WHAT'S COMING IN CEPH

OCTOPUS
FIVE THEMES

Usability
Quality
Performance
Multi-site
Ecosystem
FIVE THEMES

- Usability
- Quality
- Performance
- Multi-site
- Ecosystem
THE ORCHESTRATOR API

- Partial consensus to focus efforts on
  - Rook
  - ssh (or “Run this command on that host”)
- Expose all functionality via CLI
  - Provisioning, replacing OSDs
  - Provisioning and removing daemons
  - Blinking disk lights
- Expose key workflows via the Dashboard
  - Expanding the cluster
  - Manage hardware inventory
  - Replacing failed disks
  - ...

- Obsolete ceph-deploy
  - Bootstrap ceph-mon + ceph-mgr
  - Deploy rest of cluster via orchestrator
- Pave way for cleanup / refactor of docs.ceph.com
## UPGRADE AUTOMATION

- Codify non-trivial upgrade procedure in a mgr module
  - Define division of responsibility between orchestrator/deployer (e.g., rook) and mgr
  - Programatically gate each step on health alerts, etc.
  - Leverage orchestration API to restart daemons

1. Upgrade, restart mgr
2. `ceph upgrade start octopus`
3. Mgr module restarts mons one by one
4. Mgr module restarts osds one by one
5. ... 
6. `ceph upgrade status`, `ceph upgrade pause`
Like ‘rbd top’, ‘iotop’, but for CephFS
FIVE THEMES

Usability

Quality

Performance

Multi-site

Ecosystem
TELEMETRY AND CRASH REPORT

- Invest in backend tools to summarize, query, browse telemetry data
- Initial focus on crash reports
  - Identify crash signatures by stack trace (or other key properties)
  - Correlate crashes with ceph version or other properties
- First attempt: elasticsearch + kibana
- Encourage users to enable telemetry
- Carefully track feedback from those who opt not to
  - Ensure we are not collecting potentially identifying information
  - Adjust telemetry content accordingly
• Bi-weekly community call discussing docs
  ○ EMEA
    ■ 16:30 UTC
    ■ Monthly on second Wednesday
  ○ APAC
    ■ 01:00 UTC
    ■ Monthly on fourth Wednesday
  ○ Recordings are available on Ceph’s Youtube channel: http://bit.ly/2VAyLwW

• Discuss
  ○ Infrastructure / tooling
  ○ Content, gaps

• Find dedicated contract doc writer (funded via Ceph Foundation)
  ○ Overhaul doc structure and content
  ○ (Let us know if you are interested!)
● Ongoing per-component review of test coverage
  ○ e.g., RADOS, 2300 UTC, every 2 weeks
● Resurrect effort/investment in cross-version client/server testing
  ○ e.g., validate jewel clients work with nautilus cluster
● Reconsider downgrade test suite
  ○ Test downgrades of point releases (e.g., 14.2.2 → 14.2.1)
FIVE THEMES

- Usability
- Quality
- Performance
- Multi-site
- Ecosystem
RADOS QoS

- Partially implemented dmclock-based quality-of-service in OSD/librados
- Improvements to QoS portion in flight
  - e.g., Xie Xingguo and Yan Jun’s dmClock talk from yesterday
- Blocked because a deep queue in BlueStore obscures scheduling decisions
- Focus on understanding, tuning, and (hopefully) auto tuning bluestore queue depth
  - Device type (HDD, SSD, hybrid)
  - Workload (IO size, type)
● Sharding RocksDB show promising performance improvements
  ○ Reducing impact of compaction
  ○ Reducing required space in ‘db’ volume

● TRocksDB also promising
  ○ Segregates key and value data
  ○ Working to incorporate into upstream RocksDB
Each CephFS metadata operation is a round-trip to the MDS.
untar, rm -r tend are dominated by client/MDS network latency.
CephFS aggressively leases/delegates state/capabilities to the clients.
Allow async creates:
  - Linux client can immediately return, queue async operation with MDS.
  - Same for unlink.
  - untar etc are now blazingly fast!

Except it’s complex:
  - Current ordering between request, locks in MDS, and client capabilities.
● Focus on an end-to-end read and write path
  ○ Initially with memstore backend
  ○ Current simplifying assumption of single core per OSD
● Validate initial assumptions/design choices to inform next steps
● Multi-year effort targeting next-gen SSDs and persistent memory
FIVE THEMES

Usability
Quality
Performance
Multi-site
Ecosystem
RGW MULTISITE V3

- Bucket-granularity control of multi-site replication
- Pass-through support for S3, Azure, GCS, etc.
  - API translation
  - Enable migration of buckets to/from external object storage service
- Bi-directional replication of buckets to/from external object stores
- Tiering individual objects to external object storage
  - as part of lifecycle policy
CEPHFS MULTI-CLUSTER PLANNING

- Automate periodic snapshot + sync pattern
- Discussing more sophisticated models
  - Bidirectional, loosely/eventually consistent sync
  - Simple conflict resolution behavior?
FIVE THEMES

Usability

Quality

Performance

Multi-site

Ecosystem
ECOSYSTEMS

- Kubernetes + Rook
  - Make Ceph an obvious storage choice
  - knative serverless
- OpenStack
  - Continue supporting existing community
- Analytics
  - Make RGW a natural choice for data lakes and AI/ML platforms
  - Spark, Jupyter, etc
- ?
Do you like the 9 months cycle of @Ceph releases, asks @sweil. The #Ceph community wants your feedback: #cephalocon

- 46% Keep the 9 months
- 54% Go to 12 months

79 votes • 1 hour left

12:20 AM - 19 May 2019 from Barcelona, Spain
GET INVOLVED
USUAL FOSS TOOLS

- ceph-devel, ceph-users email list
- #ceph and #ceph-devel on irc.oftc.net
  - https://ceph.com/irc/
- Review pull requests
  - https://github.com/ceph/ceph/pulls
- Submit and/or comment on bugs
  - https://tracker.ceph.com/
WELCOME TO CEPH

Ceph uniquely delivers **object, block, and file storage in one unified system**.

**CEPH OBJECT STORE**
- RESTful Interface
- S3- and Swift-compliant APIs
- S3-style subdomains
- Unified S3/Swift namespace
- User management
- Usage tracking
- Striped objects
- Cloud solution integration
- Multi-site deployment
- Multi-site replication

**CEPH BLOCK DEVICE**
- Thin-provisioned
- Images up to 16 exabytes
- Configurable striping
- In-memory caching
- Snapshots
- Copy-on-write cloning
- Kernel driver support
- KVM/ibvit support
- Back-end for cloud solutions
- Incremental backup
- Disaster recovery (multisite asynchronous replication)

**CEPH FILESYSTEM**
- POSIX-compliant semantics
- Separates metadata from data
- Dynamic rebalancing
- Subdirectory snapshots
- Configurable striping
- Kernel driver support
- FUSE support
- NFS/CIFS deployable
- Use with Hadoop (replace HDFS)

See Ceph Object Store for additional details.
See Ceph Block Device for additional details.
See Ceph Filesystem for additional details.

Ceph is highly reliable, easy to manage, and free. The power of Ceph can transform your company's IT infrastructure and your ability to manage vast amounts of data. To try Ceph, see our Getting Started guides. To learn more about Ceph, see our Architecture section.
VIDEO MEETINGS

- Public Community Calendar
  - https://calendar.google.com/calendar/b/1?cid=OXRzOWM3bHQ3dTF2aWMyaWp2dnFxbGZwbzBAZ3JvdXAuY2FsZW5kYXJuZ29vZ2xILmNvbQ

- BlueJeans video meetings
  - Everyone is welcome

- Developer standups
  - Call attention to or ask questions about your pull request
  - Ask for guidance on a potential change, improvement
GETTING STARTED

- Proper ‘Getting Started’ page in docs is in the works
CEPH DAYS!

https://ceph.com/cephdays

- Ceph Day Netherlands - Amsterdam - Jul 2
  - CFP open until early June
- Ceph Day CERN - Geneva - September 16
  - Will be finalized soon
- Ceph Day London - October 24
- Ceph Day Poland - Wrocław - October 28
- Propose your own! Contact us at events@ceph.io