Ceph Quarterly

Issue # 1 *An overview of the past three months of Ceph upstream development.* July 2023

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 `48720` to the string `https://github.com/ceph/ceph/pull/` to make this string: `https://github.com/ceph/ceph/pull/48720`.

CephFS

1. New option added—clients can be blocked from connecting to the MDS in order to help in recovery scenarios where you want the MDS running without any client workload: **48720**
2. Clients are prevented from exceeding the xattrs key-value limits: **46357**
3. snapdiff for cephfs—a building block for efficient disaster recovery: **43546**

Cephadm

1. Support for NFS backed by virtual IP address: **47199**
2. Ingress service for nfs (haproxy/keepalived) is now redeployed when its host(s) go offline: **51120**
3. Switch from scp to sftp for better security when transferring files: **50846**
4. Ability added to preserve VG/LV for DB devices when replacing data devices: **50838**

Crimson

1. Much bug squashing and test suite expansion, successful rbd and rgw tests.
2. Groundwork for SMR HDD support for SeaStore: **48717**
3. Full snapshot support (snapshot trimming, the last outstanding piece of functionality needed to provide full snapshot support, has been merged.)
4. Simple multi-core design for SeaStore: **48717**
5. Memory usage improved by linking BlueStore with tcmalloc: **46062**

Dashboard

cephx user management

1. import/export users: **50927**
2. delete users: **50918**
3. edit users: **50183**

RGW multisite setup/config

1. multisite config creation: **49953**
2. editing zonegroups: **50557**
3. editing zones: **50643**
4. editing realms: **50529**
5. migrate to multisite: **50806**
6. delete multisite: **50600**
7. RGW role creation: **50426**

RADOS

1. mClock improvements—a number of settings have been adjusted to fix cases in which backfill or recovery ran slow. This means faster recovery for restoring redundancy than with wpq while preserving the same client performance. PRs are summarized in the reef backport: **51263**
2. Scrub costs have been adjusted similarly for mclock: **51656**
3. rocksdb updated to 7.9.2: **51737**
4. rocksdb range deletes have been optimized and made tunable online: **49748** and **49870**. These improve the performance of large omap deletes, e.g. backfilling for RGW bucket indexes, or bucket deletion.
5. osd_op_thread_timeout and suicide_timeout can now be adjusted on the fly: **49628**

RBD

1. Alternate userspace block device (out-of-tree kernel module) support added—like a better-performing rbd-nbd: **50341**
2. A number of rbd-mirror–related bug fixes. Better handling of the rbd_support mgr module being blocklisted. Failed connections to remote clusters are now handled more gracefully. Specifically:
   a. blocklist handling in the rbd_support mgr module: **49742** and **51454**
   b. Perform mirror snap removal from the local, not remote cluster: **51166**
   c. Remove previous incomplete primary snapshot after successfully creating a new one: **50324**
3. Switch to labeled perf counters (per-image stats to monitor) for rbd-mirror: **50302**
4. rbd-wnbd: optionally handle wnbd adapter restart events: **49302**
5. RADOS object map corruption in snapshots taken under I/O: **52109**. See the `e4b1e0466354942c935e9eca2ab2858e75049415` commit message for a summary. Note that this affects snap-diff operations, which means that incremental backups and snapshot-based mirroring are affected.

RGW

1. When RGW creates a new data pool, it now sets the bulk flag so that the autoscaler sets an appropriate number of PGs for good parallelism: **51497** (see also: **https://docs.ceph.com/en/latest/rados/operations/placement-groups/#automated-scaling**)
2. Bucket policies can now allow access to notifications for users who do not own a bucket: **50684**
3. Improved Trino interoperability with S3 Select and RGW: **50471**
4. An optional/zero API for benchmarking RGW was added: **50507**
5. Experimental radosgw-admin command for recreating a lost bucket index: **50348**
6. Cache for improving bucket notification CPU efficiency: **49807**
7. D4N shared cache via Redis—initial version merged: **48879**
8. Hardware cryptography acceleration improved with a batch mode: **47040**
9. Read-only role for OpenStack Keystone integration added: **45469**
10. Multisite sync speed increased by involving and coordinating work among multiple RGW daemons: **45958**

Akamai has joined the Ceph Foundation, taking over Linode's membership because Akamai acquired Linode.

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