Red Hat
Ceph Storage

# Ceph and ODF, and how they fit into your data storage needs

HVOpen October 2022

Patrick Ladd
pladd@redhat.com
https://people.redhat.com/pladd

# Red Hat Storage Timeline

| | | |
|---|---|---|
| December 18, 2003 | Sistina | GFS, LVM, DM |
| September 4, 2008 | Qumranet | KVM, RHEV, SPICE |
| October 4, 2011 | Gluster | GlusterFS |
| April 30, 2014 | Inktank | Ceph |
| July 31, 2017 | Permabit | Data deduplication and compression |
| November 28, 2018 | NooBaa | Cloud storage technology |
| January 2023 | IBM | Ceph Transition to IBM |

# Leading “ceph” related products within Red Hat portfolio

**Red Hat**
OpenStack
Platform

**Red Hat**
Ceph Storage

**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

**Red Hat**
Ceph Storage

**Ceph storage cluster**
*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

**Red Hat**
OpenShift
Data Foundation

**Ceph for OpenShift**
*Self-managing storage*

- Powered by Red Hat Ceph Storage
- Automated by Rook and completed with Multicloud object gateway (MCG)
- Advanced integration and ease of use
- Adds support for stateful workloads to OpenShift

# Red Hat Ceph Storage

# Ceph Community

# Ceph project and community overview

- Ceph is 100% open source
  - Mostly LGPL2.1/LGPL3
- Scalable, multi-protocol storage platform
- We collaborate via
  - GitHub: https://github.com/ceph/ceph
  - https://tracker.ceph.com/
  - E-mail: dev@ceph.io
  - #ceph-devel on irc.oftc.net
- We meet a lot over video chat
  - See schedule at http://ceph.io/contribute

42on.com

CERN

Deutsche Telekom

DigitalOcean

intel

ProphetStor

Red Hat

SK telecom

SUSE
We adapt. You succeed.

ZTE

# Ceph project and community overview

## Vibrant developer community

- 1000+ contributors
- 600K+ lines of code changed
- 17,000  code commits
- From multiple countries
  - USA, China, Germany, India

## Multi vendor collaboration

- 200+ organizations
  - 55% of code from Red Hat
- Variety of  vendors
  - Hardware & Software vendors and Service providers

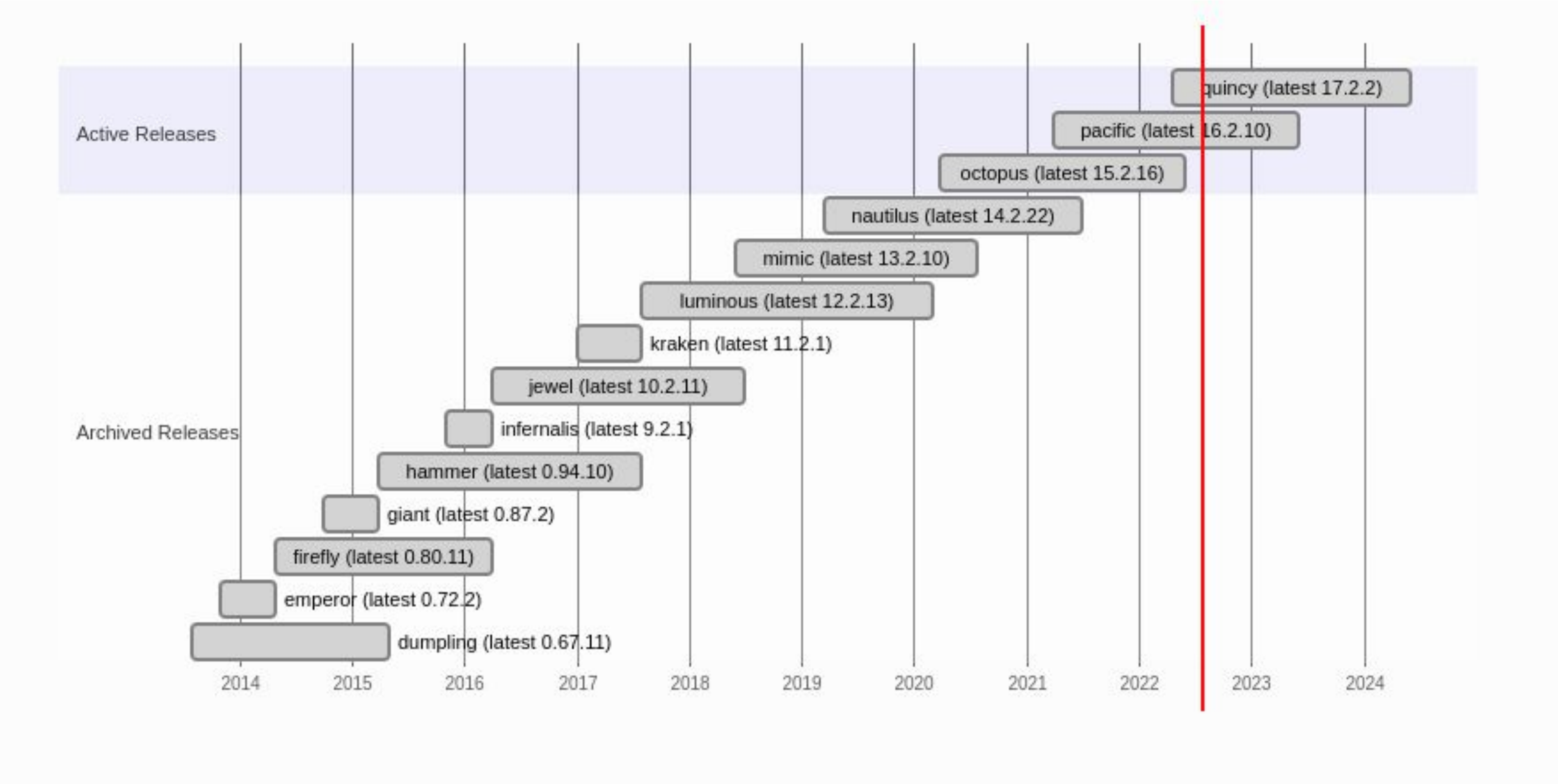
## Broad solution Ecosystem

- Vendor published solutions across diverse workloads
  - Intel
  - Samsung
  - Sandisk

## Proven Production deployments

- Multi petabyte production deployments of Ceph
  - CERN
  - NASA
  - China Mobile
  - Flipkart
  - Salesforce

# Ceph upstream community releases



# Ceph community usage from telemetry

## Community Ceph deployments

- Community Ceph survey:
  - Total of **880 PB** deployed
    - 322 respondents
  - 10 PB to 100PB: **4.92%**
    - This was the largest available cluster size option in the survey

- Community Ceph deployments:
  - CERN - 65+ PB
  - Digital Ocean - 40+ PB
  - Sanger Inst - 18 PB
  - China Mobile - 100s of PB

Upstream users, please, consider enabling it in your clusters via “`ceph telemetry on`”

# Red Hat Ceph deployment metrics

Work with Red Hat consulting team to deploy optimally sized and configured clusters

## Red Hat product deployments

★ Multi PB data center footprint

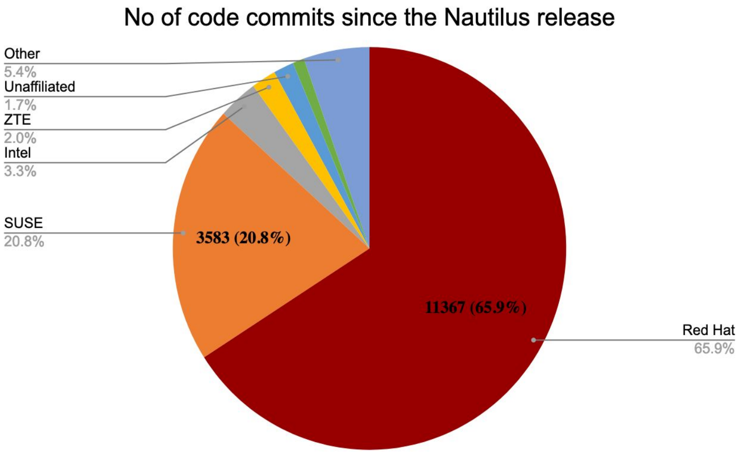
| | | |
|---|---|---|
| ○ | Retail | - 150 PB + |
| ○ | Financial | - 55 PB |
| ○ | Telco | - 35 PB |
| ○ | Govt | - 25 PB |

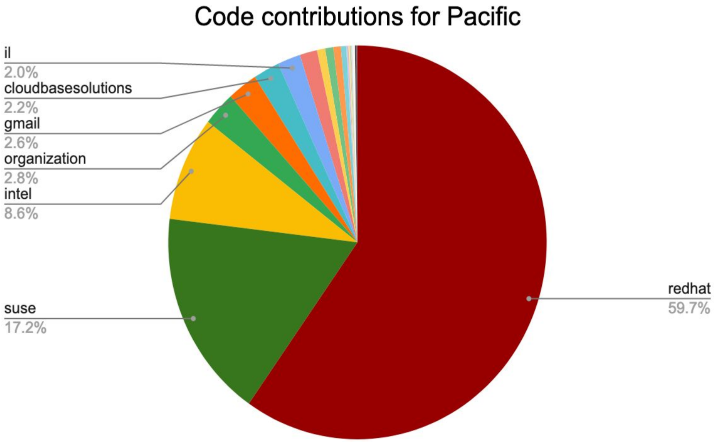
★ Large production clusters

| | | |
|---|---|---|
| ○ | Retail | - 15-20 PB |
| ○ | Financial | - 15-20 PB |
| ○ | Telco | - 10-15 PB |
| ○ | Govt | - 10-15 PB |

# Red Hat leadership in the Ceph community

Red Hat is the largest contributor to the Ceph codebase





# So you wanna learn ceph

- Read the paper - https://www.ssrc.ucsc.edu/papers/weil-osdi06.pdf
- Setup a cluster you can work with for an extended period of time
- bare-metal if you can (3 nodes minimum) ... virtual if you can’t
- Least friction approach CentOS Stream + upstream ceph
- If you want to do it with RHEL + RHCS talk to a sales person

# Red Hat Ceph Storage

## Red Hat Ceph Storage

- Software defined storage for on-premise cloud buildout
- Massively scalable to support tens of petabytes of data
- Delivers solid reliability and data durability
- Storage with industry-standard x86 servers
- Multi-site aware and disaster-recovery enabled

# Flexibility to meet the demands of tomorrow

### Delivers scalability

- Expand or shrink clusters as required
- Scale out within a cluster for capacity/speed

### Increases reliability

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

### Improves versatility

- A single cluster can support object, block, and file workloads
- Add or remove hardware while system is online —even if it’s under load
- Apply updates without interrupting service

# Ceph architecture baseline

**MONITOR PROCESS**

## Ceph uses monitors and object storage daemons

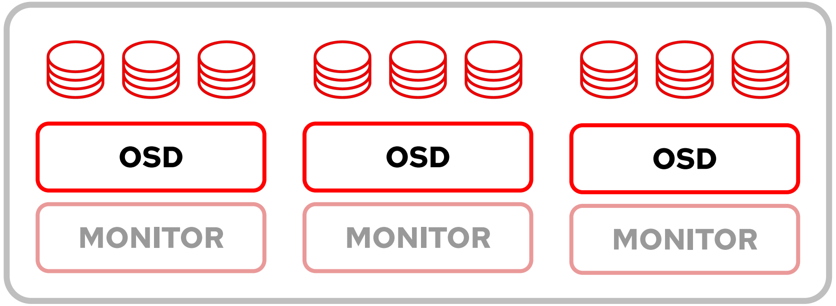
- Monitors maintain the Ceph cluster map
- Decisions are based on consensus: Paxos
- Monitors operate in a small and odd number
- Can be run as containerized processes

# Ceph architecture baseline

OSD PROCESS

## Ceph monitors and object storage daemons

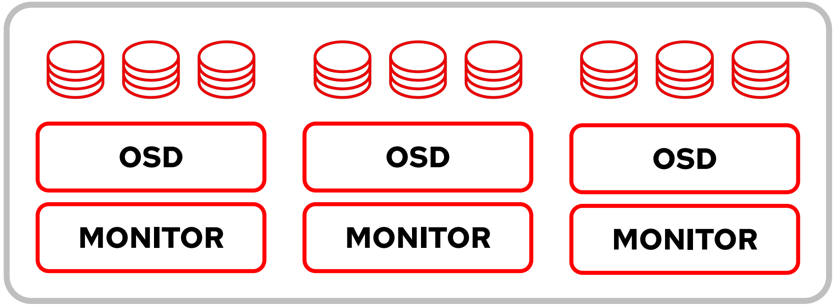
- OSD: Object storage daemon
- Provide direct data access
- Manage layout of data on media
- Peer and coordinate data distribution, integrity checking and recovery



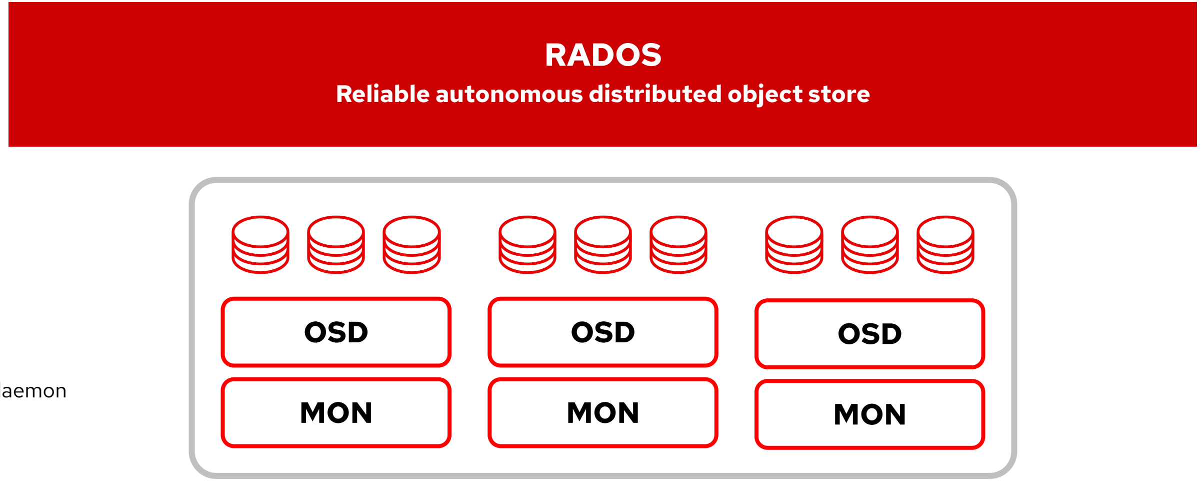
# Ceph architecture baseline

**A basic Ceph cluster setup is composed of monitors and object storage daemons (OSDs)**

- Easy command-line interface (CLI) and user interface (UI) (5.1) setup
- A minimal setup contains 3 nodes
- OSDs can scale to 10000s in a cluster
- Tune for performance, capacity or cost



# Ceph RADOS



**RADOS**

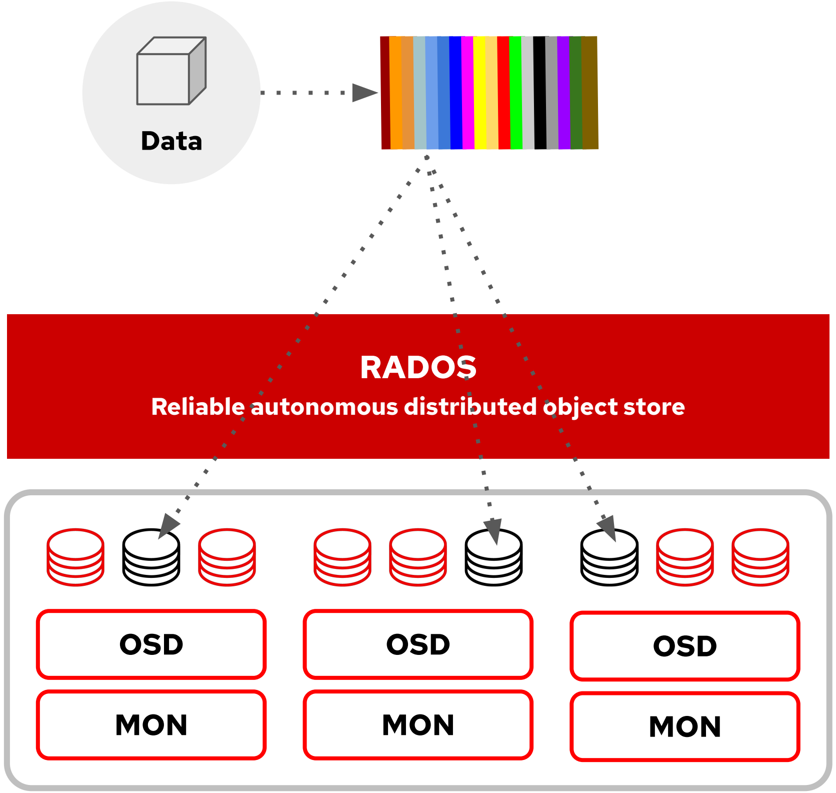
**Reliable autonomous distributed object store**

| | |
|---|---|
| MON | Monitor |
| OSD | Object storage daemon |

# Ceph CRUSH algorithm

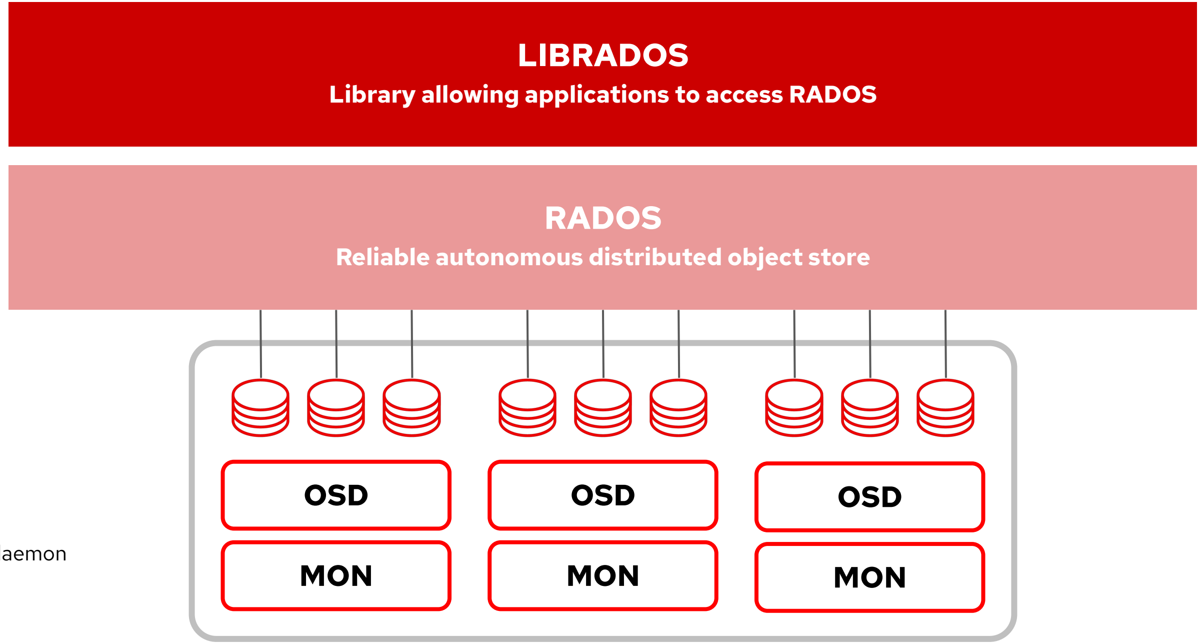
## Controlled Replication Under Scalable Hashing

- Pseudo-random placement algorithm
- Fast calculation, no lookup, no gateways
- Repeatable and deterministic
- Stable mapping
- Rule-based configuration
- Adjustable replication
- Weighting



| | |
|---|---|
| MON | Monitor |
| OSD | Object storage daemon |

# Ceph LIBRADOS



**LIBRADOS**

**Library allowing applications to access RADOS**

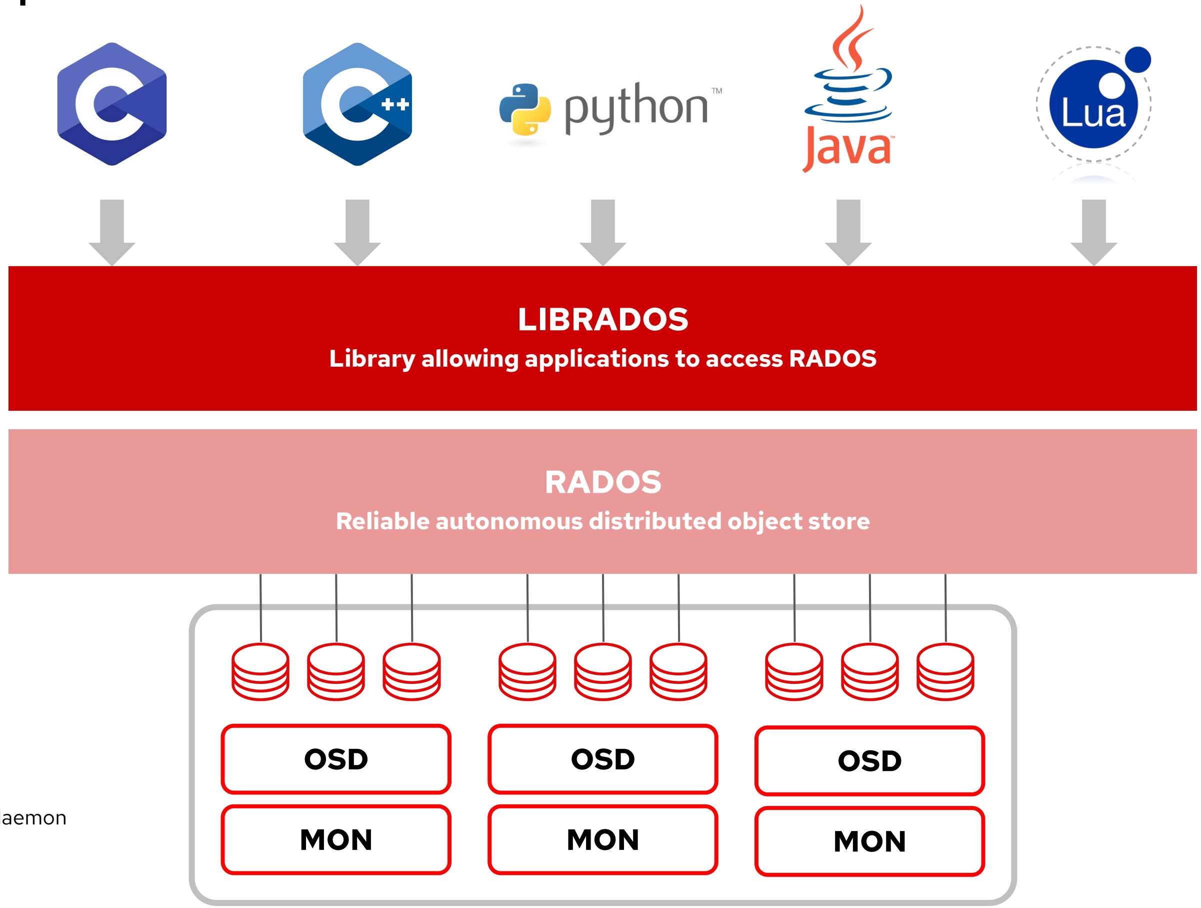
**RADOS**

**Reliable autonomous distributed object store**

**MON** Monitor

**OSD** Object storage daemon

# Accessing Ceph



**LIBRADOS**

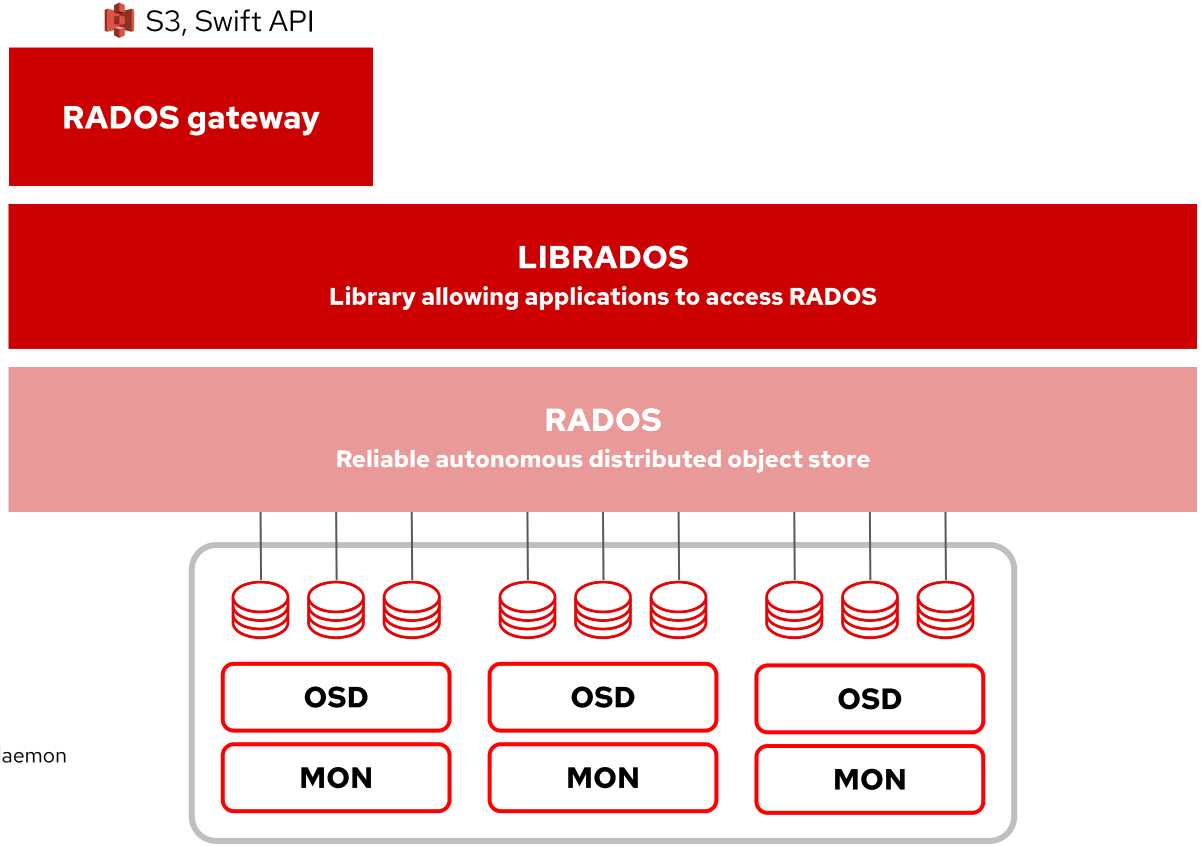
**Library allowing applications to access RADOS**

**RADOS**

**Reliable autonomous distributed object store**

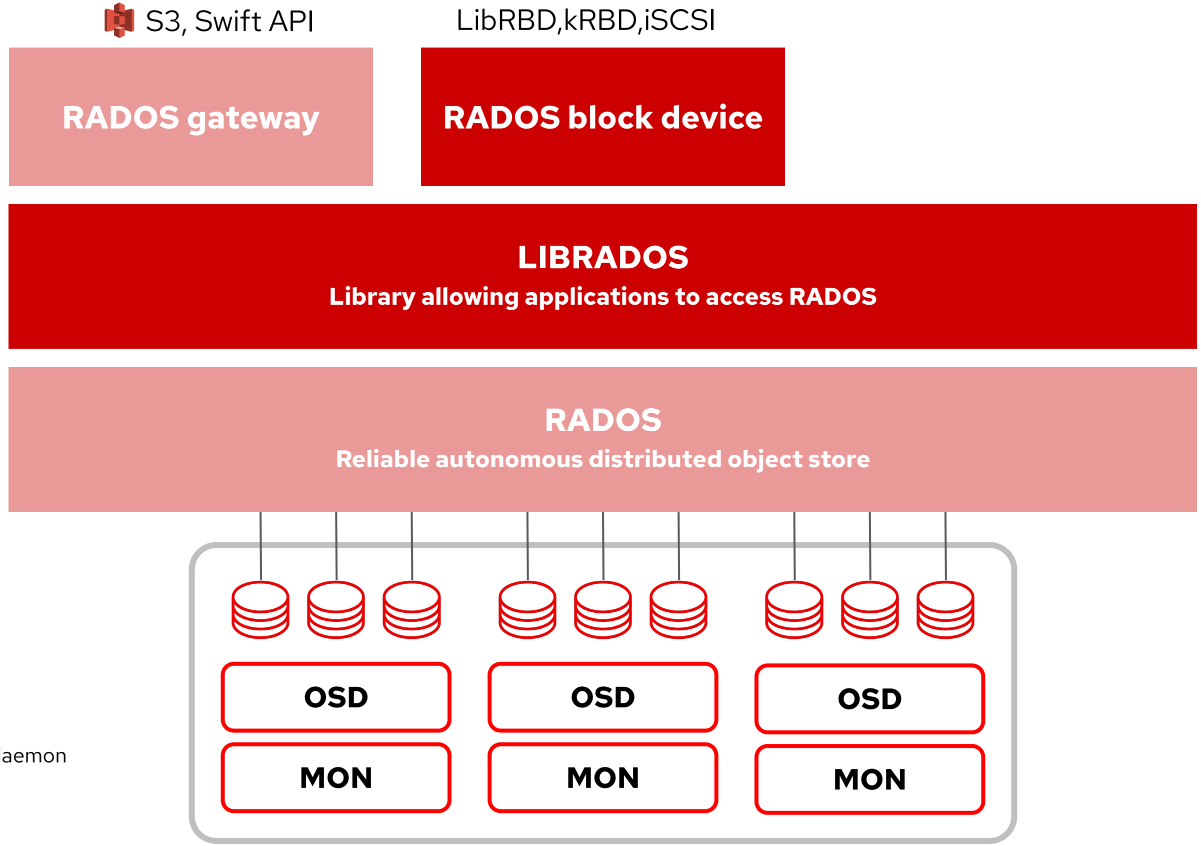
| | |
|---|---|
| MON | Monitor |
| OSD | Object storage daemon |

# Ceph architecture


S3, Swift API
RADOS gateway
LIBRADOS
Library allowing applications to access RADOS
RADOS
Reliable autonomous distributed object store
OSD
OSD
OSD
MON
MON
MON

| | |
|---|---|
| MON | Monitor |
| OSD | Object storage daemon |

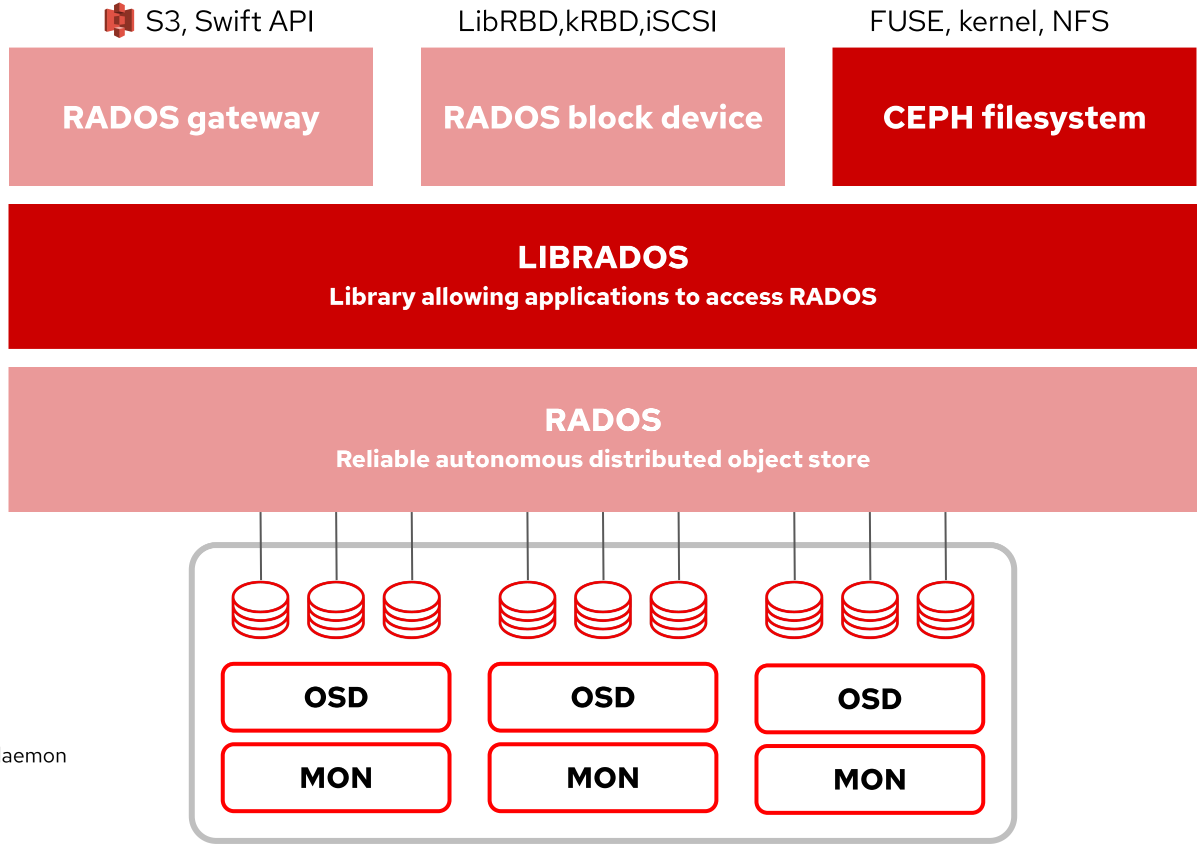
# Ceph architecture



MON Monitor

OSD Object storage daemon

# Ceph architecture



| | |
|---|---|
| MON | Monitor |
| OSD | Object storage daemon |

# What’s new?

Red Hat Ceph Storage 5

## FUNCTIONALITY

New integrated control plane
Stable management API
Network File System (NFS) support

## SECURITY

Write once, read many (WORM) (object lock)
Compliant with regulatory standards
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

# What’s new?

Red Hat Ceph Storage 5

- **FUNCTIONALITY**
- **SECURITY**
- **PERFORMANCE**
- **EFFICIENCY**

# FUNCTIONALITY

### New manageability features

Integrated control plane

Stable management application programming interface (API)

Object store daemon (OSD) replacement workflows

Object multi-site monitoring

### New Ceph filesystem capabilities

NFS access option

Erasure code option

Snapshot based geo replication

### New Rados block device capabilities

RADOS block device (RBD) snapshot based migration across clusters

# What’s new?

Red Hat Ceph Storage 5

FUNCTIONALITY

SECURITY

PERFORMANCE

EFFICIENCY

# 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 RADOS object gateway
Continued object store scalability improvements

### Better monitoring tools

Ceph file system ‘top’ joins the existing RADOS block device (RBD) top tool

# What’s new?

Red Hat Ceph Storage 5

FUNCTIONALITY

SECURITY

PERFORMANCE

EFFICIENCY

# What’s new?

Red Hat Ceph Storage 5

## SECURITY

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

**Federal Information Processing Standard (FIPS)**
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)

# What’s new?

Red Hat Ceph Storage 5

FUNCTIONALITY

SECURITY

PERFORMANCE

EFFICIENCY

# What's new?

Red Hat Ceph Storage 5

## EFFICIENCY

**Multi-site capabilities**

RADOS object gateway across sites including hybrid cloud connectivity options

**Resource consumption**

Improved internal space utilization for small files

**Improved reliability**

Erasure coding recovery with "K" shards

**Object offload to public cloud (5.1)** Future feature

Using bucket policies and AWS

# Summary of what's new RHCS 5

### Efficiency

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

### Security

- Write once, read many (WORM) object lock application programming interface (API)
- FIPS 140-2 cryptography
- Interoperate with key management interoperability protocol (KMIP) key managers
- Messenger v. 2.1 backplane encryption

### Performance

- Optimized Librados block device (LibRBD) data path: 80% faster
- Overhauled cache architecture
- 10+ billion objects in RADOS gateway (RGW)
- Ceph file system (CephFS) "Top" tool

### Manageability

- New integrated control plane–Cephadm
- Integrated monitoring and management dashboard
- OSD replacement workflow (CLI and UI)
- RGW multisite monitoring

### APIs and protocols

- Management API
- CephFS + network file system (NFS)
- CephFS geo-replication

# Workload-based configurations

| | **Edge configuration** | **Capacity optimized (big data and AI/ML)** | **Performance I/O optimized (analytics/database)** |
|---|---|---|---|
| **Options** | **Base (10 TB) or Plus (20 TB)** | **Base (30 TB) or Plus (60 TB)** | **Base (15 TB) or Plus (30 TB)** |
| **Workloads or Services** | Small footprint edge configurations | Big data workloads | Latency-sensitive workloads, such as transaction processing |
| **Red Hat OpenShift Data Foundation** | Attach to Red Hat OpenShift Container Platform cluster<br>Bare metal: RS00421 or Core pair: MCT4051 | Attach to Red Hat OpenShift Container Platform cluster<br>Bare metal: RS00421 or Core pair: MCT4051 | Attach to Red Hat OpenShift Container Platform cluster<br>Bare metal: RS00421 or Core pair: MCT4051 |
| **Platform** | 2U 1 node | 2U 1 node | 2U 1 node |
| **CPU** | **Base**: 1x Intel® Xeon® Gold 5218R processor (16 cores)<br>**Plus:** 2x Intel Xeon Gold 5218R processor (16 cores) | **Base:** 1x Intel® Xeon® Gold 6242R processor (20 cores)<br>**Plus:** 2x Intel Xeon Gold 6242R processor (20 cores) | **Base:** 2x Intel Xeon Gold 6242R processor (20 cores)<br>**Plus:** 2x Intel® Xeon® Gold 6248R processor (24 cores) |
| **Memory** | **Base**: 96 GB<br>**Plus:** 192 GB | **Base**: 96 GB<br>**Plus:** 192 GB | **Base**: 192 GB<br>**Plus:** 384 GB |
| **Data network** | **Base:** 2x Intel® Ethernet Network Adapter X710-T2L (10 GbE)<br>**Plus:** 2x Intel Ethernet Network Adapter X710-T2L (10 GbE) | **Base:** 2x Intel Ethernet Network Adapter XXV710-DA2 (25 GbE)<br>**Plus:** 2x Intel Ethernet Network Adapter XXV710-DA2 (25 GbE) | **Base:** 2x Intel Ethernet Network Adapter E810-CQDA2 (50 GbE)<br>**Plus:** 2x Intel Ethernet Network Adapter E810-CQDA2 (100 GbE) |
| **Management network** | 1x Intel® Ethernet Connection X710-DA2 (10 GbE) | 1x Intel Ethernet Connection X710-DA2 (10 GbE) | 1x Intel Ethernet Connection X710-DA2 (10 GbE) |
| **Storage cache** | **Base:** None<br>**Plus:** 1x Intel® Optane™ SSD DC P4800X (375 GB) | **Base:** 1x Intel Optane SSD DC P4800X (750 GB)<br>**Plus:** 2x Intel Optane SSD DC P4800X (750 GB) | **Base:** 2x Intel Optane SSD DC P4800X (750 GB)<br>**Plus:** 2x Intel Optane SSD DC P4800X (1.5 TB) |
| **Storage media** | **Base:** 6x Intel® SSD DC-S4510 (1.92 TB, 2.5” SATA, TLC)<br>**Plus:** 6x Intel SSD DC-S4510 (3.84 TB, 2.5” SATA, TLC) | **Base:** 8x Intel SSD DC-S4510 (3.84 TB, 2.5” SATA, TLC)<br>**Plus:** 16x Intel SSD DC-S4510 (3.84 TB, 2.5” SATA, TLC) **or** 8x Intel® SSD DC-S4510 (7.68 TB, 2.5” SATA, TLC) | **Base:** 8x Intel® SSD DC-P4610 (1.92 TB, 2.5” U.2 NVMe, TLC)<br>**Plus:** 8x Intel SSD DC-P4610 (3.84 TB, 2.5” U.2 NVMe, TLC) |

# Subscription (SKU) options

Red Hat
Ceph Storage

**Subscription model**

Pricing is based on capacity

Capacity limit is raw physical capacity of disks

**Red Hat's Certified Cloud and Service Partner program (CCSP)**

Red Hat Ceph Storage is also available through the embedded and CCSP programs

# 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

# OpenShift Data Foundation

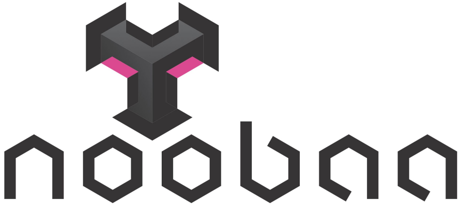
aka ODF

pka OCS

apka CNS

:) ..... :(

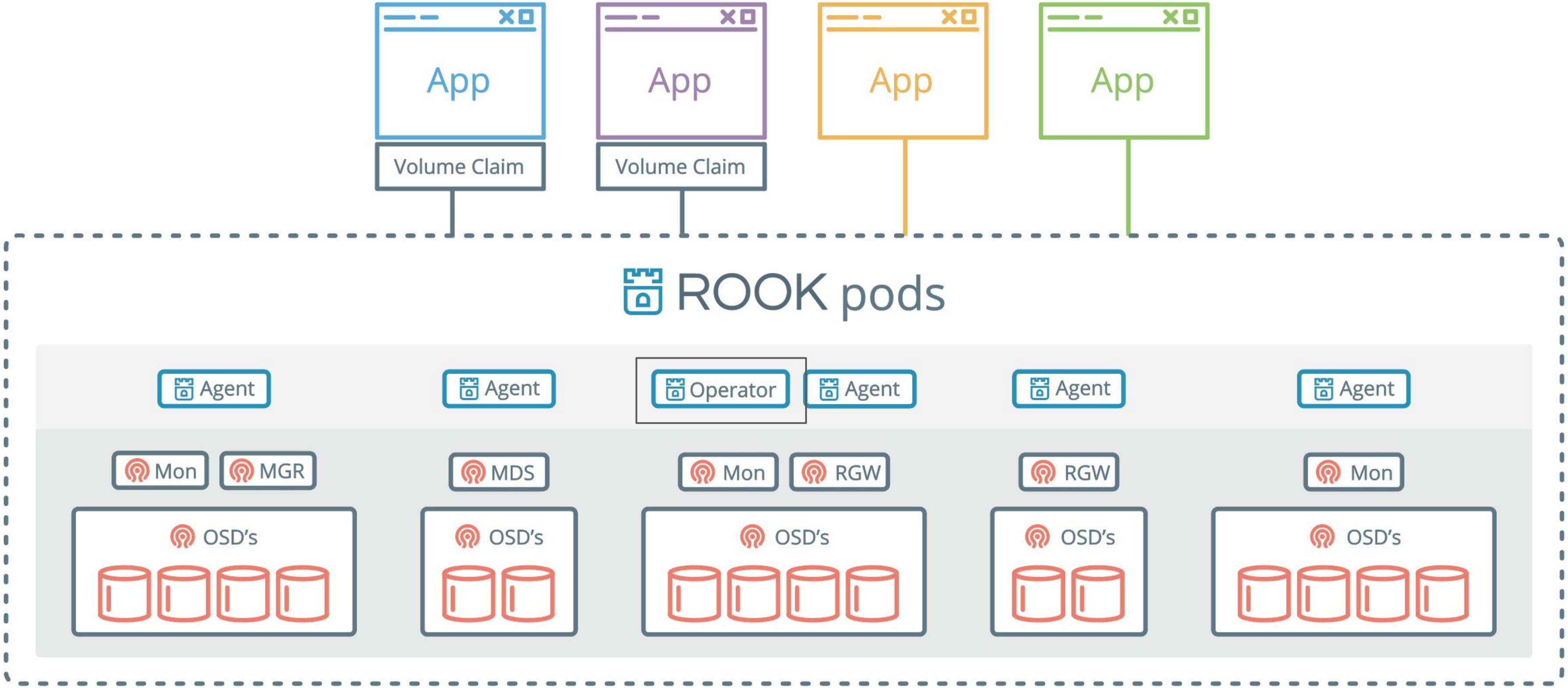
# open source upstream communities



# The benefits of Rook

- High availability with ability to handle file, block and object storage
- Increased resiliency
- Scrubs for, and repairs, inconsistent objects ensuring data is protected and coherent
- Can be deployed anywhere ensuring a consistent storage platform across the hybrid cloud

# Rook Architecture



# Multi Cloud Object Gateway technology by

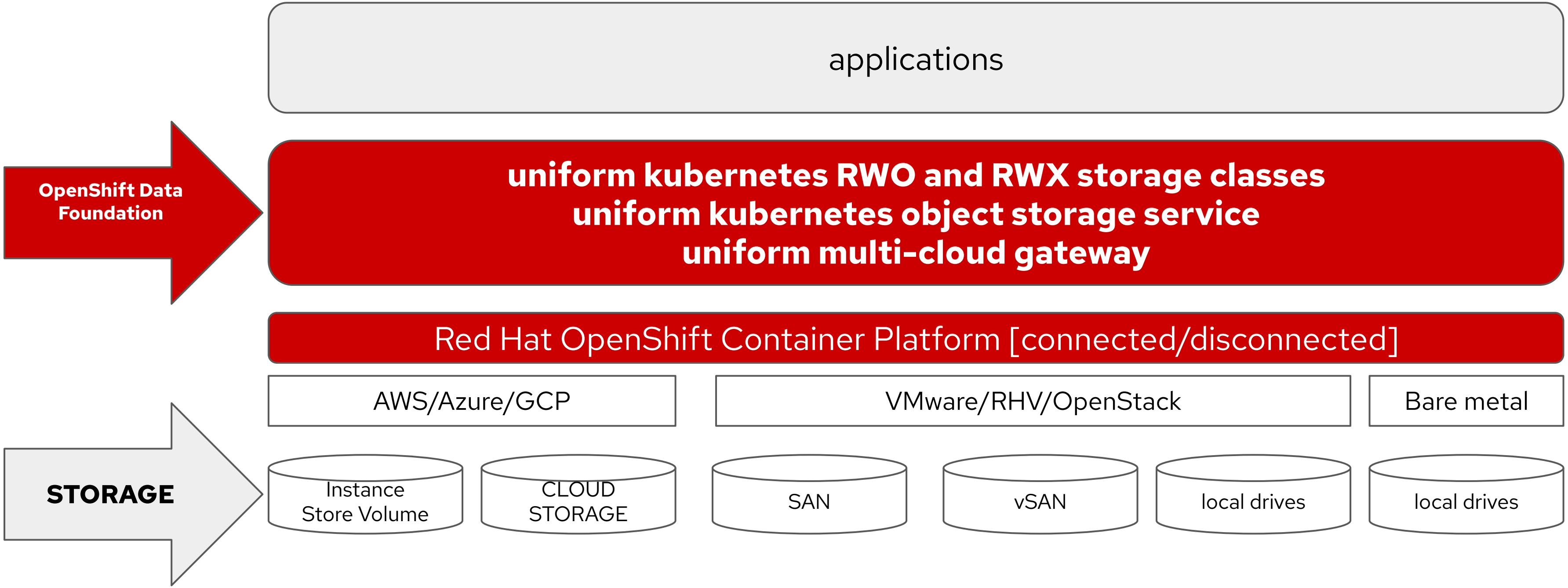
noobaa

- ODF Multi Cloud Object Gateway powered by NooBaa
    - Noobaa provides a consistent S3 endpoint across different Multi Cloud Infrastructures : AWS, AZURE, GCP, BareMetal, VMware and OpenStack
- ODF MCG Functionality
    - Read/Write access across multiple clouds
- Virtualizes and abstracts any kind of existent storage resources
    - Shared, dedicated, Physical or Virtual, Private or Public
- Full control over data placement
    - Place data based on Security, Strategy and Cost Considerations
    - All within granularity of application

**Red Hat OpenShift Data Foundation**

consumes **storage** to provide **higher-level services.**

## OpenShift Data Foundation



applications

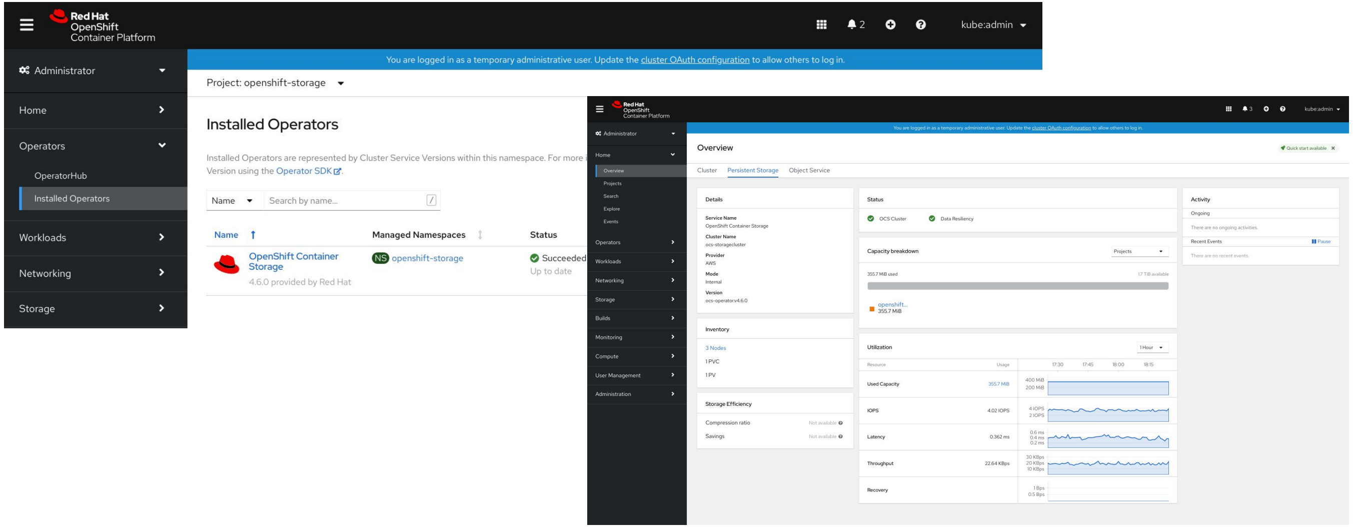
OpenShift Data Foundation

**uniform kubernetes RWO and RWX storage classes**
**uniform kubernetes object storage service**
**uniform multi-cloud gateway**

Red Hat OpenShift Container Platform [connected/disconnected]

# Focus on ease of use

- Simplified installation from the Operator Hub within OpenShift Console
- Minimize maintenance
- Integrated dashboard and configuration into OpenShift Console



You are logged in as a temporary administrative user. Update the cluster OAuth configuration to allow others to log in.

Project: openshift-storage

## Installed Operators

Installed Operators are represented by Cluster Service Versions within this namespace. For more i Version using the Operator SDK.

Name · Search by name...

| Name | Managed Namespaces | Status |
|---|---|---|
| OpenShift Container Storage<br>4.6.0 provided by Red Hat | NS openshift-storage | Succeeded<br>Up to date |

## Overview

Details

Service Name
OpenShift Container Storage
Cluster Name
ocs-storagecluster
Provider
AWS
Mode
Internal
Version
ocs-operator.v4.6.0

Status

✓ OCS Cluster ✓ Data Resiliency

Activity

Ongoing
There are no ongoing activities.
Recent Events
There are no recent events.

Capacity breakdown

355.7 MiB used 1.7 TiB available

openshift... 355.7 MiB

Inventory

3 Nodes
1 PVC
1 PV

Storage Efficiency

Compression ratio Not available
Savings Not available

Utilization

| Resource | Usage |
|---|---|
| Used Capacity | 355.7 MiB |
| IOPS | 4.02 IOPS |
| Latency | 0.362 ms |
| Throughput | 22.64 KBps |
| Recovery | |

# Persistent Volume

## Block

- Primary for DB and Transactional workloads
- Low latency
- Messaging

**Provided by Rook-Ceph**

## Shared File System

- POSIX-compliant shared file system
- Interface for legacy workloads
- CI/CD Pipelines
- AI/ML Data Aggregation

**Provided by Rook-Ceph**

# Object Service

- Media, AI/ML training data, Archiving, Backup, Health Records
- Great Bandwidth performance
- Object API (S3/Blob)

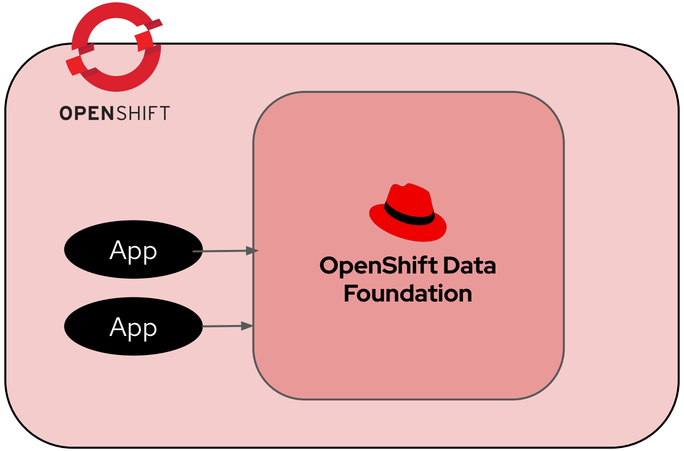
**Provided by Multicloud Object Gateway**

# Platforms

- Managed service on IBM ROKS
- Early Field Trial on ROSA and OpenShift Dedicated

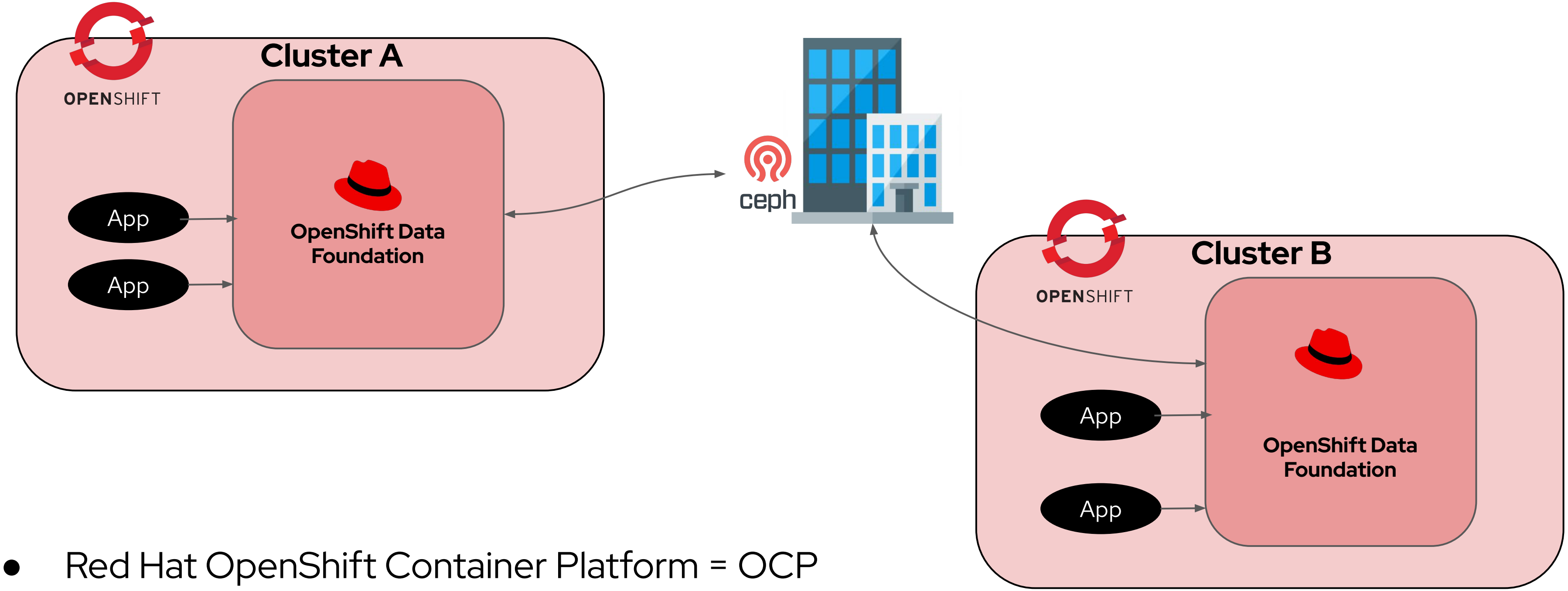
| Out of the box support | |
|---|---|
| Block, File, Object | |
| **Platforms** | |
| AWS/Azure | Google Cloud (Tech Preview) |
| ARO - Self managed OCS | **IBM ROKS & Satellite - Managed OCS (GA)** |
| RHV | OSP (Tech Preview) |
| Bare metal/IBM Z/Power | VMWare Thin/Thick IPI/UPI |
| **Deployment modes** | |
| Disconnected environment and Proxied environments | |

# ODF internal mode



- Red Hat OpenShift Container Platform = OCP
- Red Hat ODF - Internal Mode - One ODF install per single OCP
- Ceph components run internally to OpenShift Cluster as containers, using ODF operator to have an opinionated deployment

# ODF external mode



- Red Hat OpenShift Container Platform = OCP
- Red Hat ODF - External Mode - One ODF instance per single OCP)
- Red Hat Ceph Storage Cluster aka RHCS cluster
- Data actually stored in RHCS cluster

# Supported Protocols with External Mode

Which storage protocols are supported ?

Similar to OpenShift Data Foundation in internal mode equivalent

- File storage
- Block storage
- Object storage

All Storage Modes
&
All Access Modes

- RWO
- ROX
- RWX

# OpenShift Data Foundation - Essentials edition

Red Hat
OpenShift
Data Foundation

**Red Hat OpenShift Data Foundation Essentials**
Contains all basic elements that applications need to address data needs

**Basic storage classes**
Kubernetes RWO, Kubernetes RWX and
S3-compatible Object storage

**Provides** basic OpenShift **cluster level encryption**

**Batteries are included**
Red Hat OpenShift Data Foundation Essentials edition is included with Red Hat OpenShift Platform Plus—**at no additional cost**

# OpenShift Data Foundation - Advanced edition

Red Hat
OpenShift
Data Foundation

**Red Hat OpenShift Data Foundation Advanced**
Extends the essentials edition with additional capabilities

- **Enhanced level of encryption** at persistent volume level
- **Shared mode**—Share data across multiple Openshift clusters
- **Mixed use**—Workloads outside OpenShift accessing the data
- **Regional and Metropolitan disaster recovery capabilities** with Red Hat Advanced Cluster Management for Kubernetes and Red Hat OpenShift Data Foundation Advanced

# Simple ODF demo

# With Q and A

# Further Q and A

# Thank you

Red Hat is the world's leading provider of enterprise open source software solutions. Award-winning support, training, and consulting services make Red Hat a trusted adviser to the Fortune 500.

- linkedin.com/company/red-hat
- youtube.com/user/RedHatVideos
- facebook.com/redhatinc
- twitter.com/RedHat

Build out your New York community and learn about
future RHUGs, first, on Meetup.com.

Red Hat