Ceph Quarterly

Issue # 4 *An overview of the past three months of Ceph upstream development.* Apr. 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` .

cephadm

1. Allow OSDs to be created with custom CRUSH device classes and accept all custom names: **55534**

CephFS

1. MDS: The "dump dir" command now indicates that a directory is not cached. Fixes **https://tracker.ceph.com/issues/63093**: **54202**

ceph-volume

1. Avoid workqueue when using encryption on flash devices: **54423**

MGR

1. (***cephadm***) Allow idmap overrides in nfs-ganesha configuration: **54383**
2. (***dashboard***) Update babel-traverse from 7.23.0 to 7.23.2 (security fix): **54103**
3. node-proxy: Handle "None" statuses returned by RedFish: **55955**

Orchestrator

1. Introduce a new agent that can inventory a machine's hardware and provide statuses to the Dashboard and raise alerts in the Dashboard. Uses RedFish API: **54742**

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://tracker.ceph.com/issues/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**

RBD

1. When diffing against the beginning of time in fast-diff mode, 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: **55127**
2. Spawn one process per host instead of one per rbd-wnbd image mapping. This improves performance and scalability by avoiding creating excessive threads and TCP sessions: **52540**
3. "rbd children" command now accepts "--image-id" option. This allows listing children (clones) of an image that was previously moved to the RBD trash without temporarily restoring the image: (tracker number) **64376**
4. Use netlink interface by default for setting up rbd-nbd image mappings. This eliminates some post-mapping race conditions that are inherent to the legacy ioctl interface: **55234**
5. Fix resize detection for rbd-nbd image mappings set up using netlink interface: **55287**
6. Make diff-iterate account for discards that truncate the backing data object. Among other things, this makes "rbd diff" command output more precise: **56064**
7. Expand rbd-mirror test coverage: **54802**, **55797**

RGW

1. swift - Update the swift tempurl and formpost logic to support sha256 and sha512 signatures: **47723**

NEWS

Cephalocon - Cephalocon will be held on December 4th and 5th in Geneva, Switzerland - **https://events.linuxfoundation.org/cephalocon/** #cephalocon

"Auto-tiering Ceph Object Storage, Part 1" - Steven Umbehocker explains how to create multiple storage classes, which makes it possible to sort large objects from small objects, and makes it possible to boost your cluster's usable capacity and improve the speed at which the cluster retrieves objects - **https://ceph.io/en/news/blog/2024/auto-tiering-ceph-object-storage-part-1/**

Ceph reaches 1 TiB/s - Mark Nelson of Clyso explains how he drove a Ceph cluster to 1 TiB/s - **https://youtu.be/pGwwIaCXfzo**

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