Dell PowerEdge R730xd Performance and Sizing Guide for Red Hat Ceph Storage

This technical white paper provides an overview of the Dell PowerEdge R730xd server performance results with Red Hat Ceph Storage. It covers the advantages of using Red Hat Ceph Storage on Dell servers with their proven hardware components that provide high scalability, enhanced ROI cost benefits, and support of unstructured data. This paper also gives hardware configuration recommendations for the PowerEdge R730xd running Red Hat Ceph Storage.

August 2016

Revisions

Date	Description
August 2016	Initial release

Contents

Glossary

CephFS

Ceph Filesystem. The portable operating system interface (POSIX) filesystem components of Ceph.

iDRAC

The integrated Dell Remote Access Controller, an embedded server management interface.

HDD

Hard Disk Drive.

KVM

Kernel Virtual Machine, a hypervisor.

MON

The Ceph monitor software.

Node

One of the servers in the cluster.

OSD

Object Storage Device, a physical or logical unit.

RBD

Ceph RADOS Block Device.

RGW

RADOS Gateway, the S3/Swift gateway component of Ceph.

SDS

Software-defined storage, an approach to computer data storage in which software is used to manage policy-based provisioning and management of data storage, independent of the underlying hardware.

SSD

Solid State Drive.

ToR

Top-of-Rack switch.

Executive Summary

Data storage requirements are staggering, and growing at an ever-accelerating rate. These demanding capacity and growth trends are fueled in part by the enormous expansion in unstructured data, including music, image, video, and other media; database backups, log files, and other archives; financial and medical data; and large data sets, aka "big data". Not to mention the growing data storage requirements expected by the rise of the internet of things (IoT). Yet with all these demanding capacity requirements, customer expectations for high reliable and high performance are greater than ever.

As IT organizations struggle with how to manage petabytes and even exabytes of ever-growing digital information, the adoption of cloud-like storage models is becoming more common in modern data centers. One answer is the software known as Ceph.

Ceph is an open source distributed object storage system designed to provide high performance, reliability, and massive scalability. Ceph implements object storage on a distributed computer cluster, and provides interfaces for object-, block- and file-level storage. Ceph provides for completely distributed operation without a single point of failure, and scalability to the petabyte level. Ceph replicates data and makes it fault-tolerant. As a result of its design, the system is both self-healing and self-managing, helping to minimize administration time and other costs. Since Ceph uses general-purpose hardware, controlled by software whose features are exposed through application programming interfaces (APIs), it is considered to be a type of software-defined storage (SDS).

Red Hat Ceph Storage is an enterprise-ready implementation of Ceph that provides a single platform solution for software-defined storage that is open, adaptable, massively scalable, technologically advanced, and supported worldwide. Red Hat Ceph Storage combines innovation from the open source community with the backing of Red Hat engineering, consulting, and support. It includes tight integration with OpenStack services and was built from the ground up to deliver next-generation storage for cloud and emerging workloads.

This technical white paper provides performance and sizing guidelines for Red Hat Ceph Storage running on Dell servers, specifically the Dell PowerEdge R730xd server, based on extensive testing performed by Red Hat and Dell engineering teams. The PowerEdge R730xd is an award-winning server and storage platform that provides high capacity and scalability and offers an optimal balance of storage utilization, performance, and cost, along with optional in-server hybrid hard disk drive and solid state drive (HDD/SSD) storage configurations.

This paper is intended for storage architects, engineers, and IT administrators who want to explore the advantages of using Red Hat Ceph Storage on Dell PowerEdge servers and who need to design and plan implementations using proven best practices.

1 Introduction

Unstructured data has demanding storage requirements across the access, management, maintenance, and particularly the scalability dimensions. To address these requirements, Red Hat Ceph Storage provides native object-based data storage and enables support for object, block, and file storage.

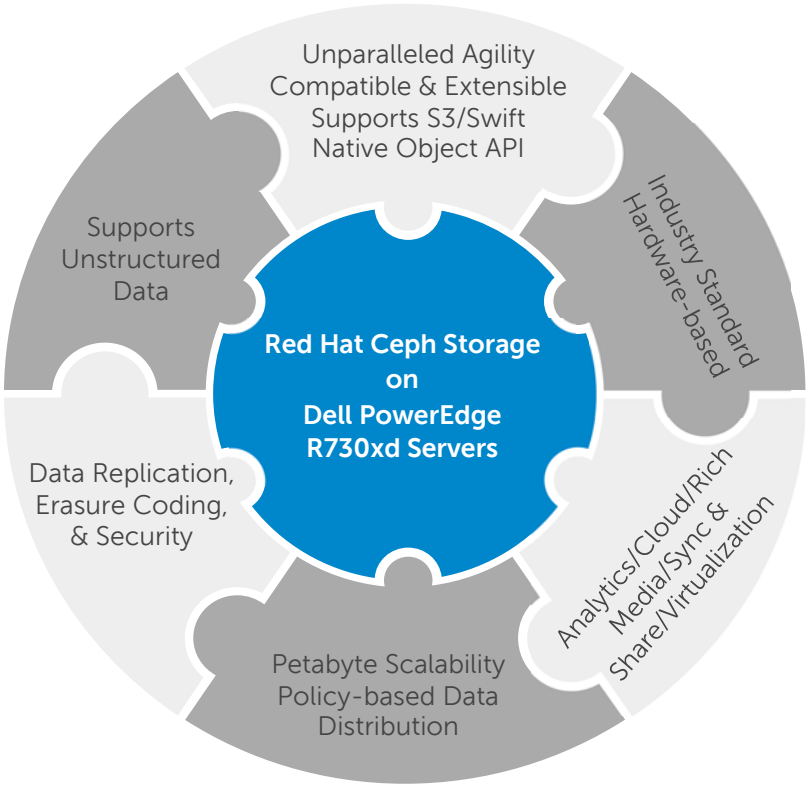


Figure 1 Key takeaways of deploying Red Hat Ceph on Dell PowerEdge R730xd servers

The Red Hat Ceph Storage environment makes use of industry standard servers that form Ceph nodes for scalability, fault-tolerance, and performance. Data protection methods play a vital role in deciding the total cost of ownership (TCO) of a solution. Ceph allows the user to set different data protection methods on different storage pools.

- Replicated storage pools make full copies of stored objects, and is ideal for quick recovery. In a replicated storage pool, the Ceph configuration defaults to a replication factor of three, where three copies of the data are placed on three separate Ceph nodes.
- Erasure-coded storage pools provide a single copy of data plus parity, and it is useful for archive storage and cost-effective durability and availability. For more information on designing scalable workload-optimized Ceph clusters, please see the configuration guide at https://access.redhat.com/documentation/en/red-hat-ceph-storage/1.3/single/ceph-configuration-guide/.

The PowerEdge R730xd is an exceptionally flexible and scalable two-socket 2U rack server, that delivers high performance processing and a broad range of workload-optimized local storage possibilities, including hybrid tiering. Designed with an incredible range of configurability, the PowerEdge R730xd is well suited for Ceph.

- PowerEdge servers allow users to construct and manage highly efficient infrastructures for data centers and small businesses.
- PowerEdge servers accelerate time-to-production with automated deployment processes that use fewer manual steps and reduce human error.
- PowerEdge servers improve IT productivity in the data center with innovative management tools, like iDRAC Quick Sync and iDRAC Direct to deliver in-depth system health status and speed deployment. For additional information on iDRAC, including Zero-Touch Auto Configuration, please see: http://en.community.dell.com/techcenter/systems-management/w/wiki/4317.white-papers-for-idrac-with-lifecycle-controller-technology
- PowerEdge servers optimize data center energy usage with improved performance-per-watt and more granular control of power and cooling.

Table 1. Dell PowerEdge R730xd Server Specifications

Feature	Dell PowerEdge R730xd Server Specifications
Form factor	2U rack
Processor	Intel® Xeon® processor E5-2600 v4 product family
Processor sockets	2
Dimensions	H: 8.73 cm (3.44 in.) x W: 44.40 cm (17.49 in.) x D: 68.40 cm (26.92 in.)
Cache	2.5 MB per core; core options: 4, 6, 8, 10, 12, 14, 16, 18, 22
Chipset	Intel C610 series chipset
Memory	Up to 1.5 TB (24 DIMM slots): 4 GB/8 GB/16 GB/32 GB/64 GB DDR4 up to 2400MT/s
I/O slots	Up to 6 x PCIe 3.0 slots plus dedicated RAID card slot
RAID controllers	Internal controllers: PERC H330, PERC H730, PERC H730P External HBAs (RAID): PERC H830 External HBAs (non-RAID): 12Gbps SAS HBA
Drive bays	Internal hard drive bay and hot-plug backplane: Up to 18 x 1.8" SSD SATA drives Up to 16 x 3.5" and 2 x 2.5" SAS or SATA drives Up to 26 x 2.5" SAS or SATA drives (optional: 4 of the 26 slots can support PCIe)
Maximum internal storage	SAS, SATA, nearline SAS, SSD, PCIe SSD: Up to 46.8TB with 24 x 2.5" 1.8TB hot-plug SAS HDD + 2 x 2.5" 1.8TB hot-plug SAS HDD Up to 81TB with 18 x 1.8" 960GB SATA SSD + 8 x 3.5" 8TB SATA/nearline SAS HDD Up to 131.6 TB with 12 x 3.5" 8TB nearline SAS HDD = 4 x 3.5" 8TB nearline SAS + 2 x 2.5" 1.8TB SAS HDD or SSD
Embedded NIC	4 x 1GbE, 2x10+2GbE, 4 x 10GbE NDC
Power supply unit (PSU)	Titanium efficiency 750 W AC PSU; 1100 W DC PSU; Platinum efficiency 495 W, 750 W, and 1100 W AC PSU
Availability	ECC memory, hot-plug hard drives, hot-plug redundant cooling, hot-plug redundant power, internal dual SD module, single device data correction (SDDC), spare rank, tool-less chassis, support for high availability clustering and virtualization, proactive systems management alerts, iDRAC8 with Lifecycle Controller
Remote management	iDRAC8 with Lifecycle Controller, iDRAC8 Express (default), iDRAC8 Enterprise (upgrade) 8GB vFlash media (upgrade), 16 GB vFlash media (upgrade)
Rack support	ReadyRails™ II sliding rails for tool-less mounting in 4-post racks with square or unthreaded round holes or tooled mounting in 4-post threaded hole racks, with support for optional tool-less cable management arm.
Recommended support	Dell ProSupport Plus for critical systems or Dell ProSupport for premium hardware and software support for your PowerEdge solution. Consulting and deployment offerings are also available. Contact your Dell representative today for more information. Availability and terms of Dell Services vary by region. For more information, visit Dell.com/ServiceDescriptions.

For the latest specifications and full spec sheet on the PowerEdge R730xd server, please see http://www.dell.com/us/business/p/poweredge-r730xd/pd.

The Dell PowerEdge R730xd offers advantages that include the ability to drive peak performance by:

- Accelerating application performance with the latest technologies and dynamic local storage.
- Scaling quickly and easily with front-accessible devices, ranging from low-cost SATA hard drives to 2.5" ultra-fast, low-latency PowerEdge Express Flash NVMe PCIe SSDs.
- Using hybrid storage configurations to tailor the R730xd to your workload needs and implement tiered-storage efficiencies. Combine 24-drive capacity with dual-PERC capability to drive storage performance and take advantage of new Dell caching technologies to boost application

The Dell PowerEdge R730xd delivers greater versatility. Unlike previous generations of servers, the R730xd can be tailored for application performance by mixing high-speed 1.8" SSDs and low-cost, high-capacity 3.5" hard drives in a single hybrid chassis, enabling accelerated data access through in-server storage tiering. Examples of this versatility include:

- In its 26-drive, dual-PERC configuration, the R730xd can be an excellent database server delivering more than one million IOPS performance.
- You can combine 26 drives with 24 DIMMs of memory and six PCI Express® (PCIe) expansion slots to provide a resource-rich virtualization environment.
- With up to four ultra-fast, ultra-low latency Express Flash NVMe PCIe SSDs, the R730xd can boost performance on critical workloads and share the cache with other servers.
- In its low-cost 16 x 3.5" drive configuration with up to 128TB capacity, the R730xd is an excellent server for unified communications and collaboration (UC&C) and delivers the scale-out storage efficiency demanded by the XaaS providers, Hadoop/big data users, and co-location hosting.

This technical white paper discusses the test results obtained by using different combinations of Dell server hardware components to achieve throughput-optimized and cost/performance configurations.

Red Hat Ceph Storage was deployed on different hardware configurations of Dell PowerEdge R730xd servers. The Ceph Benchmarking Toolkit (CBT) was used to test different parameters. Both read and write operations were performed to test throughput, price/performance, price/GB, differences between replicated and erasure-coded methods, and differences between HBA JBOD and single-disk RAID0 mode. A total of 880 tests in single-drive RAID0 and HBA JBOD modes were performed using 88 different configurations with 10 varying workloads.

Table 2 below shows a summary of the PowerEdge R730xd and Ceph Storage configurations used for benchmark tests used for this paper.

Table 2. PowerEdge R730xd and Ceph Storage Configurations used for Benchmark Tests

Configurations	Brief Description
PowerEdge R730xd 12+3, 3x Replication	PowerEdge R730xd with 12 hard disk drives (HDDs) and 3 solid state drives (SSDs), 3X data replication and single-drive RAID0 mode.
PowerEdge R730xd 12+3, EC 3+2	PowerEdge R730xd with 12 HDDs and 3 SSDs, erasure-coding and single-drive RAID0 mode.
PowerEdge R730xd 16+1, 3x Replication	PowerEdge R730xd with 16 HDDs and 1 PCIe SSD, 3X data replication method and single-drive RAID0 mode.
PowerEdge R730xd 16+1, EC 3+2	PowerEdge R730xd with 16 HDDs and 1 PCIe SSD, erasure-coding and single-drive RAID0 mode.
PowerEdge R730xd 16j+1, 3x Replication	PowerEdge R730xd with 16 HDDs and 1 PCIe SSD, 3X data replication and HBA JBOD PERC pass-through mode.
PowerEdge R730xd 16j+1, EC 3+2	PowerEdge R730xd with 16 HDDs and 1 PCIe SSD, erasure-coding and HBA JBOD PERC pass-through mode.
PowerEdge R730xd 16+0, 3x Replication	PowerEdge R730xd with 16 HDDs and no SSDs, 3X data replication and single-drive RAID0 mode.
PowerEdge R730xd 16+0, EC3+2	PowerEdge R730xd with 16 HDDs and no SSDs, erasure-coding and single-drive RAID0 mode.

Note: Though the Dell PowerEdge R730xd server provides flexibility in the layout and configuration of I/O subsystems, the combinations described in Table 2 were selected on the basis of performance data of different configuration variations tested.

After extensive performance and server scalability evaluation and testing, Red Hat and Dell have classified the benchmark results into the following five categories:

- Compare server throughput in different configurations
- Compare overall solution price/performance
- Compare overall solution price/capacity
- Compare server throughput in replication versus erasure-coded modes
- Compare server throughput in JBOD and RAID0 modes

These five comparisons are shown in section 4 of this document.

2 Overview of Red Hat Ceph Storage

A Ceph storage cluster is built from 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 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

Red Hat Ceph Storage provides an effective way for enterprise block and object storage techniques, supports archival, rich media, and cloud infrastructure workloads such as OpenStack. A few advantages of Red Hat Ceph Storage are listed in Figure 2:

- Recognized industry leadership in open source software support services and online support
- Only stable, production-ready code, vs. a mix of interim, experimental code
- Consistent quality; packaging available through Red Hat Satellite
- Well-defined, infrequent, hardened, curated, committed 3-year lifespan with strict policies
- Timely, tested patches with clearly-defined, documented, and supported migration path
- Backed by Red Hat Product Security
- Red Hat Certification & Quality Assurance Programs
- Red Hat Knowledgebase (articles, tech briefs, videos, documentation), Automated Services

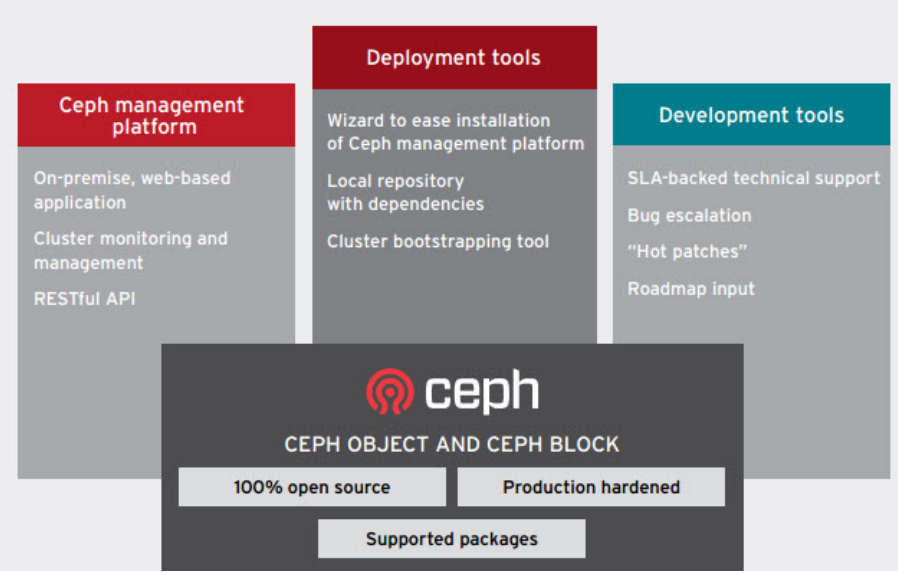
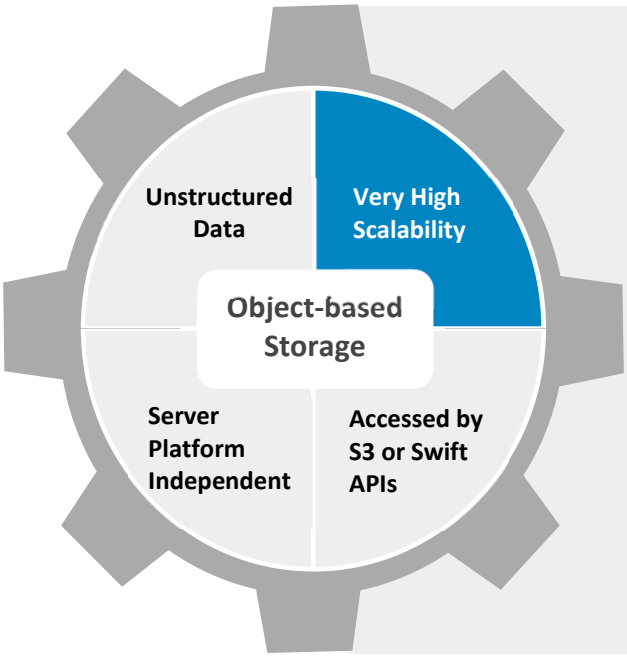


Figure 2 Red Hat Ceph Storage

Red Hat Ceph Storage significantly lowers the cost of storing enterprise data and helps 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®. 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. Some of the properties include:

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



- Flat-structured, location independent, ability to use unstructured data (i.e. PDF, video, audio, image, or Word doc) that is replicated and supports geographic failover.
- No single point of failure.
- Supports user-defined and dynamic metadata, flexible, and very high scalability (up to 1000 nodes). OSD nodes and storage can be scaled up independently.
- Can be accessed by using S3 or Swift application programming interfaces (APIs), REST, and SOAP.
- Stores bytes, terabytes, petabytes: files, videos, images, tar, backup, ISO image files, and many more.
- Well suitable for cloud computing with flat file structure (e.g. no hierarchy of folders).
- Very secure and reliable.
- Self-healing, extended metadata, policy-based capability & management.

Figure 3 Ceph Object-based Storage

Table 3 below provides a matrix of different Ceph cluster design factors, optimized by workload category. Please see https://access.redhat.com/documentation/en/red-hat-ceph-storage/1.3/single/ceph-configuration-guide/ for more information.

Table 3. Ceph cluster design considerations

Optimization Criteria	Potential Attributes	Example Uses
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
Throughput-optimized	• Lowest cost per given unit of throughput • Highest throughput • 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

2.1 Introduction to 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 logical storage pool in the Ceph cluster. Figure 4 illustrates the overall Ceph architecture, and concepts that are described in the sections that follow.

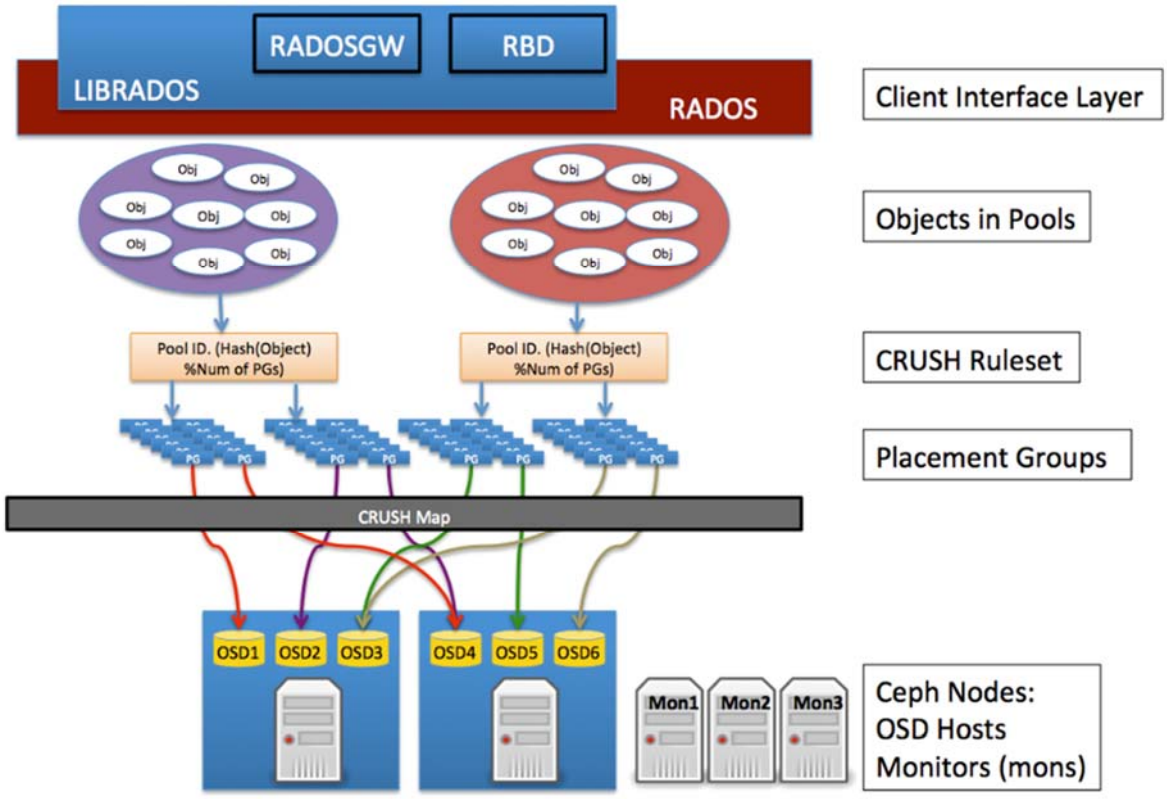


Figure 4 Ceph Storage Pools

Pools: A Ceph storage cluster stores data objects in logical dynamic partitions called pools. Pools can be created for particular data types, such as for block devices, object gateways, or simply to separate user groups. The Ceph pool configuration 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. Also, pools can "take root" at any position in the CRUSH hierarchy (see below), allowing placement on groups of servers with differing performance characteristics—allowing storage to be optimized for different workloads.

Placement groups: Ceph maps objects to placement groups (PGs). 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 provide a way to creating replication or erasure coding groups of coarser granularity than on a per-object basis. A larger number of placement groups (for example, 200/OSD or more) leads to better balancing.

CRUSH ruleset: CRUSH is an algorithm that provides controlled, scalable, and decentralized 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 OSD daemons 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 hard disk drive.

2.2 Selecting a Storage Access Method

Choosing a storage access method is an important design consideration. As discussed, all data in Ceph is stored in pools—regardless of data type. The data itself is stored in the form of objects by using the Reliable Autonomic Distributed Object Store (RADOS) layer which:

- Avoids a single point of failure
- Provides data consistency and reliability
- Enables data replication and migration
- Offers automatic fault-detection and recovery

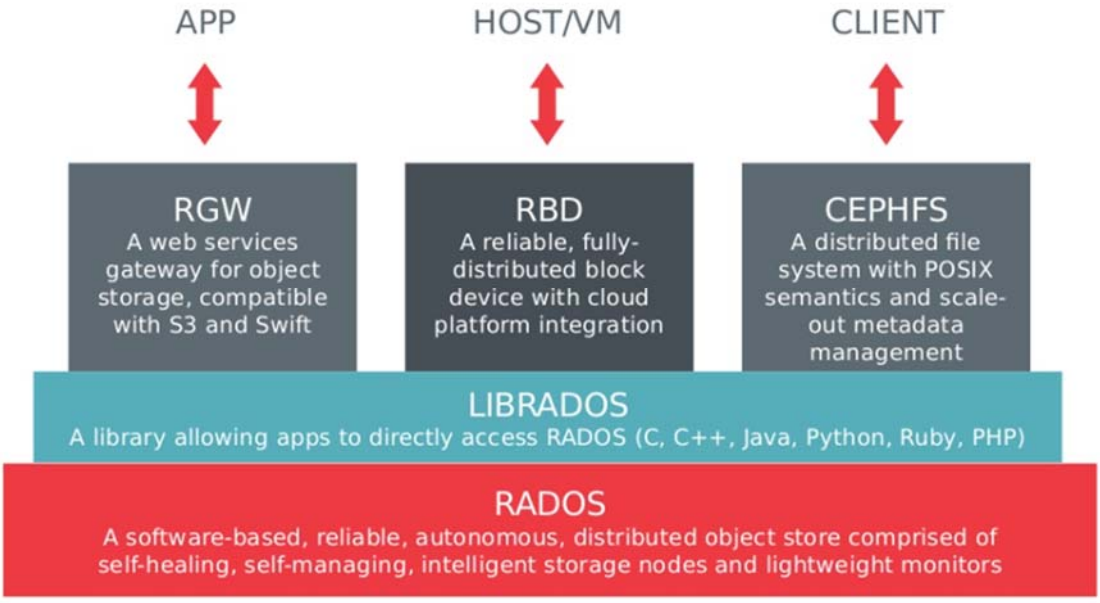


Figure 5 RADOS Layer in the Ceph Architecture

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

- RADOSGW: Bucket-based object storage gateway service with S3 compatible and OpenStack Swift compatible RESTful interfaces.
- LIBRADOS: Provides direct access to RADOS with libraries for most programming languages, including C, C++, Java, Python, Ruby, and PHP.
- 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 backend, or userspace libraries).

Storage access method and data protection method (discussed later in this technical white paper) are interrelated. For example, Ceph block storage is currently only supported on replicated pools, while Ceph object storage is supported on either erasure-coded or replicated pools. The cost of replicated architectures is categorically more expensive than that of erasure-coded architectures because of the significant difference in media costs.

2.3 Selecting a Storage 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. This is because 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 are suspended until at least two OSDs are operational.

Erasure-coded storage pools: Erasure coding provides a single copy of data plus parity, and it is useful for archive storage and cost-effective durability and availability. With erasure coding, storage pool objects are divided into chunks by 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. So for instance, *n* disks are needed to store *k* disks worth of data with data protection and fault tolerance of *m* disks.

Ceph block storage is typically configured with 3x replicated pools and is currently not supported directly on erasure-coded pools. Ceph object storage is supported on either replicated or 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 while meeting performance requirements.

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

3 Test Setup and Methodology

This section describes the Red Hat Ceph Storage on Dell PowerEdge R730xd Testbed. It also describes the testing performed on the testbed. The following subsections cover:

- Testbed hardware configuration
- Installation of Red Hat Ceph Storage software
- Benchmarking procedure

3.1 Physical setup

Figure 6 illustrates the testbed for the Red Hat Ceph Storage on Dell PowerEdge R730xd. The benchmarking testbed consists of five Ceph Storage nodes based on the Dell PowerEdge R730xd servers with up to sixteen 3.5" drives. These serve as the OSD tier. The MON servers are based on three Dell PowerEdge R630 servers. The load generators are based on Dell PowerEdge R220 servers, providing a total of 10 clients that execute various load patterns.

Each Ceph Storage node and MON server has two 10GbE links. One link is connected to the front-end network shown in Figure 5. The other link is connected to the back-end network. The load generator servers have a single 10GbE link connected to the front-end network. The ToR switching layer is provided by Dell Force10 S4048 Ethernet switches.

The subnet configuration covers two separate IP subnets, one for front-end Ceph client traffic (in orange) and a separate subnet for the back-end Ceph cluster traffic (in blue). A separate 1 GbE management network is used for administrative access to all nodes through SSH, that is not shown in the Figure 6.

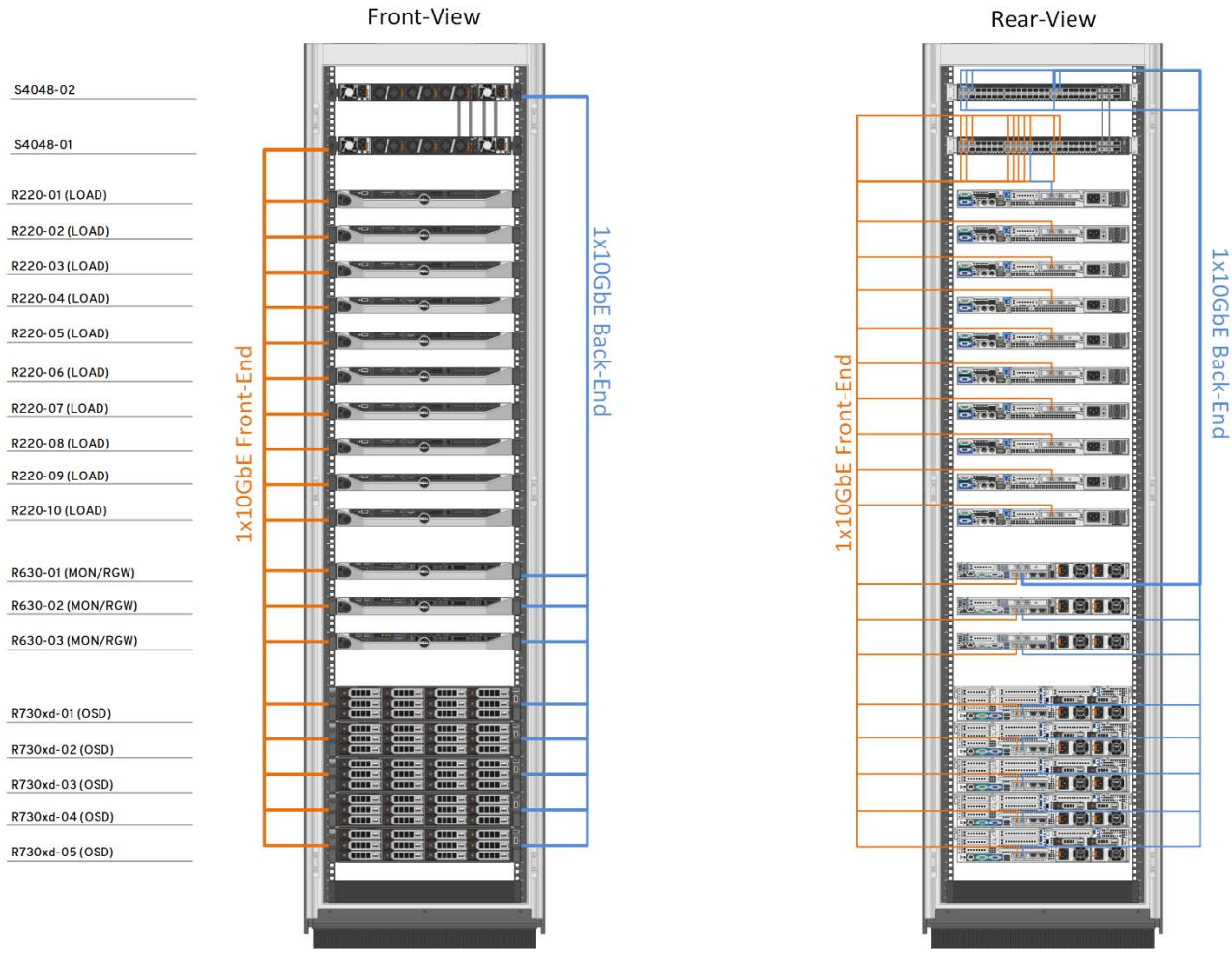


Figure 6 The 5-node Red Hat Ceph Storage cluster based on Dell PowerEdge R730xd servers

3.2 Hardware and Software Components

Tables 4 and 5 give details on the testbed hardware.

Table 4. Hardware Components used for Testbed

Testbed Details			
Ceph tier	OSD	MON/RGW	CLIENT
Platform	Dell PowerEdge R730xd	Dell PowerEdge R630	Dell PowerEdge R220
CPU	2x Intel Xeon E5-2630 v3 2.4GHz	2x Intel Xeon E5-2650 v3 2.3 GHz	1x Intel Celeron G1820 2.7 GHz
Memory	4x 16 GB 1866 MHz DDR4	8x 16 GB 2133MHz DDR4	4x 4 GB 1600 MHz DDR3
Network	1x Intel X520/2P I350 LOM	1x Intel X520/2P I350 LOM	1x Intel X520/2P I350
Storage	PERC H730 Mini / 1 GB Cache Up to 16x: SEAGATE 4 TB SAS (ST4000NM0005) Up to 3x: Intel DC S3700 SSD 200 GB SATA (SSDSC2BA20) 1x Intel DC P3700 SSD 800 GB NVMe	PERC H730 Mini / 1 GB Cache 6x SEAGATE 500 GB SAS (ST9500620SS)	1x Toshiba 50 GB SATA (DT01ACA0)

Table 5. Dell Force10 Switch Series

Dell Force10 switch configuration	
Layer	Access switch
Platform	Dell Force10 S4048
Ports	48x 10 GbE SFP+ 6x 40 GbE QSFP+

While the overall physical setup, server types, and number of systems remain unchanged, the configuration of the OSD node's storage subsystems was altered. Throughout the benchmark tests, different I/O subsystem configurations are used to determine the best performing configuration for a specific usage scenario. Table 6, Table 7, and Table 8 list the configurations used in the benchmark tests.

Table 6. Server and Ceph Storage Configurations Tested in Benchmarks

OSD to Journal Ratio [drives]	**12+3**	**16+0**	**16+1**
OSD node configuration	12+3	16+0	16+1
HDDs	12	16	16
HDD RAID mode	Single-disk RAID0	Single-disk RAID0	Single-disk RAID0
SATA SSDs	3	0	0
SSD RAID mode	JBOD[1]	JBOD	JBOD
NVMe SSDs	0	0	1
Network	1x 10 GbE Front-End 1x 10 GbE Back-End	1x 10 GbE Front-End 1x 10 GbE Back-End	1x 10 GbE Front-End 1x 10 GbE Back-End

Table 7. Software Components used for Testbed

Ceph	Red Hat Ceph Storage 1.3.2
Operating System	Red Hat Enterprise Linux 7.2
Tools	Ceph Benchmarking Tool (CBT) and FIO 2.2.8

[1] JBOD indicates PERC pass-through mode

Table 8. Server Configurations

Server configuration	**PowerEdge** R730xd 12+3, 3xRep	**PowerEdge** R730xd 16+0, EC3+2	**PowerEdge** R730xd 16r+1, 3xRep	**PowerEdge** R730xd 16+1, EC 3+2	**PowerEdge** R730xd 16j+1, 3xRep
OS disk	2x 500 GB 2.5"	2x 500 GB 2.5"	2x 500 GB 2.5"	2x 500 GB 2.5"	2x 500 GB 2.5"
Data disk type	HDD 7.2K SAS 12Gbps, 4TB	HDD 7.2K SAS 12Gbps, 4TB	HDD 7.2K SAS 12Gbps, 4TB	HDD 7.2K SAS 12Gbps, 4TB	HDD 7.2K SAS 12Gbps, 4TB
HDD quantity	12	16	16	16	16
Number of Ceph write journal devices	3	0	1	1	1
Ceph write journal device type	Intel SATA SSD S3710 (6Gb/s)	n/a	Intel P3700 PCIe NVMe HHHL AIC	Intel P3700 PCIe NVMe HHHL AIC	Intel P3700 PCIe NVMe HHHL AIC
Ceph write journal device size (GB)	200	0	800	800	800
Controller model	PERC H730, 1 GB Cache	PERC H730, 1 GB Cache	PERC H730, 1 GB Cache	PERC H730, 1 GB Cache	PERC H730, 1 GB Cache
PERC Controller configuration for HDDs	RAID	RAID	RAID	RAID	JBOD (PERC pass-through mode)
Raw capacity for Ceph OSDs (TB)	48	64	64	64	64

While the Dell PowerEdge R730xd provides a great deal of flexibility in the layout and configuration of IO subsystems, the choice was limited to the above mentioned configurations. This decision was based on performance data of different configuration variations tested during the baselining which provided the following data points:

- SATA SSDs perform better when the PERC is configured as JBOD pass-through mode rather than configured as a RAID0 single-disk devices. SATA SSDs in JBOD pass-through have higher sequential and random IO throughput and lower latency.
- SAS HDDs have higher throughput of random small-block IO when configured as RAID0 single-disk devices than as configured as JBOD pass-through devices. This is due to the PERC H730 Mini Cache being enabled for RAID0 devices.
- SAS HDDs have higher sequential write throughput when the disks are configured as Non-RAID devices. However, it was determined that the average disk access pattern of Ceph (which is more random on the individual disks) would benefit more from the presence of the RAID cache than from higher write bandwidth.

- Previous benchmark data has shown that per-disk read-ahead settings had no effect on Ceph performance.

3.3 Deploying Red Hat Enterprise Linux (RHEL)

Red Hat Ceph Storage is a software-defined object storage technology which runs on RHEL. Thus, any system that can run RHEL and offer block storage devices is able to run Red Hat Ceph Storage. For the purpose of repeated execution, the configuration of the R730xd and R630 nodes as well as the deployment of RHEL on top of them has been automated. A virtual machine running RHEL 7.2 was created to control automated installation and to coordinate benchmarks. The virtual machine will be referenced as the Admin VM throughout the remainder of this document. The Admin VM is connected to the R730xd and R630 servers via the 1 GbE management network. RHEL was installed by using the standard installation process as recommended by Red Hat. For additional information, please see https://access.redhat.com/documentation/en-US/Red_Hat_Enterprise_Linux/7/html/Installation_Guide/.

3.4 Configuring the Dell PowerEdge Servers

The Dell PowerEdge R730xd and Dell PowerEdge R630 servers are automatically configured using the iDRAC and the racadm configuration utility. The iDRAC configuration is deployed on the Admin VM and used to reset the server configuration—including the BIOS and PERC RAID controller configuration. This ensures all systems have the same configuration and were set back to known states between configuration changes.

The configuration for iDRAC is described in XML format and can be found in the GitHub repository at https://github.com/red-hat-storage/dell-ceph-psg. With the racadm command, the configuration can be retrieved and restored to and from an NFS share, which is provided by the Admin VM.

3.5 Deploying Red Hat Ceph Storage

In production environments, Red Hat Ceph Storage can be deployed with a single easy-to-use installer. This installer ships with the Red Hat Ceph Storage Distribution and is referred to as ceph-deploy. In this benchmark, for the purpose of integration into automation flows, an alternative installation routine called ceph-ansible has been selected on the basis of Ansible playbooks.

Ceph-ansible is an easy to use, end-to-end automated installation routine for Ceph clusters based on the Ansible automation framework. Relevant for this benchmarking process are mainly two configuration files in the form of Ansible variable declarations for host groups. Predefined Ansible host groups exist to denote certain servers according to their function in the Ceph cluster, namely OSD nodes, Monitor nodes, RADOS Gateway nodes, and CephFS Metadata server nodes. Tied to the predefined host groups are predefined Ansible roles. The Ansible roles are a way to organize Ansible playbooks according to the standard Ansible templating framework, which in turn, are modeled closely to roles that a server can have in a Ceph cluster. For additional information on ceph-ansible, see https://github.com/ceph/ceph-ansible.

In this benchmark, Ceph MON and RGW roles are hosted side-by-side on the Dell PowerEdge R630 servers. Although, no RGW tests were performed for this paper. The configuration files are available in the ansible-ceph-configurations directory after checkout from the GitHub repository at https://github.com/red-hat-storage/dell-ceph-psg.

3.6 Performance Baselining

Before attempting benchmark scenarios that utilize higher-layer Ceph protocols, it is recommended to establish a known performance baseline of all relevant subsystems, which are:

- HDDs and SSDs (SATA + NVMe)
- Network (10 GbE Front-End and Back-End)
- CPU

The IO-related benchmarks on storage and network subsystems will be tied into direct reference of vendor specifications whereas the CPU benchmarks are ensuring the systems are all performing equally and provide comparison to future benchmarks. As such, the following baseline benchmarks have been conducted:

Table 9. Server Subsystem Baseline Tests

Subsystem	Benchmark Tool	Benchmark Methodology
CPU	Intel-linpack-11.3.1.002	Single-Core / Multi-Core Floating-Point Calculation
Network	iperf-2.0.8	Single TCP-Stream Benchmark All-to-All
SAS HDD	fio-2.2.8	8K random read/write and 4M sequential read/write on top of XFS
SATA SSD	fio-2.2.8	4K random read/write and 4M sequential read/write on top of XFS
NVMe SSD	fio-2.2.8	4K random read/write and 4M sequential read/write on top of XFS

CPU testing has been performanced with Intel LinPACK benchmark running suitable problem sizes given each server's CPU resources.

Table 10. CPU Baseline (results are average)

Server Type	PowerEdge R730xd 2x Intel Xeon E5-2630 v3	PowerEdge R630 2x Intel Xeon E5-2650 v3	PowerEdge R220 1x Intel Pentium G1820
LinPACK Multi-Threaded	377.5473 GFlops (Problem Size = 30000)	622.7303 GFlops (Problem Size = 30000)	20.0647 GFlops (Problem Size = 22000)
LinPACK Single-Threaded	47.0928 GFlops (Problem Size = 30000)	41.2679 GFlops (Problem Size = 30000)	10.3612 GFlops (Problem Size = 22000)

Network performance measurements have been taken by running point-to-point connection tests following a fully-meshed approach; that is, each server's connection has been tested towards each available endpoint of the other servers. The tests were run one by one and thus do not include measuring the switch backplane's combined throughput. Although the physical line rate is 10000 MBit/s for each individual link, the results are within ~1.5% of the expected TCP/IP overhead. The MTU used was 1500. The TCP Window Size was 2.50 MByte as the determined default by the Iperf tool.

Table 11. Network Baseline (results are average)

Server Type	PowerEdge R730xd Intel X520 LOM	PowerEdge R630 Intel X520 LOM	PowerEdge R220 Intel X520 LOM
PowerEdge R730xd Intel X520 LOM	9889.45 MBit/s	9886.53 MBit/s	9891.02 MBit/s
PowerEdge R630 Intel X520 LOM	9888.33 MBit/s	9895.5 MBit/s	9886.07 MBit/s
PowerEdge R220 Intel X520 LOM	9892.92 MBit/s	9892.8 MBit/s	9893.08 MBit/s

Storage performance has been measured thoroughly in order to determine the maximum performance of each individual component. The tests have been run on all devices in the system in parallel to ensure the backplanes, IO hubs and PCI bridges are factored in as well and don't pose a bottleneck. The fio job spec files used in these benchmarks are found at https://github.com/red-hat-storage/dell-ceph-psg. Each job was assembled of 3 components: a global include file, a scenario specific include file stating the IO pattern and a job specific file containing the target devices.

- *include-global.fio* – the global settings for all jobs run by fio, setting access method, IO engine, run tme and ramp time for each job
- *include-<target-device>-all-<test-type>-<access-pattern>.fio* – a scenario specific include file setting benchmark specific settings like block size, queue depth, access pattern and level of parallelism:
    - *target-device* – either journal (SATA SSDs) or OSDs (4TB SAS HDDs)
    - *all* – stating that all devices of the specified type are benchmarked in parallel
    - *test-type* – denoting the nature of the test, looking to quantify either sequential or random IO or access latency
    - *access pattern* – specifying which IO access pattern this test uses, either read, write or read-write mixed

In order to try to run the tests as close as possible to the way Ceph utilizes the systems, the tests were run on files in an XFS file system on the block devices under test using the formatting options from the *ceph.conf* file: *-f –i size=2048*. Before each benchmark run, all files were filled with random data up to the maximum file system capacity to ensure steady-state performance; write benchmarks were executed before read benchmarks to alleviate for different NAND behavior during reads. The data is reported on a per-device average.

Table 12. Disk IO Baseline (results are average)

Disk Type	OSD Seagate 4TB SAS	Journal Intel DC S3700 200GB	Journal Intel DC P3700 800GB
Random Read	314 IOPS (8K blocks)	72767 IOPS (4K blocks)	374703 IOPS (4K blocks)
Random Write	506 IOPS (8K blocks)	56483 IOPS (4K blocks)	101720 IOPS (4K blocks)
Sequential Read	189.92 MB/s (4M blocks)	514.88 MB/s (4M blocks)	2201 MB/s (4M blocks)
Sequential Write	158.16 MB/s (4M blocks)	298.35 MB/s (4M blocks)	1776 MB/s (4M blocks)
Read Latency	12.676 ms (8K blocks)	0.443 ms (4K blocks)	0.682 ms (4K blocks)
Write Latency	7.869 ms (8K blocks)	0.565 ms (4K blocks)	0.652 ms (4K blocks)

As previously stated, this data was obtained to get a performance baseline of the systems in their current setup; not to get individual peak performance for every device tested out of context. With that said, individual components may perform higher in other systems or when tested in isolation.

One instance of such a case was found when testing the SAS HDDs behind the PERC H730 Mini RAID Controller. When tested in isolation, a single 4TB SAS drive is able to achieve 190 MB/s sequential read and write patterns. When tested in parallel with all drives, the write bandwidth is limited by the RAID controller and the disks in RAID0 mode.

3.7 Benchmarking with CBT

For automation of the actual Ceph benchmarks, an open-source utility called the Ceph Benchmarking Tool (CBT) was used. It is available at https://github.com/ceph/cbt.

CBT is written in Python and takes a modular approach to Ceph benchmarking. The utility is able to use different benchmark drivers for examining various layers of the Ceph Storage stack, including RADOS, RADOS Block Device (RBD), RADOS Gateway (RGW) and KVM. In this paper, storage performance on the core layer RADOS is examined for which the driver in CBT uses the 'rados bench' benchmark which ships with Ceph. CBT's architecture is depicted below.

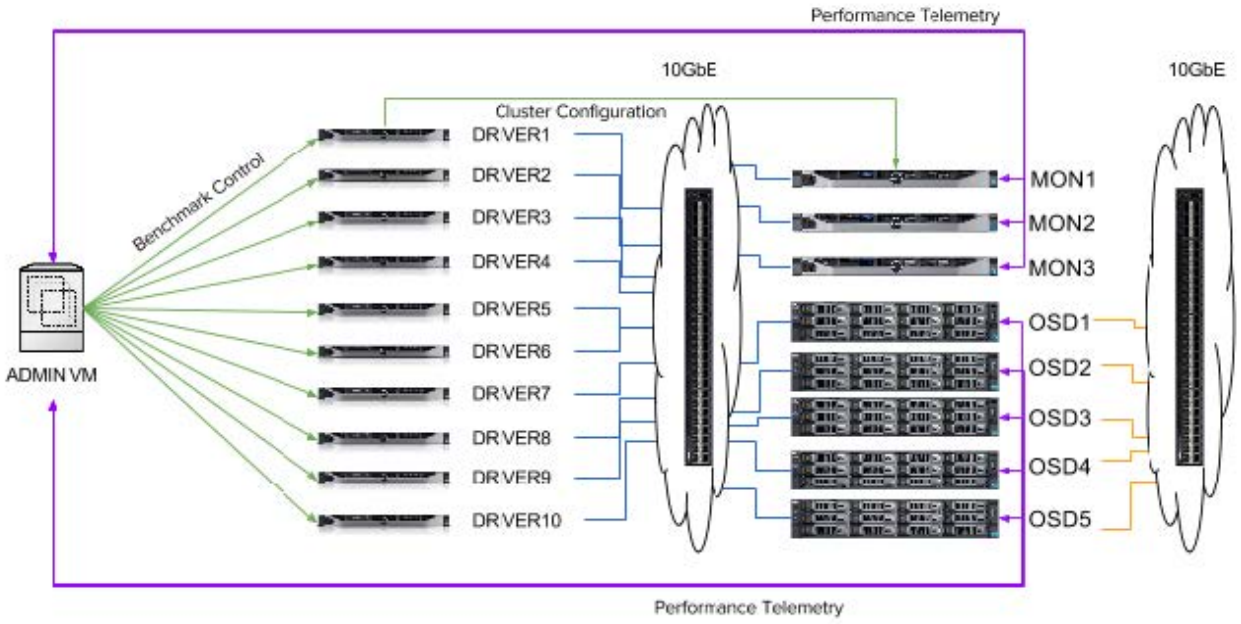


Figure 7 CBT Diagram

The utility is installed on the admin VM. From there, it communicates with various servers in different capacities via *pdsh* as follows:

- **Head Node**: a system that has administrative access to the Ceph cluster for the purpose of creating pools, rbds, change configuration or even re-deploy the entire cluster as part of a benchmark run
- **Clients**: these are the systems which have access to the Ceph cluster and from which CBT will generate load on the cluster using locally installed tools such as fio, rados or cosbench or run VMs to access the cluster
- **OSDs/MONs**: CBT triggers performance collection with various tools such as valgrind, perf, collect or blktrace on these nodes during the benchmark run and transfers their telemetry back to the head node after each execution

The CBT configuration file syntax was used to orchestrate most of the benchmarks. CBT provides flexibility to run benchmarks over multiple cluster configurations by specifying custom *ceph.conf* files. CBT also allows the user to re-deploy the cluster between benchmark runs completely.

In this benchmark, CBT was mainly used to execute the benchmarks. The cluster deployment and configuration was provided by ansible-ceph. The setup of CBT, including necessary pre-requisites and dependencies is described on the project homepage.

The CBT job files are specified in YAML syntax, for example:

```
cluster:
  user:  'cbt'
  head:  "r220-01"
  clients:  ["r220-01", "r220-02"]
  osds:  ["r730xd-01", "r730xd-02", "r730xd-03", "r730xd-04", "r730xd-05"]
  mons:
    r630-01:
      a:  "192.168.100.101:6789"
    r630-02:
      a:  "192.168.100.102:6789"
    r630-03:
      a:  "192.168.100.103:6789"
  iterations:  1
  use_existing:  True
  clustered:  "ceph"
  tmp_dir:  "/tmp/cbt"
  pool_profiles:
    replicated:
      pg_size:  4096
      pgp_size:  4096
      replications:  3
benchmarks:
  radosbench:
    op_size:  [4194304, 1048576, 524288, 131072]
    write_only:  False
    time:  300
    concurrent_ops:  [ 128 ]
    concurrent_procs:  1
    use_existing:  True
    pool_profile:  replicated
    readmode:  seq
    osd_ra:  [ 131072 ]
```

The file is divided in two sections: *cluster* and *benchmarks*. The first describes the cluster with the most essential data. The *user* specified here is a system user which needs to be present on all nodes and needs passwordless sudo access without the requirement for an interactive terminal. The *head* nodes, *clients* and *osds* are listed by their domain name of IP address. The MONs are specified in a syntax that distinguishes between a front-end and back-end network for Ceph. This is not used further here as the cluster setup is not done via CBT. This is expressed with the *use_existing* parameter set to *true*. The clusterid is provided based on what is described in the *ceph.conf* file. The *tmp_dir v*ariable specifies a directory on all the nodes that CBT access under *user* in which intermediate data is stored, mostly consisting of benchmark telemetry. The *pool_profiles* is a YAML list item which allows the user to employ different RADOS pool configurations referred to by name in the benchmark descriptions.

benchmarks enlists various benchmark runs (the amount of repeated execution is specified in *iterations* in the *clusters* section) that are processed in a sequential order. The name refers to a benchmark driver that ships with CBT. In this example, *radosbench* is the driver that executes low-level tests on the core librados layer of Ceph by using the *rados* utility. The parameters specified below are directly handed over to the rados bench call executed on the client systems, whereas list items such as *op_sizes*, *concurrent_ops* or *osd_ra* each trigger individual runs with one of their entries as the respective parameters. As a result, in the example above, the benchmark configuration will launch 4 subsequent benchmarks using the *rados* utility, each with a different *op_size* parameter. The description of these parameters can be found in the help output of the *rados* binary.

The same set of tunables were applied throughout the benchmarks. Ansible-ceph basically ships with the most recommended tunings out of the box. In this benchmark, adjustments to ceph.conf were made only to reflect the cluster setup in the test bed; which is configuration of Front-End and Back-End network and the journal size.

The *ceph.conf* used throughout this benchmark is found at https://github.com/red-hat-storage/dell-ceph-psg.

A functionality that CBT currently does not provide is to sequence the repeated benchmark execution with an increasing amount of parallelism in each run. The goal of this benchmark is also to find the high water mark of the cluster's aggregated throughput, which is the point beyond which the performance increase is becoming zero or negative. Each benchmark scenario is run in 10 iterations, with the first executing the benchmark only from a single client, the second iteration with two clients, the third with three clients in parallel and so on. To achieve this, multiple instances of a benchmark job file were created; each with an increasing amount of clients. The benchmarks were then started with CBT individually by looping over a continuous set of job files.

The job specifications can be found at https://github.com/red-hat-storage/dell-ceph-psg. The structure of the job files are as follows:

- Common configuration
  - Write benchmarks, followed by sequential read benchmarks
  - Tested block sizes: 4M, 1M, 512K and 128K
  - Execution of each benchmark run: 5 minutes
  - 128 concurrent threads per client
  - 1 rados bench instance per client
  - Pool name: cbt-librados
- rados_ec_seq_suite_<1..10>_clients.yml
  - librados-level benchmark on an erasure-coded pool
  - erasure-coding scheme used was 3:2 using the ISA driver

- rados_seq_suite_<1..10>_clients.yml
    - librados-level benchmark on a replicated pool, replication factor 3

In each benchmark, these files are called with CBT in a loop.

CAUTION

When executing multiple instances of CBT subsequently in a loop, as in this benchmark, it is important to note that CBT will delete any existing pool with the same name. This is an asynchronous process that triggers purging object structures on the backend file store. While the command 'ceph osd pool delete' returns, instantly a potential long-running IO-intensive process on the backend is started which may collide with IO issued by a subsequent benchmark run.

It is crucial to either wait for this process to finish before starting a new benchmark, or omit the pool deletion to avoid skewed benchmark results.

In this benchmark, CBT has been modified to implement a new parameter that will cause CBT to skip deletion of existing pools at the beginning of a new benchmark.

4 Benchmark Test Results

Many organizations are trying to understand how to configure hardware for optimized Ceph clusters that meet their unique needs. Red Hat Ceph Storage is able to run on a myriad of diverse industry-standard hardware configurations, however 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, Red Hat and Dell have classified the benchmark results into the following five categories:

1. Compare server throughput in different configurations
2. Compare overall solution price/performance
3. Compare overall solution price/capacity
4. Compare server throughput in replication versus erasure-coded modes
5. Compare server throughput in JBOD and RAID0 modes

These five comparisons are shown in the following sections.

4.1 Comparing Server Throughput in Different Configurations

This test compares the read and write throughputs of all configurations tested.

- Reads: The Ceph-replicated configurations generally yield higher read-throughput compared to erasure-coded configurations. This is because the erasure-coded reads must reassemble data objects from erasure-coded chunks.
- Writes: The Ceph erasure-coded configurations generally yield higher write-throughput compared to replicated configurations, because there is lower write amplification with erasure-coded writes (less MB written).

Summary: The R730xd 16+1, 3xRep configuration provided the best performance for throughput-oriented read/write workloads. However, for write-oriented workloads, the R730xd 16+1, EC3+2 configuration provides superior write performance at significantly lesser costs.

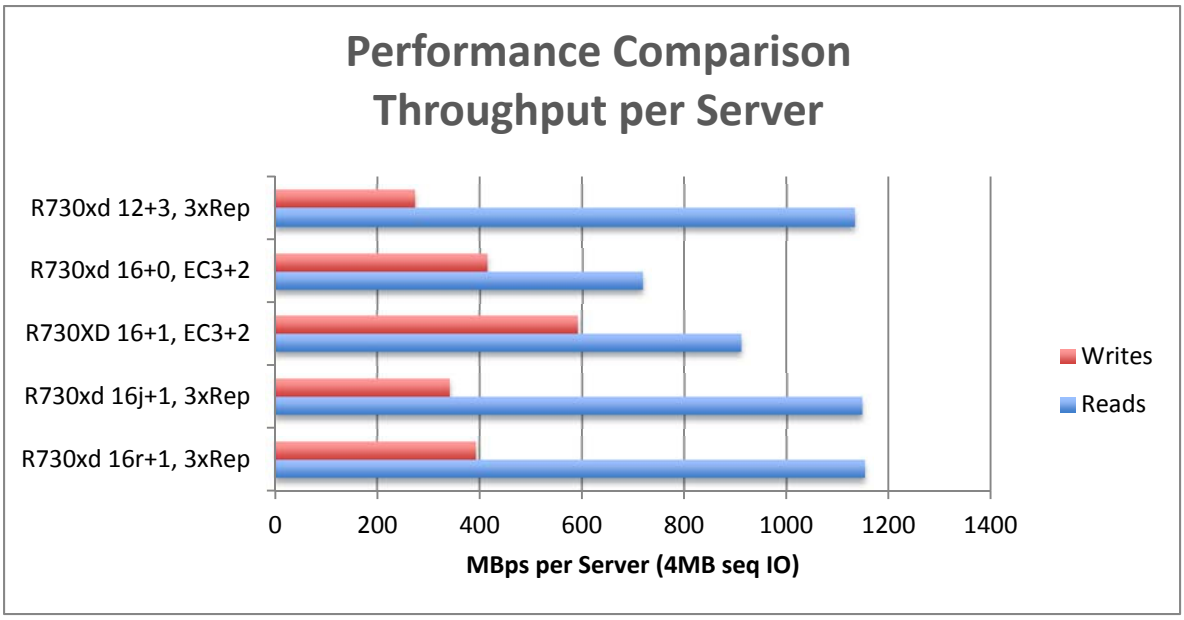


Figure 8 Throughput/server comparison by using different configurations

4.2 Comparing Overall Solution Price/Performance

Based on the highest measured write price/performance, the R730xd 16+1, 3xRep configuration yielded optimal price performance for mixed read/write, and throughput-oriented workloads. However, for read-mostly workloads, the R730xd 12+3, 3xRep configuration is an attractive alternative based on its superior read price/performance.

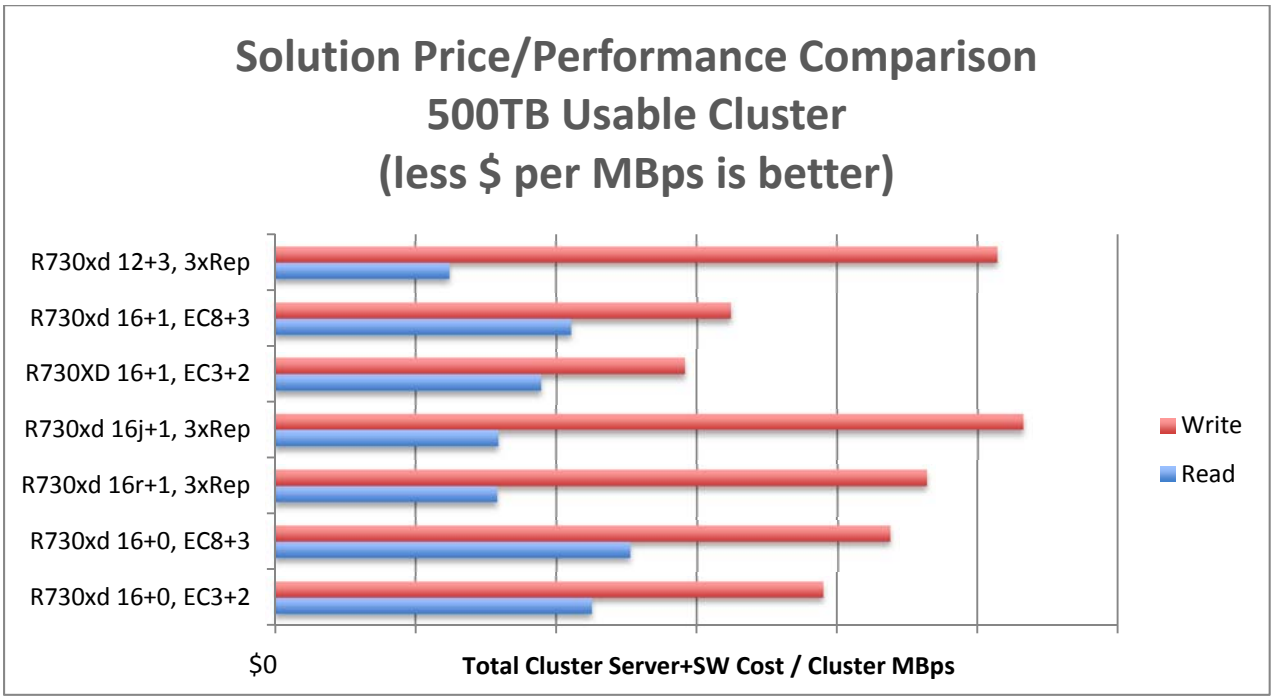


Figure 9 Total solution price/Mbps (4 MB sequential read/write)

Note: On the basis of data protection scheme selected and the average selling price (ASP) of the Red Hat Ceph Storage subscription for the raw storage capacity (TB) required:
The Cluster Server plus software cost is equal to ASP of servers required to provide 500 TB of usable capacity (based on the data protection scheme chosen), and the ASP of Red Hat Ceph Storage subscription ASP for the terabytes required.

4.3 Comparing Overall Solution Price/Capacity

For capacity-archive workloads, erasure-coded configurations are significantly less expensive per GB data archived. Write-heavy capacity-archives should use the R730xd 16+1, EC configuration, because adding an SSD write-journal increases total $/GB by only a small value and increases write performance.

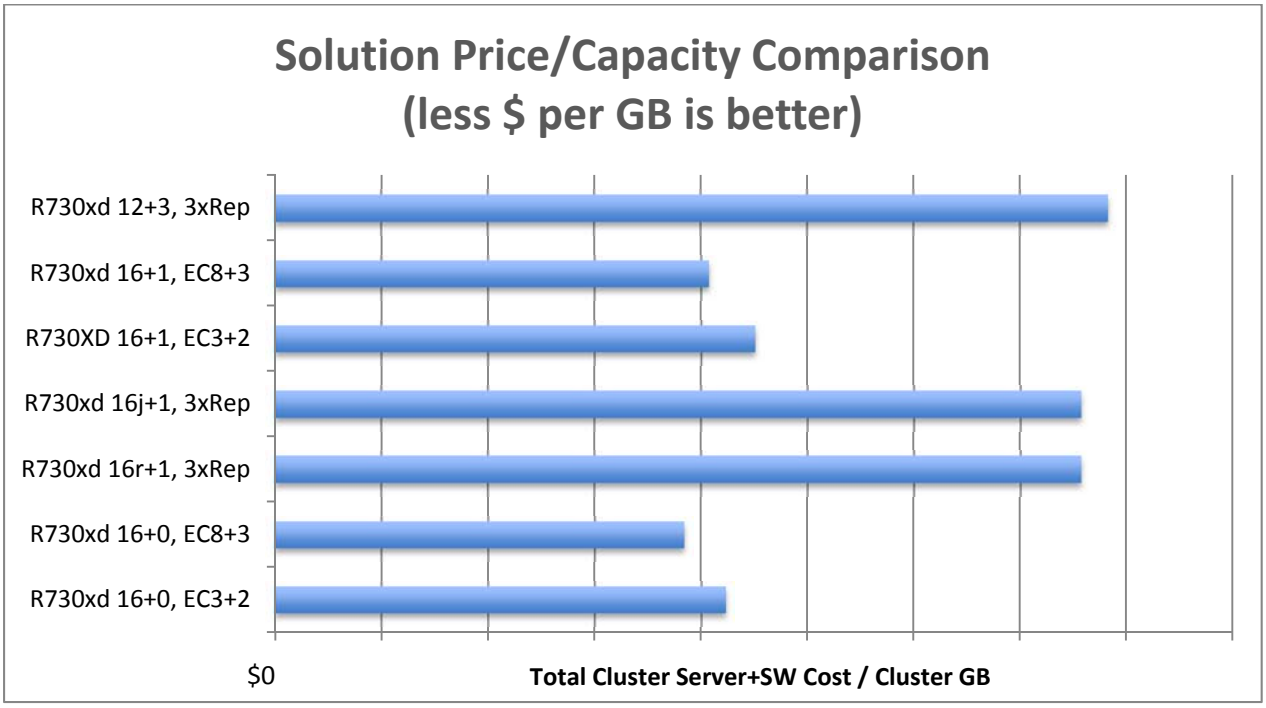


Figure 10 Total solution price/GB data protected

Note: *The Total Cluster Server plus software cost is equal to ASP of servers required to provide 1 PB of usable capacity, and the ASP of Red Hat Ceph Storage subscription ASP for the raw terabyte required.*

4.4 Comparing Server Throughput in Replication vs. Erasure-coded

Keeping everything else constant, replicated reads perform much better than erasure-coded reads. However, erasure-coded writes perform better than replicated writes.

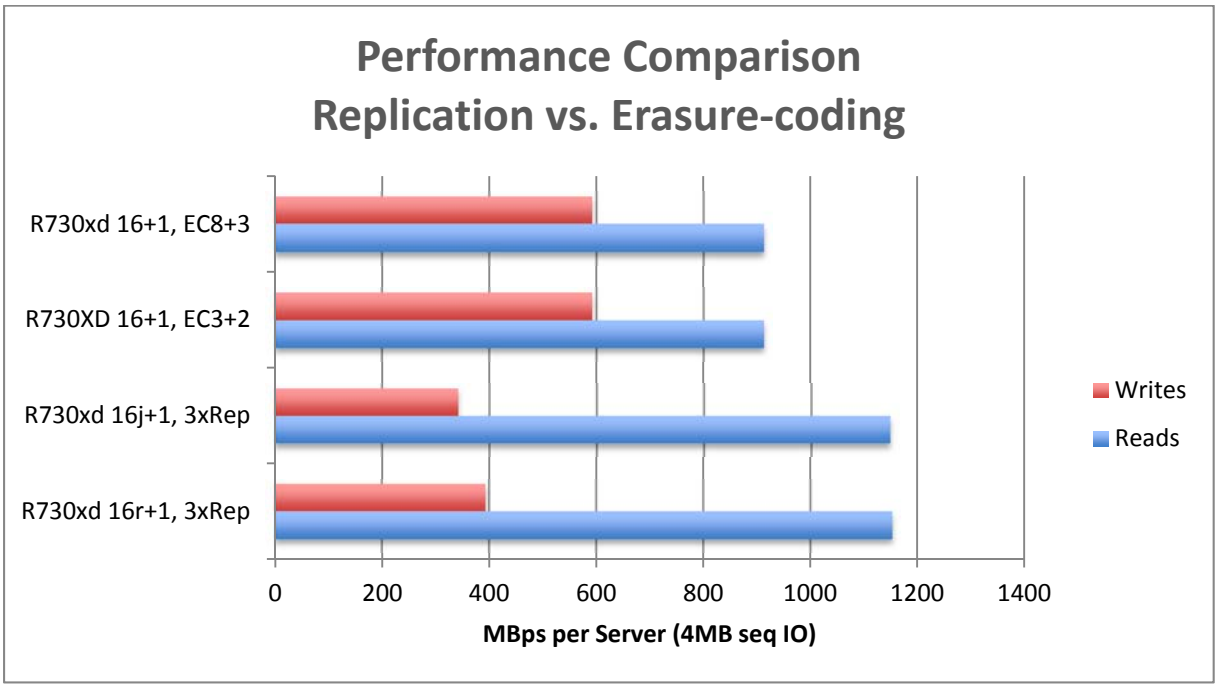


Figure 11 Comparison of server throughput in replication vs. erasure-coded modes

4.5 Comparing Server Throughput in JBOD and RAID0 Modes

The R730xd configurations in this study used the PERC H730 RAID controller. Ceph OSDs are typically configured in a 1:1 ratio with HDDs. Therefore, the RAID controller can either be configured in JBOD mode or with each HDD as single-drive RAID0 volumes.

Summary: RAID0 configurations provide better throughput than JBOD configurations. Therefore, we recommend that R730xd PERC H730 controllers be configured in single-drive RAID0 mode when used with Ceph Storage.

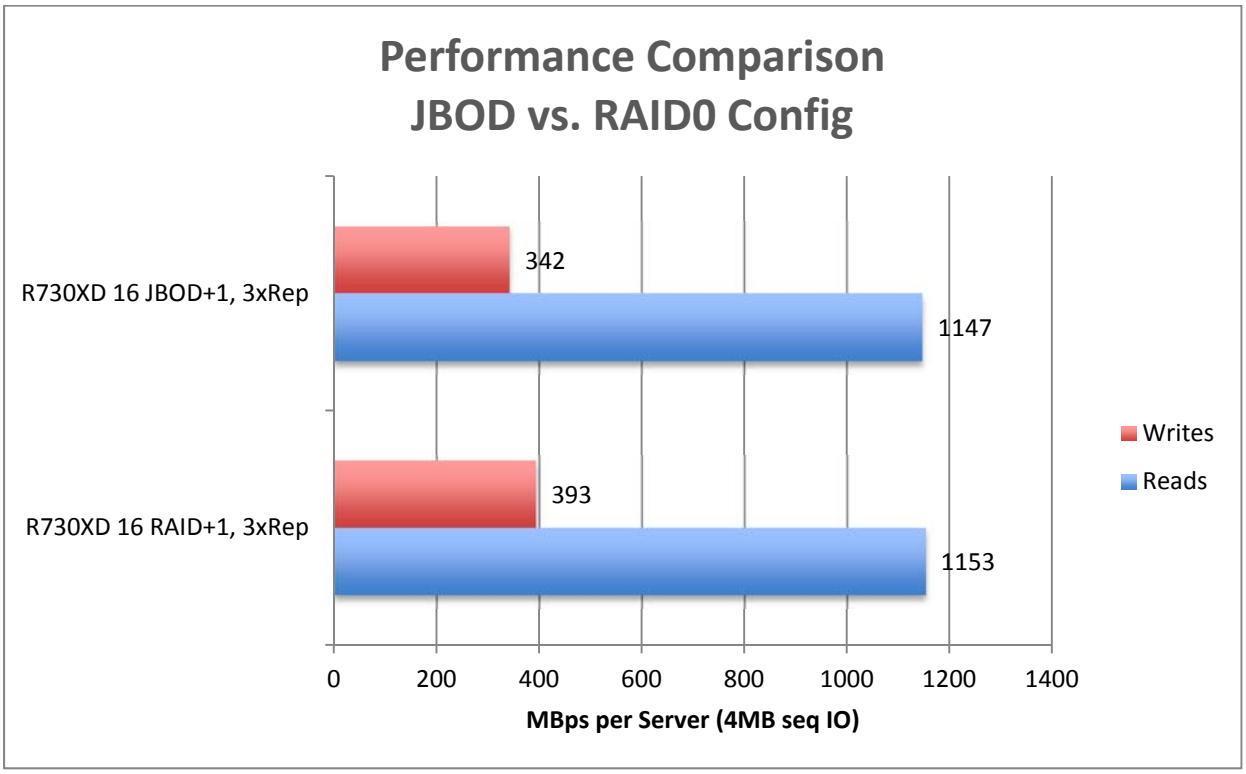


Figure 12 Comparison of server throughput in JBOD vs. RAID0 modes

5 Dell Server Recommendations for Ceph

Ceph operators frequently request simple, optimized cluster configurations for different workload types. Common requests are for throughput-optimized and capacity-optimized workloads. IOPS-intensive workloads on Ceph are also emerging. Based on extensive testing by Red Hat and Dell on various Dell PowerEdge server configurations, this matrix provides general guidance on sizing Ceph clusters built on Dell PowerEdge servers.

Table 13. Dell Server Recommendations for Ceph

Storage Capacity	Extra Small	Small	Medium
Cluster Capacity	100 TB+	500 TB+	1 PB+
Throughput-Optimized	>4x R730xd (8U)	>8x R730xd (16U)	NA
	1x server/2U chassis	1x server/2U chassis	
	16x 6 TB HDD	16x 6 TB HDD	
	1x 800 GB NVMe SSD	1x 800 GB NVMe SSD	
	2x 10 GbE	2x 10 GbE	
	3x Replication	3x Replication	
Cost/Capacity -Optimized	NA	NA	>15x R730xd (30U)
			1x server/2U chassis
			16x 8 TB HDD
			1x HHHL AIC SSD
			2x 10 GbE
			8:3 Erasure-coding

Note: Dell has other reference designs for Ceph storage. For example, Dell built the DSS 7000 to deliver maximum density and performance, providing up to 720 TB of storage capacity by using 8 TB drives. For more info, please visit http://en.community.dell.com/dell-blogs/dell4enterprise/b/dell4enterprise/archive/2016/06/22/taking-open-scale-out-object-storage-to-new-heights

6 Conclusions

After testing different combinations of Red Hat and Ceph Storage on Dell PowerEdge R730xd servers to provide a highly-scalable enterprise storage solution, the following conclusions were made:

- The 3x replication configurations provided high throughput for read operations because the erasure-coded reads have to reassemble data objects from the erasure-coded chunks. However, the erasure-coded configurations demonstrated high throughput for write operations at a lesser price, because of less write amplification.
- The PowerEdge R730xd 16+1 3x replication configuration yielded optimal price for read-write throughput-oriented workloads. The PowerEdge R730xd 12+3 3x replication configuration was superior for read-only workloads.
- The PowerEdge R730xd 16+1 erasure-coded configuration proved to be the choice for write-heavy operations because increasing the storage device does not significantly affect the total $/GB value.
- Replication mode yielded better performance for read operations and the erasure-coded mode proved better for write operations.
- When used with Ceph Storage, Dell-Red Hat recommends the usage of single-drive RAID0 mode on PowerEdge R730xd with PERC H730.

Overall, the PowerEdge R730xd 16+1 in 3x replication mode is recommended by Dell and Red Hat as the configuration that provides optimal price/performance.

7 References

Additional information can be obtained by emailing ceph_info@Dell.com. If you need additional services or implementation help, please contact your Dell sales representative.

- Red Hat Ceph Storage 1.3 Hardware Guide:
  https://access.redhat.com/webassets/avalon/d/Red_Hat_Ceph_Storage-1.3-Hardware_Guide-en-US/Red_Hat_Ceph_Storage-1.3-Hardware_Guide-en-US.pdf

- Ceph:
  http://ceph.com/
  http://docs.ceph.com/docs/master/architecture/

- Dell PowerEdge R730xd:
  http://www.dell.com/us/business/p/poweredge-r730xd/pd