Ceph Quarterly

Issue # 6 *An overview of the past three months of Ceph upstream development.* Oct. 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. Remove unused "zone_adjustments": 58588
2. Remove unused functions: 58993
3. Fixes for multiple bdev label: 59762
4. Move reservation of bdev label to proper place: 59850

cephadm

1. bootstrap: Verify orch module is running before setting cephadm backend: 59681
2. Turn off "cgroups_split" setting when bootstrapping with "--no-cgroups-split": 58379
3. smb: Determine Samba version within container: 58585
4. Do not hardcode Samba debuglevel 6: 58653
5. Implement metrics-exporter from smb service using side-car container: 58815
6. Require "group" parameter in NVMe-oF specs: 58860
7. Add SPDK log level to NVMe-oF configuration file: 58969
8. Emit warning if daemon's image is not to be used: 59485
9. Support Docker Live Restore: 59730
10. mgr/cephadm/services/nvmeof.py: Allow setting "0.0.0.0" as address in the spec file: 59755
11. Rename "whitelist_domains" field to "allowlist_domains": 59766
12. Allow services to be conditionally ranked: 58517
13. Update shared folder python packages directory: 58533
14. Update default container image for NVMe-oF gateway to latest (1.2.17) from older version (1.2.5): 59085
15. Add ability for cephadm to generate frontend cert for RGW: 59174
16. Add "original_weight" parameter to OSD class: 59318
17. Update the SPDK RPC socket fields in cephadm so that they are congruent with their counterparts in cephadm: 59362
18. Mount NVMe-oF certs into container: 59363
19. Open ceph-exporter when firewalld is enabled: 59694
20. Use host address while updating RGW zone endpoints: 59728
21. Make the discovery and gateway IPs configurable in NVMe-oF configuration: 59738
22. Update oauth2 proxy image variable name: 59843
23. Add "--no-exception-when-missing" flag to the "cert-store cert/key get" command: 59860
24. Add health check endpoint for mgmt-gateway: 59927
25. Bump Grafana to 10.4.8: 59591
26. Make "ssh keepalive" settings configurable: 59610
27. Add spec fields for "oauth2-proxy whitelist_domains". This field is required to configure the domains that are allowed to redirect during login and logout: 59634
28. Change the nginx upstream image used by mgmt-gateway: 59643

CephFS

1. mds: authpin on wrlock only when not a locallock: 58861
2. mds: Add more debug logs and log events: 58950
3. mds: Require config lists to be alphabetically sorted: 59088
4. mds: Encode quiesce payload on demand: 59176
5. mds: Move "fscrypt inode_t" metadata to "mds_co mempool": 59414
6. mds: Dump "next_snap" when checking dentry corruption: 59504
7. mds: Invalid id for client eviction is to be treated as success: 59874
8. cephfs-top: Fix exceptions on small/large sized windows (improve visual presentation when window is large): 59566
9. cephfs_mirror: Revert "sync_duration" to seconds: 59875

ceph-volume

1. Switch to new disk sorting behavior: 59170
2. Add "packaging" to "install_requires": 59204
3. Fix generic activation with raw OSDs: 59573
4. Pass "self.osd_id" to "create_id()" call: 59604
5. "ceph-volume" now calls "ceph-bluestore-tool zap-device" in addition to the existing calls when wiping a device: 59742
6. Fix OSD lvm/tpm2 activation: 59915
7. Add a new class, "UdevData", to represent udev data for a given device: 60006
8. ceph-volume: docstring and typing corrections: 60031

Client

1. Prevent race condition when printing inode in "ll_sync_inode": 59162
2. Calls to "ll_fh_exists()" should hold "client_lock": 59300
3. Use vectors for context lists to avoid an allocation for each context and to improve cache locality: 59171
4. Remove hypertable: 59810

Crimson

1. os/seastore/cached_extent: Add the "refresh" ability to LBA mappings: 58367
2. os/seastore: Track transactions/conflicts/outstanding periodically: 58467
3. os/seastore: Move OOL writes from collection lock to concurrent DeviceSubmission phase: 58913
4. os/seastore: Consider "segment_header_t::modify_time" as the segments' "modify_time" for no-tail OOL segments: 58787
5. os/seastore/lba_manager: Correct the range end of lba mappings: 58869
6. os/seastore: This PR converts "laddr_t" to a struct and makes it block-aligned, as the basic part of static 128-bits laddr design: 59182
7. os/seastore: Refine documents related to inplace rewrite: 59392
8. os/seastore/btree: Fix minor corner case issue: 59205
9. os/seastore/cache: Report LRU (Last Recently Used) usage/in/out with trans and extent type: 59212
10. os/seastore/cache: Report dirty cache usage and rewrite stats: 59391
11. os/seastore: Clean up LBA/backref node capacity: 59570
12. os/seastore: Remove unnecessary memory copy during OOL write: 59723
13. os/seastore/cache: Report cache access stats: 59553
14. Revert "crimson/os/seastore: wait ool writes in DeviceSubmission phase" (prevent data corruption in OOL write cases): 59292
15. osd/recovery_backend: Restart object pulls that are blocked by down OSDs: 59189
16. osd: Clear ondisk temp objects on startup: 58693
17. osd: Send empty transactions to backfill targets that haven't backfilled the objects yet: 59011
18. osd: Cancel ongoing pglog-based recoveries when recovery defers: 59066
19. osd/backfill_state: Support backfill cancellation: 59118
20. osd: Support backfill cancellation - part two: 59368
21. osd/pg: Implement "PG::PGLogEntryHandler::remove()" (add pglog rollback capability): 59227
22. osd/pg: Mutate only OBC for user-triggered transactions: 59407
23. osd/pg: Clear backfill_state when the PG goes clean: 59456
24. osd/pg: Fix missing priority argument: 59495
25. osd/pg: "do_osd_ops_execute" fixes: 59543
26. osd/pg: Fix wrong lambda capture of transactions in "PG::submit_error_log()": 59102
27. osd/pg: Properly propagate snap mapper updates and do clean-region-based clone objects recovery: 58868
28. osd/pg_recovery: Push the iteration forward after finding unfound objects when starting primary recoveries: 58504
29. osd: PG stats are not synced between osds after object update: 58510
30. osd: Write "require_osd_release" only when needed: 59564
31. osd/osd_operations/snaptrim_event: Increase "osd_osd_param_t::at_version" only after clone adjustments: 59652
32. osd/scrub: Fix a null pointer error: 58471
33. Audit and correct epoch captured by IOInterruptCondition: 58463
34. recovery_backend: Clean up "PGBackend::temp_contents" when PG interval changes: 58694
35. Re-queue client requests from a temporary queue other than "ClientRequest::Orderer::list": 58708
36. Access "coll_map" under alien tp with a lock: 58766
37. common/tri_mutex: avoid "hobject_t" formatting: 58897
38. common/tri_mutex: also wake up waiters when demoting: 59301
39. Preparatory work for improving cache metrics and fine-grained cache: 58983
40. Modify AlienStore and OSD to use a gate per Seastar shard instead of a shared gate, ensuring proper handling across gates: 58986
41. Clear "AlienStore::coll_map" in "umount" rather than in "stop": 59098
42. os/bluestore: Improve comments in "hybrid_allocator": 59696
43. Don't retain "InternalClientRequest" on interval change even if primary does not change: 59815

MGR

1. mgr/rgwam: Use realm/zonegroup/zone method arguments for period update: 59716
2. mgr/smb: Fix "ceph smb show" when a cluster has not associated shares: 58514
3. mgr/smb: Improvements to smb mgr module and docs: 58518
4. mgr/smb: Earmark resolver for subvolume: 59387
5. mgr/nfs: Earmark resolver for subvolume: 59726
6. mgr/smb: Stop trying to clean external store during cluster sync: 59658
7. mgr/smb: Fix condition for smb earmark when "cluster_id" doesn't match: 60044
8. mgr/status: Fix "fs status" json output: 59699
9. mgr/nfs: Do not ignore clusters from "rados pool conf" objects: 58535
10. orchestrator: Fix encrypted flag handling in "orch daemon add osd": 59175
11. pybind/mgr: Attempt to fix mypy importing from python-common: 59845
12. pybind/mgr: Drop py37 test from tox.ini: 59958
13. Verify the connection to the client before sending data: 58592
14. Add routed tabs for rgw multisite and openend modal inside router-outlet: 58656
15. Fix invalid escape sequence: 58767
16. Add support for the new Ceph VFS module in Samba to the manager module: 58994
17. Restrict subvolumes to the scope of a single protocol (either NFS-ganesha or SMB): 59111
18. Fix "daemon add osd" boolean parameter handling: 59831
19. Make "service_id" better aligned with "default/empty" group; fix "service_id" in "nvmeof daemon add": 59925

MON

1. mon: Fix "fs set down" to adjust "max_mds" only when cluster is not down: 58582
2. mon/osdmonitor: Clean up the code for "preprocess_mark_me_dead": 58672
3. Paxos service cosmetic fixes: 58976
4. Set stopping to "bool" instead of to "set()": 59166
5. Upgrade rules for NVMeoF GW monitors and gateways: 59240
6. Handle "gw fast-reboot", implement proper handling of "gw delete" scenarios: 59385
7. GW deleting state: 59579
8. Enable NVMeoF GW monitor: 59588
9. leonidc0409 blocklist fix: 59592
10. Remove duplicated "NVMeofGwMap.h": 59840
11. mon/nvmeofgw*: Fix tracking gateways in DELETING state: 59999

OSD

1. Improve scrub scheduling: 58858
2. Fix "require_min_compat_client" handling for MSR rules: 59474
3. scrub: Add configuration parameters to control delay duration: 59590
4. scrub: Decrease default deep scrub chunk size. The previous default of 25

objects per chunk proved to take too long (many hundreds of milliseconds) on a busy cluster. As the scrubber locks all objects in the chunk for the duration, a large chunk size can cause a significant impact on the client ops' latencies: 59636

5. scrub: Disable high work-queue priority for high-priority scrubs: 59641
6. osd/scrub: Remove unnecessary requested_scrub flags: 59709
7. osd/scrub: reduce "osd_requested_scrub_priority" default value to ensure that no scrub messages have a higher priority than client op messages: 59793
8. osdc: fix mutex assert for !debug builds: 60026

RBD

1. librbd: fix inconsistency between AioCompletion "is_complete()" and "wait_for_complete()": 58591
2. librbd/migration: don't instantiate NativeFormat, handle it via dispatch: 58882
3. librbd: Reduce use of atomics in librbd throttling: 58907
4. librbd: make "group snap list" async and optionally sorted by snap creation time: 59107
5. librbd/migration: prune snapshot extents in "RawFormat::list_snaps()": 59551
6. librbd: Introduce "rbd_group_snap_namespace_type_t" Enum: 59883
7. librbd: Add "LIBRBD_SUPPORTS_GROUP_SNAP_GET_INFO" definition: 59959
8. Make RBD bench write out a string of random lowercase characters instead of a single character for the size of the IO buffer, to improve test results when performing EC tests: 59000
9. Fix CLI output of "rbd group snap info" command when a group snapshot with no member images: 59153
10. rbd-mirror: Use correct ioctx for namespace: 59401
11. Allow different namespaces to be mirrored between pools: 59417
12. Make "rbd bench" write a different byte from one run to the next: 59423
13. Amend "rbd {group,} rename" and "rbd mirror pool" command descriptions: 59478
14. Make ERRNO positive: 59657
15. Set journaling feature when "—mirror-image-mode" is "journal": 59842

RGW

1. Add missing content-type for "RGWGetBucketLocation": 58442
2. Add reshard status to bucket stats; minor miscellaneous improvements: 58567
3. cmake: Work around xxhash "inlining failed" errors in debug builds: 58429
4. Clean up the only gc function that was hidden with "CLS_CLIENT_HIDE_IOCTX". This allows rgw to use it asynchronously with "rgw_rados_operate()" and "optional_yield", and warn about blocking calls that should be asynchronous: 58449
5. cls: Bump "cls_rgw_reshard_entry" decode version to match encode: 58399
6. cls/rgw: Duplicate reshard checks in all "cls_rgw" write operations: 59611
7. Correctly read a signed value as a signed value, reducing bucket sync wait times: 58403
8. datalog: LazyFIFO race fix: 58491
9. Do not allow NotPrincipal with Allow Effect: 58686
10. filestore: Remove "WITH_ZFS" option (FileStore cleanup): 58416
11. Fix the handling of HEAD requests that do not comply with RFC standards: 58572
12. Ignore EEXIST errors from local role creation if forwarded request succeeds: 58665
13. Increase quota size to avoid DoS when QuotaExceeded: 58378
14. Improve object library dependencies: 58414
15. multisite: don't retain "RGW_ATTR_OBJ_REPLICATION_TRACE" attr on "copy_object": 58519
16. msg: Insert "PriorityDispatchers" in sorted position: 58631
17. rgw/async/notifications: Use common async waiter in pubsub push: 58765
18. rgw/beast: Improve socket handling: 59014
19. rgw/kafka: Refactor topic creation to avoid "rd_kafka_topic_name()": 59741
20. rgw/lifecycle: Fix a bug in "LCOpAction_Transition::check()": 60020
21. rgw/posix: Name the "lock_guard" in "BucketCacheEntry::reclaim()": 58862
22. rgw/rados: zero-init "shard_count" in "RGWBucket::check_index_unlinked()": 59117
23. rgw/rados: "set_attrs()" falls back to existing attrs for index update: 58649
24. rgw/rados: Use "rgw_rados_operate()" instead of "operate()" to suspend the coroutine when "optional_yield" is not empty. Make a similar change for "RGWRados::Object::Read::read()": 59001
25. rgw/rados: Guard against "dir suggest" during reshard (no changes to the bucket index should be allowed while resharding): 59609
26. rgw/rados: Don't rely on "IoCtx::get_last_version()" for async ops: 59998
27. rgw/sal: "LCHead" and "LCEntry" don't need abstraction: 58603
28. rgw/auth: RemoteApplier respects implicit tenants: 58606
29. Squid Compressor: Add data format (QZ_DEFLATE_GZIP_EXT) for QAT Zlib: 58643
30. Allow restricting requests regarding SSE-C encryption with bucket policy: 58689
31. Pull in upstream fix for "zpp_bits": 59051
32. Revert account-related changes to "get_iam_policy_from_attr()": 59169
33. Increase log level for enoent caused by clients: 59172
34. Add missing closing brackets for "S3-Key notification filter" in JSON: 59218
35. rgw/http: "finish_request()" after logging errors: 59243
36. Construct tokens outside of spawn and capture them in the lambda to extend their lifetime until all spawned coroutine functions are complete: 59244
37. sts: Correct an error message to indicate that a key must be alphanumeric: 59249
38. Make "PutObj" load the source bucket's attrs, to make sure that the associated policies are loaded: 59253
39. multisite: Initialize "sync_status" in "RGWDataFullSyncSingleEntryCR": 59329
40. Decrement "qlen/qactive" perf counters on error: 59386
41. When no "rgw_realm" or "rgw_zone" options are specifically requested and we fail to load the default zone of the default realm, try to fall back to the global "default" zone/zonegroup: 59422
42. cls: Stop an integer from being interpreted as a character code when it is assigned to a string: 59489
43. Use "std::forward" with "lvalues" when dealing with forwarding references (universal references) in templates to preserve the "lvalue" or "rvalue" of the argument: 59490
44. Raise the default to "rgw_max_listing_results=5000" (allow clients to request more than the default 1000 keys per request): 59534
45. multisite: Handle errors properly in "RGWDataFullSyncSingleEntryCR()": 59536

46. Remove "rgw_data_log_obj_prefix" (ensure that "data log" is returned when the prefix is empty, and ensure that the actual prefix is returned otherwise): 59567
47. notifications: Free completion pointer using "unique_ptr"; fix access to dangling "dpp" pointer: 59607
48. rgw/rgw_iam_policy: Check for dereference of a null pointer (loaded from variable 't'): 59731
49. rgw_log_backing: Error code not returned: 59737
50. Use concurrent deletes to speed up inline garbage collection: 59864
51. Handle http options CORS with v2 auth: 59977
52. Enable RGW to decrypt the part for a "get object" request with "partNumber": 60019

Uncategorized

1. arch/s390x: Redistribute under apache2 license: 59809
2. blk/aio: Fix compile issue when "HAVE_LIBURING" isn't defined: 58745
3. cls/user: Reset stats only returns marker when truncated: 59884
4. common/async: Fix duplicate definition errors from SharedMutexImpl: 58722
5. common/async: spawn_throttle wraps call to "asio::spawn()"; prohibit child coroutines from being spawned on a different executor by wrapping the call to "asio::spawn()" in a new "spawn_throttle::spawn()" interface: 58798
6. common/gated: Enable "ceph_assert" on shard id: 59832
7. erasure-code/isa: Use ISA/RAIS's "xor_gen()" instead of the "region_xor()" optimization: 58594
8. erasure-code/clay: Clean up unused but set variable: 58703
9. kv/rocksdb: Return error for "dump_objectstore_kv_stats" asok command: 58728
10. node-proxy: Fix a regression when processing the RedFish API: 59981
11. submodule: Remove the "boost_redis" submodule again (again): 58971
12. Ceph exporter can run now as HTTP or HTTPS server (added TLSv13 support): 58440
13. Add "kv_stats" function to print all CF (column family) info: 58718
14. Fix uninitialized "discard_stop": 59026
15. Add "beacon_lock" to mitigate potential beacon delays caused by slow message handling, particularly in "handle_nvmeof_gw_map": 59053
16. Add support for enabling the smb cluster resource to create clusters of smb services instead of just multiple smb servers operating independently: 59086
17. Warn when QAT switches to software [de]compression: 59335
18. Add support for IPs managed by Samba's CTDB clustering system: 59419
19. Change code from non-PIC to PIC for ppc64 so that Ceph will build on IBM Power: 59558
20. Add "Containerfile" and "build.sh" to build it: 59868
21. README: Add OpenSSF Best Practices Badge: 60066

NEWS

Squid, the nineteenth major release of Ceph, was released on 26 September 2024. For details on this release, see the "Releases (Index)" page at docs.ceph.com.

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