Red Hat Ceph Performance & Sizing Guide

Jose De la Rosa

September 7, 2016

Agenda

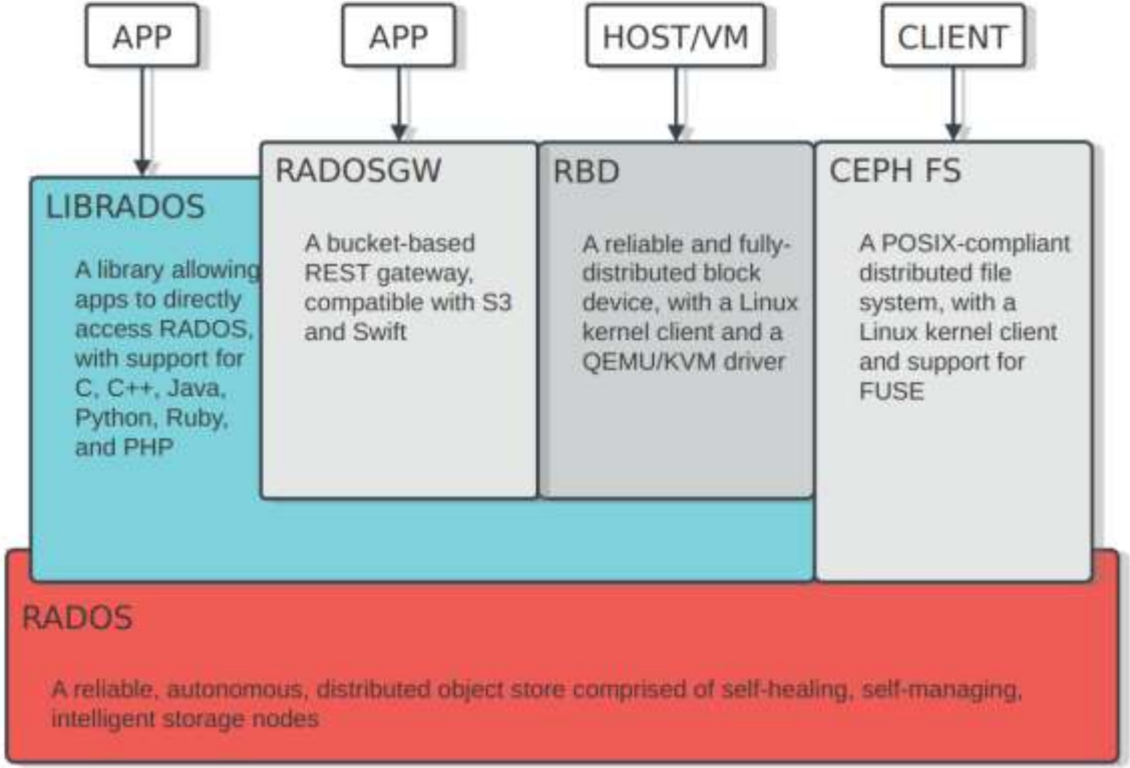
1. Ceph Overview
   a. Architecture
   b. Approach to storing data
2. Test methodology
   a. What was tested
   b. How it was tested & measured
3. Results
   a. Key findings
   b. Recommendations

Ceph Overview

Ceph Overview

1. Open source software defined storage: logical storage services and capabilities are abstracted from the underlying physical storage systems.
2. Provides object, block and file system storage. All data is stored as objects.
3. Massively scalable to thousands of storage nodes.
4. Self-healing with no single point of failure: If a node fails, it is automatically detected and data rebalances to ensure availability.

Client access



APP

APP

HOST/VM

CLIENT

LIBRADOS

A library allowing apps to directly access RADOS, with support for C, C++, Java, Python, Ruby, and PHP

RADOSGW

A bucket-based REST gateway, compatible with S3 and Swift

RBD

A reliable and fully-distributed block device, with a Linux kernel client and a QEMU/KVM driver

CEPH FS

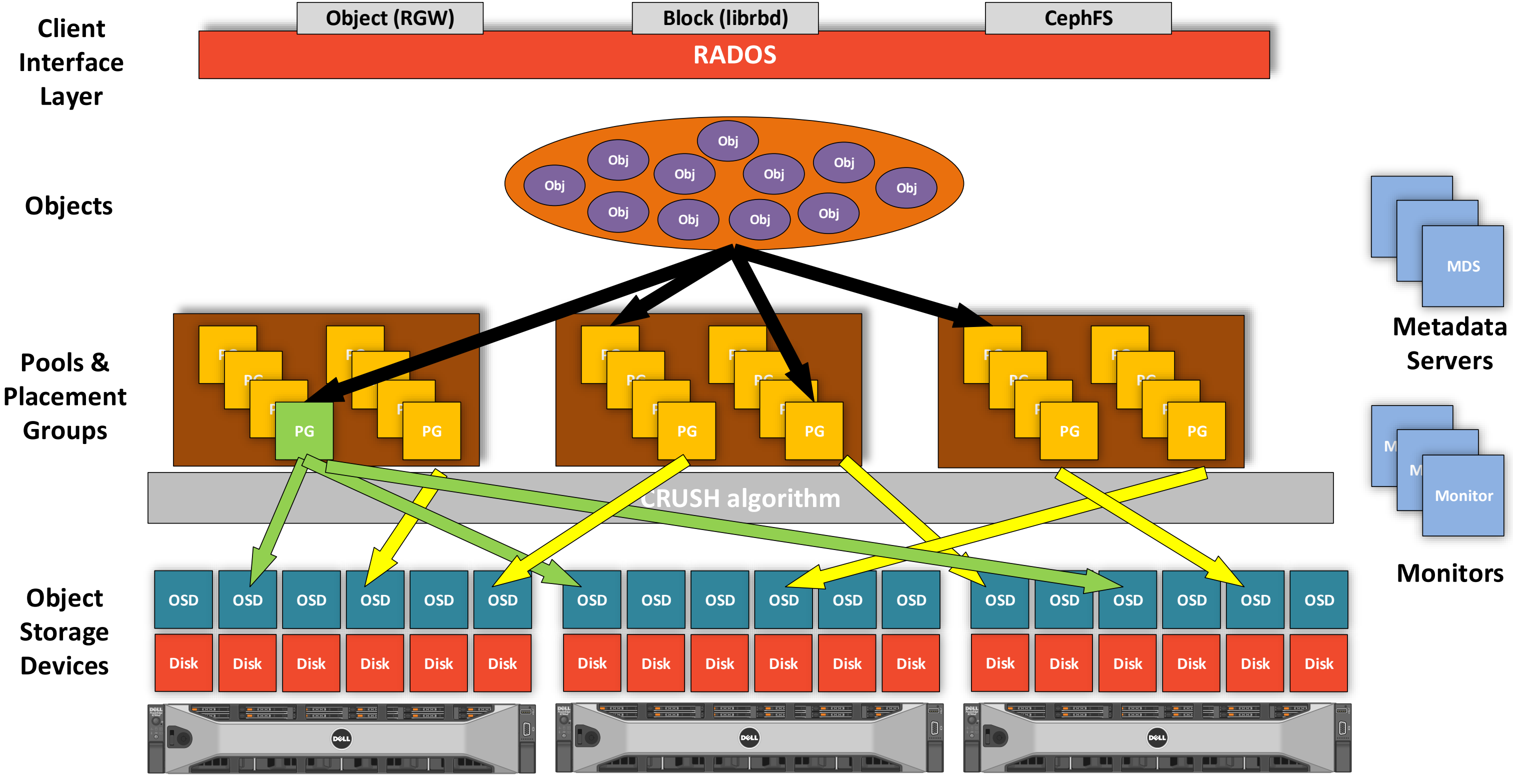
A POSIX-compliant distributed file system, with a Linux kernel client and support for FUSE

RADOS

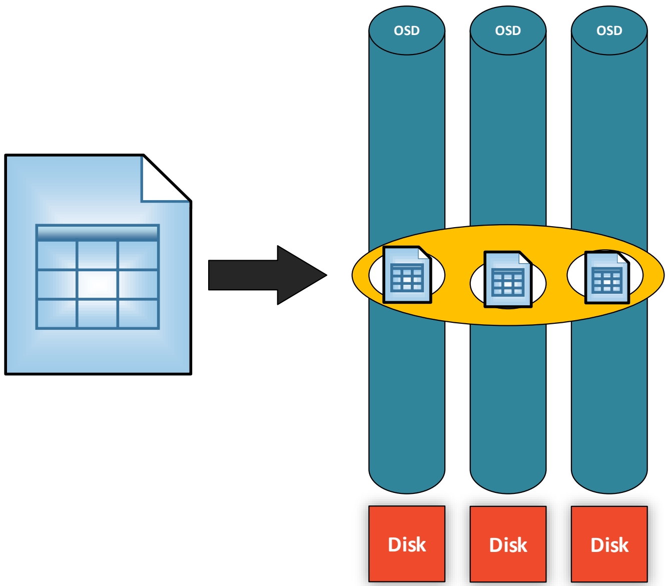
A reliable, autonomous, distributed object store comprised of self-healing, self-managing, intelligent storage nodes

Ross Turk, Red Hat - https://raw.githubusercontent.com/ceph/ceph/master/doc/images/stack.png

Architecture



Storage Protection Method

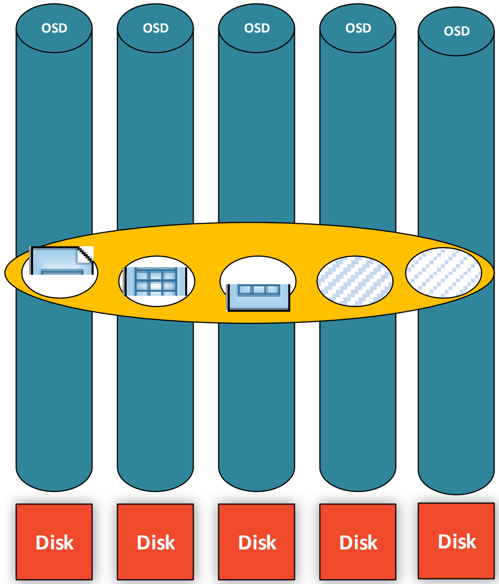


3x Replication

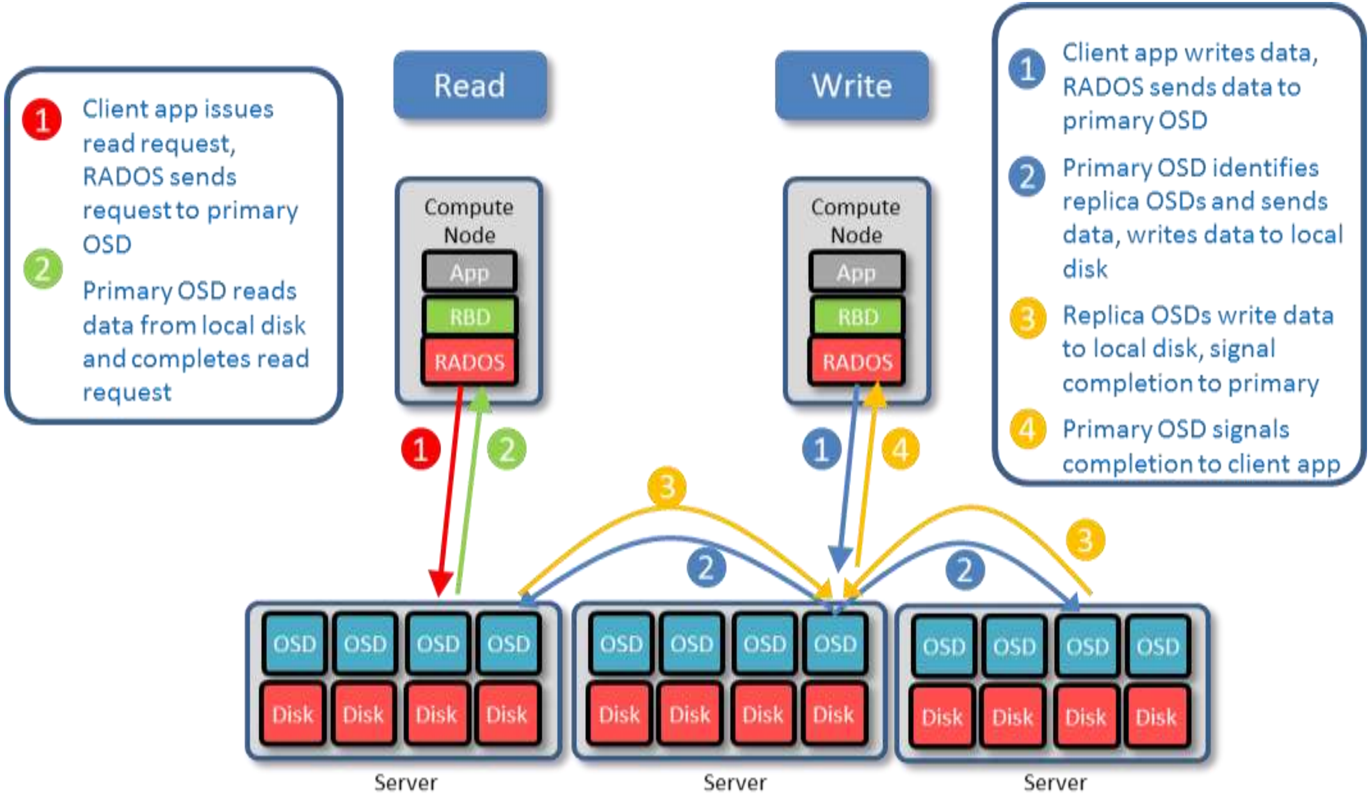
Data copied 3 times
Actual disk usage = 33%

Erasure Coded (3+2)

Data split into 3 + 2 disks used for parity
Actual disk usage = 60%



Reading and Writing Data



1 Client app issues read request, RADOS sends request to primary OSD

2 Primary OSD reads data from local disk and completes read request

1 Client app writes data, RADOS sends data to primary OSD

2 Primary OSD identifies replica OSDs and sends data, writes data to local disk

3 Replica OSDs write data to local disk, signal completion to primary

4 Primary OSD signals completion to client app

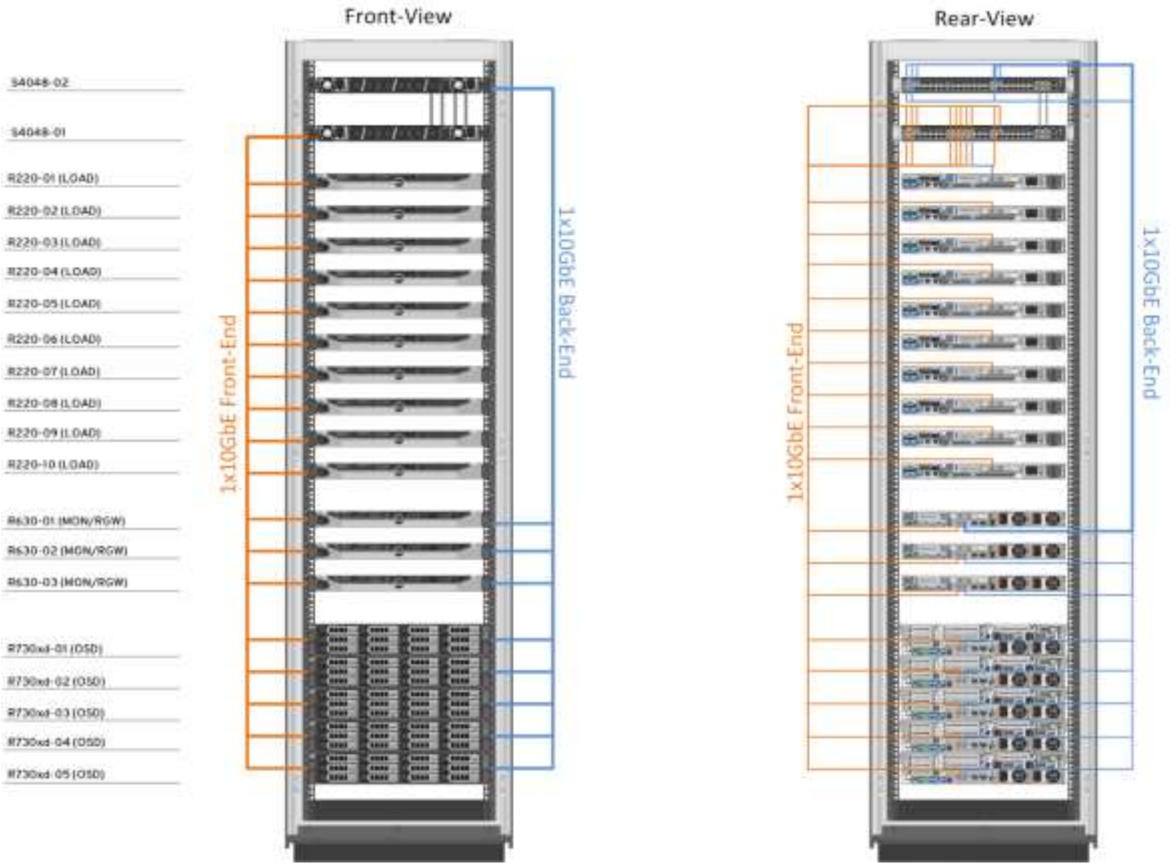
http://www.slideshare.net/LarryCover/ceph-open-source-storage-software-optimizations-on-intel-architecture-for-cloud-workloads

Test Setup & Methodology

Server configuration

Testbed Details			
Ceph tier	**Storage Nodes (5)**	**Monitors (3)**	**Clients (10)**
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 (ST9500620S)	1x Toshiba 50 GB SATA (DT01ACA0)

Network Topology



Configuration Guidance

1. General rules of thumb
   - 1 Core-GHz per OSD
   - SATA/SAS SSD to HDD Ratio: 1:4 - 1:5
   - NVME SSD to HDD Ratio: 1:17-1:18
   - 16GB RAM Baseline + 2-3GB per OSD

2. More details at https://www.redhat.com/en/resources/red-hat-ceph-storage-hardware-configuration-guide

Storage node configurations tested

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 / HBA mode
SATA SSDs	3	0	0
SSD RAID mode	JBOD	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

Benchmarking with CBT

1. For benchmark automation, the open source utility Ceph Benchmarking Tool (CBT) was used.
2. It supports different drivers for examining different layers of the storage stack:
   - **radosbench - uses librados API (used in this study)**
   - librbdfio – test block storage without KVM/QEMU instances
   - kvmrbdfio – test block volumes attached to KVM/QEMU instances
3. Available at https://github.com/ceph/cbt

Factors that influence performance

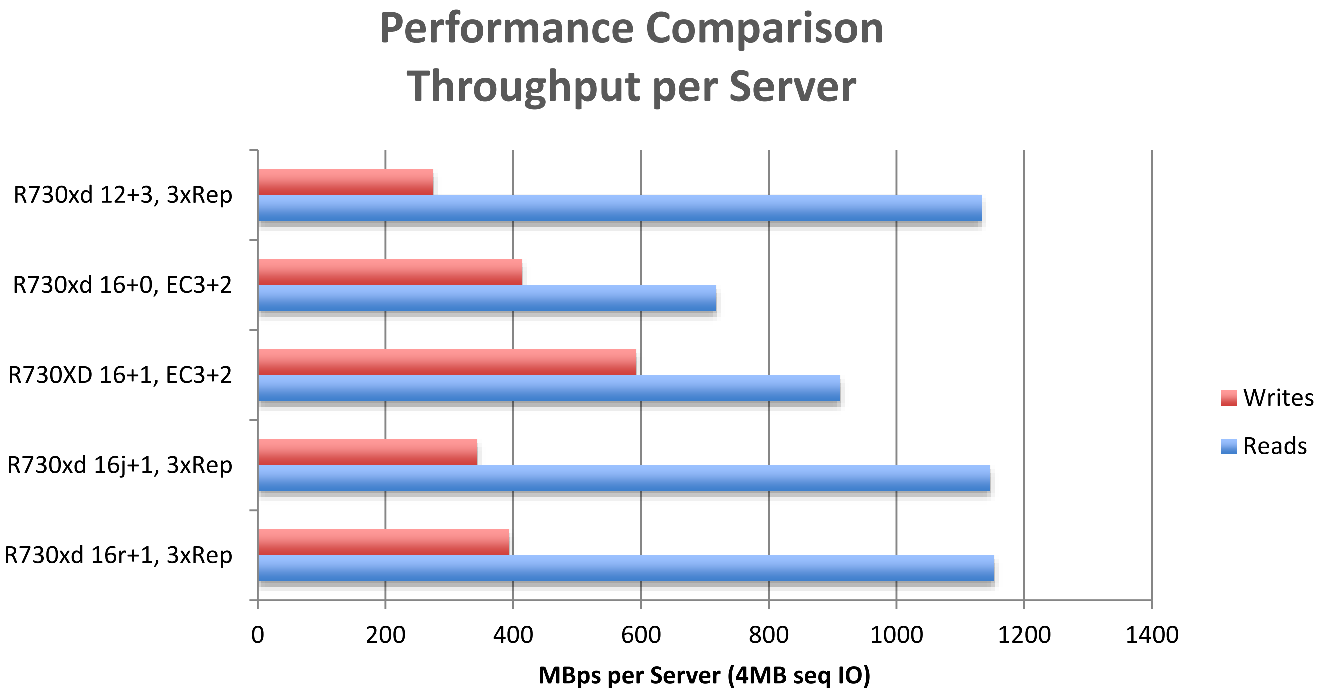
1. Device used for journaling (SSD vs. HDD)
2. RAID0 vs. pass-through (HBA) mode
3. Number of clients (single stream vs. parallel access)
4. Data-protection mode (3-way replication vs. erasure coding)

Test Matrix

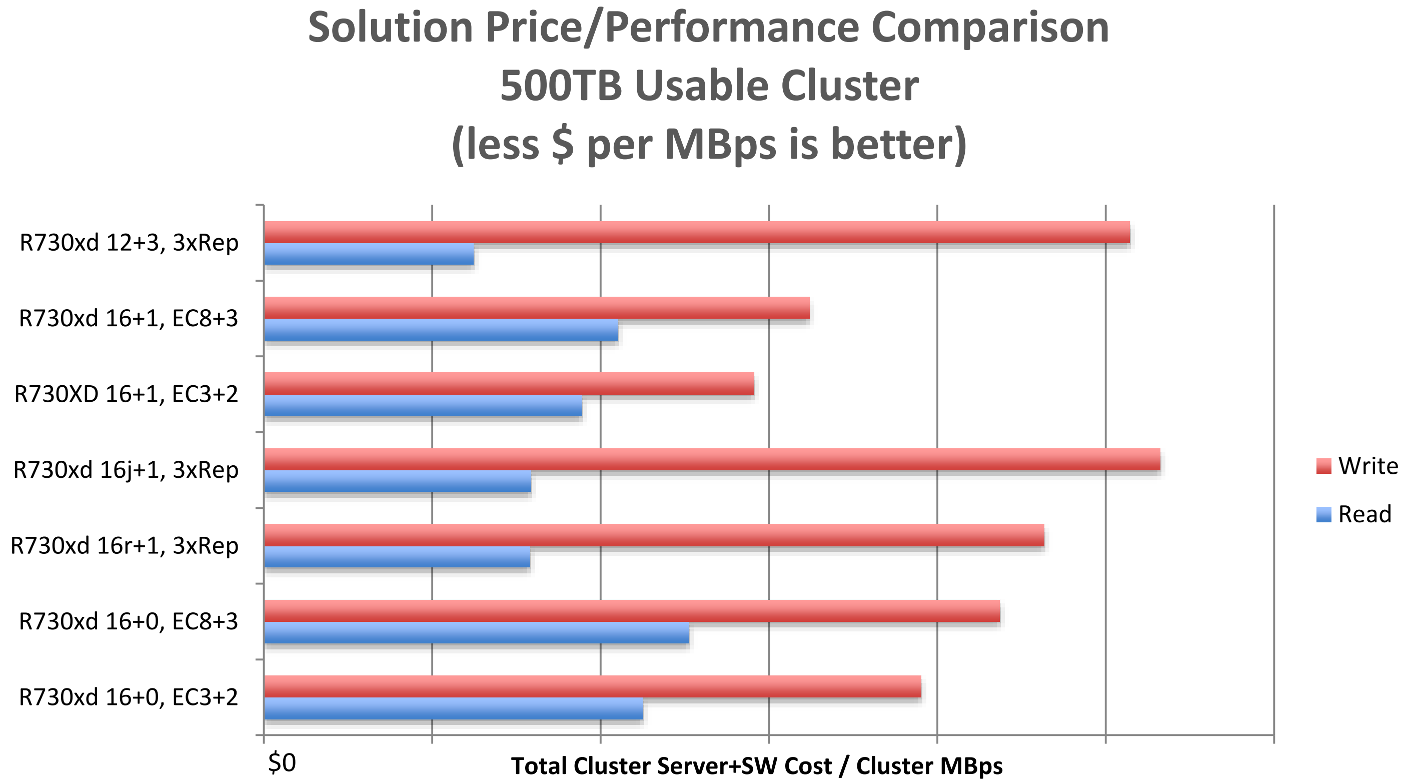
Server configuration	PowerEdge R730xd 12+3, 3xRep	PowerEdge R730xd 16+0, EC3+2	PowerEdge R730xd 16r+1, 3xRep	PowerEdge R730xd 16+1, EC 3+2	PowerEdge R730xd 16j+1, 3xRep
OS disk	2x 500 GB 2.5"	2x 500 GB 2.5"	2x 500 GB 2.5"	2x 500 GB 2.5"	2x 500 GB 2.5"
Data disk type	HDD 7.2K SAS 12Gbps, 4TB	HDD 7.2K 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

Benchmark Test Results

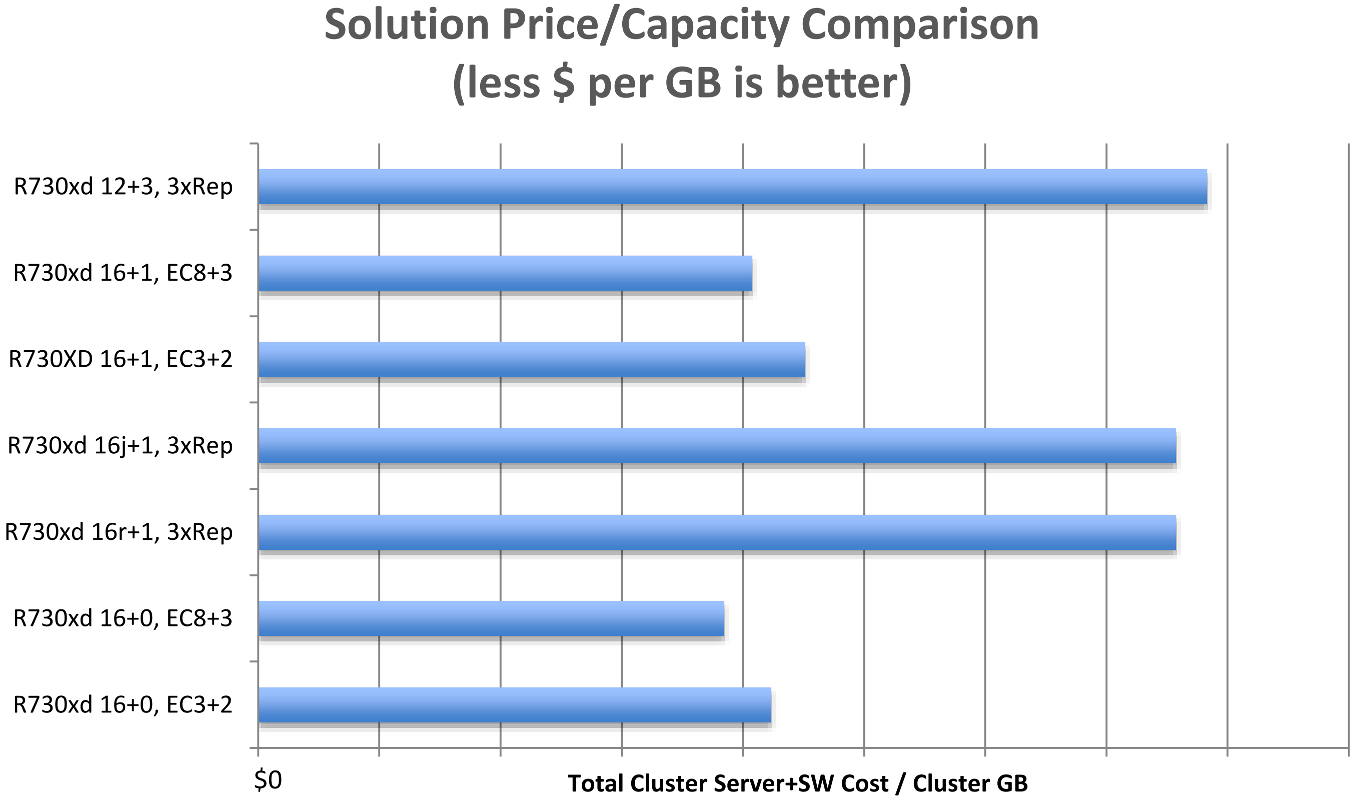
Throughput / server



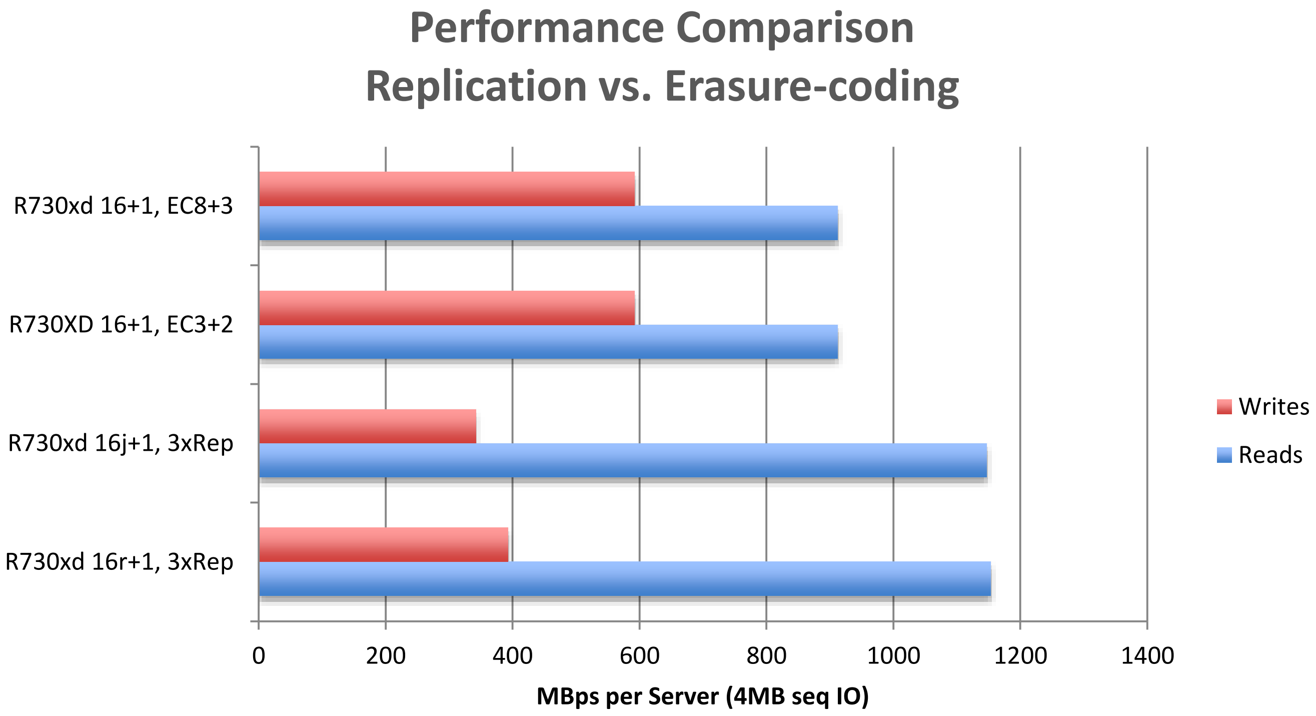
Overall Solution Price/Performance



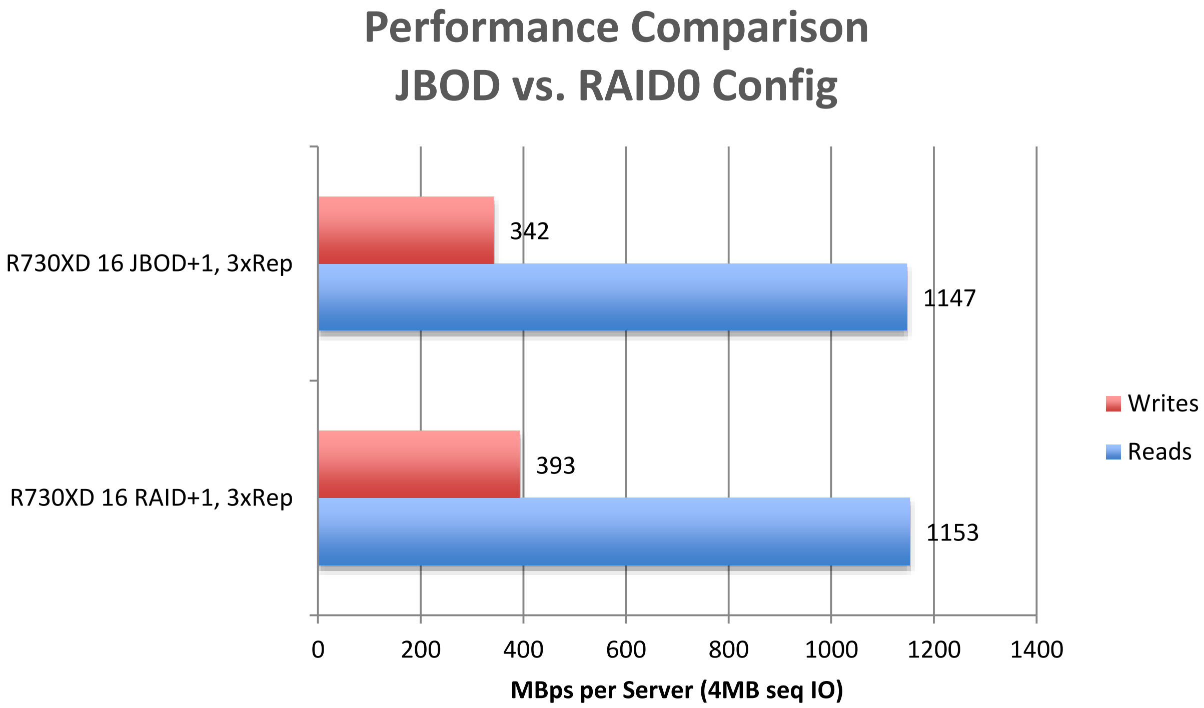
Overall Solution Price Capacity



Replication vs. Erasure Coding



JBOD vs. RAID0



Performance conclusions

1. **Replication mode** yielded better performance for **read operations** and the **erasure-coded** mode proved better for **write operations**.
2. The PowerEdge **R730xd 16+1 3x replication** configuration yielded optimal price for **read-write throughput-oriented workloads**.
3. The PowerEdge **R730xd 12+3 3x replication** configuration yielded optimal price for **read-only throughput-oriented workloads**.
4. The PowerEdge **R730xd 16+1 erasure-coded** configuration proved to be the choice for **write-heavy operations**.
5. When used with Ceph Storage, Dell & Red Hat recommend the usage of single-drive RAID0 mode on PowerEdge R730xd with PERC H730.

Sizing Recommendations

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

Observations

1. Obey SSD to HDD ratio
2. Hardware matters, look at RAID controllers if you use HDDs
3. Don’t use RAID controllers on SSDs
4. SSD sequential write bandwidth becomes a bottleneck
5. Random workloads should go on Flash-only
6. 10GbE Bonding not necessary with <=16 drives

Recommended Reading

Dell PowerEdge R730xd Red Hat Ceph Storage Performance and Sizing Guide

http://en.community.dell.com/techcenter/cloud/m/dell_cloud_resources/20442913

Enterprise Ceph: Every Way, Your Way

https://www.redhat.com/files/summit/session-assets/2016/SS88828-enterprise-ceph_every-way-your-way.pdf