Dell EMC SC5020 with Heterogeneous Virtualized Workloads on Microsoft Windows Server 2016 with Hyper-V

Abstract

This document provides best practices for configuring Microsoft® Windows Server® 2016 Hyper-V® and Dell EMC™ SC5020 storage with heterogeneous application workloads.

October 2017

Revisions

Date	Description
August 2016	Initial release
October 2017	Refreshed content for Windows Server 2016 with Hyper-V and SC5020

Table of contents

Executive summary

Virtualization technologies introduce many variables into 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 simulate a reference applications ecosystem designed to fulfill the requirements of a small- to medium-sized 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 Microsoft® Windows Server® 2016 with Hyper-V®. The tests assessed the impact of different configurations: configurations using advanced networking features, configurations with 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.

Audience

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, or managing Microsoft virtual infrastructures supporting heterogeneous applications and who would like to investigate the benefits of using Dell EMC™ SC5020 storage. This document assumes the reader is familiar with Microsoft Windows Server 2016, SC Series SAN operation, and Microsoft Hyper-V architecture and system administration. The scope of this paper is restricted to a local data center topology and does not include specific or detailed server sizing information.

We welcome your feedback along with any recommendations for improving this document. Send comments to StorageSolutionsFeedback@dell.com.

1 Introduction

Dell EMC SC Series arrays have supported Microsoft Hyper-V since the release of Microsoft Windows Server 2008, and many new feature enhancements have been added to both Hyper-V and SC Series storage over time. SC Series arrays are designed from the ground up with redundancies to avoid downtime for such events as component failures, maintenance, upgrades, and expansion. These redundancies benefit Hyper-V and other workloads as well. Hyper-V also includes similar features. When SC Series arrays host Hyper-V workloads, their feature sets complement each other.

There are several other means of presenting storage to Hyper-V, such as traditional on-board storage in host servers. Another method involves Microsoft Storage Spaces, which Microsoft has promoted for Hyper-V platforms since the release of Windows Server 2012 R2. Storage Spaces Direct and Storage Replica were introduced with Windows Server 2016. While Storage Spaces, Storage Spaces Direct, and Storage Replica offer similar basic features as traditional SAN arrays, the SC Series SAN provides a much more powerful and complete set of integrations, management, and monitoring tools that are not available with Storage Spaces or Storage Replica alone.

Determining best practices is often subject to many subtle nuances and factors that vary based on the customer environment and personal preference. These include budgetary constraints, complexity, required depth of knowledge, service-level agreements, and regulations to name a few. What works well for one environment may not work well for another. On occasion, the configuration a customer choses might not be completely harmonious with typical best practices. For example, it might be perfectly acceptable to omit costly design elements and redundancies in order to save money in a test, development, or DR environment that can suffer down time without significantly impacting the business. When different configuration options are possible, determining the best design is ultimately up to the customer to decide.

1.1 Terminology

The following definition list describes the terms used throughout this document.

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

Parent and child partitions: 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 the 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 transfers 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 processor (the local memory, accessed by channels dedicated to each processor) as opposed to symmetric multiprocessing (SMP) where memory is shared between processors.

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, a volume, an application, a file, a database entry, or a user. A GUID is commonly displayed as a 32-digit hexadecimal string.

2 Solution architecture and workload overview

2.1 Accelerate your workloads, automate your savings

The SC5020 array makes storage cost savings automatic with a modern architecture that optimizes the data center for economics while delivering transformational SSD, HDD, or hybrid performance.

SC Series storage provides the lowest effective cost per GB for flash and hybrid flash[1], giving organizations of any size the technology advantage needed to compete in today's fast-changing markets. Highlights include:

- **Data Progression:** Achieve IOPS goals with the least-expensive mix of storage media, even as performance needs evolve.
- **Deduplication and compression:** Dramatically reduce the raw capacity required to store data.
- **RAID tiering:** Eliminate manual RAID provisioning, and increase efficiency and utilization.
- **Federation:** Simplify multi-array environments with quick and seamless data movement, plus proactive load balancing assistance using Live Migrate and Volume Advisor.
- **Dell ProSupport™ services:** Reduce deployment costs with remote installation options that ensure the project is successful the first time.
- **Persistent software licensing**: Future-proof the investment, and minimize the cost of upgrades and expansions.

2.2 All-new hardware platform

Designed as the next-generation successor to the popular SC4020 array, the SC5020 array is a performance powerhouse. With dual 8-core Intel® processors, 4x more memory, and a 12Gb SAS back end, the SC5020 delivers:

- Up to 45% more IOPs[2]
- Up to 3x more bandwidth[2]
- 2x greater maximum capacity

The new 3U all-in-one chassis includes 30 drive bays plus dual hot-swappable controllers, providing up to 460 TB raw capacity in a single compact unit. A variety of expansion enclosures enables scaling up to 2 petabytes (2 PB) per array — with even larger scale-out potential in federated multi-array systems. In addition to fast hardware, the SC5020 includes all of the Dell Storage Center OS (SCOS) features to be expected from SC Series storage.

---

[1] Net usable capacity of Dell array with 5 years of support, after 4:1 data reduction, vs. major competitors net of data reduction. Street price analysis is based on a variety of sources including analyst data, price sheets when available, and public information as of January 2017.
[2] Based on April 2017 internal Dell EMC testing, compared to previous-generation SC4020. Actual performance will vary depending upon application and configuration.

2.3 Microsoft Windows Server 2016 with Hyper-V

Microsoft Hyper-V technology is implemented on the software layer as a 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 migration and transformation from one to another
- Improving the availability and resiliency of services by simplifying resource displacement and management of physical and virtual resources with tailored management tools

Since its introduction, Hyper-V has matured across multiple generations (Windows Server 2008 through Windows Server 2016), achieving a broad set of features and enhancements with the current version.

Enhancements of the latest version of Hyper-V include:

- **Discrete device assignment**: Using a device in this way bypasses the Hyper-V virtualization stack, which results in faster access.
- **Host resource protection:** This feature helps prevent a virtual machine from using more than its share of system resources.
- **Hot add and remove for network adapters and memory:** This allows adding or removing a network adapter while the virtual machine is running.
- **More memory and processors for generation 2 VMs and Hyper-V hosts:** Starting with version 8, generation 2 virtual machines can use significantly more memory and virtual processors. Hosts also can be configured with significantly more memory and virtual processors than were previously supported.
- **Remote direct memory access (RDMA) and switch embedded teaming (SET):** Set up RDMA on network adapters bound to a Hyper-V virtual switch, regardless of whether SET is also used. SET provides a virtual switch with some of same capabilities as NIC teaming.
- **Virtual machine multi queues (VMMQ):** This improves on virtual machine queue (VMQ) throughput by allocating multiple hardware queues per virtual machine. The default queue becomes a set of queues for a virtual machine, and traffic is spread between the queues.
- **Quality of Service (QoS) for software-defined networks:** This manages the default class of traffic through the virtual switch within the default class bandwidth.

More detail can be found on the Microsoft page, What's new in Hyper-V on Windows Server 2016.

Hyper-V technology is installed in Windows Server 2016 as a role and provides management tools (both GUI-based tools and PowerShell cmdlets), a management service, a virtual machine bus (VMbus), a virtualization service provider (VSP), and virtualization infrastructure drivers (VID).

Hyper-V includes a software package (integration services) for the guest operating systems designed to improve integration between the physical and virtual layers. A guest Windows Server 2016 operating system does not currently require the installation or update of the integration services.

2.3.1 Generation 1 and generation 2 Hyper-V guest VMs

With Windows Server 2012 R2 and newer, a guest VM can be designated as either a generation 1 (gen 1) or generation 2 (gen 2) guest.

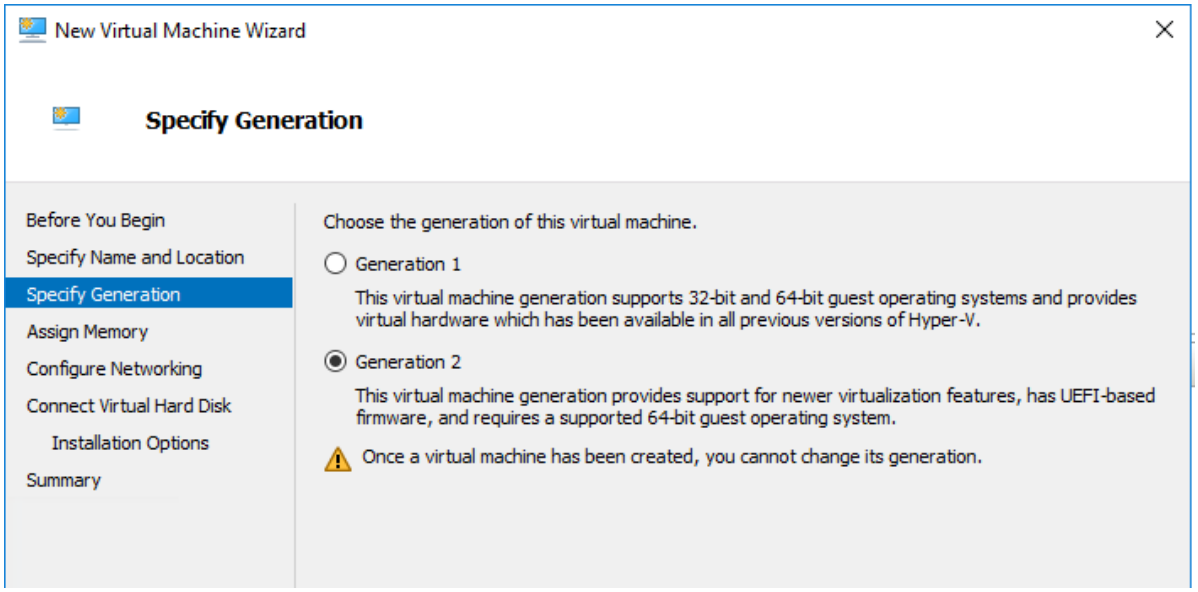


Figure 1 Specify a guest as gen 1 or gen 2

One of the most significant changes with gen 2 guests since the release of Windows Server 2012 R2 Hyper-V is the elimination of the dependency on virtual IDE for boot disks. IDE is not even a supported option with gen 2 guests.

For gen 1 guest VMs, each Hyper-V guest VM supports a maximum of only two virtual IDE controllers. Each of the two virtual IDE controllers supports a maximum of 2 virtual IDE devices, for a total of 4 virtual IDE devices per guest VM.

Each gen 1 Hyper-V guest VM also supports up to 4 virtual SCSI controllers with up to 64 devices per controller, for a maximum of 256 virtual SCSI devices per guest VM.

2.3.2 Virtual storage devices, disk formats, and disk types

Each guest VM running on Windows Server 2016 with Hyper-V has access to storage by one or more of the following devices:

- Direct-attached virtual storage by virtual SCSI or virtual IDE 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)

The following provides a detailed explanation of these devices:

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).

Virtual IDE controller is an emulated IDE disk controller (maximum of two) that supports 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 image to the VM, leaving three additional slots for virtual hard drives.

Boot from virtual disk (startup disk) in a gen 1 Hyper-V VM that requires an IDE device. SCSI devices are required for gen 2 VMs.

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:

Virtual hard disk (VHD) is a file format that represents a hard disk image. A VHD file is composed of sectors of 512 bytes each, and is addressed by a 32-bit table which allows a maximum addressable size of 2 TB (2,040 GB). 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.

Virtual hard disk extended (VHDX) is the VHD-enhanced file format representing a hard disk image, and is supported on the latest generation of Microsoft Hyper-V since Windows Server 2012. The VHDX format supports storage capacity up to 64 TB by using 4 KB 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 1 MB (from 512 KB).

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 or Live Storage Migration), and cannot be used as targets for the Hyper-V VSS provider in order to back up or take a snapshot of 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 preallocated upfront. The size is kept constant regardless of the amount of data that is added or deleted. 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 or type.

Dynamically expanding disks are small when created and grow as data is added. They allow overprovisioning 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 subject to high fragmentation that could cause slower performance due to increased latency during read activities or during 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 or type.

Differencing disks are intended for the specific purpose of parent-child relationship with another disk that should stay intact, acting as an 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 or type.

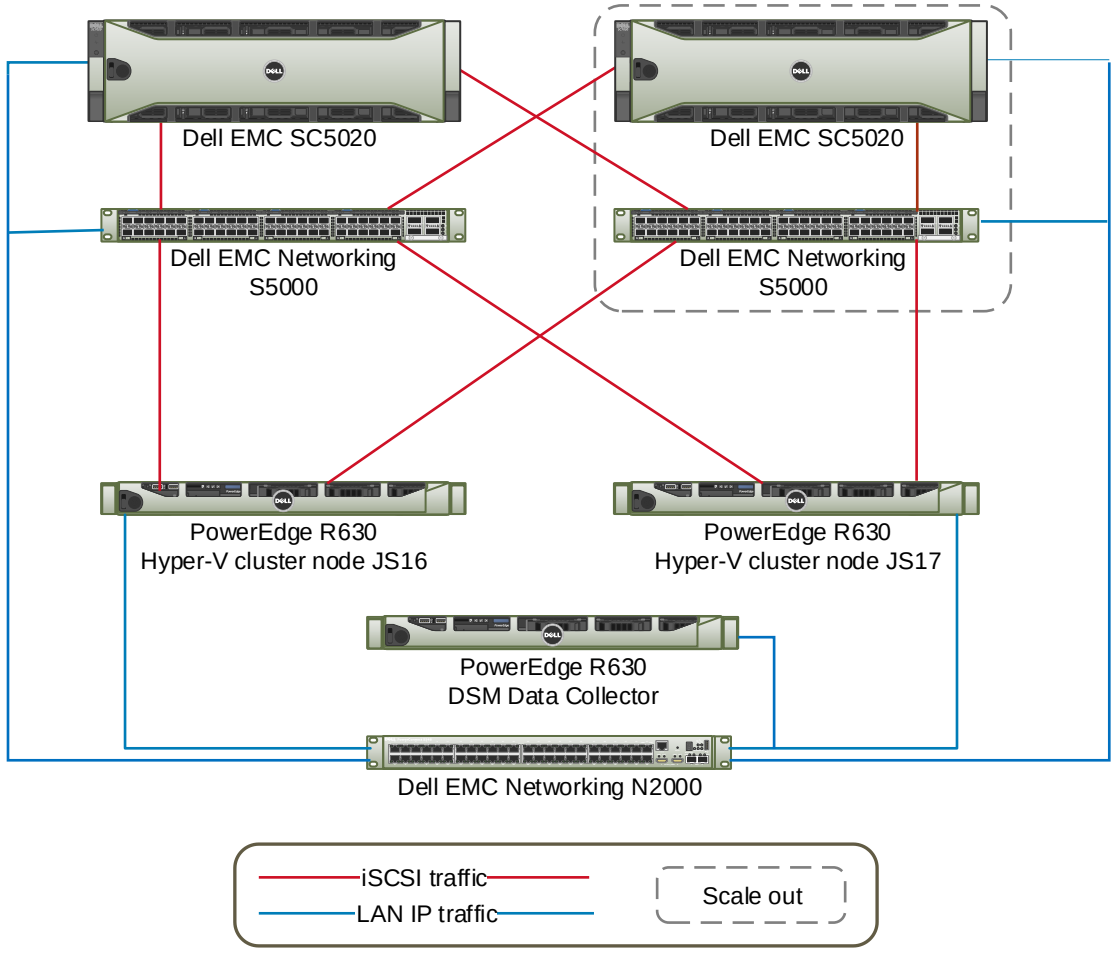


Figure 2 Physical lab design

Physical components of the solution include:

- Two Dell EMC PowerEdge™ R630 rack servers powering the hypervisors beneath the simulated application infrastructure in a dual-node Hyper-V failover cluster configuration
- One PowerEdge R630 rack server powering the hypervisor beneath the Dell Storage Manager and monitoring infrastructure
- One SC5020 Fibre Channel SAN provisioned with two 2.5-inch controller arrays (16Gb)
- One Dell EMC Networking N2000 Ethernet switch supporting LAN IP traffic
- One (two recommended) Dell EMC Networking S5000 Ethernet switches supporting the iSCSI data storage traffic on the SAN side

2.3.3 Storage disks and volumes layout

The SC5020 array and the volumes underlying the VMs and application data are distributed as:

Standalone Hyper-V configuration:

- A dedicated set of data volumes dedicated to the databases or files of each application (Microsoft Exchange, SQL Server®, DSS, and file services), uniquely assigned to the relevant VM hosting the simulation for the specific application
- Two disk storage tiers (tier 1: 1.92 TB read-intensive SSD; tier 3: 1 TB 7.2K HDD)
- Hyper-V hypervisor using two Fibre Channel adapters with Windows MPIO connection to the block storage on the SC5020

Hyper-V Windows failover cluster configuration:

- Hyper-V hypervisor using Windows MPIO-enabled Fibre Channel connections to clustered shared volumes (CSVs) on the SC5020
- Two disk storage tiers (tier 1: 1.92 TB read-intensive SSD; tier 3: 1 TB 7.2K HDD)
- Volumes storing CSVs with the storage profile, **Recommended – All Tiers**, for all volumes
- One CSV to store all Windows boot virtual disks for all VMs
- Additional volumes for classes of storage workloads: Exchange, SQL Server, file, and Microsoft SharePoint® and web services configured as CSVs

2.3.4 Disk storage assigned

Dell Storage Manager (DSM) shows the two tiers of disks installed into the SC5020 array. 30 total drives are present, filling the 2.5-inch internal storage bays. A 512 KB storage page size is defined on this system. See Figure 3.

Figure 3 DSM showing storage tiers assigned

2.3.5 Volume usage

All storage can be portrayed in an informative bar chart (Figure 4) in DSM.

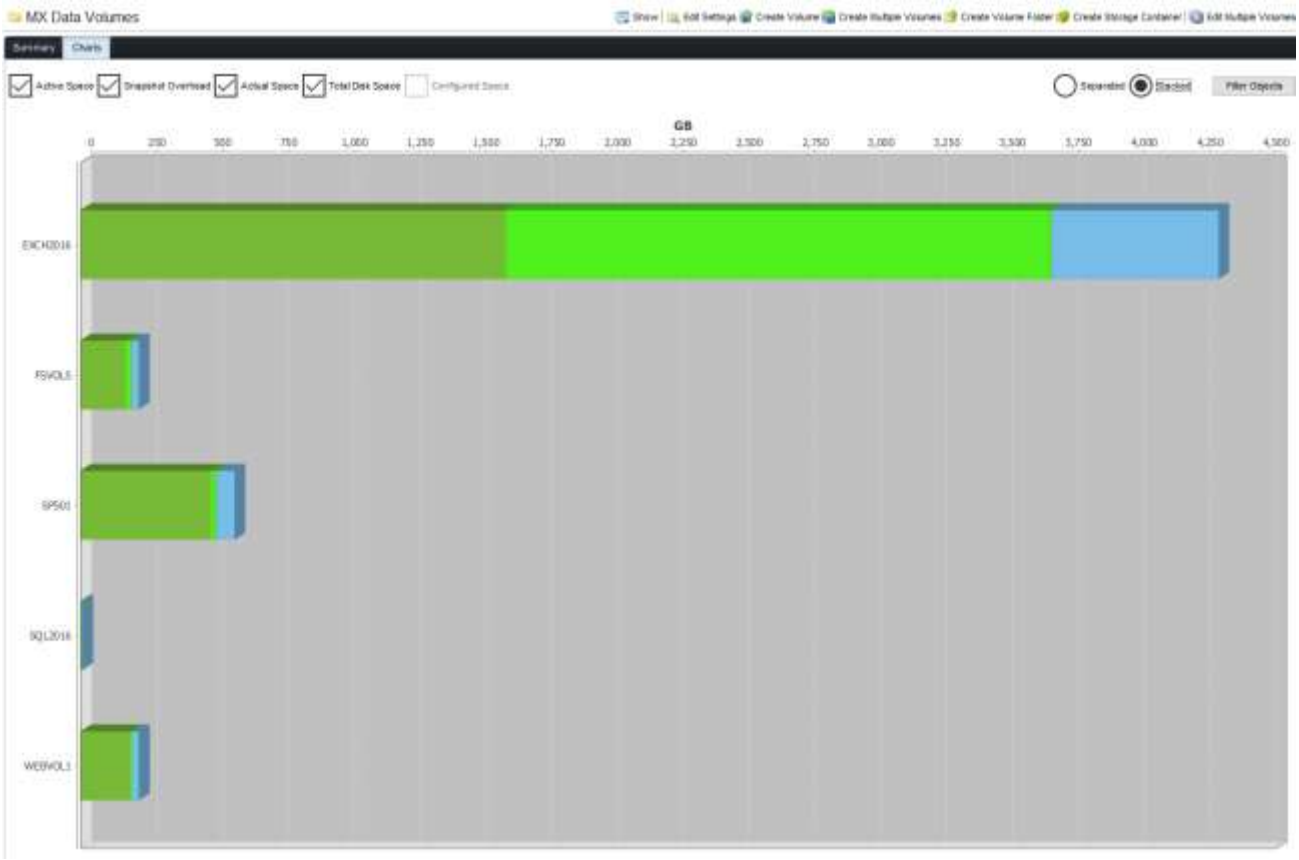


Figure 4 DSM storage bar chart view

2.3.6 Storage tiering employed for performance/capacity balance

All volumes in this mixed workload study use the **Recommended** (DSM default) storage profile that automatically tiers storage blocks to different RAID and tier levels based upon data access patterns. Figure 5 shows an example of how this heterogeneous workload's data has been tiered across the different RAID levels and tiers to provide both space efficiency and premier performance.

Note: Because tier 1 is SSD technology, the Dell Fast Track feature cannot be utilized there.

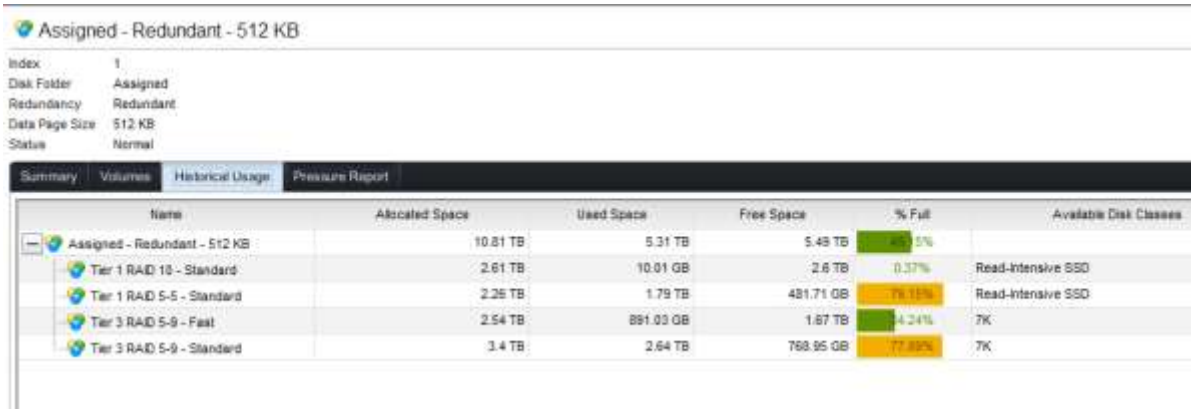


Figure 5 The assigned redundant 512KB storage type shows storage tiering

2.4 Space savings on file server volumes using SC Series data reduction

Compression is enabled on these volumes to demonstrate the file server data at rest in tier 3 shows over 82 percent savings from the File Server Capacity Tool (FSCT) test data.

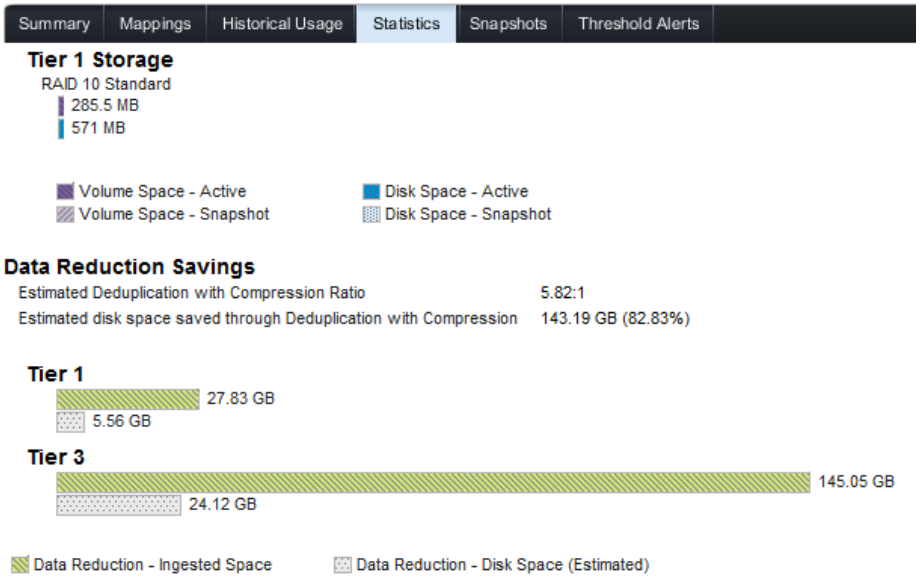


Figure 6 Data reduction savings on simulated file server data from FSCT

For more information on SC Series data reduction features see the document, *Dell Storage Center OS 7.0 Data Reduction with Deduplication and Compression.*

3 Standalone Hyper-V implementation factors

The standalone study focuses on the fundamental configuration and deployment variables that can affect the storage performances of the virtual infrastructure and applications described in section 2. The test alters one variable for each test scenario to better assess the potential value of the configuration.

Assessing the standalone hypervisor configurations provides the baseline for a more complex configuration when scaling this building block horizontally into a failover cluster. Additional scenarios unique to cluster configurations are examined in section 4.

The following constants were maintained across the tests:

- Eight VMs are started with preallocated resources and run the simulation tools.
- The startup disks of all VMs are connected to the respective virtual SCSI controllers using a VHDX format file and with a fixed disk drive type.
- All extra data volumes are connected to added virtual SCSI controllers using a VHDX format file and with a fixed disk drive type unless otherwise specified for the test.
- All VMs are configured without dynamic memory, thus allocating the entire memory at the time of startup.
- A heterogeneous storage workload is created from the aggregation of the customized simulation running in each VM, as defined previously.

The following 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 EM/DSM, or is collected by the simulation tools (for example, Iometer, Jetstress, and FSCT) and outlined in their reports. Microsoft data threshold data 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.

The following test parameters were used for the standalone Hyper-V host environment:

Table 1 Test parameters: standalone host environment

Standalone reference configuration: test variables under study	
SC5020 storage array	2 tiers of disk storage using SSDs and 7.2K drives
Volumes and VMs distribution	8 VMs with relative extra data volumes deployed per VM
Reference configuration: consistent factors across the iterations of this test	
RAID policy (SAN)	RAID 10 and RAID 5
HBA connecting the SAN	2 x QLogic® 2532 Fibre Channel HBAs connected at the Hyper-V host level to volumes containing VHDX for VM boot and data virtual disks
VMs with limited storage footprint (total of 3)	Network services (1) Content management services (1) Web server services (1)
VMs with heavy storage footprint (total of 5)	Messaging services, back end (1) OLTP relational databases (1) File services, SMB network share (3)
Ratio of VMs residing in one SAN volume	8:1

3.1 Heterogeneous workload combination

The studied 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:

VM roles in a workload category with a 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:

- Network systems: Infrastructure or network services such as domain controller, DNS
- Web server: Static or active page services including middle-tier web-based components
- 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

VM roles in a workload category with a predictable heavy footprint on storage resources and with a complex set of assigned volumes to lay out 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:

- Messaging (combined CAS and mailboxes store): Messaging store and retrieval role; for example, the Microsoft Exchange 2016 combo role server
- Relational databases (OLTP): Relational database system serving transaction-oriented applications, such as SQL Server
- File services (file sharing repositories): Client-server services providing shared file and folder capabilities, such as CIFS network shares

3.2 Workload logical design

The workload design study is based upon typical Windows Server application workloads in the both standalone hypervisor and clustered configurations. The logical design portrays the virtual application server VMs in relation to their supporting hardware.

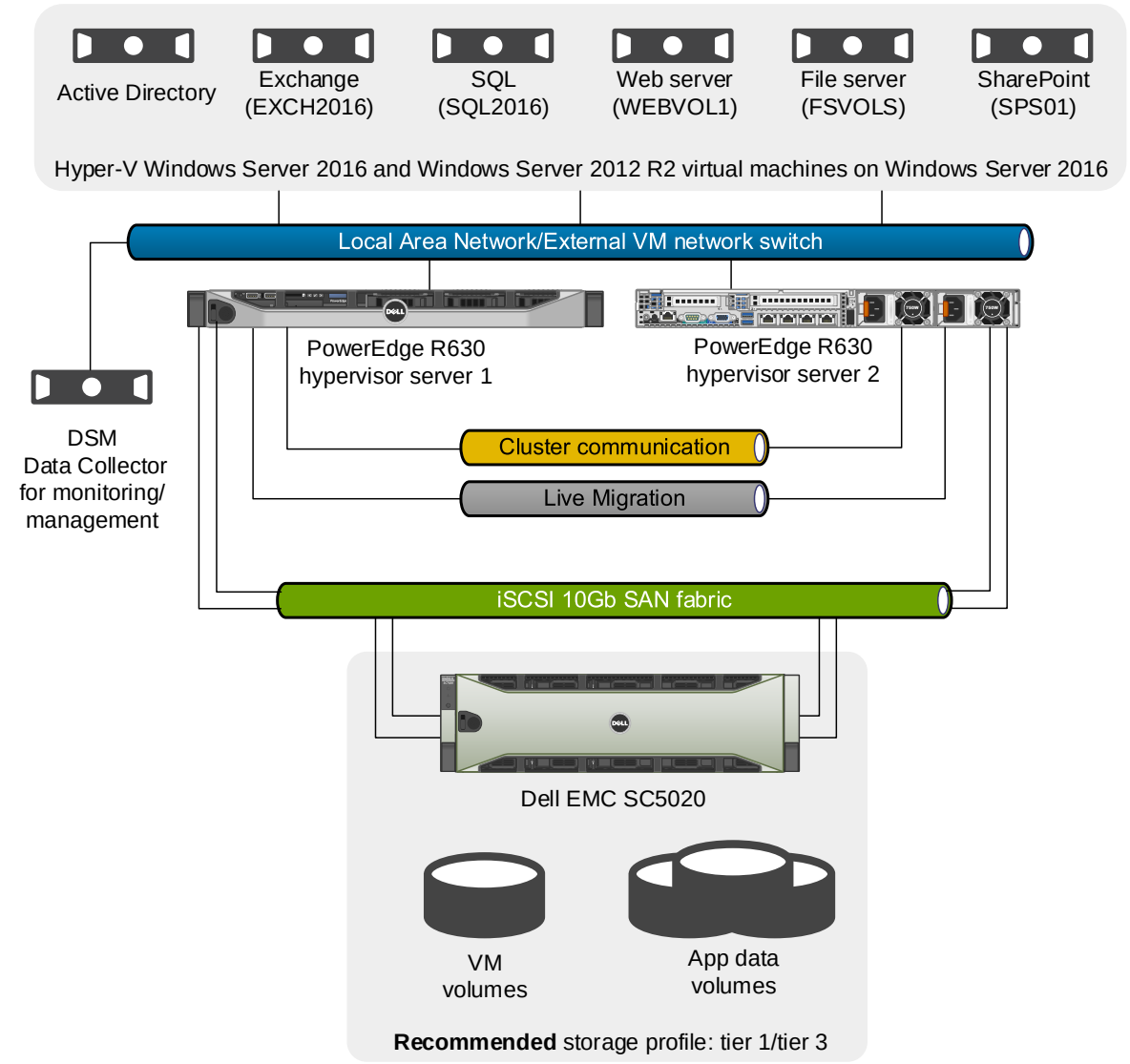


Figure 7 Hyper-V logical design for this study

Table 2 summarizes the number of VMs, the workload category and role of each VM, and the storage profile associated with it.

Table 2 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 (combo role, 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	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	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: 4–16 KB is small, 32–64 KB is medium, and over 128 KB is large.

3.3 Host resource usage

As a capacity test, one of the hypervisors hosts all workloads using the standalone configuration described previously. The host has only moderate processor usage with sufficient memory resources available for all workloads (see Figure 8).

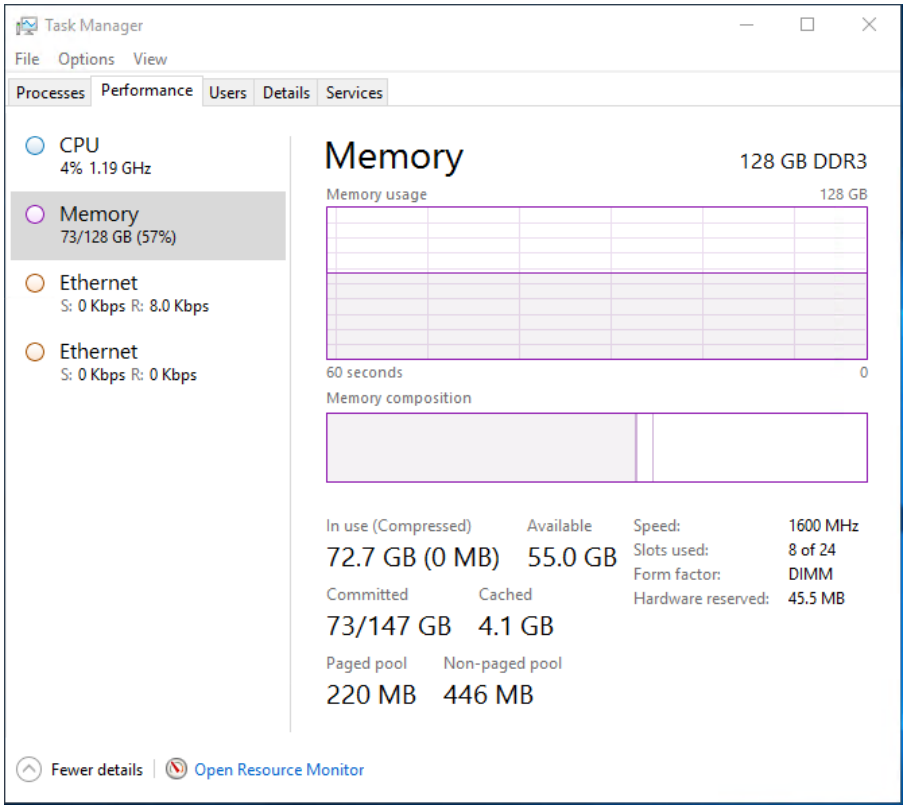


Figure 8 Windows Task Manager showing resource usage of one Hyper-V host

3.4 Standalone workload key performance indicators (KPI)

Standalone workloads were initiated simultaneously on standalone (non-clustered) Hyper-V hosts which have fewer restrictions on how storage is accessed because there are no server failover or high availability requirements. For maximum performance, the standalone environment was configured with the use of pass-through disks to directly connect to the block-level storage on the SC5020, thus bypassing any virtualized disk penalty.

3.4.1 Using pass-through disks gives best standalone performance

The following two graphs (Figure 9 and Figure 10) display the increased performance results of pass-through disks compared to virtualized disks on the same SQL Server workload.

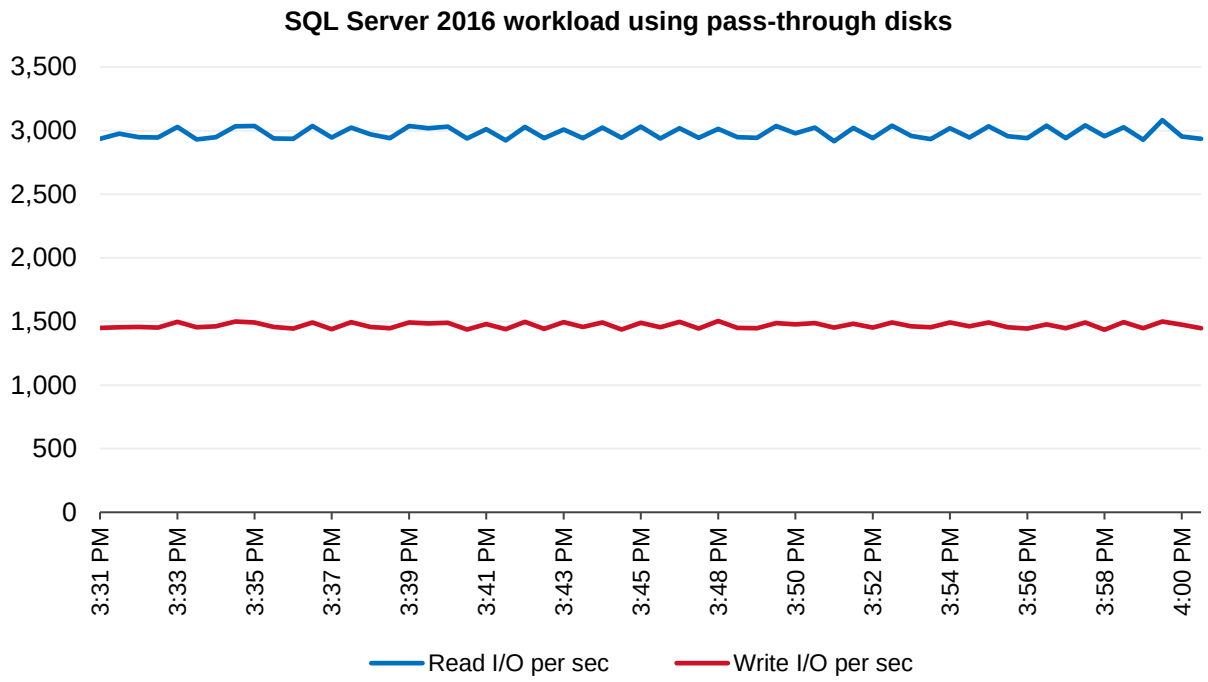


Figure 9 SQL Server workload using pass-through disks performance (~4,500 IOPS)

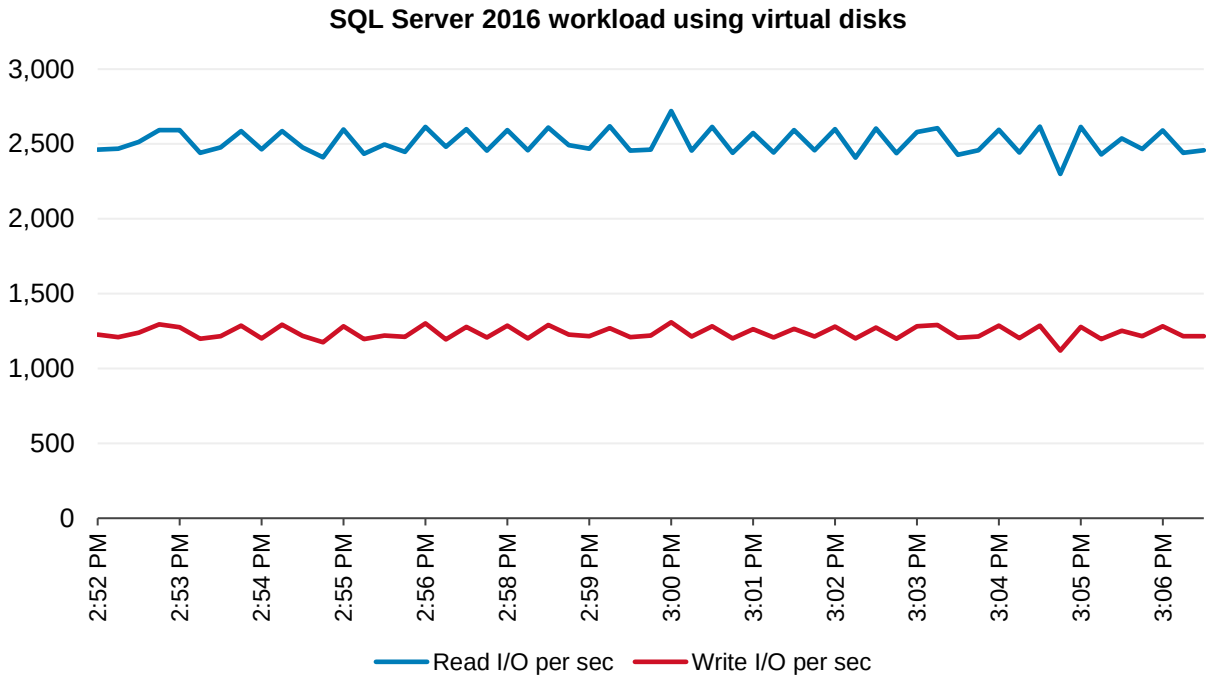


Figure 10 SQL Server workload using virtual disks performance (~3,700 IOPS)

3.4.2 Performance of tiered data

The following subsections detail the performance of data that has completed Data Progression and data that is active (non-progressed).

3.4.2.1 Performance of fully progressed data at lowest tier

To demonstrate the performance characteristics of the mixed workloads data that has been fully progressed by Data Progression to tier 3, a test run was conducted after a 12-cycle progression was completed.

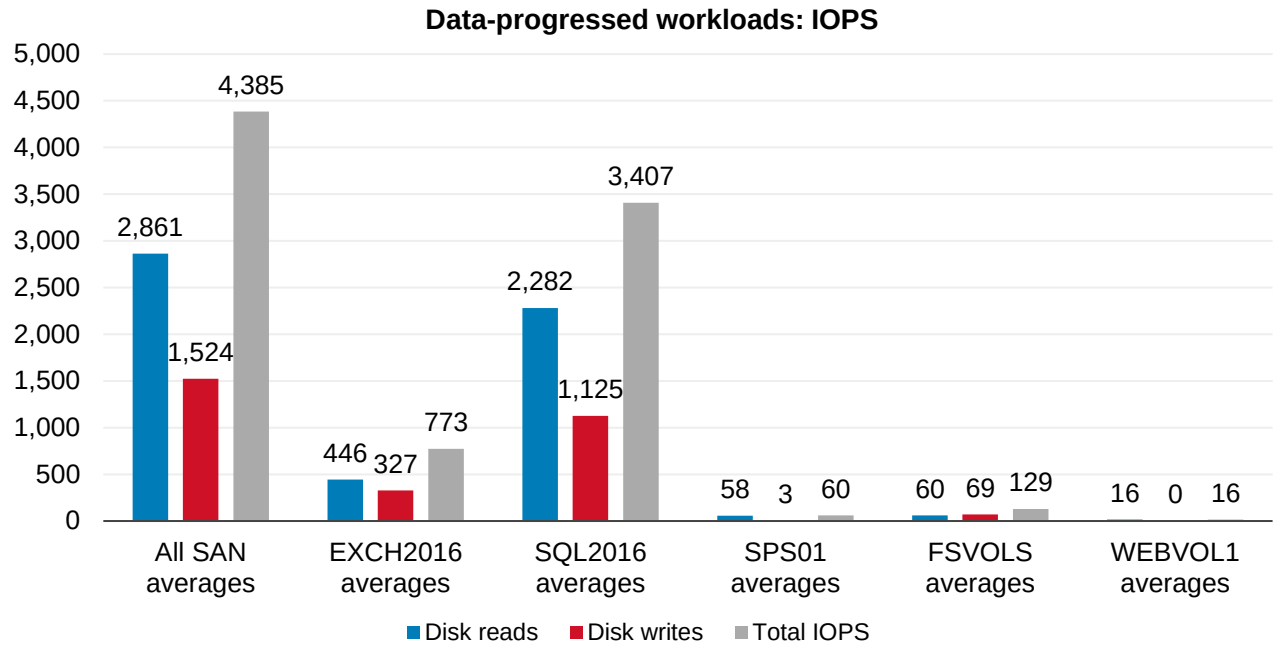


Figure 11 Fully data-progressed workloads (IOPS)

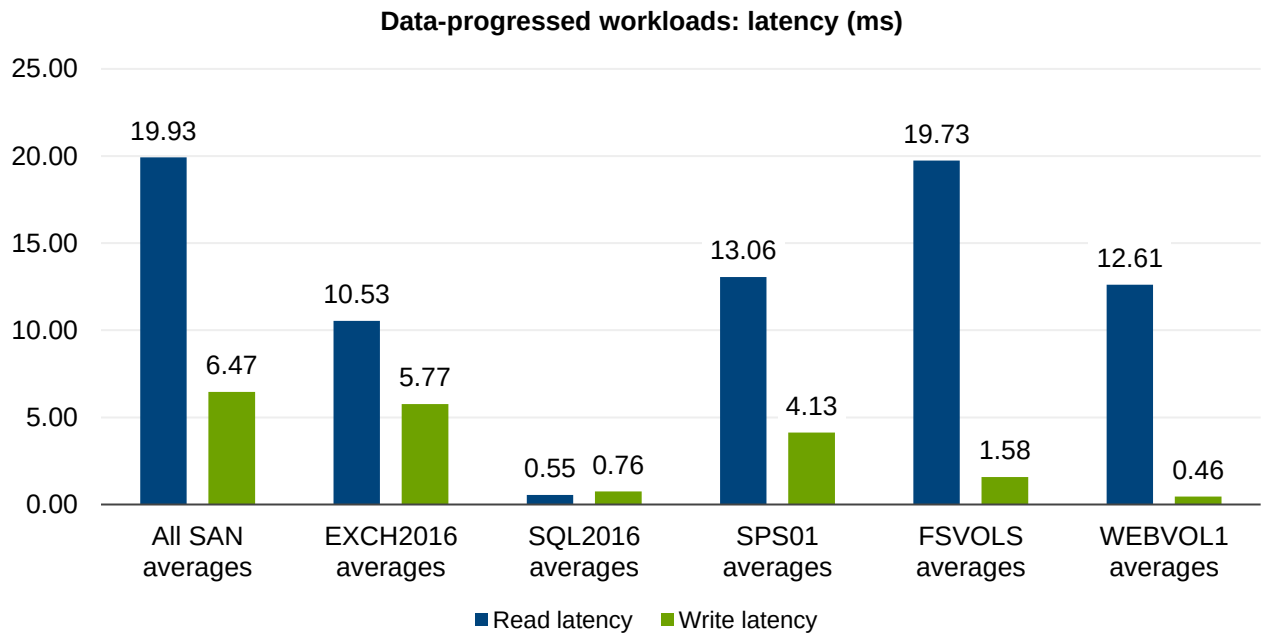


Figure 12 Fully data-progressed workloads (latency)

3.4.2.2 Performance of non-progressed (active) data

To compare the performance characteristics of the same data that had the majority of blocks in tier 1 at RAID 10, a test run was completed to show the max performance for the same 1,000-user mixed workload.

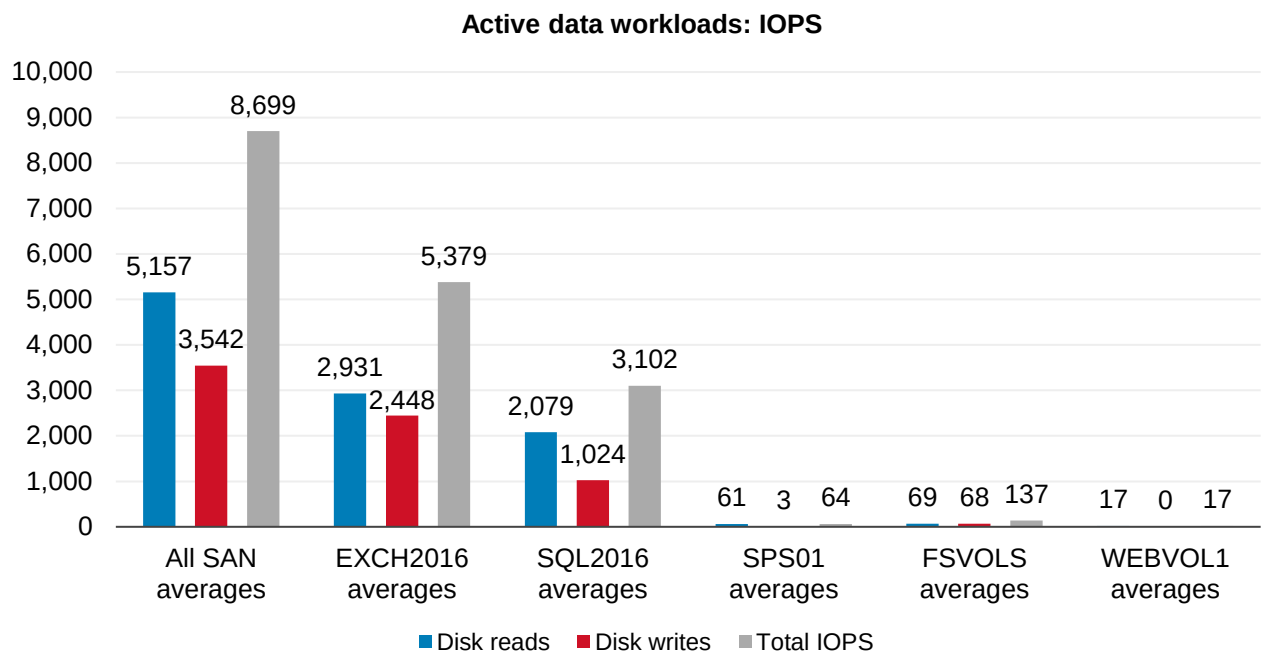


Figure 13 Active (tier 1) data performance chart

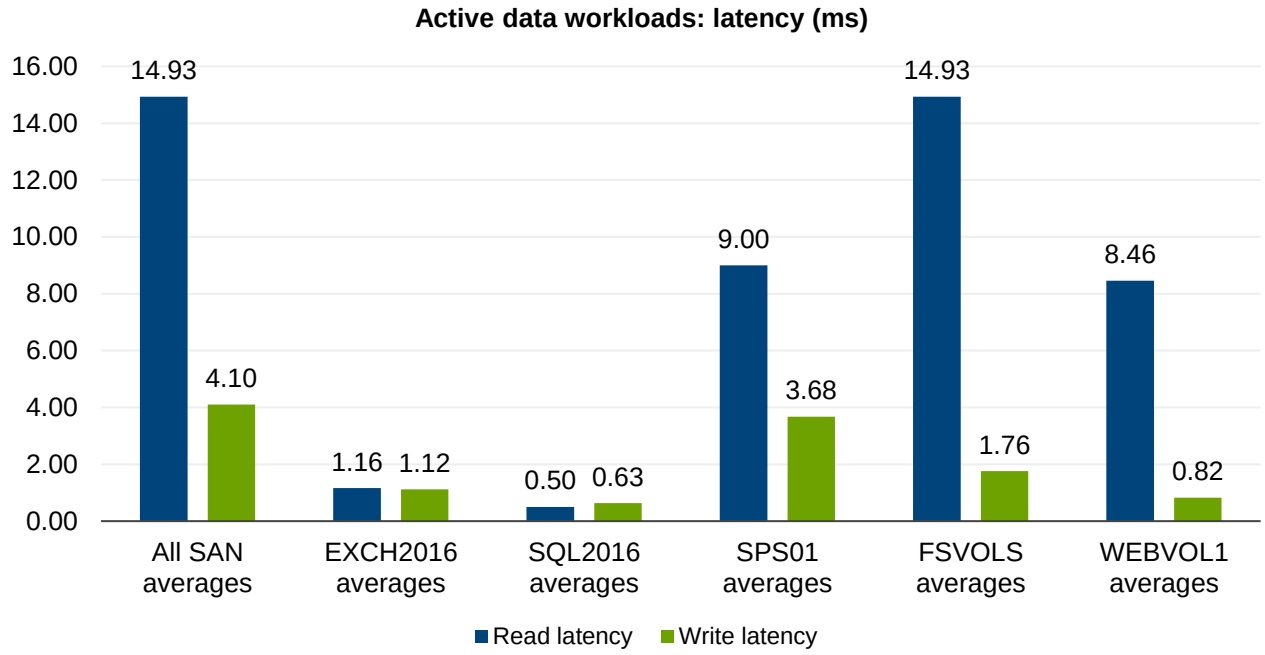


Figure 14 Active (tier 1) data performance (latency)

3.4.3 Back-end disk workload performance

Using the Recommended storage profile, tier 1 SSD disks provide more than 90 percent of the I/O operations of this workload combination. This is a desirable situation in which larger-capacity aged data such as Exchange databases, shared files, and older SQL data pages, are being stored in tier 3 on the slower, high-capacity 7.2K storage. Meanwhile, the most active data, including all write data and most read data, will be served by the fastest SSD storage.

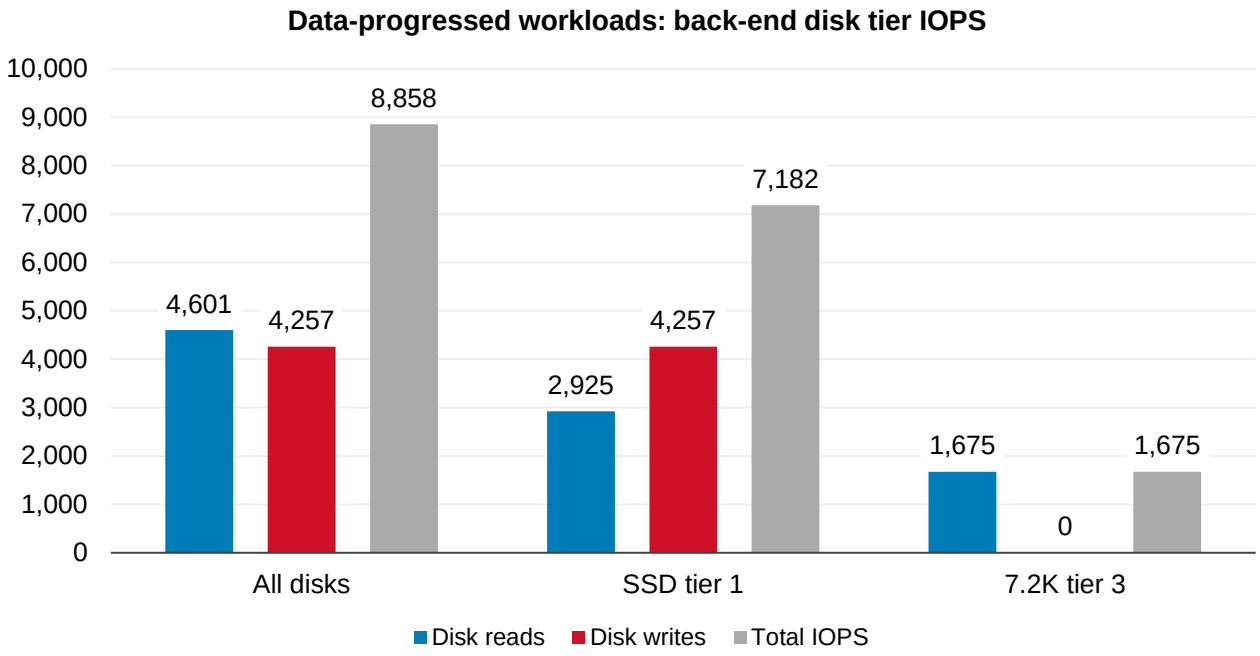


Figure 15 Disk statistics during standalone workload showing more than 90 percent of I/O served by tier 1

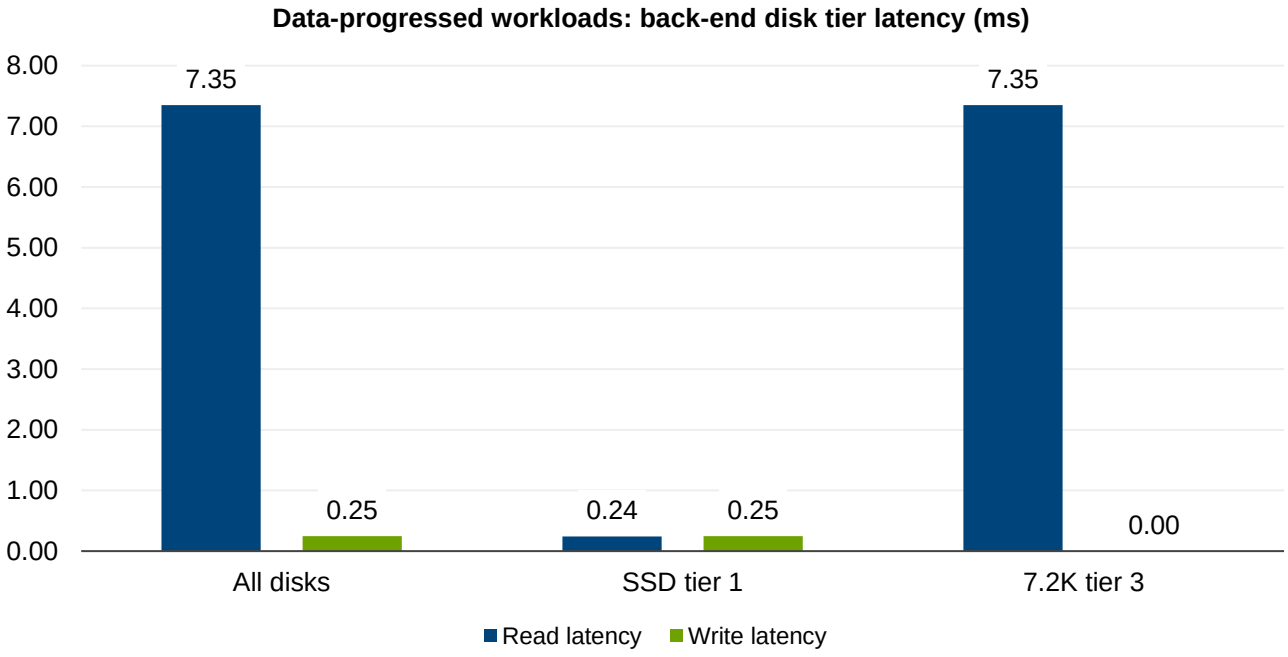


Figure 16 Disk statistics during the standalone workload showing low read and very low write latencies

4 Highly available Hyper-V clusters implementation

The second phase of this study focuses on the implementation and operation of highly available virtual environments by Windows failover cluster and Hyper-V technologies. The following configuration was built to assess application load distribution and failover cluster operations:

- Two node Windows failover cluster
- Quorum based on node and disk majority, with a dedicated disk witness on the SC5020
- 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 or 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 the VM 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 a sync before the switchover of the VM to the destination node happens. Hyper-V in Windows Server 2016 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 configured in the VM, because it must be saved to the disk first and read from there.

Storage migration is a feature native to Hyper-V in Windows Server 2016 and is previously available in a similar form only in System Center Virtual Machine Manager (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 2016 feature implemented as a combination of live migration 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.

Note: Storage migration is a feature available in a standalone Hyper-V installation as well. Its considerations are similar to other forms of VM migration and are included in this section (section 4).

4.1 Migration to a clustered environment from standalone hosts

This study includes both standalone and clustered implementations of Hyper-V, and the following procedure is a useful reference for customers looking to expand a standalone environment to include the high-availability features of a failover cluster with Hyper-V virtual machine roles and CSVs.

4.1.1 Migrating standalone hosts and storage volumes to a cluster or CSVs

The following steps describe how to move storage on existing standalone storage volumes.

1. Delete all snapshots from VMs.
2. Update Windows Server to the latest patches and hotfixes.
3. Reboot the hosts.
4. Install the Windows Failover Clustering feature on both hosts.
5. Create volumes on the SC Series array to host CSVs and create a cluster 1 GB witness disk.
6. Create a server cluster on SC Series storage using Fibre Channel HBA connections.
7. Run the **Failover Cluster Wizard** to create a cluster.
8. From the Failover Cluster Manager, click **Disk**, select the virtual machine storage, and convert the disk to clustered share volume.
9. From Server Manager, open **Hyper-V Manager**, run a storage migration, and migrate all VM data to a single location which is shared storage.
10. From the Failover Cluster Manager, run the **Configure Role Wizard**, select **Virtual Machine** from the drop-down menu, select one or more VMs, and migrate those VMs to the failover cluster node.
11. Test the live migration to ensure the Hyper-V cluster is working.

4.2 Application preferred placements on cluster nodes

Based upon workload sizing, SQL Server generates the majority of the I/O in this workload mix. The VM for SQL Server should be given a node preference that is separate from the other workloads.

The volume mappings for SQL Server should also be specifically mapped to a different SC Series controller (top or bottom) from the other workloads if possible. More discussion on this is provided later in section 4.

4.3 Cluster using VHDX virtual disks stored on CSVs

All application workloads are hosted on CSVs to allow the most flexibility in load balancing and high availability. Multiple VHDX virtual disks can be stored on a CSV. The boot C: drives of the VMs are stored on one CSV, and the data disks are stored on another CSV.

The performance of VHDX disks on a CSV is comparable to standalone volumes. The major workload in this combined workload study is SQL Server OLTP. As shown in the following figures, this performed very well because the majority of I/O consisted of writes that stayed in tier 1 SSDs.

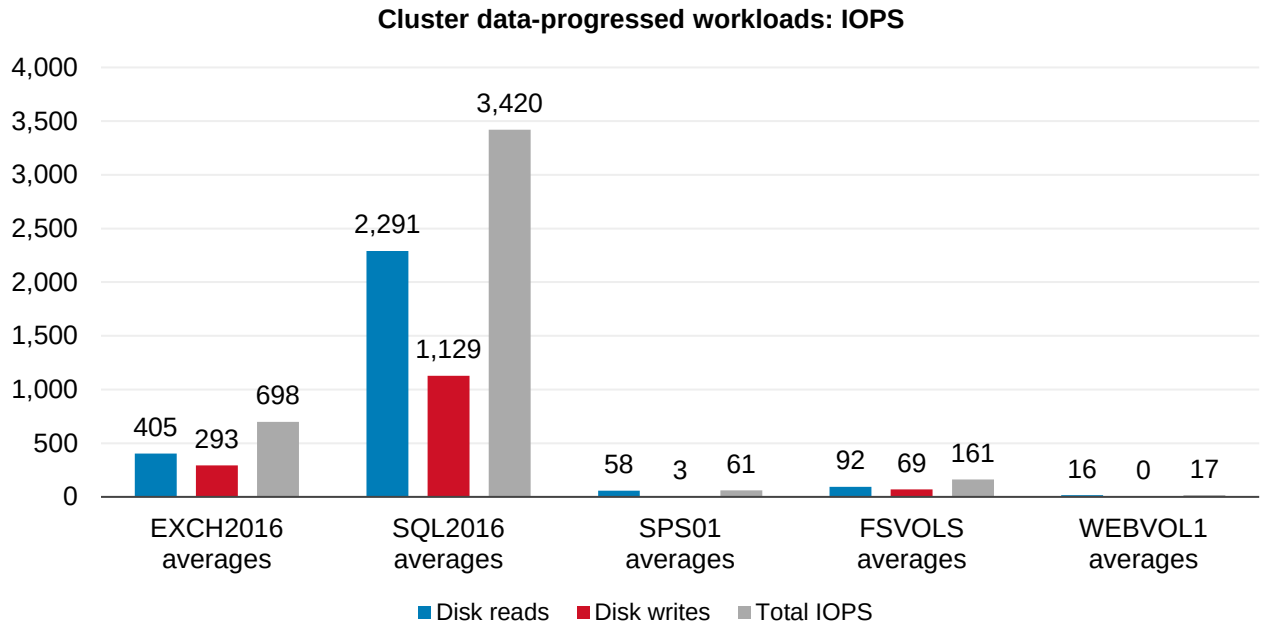


Figure 17 Clustered performance averages for each workload

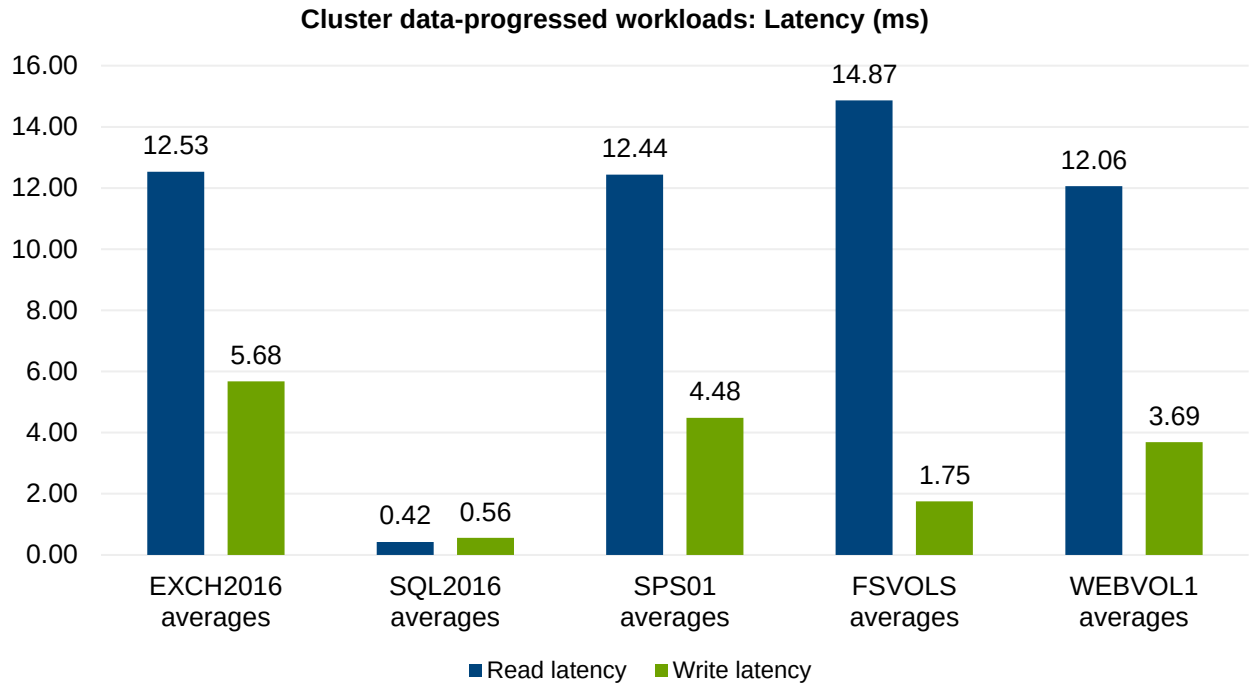


Figure 18 Clustered latency averages for each workload

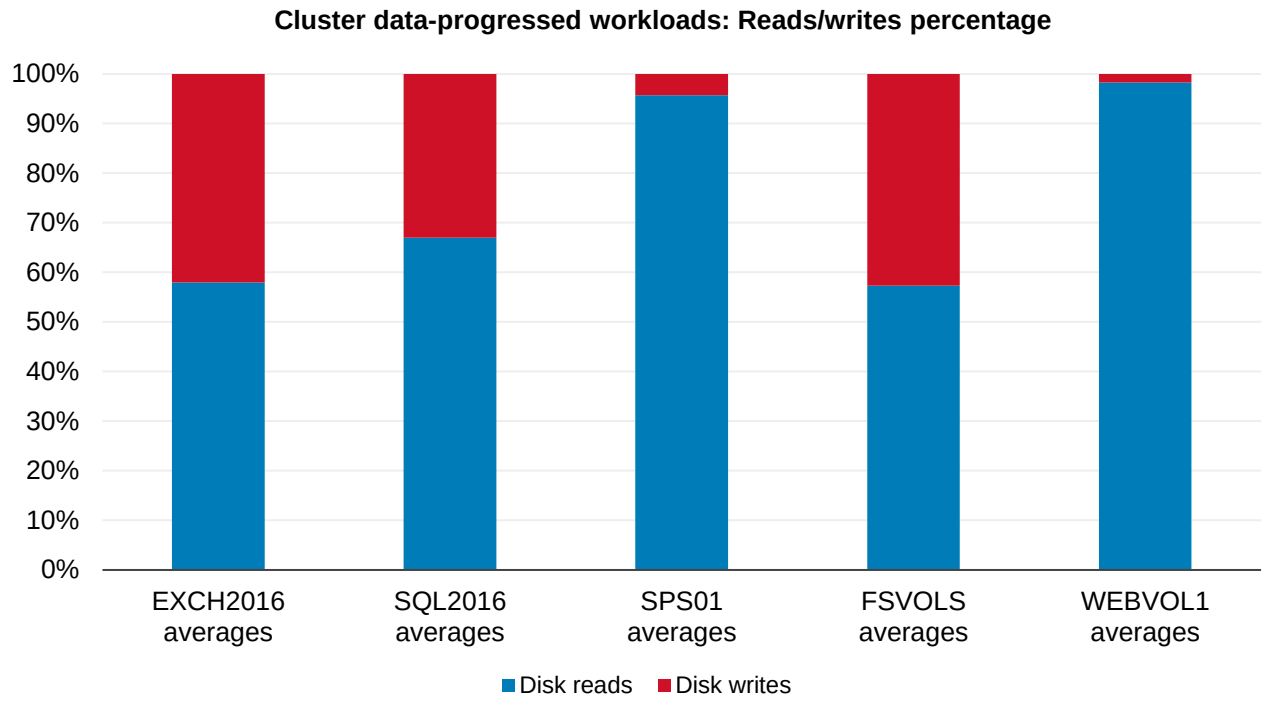


Figure 19 Clustered workload read and write average ratio statistics

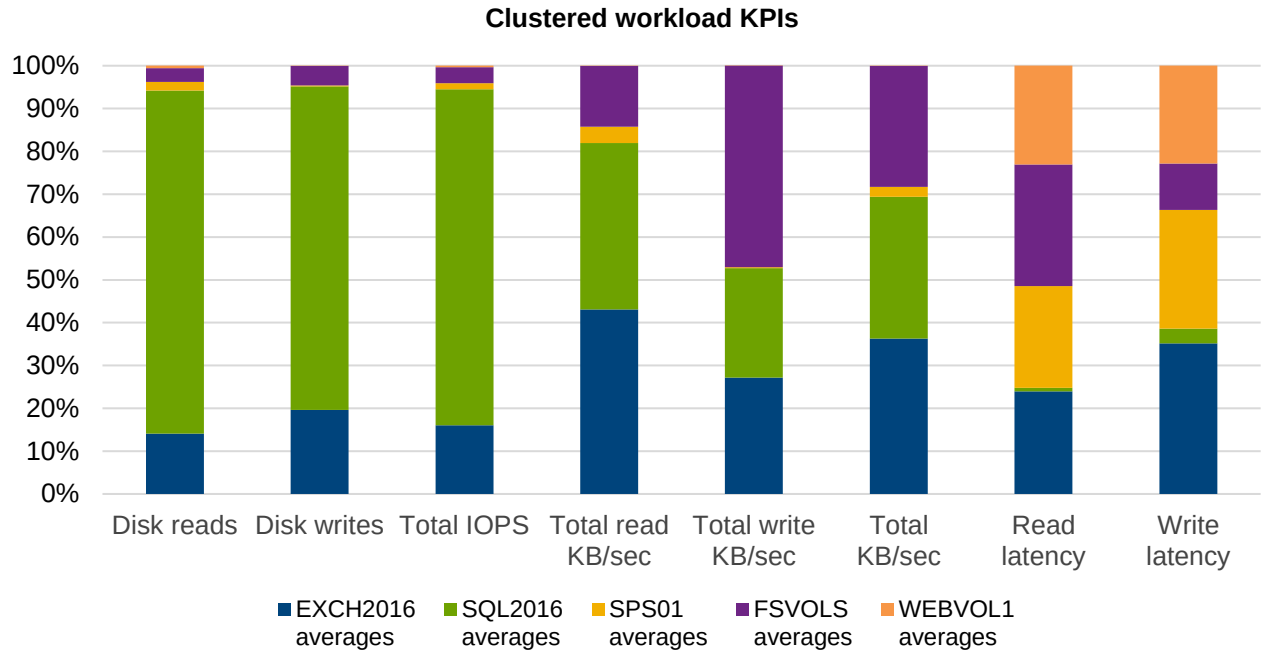


Figure 20 Clustered workload average ratios for each KPI

Table 9 Test parameters: Failover cluster scale out

Reference configuration: test variables under study	
SC5020 storage array	2 tiers of disk storage using SSDs and 7.2K drives
Volumes and VMs distribution	8 VMs with relative extra data volumes deployed per CSV
Reference configuration: consistent factors across the iterations of this test	
RAID policy (SAN)	RAID 10 and RAID 5
Fibre Channel HBA	Host software initiator, (VM startup VHDX volume, (data volumes))
VMs with limited storage footprint (total of 3)	Network services (1) Content management services (1) Web server services (1)
VMs with heavy storage footprint (total of 5)	Messaging services, back end (1) OLTP relational databases (1) File services, SMB network share (3)
Ratio of VMs residing in one SAN volume (CSV enabled)	1:1 or 2:1

4.4 Clustered volume storage efficiency with SC5020

A prerequisite to performance testing is to age the test data with multiple Data Progression cycles prior to running test suites. This moves the majority of data pages for all of the data volumes into the tier 3, 7.2K spinning media.

Several testing runs are performed to bring the simulated data into a normal operational state, with a small percentage of data pages in tier 1 and the majority remaining in tier 3, especially for the largest data set for Exchange. The resulting data placement shows the storage efficiency of the SC5020 Data Progression.

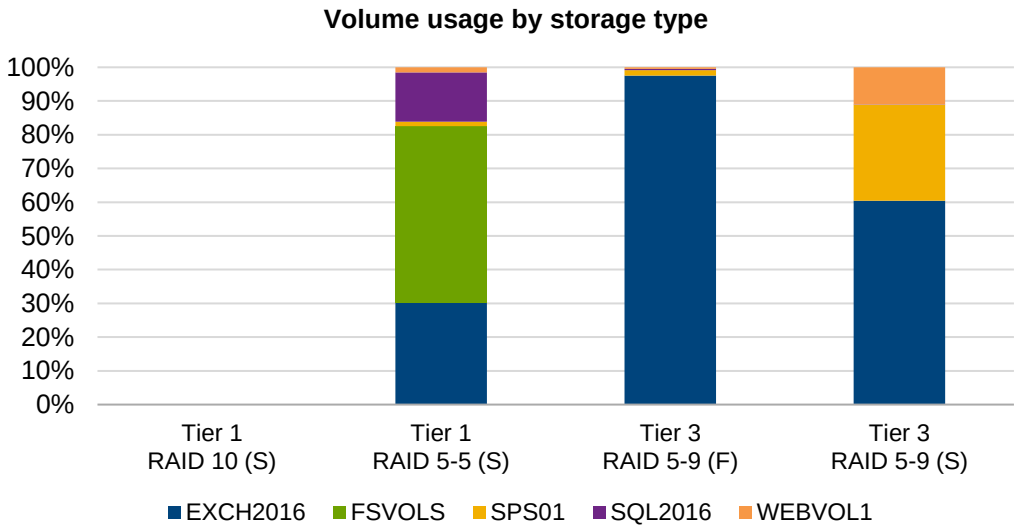


Figure 21 DSM RAID type distribution (F = Fast Track, S = Standard)

4.5 Clustered VM failover performance and factors

The clustered Hyper-V solution runs the application workloads simultaneously on both hypervisors. A live migration movement of the SQL 2016 VM is initiated from EX2 to EX1. After the live migration is completed, there is a very brief dip in I/O shown from the SC Series volumes but no discernable application performance drop (see the following figures).

Figure 22 Average of all volumes during live migration of EXCH server

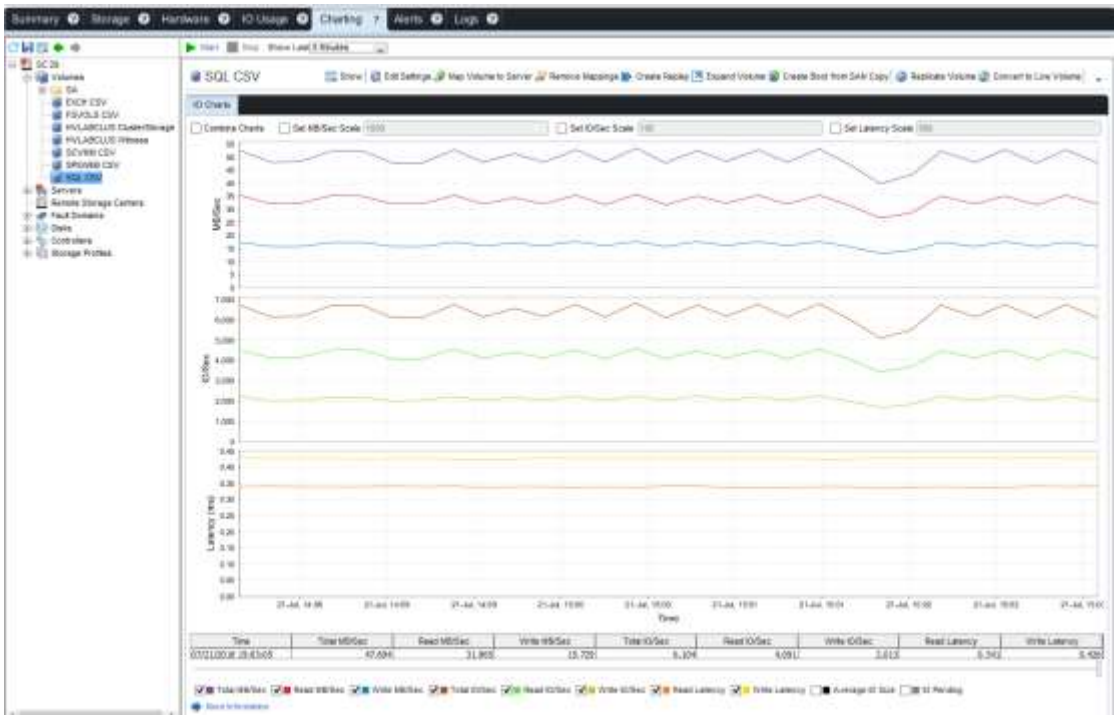


Figure 23 SQL CSV statistics during live migration of EXCH to same host

Figure 24 EXCH CSV statistics during live migration of EXCH server

4.5.1 Volume mappings controller preference

Testing showed that having the SQL OLTP and Exchange volumes mapped to their own SC Series controller is beneficial to performance, especially at peak workloads. This preference for mapping to a specific controller can be done by clicking the **Advanced** button during volume mapping operations.

MX Data Volumes

Show | Edit Settings | Cre

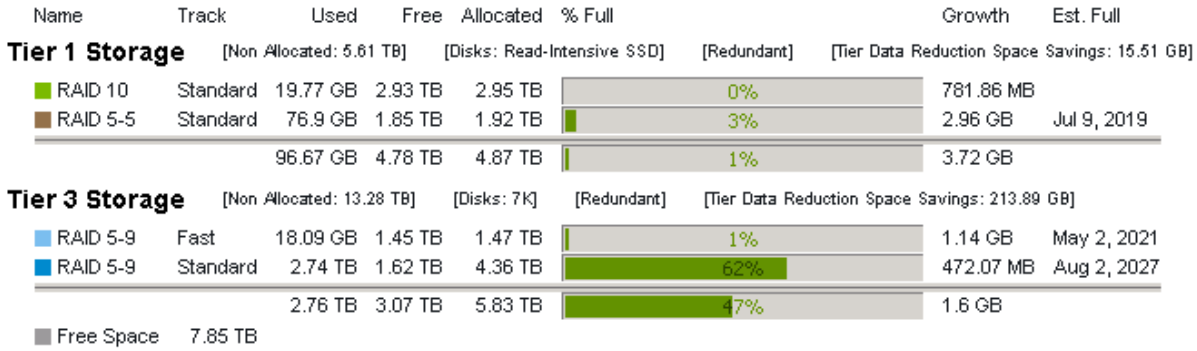
Summary | Charts

Name	Active Controller	Total Disk Space	Snapshot Overhead	Est. Data Reduction Space Savings
EXCH2016	Top Controller	1.8 TB	5 MB	0 MB
FSVOLS	Top Controller	~35.33 GB	636 MB	~137.21 GB
SPS01	Top Controller	556.75 GB	6 MB	0 MB
SQL2016	Bottom Controller	5.01 GB	163.5 MB	0 MB
WEBVOL1	Top Controller	213.79 GB	6 MB	0 MB

Figure 25 Volume mappings on the SC5020 top and bottom controllers during the live migrations

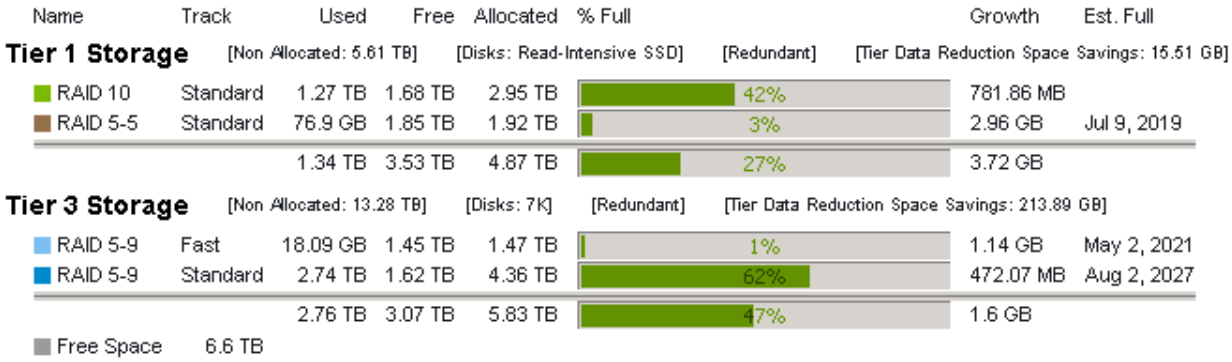
4.6 Tiering and RAID usage for best performance/capacity

After 12 days or cycles of Data Progression, data pages are transferred to RAID 5 and into tier 3. In production environments, this makes valuable, high-performance tier 1 SSD disks more available for all applications that have high I/O requirements, especially write requirements.



4.6.1 Data pages migrate back to tier 1

After running a two-hour test, all new writes over 839 GB are written into tier 1, RAID 10 (see the following screenshot). The pages in tier 1, RAID 5 are the ones most recently progressed in read-only snapshots. This provides the best performance for both write and read performance using the fastest RAID methods for each.



5 Best practices recommendations

Refer to these best practices to plan and configure the SC5020, network switches, Hyper-V servers, and VMs.

5.1 Storage best practices

- Use Multipath I/O (MPIO) to improve performance and reliability for the HBA (FC, iSCSI) connections. See *Dell EMC SC Series Storage: Microsoft Multipath I/O Best Practices.*
- 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 storage profile when designing the environment according to the performance, capacity, and failure-tolerance requirements of your environment. In most cases the **Recommended** profile will provide the best mix of performance and capacity, especially for mixed workloads.
- Thin-provisioned volumes can ease the burden of tight capacity management. Plan thin-provisioned volumes carefully and jointly with SC Series automatic reports using space usage threshold alarm reporting to avoid the risk of overprovisioning.
- 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 Availability Groups, 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 SC5020 arrays instead for each replicated copy.
- Keep the Offloaded Data Transfer (ODX) enabled (by default) on the Hyper-V hosts that are connected to the SC5020 SAN to reduce the computing footprint in the case of a large transfer of VMs. See *Windows Server ODX Solutions Guide for Dell Compellent Storage Center.*
- For environments with a very large number of volumes attached to the servers, monitor the amount of FC or iSCSI connections per host generated.
- Use mount points for the SAN volumes to simplify 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.
- Deploy multiple VMs in the same volume for Hyper-V CSVs to simplify the provisioning of the SAN volumes and the administration of the host connections. Consider separating VMs that have a larger I/O footprint from others if necessary to keep it from affecting performance of shared volume machines.
- Deploy a single VM in a volume to provide granular performance and space utilization monitoring from both the host and SAN sides, and to ease the portability of the volume to other hosts if needed.

5.2 Network best practices (iSCSI SAN)

- 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).
- Enable flow control for the switch ports hosting the SC5020 array controller connections (iSCSI).
- Enable flow control on the host network adapters dedicated to SAN traffic (iSCSI).
- Enable Jumbo frames (large MTU) for the switch ports hosting the SC5020 array controller connections.
- Evaluate Jumbo frames (large MTU) for the LAN network when appropriate (limited by the type of devices the traffic traverses).
- Use 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.

5.3 Windows Server 2016 Hyper-V and VMs best practices

- Install a Windows Server Core version in the root partition of the Hyper-V role server when it is critical to reduce maintenance, software attack surface, memory, and disk-space footprint. Otherwise, when installing a traditional Windows Server with Hyper-V technology using the GUI, minimize the use of additional software, components, or roles in the root partition.
- Use non-uniform memory access (NUMA) 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 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.
- 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).
- Do not use differencing disk image files for production environment VMs. These image files also might not be supported by the applications running in the VMs.
- While performance of dynamically expanding disks has improved, use fixed disks 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 (processor, memory, and network). Use ODX technology to help 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. Use thin-provisioned volumes in cases with space-usage challenges.

- 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 the 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 more best practices and information about Hyper-V 2016 configurations with SC Series storage, refer to the document, *Dell SC Series Storage and Microsoft Hyper-V*.

A Load simulation tools considerations

A.1 Microsoft Jetstress

Microsoft Exchange Server Jetstress 2013 is a simulation tool that reproduces the database and log I/O workload of an Exchange 2013 and 2016 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 Extensible Storage Engine (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.
- 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 have the same capacity as the one(s) planned for the Exchange Server future deployment.
- Its topology layout includes the 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 thread-execution pace).

A.2 Microsoft File Server Capacity Tool (FSCT)

The Microsoft File Server Capacity Tool (FSCT) 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.

- FSCT 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).
- FSCT 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.

B Additional resources

B.1 Technical support and customer service

Dell.com/support is focused on meeting customer needs with proven services and support.

Dell TechCenter is an online technical community where IT professionals have access to numerous resources for Dell software, hardware, and services.

Storage Solutions Technical Documents on Dell TechCenter provide expertise that helps to ensure customer success on Dell EMC storage platforms.

B.2 Related documentation

See the following referenced or recommended resources related to this document.

Referenced or recommended Dell publications:

- Dell Storage SC5020 specifications sheet
- Dell SC Series Storage and Microsoft Hyper-V
- Dell Storage Center: Microsoft Multipath I/O Best Practices
- Microsoft ESRP - Dell EMC SC Series SC5020 15,000 Mailbox Exchange 2016 Resiliency Storage Solution using 10K Drives
- Dell SC Series Storage and Microsoft Exchange Server 2016 Best Practices
- Dell Storage SC Series Arrays and Microsoft SQL Server
- Windows Server ODX Solutions Guide for Dell Compellent Storage Center
- Dell Networking S5000 documentation
- Dell PowerEdge R630 documentation

Referenced or recommended Microsoft publications:

- Windows Server 2016 Hyper-V Overview
- Microsoft Best Practices: Exchange 2016 virtualization
- Microsoft Exchange Server Jetstress 2013 (64 bit)
- Microsoft File Server Capacity Tool v1.2 (64 bit)

Referenced or recommended independent publications:

- Iometer project