Red Hat Ceph Storage 2
Red Hat Ceph Storage Hardware Guide

Hardware recommendations for Red Hat Ceph Storage

Red Hat Ceph Storage Documentation Team

Red Hat Ceph Storage 2 Red Hat Ceph Storage Hardware Guide

Hardware recommendations for Red Hat Ceph Storage

Legal Notice

Copyright © 2016 Red Hat, Inc.

The text of and illustrations in this document are licensed by Red Hat under a Creative Commons Attribution–Share Alike 3.0 Unported license ("CC-BY-SA"). An explanation of CC-BY-SA is available at
http://creativecommons.org/licenses/by-sa/3.0/
. In accordance with CC-BY-SA, if you distribute this document or an adaptation of it, you must provide the URL for the original version.

Red Hat, as the licensor of this document, waives the right to enforce, and agrees not to assert, Section 4d of CC-BY-SA to the fullest extent permitted by applicable law.

Red Hat, Red Hat Enterprise Linux, the Shadowman logo, JBoss, OpenShift, Fedora, the Infinity logo, and RHCE are trademarks of Red Hat, Inc., registered in the United States and other countries.

Linux ® is the registered trademark of Linus Torvalds in the United States and other countries.

Java ® is a registered trademark of Oracle and/or its affiliates.

XFS ® is a trademark of Silicon Graphics International Corp. or its subsidiaries in the United States and/or other countries.

MySQL ® is a registered trademark of MySQL AB in the United States, the European Union and other countries.

Node.js ® is an official trademark of Joyent. Red Hat Software Collections is not formally related to or endorsed by the official Joyent Node.js open source or commercial project.

The OpenStack ® Word Mark and OpenStack logo are either registered trademarks/service marks or trademarks/service marks of the OpenStack Foundation, in the United States and other countries and are used with the OpenStack Foundation's permission. We are not affiliated with, endorsed or sponsored by the OpenStack Foundation, or the OpenStack community.

All other trademarks are the property of their respective owners.

Abstract

This document provides high level guidance on selecting hardware for use with Red Hat Ceph Storage.

Table of Contents

CHAPTER 1. OVERVIEW

Ceph was designed to run on non-proprietary commodity hardware. Ceph supports elastic provisioning, which makes building and maintaining petabyte-to-exabyte scale data clusters economically feasible. Many mass storage systems are great at storage, but they run out of throughput or IOPS well before they run out of capacity—making them unsuitable for some cloud computing applications. Ceph scales performance and capacity independently, which enables Ceph to support deployments optimized to a particular use case.

While Ceph runs on commodity hardware, this fact **DOES NOT** mean selecting the cheapest hardware possible is necessarily a good idea. The phrase "commodity hardware" simply means that running Ceph does not require a lock-in to a particular hardware vendor. Misunderstanding the phrase "commodity hardware" can lead to common mistakes in hardware selection, including:

- Repurposing underpowered legacy hardware for use with Ceph.
- Using dissimilar hardware in the same pool.
- Using 1Gbps networks instead of 10Gbps or greater.
- Neglecting to setup both public and cluster networks.
- Using RAID instead of JBOD.
- Selecting drives on a price basis without regard to performance or throughput.
- Journaling on OSD data drives when the use case calls for an SSD journal.
- Having a disk controller with insufficient throughput characteristics.

Red Hat has performed extensive testing to characterize Red Hat Ceph Storage deployments on a range of storage servers in optimized configurations.

Important

Before purchasing hardware for use with Ceph, please read the following document.

Red Hat Ceph Storage Hardware Configuration Guide

Whereas, the RHCS Hardware Configuration Guide provides extensive detail, this guide is only intended to provide very high level guidance to avoid common hardware selection mistakes and to provide links to tested sizing and performance guides that provide significant detail with specific hardware setups, configuration, and the performance results.

CHAPTER 2. GENERAL PRINCIPLES

2.1. IDENTIFYING A PERFORMANCE USE CASE

One of the most important steps in a successful Ceph deployment is identifying a price/performance profile suitable for the cluster's use case and workload. It is important to choose the right hardware for the use case. For example, choosing IOPS-optimized hardware for a cold storage application increases hardware costs unnecessarily. Whereas, choosing capacity-optimized hardware for its more attractive price point in an IOPS-intensive workload will likely lead to unhappy users complaining about slow performance.

The primary use cases for Ceph are:

- **IOPS optimized:** IOPS optimized deployments are suitable for cloud computing operations, such as running MYSQL or MariaDB instances as virtual machines on OpenStack. IOPS optimized deployments require higher performance storage such as SAS drives and separate SSD journals to handle frequent write operations. Some high IOPS scenarios use all solid state drives, all flash storage, and sometimes RDMA over Infiniband to improve IOPS and total throughput. Additionally, the storage industry is evolving with Non-volatile Memory Express (NVMe) for SSDs, which should improve performance substantially.
- **Throughput optimized:** Throughput optimized deployments are suitable for serving up significant amounts of data, such as graphic, audio and video content. Throughput optimized deployments require networking hardware, controllers and SAS drives with acceptable total throughput characteristics. In cases where write performance is a requirement, SSD journals will substantially improve write performance.
- **Capacity-optimized:** Capacity optimized deployments are suitable for storing significant amounts of data as inexpensively as possible. Capacity optimized deployments typically trade performance for a more attractive price point. For example, capacity-optimized deployments often use slower and less expensive SATA drives and co-locate journals rather than using SSDs for journaling.

The foregoing use cases aren't exhaustive. Ceph is also evolving with its technology preview "BlueStore" for rotating disks, which may avoid the added cost of SSD journaling in some use cases when BlueStore is production ready. "PMStore," when it is production ready, should improve the performance of solid state drives and flash memory too.

2.2. CONSIDERING STORAGE DENSITY

Hardware planning should include distributing Ceph daemons and other processes that use Ceph across many hosts to maintain high availability in the event of hardware faults. This means that you will need to balance storage density considerations with the need to rebalance (backfill) your cluster in the event of hardware faults. A common mistake is to use very high storage density in small clusters, which can overload networking.

2.3. USING IDENTICAL HARDWARE

Create pools and define CRUSH hierarchies such that the OSD hardware within the pool is identical. That is:

- Same controller.
- Same drive size.

- Same RPMs.
- Same seek times.
- Same I/O.
- Same network throughput.
- Same journal configuration.

Using the same hardware within a pool provides a consistent performance profile, simplifies provisioning and streamlines troubleshooting.

2.4. USING 10GB ETHERNET-PRODUCTION MINIMUM

Carefully consider bandwidth requirements for your cluster network, be mindful of network link oversubscription, and segregate the intra-cluster traffic from the client-to-cluster traffic.

Important

1Gbps isn't suitable for production clusters.

In the case of a drive failure, replicating 1TB of data across a 1Gbps network takes 3 hours, and 3TBs (a typical drive configuration) takes 9 hours. By contrast, with a 10Gbps network, the replication times would be 20 minutes and 1 hour respectively. Remember that when an OSD fails, the cluster will recover by replicating the data it contained to other OSDs within the pool.

```
failed OSD(s)
-------------
 total OSDs
```

The failure of a larger domain such as a rack means that your cluster will utilize considerably more bandwidth. Administrators usually prefer that a cluster recovers as quickly as possible.

At a **minimum**, a single 10Gbps Ethernet link should be used for storage hardware. If your Ceph nodes have many drives each, add additional 10Gbps Ethernet links for connectivity and throughput.

Important

Set up front and backside networks on separate NICs.

Ceph supports a public (front-side) network and a cluster (back-side) network. The public network handles client traffic and communication with Ceph monitors. The cluster (back-side) network handles OSD heartbeats, replication, backfilling and recovery traffic. Red Hat recommends allocating bandwidth to the cluster (back-side) network such that it is a multiple of the front-side network using **`osd pool default size`** as the basis for your multiple on replicated pools. Red Hat also recommends running the public and cluster networks on separate NICs.

When building a cluster consisting of multiple racks (common for large clusters), consider utilizing as much network bandwidth between switches in a "fat tree" design for optimal performance. A typical 10Gbps Ethernet switch has 48 10Gbps ports and four 40Gbps ports. If you only use one 40Gbps port for connectivity, you can only connect 4 servers at full speed (i.e., 10gbps x 4). Use

your 40Gbps ports for maximum throughput. If you have unused 10G ports, you can aggregate them (with QSFP+ to 4x SFP+ cables) into more 40G ports to connect to other racks and to spine routers.

For network optimization, we recommend a jumbo frame for a better CPU/bandwidth ratio. We also recommend a non-blocking network switch back-plane.

As of Red Hat Ceph Storage v2.0, Ceph also supports RDMA over Infiniband. RDMA reduces TCP workload and thereby reduces CPU utilization while increasing throughput.

You may deploy a Ceph cluster across geographic regions; however, this is NOT RECOMMENDED UNLESS you use a dedicated network connection between datacenters. Ceph prefers consistency and acknowledges writes synchronously. Using the internet (packet-switched with many hops) between geographically separate datacenters will introduce significant write latency.

2.5. AVOIDING RAID

Ceph replicates or erasure codes objects. RAID is redundant and reduces available capacity. Consequently, RAID is an unnecessary expense. Additionally, a degraded RAID will have a **negative** impact on performance. If you have systems with RAID controllers, configure them for RAID 0 (JBOD).

CHAPTER 3. SELECTING OSD HARDWARE

Red Hat reference architectures generally involve a 3x3 matrix reflecting small, medium or large clusters optimized for one of IOPS, throughput or capacity. Each scenario balances tradeoffs between:

- Server density v. OSD ratio.
- Network bandwidth v. server density.
- CPU v. OSD ratio.
- RAM v. OSD Ratio.
- SSD Write Journal (if applicable) v. OSD ratio.
- Host Bus Adapter/controller tradeoffs.

Note

The following documents do not constitute a recommendation for any particular hardware vendor. These documents reflect Ceph storage clusters that have been deployed, configured and tested. The intent of providing these documents is to illustrate specific hardware selection, as well as Ceph configuration and performance characteristics for real world use cases.

3.1. INTEL HARDWARE GUIDE

Based on extensive testing by Red Hat and Intel with a variety of hardware providers, the following document provides general performance, capacity, and sizing guidance for servers based on Intel® Xeon® processors, optionally equipped with Intel® Solid State Drive Data Center (Intel® SSD DC) Series.

Red Hat Ceph Storage on servers with Intel® processors and SSDs.

3.2. SUPERMICRO SERVER FAMILY GUIDE

To address the need for performance, capacity, and sizing guidance, Red Hat and Supermicro have performed extensive testing to characterize optimized configurations for deploying Red Hat Ceph Storage on a range of Supermicro storage servers. For details, see the following document:

Red Hat Ceph Storage clusters on Supermicro storage servers

Note

This document was published in August of 2015. It does not contain testing for newly available features in Ceph 2.0 and beyond.

3.3. QUANTA/QCT SERVER FAMILY GUIDE

Use of standard hardware components helps ensure low costs, while QCT’s innovative

development model enables organizations to iterate more rapidly on a family of server designs optimized for different types of Ceph workloads. Red Hat Ceph Storage on QCT servers lets organizations scale out to thousands of nodes, with the ability to scale storage performance and capacity independently, depending on the needs of the application and the chosen storage server platform.

To address the need for performance, capacity, and sizing guidance, Red Hat and QCT (Quanta Cloud Technology) have performed extensive testing to characterize optimized configurations for deploying Red Hat Ceph Storage on a range of QCT servers.

For details, see the following document:

Red Hat Ceph Storage on QCT Servers

3.4. CISCO C3160 GUIDE

This document provides an overview of the use of a Cisco UCS C3160 server with Ceph in a scaling multinode setup with petabytes (PB) of storage. It demonstrates the suitability of the Cisco UCS C3160 in object and block storage environments, and the server's dense storage capabilities, performance, and scalability as you add more nodes.

Cisco UCS C3160 high Density Rack Server with Red Hat Ceph Storage

3.5. SAMSUNG SIERRA FLASH ARRAYS GUIDE

To address the needs of Ceph users to effectively deploy All-Flash Ceph clusters optimized for performance, Samsung Semiconductor Inc. and Red Hat have performed extensive testing to characterize optimized configurations for deploying Red Hat Ceph Storage on Samsung NVMe SSDs deployed in a Samsung NVMe Reference Architecture. For details, see:

Red Hat Ceph Storage on Samsung NVMe SSDs

CHAPTER 4. MINIMUM RECOMMENDATIONS

Ceph can run on non-proprietary commodity hardware. Small production clusters and development clusters can run without performance optimization with modest hardware.

Process	Criteria	Minimum Recommended
RHSC	Processor	1x AMD64 and Intel 64 quad-core
	RAM	4 GB minimum per instance
	Disk Space	10 GB per instance
	Network	2x 1GB Ethernet NICs
ceph-osd	Processor	1x AMD64 and Intel 64
	RAM	2 GB of RAM per deamon
	OS Disk	1x OS disk per host
	Volume Storage	1x storage drive per daemon
	Journal	1x SSD partition per daemon (optional)
	Network	2x 1GB Ethernet NICs
ceph-mon	Processor	1x AMD64 and Intel 64
	RAM	1 GB per daemon
	Disk Space	10 GB per daemon

Process	Criteria	Minimum Recommended
	Monitor Disk	1x SSD disk for `leveldb` monitor data (optional).