Dell EqualLogic Storage with Heterogeneous Virtualized Workloads on Microsoft Windows Server 2012 with Hyper-V

A Dell EqualLogic Reference Architecture

Dell Storage Engineering
September 2013

Table of contents

Acknowledgements

This best practice white paper was produced by the following members of the Dell Storage team:

Engineering: Danilo Feroce

Technical Marketing: Omar Rawashdeh

Editing: Margaret Boeneke

Feedback

We encourage readers of this publication to provide feedback on the quality and usefulness of this information by sending an email to SISfeedback@Dell.com.

SISfeedback@Dell.com

Executive summary

Virtualization of the data center has introduced many opportunities to increase agility and efficiency compared to the traditional way of provisioning and managing line-of-business applications' resources and infrastructure services. Meanwhile, it has raised new challenges for IT professionals and storage architects when they face the consolidation of multiple types of workloads within a single storage unit. Managing workloads with different storage access patterns has traditionally been performed by compartmentalizing storage resources in different silos. Instead, predicting and addressing the behavior of mixed storage workloads should be part of the planning of virtualization solutions.

This paper investigates how a consolidated infrastructure supporting multiple applications and producing a heterogeneous storage workload could be optimized when deployed on a Microsoft Windows Server 2012 with Hyper-V virtual environment connected to a Dell EqualLogic storage area network (SAN). Some of the most relevant topics examined are:

- New techniques to configure the physical and virtual connections to the storage have created broader choices that savvy professionals must consider carefully.
- The introduction of highly available configurations and their operative tasks may introduce unpredictable storage activities that must be planned for and monitored.
- Planning for building block deployment where horizontal and vertical scaling of the solution is assessed to provide capacity for organizations of different sizes and types.

1 Introduction

Virtualization technologies have introduced more variables in the equation that IT professionals must solve when designing and implementing a solution. The intermediary role of the hypervisor between the hardware and the software layers creates an extensive set of choices that must be addressed.

The tests in this paper simulated a reference applications ecosystem designed to fulfill the requirements of a small to medium organization or a department in a large organization with a maximum of 1,000 actively operating users. The infrastructure supporting this ecosystem is virtualized by Windows Server 2012 with Hyper-V. The tests assessed the impact of different configurations: configurations using advanced networking features, configurations using different types of iSCSI initiators and collocation in the connectivity stack, various virtual machine (VM) distributions among storage volumes, and more. Furthermore this environment was validated while scaling it horizontally to increment the workload on a building block basis or while introducing highly available configurations and operations usually performed in those scenarios.

1.1 Purpose and scope

This paper is primarily intended for IT professionals (IT managers, SAN Architects, Applications and Services administrators, and System and virtualization Engineers) who are involved in defining, deploying, and/or managing Microsoft virtual infrastructures supporting heterogeneous applications and who would like to investigate the benefits of using EqualLogic storage. This document assumes the reader is familiar with Microsoft Windows Server 2012, EqualLogic SAN operation, and Microsoft Hyper-V architecture and system administration. The scope of this paper is restricted to a local datacenter topology and does not include specific or detailed server sizing information.

1.2 Terminology

The following definition list describes terms used throughout this document.

Group: Consists of one or more EqualLogic PS Series arrays connected to an IP network that work together to provide SAN resources to host servers.

Member: Identifies a single physical EqualLogic array.

Pool: A logical collection that each member (array) is assigned to after being added to a group. The array contributes its storage space to the entire pool.

Hypervisor: Denotes the software layer that manages the access to the hardware resources, which resides above the hardware and in between the operating systems running as guests.

Parent and Child Partitions: Identifies the logical units of isolation supported by the Hyper-V hypervisor. The parent, or root, partition hosts the hypervisor itself and spawns the child partitions where the guest VMs reside.

Virtual Machine (VM): An operating system implemented on a software representation of hardware resources such as processor, memory, storage, and network. VMs are usually identified as guests in relation to the host operating system that executes the processes to allow the VMs to run over an abstraction layer of the hardware.

Synthetic drivers: Supported with the Hyper-V technology, these drivers leverage more efficient communications and data transfer between the virtual and physical hardware, as opposed to legacy emulated drivers. Synthetic drivers are only supported in newer operating system versions and allow the guest VMs to become enlightened, or aware that they run in a virtual environment.

Virtual network: A network consisting of virtual links as opposed to wired or wireless connections between computing devices. A virtual network is a software implementation similar to a physical switch, but with different limitations. Microsoft Hyper-V technology implements three types of virtual networks: connectivity between VMs and devices external to the root host (external), connectivity between the VMs and the host only (internal), or the VMs running on a specific host only (private).

VHDX: File format for a Virtual Hard Disk in a Windows Hyper-V 2012 hypervisor environment.

Non-Uniform Memory Access (NUMA): A multiprocessing architecture in which memory is associated with each CPU (the local memory, accessed by channels dedicated to each processor) as opposed to symmetric multiprocessing (SMP) where memory is shared between CPUs.

Process: An instance of a computer program or application that is being executed. It owns a set of private resources: image or code, memory, handles, security attributes, states, and threads.

Thread: A separate line of execution inside a process with access to the data and resources of the parent process. It is also the smallest unit of instructions executable by an operating system scheduler.

Key performance indicators (KPI): A set of quantifiable measures or criteria used to define the level of success of a particular activity.

Globally Unique Identifier (GUID): A unique 128-bit number generated by a Windows OS or by some Windows applications and used as the identifier for a component, an application, a file, a database entry, or a user. A GUID is commonly displayed as a 32 digit hexadecimal string.

2 Technology overview

2.1 Dell EqualLogic PS6110

Dell EqualLogic PS6110 Series arrays are designed to meet the performance and availability needs of application and virtualization environments in medium to large enterprises. These virtualized iSCSI SANs combine intelligence and automation with fault tolerance to provide simplified administration, rapid deployment, enterprise performance and reliability, and seamless scalability using innovative Fluid Data™ technology. For example, your configuration can quickly process data hungry tasks using EqualLogic PS6110XV high-performance 10GbE iSCSI arrays with easy, flexible management for highly consolidated enterprise environments.

EqualLogic PS6110 arrays automate performance and network load balancing to optimize resources — offering all-inclusive array management software, host software, and free firmware updates. By default, Offloaded Data Transfer (ODX) is enabled for any Windows Server 2012 OS, including clustered servers, Hyper-V virtual machines, and physical hosts. ODX off-loads large data transfer activities to the intelligence of the storage arrays.

EqualLogic PS Series arrays recently introduced the addition of System Center Virtual Machine Manager (SCVMM) 2012 SP1 storage management support into the Host Integration Tools for Microsoft 4.6 (HIT/Microsoft 4.6), which provides:

- Reduced complexity for storage management in infrastructures using Microsoft Windows Server and Hyper-V
- Sophisticated optimization that enables you to rapidly provision virtual environments for the Enterprise
- Automated ease with PowerShell cmdlets so you can easily automate repetitive tasks

2.2 Microsoft Windows Server 2012 with Hyper-V

Microsoft Hyper-V technology is implemented as a software layer, defined hypervisor residing above the hardware layer and in between the operating systems running as guests. It primarily manages the access from the virtualized computing environment to the underlying hardware resources (for example, processor, memory, storage, and network) and the isolation between the several VMs that it can host.

Hyper-V provides the traditional benefits of virtualization technologies, such as:

- Increasing the hardware utilization, by consolidating multiple workloads into less hardware, thus reducing the cost of the computing infrastructure and the overall power consumption footprint
- Enhancing the ability to provision new development, test, or production infrastructures and streamlining the efficiency to migrate and transform from one to another
- Improving the availability and resiliency of services by the simplification of resource displacement and the manageability of physical and virtual resources by tailored management tools

Since its introduction, Hyper-V has matured across three generations (Windows Server 2008, Windows Server 2008 R2, and then Windows Server 2012), achieving a broad set of features and enhancements with the current version. Some significant enhancements are shown in the list below:

- Host scalability: up to 320 logical processors, up to 4TB of RAM, and up to 1,024 guest VMs per host
- Guest scalability: up to 64 logical processors each, up to 1TB of RAM each, VHDX file format up to 64TB, and guest NUMA
- Highly available configuration scalability: up to 64 nodes per cluster, and up to 8,000 VMs per cluster
- Highly available VM features: Live Storage migration, and multiple Live Migration in parallel and across clusters
- Storage features: ODX, guest VMs on file storage (SMB 3.0), and 4KB disk support
- Network features: single-root I/O virtualization (SR-IOV), virtual switch extensions, dynamic VMQ, and host Data Center Bridging (DCB)

Hyper-V technology is installed in Windows Server 2012 as a role and provides management tools (both GUI-based and PowerShell cmdlets), a management service, virtual machine bus (VMbus), virtualization service provider (VSP), and virtualization infrastructure drivers (VID).

Hyper-V includes a software package (Integration Services) for the guest operating systems built with the aim of improving the integration between the physical and virtual layers. A guest Windows Server 2012 operating system does not currently require the installation or update of the integration services.

2.2.1 Virtual storage devices, disk formats, and disk types

Each guest VM running on Windows Server 2012 with Hyper-V has access to storage by one or more of the following devices:

- Direct attached virtual storage by virtual IDE or virtual SCSI controllers. The storage must be available to the host hypervisor first in the form of files, local hard drives, or volumes/LUNs mapped from a SAN.
- SAN attached network storage (for example, iSCSI targets) by virtual network adapters, maximum of 12, of which eight are VMbus network adapters and four are legacy network adapters
- SAN attached storage by virtual Fibre Channel adapters (maximum of four)

A definition and detailed explanation of these devices is:

Virtual IDE controller is an emulated IDE disk controller (maximum of two) with the support of a maximum of four virtual hard drives (two virtual hard drives for each controller). One of the virtual devices often connected to this kind of controller is a CD/DVD interface to mount a local optical drive or an ISO images to the VM, leaving three additional slots for virtual hard drives.

Virtual SCSI controller is an emulated SCSI disk controller (maximum of four) that supports a maximum of 256 hard drives (64 hard drives for each controller).

Boot from virtual disk (startup disk) in a Hyper-V VM requires an IDE device. SCSI devices are currently unsupported for this specific function.

Regardless how a virtual hard drive is attached to a VM (by an IDE or a SCSI controller), the media or formats for the Hyper-V virtual disks are as follows:

VHD (Virtual Hard Disk) is a file format that represents a hard disk image. A VHD file is composed of sectors of 512 bytes each, and addressed by a 32-bit table which allows a maximum addressable size of 2TB (or 2040GB). VHD format is supported by all three generations of Microsoft Hyper-V technologies since Windows Server 2008, as well as other virtualization platforms. VHDs can only be mounted on NTFS/ReFS volumes (not FAT/FAT32), and should not be placed within a compressed folder or volume.

VHDX (Virtual Hard Disk eXtended) is the VHD enhanced file format representing a hard disk image, and is supported only on the latest generation of Microsoft Hyper-V in Windows Server 2012. VHDX format supports storage capacity up to 64TB by using 4KB sectors and provides protection against data corruption during power failure by logging changes in its own metadata structures. VHDX also supports reclaiming unused space (“unmap/trim”) when working in combination with compatible hardware and provides better disk alignment with an increased offset of 1MB (from 512Kb).

Pass-Through Disk is a host volume presented directly to the VM. It can be a local hard drive/partition or a volume/LUN first mapped to the host from a SAN, and its size is limited only by the guest VM operating system limits. The VM requires exclusive access to the target volume/LUN while mapped through a pass-through disk. Pass-through disks do not support advanced resiliency features (Live Migration and/or Live Storage Migration), and cannot be used as targets for the Hyper-V VSS provider in order to back up or snapshot the data contained by them.

When a hard drive attached to a VM is a VHD or VHDX format, a further selection of the disk type is allowed. The virtual hard disk types and their primary characteristics are:

Fixed Disks are created with the entire designated space pre-allocated upfront. Their size is kept constant regardless of the amount of adding/deleting data activities that are performed on them. In the case of large capacity files, provisioning or moving the files can be time consuming. Fixed disks are less prone to fragmentation and provide better performance for a high level of disk activity. Fixed disks can be expanded to a larger size or converted to a different format and/or type.

Dynamically Expanding Disks are created small initially and grow as data is added. They allow over-provisioning of the host's volume capacity. Expanding disks are immediately available for use without delay at the time of provisioning, like large fixed disks. This disk type is specifically subject to high fragmentation that could cause slower performance due to increased latency during read activities or under a high rate of writes due to the automatic expansion of the disk. Expanding disks can be compacted to reclaim space or expanded to a larger maximum size, as well as converted to a different format and/or type.

Differencing Disks are intended for the specific purpose of parent-child relationship with another disk that should stay intact, as a sort of original or shared image. Parent-child files must share the same format (VHD or VHDX). The differencing disks could have a short life and are well suited for pooled virtual desktop infrastructure (VDI) VMs or for test laboratories. Hyper-V based snapshots of a guest VM take advantage of this disk type. Differencing disks can be compacted to reclaim space, expanded to a larger maximum size, merged with their respective parent disk (or discarded to roll back the changes), as well as converted to a different format and/or type.

3 Solution architecture and workload overview

The solution presented and evaluated in this paper is built upon a virtual infrastructure supported by Microsoft Windows Server 2012 with Hyper-V technology and a back-end iSCSI SAN provisioned by Dell EqualLogic Storage. The operating system of the group of VMs simulating the applications workload is a combination of Windows Server 2008 R2 and Windows Server 2012 with the intent to represent a broad mix of workload sources for storage access, which can have slightly different storage access patterns.

3.1 Conceptual systems design

The elements of the infrastructure supporting the simulation and their primary relationships and connectivity links are represented in the conceptual diagram below (Figure 1).

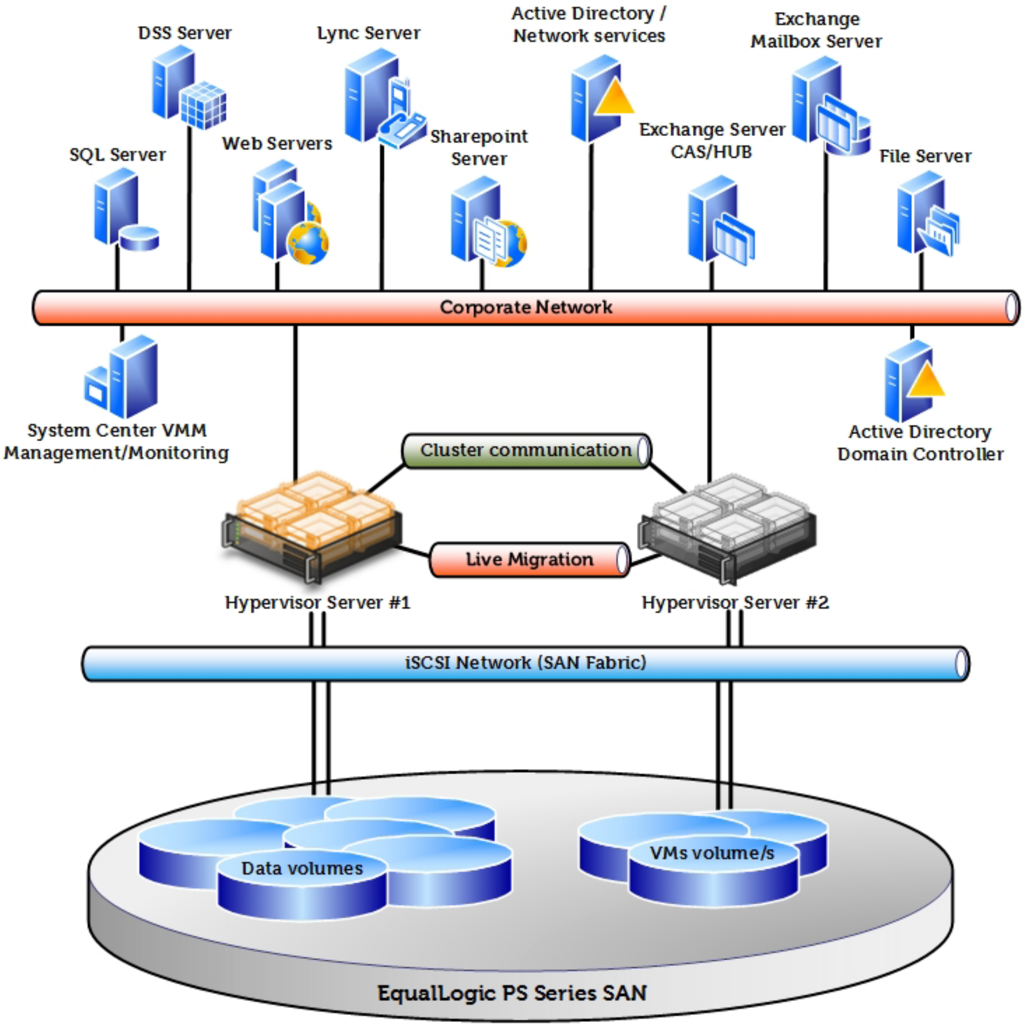


Figure 1 Conceptual systems design of the heterogeneous workload virtual environment

The key elements of this design are:

- Single Active Directory forest, single domain, single site
- Centralized management and monitoring with dedicated resources (both physical and virtual)
- Separated network design to maintain traffic isolation between traditional LAN and iSCSI access
- VM resources appropriately sized to match the related functional role of the simulated application
- Tailored workload profile for each VM to match the related functional role of the simulated application
- Hypervisor configured in a standalone server topology first, then in a dual node clustered configuration

3.2 Physical systems design

The physical components and the essential connections beneath the virtual infrastructure simulating the heterogeneous workload are shown in Figure 2.

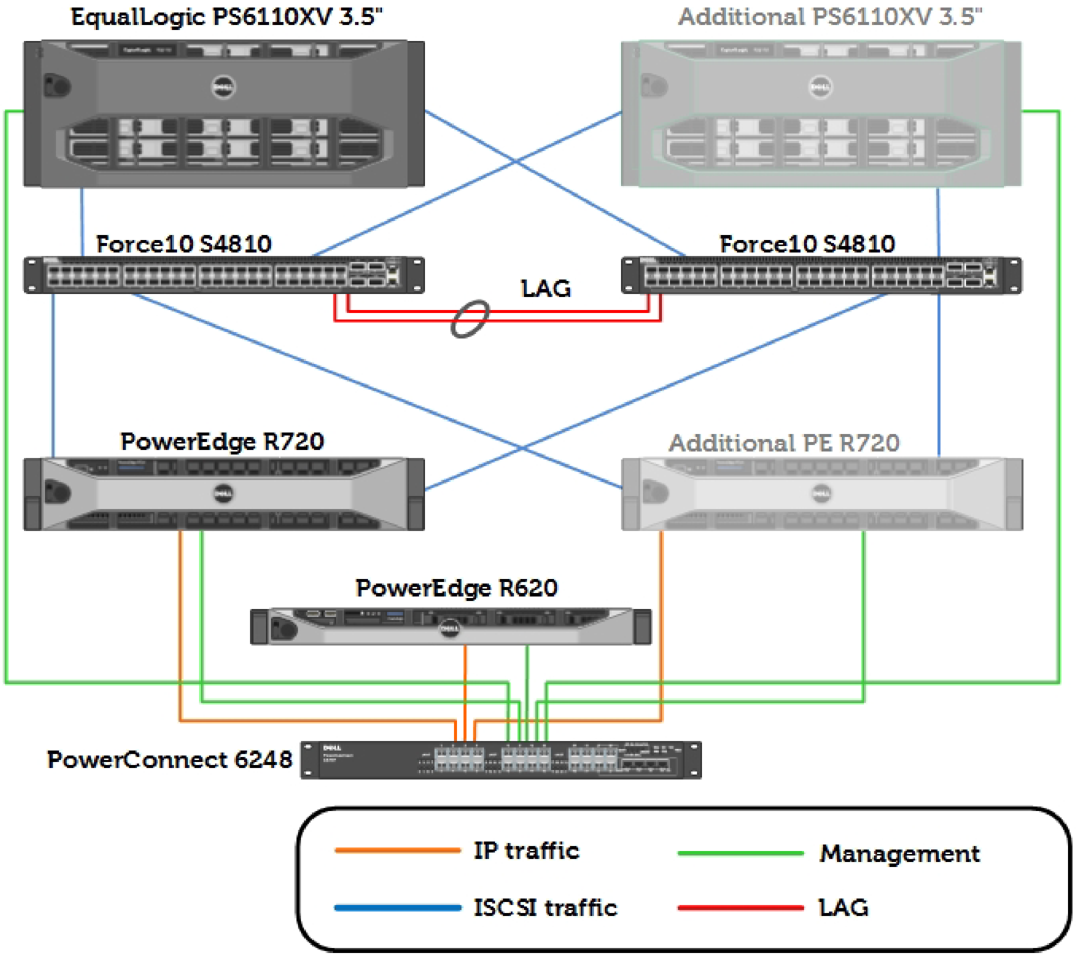


Figure 2 Physical systems design of the heterogeneous workload virtual environment

The components shown in the picture include those used for the single server scenarios and the components shown grayed out include an additional PowerEdge server required for the failover cluster scenario and the additional array unit required for the scale out exercise.

The solution architecture is deployed on Dell rack servers with top of rack (ToR) Ethernet network switches separately dedicated to IP traffic (traditional client/server, management, and hypervisor communications) and to iSCSI storage access. The hardware elements contributing in the architecture are:

- Two PowerEdge R720 rack servers to power the hypervisors beneath the simulated application infrastructure (standalone configuration first, then dual nodes cluster configurations)
- One PowerEdge R620 rack server to power the hypervisor beneath the centralized management and monitoring infrastructure
- One EqualLogic iSCSI SAN provisioned with one or two PS6110XV 3.5" arrays (10GbE)
- One PowerConnect 6248 Ethernet switch to support LAN IP traffic
- Two Force10 S4810 Ethernet switches to support the iSCSI data storage traffic on the SAN side
- Link Aggregation Group (LAG) consisting of two connections between the ToR S4810 switches

More details of the configurations used for the solution infrastructure, including a hardware and software list, SAN array characteristics, hypervisor and VMs relationship, and physical and virtual network connections are provided in Appendix A.

3.3 Storage and volumes layout

The EqualLogic SAN arrays and the volumes underlying the VMs and application data are distributed as:

- Standalone Hyper-V configuration
  - One EqualLogic group configured with one PS6110XV 3.5" unit
  - One storage pool defined within the group that includes the single member
  - One volume to host the file images and temporary files for all VMs, unless otherwise specified during the particular simulation
  - A dedicated set of data volumes dedicated to the databases or files of each application (Exchange, SQL Server, DSS, and File services), uniquely assigned to the relevant VM hosting the simulation for the specific application
- Scale out exercise with a storage building block
  - One EqualLogic group configured with two PS6110XV 3.5" array members
  - Two storage pools defined with one member dedicated to each pool
  - Two volumes, one created in each pool, to host the file images and temporary files for all VMs
  - Two set of data volumes, one created in each pool, dedicated to the correspondent VM hosting the simulation of the specific application, as mentioned above
- RAID 6 policy applied as a reference configuration, unless otherwise specified during the particular simulation

Figure 3 shows all the volumes defined on the EqualLogic SAN within the respective pool.

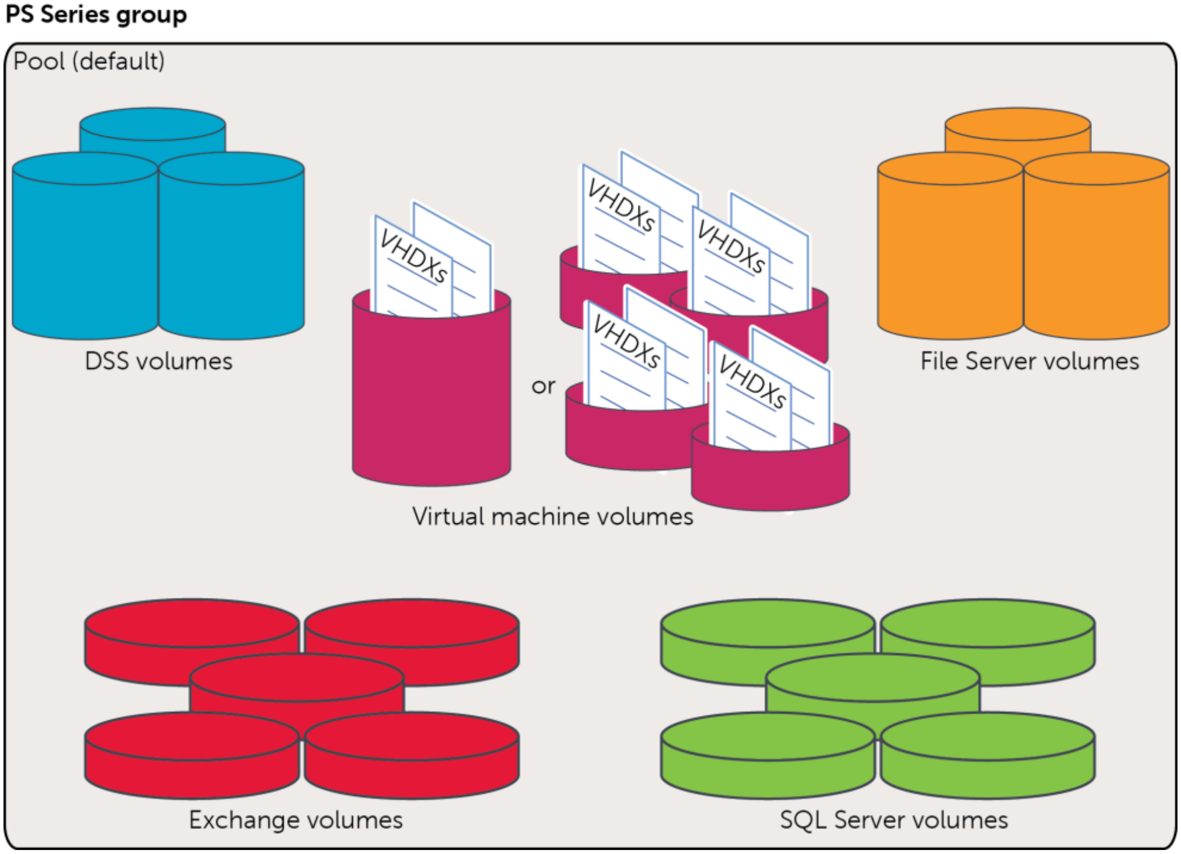


Figure 3 Storage and volumes layout

3.4 Heterogeneous workload combination

The heterogeneous workload simulation is the result of a set of application profile loads applied from each VM. A single VM assumes the specific role attributed to it and performs the amount of storage access defined by the related storage access profile.

This set of VMs can be further ranked in two different classes:

1. VM roles in a workload category with limited storage access footprint and no extra data volumes required. Usually these services are focused on quick data exchange in client-server environments (n-tiers) requiring mostly computational resources (processor, memory, and network), but with limited or less significant store capacity requirements. The workload categories of this class are:
   a. Network systems: infrastructure and/or network services such as Domain Controller, DNS, DHCP, WINS, and Radius
   b. Messaging (front-end): client access or messaging routing roles such as Microsoft Exchange HUB and Client Access Role
   c. Web server: static or active pages services including middle-tier web based components

   d. Content management server: web based documents and content management services with an external repository, usually connected to remotely attached databases or file services such as Microsoft SharePoint Portal Server
   e. Communication server: application services providing point to point or gateway functions to enable instant communications or routing toward voice networks such as Microsoft Lync/OCS

2. VM roles in a workload category with a predictable heavy footprint on storage resources and with a complex set of assigned volumes to layout the application data or files. These services are expected to have a significant computational demand (processor, memory, and network), as well as a large storage and retrieval capacity requirement. The workload categories of this class are:
   a. Messaging (back-end, mailboxes store): messaging store and retrieval role for example, the Microsoft Exchange Mailbox role
   b. Relational databases (OLTP): relational database system serving transaction oriented applications, such as Microsoft SQL Server
   c. Decision Support System (reporting): database system supporting decision making reports and queries, such as Microsoft SQL Server
   d. File services (file sharing repositories): client-server service providing shared files and folders capabilities, such as CIFS network shares

Table 1 summarizes the number of VMs, the workload category and role of each VM, and the storage profile associated with it.

Table 1 Workload category for each VM role and storage profiles

#VMs	Workload category	Storage access profile
1	Network systems	Small I/O size (*) Random I/O 80%: Read 80%, Write 20% Sequential I/O 20%: Read 80%, Write 20%
1	Messaging (front-end, connectivity)	Medium-Small I/O size (*) Random I/O 80%: Read 60-80%, Write 20-40% Sequential I/O 20%: Read 60-80%, Write 20-40%
1	Messaging (back-end, mailboxes store)	Medium-Small I/O size (*) Random I/O 80%: Read 50%, Write 50% Sequential I/O 20%: Read 50%, Write 50%
1	OLTP relational databases (transactional)	Small I/O size (*) Random I/O 100%: Read 66%, Write 34%
1	Decision Support System (reporting)	Large I/O size (*) Sequential I/O 100%: Read 95%, Write 5%
1	File services (CIFS home folders sharing)	Medium I/O size (*) Random I/O 100%: Read 70%, Write 30%
1	Content management services	Medium-Small I/O size (*) Random I/O 75%: Read 95%, Write 5% Sequential I/O 25%: Read 95%, Write 5%
1	Communication services	Medium-Small I/O size (*) Random I/O 75%: Read 95%, Write 5% Sequential I/O 25%: Read 95%, Write 5%
2	Web Server services	Medium-Small I/O size (*) Random I/O 75%: Read 95%, Write 5% Sequential I/O 25%: Read 95%, Write 5%

* The I/O size ranges are considered as follows: around 4-16KB is small, around 32-64KB is medium, and over 128KB is large.

4 Standalone Hyper-V implementation factors

The following study focuses on the fundamental configuration and deployment variables that can affect the storage performances of the virtual infrastructure and applications described in Section 3. The test approach is to only alter one variable for each test scenario, to better assess the potential value of this specific configuration.

Assessing the standalone hypervisor configurations provides the baseline for a more complex configuration when scaling this building block horizontally into a failover cluster with multiple nodes. Additional scenarios unique to cluster configurations are examined in Section 5 of this paper.

A summary of the primary constraints and factors kept constant across the tests is listed below:

- 10 VMs started with pre-allocated resources and running the simulation tools
- The startup disks of all VMs were connected to the respective virtual IDE controller, using a VHDX format file and with a fixed disk drive type
- All extra data volumes were connected by guest iSCSI initiator, unless otherwise specified for the test
- All VMs were configured without Dynamic memory, thus allocating the entire memory at the time of startup
- Heterogeneous storage workload created from the aggregation of the customized simulation running in each VM, as defined previously
- Additional simulation within each VM for processor utilization, requesting a constant 30% overall utilization through two threads configured with customized affinity to each virtual processor
- Additional simulation for memory allocation within each VM, requiring the allocation of at least 70% of available memory to multiple application instances (1GB memory for each instance)

Below is a list of metrics and pass/fail criteria recorded while completing the tests. Most of this information is captured by Windows Performance Counters and Dell EqualLogic SAN Headquarters or is collected by the simulation tools (for example, Iometer, Jetstress, and FSCT) and outlined in their reports. Microsoft data around thresholds for storage validation are reported as well.

Average Read Latency (ms) is the average length in time to wait for a disk read operation. It should be less than 20 ms according to Microsoft threshold criteria.

Average Write Latency (ms) is the average length in time to wait for a disk write operation. It should be less than 20 ms according to Microsoft threshold criteria (depending on the application).

Disk Reads per second is the total number of individual read requests retrieved from the disk over a period of one second (or the average of the values if the sample period is greater than one second).

Disk Writes per second is the total number of individual write requests sent to the disk over a period of one second (or the average of the values if the sample period is greater than one second).

Total IOPS is the rate of disk transfer (read and write operations) performed by the storage subsystem. It is measured and validated at both the host and SAN levels.

Average Disk Read Queue Length is the average number of read requests that were queued for service during the sample interval (outstanding reads). It should not exceed twice the number of disk spindles for continuous periods of time.

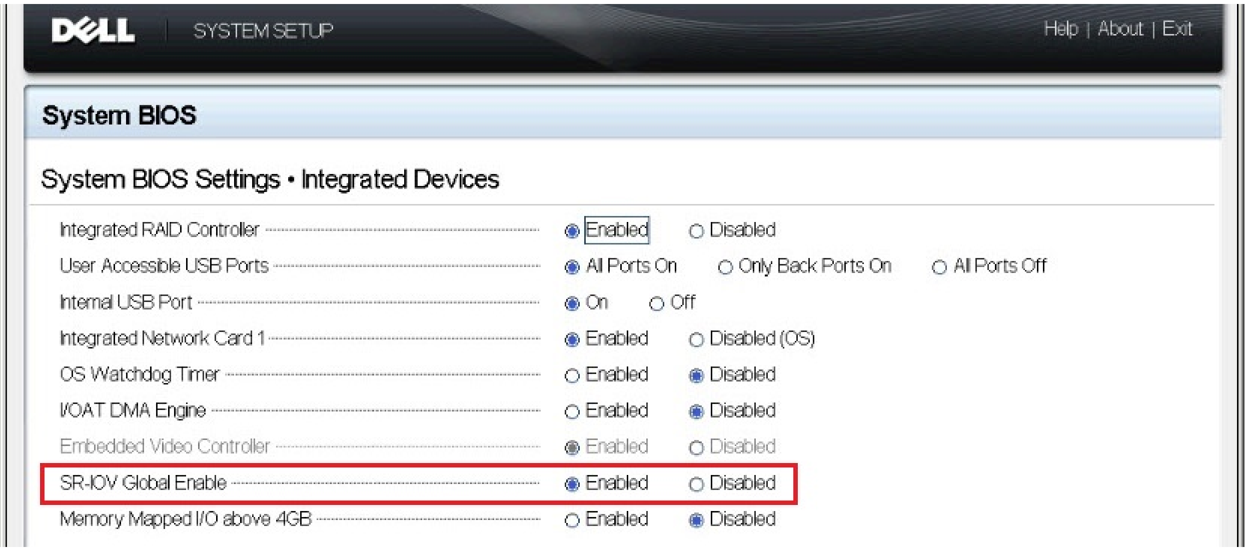
Average Disk Write Queue Length is the average number of write requests that were queued for service during the sample interval (outstanding writes). It should not exceed twice the number of disk spindles for continuous periods of time.

4.1 Advanced network technology features

The introduction of Windows Server 2012 offers additional and updated features to the VM networks in Hyper-V. Virtual network adapters now support SR-IOV and expanded VMQ support management. VMs with high network traffic requirements take advantage of these standards.

Single Root I/O Virtualization (SR-IOV) is a standard introduced by the PCI SIG consortium. The specification provides remapping abilities for interrupts and DMA right into the chipset, and allows multiple operating systems running simultaneously within the same computer to share PCI Express devices. Hyper-V in Windows Server 2012 enables SR-IOV capable network devices to be shared natively and assigned to VMs.

The SR-IOV functionality must first be enabled on the server BIOS, then verified on the host operating system. Figure 4 shows the configuration activated in a 12th generation PowerEdge server, while Figure 5 reports the output of a Windows PowerShell verification cmdlets run on the operating system.

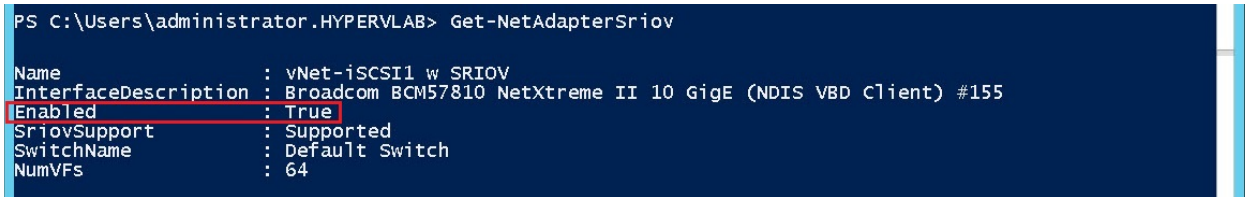


System BIOS

System BIOS Settings • Integrated Devices

Integrated RAID Controller ● Enabled ○ Disabled
User Accessible USB Ports ● All Ports On ○ Only Back Ports On ○ All Ports Off
Internal USB Port ● On ○ Off
Integrated Network Card 1 ● Enabled ○ Disabled (OS)
OS Watchdog Timer ○ Enabled ● Disabled
I/OAT DMA Engine ○ Enabled ● Disabled
Embedded Video Controller ● Enabled ○ Disabled
SR-IOV Global Enable ● Enabled ○ Disabled
Memory Mapped I/O above 4GB ○ Enabled ● Disabled

Figure 4 Dell 12th generation PowerEdge server: SR-IOV enable

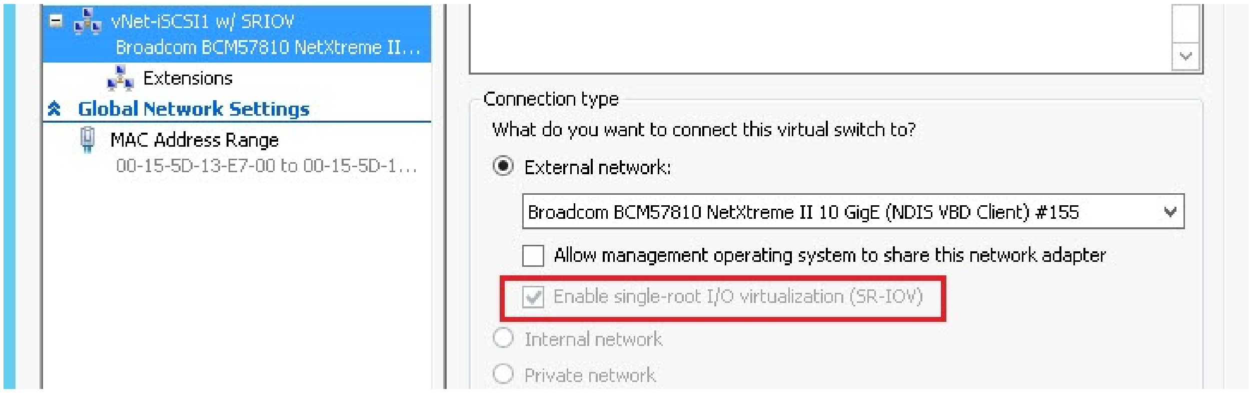


```
PS C:\Users\administrator.HYPERVLAB> Get-NetAdapterSriov

Name                 : vNet-iSCSI1 w SRIOV
InterfaceDescription : Broadcom BCM57810 NetXtreme II 10 GigE (NDIS VBD Client) #155
Enabled              : True
SriovSupport         : Supported
SwitchName           : Default Switch
NumVFs               : 64
```

Figure 5 Windows PowerShell : SR-IOV enable

The subsequent steps are to configure the Hyper-V virtual switch connected to the physical network adapter to enable SR-IOV, and then configure each virtual adapter that should use the SR-IOV “hardware acceleration” feature. Figure 6 and Figure 7 illustrate these two settings. A specific caveat for the virtual switch configuration is that Hyper-V currently supports the SR-IOV activation only at the time of the switch creation (which causes the checkbox to be grayed out after the virtual switch has been created).



Connection type

What do you want to connect this virtual switch to?

● External network:

Broadcom BCM57810 NetXtreme II 10 GigE (NDIS VBD Client) #155

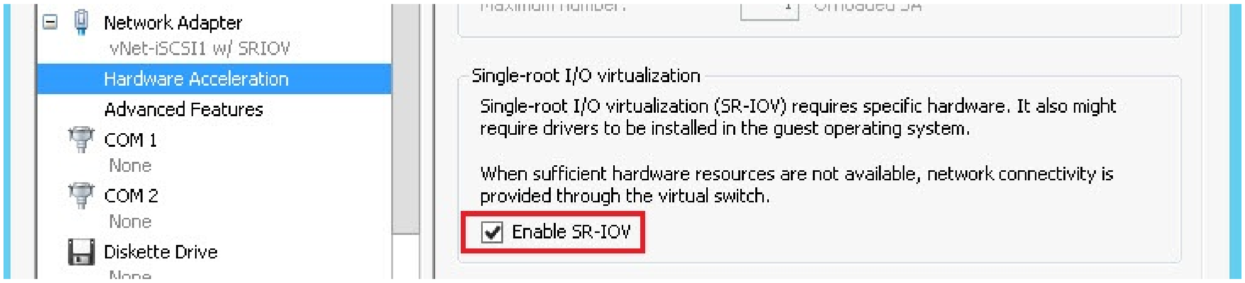
□ Allow management operating system to share this network adapter

☑ Enable single-root I/O virtualization (SR-IOV)

○ Internal network

○ Private network

Figure 6 Hyper-V Virtual Switch Manager: virtual switch with SR-IOV enable



Single-root I/O virtualization

Single-root I/O virtualization (SR-IOV) requires specific hardware. It also might require drivers to be installed in the guest operating system.

When sufficient hardware resources are not available, network connectivity is provided through the virtual switch.

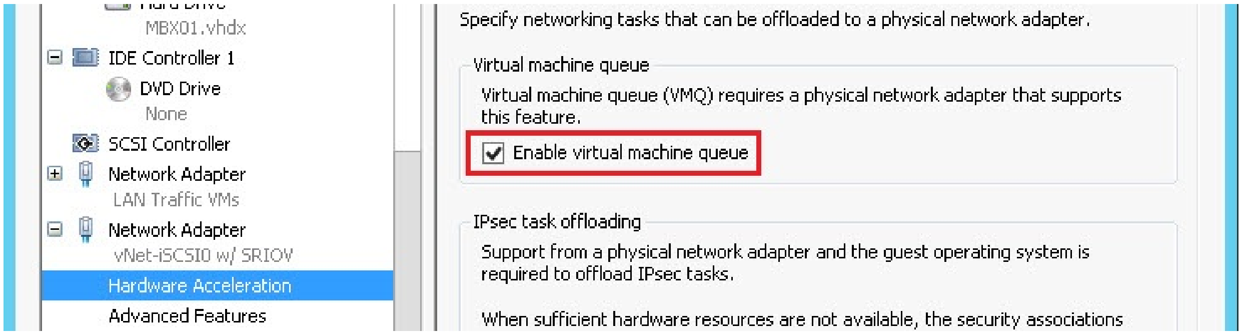
☑ Enable SR-IOV

Figure 7 Hyper-V VM settings: Hardware Acceleration SR-IOV enable

Virtual Machine Queue (VMQ) interface is introduced in Network Driver Interface Specification (NDIS) 6.2 to increase the Hyper-V network efficiency under heavy load. VMQ is a hardware technology that uses DMA to redirect all incoming network frames to a set of queues and their associated receive buffers, organized on a per queue basis. The sorting operations classify the frames using the destination MAC address and are executed from multiple processors. VMs have access to their private queues in order to receive the network frames.

Dynamic Virtual Machine Queue (D-VMQ) is introduced for Hyper-V in Windows Server 2012 and provides dynamic distribution of the frames' sorting operations among the available processors. D-VMQ only takes advantage of the physical cores of the processors; it ignores hyper-threaded cores when Hyper-Threading technology is available and enabled.

Similarly to the previous configuration of SR-IOV, VMQ capability must first be enabled at the physical network adapter layer (on the advanced properties of the Broadcom driver in the test setup), then must be enabled on each virtual network adapter connected to the virtual switch using the corresponding physical network adapter. The VM owner of this virtual network adapter will then be automatically assigned a relevant receiving queue. Figure 8 shows the virtual network adapter setting.



Specify networking tasks that can be offloaded to a physical network adapter.

Virtual machine queue

Virtual machine queue (VMQ) requires a physical network adapter that supports this feature.

☑ Enable virtual machine queue

IPsec task offloading

Support from a physical network adapter and the guest operating system is required to offload IPsec tasks.

When sufficient hardware resources are not available, the security associations

Figure 8 Hyper-V VM settings: Hardware Acceleration Virtual Machine Queue enable

The intent for the advanced networking features exercise is to assess any performance divergence when using these features for the network adapters connecting the virtualized infrastructure to the EqualLogic SAN.

Windows Server 2012 natively supports a new feature for NIC teaming called NIC Load Balancing and Failover (LBFO). Dell recommends using the Host Integration Tool kit (HIT Kit) and MPIO for any NIC that is connected to EqualLogic iSCSI storage. The use of LBFO is not recommended for NICs dedicated to SAN connectivity since it does not add any benefit over MPIO. Microsoft LBFO or other NIC vendor specific teaming can still be used for non-SAN connected interfaces. If DCB is enabled in a converged network, LBFO can also be used for NIC partitions that are not SAN connected.

Table 2 shows a summary of the configurations used for these tests.

Table 2 Test parameters: advanced network features

Reference configuration: test variables under study		
Advanced Network Feature in use	Baseline (no advanced features)	
	SR-IOV (for VMs with heavy storage footprint)	
	VMQ (for VMs with heavy storage footprint)	
	SR-IOV and VMQ (for VMs with heavy storage footprint)	
Reference configuration: consistent factors across the iterations of this test		
RAID policy (SAN)	RAID 6	
iSCSI Initiator	Host software initiator, EQL HitKit (VMs startup VHDX volume)	
	Guest software initiator, EQL HitKit (Data volumes)	
Virtual networks (iSCSI)	Two virtual switches, each created on one dedicated 10GbE port	
	Two virtual adapters for VMs with guest initiator, each adapter connected separately to a virtual switch	
VMs with limited storage footprint (*)	Network services (1) Messaging services, Front-End (1) Content Management services (1) Communication services (1) Web server services(2)	Total of 6
VMs with heavy storage footprint (*)	Messaging services, Back End (1)	Total of 4
	OLTP relational Databases (1)	
	Decision Support System (1)	
	File services, CIFS network share (1)	
Ratio of VMs residing in one SAN volume		10:1

* Limited storage footprint refers to no additional data volumes. Heavy storage footprint refers to the situation where more data volumes are required for complex applications and workloads. Refer to Section 3.4 for details.

Figure 9 illustrates the KPIs collected when implementing different combinations of advanced networking features in the virtual switches and network adapters.

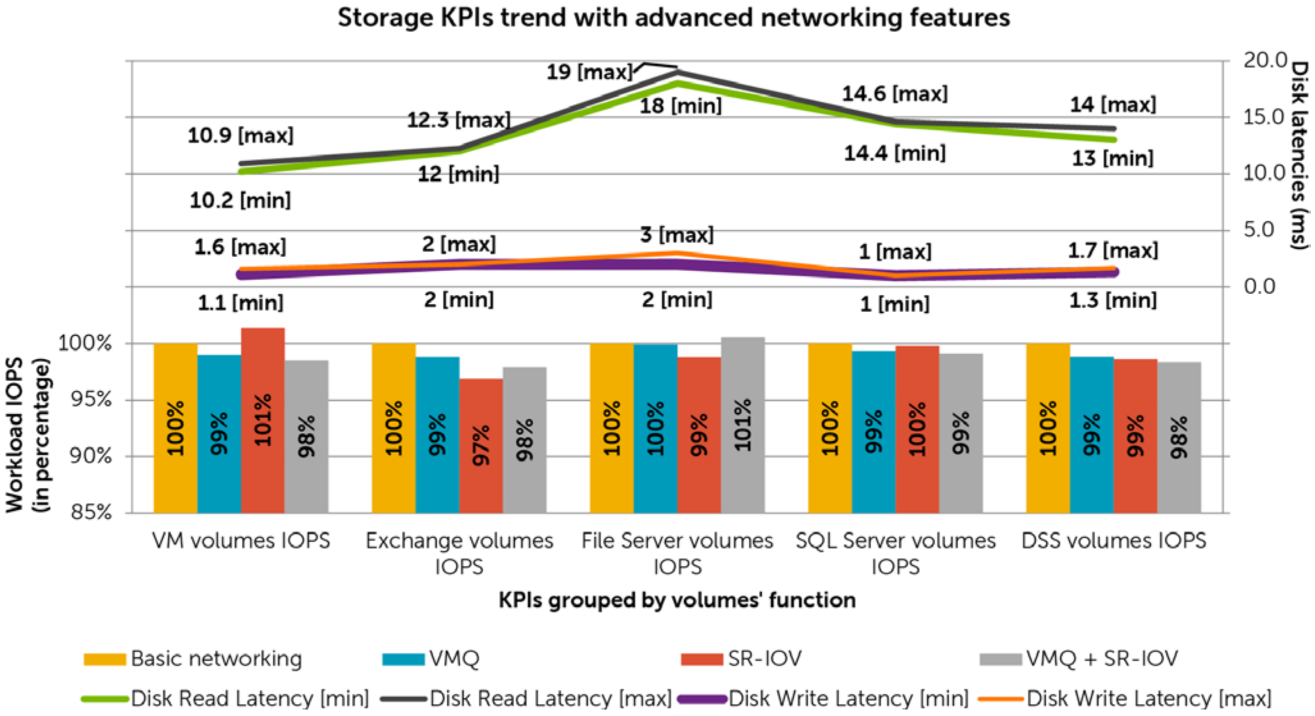


Figure 9 Advanced network features: performance result comparison chart

The results collected do not show any extensive gap of performance between any of the configurations tested. Some workload profiles provide better results in one case or in another. The average combination constantly showing the best disk latencies is the SR-IOV only. When combining SR-IOV and VMQ a slight decrease appears since the hardware resources provided by the physical network adapter are shared between these two technologies. Overall the workload does not test the boundaries of the resources, and suggests that more advantages would be noticed when the workload becomes critically heavy.

SR-IOV and VMQ are technologies built to improve network performance of the VMs and to reduce the computing overhead on the hypervisor hosting them. When both interfaces are enabled, the miniport driver allows the SR-IOV virtual ports to be used for both functions instead of generating dedicated VM queues. Furthermore, the Receive Side Scaling (RSS) interface, used for a similar purpose for the native traffic and enabled by default in the operating system, ceases to work when SR-IOV or VMQ is enabled.

For additional information on handling SR-IOV, VMQ, and RSS, refer to Microsoft Dev Center documentation at http://msdn.microsoft.com/en-us/library/windows/hardware/hh451362%28v=vs.85%29.aspx

4.2 RAID policy considerations

The EqualLogic PS Series 6110 arrays provide a range of different RAID policies and the ability to use spare drives to protect data. Each RAID level offers a distinct set of performance and availability characteristics dependent on the nature of the RAID policy and the workload applied.

- RAID 6, one or more dual parity sets
  - The highest availability at the expense of random writes performance and rebuild impact
  - A heavy impact for RAID reconstruction in the case of drive failure
  - A total of 23 available drives on PS6110, with one spare drive
- RAID 50, or striping over multiple distributed parity sets (RAID 5)
  - A balance between performance and capacity
  - A moderate impact for RAID reconstruction in case of drive failure
  - A total of 22 available drives on PS6110, with two spare drives
- RAID 10, or striping over multiple mirrored sets (RAID 1)
  - The best performance for random access at the expense of the average capacity available
  - A minimal impact for RAID reconstruction in case of drive failure
  - A total of 22 available drives on PS6110, with two spare drives
- RAID 5, while still supported, is not advised for critical data due to the limited level of resiliency.

Note: RAID implementations without spare drives for protection, even if possible on the EqualLogic PS Series, are not recommended and are not considered in this study.

EqualLogic PS6110 arrays also support RAID policy conversion, which allows you to switch from one RAID level to another without disruption of the data present on the volumes, although the performance during the conversion might be degraded. The following rules apply for RAID conversion:

- Conversion from RAID 10 to RAID 50 or RAID 6 is supported
- Conversion from RAID 50 to RAID 6 is supported
- Conversion from RAID 6 or RAID 5 to any other RAID type is not supported

The two RAID policies selected for the analysis are RAID 6 and 10. The array is reinitialized to the factory level for each RAID level run, the volumes are rebuilt, and the data is redeployed. The goal is to establish the performance differential and assess the pros and cons when implementing different RAID policies on the EqualLogic SAN with the specific workload in use.

Table 3 shows a summary of the variables and parameters used for these tests.

Table 3 Test parameters: RAID policies

Reference configuration: test variables under study		
RAID policy	RAID 6	
	RAID 10	
Reference configuration: consistent factors across the iterations of this test		
iSCSI Initiator	Host software initiator, EQL HitKit (VMs startup VHDX volume)	
	Guest software initiator, EQL HitKit (Data volumes)	
Virtual networks (iSCSI)	Two virtual switches, each created on one dedicated 10GbE port	
	Two virtual adapters for VMs with guest initiator, each adapter connected separately to a virtual switch	
Advanced Network Features in use	SR-IOV and VMQ	
VMs with limited storage footprint (*)	Network services (1) Messaging services, Front-End (1) Content Management services (1) Communication services (1) Web server services (2)	Total of 6
VMs with heavy storage footprint (*)	Messaging services, Back End (1)	Total of 4
	OLTP relational Databases (1)	
	Decision Support System (1)	
	File services, CIFS network share (1)	
Ratio of VMs residing in one SAN volume		10:1

* Limited storage footprint refers to no additional data volumes. Heavy storage footprint refers to the situation where more data volumes are required for complex applications and workloads. Refer to Section 3.4 for details.

Figure 10 illustrates the KPIs collected when different RAID policies are implemented on the EqualLogic array unit used for the SAN.

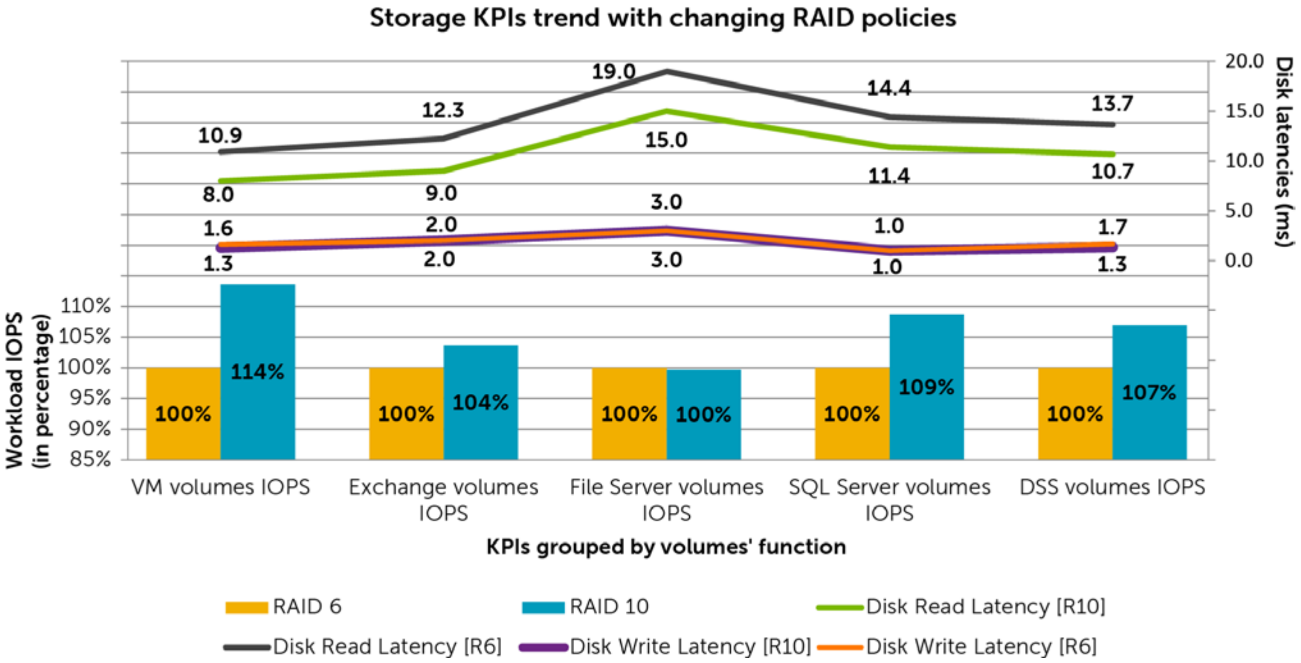


Figure 10 RAID policies: performance result comparison chart

The outcomes of this assessment show the performance fluctuation in favor of RAID 10 versus RAID 6, as expected. Different profile groups show a variation of delivered IOPS up to 14% better, while the disk read latency demonstrates a regular improved trend. The evaluation of the performance must then be weighed against additional characteristics of each policy, such the tolerance to disks failure or the total available capacity.

4.3 Selection of the iSCSI initiator type and collocation

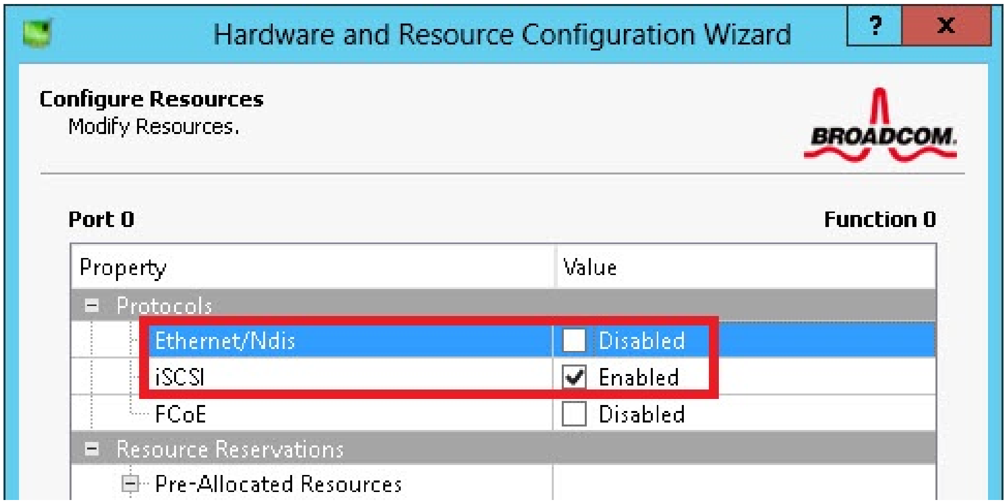
In networked storage systems, the iSCSI communications happen between two layers or end nodes through the wire. The iSCSI initiator functions as a client, which accesses the storage resources located on the target. The target acts as a server hosting the data. Messages and commands, similar to SCSI, are sent over the IP network between the two end nodes. The initiator falls into two broad categories: software or hardware initiator.

- A software initiator is an implementation of the functionalities by code. It usually runs as a device driver and reuses the network cards available in the operating system to emulate SCSI devices.
- A hardware initiator is a dedicated hardware resource, most commonly a host bus adapter. Since it is a fully dedicated resource, it generally reduces the computing impact of iSCSI traffic on the host.

While implementing a virtual infrastructure with Hyper-V technologies, an additional choice for the connectivity with the EqualLogic SAN through a software based iSCSI initiator would be the collocation of the initiator itself.

- Guest initiator: software located and running on the guest VMs, which allows direct connections to the volumes residing on the SAN through the virtual network adapters of the guest. The settings of the VMs include additional virtual network adapters dedicated to SAN traffic. The host hypervisor is not aware of the type of traffic traversing the VMBus adapters.
- Host initiator: software located and running on the host hypervisor, which allows you to connect to the volumes residing on the EqualLogic SAN from the root partition of Hyper-V through physical network adapters dedicated to the SAN traffic. VHDX/VHD files are created on the SAN volumes attached to the host and then added as SCSI disks to the settings of the VMs. The VMs are not aware of where their disks reside, either on local storage or the SAN.

The Broadcom BCM57810 network adapter, used on the PowerEdge servers for the tests in this paper, provides iSCSI Offload Engine (ISOE) capabilities together with the traditional NDIS operation in a Windows system environment. While the ISOE mode is enabled, the network adapter becomes a host bus adapter (HBA) and is used as a hardware initiator. While the traditional NDIS mode is enabled, the network adapter acts as the NIC and software initiators. The BCM57810 also allows both modes to operate at the same time, providing a two-in-one network adapter. Figure 11 shows the Broadcom network adapter resource configurations where NDIS and/or ISOE modes can be enabled or disabled (only one of the possible configurations is shown).



Configure Resources
Modify Resources.

Port 0 **Function 0**

Property	Value
⊟ Protocols	
Ethernet/Ndis	□ Disabled
iSCSI	☑ Enabled
FCoE	□ Disabled
⊟ Resource Reservations	
⊞ Pre-Allocated Resources	

Figure 11 Broadcom network adapter resource configurations: NDIS and ISOE modes enabled/disabled

The combinations of the hardware or software initiators, host or guest collocation, and Hyper-V shared virtual switch configuration create a wide spectrum of possible deployments. The configuration of the PowerEdge servers used for the tests illustrated in this paper includes four 10GbE Broadcom ports distributed across two network adapters. The following list presents how these ports are configured for this specific group of tests, and Figure 12 visualizes them:

1. NDIS host and guest (Four ports in use)
    - Two ports configured with NDIS and the host software initiator from the hypervisor
    - Two ports configured with NDIS and used by the dedicated virtual switches allowing guest software initiators from the VMs
2. NDIS host and guest reduced (Two ports in use)
    - Two ports configured with NDIS. Hyper-V virtual switches allow the host OS to share the network adapters. Host software initiator is from the hypervisor and the guest software initiators are from the VMs.
3. Mixed ISOE/NDIS ports (Four ports in use)
    - Two ports configured with ISOE only, thus hardware initiator
    - Two ports configured with NDIS and used by dedicated virtual switches allowing guest software initiators within the VMs
4. Mixed ISOE/NDIS ports reduced (Two ports in use)
    - Two ports configured with ISOE and NDIS. Hardware initiator is used for the host and dedicated virtual switches allow guest software initiators within the VMs.
5. NDIS host initiator only (Two ports in use)
    - Two ports configured with NDIS and a host software initiator for every SAN volume
6. ISOE hardware initiator only (Two ports in use)
    - Two ports configured with ISOE, thus a hardware initiator for every SAN volume

The objective of this comparison is to evaluate if the performance of the specific workload assessed is affected, and to what extent, by the use of different iSCSI initiators and multiple combinations of ports.

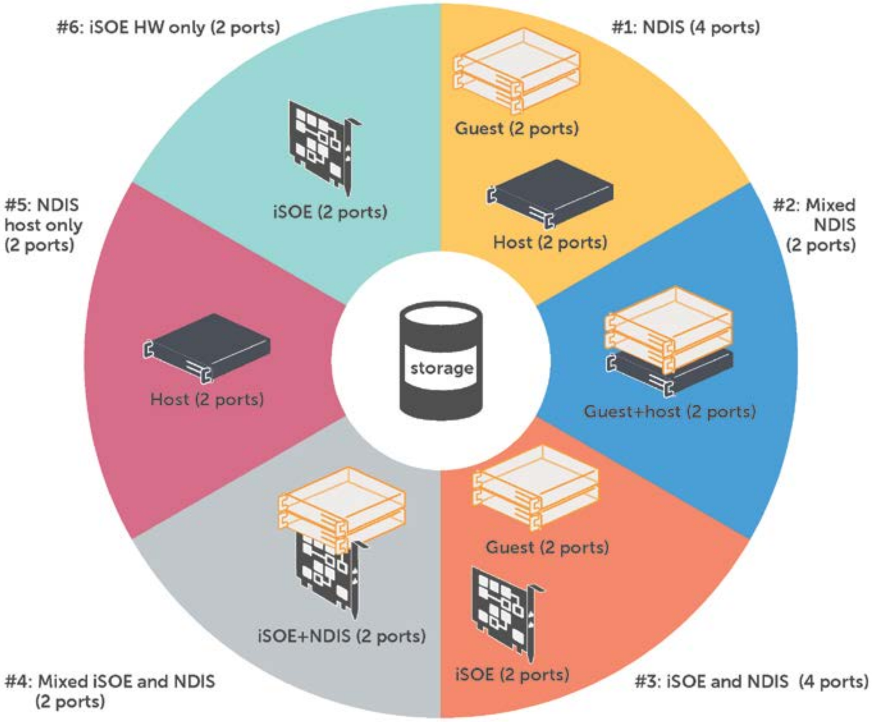


Figure 12 Graphical representation of the combinations of the initiators validated

Table 4 shows a summary of the variables and configurations used for these tests.

Table 4 Test parameters: iSCSI initiators

<table>
<tr><th colspan="4">Reference configuration: test variables under study</th></tr>
<tr><td rowspan="6">iSCSI Initiator</td><td>NDIS (Four ports)</td><td colspan="2">• Host software initiator, EQL HitKit (VMs startup VHDX volume)
• Guest software initiator, EQL HitKit (Data volumes)</td></tr>
<tr><td>NDIS (Two ports)</td><td colspan="2">• Host software initiator, EQL HitKit (VMs startup VHDX volume)
• Guest software initiator, EQL HitKit (Data volumes)</td></tr>
<tr><td>ISOE and NDIS (Four ports)</td><td colspan="2">• Hardware initiator (VMs startup VHDX volume)
• Guest software initiator, EQL HitKit (Data volumes)</td></tr>
<tr><td>ISOE and NDIS (Two ports)</td><td colspan="2">• Hardware initiator (VMs startup VHDX volume)
• Guest software initiator, EQL HitKit (Data volumes)</td></tr>
<tr><td>NDIS host only (Two ports)</td><td colspan="2">• Host software initiator, EQL HitKit (VMs startup VHDX volume and data volumes)</td></tr>
<tr><td>ISOE hardware only (Two ports)</td><td colspan="2">• Hardware initiator (VMs startup VHDX volume and data volumes)</td></tr>
<tr><td colspan="2" rowspan="2">Virtual networks (iSCSI)
* only with guest initiators in use</td><td colspan="2">Two virtual switches, each created on one dedicated 10GbE port</td></tr>
<tr><td colspan="2">Two virtual adapters for VMs with guest initiator, each adapter connected separately to a virtual switch</td></tr>
<tr><td colspan="2">Advanced Network Feature in use
* only with guest initiators in use</td><td colspan="2">SR-IOV and VMQ</td></tr>
<tr><th colspan="4">Reference configuration: consistent factors across the iterations of this test</th></tr>
<tr><td colspan="2">RAID policy (SAN)</td><td colspan="2">RAID 6</td></tr>
<tr><td colspan="2">VMs with limited storage footprint (*)</td><td>Network services (1)
Messaging services, Front-End (1)
Content Management services (1)
Communication services (1)
Web server services (2)</td><td>Total of 6</td></tr>
<tr><td colspan="2" rowspan="4">VMs with heavy storage footprint (*)</td><td>Messaging services, Back End (1)</td><td rowspan="4">Total of 4</td></tr>
<tr><td>OLTP relational Databases (1)</td></tr>
<tr><td>Decision Support System (1)</td></tr>
<tr><td>File services, CIFS network share (1)</td></tr>
<tr><td colspan="3">Ratio of VMs residing in one SAN volume</td><td>10:1</td></tr>
</table>

* Limited storage footprint refers to no additional data volumes. Heavy storage footprint refers to the situation where more data volumes are required for complex applications and workloads. Refer to Section 3.4 for details.

Figure 13 illustrates the KPIs collected when different combinations of iSCSI initiators have been configured on the host layer.

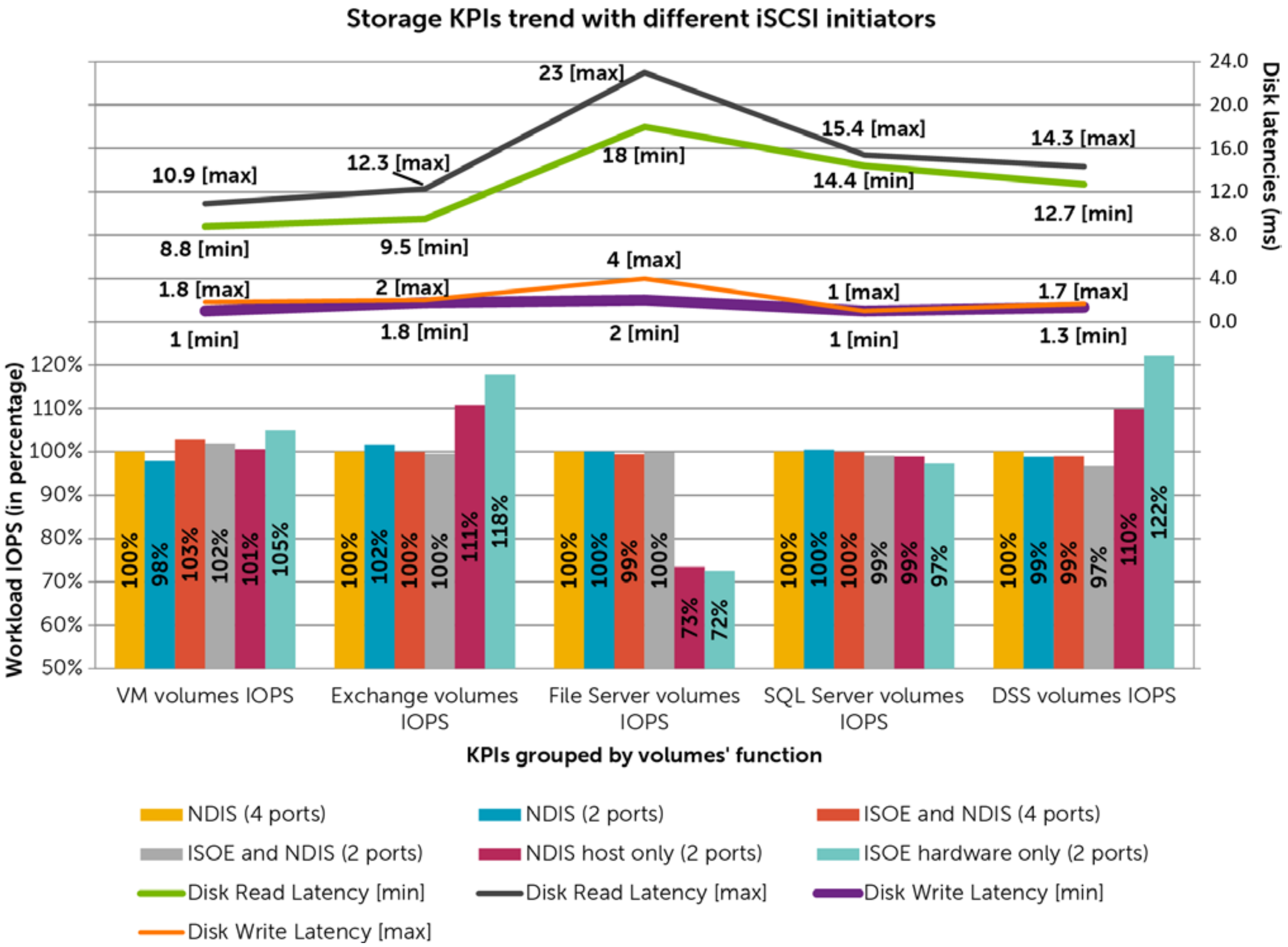


Figure 13 iSCSI initiator: performance result comparison chart

The results from the various combinations of iSCSI initiators and number of ports tested are summarized in the following points:

- The scenarios with two dedicated ports have a limited penalty in terms of measured disk latencies. Since these are more cost effective setups, they fit better in environments with no expectation of workload growth.
- The hardware initiator (ISOE) constantly gives a slightly better delivery of the IOPS and lowers the resulting disk latencies. This means it is a viable solution for configurations with increasing performance demand, but should be carefully evaluated since the monitoring of such configuration is limited from the host OS.

- The final two scenarios with the host initiator in use reveal the direct access to data stored in VHDX files provides better performance under some specific workloads (sequential and with medium to large I/O size).
- The workload profiles with the smallest I/O size and nearly complete randomness are the most challenging environment to address, with the File Server services simulation with two NDIS ports being the only solution that overcomes the threshold for disk read latency.

4.4 VMs volumes layout and provisioning

A guest VM running on Windows Server 2012 with Hyper-V is a set of files residing on the disks which support the final setup of each VM.

VM configuration file (*.XML): a structured text file containing the configuration details of the specific VM. Each VM and each Hyper-V snapshot of the VM owns one of these files. The file itself is named after the associated GUID to internally identify the VM (or snapshot).

Memory image file (*.BIN): a container of the memory of a VM (or snapshot) that is in a saved state. It has the size of the memory assigned to the VM, thus should be taken into account for capacity planning. It is used to recover the VM from a power cycle of the host or from the Quick Migration operation of the VM.

Saved state file (*.VSV): the container of information for the VM devices when in a saved state.

Virtual hard disk file (*.VHDX, *.VHD): is the file format which represents a hard disk image. It incorporates all the data written to the disk in the VM. Refer to Section 2.2.1 above for additional details about the characteristics of this file format, including its limits and structure.

Differencing disk image file (*.AVHDX, *.AVHD): used for Hyper-V snapshots. This file contains all the changes that happened in the virtual hard disk file since the snapshot was last taken. It is subject to extensive growth when the snapshot is not either reverted to the previous state or merged with the parent disk, and the VM is conditioned to a high rate of disk changes. This file is named after the parent .VHDX/.VHD file and the GUID of the snapshot itself.

To approach the design considerations required at the time of provisioning the SAN volumes and deploying the VMs, use the following points to examine the advantages and disadvantages of these configurations:

1. Many to one ratio of VMs per volume and using a fully provisioned SAN volume
   - Multiple VMs, and all their files, are hosted on the same SAN volume, appropriately sized, under a single namespace with the required nested sub-directories
   - Entirely pre-allocated physical storage space at the time of volume creation
2. One to one ratio of VMs per volume and using a set of fully provisioned SAN volumes
   - One VM, and all its files are hosted in a SAN volume appropriately sized
   - One volume per each VM under consideration
   - Entirely pre-allocated physical storage space at the time of volume creation
3. Many to one ratio of VMs per volume and using a thin provisioned SAN volume

- Multiple VMs, and all their files are hosted on the same SAN volume, appropriately sized, under a single namespace with the required nested sub-directories
- Thin provisioned SAN volume, with space allocation based on volume usage by the host

Table 5 shows a summary of the variables and configurations used while executing these tests.

Table 5 Test parameters: volume layout and provisioning

Reference configuration: test variables under study		
Ratio of VMs residing in one SAN volume / Provisioning of the volume	10:1 / fully provisioned	
	1:1 / fully provisioned	
	10:1 / thin provisioned	
Reference configuration: consistent factors across the iterations of this test		
RAID policy (SAN)	RAID 6	
iSCSI Initiator	Host software initiator, EQL HitKit (VMs startup VHDX volume)	
	Guest software initiator, EQL HitKit (Data volumes)	
Virtual networks (iSCSI)	Two virtual switches, each created on one dedicated 10GbE port	
	Two virtual adapters for VMs with guest initiator, each adapter connected separately to a virtual switch	
Advanced Network Feature in use	SR-IOV and VMQ	
VMs with limited storage footprint (*)	Network services (1) Messaging services, Front-End (1) Content Management services (1) Communication services (1) Web server services (2)	Total of 6
VMs with heavy storage footprint (*)	Messaging services, Back End (1)	Total of 4
	OLTP relational Databases (1)	
	Decision Support System (1)	
	File services, CIFS network share (1)	

* Limited storage footprint refers to no additional data volumes. Heavy storage footprint refers to the situation where more data volumes are required for complex applications and workloads. Refer to Section 3.4 for details.

Figure 14 illustrates the KPIs collected when the VMs are deployed over the varying volume layouts and provisioning types.

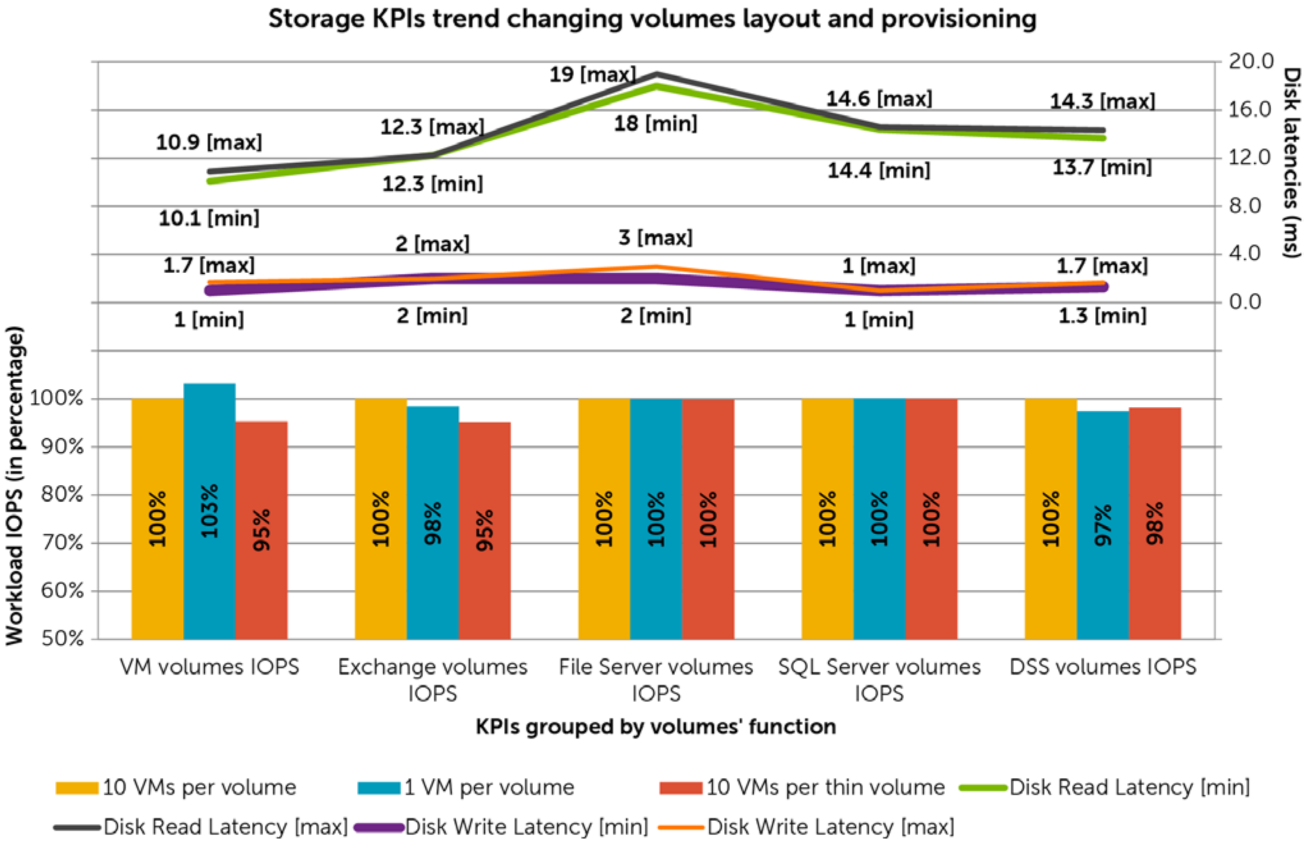


Figure 14 Volumes layout and provisioning: performance result comparison chart

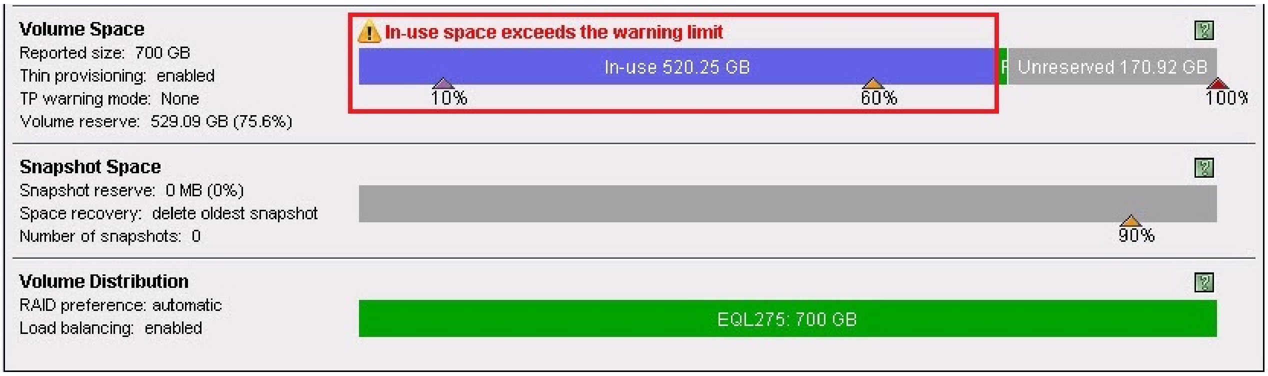
The results show similar performance among the different ways to deploy the volumes. Some minimal discrepancies are more pronounced for specific types of workload. The 1:1 ratio VM deployment reports an increase in IOPS performed with an associated cost to disk read latency. The selection of one of these two deployments must also be guided by the level of administrative complexity for both SAN volumes and VMs driven in a 1:1 ratio. The thin volume provisioning shows a slight decline of IOPS when the 'UNMAP' operations are most active, which then influences the rest of the workloads minimally.

Further scrutiny of the volume usage by the host was done for the thin provisioned volume scenario. The single thin volume was created, and then all the files of the VMs were copied over while the VMs were in the shutdown state. After the VMs were started, they were left running for 24 hours before recording the performance data. This time allowed the operating systems to execute the expected rethinning operations against the VHDX files, which were then followed by the EqualLogic array SCSI UNMAP to reclaim unused space.

The space usage of the VMs, as detailed in Appendix A.4 is:

- The total amount of space required by the startup VHDX files of the 10 VMs was about 520GB
- After the VMs are started, this expands to around 684GB which includes the memory image and the saved state files
- The space allocated for the single volume at the time of creation is 700GB (volume reserve size)

Immediately after the copy of the VM files, the space usage of the thin volume grows up to the 520GB size as shown in Figure 15. Then the storage usage first grows to 684GB when the VMs are started, then shrinks to 283GB over time, as shown in Figure 16. This is because of the automatic rethinning operations (40% of the volume reserve size, or nearly 42% of the nominal size of the files).



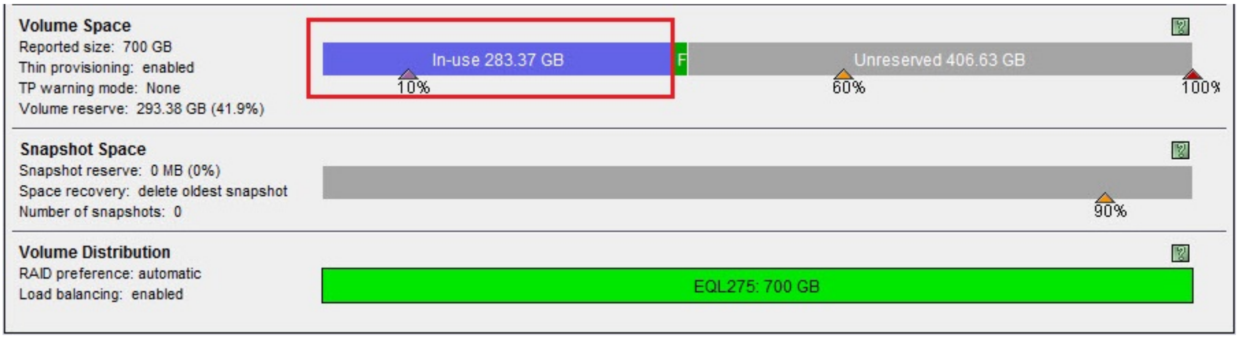
Volume Space
Reported size: 700 GB
Thin provisioning: enabled
TP warning mode: None
Volume reserve: 529.09 GB (75.6%)

In-use space exceeds the warning limit

Snapshot Space
Snapshot reserve: 0 MB (0%)
Space recovery: delete oldest snapshot
Number of snapshots: 0

Volume Distribution
RAID preference: automatic
Load balancing: enabled

Figure 15 Thin provisioned volume: reported size after static copies of the VHDX files



Volume Space
Reported size: 700 GB
Thin provisioning: enabled
TP warning mode: None
Volume reserve: 293.38 GB (41.9%)

Snapshot Space
Snapshot reserve: 0 MB (0%)
Space recovery: delete oldest snapshot
Number of snapshots: 0

Volume Distribution
RAID preference: automatic
Load balancing: enabled

Figure 16 Thin provisioned volume: reported size after 24 hours of VMs running

4.5 Scale up the number of VMs

The final analysis performed under the standalone server scenario scaled up the number of VMs. This sizing technique is usually associated with the increase of the elements under study (overall workload, VMs in this case) until one of the resources breaches its limit. The tests provide enough samples to show a VM scaling trend even though the configuration is limited by the following hardware boundaries:

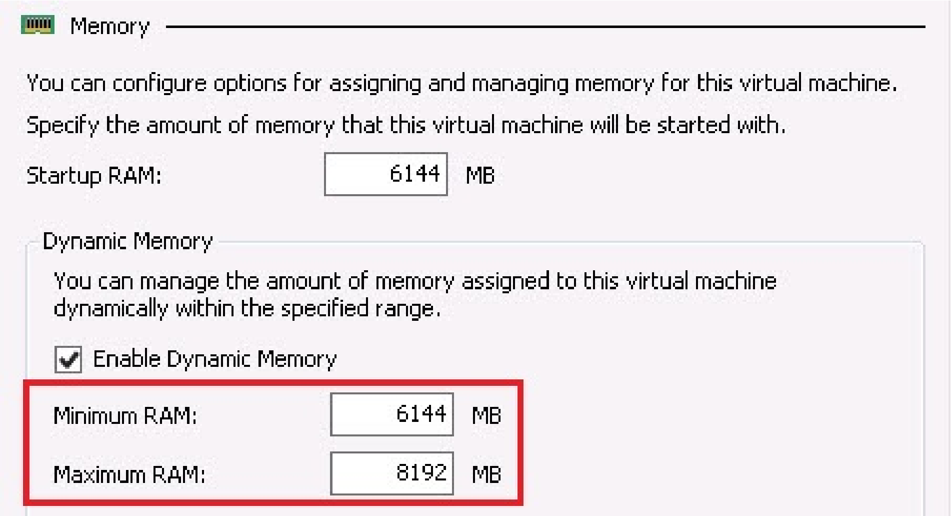
- The PowerEdge server in this test has 192GB of memory installed
- The EqualLogic array in this test has a defined capacity

To better cope with these predefined boundaries, the test setup was modified for this scenario only so that the vertical scaling exercise was based on a uniform scaling of the workload. A unique template of a VM with the following characteristics was deployed repeatedly:

- Windows Server 2012 operating system
- Virtual resources: Four CPUs, 8GB memory, One fixed-size VHDX of 50GB without extra data volumes
- Workload and profile: category of Web Server services by I/O access, with memory allocation and CPU utilization as defined previously

The progression of the workload as reported in Table 6 starts with six VMs, then 12 VMs, and finally 22 VMs which required a total memory utilization on the host close to the physical boundary of 192GB mentioned above. In order to provide a next step in the scaling, a further iteration of the test was performed using 24 VMs, but requiring the advantages of the Dynamic Memory technology provided by the Hyper-V server.

Dynamic Memory allows a VM to allocate a variable amount of memory within a predefined range as shown in Figure 17. VMs with this option enabled can start with a lower amount of memory and then grow it as needed by the requirements of their operations or decrease it and release it to the underlying host.



Memory

You can configure options for assigning and managing memory for this virtual machine.

Specify the amount of memory that this virtual machine will be started with.

Startup RAM: 6144 MB

Dynamic Memory

You can manage the amount of memory assigned to this virtual machine dynamically within the specified range.

☑ Enable Dynamic Memory

Minimum RAM: 6144 MB

Maximum RAM: 8192 MB

Figure 17 Scale up the number of VMs: Dynamic Memory

Dynamic Memory is a technology predominantly used in VDI environments, not in server-based virtual infrastructures. Some server workloads might not benefit from the use of dynamic memory and it should be reviewed on a case by case basis. This specific test iteration is for experimental purposes only and does not represent a best practice, a suggestion to use this technology in server scenarios, or a recommendation for the specific workload simulated.

Table 6 shows a summary of the variables and configurations used while executing these tests.

Table 6 Test parameters: scale up the number of VMs

Reference configuration: test variables under study	
Ratio of VMs residing in one SAN volume	6:1
	12:1
	22:1
	24:1 (VMs with Dynamic memory enabled)
Reference configuration: consistent factors across the iterations of this test	
RAID policy (SAN)	RAID 6
iSCSI Initiator	Host software initiator, EQL HitKit (VMs startup VHDX volume)
VMs with limited storage footprint (*)	Web server services

* Limited storage footprint refers to no additional data volumes. Refer to Section 3.4 for details.

Figure 18 illustrates the KPIs collected when increasing the number of running VMs. This series of tests reveals a proportional increase of both metrics of the overall workload: IOPS performed and bandwidth in use. The trend of the KPIs shown in Figure 18 while increasing the load proves the burden on the disk latencies and queue depth increases more slowly, staying well below the warning thresholds and preserving the ability to scale up far over the amount of VMs tested in the scenarios.

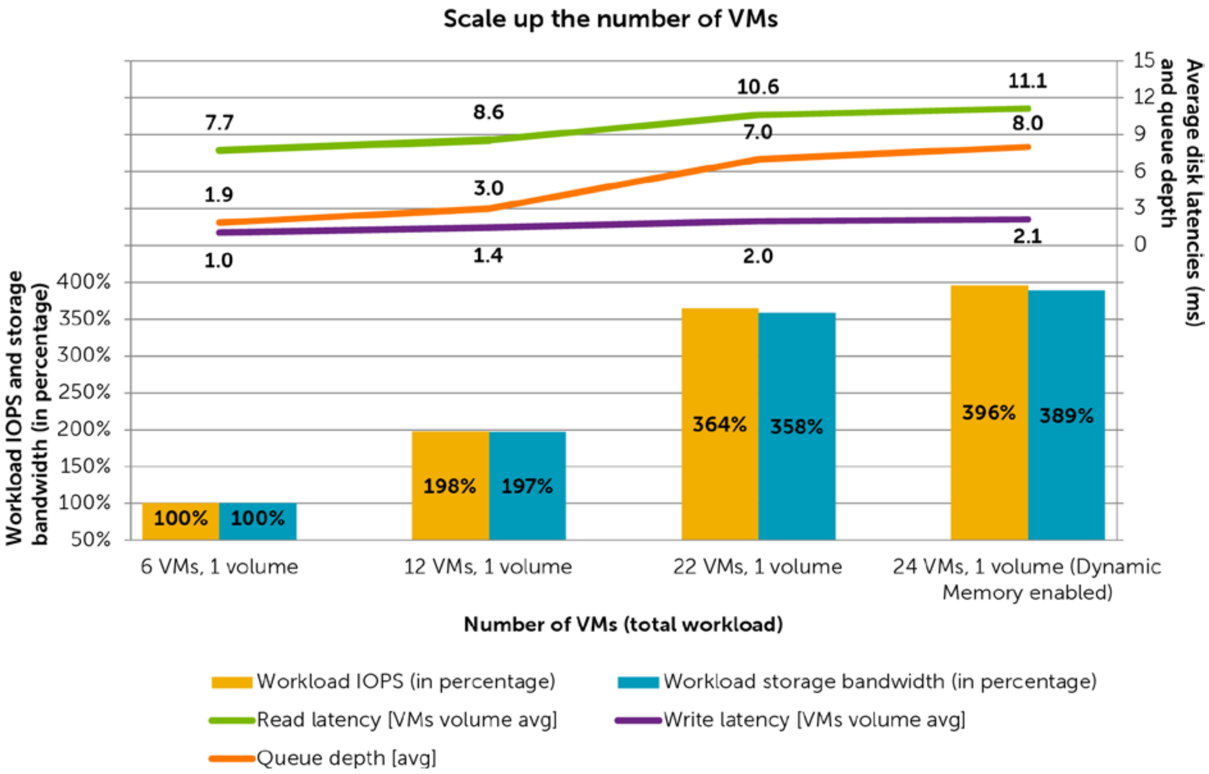


Figure 18 Scale up the number of VMs: performance result comparison chart

5 Highly available Hyper-V clusters implementation

The second phase of this study focused on the implementation and operation of highly available virtual environments by Windows Failover Cluster and Hyper-V technologies. The main configurations of the setup environment built to assess the failover cluster operations are described in the list:

- Two node Windows Failover Cluster
- Quorum based on node and disk majority, with a dedicated disk witness on the EqualLogic SAN
- Data disks with Cluster Shared Volumes (CSV)
- Redundant network for cluster network communication with a dedicated preferred network
- Dedicated network for Live Migration

Cluster Shared Volumes (CSV) is a general purpose file system layered above NTFS, and it currently supports Hyper-V and file share workloads. CSV allows shared storage resources in a Windows Failover Cluster to be accessed by read-write operations simultaneously sent from multiple nodes. It does not require a drive ownership change and volume dismount/mount operations like in former Windows cluster versions and allows quicker switchovers of the resources contained in the disk from one node to another.

Live Migration is a Hyper-V feature available in Failover Cluster configurations which provides the ability to relocate active workloads from a source node to a destination node server in the cluster without disruption of the connectivity to and the services offered from the VM. The process consists of a transfer over the Live Migration network of the VM configuration and memory, while the memory pages being modified are tracked, and then synced before the switchover of the VM to the destination node happens. Hyper-V in Windows Server 2012 allows the concurrent movement of multiple VMs or workloads.

Quick Migration is a Hyper-V feature available in Failover Cluster configurations which provides the ability to relocate active workloads from a source node to a destination node server in the cluster. Unlike Live Migration, a Quick Migration operation suits workloads that can accept short planned outages. The downtime is required to save the state of the VM on the source node and then resume it on the target node. The duration of this operation is influenced by the amount of memory of the VM, because it must be saved to the disk first and read from there.

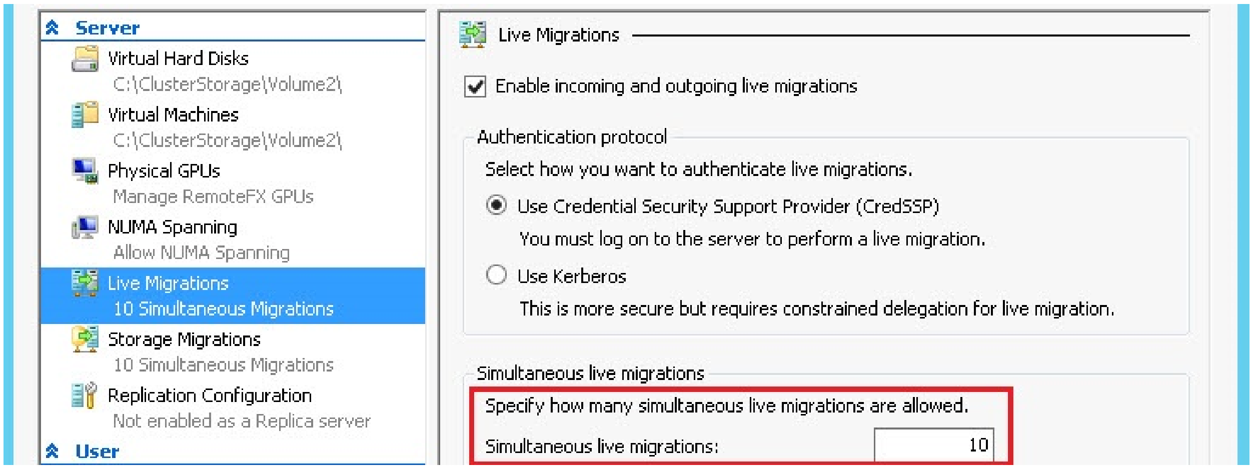
Storage Migration is a feature added natively to Hyper-V in Windows Server 2012 and previously available in a similar form only in SCVMM. It offers the capability of redistributing the VM storage elements among different physical disks within the same Hyper-V server or across nodes in a Failover Cluster. The VM remains active during the move, since the process creates the equivalent virtual hard disks and fills them with the source data while the continuous data change happening during the duration of the move is mirrored on both of the storage repositories until the final switchover.

Shared-nothing Live Migration is a Windows Server 2012 feature implemented as a combination of Live and Storage migration. It only requires network connectivity between nodes of a Failover Cluster and offers the ability to seamlessly move the storage of the VM first and the computing state and activities of the VM later within the same administrative operation.

Storage Migration is a feature available in a standalone Hyper-V installation as well. The considerations around it have been included in this section to be similar to other forms of VM migration.

5.1 Live Migration considerations

Live Migration in Windows Server 2012 offers concurrent moves of multiple VMs unlike its predecessors. The default configuration of Hyper-V manager limits the number of moves to a maximum of two. Figure 19 illustrates the changes implemented to increase the amount of simultaneous live migrations to the number of VMs running in the test environment.



Server
Virtual Hard Disks
C:\ClusterStorage\Volume2\
Virtual Machines
C:\ClusterStorage\Volume2\
Physical GPUs
Manage RemoteFX GPUs
NUMA Spanning
Allow NUMA Spanning
Live Migrations
10 Simultaneous Migrations
Storage Migrations
10 Simultaneous Migrations
Replication Configuration
Not enabled as a Replica server
User

Live Migrations

☑ Enable incoming and outgoing live migrations

Authentication protocol
Select how you want to authenticate live migrations.
● Use Credential Security Support Provider (CredSSP)
You must log on to the server to perform a live migration.
○ Use Kerberos
This is more secure but requires constrained delegation for live migration.

Simultaneous live migrations
Specify how many simultaneous live migrations are allowed.
Simultaneous live migrations: 10

Figure 19 Live Migration: simultaneous live migrations

Live Migration performance is greatly influenced by the amount of memory of the VM, the rate of memory changes, and the bandwidth of the network that sustains the transfer between the nodes. The following use cases were tested to better anticipate Live Migration operation performance, and were validated under 1GbE or 2GbE total Live Migration dedicated bandwidth:

1. Live Migration of one VM at a time, using VMs with an increasing memory footprint
   - The sequence includes 4GB, 8GB, 16GB, 24GB, 32GB, and 48GB pre-allocated memory VMs
2. Live Migration of three VMs simultaneously, using VMs with an increasing combined memory footprint
   - The sequence includes 24GB, 28GB, 48GB, and 76GB combined pre-allocated memory VMs
3. Live Migration of six VMs simultaneously, using VMs with an increasing combined memory footprint
   - The sequence includes 52GB, 72GB, and 124GB combined pre-allocated memory VMs
4. Simultaneous Live Migration of all 10 VMs in the test environment
   - The combined pre-allocated memory is 164GB

Table 7 shows a summary of the variables and configurations used while running these tests. Each specific workload defined for the VMs was running during the tests and Live Migration operations were considered successful only if no error or disconnections happened during the move.

No report of Live Migration failure or delay was recorded among all the iterations in the virtualized environment under test.

Table 7 Test parameters: Live Migration

Reference configuration: test variables under study		
Simultaneous Live Migrations	1 VM, total memory: 4GB, 8GB, 16GB, 24GB, 32GB, and 48GB	
	3 VMs, total memory: 24GB, 28GB, 48GB, and 76GB	
	6 VMs, total memory: 52GB, 72GB, and 124GB	
	10 VMs, total memory: 164GB	
Live Migration dedicated bandwidth	1GbE	
	2GbE, teamed by Broadcom BACS	
Reference configuration: consistent factors across the iterations of this test		
RAID policy (SAN)	RAID 6	
iSCSI Initiator	Host software initiator, EQL HitKit (VMs startup VHDX volume)	
	Guest software initiator, EQL HitKit (Data volumes)	
Virtual networks (iSCSI)	Two virtual switches, each created on one dedicated 10GbE port	
	Two virtual adapters for VMs with guest initiator, each adapter connected separately to a virtual switch	
Advanced Network Feature in use	SR-IOV and VMQ	
VMs with limited storage footprint (*)	Network services (1) Messaging services, Front-End (1) Content Management services (1) Communication services (1) Web server services(2)	Total of 6
VMs with heavy storage footprint (*)	Messaging services, Back End (1)	Total of 4
	OLTP relational Databases (1)	
	Decision Support System (1)	
	File services, CIFS network share (1)	
Ratio of VMs residing in one SAN volume (CSV enabled)		10:1

* Limited storage footprint refers to no additional data volumes. Heavy storage footprint refers to the situation where more data volumes are required for complex applications and workloads. Refer to Section 3.4 for details.

Figure 20 illustrates the duration of the Live Migration operation for one VM at a time, comparing two different network bandwidths implemented on the dedicated Live Migration network.

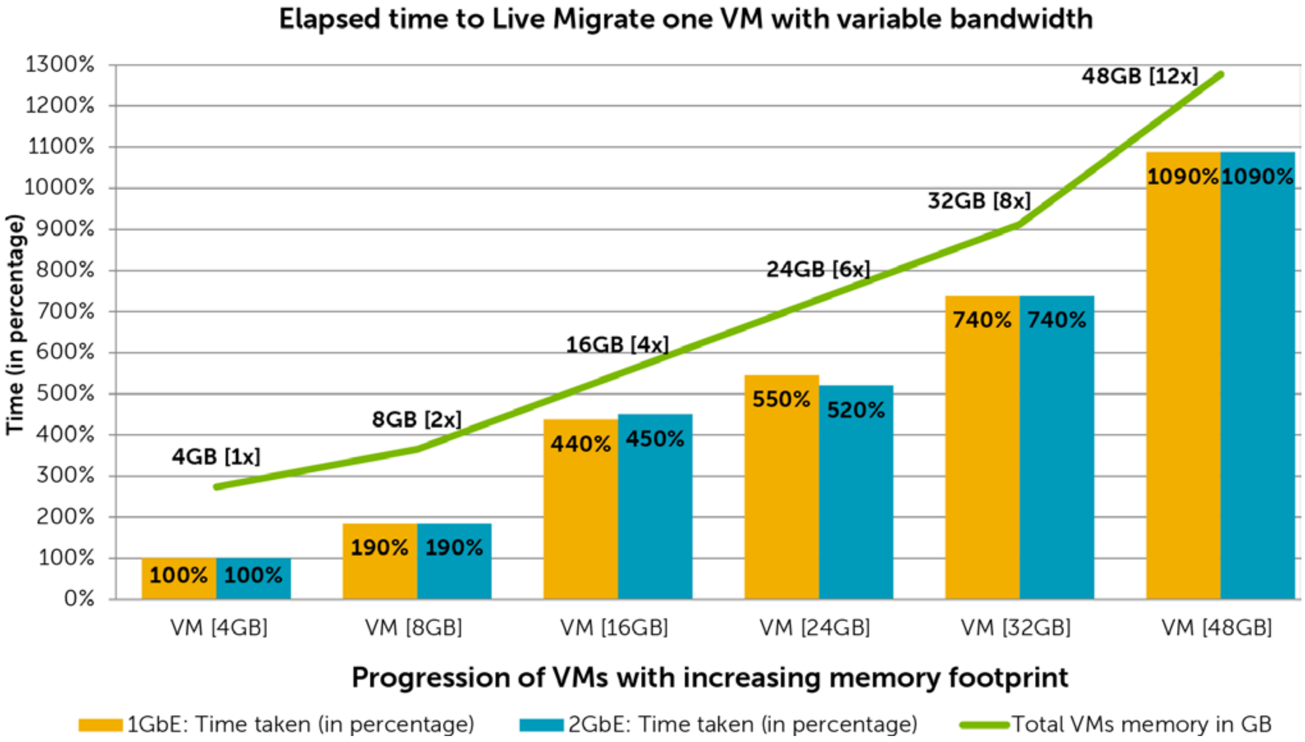


Figure 20 Live Migration: comparison of one VM live migrated across different network bandwidths

The time taken is shown to be the same across the two network configurations, providing the following insights related to this type of operation:

- The network utilization of all the completed operations, both with 1GbE and 2GbE network bandwidth, remained close to or below the edge of saturation of 1GbE. Since the Live Migration network is dedicated to this purpose, there is no risk of contention and overlapping with other processes.
- A single active Live Migration operation would not require more than 1GbE network bandwidth to be accomplished. The provision of a larger bandwidth would not provide significant benefits for a single move operation. It appears that Hyper-V internal constraints do not allow taking advantage of more bandwidth in this specific use case.
- Some small differences between the two set of tests for the same VM can be related to the running workloads within the VM itself. Some VMs have a heavier active workload then others following the VM and workloads definition proposed in Table 1 and Table 12.
- Overall the progression of VMs with increasing memory footprint shows a better than linear performance in either bandwidth scenario, since the multiplier of the memory footprint is not completely fulfilled in the time taken to live migrate.

Figure 21 illustrates the duration of the simultaneous Live Migrations of groups of VMs, comparing two network bandwidths implemented on the dedicated Live Migration network.

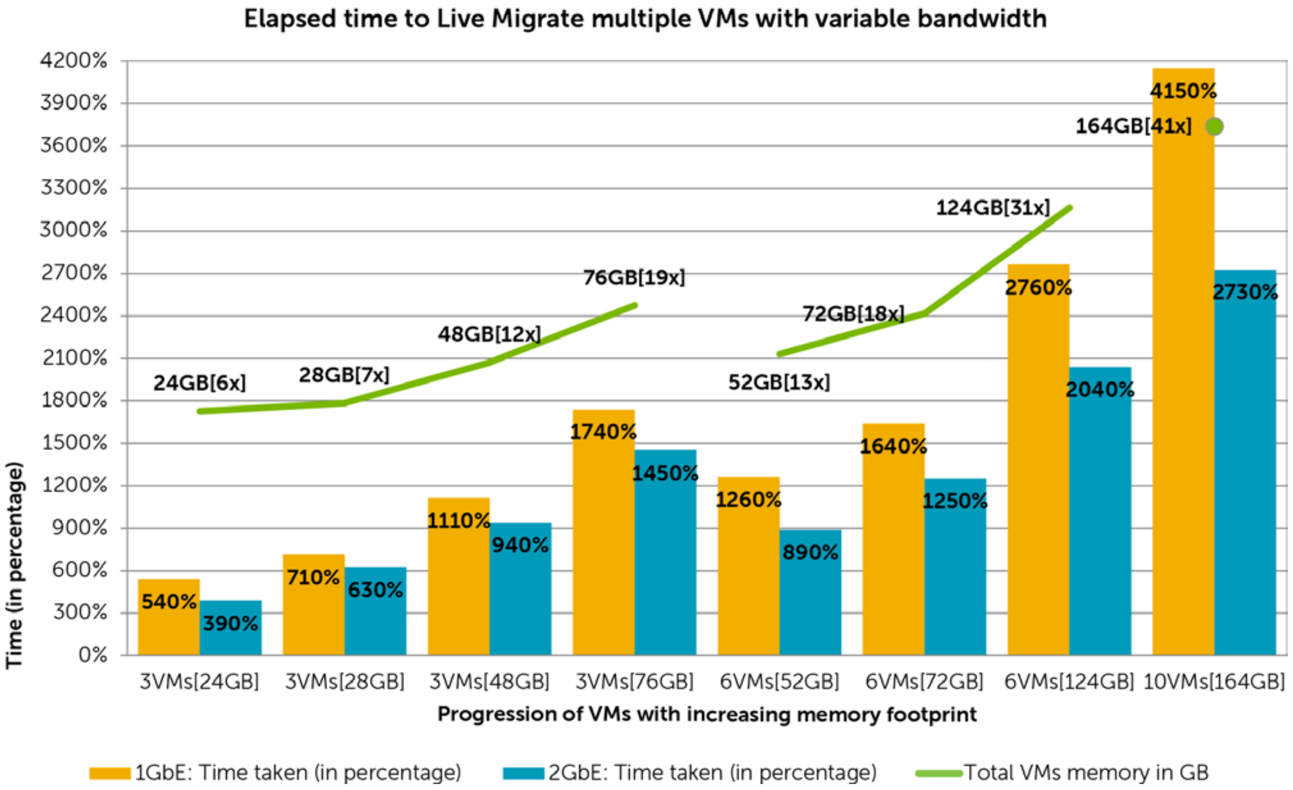


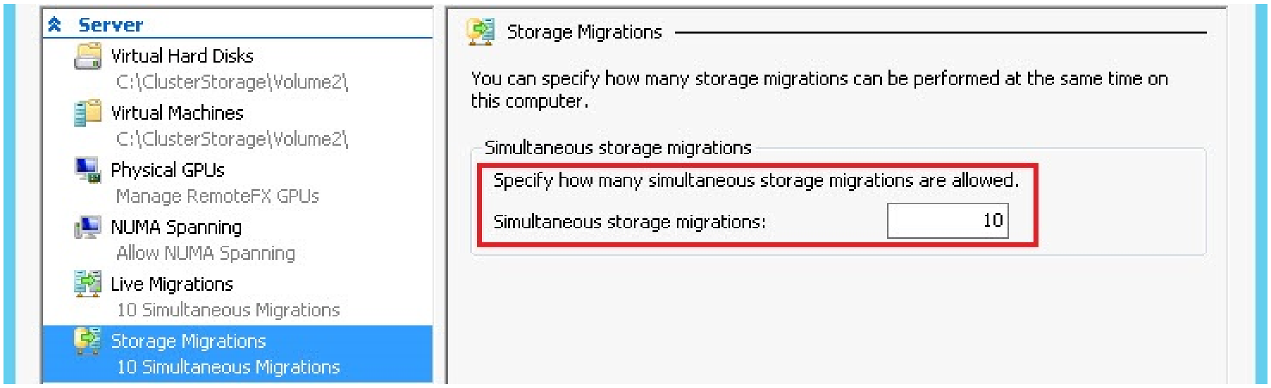
Figure 21 Live Migration: comparison of group of VMs live migrated simultaneously across different network bandwidths

Unlike the previous set of tests with one VM only, the results from the simultaneous migration of multiple VMs display a clear improvement when increasing the available bandwidth for the dedicated Live Migration network.

- All the elapsed times are calculated as a percentage compared with the reference elapsed time of a single VM with 4GB of memory from the previous set of tests (that test showed equal time between 1GbE and 2GbE).
- Even considering the 1GbE only option, the time elapsed in percentage is still lower than the proportional expected time for every iteration of the test.
- The gap between 1GbE and 2GbE increases while increasing the number of VMs, thus increasing the amount of memory to be transferred between hosts.
- The monitored utilization of the networks reports a saturated 1GbE link for every iteration of the test, while the 2GbE shows a 1.4/1.6Gbps utilization for the three simultaneous VMs test, which increases to 1.6/1.8Gbps for the six and ten simultaneous VMs use case.
- The tests show the 2GbE bandwidth to be a reasonable size for the simultaneous move of the entire workload.

5.2 Storage Migration considerations

Storage Migration is now natively available in Windows Server 2012 without the requirement to have a SCVMM installation. Similar to the Live Migration default configuration, Hyper-V limits the default amount of concurrent storage migrations to a maximum of two. Figure 22 displays the changes implemented to increase the number of simultaneous storage migrations to the number of VMs running in the test environment used.



Storage Migrations

You can specify how many storage migrations can be performed at the same time on this computer.

Simultaneous storage migrations

Specify how many simultaneous storage migrations are allowed.

Simultaneous storage migrations: 10

Figure 22 Storage Migration: simultaneous storage migrations

Storage Migration allows the Hyper-V host to rearrange the storage components of VMs among different destination devices mounted and presented to the storage subsystem of the hypervisor regardless of whether the VM is actively running or idle. The process includes first copying the file for each virtual hard disk associated with the VM (.VHDX, .VHD), then the flush of all the writes that happened since the copy of the virtual hard disk started. Some considerations about these steps are:

- The initial copy of the virtual hard disk is comparable to a sequential disk activity reading from the source and writing to the destination. An unbuffered copy of the file is an appropriate baseline to assess the duration of this task in every environment (for example, the type of copy performed by "`XCOPY /J`").
- The overall performance of a storage migration is highly affected by the rate of changes happening to the data contained in the virtual hard disk. A VM with intense disk activity will take longer to be storage migrated due to the larger amount of writes to be flushed to the destination disk image.

In an EqualLogic SAN environment, the different use cases for a storage migration can be reduced to the following list:

1. A Hyper-V host migrates the VM storage across two different groups of arrays. The disk spindles are totally separated, and the source would sustain the expected write changes from the running workload and the additional disk reading activity. All the writes to the new disk images happen on the destination, including the final writes flush.
2. A Hyper-V host migrates the VM storage across two pools within the same group of arrays. The disk spindles are separated and, as in the previous use case, the source would sustain the expected write

changes from the running workload and the additional disk reading activity. All the writes to the new disk images happen on the destination, including the final writes flush.

3. A Hyper-V host migrates the VM storage within the same pool of a group, thus the data is copied and moved between volumes or folders built on the same disk spindles. The same disk resources must sustain the workload of disk writes, and the additional sequential reading and writing activities and the final writes to flush the changes to the destination.

While the first two use cases are quite similar since they can benefit from the use of exclusive resources for the source and destination storage, the last use case is the least favorable due to contention for disk resources. The following test explores the last use case, which is the most risky operation to perform with an active workload.

Table 8 reports a summary of the variables and configurations used while comparing a storage migration sequence including the 10 VMs of the simulation. Each specific workload defined for the VMs is running during the tests, and the Storage Migration operations are considered successful only if no error or disconnections happens during the move.

Table 8 Test parameters: Storage Migration

Reference configuration: test variables under study		
Simultaneous Storage Migrations	10 VMs, total disk usage: 684GB ODX enabled (default on Windows Server 2012 and EqualLogic)	
	10 VMs, total disk usage: 684GB ODX disabled	
Reference configuration: consistent factors across the iterations of this test		
RAID policy (SAN)	RAID 6	
iSCSI Initiator	Host software initiator, EQL HitKit (VMs startup VHDX volume)	
	Guest software initiator, EQL HitKit (Data volumes)	
Virtual networks (iSCSI)	Two virtual switches, each created on one dedicated 10GbE port	
	Two virtual adapters for VMs with guest initiator, each adapter connected separately to a virtual switch	
Advanced Network Feature in use	SR-IOV and VMQ	
VMs with limited storage footprint (*)	Network services (1) Messaging services, Front-End (1) Content Management services (1) Communication services (1) Web server services(2)	Total of 6
VMs with heavy storage footprint (*)	Messaging services, Back End (1)	Total of 4
	OLTP relational Databases (1)	

	Decision Support System (1)	
	File services, CIFS network share (1)	
Ratio of VMs residing in one SAN volume (CSV enabled)		10:1

* Limited storage footprint refers to no additional data volumes. Heavy storage footprint refers to the situation where more data volumes are required for complex applications and workloads. Refer to Section 3.4 for details.

Figure 23 illustrates the duration and processor utilization of the Storage Migration operation for 10 VMs at a time, comparing the benefits of the Offloaded Data Transfer feature.

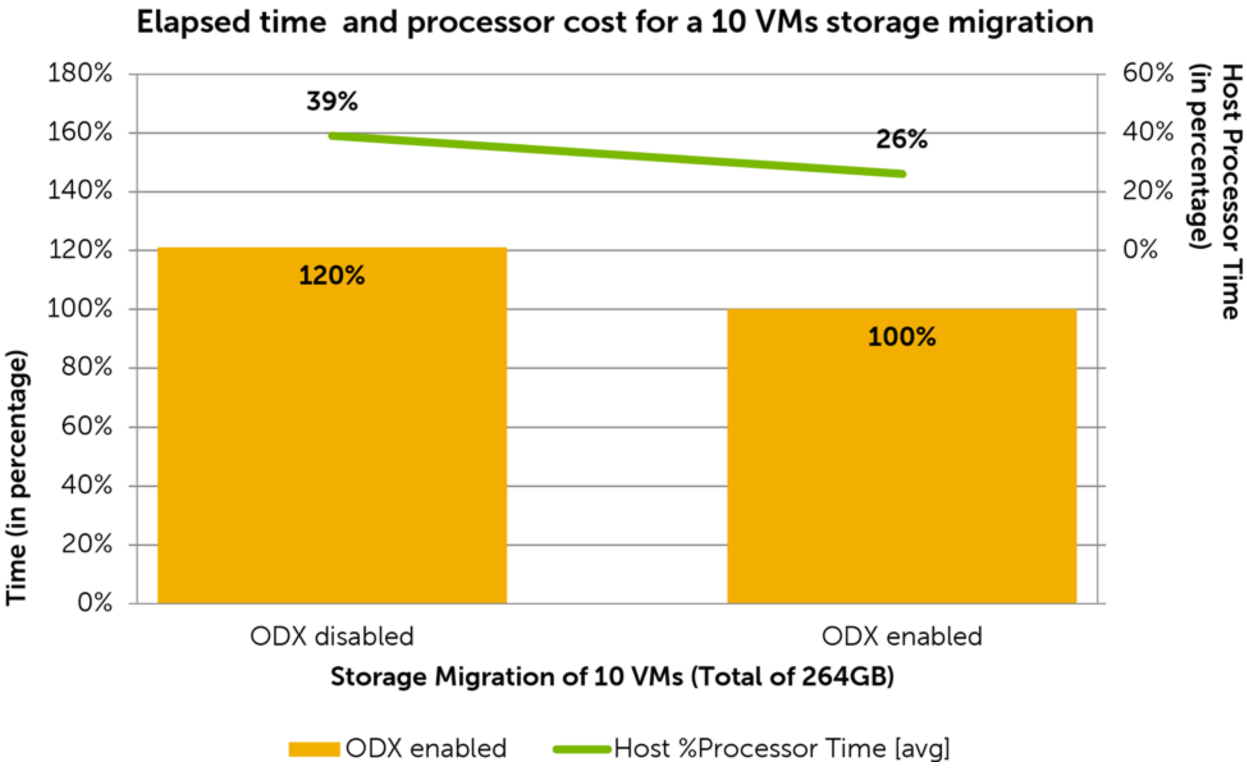


Figure 23 Storage Migration: comparison of 10 VMs migrated within the same EqualLogic array

Additional observations recorded while executing and comparing this operation with the two ODX configurations are:

- The move operation is started simultaneously on all ten VMs. The migration of the VMs with dedicated data drives and limited I/O activity on their system drive finishes before the remaining VM moves. This confirms that a sustained rate of changes on a virtual hard disk being moved lengthens the time required to migrate the VM.
- The total bandwidth in use when ODX is disabled increases significantly since the migration process records the extra read and write activities against the source and destination. With ODX enabled

instead, the host traffic from and to the SAN is kept as low as the active workloads within the VMs. Figure 24 displays a SAN HQ screen capture tracing the total bandwidth utilization among all the volumes where the difference of host-SAN utilization during the two different storage migrations of the ten VMs is conspicuous (ODX disabled on the left, enabled on the right).

- In both cases the VM workloads suffer from the contention with the storage migration operations. The VM workloads show a decrease in total IOPS performed to accomplish the planned activities for the workloads on the VM system drives, but still achieve successful results overall, including on the dedicated volumes for extra data. Nevertheless with ODX enabled, the IOPS decrease on the volume of the VMs is contained and performs 20% better compared to when the feature is disabled. The average read latency and the queues depth are 20% lower and show a more constant and predictable load without any sign of spikes.

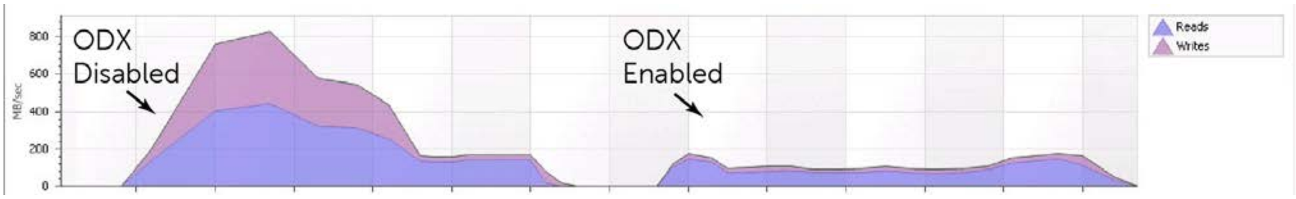


Figure 24 Storage Migration: Host-SAN bandwidth comparison with ODX disabled and enabled

For additional information on Offloaded Data Transfer (ODX) and Intelligent Storage Arrays, refer to Microsoft Devices Develop documentation at http://msdn.microsoft.com/en-us/library/windows/hardware/hh833784.aspx

5.3 Scale out the cluster and the SAN

The final considerations around the highly available deployments of Hyper-V are based on the ability to scale out the environment used for the simulations presented in this paper without affecting the performance or the end-user experience. The goal is to validate the appropriate building block to deploy when a larger amount of VMs and workloads are required.

The scenario used for this additional validation is defined by the need to implement redundant services on the virtual infrastructure, which reduces the risk factor of hosting a single instance of the storage data. Therefore it requires the duplication of both VMs and data volumes, which builds an environment where software based replication mechanisms protect from the risk of loss or corruption of data. This includes Exchange Server DAG, SQL Server Always On, Distributed File System services, and Active Directory native replication between domain controllers.

Since the simulation assumes the duplicated data is serving as data protection, the usage of multiple EqualLogic pools with distinct disk resources to guarantee the persistence of one of the copy in case of a fault is required (for example, array failure or disks failure overcoming the limit of the RAID policy implemented).

A short summary of the main characteristics of this scenario are reported below and in Table 9:

- Two node Windows Failover Cluster
- Two blocks of ten VMs each, each node of the failover cluster hosting one of the blocks
- Two EqualLogic disk arrays in the same group, with two pools and one array in each pool
- Each block of VMs deployed within one pool, including all the extra data volumes required
- Heterogeneous storage workload resulting from the customized simulation running in each VM, as defined previously
- Additional simulation within each VM for processor utilization and memory allocation, as defined previously

The number of failover cluster nodes and the distribution of VMs depicted here do not represent a reference configuration, since the memory workload of the VMs saturates the physical host used for the simulation, thus neither of the two hosts can sustain the memory workload of both blocks of VMs in case of an outage of the other node. The two node failover cluster is used as a viable environment to validate the scale out ability of the storage and the hosts.

The selection of a reference failover cluster size is coupled with the number of nodes that are planned to be down in a specific point in time (both for administrative or unpredictable causes). A three node failover cluster, for example, will maintain the availability of the 20 VMs environment with the loss of one single node, or a four node cluster can withstand a two node simultaneous outage.

Table 9 Test parameters: Failover cluster scale out

Reference configuration: test variables under study		
EqualLogic SAN	Two arrays, One group, Two pools (One array dedicated to each pool)	
Volumes and VMs distribution	10 VMs with relative extra data volumes deployed per each pool	
Reference configuration: consistent factors across the iterations of this test		
RAID policy (SAN)	RAID 6	
iSCSI Initiator	Host software initiator, EQL HitKit (VMs startup VHDX volume)	
	Guest software initiator, EQL HitKit (Data volumes)	
Virtual networks (iSCSI)	Two virtual switches, each created on one dedicated 10GbE port	
	Two virtual adapters for VMs with guest initiator, each adapter connected separately to a virtual switch	
Advanced Network Feature in use	SR-IOV and VMQ	
VMs with limited storage footprint (*)	Network services (2) Messaging services, Front-End (2) Content Management services (2) Communication services (2) Web server services (4)	Total of 12
VMs with heavy storage footprint (*)	Messaging services, Back End (2)	Total of 8
	OLTP relational Databases (2)	
	Decision Support System (2)	
	File services, CIFS network share (2)	
Ratio of VMs residing in one SAN volume (CSV enabled)		10:1

* Limited storage footprint refers to no additional data volumes. Heavy storage footprint refers to the situation where more data volumes are required for complex applications and workloads. Refer to Section 3.4 for details.

Figure 25 illustrates the trend of storage KPIs when scaling out the workloads across multiple arrays. The multiple pool selection manifests a proportional progression of the total IOPS required by the workloads that is linearly matched by the KPIs behavior, showing a predictable ability to scale out the cluster nodes.

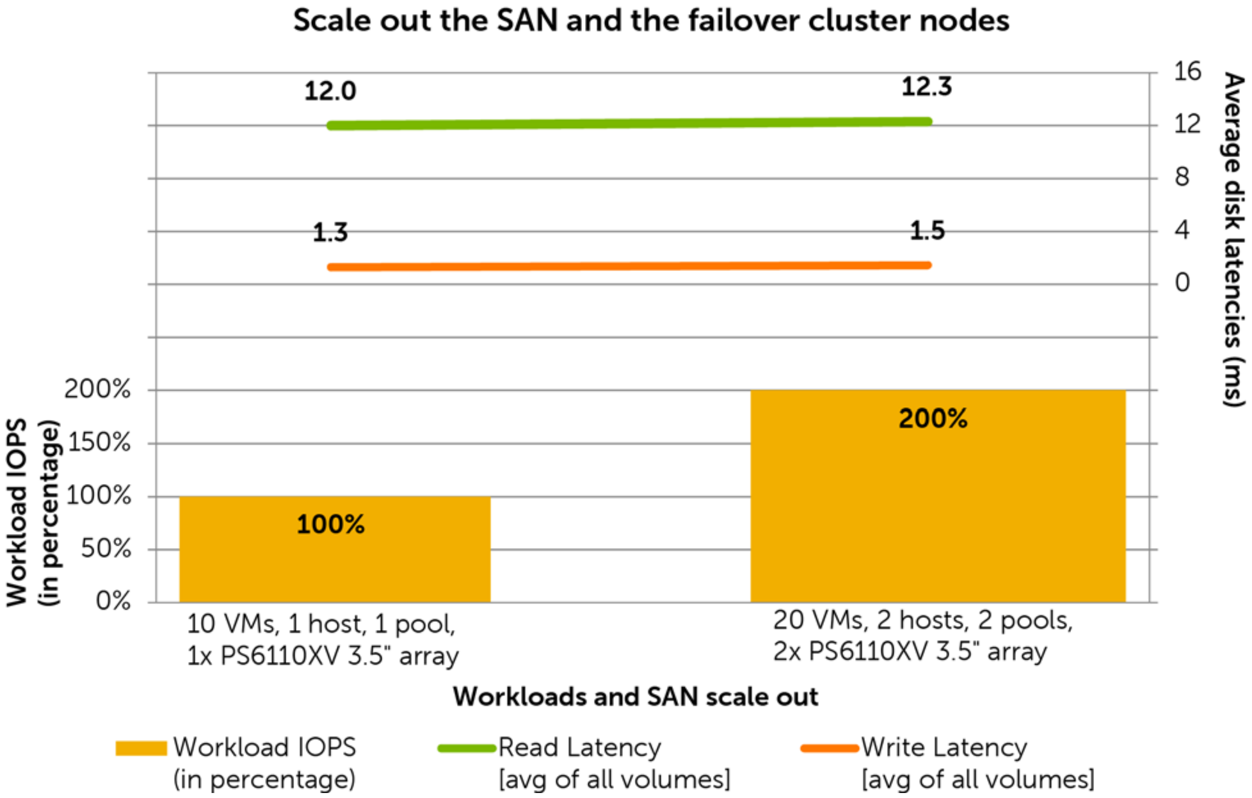


Figure 25 Failover cluster scale out: storage KPIs trend

6 Best practices recommendations

Refer to these best practices to plan and configure EqualLogic arrays, network switches, Hyper-V servers, and VMs.

Storage best practices

- Use a Multipath I/O (MPIO) Device Specific Module (DSM) provided by EqualLogic HIT Kit to improve performance and reliability for the iSCSI connections.
- Distribute network port connections on the controllers according to the port failover guidelines and the redundancy implemented on the network switches.
- Maintain a close as possible to a 1:1 ratio between the number of network ports on the active controller of the array and the number of host network adapters to maximize the utilization of the available bandwidth and to minimize oversubscription.
- Carefully choose the most appropriate RAID policy when designing the environment according to the performance, capacity, and tolerance to failure requirements of your environment.
- RAID conversion provides outstanding benefits when strictly required. Do not overuse it if a clean configuration is possible. RAID conversions of large volumes take a considerable amount of time and leave the resulting set of data pages extremely fragmented with a potential risk of lower efficiency.
- Thin provisioned volumes can ease the burden of tight capacity management. Plan thin provisioned volumes carefully and jointly with an automatic monitoring system with space usage alarm reporting to avoid the risk of overprovisioning.
- An extensive use of thin volumes for systems with a high rate of space usage change and with automatic re-thinning capabilities (Windows Server 2012) can create an excessive fragmentation of the pages allocated for the volumes.
- Do not share the disk drives for active and replicated copies of an application data repository (for example, Exchange Server DAG, SQL Server Always On, Distributed File System, and others). If there is a failure of a set of drives with multiple copies of the same data, the resilience or the perceived availability of the applications would be affected. Dedicate separate EqualLogic pools for replicated data instead.
- Keep the Offloaded Data Transfer (ODX) enabled on the Hyper-V hosts that are connected to the EqualLogic SAN to reduce the computing footprint in the case of a large transfer of VMs.
- For environments with a very large number of volumes attached to the servers, monitor the amount of iSCSI connections per host generated.
- The use of mount points for the SAN volumes simplifies the portability of the storage even within a standalone server.
- Prevent Windows Server from assigning drive letters to volumes by disabling the auto-mount option to minimize unwanted volume letter assignment in a mount point managed environment.
- The deployment of multiple VMs in the same volume simplifies the provisioning of the SAN volumes and the administration of the host connections.
- The deployment of a single VM in a volume provides granular performance and space utilization monitoring from both the host and SAN sides and eases the portability of the volume to other hosts if needed.

Network best practices

- Design separated network infrastructures to isolate the LAN traffic from the SAN traffic (iSCSI).
- Implement redundant components (switches, ISLs, and network adapters) to provision a resilient network infrastructure between the endpoints (stack, LAG, load balancing, or network card teaming).
- Disable spanning tree for the switch ports hosting the PS Series array controllers' connections and enable Rapid Spanning Tree Protocol instead.
- Enable flow control for the switch ports hosting the PS Series array controller connections.
- Enable flow control on the host network adapters dedicated to SAN traffic (iSCSI).
- Enable jumbo frames (large MTU) for the switch ports hosting the PS Series array controller connections.
- Enable jumbo frames on the host network adapters assigned to SAN traffic (iSCSI).
- Evaluate jumbo frames (large MTU) for the LAN network when appropriate (limited by the type of devices the traffic traverses).
- Strongly consider jumbo frames and SR-IOV, if supported, for the network adapters involved in Live Migration operations when in a Failover Cluster configuration.
- Enable Large Send Offload, TCP, and UDP Checksum Offload for both Rx and Tx on the host network adapters connected to the SAN traffic (iSCSI).
- Do not use the network adapters dedicated to the SAN traffic (iSCSI) for cluster communication traffic when in a Failover Cluster configuration.
- Validate the amount of bandwidth to assign for the Live Migration dedicated network. Do not grant a large amount of bandwidth by default if the operation policies do not require it.

Hyper-V and VMs best practices

- Installing a Windows Server Core version in the root partition of the Hyper-V role server is advised when reducing the maintenance, the software attack surface, the memory, and disk space footprint are critical requirements. Otherwise, when installing a traditional Windows Server with Hyper-V technology with the GUI, minimize the use of additional software, components and/or roles in the root partition.
- The use of NUMA is advised to address the management of VMs with large and very large memory settings. Verify the number of NUMA nodes available in the system, based on the number of processors, and then design and size the VMs to have their memory resources entirely contained in a single NUMA node. Spanning a VM memory across multiple NUMA nodes can result in less efficient usage of the memory and can decrease performance.
- Server applications are usually characterized by a memory intensive workload. Configure static memory in the settings of each VM to avoid allowing the dynamic memory management to create contention between different VMs running on the same host, which could penalize the storage I/O execution.
- Evaluate implementing a delayed offset when using Dynamic Memory technology to avoid startup failures at the time of boot or to avoid Smart Paging techniques which can use slow disks for the startup processes.

- Carefully plan the reserve size for the volumes hosting the hard disk files of the VMs (*.vhdx, *.vhd), and include the space required for memory image files (*.bin), saved states (*.vsv), or snapshot files (*.avhdx, *.avhd).
- Remember that the differencing disk image files are discouraged for production environment VMs and also might not be supported by the applications running in the VMs.
- While performance of dynamically expanding disks has improved, fixed disks are preferred to deploy production environment VMs, due to the risk of elevated fragmentation or high latency while disk expansion occurs.
- Standard copy and move operations of large capacity fixed sized disks heavily consume compute resources on the host (CPU, memory, and network). The ODX technology can mitigate this impact by transferring the computing execution to the SAN arrays layer.
- Do not mix both dynamically expanding disks and thin provisioned volumes in the same deployment. In cases with space usage challenges, thin provisioned volumes are recommended.
- Isolate the host management traffic from the VM traffic by using virtual switches not enabled for management.
- Select VM network bus adapters with synthetic drivers as opposed to legacy network adapters with emulated drivers.
- Avoid mixing LAN and iSCSI traffic on the same virtual adapters and enforce this with the LAN and iSCSI network isolation design.
- Configure a dedicated virtual switch for each network adapter connected to the SAN traffic (iSCSI). Maintain as close as possible to a 1:1 ratio between the number of network ports on the active array controller and the number of host network adapters configured.
- Select CSV file system for the shared storage in a Failover Cluster to allow Live Migration to move VMs independently even if deployed in the same volume.
- Use dedicated network adapters or guaranteed bandwidth for Live Migration networks in a failover cluster.
- Plan the operational practices to use Live Migration features, including how many VMs would be moved simultaneously.
- If using the ‘Intelligent Placement’ feature in SCVMM, carefully monitor the automatic movement of VMs between the nodes of the Failover clusters, especially when selecting a highly aggressive dynamic optimization.

For general recommendations and information about EqualLogic PS Series array configurations, refer to Dell EqualLogic Configuration Guide, available at:
http://www.delltechcenter.com/page/EqualLogic+Configuration+Guide

A Configuration details

A.1 Hardware components

Table 10 lists the details of the hardware components used for the configuration.

Table 10 Hardware components

Component	Description
Servers	Dell PowerEdge R620 server, Firmware 1.6.0 • 2x Eight Core Intel Xeon E5-2665 Processors, 2.4 Ghz, 20M Cache • RAM 32 GB (4x 8GB) • iDRAC7 Enterprise, Firmware 1.37.35 • PERC H710 Mini RAID controller, Firmware 21.1.0-0007 • 4x 146 GB 15K SAS (2x RAID-1, stripe 1MB) • 4x Broadcom NetXtreme 5720 Quad Port 1GbE Base-T onboard, Firmware 7.4.8 • 2x Broadcom NetXtreme II 57810 Dual Port 10GbE Base-T, Firmware 7.4.8 2x Dell PowerEdge R720 servers, Firmware 1.6.0 • 2x Eight Core Intel Xeon E5-2665 Processors, 2.4 Ghz, 20M Cache • RAM 192 GB (24x 8GB) • iDRAC7 Enterprise, Firmware 1.37.35 • PERC H710 Mini RAID controller, Firmware 21.1.0-0007 • 4x 146 GB 15K SAS (2x RAID-1, stripe 1MB) • 4x Broadcom NetXtreme 5720 Quad Port 1GbE Base-T onboard, Firmware 7.4.8 • 2x Broadcom NetXtreme II 57810 Dual Port 10GbE Base-T, Firmware 7.4.8
Network	2x Dell Force10 S4810 Ethernet switches, Firmware 8.3.12.0 • 48x 10GbE interfaces • 4x 40GbE interfaces • Installed top of the rack • Connected by 2x 40GbE redundant uplinks (LAG) Dell PowerConnect 6248 Ethernet switch, Firmware 3.3.6.4 • 48x 1GbE interfaces • 4x 10GbE interfaces • Installed top of the rack
Storage	2x Dell EqualLogic PS6110XV 3.5” • Storage Array Firmware 6.0.4 • Dual 1 port 10GbE controllers • Dual 1 port 1GbE management interface • 24x 600GB 15K 3.5” SAS disk drives, raw capacity 14.4 TB (each unit)

A.2 Software components

The environment required to perform the simulations in this paper included the following software components:

- Hypervisor: Windows Server 2012 with Hyper-V on every physical host
- Dell OpenManage Server Administrator on every physical host
- Broadcom Advanced Control Suite on every physical host
- Operating System: Windows Server 2012 or 2008 R2 on every VM
- Dell EqualLogic Host integration Toolkit to provide Dell MPIO access to the back-end SAN on each physical or VM directly accessing the SAN
- Dell EqualLogic SAN Headquarters to monitor the health and performance of the SAN
- Microsoft Exchange Jetstress to simulate the access to the storage subsystem from the mailboxes store simulated VM
- Microsoft File Server Capacity Tool to simulate the client-server activities on a CIFS file share, from multiple loading clients to the file server VM
- Iometer measurement and characterization tool to simulated disk activities under specific profile definitions on all the remaining VMs
- Processor load simulation tool to artificially generate computing workload on each VM
- Memory allocation simulation tool to artificially hold memory resources on each VM

The following software components are installed and configured to simplify the management of the environment and to support the failover cluster configuration:

- Active Directory Domain Services and DNS Server roles for the domain controller VM
- Microsoft SCVMM for the management VM (not strictly required to accomplish the tests)

Table 11 lists the details of the software components used for the configuration.

Table 11 Software components

Component	Description
Operating Systems	Host servers: Microsoft Windows Server 2012 Datacenter Edition (build 9200) with Hyper-V • Dell OpenManage Server Administrator 7.2.0 • Broadcom Advanced control Suite 4 (version 15.5.7.0) • Dell EqualLogic Host Integration Toolkit 4.5.0 (host initiator connectivity only) • MPIO enabled using EqualLogic DSM for Windows Guest VMs: Microsoft Windows Server 2012 Datacenter Edition (build 9200) Microsoft Windows Server 2008 R2 datacenter Edition Service Pack 1 (build 7601) • Hyper-V Integration Services (build 16384) • Dell EqualLogic Host Integration Toolkit 4.5.0 (only for guest initiator connectivity) • MPIO enabled using EqualLogic DSM for Windows
Applications	Microsoft System Center 2012 Virtual Machine Manager Service Pack 1 (version 3.1.6011.0)
Monitoring tools	Dell EqualLogic SAN Headquarters 2.5 (build 2.5.0.6470) Microsoft Performance Monitor from the Windows Operating System
Simulation tools	Microsoft Exchange Jetstress 2010 (build 14.01.0225.017) • Exchange 2010 Server Database Storage Engine and Library Service Pack 2 (build 14.02.0247.001) Microsoft File Server Capacity Tool version 1.2 Iometer measurement and characterization tool version 2006.07.27

A.3 Network configuration

Two physical networks provide full isolation between regular IP traffic and iSCSI data storage traffic. Also, each IP network is segregated from the others by VLANs with tagged traffic. In order to achieve network resiliency to hardware faults, at least two physical switches are stacked for the iSCSI data storage network, as well as using redundant uplinks (LAG) between the switches. Some relevant configuration aspects are:

- Flow control enabled for every port on S4810 switches
- Rapid Spanning Tree Protocol enabled for every edge port on S4810 switches
- Jumbo frames enabled for every port on S4810 and 6248 switches

Table 12 and Table 13 summarize the aspects of the physical and logical networks implemented in the reference architecture and their purpose.

Table 12 Network configuration: network switch, purpose, and networks

Network Switch	Placement	Purpose	VLAN ID
PowerConnect 6248	Top of rack	IP traffic – LAN and Management	100
		IP traffic – Cluster communication	200
		IP traffic – Live Migration	300
Force10 S4810 #1	Top of rack	iSCSI data storage traffic	1 (default)
Force10 S4810 #2	Top of rack	iSCSI data storage traffic	1 (default)

Table 13 Network configuration: host to switch connections

Server	Interface	NIC ports	Purpose	
PowerEdge R620	BCM5720 #1	1x 1GbE	LAN and Management	IP traffic
	BCM5720 #2	1x 1GbE	VM machines traffic	
	BCM5720 #3	1x 1GbE	Available for teaming	
	BCM5720 #4	1x 1GbE	Available for teaming	
	BCM57810 #1	1x 10GbE	iSCSI management	iSCSI data storage traffic
	BCM57810 #2	1x 10GbE	Available for teaming/MPIO	
	Total ports = 6 (4 onboard 1GbE, 2 slot 10GbE)			
PowerEdge R720	BCM5720 #1	1x 1GbE	LAN and Management	IP traffic
	BCM5720 #2	1x 1GbE	VM machines traffic	
	BCM5720 #3	1x 1GbE	Cluster communication	
	BCM5720 #4	1x 1GbE	Live Migration	
	BCM57810 #1	1x 10GbE	SAN access	iSCSI data storage traffic
	BCM57810 #2	1x 10GbE	SAN access	
	BCM57810 #3	1x 10GbE	SAN access	
	BCM57810 #4	1x 10GbE	SAN access	
	Total ports = 8 (4 onboard 1GbE, 4 slot 10GbE)			

A.4 Host hypervisor and VMs configurations

A virtual infrastructure built on Windows Server with Hyper-V hosted all the components of the test infrastructure. The primary elements of the virtual infrastructure configuration are:

- Windows Server 2012 with Hyper-V deployed on all hosts, managed by the Hyper-V Role Administration tools or centrally by the SCVMM server
- Up to two identical hypervisor hosts (R720s) configured as member server of the domain
- One hypervisor host (R620) configured as member server of the domain
- All guests deployed from two image templates (one per OS version) of Windows Server 2008 R2 and Windows Server 2012 operating systems
- Guest iSCSI initiator used to access volumes hosted on the EqualLogic SAN for the VMs with extra data volumes, unless otherwise specified during the specific test

Table 14 lists the relation between each hypervisor host and its respective set of VMs, with a brief summary of the virtual resources allocated for each VM.

Table 14 Configuration: guests to host placement

Host	VM	Purpose	vCPU	Memory	Storage	Network Adapters *
R620	DC01	Active Directory Domain Controller	2	4GB	25GB VHDX	1x VMBus Net Adapter
	LOAD01	FSCT simulation	2	2GB	25GB VHDX	1x VMBus Net Adapter
	LOAD02	FSCT simulation	2	2GB	25GB VHDX	1x VMBus Net Adapter
	VMM01	System Center Virtual Machine Manager	2	8GB	50GB VHDX	2x VMBus Net Adapter
R720 #1	NET01	Iometer simulation (Network systems)	2	4GB	50GB VHDX	1x VMBus Net Adapter
	EXCH01	Iometer simulation (Messaging FE)	4	8GB	50GB VHDX	1x VMBus Net Adapter
	MBX01	Jetstress simulation (Mailboxes store BE)	4	32GB	60GB VHDX 4x 350GB	3x VMBus Net Adapter
	SQL01	Iometer simulation (OLTP databases)	4	24GB	50GB VHDX 1x 50GB 1x80GB 3x200GB	3x VMBus Net Adapter
	DSS01	Iometer simulation (DSS reporting)	4	48GB	70GB VHDX 1x 50GB 1x 80GB 1x 900GB	3x VMBus Net Adapter
	FS01	FSCT simulation (CIFS folders sharing)	2	16GB	40GB VHDX 1x 1.5TB	3x VMBus Net Adapter
	SPS01	Iometer simulation (Sharepoint services)	4	8GB	50GB VHDX	1x VMBus Net Adapter
	LYNC01	Iometer simulation (Communication svc)	2	8GB	50GB VHDX	1x VMBus Net Adapter
	WEB01	Iometer simulation (Web Server services)	4	8GB	50GB VHDX	1x VMBus Net Adapter
	WEB02	Iometer simulation (Web Server services)	4	8GB	50GB VHDX	1x VMBus Net Adapter
R720 #2	Case #1: Failover Cluster node for the same 10 VMs reported above					
	Case #2: Failover Cluster node for additional 10 VMs, configured equally as the 10s above					

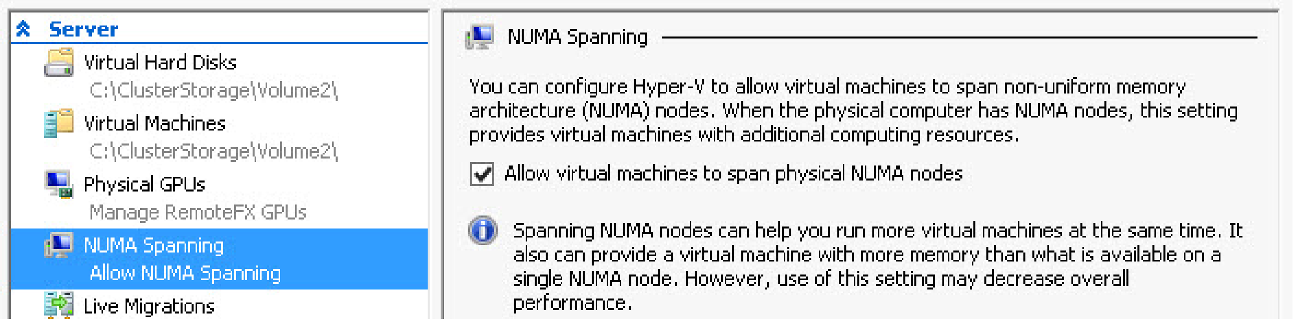
* All the network adapters configured in the VMs used VMBus network adapters with synthetic drivers as opposed to legacy network adapters with emulated drivers.

Guest VMs memory

The memory assigned to every VM in the infrastructure is configured as Static, to avoid any possible occurrence of VMs competing for the same resources.

Hyper-V configuration of NUMA

Non-Uniform Memory Access (NUMA) capabilities is enabled on the PowerEdge R720 physical hosts (Node Interleaving disabled in the server BIOS) to allow memory access across CPUs: these R720s have two NUMA nodes each managing 96GB of memory. The Hyper-V NUMA Spanning setting is left enabled (by default) as shown in Figure 26.



You can configure Hyper-V to allow virtual machines to span non-uniform memory architecture (NUMA) nodes. When the physical computer has NUMA nodes, this setting provides virtual machines with additional computing resources.

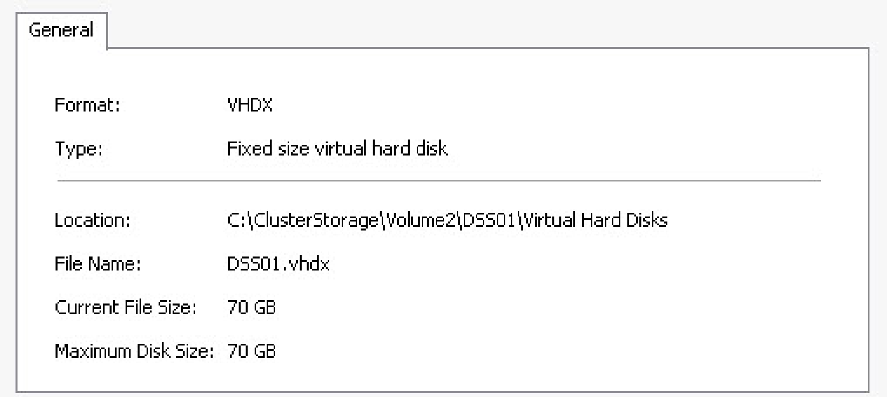
☑ Allow virtual machines to span physical NUMA nodes

Spanning NUMA nodes can help you run more virtual machines at the same time. It also can provide a virtual machine with more memory than what is available on a single NUMA node. However, use of this setting may decrease overall performance.

Figure 26 Hyper-V setting for NUMA

Guest VMs disks

The virtual disks in use to host the operating system of each VM are VHDX fixed type disks, similar to the configuration shown in Figure 27. The VHDX files for the VMs hosted from the R620 hypervisor are deployed on the second pair of local RAID-1 disks, all the VMs hosted by the R720 hypervisors are deployed on the SAN as described above.



Format:	VHDX
Type:	Fixed size virtual hard disk
Location:	C:\ClusterStorage\Volume2\DSS01\Virtual Hard Disks
File Name:	DSS01.vhdx
Current File Size:	70 GB
Maximum Disk Size:	70 GB

Figure 27 Fixed size VHDX files

Host network adapters and virtual network configuration

The host network adapters providing connectivity for the hosts and the VMs are configured as listed.

- One physical network adapter, sourced from the onboard Broadcom 1GbE ports, providing connectivity for host domain access and management
- One physical network adapter, sourced from the onboard Broadcom 1GbE ports, providing connectivity for VMs, both domain and intra-VM traffic
- One physical network adapter, sourced from the onboard Broadcom 1GbE ports, providing intra-nodes cluster communication for the failover cluster configuration
- One physical network adapter, sourced from the onboard Broadcom 1GbE ports, providing dedicated bandwidth to the Live Migration feature for the failover cluster configuration
- One (on the R620) or four (on the R720s) physical network adapters, sourced from the Broadcom 10GbE ports, individually connected to the iSCSI network, and providing access to the SAN from the host or guest environments depending on the test case configuration

Virtual network adapter configuration

The assignment of the virtual network adapters of the VMs is configured as listed below:

- One virtual network adapter to access the LAN traffic VLAN, for each VM
- One virtual network adapter to access the dedicated management subnet, for the SCVMM VM

Additional network adapters connected to the iSCSI network to access the SAN during the guest initiator test cases, and relevant configurations.

- Two virtual network adapters to access the iSCSI network for the VMs directly accessing the SAN resources
- MPIO (Multi-path I/O) enabled and provided by EqualLogic DSM module
- Jumbo frames enabled
- Large Send Offload enabled
- Send and Receive Buffer Size maximized to 8MB
- IP, TCP and UDP Checksum Offload enabled for both Rx and Tx

B Load simulation tools considerations

B.1 Microsoft Jetstress

Microsoft Exchange Server Jetstress 2010 is a simulation tool that reproduces the database and logs I/O workload of an Exchange mailbox database role server. It is usually used to verify and validate the conformity of a storage subsystem solution before the full Exchange software stack is deployed. Some elements to consider about Microsoft Jetstress are:

- Does not require and should not be hosted on a server where Exchange Server is running.
- Performs only Exchange storage access and not host processes simulations. It does not contribute in assessing or sizing the Exchange memory and processes footprints.
- Is an ESE application requiring access to the ESE dynamic link libraries to perform database access. It takes advantage of the same API used by the full Exchange Server software stack and as such it is a reliable simulation application.
- Runs on a single server. When a multiple servers' simulation is required, the orchestration of the distributed instances has to be hosted by external management tools.
- Requires, and provides, an initialization step to create and populate the database/s that will be used for the subsequent test phases. The database/s should be planned of the same capacity as the one/s planned for the Exchange Server future deployment.
- Its topology layout includes number and size of simulated mailboxes, number and placement of databases and log files, and number of database replica copies (simulates only active databases).
- While carrying out a mailbox profile test, it executes a pre-defined mix of insert, delete, replace, and commit operations against the database objects during the transactional step, then it performs a full database checksum.
- Collects Application and System Event Logs and performance counter values for both operating system resources and ESE instances. It then generates a detailed HTML-based report.
- Throttles the disk I/O generation using the assigned IOPS per mailbox, thread count per database, and SluggishSessions threads property (fine tuning for threads execution pace).

B.2 Microsoft File Server Capacity Tool

Microsoft File Server Capacity Tool is a client-server simulation tool developed to test file server (CIFS/SMB) capacity and to identify performance bottlenecks. It generates Win32 API calls to simulate the behavior of Microsoft Office applications, command line operations, and Windows Explorer.

- Is based on preconfigured test scenarios that could be expanded or modified to mimic specific behaviors (the HomeFolders workload is provided as simulation of users' home directories)
- Is based on a three layers architecture: controller, clients, and server (file server shares)
- The client/s perform multiple operations and simulate multiple sessions or active users in parallel
- The controller synchronizes the clients' activities and collects test results

Additional resources

Support.dell.com is focused on meeting your needs with proven services and support.

DellTechCenter.com is an IT Community where you can connect with Dell Customers and Dell employees for the purpose of sharing knowledge, best practices, and information about Dell products and your installations.

Referenced or recommended Dell publications:

- Dell EqualLogic Configuration Guide:
  http://en.community.dell.com/dell-groups/dtcmedia/m/mediagallery/19852516/download.aspx
- Dell EqualLogic Group Manager Administrator's Manual
  https://eqlsupport.dell.com/support/download_file.aspx?id=1630&langtype=1033
- Dell PowerConnect 6248 Documentation
  http://www.dell.com/support/Manuals/us/en/19/Product/powerconnect-6248
- Dell Force10 S4810 Documentation
  http://www.dell.com/support/Manuals/us/en/19/Product/force10-s4810
- Dell PowerEdge R720 Documentation
  http://www.dell.com/support/Manuals/us/en/19/Product/poweredge-r720

Referenced or recommended Microsoft publications:

- Windows Server 2012 Hyper-V Overview
  http://technet.microsoft.com/en-us/library/hh831531.aspx
- Understanding Exchange 2010 Virtualization
  http://technet.microsoft.com/en-us/library/jj126252%28EXCHG.141%29.aspx
- Microsoft Exchange Server Jetstress 2010 (64 bit)
  http://www.microsoft.com/en-us/download/details.aspx?id=4167#tm
- Microsoft File Server Capacity Tool v1.2 (64 bit)
  http://www.microsoft.com/en-us/download/details.aspx?id=27284

Referenced or recommended independent publications:

- Iometer project
  http://www.iometer.org/

For EqualLogic best practices white papers, reference architectures, and sizing guidelines for enterprise applications and SANs, refer to Storage Infrastructure and Solutions Team Publications at:

- http://dell.to/sM4hJT