# Easy Peasy NFS with RHEL6.5

Nirmala Sundararajan, Anthony Fernandez, June 2014

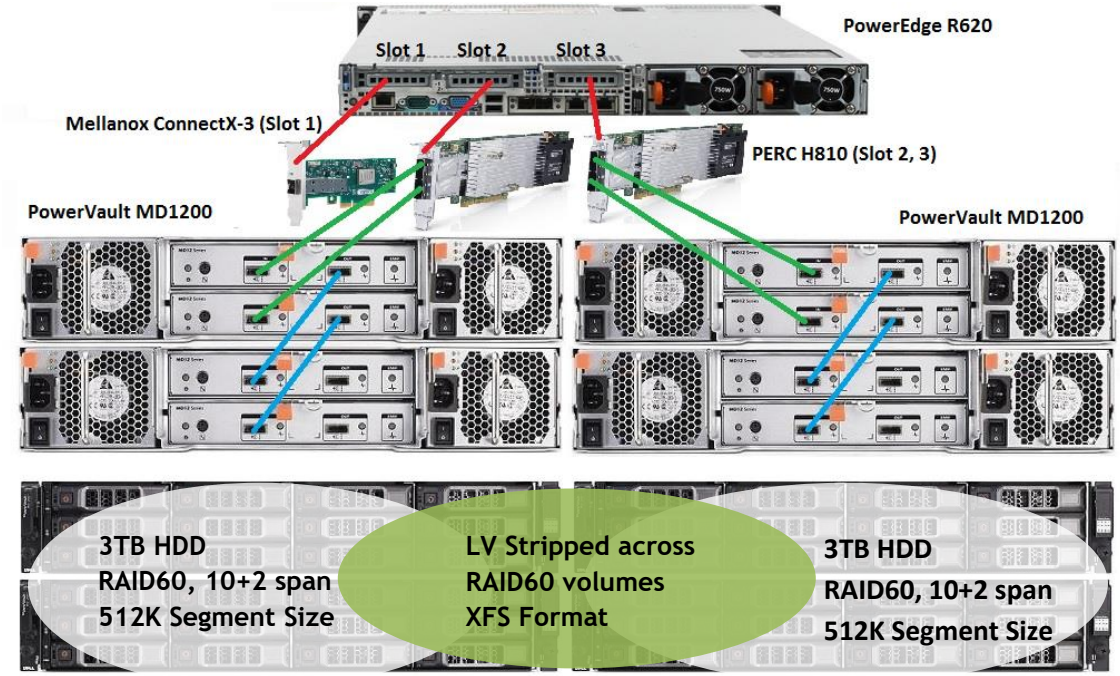
Network File System (NFS) is a central component of data processing. Multiple hardware components and interconnects as well as performance tuning options can make this process a complex one. Often times this can result in component mismatch or unsupported configurations. So were you looking for a solution that is simple yet scalable?

This blog describes a solution that we verified and benchmarked in the Dell HPC Engineering lab. It's called NSS5.5 (NFS Storage Solution, version 5.5) and aims to offer more than simplicity and scalability. Described below are the configuration and performance plus information on a step-by-step guide on best practices and how to set up such a solution. Read on for details!

For those of you familiar with our NSS-HA solutions, NSS is the non-HA version .The NSS5.5 solution includes one NFS server direct attached to standard SAS JBOD arrays. In the current 5.5 version, the NFS server is a PowerEdge R620 server with Ivy Bridge based processors (Intel Xeon E5-2600 v2 series) and the NSS storage comprises of four direct attached PowerVault MD1200 storage arrays. The NSS5.5 uses Red Hat Enterprise Linux 6.5 as the operating system on the NFS server, InfiniBand ConnectX-3 as the I/O network between the NFS server and the compute clients, and Mellanox OFED 2.1.1.0.6.

Figure 1 shows the NSS architecture as well as the high level file system layout.

## Figure 1 Architecture Diagram




The NFS server uses two PERC H810 RAID controllers, each of which is connected to two storage arrays. Each of the four storage array has 12 3TB 7.2K RPM NL-SAS drives. Each storage array was configured as RAID 60 unit with 10+2 span length, 512KB stripe element size. Two RAID 60 sets were combined as a Logical Volume (LV) as done in previous NSS configurations. This gives the NSS a total raw capacity of 144TB and 110TB usable space.

At the Virtual Disk (VD) level the Read Cache Policy is set to Read Ahead, Write Cache Policy set to Write Back and Disk Cache Policy is Disabled. The Disk Cache Policy is disabled by default to ensure data integrity in the event of power failure at the storage disk level.

Details on these and other design choices for the NSS line of solutions are described in this white paper.

The logical volume was formatted with Red Hat's Scalable File System (XFS). The XFS file system was exported with the 'async' option. This enables the NFS server to acknowledge writes before any changes are committed to disk thus providing better performance than 'sync' option. NFS protocol v3 was used between the compute clients and the NFS server.

The compute clients (in our case, a 64-node PowerEdge M420 blade compute cluster) accessed the file system over InfiniBand using the IPoIB protocol.

The IOzone benchmark was used to test sequential and random I/O performance. We conducted the test with a single thread per client.

The total I/O was kept constant at 256 GB. That is, for the 1 thread/1 client case, a single client read/wrote one 256 GB file; with 2 clients each client read/wrote a 128 GB file size and so on. 256 GB is 2x larger than the memory on the NFS server and using a large file size helps minimize cache effects and report the true performance from the storage solution.

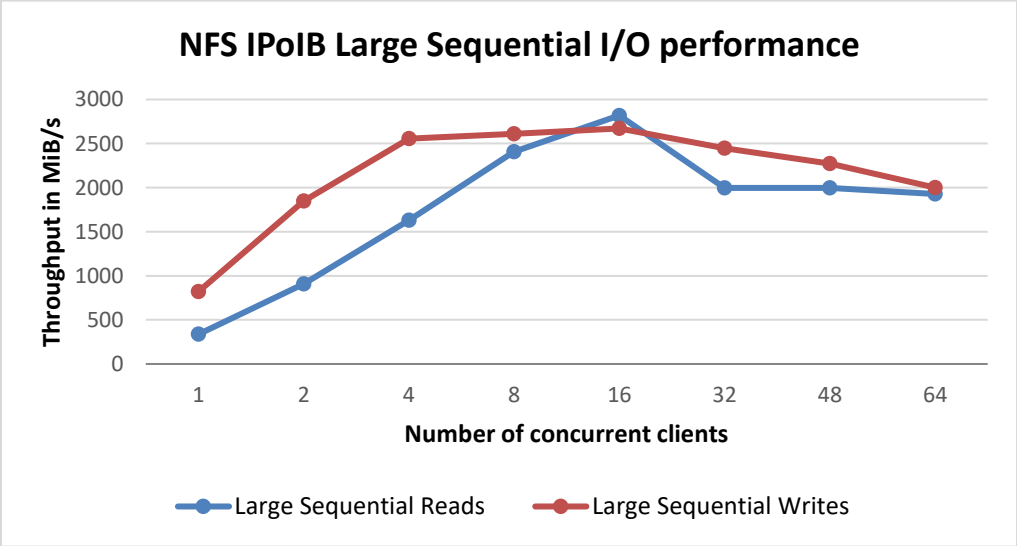
Details of the benchmark tool and methodology used are provided in Table 1.

### Table 1 Benchmark Information

| | |
|---|---|
| **Sequential Tests** | Iozone benchmark. V 3.408<br><br>1024k record size<br><br>File size varied depending on number of concurrent clients to keep total I/O at 256 GB.<br><br>Example, 1 client operated on a 256GB file, 2 clients operated on 128 GB files each, ...64 clients operated on a 4GB file each.<br><br>*iozone -i <0,1> -c -e -w -r 1024k -s <256g-4g> -t <1-64> -+n -+m ./iolist.<1-64>* |
| **Random Tests** | Iozone benchmark. V 3.408<br><br>4k record size<br><br>Each client operated on a 4GB file for all cases.<br><br>*IOZONE -I 2 -W -R 4K -I -O -S 4G -T <1-64> -+N -+M ./IOLIST.<1-64>* |

Figure 2 shows the results for large sequential I/O of block size 1024K for reads and writes.

### Figure 2 Large Sequential Test Result



As you can see from the graph a single NFS server can provide peak throughput of 2,671 MiB/s for sequential writes and 2,818 MiB/s for sequential reads.

The results of the random reads and writes are shown in Figures 3 and 4. The random tests use a block size of 4K and a file size of 4G for all test cases. Note: Depending on the data file size used, these numbers may vary.

Figure 3 Random Read Test Result

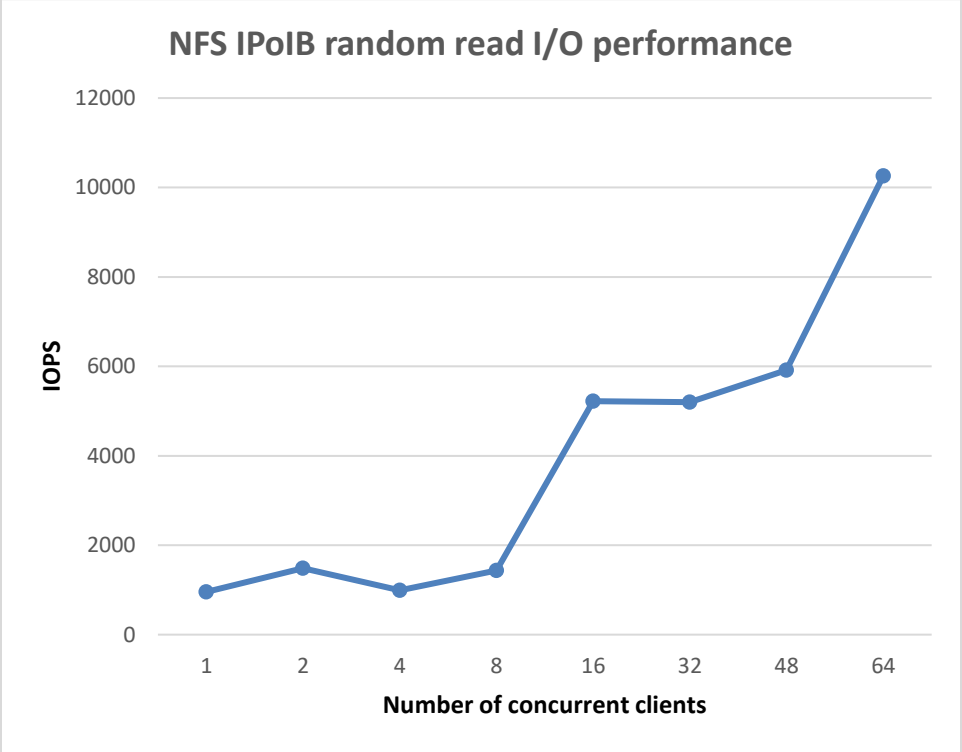
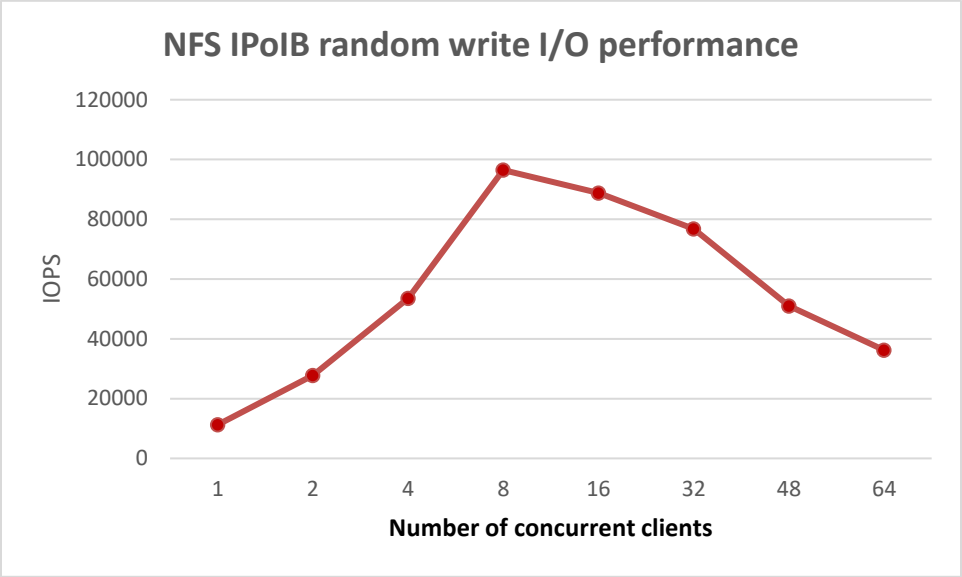


Figure 4 Random Write Test Result




A peak close to 100,000 IOPS was achieved on random writes. With 64 simultaneous clients, the peak measured with random reads is ~10,000 IOPS. From the graph in figure 3 it can be seen that the random read performance is still scaling, and has not reached saturation.

Details of the test bed are provided in the Tables 2, 3, 4 and 5.

Table 2. NSS5.5 - Hardware Configuration

| Server configuration | |
|---|---|
| **NFS server model** | Dell PowerEdge R620 |
| **Processor** | (2) Intel Xeon E5-2680 v2 @ 2.80 GHz 8 Core |
| **Memory** | 128GB (16 x 8GB 1866 MHz Single Ranked RDIMMs) |
| **Local disks and RAID controller** | (1) PERC H310 with five 300GB 10K SAS hard drives. Two drives are configured in RAID-1 for the OS, two drives are configured in RAID-0 for swap space, one drive is a hot spare for RAID-1 disk group (*)<br><br>*Note: Internal drive configuration can be configured with different RAID types and quantity of hard drives, etc. For the porpuse of the tests, the described configuration is the minimum hard drive configuration. |
| **Optional InfiniBand HCA** | (1) Mellanox ConnectX-3 FDR PCI-E card (Slot 1) |

| | |
|---|---|
| **1gbE Ethernet card (Daughter card slot)** | (1) Intel Gigabit Ethernet Quad Port I350-t Network Daughter Card |
| **External storage controller** | (2) PERC H810 RAID Controllers (Slot 2, 3) |
| **Systems Management** | iDRAC7 Enterprise |
| **Power Supply** | (2) Power Supply Units |
| **Storage configuration** | |
| **Storage Enclosure** | (4) Dell PowerVault MD1200 JBOD enclosure |
| **Hard Disk Drives** | (48) 3TB 7200 rpm NL SAS drives |

## NSS5.5 - Software Versions

| Software | |
|---|---|
| **Operating system** | Red Hat Enterprise Linux (RHEL) 6.5 x86_64 |
| **Kernel version** | 2.6.32-431.el6.x86_64 |
| **File system** | Red Hat Scalable File System (XFS) 3.1.1-14 |
| **Systems management tool** | Dell OpenManage Server Administrator 7.3.2 |

Table 4.

## NSS5.5 - Firmware and Driver Versions

| Firmware and Drivers | |
|---|---|
| **Dell PowerEdge R620 BIOS** | 2.2.2 |
| **Dell PowerEdge R620 iDRAC7 Enterprise** | 1.56.55 (Build 05) |
| **InfiniBand HCA firmware** | 2.30.8000 |
| **InfiniBand driver** | Mellanox OFED 2.1.1.0.6 |
| **PERC H810 firmware** | 21.2.0-0007 |
| **PERC H810 driver** | megaraid_sas 06.700.06.00-rh1 |

### NSS5.5 - Client / HPC Compute Cluster

| Client / HPC Compute Cluster | |
|---|---|
| **Clients**<br>Table 5. | 64 PowerEdge M420 blade servers<br>32 blades in each of two PowerEdge M1000e chassis<br>Red Hat Enterprise Linux 6.4 x86-64 |
| **Chassis configuration** | Two PowerEdge M1000e chassis, each with 32 blades<br>Two Mellanox M4001F FDR10 I/O modules per chassis<br>Two PowerConnect M6220 I/O switch modules per chassis |
| **InfiniBand** | Each blade server has one Mellanox ConnectX-3 Dual-port<br>FDR10 Mezzanine I/O Card<br>Mellanox OFED 2.0-2.0.5 |
| **InfiniBand fabric for I/O traffic** | Each PowerEdge M1000e chassis has two Mellanox M4001<br>FDR10 I/O module switches.<br>Each FDR10 I/O module has four uplinks to a rack Mellanox<br>SX6025 FDR switch for a total of 16 uplinks.<br>The FDR rack switch has a single FDR link to the NFS server. |

For more configuration details, including the step-by-step configuration instructions, please contact your Dell Sales or Services representative.