Red Hat
Data Services

Ceph Storage Roadmap
Enterprise meets Community

Red Hat Cloud Storage and Data Services

Federico Lucifredi
Product Management Director
Red Hat Data Services

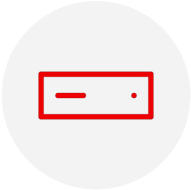
Sage Weil
Ceph Architect
Red Hat Office of the CTO

Red Hat

Intro to Ceph

Red Hat

Flexibility to meet the demands of tomorrow



Uses commodity hardware

- Runs on industry standard servers and TCP/IP networks
- HDDs, SSDs NVMe

Delivers reliability

- Fully distributed, no single point of failure
- Ensure data durability via replication or erasure coding
- Expand or shrink clusters as required
- Federate multiple clusters across sites with asynchronous replication and disaster recovery capabilities

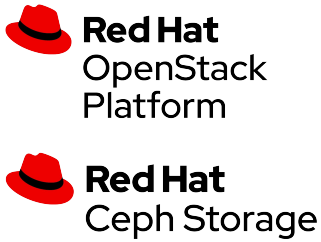
Improves versatility

- A single cluster can support object, block, and file workloads
- Scale out within a cluster for capacity/speed
- Add or remove hardware while system is online —even if it's under load
- Apply updates without interrupting service

Red Hat Ceph Storage tier 1 use cases

Ceph standalone
Leading on-prem for S3 at scale

- Object storage
- Block storage
- File storage
- Leading the on-premise object market at 10-Petabyte+ scale
- Setting the standard for S3 compatibility outside of AWS



Ceph for OpenStack
#1 in OpenStack storage

- Cinder block storage
- Nova ephemeral storage
- Glance image storage
- Swift object store
- Manila file storage
- Advanced integration
- Unified management
- Hyperconverged and Edge capabilities

Ceph for OpenShift
Self-managing storage

- Powered by Red Hat Ceph Storage
- Automated by Rook and integrated with NooBaa
- Advanced integration and ease of use
- Adds support for stateful workloads to OpenShift

Subscription Lifecycle

Red Hat
Ceph Storage

Support model

24x7 Support

Patches

Consulting services *(option)*

Base Lifecycle

12 months → bug fixes, security fixes, backports

24 months → bug fixes, security fixes

Extended Lifecycle (ELS) (option)

24 months → bug fixes, security fixes

Optional services incur additional charges; terms apply as found in support agreement

What's new in RHCS 5

Red Hat

What's new?

Red Hat Ceph Storage 5

FUNCTIONALITY

New integrated control plane
Stable management API
NFS filesystem support

SECURITY

WORM (object lock)
FIPS 140-2 Cryptography
Key management integration

PERFORMANCE

80% increase in block performance for virtual machine and container hosting..

EFFICIENCY

Reduced resource consumption for small file
Complete set of data reduction options

FUNCTIONALITY

New Manageability features

Integrated control plane
Stable management API
OSD replacement workflows
Object multi-site monitoring

New CephFS capabilities

NFS access option
Erasure code option
Snapshot based geo replication

New RBD capabilities

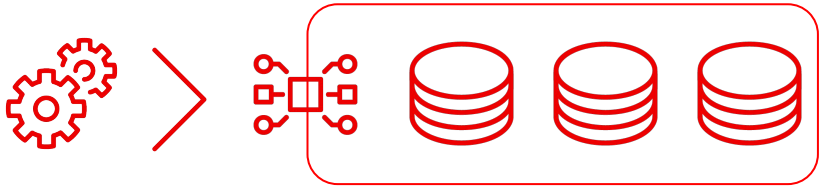
RBD snapshot based migration across clusters

FUNCTIONALITY

Manageability features

Integrated control plane (Cephadm)
Stable management API
OSD replacement workflows (UI & CLI)
Object multi-site monitoring

Management API scripts written for a major version of Red Hat Ceph Storage will continue to operate unchanged for the version's life cycle.



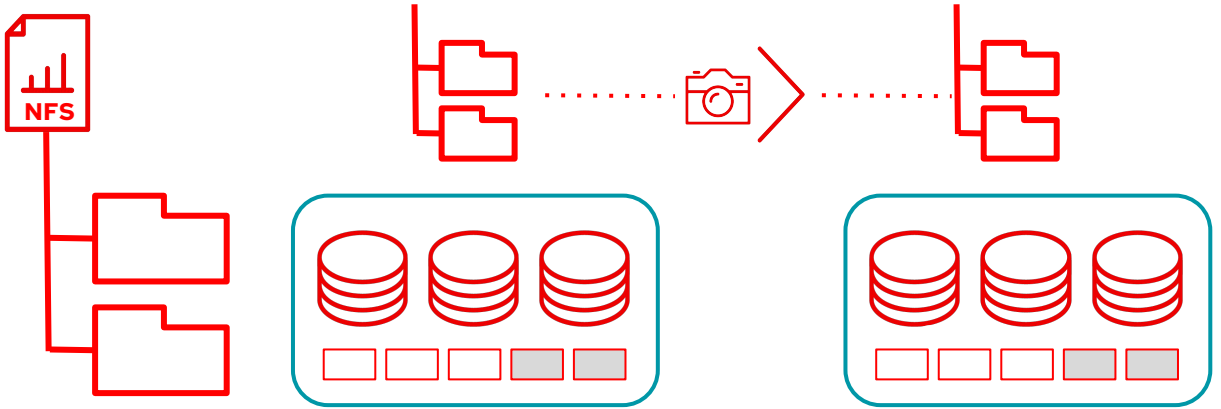
FUNCTIONALITY

New CephFS capabilities

NFS access option

Erasure code option

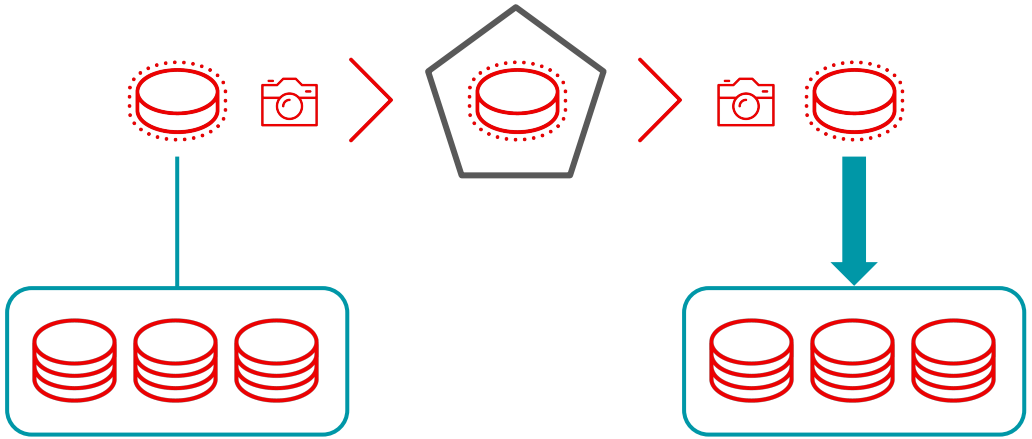
Snapshot based geo replication



FUNCTIONALITY

New RBD capabilities

RBD snapshot-based migration across clusters



FUNCTIONALITY

Business benefits

Manageability features
Reduced learning curve, simplified operations
Fast, simple scaling up or down
Additional interface capabilities
Maximized density with containerized setup

New CephFS capabilities
New access options, better monitoring, disaster recovery and data reduction

New RBD capabilities
Additional disaster recovery capabilities with lower operational cost
Moving between clusters is quick and easy

PERFORMANCE

Improved performance

Dramatic boost for virtual machines:

Improved block performance by 80%

New object benchmark HDD test results:

> 80 GB/s object aggregate throughput

Overhauled cache architecture

Improved scale

10+ billion objects in RGW

Continued object store scalability improvements

Better monitoring tools

CephFS "top" joins existing RBD top tool

PERFORMANCE

Improved performance

Dramatic boost for virtual machines:
Improved block performance by 80%

LibRBD data path optimization

Improvement shown in OpenStack benchmark

Red Hat OpenStack Platform

PERFORMANCE

Improved performance
Overhauled cache architecture

New read-only large object cache
Offloads reads of objects from cluster

Improved in-memory write-around cache
No longer serves reads (page cache), dedicated to batching writes

Optane write-back cache (tech preview)
Dramatically improving latency

Sources:
ceph.io blog article by Kyle Bader
blocksandfiles.com article (March 31, 2021)
scientific-computing.com article (April 1, 2021)

PERFORMANCE

Improved performance

New object benchmark HDD test results:
> 80 GB/s aggregate throughput

Improved scale

10+ billion objects in RGW
1.5 billion objects per node

"We were able to achieve a staggering 79.6 GiB/s aggregate throughput from the 10-node Ceph cluster utilized for our testing."
That's 85.5GB/sec from a disk-based data set composed of 350 million objects.

Kyle Bader, Red Hat

SEAGATE

PERFORMANCE

Improved performance–
CephFS ephemeral pinning
Metadata scalability enhancement

Improves the ability of multiple MDS servers to balance widely distributed workloads with a setting that distributes load in "round robin" fashion across all metadata servers.

Source: https://bit.ly/3oUwSvh

PERFORMANCE

Business benefits

Improved performance
Improved capability with same hardware, significantly faster access to data

Improved economics
Lower footprint cost

Improved scale
Start small and scale up when there is a business need - no disruptive forklift upgrades

SECURITY

Write once, read many (WORM)
S3 object lock enables WORM governance

Federal Information Processing Standard
FIPS 140-2 cryptographic libraries

Enhanced access control
Token Based with Identity Federation (STS)

External authentication integration
Key management service integration

Granular Object Encryption
Per-object encryption, key management integration (SSE-KMS)

SECURITY

WORM Object security and governance

S3 object lock provides read-only capability

S3 object lock to store objects using a write-once-read-many (WORM) model.

Object lock can help prevent objects from being deleted or overwritten for a fixed amount of time or indefinitely.

Standard certification planned for coming year.

SECURITY

Federal Information Processing Standard
FIPS 140-2 cryptography

Red Hat Ceph Storage can use FIPS 140-2 validated cryptographic modules when run on Red Hat Enterprise Linux 8.1

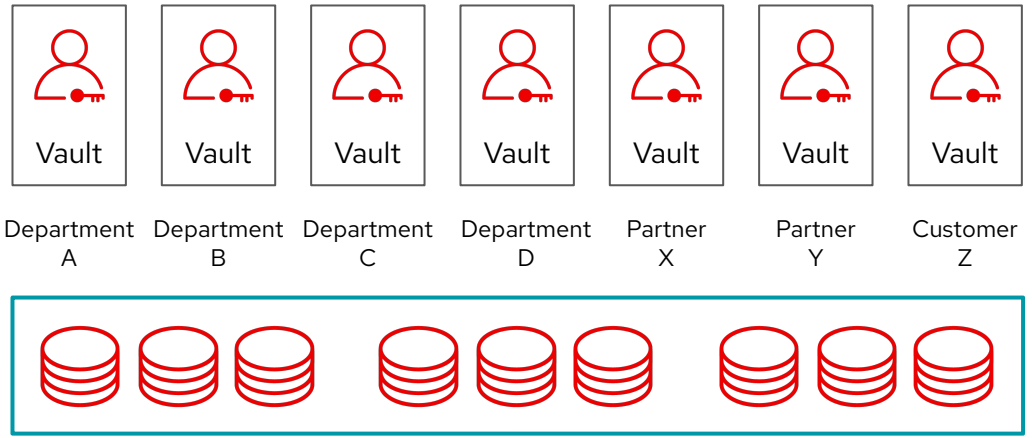
Newer versions are certified on DISA's schedule

SECURITY

External authentication integration

Key management service integration with Vault, IBM SKLM, OpenStack Barbican

Open ID Connect identity support (OIC)



SECURITY

Business benefits

Enhanced access control
Authentication and authorization can now be centralized and/or federated

WORM Object security
Object data can now be secured against alteration for archive or governance use cases
Can be used against ransomware attacks

Government-validated cryptography
Enables certification of FedRAMP environments

EFFICIENCY

Multi-site capabilities
RADOS object gateway across sites including hybrid cloud connectivity options

Resource consumption
Improved space utilization for small files

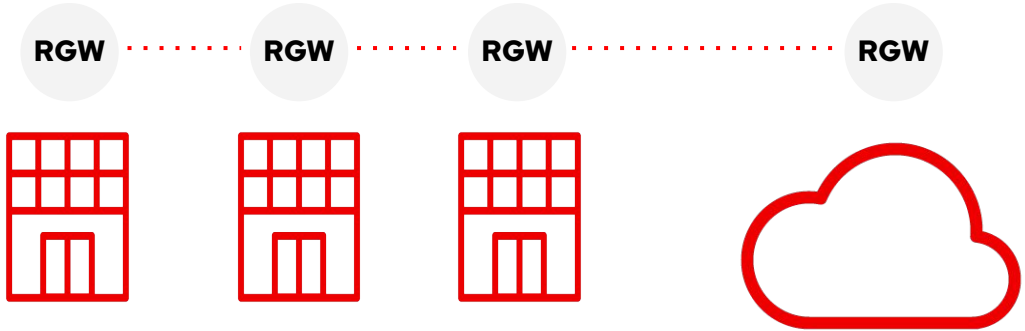
Improved reliability
Erasure coding recovery with k shards

Object offload to public cloud (5.1)
Using bucket policies and AWS

EFFICIENCY

Multi-site capabilities

RADOS object gateway across sites including hybrid cloud connectivity options

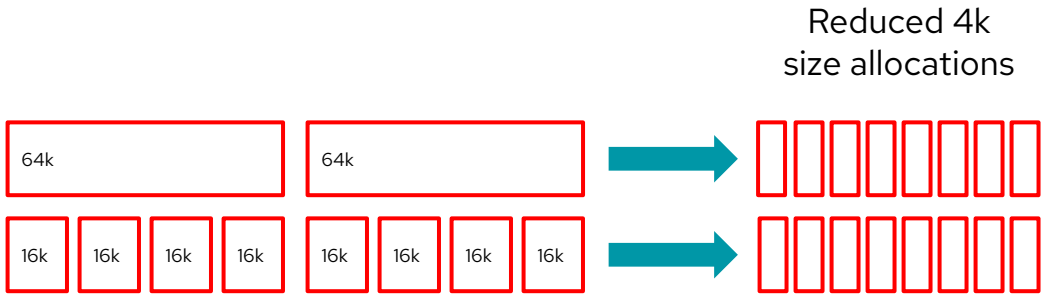


EFFICIENCY

Resource consumption

Improved space utilization for small file

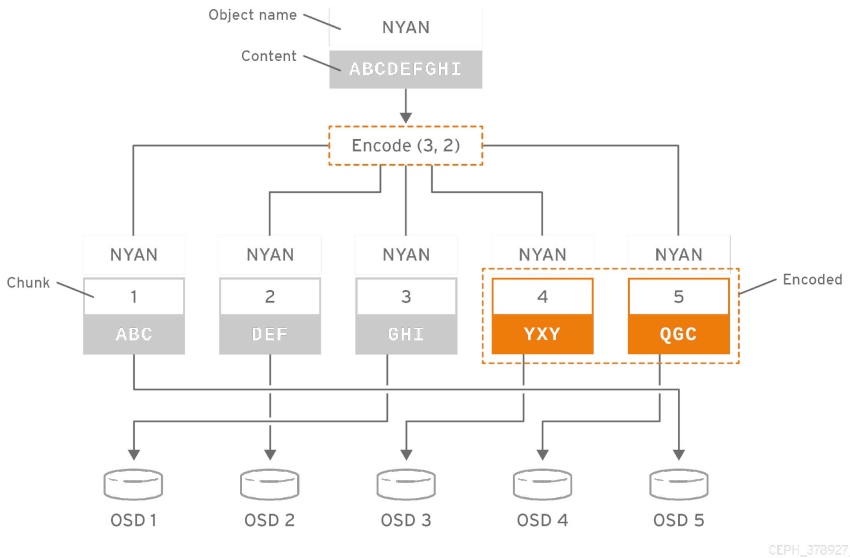
Bluestore 4k minimum allocation size replacing the current 64k (HDD) and 16k (SSD) when managing small objects, significantly reducing overhead for storage of small objects.



EFFICIENCY

Improved reliability

More robust erasure coding recovery with k shards



EFFICIENCY

Business benefits

Multi-site capabilities
Better data availability and improved accessibility, even at multiple locations

Resource consumption
Smarter resource usage and improved cost efficiency

Object offload to public cloud (5.1)
Better control by defining policies for data placement, aligning to business needs or demands

Summary

Red Hat

Summarized

Efficiency

- Full data reduction option range
- 16X better space use on HDD small file
- 4X better space use on SDD small file

Security

- WORM object lock API
- FIPS 140-2 cryptography
- Interoperate with KMIP key managers
- Messenger v. 2.1 backplane encryption

Performance

- Optimized LibRBD data path: 80% faster
- Overhauled cache architecture
- 10+ billion objects in RGW
- CephFS "Top" tool

Manageability

- New integrated control plane
- Integrated monitoring and management dashboard
- OSD replacement workflows (CLI & UI)
- RGW multisite monitoring

APIs and protocols

- Management API
- CephFS + NFS
- CephFS geo-replication

What to expect from Ceph Quincy

Sage Weil
June 2021

RELEASE SCHEDULE

- Stable, named release every 12 months
- Backports for 2 releases
  - Bug fixes and security updates
  - Nautilus reaches EOL shortly after Pacific is released
- Upgrade up to 2 releases at a time
  - Nautilus → Pacific, Octopus → Quincy
- Released as packages (deb, rpm) and container images
- Process improvements (security hotfixes; regular cadence)

QUINCY

FIVE THEMES

Usability

Quality

Performance

Multi-site

Ecosystem

CEPHADM AND ROOK

Cephadm

- RGW multisite bootstrapping
- Support for SMB
- Improved support for ingress (haproxy + keepalived)
  - NFS, SMB, mgr services (like dashboard)
- Improved scalability
  - Per-node agent, similar to kubelet
- Smarter scheduling
  - Detect and avoid port conflicts
  - Resource-aware scheduling (memory, CPU)
- Automatic tuning of memory usage
  - MDS, MON

Rook

- Improved orchestrator and dashboard integration
- Support for ingress
- Support for SMB
- OSD/device management
- Integrated testing

DASHBOARD

- Cluster installation wizard
- Usability improvements
- Improved performance
- Pagination, sorting, and field selection
- RBD snapshot mirroring
- RGW multisite
- RGW advanced features
- CephFS mirroring
- CephFS snapshot scheduling
- CephFS subvolumes
- CephFS top

- QoS management
- HA/ingress monitoring
- Customized alerts
- Customized grafana dashboards

RADOS USABILITY

- mclock scheduler used by default!
  - Eliminates the need to set sleep throttles for background ops in the OSD
  - Automated benchmarking on startup to set mclock related configuration
- PG Autoscaler profiles
  - Scale-up - existing behavior
  - Scale-down - default in new clusters, for better performance out of the box
- Balancer to consider OSD utilization, not just number of PGs per OSD
- Display degree of degradedness in ceph health

OTHER USABILITY

RBD

- Replace rbdmap init.d script with systemd unit(s)

CephFS

- mds_memory_target config
- Snapshots for subvolume groups

RGW

- radosgw-admin … -> ceph rgw …
- Automatic dashboard configuration
- RBD snapshot pseudo-bucket

FIVE THEMES

Usability

Quality

Performance

Multi-site

Ecosystem

RADOS ROBUSTNESS

- Monitors dynamically adjust trimming rate
- Further improvements to PG deletion performance
- Manager scalability
  - autotune mgr_stats_period
  - Progress module: update events at a configurable interval, not on every PGMap update
  - Insights module: avoid persisting every health update, use common .mgr pool
- Improvements to slow op logging
  - Do not log everything - OSDs to report a configurable number of slow ops
  - Avoid using cluster log, going through paxos and persisting
  - Use mgr.log to log slow ops reported by OSDs
  - Make LogMonitor more efficient

RBD ROBUSTNESS

- librbd
  - client-side encryption (support for clones)
  - import/export consistency groups
  - persistent write-back cache improvements
- rbd-nbd
  - support single daemon managing multiple images/devices
  - safe reattach after daemon upgrade (pending kernel change)

CEPHFS ROBUSTNESS

- cephfs-mirror
  - scale-out/HA (and cephadm/rook support)
- MDS scrub scheduling
- Immutable file support
- cephfs-top: multi-fs support
- Expanded test suite/test coverage
  - Multi-mds export thrashing
  - Kernel fscache

SEPIA TEST LAB

- More hardware from the Ceph Foundation
  - Expanding the lab's Ceph cluster (dog food, archived test results)
  - More build machines (braggi)
  - More test nodes (gibba)
- Improved teuthology test infrastructure
  - Moved to a single process dispatcher (Shraddha Agrawal)
  - Replaced in-memory queue with limited features with postgres (Aishwarya Mathuria)
  - Enables larger scale test clusters
  - Ability to prioritize and use lab more efficiently
- Downgrade testing (WIP)
  - Downgrade within a major release (e.g. 16.2.4 -> 16.2.3)
  - Now feasible with cephadm!

FIVE THEMES

Usability

Quality

Performance

Multi-site

Ecosystem

QUALITY OF SERVICE

- Background QoS improvements
  - Optimize performance for HDDs
  - Account for background activities like scrubbing, PG deletion etc.
  - Further testing across different types of workloads and scale
- Default profile prioritizes client I/O over background operations
- Can be configured to give higher priority to background ops like recovery using the high_recovery_ops profile
- Client vs client QoS
- Dashboard integration

BLUESTORE

- Remove allocation metadata from rocksdb
  - Significantly improved small write performance
  - Rebuild allocation map in failure scenarios
- Improvements to split_cache
- Cache age binning
- OmapBench - simple omap benchmark

RGW

- Deduplication
- Refactoring
  - sync request flow and thread consolidation
  - cls_fifo for metadata logs, bucket index logs, notification event queue
- S3 SELECT parquet
- d3n, d4n caching
  - Gateway read-only or write-back caching

TELEMETRY

- Framework for Performance channel in telemetry
  - Collect useful performance data like osd perf counters to identify areas to optimize
  - Include more metrics that can help us profile user workloads
  - Devs to use this data to adapt existing guidelines and provide recommendations based on information aggregated across different clusters
  - Definitely worth opting in for!
- Information on enabled manager modules

PROJECT CRIMSON

Why

- Not just about how many IOPS we do…
- More about IOPS per CPU core
- Current Ceph is based on traditional multi-threaded programming model
- Context switching is too expensive when storage is almost as fast as memory

- New hardware devices coming
  - DIMM form-factor persistent memory
  - ZNS - zone-based SSDs

What

- Rewrite IO path in using Seastar
  - Preallocate cores
  - One thread per core
  - Explicitly shard all data structures and work over cores
  - No locks and no blocking
  - Message passing between cores
  - Polling for IO
- DPDK, SPDK
  - Kernel bypass for network and storage IO

- Goal: Working prototype for Pacific

CRIMSON: QUINCY MILESTONES

- Benchmark: handle RBD workloads
- Scrubbing implementation
- Support for snapshots
- Multi-core support
- SeaStore
  - Sufficiently feature complete for initial teuthology testing
  - Initial performance work
  - Initial work on RandomBlockManager, persistent memory support

FIVE THEMES

Usability

Quality

Performance

Multi-site

Ecosystem

CEPHFS-MIRROR

- Snapshot-based mirroring of directory to a remote cluster
- HA / scale-out support
  - Automatically spread mirroring workload across multiple daemons
  - Cephadm and Rook support for managing daemons
- Improved efficiency of incremental updates

RBD-MIRROR

- Snapshot-based mirroring of consistency groups
- Improved monitoring + metrics

RGW MULTISITE

- Dynamic resharding in multisite
- Lifecycle transition to cloud
- Zipper refactoring
  - Reusable internal interfaces to support multisite and programmable behaviors
  - Policy with LUA
  - Database backend for metadata
- Sync from cloud

FIVE THEMES

Usability

Quality

Performance

Multi-site

Ecosystem

ECOSYSTEM EFFORTS

- Auto-generate documentation from ceph config options
- Redmine tracker integration with telemetry

NEW DEVICES

- ZNS SSDs
  - 3D NAND … dense, but the erase blocks are huge
  - Zone-based write interface
  - Combines capacity, low cost, and good performance
  - Key focus of Crimson's SeaStore!
- Multi-actuator HDDs
  - Recent devices double IOPS in existing HDD package
  - Ceph treats them as two OSDs with shared failure domain
- Persistent memory
  - Will be well-supported (but not required) by Crimson
  - Recent support in RBD client-side write-back cache

NVMe FABRICS

- Client-side
  - NVMeoF target that presents an RBD device
  - Alternative to iSCSI
  - Can be combined with new hardware (e.g., SmartNICs like Nvidia's Bluefield) to present a NVME device on PCI bus while running gateway/librbd code on the card's "DPU"
  - Useful for "metal as a service" cloud infrastructure
- Server-side
  - Some discussion around Crimson "phase 2"
  - Enable primary OSD to write directly to replica OSD's devices
  - Mechanism to reduce CPU cost per IO

INTEGRATIONS / ECOSYSTEMS

- Maturing
  - Rook
    - Key focus: Ceph orchestrator / dashboard integration with rook
  - Serverless (e.g., knative)
    - RGW event notification via Kafka, AMQP
  - Spark
    - S3 SELECT
  - Multisite
    - interop with public cloud
- New
  - Apache Arrow / Parquet
    - Data interchange formats for data pipelines

CEPH FOUNDATION UPDATE

PREMIER MEMBERS

Bloomberg®

中国移动 China Mobile | 中移软件 CMSoft

DigitalOcean

intel

OVH

Red Hat

SAMSUNG

SOFTIRON

SUSE
We adapt. You succeed.

Western Digital.

XSKY

ZTE

GENERAL MEMBERS

arm

cloudbase solutions

clyso
cloud system solutions

croit

DiDi

EasyStack
open cloud computing

Intelligent Systems Services

ubuntu
Delivered by Canonical

QCT

SRIE

VEXXHOST

ASSOCIATE MEMBERS

BOSTON UNIVERSITY

CERN

CROSS
cross.ucsc.edu

FAS RC

grnet

MONASH University

UNIVERSITY OF MICHIGAN

NRF SARAO
National Research Foundation

Science & Technology Facilities Council

SWITCH

CURRENT PROJECTS

- Ceph documentation
  - Zac Dover, full-time technical writer
- ceph.io web site update
  - Spearheaded by SoftIron
  - Static site generator; github; no more Wordpress
  - https://github.com/ceph/ceph.io
  - Planned launch next month!
- Training materials
  - Working with Linux Foundation's training group
  - Building out initial free course material
  - Can support both self-paced or instructor-led
  - Potential in future for advanced material, paid courses, and/or certifications

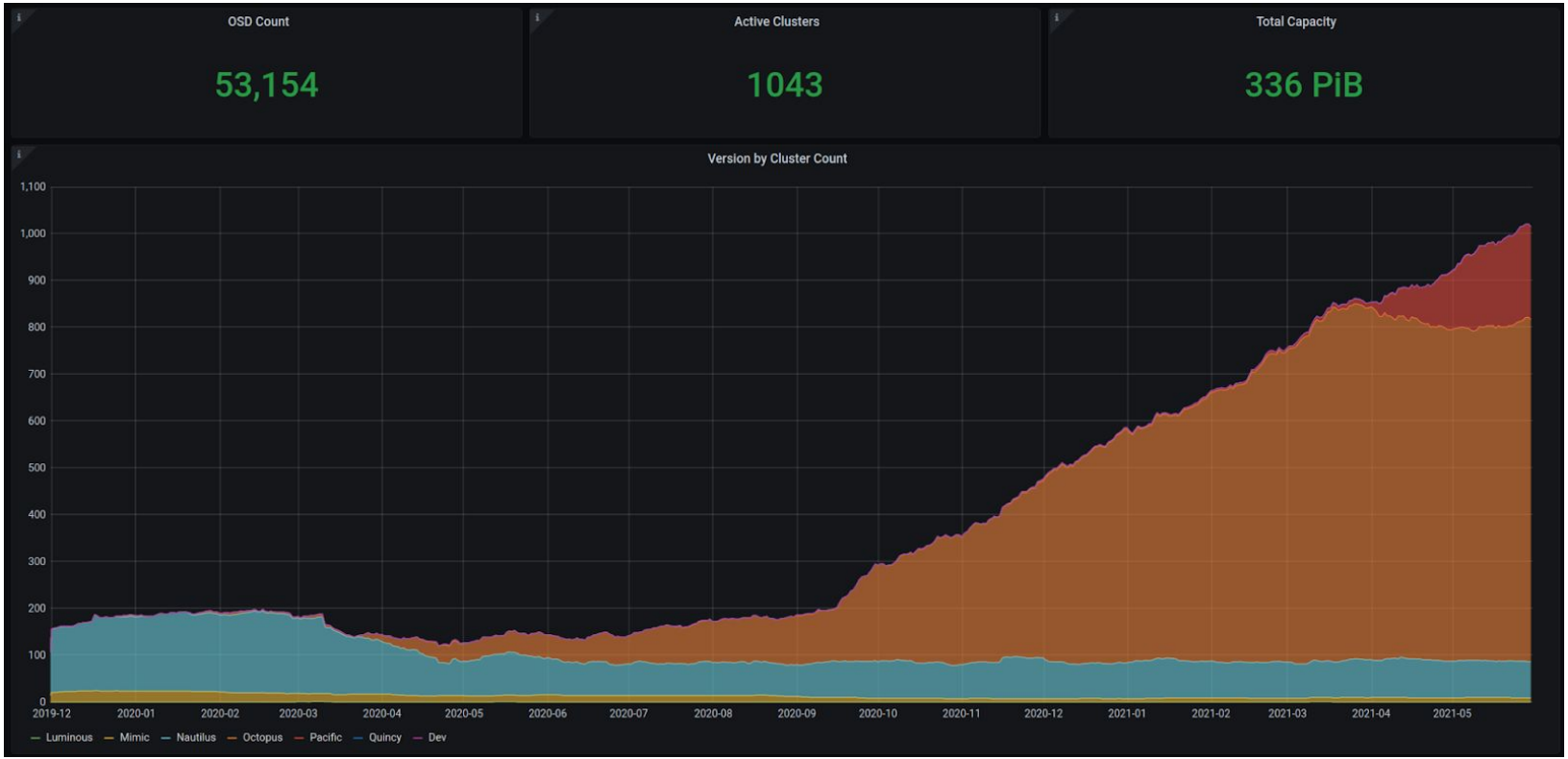
CURRENT PROJECTS

- Reducing public cloud spend
  - Build and CI hardware purchases for Sepia lab
  - We are now only hosting public-facing infra in OVH
- Lab hardware
  - Build machines
  - Expanding lab's Ceph cluster (more storage for test results, etc)
- Windows support
  - Contract with CloudBase to finish initial development, build sustainable CI infrastructure
  - RBD, CephFS
- New marketing committee

ARM AARCH64 SUPPORT

- Hardware donated by Ampere
- CI builds for teuthology, releases
  - CentOS 8 RPMs, Ubuntu Focal 20.04
  - Container images (based on CentOS)

TELEMETRY UPDATE



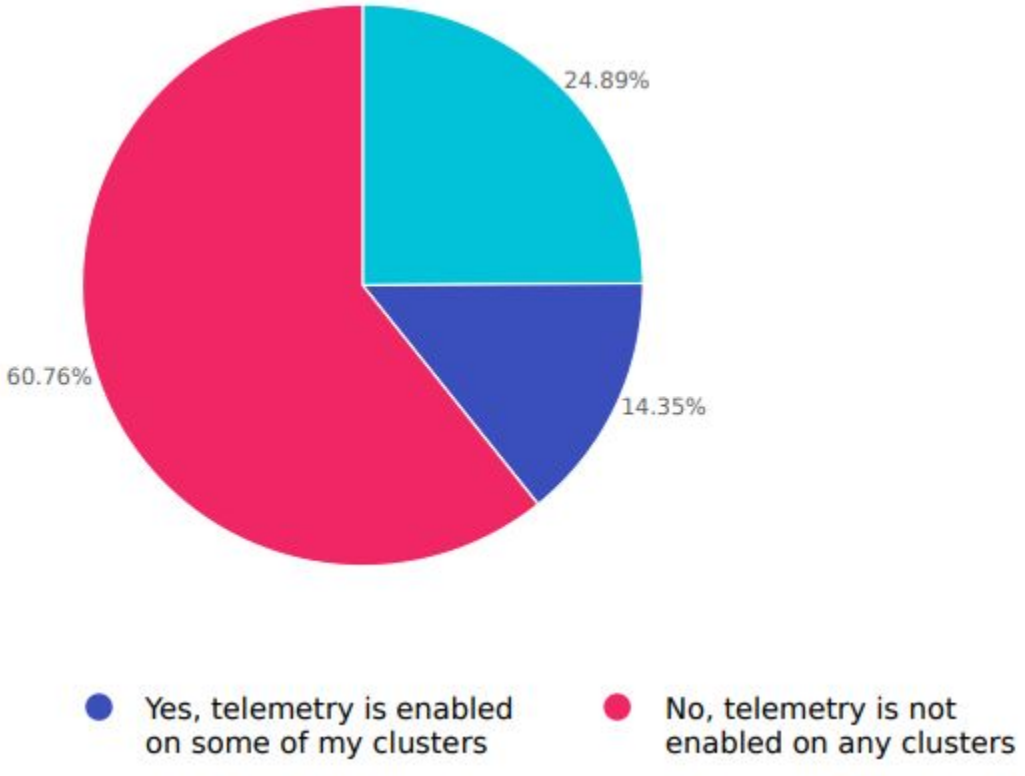
https://telemetry-public.ceph.com/

TELEMETRY AND CRASH REPORTS

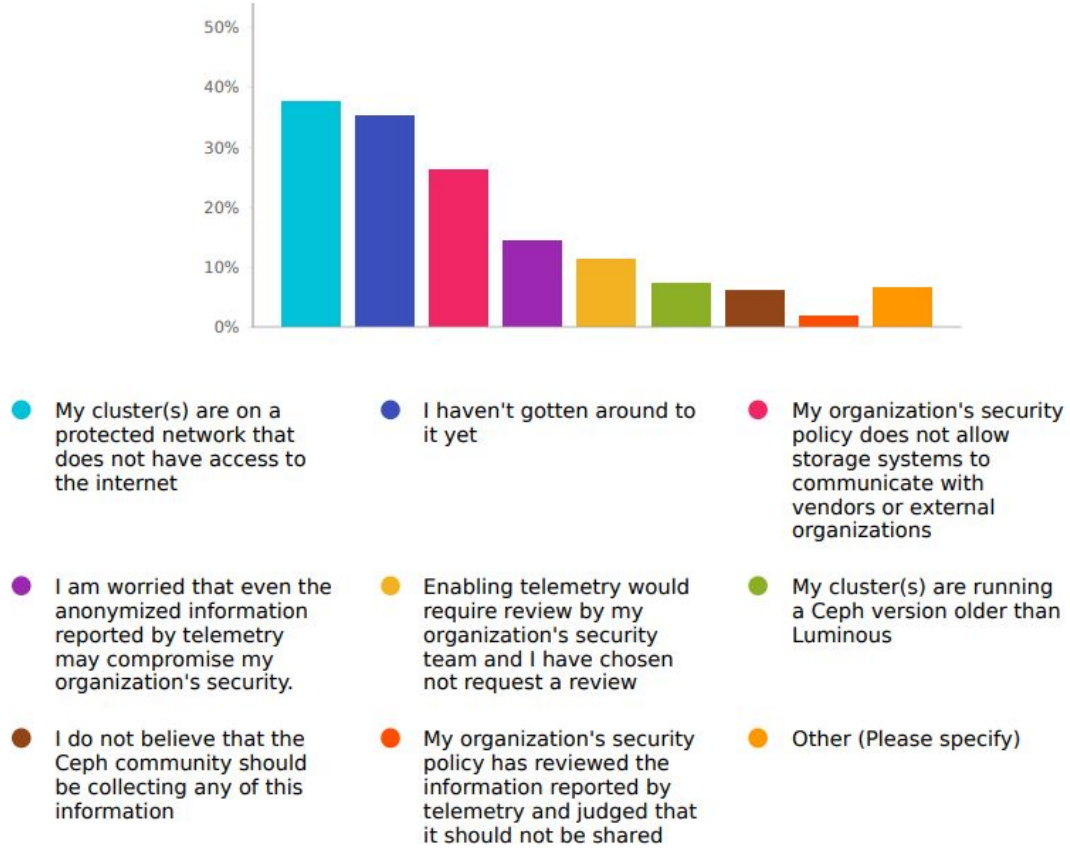
- Opt-in
  - Will require re-opt-in if telemetry content is expanded in the future
  - Explicitly acknowledge data sharing license
- Basic channel
  - Cluster size, version
  - Which features are enabled
- Crash channel
  - Anonymized crash metadata
  - Where in the code the problem happened, what version, etc.
  - Extensive (private) dashboard
  - Integration into tracker.ceph.com WIP
- Device channel
  - HDD vs SSD, vendors, models
  - Health metrics (e.g., SMART)
  - Extensive dashboard (link from top right)
- Ident channel
  - Off by default
  - Optional contact information
- Future performance channel
  - Planned for quincy
  - Optional more granular (but still anonymized) data about workloads, IO sizes, IO rates, cache hit rates, etc.
  - Help developers optimize Ceph
  - Possibly tuning suggestions for users
- Transparency!

https://telemetry-public.ceph.com/

IS TELEMETRY ENABLED?



WHY IS TELEMETRY NOT ENABLED?



BEYOND QUINCY...

CEPH R_____

https://pad.ceph.com/p/r

Vote now, or else!