



Achieving over 1-Million IOPS from Hyper-V VMs in a

Scale-Out File Server Cluster using Windows Server 2012 R2

Authors:

Liang Yang

Danyu Zhu

Jeff Woolsey

Senthil Rajaram

A Microsoft White Paper

Published: May 2014

This document is provided “as-is.” Information and views expressed in this document, including URL and other Internet Web site references, may change without notice. You bear the risk of using it.

This document does not provide you with any legal rights to any intellectual property in any Microsoft product. You may copy and use this document for your internal, reference purposes.

© 2014 Microsoft Corporation. All rights reserved.

Microsoft, Windows, Windows Server, Hyper-V are either registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries.

The names of actual companies and products mentioned herein may be the trademarks of their respective owners.

The platform this paper describes was assembled to facilitate open-ended experimentation and allow us to hit high IOPS in a controlled environment. It does not reflect a Microsoft recommended storage configuration.

# Summary

This white paper demonstrates the high performance storage and virtualization capabilities for a Hyper-V virtualized application using the Scale-Out File Server (SOFS) platform on Windows Server 2012 R2. The platform uses industry-standard servers, disks and JBOD enclosures. With the block storage made available through Storage Spaces combined with the Scale-Out File Server in a failover cluster, we can deliver extremely high levels of storage performance and resiliency similar to a traditional SAN using Microsoft Storage Solution and inexpensive shared Direct Attached Storage (DAS).

The results demonstrate the levels of storage performance that can be achieved from Hyper-V VMs in a single SOFS using shared Storage Spaces:

* 1.1 Million IOPS for 8KB random reads
* 700K IOPS for 8KB random writes
* 100Gbps aggregate read bandwidth
* 92Gbps aggregate write bandwidth

# Introduction

In the modern datacenter, storage solutions play a critical role as the storage demands continue to grow with the explosion of new application data and the users’ expectations for continuous services. Windows Server 2012 R2 introduces a wide variety of storage features and capabilities to address the storage challenges in a virtualized infrastructure with enhanced performance, flexibility, resiliency and continuous availability [1]. It provides the user with the ability to expand the storage infrastructure to keep up with growing storage needs while delivering required performance and reliability. With strong focus on storage capability and Microsoft leading virtualization technology by Hyper-V, Windows Server 2012 R2 with Hyper-V takes advantage of new hardware technology and supports most data intensive, mission-critical workloads in a virtualized environment.

In this white paper, we discuss the new storage and virtualization capabilities as well as the significant improvements made in Windows Server 2012 R2. Using the technologies and features provided in Windows Server 2012 R2, we built up a high performance enterprise-class Scale-Out File Server that can deliver extremely high levels of storage performance, resiliency, efficiency and management. Traditionally, shared direct-attached storage (DAS) provisioned using industry-standard disks and JBOD enclosures is not highly available by nature. Also, Storage Area Network (SAN) technologies such as iSCSI or Fibre Channel offers high availability but with high maintenance cost. With the block storage made available through Storage Spaces combined with Scale-Out File Server in a failover cluster, the platform can deliver much of the same performance and reliability as a SAN using Microsoft Storage Solution and inexpensive shared DAS.

Scale-Out File Server is designed to provide scale-out file shares that are continuously available for file-based server application storage. The primary applications for a Scale-Out File Server are Hyper-V VMs or SQL server. Hyper-V VMs can take advantage of Windows Server 2012 R2 storage features to store virtual machine files (VHD/VHDXs) remotely in Scale-Out File Server shares without disruption of service. Scale-Out File Server becomes an ideal file server type when deploying Hyper-V over SMB.

This white paper demonstrates workloads running in Hyper-V VMs can achieve over 1 million IOPS throughput and 100Gbps bandwidth in Scale-Out File Server over the much enhanced Microsoft SMB 3.0 on Storage Spaces.These powerful capabilities of the Microsoft Storage and Virtualization Solution in Windows Server 2012 R2 provide customers with a comprehensive platform to handle the storage demands of the modern datacenter.

# Building a high performance Scale-Out File Server

## Software Components

Windows Server 2012 R2 provides a rich set of storage features allowing you to take advantage of lower- cost industry-standard hardware without compromising performance or availability. Figure 1 presents an overview of the Microsoft software storage and virtualized components.



Figure 1: Windows Server 2012 R2 Storage Stack Diagram

Some of the features used in this white paper include the following:

**Microsoft Storage Stack**

* **Storage Spaces [2]**: Storage Spaces in Windows Server 2012 R2 provides many features including write-back cache, storage tiering, trim, mirroring, and parity with just a bunch of disks (JBOD). It gives you the ability to consolidate all of your Serial Attached SCSI (SAS) and Serial ATA (SATA) connected disks into Storage Pools. Storage Spaces are compatible with other Windows Server 2012 R2 storage features, including SMB Direct and Failover Clustering, so that a user can utilize industry-standard hardware to create powerful and resilient storage infrastructures on a limited budget and supply high-performance and feature-rich storage to servers, clusters, and applications.
* **Microsoft Multipath I/O (MPIO) [3]:** MPIO is a Microsoft-provided framework that allows storage providers to develop multipath solutions to optimize connectivity and create redundant hardware paths to storage arrays. In this whitepaper, we use Microsoft Device Specific Module (DSM) working with Microsoft Storage Spaces to provide fault tolerant connectivity to the storage array.

**Microsoft Virtualization Stack**

* **Hyper-V Storage NUMA I/O**: Windows Server 2012 R2 supports large virtual machines, up to 64 virtual processors. Any large virtual machine configuration typically also needs scalability in terms of I/O throughput. Windows Server 2012 R2 provides a Hyper-V storage NUMA I/O capability to create a number of communication channels between the guest devices and host storage stack with a specified dedicated set of VPs for the storage IO processing. Hyper-V storage NUMA I/O offers a more efficient I/O completion mechanism involving interrupts distribution amongst the virtual processors to avoid expensive inter-processor interruptions. With those improvements, the Hyper-V storage stack can provide scalability improvements in terms of I/O throughput to support the needs of large virtual machine configuration with data intensive workloads like SQL.
* **Storage Quality of Service (QoS) and Hyper-V IO Balancer [4]**: Storage QoS is a new feature provided in Windows Server 2012 R2. It offers the capability to set certain QoS parameters for storage on the virtual machines. Storage QoS and Hyper-V IO balancer provide the ability to specify maximum input/output operations per second (IOPS) values for an individual virtual hard disk. Windows Server 2012 R2 Hyper-V, with Storage QoS, can throttle storage assigned to VHD/VHDX in the same volume to prevent a single virtual machine from consuming all I/O bandwidth and help to control the balance of VM storage demand, and storage performance capacity. Currently, IO Balancer works on the host side and thus storage QoS only supports non-shared storage.
* **VHDX [5]:** VHDX is a new virtual hard disk format introduced in Windows Server 2012, which allows you to create resilient high-performance virtual disks up to 64 terabytes in size. Microsoft recommends using VHDX as the default virtual hard disk format for VMs. VHDX provides additional protection against data corruption during power failures by logging updates to the VHDX metadata structures, as well as the ability to store custom metadata. The VHDX format also provides support for the TRIM command which results in smaller file size and allows the underlying physical storage device to reclaim unused space. The support for 4KB logical sector virtual disk as well as the larger block sizes for dynamic and differential disks allows for increased performance.

**Microsoft SMB 3.0**

* **SMB Multichannel [6]**: SMB multichannel provides the capability to automatically detect multiple networks for SMB connections. It not only offers resilience against path failures and transparent failover with recovery without application service disruption, but also with much improved throughput by aggregating network bandwidth from multiple network interfaces. Server applications can then take full advantage of all available network bandwidth, making them more resistant to network failure.
* **SMB Direct**: SMB Direct (SMB over RDMA), introduced in Windows Server 2012, exposes Remote Direct Memory Access (RDMA) hardware support for SMB to provide high-performance storage capabilities. SMB Direct is intended to lower CPU consumption and latency on both the client and server while delivering high IOPS and bandwidth utilization. It can deliver enterprise-class performance without relying on expensive Fibre Channel SAN. With the CPU offloading and the ability to read and write directly against the memory of the remote storage node, RDMA network adapters can achieve extremely high performance with very low latency. Currently, SMB Direct supports InfiniBand, Internet Wide Area RDMA Protocol (iWARP) and RDMA over Converged Ethernet (RoCE). In this white paper, we demonstrates remote file server performance comparable to local storage utilizing SMB Direct over InfiniBand. SMB Direct is also compatible with SMB Multichannel to achieve load balancing and automatic failover.

**Scale-Out File Server Cluster**

* **Failover Clustering [7]:** A failover cluster is a group of independent computers that work together to increase the availability and scalability of clustered roles. If one or more of the cluster nodes fail, the service will automatically failover to other node without disruption of service.
* **Scale-Out File Server Role [8]**: The Scale-Out File Server role not only provides a continuously available SMB service, but also provides a mechanism for clustered file servers in an active-active configuration to aggregate bandwidth across the cluster nodes. In SOFS, SMB clients are transparently directed to do their IO against their owner node to achieve balancing around the cluster.
* **Cluster Shared Volumes (CSV):** CSVs in a Windows Server 2012 R2 failover cluster allow multiple nodes in the cluster to simultaneously access shared storage with a consistent and distributed namespace. To each of the cluster nodes in the cluster, the CSV provides a consistent single file namespace. Therefore, CSVs greatly simply the management of large number of LUNs in a failover cluster.
* **Continuous Availability (CA) File Share**: In this white paper, we create a failover cluster with the Scale-Out File Server role to offer continuous availability for SMB shares. In continuous availability file shares, persistent file handles are always opened with write through to guarantee that data is on stable storage and durable against cluster node failure.

## Platform Topology and Cabling Connections

Failover clustering built on top of Storage Spaces using SAS disk drives provides a resilient, high available and cost effective solution for simple deployment in a large data center. Figure 2 below shows the topology for the platform being used in this report. Both Hyper-V server and Scale-Out File Server cluster are using Dell R820 rack servers. Every server has two InfiniBand RDMA network adapters installed and both ports of each adapter are used. The client and cluster traffic between two clusters are routed through 56G InfiniBand fabric with a 56G InfiniBand switch sitting in the middle (note: the actual usable bandwidth for FDR InfiniBand is 54G). Each file server machine has 4 LSI SAS HBAs installed to connect to a shared JBOD storage using multiple paths.



Figure 2. Hyper-V Server & Scale-Out File Server Cluster Platform Topology

There are two factors to be considered when we connect file server hosts and storage target of the platform: redundancy and performance. The redundancy is best achieved at both HBA and host level. That means, if there are failures happening for a SAS HBA in a host, ideally it is better to let the same host find the available path first to the target so we can avoid triggering the failover to another host as failover is typically considered an expensive operation. That is why we split the paths from each host to the target (SAS SSDs) into two by hooking up half of SAS HBAs to the first I/O module (expander) of the JBOD and the other half of SAS HBAs on the same host to the second I/O module of JBOD. To get redundancy across host level, SAS HBAs on each host are also connected to both I/O modules of JBOD. That will preserve the access to the target if the entire host fails. To get the maximum bandwidth from the underlying PCIe slots, although each SAS HBA has two 4x wide ports, we only use one port from each HBA to connect to SAS port on the I/O module of JBOD. Figure 3 below shows the cabling method we adopt in this report.

## 

Figure 3. SAS Cabling Connection Diagram

## Hardware Components

SAS based storage has been widely adopted in enterprise due to its full support of SCSI command set and protocol while maintaining application compatibility with existing software investment. With the introduction of 12G SAS storage, SAS will become an even more appealing storage bus protocol. All the components used to build this lightning fast platform are commodity hardware.

### Server Machines: Dell R820

We use Dell R820 as both our front-end Hyper-V server and back-end file server machines. As the latest generation PowerEdge rack server offered by Dell, the R820 with PCIe 3.0 support is a high performance platform designed for both compute and storage intensive applications. The R820s being used in this report are powered by quad Intel Xeon Sandy Bridge processors with highly scalable memory and ample I/O bandwidth which enable it to readily handle very demanding and mission critical workloads in a wide range of virtualization environments.

The new family of Intel Sandy Bridge processors has embedded PCIe lanes for improved I/O performance with reduced latency and they support up to 160 lanes of PCIe 3.0 (40 per socket).

### InfiniBand Fabric: Mellanox SX6036 Switch and ConnectX-3 VPI Network Adapter

InfiniBand fabric helps to optimize the network efficiency making it a good fit for converged data centers operating a wide range of applications. Mellanox’s FDR InfiniBand based solution for data centers and high-performance computing systems includes ConnectX-3 adapters, SwitchX family of FDR InfiniBand switches and FDR copper cables ensure high interconnect performance.

* + - Mellanox Dual Port FDR InfiniBand ConnectX-3 adapter cards: Mellanox’s ConnectX-3 InfiniBand adapters provide high performing and flexible interconnect solution. ConnectX-3 delivers up to 54Gb/s throughput across the PCI Express 3.0 host bus, enables fast transaction latency, less than 1usec.
    - Mellanox SX6036 36-port FDR InfiniBand Switch: The SX6036 switch systems provide high performing fabric solutions in a 1RU form factor by delivering 4.032Tb/s of non-blocking bandwidth with 200ns port-to-port latency. Built with Mellanox's latest SwitchX-2 InfiniBand switch device, these switches deliver up to 54Gb/s full bidirectional speed per port.

### SAS Controller: LSI 9300-8e SAS HBA

We use LSI 9300-8e as the storage HBA to interface with the pooled storage in JBOD. LSI 9300-8e provides high performance for high-end servers connecting to large scale storage enclosures. As industry’s first 12G capable SAS HBA, 9300-8e is built on top of LSI SAS Fusion-MPT 3008 controller and comes with 8 external ports (2 of x4 external HD-Mini SAS ports). The 9300-8e supports 8 lanes of PCIe 3.0 and provides SAS links at transfer rates up to 12Gbps. The 9300-8e can support up to 1024 SAS/SATA end devices with the use of SAS expanders.

Although SAS HBAs here are capable of 12Gps, the actual SAS link speed of the platform is still capped at 6Gbps because SAS SSDs, SAS JBOD (expander) and SAS Cables only support 6Gbps.

### SSD Storage: HGST ZeusIOPS XE SAS SSD

We use HGST ZeusIOPS XE 300GB 6G SAS SSDs to form a storage pool with all SSDs. ZeusIOPS XE stands for eXtreme Endurance and is a Multi-Level Cell (MLC) flash based SSD. The XE SSD uses a combination of HGST's fourth-generation ASIC-based SSD controller and its proprietary CellCare technology to extend the performance, reliability and endurance capabilities.

It is worth noting that, like other enterprise SAS SSDs, HGST ZeusIOPS comes with super capacitors to protect the data saved in the volatile buffer in case of power failure. With super capacitor, the disk cache (buffer) of SSD is always on even though it appears it can be disabled successfully. Also, the presence of super capacitors make the support of FUA (Force Unit Access or Write Through) in a more optimal way than it would have to otherwise which can negatively affect performance and increase write amplification.

### Storage Enclosure: DataOn Storage 1660D SAS JBOD

We use DataOn Storage DNS-1660D 6G JBOD as our enclosure to host all the SSDs. DNS-1660D comes with a 4U 60-bay which provides a high-density 6Gbps SAS/SATA JBOD solution with massive data capacity. The DNS-1160D provides dual hot-pluggable I/O controller module (a.k.a. ESMs) and each ESM has 3 built-in PMC-Sierra 36-port expanders to connect to 60 devices.

# Configuration Settings

## Overview of Hyper-V and Scale-Out File Server Clusters

A Windows Server Failover Clustering (WSFC) cluster is a group of independent servers that work together to provide highly available applications and services. In this report, we create a Scale-Out File Server (SOFS) role in a cluster consisting of two nodes ([\\9-1109A0103](file:///\\9-1109A0103) and [\\9-1109A0104](file:///\\9-1109A0104)). Two Hyper-V servers ([\\9-1109A0101](file:///\\9-1109A0101) and [\\9-1109A0102](file:///\\9-1109A0102)) are running VMs with virtual hard disk files (VHDX) hosted in the SOFS cluster.

### **2-Node File Server Cluster:**

The Scale-Out File Server is built on top of a Windows Server Failover Cluster. Figure 4 below shows a snapshot of Failover Cluster Manager where a two node failover cluster has been created, prior to adding the Scale-Out file server role.

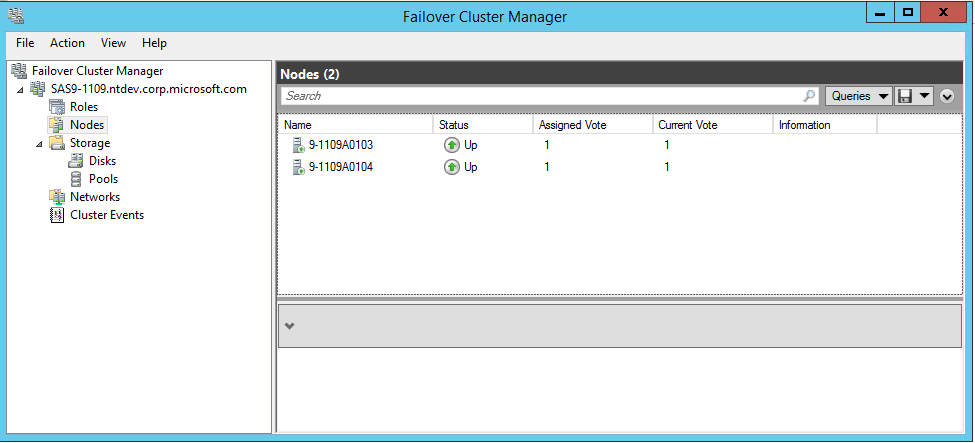


Figure 4: Two-node Failover Cluster

### **Scale-Out File Server Role added to the Failover Cluster:**

In the failover cluster, we add a scale out file server (SOFS) role ([\\SAS9-1109-SVR1](file:///\\SAS9-1109-SVR1)) as figure 5 shows.

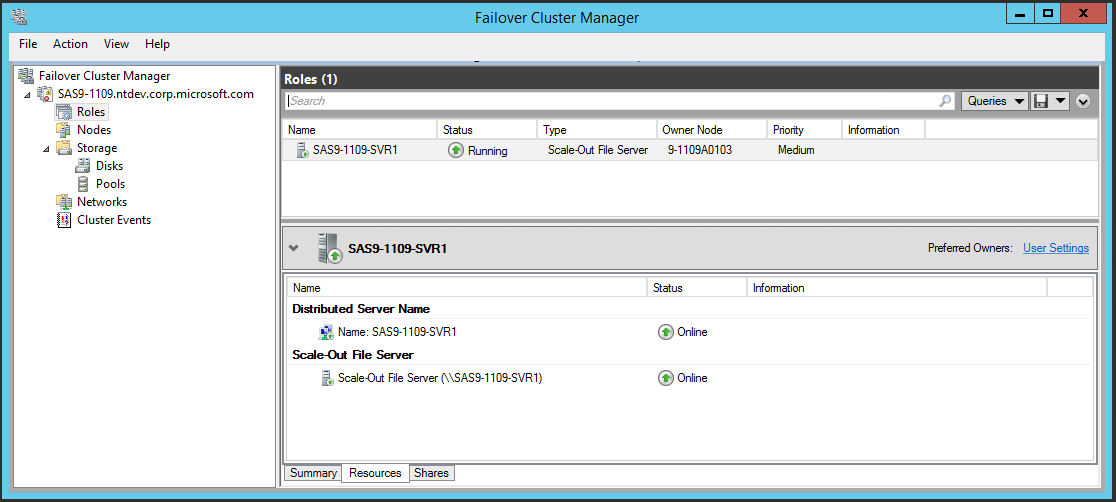


Figure 5: Scale-Out File Server Role in the Cluster

### **SMB File Shares created in SOFS:**

Using Continuously Availability (CA) file shares enables seamless service on any node in the cluster without interrupting the server applications. CA SMB share also helps redirect a server application node to a different file server node in the same cluster to facilitate better load balancing. Figure 6 shows a snapshot where all the file shares created here are continuous available with no sharing cache.

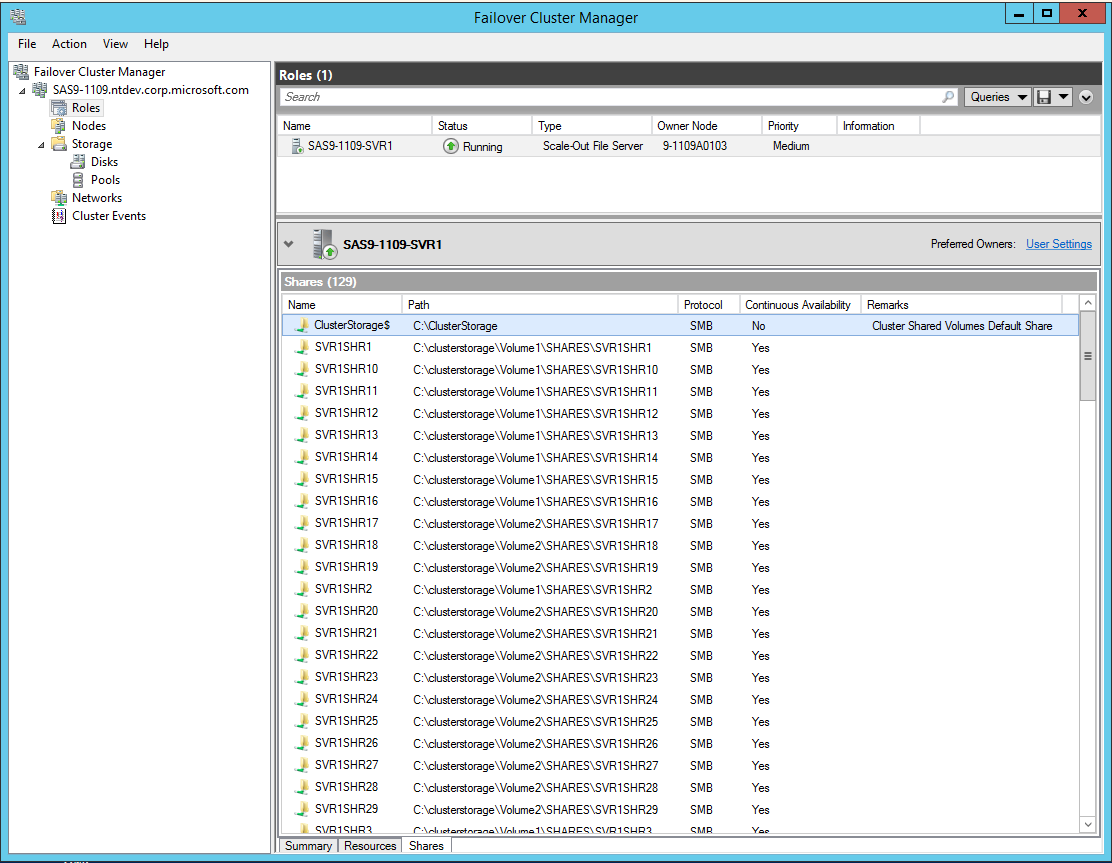


Figure 6: CA-SMB File Shares in the SOFS

### **Shared Storage with CSV in the SOFS Cluster:**

Eight virtual disks backed by Storage Spaces are added to the cluster together with one quorum disk. CSV is added on top of every shared disk in the cluster except quorum. CSV is a distributed file system access system to enable every node in the cluster to concurrently access a shared volume. By using CSV, we can unify the storage access into a single namespace for ease management and smooth VM migration purpose. SOFS only supports CSV.

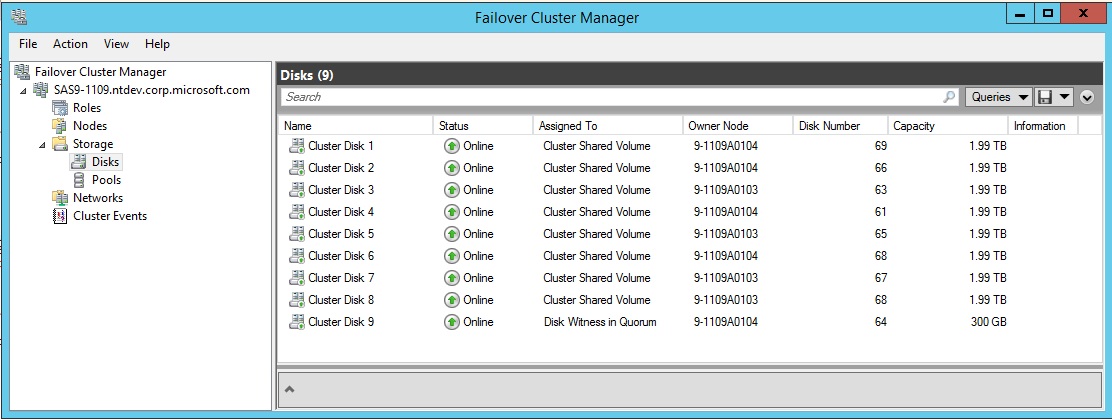


Figure 7: Shared Storage Spaces in the SOFS

### **Network Configurations in SOFS Cluster:**

SMB Multichannel allows to use multiple network interfaces for better throughput and network fault tolerance. In both traditional and scale-out file server cluster, to use the multiple paths simultaneously, a separate subnet must be configured for every NIC for SMB Multichannel as Failover Clustering will only use one IP address per subnet regardless the number of NICs on that subnet. Figure 8 shows four subnets are used in this cluster (192.168.110.0, 192.168.120.0, 192.168.130.0, 192.168.140.0) which are dedicated to four separate InfiniBand connections. Each InfiniBand network connection is assigned to allow both cluster network communication and client traffic. Figure 9 shows each InfiniBand network connection is owned by both nodes in the file server cluster.



Figure 8: Network Subnet Settings in SOFS

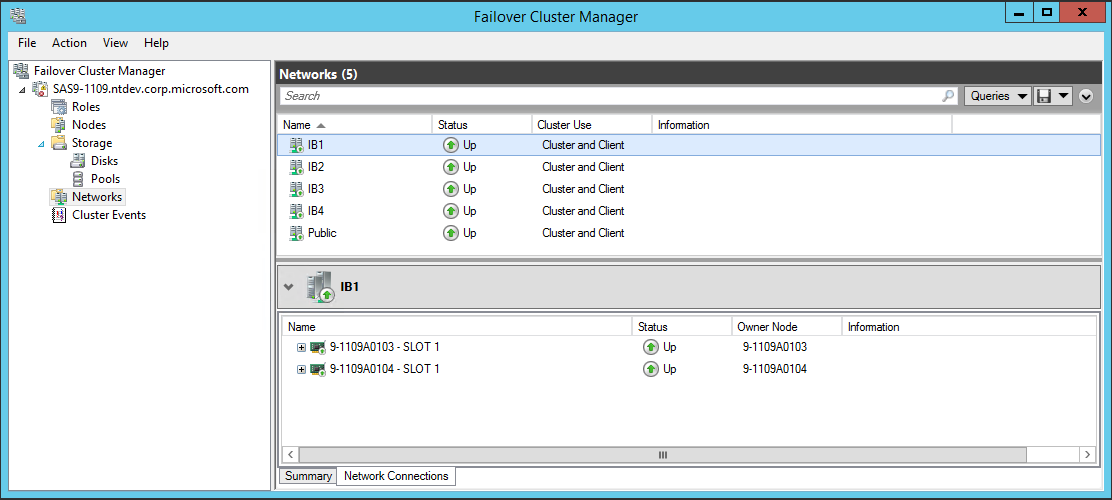


Figure 9: Network Connection Settings in SOFS

## Software Configurations

### Windows storage stack settings

#### **Windows OS**

We use Windows Server 2012 R2 Data Center SKU for both the host and VM guest OS with the latest Windows updates installed on all machines.

#### **Multi-Path I/O**

Dual ported SAS SSDs always get two active paths from the initiator when they’re hooked up to two expanders in a JBOD. The screen copy below shows default MPIO load balancing policy Round Robin was chosen using Microsoft DSM (Device Specific Module).

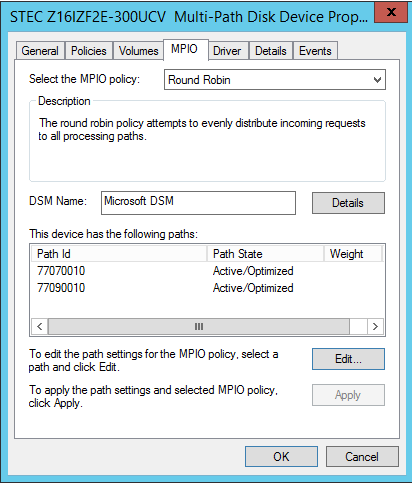


Figure 10: HGST ZeusIOPS SAS SSD with Dual Active Paths and MPIO Enabled

#### **Storage Spaces**

We created a storage pool striped across 60 SAS SSDs. On top of that storage pool, 8 Storage Spaces are created to be used as the data drive plus one to be used for quorum disk purpose in the cluster. For best performance in terms of both reads and writes and maximum capacity purposes, we choose simple storage space type (similar to RAID0) with fixed provisioning for each virtual disk.

Because this is an all SSD pool, there is no need to use portion of SSDs to create write back cache or storage tiering. The screen copy below from Windows Server Manager UI shows the settings for all the storage spaces in the cluster.

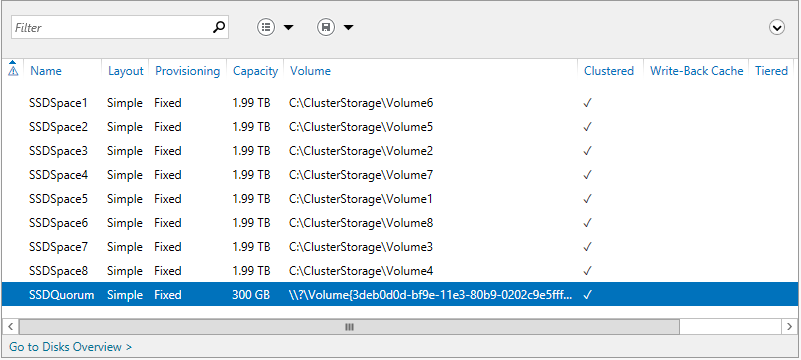


Figure 11: Storage Spaces in the cluster

#### **File System**

NTFS is the file system we used when formatting virtual disk backed by storage spaces. We used default 4K cluster size for each NTFS volume.

### Scale-Out File Server settings

#### **Symmetric Storage**

Common examples of symmetric storage are when the Scale-Out File Server is put in front of a fiber channel SAN or simple type Storage Spaces built on top of shared SAS. In this report, with all the storage spaces are created with simple type, the SMB share type created on top of storage spaces is reported as symmetric.

#### **Witness Client**

The following screen copy including the output of Get-SmbWitnessClient shows the SMB witness client registration status in this file server cluster under active workloads. The witness node and file server node associated with each client is different. This will get SMB traffic evenly distributed amongst different server nodes to avoid single server node become a bottleneck. If both clients happen to get registered with the same witness node, using PowerShell cmdlet Move-SmbWitnessClient can help us achieve active-active load balance manually.



Figure 12. SMB Witness Client Registration Status

#### **CA SMB Share**

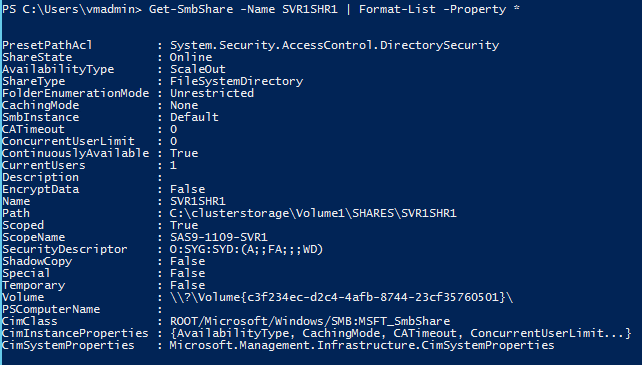
Every SMB Share is set to Continuous Availability (CA) with file caching off

Figure 13. CA SMB Share information

### Hyper-V VM Settings

Each VM has 16 virtual processors (VP) with one virtual NUMA, 16G RAM and one virtual SCSI controller with 16 VHDX files attached. Each VHDX file is about 127GB fixed type and hosted on a separate SMB share in Scale-Out file server cluster.

Table 1 list the default and maximum number of VMBus channels a VM with different number VPs support. By default, 16 VPs has one VMBus channel which will in turn has one VP involved in handling interrupt and DPC upon I/O completion. In general, the default VMBus channel setting is good enough to handle most of the workloads. In this experiment, the default single VP handling I/O completion becomes bottleneck with the extremely high I/O rate in benchmarking which causes CPU bottlenecks in VM. To mitigate that and utilize as many VPs as possible, we changed the default number of channels from 1 to 4 in each VM (HKLM\System\CurrentControlSet\Enum\VMBUS\{deviceid}\{instanceid}\Device Parameters \StorChannel\ChannelCount:4). So we can have four VPs instead of one inside VM capable of handling incoming interrupts and DPCs upon I/O completion.

**cid:image001.png@01CF2995.5AD5F130**

Table 1. Hyper-V VMBus Multi-Channel Settings(per vSCSI)

Since each VHDX file is saved on a remote SMB share and Hyper-V IO Balancer is primarily designed to balance local storage I/O traffic for VMs as of Windows Server 2012 R2, we turn off the IO Balancer here to avoid I/O throttling for VM VHDXs hosted on the remote SMB shares (HKLM\SYSTEM\CurrentControlSet \Control\StorVSP\IOBalance\Enabled:0).

Figure 14 shows the VM settings in Hyper-V VM manager UI.

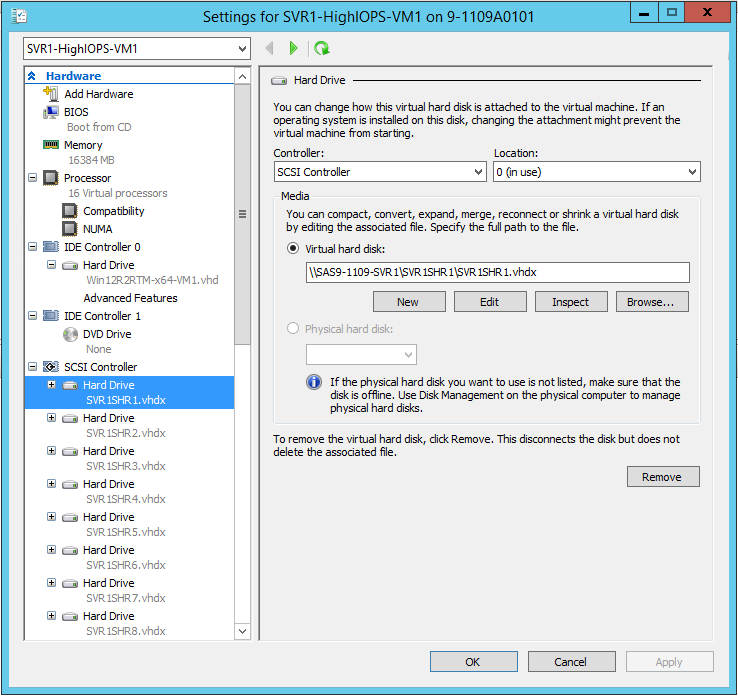


Figure 14. Hyper-V VM Settings

## Hardware Configurations

### Server Settings

Both Hyper-V servers (SMB clients) and SMB file servers are using the Dell R820 platform with Intel Sandy Bridge Xeon E5-4650L 2.60GHZ and 256G (8G x 32) 1600MHZ RDIMM RAM. With Hyper-Threading enabled, we get 64 logical processors for every host with quad socket and 8 core per socket in the system.

The Dell R820 provides 6 spare PCIe 3.0 slots. On the SMB sever machines, four slots are used for the SAS HBAs and the remaining two are used for InfiniBand network adapters. For better alignment and performance purposes, the SAS HBA and InfiniBand network cards are divided evenly between PCIe slots managed by CPU1 and CPU2, one Infiniband network adapter and two SAS HBAs per CPU[[1]](#footnote-1).

Table 2 shows Dell R820 machine settings we chose for this report:

|  |  |  |
| --- | --- | --- |
| *Settings* | | *Value* |
| *Windows (both host OS and guest OS)* | *Power Options* | *High Performance* |
| *CPU Settings in*  *BIOS (ver 2.1.0)* | *Logical Processor(HT) Support* | *Enabled* |
| *C-State* | *Disabled* |
| *Turbo Boost* | *Enabled* |
| *QPI Speed* | *Maximum data rate*  *(8.0GT/s)* |
| *Memory Settings in*  *BIOS* | *Memory Frequency* | *Maximum speed*  *(1600MHZ)* |
| *Memory Operating Mode* | *Optimizer* |
| *Node Interleaving* | *Disabled(NUMA on)* |

*Table 2. Server Machine BIOS and Power Settings*

### Network Settings

#### **MSX6036 InfiniBand Switch**

#### OS: MLNX-OS.3.3.4402

#### **InfiniBand Network Adapter**

#### Driver: 4.60.17736.0 and Firmware: 2.30.8000

All the InfiniBand network adapters on both SMB server nodes and client nodes are using the latest driver and firmware published by Mellanox as of the day this report was written. Figure 15 shows the full link speed of InfiniBand FDR at 54Gbps is achieved on the network port.

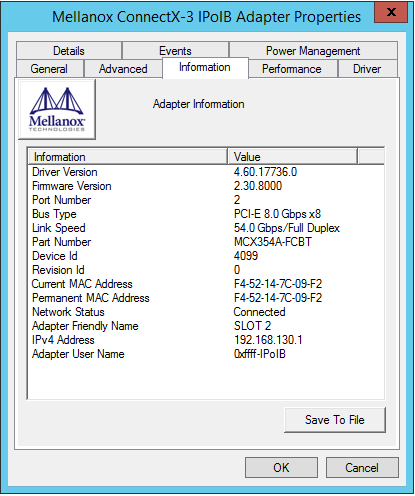


Figure 15. InfiniBand Network Adapter Information Settings

### Storage Settings

#### **SAS HBA**

All the LSI 9300-8e SAS HBAs sitting on the SMB server node are loaded with the latest public firmware and in-box driver.

##### Driver: 2.50.75.0 and firmware: Phase 3 firmware

#### **SSD**

HGST ZeusIOPS XE SAS SSD is loaded with the latest firmware from HGST.

##### Firmware: E4TD

#### **JBOD**

DataOn Storage DNS-1660D JBOD uses PMC-Sierra expanders with the latest firmware.

##### Enclosure Firmware: 0508.

# Experimental Results

## Benchmark Tool

We use IOMeter (2008.06.22.RC2) as our primary I/O benchmarking tool in this report to measure the performance. Here is a list of IOMeter settings we use for this report when running workloads within VMs:

### 1 IOMeter manager

### 16 IOMeter worker threads

### One worker per target (VHDX) and 16 targets per VM

### Ramp-up time: 5 minutes

### Run time: 5 minutes

### Queue depth per thread: 64 for Random and 1 for Sequential workloads

## Test Workloads

Different input data streams are used to get a good coverage for maximum performance in theory using both monolithic and mixed workloads. I/Os are aligned to 4K size for better performance on NAND flash.

### Synthetic monolithic workloads

##### 100% Random: 8KB 100% Reads and 8KB 100% Writes

##### 100% Sequential: 512K 100% Reads and 512KB 100% Writes

Note: 512KB I/Os are popular SQL Server DSS (Decision Support Systems) workloads.

### Simulated server workloads

#### OLTP DB Mixed: 8KB, 90% Read, 10% Write, 100% Random

#### Exchange Server Mixed: 32KB, 60% Read, 40% Write, 80% Random, 20% Sequential

## Performance Data using Monolithic Workloads

Comparison between SMB server and client side can tell us the efficiency of SMB protocol. Figure 16 and 17 compares for the maximum IOPS. Figure 18 and 19 compares the maximum bandwidth. Both figures show that SMB clients can achieve on-parity aggregated performance with the server.

### IOPS Comparison: 1.08Million IOPS (8KB 100% Random)

Figure 16. Read IOPS comparison between SMB servers and clients

Table 3 lists the CPU utilization in SOFS when running above peak IOPS workloads within VMs.

|  |  |
| --- | --- |
| Client1: 40% (64 LPs using Hyper-V LP Perf Counters) | File Server1: 60% (64 LPs) |
| Client2: 40% (64 LPs using Hyper-V LP Perf Counters) | File Server2: 60% (64 LPs) |

Table 3. CPU Utilization Comparison in SOFS for Peak IOPS

Figure 17. Write IOPS comparison between SMB servers and clients

### Bandwidth Comparison: 13.2GBps (512KB Sequential)

Figure 18. Read bandwidth comparison between SMB servers and clients

Figure 19. Write bandwidth comparison between SMB servers and clients

## Performance Data using Mixed Workloads

In reality, the server workloads always are consisted of different I/O patterns, sizes and access modes (read/writes). We use IOMeter to build and run simulated server workloads within VMs hosted in the SOFS cluster and the following tables show the mixed workload results using two clients.

### OLTP Database (8KB, 100% Random, 90% Reads, 10% Writes)

#### IOPS

|  |  |
| --- | --- |
| Aggregated | 504K IOPS + 513K IOPS = 1,017K IOPS |
| Client1 | 252K IOPS(VM1) + 252K IOPS(VM2) = 504K IOPS |
| Client2 | 256K IOPS(VM1) + 257K IOPS(VM2) = 513K IOPS |

#### Table 4. Aggregated OLTP IOPS from VMs

#### Bandwidth

|  |  |
| --- | --- |
| Aggregated | 3941MBps + 4010MBps = 7951MBps |
| Client1 | 1972MBps(VM1) + 1969MBps(VM2) =  3941MBps |
| Client2 | 2004MBps(VM1) + 2006Mbps(VM2) = 4010MBps |

*Table 5. Aggregated OLTP Bandwidth from VMs*

#### Latency

|  |  |
| --- | --- |
| Average | [4.06ms + 3.99ms] / 2 = 4ms |
| Client1 | [4.06ms(VM1) + 4.06ms(VM2)] / 2 = 4.06ms |
| Client2 | [3.99ms(VM1) + 3.99ms(VM2)] / 2 = 3.99ms |

*Table 6. Average OLTP Latency from VMs*

### Exchange Server (32KB, 80% Random, 20% Sequential, 60% Reads, 40% Writes)

#### IOPS

|  |  |
| --- | --- |
| Aggregated | 161K IOPS + 162K IOPS = 323K IOPS |
| Client1 | 80K IOPS(VM1) + 81K IOPS(VM2) = 161K IOPS |
| Client2 | 80K IOPS(VM1) + 82K IOPS(VM2) = 162K IOPS |

#### Table 7. Aggregated Exchange Server IOPS from VMs

#### Bandwidth

|  |  |
| --- | --- |
| Aggregated | 5047MBps + 5088MBps = 10.14GBps |
| Client1 | 2511MBps(VM1) + 2536MBps(VM2) = 5047MBps |
| Client2 | 2510MBps(VM1) + 2578MBps(VM2) = 5088MBps |

*Table 8. Aggregated Exchange Server Bandwidth from VMs*

#### Latency

|  |  |
| --- | --- |
| Average | [12.62ms + 12.54ms] / 2 = 12ms |
| Client1 | [12.67ms(VM1) + 12.58ms(VM2)] / 2 = 12.62ms |
| Client2 | [12.70ms(VM1) + 12.38ms(VM2)] / 2 = 12.54ms |

*Table 9. Average Exchange Server Latency from VMs*

# Conclusion

Windows Server 2012 R2 offers a wide variety of storage features and capabilities to address the storage and performance challenges in a virtualized infrastructure. This white paper demonstrates a high performance enterprise-class Scale-Out File Server platform all built with industry standard storage products. The workloads running on Hyper-V VMs can achieve over 1 million IOPS throughput and 100Gbps aggregate bandwidth.

The results presented in this white paper are by no means the upper limit for the I/O operations achievable through any of the components used for the tests. The intent of this white paper is to show the powerful capabilities of the storage and virtualization solutions in Windows Server 2012 R2. It provides customers with a comprehensive platform to address increasingly high demand for storage in modern datacenters.

# Reference

[1] Windows Storage Server Overview: <http://technet.microsoft.com/en-us/library/jj643303.aspx>

[2] Storage Spaces Overview: <http://technet.microsoft.com/en-us/library/hh831739.aspx>

[3] Multipath I/O Overview: <http://technet.microsoft.com/en-us/library/cc725907.aspx>

[4] Storage Quality of Service for Hyper-V: <http://technet.microsoft.com/en-us/library/dn282281.aspx>

[5] VHDX Format Specification: <http://www.microsoft.com/en-us/download/details.aspx?id=34750>

[6] Improve Performance of a File Server with SMB Direct: <http://technet.microsoft.com/en-us/library/jj134210.aspx>

[7] Failover Clustering Overview: <http://technet.microsoft.com/en-us/library/hh831579.aspx>

[8] Scale-Out File Server Overview: <http://technet.microsoft.com/en-us/library/hh831349.aspx>

# Acknowledgement

We want to thank the following people from Microsoft and each team behind them for their great help and support for this work:

* Hyper-V: Harini Parthasarathy, John Starks, Mike Ebersol, Jon Hagen, Mathew John, Jake Oshins, Patrick Lang, Attilio Mainetti, Taylor Brown
* Windows Fundamental: Brad Waters, Bruce Worthington, Jeff Fuller, Ahmed Talat, Tom Ootjers, Gaurav Bindlish
* File Server: Dan Lovinger, Jose Barreto, Greg Kramer, David Kruse
* Server Cluster: Elden Christensen, Claus Joergensen, Vladimir Petter
* Windows Storage: Calvin Chen, Karan Mehra, Bryan Matthew, Scott Lee, Michael Xing, Darren Moss, Matt Garson
* Networking: Sudheer Vaddi, Jeffrey Tippet, Don Stanwyck

The author would also like to thank our industry partners including Mellanox, LSI, HGST and DataOn Storage for providing their product samples to allow us to build the test infrastructure for the performance experiments discussed in this paper. The product pictures used in this report are provided by courtesy of them as well. Particularly we want to give our special thanks to the following people for their help:

* Mellanox: Motti Beck
* LSI: Thomas Hammond-Doel, Brad Besmer, Steve Hagan, Jerry Bass, Joe Koegler
* HGST: Swapna Yasarapu, Bill Katz, Craig Cooksey
* DataOn Storage: William Huang, Rocky Shek

1. See the Dell PowerEdge R820 Technical Guide, Appendix C Figure 14 “R820 system board block diagram” http://i.dell.com/sites/doccontent/shared-content/data-sheets/en/Documents/dell-poweredge-r820-technical-guide.pdf [↑](#footnote-ref-1)