Lenovo

Reference Architecture for Microsoft Storage Spaces Direct (S2D)

Introduces the Microsoft S2D Software-Defined Storage solution using Lenovo rack-based servers

Describes the hyperconverged and disaggregated deployment scenarios for S2D

Provides detailed infrastructure specifications

Discusses various options related to storage performance and sizing

David Feisthammel

David Ye

LENOVO PRESS

Abstract

This document describes the Lenovo® Reference Architecture for Microsoft Storage Spaces Direct (S2D). Lenovo Reference Architecture offerings create virtually turnkey solutions that are built around the latest Lenovo servers, networking, and storage, which takes complexity out of the solution. This Lenovo Reference Architecture combines Microsoft software, consolidated guidance, and validated configurations for compute, network, and storage.

This Lenovo solution for Microsoft S2D combines the Storage Spaces Direct and Failover Cluster features of Windows Server 2016 with Lenovo industry standard x86 servers and Lenovo RackSwitch™ network switches to provide turnkey solutions for enterprises. The architecture that is described here was validated by Lenovo and certified for Microsoft Storage Spaces Direct.

The intended audience of this document is IT professionals, technical architects, sales engineers, and consultants to assist in planning and designing, as well as to provide best practices for implementing the Lenovo Reference Architecture for Microsoft S2D.

At Lenovo Press, we bring together experts to produce technical publications around topics of importance to you, providing information and best practices for using Lenovo products and solutions to solve IT challenges.

See a list of our most recent publications at the Lenovo Press web site:

http://lenovopress.com

Do you have the latest version? We update our papers from time to time, so check whether you have the latest version of this document by clicking the **Check for Updates** button on the front page of the PDF. Pressing this button will take you to a web page that will tell you if you are reading the latest version of the document and give you a link to the latest if needed. While you're there, you can also sign up to get notified via email whenever we make an update.

Contents

Introduction 3
Requirements 5
Architectural overview 6
Component model 10
Operational model 21
Deployment considerations 28
Resources 31
Appendix: Lenovo bill of materials 32
Authors 33
Notices 34
Trademarks 35

Introduction

With the high demand for enterprise storage continuing to accelerate in recent years, Lenovo and Microsoft have teamed up to craft a software-defined storage solution leveraging the advanced feature set of Windows Server 2016, the flexibility of the Lenovo System x3650 M5 rack server and the RackSwitch G8272 network switch.

This document describes the Lenovo Reference Architecture for Microsoft Storage Spaces Direct (S2D). Lenovo reference architecture offerings create virtually turnkey solutions that are built around the latest Lenovo servers, networking, and storage, which takes complexity out of the solution. This Lenovo reference architecture combines Microsoft software, consolidated guidance, and validated configurations for compute, network, and storage.

Software-defined storage (SDS) is an evolving concept for computer data storage that manages policy-based provisioning and data storage independent of hardware. SDS definitions typically include a form of storage virtualization to separate the storage hardware from the software that manages the storage infrastructure. The software that enables an SDS environment might also provide policy management for feature options, such as deduplication, replication, thin provisioning, snapshots, and backup. The key benefits of SDS over traditional storage are increased flexibility, automated management, and cost efficiency.

Microsoft S2D uses industry-standard servers with local-attached storage devices to create highly available, highly scalable SDS for Microsoft Hyper-V, SQL Server, and other workloads at a fraction of the cost of traditional SAN or NAS arrays. Its hyperconverged or disaggregated architecture radically simplifies procurement and deployment, while features like caching, storage tiers, and erasure coding, together with the latest hardware innovation like RDMA networking and NVMe devices, deliver unrivaled efficiency and performance. S2D is included in Windows Server 2016 Datacenter.

This Lenovo solution for Microsoft S2D combines the Storage Spaces Direct and Failover Cluster features of Windows Server 2016 with Lenovo industry standard x86 servers and Lenovo RackSwitch network switches to provide turnkey solutions for enterprises. The architecture that is described here was validated by Lenovo and certified for Microsoft S2D.

The intended audience of this document is IT professionals, technical architects, sales engineers, and consultants to assist in planning and designing, as well as to provide best practices for implementing the Lenovo Reference Architecture for Microsoft S2D.

For more information regarding detailed deployment of S2D, refer to the Lenovo Press paper *Microsoft Storage Spaces Direct (S2D) Deployment Guide* available at the following URL:

https://lenovopress.com/lp0064

Business problem and value

This section describes the challenges organizations face and how this Reference Architecture for Microsoft S2D can help meet those challenges.

Business problem

The cloud and mobile innovations in the last few years present a tremendous amount of growth opportunity for those enterprises that are equipped with proper IT infrastructure. However, companies are discovering that their IT infrastructure is not always up to the task at hand; finding that budgetary constraints or outdated architectures are hindering their ability to compete. Enterprises that use proprietary systems are finding that they are locked into

© Copyright Lenovo 2017. All rights reserved.

expensive maintenance contracts and obligations that force them to continue buying expensive proprietary technologies.

With digital data growing rapidly in the enterprise, companies who deployed traditional proprietary SAN storage are seeing a significant amount of their budget being allocated for storage purchases. This is one of the inhibiting factors that limit companies' growth and competitiveness because of lack of investments in other key areas, such as new applications.

Business value

When discussing high performance and shareable storage pools, many IT professionals think of expensive SAN infrastructure. Thanks to the evolution of disk storage and server virtualization technology, as well as ongoing advancements in cost effective network throughput, it is now possible to deliver an economical, highly available and high performance storage subsystem.

The Lenovo solution for Microsoft S2D combines the skills and technologies of the leading enterprise software and hardware vendors to create a non-proprietary storage solution that lowers overall storage costs, increases storage reliability, and frees you from expensive maintenance and service contracts. Gaining access to near-zero downtime with exceptional fault tolerance, dynamic pooling, enhanced virtualization resources, end-to-end architectural and deployment guidance, predefined, out-of-box solutions and much more gives you the tools to compete today and into the future.

Table 1 lists a high-level comparison of SAN shared-storage and Microsoft S2D capabilities. Note that although S2D itself supports data deduplication, the ReFS file system does not yet support it.

Table 1 Comparison of SAN and Microsoft Scale-Out File Server with Storage Spaces Direct

FC/iSCSI SAN	**Microsoft Storage Spaces Direct**
RAID resiliency	Two-/three-way mirroring, single/dual parity
Disk pooling	Disk pooling
High availability	Continuous availability (transparent failover)
Storage tiering	Storage tiering
Snapshots	Snapshots
Persistent write-back cache	Persistent, real-time read and write cache
Data deduplication	Data deduplication (not supported on ReFS volumes)
Replication	Synchronous and asynchronous replication
FC, FCoE, iSCSI	10 GbE, 25 GbE, 40 GbE, 100 GbE SMB Direct (RDMA)

The basic building block of this Lenovo solution can scale from 2 to 16 storage nodes. This multiple storage node building block model can scale horizontally and linearly as much as your compute infrastructure requires. It also provides high performance and continuous availability. All of these features can be achieved by using standard Lenovo 2-socket x86 server hardware to ultimately realize lower total cost of ownership.

Requirements

The functional and non-functional requirements for this reference architecture are described in this section.

Functional requirements

The following functional requirements are featured:

- Integrate with Windows Server 2016 storage features:
  - Create storage pools using locally attached storage media
  - Create RAID equivalent virtual disks:
    - Enable virtual disks for storage tiering
    - Allocate cache for virtual disks from flash media (SSD or NVMe)
  - Create Cluster Shared Volumes
  - Support SMB 3.1.1 for storage access protocol:
    - Create Continuously Available (CA) File Shares
    - Create multiple SMB connections for each network adapter on-demand
    - Detect and utilize RDMA-capable network adapters
    - Encrypt storage traffic between hosts and storage nodes
    - Provide transparent failover capability
  - Support File Management:
    - Enable/Disable deduplication on per volume basis
    - Configure as DFS Namespace folder target server
    - Enable/Disable Folder Redirection
    - Support Roaming User Profiles
    - Support Home Directories
  - Support I/O intensive application workloads:
    - Microsoft Hyper-V
    - Microsoft SQL Server
- Integrate with Windows Failover Cluster for high availability (HA):
  - Create Windows Failover Cluster:
    - Create Cluster Management Network
    - Create Cluster Communication Network
    - Create Client/Server Network for Storage
    - Create Client Access Point for CA Shares
  - Single Management User Interface:
    - Manage all S2D storage functionality
    - Provide wizard-driven tools for storage related tasks
  - Enterprise Management: Support integration with Microsoft System Center

Non-functional requirements

Table 2 lists the non-functional requirements for a server-based storage solution.

Table 2 Non-functional requirements

Requirement	Description
Scalability	Scale out linearly with building block approach
Load balancing	SMB storage traffic is distributed across storage nodes

Requirement	Description
Fault tolerance	Storage nodes are redundant, fault domain at server, rack or site level
Physical footprint	Low profile servers reduce rack space
Ease of installation	Standard server hardware and software with cluster setup wizard
Ease of management/operations	Managed through Windows Failover Cluster Manager
Flexibility	Choices for various hardware components, including storage device media type; flexible deployment models, either Hyperconverged or Disaggregated
Security	Encryption of storage traffic
High performance	Low latency, high throughput RDMA over high-speed Ethernet
Certification	Hardware should be certified for Microsoft Windows 2016

Architectural overview

The initial offering of Microsoft SDS was contained in Windows Server 2012 and was called “Storage Spaces.” The next iteration of this solution has been introduced in Windows Server 2016 under the name *Storage Spaces Direct* (S2D), and continues the concept of collecting a pool of affordable drives to form a large usable and shareable storage repository. In Windows Server 2016, the solution expands to encompass support for both SATA and SAS drives, including NVMe devices, that reside internally in the server.

The Lenovo Solution for Microsoft S2D scales from a minimum of two storage nodes up to a maximum of 16 nodes. Consequently, this solution has the range and capacity to effectively address the storage needs of both small businesses and large enterprise customers.

Microsoft S2D provides resiliency for multiple drive failures. In fact, a typical 4-node cluster environment can tolerate complete failure of up to two full nodes, including all the drives they contain. When additional storage is needed, it is a simple matter to add additional storage devices in existing nodes (if empty bays exist), or to add nodes to the cluster and integrate their storage capacity into the existing S2D Storage Pool. In this manner, the S2D Solution can be scaled up or down depending on current needs.

Key considerations

There are a few considerations that are important to consider when planning an S2D solution implementation, including the following:

- S2D capacity and storage growth

  Leveraging the 14x 3.5” drive bays of the Lenovo System x3650 M5 and high-capacity storage devices, each server node is itself a JBOD (just a bunch of disks) repository. As demand for storage and/or compute resources grows, additional x3650 M5 systems are added into the environment to provide the necessary storage expansion.

- S2D performance

  Using a combination of solid-state devices (SSD or NVMe AICs) and regular hard disk drives (HDDs) as the building blocks of the storage volume, an effective method for storage tiering is available. Faster-performing flash devices act as a cache repository to

the capacity tier, which is placed on traditional HDDs in this solution. Data is striped across multiple devices, allowing for very fast retrieval from multiple read points.

At the physical network layer, 10GbE links are employed today. However, in the future, additional throughput needs can be satisfied by using higher bandwidth adapters. For now, the dual 10GbE network paths that contain both Windows Server operating system and storage replication traffic are more than sufficient in most environments to support the workloads and show no indication of bandwidth saturation.

- S2D resilience

  Traditional disk subsystem protection relies on RAID storage controllers. In S2D, high availability of the data is achieved using a non-RAID adapter and adopting redundancy measures provided by Windows Server 2016 itself. When S2D is enabled, a storage pool is automatically created using all the locally-attached storage devices from all the nodes in the cluster (not including boot devices). From this pool you can create *Storage Spaces*, which are cluster shared volumes that use mirroring, erasure coding, or both to ensure fault tolerance. In fact, if the solution is built using four or more nodes, these volumes typically have resiliency to two simultaneous drive or server failures (e.g. 3-way mirroring, with each data copy located in a different server) though rack fault tolerance is also available.

- S2D use cases

  The importance of having a SAN in the enterprise space as the high-performance and high-resilience storage platform is changing. The S2D solution is a direct replacement for this role. Whether the primary function of the environment is to provide Windows applications or a Hyper-V virtual machine farm, S2D can be configured as the principal storage provider to these environments.

  Two primary use cases for S2D are to host SQL Server database files and Hyper-V VHD(X) files. When one of the S2D storage nodes fails, the SMB client that is running on the SQL Server or Hyper-V host automatically selects the next best available S2D storage node to connect to and resumes I/O operations. From the perspective of SQL Server and Hyper-V, the I/O appears to stall for a very short amount of time before continuing as if no failure had occurred.

Deployment scenarios

S2D supports two general deployment scenarios, which are called *hyperconverged* and *disaggregated*. Microsoft sometimes uses the term “converged” to describe the disaggregated deployment scenario. Both scenarios provide storage for Hyper-V and SQL Server, specifically focusing on Hyper-V Infrastructure as a Service (IaaS) for service providers and enterprises.

For the hyperconverged approach, there is no separation between the resource pools for compute and storage. Instead, each server node provides hardware resources to support the running of virtual machines under Hyper-V, as well as the allocation of its internal storage to contribute to the S2D storage repository.

Figure 1 shows a diagram of this all-in-one configuration for a four-node hyperconverged solution. When it comes to growth, each additional node added to the environment will mean both compute and storage resources are increased together. Perhaps workload metrics dictate that a specific resource increase is sufficient to cure a bottleneck (e.g., CPU resources). Nevertheless, any scaling will mean the addition of both compute and storage resources. This is a fundamental limitation for all hyperconverged solutions.

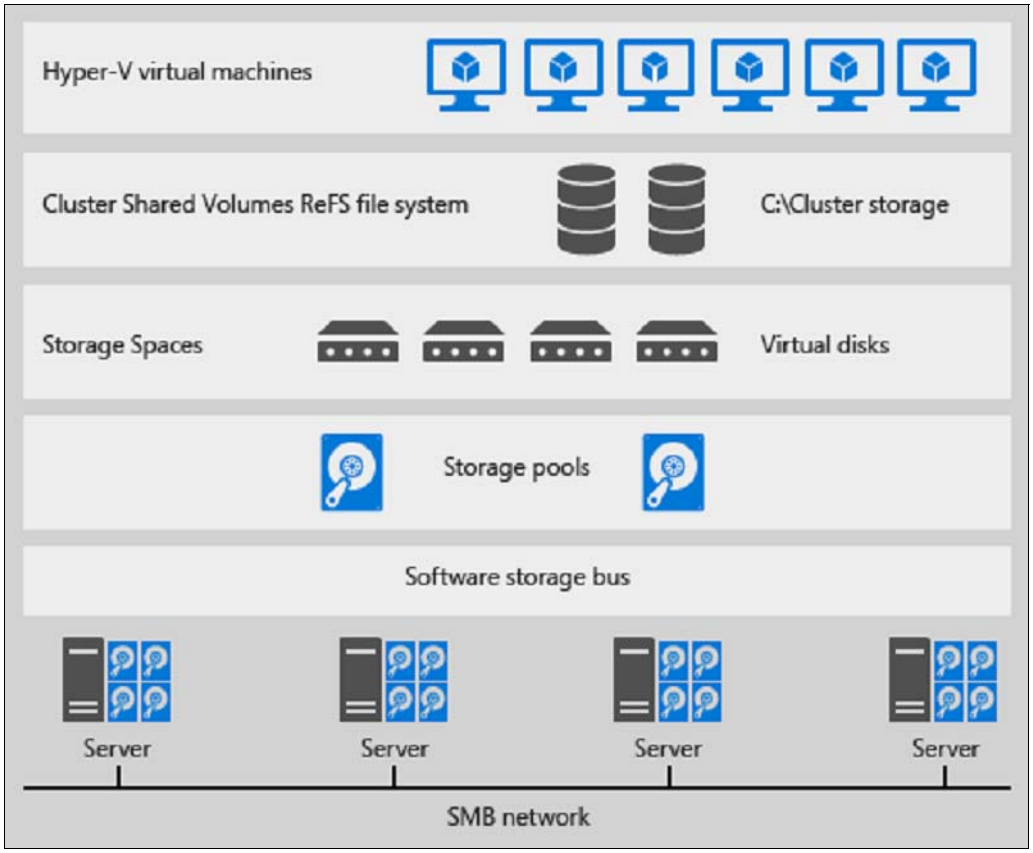


Figure 1 Diagram showing the hyperconverged deployment of S2D

In the disaggregated approach, the environment is separated into compute and storage components. An independent pool of servers running Hyper-V acts to provide the CPU and memory resources (the compute component) for the running of virtual machines that reside on the storage environment. The storage component is built using S2D and Scale-Out File Server (SOFS) to provide an independently scalable storage repository for the running of virtual machines and applications. This method, as illustrated in Figure 2, allows for the independent scaling and expanding of the compute farm (Hyper-V) and the storage farm (S2D).

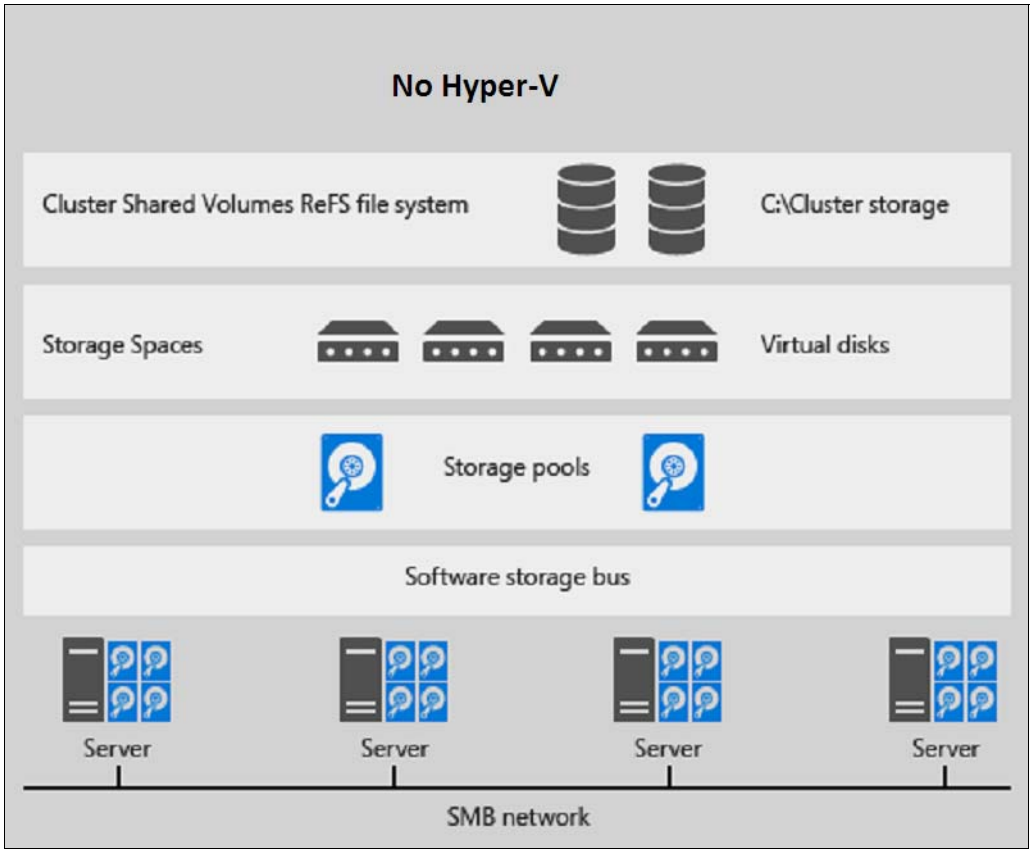


Figure 2 Diagram showing the disaggregated deployment of S2D

In the disaggregated deployment, an independent pool of servers running Hyper-V must be employed to provide the CPU and memory resources - the compute component - for the running of virtual machines whose VHD(X) files reside on the storage environment. The storage component is built using S2D and Scale-Out File Server (SOFS) to provide an independently scalable storage repository for the running of virtual machines and applications such as Microsoft SQL Server.

Figure 3 shows a Hyper-V cluster connected to a disaggregated S2D cluster through Server Message Block (SMB 3) protocol over a high-speed Ethernet network. In addition to Hyper-V, SQL Server can also use directly any storage made available by S2D via SOFS.

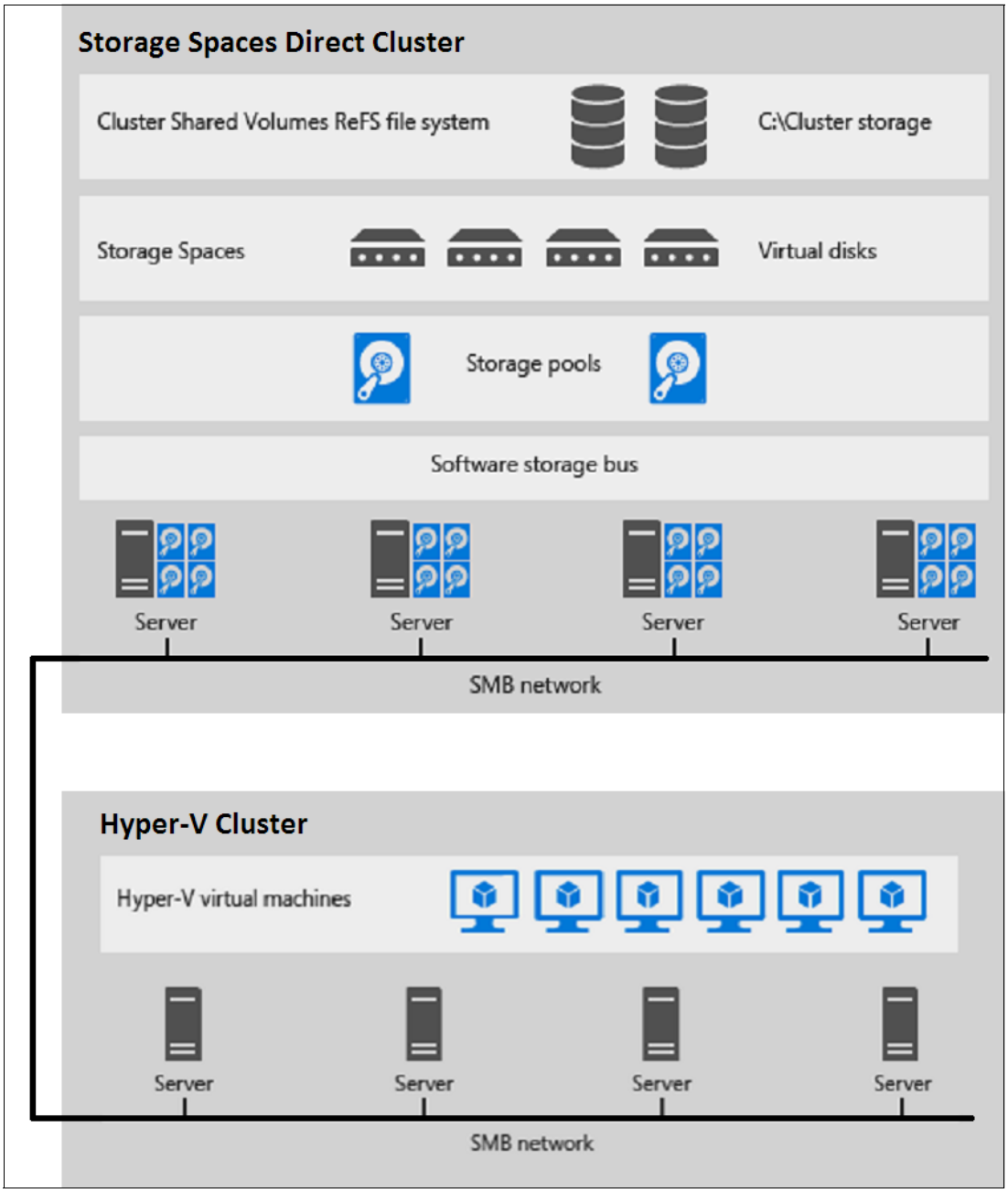


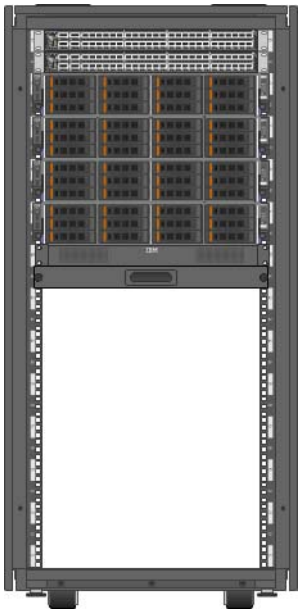
Figure 3 Diagram showing the disaggregated deployment of S2D connected to a Hyper-V cluster

Component model

We will describe this solution by first discussing the hardware components, broken into the classic categories of Compute, Storage, and Network. This includes all the physical hardware, including rack, servers, storage devices, and network switches. We will then move on to discuss the software components, which includes the Windows Server 2016 operating system and its built-in software features that enable the physical hardware to provide storage functionality, as well as systems management components.

Hardware components

This section describes the hardware components of this solution. Figure 4 shows high-level details of a 4-node S2D configuration. The four server/storage nodes and two switches take up a combined total of 10 rack units of space.



Networking: Two Lenovo RackSwitch G8272 switches, each containing:

- 48 ports at 10Gbps SFP+
- 4 ports at 40Gbps QSFP+

Compute: Four Lenovo System x3650 M5 servers, each containing:

- Two Intel Xeon E5-2680 v4 processors
- 256 GB memory
- One dual-port 10GbE Mellanox ConnectX-4 PCIe adapter with RoCE support

Storage in each x3650 M5 server:

- Twelve 3.5” HDD at front
- Two 3.5” HDD + Two 2.5” HDD at rear
- ServeRAID M1215 SAS RAID adapter
- N2215 SAS HBA (LSI SAS3008 at 12 Gbps)

Figure 4 Solution rack configuration using System x3650 M5 systems

Compute

System x3650 M5

The Lenovo System x3650 M5 server (shown in Figure 5) is an enterprise class 2U two-socket versatile server that incorporates outstanding reliability, availability, serviceability, security, and high efficiency for business-critical applications and cloud deployments. It offers a flexible, scalable design and simple upgrade path to 14 3.5” storage devices, with doubled data transfer rate via 12 Gbps SAS internal storage connectivity and up to 1.5 TB of TruDDR4™ Memory. Its onboard Ethernet solution provides four standard embedded Gigabit Ethernet ports and two optional embedded 10 Gigabit Ethernet ports without occupying PCIe slots.

Figure 5 Lenovo System x3650 M5 rack server

Combined with the Intel Xeon processor E5-2600 v4 product family, the Lenovo x3650 M5 server supports high density workloads and performance that is targeted to lower the total cost of ownership (TCO) per virtual machine. Its flexible, pay-as-you-grow design and great expansion capabilities solidify dependability for any kind of virtualized workload with minimal downtime.

The Lenovo x3650 M5 server provides internal storage density with up to 14 x 3.5" storage devices in a 2U form factor with its impressive array of workload-optimized storage configurations. The x3650 M5 offers easy management and saves floor space and power consumption for the most demanding storage virtualization use cases by consolidating the storage and server into one system.

Figure 6 shows the layout of the drives. There are 14x 3.5” drive bays in the server, 12 at the front of the server and two at the rear of the server. Four are 800 GB SSD devices, while the remaining ten drives are 4 TB SATA HDDs. These 14 drives form the tiered storage pool of S2D and are connected to the N2215 SAS HBA. Two 2.5” drive bays at the rear of the server contain a pair of 600 GB SAS HDDs that are mirrored (RAID-1) for the boot drive and connected to the ServeRAID™ M1215 SAS RAID adapter.

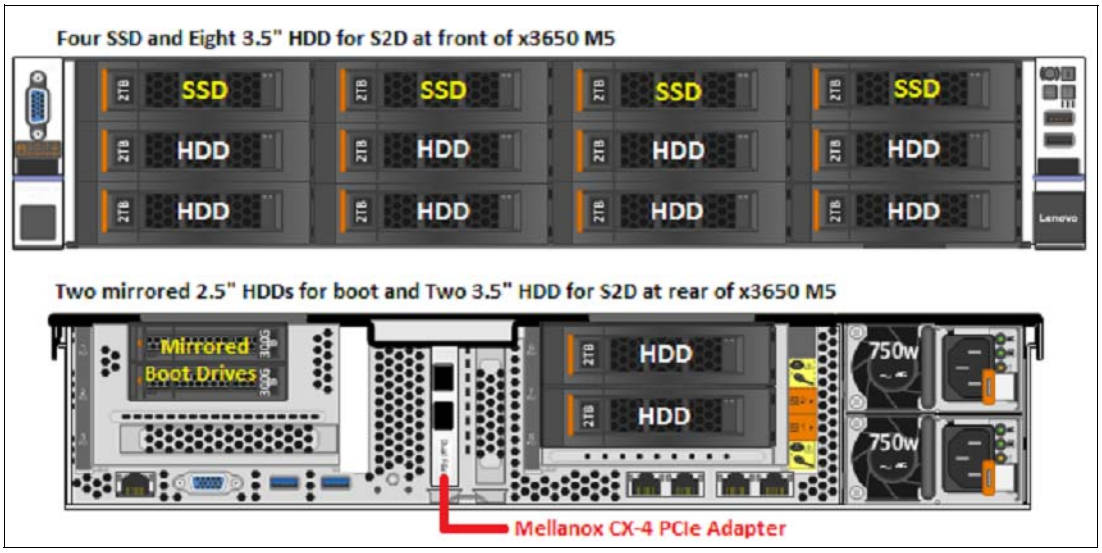


Figure 6 x3650 M5 storage subsystem

One of the requirements for this solution is that a non-RAID storage controller is used for the S2D data devices. Note that using a RAID storage controller set to pass-through mode is not supported at the time of this writing. The ServeRAID adapter is required for high availability of the operating system and is not used by S2D for its storage repository.

For more information, see the following web pages:

- Lenovo System x3650 M5 overview

  http://shop.lenovo.com/us/en/systems/servers/racks/x3650-m5

- Lenovo System x3650 M5 Product Guide

  https://lenovopress.com/lp0068-lenovo-system-x3650-m5-machine-type-8871

Intel Xeon E5-2600 v4 processor family

The powerful Intel Xeon processor E5-2600 v4 product family offers versatility across diverse data center workloads. These processors are designed to help IT address the ever-growing demand for increased performance, agility, operational efficiency, scale-on demand and security throughout their server, storage and network infrastructure - while minimizing their total cost of ownership. These processors take performance and efficiency to new heights across the widest range of workloads including cloud-native, as well as traditional applications (such as business processing, decision support, high performance computing, storage, networking, application development, web infrastructure and collaboration), while providing an array of new technologies for more efficient virtualization, smarter resource orchestration, and enhanced end-to-end platform and data security. The Intel Xeon processor

E5-2600 v4 product family can help enterprises, telecommunication service providers, and cloud service providers get high performance and value from every new server, while accelerating their move toward the next-generation efficiencies of software defined infrastructure (SDI).

For more information, see the following web page:

- Intel Xeon Processor E5 v4

  http://www.intel.com/content/www/us/en/processors/xeon/xeon-processor-e5-family.html

Storage devices

This Lenovo solution uses multiple types of storage media, including Hard Disk Drive (HDD), Solid State Drive (SSD), and Non-Volatile Memory express (NVMe) add-in cards (AIC), which are installed in each of the S2D cluster nodes.

HDD

HDDs are the classic spinning disks that characteristically provide capacity to any storage solution that is not configured as all-flash. HDD I/O performance is typically much lower than SSD and NVMe flash devices, while latency is significantly higher. However, S2D circumvents some of the limitations of HDDs by managing the flow of data to/from these devices.

The Software Storage Bus, new in Storage Spaces Direct, spans the cluster and establishes a software-defined storage fabric in which all the cluster nodes can see all of each other's local devices. The Software Storage Bus dynamically binds the fastest devices present (e.g. SSD) to slower devices (e.g. HDDs) to provide server-side read/write caching that accelerates IO and boosts throughput to the slower devices.

Lenovo offers multiple HDD choices for the System x3650 M5 server, including 4 TB and 6 TB drives that are suitable for S2D. Although higher capacity HDDs are also available, rebuild times can become very lengthy if a drive (or node) fails.

SSD

The Intel SSD DC S3710 Series offers the next generation of data center SSDs optimized for write intensive performance with high endurance and strong data protection. The Intel SSD DC S3710 Series accelerates data center performance with read/write throughput speeds up to 550/520 megabytes per second (MB/s) and 4K random read/write IOPS of up to 85,000/45,000. Applications benefit from 55 microsecond typical latency with max read latencies of 500 microseconds 99.9 percent of the time. Combining performance with low typical active power (less than 6.9 watts), the Intel SSD DC S3710 Series improves data center efficiency with superior quality of service and reduced energy costs.

For more information, see the following website:

- Intel DC S3710 series SSD

  http://www.intel.com/content/www/us/en/solid-state-drives/solid-state-drives-dc-s3710-series.html

NVMe

The consistently high performance of the Intel NVMe Data Center Family for PCIe provides fast, unwavering data streams directly to Intel Xeon processors making server data transfers efficient. NVMe performance consistency provides scalable throughput when multiple NVMe devices are unified into a single storage volume. The massive storage bandwidth increase feeds Intel Xeon processor-based systems giving data center servers a performance boost.

Servers can now support more users simultaneously, compute on larger data sets, and address high-performance computing at a lower TCO.

For more information, see the following website:

- Intel DC P3700 series NVMe AIC

  http://www.intel.com/content/www/us/en/solid-state-drives/ssd-dc-p3700-spec.html

Network devices

The network devices incorporated in this S2D solution include a choice of two 10GbE RDMA network adapters installed in each cluster node as well as the Lenovo RackSwitch G8272 network switch.

10 GbE RDMA capable network adapters

This Lenovo solution for Microsoft S2D offers the choice of two 10GbE RDMA network adapters. These adapters provide high throughput and low latency capabilities for an S2D solution.

- Mellanox ConnectX-4

  The Mellanox ConnectX-4 (CX-4) Lx EN Ethernet Network Controller with 10/25/40/50 Gb/s ports enables the most cost-effective Ethernet connectivity for Hyperscale, cloud infrastructures, enterprise data centers and more, delivering best- in-class performance with smart offloading. ConnectX-4 Lx enables data centers to migrate from 10G to 25G and from 40G to 50G speeds at similar power consumption, cost, and infrastructure needs. With ConnectX-4 Lx, IT and applications managers can enjoy greater data speeds of 25G and 50G to handle the growing demands for data analytics today.

  For more information, see the following web pages:

  - Mellanox ConnectX-4 Lx EN 10/25/40/50 GbE network adapter

    http://www.mellanox.com/page/products_dyn?product_family=214&mtag=connectx_4_lx_en_ic

  - Mellanox ConnectX-4 EN 100 GbE network adapter

    http://www.mellanox.com/page/products_dyn?product_family=204&mtag=connectx_4_en_card

- Chelsio iWARP

  Lenovo also supports Chelsio T520-LL-CR dual-port 10GbE network cards that use the iWARP protocol. This Chelsio NIC can be ordered via the CORE (formerly known as SPORE) process as Lenovo part number 46W0609 - contact your local Lenovo client representative for more information.

  For more information, see the following web pages:

  - Chelsio T520-LL-CR network adapter

    http://www.chelsio.com/nic/unified-wire-adapters/t520-ll-cr/

Lenovo RackSwitch G8272

The Lenovo RackSwitch G8272 switch offers 48 x 10 Gb SFP+ ports and 6 x 40 Gb QSFP+ ports in a 1U form factor. It is an enterprise-class and full-featured data center switch that delivers line-rate, high-bandwidth switching, filtering, and traffic queuing without delaying data. Large data center grade buffers keep traffic moving. Redundant power and fans and numerous high availability features equip the switches for business-sensitive traffic.

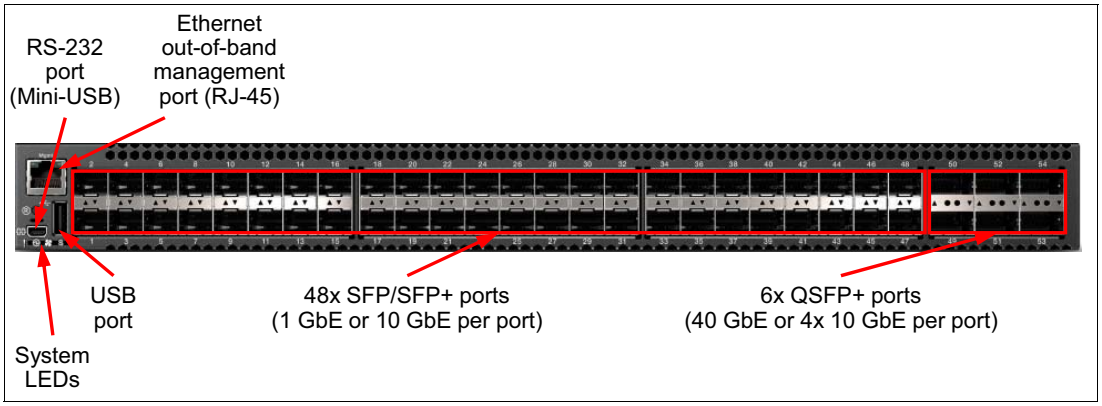


Figure 7 Port descriptions of the Lenovo RackSwitch G8272 high-speed Ethernet switch

Figure 8 Lenovo RackSwitch G8272 high-speed Ethernet switch

The G8272 switch is ideal for latency-sensitive applications, such as client virtualization. It supports Virtual Fabric to help clients reduce the number of I/O adapters to a single dual-port 10 Gb adapter, which helps reduce cost and complexity. Designed with ultra-low latency and top performance in mind, the RackSwitch G8272 also supports Converged Enhanced Ethernet (CEE) and Data Center Bridging (DCB) for support of FCoE and can also be used for NAS or iSCSI.

The G8272 is easier to use and manage, with server-oriented provisioning via point-and-click management interfaces. Its industry-standard command-line interface (CLI) and easy interoperability simplifies configuration for those users who are familiar with Cisco environments.

In addition to the Lenovo RackSwitch G8272 10 GbE switch, the G8124E can be used as a lower cost option and the G8332 40 GbE switch can be used if a high performance solution is required.

For more information, see the following web pages:

- Lenovo RackSwitch G8272 Overview

  http://shop.lenovo.com/us/en/systems/networking/ethernet-rackswitch/g8272/
- Lenovo RackSwitch G8272 Product Guide

  https://lenovopress.com/tips1267-lenovo-rackswitch-g8272
- Lenovo RackSwitch G8124E Overview

  https://www.lenovo.com/images/products/system-x/pdfs/datasheets/rackswitch_g8124e_ds.pdf
- Lenovo RackSwitch G8124E Product Guide

  https://lenovopress.com/tips1271-lenovo-rackswitch-g8124e

- Lenovo RackSwitch G8332 Overview

  https://www.lenovo.com/images/products/system-x/pdfs/datasheets/rackswitch_g8332_ds.pdf

- Lenovo RackSwitch G8332 Product Guide

  https://lenovopress.com/tips1274-lenovo-rackswitch-g8332

Software components

This section describes some of the key Microsoft technology components that are integral parts of any S2D solution, regardless of whether the deployment is hyperconverged or disaggregated. These technologies are used by S2D itself as well as by systems that run workloads that require access to the S2D storage pool. They are included in Windows Server 2016, which provides a rich set of storage features with which you can use lower-cost, industry-standard hardware without compromising performance or availability.

Microsoft Active Directory Services

Microsoft Active Directory (AD), Domain Name Servers (DNS), and Dynamic Host Configuration Protocol (DHCP) servers provide user authentication, domain name service, and dynamic IP service infrastructure for Hyper-V and S2D storage nodes. Improvements to AD in Windows Server 2016 include privileged identity management, time-bound memberships, Azure AD join, and Microsoft Passport.

Microsoft SMB 3 protocol

The Server Message Block (SMB) Protocol is a network file sharing protocol, and as implemented in Microsoft Windows is known as Microsoft SMB Protocol. The set of message packets that defines a particular version of the protocol is called a dialect. The Windows Server 2016 SMB dialect is v3.1.1.

SMB Multichannel

SMB multichannel provides the capability to automatically detect multiple networks for SMB connections. It offers resilience against path failures and transparent failover with recovery without application service disruption with much improved throughput by aggregating network bandwidth from multiple network interfaces. Server applications then can use all available network bandwidth, which makes them more resistant to network failure.

Microsoft S2D leverages the SMB protocol, which detects whether a network adapter has the Remote Direct Memory Access (RDMA) capability, and then creates multiple RDMA connections for that single session (two per interface). This allows SMB to use the high throughput, low latency, and low CPU utilization offered by RDMA-capable network adapters. It also offers fault tolerance if you are using multiple RDMA interfaces.

SMB Direct

SMB Direct (SMB over RDMA) makes available RDMA hardware support for SMB to provide high-performance storage capabilities. SMB Direct is intended to lower CPU usage and latency on the client and server while delivering high IOPS and bandwidth utilization. It can deliver enterprise-class performance without relying on an expensive Fibre Channel SAN. With the CPU offloading and the ability to read and write directly against the memory of the remote storage node, RDMA network adapters can achieve extremely high performance with low latency. SMB Direct is also compatible with SMB Multichannel to achieve load balancing and automatic failover.

For more information, see *What's new in SMB 3.1.1 in Windows Server 2016*:

https://blogs.technet.microsoft.com/josebda/2015/05/05/whats-new-in-smb-3-1-1-in-the-windows-server-2016-technical-preview-2/

Windows Failover Cluster

A failover cluster is a group of independent computers that work together to increase the availability and scalability of clustered roles. The clustered servers (called nodes) are connected by physical network cables and by software. If one or more of the cluster nodes fail, other nodes begin to provide service (a process known as failover). In addition, the clustered roles are proactively monitored to verify that they are working properly. If they are not working, they are restarted or moved to another node.

Windows Failover Clustering includes the following key features, all of which are utilized by this solution:

Cluster Shared Volume (CSV)

A CSV is a shared NTFS or ReFS volume that is made accessible for read and write operations by all nodes within a Windows Server Failover Cluster. To each of the cluster nodes, the CSV provides a consistent single file namespace. Therefore, CSVs simplify the management of many LUNs in a failover cluster. Microsoft SQL Server and Hyper-V data can be stored in CSVs.

Continuous Availability (CA) file share

A feature in Windows Failover Clustering, this technology tracks file operations on highly available file shares so that storage clients (such as those running under Hyper-V or SQL Server) can failover to another node of the storage cluster without interruption. In addition, this capability is used by Microsoft Scale Out File Server (SOFS) to provide scale-out file shares that are continuously available for file-based server application storage. Microsoft SOFS also provides the ability to share the same folder from multiple nodes of the same cluster using a CSV.

Failover Cluster Manager

Failover Cluster Manager provides a user interface with which you can manage all aspects of a Windows Failover Cluster. By using this feature, you can run tasks to create and manage Failover Clusters, Storage Pools, CSVs, and CA File Share servers.

For more information, see the following web pages:

- Microsoft Failover Clustering Overview

  https://technet.microsoft.com/en-us/library/hh831579(v=ws.11).aspx

- What's new in Failover Clustering in Windows Server 2016

  https://technet.microsoft.com/en-us/windows-server-docs/failover-clustering/whats-new-in-failover-clustering

Fault Domain awareness

Failover Clustering enables multiple servers to work together to provide high availability, to provide node fault tolerance. But today's businesses demand ever-greater availability from their infrastructure. For this reason, Failover Clustering in Windows Server 2016 introduced chassis, rack, and site fault tolerance in addition to node fault tolerance.

Fault domains and fault tolerance are closely related concepts. A fault domain is a set of hardware components that share a single point of failure. To be fault tolerant at a certain level, multiple fault domains are required at that level. For example, to be rack fault tolerant, servers and data must be distributed across multiple racks.

There are four levels of fault domains - site, rack, chassis, and node. Nodes are discovered automatically; each additional level is optional.

For more information about fault domains and how to configure them, see the Microsoft web page, *Fault domain awareness in Windows Server 2016*:

https://technet.microsoft.com/en-us/windows-server-docs/failover-clustering/fault-domains

Microsoft virtualization stack

Hyper-V is Microsoft's hardware virtualization product. It lets you create and run a software version of a computer, called a virtual machine. Each virtual machine acts like a complete computer, running an operating system and programs. When you need computing resources, virtual machines give you more flexibility, help save time and money, and are a more efficient way to use hardware than just running one operating system on physical hardware.

Hyper-V runs each virtual machine in its own isolated space, which means you can run more than one virtual machine on the same hardware at the same time. You might want to do this to avoid problems such as a crash affecting the other workloads, or to give different people, groups or services access to different systems.

Improvements to Hyper-V in Windows Server 2016 include nested virtualization, Hyper-V VMCX binary configuration file format, production checkpoints, hot-add/remove memory for Gen 1 and Gen 2 virtual machines, hot-add/remove NICs for Gen 2 virtual machines, PowerShell Direct, virtualized TPM for Gen 2 virtual machines, shielded virtual machines, rolling Hyper-V Cluster upgrade and more.

New networking features in Windows Server 2016 Hyper-V are particularly important to S2D. These include:

- Support for RDMA and Switch Embedded Teaming (SET)

  RDMA can now be enabled 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)

  Improves on 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

  Manages the default class of traffic through the virtual switch within the default class bandwidth. This is in addition to the storage-specific QoS discussed below under “Software-Defined Storage stack”.

Hyper-V Storage NUMA I/O

Windows Server 2016 supports large virtual machines, up to 240 virtual processors (vCPUs) and 12 TB of memory for Generation 2 virtual machines. Any large virtual machine configuration typically also needs scalability in terms of I/O throughput. The Hyper-V storage NUMA I/O capability creates multiple communication channels between the available guest devices and the host storage stack with a specified dedicated set of vCPUs for storage I/O processing. Hyper-V storage NUMA I/O offers a more efficient I/O completion mechanism that performs distribution amongst the virtual processors to avoid expensive inter-processor interruptions. With these improvements, the Hyper-V storage stack can provide scalability improvements in terms of I/O throughput to support the needs of large virtual machine configurations with data intensive workloads like Exchange, SQL Server, and SharePoint.

Hyper-V I/O Balancer

In conjunction with Storage QoS, Hyper-V I/O balancer provides the ability to specify maximum I/O operations per second (IOPS) values for an individual virtual hard disk. These two Windows Server technologies help to control the balance of virtual machine storage demand and performance.

VHDX file format

VHDX is a virtual hard disk format with which you can create resilient high-performance virtual disks up to 64 TB. Microsoft recommends the use of VHDX as the default virtual hard disk format for virtual machines rather than the original VHD format. The VHDX format provides more protection against data corruption during power failures by logging updates to the VHDX metadata structures and the ability to store custom metadata. The VHDX format also provides support for the TRIM command, which results in smaller file size and allows the underlying physical storage device to reclaim unused space. The support for 4KB logical sector virtual disks and the larger block sizes for dynamic and differential disks allows for increased performance.

For more information, see the following web pages:

- Microsoft Hyper-V Technology Overview

  https://technet.microsoft.com/en-us/windows-server-docs/compute/hyper-v/hyper-v-technology-overview

- Microsoft Hyper-V on Windows Server 2016

  https://technet.microsoft.com/en-us/windows-server-docs/compute/hyper-v/hyper-v-on-windows-server

Software-Defined Storage stack

Figure 9 shows a block diagram of the Microsoft Software-Defined Storage stack in Windows Server 2016. Not surprisingly, this chunk of the operating system provides critical functionality to the S2D solution being discussed.

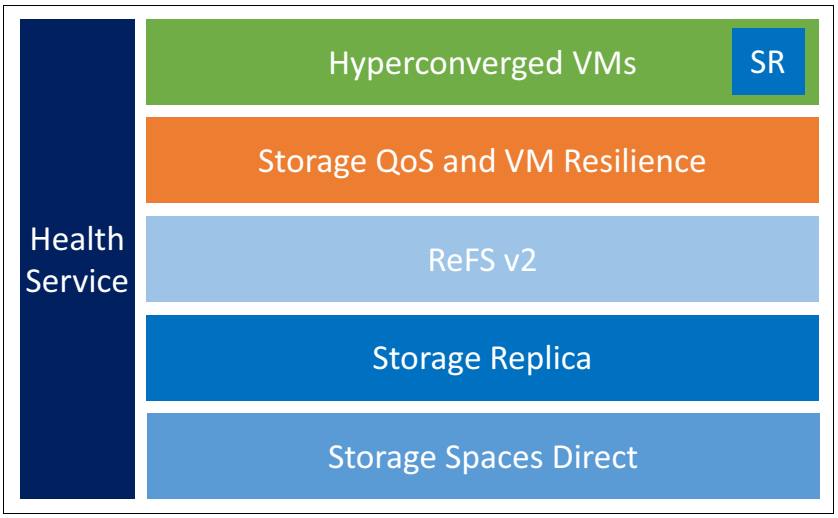


Figure 9 Windows Server 2016 Software-Defined Storage stack diagram

Storage Quality of Service (QoS)

Storage QoS offers the capability to set certain QoS parameters for storage on virtual machines. Storage QoS policies can be created and applied to CSVs and one or more virtual disks on Hyper-V virtual machines. Storage performance is automatically readjusted to meet policies as the workloads and storage loads fluctuate. Storage QoS and Hyper-V I/O balancer

provide the ability to specify maximum I/O throughput values for an individual virtual hard disk. Hyper-V with Storage QoS can throttle storage that is assigned to VHD/VHDX files in the same CSV to prevent a single virtual machine from using all I/O bandwidth and help to control the balance of virtual machine storage demand and performance.

Resilient File System (ReFS)

ReFS is Microsoft's newest file system, designed to maximize data availability, scale efficiently to large data sets across diverse workloads, and provide data integrity by means of resiliency to corruption. It seeks to address an expanding set of storage scenarios and establish a foundation for future innovations. ReFS v2 is the preferred data volume for Windows Server 2016. This updated version provides many new capabilities for private cloud workloads, including improvements to data integrity, resiliency and availability, speed and efficiency, and others.

Note: ReFS does not currently support Windows Server 2016 data deduplication. For those volumes that require data deduplication, the NTFS file system should be specified at volume creation.

Storage Replica

Storage Replica is Windows Server technology that enables storage-agnostic, block-level, synchronous replication of volumes between servers or clusters for disaster recovery. It also enables asynchronous replication to create failover clusters that span two sites, with all nodes staying in sync. Storage Replica supports synchronous and asynchronous replication. Synchronous replication mirrors data within a low-latency network site with crash-consistent volumes to ensure zero data loss at the file-system level during a failure. Asynchronous replication mirrors data across sites beyond metropolitan ranges over network links with higher latencies, but without a guarantee that both sites have identical copies of the data at the time of a failure.

Storage Spaces Direct

S2D is a storage virtualization feature in Windows Server 2016 that provides many features including persistent real-time read and write cache, storage tiering, thin provisioning, mirroring, and parity with just a bunch of disks (JBOD). It gives you the ability to consolidate all of your SAS and SATA connected storage devices into Storage Pools, which are groupings of industry-standard storage devices, such as HDDs, SSDs, or NVMe add-in cards and provides the aggregate performance of all devices in the pool.

For more information, see the following web pages:

- Storage in Windows Server 2016

  https://technet.microsoft.com/en-us/windows-server-docs/storage/storage

- Storage Spaces Direct in Windows Server 2016

  https://technet.microsoft.com/en-us/windows-server-docs/storage/storage-spaces/storage-spaces-direct-overview

- The storage pool in Storage Spaces Direct

  https://blogs.technet.microsoft.com/filecab/2016/11/21/deep-dive-pool-in-spaces-direct/

Operational model

The operational model for this solution includes the infrastructure required to build the solution itself, as well as components required to perform systems management functions for the solution.

Infrastructure

This section describes at a high level the physical infrastructure components of the solution, including the cluster nodes as well as connectivity and networking.

Cluster nodes

This solution uses the Lenovo System x3650 M5 rack server for each cluster node. Each server features the following components:

- Two E5-2680 v4 Intel Xeon processors
- 256 GB of memory (or 128 GB for a disaggregated deployment)
- Two 2.5" HDDs in a RAID-1 pair for boot
- Four on-board 1 Gbps Ethernet ports
- One dual-port 10/25 Gbps Ethernet adapter with RDMA capability
- Two dual-port SAS HBAs

The storage nodes are clustered to provide continuous availability back-end storage for Microsoft Hyper-V or SQL Server. Microsoft S2D is used to pool the disk resources installed in all of the cluster nodes. Cluster virtual disks are created from the pool and presented as CSVs. S2D provides RAID resiliency options of mirror and parity, as well as a combination of these two methods, during the creation of a volume.

SSDs and NVMe add-in cards can be used for storage tiering or as a cache for virtual disks. Virtual disks are formatted using Microsoft ReFS as a default and presented as CSVs. CSVs are accessed by all storage nodes in the S2D cluster. Server Message Block (SMB) v3 file shares are created by using CSV volumes on the S2D cluster.

Note: ReFS does not currently support Windows Server 2016 data deduplication. For those volumes that require data deduplication, the NTFS file system should be specified at volume creation.

Each cluster node is equipped with a dual-port 10/25 Gbps Ethernet adapter. The Ethernet adapter is RDMA-capable, which is used by the SMB-Direct feature of the SMB v3 protocol to provide high speed and low latency network access with low CPU utilization. These RDMA adapters are dedicated for SMB file share storage traffic.

With the file shares in place, Hyper-V or SQL Server hosts can use the files shares as back-end storage to store VHD(X) virtual disk or SQL Server databases files. If there is a storage node failure while Hyper-V or SQL Server is accessing the storage, the host's SMB and witness client are notified immediately by the cluster witness service and the host automatically switches to the next best available storage node, re-establishing the session, and continuing I/O operations from the point of failure. This process is transparent to Hyper-V and SQL Server and appears merely as a brief pause in I/O activity. This transparent failover capability or continuous availability is what makes S2D a highly available storage solution.

Connectivity and networking

This solution uses two Lenovo RackSwitch G8272 network switches for high availability and performance capabilities and is ideal for I/O-intensive storage software applications, such as Microsoft SOFS, Hyper-V, and SQL Server. These Lenovo top-of-rack switches support the latest CEE standard, which is required for network adapters that have RDMA capability. Network adapters that have RDMA can function at full speed with low latency while using little CPU. For workloads (such as Hyper-V or Microsoft SQL Server), this feature enables S2D storage to resemble local direct-attached block storage.

The block diagram in Figure 10 shows network connectivity between a disaggregated S2D solution and a Hyper-V (or SQL Server) Failover Cluster. As the diagram shows, each S2D storage node has one independent 10 Gbps Ethernet connection to each of the two Lenovo RackSwitch G8272 network switches.

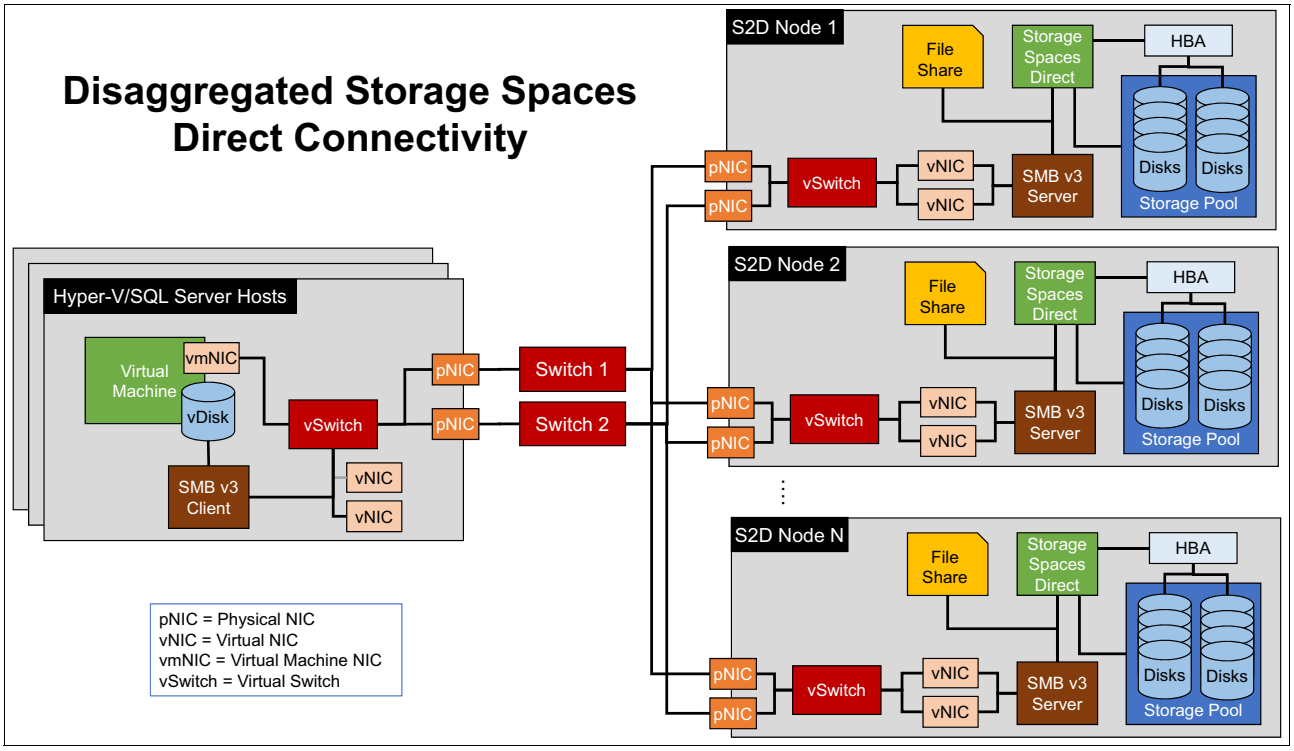


Figure 10 S2D Logical networking block diagram

Data Center Bridging (DCB)

Significant advances in the Windows Server 2016 network stack enable S2D. One particular enhancement vital to S2D performance is support for Data Center Bridging. DCB is a collection of standards that defines a unified 802.3 Ethernet media interface, or fabric, for LAN and SAN technologies. DCB extends the 802.1 bridging specification to support the coexistence of LAN-based and SAN-based applications over the same networking fabric within a data center. DCB also supports technologies, such as FCoE and iSCSI, by defining link-level policies that prevent packet loss.

DCB consists of the following 802.1 standards that specify how networking devices can interoperate within a unified data center fabric:

- Priority-based Flow Control (PFC)

  PFC is specified in the IEEE 802.1Qbb standard. This standard is part of the framework for the DCB interface. PFC supports the reliable delivery of data by substantially reducing

packet loss due to congestion. This allows loss-sensitive protocols, such as FCoE, to coexist with traditional loss-insensitive protocols over the same unified fabric.

PFC specifies a link-level flow control mechanism between directly connected peers. PFC is similar to IEEE 802.3 PAUSE frames but operates on individual 802.1p priority levels instead. This allows a receiver to pause a transmitter on any priority level.

For more information on PFC, see the following web page:

https://msdn.microsoft.com/en-us/windows/hardware/drivers/network/priority-based-flow-control--pfc

- Enhanced Transmission Selection (ETS)

  ETS is a transmission selection algorithm that is specified in the IEEE 802.1Qaz standard. This standard is part of the framework for the DCB interface. ETS allocates bandwidth between traffic classes that are assigned to different IEEE 802.1p priority levels. Each traffic class is allocated a percentage of available bandwidth on the data link between directly connected peers. If a traffic class doesn't use its allocated bandwidth, ETS allows other traffic classes to use the available bandwidth that the traffic class is not using.

  For more information on ETS, see the following web page:

  https://msdn.microsoft.com/en-us/windows/hardware/drivers/network/enhanced-transmission-selection--ets--algorithm

- Data Center Bridging Exchange (DCBX) Protocol

  The Data Center Bridging Exchange (DCBX) protocol is also specified in the IEEE 802.1Qaz standard. DCBX allows DCB configuration parameters to be exchanged between two directly connected peers, allowing these peers to adapt and tune QoS parameters to optimize data transfer over the connection.

  DCBX is also used to detect conflicting QoS parameter settings between the network adapter (local peer) and the remote peer. Based on the local and remote QoS parameter settings, the miniport driver resolves conflicts and derives a set of operational QoS parameters. The network adapter uses these operational parameters for the prioritized transmission of packets to the remote peer.

  For more information about priority levels, see the following web page:

  https://msdn.microsoft.com/en-us/windows/hardware/drivers/network/ieee-802-1p-priority-levels

This solution focuses on the hardware components to Microsoft S2D. It is assumed that there is a network infrastructure that supports client connections, and that Active Directory, DNS, and DHCP servers are in place at the customer site.

As a best practice, the switches described in this solution can be dedicated for storage traffic between Hyper-V or SQL Server and storage nodes. This configuration ensures the best throughput and lowest I/O latency for storage.

Systems management

In addition to the infrastructure required for the S2D solution itself, it is important to provide a method to manage the systems and processes that make up the solution. This section describes some options for providing this systems management functionality.

Lenovo XClarity Administrator

Lenovo XClarity™ Administrator is a centralized resource management solution that reduces complexity, speeds up response, and enhances the availability of Lenovo server systems and solutions.

Lenovo XClarity Administrator provides agent-free hardware management for Lenovo servers, storage, network switches, and HX Series appliances. Figure 11 on page 24 shows the Lenovo XClarity administrator dashboard. Lenovo XClarity Administrator is a virtual appliance that is quickly imported into a virtualized environment server configuration.

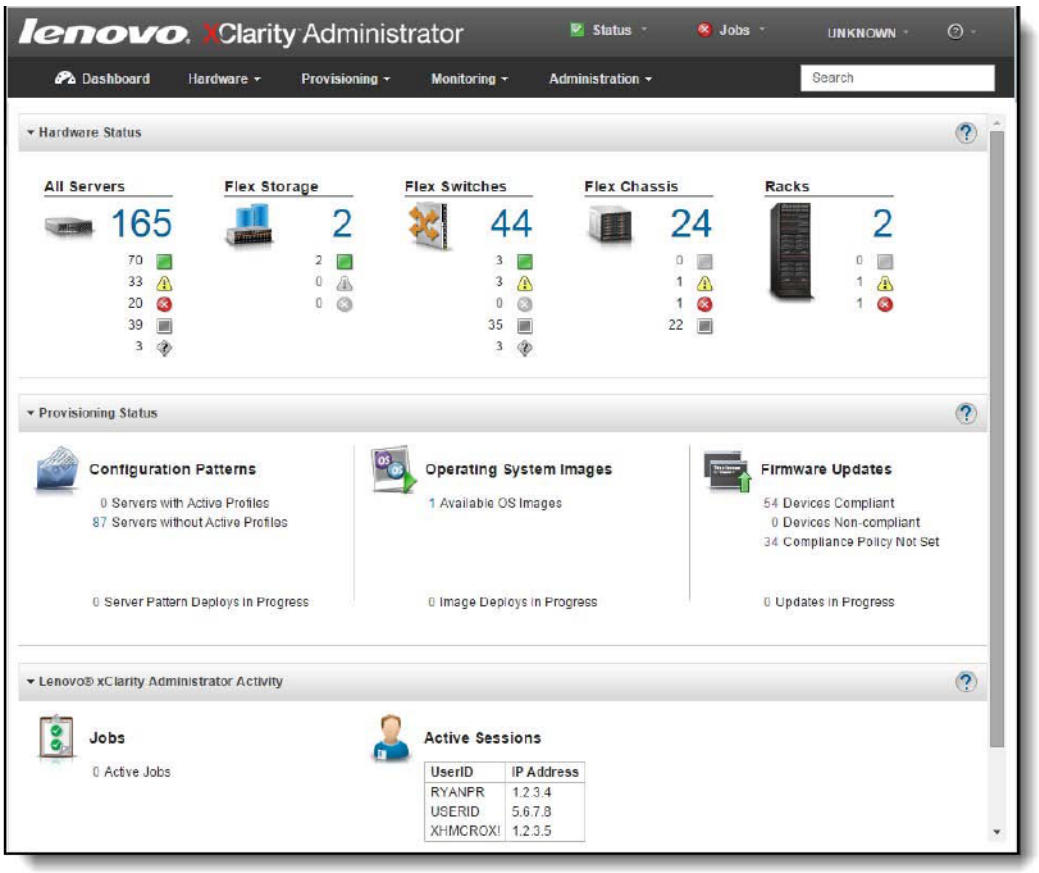


Figure 11 XClarity Administrator dashboard

Lenovo XClarity Administrator delivers Lenovo resources faster. With a simplified administration dashboard, the following functions can be easily achieved:

- Discovery
- Inventory
- Monitoring
- Firmware updates
- Firmware compliance
- Configuration management
- Deployment of operating systems and hypervisors to bare metal servers
- Lenovo XClarity mobile app for Android and iOS devices

Note: If using XClarity Administrator, it is a best practice to not install it inside an S2D Hyperconverged environment. Doing so opens the possibility of rebooting the host node on which the XCLA virtual machine is running during host firmware updates.

XClarity Integrator for Microsoft System Center

The Lenovo XClarity Integrator for Microsoft System Center (LXCI for MSSC) is an offering that provides IT administrators the ability to integrate the management features of Lenovo XClarity Administrator and System x® servers with Microsoft System Center. Lenovo expands Microsoft System Center server management capabilities by integrating Lenovo hardware management functionality, providing affordable, basic management of physical and virtual environments to reduce the time and effort required for routine system administration. It provides the discovery, configuration, monitoring, event management, and power monitoring needed to reduce cost and complexity through server consolidation and simplified management.

The Lenovo XClarity Integrator for Microsoft System Center provides:

- Seamless Integration with Lenovo XClarity Administrator, providing a physical hierarchy view of Lenovo infrastructure, visualizing inventory map views of managed servers and conveniently configuring servers with configuration patterns
- Integrated end-to-end management of System x hardware with monitoring of both physical and virtual server health
- Operating system and hypervisor deployment with the latest Lenovo firmware and driver update management
- Minimized maintenance downtime through non-disruptive rolling firmware updates or server reboots which automates the VM migration and update process of the hosts in a cluster environment without any workload interruption
- Minimized unplanned downtime by monitoring user-defined hardware events and automatically evacuating VMs in response to selected events to protect your workloads
- Collection of Lenovo specific hardware inventory of Lenovo System x
- Remote power control of servers via the Microsoft System Center Console
- Ability to write and edit configuration packs to perform compliance checking on Lenovo System x servers
- Remote server management, independent of operating system state

Microsoft System Center 2016

Microsoft System Center 2016 delivers a simplified datacenter management experience that provides control of the IT environment for on-premises private and hybrid cloud solutions. Microsoft System Center provides visibility and control of data and applications that live across multiple systems, from a single solution. With the Lenovo LXCI for MSSC integration into Microsoft System Center 2016 Operations Manager and Virtual Machine Manager, the combined solution provides end-to-end IT infrastructure management for private and hybrid cloud solutions. Operations Manager and Virtual Machine Manager provide the following capabilities:

Operations Manager

Monitor health, capacity, and usage across applications, workloads and infrastructure, including Microsoft Public Azure and Office 365. Benefit from broader support for Linux environments, improved monitoring with management pack discoverability, data-driven alert management, and integration with Operations Management Suite for rich analytics and insights.

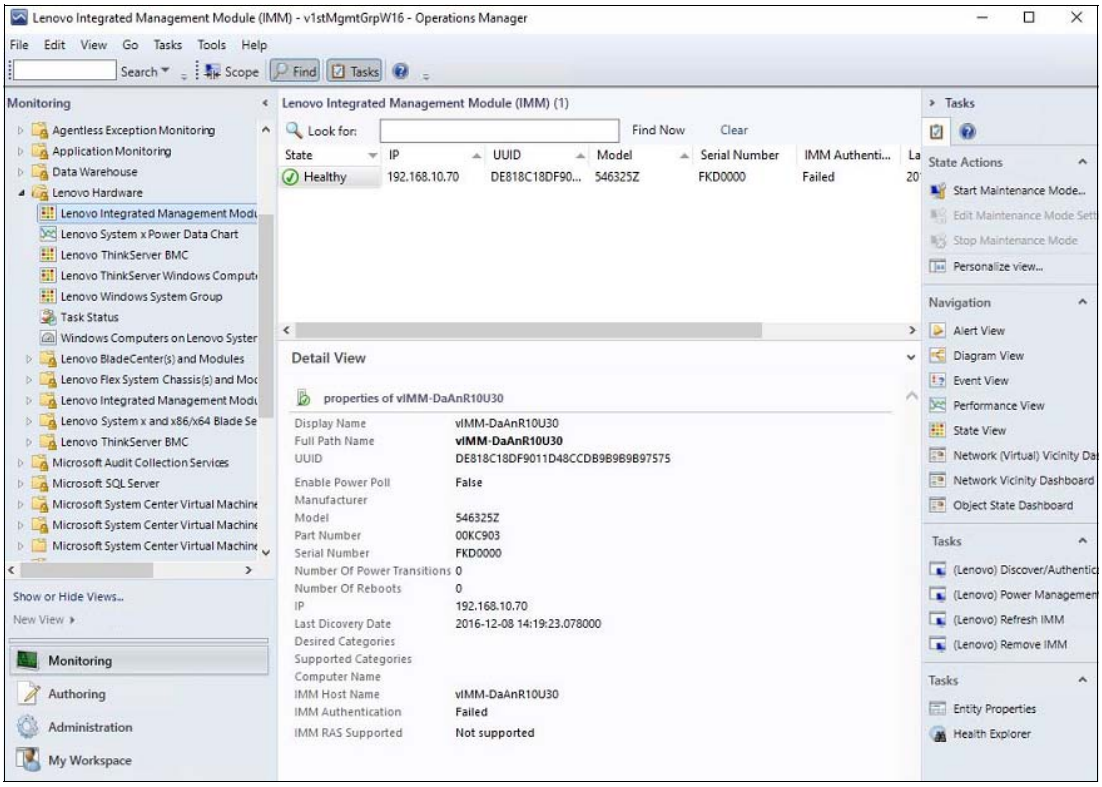


Figure 12 XClarity Integrator for Microsoft System Center showing integration into Operations Manager

Virtual Machine Manager

Deploy and manage a virtualized, software-defined datacenter with a comprehensive solution for networking, storage, compute, and security. Easily manage Windows 2016 capabilities, including Nano Server, shielded VMs, hyperconverged clusters, and more.

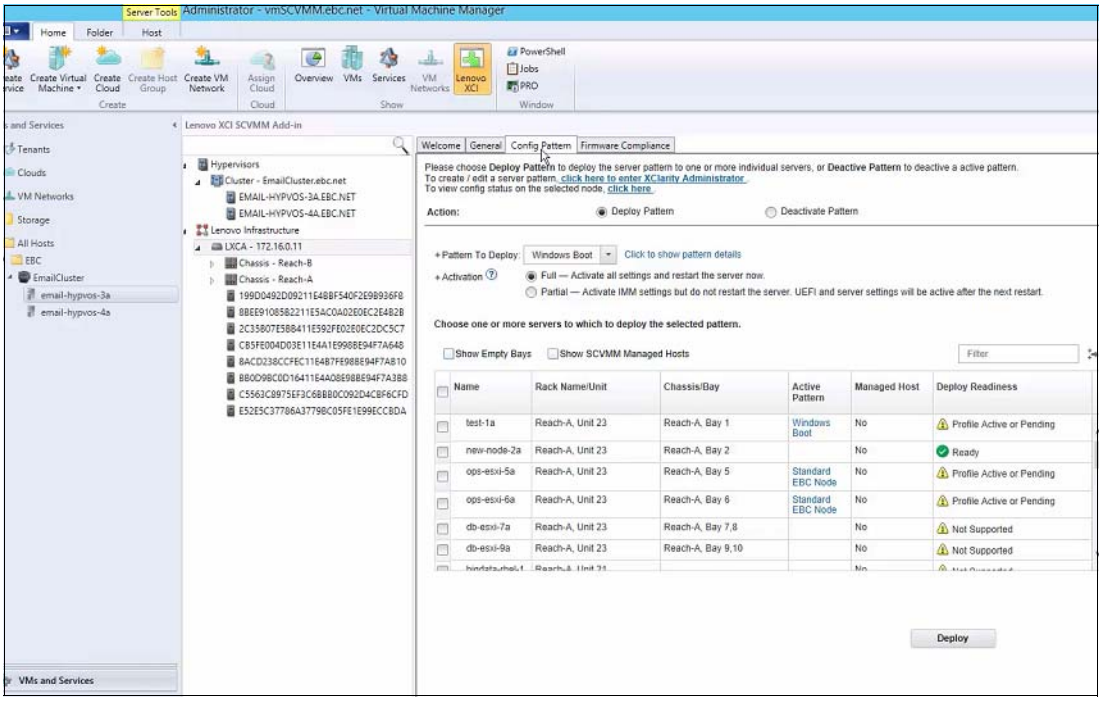


Figure 13 XClarity Integrator for System Center showing integration into Virtual Machine Manager

Microsoft Windows Azure Pack

The advent of cloud computing has transformed the datacenter. To provide more flexibility and reduce costs for private and hybrid cloud operations of the enterprise customers, similar to Microsoft Public Azure, Microsoft offers enterprise customers a familiar technology stack, Windows Azure Pack (WAP), which enterprises can trust and integrate with their on-premises data centers.

WAP provides a multi-tenant, self-service cloud that works on top of existing software and hardware investments. Building on the familiar foundation of Windows Server and System Center, WAP offers a flexible and familiar solution that IT departments can take advantage of to deliver self-service provisioning and management of infrastructure - Infrastructure as a service (IaaS), and application services - Platform as a Service (PaaS), such as Web Sites and Virtual Machines.

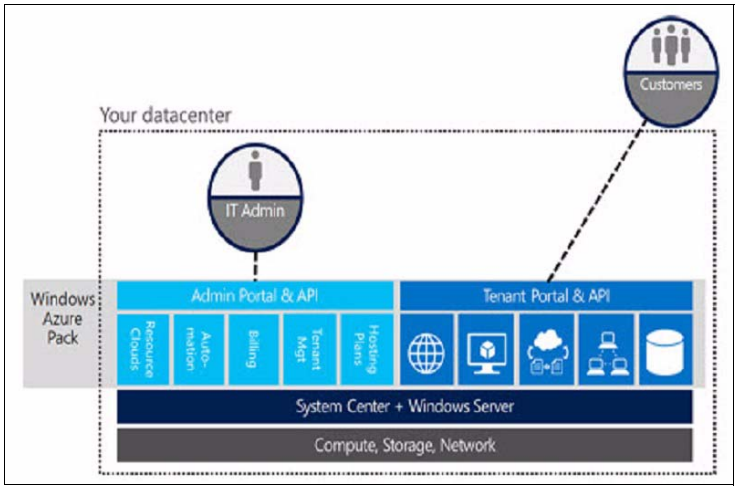


Figure 14 View of Microsoft Windows Azure Pack in the datacenter

Similar to Microsoft Public Azure, WAP offers consistent self-service management experiences. The Management Portal in WAP allows control of how IT services are offered to tenants while also providing tenants with a rich, self-service user experience for provisioning and managing resources. To enable this functionality, WAP offers the following Management portals:

Tenant Management Portal - This portal, consistent with the Windows Azure Developer portal experience found in Microsoft Public Azure, offers self-service provisioning and management capabilities for tenants. Multiple authentication technologies include Active Directory Federation Services (ADFS).

Administrator Management Portal - This portal enables administrators to configure and manage the services and resource clouds that are made available to tenants.

Note: Windows Azure Pack is not the same as Microsoft Azure Stack. While WAP is an add-on to System Center, Azure Stack is actually an extension of Microsoft Public Azure into a customer's on-premises datacenter and does not require System Center. Microsoft plans to release Azure Stack in the mid-2017 timeframe.

Deployment considerations

While planning an S2D deployment, several considerations must be made. Among these, storage performance and sizing are likely at the top of the list. In addition and related to performance and sizing, it is important to decide whether an all-flash or hybrid deployment is preferred. This section discusses details that should help make these decisions.

Storage performance and sizing

S2D storage performance and sizing are important elements of providing an optimal storage solution for Microsoft Hyper-V and SQL Server, as well as other workloads. The following outlines some of the key aspects of storage performance and sizing for S2D.

Storage performance is all about the cache

S2D dynamically binds the fastest devices present (SSD or NVMe) to slower devices (HDDs) to provide server-side read/write caching that accelerates I/O and boosts throughput. This cache mechanism improves performance by handling bursty small random write I/O operations to virtual disks.

It is important to note that the cache is independent of the storage pool and volumes and is handled automatically by S2D except when the entire solution uses only a single media type (such as SSD). In this case, no cache is configured automatically. You have the option to manually configure higher-endurance devices to cache for lower-endurance devices of the same type.

In deployments using multiple storage media types, S2D consumes all of the fastest performing devices installed on each of the cluster nodes and assigns these devices to be used as cache for the storage pool. These devices do not contribute to the usable storage capacity of the solution.

The behavior of the cache is determined automatically based on the media type of the devices that are being cached for. When caching for solid-state devices (such as NVMe caching for SSDs), only writes are cached. When caching for rotational devices (such as SSDs caching for HDDs), both reads and writes are cached.

Even though performance of both read and write operations is very good on solid-state devices, writes are cached in all-flash configurations in order to reduce wear on the capacity devices. Many writes and re-writes can coalesce in the cache and then de-stage only as needed, reducing the cumulative number and volume of write operations to the capacity devices, which can extend their lifetime significantly. For this reason, we recommend selecting higher-endurance, write-optimized devices for the cache.

Because reads do not significantly affect the lifespan of flash devices, and because solid-state devices universally offer low read latency, reads are not cached. This allows the cache to be dedicated entirely to writes, maximizing its effectiveness. This results in write characteristics being dictated by the cache devices, while read characteristics are dictated by the capacity devices.

When caching for HDDs, both reads and writes are cached to provide flash-like latency for both. The read cache stores recently and frequently read data for fast access and to minimize random traffic to the HDDs. Writes are cached to absorb bursts and to coalesce writes and re-writes, minimizing the cumulative traffic to the capacity devices.

S2D implements an algorithm that de-randomizes writes before de-staging them, to emulate an I/O pattern to disk that seems sequential even though the actual I/O coming from the workload is random. This minimizes latency and maximizes throughput to the HDDs.

One last point regarding S2D cache: Earlier we stated that the device binding in S2D is dynamic. Knowing the rules for how S2D assigns cache devices means that cache and capacity devices can be added independently, whenever necessary. Although there is no restriction regarding the ratio of cache to capacity devices, it is a best practice to ensure that the number of capacity devices is a multiple of the number of cache devices. This ensures balanced and symmetrical storage performance across the cluster.

Storage sizing

Given the preceding discussion of storage performance, it is obvious that any exercise in sizing the solution must consider both the cache and the storage pool itself. The ideal use case for a Microsoft S2D solution is that the daily application working set is stored entirely in the S2D cache layer, while infrequently used data is de-staged to the storage pool (capacity layer).

The best approach is to monitor the production environment using the Windows Performance Monitor (PerfMon) tool for some time to analyze how much data is used daily. In addition, it can be very useful (and illuminating) to monitor the **Cache Miss Reads/sec** counter, which is under the **Cluster Storage Hybrid Disk** counter. This counter will reveal the rate of cache misses, which increases if the cache becomes overwhelmed (i.e. the working set is too large for the cache size).

Each environment is unique; however, as a general best practice and recommendation, the cache-to-pool ratio recommended by Microsoft is approximately 10%. For example, if the S2D storage pool has a total usable capacity of 100 TB, the cache layer should be configured at about 10 TB.

All-flash vs. hybrid deployments

One of the key decisions that needs to be made during the design phase of an S2D solution is to determine which media types will be used in the solution. S2D can be configured using a single media type (such as SSD), or up to three media types (NVMe, SSD, and HDD). This section provides information that can help determine how many and which media types to use.

All-flash deployments

All-flash deployments rely on NVMe and/or SSD media types for the entire storage environment. To be clear, they can use a single media type or both media types, depending on customer requirements. The more typical option is to use NVMe devices for the cache layer and SSD devices for the storage pool (capacity layer) of the solution.

As noted previously, an S2D solution built using all NVMe or all SSD devices can benefit from some manual configuration, since the cache layer is not automatically configured when S2D is enabled in this type of deployment. Specifically, it is useful to use higher-endurance devices to cache for lower-endurance devices of the same type. To do this, the device model that should be used for cache is specified via the **-CacheDeviceModel** parameter of the **Enable-ClusterS2D** cmdlet. Once Storage Spaces Direct is enabled, all drives of that model will be used for the cache layer.

If using both NVMe and SSD devices, S2D automatically assigns the NVMe devices to be used for caching (since they are faster than SSDs), while the SSD devices are consumed by the S2D storage pool.

Hybrid deployments

For the purpose of this discussion, a “hybrid” deployment is any deployment that includes both flash and rotational storage media. One of the main goals of this type of deployment is to balance the performance of flash media with the relatively inexpensive capacity provided by HDDs. The most popular hybrid deployments use NVMe or SSD devices to provide caching for a storage pool made up of HDD devices.

The final hybrid deployment type uses all three supported media types, NVMe, SSD, and HDD. In this scenario, NVMe devices are used for cache, while SSD and HDD devices are used for the storage pool. A key benefit of this environment is its flexibility, since it can be made to behave as two independent storage pools. Volumes can be created using only SSD media (a Performance volume), only HDD media (a Capacity volume), or a combination of SSD and HDD media (a Multi-Resilient volume or MRV).

Performance volume

A Performance volume uses three-way mirroring (assuming a minimum of four nodes in the S2D cluster) to keep three separate copies of all data, with each copy being automatically stored on drives of different nodes. Since three full copies of all data are stored, the storage efficiency of a Performance volume is 33.3%—to write 1 TB of data, you need at least 3 TB of physical storage available in the volume. Three-way mirroring can safely tolerate two hardware problems (drive or server) at a time.

This is the recommended volume type for any workload that has strict latency requirements or that generates significant mixed random IOPS, such as SQL Server databases or performance-sensitive Hyper-V virtual machines.

Capacity volume

A Capacity volume uses dual parity (again assuming four nodes or more) to provide the same fault tolerance as three-way mirroring but with better storage efficiency. With four storage nodes, storage efficiency is 50.0%—to store 2 TB of data, you need 4 TB of physical storage capacity available in the volume. Storage efficiency increases as nodes are added, providing 66.7% efficiency with seven nodes. However, this efficiency comes with a performance penalty, since parity encoding is more compute-intensive. Parity calculations inevitably increase CPU utilization and I/O latency, particularly on writes, compared to mirroring.

This is the recommended volume type for workloads that write infrequently, such as data warehouses or “cold” storage, since storage efficiency is maximized. Certain other workloads, such as traditional file servers, virtual desktop infrastructure (VDI), or others that don’t create lots of fast-drifting random I/O traffic or don’t require top performance may also use this volume type, at your discretion.

Multi-Resilient volume (MRV)

An MRV mixes mirroring and parity. Writes land first in the mirrored portion and are gradually moved into the parity portion later. This accelerates ingestion and reduces resource utilization when large writes arrive by allowing the compute-intensive parity encoding to happen over a longer time. When sizing the portions, consider that the quantity of writes that happen at once (such as one daily backup) should comfortably fit in the mirror portion.

This is the recommended volume type for workloads that write in large sequential passes, such as archival or backup targets.

Important: It is not recommended to use MRVs for typical random I/O workloads.

Deployment best practices

To be successful in deploying Microsoft S2D, adhere to the following general deployment best practices:

Verify purchased hardware

Verify the disks to make sure that the disk performance of the same disk model is consistent. This verification can be useful in isolating issues, such as hardware components that might have different firmware levels because of different manufacturing and purchase dates.

For more information, see the following web pages:

- Storage Spaces Physical Disk Validation Script

  http://gallery.technet.microsoft.com/scriptcenter/Storage-Spaces-Physical-7ca9f304
- Validate Hardware for a Failover Cluster

  http://technet.microsoft.com/en-us/library/jj134244.aspx

Deploy and configure S2D

For details on how to deploy S2D on Lenovo hardware, see the *Microsoft Storage Spaces Direct (S2D) Deployment Guide*, available from:

https://lenovopress.com/lp0064

Validate S2D performance

Use an I/O performance tool to verify the S2D performance.

Test workloads against S2D

Run your application workloads against SOFS to ensure S2D meets your I/O requirements.

Resources

For more information about the topics that are described in this document, see the resources listed in this section.

- Microsoft Storage Spaces Direct (S2D) Deployment Guide

  https://lenovopress.com/lp0064.pdf
- Storage Spaces Direct in Windows Server 2016

  https://technet.microsoft.com/en-us/windows-server-docs/storage/storage-spaces/storage-spaces-direct-overview
- Server Storage at Microsoft, the official blog of the Windows Server storage engineering team

  https://blogs.technet.microsoft.com/filecab
- Lenovo System x3650 M5 Overview

  http://shop.lenovo.com/us/en/systems/servers/racks/x3650-m5
- Lenovo System x3650 M5 Product Guide

  https://lenovopress.com/lp0068-lenovo-system-x3650-m5-machine-type-8871
- Intel Xeon Processor E5 v4

  http://www.intel.com/content/www/us/en/processors/xeon/xeon-processor-e5-family.html

- Lenovo RackSwitch G8272 Overview

  http://shop.lenovo.com/us/en/systems/networking/ethernet-rackswitch/g8272
- Lenovo RackSwitch G8272 Product Guide

  https://lenovopress.com/tips1267-lenovo-rackswitch-g8272
- Lenovo RackSwitch G8124E Overview

  https://www.lenovo.com/images/products/system-x/pdfs/datasheets/rackswitch_g8124e_ds.pdf
- Lenovo RackSwitch G8124E Product Guide

  https://lenovopress.com/tips1271-lenovo-rackswitch-g8124e
- Lenovo RackSwitch G8332 Overview

  https://www.lenovo.com/images/products/system-x/pdfs/datasheets/rackswitch_g8332_ds.pdf
- Lenovo RackSwitch G8332 Product Guide

  https://lenovopress.com/tips1274-lenovo-rackswitch-g8332

Lenovo Professional Services

Lenovo offers an extensive range of solutions, from the simple OS-only laden product to much more complex solutions running cluster and cloud technologies. For customers looking for assistance in the form of design, deploy or migrate, Lenovo Professional Services is your go-to partner.

Our worldwide team of IT Specialists and IT Architects can help customers scope and size the right solutions to meet their requirements, and then accelerate the implementation of the solution with our on-site and remote services. For customers also looking to elevate their own skill sets, our Technology Trainers can craft services that encompass solution deployment plus skills transfer, all in a single affordable package.

To inquire about our extensive service offerings and solicit information on how we can assist in your new Storage Spaces Direct implementation, please contact us at: x86svcs@lenovo.com

For more information about our service portfolio, please see our website:

http://shop.lenovo.com/us/en/systems/services/?menu-id=services

Appendix: Lenovo bill of materials

For a Bill of Materials (BOM) for different configurations of hardware for Microsoft S2D deployments, refer to the *Microsoft Storage Spaces Direct (S2D) Deployment Guide* available from:

https://lenovopress.com/lp0064

Authors

This paper was produced by the following team of specialists:

Dave Feisthammel is a Microsoft Solutions Architect working at the Lenovo Center for Microsoft Technologies in Kirkland, Washington. He has over 24 years of experience in the IT field, including four years as an IBM client and 14 years working for IBM. His areas of expertise include systems management, as well as virtualization, storage, and cloud technologies.

David Ye is a Senior Solutions Architect and has been working at Lenovo Center for Microsoft Technologies for 15 years. He started his career at IBM as a Worldwide Windows Level 3 Support Engineer. In this role, he helped customers solve complex problems and was involved in many critical customer support cases. He is now a Senior Solutions Architect in the System x Enterprise Solutions Technical Services group, where he works with customers on Proof of Concepts, solution sizing, performance optimization, and solution reviews. His areas of expertise are Windows Server, SAN Storage, Virtualization, and Microsoft Exchange Server.

Thanks to the following people for their contributions to this project:

- Valentin Danciu
- Vinay Kulkarni
- Michael Miller
- Zhi (Paul) Wang
- David Watts

Notices

Lenovo may not offer the products, services, or features discussed in this document in all countries. Consult your local Lenovo representative for information on the products and services currently available in your area. Any reference to a Lenovo product, program, or service is not intended to state or imply that only that Lenovo product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any Lenovo intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any other product, program, or service.

Lenovo may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

Lenovo (United States), Inc.
1009 Think Place - Building One
Morrisville, NC 27560
U.S.A.
Attention: Lenovo Director of Licensing

LENOVO PROVIDES THIS PUBLICATION “AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. Lenovo may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

The products described in this document are not intended for use in implantation or other life support applications where malfunction may result in injury or death to persons. The information contained in this document does not affect or change Lenovo product specifications or warranties. Nothing in this document shall operate as an express or implied license or indemnity under the intellectual property rights of Lenovo or third parties. All information contained in this document was obtained in specific environments and is presented as an illustration. The result obtained in other operating environments may vary.

Lenovo may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any references in this publication to non-Lenovo Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this Lenovo product, and use of those Web sites is at your own risk.

Any performance data contained herein was determined in a controlled environment. Therefore, the result obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

© Copyright Lenovo 2017. All rights reserved.
Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by Global Services Administration (GSA) ADP Schedule Contract

This document was created or updated on March 24, 2017.

Send us your comments via the **Rate & Provide Feedback** form found at
http://lenovopress.com/lp0569

Trademarks

Lenovo, the Lenovo logo, and For Those Who Do are trademarks or registered trademarks of Lenovo in the United States, other countries, or both. These and other Lenovo trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by Lenovo at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of Lenovo trademarks is available on the Web at http://www.lenovo.com/legal/copytrade.html.

The following terms are trademarks of Lenovo in the United States, other countries, or both:

Lenovo®	Lenovo(logo)®	TruDDR4™
Lenovo XClarity™	ServeRAID™	
RackSwitch™	System x®	

The following terms are trademarks of other companies:

Intel, Xeon, and the Intel logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Active Directory, Azure, Hyper-V, Microsoft, Microsoft Passport, Office 365, PowerShell, SharePoint, SQL Server, Windows, Windows Azure, Windows Server, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.