RED HAT® STORAGE

redhat. | QCT Quanta CLOUD TECHNOLOGY

TECHNOLOGY DETAIL

PERFORMANCE AND SIZING GUIDE: RED HAT CEPH STORAGE ON QCT SERVERS

QCT (Quanta Cloud Technology) offers a family of servers for building different types of scale-out storage clusters based on Red Hat Ceph Storage—each optimized to suit different workload and budgetary needs.

Throughput-optimized configurations offer impressive performance with both standard and high-density servers.

Cost or capacity-optimized configurations provide industry-leading price and density with innovative QCT server platforms that are easy to deploy rapidly at scale.

Extensive Red Hat and QCT testing helps take the risk out of deploying scale-out storage solutions based on Ceph.

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

redhat.com

EXECUTIVE SUMMARY

Ceph users frequently request simple, optimized cluster configurations for different workload types. Common requests are for throughput-optimized and capacity-optimized workloads, but IOPS-intensive workloads on Ceph are also emerging. To address the need for performance, capacity, and sizing guidance, Red Hat and QCT (Quanta Cloud Technology) have performed extensive testing to characterize optimized configurations for deploying Red Hat Ceph Storage on a range of QCT servers.

TABLE OF CONTENTS

DOCUMENT PURPOSE

The purpose of this document is to characterize and compare the performance of Red Hat® Ceph Storage on various QCT (Quanta Cloud Technology) 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 combined with performance results. The testing methodology is also provided, and is based on the standardized Ceph Benchmarking Tool, available in a GitHub repository under the Ceph organization.[1] The study described herein largely used off-the-shelf hardware and software components, and did not make a detailed study of changing various configuration settings within the kernel, Ceph, XFS®, or the network.

INTRODUCTION

As the need for storage escalates, enterprises of all kinds are seeking to emulate efficiencies achieved by public cloud providers—with their highly successful software-defined cloud datacenter models based on standard servers and open source software. At the same time, the $35 billion storage market is undergoing a fundamental structural shift, with storage capacity returning to the server following decades of external NAS and SAN growth.[2] Software-defined scale-out storage has emerged as a viable alternative, where standard servers and independent software unite to provide data access and highly available services across the enterprise.

The combination of QCT servers and Red Hat Storage software squarely addresses these industry trends, and both are already at the heart of many public cloud datacenters. QCT is reinventing datacenter server technology to boost storage capacity and density, and redesigning scalable hardware for cloud applications. As the world's largest enterprise software company with an open source development model, Red Hat has partnered with several public cloud providers to provide Ceph and Gluster storage software in production environments. Together, QCT servers and Red Hat Ceph Storage provide software-defined storage solutions for both private and public clouds, helping to accelerate the shift away from costly, proprietary external storage solutions.

Red Hat Ceph Storage significantly lowers the cost of storing enterprise data and helps enterprises manage exponential data growth. The software is a robust, petabyte-scale storage platform for enterprises 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 archival, 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.

Running Red Hat Ceph Storage on QCT servers provides open interaction with a community-based software development model, backed by the 24x7 support of the world's most experienced open source software company. Use of standard hardware components helps ensure low costs, while QCT's innovative development model lets organizations iterate more rapidly on a family of server designs optimized for different types of Ceph workloads. Unlike scale-up storage solutions, Red Hat Ceph Storage on QCT servers lets organizations scale out to thousands of nodes, with the ability to scale storage performance and capacity independently, depending on the needs of the application and the chosen storage server platform.

1 *https://github.com/ceph/cbt*
2 *IDC Worldwide Quarterly Disk Storage Systems Tracker, June 5, 2015*

WORKLOAD-OPTIMIZED SCALE-OUT STORAGE CLUSTERS

Red Hat Ceph Storage on QCT servers can be easily optimized and sized to serve specific workloads through a flexible choice of systems and components.

CHARACTERIZING STORAGE WORKLOADS

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.
- As IOPS-optimized workloads emerge on Ceph, high-IOPS server pools can also be added to a Ceph cluster.

Table 1 provides the criteria used to identify optimal Red Hat Ceph Storage cluster configurations on QCT-based storage servers. These categories are provided as general guidelines for hardware purchase and configuration decisions, and 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.

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

SIZING SUMMARY FOR WORKLOAD-OPTIMIZED CLUSTERS

Red Hat Ceph Storage is able to run on myriad diverse hardware configurations. The purpose of this reference architecture document is to help organizations evaluate key architectural concepts with corresponding test results in order to architect appropriately sized and optimized Red Hat Ceph Storage clusters on QCT servers. To this end, Red Hat and QCT architects conducted extensive Ceph testing on various configurations of two QCT servers.

- **QCT QuantaGrid D51PH-1ULH server.** Ideal for smaller-capacity clusters, the compact 1 rack unit (1U) QuantaGrid D51PH-1ULH server provides 12 hot-swappable disk drives and four additional hot-swappable solid state drives (SSDs).
- **QCT QuantaPlex T21P-4U server.** The QuantaPlex T21P-4U server is configurable as a single-node (up to 78 HDDs) or dual-node system (up to 35 HDDs per node), maximizing storage density to meet the demand for growing storage capacity in hyperscale datacenters.

Through testing, engineers identified a variety of throughput-optimized and cost/capacity-optimized configurations, sized to fit the needs of different cluster sizes. Table 2 summarizes different workload-optimized configurations with usable storage capacities ranging from 100 TB to more than 2 petabytes (PB). The remainder of this document describes how these configurations were selected and tested.

TABLE 2. WORKLOAD OPTIMIZED CONFIGURATIONS OF QCT STORAGE SERVERS.

	EXTRA SMALL (100 TB*)	SMALL (500 TB*)	MEDIUM (>1 PB*)	LARGE (>2 PB*)
IOPS- optimized	Future direction	Future direction	Future direction	NA
Throughput-optimized	7x QuantaGrid D51PH-1ULH • 7U • 12x 4 TB HDDs • 3x SSDs • 2x 10 GbE • 3x replication	32x QuantaGrid D51PH-1ULH • 32U • 12x 4 TB HDDs • 3x SSDs • 2x 10 GbE • 3x replication	11x QuantaPlex T21P-4U/Dual • 44U • 2x 35x 4 TB HDDs • 2x 2x PCIe SSDs • 2x 1x 40 GbE • 3x replication	22x QuantaPlex T21P-4U/Dual • 88U • 2x 35x 4 TB HDDs • 2x 2x PCIe SSDs • 2x 1x 40 GbE • 3x replication
Cost/capacity-optimized	NA	8x QuantaGrid D51PH-1ULH • 8U • 12x 8 TB HDDs • 0x SSDs • 2x 10 GbE • Erasure coding (4:2)	4x QuantaPlex T21P-4U/dual • 16U • 2x 35x 6 TB HDDs • 0x SSDs • 2x 2x 10 GbE • Erasure coding (4:2)	7x QuantaPlex T21P-4U/mono • 28U • 78x 6 TB HDDs • 0x SSDs • 2x 10 GbE • Erasure coding (4:2)

* *Usable storage capacity*

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, and complexity can compromise cost-effectiveness. In contrast, scale-out storage infrastructure based on clustered storage servers has emerged as a way to deploy cost-effective and manageable storage at scale, with Ceph among the leading solutions.[3] In fact, cloud storage companies are already using Ceph at near exabyte scale, with expected continual growth. 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 and retrieve data
- Replicate data
- Monitor and report on cluster health
- Redistribute data dynamically (remap and backfill)
- Ensure data integrity (scrubbing)
- Recover from failures

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 storage and retrieval of objects.

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 1) which:

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

3 Ceph is and has been the leading storage for OpenStack according to several semi-annual OpenStack user surveys.
4 http://yahooeng.tumblr.com/post/116391291701/yahoo-cloud-object-store-object-storage-at

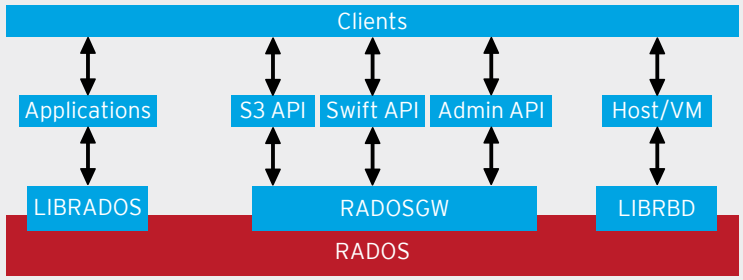

Figure 1. The Reliable Autonomic Distributed Object Store (RADOS) is the foundation of the Ceph storage cluster.

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 2 illustrates the overall Ceph architecture, with concepts that are described in the sections that follow.

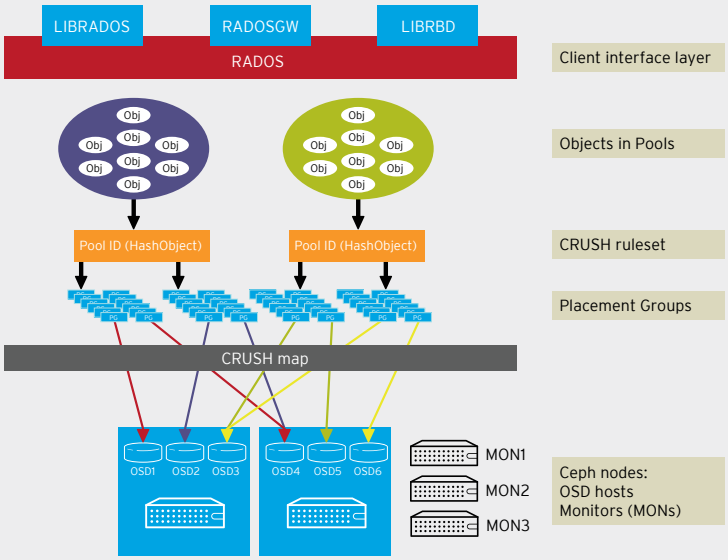

Figure 2. Clients write to Ceph storage pools while the CRUSH ruleset determines how placement groups are distributed across object storage daemons (OSDs).

- **Pools.** A Ceph storage cluster stores data objects in logical 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 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 "take root" at any position in the CRUSH hierarchy, allowing placement 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 provide a means of creating 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 (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 filesystem on a physical hard disk drive.
- **Ceph monitors (MONs)**. Before Ceph clients can read or write data, they must contact a Ceph 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, and 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-coded storage pools**. Erasure coding provides a single copy of data plus parity, and it 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 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 changes.
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 data and parity chunks to secondary OSDs, while storing one data chunk locally.

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

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

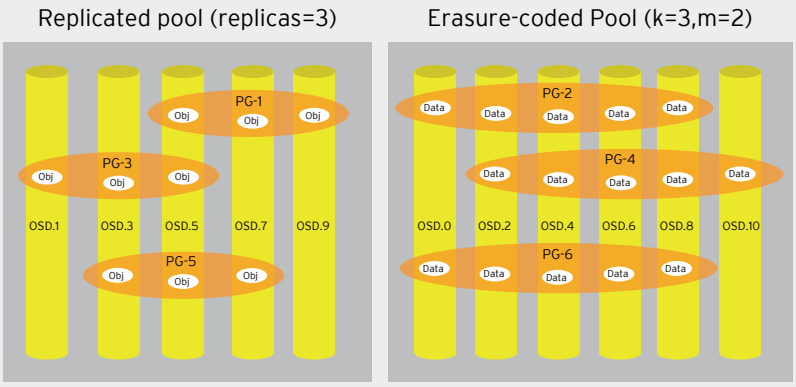

Figure 3. Graphical representation of relationships among Ceph storage pools, PGs, and OSDs, for both replicated pools and erasure-coded pools.

REFERENCE ARCHITECTURE ELEMENTS

The following sections discuss the overall architecture of the Red Hat and QCT reference architecture, as well as key technical aspects of the principal components.

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 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 object storage, or vice versa. Ceph provides seamless access to objects with either native language bindings or through the RADOS gateway, a REST interface that's compatible with a large ecosystem of 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, but with hardware flexibility, massive scalability, and fault tolerance.
- **Industry-leading expertise.** Red Hat Ceph Storage is backed by the experience and expertise of Ceph's creators and primary sponsors, and engineers with decades of experience with the Linux kernel and filesystems. Red Hat also offers hands-on training and professional services.

QCT SERVERS FOR CEPH

Scale-out storage requires capable and scalable server platforms that can be selected and sized to meet the needs of specific workloads. Driven by social media, mobile applications and the demands of hyperscale datacenters, storage servers and storage platforms must provide increasing capacity and performance to store increasing volumes of data with ever-longer retention periods. QCT servers are offered in a range of configurations to allow optimization for diverse application workloads. Servers range from dense, single rack unit (1U) systems to models providing massive storage capacity using only four rack units. Servers enable Ceph journaling by providing expansion slots for PCIe SSDs or specially designed spaces for SSDs. SSD caching is important to accelerate IOPS and throughput in many software-defined storage technologies. Illustrated in Figure 4, Red Hat and QCT tested two QCT servers optimized for Ceph workloads:

- **QCT QuantaGrid D51PH-1ULH server**. With an ideal compact design for smaller Ceph clusters, the QCT QuantaGrid D51PH-1ULH is delivered in an ultra-dense 1U package, with 12 hot-swappable disk drives and four hot-swappable SSDs. The system is the first worldwide to employ the Intel Xeon Processor E5-2600 v3 family in a 1U, 12-drive platform. Importantly, the four SSDs are supported without sacrificing space for 12 disk drives. The QCT QuantaGrid D51PH-1ULH provides a perfect ratio of regular disk drives to SSDs—4:1 when enabling Ceph OSD journals. The server's innovative hot-swappable drive design means no external cable management arm is required—significantly reducing system deployment and rack assembly time. As a result, IT administrators can service drives with minimal effort or downtime.
- **QCT QuantaPlex T21P-4U server**. Capable of delivering up to 620TB of storage in just one system, the QuantaPlex T21P-4U efficiently serves the most demanding cloud storage environments. The server maximizes storage density to meet the demand for growing storage capacity in hyperscale datacenters. Two models are available: a single storage node can be equipped with 78 hard disk drives (HDDs) to achieve ultra-dense capacity and low cost per gigabyte, or the system can be configured as dual nodes, each with 35 HDDs to optimize rack density. Along with support for two PCIe Gen3 slots for PCIe-based SSD, the server offers flexible and versatile I/O expansion capacity. The ratio of regular drives to SSDs of 17.5:1 in a dual-node configuration boosts Ceph performance and IOPS. The QCT QuantaPlex T21P-4U server features a unique, innovative screw-less hard drive carrier design to let operators rapidly complete system assembly, significantly reducing deployment and service time.

Both servers offer flexible networking options to support workload requirements based on application needs. QCT mezzanine cards provide varied Ethernet connectivity, ranging from Gigabit Ethernet (GbE) to 10 GbE and 40 GbE.

QCT QuantaGrid D51PH-1ULH

QCT QuantaPlex T21P-4U

Figure 4. QCT storage servers are ideal for running Red Hat Ceph Storage.

ARCHITECTURAL DESIGN CONSIDERATIONS

Key design considerations can help organizations shape Ceph cluster server and network architectures. Each of these topics is intended to be a conversation between peer architects. Later sections describe testing results that help to illustrate how some of these design considerations affect Ceph cluster price/performance.

QUALIFYING THE NEED FOR SCALE-OUT STORAGE

Not every storage situation calls for scale-out storage. When requirements include several of the following needs, they probably point to a good fit for scale-out storage.

- **Dynamic storage provisioning**. By dynamically provisioning capacity from a pool of storage, organizations are typically building a private storage cloud, mimicking services such as Amazon Simple Storage Service (S3) for object storage or elastic block storage (EBS).
- **Standard storage servers**. Scale-out storage employs storage clusters built from industry-standard x86 servers rather than proprietary storage appliances, allowing incremental growth of storage capacity and/or performance without forklift appliance upgrades.
- **Unified name spaces**. Scale-out storage allows pooling storage across tens, hundreds, or even thousands of storage servers in one or more unified namespaces.
- **High data availability**. Scale-out storage provides high-availability of data across what would otherwise be "server storage islands" within the storage cluster.
- **Independent multidimensional scalability**. Unlike typical NAS and SAN devices that may exhaust throughput or IOPS before they run out of capacity, scale-out storage allows organizations to add storage performance or capacity incrementally by independently adding more storage servers or disks as required.

DESIGNING FOR THE TARGET WORKLOAD

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

If needs are more performance-oriented, IOPS and throughput are often taken into consideration. Historically, Ceph has performed very well with high-throughput workloads, and has been widely deployed for these use cases. Use cases are frequently characterized by large-block, asynchronous, sequential I/O (e.g., 1 Mb sequential I/O).

In contrast, high IOPS workloads are frequently characterized by small-block synchronous random I/O (e.g., 4 Kb random I/O). At present, the use of Ceph for high IOPS workloads is emerging. Though historically, Ceph has not been recommended for traditional OLTP workloads, the use of Ceph for MySQL and PostgreSQL workloads is gaining momentum.

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

To simplify configuration and testing choices and help structure optimized cluster configurations, Red Hat categorizes workload profiles as follows:

- IOPS-optimized clusters
- Throughput-optimized clusters
- Cost/capacity-optimized clusters

CHOOSING A STORAGE ACCESS METHOD

Choosing a storage access method is another important design consideration. For example, Ceph block storage is 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 due to the significant difference in media costs. Note that while CephFS distributed file storage is not yet supported on Red Hat Ceph Storage as of this writing, file systems are routinely created on top of Ceph block devices.

IDENTIFYING STORAGE CAPACITY

Identifying storage capacity may seem trivial, but it can have a distinct effect on the chosen target server architecture. In particular, storage capacity must be weighed in concert with considerations such as fault domain risk tolerance. For example, if an organization is designing a small, half-petabyte cluster, minimum server fault domain recommendations will preclude the use of ultra-dense storage servers in the architecture, to avoid unacceptable failure domain risk on a small number of very large nodes.

SELECTING A DATA PROTECTION METHOD

Ceph offers two data protection schemes: replication and erasure coding. 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.

Ceph block storage defaults to 3x replicated pools and is not supported directly on erasure-coded pools. Ceph object storage is supported on either replicated or erasure-coded pools. Depending upon 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.

DETERMINING FAULT DOMAIN RISK TOLERANCE

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

Primary factors for weighing fault domain risks include:

- **Reserving capacity for self-healing.** When a storage node fails, Ceph self-healing begins after a configured time period. 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 due to using algorithmic placement). As a result, each node in a cluster should operate at less than 80% of total capacity.
- **Accommodating impact on performance.** During self-healing, a percentage of cluster throughput capacity will be diverted to reconstituting object copies from the failed node on the surviving nodes. The percentage of cluster performance degradation is a function of the number of nodes in the cluster and how Ceph is configured.

Red Hat and QCT recommend the following minimum cluster sizes:

- **Supported minimum cluster size:** Three storage (OSD) servers, suitable for use cases with higher risk tolerance for performance degradation during recovery from node failure
- **Recommended minimum cluster size (throughput-optimized cluster):** 10 storage (OSD) servers
- **Recommended minimum cluster size (cost/capacity-optimized cluster):** 7 storage (OSD) servers

TESTED CONFIGURATIONS

Two separate cluster configurations were constructed and tested by Red Hat and QCT.

QUANTAGRID D51PH-1ULH CONFIGURATION

As shown in Figure 5, five 1U QuantaGrid D51PH-1ULH servers were connected to both a 10 GbE cluster network as well as a 10 GbE public network. Ten client nodes were likewise attached to the public network for load generation. Each QuantaGrid D51PH-1ULH server was configured with:

- Processors: Two Intel Xeon processor E5-2660 V3 10-core 2.66 GHz
- Memory: 4x 16 GB 2133 MHz DDR4 RDIMM 1.2V
- SAS controller: QCT SAS Mezz LSI 3008
- Network controller: QCT Intel 82599ES dual-port 10 GbE SFP+ OCP mezzanine
- Onboard storage: Flash/M SATADOM 32 GB
- Write journal: 0x or 3x Intel SSD DC S3710 200 GB, 2.5-inch SATA 6 Gb/s SSD, MLC flash
- Hard disk drives: 12x Seagate 3.5-inch SAS 2 TB 7.2K RPM

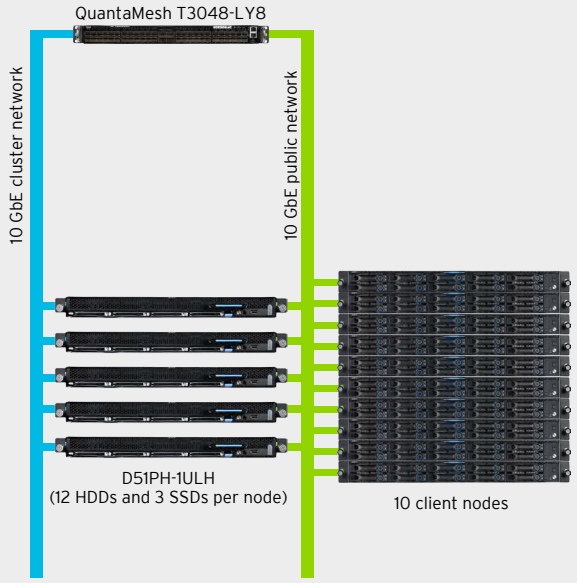

Figure 5. QuantaGrid D51PH-1ULH servers were tested with cluster and public 10 GbE networks.

QUANTAPLEX T21P-4U CONFIGURATION

As shown in Figure 6, two 4U QuantaPlex T41P-4U/Dual Node servers were connected, each with dual 40 GbE interfaces (one per node) to a shared public network. Sixteen client nodes were likewise attached to the public network via 10 GbE for load generation. Each QuantaPlex T41P-4U/Dual Node server was configured with:

- Processors: (2x 2) Intel Xeon processor E5-2660 V3 10-core 2.3 GHz
- Memory: (2x 8) 16 GB 2133 MHz DDR4 RDIMM 1.2V
- SAS controller: 2x QCT SAS Mezz LSI 3008 SAS controller
- Network controller: 2x QCT Intel 82599ES dual-port 10 GbE SFP+ OCP mezzanine or 2x QCT Mellanox ConnectX-3 EN 40 GbE SFP+ single-port OCP mezzanine
- Onboard storage: 2x Intel SSD DC S3510 120 GB, 2.5-inch SATA 6 Gb/s, MLC flash
- Write journal: 0x or 3x Intel SSD DC P3700 800 GB, 1/2 height PCIe 3.0, MLC flash
- Hard disk drives: (2x 35) Seagate 3.5-inch SAS 6 TB 7.2K RPM

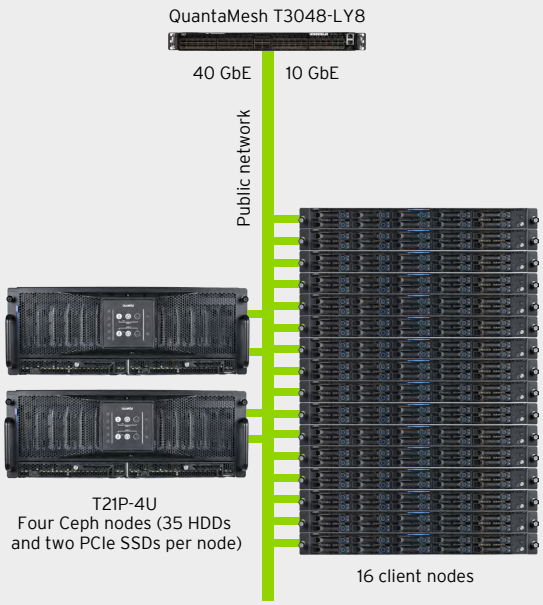

Figure 6. Two QuantaPlex T21P-4U/Dual Node servers (four Ceph nodes) were configured, each node with a 40 GbE interface, with the configuration driven by 16 client nodes.

SOFTWARE CONFIGURATION

Server systems were configured with the following storage server software:

- Red Hat Ceph Storage 1.3
- Red Hat Enterprise Linux 7.1
- Kernel version 3.10.00

Client systems were configured with the following software:

- Red Hat Enterprise Linux 7.1
- Kernel version 3.10.00

PERFORMANCE SUMMARY

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

CEPH BENCHMARK TOOL (CBT)

Testing in this reference architecture was conducted using multiple tools, including the Ceph Benchmark Tool (CBT). CBT is a Python tool for building a Ceph cluster and running benchmarks against the cluster. 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, FIO, COSbench, and MySQL sysbench.

Ceph includes the RADOS bench load test utility (a standard part of Red Hat Ceph Storage) and industry standard FIO to measure sequential and random throughput performance and latency. 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.[5]

CLUSTER SCALE-OUT PERFORMANCE

RADOS bench results on 12-bay QCT QuantaGrid D51PH-1ULH servers and 35-bay QCT QuantaPlex T21P-4U servers demonstrated that cluster throughput scales linearly for sequential reads and writes as the number of QCT storage nodes is increased (Figure 7). This capability is key because it allows organizations to scale out storage predictably by simply adding QCT servers running Red Hat Ceph Storage.

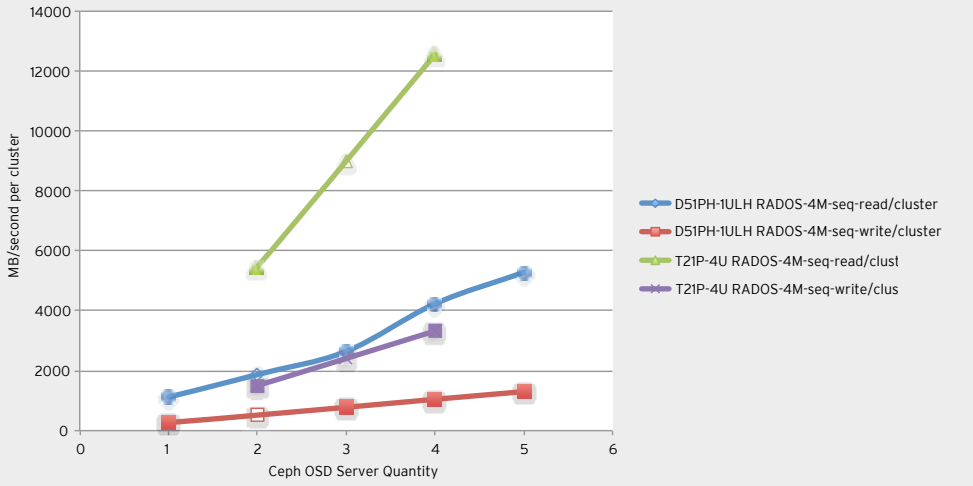

Figure 7. QCT servers running Red Hat Ceph Storage provide linear scale-out performance scaling as additional servers are added to the cluster.

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

SERVER SCALE-UP PERFORMANCE

Just as it is important for performance to scale out, so individual servers within the cluster must also be able to scale up as disks are added. Again using 4 MB RADOS bench sequential reads and writes, Figure 8 demonstrates that Ceph can achieve the same performance per disk on dense servers as it can on sparse servers. In other words, as more disks per server are deployed, Ceph can achieve consistent performance per disk.

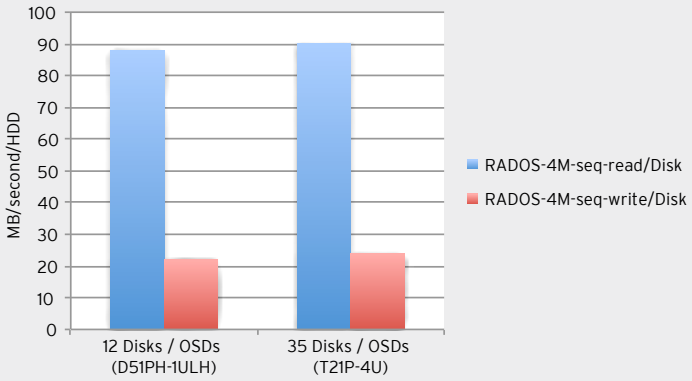

Figure 8. QCT servers running Red Hat Ceph Storage demonstrated that Ceph can achieve consistent performance per disk.

PRICE/PERFORMANCE

For throughput-oriented workloads, both servers configured with Ceph replicated pools provided excellent price/performance, enabling a range of different cluster sizes (Figure 9). Tests showed the more dense QCT T21P-4U Dual Node server achieved slightly better price/performance than the QCT D51PH-1ULH server for both reads and writes.

Figure 9. Both QCT servers offer excellent price/performance, enabling a range of cluster sizes.

COMPARING DIFFERENT REPLICATION SCHEMES

Both journaling and replication schemes can have significant impact on the performance of a Ceph Cluster. To evaluate these effects, Red Hat and QCT engineers varied both the number of SSDs for Ceph write journaling as well as replication methods. Figures 10 and 11 illustrate read and write throughput and latency respectively on different cluster configurations of QCT D51PH-1ULH servers with 12 OSDs. Read throughput is higher with replication compared to erasure-coding, while write throughput is higher with erasure-coding. Erasure-coding with no write journals resulted in the lowest performance (which may be acceptable for cost/capacity-optimized workloads).

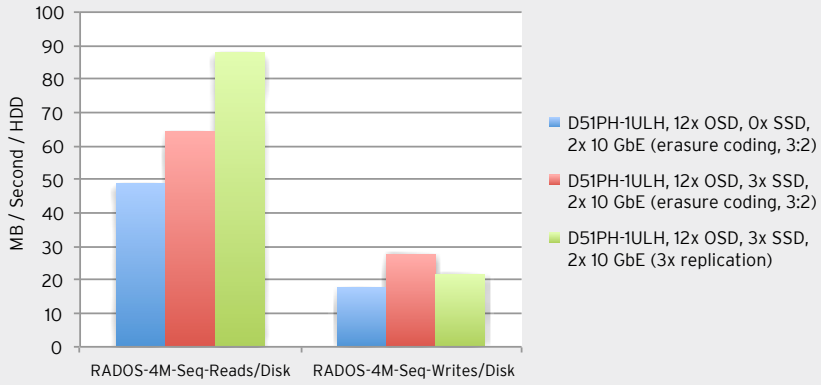

Figure 10. Effect on read and write throughput by SSDs for Ceph write journaling and replication scheme.

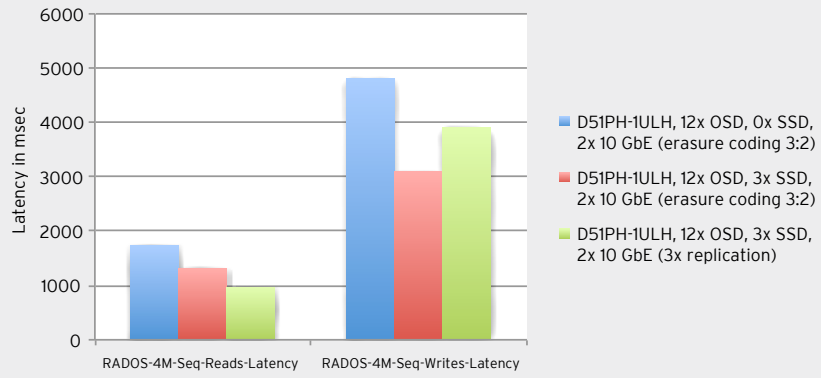

Figure 11. Effect on latency by SSDs for Ceph write journaling and replication scheme.

COMPARING DIFFERENT JOURNALING CONFIGURATIONS

Figure 12 and 13 compare the throughput and latency respectively for different journaling configurations. The dual-node QCT QuantaPlex T21P-4U server equipped with 40 GbE network interfaces, adding two PCIe SSDs for Ceph write journaling demonstrated a marked improvement in sequential writes per OSD. PCIe SSDs also showed an improvement in write latency as well.

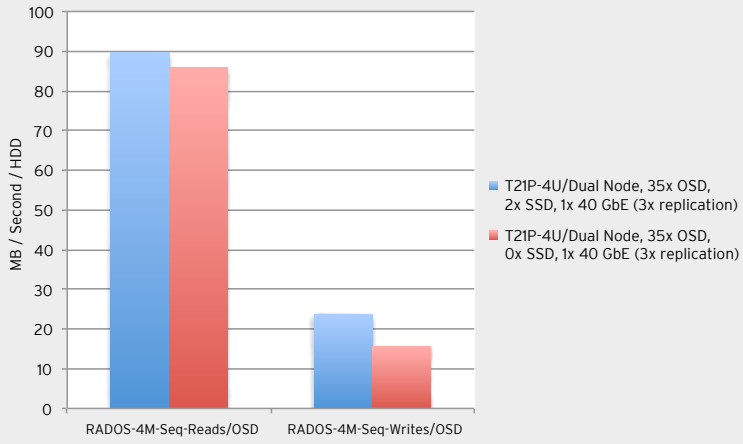

Figure 12. In QCT Quantaplex T21P-4U servers, two PCIe SSDs for Ceph write journaling improve both read and write throughput per OSD.

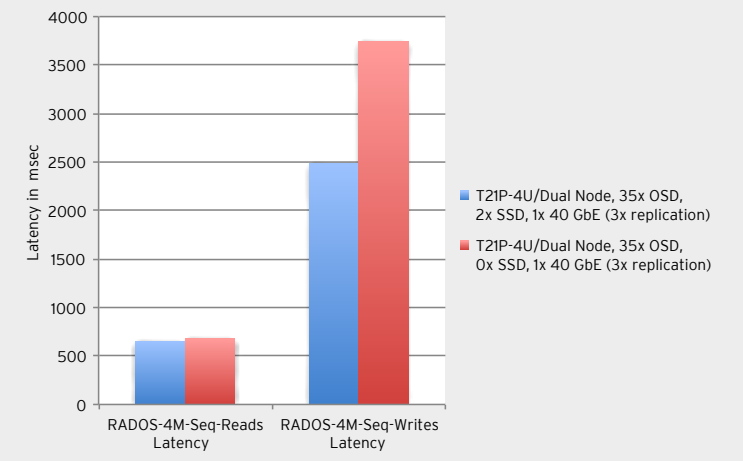

Figure 13. PCIe SSDs for Ceph write journaling also improved latency.

40 GIGABIT ETHERNET NETWORKING FOR HIGH THROUGHPUT WORKLOADS

Networking can have a significant influence on both throughput and latency. As shown in Figures 14 and 15, dense QCT QuantaPlex T21P-4U servers with Mellanox ConnectX-3 Pro 40 GbE NICs and SSDs provide three times the throughput compared to the same server configured with dual 10 GbE interfaces.

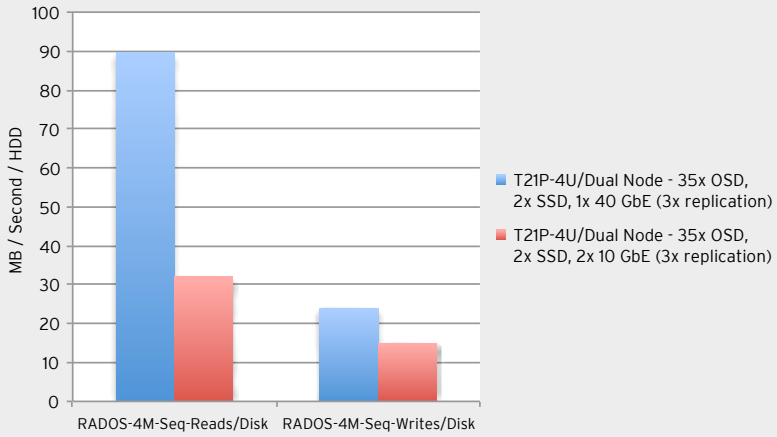

Figure 14. On the dual-node QCT QuantaPlex T21P-4U server, 40 GbE and SSDs provide dramatic improvements in throughput over the system with dual 10 GbE interfaces and SSDs.

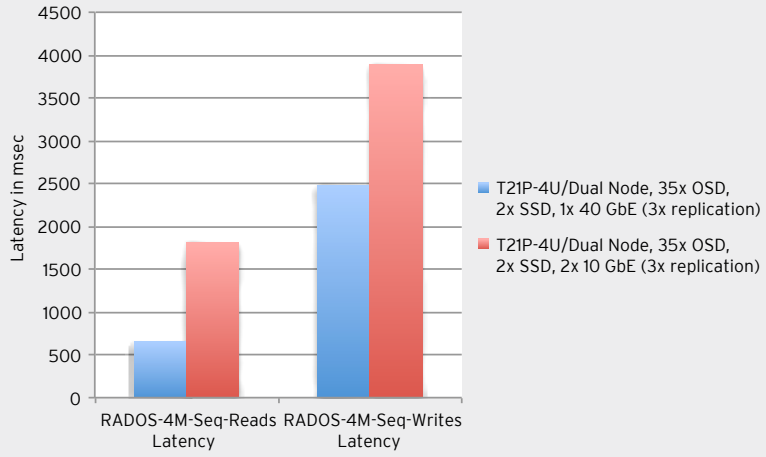

Figure 15. Both read and write latency were also improved with a 40 GbE interface as compared to dual 10 GbE network interfaces.

COST/CAPACITY OPTIMIZATION: RELATIVE COST PER TERABYTE

When optimizing for capacity in archiving scenarios, the relative cost per terabyte of various cluster configurations is an important factor. Red Hat and QCT evaluated both QuantaGrid D51PH-1ULH and QuantaPlex T21P-4U servers in terms of relative cost per terabyte. As shown in Figure 16, 8:3 erasure-coded configurations of both servers provided a cost advantage as compared to replicated storage pools.

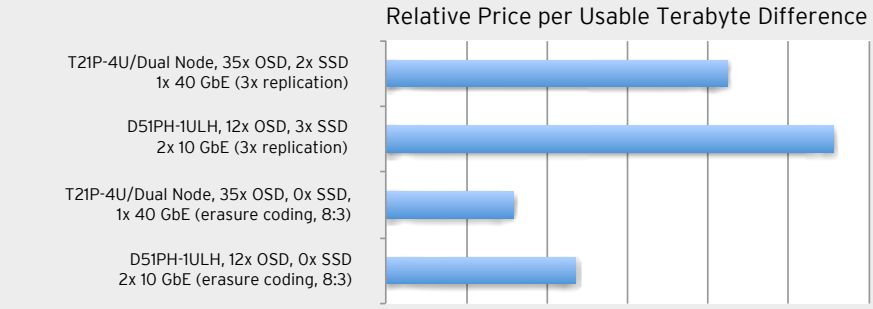

Figure 16. For cost/capacity optimization, erasure coding (8:3) provides distinct advantages.

For more performance details, including sequential read and write throughput results for all of the configurations tested, see Appendix C: Performance Details.

CONCLUSION

Red Hat's approach is to evaluate, test, and document reference configurations that depict real-world deployment scenarios, giving organizations specific and proven configuration advice that fits their application needs. Red Hat Ceph Storage combined with QCT storage servers 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 workload 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 QCT engineers.

APPENDIX A: RECOMMENDED THROUGHPUT-OPTIMIZED CONFIGURATIONS

The following list is a recommended throughput-optimized configuration for the QuantaGrid D15PH-1ULH server:

- **Processor:** 2x Intel Xeon processor E5-2630 V3 8-core 2.4 GHz
- **Memory:** 4x 16 GB 2133 MHz DDR4 RDIMM 1.2V
- **SAS controller:** QCT SAS Mezz LSI 3008
- **Network controller:** 1x QCT Intel 82599ES dual-port 10 GbE SFP+ OCP mezzanine or,
  1x QCT Intel X540 dual-port 10 GbE BASE-T OCP mezzanine
  1x dedicated management 10/100/1000 port
- **Onboard storage:** 1x Flash/M SATADOM 32 GB
- **Write-journal:** 3x Intel SSD DC S3710 200 GB, 2.5-inch SATA 6 Gb/s, MLC
- **Hard disk drive:** 12x 3.5-inch SAS 4TB 7.2 K RPM

The following list is a recommended throughput-optimized configuration for the QuantaPlex T21P-4U/Dual Node server:

- **Processor:** (2x 2) Intel Xeon processor E5-2650 V3 10-core 2.3 GHz
- **Memory:** (2x 8) 16 GB 2133 MHz DDR4 RDIMM 1.2V
- **Form factor:** 4U Rack Mount
- **SAS controller:** (2x 1) QCT SAS Mezz LSI 3008
- **Network controller:** (2x 1) QCT Mellanox ConnectX-3 EN 40 GbE SFP+ single-port OCP Mezzanine
  1x dedicated management 10/100/1000 port
- **Onboard storage:** (2x 1) Intel SSD DC S3510 120 GB, 2.5-inch SATA 6 Gb/s, MLC
- **Write-journal:** (2x 2) Intel SSD DC P3700 800 GB, 1/2 Height PCIe 3.0, MLC
- **Hard disk drive:** (2x 35) 3.5-inch SAS 4 TB 7.2 K RPM

APPENDIX B: RECOMMENDED COST/CAPACITY-OPTIMIZED CONFIGURATIONS

The following list is a recommended cost/capacity configuration for the QuantaGrid D51PH-1ULH server:

- **Processor:** 2x Intel Xeon processor E5-2630 V3 8-core 2.4 GHz
- **Memory:** 4x 16 GB 2133MHz DDR4 RDIMM 1.2V
- **SAS controller:** 1x QCT SAS Mezz LSI 3008
- **Network controller:** 1x QCT Intel 82599ES dual-port 10 GbE SFP+ OCP mezzanine, or
  1x QCT Intel X540 dual-port 10 GbE BASE-T OCP mezzanine
  1x dedicated management 10/100/1000 port
- **Onboard storage:** 1x Flash/M SATADOM 32 GB
- **Hard disk drive:** 12x 3.5-inch SAS 8 TB 7.2 K RPM

The following list is a recommended cost/capacity configuration for the QuantaPlex T21P-4U/Dual Node server:

Processor: (2x 2) Intel Xeon processor E5-2650 V3 10-core 2.3GHz

Memory: (2x 8) 16 GB 2133 MHz DDR4 RDIMM 1.2V

SAS controller: (2x 1) QCT SAS Mezz LSI 3008

Network controller: (2x 1) QCT Intel 82599ES dual-port 10 GbE SFP+ OCP mezzanine, or
(2x 1) QCT Intel X540 dual-port 10GbE BASE-T OCP mezzanine
(2x 1) dedicated management 10/100/1000 port

Onboard storage: (2x 2) Intel SSD DC S3510 120GB, 2.5-inch SATA 6 Gb/s, MLC

Hard disk drive: (2x 35) 3.5-inch SAS 8 TB 7.2 K RPM

APPENDIX C: PERFORMANCE DETAIL

I/O block size can have an additional impact on both throughput and latency. To evaluate this variable, Red Hat and QCT engineers tested all of the server configurations with multiple I/O block sizes.

SEQUENTIAL READ THROUGHPUT

Sequential read throughput was evaluated at multiple I/O block sizes for all server configurations, as shown in Figure 17. Various combinations of SSDs for Ceph write journaling were used as well as different replication strategies.

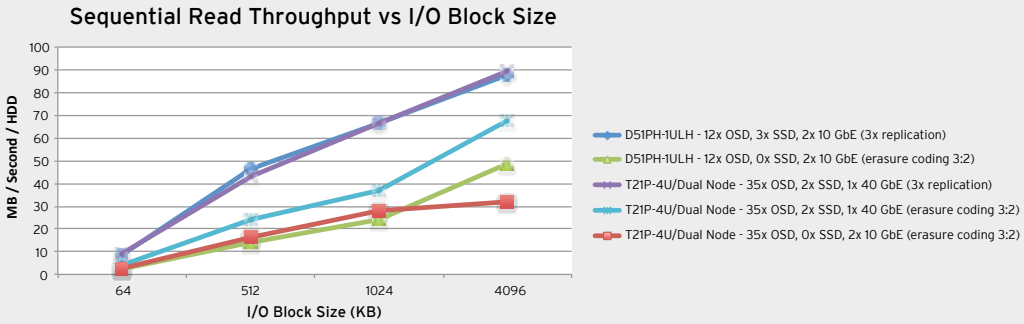

Figure 17. Sequential read throughput scales depending on the I/O block size.

Sequential write throughput for the same set of servers is shown in Figure 18. Note that sequential writes were measured from a client perspective. The amount of data written, and the corresponding throughput rate, is actually higher due to back-end write amplification from data protection. For example, with 3x replication pools, the Ceph cluster actually was writing data at three times the rate shown.

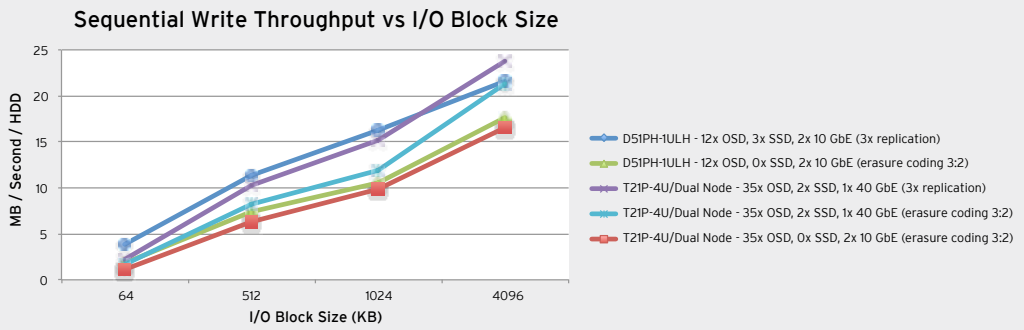

Figure 18. Sequential write throughput for tested servers with various I/O block sizes.

Sequential read and write latency for all of the server configurations is shown in Figure 19 and 20, respectively.

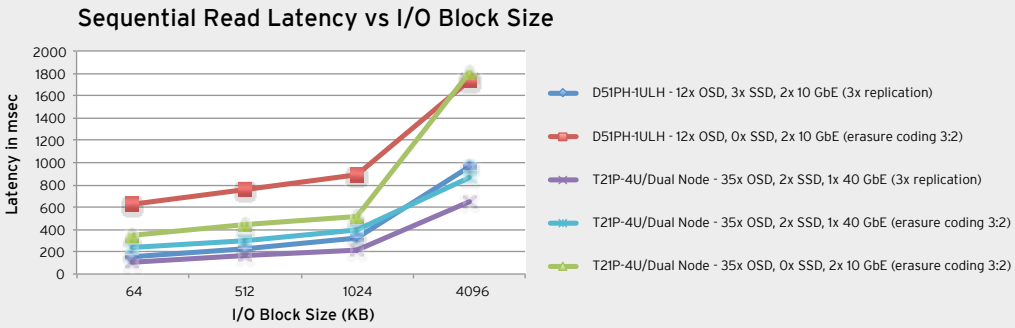

Figure 19. Sequential read latency compared to I/O block size.

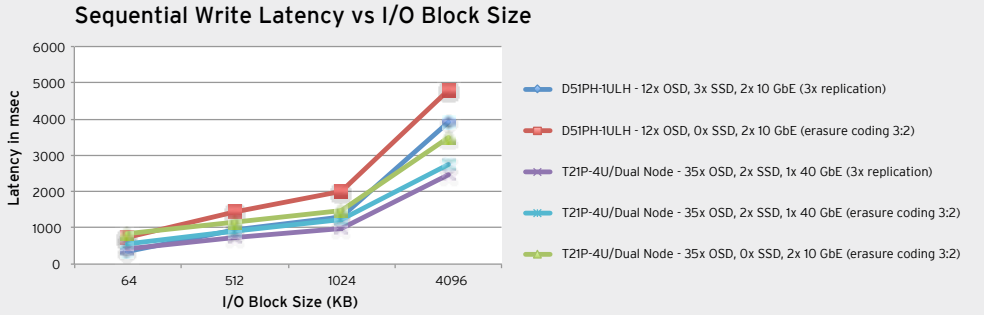

Figure 20. Sequential write latency for various I/O block sizes.

RED HAT®
STORAGE

redhat.

QCT
Quanta CLOUD TECHNOLOGY

TECHNOLOGY DETAIL Red Hat Ceph Storage on QCT Servers

ABOUT QCT (QUANTA CLOUD TECHNOLOGY)

QCT (Quanta Cloud Technology) is a global datacenter solution provider extending the power of hyperscale datacenter design in standard and open SKUs to all datacenter customers. Product lines include servers, storage, network switches, integrated rack systems and cloud solutions, all delivering hyperscale efficiency, scalability, reliability, manageability, serviceability and optimized performance for each workload. QCT offers a full spectrum of datacenter products and services from engineering, integration and optimization to global supply chain support, all under one roof. The parent of QCT is Quanta Computer Inc., a Fortune Global 500 technology engineering and manufacturing company.
www.qct.io

ABOUT RED HAT

Red Hat is the world's leading provider of open source solutions, using a community-powered approach to provide reliable and high-performing cloud, virtualization, storage, Linux, and middleware technologies. Red Hat also offers award-winning support, training, and consulting services. Red Hat is an S&P company with more than 80 offices spanning the globe, empowering its customers' businesses.

NORTH AMERICA
1 888 REDHAT1

EUROPE, MIDDLE EAST, AND AFRICA
00800 7334 2835
europe@redhat.com

ASIA PACIFIC
+65 6490 4200
apac@redhat.com

LATIN AMERICA
+54 11 4329 7300
info-latam@redhat.com

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

redhat.com

#INC0347490 0116