Dell HPC NFS Storage Solution High Availability (NSS-HA) Configurations with Dell PowerEdge 12th Generation Servers

Xin Chen, Garima Kochhar

and Mario Gallegos

Dell HPC Engineering

July 2012| Version 1.0

July 2012| Rev 1.0

Contents

Executive summary 5
1. Introduction 6
2. NSS-HA solution review 6
2.1. A brief introduction to NSS-HA solutions 6
2.2. NSS-HA offerings from Dell 8
2.3. General comparisons among NSS-HA solutions 10
3. Dell PowerEdge 12th generation servers in NSS4-HA 11
3.1. Dell PowerEdge R620 vs. Dell PowerEdge R710 11
3.2. PCI-E slots recommendations in the Dell PowerEdge R620 12
3.3. Comprehensive comparisons between NSS4-HA and NSS3-HA releases 13
4. Evaluation 16
4.1. Method 16
4.2. Test bed 16
4.3. HA functionality 19
4.3.1. Potential failures and fault tolerant mechanisms in NSS-HA 19
4.3.2. HA tests for NSS-HA 20
5. NSS-HA Performance with Dell PowerEdge 12th generation servers 22
5.1. IPoIB sequential writes and reads 23
5.2. Random writes and reads 24
5.3. Metadata tests 25
6. Conclusion 27
7. References 27
Appendix A: Benchmarks and test tools 28
A.1. IOzone 28
A.2. mdtest 30
A.3. Checkstream 31
A.4. The dd Linux utility 32

Tables

Table 1. NSS-HA Solutions[2, 3] 9
Table 2. The comparisons among NSS-HA Solutions[2,3] 10
Table 3. Dell PowerEdge R620 vs. Dell PowerEdge R710 12
Table 4. Storage components in NSS-HA 14

Table 5. Server components in NSS-HA 15
Table 6. NSS4-HA hardware configuration details 17
Table 7. NSS4-HA software configuration details 18
Table 8. NSS4-HA firmware and driver configuration details 18
Table 9. NSS4-HA client configuration details 19
Table 10. NSS-HA mechanisms to handle failures 20

Figures

Figure 1. The infrastructure of the NSS-HA solution 7
Figure 2. A failure scenario in NSS-HA 8
Figure 3. PCI-E slots usage in the Dell PowerEdge R620: option 1 vs. option 2 13
Figure 4. NSS4-HA test bed 17
Figure 5. IPoIB large sequential write performance: NSS4-HA vs. NSS3-HA 23
Figure 6. IPoIB large sequential read performance: NSS4-HA vs. NSS3-HA 24
Figure 7. IPoIB random write performance: NSS4-HA vs. NSS3-HA 25
Figure 8. IPoIB random read performance: NSS4-HA vs. NSS3-HA 25
Figure 9. IPoIB file create performance: NSS4-HA vs. NSS3-HA 26
Figure 10. IPoIB file stat performance: NSS4-HA vs. NSS3-HA 26
Figure 11. IPoIB file remove performance: NSS4-HA vs. NSS3-HA 27

Executive summary

This solution guide describes the Dell NFS Storage Solution High Availability configuration (NSS-HA) with Dell PowerEdge 12th generation servers. It presents a comparison between all available NSS-HA offerings so far, and provides tuning best practices and performance details for a configuration with a storage system capacity of 288TB and two Dell PowerEdge 12th generation servers (R620).

The NSS-HA solution described is designed for high availability using a pair of NFS servers in active-passive configuration to provide storage service to the HPC compute cluster. As in previous versions of Dell NSS-HA solution guides, the goal is to improve service availability and maintain data integrity in the presence of possible failures or faults, and to maximize performance in the failure-free case.

1. Introduction

This solution guide provides information on the latest Dell NFS Storage Solution High Availability configuration (NSS-HA) with Dell PowerEdge 12th generation servers. The solution uses the NFS file system along with the Red Hat Scalable File system (XFS) and Dell PowerVault storage to provide an easy to manage, reliable, and cost effective storage solution for HPC clusters. With this offering, NSS-HA configuration leverages with the most advanced Dell servers (Dell PowerEdge 12th generation servers) to deliver better performance than in previous NSS-HA solutions.

The philosophy and design principles for this release remain the same as previous Dell NSS-HA configurations. This version of the solution guide describes the deltas in configuration and performance. For complete details, review this document along with the previous version titled "Dell HPC NFS Storage Solution High Availability Configurations, Version 1.1" and "Dell HPC NFS Storage Solution – High availability with large capacities, Version 2.1". The main change in this version of the solution is the change from Dell PowerEdge 11th generation servers to the Dell PowerEdge 12th generation servers.

The following sections describe the technical details, evaluation method, and the expected performance of the solution.

2. NSS-HA solution review

Along with the current solution, three versions of NSS-HA solutions have been released since 2011. This section provides a brief description of the NSS-HA solution, lists the available Dell NSS-HA offerings, and gives general comparisons among those offerings.

2.1. A brief introduction to NSS-HA solutions

The design of the NSS-HA solution for each release is almost identical. This section only provides a brief description of NSS-HA solution. For more detailed information, refer to the two white papers "Dell HPC NFS Storage Solution High Availability Configurations, Version 1.1" and "Dell HPC NFS Storage Solution – High availability with large capacities, Version 2.1". If you are already familiar with the NSS-HA architecture, skip this section.

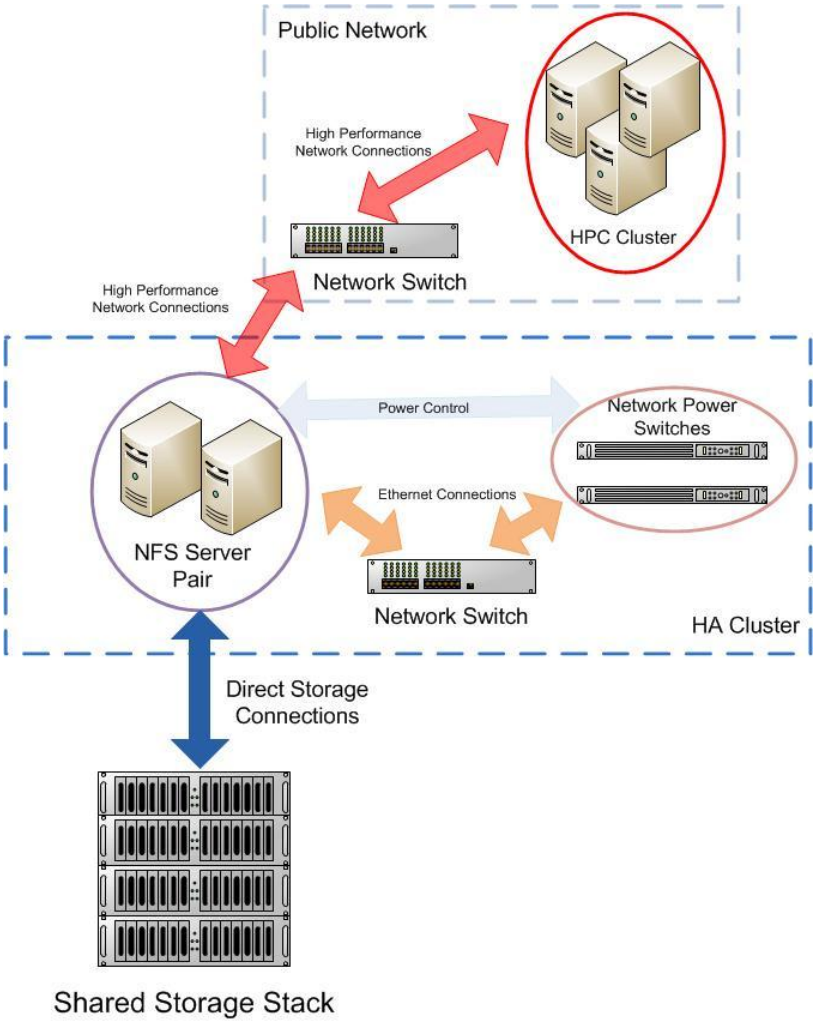
The core of the solution is a high availability (HA) cluster(1), which provides a highly reliable and available storage service to the HPC compute cluster using a high performance network connection such as InfiniBand (IB) or 10 Gigabit Ethernet (10GbE). The HA cluster has shared access to disk-based Dell PowerVault storage in a variety of capacities. Figure 1 depicts the general infrastructure of the NSS-HA solution.

The HA cluster consists of several components as listed below:

- High Availability nodes - These are servers configured with the Red Hat Enterprise Linux high availability cluster software stack. In the NSS-HA solution, two systems are deployed as a pair of NFS severs; they are configured in an active/passive mode, and have direct access to the shared storage stack.
- Network switch for the HA cluster (or the private network) – The private network is used for communication between the HA cluster nodes and other cluster hardware, such as network power switches and the fence devices, which are installed in the cluster nodes.

- Fence devices – Fence devices are required for fencing (rebooting) the failed or misbehaving cluster node in the HA cluster. In the NSS-HA solution, two types of fence devices are configured: Switched Power Distribution Units (PDU) and the Dell server management controller, the iDRAC.

Figure 1. The infrastructure of the NSS-HA solution



A major goal of the NSS-HA solution is to improve storage service availability in the presence of possible failures or faults. This goal is achieved by a *failover* process implemented by Red Hat Enterprise High Availability Cluster software stack. The failover process is divided into three stages: failure detection, fencing, and service failover.

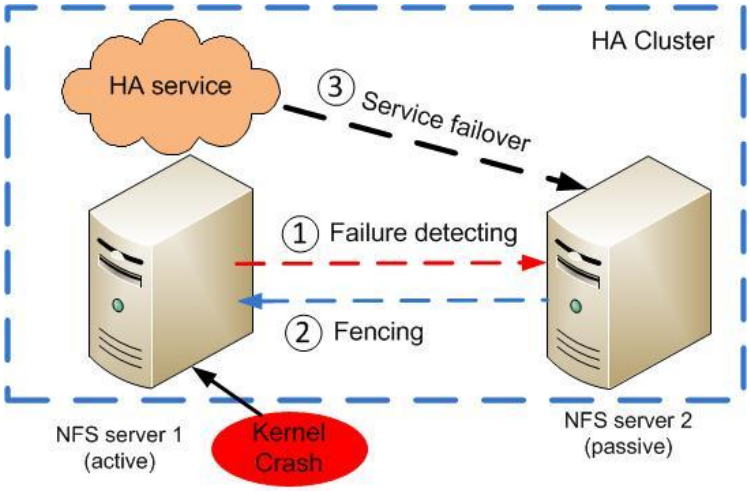
Figure 2 shows a typical scenario of how storage service availability is guaranteed in the NSS-HA solution. In this scenario, a kernel crash occurs on an NFS server (the active one), which is the NFS gateway for the HA cluster.

1) Failure detection – Resources related to the storage service, such as file system, service IP address, and so on, are defined, configured and monitored for health by the HA cluster. Any interruption in access to the storage is detected. In this case, once a kernel crash occurs at NFS server 1 (the active one), a message in terms of loss of heartbeat signal is passed to the NFS server 2, and server 2 recognizes that the server 1 has failed.

2) Fencing – In the HA cluster, once a node notices that the other node has failed, it fences (reboot) the failed node using a fence device. This is to make sure that only one server accesses the data at any point to protect data integrity. In NSS-HA, a node can fence the other using the Dell iDRAC or an APC PDU. The fence devices and corresponding fence commands are configured as part of the HA cluster configuration process. In this case, NFS server 2 fences NFS server 1.

3) Service failover – In the HA cluster, only after a node successfully fences the other can the service failover process start. Failover means that the HA service running previously on the failed server is transferred to the healthy one. In this case, once NFS server 2 has successfully fenced server 1, the HA service is transferred to and started on NFS server 2.

Figure 2. A failure scenario in NSS-HA



From the perspective of the HA cluster, degradation in performance occurs during the actual HA failover process. But the failover is transparent to the cluster as far as possible and user applications continue to function and access data as before.

The HA service is defined and configured in the cluster configuration process. In the NSS-HA, NFS export, the service IP on which the compute nodes access the NFS server, and logical volume manager (LVM) are configured as a HA service.

2.2. NSS-HA offerings from Dell

The current Dell NSS-HA solution continues the evolutionary growth of the Dell NFS Storage Solution HA family, the NSS-HA. The first NSS-HA solution for HPC from Dell used the Dell PowerEdge R710 servers, the Dell PowerVault MD3200 RAID array and the Dell PowerVault MD1200 expansion chassis. At the time, 2-TB drives provided the best value in terms of capacity for each dollar. The second version of NSS-HA used the same Dell PowerEdge R710 servers, but integrated 3-TB hard drives for improved capacity and broke the 100-TB supported capacity limit for Red Hat-based scalable file system.

With the introduction of the third version of NSS-HA (described in this document), Dell is upgrading the servers to a smaller form factor Dell PowerEdge R620 server, which features the Intel E5-2600 processors (codenamed Sandy Bridge EP). The 1U Dell PowerEdge R620 server, with the integrated PCIe Gen-3 I/O capabilities of the Intel E5-2600 processor, allows for a faster interconnect using the

56Gb fourteen data rate (FDR) as well as increased memory capacity and bandwidth. The storage subsystem hardware remains unchanged.

Table 1 lists all the Dell NSS-HA solutions with standard configurations. In addition to the standard configurations in Table 1, a special NSS-HA configuration, *XL configuration* is also available. The NSS-HA-XL provides two storage system services running concurrently, along with the high availability functionality, but instead of one active-passive failover pair, the solution is designed as two active-passive failover pairs that host the two storage services. The NSS-HA-XL solution uses the same hardware and software as required in the standard configurations but with twice the number of PowerVault storage enclosures compared to the NSS-HA 288TB standard configuration. For more information about this special configuration, refer to the blog post titled "Dell NFS Storage Solution with High Availability – XL configuration".

Table 1. NSS-HA Solutions[2], [3]

	NSS2-HA Release (April 2011)	**NSS3-HA Release (February 2012) "Large capacity configuration"**	**NSS4-HA Release (July 2012) "PowerEdge 12G based solution"**
Storage Capacity	48 TB and 96 TB of raw storage space	144 TB and 288 TB of raw storage space	
Network Connectivity	QDR InfiniBand or 10Gb Ethernet connectivity.		FDR InfiniBand or 10Gb Ethernet Connectivity
NFS servers	Dell PowerEdge R710 servers		Dell PowerEdge R620 servers
Software	Red Hat Enterprise Linux 5.5 Red Hat Scalable File system (XFS) v2.10.2-7	Red Hat Enterprise Linux 6.1 Red Hat Scalable File system (XFS) v3.1.1-4	
Storage Devices	Dell PowerVault MD3200 and MD1200s 2 TB NL SAS drives	Dell PowerVault MD3200 and MD1200s 3 TB NL SAS drives	
Switch and Power Distribution Units (PDUs)	PowerConnect 5424 Two APC switched PDUs to manage high availability. Please refer to "Fence device and Agent Information for Red Hat Enterprise Linux" for the supported model of APC PDUs.		
Support and Services	3 years of Dell PRO Support for IT and Mission Critical 4HR 7x24 onsite pack. Dell deployment service to speed up installation, optimize performance and integrate with your environment.		

Note: 1. There is no NSS1-HA release available on market. 2. Contact a Dell Sales Representative to discuss which offering works best in your environment, and the ordering process. You can order any of the pre-configured solutions or a customized solution to fit specific needs. Based on the customization selected, some of the best practices discussed in this document may not apply.

2.3. General comparisons among NSS-HA solutions

Table 2 gives general comparisons among the three releases of NSS-HA solutions, focusing on four aspects: storage capacity, performance, configuration, and HA functionalities. It is worth pointing out that there are significant changes in configuration steps between the first and second release, while there is a few configuration step changes between the second and third release. Furthermore, with the help of Dell PowerEdge 12th generation servers and FDR InfiniBand network connection, the NSS-HA solution now achieves sequential read peak performance up to 4508 MB/sec.

Table 2. The comparisons among NSS-HA Solutions[2],[3]

	NSS2-HA Release (April 2011)	**NSS3-HA Release (February 2012) "Large capacity configuration"**	**NSS4-HA Release (July 2012) "PowerEdge 12G based solution"**
Release Purpose	Initial Release	Add the ability to support greater than 100TB storage capacity.	Move to latest server technology. Take advantage of the performance improvement with Dell PowerEdge 12th generation servers
Storage Capacity	The maximum supported size is 96 TB in a standard configuration The XL configuration supports 2x96TB.	The maximum supported size is 288 TB in a standard configuration. The XL configuration supports 2x288TB (two file systems).	
Sequential Performance (standard configuration)	Peak write: 1275 MB/sec Peak read: 2430 MB/sec	Peak write: 1495 MB/sec Peak read: 2127 MB/sec	Peak write: 1535 MB/sec Peak read: 4058 MB/sec
Configuration Details	The complete configuration steps can be found at Dell HPC NFS Storage Solution High Availability Configurations, Version 1.1	Compared to the 1st release, there are significant changes in NSS-HA configuration steps. The complete configuration steps can be found at Dell HPC NFS Storage Solution – High availability with large capacities, Version 2.1	Compared to the 2nd release, there are few changes in NSS-HA configuration steps. The complete configuration steps will be published in July 2012 with an update to this white paper.
HA Functionalities	All three releases use the same mechanisms to tolerate or recover the following failures: • Single local disk failure on a server • Single server failure • Power supply or power bus failure • Fence device failure • SAS cable/port failure • Dual SAS cable/card failure • InfiniBand /10GbE link failure • Private switch failure • Heartbeat network interface failure • RAID controller failure on Dell PowerVault MD3200 storage array		

Note: For detailed technical comparisons, refer to Table 4 and Table 5 in section title, *Comprehensive comparisons between NSS4-HA and NSS3-HA releases*.

3. Dell PowerEdge 12th generation servers in NSS4-HA

As compared to the NSS2-HA solution releases, the biggest change in the NSS4-HA release is that the new Dell PowerEdge R620 12th generation server is deployed as an NFS server, while the Dell PowerEdge R710 11th generation server was used as the NFS server in the previous releases.

The Dell PowerEdge R620 server is designed to meet the very high demands of high-performance computing; it maximizes computing power without compromise in space-sensitive environments. This section highlights the technology advantages of Dell PowerEdge R620 server and demonstrates that the Dell PowerEdge R620 is the best-fit server for NSS-HA solution so far. This section also discusses the PCI-E slot configuration for the Dell PowerEdge R620, which may impact the overall system performance. Finally, a detailed technical comparison among all three NSS-HA solution releases is provided.

3.1. Dell PowerEdge R620 vs. Dell PowerEdge R710

Dell NSS-HA solution is designed to provide storage service to HPC clusters. Besides the goal of high availability and reliability, it is also very important and necessary for a storage-based solution to deliver a good I/O performance for HPC clusters.

In NSS-HA solution, NFS servers are directly connected to a shared storage stack, and receive/send I/O requests from an HPC cluster using high speed network connections, such as InfiniBand and 10GbE Ethernet. To process high volume I/O data in a short period, an NFS server must have balanced network and disk I/O processing capabilities.

The Dell PowerEdge R620 leverages the current state-of-art technologies to enhance the existing network and disk I/O processing capabilities to a new degree, as compared to Dell PowerEdge R710. The key features of Dell PowerEdge R620 are listed below, which make the Dell PowerEdge R620 a better fit NFS server in NSS-HA than the Dell PowerEdge R710:

- Faster processor: The Dell PowerEdge R620 is equipped with the new Intel Xeon E5-2680 processor, which provides faster processing speed and more cores than the Intel Xeon E5630 used in Dell PowerEdge R710.
- Bigger and faster memory: in the Dell PowerEdge R620, the maximum memory size can be 768 GB, and the memory frequency increased to 1600MHz. With this release of the NSS-HA, the NFS server is equipped with 128GB of memory running at 1600MT/s versus 96GB of 1333MT/s memory in the previous solution. Larger memory size and high frequency are critical to server performance.
- Fast internal connection: faster connections are provided throughout the system with 8.0 GT/s with Intel Quick Path Interconnect (QPI) compared to 5.86GT/s supported with the Intel Xeon E5630 in the Dell PowerEdge R710.
- Faster InfiniBand link support: In R620, a PCI-E Gen 3 based fourteen data rate (FDR) card is supported, which can provide up to 56-Gb/sec bandwidth, while the Dell PowerEdge R710 can only support PCI-E Gen 2 speeds and uses quad data rate (QDR). QDR links have a maximum bandwidth of 40 Gb/sec.
- Smaller form factor: The Dell PowerEdge R620 is a 1U rack server, while Dell PowerEdge R710 is a 2U rack server, which results in a denser solution with this release of the NSS-HA solution.

- The Dell PowerEdge R620 can support an onboard 10Gigabit Ethernet network daughter card for clusters that require 10GbE connectivity, which frees up a PCI-E slot in the NFS server.

Table 3 gives a detailed comparison between the Dell PowerEdge R620 and the Dell PowerEdge R710 used in NSS-HA solutions.

Table 3. Dell PowerEdge R620 vs. Dell PowerEdge R710

	Dell PowerEdge R620	Dell PowerEdge R710
Processor	Intel Xeon processor E5-2680 @2.70GHz	Intel Xeon processor E5630 @2.53GHz
Form factor	1U rack	2U rack
Memory	Recommend: 128GB 16 x 8GB DDR3 1600MHz	Recommend: 96GB 12 x 8GB DDR3 1333MHz
Slots	3 PCI-E Gen 3 slots: Two x16 slots with x16 bandwidth, half-height, half-length One x16 slot with x8 bandwidth, half-height, half-length	4 PCI-E Gen 2 slots+ 1 storage slot Two x8 slots Two x4 slots One x4 storage slot
Drive Bays	Up to ten 2.5'' hot-plug SAS, SATA, or SSD	8 x 2.5'' hard drive option
Internal RAID controller	PERC H710P	PERC H700
InfiniBand support	QDR/FDR Links	QDR links

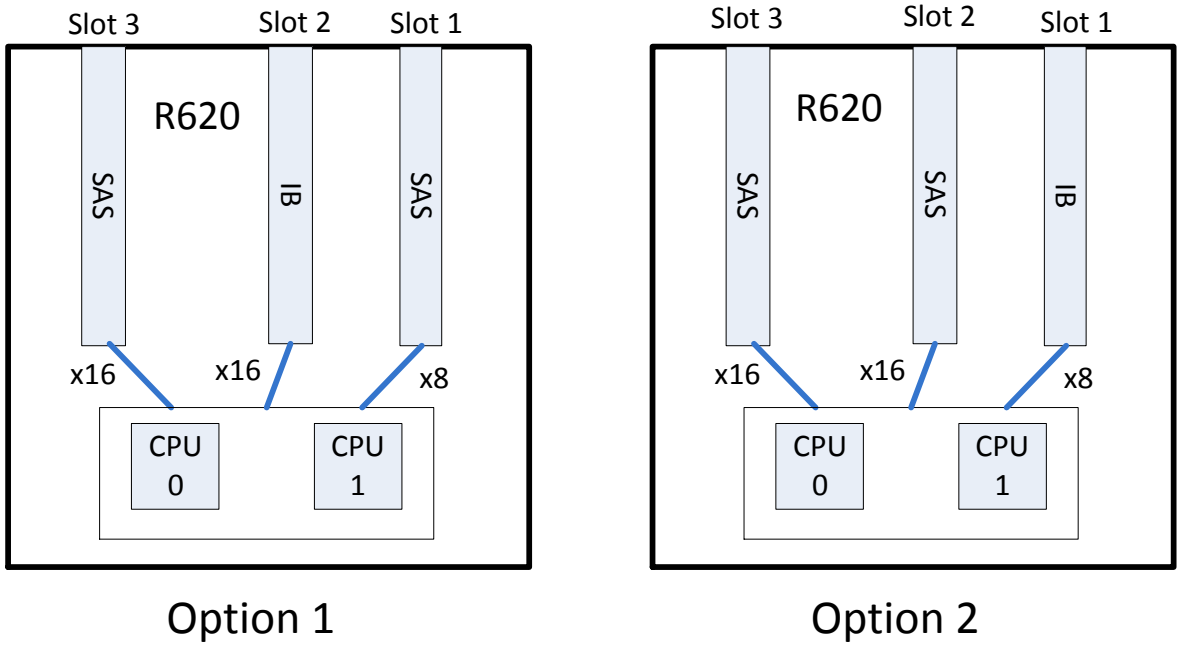
3.2. PCI-E slots recommendations in the Dell PowerEdge R620

In NSS-HA IP over InfiniBand (IPoIB) based solutions, two SAS HBA cards and one InfiniBand HCA card are required to directly connect to a shared storage stack and the InfiniBand switch, respectively. However, given the PCI-E slot design (two x16 slots with x16 bandwidth, one x16 slot with x8 bandwidth) in the Dell PowerEdge R620 poses a question of how to distribute SAS HBA cards and InfiniBand HCA cards in a Dell PowerEdge R620 to achieve the best overall system performance.

There are two options as shown in Figure 3:

- Option 1: InfiniBand HCA card and one SAS HBA card occupy slot 2 and slot 3 with x16 bandwidth; the other SAS HBA card is installed in the slot 1 with x8 bandwidth.
- Option 2: Two SAS HBA cards are installed in the slot 2 and slot 3 with x16 bandwidth, while the InfiniBand HCA card occupies the slot 1 with x8 bandwidth.

Figure 3. PCI-E slots usage in the Dell PowerEdge R620: option 1 vs. option 2



In NSS3-HA, both SAS HBA cards and the InfiniBand HCA card only work with PCI-E x8; thus, it seems that there should be no difference between the two options, as the PCI-E bandwidth provided in Dell PowerEdge R620 can satisfy the bandwidth requirement for both SAS cards and InfiniBand card. However, as the I/O controller is integrated with the processor in the Intel Sandy Bridge microarchitecture [4]; it implies that a processor will get privileged access to some I/O slots due to the shorter physical distance, which is very similar to the well-known NUMA (Non-Uniform Memory Access) effect. For the other processor, the access to I/O slots requires an extra hop using the QPI link between the sockets. Given the non-uniform IO access to the different PCI slots, it is hard to decide if the two options will have similar performance as the two options present different communication paths between processors and I/O devices.

Based on more than twenty performance benchmarking tests in the Dell HPC lab, option 1 and option 2 have very similar write performance, but option 1 can deliver up to 4GB/sec sequential read performance, while the maximum read performance of option 2 is only close to 3.3GB/sec. Given the higher read performance with option 1, option 1 is recommended in the current NSS-HA solution.

NOTE: Given the affinities between processes and data, some level of performance variations may be introduced by both options, as a processor may not always fetch data from nearby I/O slots.

For NSS-HA with 10GbE, option 1 is the recommended configuration. In this case, the two SAS cards should be installed as described in Option 1 above. No InfiniBand card is needed, and an onboard 10GbE Network Daughter Card (NDC) provides the network to the client.

3.3. Comprehensive comparisons between NSS4-HA and NSS3-HA releases

The previous sections introduced new features in the current NSS-HA release, which are mainly introduced by the Dell PowerEdge R620 server. This section provides a comprehensive examination

between the current and previous NSS-HA releases to identify their major similarities and differences using the following two tables:

- Table 4 provides the information about storage subsystem.
- Table 5 lists the major similarities and differences in the NFS servers.

Table 4. Storage components in NSS-HA

Storage components	NSS3-HA release (February 2012) "Large capacity configuration"	NSS4-HA release (July 2012) "PowerEdge 12th Generation based solution"
Storage array	Dell PowerVault MD3200 and MD1200s 144TB solution: 4 arrays (1 Dell PowerVault MD3200 + 3 Dell PowerVault MD1200s): 48 drives (12 disks/array x 4 arrays x 3TB/disk), 109TB usable space. 288TB solution: 8 arrays (1 Dell PowerVault MD3200 + 7 Dell PowerVault MD1200s): 96 drives (12 disks/array x 8 arrays x 3TB/disk), 218TB usable space.	
Dell PowerVault MD3200 firmware	Latest version to support 3TB drives. New Dell PowerVault MD3200 firmware needed for 3TB drive support.	
Disks in storage array	3TB NL SAS	
Virtual disk configuration	RAID 6 10+2, segment size 512K Write cache mirroring enabled Read cache enabled Dynamic cache read prefetch enabled	
Storage enclosure cabling	The asymmetric ring cabling scheme	
Capacities supported	Up to 288TB for standard configurations. Up to 2x288TB for XL configurations Increased capacities, denser solution.	
LVM configuration for HA	Clustered LVM. Clustered LVM is supported in RHEL 6.1 and is the recommended method.	

Note: NSS-HA XL configuration is also available for each NSS-HA release; contact your Dell sales representative for detailed information.

Table 5. Server components in NSS-HA

Server components	NSS3-HA release (February 2012) "Large capacity configuration"	NSS4-HA release (July 2012) "PowerEdge 12th generation server based solution"
NFS server	Dell PowerEdge R710	Dell PowerEdge R620
Hardware components		
Memory per NFS server	96 GB, 12 x 8 GB 1333MHz RDIMMs	128 GB, 16 x 8GB 1600MHz RDIMMs
SAS cards per server	1 SAS 6Gbps HBA	2 SAS 6Gbps HBAs
RAID Disk controller	PERC H700 with five 146GB 10K SAS hard drives: Two disks in RAID1 with 1 hot spare for OS; two disks in RAID0 for swap space.	PERC H710P with five 300GB 15K SAS hard drives: Two disks in RAID1 with 1 hot spare for OS; two disks in RAID0 for swap space.
InfiniBand HCA	Mellanox ConnectX-2 QDR PCI-E card	Mellanox ConnectX-3 FDR PCI-E card
10 Gigabit Ethernet card	Intel X520 DA 10Gb Dual Port SFP+ Advanced	Intel Ethernet X540 DP 10Gb BT + I350 1Gb BT DP, one custom rack network daughter card
Systems Management device	iDRAC6 Enterprise	iDRAC7 Enterprise
Software components		
Operating System	RHEL 5.5 x86_64	RHEL 6.1 x86_64
Kernel version	2.6.18-194.el5	2.6.32-131.0.15.el6
File system (XFS) version	2.10.2-7	3.1.1-4
InfiniBand driver	Mellanox OFED 1.5.1	Mellanox OFED 1.5.3-3.0.0
HA cluster suite	Red Hat Cluster Suite from RHEL 5.5	Red Hat Cluster Suite from RHEL 6.1
Systems Management	Dell OpenManage Server Administrator 6.5.0	Dell OpenManage Server Administrator 7.0.0

Note: Details of the complete test bed are provided in the section called "Test bed."

4. Evaluation

The architecture proposed in this white paper was evaluated in the Dell HPC lab. This section describes the test methodology and the test bed used for verification. It also contains details on the functionality tests.

4.1. Method

The NFS Storage Solution described in this solution guide was tested for HA functionality and performance. A 288TB NSS4-HA configuration was used to test the HA functionality of the solution. Different types of failures were introduced and the fault tolerance and robustness of the solution was measured. "HA functionality" describes these functionality tests and their results. Functionality testing was similar to the work done in the previous versions of the solution [3].

4.2. Test bed

The test bed used to evaluate the NSS4-HA functionality and performance is shown in Figure 4.

- A 64 node HPC compute cluster was used to provide I/O traffic for the test bed.
- A pair of Dell PowerEdge R620 servers are configured as an active-passive HA pair and function as an NFS gateway for the HPC compute cluster (also called the clients).
- Both NFS servers are connected shared Dell PowerVault MD3200 SAS storage extended with Dell PowerVault MD1200 arrays (Figure 4 shows a 288TB solution with eight MD storage arrays) at the backend. The user data resides on an XFS file system created on this storage. The XFS file system is exported using NFS to the clients.
- The NFS servers are connected to the clients using the public network. This network is either InfiniBand or 10 Gigabit Ethernet.
- For the HA functionality of the NFS servers, a private Gigabit Ethernet network is configured to monitor server health and heartbeat, and to provide a route for the fencing operations using a PowerConnect 5424 Gigabit Ethernet switch.
- Power to the NFS servers is driven by two APC switched PDUs on two separate power buses.

Complete configuration details are provided in Table 6, 0, and Table 8.

Figure 4. NSS4-HA test bed

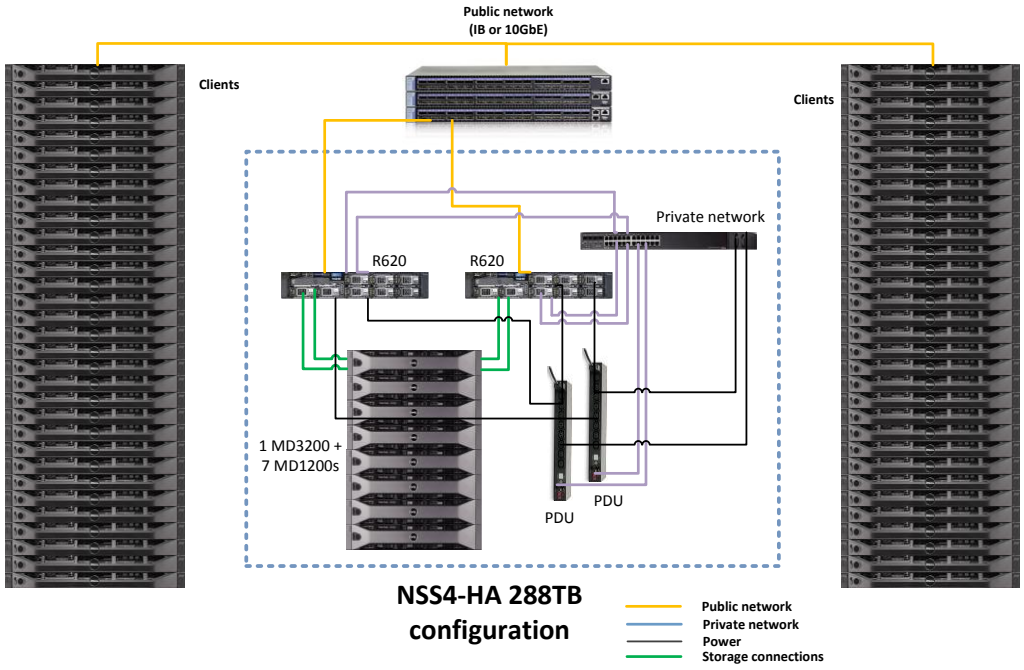


Table 6. NSS4-HA hardware configuration details

Server configuration	
NFS server model	Two Dell PowerEdge R620
Processor	Dual Intel Xeon E5-2680 @ 2.70GHz
Memory	8 * 8GB 1600MHz RDIMMs (The test bed used 64GB; the recommendation for production clusters is to use 128GB).
Local disks and RAID controller	PERC H710P with five 300GB 15K SAS hard drives
Optional InfiniBand HCA (slot 2)	Mellanox ConnectX-3 FDR PCI-E card
Optional 10 Gigabit Ethernet card (Onboard)	Intel Ethernet X540 DP 10Gb BT + I350 1Gb BT DP network daughter card
External storage controller (slot 3 and slot 1)	Two 6Gbps SAS HBA
Systems Management	iDRAC Enterprise
Power Supply	Dual PSUs

Storage configuration	
Storage Enclosure	One Dell PowerVault MD3200 array. Seven Dell MD1200 expansion arrays for the 288TB solution. High Performance Tier feature enabled on the Dell PowerVault MD3200
RAID controllers	Duplex RAID controllers in the Dell MD3200
Hard Disk Drives	Twelve 3TB 7200 rpm NL SAS drivers per array
Other components	
Private Gigabit Ethernet switch	Dell PowerConnect 5424
Power Distribution Unit	Two APC switched Rack PDUs, model AP7921

Table 7. NSS4-HA software configuration details

Software	
Operating system	Red Hat Enterprise Linux (RHEL) 6.1 x86_64
Kernel version	2.6.32-131.0.15.el6.x86_64
Cluster Suite	Red Hat Cluster Suite from RHEL 6.1
File system	Red Hat Scalable File System (XFS) 3.1.1-4
Systems management	Dell OpenManage Server Administrator 7.0.0
Storage management	Dell Modular Disk Storage Manager 3.0.0.18

Table 8. NSS4-HA firmware and driver configuration details

Firmware and Drivers	
Dell PowerEdge R620 BIOS	1.1.2
Dell PowerEdge R620 iDRAC	1.06.06
InfiniBand firmware	2.10.700
InfiniBand driver	Mellanox OFED 1.5.3-3.0.0

Firmware and Drivers	
10 Gigabit Ethernet driver	ixgbe 3.6.7-NAPI
PERC H710P firmware	21.0.2-0001
PERC H710P driver	megaraid_sas 00.00.05.34-rc1
6Gbps SAS firmware	07.03.05.00
6Gbps SAS driver	mpt2sas 08.101.00.00

Table 9. NSS4-HA client configuration details

Client / HPC Compute Cluster	
Clients	64 PowerEdge R410 compute nodes Red Hat Enterprise Linux 6.1 x86-64
InfiniBand	Mellanox ConnectX-2 QDR HCA Mellanox OFED 1.5.3-3.0.0
InfiniBand fabric	All clients connected to a single large port count InfiniBand switch (Mellanox IS5100). Both R620 NSS-HA servers also connected to the InfiniBand switch.
Ethernet	Onboard 1 GbE Broadcom 5716 network adapter. bnx2 driver v2.1.6.
Ethernet fabric	Two sets of 32 compute nodes connected to two Dell PowerConnect 6248 Gigabit Ethernet switches. Both Dell PowerConnect 6248 switches have four 10GbE links each to a 10GbE Dell PowerConnect 8024 switch. Both R620 NSS-HA servers connected directly to the Dell PowerConnect 8024 switch. Flow control was disabled on the Dell PowerConnect 8024 switch and two Dell PowerConnect 6248 switches.

4.3. HA functionality

The HA functionality of the solution was tested by simulating several component failures. The design of the tests and the test results are similar to previous versions of the solution since the broad architecture of the solution has not changed with this release. The section reviews the failures and fault tolerant mechanisms in NSS-HA solutions, then presents the HA functionality tests with regarding to different potential failures and faults.

4.3.1. Potential failures and fault tolerant mechanisms in NSS-HA

In the real world, there are many different types of failures and faults that can impact the functionality of NSS-HA. Table 10 lists the potential failures that are tolerated in NSS-HA solutions.

Note: The analysis below assumes that the HA cluster service is running on the *active* server; the *passive* server is the other component of the cluster.

Table 10. NSS-HA mechanisms to handle failures

Failure type	Mechanism to handle failure
Single local disk failure on a server	Operating system installed on a two-disk RAID 1 device with one hot spare. Single disk failure is unlikely to bring down server.
Single server failure	Monitored by the cluster service. Service fails over to passive server.
Power supply or power bus failure	Dual power supplies in each server. Each power supply connected to a separate power bus. Server continues functioning with a single power supply.
Fence device failure	iDRAC used as primary fence device. Switched PDUs used as secondary fence devices.
SAS cable/port failure	Two SAS cards in each NFS server. Each card has a SAS cable to storage. A single SAS card/cable failure will not impact data availability.
Dual SAS cable/card failure	Monitored by the cluster service. If all data paths to the storage are lost, service fails over to the passive server.
InfiniBand /10GbE link failure	Monitored by the cluster service. Service fails over to 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 required.
Heartbeat network interface failure	Monitored by the cluster service. Service fails over to passive server.
RAID controller failure on Dell PowerVault MD3200 storage array	Dual controllers in the Dell PowerVault MD3200. The second controller handles all data requests. Performance may be degraded but functionality is not impacted.

4.3.2. HA tests for NSS-HA

Functionality was verified for an NFSv3 based solution. The following failures were simulated on the cluster with the consideration of the failures and faults listed Table 10.

- Server failure
- Heartbeat link failure
- Public link failure

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

The NSS-HA behaviors are outlined below in response to these failures.

Server response to a failure

The server response to a failure event within the HA cluster was recorded. Time to recover from a failure was used as a performance metric. Time was measured from the point when the fault was injected in the server running the HA service (active) until the service was migrated and running on the other server (passive).

- Server failure - simulated by introducing a kernel panic.
  When the active server fails, the heartbeat between the two servers is interrupted. The passive server waits for a defined period of time and then attempts to fence the active server. Once fencing is successful, the passive server takes ownership of the cluster service. Clients cannot access the data until the failover process is complete.

- 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 since 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 InfiniBand or 10 Gigabit Ethernet 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 powering 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 cable from the server.
  If the iDRAC on a server fails, the server is fenced using the network PDUs, which are defined as secondary fence devices in the cluster configuration files.

- One SAS link failure - simulated by disconnecting one SAS link between the Dell PowerEdge R620 server and the Dell PowerVault MD3200 storage.
  In the case where only one SAS link fails, the 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.

  For the above cases, it was observed that the HA service failover takes in the range of a 30 to 60 seconds. This reaction time is faster with this version of the cluster suite than with the previous

version [4]. In a healthy cluster, any failure event should be noted by the Red Hat cluster management daemon and acted upon within minutes. Note that this is the failover time on the NFS servers; the impact to the clients could be longer.

- Multiple SAS link failures - simulated by disconnecting all SAS links between one Dell PowerEdge R620 server and the Dell PowerVault MD3200 storage.
  When all SAS links on the active server fail, the multipath daemon on the active server retries the path to the storage based on the parameters configured in the `multipath.conf` file. This is set to 150 seconds by default. After this process times out, the HA service will attempt to failover to the passive server.

  If the cluster service is unable to cleanly stop the LVM and the file system because of the broken path, a watchdog script reboots the active server after five minutes. At this point the passive server fences the active server, restart the HA service and provide the data path to the clients. This failover can therefore take anywhere in the range of three to eight minutes.

Impact to clients

Clients mount the NFS file system exported by the server using the HA service IP. This IP is associated with either an IPoIB or a 10 Gigabit Ethernet network interface on the NFS server. To measure any impact on the client, the `dd` utility and the `iozone` benchmark were used to read and write large files between the client and the file system. Component failures were introduced on the server while the client was actively reading and writing data from the file system.

In all scenarios, it was observed that the client processes complete 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 is temporarily unavailable, the client process was observed to be in an uninterruptible sleep state.

Depending on the characteristics of the client process, it can be expected to abort or sleep while the NFS share is temporarily unavailable during the failover process. Any data that has already been written to the file system will be available.

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

5. NSS-HA Performance with Dell PowerEdge 12th generation servers

This section presents the results of the performance related tests conducted on the current NSS-HA solution. All performance tests were performed in a failure free scenario to measure the maximum capability of the solution. The tests focused on three types of IO patterns: large sequential reads and writes, small random reads and writes, and three metadata operations (file create, stat, and remove).

A 288TB configuration is benchmarked with IPoIB network connectivity. The 64-node compute cluster described in Test bed was used to generate IO load to the NSS-HA solution. Each test was run over a range of clients to test the scalability of the solution.

The `iozone` and `mdtest` utilities were used in this study. `Iozone` was used for the sequential and random tests. For sequential tests, a request size of 1024KB was used. The total amount of data transferred was 256GB to ensure that the NFS server cache was saturated. Random tests used a 4KB request size and each client read and wrote a 2GB file. Metadata tests were performed using the

`mdtest` benchmark and include file create, stat and remove operations. Refer to Appendix A for the complete command lines used in the tests.

5.1. IPoIB sequential writes and reads

This section compares the random write and read performance between the current NSS4-HA release with a Dell PowerEdge R620 server and the previous NSS3-HA release[3] with a Dell PowerEdge R710 server. The only difference between the tests for the two releases is that the total amount of data for I/O was 256GB to eliminate the cache effect for the current test, while the amount was set to 128GB when testing the previous NSS-HA release (which had only 48 GB installed during the tests). The storage hardware and software versions used for the two solutions is identical; the only update in server hardware is the use of the Dell PowerEdge R620 as a NFS server instead of the Dell PowerEdge R710.

Due to the many powerful features of the Dell PowerEdge R620 listed in Dell PowerEdge R620 vs. Dell PowerEdge R710, the current NSS-HA release gains a huge read performance improvement, about 75 percent increment on average, and the maximum read performance is up to 4058 GB/sec. However, the write performance does not change much between the current and previous release, as the RAID6 write performance is largely determined by the storage system itself (disk drives in the storage subsystem are configured with RAID6).

Figure 5 and Figure 6 show the sequential write and read performance, respectively.

Note: Figure 5 and Figure 6 just give the measured maximum performance numbers for each client case. The actual performance number depends on many factors including but not limited to number of clients, file size, switch used, and firmware and driver versions.

Figure 5. IPoIB large sequential write performance: NSS4-HA vs. NSS3-HA

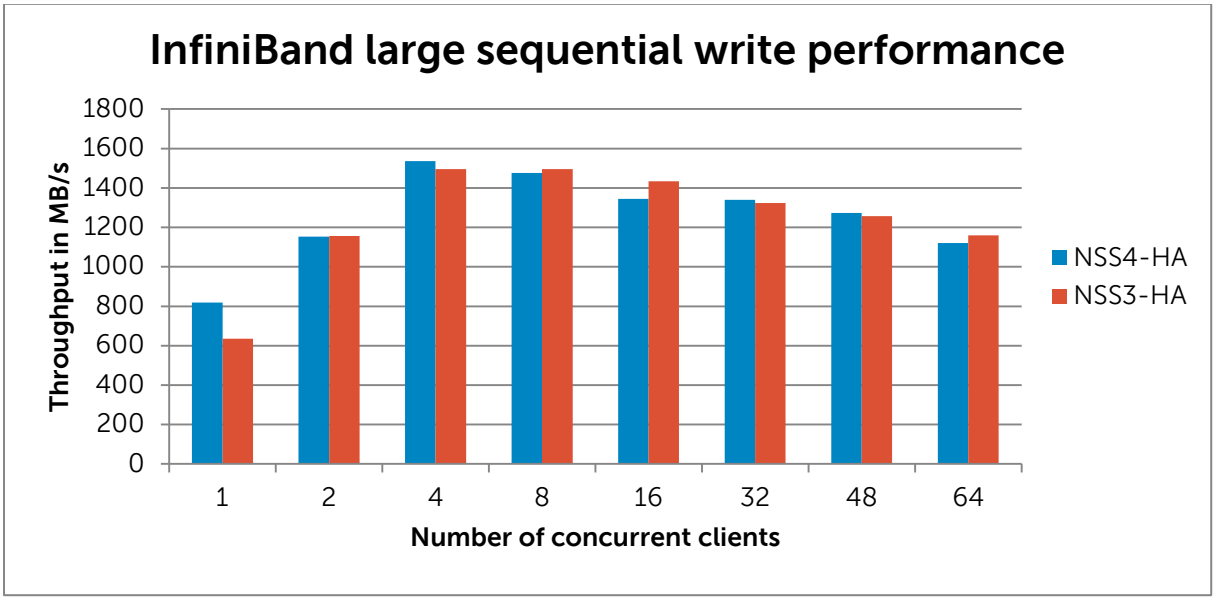
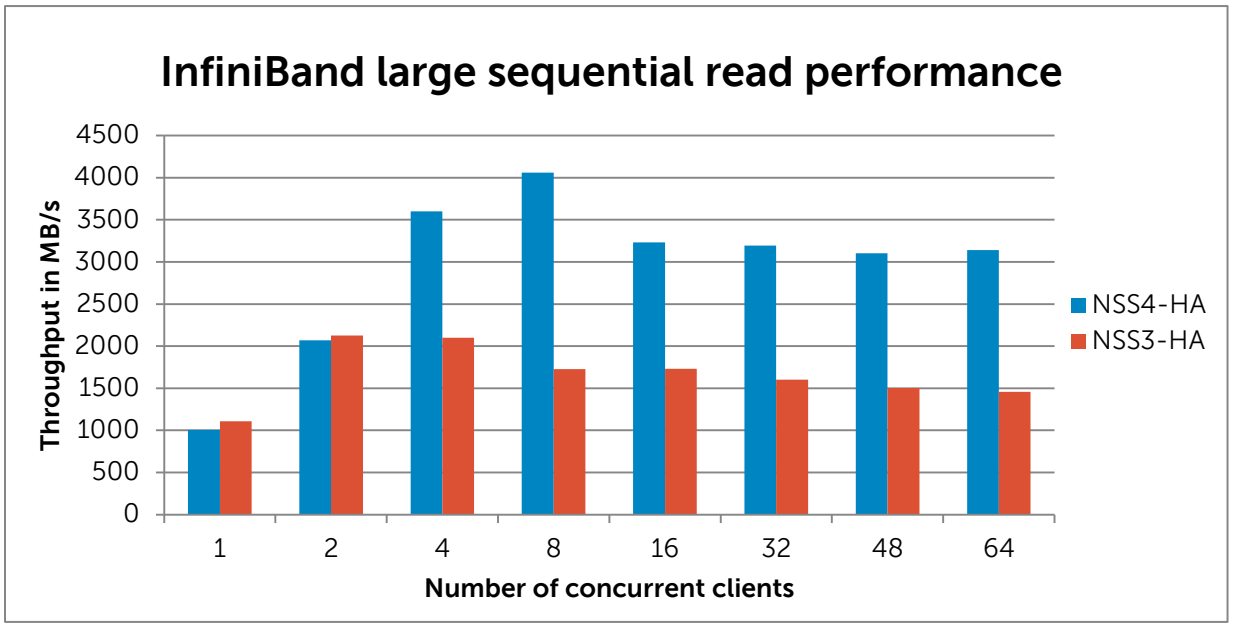


Figure 6. IPoIB large sequential read performance: NSS4-HA vs. NSS3-HA



5.2. Random writes and reads

This section compares the random write and read performance between the current NSS4-HA release with a Dell PowerEdge R620 server and the previous NSS3-HA release [3] with Dell PowerEdge R710 server. The test parameters, such as number of files, number of runs, and so on, are the same for both releases.

As disk seek latency is the major factor to determine random write and read performance, and there is no change in the storage components (hardware and software) between the current and previous release, it was expected that the test results between the two releases should be similar. However, on average, about 17 percent improvement for random writes and 23 percent improvement for random reads were observed during the tests. Such significant improvement is mainly due to the bigger memory (64GB) deployed in the current release.

Figure 7 and Figure 8 show the random write and read performance, respectively.

Figure 7. IPoIB random write performance: NSS4-HA vs. NSS3-HA

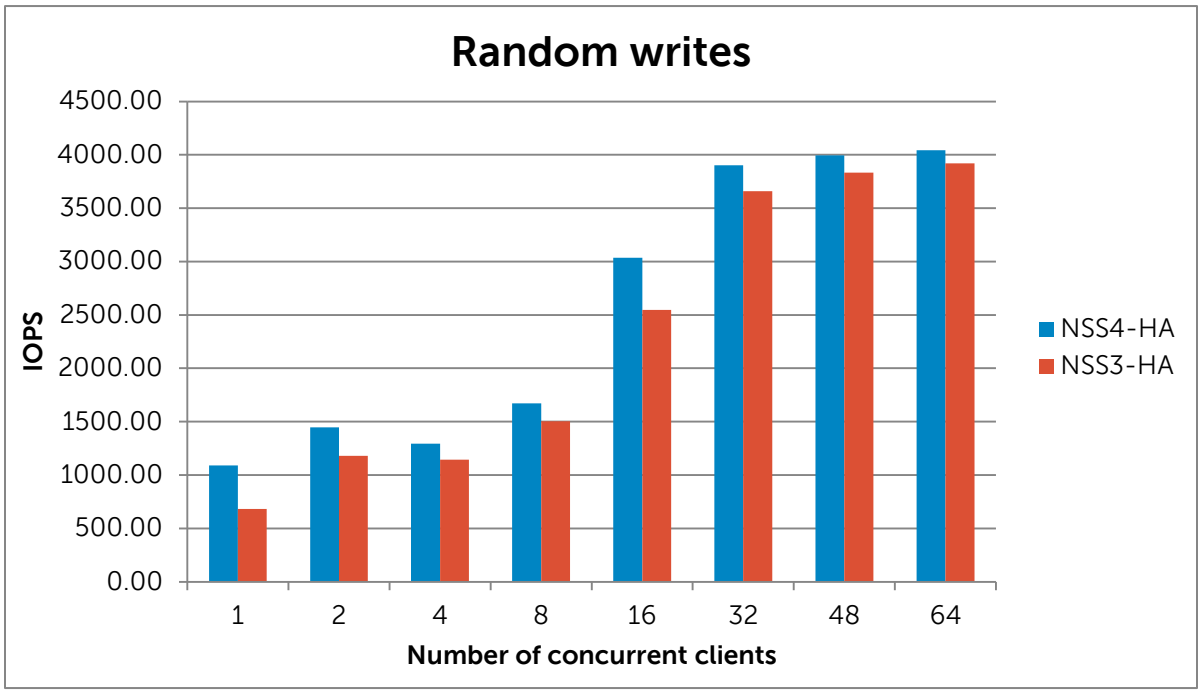
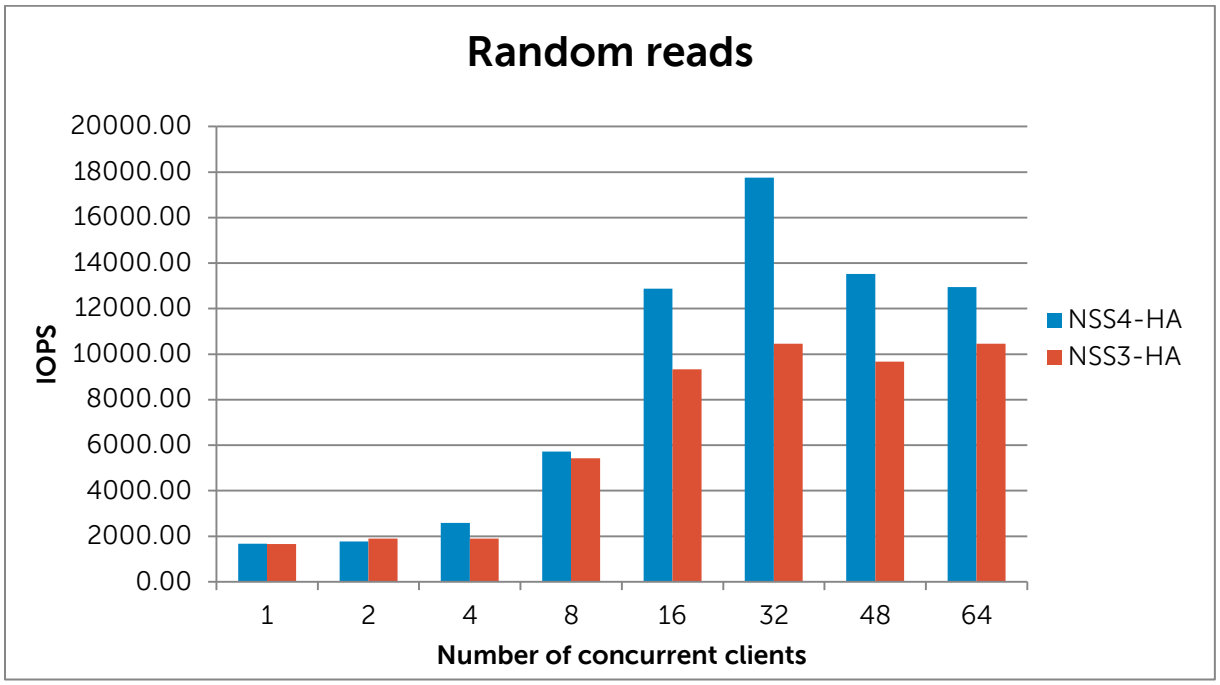


Figure 8. IPoIB random read performance: NSS4-HA vs. NSS3-HA



5.3. Metadata tests

This section compares the metadata operation performance between the current NSS4-HA release with the Dell PowerEdge R620 server and the previous NSS3-HA release [3] with the Dell PowerEdge R710 server. The test parameters, such as number of files, number of runs, and so on, are same for both releases. As similar as sequential and random I/O tests, the metadata operation performance with the

Dell PowerEdge R620 is still better than with the Dell PowerEdge R710. The improvement on average is more than 20 percent.

Figure 9, Figure 10, and Figure 11 show the results of file create, stat and remove operations, respectively. As the HPC compute cluster has 64 compute nodes, in the graphs below each client executed a maximum of one thread for client counts up to 64. For client counts of 128, 256 and 512, each client executed 2, 3 or 4 simultaneous operations.

Figure 9. IPoIB file create performance: NSS4-HA vs. NSS3-HA

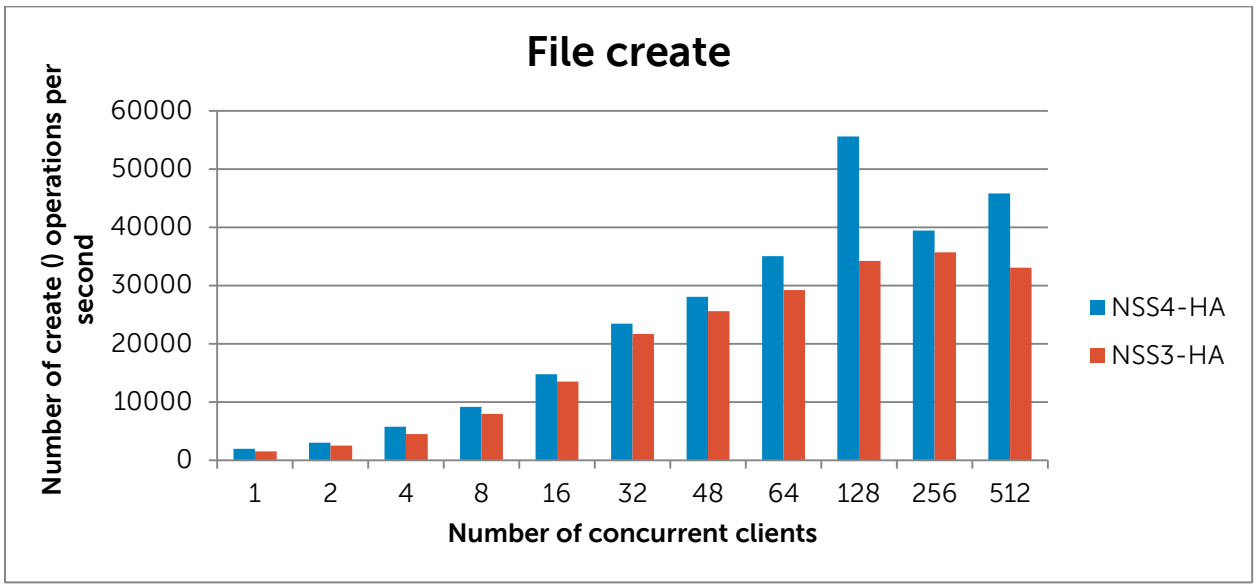


Figure 10. IPoIB file stat performance: NSS4-HA vs. NSS3-HA

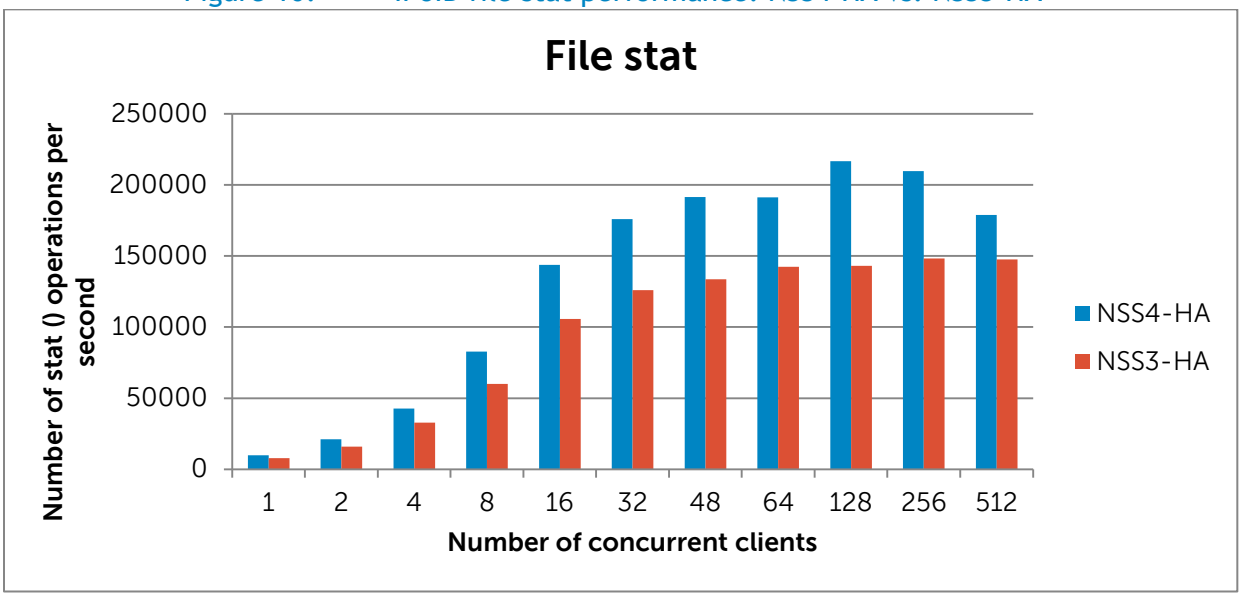
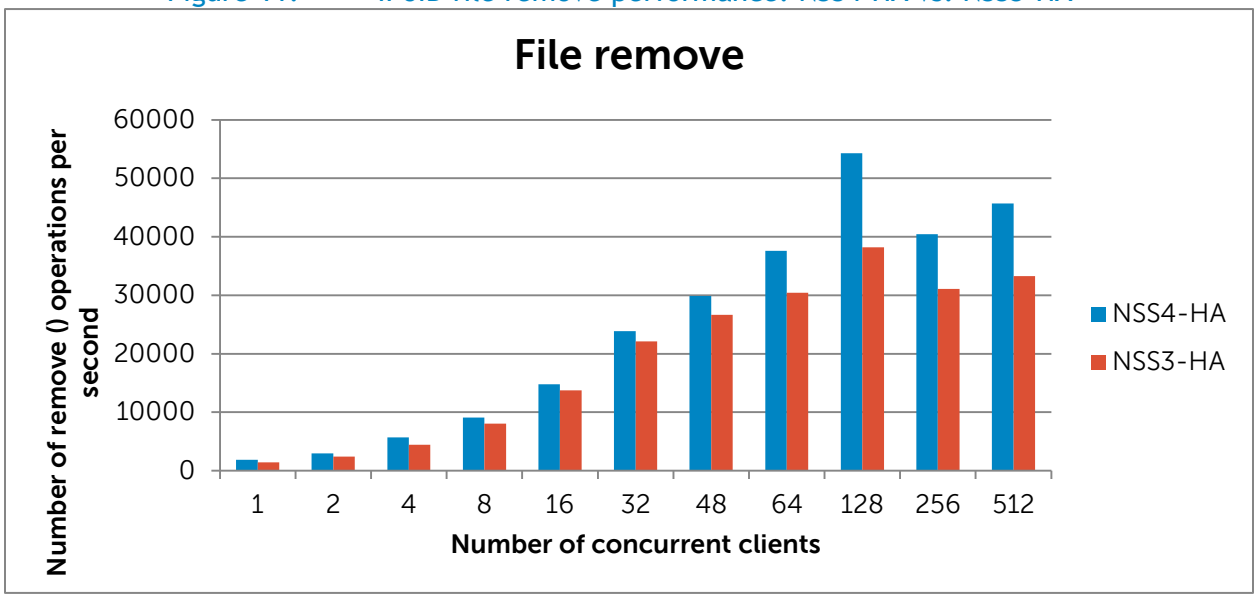


Figure 11. IPoIB file remove performance: NSS4-HA vs. NSS3-HA



6. Conclusion

This solution guide provides details the latest NSS-HA Solution for HPC from Dell. With this release, the Dell NSS-HA solution gains huge performance improvement with Dell PowerEdge 12th generation servers than the previous releases. The Dell NSS-HA solution is available with deployment services and full hardware and software support from Dell. This document provides complete technical details on architecture, design, configuration, and performance analysis of such solutions.

7. References

1. Red Hat Enterprise Linux 6 Cluster Administration -- Configuring and Managing the High Availability Add-On.

   http://docs.redhat.com/docs/en-US/Red_Hat_Enterprise_Linux/6/pdf/Cluster_Administration/Red_Hat_Enterprise_Linux-6-Cluster_Administration-en-US.pdf

2. Dell HPC NFS Storage Solution High Availability Configurations, Version 1.1 http://i.dell.com/sites/content/business/solutions/whitepapers/en/Documents/dell-hpc-nssha-sg.pdf

3. Dell HPC NFS Storage Solution – High availability with large capacities, Version 2.1 http://i.dell.com/sites/content/business/solutions/engineering-docs/en/Documents/hpc-nfs-storage-solution.pdf

4. Intel® Xeon® Processor E5-2600 Product Family http://download.intel.com/newsroom/kits/xeon/e5/pdfs/Intel_Xeon_E5_Factsheet.pdf

Appendix A: Benchmarks and test tools

The `iozone` benchmark was used to measure sequential read and write throughput (MB/sec) as well as random read and write I/O operations per second (IOPS).

The `mdtest` benchmark was used to test metadata operation performance.

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 and to measure data throughput as well as time to complete for file copy operations.

A.1. IOzone

You can download the `IOzone` from http://www.iozone.org/. Version 3.397 was used for these tests and installed on both the NFS servers and all the compute nodes.

The `IOzone` tests were run from 1-64 nodes in clustered mode. All tests were N-to-N, that is N clients would 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. This umounts the XFS file system on the server.
- Start the cluster service on the server.
- Mount NFS Share on clients.

The following table describes the 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

IOzone Argument	Description
-t	Number of threads
+m	Location of clients to run IOzone on when in clustered mode
-w	Does not unlink (delete) temporary file
-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 such that the total amount of data written was 128G (number of clients * file size per client = 128G).

IOzone Sequential Writes

```
# /usr/sbin/iozone -i 0 -c -e -w -r 1024k -s 4g -t 64 -+n -+m ./clientlist
```

IOzone Sequential Reads

```
# /usr/sbin/iozone -i 1 -c -e -w -r 1024k -s 4g -t 64 -+n -+m ./clientlist
```

For the random tests, each client read or wrote a 2G file. The record size used for the random tests was 4k to simulate small random data accesses.

`IOzone` IOPs Random Access (Reads and Writes)

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

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

A.2. mdtest

You can download `mdtest` can be downloaded from http://sourceforge.net/projects/mdtest/. Version 1.8.3 was used in these tests. It was compiled and installed on a NFS share that was accessible by compute nodes. `dtest` is launched with `mpirun`. For these tests, MPICH2 version 1.3.2 was used. The following table describes the `mdtest` command-line arguments.

mpirun ARGUMENT	DESCRIPTION
-np	Number of Processes
--nolocal	Instructs mpirun not to run locally
--hostfile	Tells mpirun where the hostfile is
▪ mdtest ARGUMENT	▪ DESCRIPTION
-d	The directory mdtest should run in
-i	The number of iterations the test will run
-b	Branching factor of directory structure
-z	Depth of the directory structure
-L	Files only at leaf level of tree
-I	Number of files per directory tree
-y	Sync 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 run

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

- Unmount NFS share on clients.
- Stop the cluster service on the server. This umounts the XFS file system on the server.

- Start the cluster service on the server.
- Mount NFS Share on clients.

Metadata file and directory creation test:

```
# 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
```

Metadata file and directory stat test:

```
# 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
```

Metadata file and directory removal test:

```
# 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
```

A.3. Checkstream

The `checkstream` utility is available at http://sourceforge.net/projects/checkstream/. Version 1.0 was installed and compiled on the NFS servers and used for these tests.

First, a large file was created using the `genstream` utility. This file was copied to and from the NFS share by a client using `dd` to mimic 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 resultant output files were checked using the `checkstream` utility to test for data correctness and ensure that there was no data corruption.

Given below is 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, here is an example of a failing test with data corruption in the copied file. For example, if the file system is exported via the NFS async operation and there is an HA service failover during a write operation, data corruption is likely to 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
```

A.4. The dd Linux utility

`dd` is a Linux utility provided by the coreutils rpm distributed with RHEL 6.1. It is used to copy a file. The file system was mounted at `/mnt/xfs` on the client.

To write data to the storage, the following command line was used.

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

To read data from the storage, the following command line was used.

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