RED HAT®
STORAGE

redhat.

TECHNOLOGY DETAIL

# RED HAT CEPH STORAGE HARDWARE SELECTION GUIDE

Choosing hardware for a workload-optimized Ceph Cluster

Deliver object, file, and block storage in one self-managing, self-healing platform with no single point of failure.

Gain multi-petabyte software-defined enterprise storage across a range of industry-standard hardware.

Support IOPS-, throughput-, or cost/capacity-optimized workloads.

facebook.com/redhatinc
@redhatnews
linkedin.com/company/red-hat

redhat.com

## EXECUTIVE SUMMARY

Many hardware vendors now offer both Ceph-optimized servers and rack-level solutions designed for distinct workload profiles. To simplify the hardware selection process and reduce risk for organizations, Red Hat has worked with multiple storage server vendors to test and evaluate specific cluster options for different cluster sizes and workload profiles. Red Hat's exacting methodology combines performance testing with proven guidance for a broad range of cluster capabilities and sizes. With appropriate storage servers and rack-level solutions, Red Hat® Ceph Storage can provide storage pools serving variety of workloads – from throughput-sensitive and cost/capacity-focused workloads to emerging IOPS-intensive workloads.

## TABLE OF CONTENTS

## INTRODUCTION

Red Hat Ceph Storage significantly lowers the cost of storing enterprise data and helps organizations manage exponential data growth. The software is a robust and modern petabyte-scale storage platform for public or private cloud deployments. Red Hat Ceph Storage offers mature interfaces for enterprise block and object storage, making it an optimal solution for active archive, rich media, and cloud infrastructure workloads characterized by tenant-agnostic OpenStack® environments[1]. Delivered as a unified, software-defined, scale-out storage platform, Red Hat Ceph Storage lets businesses focus on improving application innovation and availability by offering capabilities such as:

- Scaling to hundreds of petabytes[2].
- No single point of failure in the cluster.
- Lower capital expenses (CapEx) by running on commodity server hardware.
- Lower operational expenses (OpEx) with self-managing and self-healing properties.

Red Hat Ceph Storage can run on myriad industry-standard hardware configurations to satisfy diverse needs. To simplify and accelerate the cluster design process, Red Hat conducts extensive performance and suitability testing with participating hardware vendors. This testing allows evaluation of selected hardware under load and generates essential performance and sizing data for diverse workloads – ultimately simplifying Ceph storage cluster hardware selection.

As discussed in this guide, multiple hardware vendors now provide server and rack-level solutions optimized for Red Hat Ceph Storage deployments with IOPS-, throughput-, and cost/capacity-optimized solutions as available options. For more information on configuring a Red Hat Ceph Storage cluster, see the Red Hat Ceph Storage hardware configuration guide. Full performance and sizing guides for several vendors are also available, providing complete and detailed information on the systems tested and results achieved.

## WORKLOAD-OPTIMIZED CEPH PERFORMANCE DOMAINS

One of the key benefits of Ceph storage is the ability to support different types of workloads within the same cluster using Ceph performance domains. Dramatically different hardware configurations can be associated with each performance domain. Storage pools can then be deployed on the appropriate performance domain, providing applications with storage tailored to specific performance and cost profiles. Selecting appropriately sized and optimized servers for these performance domains is an essential aspect of designing a Red Hat Ceph Storage cluster.

Table 1 provides the criteria Red Hat uses to identify optimal Red Hat Ceph Storage cluster configurations on storage servers. These categories are provided as general guidelines for hardware purchases and configuration decisions, and can be adjusted to satisfy unique workload blends. Actual hardware configurations chosen will vary depending on specific workload mix and vendor capabilities.

1 *Ceph is and has been the leading storage for OpenStack according to several semi-annual OpenStack user surveys.*
2 *yahooeng.tumblr.com/post/116391291701/yahoo-cloud-object-store-object-storage-at*

TABLE 1. CEPH CLUSTER OPTIMIZATION CRITERIA

| CLUSTER OPTIMIZATION CRITERIA | PROPERTIES | EXAMPLE USES |
|---|---|---|
| **IOPS-OPTIMIZED** | • Lowest cost per IOPS<br>• Highest IOPS per GB<br>• 99th percentile latency consistency | • Typically block storage<br>• 3x replication for hard disk drives (HDDs) or 2x replication for solid-state drives (SSDs)<br>• MySQL on OpenStack clouds |
| **THROUGHPUT-OPTIMIZED** | • Lowest cost per MBps (throughput)<br>• Highest MBps per TB<br>• Highest MBps per BTU<br>• Highest MBps per Watt<br>• 97th percentile latency consistency | • Block or object storage<br>• 3x replication<br>• Active performance storage for video, audio, and images<br>• Streaming media |
| **COST/CAPACITY-OPTIMIZED** | • Lowest cost per TB<br>• Lowest BTU per TB<br>• Lowest Watts required per TB | • Typically object storage<br>• Erasure coding common for maximizing usable capacity<br>• Object archive<br>• Video, audio, and image object repositories |

To the Ceph client interface that reads and writes data, a Ceph cluster appears as a simple pool where data is stored. However, the storage cluster performs many complex operations in a manner that is completely transparent to the client interface. Ceph clients and Ceph object storage daemons (Ceph OSDs, or simply OSDs) both use the controlled replication under scalable hashing (CRUSH) algorithm for storage and retrieval of objects. OSDs run on OSD hosts — the storage servers within the cluster.

A CRUSH map describes a topography of cluster resources, and the map exists both on client nodes as well as Ceph Monitor (MON) nodes within the cluster. Ceph clients and Ceph OSDs both use the CRUSH map and the CRUSH algorithm. Ceph clients communicate directly with OSDs, eliminating a centralized object lookup and a potential performance bottleneck. With awareness of the CRUSH map and communication with their peers, OSDs can handle replication, backfilling, and recovery—allowing for dynamic failure recovery.

The CRUSH map is also used to implement both failure domains and performance domains. Performance domains are simply a hierarchy that takes the performance profile of the underlying hardware into consideration. The CRUSH map describes how Ceph stores data, and it is implemented as a simple hierarchy (acyclic graph) and a ruleset. The CRUSH map can support multiple hierarchies to separate one type of hardware performance profile from another. For example:

- Hard disk drives (HDDs) are typically appropriate for cost/capacity-focused workloads.
- HDDs with Ceph write journals on solid state drives (SSDs) are often used for throughput-sensitive workloads.
- SSDs are used for IOPS-intensive workloads such as MySQL and MariaDB.

All of these performance domains can coexist in a Ceph cluster, supported by differently-configured servers. Figure 1 illustrates how different HDD-based nodes and SSD-based nodes can serve as OSD hosts for differently optimized performance domains – spread across multiple racks in the data center. Actual server distribution of across racks would be done with specific failure domains in mind as well.

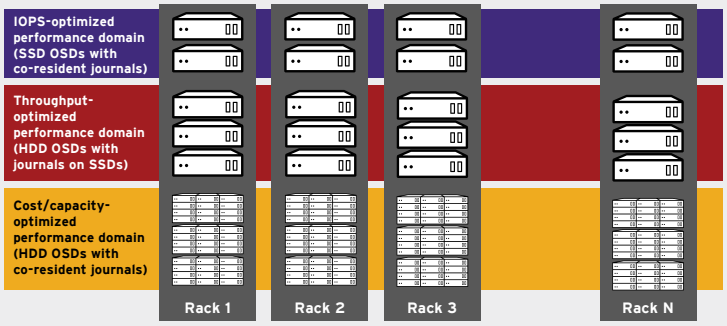



*Figure 1. Multiple performance domains can coexist in a single Ceph cluster, supported by dramatically different server configurations.*

## SERVER AND RACK-LEVEL CEPH SOLUTIONS

Hardware vendors have responded to the enthusiasm around Ceph by providing both optimized server-level and rack-level solution SKUs. Validated through joint testing with Red Hat, these solutions offer predictable price/performance ratios for Ceph deployments, with a convenient modular approach to expand Ceph storage for specific workloads. As shown in Figure 2, typical rack-level solutions include:

- **Network switching**. Redundant network switching interconnects the cluster and provides access to clients.
- **Ceph MON nodes**. The Ceph monitor is a datastore for the health of the entire cluster, and contains the cluster log. A minimum of three monitor nodes are strongly recommended for a cluster quorum in production.
- **Ceph OSD hosts**. Ceph OSD hosts house the storage capacity for the cluster, with one or more OSDs running per individual storage device. OSD hosts are selected and configured differently depending on both workload optimization and the data devices installed: HDDs, SSDs, or NVMe SSDs.
- **Red Hat Ceph Storage**. Many vendors provide a capacity-based subscription for Red Hat Ceph Storage bundled with both server and rack-level solution SKUs.

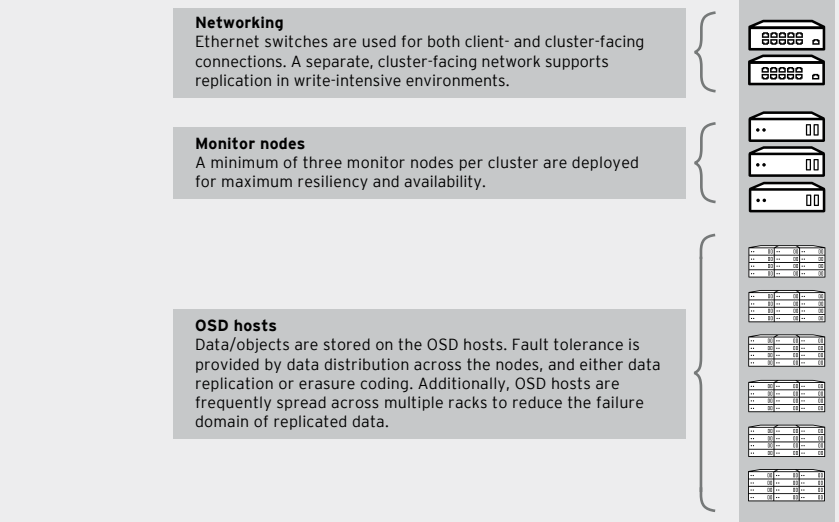



*Figure 2. Rack-level Ceph options typically contain networking, MON nodes, and OSD hosts.*

## IOPS-OPTIMIZED SOLUTIONS

With the growing use of flash storage, IOPS-intensive workloads are increasingly being hosted on Ceph clusters to let organizations emulate high-performance public cloud solutions with private cloud storage. These workloads commonly involve structured data from MySQL-, MariaDB-, or PostgreSQL-based applications. OSDs are typically hosted on NVMe SSDs with co-located Ceph write journals. Typical servers are listed in Table 2, and include the following elements:

- **CPU.** 10 cores per NVMe SSD, assuming a 2 GHz CPU.
- **RAM.** 16GB baseline, plus 2GB per OSD.
- **Networking.** 10 Gigabit Ethernet (GbE) per 12 OSDs (each for client- and cluster-facing networks).
- **OSD media.** High-performance, high-endurance enterprise NVMe SSDs.
- **OSDs.** Four per NVMe SSD.
- **Journal media.** High-performance, high-endurance enterprise NVMe SSD, co-located with OSDs.
- **Controller.** Native PCIe bus.

### TABLE 2. SOLUTIONS SKUs FOR IOPS-OPTIMIZED CEPH WORKLOADS, BY CLUSTER SIZE

| VENDOR | SMALL (250TB+) | MEDIUM (1PB+) | LARGE (2PB+) |
|---|---|---|---|
| **INDIVIDUAL OSD SERVERS** | | | |
| **SUPERMICRO** | SYS-5038MR-OSD006P | N/A | N/A |

### THROUGHPUT-OPTIMIZED SOLUTIONS

Throughput-optimized Ceph solutions are usually centered around semi-structured or unstructured data. Large-block sequential I/O is typical. Storage media on OSD hosts is commonly HDDs with write journals on SSD-based volumes. Typical server elements include:

- **CPU.** 0.5 cores per HDD, assuming a 2 GHz CPU.
- **RAM.** 16GB baseline, plus 2GB per OSD.
- **Networking.** 10 GbE per 12 OSDs (each for client- and cluster-facing networks).
- **OSD media.** 7,200 RPM enterprise HDDs.
- **OSDs.** One per HDD.
- **Journal media.** High-endurance, high-performance enterprise serial-attached SCSI (SAS) or NVMe SSDs.
- **OSD-to-journal ratio.** 4-5:1 for an SSD journal, or 12-18:1 for an NVMe journal.
- **Host bus adapter (HBA).** just a bunch of disks (JBOD).

Several vendors provide pre-configured server and rack-level solutions for throughput-optimized Ceph workloads. Red Hat has conducted extensive testing and evaluation of servers from Supermicro and Quanta Cloud Technologies (QCT). Table 3 lists pre-configured rack and individual server SKUs for small, medium, and large Ceph clusters.

**TABLE 3. PRE-CONFIGURED RACK- AND SERVER-LEVEL SKUs FOR THROUGHPUT-OPTIMIZED CEPH WORKLOADS, BY CLUSTER SIZE**

| VENDOR | SMALL (250TB+) | MEDIUM (1PB+) | LARGE (2PB+) |
|---|---|---|---|
| **RACK-LEVEL SKUs: CEPH OSDS, MONS, AND TOP OF RACK (TOR) SWITCHES** | | | |
| **SUPERMICRO** | • SRS-42E112-Ceph-03 | • SRS-42E136-Ceph-03 | • SRS-42E136-Ceph-03 |
| **INDIVIDUAL OSD SERVERS** | | | |
| **SUPERMICRO** | • SSG-6028R-OSD072P | • SSG-6048-OSD216P | • SSG-6048-OSD216P |
| **QCT** | • QxStor RCT-200 | • QxStor RCT-400 | • QxStor RCT-400 |

In addition, several other vendors offer base servers that can be configured to fit the general requirements of throughput-optimized OSD servers (Table 4).

**TABLE 4. ADDITIONAL SERVERS THAT CAN BE CONFIGURED FOR THROUGHPUT-OPTIMIZED CEPH OSD WORKLOADS, BY CLUSTER SIZE**

| VENDOR | SMALL (250TB+) | MEDIUM (1PB+) | LARGE (2PB+) |
|---|---|---|---|
| **DELL** | • PowerEdge R730XD | • DSS 7000, twin node | • DSS 7000, twin node |
| **CISCO** | • UCS C240 M4 | • UCS C3260 | • UCS C3260 |
| **LENOVO** | • System x3650 M5 | • System x3650 M5 | N/A |

## COST/CAPACITY-OPTIMIZED SOLUTIONS

Cost/capacity-optimized solutions typically focus on higher capacity, or longer archival scenarios. Data can be either semi-structured or unstructured. Workloads include media archives, big data analytics archives, and machine image backups. Large-block sequential I/O is typical. For greater cost effectiveness, OSDs are usually hosted on HDDs with Ceph write journals co-located on the HDDs. Solutions typically include the following elements:

- **CPU.** 0.5 cores per HDD, assuming a 2 GHz CPU.
- **RAM.** 16GB baseline, plus 2GB per OSD.
- **Networking.** 10 GbE per 12 OSDs (each for client- and cluster-facing networks).
- **OSD media.** 7,200 RPM enterprise HDDs.
- **OSDs.** One per HDD.
- **Journal media.** Co-located on the HDD.
- **HBA.** JBOD.

Supermicro and QCT provide pre-configured server and rack-level solution SKUs for cost/capacity-focused Ceph workloads. Table 5 lists pre-configured SKUs that are appropriate for small, medium, and large Ceph clusters.

### TABLE 5. PRE-CONFIGURED RACK AND SERVER LEVEL SKUs FOR COST/CAPACITY-OPTIMIZED CEPH WORKLOADS, BY CLUSTER SIZE

| VENDOR | SMALL (250TB+) | MEDIUM (1PB+) | LARGE (2PB+) |
|---|---|---|---|
| **RACK-LEVEL SKUs: CEPH OSDS, MONS, AND TOR SWITCHES** | | | |
| **SUPERMICRO** | N/A | • SRS-42E136-Ceph-03 | • SRS-42E172-Ceph-03 |
| **INDIVIDUAL OSD SERVERS** | | | |
| **SUPERMICRO** | N/A | • SSG-6048R-OSD216P | • SSD-6048R-OSD360P |
| **QCT** | N/A | • QxStor RCC-400 | • QxStor RCC-400 |

In addition, several other vendors offer servers that fit the general requirements of cost/capacity-optimized OSD servers (Table 6).

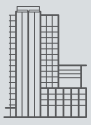
### TABLE 6. ADDITIONAL SERVERS THAT CAN BE CONFIGURED FOR COST/CAPACITY-OPTIMIZED CEPH WORKLOADS, BY CLUSTER SIZE

| VENDOR | SMALL (250TB+) | MEDIUM (1PB+) | LARGE (2PB+) |
|---|---|---|---|
| **DELL** | N/A | • DSS 7000, twin node | • DSS 7000, twin node |
| **CISCO** | N/A | • UCS C3260 | • UCS C3160 |
| **LENOVO** | N/A | • System x3650 M5 | NA |

## CONCLUSION

Software-defined storage presents many advantages to organizations seeking scale-out solutions to meet demanding applications and escalating storage needs. With a proven methodology and extensive testing performed with multiple vendors, Red Hat simplifies the process of selecting hardware to meet the demands of any environment. Importantly, the guidelines and example systems listed in this document are not a substitute for quantifying the impact of production workloads on sample systems.

For specific information on configuring servers for running Red Hat Ceph Storage, refer to the methodology and best practices documented in the Red Hat Ceph Storage hardware configuration guide. Detailed information, including Red Hat Ceph Storage test results, can be found in the performance and sizing guides for popular hardware vendors.



### ABOUT RED HAT

Red Hat is the world's leading provider of open source solutions, using a community-powered approach to provide reliable and high-performing cloud, virtualization, storage, Linux, and middleware technologies. Red Hat also offers award-winning support, training, and consulting services. Red Hat is an S&P company with more than 80 offices spanning the globe, empowering its customers' businesses.

facebook.com/redhatinc
@redhatnews
linkedin.com/company/red-hat

redhat.com
# INC0232826_0716

NORTH AMERICA
1 888 REDHAT1

EUROPE, MIDDLE EAST, AND AFRICA
00800 7334 2835
europe@redhat.com

ASIA PACIFIC
+65 6490 4200
apac@redhat.com

LATIN AMERICA
+54 11 4329 7300
info-latam@redhat.com