RED HAT® STORAGE

redhat.

SanDisk®
a Western Digital brand

REFERENCE ARCHITECTURE

RED HAT CEPH STORAGE ON THE INFINIFLASH ALL-FLASH STORAGE SYSTEM FROM SANDISK

Combining Red Hat Ceph Storage with the InfiniFlash system from SanDisk yields software-defined all-flash storage without traditional limitations.

The scalable InfiniFlash system effectively increases flash density at highly competitive price points, while reducing operational expenses (OpEx) associated with traditional clusters based on hard disk drives (HDDs).

Extensive testing by Red Hat and SanDisk has shown the solution effective for both IOPS- and throughput-intensive workloads, while demonstrating very low latency..

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

redhat.com

ABSTRACT

Based on the successes of public cloud and hyperscale deployments, software-defined storage solutions like Red Hat® Ceph Storage have become a popular alternative to traditional proprietary storage. While solid state flash technology has appeared in some of these solutions, it has usually been reserved for the most demanding application requirements due to cost. The InfiniFlash system from SanDisk, however, provides a cost-effective complement for Red Hat Ceph Storage, offering a dense, reliable, efficient, and high-performance platform for both IOPS- and throughput-intensive workloads. Extensive testing by Red Hat and SanDisk has demonstrated that flash is no longer limited to top-tier applications.

TABLE OF CONTENTS

INTRODUCTION

Countless organizations are demanding storage solutions that deliver performance at increasingly massive scale. The success of hyperscale cloud deployments such as Facebook, Google, and Amazon is inspirational to many. At the same time, most enterprises lack the size of these organizations, and very few have the substantial development and operational resources to deploy storage infrastructure in the same ways. Most enterprises also have broad heterogeneous application requirements that dictate agility and flexibility from storage infrastructure.

The compelling rise of flexible software-defined storage is a common thread for both hyperscale and enterprise IT. First deployed in the cloud, software-defined storage has now emerged as a viable alternative to costly monolithic proprietary storage solutions in the enterprise. Ceph in particular has proved to be a flexible and capable storage platform that can serve a wide variety of IOPS- and throughput-intensive workloads alike. With Ceph, organizations can employ a wide range of industry-standard servers to avoid the scalability issues and storage silos of traditional solutions.

As organizations encounter the physical limitations of hard disk drives (HDDs) flash technology has emerged as a component of many storage strategies and even hybrid approaches. Unfortunately, despite its superior performance profile, many of the attempts to integrate flash into storage infrastructure have proven problematic, expensive, or limiting in important ways. As a result, the high performance of flash has only been accessible to relatively few workloads.

As a manufacturer of flash memory, SanDisk is in a unique position to innovate at all layers of the technology stack. Coupling Red Hat Ceph Storage with the InfiniFlash system from SanDisk offers new possibilities that allow organizations to combine robust open source storage services software with a proven all-flash storage system . Red Hat Ceph Storage is software that is designed to accommodate a wide range of storage server hardware while the InfiniFlash system has emerged as a versatile, highly scalable, and cost-effective clustered flash platform that is easily deployed in software-defined storage architectures (Figure 1).

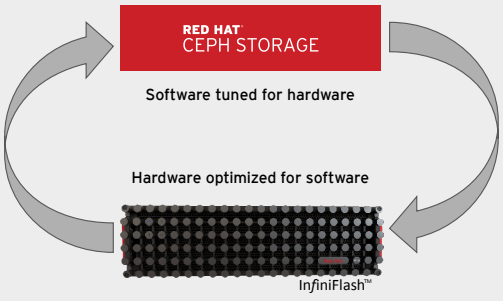

Figure 1. Red Hat Ceph Storage and the InfiniFlash system represent complementary technologies.

InfiniFlash offers 50x the performance, 5x the density, and 4x the reliability of traditional HDDs, while consuming 80% less energy.[1] Moreover, by decoupling the storage from the compute and networking elements of storage infrastructure, InfiniFlash strongly complements a software-defined approach like Red Hat Ceph Storage — offering flexible and cost-effective infrastructure that can deliver scalable storage for a range of enterprise needs.

1 *Based on published specification and internal testing at SanDisk.*

RED HAT CEPH STORAGE ON INFINIFLASH FROM SANDISK

Red Hat Ceph Storage allows organizations to choose the optimized hardware platforms that best suit their application and business needs. For instance, individual storage servers from multiple vendors can be added in conjunction with Red Hat Ceph Storage to accelerate performance for IOPS-intensive workloads.[2] Flash storage is currently deployed in multiple forms in the datacenter, with varying levels of success.

- ***Server-based storage.*** Server-based flash storage provides the lowest latency and highest performance, but it is limited to individual hosts. This approach uses a server platform with a large number of internal drive slots full of high-capacity solid-state drives (SSDs) or NVM Express (NVMe) devices. These systems are often cost-effective, but can exhibit performance bottlenecks due to the fixed amount of server CPU and RAM resources that access flash.
- ***All-flash arrays.*** All-flash arrays are dedicated systems with their own flash storage and software, usually integrating server form-factor solid-state drives (SSDs) into custom enclosures. While these systems can offer high performance and enterprise-class features they can also bring the cost, complexity, and siloed storage attributes that many organizations are trying to escape.
- ***Clustered flash storage.*** Clustered flash storage like InfiniFlash takes an entirely different approach. By design, InfiniFlash is not engineered with computational CPU cores inside the storage unit. Instead, up to eight servers connect to InfiniFlash via high-speed 12Gbps SAS connections.

With InfiniFlash, storage compute is thus effectively separated and decoupled from storage capacity, so that the right amount of compute power can be flexibly matched to the right amount of storage capacity for a given workload. Not only does this approach allow vastly improved utilization, but high-performance flash storage infrastructure can be tuned to specific workloads through multidimensional scaling:

- More InfiniFlash units can be added to extend storage capacity.
- The number, density, and character of individual flash cards can be varied to suit the application.
- Ceph OSD hosts can be added or removed as desired to match the desired storage workload with additional amounts of storage and compute.
- Scaling out can be accomplished by growing each dimension while maintaining a fixed ratio.

USE CASE: A UNIFIED PLATFORM FOR OPENSTACK

According to semi-annual OpenStack® user surveys Ceph is the leading storage platform for OpenStack.[3] Figure 2 illustrates a typical Ceph configuration for OpenStack based on the InfiniFlash system. Importantly, even when using Ceph for OpenStack, the ability to independently vary CPU and storage ratios is paramount.

The high-density InfiniFlash system requires much less CPU compared to sparser SSDs or HDDs. In addition, as users scale the cluster, they can scale the storage compute and storage capacity independently, making a perfectly balanced cluster for the workload, and delivering a greater cost advantage for large-scale cluster deployments. The superior reliability of an InfiniFlash storage node requires less hardware and provides higher reliability (1.5 million hours of MTBF) than a typical HDD

2 https://www.redhat.com/en/resources/mysql-databases-ceph-storage-reference-architecture
3 Ceph is and has been the leading storage for OpenStack according to several semi-annual OpenStack user surveys.

node. Three full copies (3x replication, as is typically used on HDD nodes) are no longer required due to the higher reliability of InfiniFlash. For this reason, 2x replication plus a reliable backup is typically sufficient, saving both acquisition cost, rack space, and ongoing operational costs.

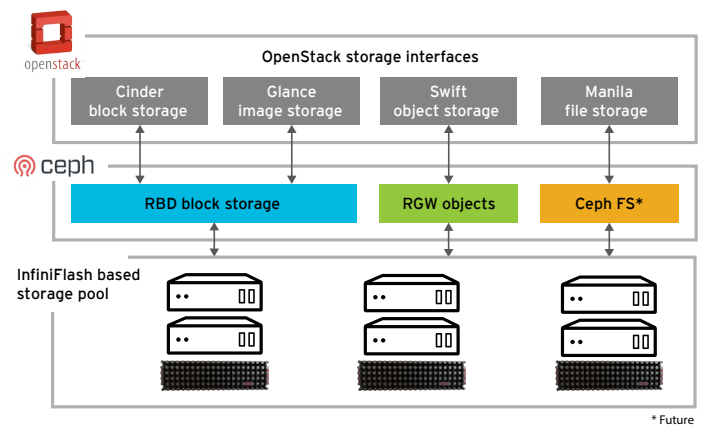

Figure 2. Red Hat Ceph Storage and InfiniFlash represent a high-performance storage solution for OpenStack workloads.

USE CASE: STAND-ALONE OBJECT STORAGE

Stand-alone object stores work well for active archives, data warehouses or big data lakes, and content libraries. Red Hat Ceph Storage coupled with InfiniFlash enables unstructured and semi-structured data storage at web scale, with scalable throughput (Figure 3). Using the Ceph object gateway, the solution works with standard S3 and Swift APIs for a wide range of compatibility. A Ceph cluster can span multiple active-active geographic regions with no single point of failure.

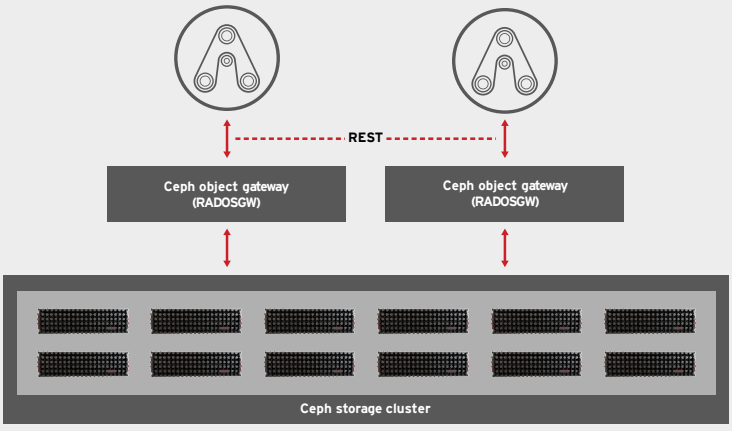

Figure 3. Red Hat Ceph Storage and InfiniFlash can support stand-alone object storage by using the Ceph object gateway.

USE CASE: STORAGE FOR MYSQL DATABASES

Administration of storage for proliferating MySQL databases has become a challenge, especially as administrators seek to consolidate database workloads. Those managing hundreds or thousands of MySQL database instances seek the flexibility of allocating storage capacity from a global, elastic storage pool, without the concerns of moving instances between storage islands. Traditional all-flash arrays provide high IOPS in storage islands, in contrast to Ceph-based storage that aggregates storage capacity from multiple servers or InfiniFlash systems into a single global namespace. Alternatively, a hyperconverged appliance could generate plentiful IOPS. However, those SQL workloads that don't require high IOPS or don't need the provided capacity leave capacity and/or performance under-utilized in each appliance.

In contrast, InfiniFlash allows storage compute and storage capacity to be varied independently for individual workloads. This flexibility effectively avoids the unused and underutilized capacity problems of hyperconverged appliances. From a performance standpoint, each InfiniFlash system can typically achieve over 800K IOPS and provide 64-512TB of raw capacity.

USE CASE: CUSTOM STORAGE WORKLOADS

Because InfiniFlash is decoupled from CPU resources, it can effectively serve custom workload environments. Multiple servers can connect to an individual InfiniFlash system. This flexibility allows organizations to size CPUs and numbers of servers to the workload, independently of the desired capacity. Different numbers of I/O connections can be provided to different application servers. InfiniFlash provides sufficient IOPS and throughput to handle multiple SQL workloads, serve as a media repository, and provide big data throughput.

CEPH ARCHITECTURE OVERVIEW

A Ceph storage cluster is built from large numbers of nodes for scalability, fault-tolerance, and performance. Each node is based on industry-standard server 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 upon cluster expansion or hardware failure (remap and back fill)
- Ensure data integrity (scrubbing)
- Detect and recover from faults and failures

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

When a Ceph client reads or writes data (referred to as an I/O context), it connects to a logical storage pool in the Ceph cluster. Figure 4 illustrates the overall Ceph architecture, with concepts that are described in the sections that follow.

- ***Client interface layer***. Writing and reading data in a Ceph storage cluster is accomplished using the Ceph client architecture. Ceph supports a range of storage methods. RADOSGW is an 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. RBD offers a Ceph block storage device that mounts like a physical storage drive for both physical and virtual systems.
- ***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. For data protection, 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—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. Peer OSDs each receive an object replica (or erasure-coded chunk) upon a write. Fault-domain policies within the CRUSH ruleset can force OSD peers to be selected on different servers, racks, or rows. 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., 200 per OSD or more) leads to better balancing.

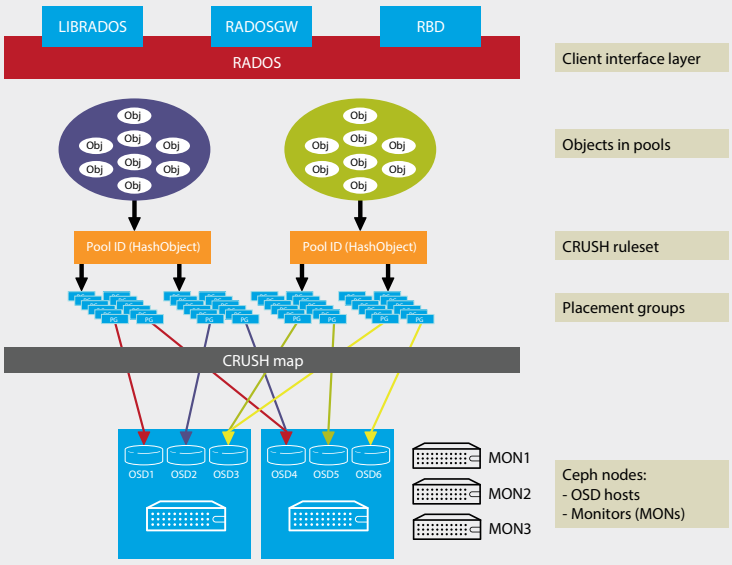

Figure 4. Clients write to Ceph storage pools while the CRUSH ruleset determines how placement groups are distributed across object storage daemons (OSDs).

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

REFERENCE ARCHITECTURE ELEMENTS

In Red Hat and SanDisk testing, the solution architecture included Red Hat Ceph Storage installed on multiple industry-standard servers connected to one or more InfiniFlash systems.

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, hybrid, 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 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, with properties that 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

THE INFINIFLASH SYSTEM FROM SANDISK

From a price/performance perspective, InfiniFlash fits between high-performance, high-cost conventional all-flash arrays and server-based flash systems. The solution provides scalable capacity and breakthrough economics in both CapEx and OpEx easily surpassing economics for high-performance HDD-based storage systems, while providing additional benefits:

- ***Reliability and availability.*** Due to the higher performance of flash, less infrastructure is needed to provide the same level of performance, leading to much higher system reliability. Better reliability reduces maintenance costs and lowers required spares inventories, while contributing strongly to more availability for applications.
- ***TCO savings.*** A fully-populated 512TB InfiniFlash System 150 consumes approximately 450 watts of power. When combined in operation with servers, it will use up to 80% less power than a comparable HDD array requires. The TCO savings improve dramatically for IOPS-intensive workloads, where large numbers of HDDs are required to achieve the spindle count needed for performance. For high-density configurations, InfiniFlash can provide up to 6PB of storage in a single rack.
- ***Software deployment flexibility.*** As a building block for a software-defined Red Hat Ceph Storage solution, InfiniFlash allows organizations considerable flexibility. Compute and storage resources can be precisely tuned to provide the needed performance, without compromising utilization.

Shown in Figure 5, the InfiniFlash System IF150 delivers from 64TB to 512TB of resilient flash storage in a highly dense form factor for petabyte-scale capacity, high density, and high-performance storage environments. Each IF150 system can be configured with up to sixty-four 4TB or 8TB hot-swappable solid state SanDisk flash cards – delivering up to half a petabyte of raw flash storage in a 3 rack unit (3U) enclosure and up to 6PB in a single rack. The system scales easily, and each InfiniFlash System IF150 can connect up to eight individual servers with 12Gbps SAS connections.

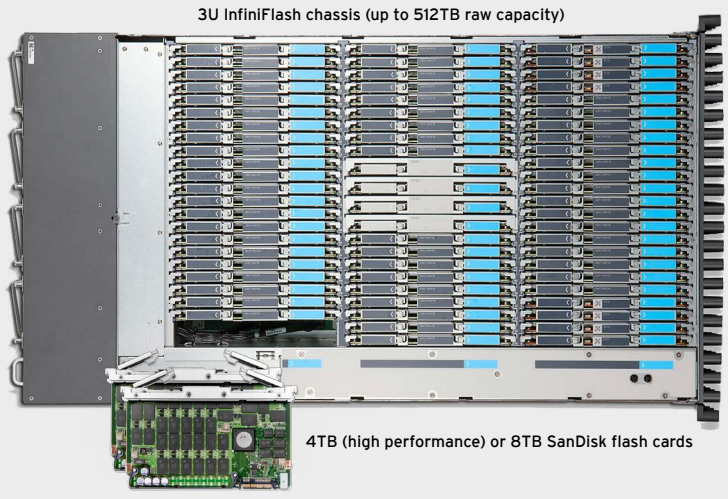

Figure 5. With custom 4TB or 8TB flash cards, the InfiniFlash supports up to 512TB in one three-rack unit enclosure.

Unlike many all-flash array systems, InfiniFlash does not use off-the-shelf SSDs, which can needlessly impose limitations on flash storage. Instead, the InfiniFlash system employs innovative custom SanDisk flash cards that deliver excellent flash performance, reliability, and storage density with low latency. The individual state of each flash cell is visible to the system's firmware. This capability enables the system to better manage I/O, more efficiently schedule housekeeping operations such as free space management (i.e., garbage collection), and delivers consistent performance in the face of changing workloads and rigid SLA requirements.

COMMODITY SERVERS

Eight SAS connections are available on the InfiniFlash chassis. Two to eight individual servers are typically attached via SAS expansion connectivity. OSDs running on the servers thus connect to InfiniFlash flash devices as if they were resources local to the server.

TESTING AND PERFORMANCE SUMMARY

Red Hat and SanDisk testing exercised the flexibility of the joint solution in the interest of understanding how it performed, both in terms of IOPS and throughput.

TESTED CONFIGURATIONS

Befitting the flexibility of a software-defined approach, testing evaluated different numbers of OSD servers, OSDs, InfiniFlash chassis, and flash devices. Several variations on two basic configurations were tested:

- **A4 configuration**. An InfiniFlash chassis with four attached servers (OSD nodes)
- **A8 configuration**. An InfiniFlash chassis with eight attached servers (OSD nodes)

Configurations with one and two InfiniFlash chassis were evaluated. The single-chassis A4 and A8 configurations were also evaluated both half-full and full of flash devices. Table 3 lists the details of each configuration, including the number of InfiniFlash chassis, number of OSD nodes, and number of OSD nodes and number of flash devices. Full detailed test results for all configurations can be later in this document.

TABLE 3. TESTED INFINIFLASH CONFIGURATIONS

SETUP	NUMBER OF INFINIFLASH	AMOUNT OF FLASH	FLASH DEVICES PER INFINIFLASH	OSD NODES PER INFINIFLASH	FLASH DEVICES / OSDS PER SERVER
1x A4	1	256TB	32	4	8
2x A4	2	512TB	32	4	8
1x A4 full	1	512TB	64	4	16
1x A8	1	256TB	32	8	4
2x A8	2	512TB	32	8	4
1x A8 full	1	512TB	64	8	8

SCALABILITY SUMMARY

Performance scalability is a key consideration for any storage solution. Ideally the solution should also be able to scale in terms of both IOPS and throughput. Figure 6 summarizes IOPS performance when scaling from a single InfiniFlash chassis to two. Data points are provided for both four and eight servers per InfiniFlash chassis. While providing a dense decoupled Ceph storage cluster, the InfiniFlash solution enables organizations to scale out both in capacity and in performance.

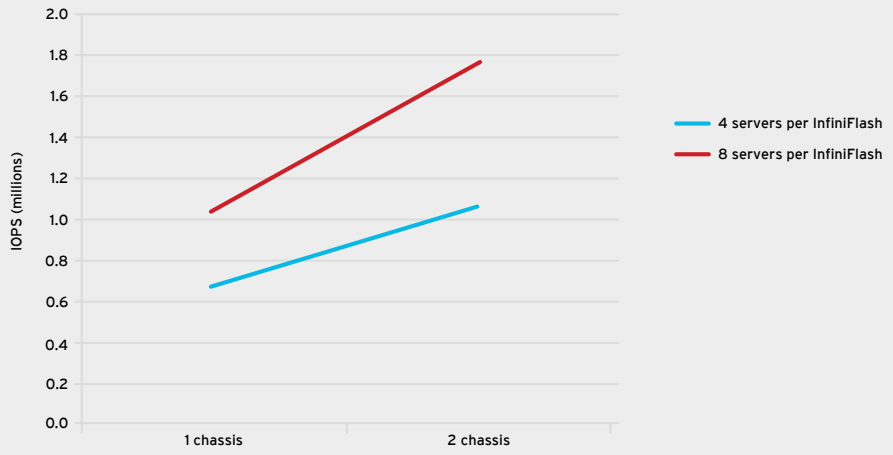

Figure 6.IOPS performance scales when moving from one to two InfiniFlash chassis running Red Hat Ceph Storage.

Figure 7 illustrates throughput scalability in terms of GB/second when moving from a single InfiniFlash chassis to two. Again, scalability is shown for both four and eight servers per InfiniFlash.

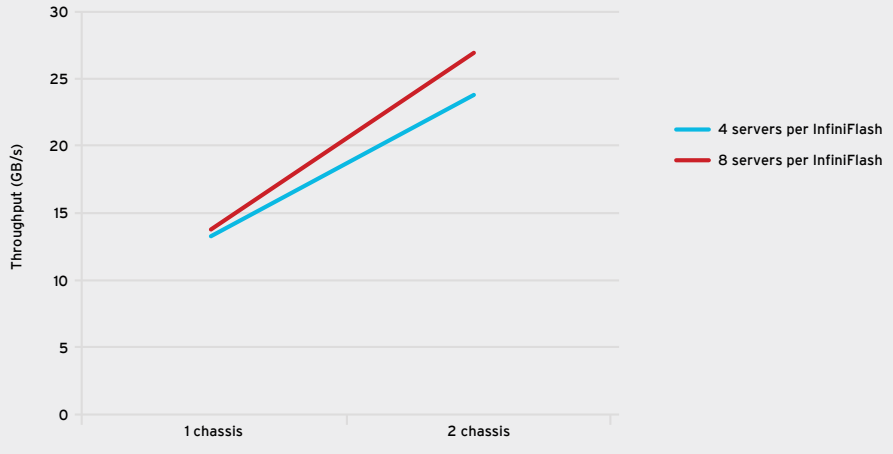

Figure 7. Throughput also scales well when moving from one to two InfiniFlash chassis running Red Hat Ceph Storage.

These two graphs illustrate the flexibility of InfiniFlash to vary the ratio between storage compute and storage capacity to fit workload needs. Figure 6 illustrates significantly higher IOPS performance by doubling the amount of storage compute per InfiniFlash chassis. On the other hand, throughput-intensive workloads do not benefit much from adding additional compute resources, as illustrated in Figure 7.

COST VERSUS PERFORMANCE

With the Red Hat Ceph Storage and InfiniFlash solution, performance and cost are tunable by varying OSD nodes, the number of chassis, and the number of OSDs and flash devices in each InfiniFlash chassis. As such, organizations can choose optimized configurations to meet specific IOPS, bandwidth, or mixed workload requirements. Figure 8 and Table 4 show how performance-optimized and capacity-optimized InfiniFlash configurations can complement both traditional HDD-based servers and servers with integral NVMe solutions in terms of both cost and performance.

TABLE 4. MULTI-TIER STORAGE OPTIONS FOR CEPH

	LOW-MEDIUM PERFORMANCE TIER	HIGH-PERFORMANCE TIER	ULTRA-PERFORMANCE TIER
Ceph-based storage solution	Servers with HDD drives	InfiniFlash	Servers with NVMe drives
Use cases	Archive storage, batch analytics, and object storage	VMs for tier-1 applications, high-performance analytics, rich media streaming	Low-latency databases, write-heavy OLTP applications
Capacity	500TB - 10PB+	100TB - 10PB+	10-50TB
Relative cost	$	$$	$$$$

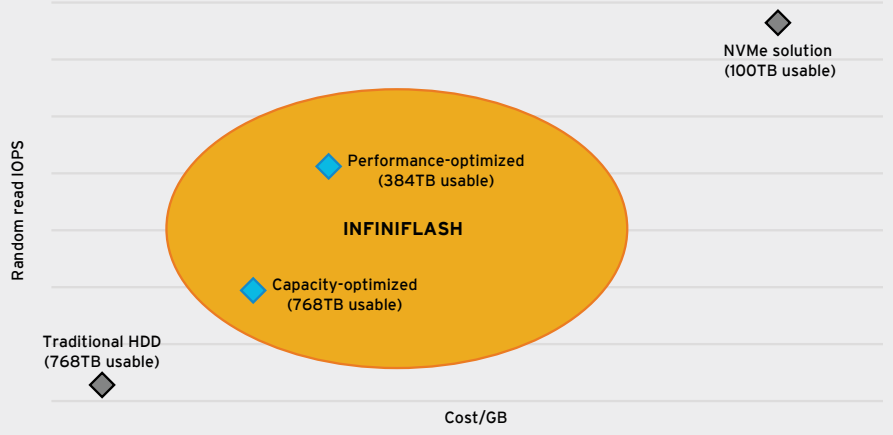

Figure 8. Red Hat Ceph Storage on InfiniFlash can serve both capacity-optimized and performance-optimized configurations, complementing traditional HDD-based solutions as well as more costly PCIe-based solutions.

Beyond overall cost/performance, workload-specific metrics are important for any storage deployment. Red Hat and SanDisk testing demonstrated that the ability to vary the number of OSD nodes as well as the number of flash devices provides a range of cost and performance points that can be tailored to suit the needs of the workload (Figure 9, lower is better).

- For IOPS-sensitive workloads, additional OSD nodes can provide a significant performance boost while only adding a small incremental cost. This difference is shown by comparing the lower cost per IOPS of having eight servers attached to a single chassis (1A8) versus having four servers attached to a single chassis (1A4).
- For bandwidth-sensitive workloads, increasing the ratio of InfiniFlash chassis and spreading the storage across both chassis can provide a better improvement in cost per unit of throughput (MB/s). Having eight servers connected to two chassis provides (2A4) provides a lower cost as compared to having eight servers connected to one chassis (1A8).
- The lowest overall cost per IOPS can be achieved by increasing the ratio of servers to InfiniFlash chassis (2A8). However, this configuration does come at the cost of requiring additional rack space.

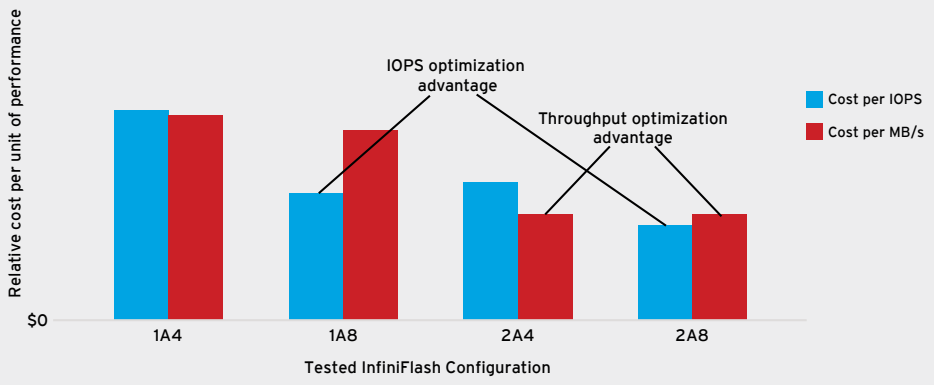

Figure 9. IOPS and throughput cost are best optimized by different configurations (lower is better).

SERVER SIZING

Combined with Red Hat Ceph Storage, the InfiniFlash platform allows both scale-up and scale-out approaches. Because InfiniFlash storage is decoupled from the server, a wide range of industry-standard servers can be used. Table 5 provides example server and InfiniFlash sizing guidance for addressing a range of workloads for various cluster sizes. All configurations would be configured to use the dual-port LSI 9300-8e PCI-Express 3.0 SATA/SAS 8-port SAS3 12Gb/s host bus adapter (HBA) or similar.

TABLE 5. RECOMMENDED CONFIGURATIONS AND COMPONENTS FOR VARIOUS WORKLOADS

WORKLOAD	SYSTEM COMPONENTS
IOPS-OPTIMIZED (SMALL-BLOCK I/O)	1x Ceph OSD server per 4-8 InfiniFlash cards: • Dual Intel Xeon Processor E5-2687 • 128GB Ram
	Two InfiniFlash System IF150 (128TB to 256TB per enclosure using 4TB performance flash cards)
THROUGHPUT-OPTIMIZED	1x Ceph OSD server per 16 InfiniFlash cards: • Dual Intel Xeon Processor E5-2680 • 128GB Ram
	Two InfiniFlash System IF150 (128TB to 512TB per enclosure)
MIXED WORKLOADS	1x Ceph OSD server per 8-16 InfiniFlash cards: • Dual Intel Xeon Processor E5-2690+ • 128GB Ram
	Two InfiniFlash System IF150 (128TB to 512TB per enclosure)*

* Note: Optional 4TB performance cards can be used for the larger configurations if desired for workload.

DETAILED PERFORMANCE RESULTS

The sections that follow provide the testing methodology, test harness, and detailed results for testing performed by Red Hat and SanDisk.

SOFTWARE VERSIONS AND PERFORMANCE MEASUREMENT

All testing in this reference architecture was conducted using Red Hat Ceph Storage 1.3.2 along with the Ceph Benchmark Tool (CBT).[4] CBT is a Python tool for building a Ceph cluster and running benchmarks against it. As a part of SanDisk testing, CBT was used to evaluate the performance of actual configurations. CBT automates key tasks such as Ceph cluster creation and tear-down and also provides the test harness for automating various load-test utilities. The flexible file I/O (FIO) utility was called from within CBT as a part of this testing.[5] FIO version 2.8.19 used in testing was built from the source to include the RBD engine. The Parallel Distributed Shell (PDSH version 2.31-4) and Collectl (version 4.0.2) were also used in testing and monitoring.

It is important to note that while the CBT utility can save time on benchmarking a Ceph environment, it does have some limitations, leading to some results that may seem suboptimal. In particular, the queue depth (QD) is not configurable for a given block size. For example, it is not possible to set a QD of 128 with a 4Kb block size and a QD of 32 for a 1Mb block size. Without the ability to vary queue depth, driving the high IOPS figures for small block sizes leads to high latency on larger block sizes. Given an ability to adjust queue depth, lower latencies could be generated for large-block throughput tests. In the interest of not having customized runs for every test, a generic CBT YAML script was used for all block sizes.

4 *The Ceph Benchmarking Tool was retrieved from https://github.com/ceph/cbt.*
5 *FIO was retrieved from https://github.com/axboe/fio*

TESTING ARCHITECTURE

Performance for each storage cluster configuration was collected over five runs, with each run 20 minutes in length. A single 2TB RBD volume was used for each client. Enough data to was used to ensure the usage of OSD server disk caches. CBT automatically drops the file and slab caches before each run to make sure that the caches are built organically during the test. To achieve a representative performance value, high and low results from each collection of five runs were discarded, and the remaining three runs averaged together.

Figure 10 illustrates the A4 test configuration with four servers connected to each InfiniFlash chassis. For this test, a single Ceph MON node was provided in addition to eight client nodes. Clients were connected to the storage cluster by a 40Gb Ethernet (40GbE) switch. In this configuration, two SAS host bus adapters were connected to each of the two host SAS expanders (HSEs) on each InfiniFlash chassis.

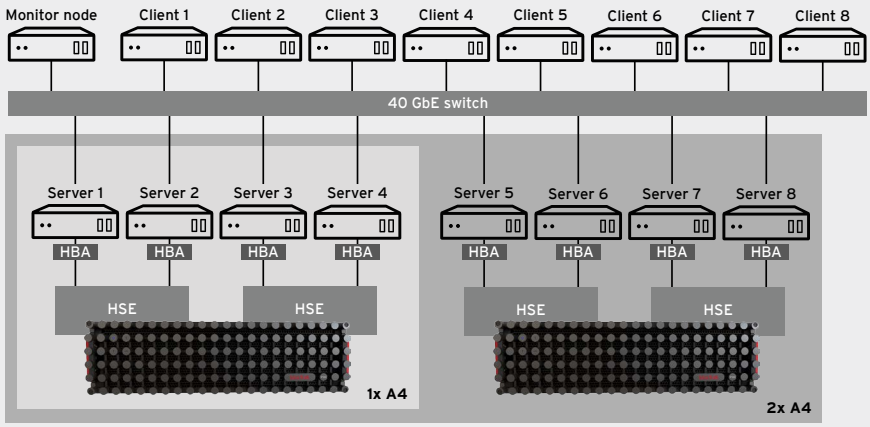

Figure 10. Four-server-per-InfiniFlash test configuration (single- and dual-chassis)

Figure 11 illustrates the eight-server-per-InfiniFlash test configuration (A8).

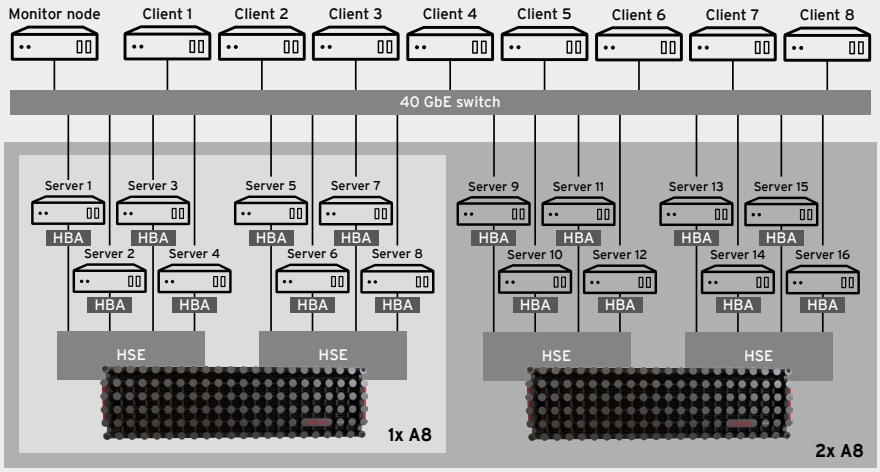

Figure 11. Eight-server-per-InfiniFlash test configuration (single- and dual-chassis)

IOPS PERFORMANCE

Testing used a range of block sizes to evaluate impact across all of the tested configurations. In general, smaller block sizes are representative of IOPS-intensive workloads while larger block sizes are used for more throughput-intensive workloads. Figure 12 illustrates read IOPS results. Performance scales reliably across the six cluster configurations. As expected, doubling the number of chassis and OSD nodes provides a substantial performance boost in read IOPS.

As illustrated by the 1x A8 (half) and the (1x) A8 full data, adding more flash cards without increasing compute does not improve 4K random read IOPS performance. However, as illustrated in Figure 13 (write IOPS), this same comparison shows significantly increased 4K random write IOPS performance when adding more flash cards while keeping server compute constant. These results show that for IOPS-oriented workloads, the optimal ratio of flash to compute varies based on read/write mix. Both charts illustrate that IOPS scalability becomes increasingly limited as block size increases.

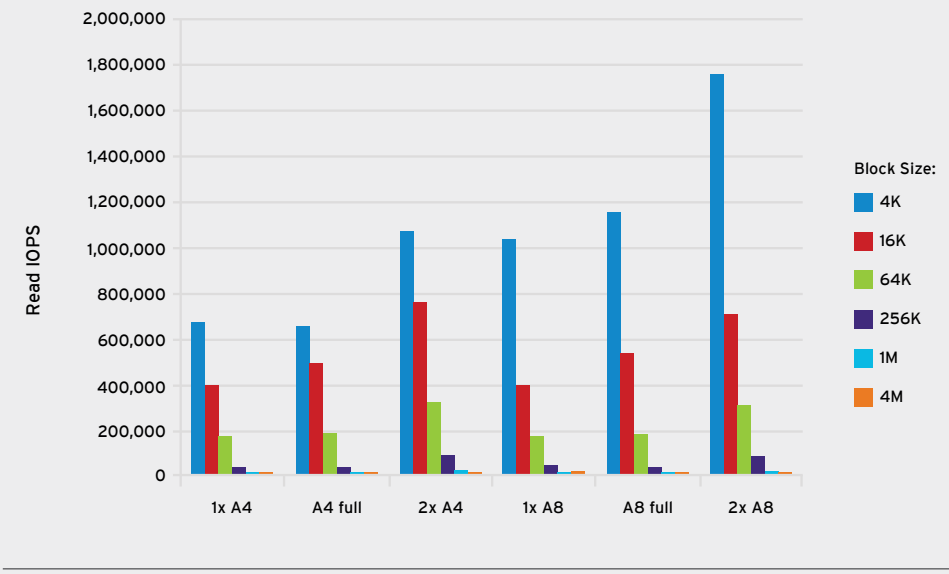

Figure 12. Read IOPS scales well as cluster size increases.

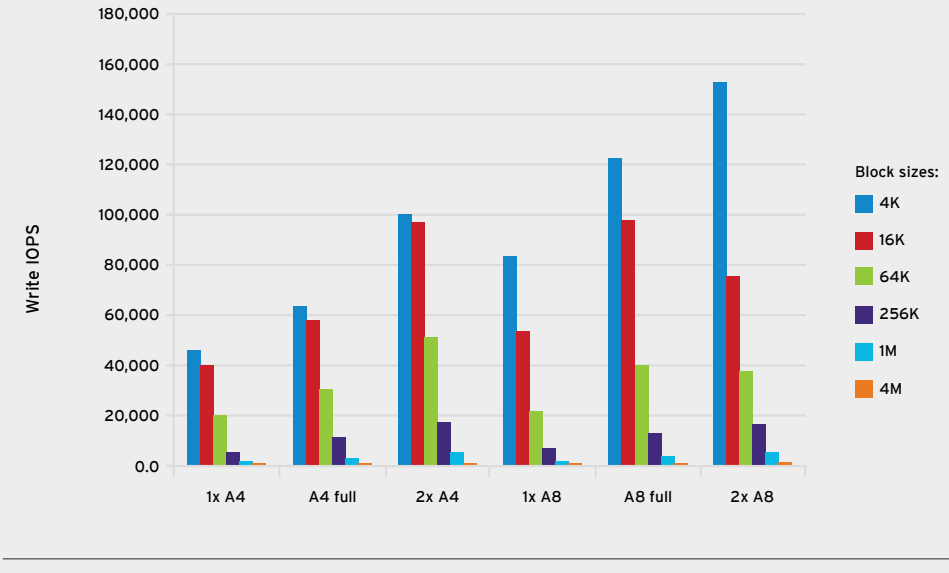

Figure 13. Write IOPS scalability

THROUGHPUT PERFORMANCE

Read and write throughput are shown in Figures 14 and 15 respectively. As expected, throughput increases with block size, and approximately doubles with cluster size. The charts also illustrate that optimal price/performance for large block throughput-oriented workloads occurs with four servers per InfiniFlash chassis (A4) vs. 8 servers per InfiniFlash chassis (A8). In other words, throughput performance does not increase substantially between 2x A4 and 2x A8.

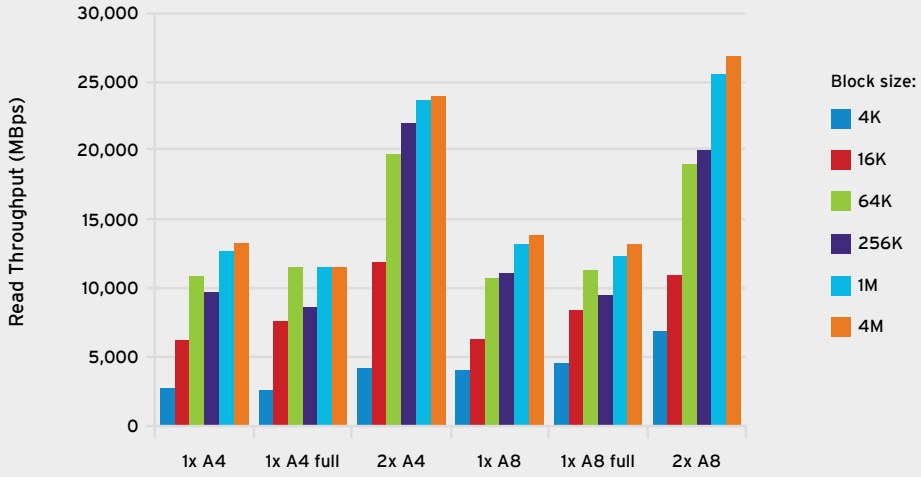

Figure 14. Read throughput across the test clusters

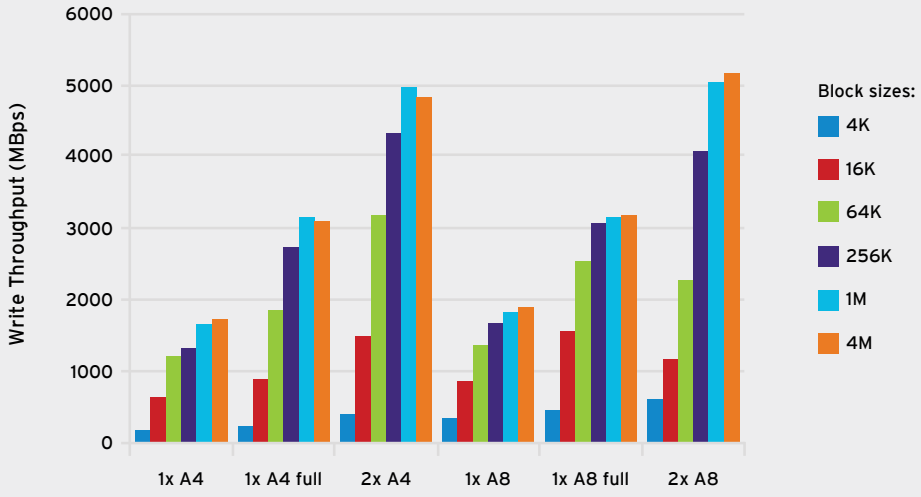

Figure 15. Write throughput scalability

LATENCY

Low latency storage is particularly important for IOPS-intensive workloads as longer latency can impact transaction completion and slow database performance. As shown in Figure 16, the tested InfiniFlash clusters uniformly demonstrated very low latency for smaller block sizes, with latency gradually rising for larger block sizes. Note that larger block sizes have been omitted from these data in the interest of practicality and clarity.

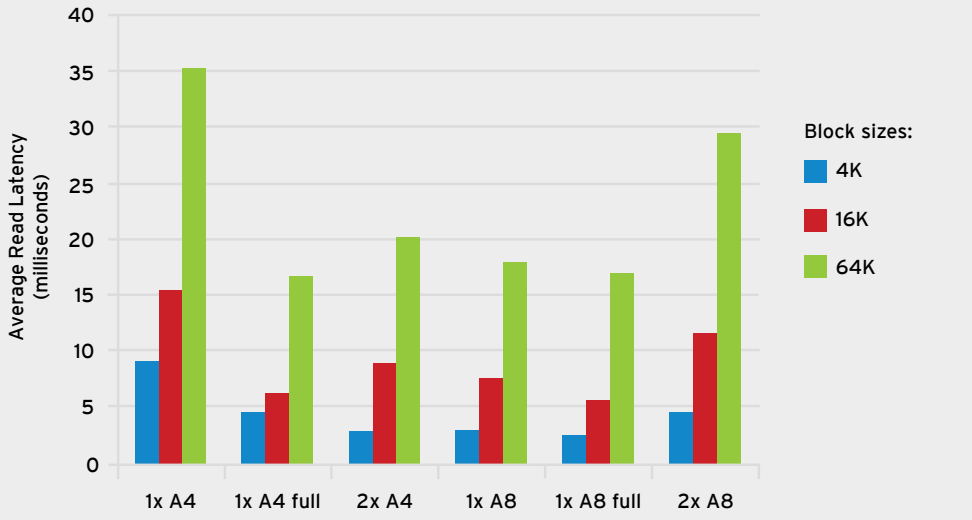

Figure 16. Average read latency remains very low, but grows with block size as expected.

99th-percentile latency reflects the upper bound of latencies experienced by 99% of transactions. In other words, 99% of the transactions experienced less than the 99th-percentile latency. Figure 17 shows 99th-percentile read latency across all of the clients in the tested cluster configurations.

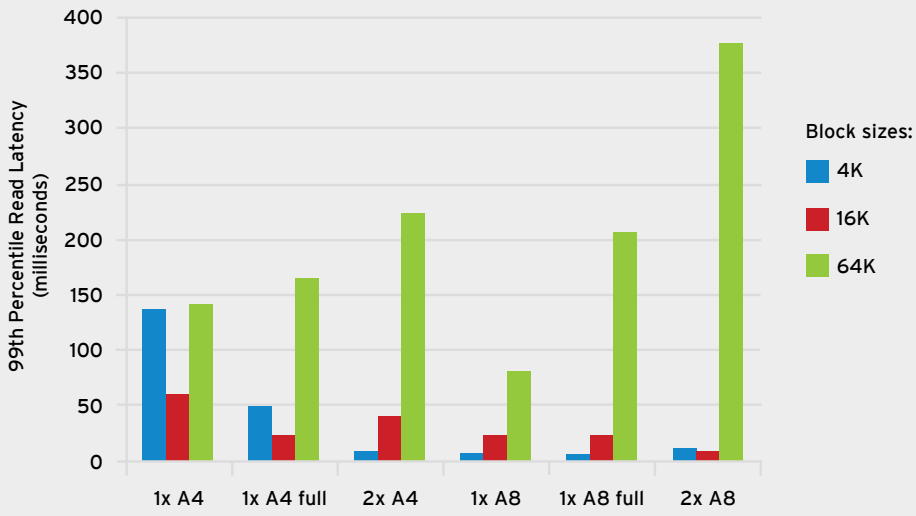

Figure 17. 99th-percentile read latency

Average write latency is shown in Figure 18, again demonstrating very low latency for smaller block sizes with growing latency for larger block sizes. Latency is reduced as the cluster scales to multiple InfiniFlash chassis and larger numbers of OSD nodes.

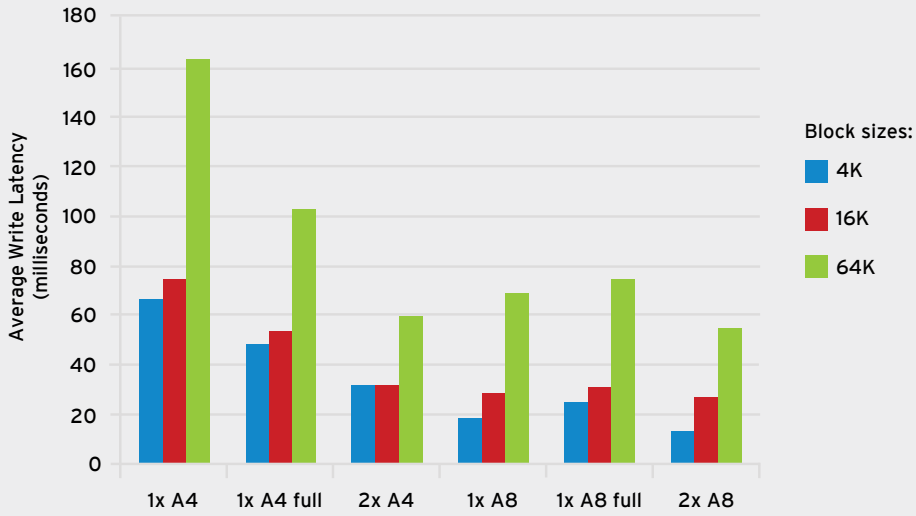

Figure 18. Average write latency.

99th-percentile write latency is shown in Figure 19. Block size ultimate overwhelms the natural scaling patterns of the various cluster sizes.

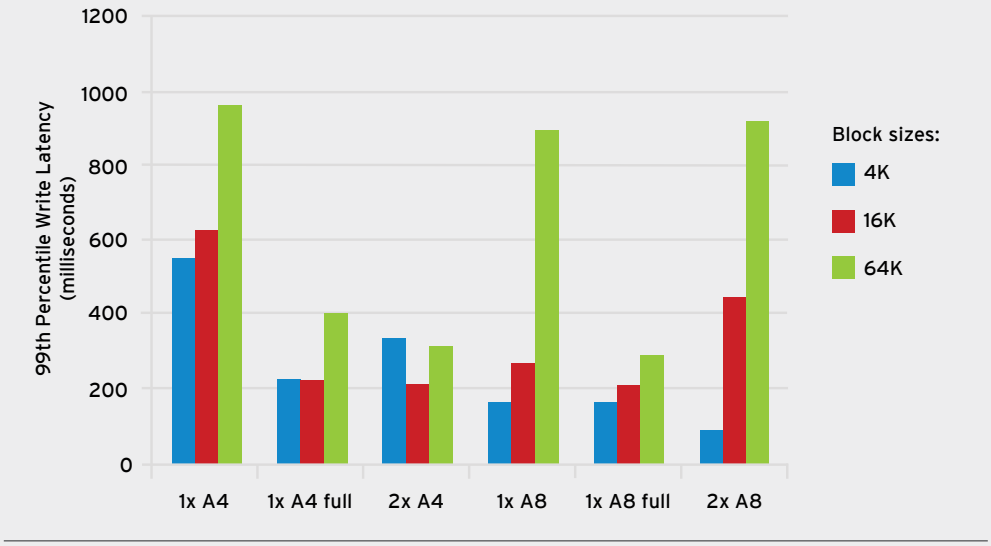

Figure 19. 99th-percentile write latency.

FAILURE AND RECOVERY SCENARIOS

Ceph provides a choice of data protection methods, supporting both replication and erasure coding. Replicated Ceph pools were used for Red Hat and SanDisk testing. Replication makes full copies of stored objects, and is ideal for quick recovery. In a replicated storage pool, Ceph typically defaults to making three copies of an object with a minimum of two copies for clean write operations. However, the greater reliability of InfiniFlash HDD-based storage makes 2x replication a viable alternative, coupled with a data backup strategy. Replication with two copies also saves on the overall storage required for a given solution.

Two recovery configurations were tested with different Ceph performance settings. A "default" configuration represents typical default settings for Ceph. A "minimum" configuration was also tested. that represents minimum impact to I/O performance and an absolute worst case for recovery time. The minimum configuration represents minimum impact to I/O performance and an absolute worst case for recovery time. Table 6 provides system parameters for the two scenarios.

TABLE 6. TESTED RECOVERY SETTINGS

SYSTEM PARAMETER	DEFAULT RECOVERY	MINIMUM I/O IMPACT
osd_recovery_max_active	15	1
osd_max_backfills	10	1
osd_recovery_threads	1	1
osd_recovery_op_priority	10	1

Failure and recovery tests were performed on a Red Hat Ceph Storage cluster utilizing eight OSD servers connected to a single fully populated InfiniFlash IF150. The cluster supported 108TB of data with roughly 1.7TB of data per flash card within the InfiniFlash system, and per OSD on the connected OSD node.

SINGLE OSD DOWN AND OUT

Figure 20 illustrates the effects of removing single OSD and its corresponding flash card with ~1.7TB of data from the cluster, without replacing it. The yellow line shows that the time required to regain optimal I/O performance was roughly 13 minutes, accompanied by a large drop in workload I/O performance. The minimal I/O impact configuration with minimal threads required only 45 minutes for the cluster to regain optimal I/O performance, accompanied by very little impact to client workload performance.

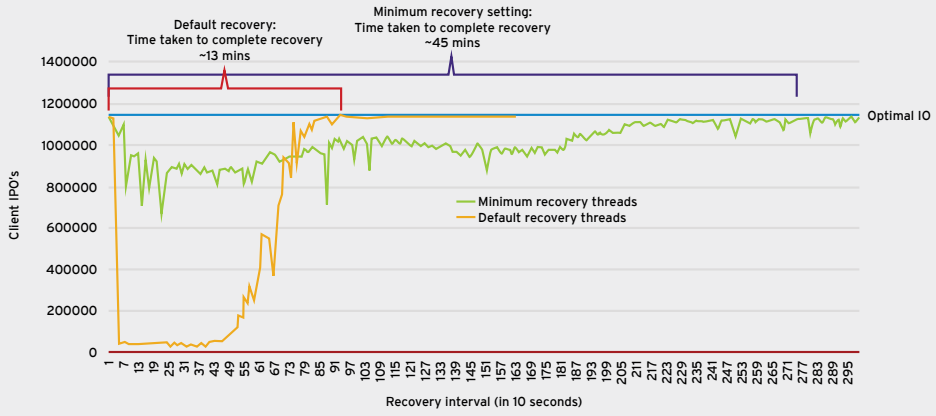

Figure 20. With a single OSD down and out, the cluster takes roughly 13 minutes to recover to full I/O performance.

SINGLE OSD BACK IN

Figure 21 shows the effect of replacing the missing OSD and allowing the cluster to rebuild. Default recovery of the 1.7TB of data was slightly more than an hour. Minimum-impact recovery was ~2.45 hours. Importantly, I/O performance was very close to optimal even for much of the recovery period for both scenarios, with only a brief dip early in the recovery process.

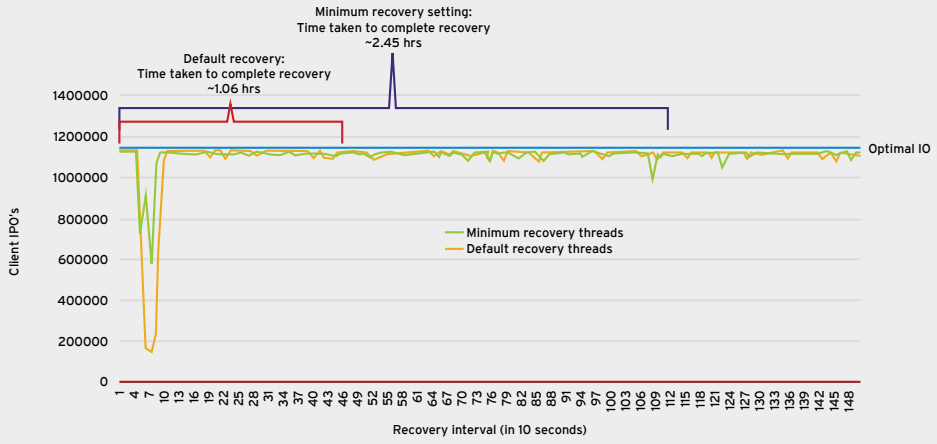

Figure 21. With the single OSD replaced into the cluster, recovery to optimal I/O takes approximately an hour.

FULL OSD NODE DOWN AND OUT

The removal of a full host from the cluster is a more significant event. As tested, a full host represented 13.6TB of data, with eight flash devices connected to each OSD node, each with 1.7TB of data. As shown in Figure 22, recovery for the default configuration was roughly one hour, accompanied by a large drop in client workload performance. Minimum-impact recovery required ~4.5 hours, but was accompanied by a much smaller drop in client workload performance.

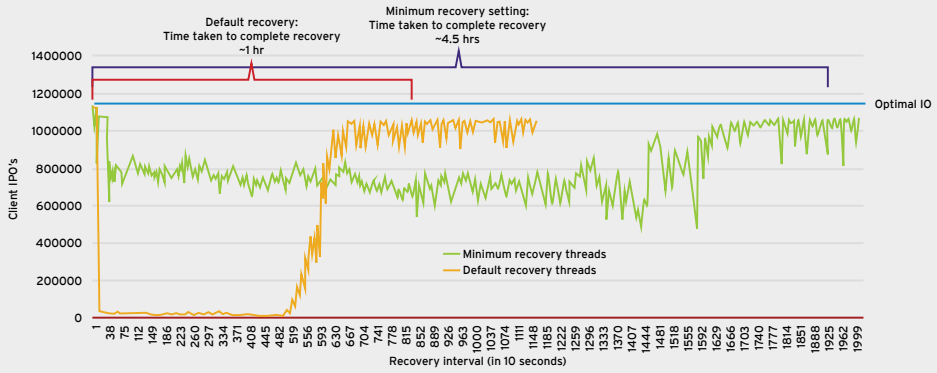

Figure 22. Recovery to optimal I/O performance takes approximately an hour with a full OSD node down and out of the cluster.

FULL OSD NODE BACK IN

Figure 23 illustrates the effect on I/O performance of re-inserting the full OSD node back into the cluster. Full rebuild of the 13.6TB of data required 1.3 hours for the default configuration and a worst-case of ~4.3 hours for the minimum-impact configuration.

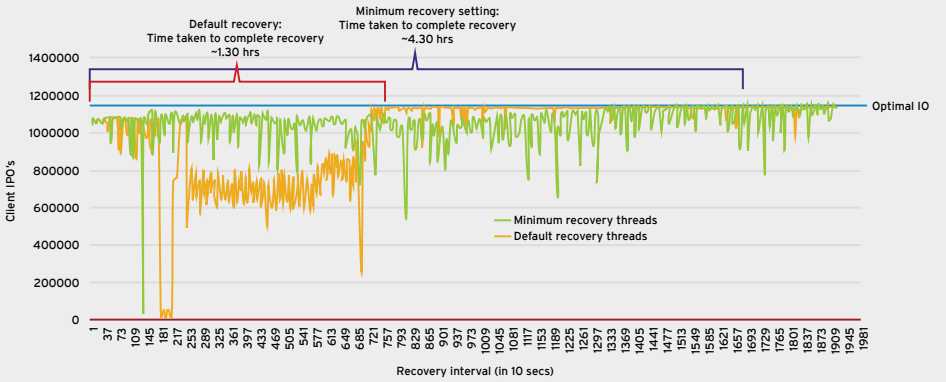

Figure 23. Replacing a full OSD node (13.6TB of data) requires roughly 1.3 hours for the cluster to return to optimal I/O performance.

SUMMARY

Red Hat Ceph Storage and the InfiniFlash system represent an excellent technology combination. The flexibility of decoupled flash storage allows CPU and storage to be adjusted independently, enabling clusters to be optimized for specific workloads. Importantly, Red Hat and SanDisk testing has shown that the InfiniFlash system can effectively serve both IOPS- and throughput-intensive workloads alike, by intelligently altering the ratio between InfiniFlash flash cards and OSD compute server hosts. With dramatically improved reliability, InfiniFlash-based Red Hat Ceph Storage clusters can be configured with 2x replication for considerable savings over 3x replicated configurations using other media.

ABOUT RED HAT

Red Hat is the world's leading provider of open source software solutions, using a community-powered approach to provide reliable and high-performing cloud, Linux, middleware, storage, and virtualization technologies. Red Hat also offers award-winning support, training, and consulting services. As a connective hub in a global network of enterprises, partners, and open source communities, Red Hat helps create relevant, innovative technologies that liberate resources for growth and prepare customers for the future of IT.

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
#US114575_0017