RED HAT® STORAGE

TECHNOLOGY DETAIL

RED HAT CEPH STORAGE ON SERVERS WITH INTEL® PROCESSORS AND SSDs

Configuring scalable, workload-optimized Ceph clusters

Ceph has been developed to deliver object, file, and block storage in one self-managing, self-healing platform with no single point of failure.

Red Hat Ceph Storage offers multipetabyte software-defined storage for the enterprise, across a range of industry-standard hardware.

Intel Xeon Scalable Processor-based servers equipped with Intel SSD DC Series are ideally suited for Red Hat Ceph Storage clusters.

With proper configuration, Red Hat Ceph Storage clusters can be designed for IOPS-optimized, throughput-optimized, or cost/capacity-optimized workloads.

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

redhat.com

EXECUTIVE SUMMARY

Ceph users frequently request simple, optimized cluster configurations for different workload types. Common requests are for throughput-optimized and capacity-optimized workloads, but input/output operations per second (IOPS)-intensive workloads on Ceph are also emerging. Based on extensive testing by Red Hat and Intel with a variety of hardware providers, this document provides general performance, capacity, and sizing guidance for servers based on Intel® Xeon® Scalable processors, optionally equipped with the latest Intel Solid State Drive (SSD) Data Center (DC) Series, including Intel Optane™ SSDs.

TABLE OF CONTENTS

INTRODUCTION

Storage infrastructure is undergoing tremendous change, particularly as organizations deploy storage to support big data and private clouds. Traditional scale-up arrays are limited in scalability, and their complexity at scale can compromise cost-effectiveness. In contrast, software-defined storage infrastructure based on clustered storage servers has emerged as a way to deploy cost-effective and manageable storage at scale, with Ceph among the leading solutions. In fact, cloud storage companies are already using Ceph at near exabyte scale, with expected continual growth.

Deploying Red Hat® Ceph Storage on servers with Intel Xeon processors and Intel SSD DC Series can significantly lower the cost of storing enterprise data and can help organizations manage exponential data growth. The software is a robust, petabyte-scale storage platform for those deploying public or private clouds. As a modern storage system for cloud deployments, Red Hat Ceph Storage offers mature interfaces for enterprise block and object storage, making it well suited for active archive, rich media, and cloud infrastructure workloads like OpenStack®.[1] Delivered in a unified self-healing and self-managing platform with no single point of failure, Red Hat Ceph Storage handles data management so businesses can focus on improving application availability, with properties that include:

- Scaling to exabytes.
- No single point of failure in the cluster.
- Lower capital expenses (CapEx) by running on industry-standard server hardware.
- Lower operational expenses (OpEx) by self-managing and self-healing.

Many organizations are trying to understand how to configure Intel Xeon processor-based servers for optimized Ceph clusters that meet their unique needs. Red Hat Ceph Storage is able to run on myriad diverse hardware configurations, but designing a successful Ceph cluster requires careful analysis of issues related to application, capacity, and workload. The ability to address dramatically different kinds of I/O workloads within a single Ceph cluster makes understanding these issues paramount to a successful deployment.

After extensive performance and server scalability evaluation and testing with many vendors, Red Hat has developed a proven methodology that helps ask and answer key questions that lead to properly sized and configured scale-out storage clusters based on Red Hat Ceph Storage. Described in greater detail in this guide, the methodology includes:

- Qualifying the need for scale-out storage.
- Identifying target workload I/O profiles.
- Choosing a storage access method.
- Identifying capacity.
- Selecting a data protection method.
- Determining fault domain risk tolerance.
- Choosing an object storage daemon (OSD) backing store.

1 openstack.org/assets/survey/April2017SurveyReport.pdf, *pp. 53*

CEPH ARCHITECTURE OVERVIEW

A Ceph storage cluster is built from large numbers of Ceph nodes for scalability, fault-tolerance, and performance. Each node is based on industry-standard hardware and uses intelligent Ceph daemons that communicate with each other to:

- Store and retrieve data.
- Replicate data.
- Monitor and report on cluster health.
- Redistribute data dynamically (remap and backfill).
- Ensure data integrity (scrubbing).
- Detect and recover from faults and failures.

To the Ceph client interface that reads and writes data, a Ceph storage 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 OSD daemons, or OSDs) both use the CRUSH (controlled replication under scalable hashing) algorithm for storage and retrieval of objects.

When a Ceph client reads or writes data, referred to as an I/O contex, it connects to a logical storage pool in the Ceph cluster. Figure 1 illustrates the overall Ceph architecture, with concepts that are described in the sections that follow.

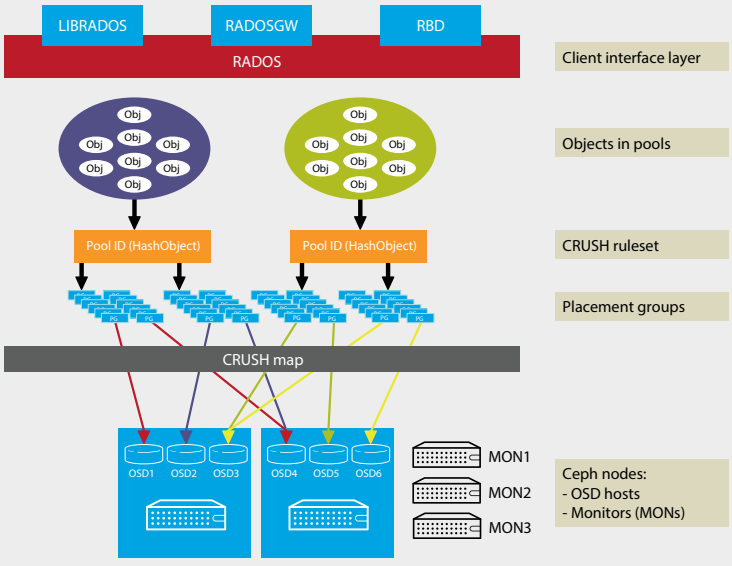

Figure 1. Clients write to Ceph storage pools while the CRUSH ruleset determines how placement groups (PGs) are distributed across object storage daemons (OSDs).

- ***CRUSH ruleset***. The CRUSH algorithm provides controlled, scalable, and declustered placement of replicated or erasure-coded data within Ceph and determines how to store and retrieve data by computing data storage locations. CRUSH empowers Ceph clients to communicate with OSDs directly, rather than through a centralized server or broker. By determining a method of storing and retrieving data by algorithm, Ceph avoids a single point of failure, a performance bottleneck, and a physical limit to scalability.
- ***Ceph monitors (MONs)***. Before Ceph clients can read or write data, they must contact a Ceph MON to obtain the current cluster map. A Ceph storage cluster can operate with a single monitor, but this introduces a single point of failure. For added reliability and fault tolerance, Ceph supports an odd number of monitors in a quorum (typically three or five for small to mid-sized clusters). Consensus among various monitor instances ensures consistent knowledge about the state of the cluster.
- ***Ceph OSD daemons***. In a Ceph cluster, Ceph OSDs store data and handle data replication, recovery, backfilling, and rebalancing. They also provide some cluster state information to Ceph monitors by checking other Ceph OSD daemons with a heartbeat mechanism. A Ceph storage cluster configured to keep three replicas of every object requires a minimum of three Ceph OSD daemons, two of which need to be operational to successfully process write requests. Ceph OSD daemons roughly correspond to a file system on a physical hard disk drive (HDD) or flash storage. Multiple OSDs can exist on a physical OSD node.

CLUSTER CONFIGURATION GUIDANCE

Despite the flexibility of Ceph, no one cluster or pool configuration fits all applications or situations. Instead, successful configuration of a Ceph cluster requires answering key questions about how the cluster will be used and the applications it will serve.

QUALIFYING THE NEED FOR SOFTWARE-DESIGNED STORAGE

Not every storage situation calls for software-defined storage. When requirements include several of the following needs, scale-out storage is likely the best solution.

- **Dynamic storage provisioning**. By dynamically provisioning capacity from a pool of storage, organizations are typically building a private storage cloud, emulating services such as Amazon Simple Storage Service (S3) for object storage or Amazon Elastic Block Store (EBS).
- **Standard storage servers**. Scale-out storage employs storage clusters built from industry-standard Intel Xeon processor-based servers rather than proprietary storage appliances, allowing incremental growth of storage capacity and/or performance without forklift appliance upgrades.
- **Unified namespaces**. Scale-out storage allows pooling storage across tens, hundreds, or even thousands of storage servers in one or more unified namespaces, ideal for access by analytics tools.
- **High data availability**. Scale-out storage provides high data availability across what would otherwise be "server storage islands" within the storage cluster.
- **Independent scalability of performance and capacity**. Unlike typical scale-up network-attached storage (NAS) and storage area network (SAN) devices that frequently run out of performance before running out of capacity, scale-out storage allows organizations to add storage performance or capacity incrementally by independently adding more storage servers or disks as required.

IDENTIFYING TARGET WORKLOAD I/O PROFILES

Accommodating the target workload I/O profile is perhaps the most crucial design consideration. As a first approximation, organizations need to understand if they are simply deploying low-cost archive storage or if their storage needs to meet specific performance requirements. If the lowest cost per terabyte is the overriding need, a Ceph cluster architecture can be designed at dramatically lower costs. For example, Ceph object archives with erasure-coded pools and without dedicated SSD write journals can be dramatically lower in cost than Ceph block devices on 3x-replicated pools with dedicated flash write journals.

For performance-oriented Ceph clusters, however, IOPS, throughput, and latency requirements must be clearly defined. Historically, Ceph has performed very well with high-throughput workloads and has been widely deployed for these use cases. Applications are frequently characterized by large-block, asynchronous, sequential I/O (such as digital media performance nodes). In contrast, high IOPS workloads are frequently characterized by small-block synchronous random I/O (for example, 4KB random I/O). The use of Ceph for high IOPS open source database workloads is emerging with MySQL, MariaDB, PostgreSQL, and other offerings. Moreover, when Ceph is deployed as Cinder block storage for OpenStack virtual machines (VMs), it typically serves a mix of IOPS- and throughput-intensive I/O patterns.

Additionally, understanding the workload read/write mix can affect architecture design decisions. For example, erasure-coded pools can perform better than replicated pools for sequential writes, but perform worse than replicated pools for sequential reads. As a result, a write-mostly object archive workload (such as video surveillance archival) may perform similarly between erasure-coded pools and replicated pools, although erasure-coded pools may be significantly less expensive.

To simplify configuration and testing options and optimally structure cluster configurations, Red Hat categorizes workload profiles as:

- IOPS-optimized.
- Throughput-optimized.
- Cost/capacity-optimized.

Table 1 provides the criteria used to identify optimal Red Hat Ceph Storage cluster configurations, including their properties and example uses. These categories are provided as general guidelines for hardware purchases and configuration decisions and can be adjusted to satisfy unique workload blends. As the workload mix varies from organization to organization, actual hardware configurations chosen will vary.

As previously mentioned, a single Ceph cluster can be configured to have multiple pools that serve different workloads. For example, OSDs on IOPS-optimized servers can be configured into a pool serving MySQL workloads, while OSDs on throughput-optimized servers can be configured into a pool serving digital media performance workloads.

TABLE 1. CEPH CLUSTER OPTIMIZATION CRITERIA

OPTIMIZATION CRITERIA	PROPERTIES	EXAMPLE USES
IOPS-OPTIMIZED	• Lowest cost per IOPS • Highest IOPS • Meets minimum fault domain recommendation (single server is less than or equal to 10% of the cluster)	• Typically block storage • 3x replication on HDD or 2x replication on Intel SSD DC Series • MySQL on OpenStack clouds
THROUGHPUT-OPTIMIZED	• Lowest cost per given unit of throughput • Highest throughput • Highest throughput per BTU • Highest throughput per watt • Meets minimum fault domain recommendation (single server is less than or equal to 10% of the cluster)	• Block or object storage • 3x replication • Active performance storage for video, audio, and images • Streaming media
CAPACITY-OPTIMIZED	• Lowest cost per TB • Lowest BTU per TB • Lowest watt per TB • Meets minimum fault domain recommendation (single server is less than or equal to 15% of the cluster)	• Typically object storage • Erasure coding common for maximizing usable capacity • Object archive • Video, audio, and image object archive repositories

CHOOSING A STORAGE ACCESS METHOD

Choosing a storage access method is another important design consideration. All data in Ceph is stored in pools—regardless of type. The data itself is stored in the form of objects via the Reliable Autonomic Distributed Object Store (RADOS) layer (Figure 2) to:

- Avoid a single point of failure.
- Provide data consistency and reliability.
- Enable data replication and migration.
- Offer automatic fault detection and recovery.

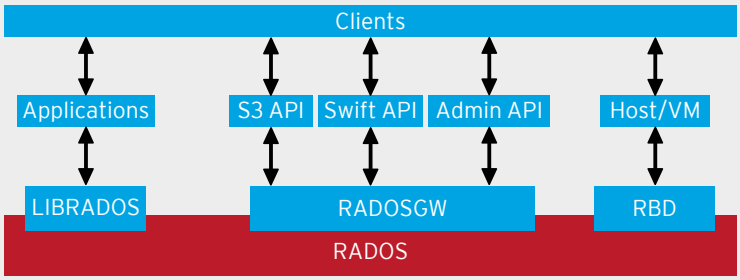

Figure 2. RADOS is the foundation of the Ceph storage cluster.

Writing and reading data in a Ceph storage cluster is accomplished using the Ceph client architecture. Ceph clients differ from competitive offerings in how they present data storage interfaces. A wide range of access methods are supported, including:

- **RADOS Gateway (RADOSGW or RGW).** RADOSGW is a bucket-based object storage gateway service with S3-compatible and OpenStack Swift-compatible RESTful interfaces.
- **LIBRADOS.** LIBRADOS provides direct access to RADOS with libraries for most programming languages, including C, C++, Java™, Python, Ruby, and PHP.
- **RADOS Block Device (RBD).** RBD offers a Ceph block storage device that mounts like a physical storage drive for use by both physical and virtual systems (with a Linux® kernel driver, KVM/QEMU storage back-end, or user-space libraries).

Storage access method and data protection method (discussed later in this document) are interrelated. For example, Ceph block storage is currently only supported on replicated pools, while Ceph object storage is supported on both erasure-coded and replicated pools.[2] Due to a significant difference in media costs, replicated architectures are categorically more expensive than erasure-coded architectures. Note that although CephFS file system support is provided with Red Hat Ceph Storage 3, performance and sizing characterization in this guide is focused on block and object storage.

IDENTIFYING CAPACITY NEEDS

Identifying storage capacity may seem trivial, but it can have a significant effect on the chosen target server architecture. In particular, the cluster's predicted storage capacity needs must be considered together with fault domain risk tolerance and other capabilities. For example, minimum server fault domain recommendations for a small half-petabyte cluster will prevent the use of ultra-dense storage servers in the architecture. This avoids unacceptable fault domain risk on a small number of very large nodes. Table 2 lists broad server sizing trends, with typical types of servers categorized by both workload optimization and overall cluster size.

2 *Note: Block storage (Ceph RBD) can use erasure-coded pools, but requires the BlueStore filestore. BlueStore is in Technology Preview for Red Hat Ceph Storage 3 and is thus not yet supported in production environments.*

TABLE 2. BROAD SERVER SIZING TRENDS

OPTIMIZATION CRITERIA	OPENSTACK STARTER (64TB)	SMALL (250TB)	MEDIUM (1PB)	LARGE (2PB)
IOPS-OPTIMIZED	• Servers with 2-4x PCIe/NVMe slots, or • Servers with 8-12x 2.5-inch SSD bays (SAS/SATA)		• Servers with 16-24x 2.5-Inch SSD bays (SAS/SATA/PCIe)	• Not typical
THROUGHPUT-OPTIMIZED	• Servers with 12-16x 3.5-inch drive bays		• Servers with 24-36x 3.5-inch drive bays	• Servers with 24-36x 3.5-inch drive bays
CAPACITY-OPTIMIZED				• Servers with 60-72x 3.5-inch drive bays

SELECTING A DATA PROTECTION METHOD

As a design decision, choosing the data protection method can affect the solution's total cost of ownership (TCO) more than any other factor. The chosen data protection method strongly affects the amount of raw storage capacity that must be purchased to yield the desired amount of usable storage capacity.

Applications have diverse needs for performance and availability. As a result, Ceph provides data protection at the storage pool level.

- **Replicated storage pools**. Replication makes full copies of stored objects and is ideal for quick recovery. In a replicated storage pool, Ceph configuration defaults to a replication factor of three, involving a primary OSD and two secondary OSDs. If two of the three OSDs in a placement group become unavailable, data may be read, but write operations will be suspended until at least two OSDs are operational.
- **Erasure-coded storage pools**. Erasure coding provides a single copy of data plus parity and is useful for archive storage and cost-effective durability and availability. With erasure coding, storage pool objects are divided into chunks using the **n=k+m** notation, where **k** is the number data chunks that are created, **m** is the number of coding chunks that will be created to provide data protection, and **n** is the total number of chunks placed by CRUSH after the erasure coding process.

Ceph block and object storage is supported on both replicated and erasure-coded pools. Depending on the performance needs and read/write mix of an object storage workload, an erasure-coded pool can provide an extremely cost-effective solution that meets performance requirements. Figure 3 illustrates the relationships among Ceph storage pools, placement groups (PGs), and OSDs for both replicated and erasure-coded pools.

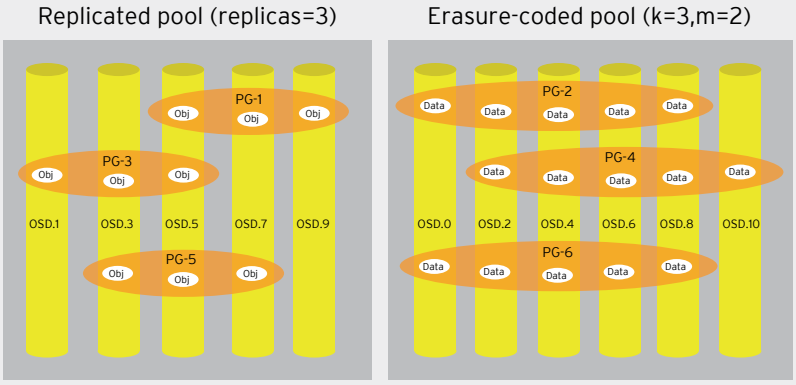

Figure 3. Ceph storage pools, PGs, and OSDs, for both replicated and erasure-coded pools.

DETERMINING FAULT DOMAIN RISK TOLERANCE

It may be tempting to deploy the largest servers possible in the interest of economics. However, production environments need to provide reliability and availability for the applications they serve, and this necessity extends to the storage upon which they depend. The fault domain that a single OSD server represents is key to cluster design. As a result, dense servers should be reserved for multi-petabyte clusters where the capacity of an individual server accounts for less than 10-15% of the total cluster capacity. This recommendation may be relaxed for less critical pilot projects. Primary factors for weighing fault domain risks include:

- **Reserving capacity for self-healing**. When a storage node fails, Ceph self-healing begins after a configured time period. For successful self-healing, the unused storage capacity of the surviving cluster nodes must be greater than the used capacity of the failed server. For example, in a 10-node cluster, each node should reserve 10% unused capacity for self-healing of a failed node (in addition to reserving 10% for statistical deviation due to using algorithmic placement). As a result, each node in a cluster should operate at less than 80% of total capacity.
- **Accommodating impact on performance**. During self-healing, a percentage of cluster throughput will be diverted to reconstituting object copies from the failed node on the surviving nodes. The percentage of cluster performance degradation is a function of the number of nodes in the cluster and how Ceph is configured. More nodes in the cluster results in less impact per node.

Ceph will automatically recover by re-replicating the data from the failed node using secondary copies on other nodes in the cluster. As a result, a node failure has several effects:

- Total cluster capacity is reduced by some fraction.
- Total cluster throughput is reduced by some fraction.
- The cluster enters an I/O-heavy recovery process, temporarily diverting an additional fraction of the available throughput.

For more information on Ceph architecture, see the Ceph documentation at docs.ceph.com/docs/master/architecture.

The time required for the recovery process is directly proportional to how much data was on the failed node and how much throughput the rest of the cluster can sustain. A general formula for calculating recovery time in a Ceph cluster given one disk per OSD is:

```
Recovery time seconds = (disk capacity in gigabits / network speed) / (nodes – 1)
```

For example, if a 2TB OSD node fails in a 10-node cluster with a 10 GbE (Gigabit Ethernet) back-end, the cluster will take approximately three minutes to recover with 100% of the network bandwidth and no CPU overhead. In practice, using 20% of the available 10 GbE network, the cluster will take approximately 15 minutes to recover, and that time will double with a 4TB drive.

Red Hat recommends the following minimum cluster sizes:

- **Supported minimum cluster size:** Three storage (OSD) servers, suitable for use cases with higher risk tolerance for performance degradation during node failure recovery
- **Recommended minimum cluster size (IOPS- and throughput-optimized cluster):** 10 storage (OSD) servers
- **Recommended minimum cluster size (cost/capacity-optimized cluster):** Seven storage (OSD) servers
- **Recommended minimum cluster size (erasure-coded configurations)**: At least k+m+1 nodes, e.g., using EC:4+2 would require a minimum of seven nodes.

There are also other considerations related to fault domain risk and performance that must be considered. Ceph replicates objects across multiple nodes in a storage cluster to provide data redundancy and higher data availability. When designing a cluster, it is important to ask:

- Should the replicated node be in the same rack or multiple racks to avoid a single rack failure?
- Should Ceph OSD traffic stay within the rack or span across racks in a dedicated or shared network?
- Are the application servers in the rack or datacenter proximate to the storage nodes?
- How many concurrent node failures can be tolerated?

Automatic and intelligent placement of object replicas across server, rack, row, and datacenter fault domains can be governed by CRUSH ruleset configuration parameters.

CHOOSING AN OSD BACKING STORE

Ceph now offers a choice of backing store for OSD daemons, including FileStore or BlueStore. With FileStore, OSDs store Ceph objects on a native file system (e.g. XFS). With BlueStore, OSDs store Ceph objects directly to a block device, enhancing performance with some workloads.

FileStore OSD backing store

With a traditional FileStore backing store, Ceph uses a journal to allow it to create atomic updates—required to ensure data consistency. To complete these updates, an OSD writes the data payload and metadata to a write journal before writing to the OSD's data partition. A write is acknowledged to the client after all OSD peers in the PG have successfully written their assigned replica or shard to their write journal.

A beneficial side-effect of write journaling may be some potential coalescing of small writes. For performance-optimized clusters, journals are typically located on a partition of a faster media type than the OSD media. For example, a throughput-optimized OSD server typically has HDD-based OSDs, and a dedicated write journal based on an SSD.

BlueStore OSD backing store

OSD BlueStore is a new back-end for OSD daemons that allows for storing objects directly on the Ceph block devices without any file system interface. BlueStore stores OSD metadata in a RocksDB key-value database that contains:

- Object metadata
- A write-ahead log (WAL)
- Ceph `omap` data
- Allocator metadata

The BlueStore backing store provides a number of advantages:

- **No large double-writes.** BlueStore first writes any new data to unallocated space on a block device, and then commits a RocksDB transaction that updates the object metadata to reference the new region of the disk. When the write operation is below a configurable size threshold, the system falls back to a write-ahead journaling scheme, similar to the FileStore OSD backing store.
- **Multidevice support**. BlueStore can use multiple block devices for storing different data. For example, HDDs can be used for data, SSDs for metadata, and non-volatile memory (NVM) or non-volatile random-access memory (NVRAM) can be used for persistent storage of the RocksDB write-ahead log (WAL).
- **Efficient block device usage**. Because BlueStore does not use a file system, it minimizes the need to clear the storage device cache.
- **Flexible allocator**. The block allocation scheme is pluggable, allowing BlueStore to implement different policies for different types of storage devices. For example, there is a different behavior for HDDs and SSDs.
- **Inline compression**. BlueStore supports inline compression using snappy, zlib, or iz4. Whether data in BlueStore is compressed is determined by a combination of the compression mode, and any hints associated with a write operation.
- **Checksums**. BlueStore checksums all metadata and data written to disk. Full data checksumming does increase the amount of metadata that BlueStore must store and manage.

INTEL HARDWARE CONFIGURATION GUIDELINES

The sections that follow provide broad guidelines for the selection of monitor nodes and OSD hosts. Actual configuration of OSD servers can vary based on application and workload optimization.

MONITOR NODES

The Ceph monitor is a datastore for the health of the entire cluster and also contains the cluster log. A minimum of three monitors are recommended for a cluster quorum. Monitor nodes typically have fairly modest CPU and memory requirements. A single rack unit (1U) server with a low-cost Intel processor (such as an Intel Xeon Bronze or Silver Processor), 16GB of RAM, and GbE networking should suffice in most cases. Since logs are stored on local disk(s) on the monitor node, it is important to make sure that sufficient disk space is provisioned. In addition, the monitor store should be placed on an Intel SSD DC Series, because the leveldb store can become I/O bound.

For example, when monitoring 100 OSDs in a healthy Ceph cluster, each monitor will collect data for all of the OSDs until all the monitors are synchronized with the same Ceph cluster information. The size of the datastore can vary, but 200MB up to 1-2GB are to be expected, depending not only on the size of the cluster, but the state change (churn) that the cluster undergoes. Logs for recovering clusters can grow quickly, reaching dozens and even hundreds of gigabytes. Abnormal monitor datastore growth should be investigated by an operator, as there is usually an underlying condition that should be remedied. Log rotation is a good practice that can guarantee that available disk space is not blindly consumed, especially if verbose debugging output is set on the monitors, since they will generate a large amount of logging information.[3] In most situations, monitors should be run on distinct nodes or on VMs that reside on physically separate machines to prevent a single point of failure.

OSD HOSTS

Ceph OSD hosts are configured differently depending on both workload optimization and the data devices installed: HDDs, SSDs, or NVMe SSDs.

CPU specifications

CPU recommendations for OSD hosts differ depending on the media that is employed for the OSDs.

- For HDD-based OSDs, one core-GHz is recommended for each OSD. For example, 16 HDD-based OSDs can be supported with an Intel Xeon Silver 4110 Processor:
  `8 cores * 2.10 GHz = 16.8 core-GHz`.
- For OSDs based on 3D NAND NVMe SSDs, such as Intel SSD Data Center P4500 Series NVMe drives, the recommendation is for four OSDs per SSD, and four SSDs per server (16 OSDs per server, total). Intel Xeon Gold 6152 Processors are recommended for these 16 OSDs:
  `22 cores * 2.1GHz * two sockets = 80 core-GHz`.
- In addition, use of checksums and inline compression may require additional processing power (cores), while providing a significant savings on capacity (in the case of compression).

Memory specifications

Red Hat typically recommends a baseline of 16GB of RAM, with an additional 2GB of RAM per OSD. When sizing memory requirements, it is important to consider:

- The number of OSDs per node.
- The number of memory banks available.
- The number of memory channels per bank.
- The cost of DRAM (DIMMs).

3 Refer to Ceph documentation on monitor log settings for additional details.

Data devices

OSDs and OSD data drives are independently configurable and Ceph OSD performance is naturally dependent on the throughput and latency of the underlying media. The actual number of OSDs configured per drive depends on the type of media configured on the OSD host. For magnetic storage media, one OSD should be configured per HDD. On the other hand, an IOPS-optimized OSD host with a smaller number of high-speed SSDs might be configured with 2-4 OSDs per SSD to exploit the available I/O bandwidth.

- **Write journal ratio recommendations for HDD-based OSDs**:
  - 4-5 HDD-OSDs for each SSD-based journal drive. To help decrease cost for cost/capacity-optimized clusters, journals can be co-located with OSDs on the same HDD via partitions.
  - 12-18 HDD-OSDs for each NVMe-based journal/metadata/caching drive.
- **Write journal recommendations for SSD-based OSDs**:
  - When using an all-NVMe OSD cluster for data storage, Intel SSD Data Center P4800X are recommended as a journal/metadata/caching drive, as they provide the highest IOPS and lowest tail latency (4x NVMe OSDs per 1x P4800X write journal).

Storage Media

Performance and economics for Ceph clusters both depend heavily on an effective choice of storage media. For throughput- and cost/capacity-optimized clusters, magnetic media currently accounts for the bulk of the deployed storage capacity, but the economics of SSDs are changing rapidly.

- **Magnetic media.** Enterprise-, or cloud-class HDDs should be used for Ceph clusters. Desktop-class disk drives are not well suited for Ceph deployments as they lack sufficient rotational vibration (RV) compensation for high-density, high-duty-cycle applications and use cases. When dozens or hundreds of rotating HDDs are installed in close proximity, RV quickly becomes a challenge. Failures, errors, and even overall cluster performance can be adversely affected by the rotation of neighboring disks interfering with the rapidly spinning platters in high-density storage enclosures. Enterprise-class HDDs contain higher quality bearings and RV compensation circuitry to mitigate these issues in multispindle applications and use cases—especially in densities above 4-6 HDDs in a single enclosure. Both SAS and SATA interface types are acceptable.
- **Solid-state media.** Solid-state media can be used to host OSDs directly (e.g., for IOPS-intensive configurations), to accelerate Ceph write journals (for FileStore backing store), or for XFS file system caching using Intel Cache Acceleration Software (CAS).[4] Ceph is strongly consistent storage, so every write to the Ceph cluster must be written to Ceph journals (when using a FileStore backing store) before the write is acknowledged to the client. The data remain in the journal until all replicas or shards are acknowledged as fully written. Only then will the next write happen. SSD journals let the OSDs write faster, reducing the time before a write acknowledgment is sent to the client. In some cases, several small writes can be coalesced during a single journal flush, which can also improve performance. SSDs can also be used for OSD data as well, as recommended in IOPS-optimized Ceph configurations. NVMe SSDs provide advantages above SSDs, removing legacy instruction sets and streamlining communications to the CPU. While the price of NVMe SSDs may be higher, the price/performance is typically lower.

4 See the section titled "Intel CAS and Red Hat Ceph Storage" in this document.

SSD choice is an essential factor and a key area of focus for Red Hat and Intel. Important criteria to consider when selecting solid-state media for Ceph include:

- **Classification**. Only enterprise-class SSDs should be deployed with Ceph. Consumer-class SSDs should not be used. Intel recommends Intel SSD Data Center Series drives.
- **Endurance**. Write endurance is important, as Ceph write journals are heavily used and could exceed recommended program/erase cycles of an SSD rated for lower endurance.
- **Power fail protection**. Supercapacitors for power failure protection are vital. In the event of a power failure, supercapacitors must be properly sized to allow the drive to persist all in-flight writes to non-volatile NAND storage. Intel SSD Data Center Series drives offer greater than 2 million Power Loss Imminent (PLI) cycles.[5]
- **Performance**. For Ceph write journaling, the write throughput rating of the journal device should exceed the aggregate write throughput rating of all underlying OSD devices that are served by that journal device.
- **Reliability**. Intel SSD Data Center Series drives offer self-testing and trusted protection from data loss. With 10^{17} uncorrectable bit-error rate (UBER), these drives are proven to be 100-fold more reliable than consumer SSDs, helping to prevent silent data corruption (SDC).[6] These devices also supersede the Joint Electonic Device Engineering Council (JEDEC) annual failure rate (AFR).[7]

I/O controllers

Servers with JBOD (just a bunch of disks) host bus adapters (HBAs) are generally appropriate for OSD hosts, depending on workload and application expectations. For large-block sequential I/O workload patterns, HDDs typically perform better when configured in JBOD mode than as redundant array of independent disk (RAID) volumes. However, for small-block random I/O workload patterns, HDD-based OSDs configured on single-drive RAID 0 volumes provide higher IOPS than when configured in JBOD mode.

Many modern systems that house more than eight drives have SAS expander chips on the drive hot swap backplane. Similar to network switches, SAS expanders often allow connection of many SAS devices to a controller with a limited number of SAS lanes. Ceph node configurations with SAS expanders are well suited for large capacity-optimized clusters. However, when selecting hardware with SAS expanders, consider how the following will affect performance:

- Adding extra latency.
- Oversubscribed SAS lanes.
- Spanning Tree Protocol (STP) overhead of tunneling SATA over SAS

Due to backplane oversubscription or poor design, some servers used for Ceph deployments may encounter sub-par performance in systems that use SAS expanders. The type of controller, expander, and even brand of drive and firmware all play a part in determining performance.

5 intel.com/content/www/us/en/solid-state-drives/data-center-class-solid-state-drive-brief.html?wapkw=benefits+of+ssd
6 intel.com/content/www/us/en/solid-state-drives/data-center-class-solid-state-drive-brief.html?wapkw=benefits+of+ssd
7 *Intel Data Center SSDs deliver 10^{17} UBER, see* jedec.org/standards-documents/focus/flash/solid-state-drives.

Network interfaces

Providing sufficient network bandwidth is essential for an effective and performant Ceph cluster. Fortunately, network technology is improving rapidly. Standard Ethernet-based interfaces are now available with ever-increasing bandwidth. In servers employed as OSD hosts, network capacity should generally relate to storage capacity. For smaller OSD hosts with 12-16 drive bays, 10 GbE is typically sufficient. For larger OSD hosts with 24-72 drive bays, 25, 40, or 50GbE may be preferred to provide the required bandwidth and throughput.

Physical deployment characteristics must also be taken into account. If the nodes are spread across multiple racks in the datacenter, the network design should ensure high bisectional bandwidth and minimal network diameter. For 10GbE connections, each OSD server should have two network interfaces for data traffic: one connected to the client systems and another for the private network connecting the OSD servers. For 25GbE connections and above, the simplicity of a single link may be preferable (and may also result in a better transaction balance).

CONFIGURATION GUIDANCE FOR INTEL PROCESSOR-BASED SERVERS

Table 3 provides general guidance for configuring Intel-based OSD hosts for Red Hat Ceph Storage.

TABLE 3. CONFIGURING INTEL-BASED SERVERS FOR RED HAT CEPH STORAGE

OPTIMIZATION CRITERIA	CONFIGURATION RECOMMENDATIONS
IOPS-OPTIMIZED	• Ceph RBD (block) pools • OSDs on SSDs 4x Intel SSD DC P4500 per server: Dual-socket Intel Xeon Gold 6152 processor • Four OSDs per SSD or NVMe drive, 16 OSDs per server • Data protection: Replication (2x on SSD-based OSDs) with regular backups to the object storage pool • All-NVMe cluster: Intel SSD DC P4800X for jounal/metadata/caching drive for highest IOPS and lowest tail latency

OPTIMIZATION CRITERIA	CONFIGURATION RECOMMENDATIONS

THROUGHPUT-OPTIMIZED	• Ceph RBD (block) or Ceph RGW (object) pools
	• OSDs on HDDs:
	  • Good: Write journals on Intel SSD DC S4600 480GB drives, with a ratio of 4-5 HDDs to each SSD
	  • Better: Write journals on Intel SSD DC P4600 1TB NVMe drives, with a ratio of 12-18 HDDs to each SSD
	  • Best: Write journals and Intel Cache Acceleration Software (CAS) on an Intel SSD DC P4600 2TB
	• One CPU core-GHz per OSD. For example:
	  • 12 OSD/HDDs/server: Dual-socket Intel Xeon Silver 4110 Processor (8 cores per socket @2.1 GHz)
	  • 36 OSD/HDDs/server: Dual-socket Intel Xeon Gold 6138 Processor (20 cores per socket @2.0 GHz)
	  • 60 OSD/HDDs/server: Dual-socket Intel Xeon Gold 6130 (16 cores * 2.1 GHz * 2 sockets)
	  • Data protection: Replication (read-intensive or mixed read/write) or erasure-coded (write-intensive, more CPU cores recommended)
	• High-bandwidth networking: Greater than 10GbE for servers with more than 12-16 drives

CAPACITY-OPTIMIZED	• Ceph RGW (object) pools • OSDs on HDDs with write journals co-located on HDDs in a separate partition • One CPU core-GHz per OSD. See throughput-optimized section above for examples. Intel Xeon E3 processor-based servers would suffice. • Data protection: Erasure coding

* All SSDs should be enterprise-class, meeting the requirements noted above.

EVALUATING THE LATEST INTEL AND CEPH TECHNOLOGY

One of the benefits of software-defined storage is the ability to rapidly exploit the latest advances in technology. In pursuit of this interest, Intel evaluated performance on an all-NVMe Ceph cluster with storage nodes based on:

- Intel Xeon Scalable Processors
- Intel Data Center Family 3D NAND SSDs for data storage
- Intel Optane™ SSD DC P4800X series SSDs with 3D XPoint™ technology for metadata storage
- Ceph BlueStore backing store[8]

8 *Red Hat Ceph Storage 3 and prior releases support a FileStore backing store. BlueStore is a Technology Preview feature in Red Hat Ceph Storage 3, and is thus not supported for production use. Red Hat believes that the general sizing guidance described herein should be appropriate for both BlueStore and FileStore backends.*

All-NVMe configurations are well suited to more demanding workloads or situations where low latency is desired. These configurations not only deliver higher performance, but typically offer attractive price-performance ratios as well as high reliability. Testing was performed on a six-node server configuration driven by six clients, as shown in Figure 4 and described in Table 4.

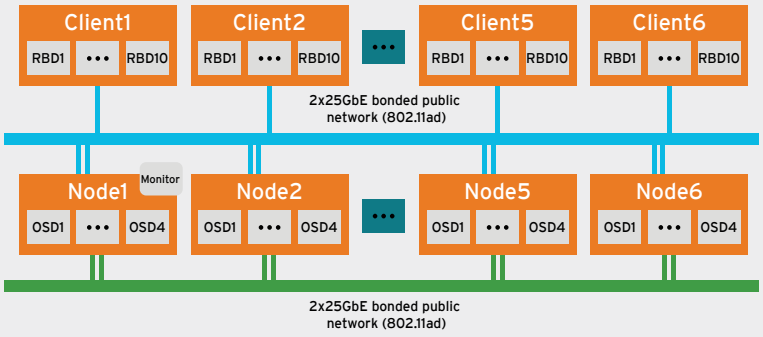

Figure 4. Red Hat Ceph Storage all-NVMe cluster.

TABLE 4. ALL-NVME CLUSTER CONFIGURATION

CLIENT NODES (6X)	• Intel Xeon processor E5-2699 v4 @2.2GHz • 128GB memory • Mellanox 100Gb Ethernet networking
STORAGE NODES (6X)	• Intel Xeon Platinum 8176 Processor @ 2.1 GHz • 384GB memory • 1x Intel Optane SSD DC P4800X 375GB as database (DB)/write-ahead log (WAL) drive • 4x Intel SSD DC P4500 4TB as data drive • 2x dual-port Mellanox 25Gb Ethernet
CONFIGURATION	• Ceph 12.1.1-175 (Luminous rc) BlueStore • 2x replication pool • 8192 placement groups (PGs) • Tested with 1, 2, and 4 OSDs per NVMe SSD

In order to evaluate the Ceph cluster in a configurable and repeatable fashion, engineers used the Ceph Benchmarking Tool (CBT). CBT is a test harness written in Python that can automate a variety of tasks related to testing the performance of Ceph clusters. CBT has various benchmark modules that run against different layers of the Ceph storage stack, including RADOS, RADOS Block Device (RBD), and RADOS Gateway (RGW).

The "librbdfio" benchmark module in CBT was used to test the performance of the RBD storage layer of the Ceph cluster. The module is the simplest way of testing the block storage performance of a Ceph cluster. It uses the benchmark tool known as Flexible I/O (FIO) to generate load to a given block device. Recent releases of FIO provide a RBD I/O engine. This ability allows FIO to test block storage

performance of RBD volumes through the userspace libRBD libraries, without KVM/QEMU configuration. These libraries are the same ones used by the QEMU back-end, so it allows an approximation to KVM/QEMU performance.

SMALL-BLOCK RANDOM PERFORMANCE

In general, hosting multiple OSD processes on a single NVMe SSD is recommended in order to fully utilize the available I/O bandwidth of the NVMe SSD. Figure 5 depicts performance differences when increasing the number of OSD processes per NVMe device, illustrating how engineers arrived at the general recommendation of configuring four OSDs per NVMe SSD.

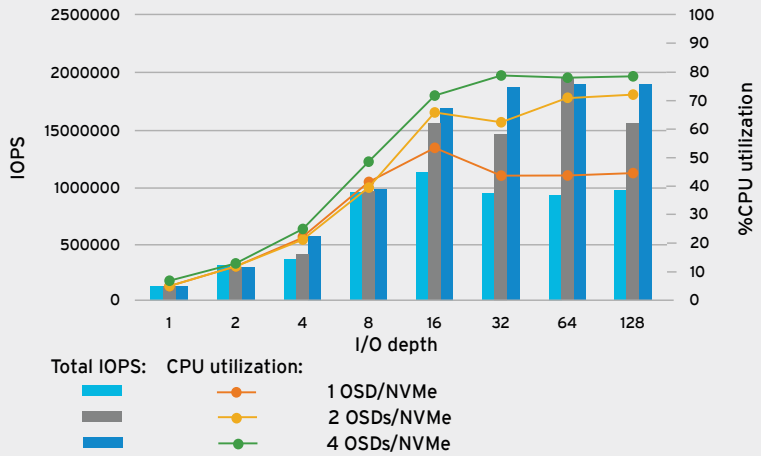

Figure 5. OSDs per NVMe SSD, read IOPS and CPU utilization comparison (60 libRBD clients, 4K random reads).

For 4KB random reads, the cluster was able to deliver 1.9M IOPS. Four OSDs mapped to each NVMe drive provided the optimal performance in terms of both IOPS and CPU utilization, with performance scaling from 138K IOPS to 1.9M IOPS as the I/O depth was increased from 1 to 128. From an I/O depth of one to eight, the different configurations of OSDs per NVMe performed similarly. For I/O depths of 16 and higher, increasing the number of OSDs per NVMe provided increasingly higher performance. Both IOPS and CPU utilization fluctuated for the one and two OSD per NVMe configurations at higher I/O depths. The four OSD per NVMe configuration utilized up to 80% of the total host CPU.

For 4KB random writes, the cluster was able to deliver 350K IOPS (Figure 6). Again, four OSDs mapped to each NVMe drive provided the optimal performance for 4KB random writes, in terms of both IOPS and CPU utilization. Using four OSDs per NVMe, the IOPS performance scaled from 50K IOPS to 350K IOPS as the I/O depth was increased from 1 to 128. From an I/O depth of one to two, the different configurations of OSDs per NVMe performed similarly. For I/O depths of four and higher, increasing numbers of OSDs per NVMe provided better performance. Importantly, at very large I/O depths of 128, four OSDs per NVMe did not always deliver the maximum IOPS. In some of these cases, performance was reduced due to increased overhead associated with context switching. The configuration with four OSDs per NVMe utilized up to 70% of the total host CPU.

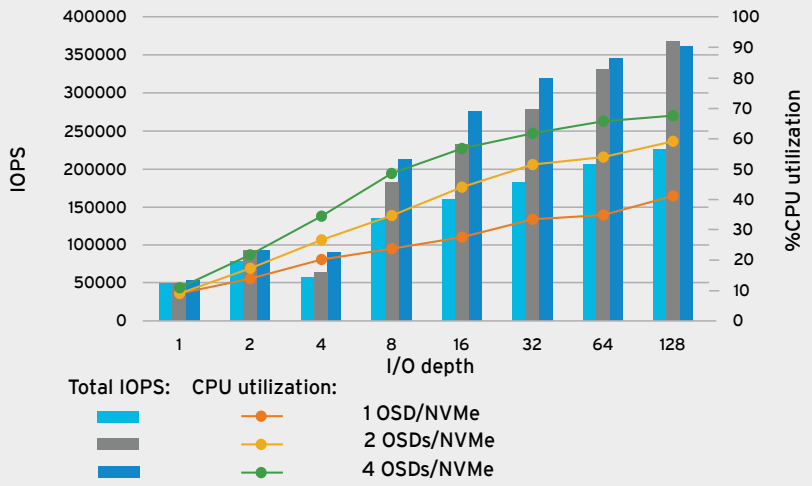

Figure 6. OSDs per NVMe, write IOPS and CPU comparison (60 libRBD clients, 4KB random writes).

For a 4KB random 70/30 mixed read/write workload, the cluster was able to deliver 998K IOPS (Figure 7). Depending on I/O depth, two or four OSDs mapped to each NVMe provided the optimal performance, in terms of both IOPS and CPU utilization. From I/O depths of one to 16, four OSDs per NVMe performed slightly better than the other configurations. For I/O depths of 32 to 128, however, two OSDs per NVMe outperformed all other configurations. Similar to the 4KB write results in Figure 6, increased overhead from the additional OSDs reduced the performance benefit gained from co-locating OSD processes. The configuration with four OSDs per NVMe utilized up to 70% of the total host CPU at the highest I/O depths, while the configuration with two OSDs per NVMe was slightly lower.

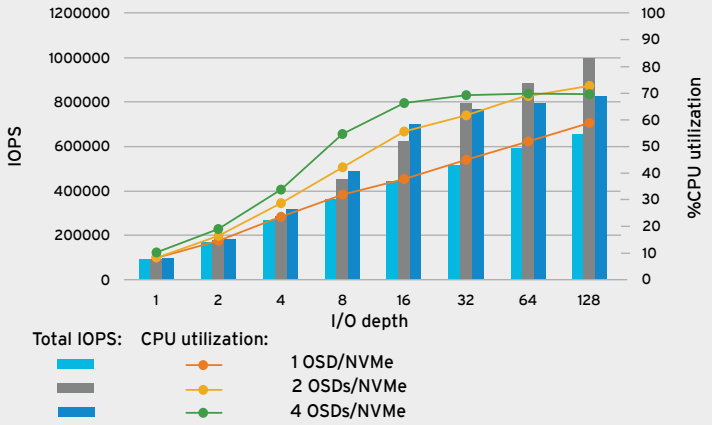

Figure 7. OSDs per NVMe, 70/30 read/write IOPS and CPU comparison (60 libRBD clients, 4KB random 70/30 r/w).

Hosting more OSD processes on an NVMe SSD poses a trade-off in terms of higher CPU utilization on the host. Configuring four OSD processes per NVMe device allows the Ceph cluster to serve more IOPS in-flight in most small-block random workloads.

LARGE-BLOCK SEQUENTIAL PERFORMANCE

Large-block sequential performance is another important factor when evaluating a Ceph cluster. Overall throughput is measured in terms of MB/s. Using CBT and the libRBD FIO module, engineers ran sequential read and write tests with a 1MB block size. Similar to the random 4KB workloads discussed above, engineers compared aggregate bandwidth from varying numbers of OSDs per NVMe.

For 1MB sequential read workloads, the cluster was able to deliver throughput of 31 GB/s (Figure 8). Mapping one OSD to each NVMe SSD is recommended in this case, as it provided the best performance across a range of I/O depths. The number of OSDs configured per NVMe drive did not affect the aggregate bandwidth until higher I/O depths were explored (e.g., I/O depths of 16 and 32). As shown, the performance for four OSDs per NVMe was lower than for one OSD per NVMe at these high I/O depths. This reversal is due to the 50Gbps cluster networking beginning to saturate before reaching the performance limit of the NVMe SSDs.

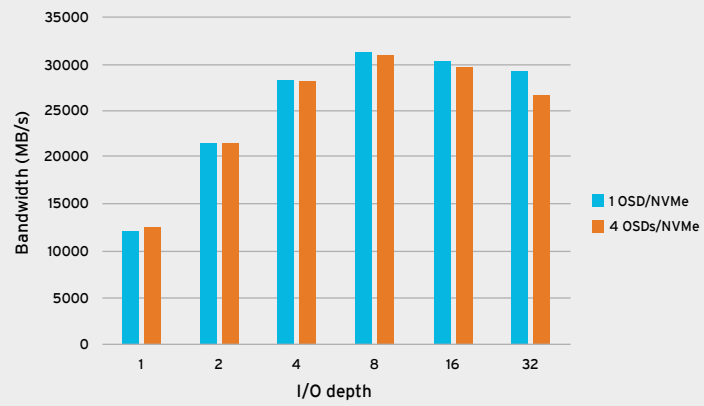

Figure 8. Large-block sequential read bandwidth, OSDs per NVMe comparison (60 libRBD clients, 1MB sequential reads).

For 1MB sequential writes, the Ceph cluster achieved 21 GB/s (Figure 9). In this case, using four OSDs per NVMe provided higher aggregate bandwidth than the default 1:1 mapping across the range of I/O depths. This configuration also delivers higher utilization of the NVMe SSD. As such, mapping four OSDs to each NVMe SSD is recommended for write-intensive large-block sequential workloads.

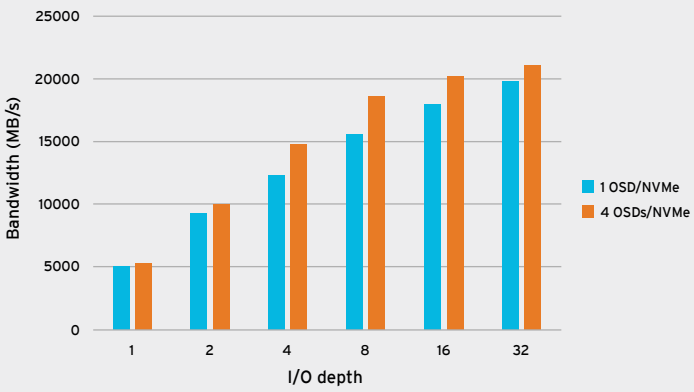

Figure 9. Large-block sequential write bandwidth, OSDs per NVMe comparison (60 libRBD clients, 1MB sequential writes).

INTEL CAS AND RED HAT CEPH STORAGE

Throughput-optimized performance can benefit from Intel Cache Acceleration Software (CAS). Intel CAS increases storage performance via intelligent caching and is designed to work with high-performance Intel SSDs.[9] Unlike inefficient conventional caching techniques that rely solely on data temperature, Intel CAS employs a classification-based solution that intelligently prioritizes I/O. This unique capability allows organizations to further optimize their performance based on I/O types (e.g., data versus metadata), size, and additional parameters.

The advantage of Intel's approach is that logical block-level storage volumes can be configured with multiple performance requirements in mind. For example, the file system journal could receive a different class of service than regular file data, allowing workloads to be better tuned for specific applications. Efficiencies can be increased incrementally as higher-performing Intel NVMe SSDs are used for data caching. Intel CAS is depicted logically in Figure 10, in the context of Ceph OSDs.

9 *Intel Cache Acceleration Software is licensed on a perpetual basis per SSD and includes one year of support at the time of purchase. To learn more, visit* intel.com/content/www/us/en/products/memory-storage/solid-state-drives.html.

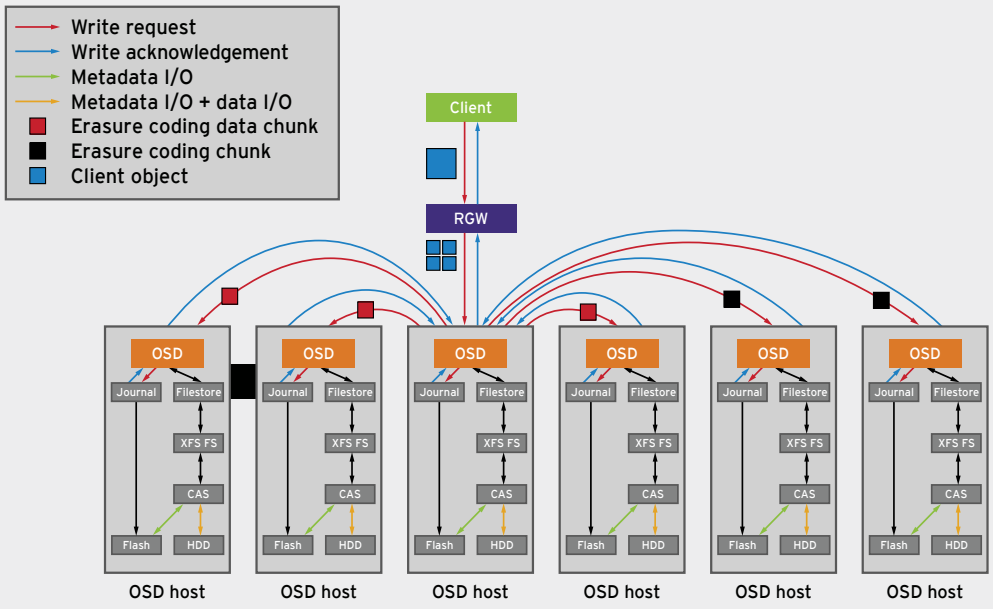

Figure 10. Intel CAS helps accelerate Ceph object stoarge workloads via the RGW interface.

Intel CAS also features several enhancements to improve deployment and usability for Red Hat Ceph Storage installations. The software uses a small default memory footprint that can be further reduced by using a feature called selective cache line size, offering potential cost savings benefits with higher-density storage servers. Caching mode can be changed on the fly, with immediate effect. Other features, such as in-flight upgrades, allow Intel CAS to be upgraded without stopping applications or rebooting servers, yielding further operational efficiencies.

CONCLUSION

Selecting the right hardware for target Ceph workloads can be a challenge, and this is especially true for software-defined storage solutions that run on industry-standard hardware. Because every environment differs, the general guidelines for sizing CPU, memory, and storage media per node in this document should be mapped to a preferred vendor's product portfolio for determining appropriate server hardware. Additionally, the guidelines and best practices highlighted in this document are not a substitute for running baseline benchmarks before going into production.

Coupled with innovative Intel SSDs Data Center, Intel Xeon processors power a wide range of industry-standard server platforms with diverse capabilities. Red Hat and Intel have conducted extensive testing with a number of vendors that supply hardware optimized for Ceph workloads. For specific information on selecting servers for running Red Hat Ceph Storage, detailed information including Red Hat Ceph Storage test results can be found in performance and sizing guides for popular hardware vendors.

RED HAT®
STORAGE

TECHNOLOGY DETAIL Red Hat Ceph Storage on Servers with Intel Processors and SSDs

Intel disclaimer: Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at www.intel.com.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more information go to www.intel.com/performance.

Benchmark results were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown". Implementation of these updates may make these results inapplicable to your device or system.

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.

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

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

redhat.com
#F11532_0418