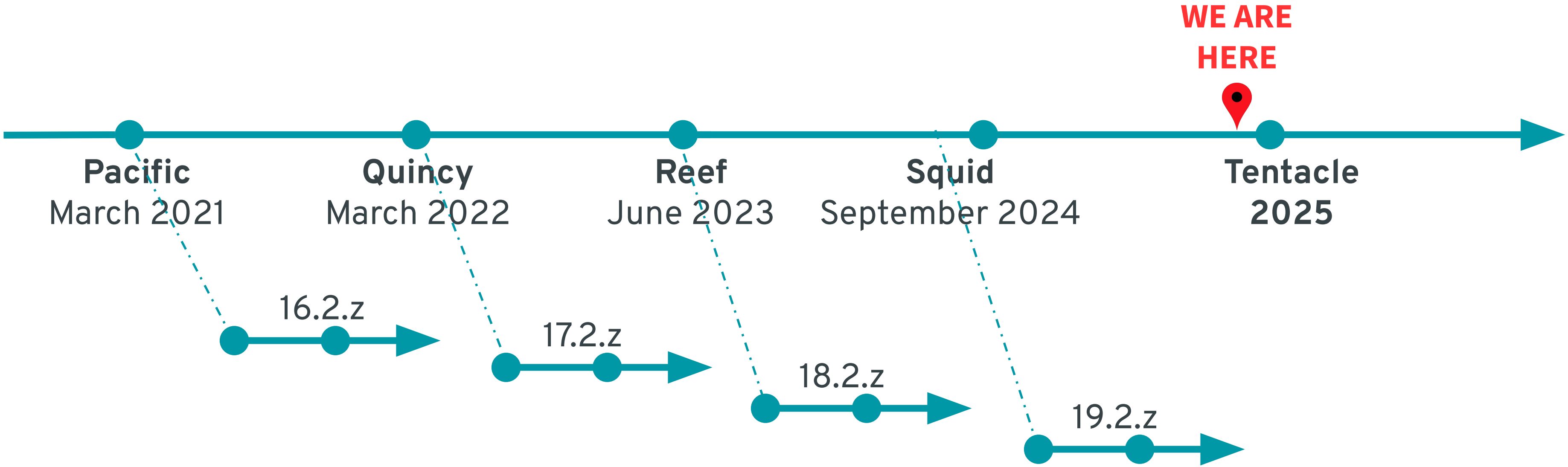
Keynote - State of Ceph

Neha Ojha
Ceph Day Silicon Valley - 2025.03.25

Where we stand in terms of contributions…

- ~1400 contributors
- 700K+ lines of source code changed
- 150K+ overall commits

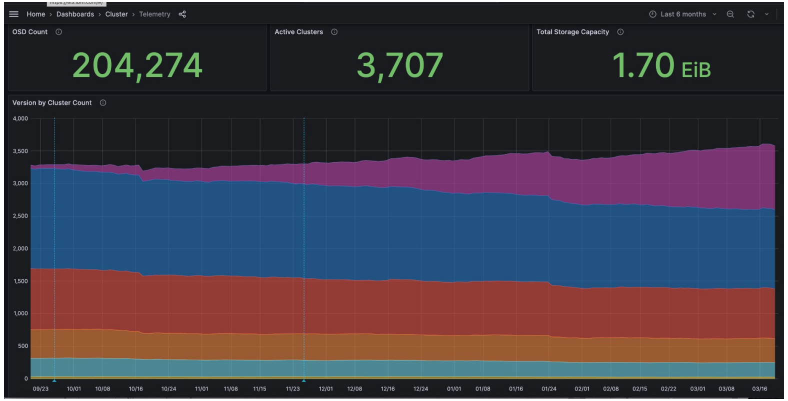
Ceph Release Schedule



- Stable, named release every 12 months
- Backports for 2 releases
  - 17.2.8 final Quincy release
- Upgrade up to 2 releases at a time
  - Pacific ➝ Reef, Quincy ➝ Squid, Reef ➝ **Tentacle**

PROJECT MILESTONES

- Cephalocon 2024 CERN, Geneva was a huge success!
- Ceph footprint reaches ~ 1.7 exabyte in Telemetry!
- Releases
  - Squid is the latest GA release
  - Tentacle (20.x.x!) is in dev phase
- Ceph upstream community engagement
  - **Ceph User Council** initiative
  - Slack for community engagement
  - Ceph Meetup group
- Community outreach continues
  - Google Summer of Code
  - Grace Hopper Open Source Day



Home › Dashboards › Cluster › Telemetry

Last 6 months

OSD Count: 204,274

Active Clusters: 3,707

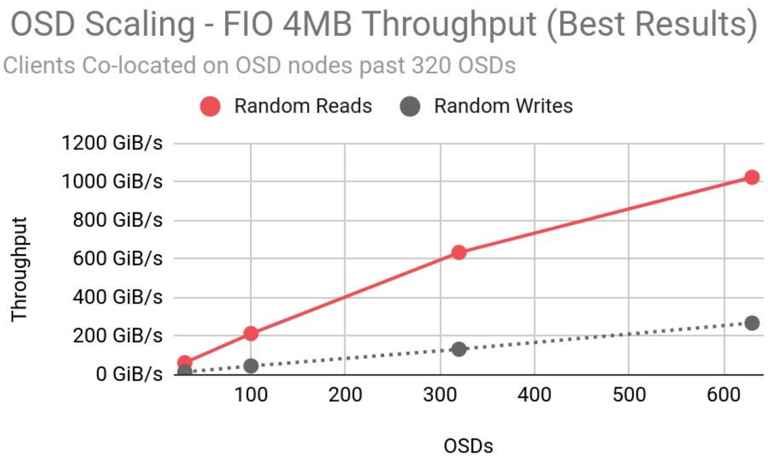
Total Storage Capacity: 1.70 EiB

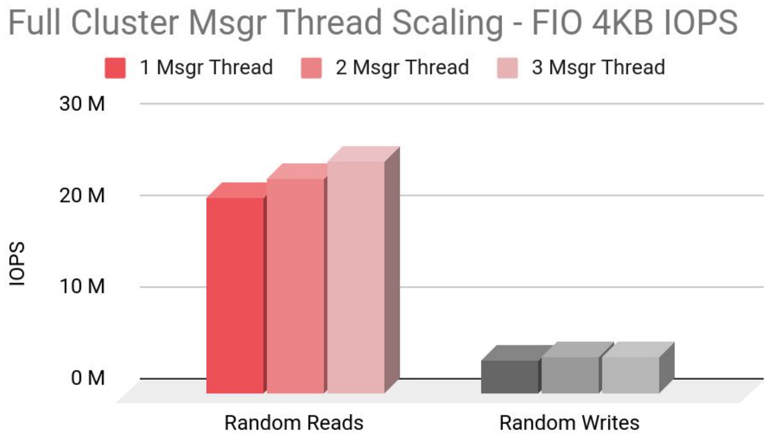
Version by Cluster Count

PERFORMANCE - NEW RECORDS

Upstream Ceph has achieved 1 TiB/s and 25M+ IOPS!

Congratulations everyone for your hard work!





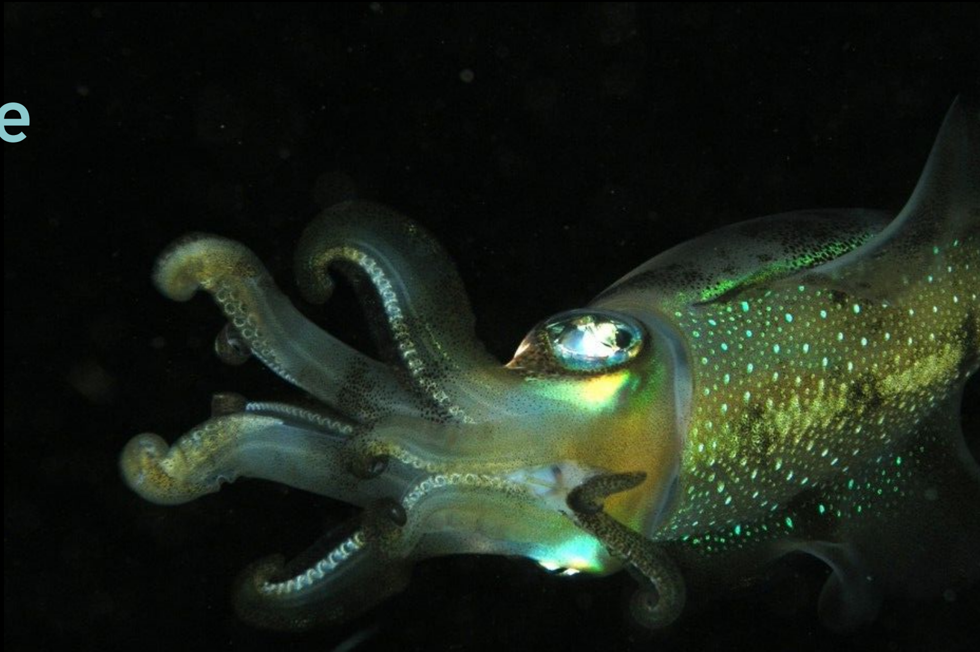
COMMUNITY UPDATES

- New Ceph Foundation Structure
  - new Diamond Members (Bloomberg, IBM, 45Drives) and revamped membership tiers
  - DigitalOcean joined as a silver member!
- Events Updates - 2025
  - Ceph Day India, Jan 23, 2025
    - Second year in a row success!
  - More Ceph Days
    - Ceph Day London, June 4, 2025
    - Ceph Day Berlin, Nov 13 - 14, 2025
    - Ceph Day NYC, TBA
    - Ceph Day Seattle, TBA
  - Cephalocon 2025!
    - In planning stage

TECHNICAL FOCUS AREAS

Consistent performance at scale	Ease of Use
Expanded Protocol Support and Use Cases	Storage Efficiency

What’s coming in Tentacle



RADOS - TENTACLE

- Erasure Coding improvements to reduce TCO and improve performance
    - EC Partial stripe reads (8x performance improvement for small reads)
    - EC Stop padding objects to stripe size (capacity saving + performance optimization)
    - EC Partial stripe write optimization (3.5x performance improvement for small writes)
    - EC Change default plugin for erasure coding from Jerasure to ISA-L (performance)
    - Align buffers to eliminate memcpy's on EC I/O path (performance)
- OMAP listing performance – improvements for RGW bucket listing
    - (5x performance improvement for CEPH_OSD_OP_OMAPGETVALS)
- Scrub - simplify control over deep scrub and scrub scheduling
- Ceph-mgr robustness
    - Better handling of module loading sequence
    - More granular performance counters
- Availability score of Ceph over time
- Ability to unstretch a cluster in stretch mode

BLUESTORE - TENTACLE

- Compression performance improvements
- Performance - Bluefs WAL v2, hybdrid_btree2 disk allocator
- Scraper for analyzing and replaying real-world workloads
- Health warnings for highly fragmented OSDs
- Object content recompression/defragmentation during scrubbing

CRIMSON - TENTACLE

- Crimson
  - Groundwork to switch backend to an architecture native backend **SeaStore**
  - Performance focus
  - Filling out OSD features:
    - Background scrub, PG splitting, EC, QoS
- SeaStore:
  - Performance profiling and optimizations:
    - Partial extent caching
    - PGLog optimization
    - LBA operation batching
  - Support clone range
  - More efficient device tiering
  - Random block manager for SLC/HDD

RGW - TENTACLE

- S3 Additional Checksums and GetObjectAttributes
- D4N data cache MVP
  - Distributed writeback cache, can use local SSD or Redis
  - Collaboration with Massachusetts Open Cloud, Boston University, Northeastern University
- In-Order Bucket-Index Sharding
  - Permit arbitrary sharding of index pages preserving ListObjects in linear time
- Bucket resharding: optimize and minimize blocking with log-based approach
- Multisite
  - Bucket-Index cleanup after re-sharding in the general case
  - Fully-consistent data-log updates

RGW - TENTACLE

- More CloudTier and Backup Milestones
  - S3 RestoreObject against cloud/archive targets (e.g., Tape)
- S3 Inventory
- Deduplication for large objects
- Improve end-to-end distributed tracing with OpenTelemetry
- New APIs
  - S3 bucket logging: initial support for journaling notifications for backup
  - PutBucketOwnershipControls to disable bucket/object ACLs

RBD - TENTACLE

- Disaster recovery
  - Snapshot-based mirroring of consistency groups
    (operate on “rbd group” groups of images as a whole, including failover/failback)
  - Mirroring to a differently named RBD namespace for multi-tenancy support
- Live migration
  - Import image from another Ceph cluster
  - Import from a NBD export
  - Improving support for importing encrypted images
- Kernel client support for upmap read balancer and CRUSH MSR rules

NVMe-oF - TENTACLE

- HA support
  - Failover and Load balancing
- GW Groups
  - Up to 4 GW groups
  - 8 GWs in a group
- Security
  - NVMe in-band authentication
  - Namespace masking
  - mTLS, and TLS PSK
- Scalability
  - Up to 128 Subsystems in a GW group
  - Up to 1024 namespaces in a GW group
  - Up to 32 GWs in a cluster (4 GW groups)
- Dashboard
- Alerts
- Testing and Teuthology suite

CEPHFS - TENTACLE

- Case-insensitive directory trees, for interoperability with Samba
- Store subvolume metadata with libcephsqlite for better handling of full clusters
- Proactive monitoring and alerts for metadata-heavy workloads
- Further stabilization of fscrypt for userspace clients (fuse, libcephfs)
- Improved NFS and SMB integrations

DASHBOARD - TENTACLE

- Management and Workflows
  - Multi-cluster management
  - NVMe/TCP Management
  - RGW Multisite automation
  - OAuth2 SSO Integration
  - SMB Management
- Monitoring
  - Multi-cluster dashboard
- UI/UX Improvements
  - Carbon Design System
- Continuous improvement
  - Upgrading Angular
  - Upgrading the monitoring stacks

CEPHADM - TENTACLE

- SMB orchestration
- Multi-arch image support
- Simplified OSD management
  - One-shot orchestration
  - No-replace marker for OSDs
- Disconnected install improvements
  - Simpler initial config of images
  - Mirror images locally (e.g. with skopeo)
- New services:
  - **mgmt-gateway:** enhanced security for the cluster access
  - **oauth2-proxy** integration for SSO/OIDC support
- Monitoring high-availability support
- Automated certificates management

ROOK - TENTACLE

- Mirroring for RADOS namespaces
- OSD migration to enable encryption as a day 2 operation

Ceph-CSI - TENTACLE

- Mirroring of VolumeGroups
- Space efficient replication for cloned volumes (RBD)
- Change Block Tracking
- Shallow cloning for Read-Only-Many volumes (RBD and NFS)
- Cross RADOS namespace mirroring

TELEMETRY - TENTACLE

- Add data about cluster usage
  - Stretch cluster
  - Multisite
- Scrub metrics collection
- Perf counters analysis

DOCS

- Beginner’s Guide
  - https://docs.ceph.com/en/latest/start/beginners-guide/
- IcePic initiative - improving online help, starting with most common CLI
- Line-editing across the docs
- Ceph Quarterly - summary of recent Ceph developments
- See something that could be improved? Add to the pad or email Zac:
  - https://pad.ceph.com/p/Report_Documentation_Bugs
  - ZAC.DOVER <at> PROTON.ME

TEUTHOLOGY / BUILD / CI

- ceph-devstack
  - Run teuthology integration tests using local containers as test nodes
  - Can run on CentOS, MacOS, Fedora
- pulpito-ng - New UI and API for interacting with test system
  - Easier to use scheduling, test run management, results viewing and analysis
- Sepia lab build improvements
  - Sccache for distributed build cache: large speedup in build time
  - Better pipelining, less compression for faster dev builds
  - Container-based build environments for more isolation, dev containers

PERFORMANCE - BENCHMARKING TOOLS

- Benchmarking improvements
  - For CBT:
    - Standardised benchmark performance tool for the community to measure Ceph performance using a consistent and deterministic approach.
    - Support for block and object using FIO, hsbench and elbencho. (File to come).
    - Single click, point CBT at your cluster and it will determine the configuration and how to configure it.
    - Automatic report generator with 30+ response curves. Random, sequential, mixed read/write.
    - Comparison tools to compare your results. and optional upload results to github repository to compare results with other systems. Useful for the community to use for performance sizing of Ceph clusters.
    - Full statistics collection and profiling capabilities for deep analysis.
  - Extending OMAP testing

Thanks!