

Ceph Quarterly

Issue # 5 *An overview of the past three months of Ceph upstream development.* Jul. 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. Add a drain wait period that waits for all discarded entries in the worker-threads private space to be committed. Limit the private-space entries to ten entries in order to speedily end the drain wait: **56744**
2. Check 'it' valid before using: **56854**
3. Improve BlueFS allocation handling: **57015**
4. Correct the description of the bluefs stats output: **58079**
5. Remove zoned namespace support: **58192**

CephFS

1. Initialize mds_cache_quiesce_decay_rate from the value in quiesce_counter: **56723**
2. (MDS) - Reduce log level from 7 to 5 while the MDS is stopping. This reduces the likelihood of the MDS lagging while stopping: **56629**
3. (MDS) - Let clients keep their buffered writes for a quiesced file: **56755**
4. (MDS) - Don't assume all non-auth xlocks are remote: **57020**
5. (MDS) - Add missing policylock to test F_QUIESCE_BLOCK: **57010**
6. (MDS) - Remove the incorrect std::move for fsname and path: **56981**
7. (MDS) - agent: Avoid a race condition with rapid db updates: **56956**
8. (MDS) - Find a new head for the batch ops when the head is dead: **56941**
9. (MDS) - Regular file inode flags are not replicated by the policylock: **56935**
10. (MDS) - inode_t flags not protected by the policylock during set_vxattr: **56934**
11. (MDS) - db: quiesce-await should EPERM if a set is past QS_QUIESCED: **57099**
12. (MDS) - Use regular dispatch to process metrics to avoid MetricAggregator::lock contention during fast dispatch of client metrics: **57081**
13. (MDS) - Abort fragment/export when quiesced: **57059**
14. (MDS) - Improve quiesce: **57332**
15. (MDS) - Don't stall the asok thread for flush commands: **57274**
16. (MDS) - Check relevant caps for fs include root_squash: **57192**
17. (MDS) - quiesce-db: Track DB epoch separately from the membership epoch: **57454**
18. (MDS) - Relax divergent backtrace scrub failures for replicated ancestor inodes: **57354**
19. (MDS) - Fix timeouts, a crash, and overdrive a tree export when possible: **57579**
20. (MDS) - When head is dead, try to choose a new batch head in request_clientup(): **57553**
21. (MDS) - quiesce-db: Collect acks while bootstrapping to avoid unnecessary set state transitions: **57552**
22. (MDS) - Use regular dispatch for processing beacons: **57469**
23. (MDS) - Dump formatter when errors occur: **57673**
24. (MDS) - Set the proper extra bl for the create request: **57754**
25. (MDS) - Use init instead of swap to initialize: **57813**
26. (MDS) - Add debug message when conf changes are processed: **57882**
27. (MDS) - QuiesceDbRequest: Update the internal encoding of ops by making "exclude" and "cancel" share the same op code: **57912**
28. (MDS) - Properly initialize epoch for quiescedb: **57993**
29. quiesce-db: calculate_quiesce_map: Aggregate quiesce roots' TTL as "max": **57980**

Client

1. Preserve dentry trimming by running "set LIBMOUNT_FORCE_MOUNT2=always": **57170**
2. Clear resend_mds only after sending request: **57043**

Crimson

1. Make the refcount of split extents consistent with the original refcount: **56627**
2. crimson/osd/replicated_recovery_backend: prepare_pull use pg_info: **56611**
3. Let cloned OBC reference its head SSC: **56610**
4. crimson/osd/ops_executer: Fix snap overlap range error: **56606**
5. Improve osdmap handling, preventing MGR crashes: **56875**
6. Discard outdated recovery ops, improving OSD performance: **56848**
7. Make locking/promotion atomic if possible: **56844**
8. pg_recovery: Backoff if the recovery/backfill is deferred: **56806**
9. Implement basic reactor-utilization stats report to log: **56775**
10. Permit snap trimming only when PGs are clean: **56998**
11. Improve the dump_historic_slow_ops command: **56994**
12. Various fixes (recovery): **56916**
13. common/operation: Fix and move exit() after entering the next phase: **56912**
14. osd/osd_operations/client_request: Retrieve the correct version for objects to be recovered urgently: **56892**
15. osd/pg_recovery: Skip unfound objects when recovering the primary: **57147**
16. Seastore - transaction_manager: Remove incorrect assertions: **57135**
17. Seastore - transaction_manager: Fix write pipeline phase leak: **57129**

18. Seastore - Update onode sizes only when necessary: **57088**
 19. Detach blockers from blocking events when they are destroyed: **57069**
 20. ops_executor: Calculation of clone_overlap shouldn't consider snap contexts: **57313**
 21. crimson/osd/pg: Trigger wait_for_active_blocker on replica osds when the activate event is committed: **57279**
 22. crimson/osd/pg_backend: Do not modify OSDOp::indata when handling CEPH_OSD_OP_CHECKSUM: **57276**
 23. crimson/os/seastore/lba_manager: Don't increase intermediate mappings' refcount if LBAManager::clone_mapping() is called to remap mappings: **57262**
 24. Clamp reads to object size and bring full read trimming: **57204**
 25. os/seastore/record_scanner: Replace [=] capturing: **57434**
 26. os/seastore/object_data_handler: Clean up read(): **57432**
 27. Hold PGs' references to the last minute of snap trim events executions: **57416**
 28. osd/recovery_backend: Change recovery waiters' promises into optional ones: **57386**
 29. os/seastore: Add is_data_stable() to allow delta-overwrite on EXIST_CLEAN: **57368**
 30. osd/osd_operations/client_request: Check "can_serve_replica_reads" before getting obc: **57367**
 31. Allow basic cluster deployments: **57593**
 32. osd/ops_executor: LIST_SNAPS only on CEPH_SNAPDIR: **57561**
 33. os/seastore/transaction_manager: Drop unused code: **57476**
 34. os/seastore/transaction_manager: Correct the offset of the data copied from the original extents: **57474**
 35. os/seastore: Fix compilation error in release build: **57654**
 36. osd/osd_operations/client_request_common: PeeringState::needs_recovery() may fail if the object is under backfill: **57691**
 37. crimson/os/seastore/rbm/avlallocator: Return enough regions when request size is greater than max_alloc_size: **57694**
 38. os/seastore: Avoid getting wrong logical extents through "parent-invalid" lba mappings: **57709**
 39. os/seastore/async_cleaner: Fix incorrect get_num_rolls(): **57711**
 40. osd/object_context_loader: Fix obc cache existence usage: **57725**
 41. os/seastore: Implement disk and writer level stats reporting: **57788**
 42. os/seastore/lba_manager: Do batch mapping allocs when remapping multiple mappings: **57818**
 43. os/seastore/btree: Improve lba pointer related UT checks: **57828**
 44. osd/osd_operations: Correct connection pipelines for osd operations: **57908**
 45. Simplify obc loading by locking excl for load and demoting to needed lock: **57977**
 46. Simplify snaptrim operation pipeline usage: **57978**
 47. os/seastore: Add writer level stats to RBM: **58083**
 48. Fix ObjectContext::with_lock to only unlock if lock is taken: **58099**
 49. Revert "crimson/osd/osd_operation: fix 'dump_historic_slow_ops' command": **58223**
 50. os/seastore/onode: Add hobject_t into Onode: **58356**
- MGR**
1. (Dashboard) - "401 Unauthorized" message when anonymous access is disabled: **56939**
 2. (Dashboard) - Support Description and AccountId in rgw roles: **56919**
 3. (Dashboard) - Remove minutely from retention policy dropdown menu: **56907**
 4. (Dashboard) - Exclude cloned-deleted RBD snaps: **57151**
 5. (Dashboard) - Add RGW policy group management api: **57462**
 6. (Dashboard) - Fix host count per cluster and total hosts count on multi-cluster overview page: **57497**
 7. (cephadm) - Make the following NVMe-oF gateway fields configurable: "allowed_consecutive_spdk_ping_failures", "spdk_ping_interval_in_seconds", and "ping_spdk_under_lock": **56628**
 8. (cephadm) - Add more debug logging for autotuner: **56823**
 9. (cephadm) - Check if file exists when passing "--apply_spec": **56817**
 10. (cephadm) - Make enable_monitor_client configurable for NVMe-oF: **56791**
 11. (cephadm) - Have agent check for errors before json.loading mgr response: **56961**
 12. (cephadm) - Don't mark daemons created/removed in the last minute as stray: **56957**
 13. (cephadm) - Extend timeout from 60s to 300s for arm64 make check: **56942**
 14. (cephadm) - Clean up service size logic block (no functional change): **56933**
 15. (cephadm) - Update default NVMe-oF container image version: **57182**
 16. (cephadm) - Clean up iscsi and NVMe-oF keyrings upon daemon removal: **57181**
 17. (cephadm) - Update loki and promtail containers: **57164**
 18. (cephadm) - Set OSD cap for NVMe-oF daemon to "profile rbd" (otherwise snapshotting is impossible): **57143**
 19. (cephadm) - Change some omap_file_lock defaults: **57033**
 20. (cephadm) - Make SMB and NVMe-oF upgrade last in staggered upgrade: **57292**
 21. (cephadm) - Mark progress events as complete/fail only if they are initialized: **57259**
 22. (cephadm) - Make setting --cgroups=split configurable for adopted daemons: **57205**
 23. (cephadm) - Use pyyaml so that proper YAML format can be used to improve data processing: **57601**
 24. (cephadm) - Prioritize user-specified config during bootstrapping: **57829**
 25. (cephadm) - Configure security user in keepalived template: **57847**
 26. (cephadm) - Fix a keepalived config bug that caused buggy config to generate: **57848**
 27. (cephadm) - Make the host_facts class kernel_security method correctly read AppArmor profile names that have spaces in them: **57955**
 28. (cephadm) - Disable ms_bind_ipv4 when ms_bind_ipv6 is enabled: **57975**
 29. (cephadm) - Fix flake8 test failures: **58062**
 30. (cephadm) - Redeploy when some dependency daemon is added or removed: **58230**
 31. Use importlib.metadata for querying ceph_iscsi's version: **57685**
 32. (SMB) - Store potentially sensitive information in a resource separate from the less-sensitive general cluster config: **57180**
 33. (SMB) - Alters the output of the smb commands so that the JSON/YAML outputs make sense to human readers (rather than sorted by the key names): **57096**
 34. (SMB) - Allow devs/testers/experimenters to add custom smb config params to managed shares or clusters. USE AT YOUR OWN RISK: **57294**
 35. (SMB) - Make "create" commands create-only, not create-or-update: **57293**
 36. (prometheus) - s/pkg_resources.packaging/packaging: **57700**
 37. Add the "rgw mgr module" to ceph-mgr-modules-core in Debian: **57811**
 38. (MGR) - Fix error handling in "rgw zone create": **58142**
 39. (MGR) - Improve "ceph orch apply osd" error message: **58154**

40. (MGR) - Use un-deprecated APIs to initialize Python interpreter: **58199**

MON

1. Check all subnets (not just the first listed) when an OSD is attempting to join within a subnet address: **56640**
2. AuthMonitor: Provide command to rotate the key for a user credential: **58121**

OSD

1. Call stat/getattrs only once per object during deep-scrub: **56995**
2. ECTransaction: Remove incorrect asserts in generate_transactions: **56924**
3. CEPH_OSD_OP_FLAG_BYPASS_CLEAN_CACHE flag is passed from ECBackend (improve deep scrubbing): **57137**

RADOS

1. cls_fifo_legacy: 'oid' used after it was moved. Fixes <https://tracker.ceph.com/issues/66223>: **57702**

RBD

1. Make librbd::Image moveable: **56801**
2. Make group IDs and group snapshot IDs more random: **56987**
3. rbd-mirror - Improve stale pool-replayer cleanup and callout cleanup: **57082**
4. pybind/rbd: Expose CLONE_FORMAT and FLATTEN image options: **57212**
5. librbd: Don't crash on a zero-length read if the buffer is NULL: **57433**
6. rbd-wnbd: Wait for the disk cleanup to complete when stopping the service: **57697**
7. Add "rbd group info <group-spec>" command: **57759**
8. pybind/rbd: Parse access and modify timestamps in UTC: **57889**
9. librbd: Allow cloning from non-user snapshots: **57954**
10. librbd: Prevent diff-iterate from crashing on empty byte ranges: **57973**
11. librbd: Disallow group snap rollback if memberships don't match: **58074**
12. librbd: Make diff-iterate in fast-diff mode aware of encryption: **58201**

OSD

1. Reply with "pg_created" when a PG is peered and it is active+clean. This makes it possible for monitors to trim OSD maps as intended and fixes <https://github.com/ceph/ceph/pull/63912>: **55239**
2. Add a "clean primary" base state to the scrubber state machine. This state is entered after the peering is concluded and the PG is set to be Primary and is active+clean: **54996**

3. scrub - Remove "scrub_clear_state()", the functionality of which is now handled by the FSM: **55009**
4. scrub - Improve the scrub scheduler by removing the "penalty queue" from the scrubber and introducing the "not before" delay mechanism: **55107**
5. Distinguish between "osd_stat_report_max_epoch" and "pg_stat_report_max_seconds" and make "PeeringState::prepare_stats_for_publish" check for both. Fixes <https://tracker.ceph.com/issues/63520>: **54491**
6. scrub - A part of a reimplementation of scrub resource reservation requests that will no longer immediately grant or refuse scrub reservation requests but will instead queue them in an async reserver, similar to the way that backfill reservations are handled: **55131**
7. scrub - Compare a token (nonce) carried in the reservation reply with the remembered token of the reservation request. When they don't match, ignore them and log a stale reply: **55217**
8. scrub - Improve scheduling decision logs: **55453**
9. scrub - Use an AsyncReserver to handle scrub reservations on the replica side. The primary sends a reservation request with a 'queue this request' flag set. That request is queued at the scrub-reserver, and granted after the number of concurrent 'remote reservations' falls below the configured threshold: **55340**
10. Improve hobject_t::to_str() performance: **55583**
11. Directly display oldest_map and newest_map: **54913**

RGW

1. Improve handling of "WITH_RADOSGW_D4N=OFF": **56728**
2. Improve various "rgw_d4n" definitions: **56735**
3. Fix lifecycle crashes that occurred during bucket reloads: **56712**
4. Do not map normal HTTP errors to EIO. This avoids incorrectly marking endpoints as failed. Default to ERR_INTERNAL_ERROR (500 InternalError) instead: **56704**
5. Handle RGWRESTStreamS3PutObj initialization failures: **56657**
6. Prevent request payers from incurring charges for 403 requests (AWS): **56868**
7. Increase log level on abort_early to exclude client errors: **56866**
8. Eliminate SSL enforcement for SSE-S3 encryption: **56860**
9. Allow disabling mdsearch APIs: **56802**

10. multisite: Use dump_time_header() instead of dump_time(): **56765**
11. Improve the handling of bucket-instance metadata writes: **57004**
12. Add bucket_quota to RGWAccount-Info: **56986**
13. Start/stop endpoint managers in notification manager: **56979**
14. rgw_lua_utils: Free std::string: **56978**
15. Test for exact match if the test is not persistent: **56943**
16. boost/redis: Point to 1.85 tag: **56926**
17. Remove potential string overflow in POSIXDriver: **56906**
18. rgw/beast: Fix crash observed in SSL stream.async_shutdown(): **57155**
19. s3: Placement target is added in GetBucketLocation api response as header: **57152**
20. Do not log endpoint as it could contain broker user-id & password: **57098**
21. Improve bucket check efficiency during incomplete multipart uploads: **57083**
22. lifecycle-notification: Do not block lifecycle processing for notification errors: **57079**
23. Print no error when LDAP isn't configured (lack of LDAP config is the default, not an error): **57075**
24. Fix CompleteMultipart error handling regression: **57257**
25. RGW/notify: Allow persistent topics to be added/removed even if no rgw::notify::Manager is running. Hide the rados dependency behind new (hopefully temporary) sal::Driver interfaces: **57249**
26. Add new sync-policy related params to boto3 extension: **57391**
27. Remove RGWObjState and get_obj_state()/put_obj_state() from the SAL API: **57377**
28. Fetch obj_state after cloud transition to keep object head/attrs current: **57356**
29. Add shard reduction ability to dynamic resharding: **57538**
30. notification: Fix the caching issues of notification brokers, where the cache was not invalidated if topic attributes were changed: **57537**
31. notification: Store the value of persistent_queue for existing topics and continue committing events for all topics subscribed to given bucket: **57536**
32. Repair memory allocation issue: **57612**
33. Code de-duplication for administrative commands: **57626**
34. Fix a potential date race with finish() by checking the 'done' flag under the mutex() in wait(): **57632**

35. Add concurrency to RGWDeleteMulti-Obj in the null_yield case: **57641**
36. Bucket notification: Reload realm correctly. Fixes <https://tracker.ceph.com/issues/66206>: **57655**
37. Print new filters and newer-noncurrent elements in "radosgw-admin lc get": **57774**
38. multisite: Fix use-after-move in retry logic in logbacking: **57858**
39. Make HTTP tests run on a separate task: **57894**
40. Add "radosgw-admin topic dump --topic" command to dump notifications: **57898**
41. Do not assert on thread name-setting failures: **57969**
42. adminops: Add option to provide storageclass adminops user apis, so that the default storage class can be defined by the user: **57985**
43. notifications: Report an error when persistent queue deletion fails: **58081**
44. amqp: Fix valgrind errors on uninitialized memory: **58102**
45. rgw_user: Ensure that 'keys' is not in an undefined state after being moved: **58119**
46. Fix deleting an object with null version: **58133**
47. Test that no asio threads are blocked on "null_yield": **58179**
48. Fix multipart get part when count==1: **58288**
49. Use s3website REST when the s3website API is enabled and has a higher priority and no host header: **58339**
8. node-proxy: Make the daemon discover endpoints, allowing node-proxy to be compatible with more hardware: **57138**
9. Use free() to free the memory chunk allocated by reed_sol_vandermonde_coding_matrix(): **57112**
10. common: Mark assert-only variables as unused to remove complier warnings: **57226**
11. common: Add output file switch for JSON dumps: **57215**
12. common: Formatter: Use Cached-StackStringStream for efficiency: **57392**
13. common: Formatter: Trivial cons/des should be default: **57374**
14. common: options: Link to mon_osd_blocklist_default_expire from RBD: **57498**
15. common: TrackedOp: Do not count the ops marked as nowarn: **58125**
16. neorados: Remove unused symlink to completion.h: **57360**
17. vstart.sh: Add options to set number of alien threads, and number of CPU cores for alien threads: **57359**
18. (cmake) - Improve handling when ENABLE_COVERAGE is set to "ON": **57594**
19. (cmake) - Link rados_snap_set_diff_obj and krbd against legacy-option-headers: **57583**
20. (cmake) - Disable WITH_QATLIB/ZIP on non-x86 (allows building on s390x and ppc64le): **57479**
21. (ceph-volume) - Use importlib from stdlib on Python 3.8 and up: **57650**
22. (tools/first-damage) - Don't skip stray directory object: **57688**
23. (tools/first-damage) - Make CEPH_NOSNAP int64: **57696**
24. (ceph-volume) - Ensure that "lvm migrate" zaps source WAL/DB devices when they are removed or replaced: **57807**
25. (ceph-volume) - Fix a regression involving a call to entry_points(group=group): **57830**
26. (ceph-volume) - Fix set_dmccrypt_no_workqueue to allow flexible version handling: **57925**
27. (ceph-volume) - Fix regex usage in "set_dmccrypt_no_workqueue": **58138**
28. osd/scrub: Move more of the scrub initiation login into the scrubber: **58003**
29. osd: Suppress two false clang-tidy detections: **58036**
30. src/cephadm/box: Remove unused imports: **58053**
31. (nasm-wrapper) - Improve handling of "--coverage" flag: **58209**
32. dout: Add macro for libfmt-style logging: **58229**

NEWS

Neha and Dan Give Magazine Interview - Neha Ojha and Dan van der Ster were interviewed in FedTech magazine: <https://fedtechmagazine.com/article/2024/07/what-cluster-computing-and-how-can-it-help-federal-it-perfcon>

Reef Backport Misnumbering Alert - There is news about the fourth backport release in the Reef series. An early build of this release was accidentally exposed and packaged as version 18.2.3 by the Debian project in April. The version 18.2.3 release should not be used. The official release was re-tagged as version 18.2.4 to avoid further confusion.

Uncategorized

1. Use "tell pg deep-scrub pgid" instead of "tell pgid deep-scrub": **56745**
2. osd: Fix for segmentation fault on OSD fast shutdown: **56804**
3. osd/scrub: Allow new scrub sessions to be initiated by an OSD even while a PG is reserving scrub resources: **57865**
4. osd: ECBackend.cc: Fix double increment of "num_shards_repaired" stat: **58139**
5. python-common: Handle "anonymous_access: false" in to_json of Grafana spec: **56928**
6. (ceph-volume) Create LVs when using partitions: **56882**
7. Adds spawn_throttle class for bounded concurrency with stackful coroutines from boost::asio::spawn(). This relies on new support for per-op cancellation to guarantee that the lifetime of child coroutines won't exceed the lifetime of their throttle, making it safe for children to access memory from their parent's stack: **57188**

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