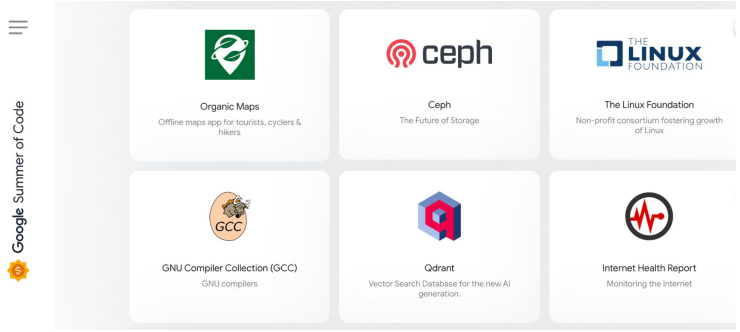
Diving Deep with Squid

2024.04.26

PROJECT MILESTONES

- Cephalocon after 3 years
- Celebrating one exabyte of Ceph storage in Telemetry!
- Reef widely used
  - ~1/3 of clusters reporting via telemetry are running Reef
  - ~1/3 Quincy
  - Rest split between Octopus, Pacific, and Nautilus
- Squid RC coming out soon

- Ceph upstream community engagement in Slack
  - #ceph-at-scale new channel for community engagement
- Ceph Infrastructure Upstream meetings
- Community outreach continues
  - Google Summer of Code
  - Grace Hopper Open Source Day



COMMUNITY UPDATES

- New Ceph Foundation Structure
  - new Diamond Members (Bloomberg, IBM, 45Drives) and revamped membership tiers
- Community Management Updates
  - Thanks to Mike Perez for his work in this role over the past several years
  - Foundation funding services from the Linux Foundation to fulfill some of these responsibilities
    - Noah Lehman - Marketing
    - Casey Cain - Program Management
  - Gaurav Sitlani organizing Ceph Ambassadors (event organization around the world)
- Events Updates
  - Cephalocon 2024!
    - December 4 - 5, 2024 in CERN, Geneva, Switzerland
  - More Ceph Days
    - Ceph Day India, May 31, 2024, Bengaluru, India
    - Ceph Day Asia, Sep 3 - 4, 2024, Suwon, South Korea (collocated with OpenInfra Summit Asia)

What’s coming in Squid

RADOS - SQUID

- The Elastic Shared Blob optimization in BlueStore
  - lower latency and CPU usage with snapshot intensive workloads
- Support for more flexible EC configurations
  - new CRUSH rule: MSR (Multi-Step Retry)
- Bluestore now defaults to using RocksDB LZ4 compression for increased average performance and reduced “fast device” space usage.
- Scrub scheduling and usability improvements
- OpTracker for ceph-mgr
  - it will be far easier to debug mgr module issues
- MGR Online Read Balancer
  - the read balancer can now be enabled automatically via the balancer module
- Client vs client QoS foundation - ongoing

CRIMSON - SQUID

- Testing/Stability
  - Expand crimson-rados suite (EC, scrub, OSD thrashing)
  - Add seastore tests to crimson-rados
- OSD Features:
  - Scrub
  - Erasure Coding - ongoing
- Seastore
  - Multicore scaling optimizations
  - Clone
  - Partial extent caching
- Performance - measurements and optimizations
  - Multicore scaling optimizations (including messenger)

TELEMETRY - SQUID

- New pool flags collection
  - Is it a Crimson cluster?
- Further analysis of metrics
- New Storage Capacity Utilization Percentiles by Class panel

DASHBOARD - SQUID

- Ceph FS Management
  - Subvolume/Subvolume group management
  - Snapshots and clone management
  - Access management
- RGW Advanced Workflows (user roles/policies, bucket policies...)
- RGW Multi-site Workflow
- Continuous UI/UX Improvements
- More changes and improvements to Landing page

CEPHADM - SQUID

- Samba based SMB support
  - Work includes support for “sidecar” containers
  - Orchestration deployment work completed
  - SMB mgr module also in progress
    - To allow making clusters, exports, etc. similar to NFS mgr module
    - Will possibly see this in a squid point release
- YAML, Jinja2 templating library support in cephadm binary
  - Previous work allows us to add a few pure python deps to the binary
- Regex based host patterns for placements
- NVMEoF deployment
  - Still a lot of churn here, but general deployment work completed
- Support for CA signed SSH keys
- Now possible to have cephadm managed hosts where daemon deployment is blocked but client keyrings can still be distributed
- Can now zap OSD devices as part of host drain procedure

ROOK - SQUID

- Object stores
  - DNS virtual-style hosting for buckets
  - Allow creating multiple object stores in shared pools
    - Isolation of each store is via RADOS namespaces
- Additions to the Rook kubectl plugin for troubleshooting and managing Ceph
- Azure Key Vault KMS support for encrypting OSDs

RBD - SQUID

- diff-iterate performance optimizations
  - Needed for QEMU block-status hook
  - Synthetic test based on QEMU blockdev-backup job showed 100x speedup!
  - Also made a number of correctness fixes in the process: report HOLE only where DATA was reported previously, account for discards that truncate, etc
- NVMe-oF target gateway maturity
  - Multipathing and gateway groups (ANA states, active/passive HA in progress)
  - Add (more) unit tests, set up integration tests
  - Monitoring and observability capabilities
  - Scale testing

RBD - SQUID

- Snapshot-based mirroring hardening and enhancements (ongoing)
  - Remaining sync correctness edge cases: full-object discards, copyup in clone images, some issues with object maps
  - Consistency group support (operate on “rbd group” groups of images as a whole, including failover/failback)
- rbd-wnbd (Windows) enhancements
  - A single daemon process per host (vs a process per mapped disk!)
  - Allow importing from/exporting to block devices (e.g. \\.\PhysicalDrive1)

RGW - SQUID

- User Accounts and AWS Identity and Access Management (IAM) API support
  - For self-service management of users, keys, groups, roles, and associated IAM policy
- Bucket Notification and Topic enhancements
  - Added permission model with support for IAM resource policy
  - Multisite replication for associated metadata
- AWS Signature v4 support for streaming payloads with trailing headers
- Performance counters can now track per-bucket and per-user metrics

CEPHFS - SQUID

- Improved efficiency of subtree state journaling to improve performance and scalability in the presence of hundreds or more subtrees.
- New “quiesce” API for backup systems to take multiple snapshots fully consistent with each other
- Kernel FSCrypt support has arrived
- NFS-Ganesha memory improvements: can use a single CephFS client across all shares (instead of one per share) *TBM
- mgr/volumes plugin is now much more scalable
- Mounds of UX papercuts have been identified and fixed
- Significant system hardening and bug fixing as deployments ramp up