Due to this, the number of **omap** entries in a directory object would exceed a certain limit, and result in cluster warnings.

With this release, non-head **omap** entries are included in the number of **omap** entries when deciding to merge or split a directory fragment to never exceed the limit.

Bugzilla:2159294

3.4. CEPH OBJECT GATEWAY

Objects replicated from another zone now returns the header

With this release, in a multi-site configuration, objects that have replicated from another zone, return the header **x-amz-replication-status=REPLICA**, to allow multi-site users to identify if the object was replicated locally or not.

Bugzilla:1467648

Support for AWS PublicAccessBlock

With this release, Ceph Object Storage supports the AWS public access block S3 APIs such as **PutPublicAccessBlock**.

Bugzilla:2064260

Swift object storage dialect now includes support for SHA-256 and SHA-512 digest algorithms

Previously, support for digest algorithms was added by OpenStack Swift in 2022, but Ceph Object Gateway had not implemented them.

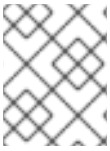
With this release, Ceph Object Gateway's Swift object storage dialect now includes support for **SHA-256** and **SHA-512** digest methods in **tempurl** operations. Ceph Object Gateway can now correctly handle **tempurl** operations by recent OpenStack Swift clients.

Bugzilla:2105950

3.5. MULTI-SITE CEPH OBJECT GATEWAY

Bucket notifications are sent when an object is synced to a zone

With this release, bucket notifications are sent when an object is synced to a zone, to allow external systems to receive information into the zone syncing status at object-level. The following bucket notification event types are added - **s3:ObjectSynced:*** and **s3:ObjectSynced:Created**. When configured with the bucket notification mechanism, a notification event is sent from the synced Ceph Object Gateway upon successful sync of an object.



NOTE

Both topics and the notification configuration should be done separately in each zone from which you would like to see the notification events being sent.

Bugzilla:2053347

Disable per-bucket replication when zones replicate by default

With this release, the ability to disable per-bucket replication when the zones replicate by default, using multisite sync policy, is introduced to ensure that selective buckets can opt out.

Bugzilla:2132554

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. THE CEPHADM UTILITY

Bootstrap no longer fails if a comma-separated list of quoted IPs are passed in as the public network in the initial Ceph configuration

Previously,**cephadm** bootstrap would improperly parse comma-delimited lists of IP addresses, if the list was quoted. Due to this, the bootstrap would fail if a comma-separated list of quoted IP addresses, for example, '172.120.3.0/24,172.117.3.0/24,172.118.3.0/24,172.119.3.0/24', was provided as the **public_network** in the initial Ceph configuration passed to bootstrap with the **--config** parameter.

With this fix, you can enter the comma-separated lists of quoted IPs into the initial Ceph configuration passed to bootstrap for the **public_network** or **cluster_network**, and it works as expected.

Bugzilla:2111680

cephadm no longer attempts to parse the provided yaml files more than necessary

Previously, **cephadm** bootstrap would attempt to manually parse the provided yaml files more than necessary. Due to this, sometimes, even if the user had provided a valid yaml file to **cephadm** bootstrap, the manual parsing would fail, depending on the individual specification, causing the entire specification to be discarded.

With this fix, **cephadm** no longer attempts to parse the yaml more than necessary. The host specification is searched only for the purpose of spreading SSH keys. Otherwise, the specification is just passed up to the manager module. The **cephadm bootstrap --apply-spec** command now works as expected with any valid specification.

Bugzilla:2112309

host.containers.internal entry is no longer added to the /etc/hosts file of deployed containers

Previously, certain podman versions would, by default, add a **host.containers.internal** entry to the **/etc/hosts** file of deployed containers. Due to this, issues arose in some services with respect to this entry, as it was misunderstood to represent FQDN of a real node.

With this fix, Cephadm mounts the host's **/etc/hosts** file when deploying containers. The **host.containers.internal** entry in the **/etc/hosts** file in the containers, is no longer present, avoiding all bugs related to the entry, although users can still see the host's **/etc/hosts** for name resolution within the container.

Bugzilla:2133549

Cephadm now logs device information only when an actual change occurs

Previously, **cephadm`would compare all fields reported for OSDs, to check for new or changed devices. But one of these fields included a timestamp that would differ every time. Due to this, `cephadm** would log that it 'Detected new or changed devices' every time it refreshed a host's devices, regardless of whether anything actually changed or not.

With this fix, the comparison of device information against previous information no longer takes the timestamp fields into account that are expected to constantly change. **Cephadm** now logs only when there is an actual change in the devices.

Bugzilla:2136336

The generated Prometheus URL is now accessible

Previously, if a host did not have an FQDN, the Prometheus URL generated would be *http://_host-shortname*:9095_, and it would be inaccessible.

With this fix, if no FQDN is available, the host IP is used over the shortname. The URL generated for Prometheus is now in a format that is accessible, even if the host Prometheus is deployed on a service that has no FQDN available.

Bugzilla:2153726

cephadm no longer has permission issues while writing files to the host

Previously, cephadm would first create files within the /**tmp** directory, and then move them to their final location. Due to this, in certain setups, a permission issue would arise when writing files, making cephadm effectively unable to operate until permissions were modified.

With this fix, cephadm uses a subdirectory within`/tmp` to write files to the host that do not have the same permission issues.

Bugzilla:2182035

4.2. CEPH DASHBOARD

The default option in the OSD creation step of *Expand Cluster wizard* works as expected

Previously, the default option in the OSD creation step of *Expand Cluster wizard* was not working on the dashboard, causing the user to be misled by showing the option as “selected”.

With this fix, the default option works as expected. Additionally, a “Skip” button is added if the user decides to skip the step.

Bugzilla:2111751

Users can create normal or mirror snapshots

Previously, even though the users were allowed to create a normal image snapshot and mirror image snapshot, it was not possible to create a normal image snapshot.

With this fix, the user can choose from two options to select either normal or mirror image snapshot modes.

Bugzilla:2145104

Flicker no longer occurs on the Host page

Previously, the host page would flicker after 5 seconds if there were more than 1 hosts, causing a bad user experience.

With this fix, the API is optimized to load the page normally and the flicker no longer occurs.

Bugzilla:2164327

4.3. CEPH METRICS

The metrics names produced by Ceph exporter and prometheus manager module are the same

Previously, the metrics coming from the Ceph daemons (performance counters) were produced by the Prometheus manager module. The new Ceph exporter would replace the Prometheus manager module, and the metrics name produced would not follow the same rules applied in the Prometheus manager module. Due to this, the name of the metrics for the same performance counters were different depending on the provider of the metric (Prometheus manager module or Ceph exporter)

With this fix, the Ceph exporter uses the same rules as the ones in the Prometheus manager module to generate metric names from Ceph performance counters. The metrics produced by Ceph exporter and Prometheus manager module are exactly the same.

Bugzilla:2186557

4.4. CEPH FILE SYSTEM

mtime and *change_attr* are now updated for snapshot directory when snapshots are created

Previously, **libcephfs** clients would not update *mtime*, and would change the attribute when snaps were created or deleted. Due to this, NFS clients could not list CephFS snapshots within a CephFS NFS-Ganesha export correctly.

With this fix, *mtime* and *change_attr* are updated for the snapshot directory, **.snap**, when snapshots are created, deleted, and renamed. Correct *mtime* and *change_attr* ensure that listing snapshots do not return stale snapshot entries.

Bugzilla:1975689

cephfs-top -d [--delay] option accepts only integer values ranging between 1 to 25

Previously, **cephfs-top -d [--delay]** option would not work properly, due to the addition of a few new curses methods. The new curses method would accept only integer values, due to which an exception was thrown on getting the float values from a helper function.

With this fix, **cephfs-top -d [--delay]** option accepts only integer values ranging between 1 and 25, and **cephfs-top** utility works as expected.

Bugzilla:2136031

Creating same dentries after the unlink finishes does not crash the MDS daemons

Previously, there was a racy condition between unlink and creating operations. Due to this, if the previous unlink request was delayed due to any reasons, and creating same dentries was attempted during this time, it would fail by crashing the MDS daemons or new creation would succeed but the written content would be lost.

With this fix, users need to ensure to wait until the unlink finishes, to avoid conflict when creating the same dentries.

Bugzilla:2140784

Non-existing cluster no longer shows up when running the **ceph nfs cluster info *CLUSTER_ID*** command.

Previously, existence of a cluster would not be checked when **ceph nfs cluster info *CLUSTER_ID*** command was run, due to which, information of the non-existing cluster would be shown, such as **virtual_ip** and **backend**, null and empty respectively.

With this fix, the `ceph nfs cluster info *CLUSTER_ID*` command checks the cluster existence and an *Error ENOENT: cluster does not exist* is thrown in case a non-existing cluster is queried.

Bugzilla:2149415

The snap-schedule module no longer incorrectly refers to the volumes module

Previously, the snap-schedule module would incorrectly refer to the volumes module when attempting to fetch the subvolume path. Due to using the incorrect name of the volumes module and remote method name, the **ImportError** traceback would be seen.

With this fix, the untested and incorrect code is rectified, and the method is implemented and correctly invoked from the snap-schedule CLI interface methods. The snap-schedule module now correctly resolves the subvolume path when trying to add a subvolume level schedule.

Bugzilla:2153196

Integer overflow and ops_in_flight value overflow no longer happens

Previously, **_calculate_ops** would rely on a configuration option **filer_max_purge_ops**, which could be modified on the fly too. Due to this, if the value of **ops_in_flight** is set to more than **uint64**'s capability, then there would be an integer overflow, and this would make **ops_in_flight** far more greater than **max_purge_ops** and it would not be able to go back to a reasonable value.

With this fix, the usage of **filer_max_purge_ops** in **ops_in_flight** is ignored, since it is already used in **Filer::_do_purge_range()**. Integer overflow and **ops_in_flight** value overflow no longer happens.

Bugzilla:2159307

Invalid OSD requests are no longer submitted to RADOS

Previously, when the first dentry had enough metadata and the size was larger than **max_write_size**, an invalid OSD request would be submitted to RADOS. Due to this, RADOS would fail the invalid request, causing CephFS to be read-only.

With this fix, all the OSD requests are filled with validated information before sending it to RADOS and no invalid OSD requests cause the CephFS to be read-only.

Bugzilla:2160598

MDS now processes all stray directory entries.

Previously, a bug in the MDS stray directory processing logic caused the MDS to skip processing a few stray directory entries. Due to this, the MDS would not process all stray directory entries, causing deleted files to not free up space.

With this fix, the stray index pointer is corrected, so that the MDS processes all stray directories.

Bugzilla:2161479

Pool-level snaps for pools attached to a Ceph File System are disabled

Previously, the pool-level snaps and mon-managed snaps had their own snap ID namespace and this caused a clash between the IDs, and the Ceph Monitor was unable to uniquely identify a snap as to whether it is a pool-level snap or a mon-managed snap. Due to this, there were chances for the wrong

snap to get deleted when referring to an ID, which is present in the set of pool-level snaps and mon-managed snaps.

With this fix, the pool-level snaps for the pools attached to a Ceph File System are disabled and no clash of pool IDs occurs. Hence, no unintentional data loss happens when a CephFS snap is removed.

Bugzilla:2168541

Client requests no longer bounce indefinitely between MDS and clients

Previously, there was a mismatch between the Ceph protocols for client requests between CephFS client and MDS. Due to this, the corresponding information would be truncated or lost when communicating between CephFS clients and MDS, and the client requests would indefinitely bounce between MDS and clients.

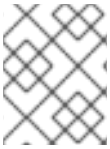
With this fix, the type of the corresponding members in the protocol for the client requests is corrected by making them the same type and the new code is made to be compatible with the old Cephs. The client request does not bounce between MDS and clients indefinitely, and stops after being well retried.

Bugzilla:2172791

A code assert is added to the Ceph Manager daemon service to detect metadata corruption

Previously, a type of snapshot-related metadata corruption would be introduced by the manager daemon service for workloads running Postgres, and possibly others.

With this fix, a code assert is added to the manager daemon service which is triggered if a new corruption is detected. This reduces the proliferation of the damage, and allows the collection of logs to ascertain the cause.



NOTE

If daemons crash after the cluster is upgraded to Red Hat Ceph Storage 6.1, contact *Red Hat support* for analysis and corrective action.

Bugzilla:2175307

MDS daemons no longer crash due to sessionmap version mismatch issue

Previously, MDS sessionmap journal log would not correctly persist when MDS failover occurred. Due to this, when a new MDS was trying to replay the journal logs, the sessionmap journal logs would mismatch with the information in the MDCache or the information from other journal logs, causing the MDS daemons to trigger an assert to crash themselves.

With this fix, trying to force replay the sessionmap version instead of crashing the MDS daemons results in no MDS daemon crashes due to sessionmap version mismatch issue.

Bugzilla:2182564

MDS no longer gets indefinitely stuck while waiting for the cap revocation acknowledgement

Previously, if **__setattrx()** failed, the **_write()** would retain the **CEPH_CAP_FILE_WR** caps reference, the MDS would be indefinitely stuck waiting for the cap revocation acknowledgment. It would also cause other clients' requests to be stuck indefinitely.

With this fix, the **CEPH_CAP_FILE_WR** caps reference is released if the **__setattrx()** fails and MDS' caps revoke request is not stuck.

Bugzilla:2182613

4.5. THE CEPH VOLUME UTILITY

The correct size is calculated for each database device in ceph-volume

Previously, as of RHCS 4.3, **ceph-volume** would not make a single VG with all database devices inside, since each database device had its own VG. Due to this, the database size was calculated differently for each LV.

With this release, the logic is updated to take into account the new database devices with LVM layout. The correct size is calculated for each database device.

Bugzilla:2185588

4.6. CEPH OBJECT GATEWAY

Topic creation is now allowed with or without trailing slash

Previously, http endpoints with one trailing slash in the push-endpoint URL, failed to create a topic.

With this fix, topic creation is allowed with or without trailing slash and it creates successfully.

Bugzilla:2082666

Blocksize is changed to 4K

Previously, Ceph Object Gateway GC processing would consume excessive time due to the use of a 1K blocksize that would consume the GC queue. This caused slower processing of large GC queues.

With this fix, blocksize is changed to 4K, which has accelerated the processing of large GC queues.

Bugzilla:2142167

Timestamp is sent in the multipart upload bucket notification event to the receiver

Previously, no timestamp was sent on the multipart upload bucket notification event. Due to this, the receiver of the event would not know when the multipart upload ended.

With this fix, the timestamp when the multipart upload ends is sent in the notification event to the receiver.

Bugzilla:2149259

Object size and etag values are no longer sent as 0/empty

Previously, some object metadata would not be decoded before dispatching bucket notifications from the lifecycle. Due to this, object size and **etag** values were sent as **0**/**empty** in notifications from lifecycle events.

With this fix, object metadata is fetched and values are now correctly sent with notifications.

Bugzilla:2153533

Ceph Object Gateway recovers from kafka broker disconnections

Previously, if the kafka broker was down for more than 30 seconds, there would be no reconnect after the broker was up again. Due to this, bucket notifications would not be sent, and eventually, after queue fill up, S3 operations that require notifications would be rejected.

With this fix, the broker reconnect happens regardless of the time duration the broker is down and the Ceph Object Gateway is able to recover from kafka broker disconnects.

Bugzilla:2184268

S3 PUT requests with chunked Transfer-Encoding does not require content-length

Previously, S3 clients that PUT objects with **Transfer-Encoding:chunked**, without providing the **x-amz-decoded-content-length** field, would fail. As a result, the S3 PUT requests would fail with **411 Length Required** http status code.

With this fix, S3 PUT requests with chunked Transfer-Encoding need not specify a **content-length**, and S3 clients can perform S3 PUT requests as expected.

Bugzilla:2186760

Users can now configure the remote S3 service with the right credentials

Previously, while configuring remote cloud S3 object store service to transition objects, access keys starting with digit were incorrectly parsed. Due to this, there were chances for the object transition to fail.

With this fix, the keys are parsed correctly. Users cannot configure the remote S3 service with the right credentials for transition.

Bugzilla:2187394

4.7. MULTI-SITE CEPH OBJECT GATEWAY

Bucket attributes are no longer overwritten in the archive sync module

Previously, bucket attributes were overwritten in the archive sync module. Due to this, bucket policy or any other attributes would be reset when **archive zone sync_object()** was executed.

With this fix, ensure to not reset bucket attributes. Any bucket attribute set on source replicates to the archive zone without being reset.

Bugzilla:1937618

Zonegroup is added to the bucket ARN in the notification event

Previously, zonegroup was missing from bucket ARN in the notification event. Due to this, while the notification events handler received events from multiple zone groups, it would cause confusion in the identification of the source bucket of the event.

With this fix, zonegroup is added to the bucket ARN and the notification events handler receiving events from multiple zone groups has all the required information.

Bugzilla:2004175

bucket read_sync_status() command no longer returns a negative ret value

Previously, **bucket read_sync_status()** would always return a negative ret value. Due to this, the bucket sync marker command would fail with **: ERROR: sync.read_sync_status() returned error=0**.

With this fix, the actual ret value from the **bucket read_sync_status()** operation is returned and the bucket sync marker command runs successfully.

Bugzilla:2127926

New bucket instance information are stored on the newly created bucket

Previously, in the archive zone, a new bucket would be created when a source bucket was deleted, in order to preserve the archived versions of objects. The new bucket instance information would be stored in the old instance rendering the new bucket on the archived zone to be in accessible

With this fix, the bucket instance information is stored in the newly created bucket. Deleted buckets on source are still accessible in the archive zone.

Bugzilla:2186774

Segmentation fault no longer occurs when bucket has a num_shards value of 0

Previously, multi-site sync would result in segmentation faults when a bucket had **num_shards** value of **0**. This resulted in inconsistent sync behavior and segmentation fault.

With this fix, **num_shards=0** is properly represented in data sync and buckets with shard value **0** does not have any issues with syncing.

Bugzilla:2187617

4.8. RADOS

Upon querying the IOPS capacity for an OSD, only the configuration option that matches the underlying device type shows the measured/default value

Previously, the **osd_mclock_max_capacity_iops_[ssd|hdd]** values were set depending on the OSD's underlying device type. The configuration options also had default values that were displayed when queried. For example, if the underlying device type for an OSD was SSD, the default value for the HDD option, **osd_mclock_max_capacity_iops_hdd**, was also displayed with a non-zero value. Due to this, displaying values for both HDD and SSD options of an OSD when queried, caused confusion regarding the correct option to interpret.

With this fix, the IOPS capacity-related configuration option of the OSD that matches the underlying device type is set and the alternate/inactive configuration option is set to **0**. When a user queries the IOPS capacity for an OSD, only the configuration option that matches the underlying device type shows the measured/default value. The alternative/inactive option is set to 0 to clearly indicate that it is disabled.

Bugzilla:2111282

4.9. RBD MIRRORING

Error message when enabling image mirroring within a namespace now provides more insight

Previously, attempting to enable image mirroring within a namespace would fail with a "cannot enable mirroring in current pool mirroring mode" error. The error would neither provide insight into the problem nor provide any solution.

With this fix, to provide more insight, the error handling is improved and the error now states "cannot enable mirroring: mirroring is not enabled on a namespace".

Bugzilla:2024444

Snapshot mirroring no longer halts permanently

Previously, if a primary snapshot creation request was forwarded to rbd-mirror daemon when the rbd-mirror daemon was axed for some practical reason before marking the snapshot as complete, the primary snapshot would be permanently incomplete. This is because, upon retrying that primary snapshot creation request, **librbd** would notice that such a snapshot already existed. It would not check whether this "pre-existing" snapshot was complete or not. Due to this, the mirroring of snapshots was permanently halted.

With this fix, as part of the next mirror snapshot creation, including being triggered by a scheduler, checks are made to ensure that any incomplete snapshots are deleted accordingly to resume the mirroring.

Bugzilla:2120624

CHAPTER 5. KNOWN ISSUES

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

5.1. CEPH OBJECT GATEWAY

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

5.1.1. Ceph Object Gateway

Bucket lifecycle processing might get delayed if the Ceph Object Gateway instances are killed due to a crash or kill

Presently, if a Ceph Object Gateway instance is killed (ungraceful shutdown) due to a crash or kill -9 while lifecycle processing for one or more buckets is taking place, processing might not continue on those buckets until two scheduling periods have elapsed, for example, two days. At this point, the buckets are marked stale and reinitialized. There is no workaround for this issue.

Bugzilla:2072680

CHAPTER 6. ASYNCHRONOUS ERRATA UPDATES

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

6.1. RED HAT CEPH STORAGE 6.1Z7

Red Hat Ceph Storage release 6.1z7 is now available. The bug fixes that are included in the update are listed in the [advisory links] advisories.

6.1.1. Enhancements

6.1.1.1. Ceph File System

New clone creation no longer slows down due to parallel clone limit

Previously, upon reaching the limit of parallel clones, the rest of the clones would queue up, slowing down the cloning.

With this enhancement, upon reaching the limit of parallel clones at a time, the new clone creation requests are rejected. This feature is enabled by default but can be disabled.

Bugzilla:2196829

The Python librados supports iterating object omap key/values

Previously, the iteration would break whenever a binary/unicode key was encountered.

With this release, the Python librados supports iterating object omap key/values with unicode or binary keys and the iteration continues as expected.

Bugzilla:2232161

6.1.1.2. Ceph Object Gateway

Improved temporary file placement and error messages for the with the '/usr/bin/rgw-restore-bucket-index' tool

Previously, the '/usr/bin/rgw-restore-bucket-index' tool only placed temporary files into the '/tmp' directory. This could lead to issues if the directory ran out of space, resulting in the following error messing being emitted: "ln: failed to access '/tmp/rgwrbi-object-list.XXX': No such file or directory".

With this enhancement, users can now specify a specific directory to place temporary files, by using the '-t' command-line option. Additionally, if the specified directory is full, users now receive an error message specifying the problem: "ERROR: the temporary directory's partition is full, preventing continuation".

Bugzilla:2270322

S3 requests are no longer cut off in the middle of transmission during shutdown

Previously, a few clients faced issues with the S3 request being cut off in the middle of transmission during shutdown without waiting.

With this enhancement, the S3 requests can be configured to wait for the duration defined in the **rgw_exit_timeout_secs** parameter for all outstanding requests to complete before exiting the Ceph Object Gateway process unconditionally. Ceph Object Gateway will wait for up to 120 seconds

(configurable) for all on-going S3 requests to complete before exiting unconditionally. During this time, new S3 requests will not be accepted. This configuration is off by default.

NOTE

In containerized deployments, an additional **extra_container_args** parameter configuration of **--stop-timeout=120** (or the value of **rgw_exit_timeout_secs** parameter, if not default) is also necessary.top-timeout=120` (or the value of **rgw_exit_timeout_secs** parameter, if not default) is also necessary.

Bugzilla:2298708

6.1.1.3. RADOS

New 'mon_cluster_log_level' command option to control the cluster log level verbosity for external entities

Previously, debug verbosity logs were sent to all external logging systems regardless of their level settings. As a result, the '/var/' filesystem would rapidly fill up.

With this enhancement, 'mon_cluster_log_file_level' and 'mon_cluster_log_to_syslog_level' command options have been removed. From this release, use only the new generic 'mon_cluster_log_level' command option to control the cluster log level verbosity for the cluster log file and all external entities.

Bugzilla:2053021

6.2. RED HAT CEPH STORAGE 6.1Z6

Red Hat Ceph Storage release 6.1z6 is now available. The bug fixes that are included in the update are listed in the RHBA-2024:2631 and RHSA-2024:2633 advisories.

6.3. RED HAT CEPH STORAGE 6.1Z5

Red Hat Ceph Storage release 6.1z5 is now available. The bug fixes that are included in the update are listed in the RHBA-2024:1580 and RHBA-2024:1581 advisories.

6.3.1. Enhancements

6.3.1.1. The Ceph Ansible utility

All bootstrap CLI parameters are now made available for usage in the cephadm-ansible module

Previously, only a subset of the bootstrap CLI parameters were available and it was limiting the module usage.

With this enhancement, all bootstrap CLI parameters are made available for usage in the **cephadm-ansible** module.

Bugzilla:2269514

6.3.1.2. RBD Mirroring

RBD diff-iterate now executes locally if exclusive lock is available

Previously, when diffing against the beginning of time (**fromsnapname == NULL**) in fast-diff mode (**whole_object == true** with **fast-diff** image feature enabled and valid), RBD diff-iterate wasn't guaranteed to execute locally.

With this enhancement, **rbd_diff_iterate2()** API performance improvement is implemented and RBD diff-iterate is now guaranteed to execute locally if exclusive lock is available. This brings a dramatic performance improvement for QEMU live disk synchronization and backup use cases assuming **fast-diff** image feature is enabled.

Bugzilla:2259053

6.3.1.3. Ceph File System

Snapshot scheduling support is now provided for subvolumes

With this enhancement, snapshot scheduling support is provided for subvolumes. All snapshot scheduling commands accept **--subvol** and **--group** arguments to refer to appropriate subvolumes and subvolume groups. If a subvolume is specified without a subvolume group argument, then the default subvolume group is considered. Also, a valid path need not be specified when referring to subvolumes and just a placeholder string is sufficient due to the nature of the argument parsing employed.

Example

```
# ceph fs snap-schedule add - 15m --subvol sv1 --group g1
# ceph fs snap-schedule status - --subvol sv1 --group g1
```

Bugzilla:2243783

6.4. RED HAT CEPH STORAGE 6.1Z4

Red Hat Ceph Storage release 6.1z4 is now available. The bug fixes that are included in the update are listed in the RHBA-2024:0747 advisory.

6.4.1. Enhancements

6.4.1.1. Ceph File System

MDS dynamic metadata balancer is off by default.

With this enhancement, MDS dynamic metadata balancer is **off** by default to improve the poor balancer behavior that could fragment trees in undesirable or unintended ways simply by increasing the **max_mds** file system setting.

Operators must turn **on** the balancer explicitly to use it.

Bugzilla:2255435

The resident segment size perf counter in the MDS is tracked with a higher priority.

With this enhancement, the MDS resident segment size (or RSS) **perf counter** is tracked with a higher priority to allow callers to consume its value to generate useful warnings which is useful for rook to ascertain the MDS RSS size and act accordingly.

Bugzilla:2256731

ceph auth commands give a message when permissions in MDS are incorrect

With this enhancement, permissions in MDS capability now either start with **r**, **rw,** `* or **all**. This results in the **ceph auth** commands like **ceph auth add**, **ceph auth caps**, **ceph auth get-or-create** and **ceph auth get-or-create-key** to generate a clear message when the permissions in the MDS caps are incorrect..

Bugzilla:2218189

6.4.1.2. Ceph Object Gateway

The radosgw-admin bucket stats 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:2256364

6.5. RED HAT CEPH STORAGE 6.1Z3

Red Hat Ceph Storage release 6.1z3 is now available. The bug fixes that are included in the update are listed in the RHSA-2023:7740 advisory.

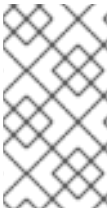
6.5.1. Enhancements

6.5.1.1. Ceph File System

The snap schedule module now supports a new retention specification

With this enhancement, users can define a new retention specification to retain the number of snapshots.

For example, if a user defines to retain 50 snapshots irrespective of the snapshot creation cadence, the snapshots retained is 1 less than the maximum specified as the pruning happens after a new snapshot is created. In this case, 49 snapshots are retained so that there is a margin of 1 snapshot to be created on the file system on the next iteration to avoid breaching the system configured limit of **mds_max_snaps_per_dir**.



NOTE

> Configure the `mds_max_snaps_per_dir` and snapshot scheduling carefully to avoid unintentional deactivation of snapshot schedules due to file system returning a "Too many links" error if the `mds_max_snaps_per_dir limit` is breached.

Bugzilla:2227807

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 (cannot flush dirty data for cap revokes).

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

Bugzilla:2228065

6.5.1.2. 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:2182385

6.5.1.3. NFS Ganesha

NFS Ganesha version updated to V5.6

With this enhancement of updated version of NFS Ganesha, following issues have been fixed: * FSAL's **state_free`** function called by **free_state** did not actually free. * CEPH: Fixed **cmount_path**. * CEPH: Currently, the **client_oc true** was broken, forced it to **false**.

Bugzilla:2249958

6.5.1.4. RADOS

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:2240838

Turning the noautoscale flag on/off now retains each pool's original autoscale mode configuration.

Previously, the **pg_autoscaler** did not persist in each pool's **autoscale mode** configuration when the **noautoscale** flag was set. Due to this, after turning the **noautoscale** flag on/off, the user would have to go back and set the autoscale mode for each pool again.

With this enhancement, the **pg_autoscaler** module persists individual pool configuration for the autoscaler mode after the **noautoscale flag** is set.

Bugzilla:2241201

BlueStore instance cannot be opened twice

Previously, when using containers, it was possible to create unrelated inodes that targeted the same block device **mknod b** causing multiple containers to think that they have exclusive access.

With this enhancement, the reinforced advisory locking with **O_EXCL** open flag dedicated for block devices is now implemented,thereby improving the protection against running OSD twice at the same time on one block device.

Bugzilla:2239449

6.6. RED HAT CEPH STORAGE 6.1Z2

Red Hat Ceph Storage release 6.1z2 is now available. The bug fixes that are included in the update are listed in the RHSA-2023:5693 advisory.

6.6.1. Enhancements

6.6.1.1. Ceph Object Gateway

Additional features and enhancements are added to rgw-gap-list and rgw-orphan-list scripts to enhance end-users' experience

With this enhancement, to improve the end-users' experience with rgw-gap-list and rgw-orphan-list scripts, a number of features and enhancements have been added, including internal checks, more command-line options, and enhanced output.

Bugzilla:2228242

Realm, zone group, and/or zone can be specified when running the rgw-restore-bucket-index command

Previously, the tool could only work with the default realm, zone group, and zone.

With this enhancement, realm, zone group, and/or zone can be specified when running the **rgw-restore-bucket-index** command. Three additional command-line options are added:

- "-r <realm>"
- "-g <zone group>"
- "-z <zone>"

Bugzilla:2183926

6.6.1.2. Multi-site Ceph Object Gateway

Original multipart uploads can now be identified in multi-site configurations

Previously, a data corruption bug was fixed in 6.1z1 release that effected multi-part uploads with server-side encryption in multi-site configurations.

With this enhancement, a new tool, **radosgw-admin bucket resync encrypted multipart**, can be used to identify these original multipart uploads. The LastModified timestamp of any identified object is incremented by 1ns to cause peer zones to replicate it again. For multi-site deployments that make any use of Server-Side encryption, users are recommended to run this command against every bucket in every zone after all zones have upgraded.

Bugzilla:2227842

6.6.1.3. Ceph Dashboard

Dashboard host loading speed is improved and pages now load faster

Previously, large clusters of five or more hosts have a linear increase in load time on hosts page and main page.

With this enhancement, the dashboard host loading speed is improved and pages now load orders of magnitude faster.

Bugzilla:2220922

6.7. RED HAT CEPH STORAGE 6.1Z1

Red Hat Ceph Storage release 6.1z1 is now available. The bug fixes that are included in the update are listed in the RHBA-2023:4473 advisory.

6.7.1. Enhancements

6.7.1.1. Ceph File System

Switch the unfair Mutex lock to fair mutex

Previously, the implementations of the *Mutex*, for example, **std::mutex** in *C++*, would not guarantee fairness and would not guarantee that the lock would be acquired by threads in the order called **lock()**. In most cases, this worked well but in an overloaded case, the client requests handling thread and *submit* thread would always successfully acquire the *submit_mutex* in a long time, causing **MDLog::trim()** to get stuck. That meant the MDS daemons would fill journal logs into the metadata pool, but could not trim the expired segments in time.

With this enhancement, the unfair Mutex lock is switched to fair mutex and all the *submit_mutex* waiters are woken up one by one in FIFO mode.

Bugzilla:2158304

6.7.1.2. Ceph Object Gateway

The bucket listing feature enables the *rgw-restore-bucket-index* tool to complete reindexing

Previously, the *rgw-restore-bucket-index* tool would restore the bucket's index partially until the next user listed out the bucket. Due to this, the bucket's statistics would report incorrectly until the reindexing completed.

With this enhancement, the bucket listing feature is added which enables the tool to complete the reindexing and the bucket statistics are reported correctly. Additionally, a small change to the build process is added that would not affect end-users.

Bugzilla:2182456

Lifecycle transition no longer fails for objects with modified metadata

Previously, setting an ACL on an existing object would change its **mtime** due to which lifecycle transition failed for such objects.

With this fix, unless it is a copy operation, the object's **mtime** remains unchanged while modifying just the object metadata, such as setting ACL or any other attributes.

Bugzilla:2213801

Blocksize is changed to 4K

Previously, Ceph Object Gateway GC processing would consume excessive time due to the use of a 1K blocksize that would consume the GC queue. This caused slower processing of large GC queues.

With this fix, blocksize is changed to 4K, which has accelerated the processing of large GC queues.

Bugzilla:2212446

Object map for the snapshot accurately reflects the contents of the snapshot

Previously, due to an implementation defect, a stale snapshot context would be used when handling a write-like operation. Due to this, the object map for the snapshot was not guaranteed to accurately reflect the contents of the snapshot in case the snapshot was taken without quiescing the workload. In differential backup and snapshot-based mirroring, use cases with object-map and/or fast-diff features enabled, the destination image could get corrupted.

With this fix, the implementation defect is fixed and everything works as expected.

Bugzilla:2216186

6.7.1.3. The Cephadm Utility

public_network parameter can now have configuration options, such as global or mon

Previously, in **cephadm**, the **public_network** parameter was always set as a part of the **mon** configuration section during a cluster bootstrap without providing any configuration option to alter this behavior.

With this enhancement, you can specify the configuration options, such as **global** or **mon** for the **public_network** parameter during cluster bootstrap by utilizing the Ceph configuration file.

Bugzilla:2156919

The Cephadm commands that are run on the host from the cephadm Manager module now have timeouts

Previously, one of the Cephadm commands would occasionally hang indefinitely, and it was difficult for users to notice and sort the issue.

With this release, timeouts are introduced in the Cephadm commands that are run on the host from the Cephadm mgr module. Users are now alerted with a health warning about eventual failure if one of the commands hangs. The timeout is configurable with the **mgr/cephadm/default_cephadm_command_timeout** setting, and defaults to 900 seconds.

Bugzilla:2151908

cephadm support for CA signed keys is implemented

Previously, CA signed keys worked as a deployment setup in Red Hat Ceph Storage 5, although their working was accidental, untested, and broken in changes from Red Hat Ceph Storage 5 to Red Hat Ceph Storage 6.

With this enhancement, **cephadm** support for CA signed keys is implemented. Users can now use CA signed keys rather than typical *pubkeys* for SSH authentication scheme.

Bugzilla:2182941

6.7.2. Known issues

6.7.2.1. Multi-site Ceph Object Gateway

Deleting objects in versioned buckets causes statistics mismatch

Due to versioned buckets having a mix of current and non-current objects, deleting objects might cause bucket and user statistics discrepancies on local and remote sites. This does not cause object leaks on either site, just statistics mismatch.

Bugzilla:1871333

Multisite replication may stop during upgrade

Multi-site replication may stop if clusters are on different versions during the process of an upgrade. We would need to suspend sync until both clusters are upgraded to the same version.

Bugzilla:2178909

CHAPTER 7. SOURCES

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

- For Red Hat Enterprise Linux 8:
  http://ftp.redhat.com/redhat/linux/enterprise/8Base/en/RHCEPH/SRPMS/
- For Red Hat Enterprise Linux 9:
  https://ftp.redhat.com/redhat/linux/enterprise/9Base/en/RHCEPH/SRPMS/