Ceph Quarterly

Issue # 3 *An overview of the past three months of Ceph upstream development.* Jan. 2024

Pull request (PR) numbers are provided for many of the items in the list below. To see the PR associated with a list item, append the PR number to the string `https://github.com/ceph/ceph/pull/` . For example, to see the PR for the first item in the left column below, append the string `53597` to the string `https://github.com/ceph/ceph/pull/` to make this string: `https://github.com/ceph/ceph/pull/53597` .

BlueStore

1. Improvements to BlueStoreFS performance counters, improvements to file_map handling: **53597**
2. Ensure correct sequence number during log extensions: **53732**
3. Handle situations where zone_ size is set to 0 without crashing: **54155**
4. Free space is now updated after "bedev-expand" in NCB mode: **54990**

CephFS

1. Upgrades are now allowed even if no MDSes are up: **53600**
2. Improve journal handling around segment boundaries: **53494**
3. MDS: Add an option for configuring the number of epochs the overload layer lasts before migrating, to avoid frequent migrations: **53332**
4. Disallow delegating preallocated inode ranges to clients: **53836**
5. Clean up FS commands: **53719**
6. Improve fragset handling during scrubs: **53636**
7. Add an FS command that enables users to swap names of two file systems in a single PAXOS transaction: **50212**
8. Ensure peon pending_metadata consistency with mon's db: **53883**
9. Increase accepted xattr value length to 64K (cephfs-shell): **53126**
10. Improve laggy client reporting and hygiene due to laggy OSD: **53839**
11. Improve MDS understanding of filesystems with respect to mountpoints: **53887**
12. Improve client request replay handling: **47121**
13. Ignore queued callbacks when MDS is inactive: **54178**
14. Require CephFS to be offline during renaming: **53899**
15. More CephFS renaming-while-FS-is-offline changes: **53899**
16. Add an FS command that enables users to swap the names of two file systems in a single PAXOS trans: **50212**
17. Improve loner capability defaults, setting loner to true for LOCK_ EXCL_XSYN: **54669**
18. Improve lock handling during transition from LOCK_MIX to LOCK_ SYNC: **54687**
19. Improve MDS metrics: **54825**
20. Improve handling when TID is not present in pending_notifies: **54958**
21. Correct misleading error message (kernel mount command-related): **54972**
22. MDS: Improve journal handling around segment boundaries: **53494**
23. MDS: Improve fragset handling during scrubs: **53636**
24. MDS: Improve laggy client reporting and hygiene due to laggy OSD: **53839**
25. MDS: Improve client request replay handling: **47121**

Cephadm

1. Introduce "Daemon Forms" to standardize "daemon-type" classes and instances: **53621**
2. Improve behavior of nvmeof during upgrades: **53862**
3. Improve handling of Podman-related operations: **54081**
4. Add command "cephadm unit-install" to install systemd units only: **53987**
5. Add a --dry-run option to cephadm shell and improve debugging: **53976**
6. Improve firewalld handling by respecting the --skip-firewalld flag: **54158**
7. Add the ability to bundle in dependencies derived from installed RPMs as well the ability to disable bundling: **54173**
8. Make sure that OSD weights are restored when OSD removal is stopped: **54401**
9. Warn when a user drains a host that is explicitly listed in the placement: **54046**
10. Move logrotate configs to jinja2 template: **54378**
11. Provide support for regex-based host patterns: **53803**
12. Reorganize container setup methods (cephadm refactoring project): **54403**
13. Reorganize container mounts functions (cephadm refactoring project): **54399**
14. Expose nvmeof gateway configuration parameters through specifications: **54490**
15. Improve handling relative paths when jinja2 is in play: **54487**
16. Remove circular dependencies that prevented moving the daemon-type classes to files in cephadmlib: **54424**
17. Ensure that drivespec counts existing devices correctly: **54681**

18. Improve the core of cephadm (many commits): **54441**
19. Simplify loading jinja2 templates from zipapp: **54773**
20. Add rudimentary SMB daemon: **54817**
21. Improve build.py script's handling of dependencies: **54889**
22. Improve handling of Python when multiple versions of Python are installed: **54948**
23. Ensure that jaeger-collector is a dependency of jaeger-agent, so that jaeger-agents are aware of the URLs of collectors: **51416**
24. Improve handling of removal of host entries from the CRUSH map: **53737**
25. Improve discovery service for ipv6-only clusters: **54285**

ceph-volume

1. Improve the listing of Logical Volumes within multiple Volume Groups: **53841**
2. Improve partition reporting: **53804**
3. ceph-volume no longer crashes when given a nonexistent volume: **54392**

Client

1. Improve handling of client_lock: **53822**
2. Properly handle ceph_mds_request_head_legacy (avoid crashes in legacy kernels): **54149**
3. Improve copying bufferlist to iovec structures: **54808**

Common

1. Resolve config proxy deadlock by proper refcounting "uses" of observers (improve lock handling): **53568**
2. Add "remove key" and "get key" commands to ceph-monstore-tool: **52855**
3. Improve KVTest.RocksDV_estimate_size tests: **53909**
4. Make it possible to turn off RocksDB's use of liburing: **54042**
5. Allow install_deps.sh to continue even if ceph-libbost install fails: **54226**
6. Improve the use of fmt during the build process: **54216**
7. Add "--progress" flag to git submodule update commands: **54210**
8. Fix the behavior of blkid() in cases where the key doesn't exist: **54453**
9. Fix a regression in "ceph-volume raw list" that broke Rook's OSD creation: **54514**
10. Distinguish between device and partition: **53798**
11. Support Windows Unicode CLI (translate between ANSI or UTF-16 and UTF-8): **53291**
12. Build dependencies before using RocksDB: **54626**
13. Improve LogClient efficiency: **54780**

Documentation

1. Add a "Monitoring" section explaining reported metrics: **54099**

Erasure Code

1. Add neon-based region_xor implementation: **54882**

MGR

1. (*dashboard*) Improve layout of the landing page: **53522**
2. (*dashboard*) Add configuration option to allow TLS 1.2: **53699**
3. (*dashboard*) Hide usage bar when disk usage is not provided: **53746**
4. (*dashboard*) Remind users to restart RGW daemons after migrating from "single site" to "multi site": **53673**
5. (*dashboard*) A "protect option" is enabled when layering is enabled: **53671**
6. (*dashboard*) Improve listing of CephFS snapshots in directories: **53700**
7. (*dashboard*) Improve text for "Multi-site": **53307**
8. (*dashboard*) Improve reporting of "out" or "down" OSDs in Grafana dashboard: **53650**
9. (*dashboard*) Fix broken alert generator when running "tox -ealerts-fix": **53653**
10. (*mgr/nfs*) Enable reporting err status in Responder decorator, enhance err reporting in "nfs export apply": **53431**
11. (*Orchestrator*) block OSD specs that have no service ID: **54812**
12. Add a throttle policy for DaemonServer: **52509**
13. Improve service_to_daemon_types lookup (to improve Rook-related lookups): **53910**
14. Improve "fs subvolume group rm" error message: **53651**
15. Clean up stale OSD "out" and "down" warnings in "ceph -s" output: **53993**
16. Improve timestamping of similar daemon and service events: **54077**
17. Remove references to hard-coded "ceph-rook" namespace: **54151**
18. Improve handling of CreateSnapshotRequests lock: **54251**
19. Remove unreachable and therefore unnecessary shutdown code: **54319**
20. Improve comments related to cloning: **54616**
21. Process monitor commands before lower-priority issues: **54627**
22. Improve flake8 warnings: **54308**
23. Improve handling of Debian bcrypt package requirements: **53290**
24. Various minor improvements to mgr/vol (grab bag): **54616**
25. Improve the listing of physical devices on Rook: **54676**
26. Drop mon_host from peer_list to make it similar to rbd_mirror peer_list: **54682**
27. Add check for "norecover" flag: **54708**
28. Improve Ceph Balancer status output: **54801**
29. Improve handling of period realm_name: **54864**
30. Show correct number of CPU threads when "host ls --detail" is run: **54977**

MON

1. Prevent newly-revived monitors from getting stuck in election because they have different "disallowed_leaders" set: **53979**
2. A new "raise" command allows sending a signal to a daemon: **53689**
3. ConfigMap unit testing and cleanups: **52406**
4. Improve logging to improve handling of CRC mismatch: **54890**

RADOS

1. Improve "rados clearomap" error messages: **54518**
2. Improve handling of nonexistent pgmeta object attributes in empty pools: **54663**

RBD

1. Improve lock handling in librbd: **53829**
2. (*librbd*) Add image and object information to deep_copy logs to ease debugging: **54222**
3. (*librbd*) Improve handling of http stream imports (include ASIO strand headers in librbd.cc): **54839**
4. C-related parts of Ceph can now be built using GCC 14: **54974**
5. Recovery from "watch errors" improved: **53735**

RGW

1. Migration has been improved through the addition of timestamps: **53580**
2. Improved bucket validation against POST policies – addresses **https://nvd.nist.gov/vuln/detail/CVE-2023-34040**: **53714**
3. Repair "unwatch crash" at RGW startup: **53691**
4. (*rgw/async*) use optional_yield for keystone and kms requests: **53684**
5. Improve S3 test case handling when access keys are found in Keystone: **53680**
6. Fix decode_json max_read_bytes and max_write_bytes field mismatch: **53614**
7. Improve S3website data handling by avoiding prefetching data: **53602**
8. Improve buffer list utilization in the "chunkupload scenario": **53266**
9. Improve the handling of bucket deletion: **52960**
10. Remove settable attributes from buckets: **52144**
11. Bucket and object names are now part of the log line collected by "rgw ops log": **50350**
12. Modular restructure of cephadm: **53703**
13. Add versioning information to radosgw-admin bucket output: **53813**
14. Fix http error checks in Keystone, Barbican, and Vault clients: **53846**
15. Support the reloading of Lua packages on all RGWs: **52326**
16. Add request context structure: **53547**
17. Add a perf counters cache per CephContext that acts as a wrapper around perf counters for storing and modifying labeled perf counters per CephContex, and add RGW-related instrumentation: **53003**
18. Improve s3select output: **53351**
19. Improve the RGW "swift info" command: **54050**
20. Handle indexless buckets during dynamic resharding: **53929**
21. Improve RabbitMQ and Kafka handling in Centos 9: **54025**
22. Honor maximum session duration when creating roles: **53842**
23. Improve topic policies: **53848**
24. Improve size accounting in bucket index for compressed or encrypted objects: **54174**
25. Remove rgw-policy-check for non-empty tenant: **54213**
26. Add "index generation" field to bucket statistics: **54197**
27. Improve Kafka processing, especially when the broker is down (Bloomberg): **54215**
28. Improve the handling of leading hyphens in handle instance names: **54257**
29. Remove unnecessary realm name from period configuration: **54264**
30. Prevent illegal behavior due to mismatched allocators when run under nfs-ganesha or other consumers: **54297**
31. Make buckets discoverable across tenants: **54299**
32. Improve error logging after "bi list": **54196**
33. Add perf counters for persistent topic stats: **54147**
34. Zipper project: Make subclass declarations explicit: **54448**
35. Improve flight load_bucket call: **54447**
36. Clean up croutines after sending notifications: **54528**
37. Prevent lock_lambda from overwriting the value of "ret": **47208**
38. Improve tracking of AWS S3 multi-part uploads: **50148**
39. Improve handling of lifecycle work times when work time spans two calendar days: **54622**
40. Wait for Kafka before publishing RGW notifications: **54637**
41. Improve handling of dashes and underscores in Swift user metadata names: **50790**
42. Improve cache updating when 304 responses are received: **54587**
43. Improve RADOS's current support for the old "RGW 2pc remove": **54459**
44. "cephadm ls" now lists legacy RGWs: **54679**
45. Kafka: fix race condition in asynchronous completion handlers: **54697**
46. Remove RGWSI_RADOS to eliminate RADOS-handle redundancy: **50359**
47. Make "post obj" return the response etag field (AWS-related): **54546**
48. Handle invalid IP raw endpoint data: **54668**

49. Improve handling of configuration of old pools: **54749**
50. Improve logging of lifecycle transitions and deletes: **54759**
51. Fix a race condition in RGW shutdown: **54810**
52. Support GetBucketLogging and PutBucketLogging: **54815**
53. Improve propagation of role deletion from primary zone: **54829**
54. Prevent "radosgw-admin zone set" from overwriting an existing default placement target: **54903**
55. The "connection idle" timeout is now configurable: **55022**

OSD

1. Improve handling of "fast shutdown" timeouts: **53530**
2. Improve read balancer logic w/r/t PGs: **53449**
3. Add dump_osd_pg_stats admin socket command: **53563**
4. Improve speed of scheduling when item_cost is large: **53417**
5. (*rgw*) Add a wrapper for librados::AioCompletion to prevent memory leaks: **52276**
6. Improve OSD handling of purged snapshots: **53579**
7. Don't increment snap_seq when a snapshot is removed: **54024**
8. Improve user policy attribute handling: **53750**
9. Add subuser to user policy condition check: **53997**
10. Improve handling of replica reservation at the start of scrubs: **53843**
11. Improve _set_cache_sizes ratio: **51784**
12. Extend scrub reservation timeout: **53843**
13. Improve performance of OSD SnapMapper: **53132**
14. Improve deep scrub functionality: **54363**
15. Improve scheduler behavior: **53524**
16. Fix ambiguous call to format_to(): **54529**
17. Improve scrub behavior (replica reservation process): **54482**
18. Log the number of extents: **54571**
19. Allow auto-repair on operator-initiated scrubs: **54615**
20. Improve PG object size estimates for mClock scheduler uses: **54597**
21. Decouple being reserved from the handling of scrub requests: **54482**
22. Improve OSD trim maps: **54686**
23. Improve deduplication by making data promotion smarter: **54740**
24. Improve handling of local scrubs: **54828**
25. Display oldest_map and newest_map in "ceph daemon osd.x status" output: **54913**
26. Unhandled scrub backend errors now cause an abort: **54982**
27. Add a "clean primary" base state to the scrubber state machine: **54996**

Upstream Infrastructure

1. Improve test output by adding an explicit "return 0" in a case where it was needed. This was part of Grace Hopper Open Source Day 2023: **62855**, **53606**

Vstart

1. Exclude default route during vstart cluster setup: **53782**
2. Improve vstart networking: **53319**
3. Remove the redundant RGWSI_RADOS: **50359**

In Development

The PRs listed here refer to features that are in development during the first quarter of 2024. We list them here to attract attention to them and to invite interest in their development.

1. Efficient hardlink management as a precursor work for snapshot layering support: **49945**
2. Crash-consistent snapshot support: **54485**, **54581**

Grace Hopper Open Source Day 2023:

On 22 Sep 2023, Ceph participated in Grace Hopper Open Source Day, an all-day hackathon for women and nonbinary developers. Laura Flores led the Ceph division, and Yaarit Hatuka, Shreyansh Sancheti, and Aishwarya Mathuria participated as mentors. From 12pm EST to 7:30pm EST, Laura showed more than 40 attendees how to run a Ceph vstart cluster in an Ubuntu Docker container. Yaarit, Shreyansh, and Aishwarya spent the day working one-on-one with attendees, helping them troubleshoot and work through a curated list of low-hanging-fruit issues. By the day's end, Grace Hopper attendees submitted eight pull requests. As of the publication of this sentence, two have been merged and the others are expected to be merged soon.

For more information about GHC Open Source Day, see **https://ghc.anitab.org/awards-programs/open-source-day/**

Ceph partners with RCOS:

Ceph has partnered for the first time with the Rensselaer Center for Open Source (RCOS), an organization at Rensselaer Polytechnic Institute that helps students jumpstart their careers in software by giving them the opportunity to work on various open source projects for class credit.

Laura Flores, representing Ceph, is mentoring three RPI students on a project to improve the output of the `ceph balancer status` command.

For more information about RCOS, see **https://rcos.io/**

User+Dev Monthly Talks:

October

- "Crush Changes at Scale" by Joshua Baergen of Digital Ocean
- "CephFS Management with Ceph Dashboard" by Pedro Gonzalez Gomez

November

- "Ceph Beginner's Guide" by Zac Dover of The Ceph Foundation

December

- [No talk in December]

For more information about these talks, see **https://pad.ceph.com/p/ceph-user-dev-monthly-minutes**

CQ is a production of the Ceph Foundation. To support or join the Ceph Foundation, contact **membership@linuxfoundation.org**.

Send all inquiries and comments to Zac Dover at zac.dover@proton.me