RED HAT®
STORAGE

REFERENCE ARCHITECTURE

DEPLOYING RED HAT CEPH STORAGE CLUSTERS BASED ON SUPERMICRO STORAGE SERVERS

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 IOPS-intensive workloads on Ceph are also emerging. 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[1].

TABLE OF CONTENTS

Deploy cost-effective Ceph scale-out storage for block- or object-based applications on a tested and proven platform.

Optimize Ceph deployments for throughput or capacity, choosing the appropriate hardware for price/performance optimization.

Select from a wide range of Supermicro storage servers with the latest Intel Xeon processors and flexible storage and network options to optimize configurations.

Reduce risk by deploying Red Hat® Ceph Storage with support from the experts in Ceph technology.

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

redhat.com

[1] The results reported in this reference architecture are based on the configurations and specifications described. Your results may differ.

DOCUMENT PURPOSE

The purpose of this document is to characterize and compare the performance of Red Hat Ceph Storage on various Supermicro servers. Optimal Ceph cluster configurations are identified for general workload categories. As a reference architecture, this document provides details on cluster hardware, software, and network configuration, with performance results. The testing methodology is also provided and is based on the standardized Ceph Benchmarking Tool (CBT), available in a GitHub repository under the Ceph organization.[2] This particular study largely used off-the-shelf hardware and software components and did not study changing various configuration settings within the kernel, Ceph, XFS®, or the network in detail.

INTRODUCTION

To achieve the best cost-effective performance, different workload types require distinct approaches to storage infrastructure. For example, video on demand applications require throughput- and latency-optimized storage infrastructure, while an object archive might require capacity optimization. Relational database management system (RDBMS) workloads running on OpenStack clouds are an example of emerging IOPS-optimized workloads running on Ceph. Ceph lets organizations deliver object storage, block storage, and file systems through a unified and distributed cluster. Red Hat Ceph Storage provides enterprise support for object and block storage, and CephFS remains in active development within the Ceph community. To deploy efficient Ceph clusters, organizations need simple and tested cluster configurations, optimized for different workload types.

Ceph is extremely flexible, but deployments should be carefully designed for fault domain risk tolerance, depending on the needs of the application and the enterprise. Organizations need optimized configurations that allow scaling from a cluster measured in hundreds of terabytes to one measured in many petabytes. Red Hat's approach is to evaluate, test, and document reference configurations optimized for different workload categories, giving organizations specific and proven configuration advice. Figure 1 highlights a few of the key factors used to determine optimal configurations evaluated in this study.

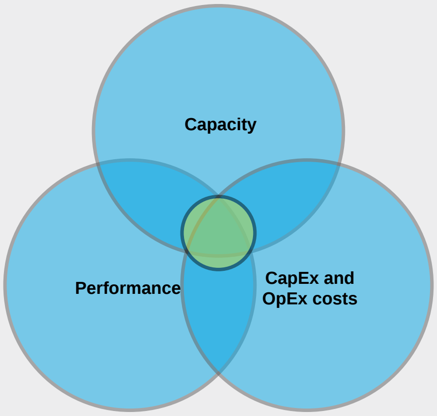

Figure 1. Different storage workloads require balancing factors such as performance and capacity as well as CapEx and OpEx costs.

[2] *https://github.com/ceph/cbt*

CHARACTERISTICS OF WORKLOAD-OPTIMIZED SCALE-OUT STORAGE CLUSTERS

One of the key benefits of Ceph storage is the ability to provision different types of storage pools within the same cluster, targeted for different workloads. This ability allows organizations to tailor storage infrastructure to their changing needs.

- Block storage pools typically use triple replication for data protection on throughput-optimized servers
- Object storage pools typically use erasure coding for data protection on capacity-optimized servers
- High-IOPS server pools can also be added to a Ceph cluster, as IOPS-optimized workloads emerge on Ceph.

Table 1 provides the criteria used to identify optimal Red Hat Ceph Storage cluster configurations on Supermicro storage servers. These categories are provided as general guidelines for hardware purchase and configuration decisions that can be adjusted to satisfy unique workload blends of different operators. As the workload mix varies from organization to organization, actual hardware configurations chosen will vary.

TABLE 1. CEPH CLUSTER OPTIMIZATION CRITERIA.

CLUSTER OPTIMIZATION CRITERIA	PROPERTIES	EXAMPLE USES
IOPS-OPTIMIZED	• Lowest cost per IOP • Highest IOPS • Meets minimum fault domain recommendation (single server is less than or equal to 10% of the cluster)	• Typically block storage • 3x replication • 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 for higher read throughput
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

STUDY CRITERIA AND FUTURE WORK

The reference architecture described in this document is the result of extensive testing by Supermicro and Red Hat to evaluate the performance of Red Hat Ceph Storage on Supermicro storage servers across a number of different configurations. The goals were to provide optimized, repeatable configurations for the throughput- and capacity-optimized criteria listed above. All-flash and IOPS-optimized configurations are an area of rapidly growing interest for Ceph.

The resulting optimized configurations can be used as starting points to build a range of cluster sizes, from hundreds of terabytes to multiple petabytes in size. Future work is anticipated in the following areas:

- Most of the test results were produced via the Reliable Automatic Distributed Object Store (RADOS) bench load test utility, which exercised the Ceph clusters at the lowest level using Ceph's native RADOS protocol. A more complete set of results would contain results from RADOS bench, FIO (RADOS Block Device (RBD) block level), and COS Bench (at the RADOS gateway (RGW) RESTful object level) for all configurations.
- The latency measurements in this study were average latencies. Assessing 95th or 99th percentile latency coupled with maximum latency would be more useful.
- The test run durations in this study were measured in minutes. A more complete study of steady-state operation would be provided by multihour test iterations.
- These performance results were generated under normal operating conditions. A more complete set of results would contain performance results captured while a cluster is in a degraded state.
- The largest cluster tested contained six Object Storage Device (OSD) nodes. Future work anticipates testing on 20-node or larger clusters.

CEPH DISTRIBUTED STORAGE ARCHITECTURE OVERVIEW

Storage infrastructure is undergoing tremendous change, particularly as organizations deploy infrastructure to support big data and private clouds. Traditional scale-up arrays are limited in scalability, complexity, and cost-effectiveness. Scale-out storage infrastructure based on clustered storage servers has emerged as a way to deploy cost-effective, manageable storage at scale, with Ceph among the leading solutions, as evidenced by the May 2015 OpenStack user survey.[3] In fact, cloud storage companies are already using Ceph at near exabyte scale. For example, Yahoo estimates that their Ceph-based Cloud Object Store will grow 20-25% annually.[4]

INTRODUCTION TO CEPH

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

- Store, retrieve, and replicate data.
- Monitor and report on cluster health.
- Redistribute data dynamically remap and backfill).
- Ensure data integrity (scrubbing).
- Recover from failures.

[3] *http://superuser.openstack.org/articles/openstack-users-share-how-their-deployments-stack-up*

[4] *http://yahooeng.tumblr.com/post/116391291701/yahoo-cloud-object-store-object-storage-at*

To the Ceph client interface that reads and writes data, a Ceph storage cluster looks like 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) both use the CRUSH (Controlled Replication Under Scalable Hashing) algorithm for object storage and retrieval.

CEPH ACCESS METHODS

All data in Ceph, regardless of data type, is stored in pools. The data itself is stored in the form of objects via the 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.

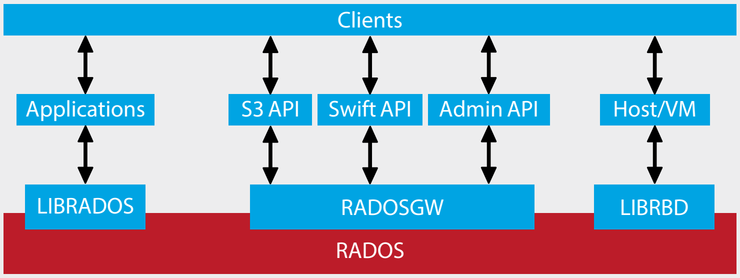

Figure 2. The Reliable Autonomic Distributed Object Store (RADOS) is the foundation of the Ceph storage cluster.

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:

- **RADOSGW.** A bucket-based object storage gateway service with S3-compliant and OpenStack Swift-compliant RESTful interfaces.
- **LIBRADOS**. A method providing direct access to RADOS with libraries for most programming languages, including C, C++, Java, Python, Ruby, and PHP.
- **RBD**. 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 backend, or user space libraries).

CEPH STORAGE POOLS

For a Ceph client, the storage cluster is very simple. When a Ceph client reads or writes data (referred to as an I/O context), it connects to a storage pool in the Ceph cluster. Figure 3 illustrates the overall Ceph architecture, with concepts that are described in the sections that follow.

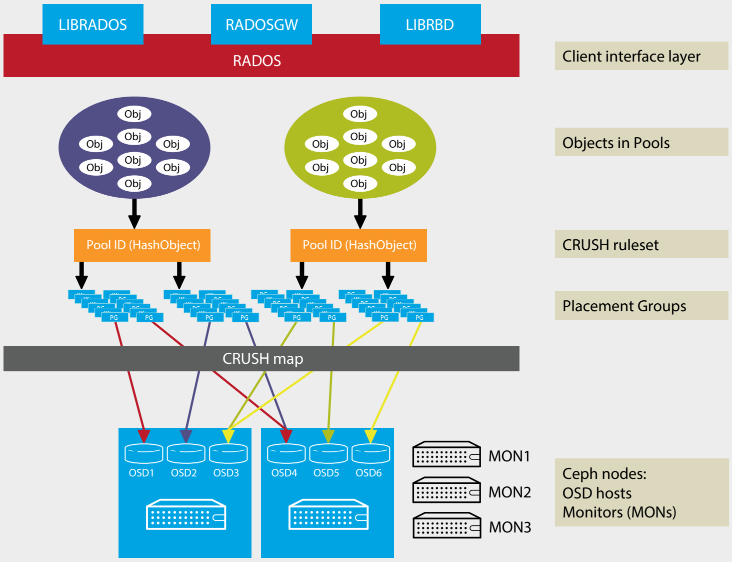

Figure 3. Clients write to Ceph Storage Pools while the CRUSH ruleset determines how Placement Groups are distributed across Object Storage Daemons.

- **Pools**. A Ceph storage cluster stores data objects in logical partitions called pools. Pools can be created for particular data types – such as block devices or object gateways – or simply to separate user groups. The Ceph pool dictates the number of object replicas and the number of placement groups (PGs) in the pool. Ceph storage pools can be either replicated or erasure coded, as appropriate for the application and cost model. Additionally, pools can be placed at any position in the CRUSH hierarchy, allowing them on groups of servers with differing performance characteristics.
- **Placement groups**. Ceph maps objects to placement groups. PGs are shards or fragments of a logical object pool that are composed of a group of Ceph OSD Daemons that are in a peering relationship. Placement groups allow the creation of replication or erasure coding groups of coarser granularity than on a per-object basis. A larger number of placement groups (e.g., 100 per OSD) leads to better balancing.
- **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 OSD Daemons**. In a Ceph cluster, Ceph OSD Daemons store data and handle data replication, recovery, backfilling, and rebalancing. They also provide some monitoring information to Ceph Monitors by checking other Ceph OSD Daemons with a heartbeat mechanism. A Ceph storage cluster requires at least two Ceph OSD Daemons (the default is three) to achieve an active and clean state when the cluster makes two copies of stored data. Ceph OSD Daemons roughly correspond to a file system on a physical hard disk drive.

- **Ceph Monitors (MONs)**. Before Ceph clients can read or write data, they must contact a Ceph Monitor (MON) to obtain the most recent copy of the 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 a cluster of monitors. Consensus among various monitor instances ensures consistent knowledge about the state of the cluster.

CEPH DATA PROTECTION METHODS

Applications have diverse needs for durability and availability, as well as different sensitivities to data loss. 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 defaults to making three copies of an object, with a minimum of two copies for clean write operations. If two of the three OSDs fail, the data will still be preserved, but write operations will be interrupted.

- **Erasure coding**. Erasure coding provides a single copy of data plus parity and is useful for archive storage and cost-effective durability. With erasure coding, storage pool objects are divided into chunks using the **n=k+m** notation, where k is the number of 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..

Typical Ceph read/write operations follow the steps below:

1. Ceph clients contact a Ceph Monitor to verify that they have an up-to-date version of the cluster map, and if not, retrieve the most recent version.

2. Data is converted into objects containing object/pool IDs.

3. The CRUSH algorithm determines the PG and primary OSD.

4. The client contacts the primary OSD directly to store/retrieve data.

5. The primary OSD performs a CRUSH lookup to determine the secondary PGs and OSDs.

6. In a replicated pool, the primary OSD copies the object(s) and sends them to the secondary OSDs.

7. In an erasure-coded pool, the primary OSD breaks up the object into chunks, generates parity chunks, and distributes the data and parity chunks to secondary OSDs, while storing one data chunk locally.

Figure 4 illustrates the relationships between Ceph storage pools, PGs, and OSDs for both replicated and erasure-coded pools.

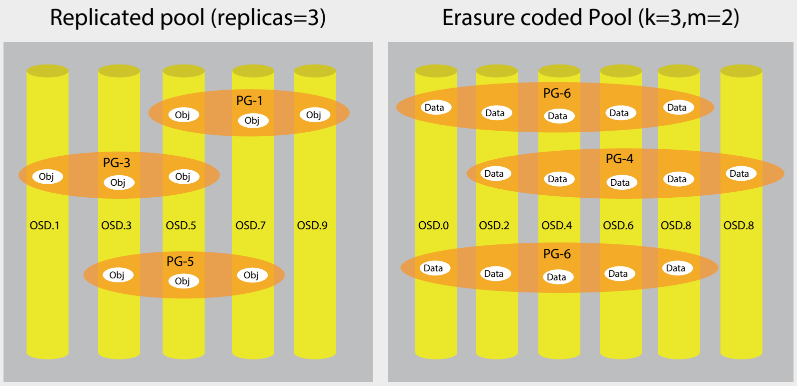

Figure 4. Graphical representation of relationships between Ceph Storage Pools, PGs, and OSDs, for both replicated pools and erasure coded pools.

REFERENCE ARCHITECTURE ELEMENTS

The following sections discuss the overall architecture of the solution, as well as key technical aspects of the principal components as they contribute to the reference architecture.

RED HAT CEPH STORAGE

Red Hat Ceph Storage provides a complete Ceph distribution with full support under subscription-based licensing. By providing block device storage and object gateway storage in a single solution, Red Hat Ceph Storage can be integrated easily into existing infrastructure. Red Hat Ceph Storage offers robust, multi-tenant storage for cloud and virtualization platforms such as Red Hat Enterprise Linux OpenStack Platform and provides an Amazon Web Services (AWS) S3 interface. Red Hat Ceph Storage offers distinct advantages that asdfasdfsdfa include:

- **Value**. With low data storage costs and enterprise-class support, Red Hat Ceph Storage lays a foundation for managing exponential data growth at a low cost per gigabyte.
- **Longevity**. Organizations can start with block storage and transition into objects storage, or vice versa. Ceph provides seamless access to objects either with native language bindings or through the RADOS Gateway, a RESTful interface that is compatible with many applications written for S3 and Swift. The Ceph RBD provides access to block device images that are striped and replicated across the entire storage cluster.
- **Enterprise-readiness**. Red Hat Ceph Storage integrates tightly with OpenStack to provide the block storage capabilities of a traditional block storage device with added hardware flexibility, massive scalability, and fault tolerance capabilities.
- **Industry-leading expertise**. Red Hat Ceph Storage is backed by the experience and expertise of Ceph's creators and primary sponsors, as well as engineers with decades of experience with the Linux kernel and filesystems. Red Hat also offers hands-on training and professional services.

The benchmarking tests described in this reference architecture were performed with Red Hat Ceph Storage 1.2.2, running a 3.10.x kernel on Red Hat Enterprise Linux 7.1.

SUPERMICRO STORAGE SERVERS

Supermicro storage servers use components available in workload optimized form factors from one to four rack units (1U to 4U). These solutions offer high storage density coupled with up to 96% power efficiency and provide advantages in both procurement and operational costs for deployments of all sizes. The Supermicro configurations below were the result of testing with different quantities of disk drives and flash media devices for OSD journaling (shown as x disk drives + y flash media devices). Supermicro storage servers optimized for Ceph storage infrastructure feature the latest Intel Xeon E5-2600 CPUs and include:

- **Supermicro SYS-6018R-MON2.** This server is a 1U, four-bay, dual processor server with dual Small Form Factor Pluggable Plus (SFP+) 10 Gigabit Ethernet interfaces. The general-purpose server is populated with sufficient processing power to serve in either a cluster monitor or an administrative role. Supermicro MON hardware configurations also work well for RADOS gateways. As mentioned, a minimum of three of these servers are required to form a Ceph cluster quorum.
- **Supermicro SSG-6028R-OSD072P.** This 2U, 12-bay server features 6TB 7200 RPM SATA hard disk drives (HDDs) and a single NVM Express (NVMe) flash device (12+1). The flexible 72TB server is a great choice for throughput-optimized configurations with a single processor installed. A second processor could be installed for co-locating Ceph monitors in smaller environments. This server provides an optimized component for high-throughput needs.
- **Supermicro SYS-F618H-OSD288P.** his enclosure offers four nodes and 288TB of raw storage in a single 4U chassis, delivering increased power and cooling efficiency compared to four standalone servers. It is ideal for high-density, throughput-optimized configurations. Four 1U, 12-bay nodes are fully populated with 6TB 7200 RPM SATA drives and a single NVMe device per node ((12+1)x4). This server is optimal for high-throughput, high-density deployments as well as lab or proof of concept (POC) environments.
- **Supermicro SSG-6048R-OSD216.** This storage server is a capacity-optimized system for large object sequential applications, such as backup and archive. Under load, it has enough hard drives to saturate the 10 Gigabit Ethernet client interface. The 4U server contains 36 drive bays equipped with 6TB 7200 RPM SATA drives (36+0).
- **Supermicro SSG-6048R-OSD216P.** This server is a capacity-optimized system with solid-state drive (SSD) journaling to improve I/O performance under mixed workloads. The 4U server features 36 drive bays with 6TB 7200 RPM SATA drives and two NVMe SSDs (36+2).
- **Supermicro SSG-6048R-OSD432.** This 432TB capacity-optimized system is ideal for multi-petabyte use cases with less frequently accessed data, such as backup and archive. The 4U server features 72 drive bays for 6TB 7200 RPM SATA drives (72+0).
- **Supermicro SSG-6048R-OSD360P.** This server is a 360TB capacity-optimized system with SSD journaling to improve I/O performance under mixed workloads. The 4U server features 72 drive bays with 60 6TB 7200 SATA drives and 12 SATA SSDs for journaling (60+12).

Server combinations both with and without flash acceleration for journaling were tested as a part of the reference architecture, in the following configurations:

- 12 HDDs, zero flash devices (12+0). Single Intel Xeon E5-2630v2 CPU, 64 GB memory, 12 HDDs, and two 10 Gigabit Ethernet interfaces (SFP+).

- 12 HDDs, one flash device (12+1). Single Intel Xeon E5-2630v2 CPU, 128 GB memory, 12 HDDs, one NVMe, and two 10 Gigabit Ethernet interfaces (SFP+)
- 18 HDDs, zero flash devices (18+0). Single Intel Xeon E5-2630v2 CPU, 128 GB memory, 18 HDDs, and two 10 Gigabit Ethernet interfaces (SFP+)
- 18 HDDs, one flash device (12+1). Single Intel Xeon E5-2630v2 CPU, 128 GB memory, 18 HDDs, one NVMe, and two 10 Gigabit Ethernet interfaces (SFP+)
- 36 HDDs, zero flash devices (36+0). Dual Intel Xeon E5-2630v2 CPUs, 128 GB memory, 36 HDDs, and two 10 Gigabit Ethernet interfaces (SFP+)
- 36 HDDs, 2 flash devices (36+2). Dual Intel Xeon E5-2630v2 CPU, 128 GB memory, 36 HDDs, two NVMe flash drives, and two 10 Gigabit Ethernet interfaces (SFP+) (also tested with a single shared Mellanox 40 Gigabit Ethernet interface)
- 72 HDDs, zero flash devices (72+0). Dual Intel Xeon E5-2697v2 CPUs, 256 GB memory, 72 HDDs, and one Mellanox 40 Gigabit Ethernet interface
- 60 HDDs, 12 flash devices (60+12). Dual Intel Xeon E5-2697v2 CPUs, 256 GB memory, 60 HDDs, 12 SSDs, and one Mellanox 40 Gigabit Ethernet interface

STORAGE MEDIA

Performance and economics for Ceph clusters both depend heavily on an effective choice of storage media. In general, rotating media account for the bulk of the deployed storage capacity, and must be selected for their ability to function well in enterprise and dense datacenter environments. For throughput-optimized configurations, solid-state storage media are typically used for Ceph journaling to accelerate writes. For capacity-optimized configurations, write journaling is co-resident on the HDDs.

Seagate HDD storage media

Enterprise-, cloud-, and archive-class HDDs should be used for Ceph clusters. Desktop-class disk drives are typically not suited for Ceph deployments, as they lack sufficient rotational vibration compensation for high density, high duty-cycle applications and use cases. When dozens (or hundreds) of rotating HDDs are installed in close proximity, rotational vibration 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 four to six HDDs in a single enclosure. Both SAS and SATA interface types are acceptable. The RV Compensation circuitry and media may be similar in both SAS and SATA enterprise- or cloud-class HDDs.

The choice of interface, however, can make a difference in many use cases and applications. SAS Near Line HDDs have dual SAS 12 GB/s ports and are typically higher performing than single-ported 6 GB/s SATA interface HDDs. Dual SAS ports offer redundant connectivity all the way from controller to disk drive—and can also allow simultaneous reading and writing. Additionally, SAS devices generally have much lower unrecoverable read error (URE) rates, sometimes by an order of magnitude. Lower URE rates result in fewer scrubbing errors, translating to fewer placement group repair operations.

The rotating media used in this study are Seagate ST6000MN0034 6TB SAS Near-Line HDDs.[5] These drives were chosen for this study for the following reasons:

- They are able to store up to 6TB of data without sacrificing performance.
- Fast random and sequential read/write performance means that they can quickly access and store data.
- Eighth-generation drive technology provides reliable access to bulk storage of data via the dual-port 12 Gb/s SAS interfaces.
- Rotational vibration tolerance and compensation help ensure consistent performance and reliability in servers with high spindle density.
- Both 12 Gb/s SAS dual-port (used in these tests) and 6 Gb/s SATA interface options are available.

Solid state storage media for Ceph write journals

Ceph is strongly consistent storage, so every write to the Ceph cluster must be written to Ceph journals before the write is acknowledged to the client. The data remain in the journal until all copies are acknowledged as fully written (with three replicas being typical). Only then will the next write happen. With SSD journals, the secondary OSDs are able to write their bits 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. File systems like XFS are able to use internal heuristics to best effect when they are the sole consumer of a media device. Moving journals to separate media ensures the highest efficacy of these heuristics.

Two solid state storage media types were evaluated in testing performed by Red Hat and Supermicro:

- **SSDs** installed in the drive bays of the Supermicro storage servers, typically a ratio of one SSD to each five HDDs (1:5 ratio). Either SAS or SATA SSDs can be used.
- **PCIe or NVMe flash devices** installed directly in the OSD storage server PCIe slots instead of in the drive bays, typically in a ratio of one SSD to each 12-18 HDDs (1:12-18 ratio). This approach allows all the drive bays to be populated with disks for data storage, increasing the capacity of each server. In addition to saving space, PCIe devices do not have intermediating components, like SAS expanders and host bus adapters, that can increase latencies.

Certain important criteria must be considered when selecting solid state media for scale-out, software-defined storage clusters using Ceph. These criteria apply whether SSDs or PCIe/NVMe-based solid state storage is chosen, and include:

- **Endurance**. Write endurance is important as Ceph write journals are heavily used and could wear out an SSD of lower endurance.
- **Power fail protection**. Super capacitors for power fail protection are vital. In the event of a power failure, super capacitors must be properly sized to allow the drive to persist to non-volatile storage data that is still residing in caches.
- **Performance**. uper capacitors for power fail protection are vital. In the event of a power failure, super capacitors must be properly sized to allow the drive to persist to non-volatile storage data that is still residing in caches.

[5] *In one configuration, the HGST SATA 3TB HDD was used, as it was also Supermicro/expander qualified and used in 10-node tests with the 4U 36-bay enclosures.*

The SAS SSDs used in the drive bays of the Supermicro storage server test configurations were the Seagate ST200FM0053 200GB dual-port 12 Gb/s SAS SSD. The Seagate SSD is designed for storage array use cases with write-intensive workloads typical of 24x7 datacenters. The Seagate SSDs provide:

- Random read/write performance of 110K / 40K IOPS
- Sequential read/write performance of 750 / 500 MB/s

Although not used in this particular set of benchmarks and tests, the Intel DC S3700 Series SSD has been previously qualified for Ceph journal solid state media use cases with adequate endurance and power fail protection. The Intel SSD offers

- Random read/write performance of 75K / 32-36K IOPS
- Sequential read/write performance of 500 / 365-460 MB/s

In addition to SSDs, PCIe and NVMe based solid state storage was used for Ceph OSD Daemon journaling for the reference architecture. NVMe and PCIe flash cards can both install in the PCIe bus slots of Supermicro storage servers. Red Hat and Supermicro observed similar performance and capabilities between NVMe and PCIe cards, since both use the higher performance, lower latency PCIe bus to pass data.

Intel DC P3700 NVMe flash SSD and the Seagate XP6209 PCIe flash SSD were both used in OSD server configurations for Ceph journaling. Both devices offer super capacitor power failure protection intended to protect Ceph OSD Daemon journal data. Both devices also claim to offer enterprise-class datacenter use case write endurance.

Testing showed that using in-server solid state flash for Ceph journals offers the following advantages compared to SSDs (disk drive form factor) installed in the server's drive bays:

- The NVMe/PCIe approach yielded an average 20% higher cluster capacity when compared with drive form factor SAS/SATA SSD installed in drive bays.
- The NVMe/PCIe typically provided lower latency and higher performance on each journal write via PCIe data bus, because the PCIe bus is much simpler, lacking the complexities of SAS/SATA topographies.
- The PCIe and NVMe flash SSD storage media transfer data to and from the processor with up to 6x better throughput compared to SAS/SATA SSDs.
- Configurations of 36 HDDs with two PCIe SSDs for Ceph OSD Daemon journal data stores (36+2) performed favorably compared to earlier Supermicro testing of 30 HDDs with six SAS/SATA SSDs in drive bays (30+6).

For more information on using solid state technology for Ceph write journals, see the report "Ceph PCIe SSD Performance Part 1: OSD Journals", written in cooperation with Seagate.[6]

Though not explored in this study, PCIe/NVMe flash SSD storage media can also be used for high performance storage pools and the leveldb stores for Ceph monitors. In fact, SSDs should always be included in monitor configurations.

[6] *https://www.redhat.com/en/resources/ceph-pcie-ssd-performance-part-1*

MELLANOX NETWORKING

The Mellanox ConnectX-3 EN 40/56 Gigabit Ethernet Network Interface Cards (NIC) used in Red Hat and Supermicro testing deliver high-bandwidth, low latency, and industry-leading Ethernet connectivity for performance-driven server and storage applications. Available in 10-gigabit or 40/56-gigabit speeds and in single- or dual-port configurations, these NICs offer high throughput, low latency, and low power consumption. In addition the ConnectX-4 adapter family supports Ethernet connectivity at 10, 25, 40, 50, 56, and 100 gigabit speeds.

Mellanox SwitchX-2 switch systems deliver high-performing top-of-rack (ToR) solutions with non-blocking throughput from 12 to 64 ports. Offering some of the lowest latency and power consumption in the industry, Mellanox Ethernet switches help accelerate application performance and are ideal for connecting pods or clusters of servers and storage nodes. The Mellanox SX1012 used in this reference architecture is a half-width 1U high-form-factor Ethernet switch that delivers up to 1.3 Tb/s of non-blocking throughput with up to 12 40/56 Gigabit Ethernet ports. Other Mellanox switch models support 10, 25, 50, and 100 Gigabit Ethernet line speeds.

OPERATIONAL PLANNING CONSIDERATIONS

The following sections provide general guidance on operational planning for deploying Red Hat Ceph Storage on Supermicro storage servers.

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 scale-out storage upon which they depend. The fault domain that a single OSD server represents is most important, so dense servers should be reserved for multi-petabyte clusters where the capacity of an individual server is less than 10-15% of the total cluster capacity. This recommendation may be relaxed for less critical pilot projects.

There are three primary factors for weighing fault domain risks

- **Reserved capacity for self healing.** When a storage node fails, Ceph self-healing begins after a configured time period. The unused storage capacity of the surviving cluster nodes must be greater than the used capacity of the failed server for successful self-healing. 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 from algorithmic placement). As a result, each node in a cluster should operate at less than 80% of total capacity.
- **Impact to performance.** During self-healing, a percentage of cluster throughput capacity will be diverted to recreating object copies from the failed node on the surviving nodes. The percentage of cluster performance degradation depends on the number of nodes in the cluster and how Ceph is configured.
- **Declustering of replicas and erasure coded chunks across hosts.** When configuring CRUSH, it is important to use `ruleset-failure-domain=host` to ensure that loss of a single node does not result in unavailability of multiple OSDs in any given placement group. For instance, in order for a 3x replicated pool to tolerate concurrent unavailability of two hosts (or even a single host and a failed disk in another host) would require a minimum of four hosts. The number of required hosts is greater for erasure-coded pools, where the sum of m+k must be less than the number of hosts

in the system. For instance, in order for a m=6, k=2 erasure-coded pool to tolerate concurrent unavailability of two hosts (or even a single host and a failed disk in another host) would require a minimum of 10 hosts.

Red Hat and Supermicro recommend the following minimum cluster sizes:

- **Supported minimum cluster size:** Three storage (OSD) servers, suitable for use cases with risk tolerance for availability
- **Recommended minimum cluster size:** 10 storage (OSD) servers

POWER, COOLING, AND RACK DENSITY

Supermicro servers feature 80PLUS Platinum (95%+ efficiency) and 80PLUS Titanium (96%+ efficiency) certified power supplies, assuring that a minimum amount of power is lost as heat when converting AC to DC. CPU, memory, and hard disk drives have also improved thermal performance compared to previous generations. However, these improvements have resulted in higher density server hardware with greater overall power and thermal density at the rack level.

Power, cooling, and rack density are interrelated attributes when selecting server form factors. Instead of requiring the highest density server hardware available, data center infrastructure will often dictate the best value and overall total cost of ownership (TCO.) Table 2 below provides some high-level guidance when considering Supermicro hardware configurations for Ceph.

TABLE 2. MOST COMMON RACK POWER BUDGET, AND CORRESPONDING SUPERMICRO CEPH OPTIMIZED CONFIGURATIONS.

SUPERMICRO MODELS (NODES)	5 KVA	8.5 KVA	10 KVA	16 KVA
2 RU / 12-BAY SERVERS				
SSG-6028R-OSD072 SSG-6028R-OSD072P	• 13 nodes • 26 RU • 156 HDDs	• 20 nodes • 40 RU • 240 HDDs		
4 RU / 36-BAY SERVERS				
SSG-6048R-OSD216 SSG-6048R-OSD216P	• 5 nodes • 20 RU • 180 HDDs	• 9 nodes • 36 RU • 324 HDDs		
4 RU / 72-BAY SERVERS				
SSG-6048R-OSD432	• 3 nodes • 12 RU • 216 HDDs	• 5 nodes • 20 RU • 360 HDDs	• 6 nodes • 24 RU • 432 HDDs	• 9 nodes • 36 RU • 648 HDDs
SSG6048R-OSD360P	• 3 nodes • 12 RU • 180 HDDs	• 5 nodes • 20 RU • 300 HDDs	• 6 nodes • 24 RU • 360 HDDs	• 9 nodes • 36 RU • 540 HDDs

As shown in the table, the deployment of 2 RU/12-bay Supermicro throughput-optimized configurations and the 4 RU/36-bay capacity-optimized configurations are a good fit for the most common rack power envelope for existing co-location facilities (8.5 KVA). The higher density server configurations (1 RU/12-bay and 4 RU/72-bay) will require more power to fill 42 RU of rack space. These configurations achieve optimal rack density when deploying to modern co-location facilities with greater power and cooling infrastructure.

Table 3 lists the standard optimized configurations for deploying Ceph on Supermicro hardware, informed by Red Hat and Supermicro lab testing. Rack level configurations are also available, providing rapidly deployable datacenter components tuned for Ceph.

TABLE 3. COMMON OPTIMIZED CONFIGURATIONS FOR DEPLOYING CEPH ON SUPERMICRO HARDWARE

X10 MODELS	CPU	DRIVE CONFIG	X9 MODELS	CPU	DRIVE CONFIG
SSG-F618H-OSD288P	Single Intel Xeon E5-2620v3	(12+1) 12x 6 TB HDD + 1x NVMe	NA	NA	NA
SSG-6028R-OSD072	Single Intel Xeon E5-2620v3	(12+0) 12x 6 TB HDD	NA	NA	NA
SSG-6028R-OSD072P	Single Intel Xeon E5-2620v3	(12+1) 12x 6 TB HDD + 1x NVMe	SSG-6027R-OSD040H	Single Intel Xeon E5-2630v2	(10+2) 10x 4 TB HDD + 2x SSD
SSG-6048R-OSD216	Dual Intel Xeon E5-2630v3	(36+0) 36x 6 TB HDD	NA	NA	NA
SSG-6048R-OSD216P	Dual Intel Xeon E5-2630v3	(36+2) 36x 6 TB HDD +2x NVMe	SSG-6047R-OSD120H	Dual Intel Xeon E5-2630v2	(30+6) 30x 4 TB HDD + 6x SSD
SSG-6048R-OSD432	Dual Intel Xeon E5-2690v3	(72+0) 72x 6 TB HDD	NA	NA	NA
SSG-6048R-OSD360P	Dual Intel Xeon E5-2690v3	(60+12) 60x 6 TB HDD + 12x SSD	SSG-6047R-OSD320H	Dual Intel Xeon E5-2670v2 (E5-2697 recommended)	(60+12) 60x 4 TB HDD + 12x SSD

OPTIMIZED CONFIGURATIONS

Red Hat and Supermicro realize that every organization is different. This section provides general sizing guidelines, as well as specific configurations that can be extended and scaled as required.

GENERAL SIZING AND CAPACITY GUIDELINES

When sizing a Ceph cluster, organizations should take likely workloads and how the cluster will be used into account. Similar to storage and server sizing, cluster sizing is also sensitive to workload. Throughput-optimized and capacity-optimized clusters have distinct sizing requirements:

- **Throughput-optimized configurations**. Servers with 12-16 3.5-inch HDDs with a single PCIe SSD or NVMe generally provide the most throughput per unit of cost (CapEx). Current results show that the amount of Ceph throughput per drive is lower on dense servers (defined as greater than 16 drives) than on sparse servers. However, other factors may lead organizations to deploy more dense servers, such as TB per rack unit cost when datacenter square footage is at a premium. It is important to remember that read throughput-optimized configurations typically run on replicated storage pools. Three replicas are typical, yielding 33% of total capacity usable. For write throughput optimization, erasure coding often provides more throughput.
- **Capacity-optimized configurations**. The primary factors affecting choice of server size for capacity-optimized configurations are cluster size, cost per TB (CAPEX), and corresponding failure domain risk tolerance (see above). Dense servers are best deployed in clusters with multiple petabytes. Capacity-optimized configurations typically use erasure-coded pools, yielding 60-80% of capacity usable, depending on the erasure-coded parameters chosen.

OSD throughput

A variety of benchmarks from independent sources provide some background for estimating the appropriate number of disk drives for various configurations. A baseline range for OSD read throughput of 50-75 MB/sec is typical (assuming sequential 4MB reads on 7200rpm SATA drives). This throughput equates to 33% of an HDD's published sustained throughput specification. This baseline throughput per OSD provides a simple means for estimating the number of OSDs per server that would saturate a server's network bandwidth. In this study, however, it is important to note that throughput per OSD/HDD declined as the quantity of OSD/HDDs per server increased.

SSD write journals

For IOPS- and throughput-optimized configurations, SSD write journals typically increase performance. For cost/capacity-optimized configurations that are not latency sensitive, SSD write journals are unlikely to be cost effective and can be eliminated (write journaling would be co-located on the HDDs in this case).

- SSDs used for OSD write journaling should be enterprise-grade, provide power loss protection, and demonstrate good write endurance (matching the anticipated server refresh cycle when used as an OSD write journal). For instance, if the server refresh cycle is three years, architects should ensure that a SATA/SAS SSD drive will last three years with an anticipated 30 MB/s average write rate per OSD Ceph workload and a 5:1 HDD:SSD ratio.
- Systems should be configured for performance in the recommended HDD:SSD ratio. Generally, the write throughput of the SSD should be greater than or equal to the aggregate write capacity of the HDDs journaling to that SSD. Typical configurations deploy four to five OSDs to a single SATA SSD providing 300 MB/s write throughput. A configuration could alternately deploy 12-18 OSDs with a single PCIe SSD providing 1900 MB/sec write throughput.

CPU and memory

CPU processor count, core count, and frequency must be selected to ensure adequate processing power for OSD daemons. It is generally recommended that each OSD have a minimum of one CPU-core-GHz. For a workable system, the following equation should remain greater or equal to 1.

((CPU sockets * cores * CPU clock speed) / OSDs) >= 1

For example, the following equation demonstrates a workable system:

2 sockets * 8 cores * 2.6 Ghz = 41.6 / 30 OSDs = 1.39

For dense servers (greater than 18 OSDs per server), Red Hat and Supermicro are currently studying the ratio of cores (with and without hyperthreading) and plan to publish the results when complete. Two gigabytes of memory is typically added to each OSD server for each OSD. The minimum recommended memory capacity for an OSD server is 32GB.

Networking

For performance-optimized clusters using 10 Gigabit Ethernet, Ceph documentation recommends separate networks for communication between Ceph clients and Ceph storage servers (client-facing network), and between different Ceph storage servers (cluster-facing). For performance-optimized clusters using 40 Gigabit Ethernet, either a single shared network or dual networks may be deployed, depending on throughput requirements. Additional out-of-band networks are recommended for management and the intelligent platform management interface (IPMI).

Current tests on most HDD-based servers show that dual 10 Gigabit Ethernet networks (one client-facing, one cluster-facing) are sufficient to accommodate Ceph throughput. As per-drive throughput has tested lower on dense servers, many of these servers are unable to take full advantage of network links greater than 10 Gb/s, although some dense servers have shown the ability to do so. Tests with 40 Gigabit Ethernet networks have shown significantly lower latencies. Preliminary results from all-flash Ceph servers illustrate their ability to consume network bandwidth well beyond 10 Gigabit Ethernet.

THROUGHPUT-OPTIMIZED CONFIGURATIONS

To evaluate throughput-optimized configurations, Red Hat tested three throughput-optimized configurations:

- 36-bay servers with dual 10 Gigabit Ethernet networking
- 36-bay servers with 40 Gigabit Ethernet networking
- Single-socket 12-bay servers with dual 10 Gigabit Ethernet networking[7]

Due to hardware availability, the 36-drive bay servers were also used for 12- and 18-bay testing using Intel Xeon E5-2630v2 processors. For testing these smaller configurations, one CPU was disabled together with 64MB or 128MB of RAM, while subsets of drive bays were used as shown in Figure 5.

The 36-drive bay storage servers were configured for throughput. Each system was configured with 36 HDDs and two Intel P3700 PCIe NVMe devices, with the latter used for OSD write journaling.

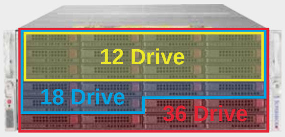

Figure 5. 36-drive bay servers were used for 12-drive, 18-drive, and 36-drive testing.

Figure 6 illustrates the topology of the five OSD node test bed, configured as follows:

- Three monitor nodes were deployed with the configuration, with one doubling as a management node.
- Each of the five OSD nodes was configured with 36 3.5-inch drive bays (tested as 12-, 18-, and 36-drive configurations).
- Eight client nodes were configured.
- 10 Gigabit Ethernet networking was employed for client-facing and cluster-facing networks in one configuration, with a separate configuration employing a single shared 40 Gigabit Ethernet network.
- OSD write journaling was performed on PCIe flash media devices.

[7] *The 36-bay server chassis were also used for 12-bay testing by disabling one CPU, 64MB or 128MB of RAM, and 24 HDDs.*

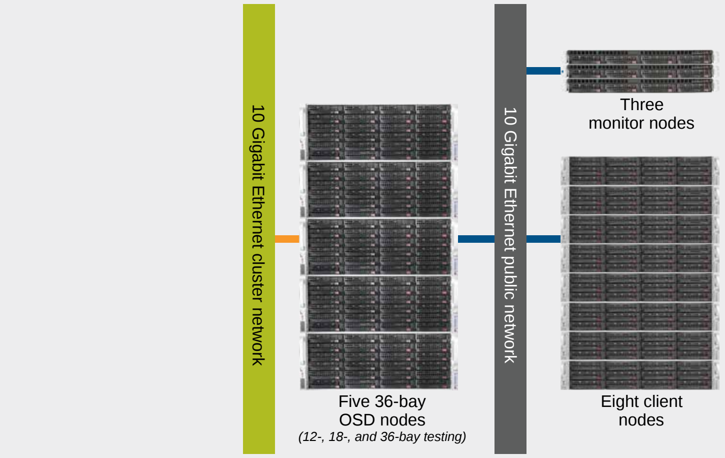
10 Gigabit Ethernet cluster network
10 Gigabit Ethernet public network
Three monitor nodes
Five 36-bay OSD nodes
(12-, 18-, and 36-bay testing)
Eight client nodes

Figure 6.A throughput-optimized configuration with five OSD node, with each server configured 36+2, utilizing 12-, 18-, and 36-drive configurations.

Throughput-optimized configurations can be scaled easily. Figure 7 illustrates a 10-node throughput-optimized configuration, accomplished by simply adding additional OSD nodes.

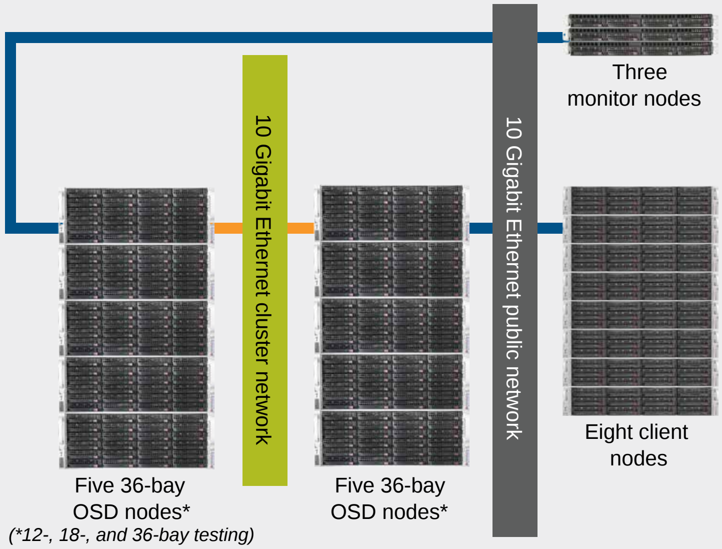
10 Gigabit Ethernet cluster network
10 Gigabit Ethernet public network
Three monitor nodes
Eight client nodes
Five 36-bay OSD nodes*
(*12-, 18-, and 36-bay testing)
Five 36-bay OSD nodes*

Figure 7. 10-node throughput-optimized configuration, with each node configured 36+2 utilizing 12-, 18-, and 36-drive configurations.

The component configuration for each 36-bay OSD node is shown in Table 4.

TABLE 4. SYSTEM DETAILS FOR 36-BAY THROUGHPUT-OPTIMIZED OSD NODES.

	COMPONENTS	CONFIGURATION DETAILS
NETWORKING	• Intel 82599ES 10 Gigabit Ethernet Controller • Intel 82599ES 10 Gigabit Ethernet Controller	• Single port used for public network • Single port used for cluster network
PROCESSOR	• Intel Xeon processor E5-2630 v2	• Two per server
MEMORY	• 128 GB DDR3 1333 MHz	• Eight slots used
STORAGE CONTROLLER	• LSI Logic SAS2308	• SAS-2 controller (x2)
FLASH STORAGE	• Intel 800 GB PCIe SSD DC P3700 NVMe device	• NVMe journal devices (x2)
HARD DISK DRIVES	• HGST 3 TB 7200 RPM SATA	• OSD data disk drives (x36)

CAPACITY-OPTIMIZED CONFIGURATIONS

Capacity-optimized Ceph workload testing was performed with larger 72-drive bay OSD nodes, in both three-node and single-node configurations.

High-capacity testing

High capacity testing was performed using a total of three 72-bay OSD nodes. Single-node baseline testing was performed on two CPU configurations (Ivy Bridge). Up to four clients were connected to allow for incremental loading, and three monitor nodes were deployed for redundancy. The networking component employed 40 Gigabit Ethernet cards from Mellanox accessing a common network for public and cluster traffic. Two 72-bay configurations were tested, 60+12 and 72+0. Three LSI host bus adapters were used to connect the HDDs and SSDs. In the 60+12 configuration, 12 Seagate SSDs were used for journaling. The components used are listed in Table 4. An architectural diagram of the three-node configuration is shown in Figure 8.

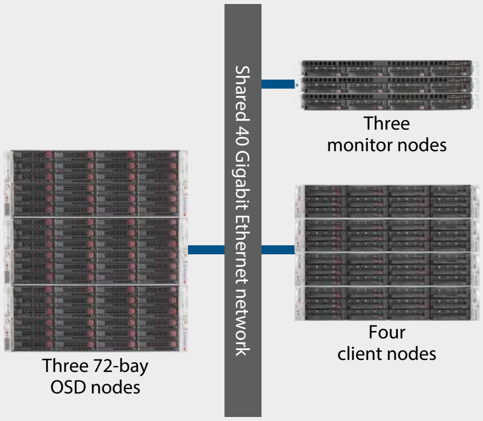

Figure 8. Three-node capacity-optimized storage server configuration, with each 72-bay server configured 60+12.

Single-node testing

Single node testing was also conducted for reference. The configuration is shown in Figure 9.

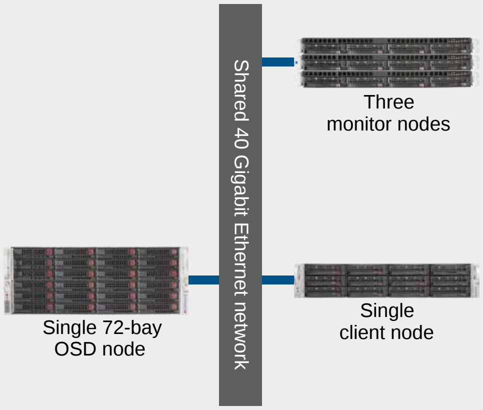

Figure 9. Single-node capacity optimized configuration also included 12-, 18-, and 36-bay testing.

The component configuration for each 72-bay OSD node is shown in Table 5.

TABLE 5. SYSTEM DETAILS FOR 72-BAY CAPACITY-OPTIMIZED OSD NODES.

	COMPONENTS	CONFIGURATION DETAILS
NETWORKING	• Mellanox ConnectX-3 MCX314A 40 Gigabit Ethernet card	• Single port used for public network (X9 and X10)
PROCESSOR	• Intel Xeon E5-2697 v2 @ 2.7 GHz	• X9 only
MEMORY	• 256 GB DDR3 1333 MHz	• X9 and X10
STORAGE CONTROLLER	• LSI Logic SAS2308 (x3)	• X9 only
FLASH STORAGE	• Seagate ST200FM0053 200 GB	• SSD journals x12 (X9 and X10)
HARD DISK DRIVES	• Seagate ST6000MN0034 7200 RPM SAS	• OSD data disk drives x60 (X9 and X10)

BENCHMARK RESULTS

To characterize performance, Red Hat and Supermicro ran a series of benchmarks across different cluster configurations, varying the servers involved, the number of spinning and solid state storage devices, data protection schemes, and benchmark workloads.

CEPH BENCHMARK TOOL

All testing in this reference architecture was conducted using the Ceph Benchmark Tool (CBT). CBT is a Python tool for building a Ceph cluster and running benchmarks against it. As a part of Red Hat testing, CBT was used to evaluate the performance of actual configurations, as described in the following sections. CBT automates key tasks such as Ceph cluster creation and tear-down and also provides the test harness for automating various load-test utilities such as rados bench.

Ceph includes the rados bench load-test utility, designed specifically to benchmark a RADOS storage cluster. Testing involves creating a pool of the desired configuration, and then performing operations (writes and reads) against it as desired. Both sequential and random reads and writes can be evaluated. As described in the sections that follow, rados bench was used to test a wide range of configurations with Red Hat Ceph Storage deployed on Supermicro storage servers.[8]

SEQUENTIAL READS

Sequential read performance and latency were evaluated both on a per-server basis, and on a per-OSD basis across a broad variety of cluster configurations with varying server configurations and networks. Both replicated and erasure-coded pools were tested.

Sequential read throughput per server

Figure 10 illustrates sequential read throughput (in MB/s) at the librados level on a per server basis using the rados bench command. This test used a replicated pool with three replicas. Servers accommodating 12 to 72 bays were tested. Separate 10 Gigabit networks and shared 40 Gigabit networks were both evaluated.

[8] *For general information on Ceph benchmarking, see https://wiki.ceph.com/Guides/How_To/Benchmark_Ceph_Cluster_Performance*

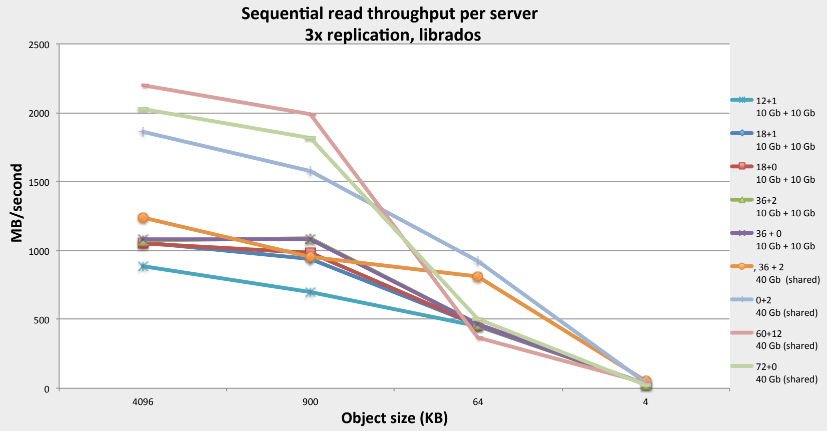

Figure 10. Sequential read throughput per server at the librados level with 3x replicated pools for tested Supermicro server configurations.

Sequential read latency at the librados level for the same set of servers is shown in Figure 11. Again, a replicated pool with three replicas was used.

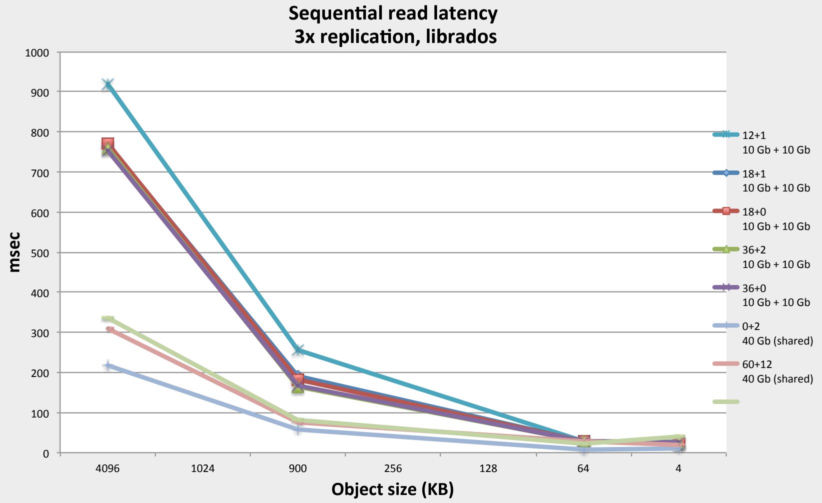

Figure 11. Sequential read latency per server at the librados level with 3x replication for tested Supermicro server configurations.

Figure 12 illustrates sequential read throughput on a per-server basis using erasure coding (3+2) across the same group of servers and cluster configurations. This testing also evaluates throughput at the librados level.

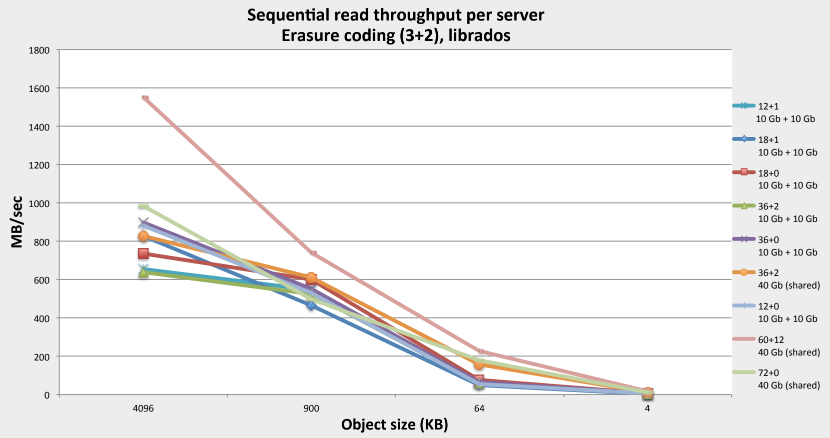

Figure 12. Sequential read throughput per server at the librados level with erasure coding (3+2) for tested Supermicro server configurations.

Sequential read latency for the erasure-coded configuration is shown in Figure 13.

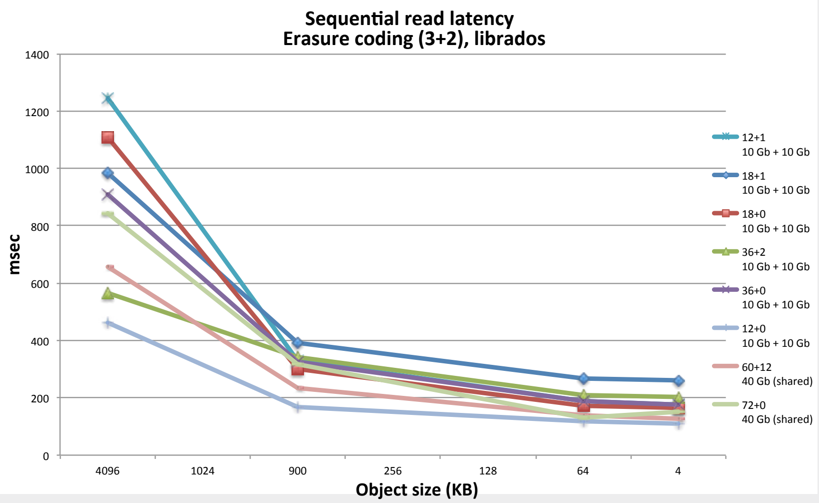

Figure 13. Sequential read latency per server at the librados level with erasure coding (3+2) for tested Supermicro server configurations.

Sequential read throughput per OSD

The rados bench command was also used to evaluate read throughput at the OSD level. Figure 14 shows sequential read throughput per OSD at the librados level, with 3x replication.

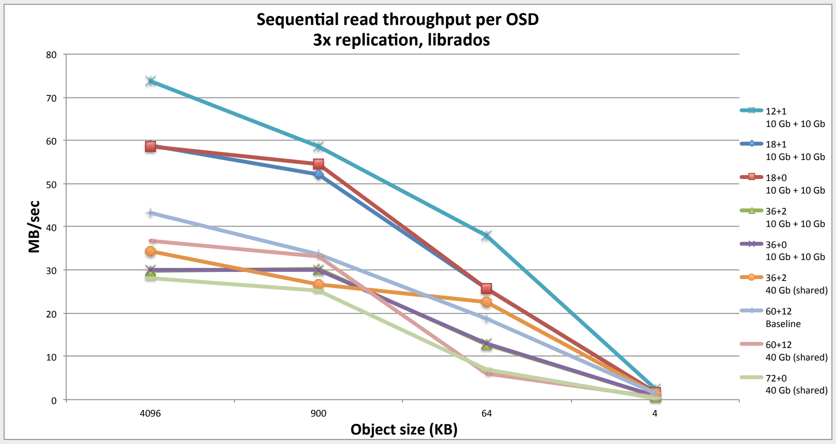

Figure 14. Sequential read throughput per OSD at the librados level with 3x replication for tested Supermicro server configurations.

Figure 15 shows sequential read throughput per OSD using an erasure-coded pool (3+2) for the same set of cluster configurations.

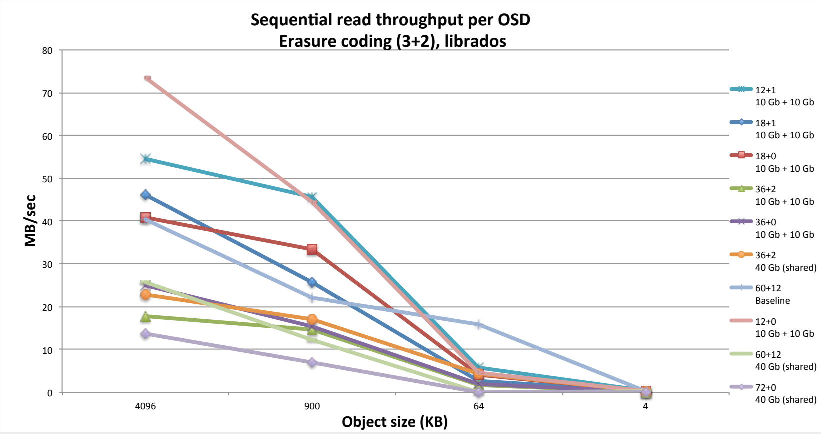

Figure 15. Sequential read throughput per OSD at the librados level with erasure coding (3+2) for tested Supermicro server configurations.

SEQUENTIAL WRITES

Sequential write performance and latency was also evaluated both on a per-server basis, as well as per OSD. Both replicated and erasure-coded pools were evaluated using the rados bench load-test utility.

Sequential write throughput per server

Figure 16 shows sequential write throughput per server at the librados level with a 3x replicated pool.

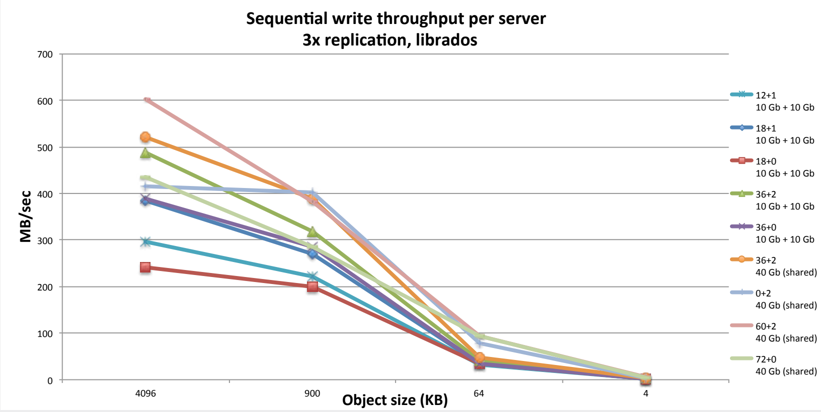

Figure 16. Sequential write throughput per server at the librados level with 3x replication for tested Supermicro server configurations.

Figure 17 illustrates sequential write latency per server at the librados level across a 3x replicated pool.

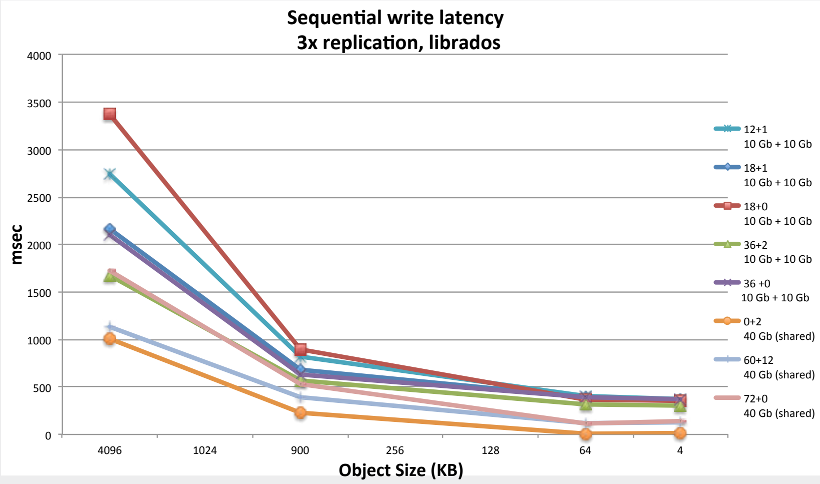

Figure 17. Sequential write latency per server at the librados level with 3x replication for tested Supermicro server configurations.

Figure 18 shows sequential write throughput per server at the librados level using an erasure-coded pool (3+2).

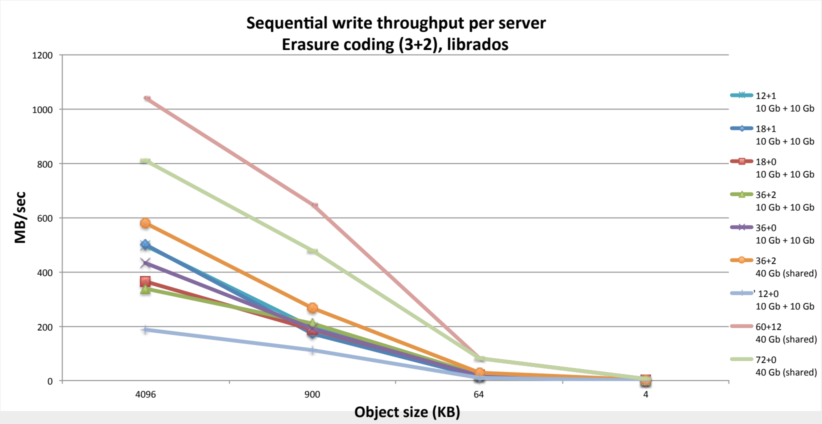

Figure 18. Sequential write throughput per server at the librados level with erasure coding (3+2) for tested Supermicro server configurations.

Figure 19 shows sequential write latency with an erasure coded pool at the librados level.

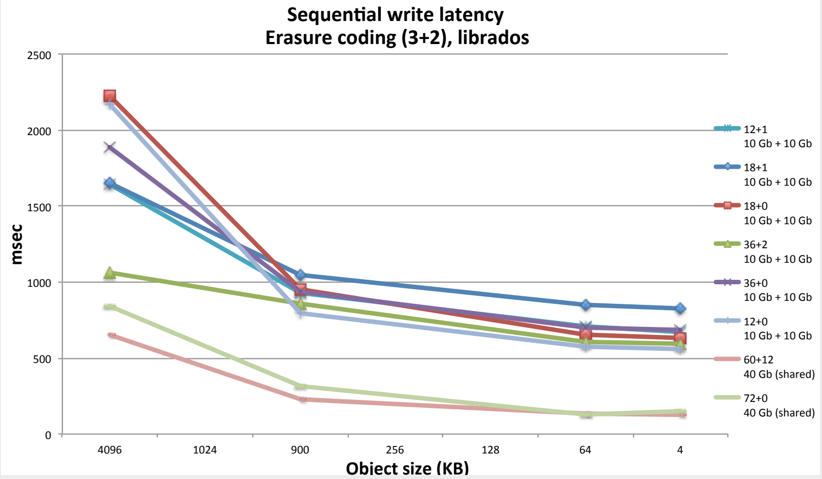

Figure 19. Sequential write latency per server at the librados level with erasure coding (3+2) for tested Supermicro server configurations.

Sequential write throughput per OSD

Sequential write throughput was also evaluated per OSD. Figure 20 illustrates sequential write throughput per OSD using a 3x replicated pool.

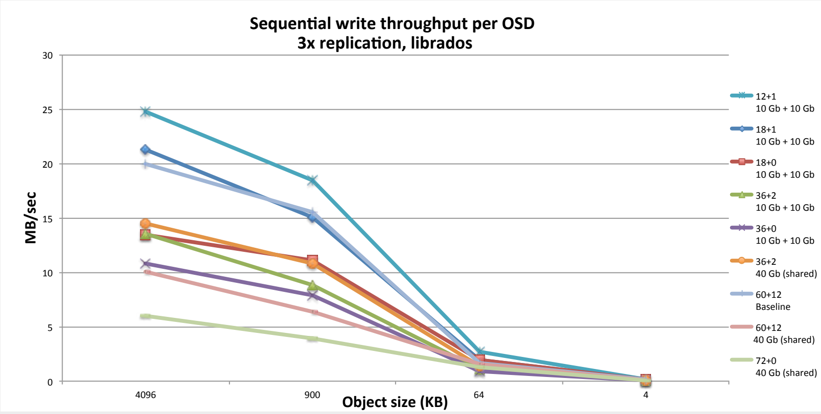

Figure 20. Sequential write throughput per OSD at the librados level with 3x replication for tested Supermicro server configurations.

Sequential write throughput per OSD using an erasure-coded pool (3+2) is shown in Figure 21.

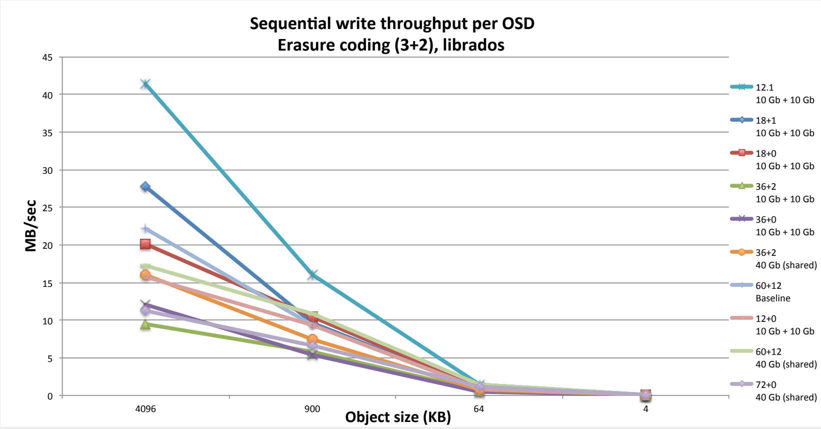

Figure 21. Sequential write throughput per OSD at the librados level with erasure coding (3+2) for tested Supermicro server configurations.

TESTING METHODOLOGY AND DETAILS

The sections that follow detail the testing methodology used by Red Hat and Supermicro in establishing the optimized configurations.

BASELINE SINGLE SERVER

Before conducting complicated testing of a distributed system, it is helpful to understand hardware capabilities and limits. Specifically, it is important to characterize the performance of the storage media and its associated components. Ideally, the aggregate disk performance is equal to the product of the number of disks and the performance obtained by a single disk.

FIO independent disk analysis

Typically, there is either a single class or two classes of storage media in an OSD host. When there are two classes of storage media, one is typically slower and used for OSD data volumeswhile the other, faster media is used for OSD journals. For each type of media, it is important to establish a baseline, both for comparison against the system aggregate performance and to determine whether the ratio of data volumes to journaling devices is too high.

FIO aggregate disk analysis

Once the baseline performance of each type of disk media is known, the next step is to test all media devices concurrently. Sometimes, the performance exhibited by each device will be lower than what was achievable during independent device testing. Performance degradation when all devices are being stressed concurrently may be due to a bottleneck in the SAS/PCI topography. Determining where in the SAS/PCI topography the bottleneck lies is beyond the scope of this document.

RADOS SINGLE HOST BENCHMARK

Before testing a Ceph cluster in aggregate, it is advisable to test a single host for baseline comparability against the results of cluster testing. During single host testing, the ruleset-failure-domain must be relaxed to the OSD level, since there will not be sufficient systems to which replicas and erasure-coded chunks can be distributed. Single host benchmarks still involve the use of multiple client systemsto avoid conflating load induced by load generation with Ceph OSDs system load. Because replicas and chunks are separated only by internal system buses, system performance in terms of writes (in the case of replication and erasure coding) and reads (in the case of erasure coding) could be significantly higher for both throughput and latency.

BIDIRECTIONAL COMPLETE GRAPH NETWORK TESTING

To independently validate the network configuration, it is important to verify that both clients and OSD hosts are able to saturate their network interfaces.

Client to cluster

The network utility iperf should be used to verify that clients are able to saturate their network interface on the Ceph public network. To more closely characterize the behavior of Ceph network traffic, we passed a flag to iperf clients to disable congestion control. A iperf server should be started on each OSD host for each client, and each client should create a iperf client for each OSD host.

Cluster to cluster

Likewise, iperf should be used to verify that OSD hosts are able to saturate their network interface on the Ceph cluster network. A iperf server should be started on each OSD host for each other OSD host, and each OSD host should create a iperf client for each other OSD host. If a distinct Ceph cluster network is absent, OSD host to OSD host testing should instead be conducted on the shared Ceph public network.

SYSTEM TUNING

The systems in the CBT environment were subject to a number of system tuning efforts. Each adjustment is described in the sections that follow, along with the rationale for making the adjustment.

OSD system tuning

Red Hat Enterprise Linux offers a number of system profiles that are well-suited for a number of common use cases. These profiles can be selected with the tuned-adm system utility. Our testing has revealed that the default, throughput-performance, is the ideal starting point. A system administrator can ensure that their system is using the throughput-performance tuned profile by running the following command:

```
# tuned-adm list
```

OSD systems run one OSD server process per data disk in the system. Each of these processes generates a multitude of threads for messaging, filestore operations, and similar operations. On dense systems, the maximum number of process identifiers can be quickly met, because threads are indistinguishable from distinct processes to the scheduler in the Linux kernel. In order to support the number of threads required for dense systems, administrators should increase the maximum number of process identifiers using a system control configuration parameter, as shown:

```
# sysctl -w kernel.pid_max=131072
```

Each Ceph OSD server process stores RADOS objects as files in a file system on the OSDs data device. Reads—including reads for modification—require identifying where a file is stored within the file system. To avoid having to seek once for file system metadata and a second time to access the contents of a file, system administrators should adjust a system control parameter so that the Linux kernel prefers to retain inode/dentry caches when the system is under memory pressure, as shown:

```
# vm.vfs_cache_pressure=1
```

In addition to adjusting how the Linux kernel handles various caches when under memory pressure, adjusting the swappiness system control parameter lets the kernel gives priority to runtime memory allocations over page cache, as shown:

```
# sysctl -w vm.swappiness=1
```

The recommended file system for use with Ceph is XFS. All journaling file systems, including XFS, require that acknowledged writes are persisted to non-volatile storage. Most drives have a small writeback cache that can increase write performance by coalescing. These caches should be disabled on Ceph OSD data and journal devices to prevent journal or file system corruption. The `hdparm` tool provided with Red Hat Enterprise Linux can be used to adjust the cache setting on each drive, as shown:

```
# hdparm -W0 /dev/sdX
```

Modern disk drives read a configurable number of sectors at once in order to increase read throughput or avoid extra seeks when there are multiple read/modify/write operations acting on adjacent sectors. In our tests the read ahead was configured for 256 sectors, a system administrator can adjust the read ahead configuration of their OSD data devices using the `hdparm` tool.

```
# hdparm -a256 /dev/sdX
```

Ceph OSD systems have many devices that generate system interrupts, including host bus adapters and/or RAID controllers, PCIe flash cards, and high speed network adapters. The total number of system interrupts can easily prevent processor core(s) from doing system or user tasks. The kernel drivers for most devices will provide a number of MSI-X queues on which they can raise system interrupts. Each MSI-X queue is assigned to a single CPU core. It is advisable to ensure that the core handling a MSI-X queue be affliated with the device presenting the queue. The sys virtual file system provides a way of finding which processor cores share CPU affinity, as shown:

```
# /sys/class/pci_bus/<pci bus id>/cpulistaffinity
```

Before tuning interrupt handling, the *irqbalance* system daemon should be stopped and disabled to prevent it from altering MSI-X queue assignments. It is also important to script MSI-X tuning and include it in a system startup script, since interrupt queue assignments do not persist across system reboots. Knowing which processor cores share affinity, administrators can balance the MSI-X queues by taking the MSI-X IRQ ids from /proc/interrupts and using them to specify cores to handle each IRQ, as shown:

```
# /bin/echo $core_id > /proc/irq/<irq id>/smp_affinity
```

Engineers found a number of XFS parameters to be beneficial for OSD data volume file system performance. During file system creation, the fundamental block size should be set to 2KB, and the directory area block size should be set to 64KB. When mounting OSD XFS filesystems, enable 64-bit inodes, both for performance and because filesystems larger than 1TB may prematurely run out of inode identifiers. File access timestamps are not as useful for Ceph filestores, because replication and erasure coding do not use this extra metadata. Therefore, it is safe to disable access timestamps when mounting by passing the `noatime` option, which could provide marginal performance gains.

XFS is a journaling file system, and as such, journal performance is paramount to the performance of the overall file system. To reduce the number of journal I/O operations, the journal log buffer size should be increased to 256K. Ceph data volumes should be prepared with the ceph-disk utility, allowing the init scripts bundled with Ceph to mount the OSD data volume file systems. To ensure the proper settings are used during file system creation and mounting, the following can be added to /etc/ceph/ceph.conf:

```
osd mkfs options = -f -i size=2048 -n size=64k

osd mount options xfs = inode64,noatime,logbsize=256k
```

By default, Ceph uses extended attributes to store metadata about RADOS objects. In some cases, the metadata exceeds what is possible to store as extended attributes, necessitating storing this metadata as key/value pairs in an omap (a key/value store inside RADOS). The omap method of storing extended attributes tends to perform better than relying on file system extended attributes. To store extended attributes as omap entries, add the following to /etc/ceph/ceph.conf:

```
filestore xattr use omap = true
```

Ceph creates a series of directory hierarchies on each OSD disk to store objects. A hierarchy of directories is formed for each PG in every storage pool in the cluster. By default, a directory will be split into sub directories when 320 objects are placed in a directory. These tuneable options increase the threshold for repartitioning. Adding these values to /etc/ceph/ceph.conf should improve read and write performance at the possible expense of lower backfill performance:

```
filestore merge threshold = 40

filestore split multiple = 8
```

Slightly improved read performance has been observed during synthetic testing by adjusting the number of OSD op threads as follows:

```
osd op threads = 12
```

RADOS MULTI-HOST BENCHMARK

Once single-server baseline network graph testing and single-host RADOS testing were completed, rados cluster testing was conducted, as described above, using the CBT.

CEPH BENCHMARK TOOL

The Ceph Benchmark Tool (CBT) is a framework to aid in evaluating the performance of Ceph storage clusters. A CBT environment consists of a head host, any number of client hosts, Ceph monitor hosts, and any number of Ceph OSD hosts. Cluster tests are defined in a YAML configuration file that outlines the role of each host in the environment: whether it should create or use an existing cluster, which driver to use to load the cluster, and what testing parameters to pass to the drivers. CBT drivers launch a load generation tool on each client and allow some parameters to be specified in the benchmark YAML. These parameters control which flags are passed to the oad generation tools. Some parameters exposed by CBT allow arrays of values, enabling the benchmark to iterate a number of times with different testing parameters.

NETWORK CONFIGURATION PARAMETERS

Testing included both 10 Gigabit Ethernet and 40 Gigabit Ethernet network adaptors from Intel and Mellanox respectively. In each case, it is advisable to follow the adapter tuning guides.

Intel network adapter tuning

The following system control parameters are recommended for tuning Intel network adapters:

```
# sysctl -w net.core.rmem_default=524287

# sysctl -w net.core.wmem_default=524287

# sysctl -w net.core.rmem_max=524287

# sysctl -w net.core.wmem_max=524287

# sysctl -w net.core.optmem_max=524287

# sysctl -w net.core.netdev_max_backlog=300000

# sysctl -w net.ipv4.tcp_rmem=”10000000 10000000 10000000”

# sysctl -w net.ipv4.tcp_wmem=”10000000 10000000 10000000”

# sysctl -w net.ipv4.tcp_mem=”10000000 10000000 10000000”
```

Mellanox tuning

Mellanox has published a guide that describes how to tune the performance of their network adapters. The guide can be downloaded from the Mellanox website[9]. The recommended system control parameters suggested in the Mellanox tuning document are, as follows:

[9] *http://www.mellanox.com/related-docs/prod_software/Performance_Tuning_Guide_for_Mellanox_Network_Adapters.pdf*

```
# sysctl -w net.ipv4.tcp_timestamps=0

# sysctl -w net.ipv4.tcp_sack=1

# sysctl -w net.core.netdev_max_backlog=250000

# sysctl -w net.core.rmem_max=4194304

# sysctl -w net.core.wmem_max=4194304

# sysctl -w net.core.rmem_default=4194304

# sysctl -w net.core.wmem_default=4194304

# sysctl -w net.core.optmem_max=4194304

# sysctl -w net.ipv4.tcp_rmem=”4096 87380 4194304”

# sysctl -w net.ipv4.tcp_wmem=”4096 65536 4194304”

# sysctl -w net.ipv4.tcp_low_latency=1

# sysctl -w net.ipv4.tcp_adv_win_scale=1
```

COST, OBSERVATIONS, AND OPTIMAL CONFIGURATIONS

Building scale-out storage infrastructure ultimately involves considering multiple factors. The sections that follow provide price/performance comparisons for running Red Hat Ceph Storage on Supermicro storage servers, along with optimal small, medium, and large configurations. Observations are also provided for deployment configuration and sizing.

PRICE/PERFORMANCE COMPARISONS

System configurations were tested using both throughput-optimization and capacity-optimization criteria. However, the requirements of many organizations include a combination of throughput and capacity-optimized criteria, so optimal configurations will vary accordingly.

Throughput-optimized configurations

The following criteria were used to identify throughput optimized configurations:

- Lowest cost per unit of throughput
- Highest throughput
- Highest throughput per BTU
- Highest throughput per watt
- Meets minimum fault domain recommendation (one server less than or equal to 10% of the cluster)

Table 6 compares a variety of Supermicro server configurations against these criteria. Absolute pricing information was omitted.

TABLE 6. PERFORMANCE DATA FOR THROUGHPUT-OPTIMIZED CONFIGURATIONS.

DRIVES (SSDS)	12x 4TB (1x PCIe)	12x 6 TB (1x PCIe)	30x 4 TB (6x SSD)	36x 4 TB (2x PCIe)	36x 6 TB (2x PCIe)	60x 4 TB (12x SSD)	72x 6 TB
INDIVIDUAL SERVER SPECIFICATIONS							
RAW TB	48	72	120	144	216	240	432
USABLE TB	16	24	40	48	72	80	144
READ THROUGHPUT (MB/SEC)	885	885	1,020	1,241	1,241	2,202	2,027
WRITE THROUGHPUT (MB/SEC)	297	297	330	521	521	601	434
HEAT (BTU/HR)	1,274	1,274	2,858	2,551	2,551	5,534	5,839
WATTS	373	373	868	747	747	1,621	1,711
RACK UNITS	2	2	4	4	4	4	4
WRITE THROUGHPUT PER BTU	0.23	0.23	0.12	0.20	0.20	0.11	0.07
WRITE THROUGHPUT PER WATT	0.80	0.80	0.38	0.70	0.70	0.37	0.25
OPENSTACK STARTER (50 TB) CLUSTER							
CLUSTER SIZE (USABLE TB)	50	50	50	50	50	50	50
SERVERS	4	3	2	2	1	1	1
READ THROUGHPUT (ESTIMATED)	3,540	–	–	–	–	–	–
WRITE THROUGHPUT (ESTIMATED)	1,118	–	–	–	–	–	–
HEAT (BTU/HR)	5,096	–	–	–	–	–	–
WATTS	11,936	–	–	–	–	–	–
RACK UNITS	8	–	–	–	–	–	–
SMALL (500 TB USABLE) CLUSTER							
SERVERS	32	21	13	11	7	7	4
READ THROUGHPUT (ESTIMATED)	28,320	18,585	13,260	13,651	8,687	15,414	–
WRITE THROUGHPUT (ESTIMATED)	9,504	6,237	4,290	5,731	3,647	4,207	–
HEAT (BTU/HR)	40,768	26.754	37,154	28,061	17,857	38,738	–
WATTS	11,936	7,833	11,284	8,217	5,229	11,347	–

DRIVES (SSDS)	12x 4TB (1x PCIe)	12x 6 TB (1x PCIe)	30x 4 TB (6x SSD)	36x 4 TB (2x PCIe)	36x 6 TB (2x PCIe)	60x 4 TB (12x SSD)	72x 6 TB
RACK UNITS	64	42	52	44	28	28	–
MEDIUM (1 PB USABLE) CLUSTER							
SERVERS	63	42	25	21	14	13	7
READ THROUGHPUT (ESTIMATED)	55,755	37,170	25,500	26,061	17,374	28,626	14,189
WRITE THROUGHPUT (ESTIMATED)	18,711	12,474	8,250	10,941	7,294	7,813	3,038
HEAT (BTU/HR)	80,262	53,508	71,450	53,571	35,714	71,942	40,873
WATTS	11,936	7,833	11,284	8,217	5,229	11,347	6,844
RACK UNITS	126	84	100	84	56	52	28
LARGE (2 PB USABLE) CLUSTER							
SERVERS	125	84	50	42	28	25	14
READ THROUGHPUT (ESTIMATED)	110,625	74,340	51,000	52,122	34,748	55,050	28,378
WRITE THROUGHPUT (ESTIMATED)	37,125	24,948	16,500	21,882	14,588	15,025	6,076
HEAT (BTU/HR)	159,250	107,016	142,900	107,142	71,428	138,350	81,746
WATTS	11,936	7,833	11,284	8,217	5,229	11,347	6,844
SELECTED SERVER SPECIFICATIONS							
RAM (GB)	64	64	128	128	128	256	256
HDD QUANTITY	12	12	30	36	36	60	72
HDD SIZE (TB)	4	6	4	4	6	4	6
SSD MODEL	Intel P3700	Intel P3700	Intel P3700	Intel P3700	Intel P3700	Intel S3710	–
SSD INTERFACE	PCIe	PCIe	SATA	PCIe	PCIe	SATA	–
SSD JOURNAL QUANTITY	1	1	6	2	2	12	–
SSD JOURNAL SIZE (GB)	800	800	200	800	800	200	–
SSD MB/S WRITE SPEC.	1900	1900	300	1900	1900	300	–
CAPACITY EFFICIENCY (3X REPLICATION)	33%	33%	33%	33%	33%	33%	33%

Figures 23 and 24 provide cost comparisons between various system configurations.

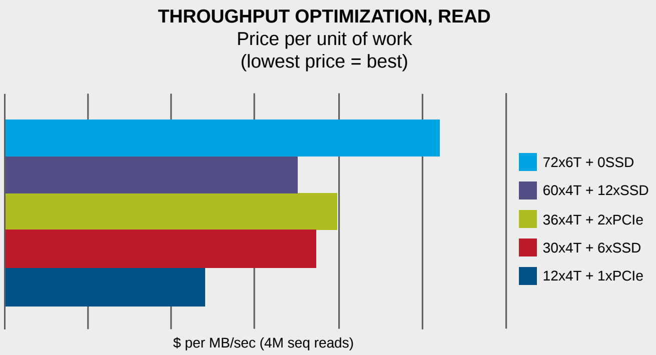

Figure 23. System comparison for read throughput optimization based on lowest price per unit of work.

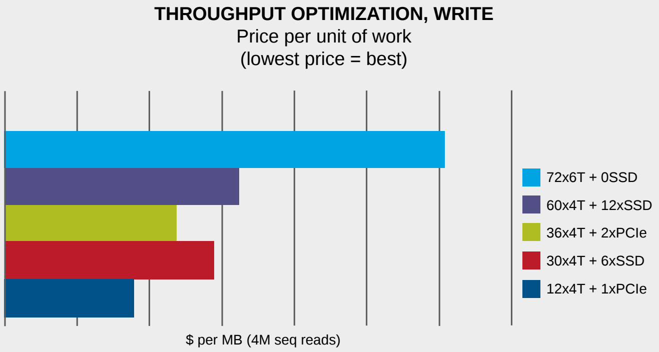

Figure 24. System comparison for write throughput optimization based on lowest price per unit of work.

Capacity-optimized configurations

The following criteria were used to identify throughput optimized configurations:

- Lowest cost per TB
- Lowest BTU per TB
- Lowest watt per TB
- Highest TB per rack unit
- Meets minimum fault domain recommendation(1 server less than or equal to 15% of the cluster)

Table 7 and Figure 25 compare a variety of configurations against these criteria. Absolute pricing information was omitted.

TABLE 7. PERFORMANCE DATA FOR CAPACITY-OPTIMIZED CONFIGURATIONS.

DRIVES (SSDS)	12x 6TB (no SSD)	36x 6 TB (no SSD)	72x 6 TB (no SSD)
INDIVIDUAL SERVER SPECIFICATIONS			
RAW TB	72	216	432
USABLE TB (ERASURE CODED 6+2)	54	162	324
READ THROUGHPUT (MB/S 4M)	881	896	981
WRITE THROUGHPUT (MB/S 4M)	189	432	809
HEAT (BTU/HR)	1,274	2,366	5,534
WATTS	373	693	1,621
RACK UNITS	2	4	4
TB PER RACK UNIT (RAW)	36	54	108
WRITE THROUGHPUT PER BTU	0.15	0.18	0.15
WRITE THROUGHPUT PER WATT	0.51	0.62	0.50
BTUS PER TB	17.69	10.95	12.81
WATTS PER TB	5.18	3.21	3.75
SMALL (500 TB USABLE)CLUSTER			
SERVERS	10	4	2
READ THROUGHPUT (ESTIMATED)	8,810	–	–
WRITE THROUGHPUT (ESTIMATED)	1,890	–	–
HEAT (BTU/HR)	12,740	–	–
WATTS	3,730	–	–
RACK UNITS	20	–	–
MEDIUM (1 PB USABLE) CLUSTER			
SERVERS	19	7	4
READ THROUGHPUT (ESTIMATED)	16,739	6,272	–
WRITE THROUGHPUT (ESTIMATED)	3,591	3,024	–
HEAT (BTU/HR)	24,206	16,562	–
WATTS	3,730	2,772	–
RACK UNITS	38	28	–

DRIVES (SSDS)	12x 6TB (no SSD)	36x 6 TB (no SSD)	72x 6 TB (no SSD)
LARGE (2 PB USABLE) CLUSTER			
SERVERS	38	13	7
READ THROUGHPUT (ESTIMATED)	33,478	11,648	6,867
WRITE THROUGHPUT (ESTIMATED)	7,182	5,616	5,663
HEAT (BTU/HR)	48,412	30,758	38,738
WATTS	3,730	2,772	3,242
RACK UNITS	76	52	28
SELECTED SERVER SPECIFICATIONS			
RAM (GB)	64	128	256
HDD model	Seagate C. ES.3	Seagate C. ES.3	Seagate C. ES.3
HDD quantity	12	36	72

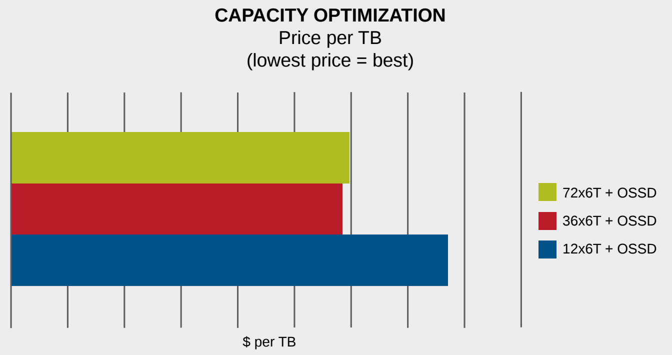

Figure 25. System comparison for capacity optimization based on lowest price per terabyte.

OBSERVATIONS

Testing conducted by Red Hat and Supermicro allowed engineers to make a number of observations relative to sizing and appropriate architectures:

- **Data protection**. All throughput-optimized configurations were configured with replicated storage pools (3x replication). Capacity-optimized configurations employed erasure coding. Erasure coding with 6+2 was used for optimal configuration modeling. The tests used 4+2, 3+2, or 2+1, due to system availability for testing and to require a ruleset-failure-domain of host (to properly characterize the impact of storing and retrieving replicas/chunks across the network fabric). Note that Ceph RBD is supported on replicated pools only, while Ceph RGW is supported on either replicated or erasure-coded pools. Future benchmarking work is planned with RBD volumes using cache tiering on small replicated pools backed by larger erasure-coded pools for price/ performance optimization.

- **Server size.** The 12-bay server provided optimal results for throughput-optimized clusters of all sizes, based on the optimization criteria. Many organizations prefer the simplicity and commodity of 12-bay servers as well. It is important to note, however, that the 36-bay server offers more terabytes per rack unit and lower watts and BTUs per terabyte. As such, the 36-bay server configuration may be optimal for write-heavy, latency-sensitive workloads for 500TB+ throughput-optimized clusters.
- **SSD write journaling.** SSD write journals did materially enhance throughput-optimized write performance (using replication) but showed only marginal benefit on capacity-optimized configurations (using erasure coding).
- **PCIe versus SATA SSDs.** For throughput-optimized configurations, price, capacity, and performance analysis typically favored populating all SAS/SATA drive bays with HDDs while journaling to 1-2 PCIe SSDs. This configuration choice is preferable to replacing some of the HDDs with SAS/ SATA SSD journals as it provides more usable storage capacity per node.
- **Throughput per OSD/HDD.** OSDs on 12-bay servers produced significantly more throughput than OSDs on larger servers. Analysis of CPU utilization, I/O subsystem, memory, and 40 Gigabit Ethernet networking suggested that these subsystems were not bottlenecks. Further investigation is underway to evaluate the implications of non-uniform memory access (NUMA) systems, along with core to Ceph OSD ratios (with and without hyperthreading).
- **Server fault domains.** As discussed, a supported minimum Ceph cluster should contain three storage servers, with 10 storage servers recommended. During self-healing, a percentage of cluster throughput capacity will be diverted to creating object copies from the failed node on the remaining nodes. The percentage of cluster performance degradation is inversely proportional to the number of nodes in the cluster. See the section on fault domain risk tolerance above for further explanation.

OPTIMIZED CONFIGURATIONS (SMALL, MEDIUM, AND LARGE)

Tables 8 and 9 below contain optimal small, medium, and large Supermicro server configurations for both throughput and capacity optimization, according to the criteria listed in this paper. Work is ongoing for IOPS-optimized configurations. The performance figures in these tables are estimates based upon actual benchmark results produced on smaller clusters published in this document.

TABLE 8. THROUGHPUT-OPTIMIZED SUPERMICRO SERVER CONFIGURATIONS.

	CEPH CLUSTER SIZE (USABLE CAPACITY)			
	STARTER (50 TB)	SMALL (500 TB)	MEDIUM (1 PB)	LARGE (2 PB)
OSD SERVER QUANTITY	• 4	• 32	• 63	• 125
PERFORMANCE (ESTIMATED)	• Read: 3,500 MB/s • Write: 1,200 MB/s	• Read: 28,000 MB/s • Write: 9,500 MB/s	• Read: 55,000 MB/s • Write: 19,000 MB/s	• Read: 110,000 MB/s • Write: 37,000 MB/s
SUPERMICRO SERVERS	SSG-2028R-OSD072 (w/ 4 TB HDDs), or SSG-F618H-OSD288 (w/ 4 TB HDDs): 1x E5-2620v3 64 GB RAM 12x 4T HDD 1x 800 GB PCIe			
NETWORKING	10 Gigabit Ethernet 10 Gigabit Ethernet Gigabit Ethernet			

TABLE 9. CAPACITY-OPTIMIZED SUPERMICRO SERVER CONFIGURATIONS.

	CEPH CLUSTER SIZE (USABLE CAPACITY)		
	SMALL (500 TB)	MEDIUM (1 PB)	LARGE (2 PB)
OSD SERVER QUANTITY	• 10	• 7	• 13
PERFORMANCE (ESTIMATED)	• Read: 8,000 MB/s • Write: 2,000 MB/s	• Read: 6,000 MB/s • Write: 3,000 MB/s	• Read: 11,000 MB/s • Write: 5,000 MB/s
SUPERMICRO SERVERS	SSG-2028R-OSD072, or SSG-F618H-OSD288 including: • 1x E5-2620v3 • 64 GB RAM • 12 x 6TB HDD	SSG-6048R-OSD216 including: • 2x E5-2630v3 • 128 GB RAM • 36 x 6TB HDD	SSG-6048R-OSD432 including: • 2x E5-2695v3 • 256 GB RAM • 72 x 6TB HDD
NETWORKING	10 Gigabit Ethernet 10 Gigabit Ethernet Gigabit Ethernet		

CONCLUSION

Red Hat evaluates, tests, and documents reference configurations that depict real-world deployment scenarios, giving organizations specific and proven configuration advice that fits their application needs. Red Hat Ceph Storage used with Supermicro storage servers, Seagate HDDs, Intel SSDs, and Mellanox and Supermicro networking represents an ideal technology combination. The architectures described herein afford a choice of throughput or cost/capacity workload focus, allowing organizations to customize their Ceph deployments to their needs. The architectures also provide a range of cluster sizes, ranging from hundreds of terabytes to multiple petabytes, with repeatable configurations that have been tested and verified by Red Hat and Supermicro engineers.

APPENDIX A: BILL OF MATERIALS FOR SELECT SYSTEMS

Tables 10-14 describe Supermicro server configurations optimized for Red Hat Ceph Storage.

TABLE 10. MONITOR NODE CONFIGURATION

	SYS-6017R-MON1
KEY FEATURES	• 4 x 3.5-inch HDD bays • Dual 10 Gigabit Ethernet (SFP+)
PROCESSOR	• Dual Intel Xeon E5-2630 v2 6-core 2.6 GHz 15M 7.2 GT/s QPI
MEMORY	• 64GB per node
NETWORKING	• On-board dual-port 10 Gigabit Ethernet (SFP+)
DRIVE CONFIGURATION	• 4 x 300GB HDDs (SAS3)
FORM FACTOR	• 1U with redundant hot-swap 700W power supplies

TABLE 11. 2U, 12-DRIVE OSD NODE FOR THROUGHPUT-OPTIMIZED WORKLOADS

	SSG-6028R-OSD072 & SSG-6028R-OSD72P
KEY FEATURES	• 12 x 3.5-inch HDD bays + mirrored 80GB OS drive (rear hot-swap) • x8 SAS3 connectivity • Dual 10 Gigabit Ethernet (SFP+)
PROCESSOR	• Single Intel Xeon E5-2620 v3 6-core 2.4 GHz 15M 8 GT/s QPI
MEMORY	• 64GB per node
NETWORKING	• AOC-STGN-12S dual-port 10 Gigabit Ethernet (SFP+)
DRIVE CONFIGURATION	• 1 x 800GB PCIe Flash card (OSD72P configuration) • 12 x 6TB HDDs, Seagate ST6000NM0004 (72TB raw capacity)
FORM FACTOR	• 2 RU with redundant 920W platinum power supplies

TABLE 12. 4X 12-DRIVE OSD NODE FOR THROUGHPUT-OPTIMIZED WORKLOADS

SYS-F618H-OSD288P	
KEY FEATURES	• 12 x 3.5-inch HDD bays + mirrored SATADOM • 4x node front I/O • Dual 10 Gigabit Ethernet (SFP+)
PROCESSOR	• Single Intel Xeon E5-2630 v2 8-core 2.4 GHz 20M 8 GT/s QPI
MEMORY	• 64GB per node
NETWORKING	• AOC-STGN-12S dual-port 10 Gigabit Ethernet (SFP+)
DRIVE CONFIGURATION	• 1 x 400GB PCIe Flash card (internal) • 12 x 6TB HDDs (72 TB raw capacity)
FORM FACTOR	• 4 RU 4x node, with redundant 1680W Titanium power supplies

TABLE 13. 36-DRIVE OSD NODES FOR CAPACITY-OPTIMIZED WORKLOADS

SSG-6048R-OSD216 AND SSG-6048R-OSD216P	
KEY FEATURES	• 36 x 3.5-inch HDD bays + 2 x internal 2.5-inch OS drives (mirrored, 80 GB SSD) • x8 SAS3 connectivity • Quad 10 Gigabit Ethernet (SFP+)
PROCESSOR	• Dual Intel Xeon E5-2630 v3 8-core 2.4 GHz 20M 8 GT/s QPI
MEMORY	• 128GB per node
NETWORKING	• AOC-STGN-12S qual port 10 Gigabit Ethernet (SFP+)
DRIVE CONFIGURATION	• OSD216: 36 x 6TB HDDs (216TB raw capacity) • OSD216P: 36 x 6TB HHDs (216TB raw capacity) with 2 x 400GB PCIe Flash cards (internal)
FORM FACTOR	• 4 RU with redundant hot-swap 1280W Platinum power supplies

TABLE 14. 72-DRIVE OSD NODES FOR CAPACITY-OPTIMIZED WORKLOADS

SSG-6048R-OSD32 AND SSG-6048R-OSD360P	
KEY FEATURES	• 72 x 3.5-inch HDD bays + 2 x hot-swap OS drives (mirrored, 80 GB SSD) • x8 SAS3 connectivity • Quad 10 Gigabit Ethernet (SFP+)
PROCESSOR	• Dual Intel Xeon E5-2690 v3 12-core 2.6 GHz 30M 9.6 GT/s QPI
MEMORY	• 256GB per node
NETWORKING	• AOC-STGN-12S quad-port 10 Gigabit Ethernet (SFP+)
DRIVE CONFIGURATION	• OSD432: 72 x 6TB HDDs (432TB raw capacity) • OSD360P: 60 x 6TB HHDs (360TB raw capacity) with 12 x 400 SATA3 SSDs
FORM FACTOR	• 4 RU with redundant hot-swap 2000W Titanium power supplies

ABOUT SUPERMICRO COMPUTER, INC.

Supermicro (NASDAQ: SMCI), a global leader in high-performance, high-efficiency server technology and innovation, is a premier provider of end-to-end green computing solutions for datacenter, cloud computing, enterprise IT, Hadoop/big data, HPC, and embedded systems worldwide. Supermicro's advanced Server Building Block Solutions® offer a vast array of components for building energy-efficient, application-optimized, computing solutions.

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