Dell HPC NFS Storage Solution

High Availability (NSS7.3-HA) Configuration

February 2019

H17571

White Paper

Abstract

This white paper describes the configuration for Dell NFS Storage Solution (NSS)-High Availability (HA) for version NSS7.3-HA of the solution. It describes the technical details, evaluation method, and expected performance of the solution.

Dell EMC Solutions

Copyright

Contents

Revisions

Table 1. Revisions

Date	Description
January 2019	Initial version

Executive summary

This white paper describes the Dell NFS Storage Solution-High Availability (NSS7.3-HA) configuration. It compares all available NSS-HA offerings and provides performance results for a configuration with a storage system providing 1,008 TB of raw storage capacity (768 TB of formatted space).

The NSS-HA solution that is described in this white paper is designed to enhance availability of storage services to the high performance computing (HPC) cluster by using a pair of Dell PowerEdge servers and Dell EMC PowerVault storage arrays with the Red Hat Enterprise Linux High Availability Add-On. The goal of the solution is to improve storage service availability and maintain data integrity if failures or faults occur and to optimize performance in a failure-free situation.

Introduction

This white paper provides information about the latest Dell NFS Storage Solution-High Availability (NSS7.3-HA) configurations with Dell PowerEdge 14th-generation servers. The solution uses the PowerEdge R740 servers and the Red Hat Enterprise Linux (RHEL) 7.5 operating system to deliver a more powerful storage solution than the earlier NSS-HA solutions. This release continues to provide an easy to manage, reliable, and cost-effective storage solution for HPC clusters.

The design principle for this release remains the same as previous Dell NSS-HA solutions. The major updates between the current and the previous release (NSS7.2-HA) of the NSS-HA solution are:

- PowerVault ME4 Series Storage replaces PowerVault MD3 Series Storage.
- The RHEL 7.5 operating system replaces the RHEL 7.4 operating system.

For more information about the NSS-HA solution family, see References, which includes NSS-HA white papers for previous releases.

We value your feedback

Dell EMC and the authors of this document welcome your feedback on the solution and the solution documentation. Contact the Dell EMC Solutions team by email or provide your comments by completing our documentation survey.

Authors: Mario Gallegos, Penelope Howe-Mailly

NSS-HA solutions overview

This section provides a brief description of the NSS-HA solution and contrasts the currently available and previous NSS-HA offerings.

Introduction to NSS-HA solutions

The design of the NSS-HA solution for each version is similar. The core of the solution is an HA cluster(6), which provides a highly reliable and available storage service to HPC compute clusters by using a high performance network connection such as Intel Omni-Path (OPA), InfiniBand (IB), or 10 Gigabit Ethernet (10 GbE).

The HA cluster consists of a pair of Dell PowerEdge servers and a network switch. The two PowerEdge servers have shared access to disk-based Dell EMC PowerVault storage arrays in a variety of capacities. Both servers are directly connected to the HPC cluster by using OPA, IB or 10 GbE. The two servers are equipped with two fence devices:

- iDRAC 9 Enterprise
- An APC Power Distribution Unit (PDU)

If failures, such as storage or network disconnection and a nonfunctioning system, occur on one server, the HA cluster fails over the storage service to the healthy server with the assistance of the two fence devices. It also ensures that the failed server does not take back control without the administrator's knowledge or control.

The disk-based storage array is formatted as a Red Hat Scalable file system (XFS) and is exported to the HPC cluster by using the NFS service of the HA cluster. Note that large-capacity file systems (greater than 100 TB) have been supported since the second release of NSS-HA solution (2).

The following figure shows the general infrastructure of the NSS-HA solution. For more information about the NSS-HA solution, see the previous NSS-HA white papers listed in References.

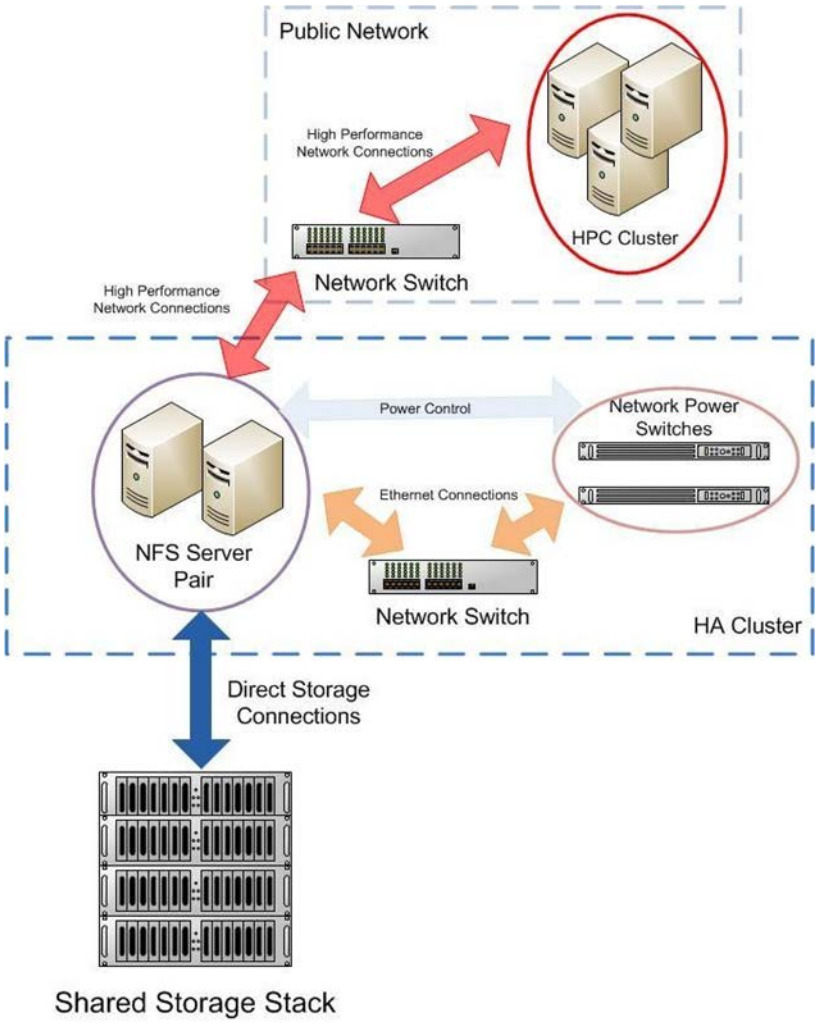

Figure 1. NSS-HA solution infrastructure

Note: The figure does not show the iDRAC 9 Enterprise, which is installed on each NFS server for NSS-HA solutions. The term “Network Power Switches” refers to the APC PDU in NSS-HA solutions.

Dell NSS-HA offerings

This NSS-HA release incorporates Dell EMC PowerVault ME4084 storage arrays (9) and RHEL 7.5. It continues to use the Intel Xeon Scalable Processor Family CPUs (code name Skylake) to offer higher overall system performance than previous NSS-HA solutions.

A major improvement from the NSS7.2-HA solution to the NSS7.3-HA solution is the increase in maximum capacity. The NSS7.2-HA solution is limited by the Red Hat XFS current support limit of 500 TB. After extensive testing and validation in our labs, Dell EMC and Red Hat reached a cooperative agreement that supports NSS7.3-HA configurations with up to 768 TB of usable space, that is, a PowerVault ME4084 array that is fully populated with 12 TB HDDs or 1,008 TB of raw storage space.

The following table contrasts the last two NSS-HA solutions with standard configurations.

Note: The table only lists NSS-HA versions currently available. Previous NSS-HA versions are no longer available.

Table 2. NSS-HA solutions comparison (1)(2)(3)(4)(5)(6)

Component	NSS7.2-HA Release (April 2018) PowerEdge 14th-generation servers and MD3460 + MD3060e-based solution	NSS7.3-HA Release (October 2018) PowerEdge 14th-generation servers and ME4084-based solution
Software	Red Hat Enterprise Linux 7.4, Kernel 3.10.0-693.el7.x86_64 Red Hat Scalable File system (XFS) v4.5.0-12	Red Hat Enterprise Linux 7.5, Kernel 3.10.0-862.el7.x86_64 Red Hat Scalable File system (XFS) v4.5.0-15
NFS servers	Two Dell PowerEdge R740 servers CPU: Dual Intel Xeon Gold 6136 @ 3.0 GHz, 12 cores per processor Memory: 12 x 16 GiB 2,666 MT/s RDIMMs	
External network connectivity	IB EDR, 10 GbE, or Intel Omni-Path For this document, Mellanox ConnectX-4 IB EDR For purchase orders, CX-5 IB EDR	
Internal connectivity	Gigabit Ethernet, switch Dell Networking S3048-ON	
OpenFabrics Enterprise Distribution (OFED) version	Mellanox OFED 4.3-1.0.1.0	Mellanox OFED 4.4-1.0.0
Direct Storage connection	12 Gbps SAS connections.	
Storage subsystem	• Dell EMC MD3460 + optional MD3060e arrays • 60 to 120 3.5 in. NL SAS 4 TB drives • 2 configurations with 240 TB or 480 TB (raw space) • 6 or 12 LUNs, 8+2 RAID 6, segment size 512 KiB • No spares	• Dell EMC PowerVault ME4084 • 84 3.5 in. NL SAS drives, up to 12 TB • 1 configuration of up to 1,008 TB (raw space) • 8 LUNs, linear 8+2 RAID 6, chunk size 128 KiB • 4 Global HDD spares
PDUs	2 APC-switched PDUs to manage high availability. For information about supported APC PDU models, see Knowledge Base article 82603: Fence device and Agent Information for Red Hat Enterprise Linux (7).	
Support and services	3 years of Dell ProSupport for IT and Mission Critical 4-hour 7x24 onsite options. Dell deployment services are available to accelerate installation, optimize performance, and integrate the NSS-HA solution with customers' HPC cluster environment.	

Note: Contact your Dell Sales Representative to discuss which HPC storage solution is suited for your environment. You can order any of the preconfigured solutions or a customized solution that is designed to fulfill your requirements. Some of the best practices that are discussed in this white paper might not apply to customized solutions.

Dell EMC PowerVault ME4084 storage array

The PowerVault ME4084 (9) storage array is a 5U, 84-drive dense enclosure with dual active-active RAID controllers and is used as a RAID Bunch of Disks (RBOD).

The NSS7.3-HA solution has a new storage configuration that takes advantage of the new storage array, while offering more flexibility compared to previous versions of the solution. The following table summarizes the supported NSS7.3-HA storage configurations.

Table 3. NSS7.3-HA solution storage configurations

Components	Configuration details	
Storage configurations	Up to 768 TB of usable space using 1 PowerVault ME4084 array	
3.5 in. NL SAS disks in the storage array	**HDD-supported size**	**Usable space (raw space)**
	4 TB	256 TB (336 TB)
	8 TB	512 TB (672 TB)
	10 TB	640 TB (840 TB)
	12 TB	768 TB (1,008 TB)
Virtual disk configuration	• 8 linear RAID6 8+2 per enclosure • Chunk (segment) size 128 KiB • Write cache mirroring enabled • Read ahead enabled with a value of “Stripe Size” • 4 spare disks that are enabled for use as dynamic spares, which immediately replace failed disks without intervention • 2 spares disks per tray	
Storage enclosure cabling	• Expansion (ME4084 to M484) SAS cabling is not required for this release because only one enclosure is supported. • Future releases might support an ME484 expansion and require SAS cabling.	
Logical volume configuration	• Stripe element size: 512 KiB • Number of stripes: 8 • Number of virtual disks per logical volume: 8	

The new PowerVault ME4084 storage array continues to use linear 8+2 RAID 6 as the basic building unit with a new chunk size (segment size) of 128 KiB and a read ahead value of “stripe size” that is selected for optimum performance. Because we now have 84 drives, we have eight LUNs that are based on the RAID 6 arrays and four global spare HDDs configured to replace any failed disk immediately. Therefore, this solution can have up to 768 TB of usable space (when using 12 TB HDDs).

Evaluation

The Dell HPC lab evaluated the architecture that is proposed in this white paper. This section describes the test methodology and the test bed that was used for verification. It provides information about the functionality tests. Performance tests and results are described in NSS7.3-HA I/O performance.

Test method

We used a 640 TB NSS7.3-HA configuration to test the HA functionality and performance of the NFS Storage Solution that is described in this white paper. We introduced different types of failures and verified the fault-tolerance and robustness of the solution. HA functionality describes these HA functionality tests and their results. The HA functionality testing was similar to that of the previous releases of the solution.

Evaluation test bed

The following figure shows the test bed that was used to evaluate NSS7.3-HA functionality and performance.

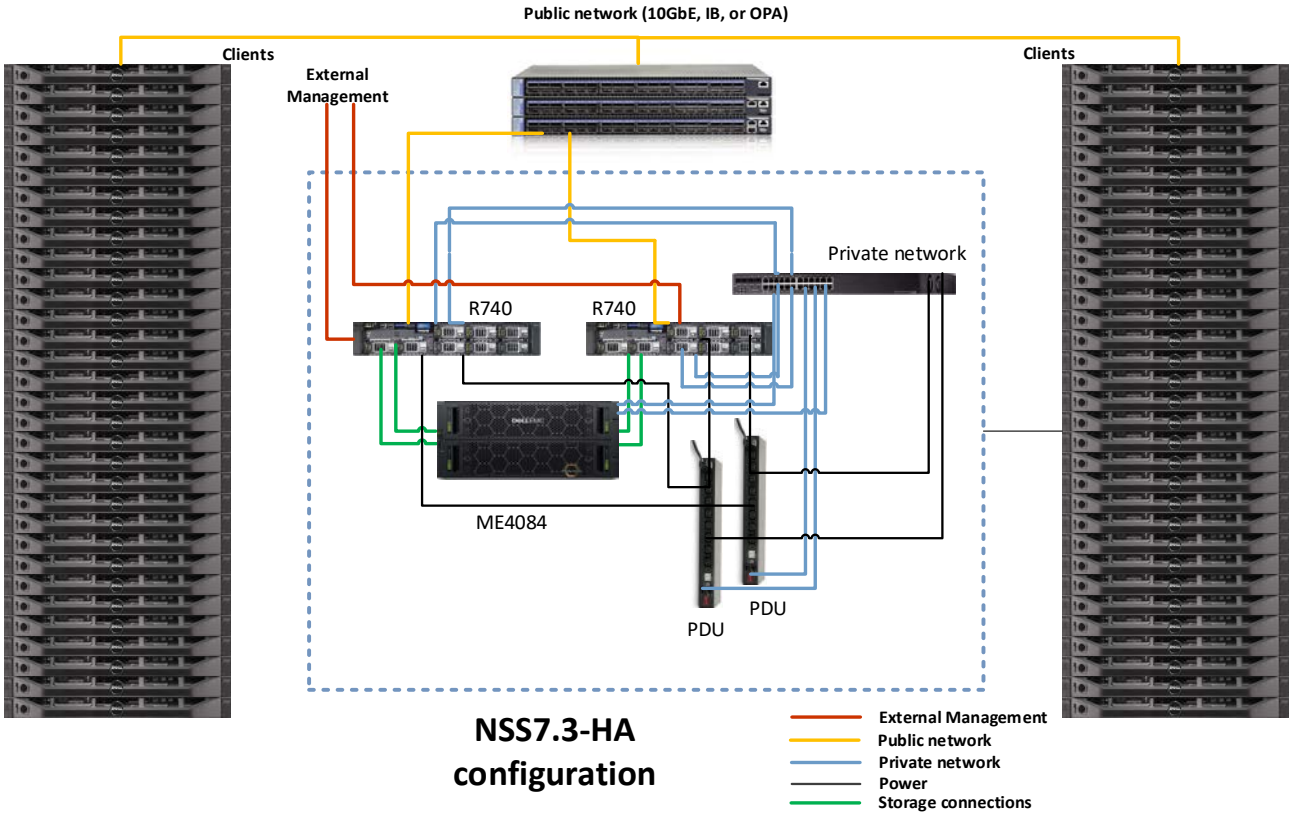

Figure 2. NSS test bed

For HA functionality testing, we set up the following configuration:

- Used a 32-node HPC compute cluster (also known as “the clients”) to provide I/O network traffic for the test bed.
- Configured a pair of Dell PowerEdge R740 servers as an active-passive HA pair that functioned as an NFS server for the HPC compute cluster.
- Connected both NFS servers to a shared PowerVault ME4084 storage enclosure at the back-end. The user data is stored on an XFS file system that is created on this storage. The XFS file system was exported to the clients by using NFS.
- Connected the NFS servers to the clients by using the public network, which was IB EDR.
- For the HA functionality of the NFS servers, configured a private 1 Gigabit Ethernet network to monitor server health and heartbeat, and to provide a route for the fencing operations by using a Dell Networking 3048-ON Gigabit Ethernet switch.

- Used two APC-switched PDUs on two separate power buses to provide power to the NFS servers.

The following tables provide complete configuration details.

Table 4. NSS7.3-HA hardware configuration

Component	Description
NFS server model	Two Dell PowerEdge R740 servers
Processor	Dual Intel Xeon Gold 6136 @ 3.0 GHz, 12 cores per processor
Memory	12 x 16 GiB 2,666 MT/s RDIMMs
Local disks and RAID controller	PERC H730 with five 300 GB 15K SAS hard drives. Two drives are configured in RAID1 for the operating system, two drives are configured in RAID0 for swap space, and the fifth drive is a hot spare for the RAID1 disk group.
Optional IB EDR CX5 HCA (slot 4)	Mellanox IB EDR CX5 Host Channel Adapter (HCA)
Optional OPA HCA (slot 4)	Intel Omni-Path Host Fabric Interface (HFI)
1 GbE Ethernet Card (daughter card slot)	Broadcom 5720 QP 1 Gigabit Ethernet network daughter card
SAS HBAs (slot 1 and slot 6)	Two 12 Gbps SAS Dell HBAs
Systems management	iDRAC 9 Enterprise version
Power supply	Dual PSUs

Table 5. NSS7.3-HA software versions

Component	Description
Operating system	Red Hat Enterprise Linux 7.5
Kernel version	3.10.0-862.el7.x86_64
Cluster Suite	Red Hat Cluster Suite from RHEL 7.5
File system	Red Hat Scalable File System (XFS) v4.5.0-15
Systems management tool	Dell Open Manage Server Administrator 9.1.0
OFED version	Mellanox OFED 4.4-1.0.0

Table 6. NSS7.3-HA client cluster configuration

Clients	32 PowerEdge C6420 servers Each compute node has: • CPU: Intel(R) Xeon(R) Gold 6148 CPU @ 2.40GHz, 20 cores per processor • Memory: 192 GiB • Red Hat Enterprise Linux 7.4, kernel 3.10.0-693.17.1.el7.x86_64
HCA card	Mellanox ConnectX-4 VPI IB EDR/100 GbE Single Port QSFP28
OFED	MLNX_OFED -4.3-1.0.1.0

HA functionality

The solution's HA functionality was tested by simulating several component failures. The test designs and test results are similar to previous releases of the solution because the general architecture of the solution has not changed. This section reviews the failures and fault tolerant mechanisms in NSS-HA solutions, and then presents the HA functionality tests that are relative to different potential failures and faults.

Potential failures and fault tolerant mechanisms in NSS-HA

There are many different types of failures that can impact the functionality of the NSS-HA solutions. The following table lists the potential failures that are tolerated in NSS-HA solutions.

Note: The following analysis assumes that the HA cluster service is running on the active server and the passive server is the other standby component of the cluster.

Table 7. NSS-HA mechanisms to handle failures

Failure type	If a failure occurs
Single local disk failure on a server	The operating system is installed on a two-disk RAID1 device with 1 hot spare. A single-disk failure is unlikely to cause the server not to function.
Single server failure	The cluster service monitors this failure. Service fails over to passive server.
Power supply or power bus failure	There are dual PSUs in each server. Each PSU is connected to a separate power bus. The server continues to function with a single PSU.
Fence device failure	iDRAC 9 Enterprise is used as the primary fence device. The switched PDUs are used as secondary fence devices.
SAS cable/port failure	There are 2 SAS cards in each NFS server. Each card has a SAS cable to each controller in the shared storage. A single SAS card or cable failure does not impact data availability.
Dual SAS cable/card failure	The cluster service monitors this failure. If all data paths to the shared storage are lost, service fails over to the passive server.
OPA/IB/10 GbE link failure	The cluster service monitors this failure. Service fails over to the passive server.
Private switch failure	Cluster service continues on the active server. If there is an additional component failure, service is stopped and system administrator intervention is required.
Heartbeat network interface failure	The cluster service monitors this failure. Service fails over to the passive server.
RAID controller failure on the PowerVault ME4084 storage array	There are dual controllers in the PowerVault ME4084 array. The second controller manages all data requests.

HA tests for the NSS-HA solution

We verified functionality for an NFSv4-based solution. The following failures were simulated on the cluster with consideration of the failures that are listed in Table 7:

- Server failure
- Heartbeat link failure
- Public link failure

- Private switch failure
- Fence device failure
- Single SAS link failure
- Multiple SAS link failures

The NSS-HA behaviors in response to these failures include:

- **Server failure**—Simulated by introducing a kernel panic. When the active server stops functioning, the heartbeat between the two servers is interrupted. The passive server waits for a defined period, and then attempts to fence the active server. After fencing is successful, the passive server takes ownership of the cluster service. Clients cannot access the data until the failover process is completed.
- **Heartbeat link failure**—Simulated by disconnecting the private network link on the active server. When the heartbeat link is removed from the active server, both servers detect the missing heartbeat and attempt to fence each other. The active server is unable to fence the passive server because the missing link prevents it from communicating over the private network. The passive server successfully fences the active server and takes ownership of the HA service.
- **Public link failure**—Simulated by disconnecting the Intel OPA, IB, or 10 GbE link on the active server. The HA service is configured to monitor this link. When the public network link is disconnected on the active server, the cluster service stops on the active server and is relocated to the passive server.
- **Private switch failure**—Simulated by turning off the private network switch. When the private switch fails, both servers detect the missing heartbeat from the other server and attempt to fence each other. Fencing is unsuccessful because the network is unavailable and the HA service continues to run on the active server.
- **Fence device failure**—Simulated by disconnecting the iDRAC 9 Enterprise cable from a server. If the iDRAC on a server stops responding, the server is fenced by using the network PDUs, which are defined as secondary fence devices during the configuration.
- **Single SAS link failure**—Simulated by disconnecting one SAS link between the Dell PowerEdge R740 server and the Dell EMC PowerVault ME4084 storage array. When only one SAS link fails, cluster service is not interrupted. Because there are multiple paths from the server to the storage, a single SAS link failure does not break the data path from the clients to the storage and does not trigger a cluster service failover.
- **Multiple SAS link failures**—Simulated by disconnecting all SAS links between one PowerEdge R740 server and the PowerVault ME4084 storage array. When all SAS links on the active server fail, the HA service attempts to fail over to the passive server. At this point, the passive server fences the active server, restarts the HA service, and provides a data path again to the clients. This failover can usually take less than two minutes.

For all of behaviors, other than for the SAS link failures, we observed that the HA service failover takes between 30 seconds to 60 seconds. In a healthy cluster, the Red Hat cluster management daemon notes any failure event and acts on it within minutes. Note that this failover time is on the NFS servers. The impact to the clients might be longer.

Impact to clients

Clients mount the NFS file system that the server exports by using the HA service IP address. This IP address is associated with either OPA, IP over IB (IPoIB), or a 10 GbE network interface on the NFS server. To measure any impact to the client, we used the dd utility and the IOzone benchmark tool to read and write large files between the clients and the file system. We introduced component failures into the server while the clients were actively reading and writing data from and to the file system.

In all tests, the client processes completed the read and write operations successfully. As expected, the client processes take longer to complete if the process is actively accessing data during a failover event.

During the failover period, when the data share was temporarily unavailable, the client processes were in an uninterruptible sleep state.

Depending on the characteristics of the client processes, the processes can be expected to either stop abruptly or sleep while the NFS share is temporarily unavailable during the failover process. Any data that has already been written to the file system is available after the failover is completed.

For read and write operations during the failover case, data accuracy was successfully verified by using the checkstream utility.

See Appendix A: Benchmarks and test tools for information about the IOzone benchmark tool and the checkstream and dd utilities.

NSS7.3-HA I/O performance

This section presents the results of the I/O performance tests for the current NSS-HA solution. All performance tests were conducted in a failure-free scenario to measure the performance of the solution. The tests focused on two types of I/O patterns:

- Large sequential read and write operations
- Small random read and write operations

The 640 TB configuration was benchmarked with IB EDR 100 Gb/s network connectivity. The 32-node compute cluster that is described in Evaluation test bed was used to generate workload for the benchmarking tests. Each test was run over a range of clients to test the scalability of the solution.

The IOzone benchmark tool was used for sequential and random tests. For sequential tests, we used a request size of 1,024 KiB. The total size of data transferred was 512 GiB to ensure that the NFS server cache was saturated. Random tests used a 4 KiB request size and each client read and wrote a 4 GiB file. Metadata tests were performed by using the mdtest benchmark and included file create, stat, and remove operations. All of the test results were based on NFSv4. For a list of commands that were used in the tests, see Appendix A: Benchmarks and test tools.

Sequential writes and reads

In the sequential write and read tests, the I/O access patterns are N-to-N. That is, each client reads and writes to its own file. The IOzone tool was run in clustered mode and one thread was launched on each compute node for values of N up to 32, the number of compute nodes. The 64-thread data points were obtained by using 32 clients running two threads each. Because the total transferred data size was kept constant at 512 GiB, the file size per client varied accordingly for each test case. For example, a 512 GiB file was read or written in a one-client test case and a 256 GiB file was read or written per client node in a two-client test case.

The following figures contrast the sequential write and read performance for the NSS-HA 7.2 release and NSS7.3-HA release of the solution. The figures show the aggregate throughput that can be achieved when a number of clients are simultaneously writing or reading from storage over the IB EDR network fabric.

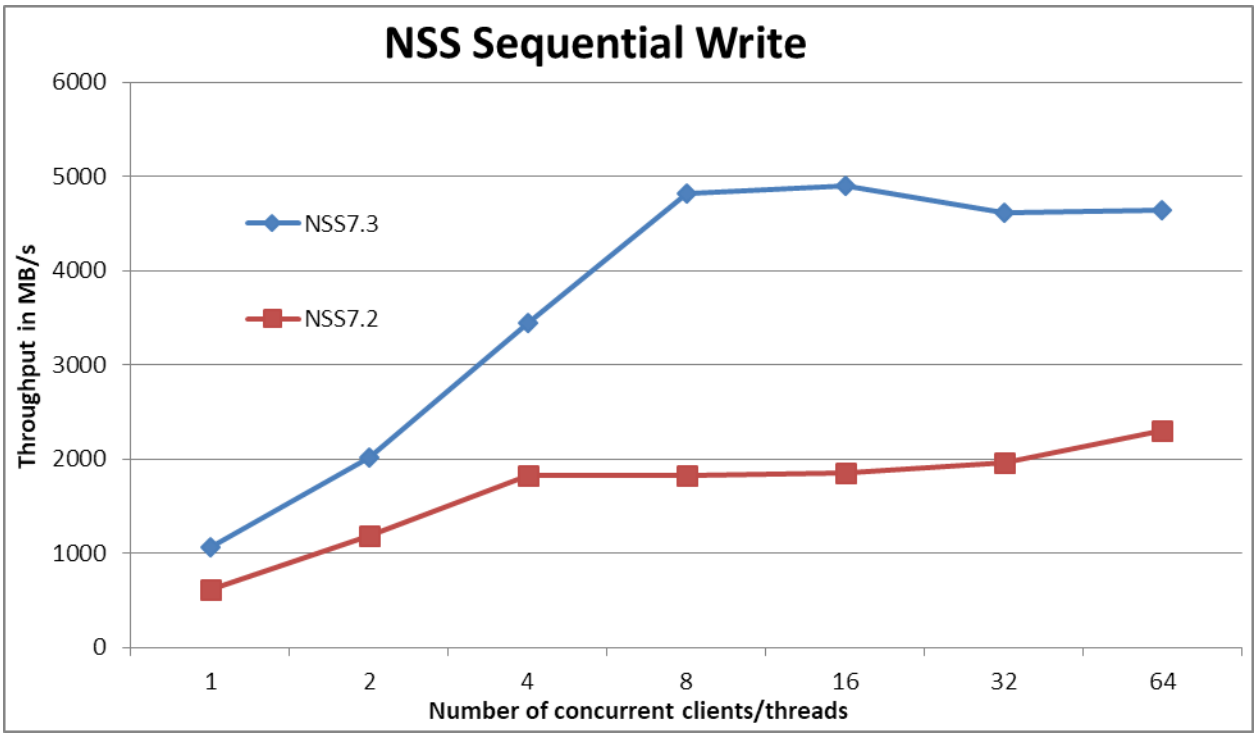

Figure 3. Large sequential write performance

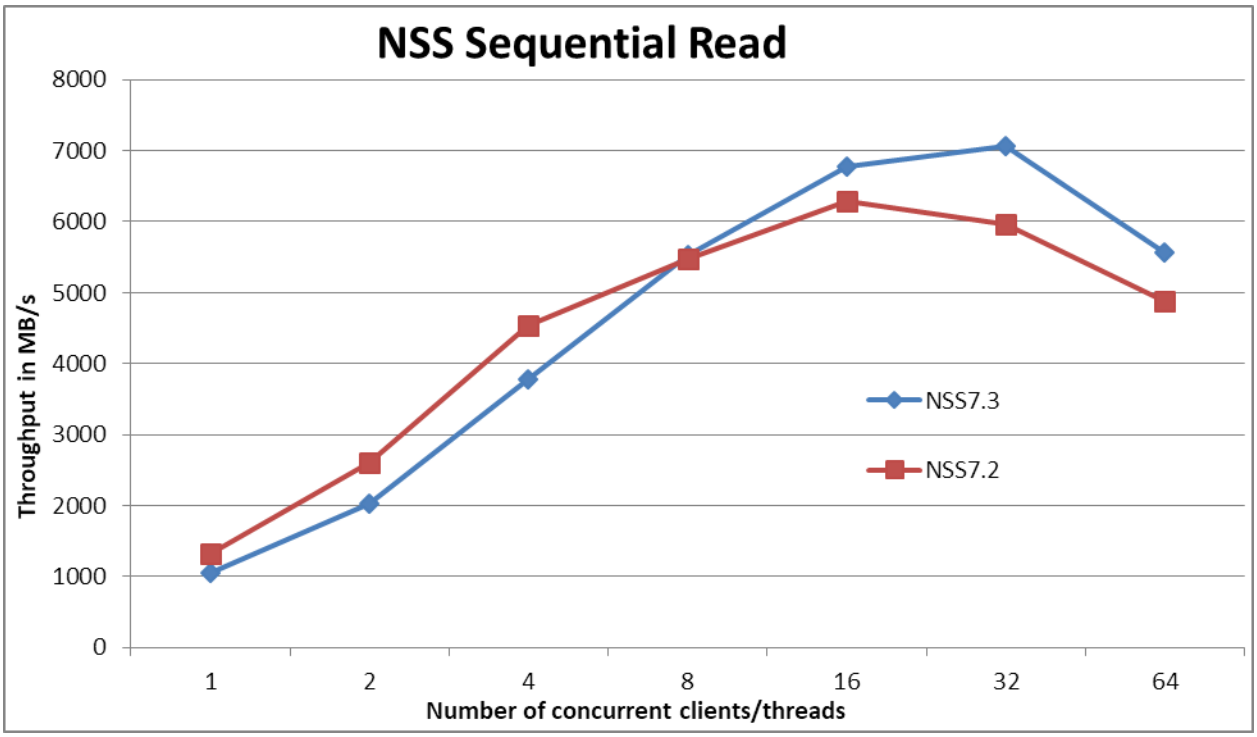

Figure 4. Large sequential read performance

During the benchmark tests, we observed that the peak read performance is about seven GB/sec and the peak write performance is almost 5 GB/sec. The figure makes it obvious that the current NSS7.3-HA solution has higher sequential performance numbers than the previous release. Reads are up to 18.7 percent better, but write performance is especially improved with up to 2.65 times (at 16 threads) the performance of the previous release. Comparing peak performance values, writes on the NSS7.3-HA solution are 2.13 times faster and reads are 12.5 percent better.

This performance response is partially due to the higher SAS speed of 12 Gbps for all PowerVault ME4084 internal components including HDDs (the internal speed for PowerVault MD3460 internal components was 6 Gbps). This internal speed enables a higher throughput per LUN. Also, the new storage controllers can process information faster than the previous generation of PowerVault MD3460 arrays.

Random write and read operations

The following figures show the random write and read performance. They show the aggregate I/O operations per second when a number of clients are simultaneously writing or reading to or from the storage over the IB EDR fabric.

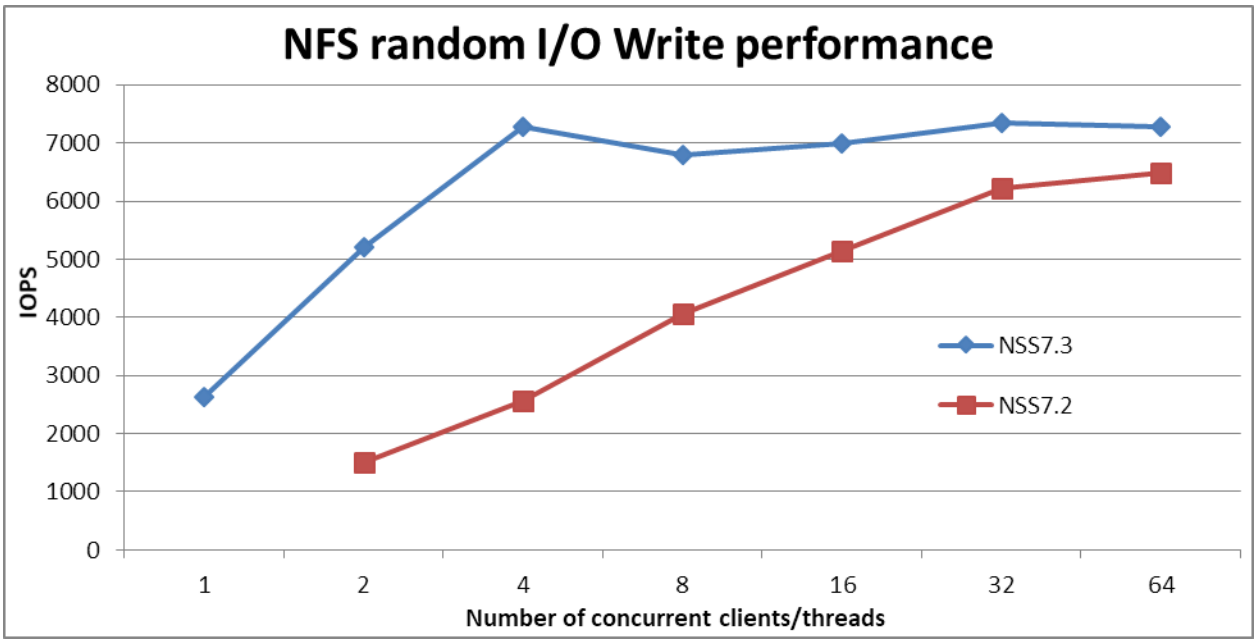

Figure 5. Random write performance

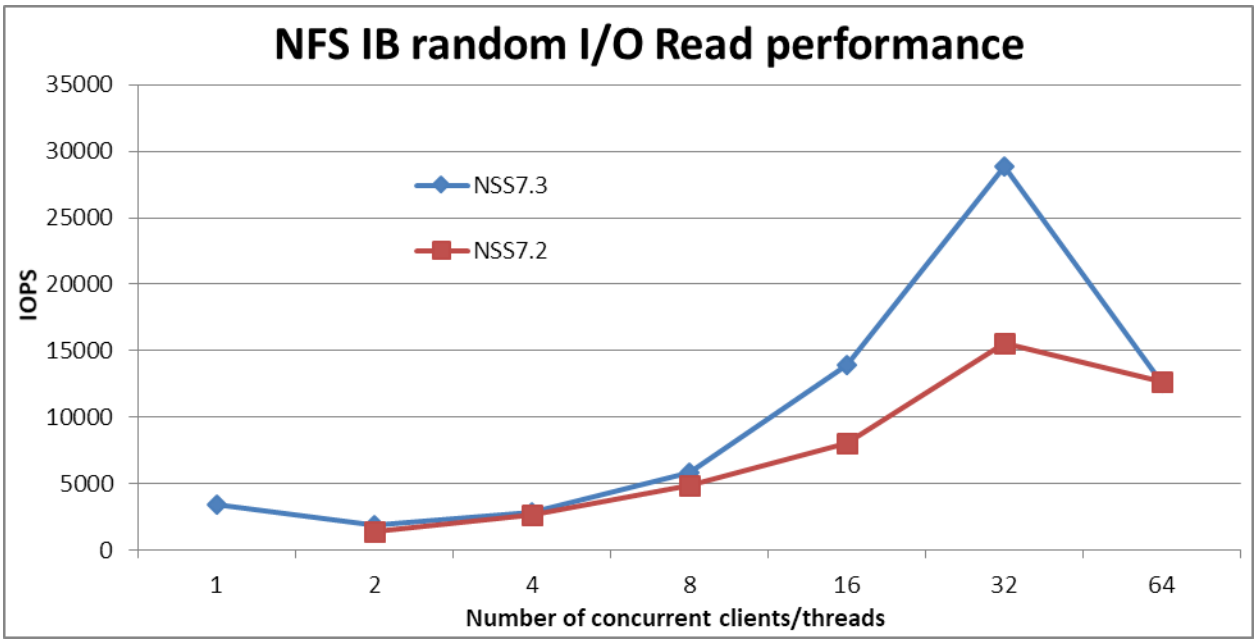

Figure 6. Random read performance

During the benchmark tests, we observed that the random write achieves peak performance at 32 threads, while the previous release of the solution peaked at 64 threads. The random read performance increases steadily on the NSS7.3 solution up to 32 clients and for the previous solution the peak was at 16 clients. Again, the new storage shows its superior performance with up to 3.44 times improvement on writes (at two threads) and 85 percent higher read performance (at 32 threads) over the previous

release. Comparing peak performances, the difference is about 13 percent on random writes and 85 percent on random reads. These improvements are primarily due to the new PowerVault ME4084 controllers that have faster processing capabilities than the PowerVault MD3460 controllers.

Metadata operations

The following figures show the results of file create, stat, and remove operations, respectively. Because the HPC compute cluster has only 32 compute nodes, each client executed a maximum of one thread for client counts up to 32. For thread counts of 64, 128, 256, and 512, each client executed 2, 4, 8, or 16 simultaneous operations (threads).

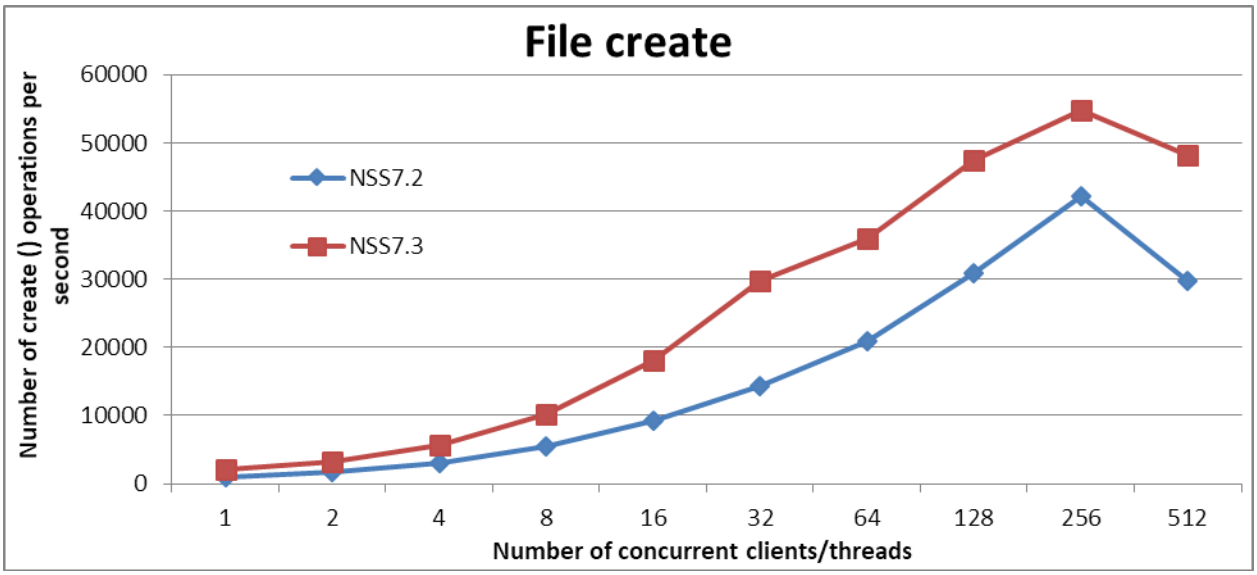

Figure 7. File create performance

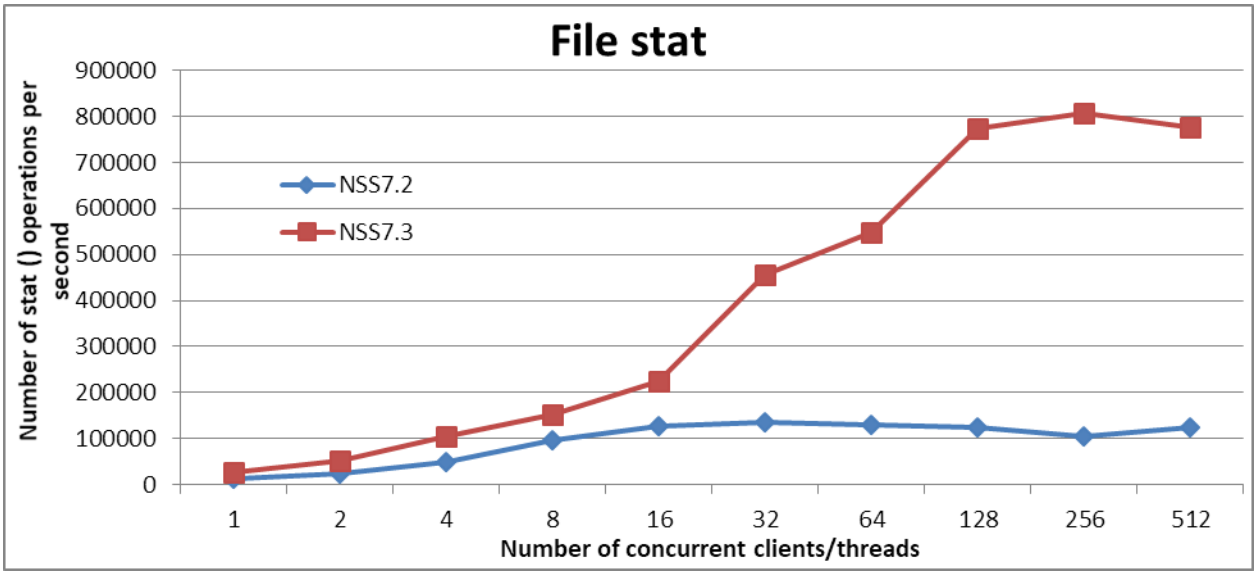

Figure 8. File stat performance

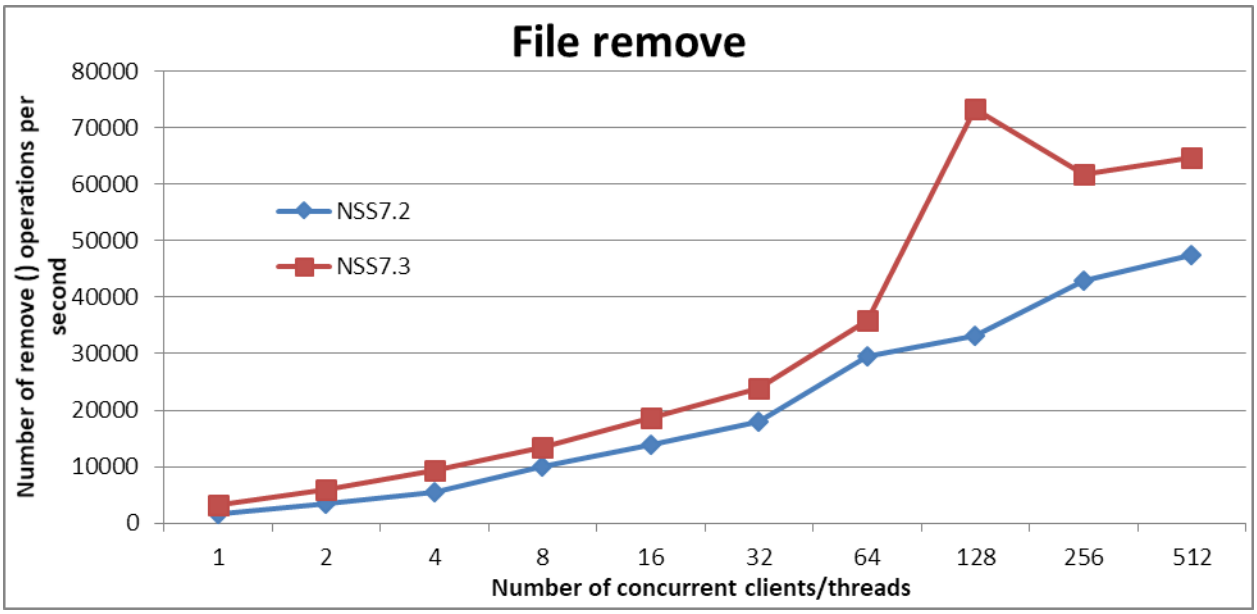

Figure 9. File remove performance

For file create operations, compared to the previous solution, the new solution shows a sustained improvement of about twice the performance with a peak difference (208 percent) at 32 clients, and then decreases slightly. However, comparing the peak performance for both solutions at 256 threads, the new solution is 30 percent faster.

Stat operations are the most improved by the new storage. Improvements are as high as 7.7 times that of the previous release at 256 threads. Comparing the peak performances, the NSS7.3 solution shows almost six times the number of stat operations per second than the previous release.

Finally, remove operations have comparatively marginal improvement with most data points at 33 percent or better performance than the previous release. However for 128 threads, performance is 2.21 times better. At peak performance, the new solution achieves almost 55 percent higher performance compared to the previous release.

These improvements are due to the faster HDDs using SAS3 speeds (12 Gbps), as well as to the new PowerVault ME4084 controllers that are capable of higher IOPs and bandwidth.

Summary

This white paper provides information about the latest Dell HPC NSS-HA solution, including the solution configuration, HA functionality evaluation, and performance evaluation. With this release, the Dell NSS7.3-HA solution supports the IB EDR network connection and delivers better sequential and random I/O performance than the previous release. The Dell NSS-HA solution is available with deployment services and full hardware and software support from Dell.

To show the performance difference between the NSS7.3-HA solution and the previous release (the NSS7.2-HA solution), the performance numbers of both solutions were contrasted, showing the superior performance of the latest release that is based on the PowerVault ME4084 array:

- Up to 2.65 times the sequential write performance and up to 18.7 percent faster read performance
- Up to 3.44 times the random write performance and up to 85 percent faster random read performance
- Up to 2.1 times the create rate, 7.7 times the stat rate, and 2.2 times the remove rate

Appendix A: Benchmark and test tools

The following tools were used to test the solution:

- The IOzone benchmark tool was used to measure sequential read and write throughput (MB/s) and random read and write I/O operations per second (IOPS).
- The mdtest HPC Benchmark tool was used to test the metadata performance of the file system.
- The checkstream utility was used to test for data correctness under failure and failover cases.
- The Linux dd utility was used for initial failover testing, to measure data throughput, and the time to complete file copy operations.

IOzone benchmark tool

We used IOzone version 3.420 for these tests. It was installed on both the NFS servers and all the compute nodes. Download IOzone from IOzone Filesystem Benchmark.

The IOzone tests were run from one to 32 nodes in clustered mode. All tests were N-to-N, which means that N clients read or write N independent files.

Between tests, the following procedure was followed to minimize cache effects:

- Unmount NFS share on clients.
- Stop the cluster service on the server, which unmounts the XFS file system on the server.
- Start the cluster service on the server.
- Mount the NFS share on clients.

The following table describes the IOzone command line arguments.

Table 8. IOzone command line arguments

IOzone argument	Description
-i 0	Write test
-i 1	Read test
-i 2	Random Access test
-+n	No retest
-c	Includes close in the timing calculations
-t	Number of threads
-e	Includes flush in the timing calculations
-r	Records size
-s	File size
-t	Number of threads
+m	Location of clients to run IOzone when in clustered mode
-w	Does not unlink (delete) temporary file

IOzone argument	Description
-I	Use O_DIRECT, bypass client cache
-O	Give results in ops/sec

For the sequential tests, file size was varied along with the number of clients so that the total amount of data written was 512 GiB (number of clients * file size per client = 512GiB).

- For IOzone sequential writes, we used the following command line:

```
# /usr/sbin/iozone -i 0 -c –e –w –r 1024k –s 16g –t 32 -+n -+m ./clientlist
```

- For IOzone sequential reads, we used the following command line:

```
# /usr/sbin/iozone -i 1 -c -e -w -r 1024k -s 16g -t 32 -+n -+m ./clientlist
```

For the random tests, each client read or wrote a 4 GiB file. The record size that is used for the random tests was 4 KiB to simulate small random data accesses.

- For IOzone IOPs random access (reads and writes), we used the following command line:

```
# /usr/sbin/iozone -i 2 -w -r 4k -I -O -w -+n -s 4g -t 1 -+m ./clientlist
```

By using the -c and `-e` arguments in the test, IOzone provides a more realistic view of what a typical application does. The `O_Direct` command line parameter enables us to bypass the cache on the compute node on which we are running the IOzone thread.

mdtest HCP Benchmark tool

We used version 1.9.3 for these tests. It was compiled and installed on an NFS share that was accessible by compute nodes. The mdtest HCP Benchmark tool is used with mpirun. For these tests, we used OpenMPI version 1.10.0. Download the mdtest HCP Benchmark from SourceForge.

The following table describes the mdtest command-line arguments.

Table 9.

Argument	Description
mpirun arguments	
-np	Number of processes
--nolocal	Instructs mpirun not to run locally
--hostfile	Tells mpirun where the hostfile is located
mdtest arguments	
-d	The directory in which mdtest must run
-i	The number of iterations that the test runs
-b	Branching factor of the directory structure
-z	Depth of the directory structure

Argument	Description
-L	Files only at the leaf level of the tree
-I	Number of files per directory tree
-y	Synchronize the file after writing
-u	Unique working directory for each task
-C	Create files and directories
-R	Randomly stat files
-T	Only stat files and directories
-r	Remove files and directories left over from the run

As with the IOzone random access patterns, the following procedure was followed to minimize cache effects during the metadata testing:

1. Unmount NFS share on clients.
2. Stop the cluster service on the server, which unmounts the XFS file system on the server.
3. Start the cluster service on the server.
4. Mount NFS Share on clients.

- For the metadata file and directory creation test, we used the following command line:

```
# mpirun -np 32 --nolocal --hostfile ./hosts /nfs/share/mdtest -d
/nfs/share/filedir -i 6 -b 320 -z 1 -L -I 3000 -y -u -t -C
```

- For the metadata file and directory stat test, we used the following command line:

```
# mpirun -np 32 --nolocal --hostfile ./hosts /nfs/share/mdtest -d
/nfs/share/filedir -i 6 -b 320 -z 1 -L -I 3000 -y -u -t -R -T
```

- For the metadata file and directory removal test, we used the following command line:

```
# mpirun -np 32 --nolocal --hostfile ./hosts /nfs/share/mdtest -d
/nfs/share/filedir -i 6 -b 320 -z 1 -L -I 3000 -y -u -t –r
```

checkstream utility

Version 1.0 was installed and compiled on the NFS servers and used for these tests. The checkstream utility is available from SourceForge.

First, the genstream utility created a large file. Each client copied this file to and from the NFS share by using the dd utility to simulate write and read operations. Failures were simulated during the file copy process and the NFS service was failed over from one server to another. The checkstream utility tested the resulting output files for data accuracy and to ensure that there was no data corruption.

The following example shows sample output of a successful test with no data corruption:

```
checkstream[genstream.file.100G]: ----------------------------------------------
checkstream[genstream.file.100G]: valid data for 107374182400 bytes at offset 0
checkstream[genstream.file.100G]: ----------------------------------------------
checkstream[genstream.file.100G]: end of file summary
checkstream[genstream.file.100G]: [valid data] 1 valid extents in 261.205032
seconds (0.00382841 err/sec)
checkstream[genstream.file.100G]: [valid data] 107374182400/107374182400 bytes
(100 GiB/100 GiB)
checkstream[genstream.file.100G]: read 26214400 blocks 107374182400 bytes in
261.205032 seconds (401438 KiB/sec), no errors
```

For comparison, the following example shows sample output of a failing test with data corruption in the copied file. For example, if the file system is exported by using an NFS async operation and there is an HA service failover during a write operation, data corruption might occur.

```
checkstream[compute-00-10]: ----------------------------------------------
checkstream[compute-00-10]: valid data for 51087769600 bytes at offset 45548994560
checkstream[compute-00-10]:
checkstream[compute-00-10]: ----------------------------------------------
checkstream[compute-00-10]: end of file summary
checkstream[compute-00-10]: [valid data] 1488 valid extents in 273.860652 seconds
(5.43342 err/sec)
checkstream[compute-00-10]: [valid data] 93898678272/96636764160 bytes (87 GiB/90
GiB)
checkstream[compute-00-10]: [zero data] 1487 errors in 273.860652 seconds (5.42977
err/sec)
checkstream[compute-00-10]: [zero data] 2738085888/96636764160 bytes (2 GiB/90
GiB)
checkstream[compute-00-10]: read 23592960 blocks 96636764160 bytes in 273.860652
seconds (344598 KiB/sec)
checkstream[compute-00-10]: ----------------------------------------------
checkstream[compute-00-10]: encountered 1487 errors, failing
```

dd utility

The coreutils rpm, which is distributed with RHEL 7.5, provides the Linux dd utility. The dd utility was used to copy a file. The NFS file system was mounted at `/mnt/xfs` on the clients.

- To write data to the storage, we used the following command line:

  ```
  # dd if=/dev/zero of=/mnt/xfs/file bs=1M count=90000
  ```

- To read data from the storage, we used the following command line:

  ```
  # dd if=/mnt/xfs /file of=/dev/null bs=1M
  ```

References

1. Dell HPC NFS Storage Solution High Availability Configurations, Version 1.1
2. Dell HPC NFS Storage Solution—High Availability Configurations with Large Capacities, Version 2
3. Dell HPC NFS Storage Solution - High Availability (NSS4-HA) Configuration with Dell PowerVault MD3260/MD3060e Storage Arrays
4. Dell HPC NFS Storage Solution High Availability (NSS5.5-HA) Configuration with Dell PowerVault MD3460 and MD3060e Storage Arrays, Version 1.0
5. Dell HPC NFS Storage Solution High Availability (NSS6.0-HA) Configuration with Dell PowerEdge 13th Generation Servers, Version 1.0
6. Dell HPC NFS Storage Solution High Availability (NSS7.0-HA) Configuration
7. Knowledge Base article 28603: Fence device and Agent Information for Red Hat Enterprise Linux
8. Red Hat Enterprise Linux 7 High Availability Add-On Reference
9. Support for Dell EMC PowerVault ME4084 Manuals & Documents
10. Support for PowerVault MD3460 Manuals & Documents
11. Support for PowerVault MD3060e Manuals & Documents