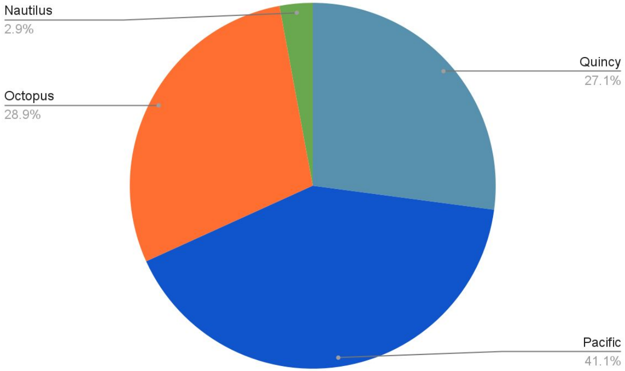
# State of the Cephalopod

2022.11.03

# PROJECT UPDATE

- New Ceph Governance Model
  - 3-member elected Executive Council
    - Project coordination, single contact point
    - Interface with the foundation
  - Driven by the Ceph Leadership Team - group effort, shared leadership, open meetings
- Recent Focus Areas
  - Release process
    - Publishing RC candidates
    - Multiple real-world upgrades before release
  - Performance and scalability hardening
    - Pawsey, other major scale tests with 1000s of OSDs
    - Logical large scale tests in teuthology

- Telemetry data:
  - More than 2K reporting clusters
  - > 800 PB total capacity, > 100K OSDs



- Public dashboards:
  - telemetry-public.ceph.com

# COMMUNITY

Healthy Ceph Foundation membership

- 10 Premier, 13 General, 11 Associate
- Upstream events, labs, docs, marketing
- Board meets monthly

User/Dev meeting to get more interaction and feedback, monthly virtual event

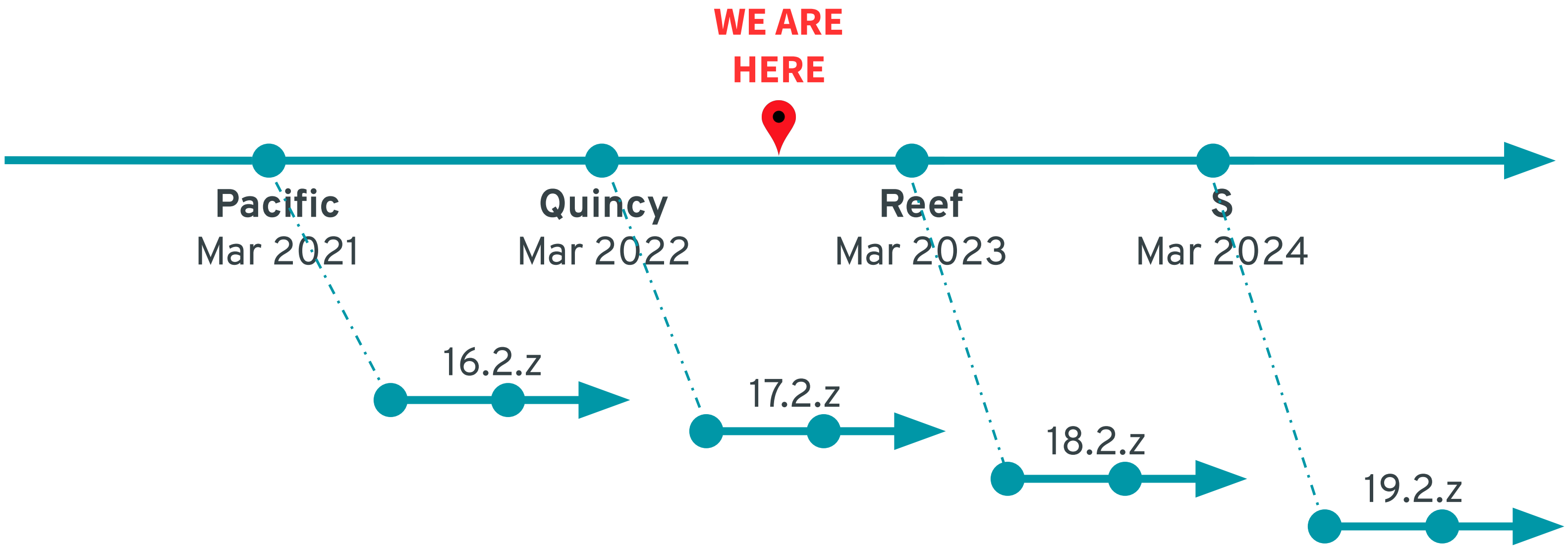
Continue to participate in GSoC, Outreachy

Grace Hopper Open Source Day

Regular tech talks

Early discussions on Cephalocon 2023

# RELEASE SCHEDULE



- Stable, named release every 12 months
- Backports for 2 releases
  - Pacific reaches EOL shortly after Reef is released
- Upgrade up to 2 releases at a time
  - Nautilus → Pacific, Pacific → Reef, Quincy → S

# Quincy and Beyond

# PERFORMANCE & SCALE

- Pawsey Supercomputing Centre
  - 4000 OSDs - real world setup
  - 64PB raw capacity
  - Several bottlenecks in cephadm/dashboard/mgr fixed
- Gibba - upstream sepia lab
  - 1000 OSDs - logical scale (limited per-OSD resources)
  - Verifying Quincy works at scale, e.g. upgrades and QoS
- Red Hat scale lab
  - 8000 OSDs - logical scale (limited per-OSD resources)
  - Exploring monitoring bottlenecks and solutions
- Upstream High Performance Cluster

# RADOS - QUINCY

- QoS in the OSD, by default
    - Different profiles to prioritize client I/O, recovery and other background tasks
    - Extended testing across different types of workloads and performance evaluation at scale
- BlueStore
    - Remove allocation metadata from rocksdb for significantly improved small write performance
    - BlueFS finer-grained locking
    - Cache age binning
- Msgr on-the-wire compression for osd-osd communication
- More concise and meaningful reporting of slow operations in the cluster log
- Health warnings for
    - Filestore OSDs, indicating deprecation in Reef
    - 'require-osd-release' flag does not match current release
- Improved opt-in flow in telemetry module

# RADOS - REEF

- client vs client QoS
  - Initial implementation in librados and testing
- Support high priority operations with mclock better, e.g. user-initiated object repair, force recovery
- BlueStore
  - custom WAL for RocksDB
  - New one-tracker-per-object mode is to replace shared blobs logic
  - 4K allocation unit for bluefs, expandable superblocks
- Balancer - workload (primary) balancer
  - Optimizes for balance of reads, in addition to writes
- Data availability score based on PG state
- PG log improvements to avoid and detect memory growth

# CRIMSON PROJECT

- High-performance rewrite of the OSD, currently supports RBD workloads on replicated pools with BlueStore
- New in Reef:
  - Initial multi-reactor support, S-release will add further improvements to the messenger
  - Usability improvements (set-allow-crimson, crimson pool type)
  - Initial snapshot support, snapshot trimming work ongoing
  - Improvements to SeaStore (next generation ObjectStore implementation) focused on efficiency and tiering
- Planned for S:
  - SeaStore multi-reactor support, tiering support
  - Scrub
  - Performance optimization

# TELEMETRY - QUINCY

- Enhanced opt-in flow
  - Allows devs data collection flexibility
  - Allows users to keep sending only what they opted-in to, whenever new data is collected (no forced re-opt-in)
  - Explicitly acknowledge data sharing license
- Telemetry channels
  - **basic** - cluster size, version, etc.
  - **crash** - anonymized crash metadata
  - **device** - device health (SMART) data
  - **ident** - contact info (off by default!)
  - **perf** - various performance metrics (off by default)

- Focus on crash reports analysis
  - Integration with bug tracker
  - Daily reports on top crashes in wild
  - Fancy (internal) dashboard
- Extensive device dashboard
  - See which HDD and SSD models ceph users are deploying
- Public dashboards!
  - https://telemetry-public.ceph.com/
  - Clusters, devices

# TELEMETRY - REEF

- Work continues on backend analysis of telemetry data
  - Tools for developers to use crash reports to identify and prioritize bug fixes
  - Perf counters analysis
- Adjustments in collected data
  - Adjust what data is collected for Reef
  - Periodic backport to Quincy (we re-opt-in)
  - e.g., which orchestrator module is in use (if any)
- Drive failure prediction
  - Building improved models for predictive drive failures
    - Collaborating with drive manufacturers
    - Initial focus on flash failure prediction models
  - Expanding data set via Ceph collector, standalone collector, and other data sources

# DASHBOARD - QUINCY

- **Cluster Expansion Wizard:**
  - After bootstrapping a minimal cluster, the Dashboard guides users to expand their clusters
- **Unified NFS management**
- **Integrated reporting of issues**
  - Ceph defects can now be directly reported from the Dashboard
- **Host management**
  - Patterns to add multiple hosts
  - Improved labels
  - Support for host draining
- **Service management**
  - Ingress (keepalived/HAProxy)
  - SNMP gateway
- **Monitoring**
  - SNMP support
  - 39 new alerts added to the existing 18
  - Improved highlighting of nearfull/full events

# DASHBOARD - REEF

- **RGW Advanced Workflows (user roles/policies, bucket policies, lifecycle, notifications...)**
- **RGW Server-side encryption**
- **Multi-site:**
  - Complete support for RBD
  - RGW Multi-site set-up
- **Operational improvements**
  - 1-click OSD creation
  - Improved capacity planning
  - Ceph auth and user management
  - Distributed QoS profiles
  - Cluster upgrades
- **Observability**
  - Centralized Logging (ELK-Loki based)
  - Multi-cluster monitoring

# CEPHADM - QUINCY

New Features

- SNMP Support
- Colocation of Daemons (mgr, mds, rgw)
- osd memory autotuning
- Integration with new NFS mgr module
- Ability to zap osds as they are removed
- cephadm agent for increased performance/scalability

Robustness

- Lots of small usability improvements
- Lots of bug fixes
  - Backported into Pacific already
- Ongoing cleanup of docs.ceph.com

# CEPHADM - REEF

- OS Tuning Profiles
  - Manage sysctl settings across hosts using cephadm
- Staggered Upgrades
  - Allow upgrading by one daemon type/service at a time
  - Can tell cephadm to only upgrade X number of daemons then stop
- Simplified rgw multisite workflow
  - Still WIP, should be done for Reef release
- Cephadm is now “compiled” (by py zipapp)
  - Will allow splitting the (nearly 10000 line) cephadm binary into multiple files.
  - Should have minimal user impact
  - Will be publishing the “compiled” version with the release instead of expecting users to curl from github
  - Should also be simple for users to “compile” on their own from the source tree as long as they have Python >= 3.5 (just run the “build.py” python script)
- Auth Key rotation for ceph daemons
  - ceph orch daemon rotate-key <daemon-name>

# ROOK

- Rook v1.10
  - Supports Pacific and Quincy
  - Removed support for Octopus
- New Krew plugin to aid with troubleshooting scenarios
  - Start mon or OSD daemons in maintenance mode to run ceph-bluestore-tool, etc
  - Repair mon quorum from a single healthy mon after quorum is lost
  - Show detailed cluster status
- Support for NFS snapshots, restore, clone, and resize

# RBD

- NVMeoF target gateway
  - Initial single-gateway-in-single-gateway-group implementation
  - Discovery service, deployment implementation in progress
- librbd migration to boost::asio reactor
  - Event driven; uses neorados
  - May eventually allow tighter integration with SPDK
  - Much higher throughput per client
- Persistent write-back cache stabilization
- rbd-mirror stabilization and hardening
  - Ensure correct operation when daemon restarts - pick up where it left off
  - Consistent per-image metrics (using new per-node exporter framework)
  - Scale testing
- Research into log-structured data format - https://github.com/CCI-MOC/lsvd-rbd

# RGW

- Policy-Based Rate Limiting
    - Per-user and per-bucket rate limit policies
        - enforced independently by radosgw instances in current implementation
- SSE-S3
    - transparent, server-managed encryption
    - PutBucketEncryption (S3 API)
- S3Select Enhancements
    - Parquet object format support
- Multisite Replication Enhancements (In Progress)
    - Dynamic Bucket-Index Resharding
    - Optimized Replication Intent Logs (OMAP Offload)
    - Sync fairness - load balancing across RGWs

# RGW

- Cloud Tiering
  - S3 Lifecycle Transition to S3 Cloud Targets
    - e.g., tier to AWS or other remote by storage class transition
- RGW Standalone (Prototype)
  - Based on Zipper flexible backing store abstraction
  - Building block for disconnected S3 object storage profiles (e.g. edge computing)
- Observability - distributed tracing using OpenTelemetry + Jaeger
  - Helpful for troubleshooting
  - Deployable with cephadm

# CEPHFS

- cephfs-top
  - more useful metrics added (avg latency, IO bandwidth of clients)
- improved hardlink tracking
  - https://tracker.ceph.com/issues/54205
- layering support for cloning from snapshots
  - (fast) clone a snapshot