# Dell EMC Integrated System for Microsoft Azure Stack HCI: Stretched Cluster Deployment

Reference Architecture Guide


**Abstract**

This reference architecture guide provides an overview of the Microsoft Azure Stack HCI operating system and guidance on how to deploy stretched clusters in your environment.


**Dell Technologies Solutions**


**Part Number: H18573**
**February 2021**


DELL Technologies

## Notes, cautions, and warnings

**NOTE:** A NOTE indicates important information that helps you make better use of your product.

**CAUTION: A CAUTION indicates either potential damage to hardware or loss of data and tells you how to avoid the problem.**

**WARNING: A WARNING indicates a potential for property damage, personal injury, or death.**

# Contents

# 1

# Introduction

This chapter presents the following topics:

**Topics:**



## Document overview

This reference architecture guide provides an overview of the Microsoft Azure Stack HCI operating system and guidance on how to deploy stretched clusters in your environment. The guide provides network topology references and best practices to consider during a stretched cluster deployment.

The Microsoft Azure Stack HCI operating system can be deployed in both standalone and stretched cluster environments. For the deployment steps for a standalone cluster and end-to-end deployment steps with network and host configuration options, see the Network Integration and Host Network Configuration Options article.

Dell Technologies offers integrated systems with the new Azure Stack HCI operating system. This guide applies to select configurations of the integrated systems built using AX nodes.

## Audience and scope

This guide is for systems engineers, field consultants, partner engineering team members, and customers with knowledge of deploying hyperconverged infrastructures (HCIs) with Windows Server operating systems and the newly released Azure Stack HCI operating system (20H2).

Customer site-to-site networking configuration and guidance is outside the scope of this document.

### Assumptions

This guide assumes that deployment personnel have:

- Knowledge of AX nodes from Dell Technologies.
- Experience of configuring BIOS and integrated Dell Remote Access Controller (iDRAC) settings.
- Advanced knowledge of deploying and configuring Windows Servers and Hyper-V infrastructure.
- Experience with deploying and configuring Storage Spaces Direct Solutions with Windows Server or Azure Stack HCI.
- Familiarity with customer site-to-site networking, including enabling and configuring the necessary static routes or inter-site bandwidth throttling (if needed) according to the stretched cluster requirement.

# 2

# Solution overview

This chapter presents the following topics:

**Topics:**



# Introduction

Dell EMC Solutions for Azure Stack HCI offers stretched cluster solutions with AX nodes from Dell Technologies. Built using industry-leading PowerEdge servers, AX nodes offer fully validated HCI nodes for a variety of use cases. A robust set of configurations and different models allows you to customize your infrastructure for application performance, capacity, or deployment location requirements.

## Stretched clusters and Storage Replica

An Azure Stack HCI stretched cluster solution is a disaster recovery solution that provides an automatic failover capability to restore production quickly, with little or no manual intervention. Storage Replica, a Windows Server technology, enables replication of volumes between servers across sites for disaster recovery. For more information, see Storage Replica overview.

A stretched cluster with Azure Stack HCI consists of servers residing at two different locations or sites, with each site having two or more servers, replicating volumes either in synchronous or asynchronous mode. For more information, see Stretched clusters overview.

A stretched cluster can be set up as either Active-Active or Active-Passive. In an Active-Active setup, both sites will actively run VMs or applications; therefore the replication is bidirectional. In an Active-Passive setup, one site is always dormant unless there is a failure or planned downtime.

Sites can be on the same campus or in different places. Stretched clusters using two sites provides disaster recovery if a site experiences an outage or failure.

The following figure shows an Active-Active setup:

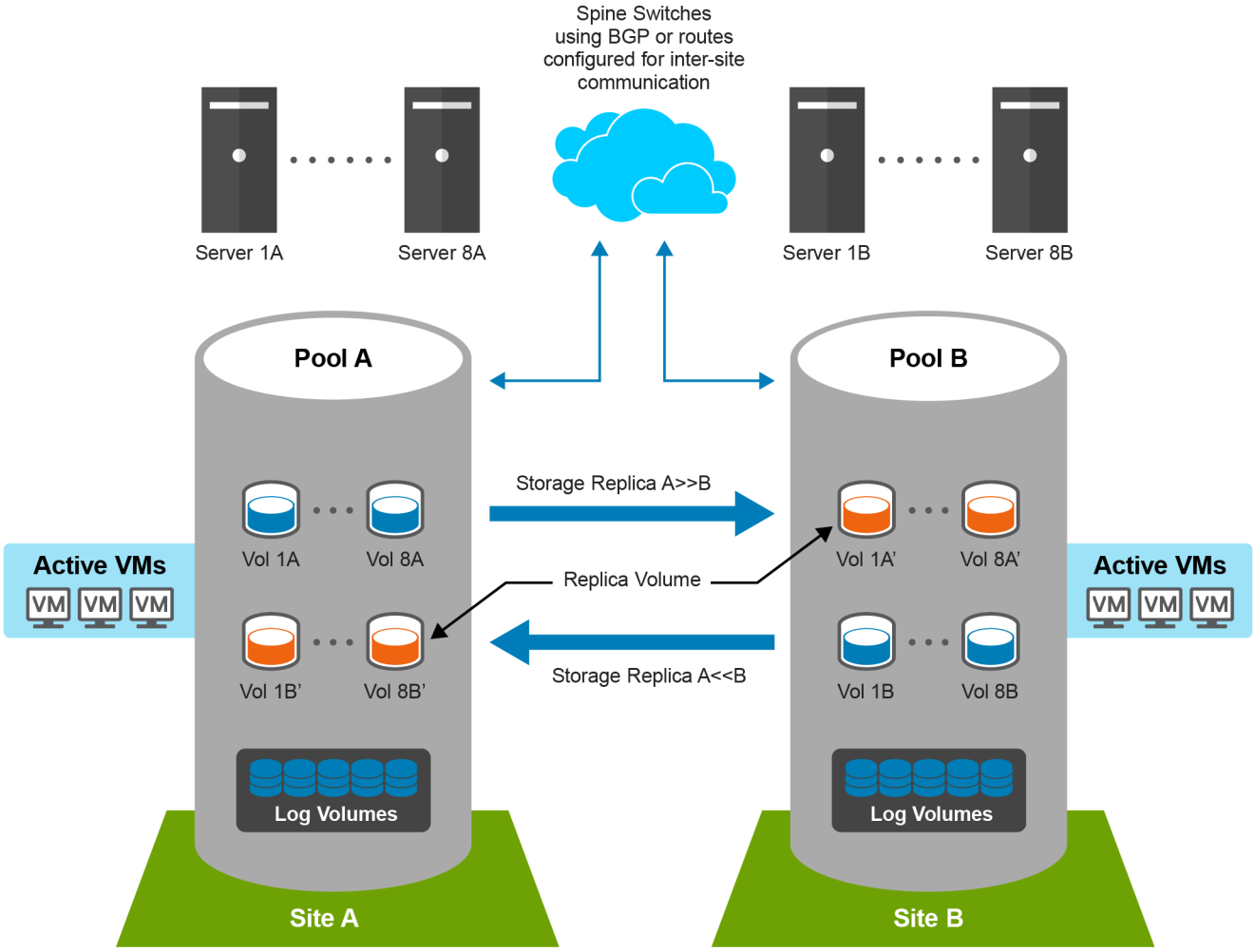



**Figure 1. An Active-Active setup**

Sites can be logical or physical. For logical sites, a stretched cluster can exist on single or multiple racks or in different rooms in the same data center. For physical sites, the stretched cluster can be in different data centers on the same campus or in different cities or regions. Stretched clusters using two physical sites provide disaster recovery and business continuity should a site suffer an outage.

# Solution integration and network architecture

Dell EMC Solutions for Azure Stack HCI stretched clusters offer distinct network topologies that are validated with the following stretched cluster configurations:

- Basic configuration
- High throughput configuration

**Basic configuration** refers to a network topology that requires minimal changes to a traditional single-site Azure Stack HCI configuration. This configuration uses a single network/fabric for management, VM, and replication traffic, keeping host networking simple. The customer network team must configure quality of service (QoS) on an external firewall or routers to throttle inter-site bandwidth and thereby ensure that Replica/VM traffic does not saturate the Management network.

**High throughput configuration** suits customer environments that are dense and involves higher write IOPs compared to a basic configuration. This configuration requires a dedicated channel (network interface cards (NICs) or fabric) for Replica traffic (using SMB-Multichannel). This network topology should be used only if inter-site bandwidth is higher than 10 Gbps. The network team must configure multiple static routes on the host to ensure that Replica traffic uses the dedicated channel that has been created for it. If the customer environment does not use Border Gateway Protocol (BGP) at the ToR layer, static

routes are needed on the L2/L3 to ensure that the Replica networks reach the intended destination. Subsequent sections of this guide provide more information about the expectations of customer networking teams.

A stretched cluster environment has two storage pools, one per site. In both topologies described in the preceding section, storage traffic requires Remote Direct Memory Access (RDMA) to transfer data between nodes within the same site. Because Storage and Replica traffic produces heavy throughput on an all-flash or NVMe configuration, we recommend that you put Storage traffic on separate redundant physical NICs.

This table shows the types of traffic, the protocol used, and the recommended bandwidth:

**Table 1. Types of traffic**

| Types of traffic | Protocol used | Recommended bandwidth |
|---|---|---|
| Management | TCP | 1/10/25 Gb |
| Replica | TCP | 1/10/25 Gb |
| Intra-site storage | RDMA | 10/25 Gb |
| Compute Network | TCP | 10/25 Gb |

Here are some points to consider about network configuration:

- Management traffic uses Transmission Control Protocol (TCP). Because management traffic uses minimal bandwidth, it can be combined with Storage Replica traffic or even use the LOM, OCP or rNDC ports.
- VM Compute traffic can be combined with management traffic.
- Inter-site Live Migration traffic will use the same network as Storage Replica.
- Storage Replica uses TCP as RDMA is not supported for replica traffic over L3 or WAN links. Depending on the bandwidth and latency between sites and the throughput requirements of the cluster, consider using separate redundant physical NICs for Storage Replica traffic.

# 3

# Solution deployment

This chapter presents the following topics:

**Topics:**



## Introduction

Stretched clusters with Dell EMC Solutions for Azure Stack HCI can be configured using PowerShell. This guide describes the prerequisites for this deployment.

**NOTE:** The instructions in this guide are applicable only to the Microsoft Windows Azure Stack HCI operating system.

Each task in this deployment guide requires running one or more PowerShell commands. On some occasions you might have to use Failover Cluster Manager or Windows Admin Center from a machine that supports Desktop Experience.

## Deployment prerequisites for stretched clusters

Dell Technologies assumes that the management services required for the operating system deployment and cluster configuration are present in the existing infrastructure. An internet connection is required to license and register the cluster with Azure. Because Microsoft Azure Stack HCI operating system is a Server Core operating system, you require a system that supports Desktop Experience to access Failover Cluster Manager and Windows Admin Center. For more information, see the Windows Admin Center FAQ.

**Table 2. Deployment prerequisites for stretched clusters**

| Component | Requirements |
|---|---|
| Active Directory Sites & Subnets | Configure two sites and their corresponding subnets in Active Directory so that the correct sites appear on Failover Cluster Manager on configuration of stretched clusters.<br>Configure Fault domains for each cluster if the IP subnets are the same across both sites. |
| Witness | Customers can choose to have a File Share witness either at a tertiary site or on Azure Cloud. |
| Windows features | Hyper-V<br>Failover-Clustering<br>Data-Center-Bridging<br>Storage-Replica with PowerShell Management Tools on all nodes<br>File server role |
| Network | The following requirements apply: |

**Table 2. Deployment prerequisites for stretched clusters (continued)**

| Component | Requirements |
|---|---|
| | • If two sites have host networks in different subnets, no additional configuration is needed for creating clusters. Otherwise, manual configuration of the cluster fault domain is required.<br>• RDMA Adapters for Storage/SMB traffic.<br>• RDMA is not supported for Replica traffic across WAN.<br>• At least a 1 Gb network between sites for Replication and inter-site Live Migration is required.<br>• The bandwidth between sites should be sufficient to meet the write I/Os on the primary site.<br>• An average latency of 5 ms or lesser for Synchronous Replication.<br>• There are no latency requirements or recommendations for Asynchronous replication.<br>• There is no recommendation from Microsoft regarding the maximum distance between sites that a stretched cluster can support. Longer distances normally translate into higher network latency. |
| Windows Admin Center Node | Windows features required:<br>RSAT-Clustering<br>RSAT-Storage-Replica |
| Number of nodes supported | Minimum: 4 (2 Nodes per site)<br>Maximum: 16 (8 Nodes Per Site) |
| Number of drives supported | Minimum of 4 drives per node. Both sites should have the same capacity and number of drives. Dell Technologies currently supports only an All-Flash configuration for stretched clusters. |
| Tuning of cluster heartbeats | (get-cluster).SameSubnetThreshold = 10<br>(get-cluster).CrossSubnetThreshold = 20 |
| SDN/VM Network | SDN on multi-site clusters is not supported at this time. |

For the maximum supported hardware configuration, see Review maximum supported hardware specifications.

# Customer network team requirements

Depending on the network configuration chosen, customers should ensure that the requisite end-to-end routing is enabled for inter-site communication. A minimum of one IP route or three IP routes based on Basic or High Throughput configuration is required for the environment.

Depending on the network configuration, the customer network team may also need to add static routes on the switches or on Layer-3 to ensure site-to-site connectivity.

# Design principles and best practices

## Stretched clusters and Storage Replica

A stretched cluster setup has two sites and two storage pools. Replicating data across WAN and writes on both sites results in lower performance compared to a standalone Storage Spaces Direct Cluster. Low latency inter-site links are necessary for

optimum performance of workloads. Low bandwidth and high latency between sites can result in very poor performance on the primary site in the case of both synchronous and asynchronous replication.

Synchronous replication involves data blocks being written to log files on both sites before being committed. In asynchronous replication, the remote node accepts the block of replicated data and acknowledges back to the source copy. Application performance is not affected unless the rate of change of data is faster than the bandwidth of the replica link between the sites for large periods of time. This point is critical and must be taken into consideration when you are designing the solution.

The size of the log volume has no bearing on the performance of the solution. A larger log collects and retains more write I/Os before they are wrapped out. This allows for an interruption in service between the two sites (such as a network outage or the destination site being offline) to go on for a longer period.

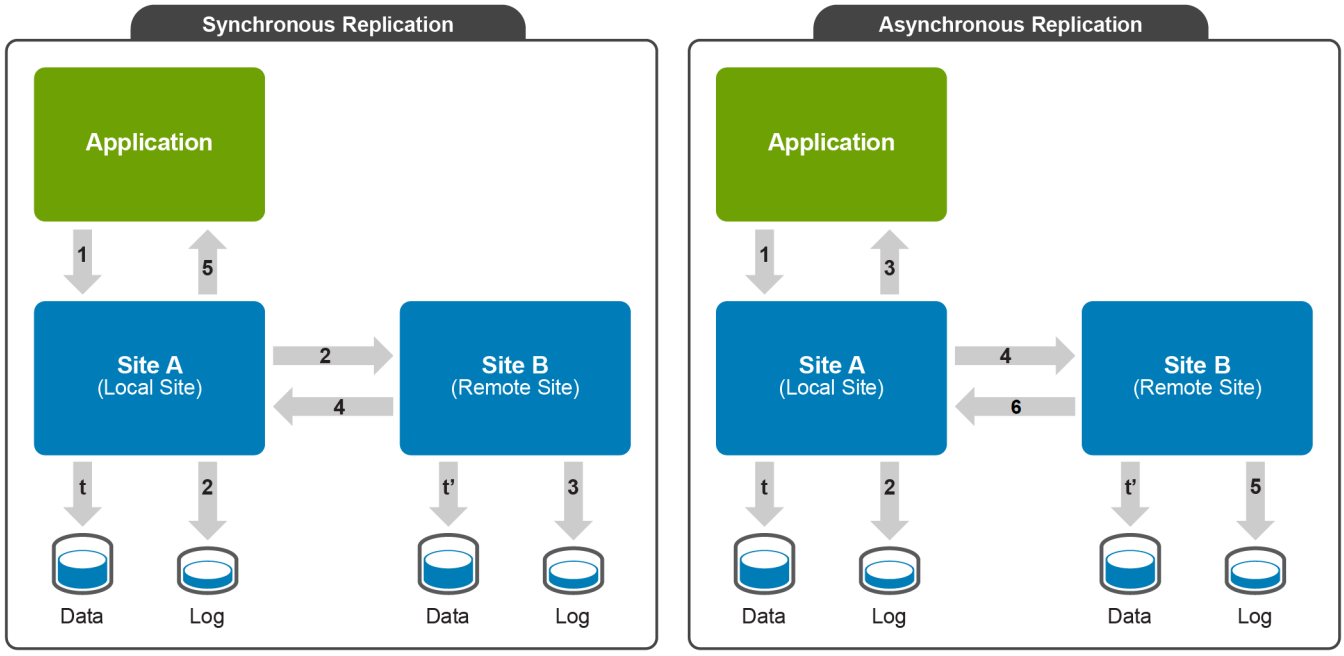
### Table 3. Disk writes

| Scenario | Writes in two-way mirrored volumes | Writes in three-way mirrored volumes |
|---|---|---|
| Standalone storage spaces | 2x | 3x |
| Replication to secondary site | 4x | 6x |

**NOTE:** WAN latency and additional writes to log volumes on both sites causes higher write latency. Along with writes to the log and data disks, the inter-site bandwidth and latency also play a role in limiting the IOPs in the environment. For this reason, we highly recommend using all-flash configurations for stretched clusters.

**NOTE:** In a Storage Spaces Direct environment both data and log volumes eventually reside on the same SSD pool because multiple storage pools per site are not supported.

The following figure illustrates the difference between synchronous and asynchronous replication:




### Figure 2. Synchronous and asynchronous replication

**Synchronous replication**: A block of data written by an application to a volume on Site A (1) is written first to the corresponding log volume on the same site (2), and is then replicated to Site B (2). At site B, the block of data is written to the Replica log volume (3) before a commit is sent back to the application using the same route (4 and 5). The block is subsequently pushed to the data volumes on both sites. For each block of data that the application writes, the commit is issued only after data is written to the secondary site. Thus there is no data loss at file system level in the event of a site failure. This results in a lower application write performance compared to a standalone deployment.

**Asynchronous replication**: A block of data written by an application to a volume on Site A (1) is written first to the corresponding log volume on the same site (2). A commit is immediately sent back to the application. At the same time, the block of data is replicated to Site B and written to the Replica log volume. In the case of a site failure, the cluster ensures that no data is lost beyond the configured Recovery Point Objective (RPO). Application performance is not affected unless the

rate of change of data is faster than the bandwidth of the replica link between the sites for large periods of time. This is critical and must be taken into consideration when designing the solution.

**NOTE:** Both replication scenarios affect application performance because each data block has to be written multiple times, assuming that all volumes are configured for replication.

**NOTE:** Stretched cluster with Storage Replica is not a substitute for a backup solution. Stretched cluster is a disaster recovery solution that keeps a business running in the event of a site failure. Customers should still rely on application and infrastructure backup solutions to recover lost data due to user error or application/data corruption.

# Validated network topology

## Basic configuration

This section describes the host network configuration and network cards that are required to configure a basic stretched cluster. The purpose of this topology is to keep the host and inter-site configuration simple with little or no change to a standard standalone cluster networking architecture.

Here we use two 25 GbE NICs for each host on both sites. One NIC is dedicated to intra-site storage traffic, similar to a standalone Storage Spaces Direct environment. The second NIC is used for management, compute, and Storage Replica traffic. To ensure management traffic is not bottlenecked due to high traffic on the Replica network, we request the customer network team to throttle traffic between the two sites using firewall or router QoS rules. It is recommended that the network is throttled to 50 percent of the capacity of the total number of network cards supporting the management NIC team.

The management network is the only interface between the two sites. Because only one network pipe is available between the hosts on Site A and Site B, you will see the following warning in the cluster validation. This is an expected behavior.

```
Node SiteANode1.Test.lab is reachable from Node SiteBNode1.Test.lab by only one pair of
network interfaces. It is possible that this network path is a single point of failure
for communication within the cluster. Please verify that this single path is highly
available, or consider adding additional networks to the cluster.
```

**Table 4. Sample IP address schema**

| | Site A | Site B | Type of traffic |
|---|---|---|---|
| Management/Replica Traffic | 192.168.100.0/24 | 192.168.200.0/24 | L2/L3 |
| Intra-site Storage (RDMA) - 1 | 192.168.101.0/24 | 192.168.201.0/24 | L2 |
| Intra-site Storage (RDMA) - 2 | 192.168.102.0/24 | 192.168.202.0/24 | L2 |
| VMNetwork/Compute Network | As per customer environment | As per customer environment | L2/L3 |

The following figure shows the network topology of a basic stretched cluster:

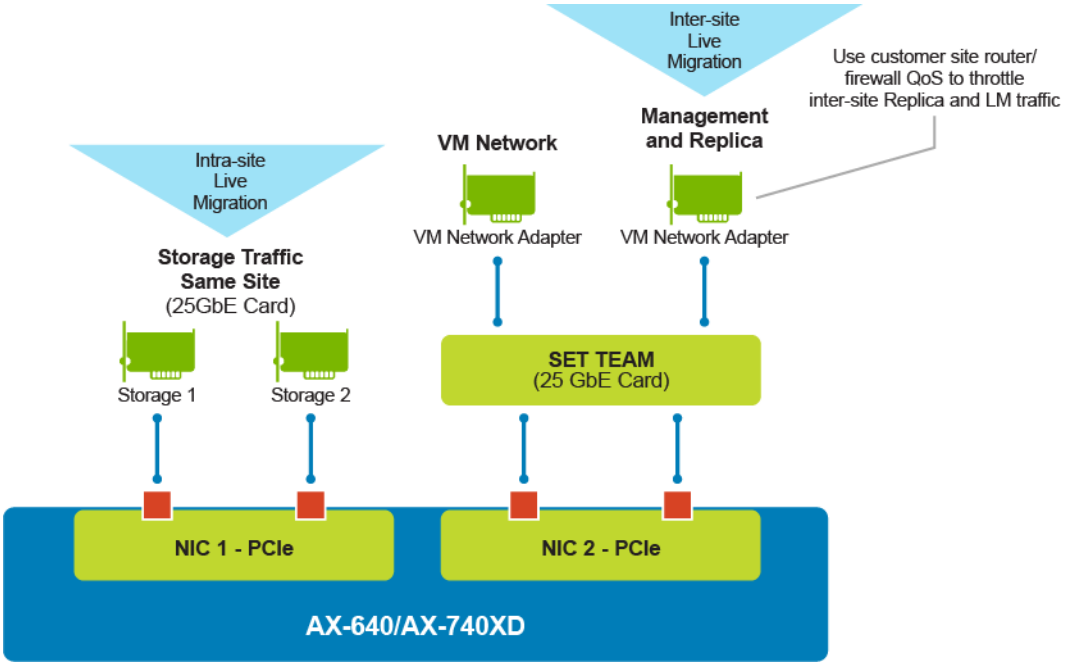



**Figure 3. Network topology for a stretched cluster (basic)**

## High throughput configuration

In this topology we use two 25 GbE and one 1/10 GbE/25 GbE NICs for each host to configure a high throughput stretched cluster. One NIC is dedicated for intra-site RDMA traffic, similar to a standalone Storage Spaces Direct environment. The second NIC is used for replica traffic. SMB Multichannel is used to distribute traffic evenly across both replica adapters and it increases network performance and availability. SMB Multichannel enables the use of multiple network connections simultaneously, and facilitates the aggregation of network bandwidth and network fault tolerance when multiple paths are available. For more information, see Manage SMB Multichannel.

The `Set-SRNetworkConstraint` cmdlet is used to ensure replica traffic flows only through the dedicated interfaces and not through the management interface. Run this cmdlet once for each volume.

## IP Address schema

The following table shows the IP Address schema:

**Table 5. IP Address schema**

| | Site A | Site B | Type of traffic |
|---|---|---|---|
| Management | 192.168.100.0/24 | 192.168.200.0/24 | L2/L3 |
| Intra-site Storage (RDMA) - 1 | 192.168.101.0/24 | 192.168.201.0/24 | L2 |
| Intra-site Storage (RDMA) - 2 | 192.168.102.0/24 | 192.168.202.0/24 | L2 |
| Replica - 1* | 192.168.111.0/24 | 192.168.211.0/24 | L2/L3 |
| Replica - 2* | 192.168.112.0/24 | 192.168.212.0/24 | L2/L3 |
| VMNetwork | As per customer environment | As per customer environment | L2/L3 |
| Cluster IP | 192.168.100.100 | 192.168.200.100 | L2 |

*Static routes are needed on all hosts on both sites to ensure the 192.168.111.0/24 network can reach 192.168.211.0/24 and the 192.168.112.0/24 network can reach 192.168.212.0/24. Static routes are needed in this network topology because we have three network pipes between Site A and Site B. Network traffic on Management uses the default gateway to traverse the network, while Replica network uses static routes on the hosts to reach the secondary site. If your ToR switches do not have BGP configured, static routes are needed on them also.

The following figure shows the network topology of an advanced stretched cluster:

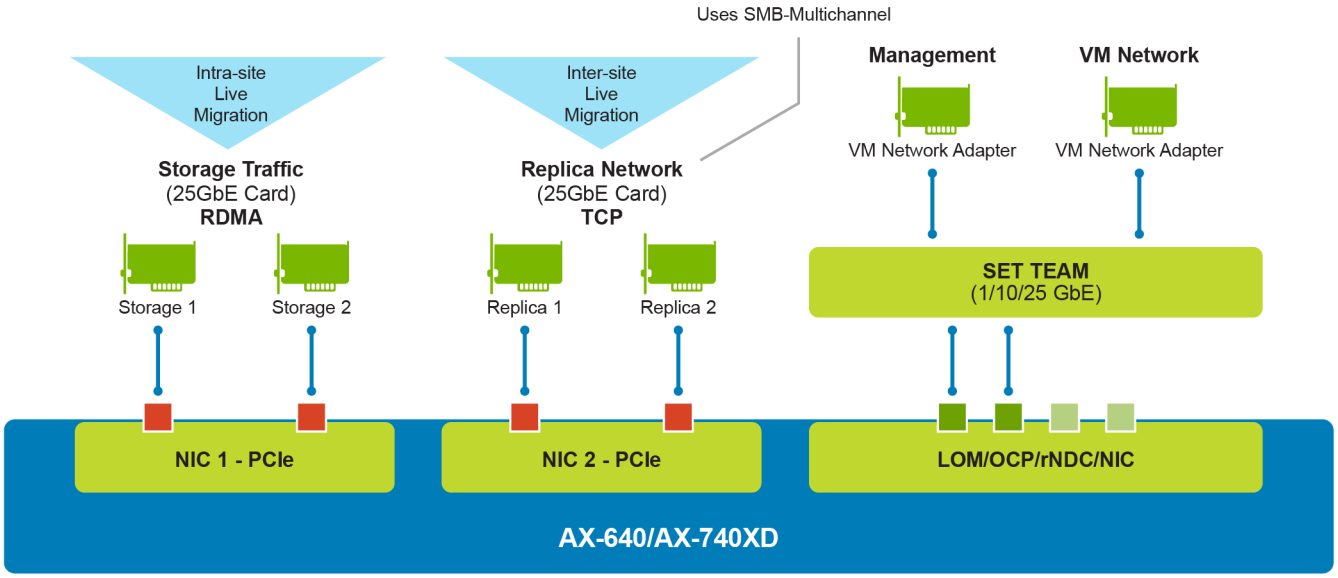



**Figure 4. Network topology for a stretched cluster (advanced)**

# 4

# Creating a stretched cluster

This chapter presents the following topics:

**Topics:**



## Introduction

This section outlines the steps that are needed for configuring a stretched cluster. Complete the network configuration on all nodes for the network topology applicable to you. A sample IP address schema is provided for both supported network topologies in the previous section of this guide. Consider these points before you begin:

- Ensure management IPs of all nodes are reachable from any host
- Ensure static routes are configured on all hosts for inter-site communication using the Replica network
- Ensure all nodes from Site A can reach corresponding Replica IPs on Site B using the Replica path

## Test-Cluster

Test-Cluster is a fundamental function that is needed to ensure that the cluster to be created meets Microsoft's recommendations regarding Failover Clustering. It also ensures that your hardware and settings are compatible. Run Test-Cluster with all nodes and include All Tests (namely, 'Storage Spaces Direct', 'Inventory', 'Network' and 'System Configuration').

Ensure there are test-cluster passes without warnings for the 'High Throughput Configuration', while 'Basic Configuration' will receive a warning as mentioned in the previous section of this guide.

## Cluster creation

This section looks at creating a cluster using PowerShell cmdlets.

### Manual cluster creation

Once Test-Cluster completes successfully, use the `New-Cluster` cmdlet to create a new stretched cluster. Because the nodes specified are part of different IP schemas, Enable-ClusterS2D understands that the cluster is part of a multi-site topology. It automatically creates two storage pools and corresponding ClusterPerformanceHistory volumes and their replica volumes.

After a cluster is created, you will see a warning similar to the one shown below. This is an expected behavior.

```
No matching network interface found for resource 'Cluster IP Address 172.18.160.160' IP
address '192.168.200.100' (return code was '5035'). If your cluster nodes span different
subnets, this may be normal.
```

Configure cluster witness and Enable Storage Spaces Direct on the cluster.

**NOTE:** Cluster witness can be either on a tertiary site or on Azure Cloud. Ensure that the "Storage Replica" feature is installed on all nodes in the cluster.

If Sites and Services with IP Subnets are configured on Active Directory, Failover Cluster Manager correctly shows a node to Site mapping, under **Cluster Name >> Nodes**.

The following is a sample image of IP subnets defined in an Active Directory:

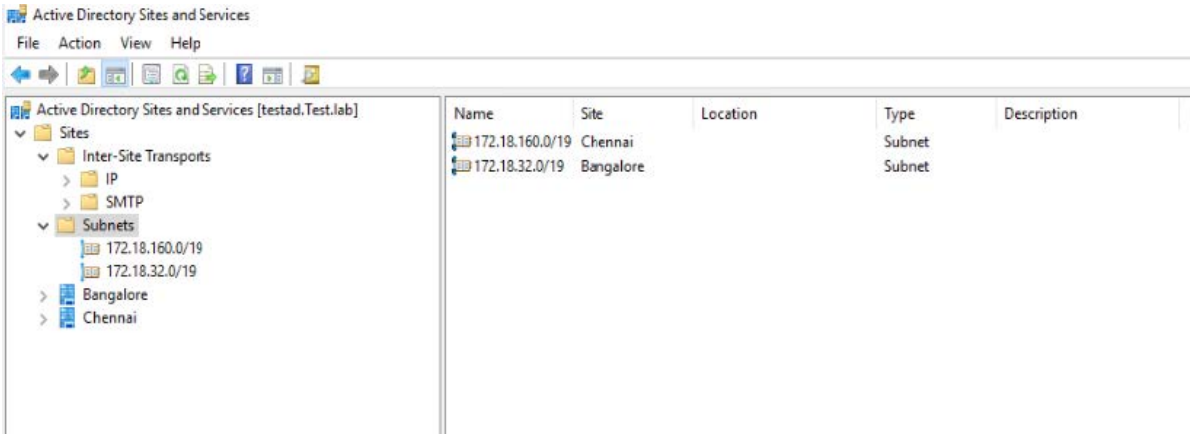



**Figure 5. IP subnets in an Active Directory**

If both sites are in the same IP network, use the `New-ClusterFaultDomain` cmdlet to define the two site names. Site names defined using `New-ClusterFaultDomain` override the names given in Active Directory.

The following is a sample image of how sites appears in Failover Cluster Manager:

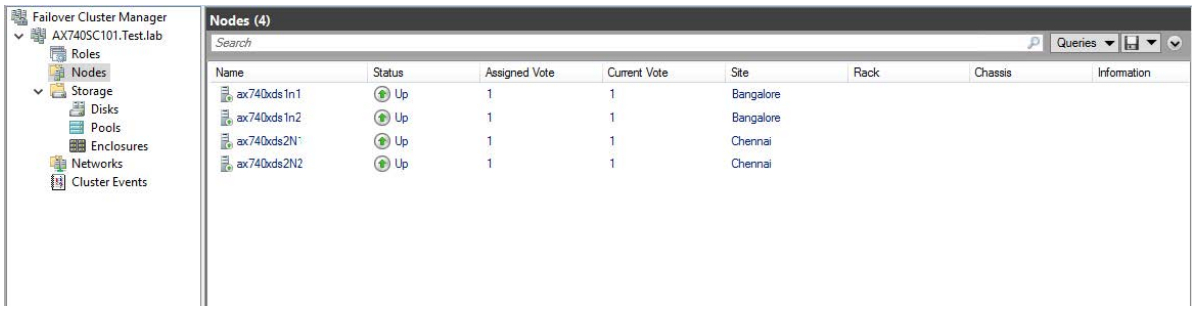



**Figure 6. Sites in Failover Cluster Manager**

Once a cluster is created, use Failover Cluster Manager to rename the cluster networks.

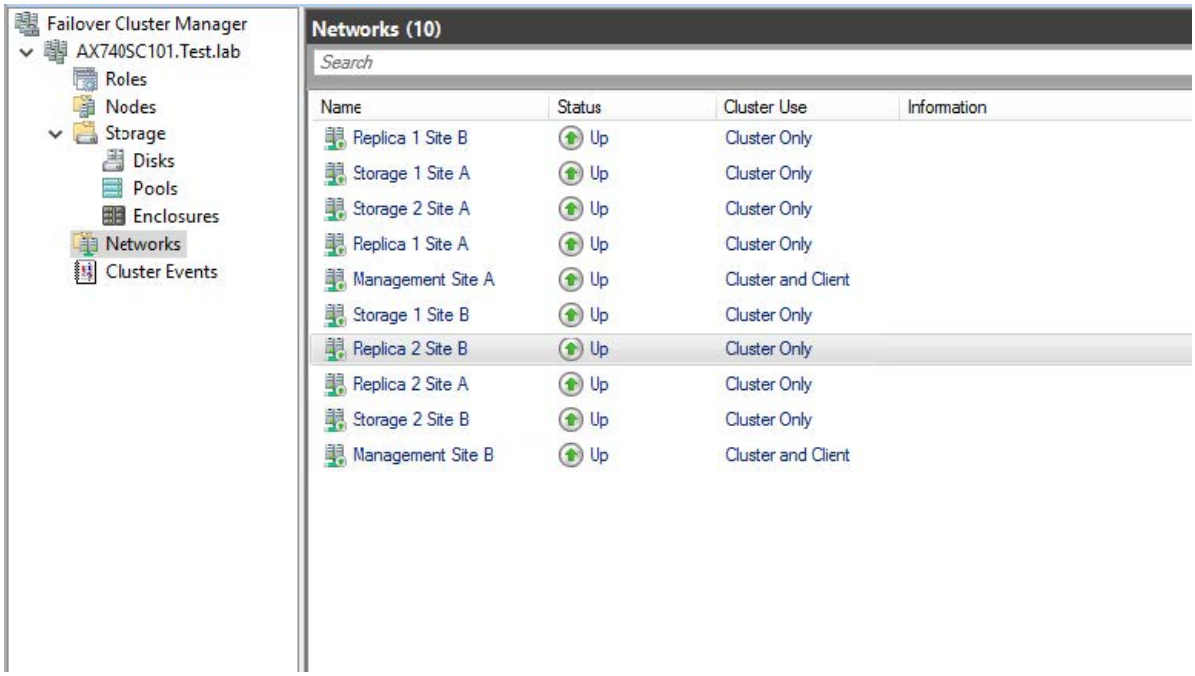



**Figure 7. Cluster networks**

# Volumes

Replication-enabled volumes can be created using a combination of PowerShell and Failover Cluster Manager or by using Windows Admin Center.

**NOTE:** Install Storage Replica Module for Windows PowerShell (RSAT-Storage-Replica) on the management node with Desktop Experience that is used for installing Windows Admin Center and Failover Cluster Manager to access the cluster.

For each replica-enabled volume, you need a corresponding log volume on both sites (with a minimum of 8 GB in size) and an equivalent replica volume on the secondary site. The log volume is used to serialize writes for replication.

The following table shows the volumes that are needed to create a 1 TB replica volume:

**Table 6. Volumes in a 1 TB replica volume**

| Site A | Size | Site B | Size |
|---|---|---|---|
| VolumeA | 1 TB | VolumeA-Replica | 1 TB |
| VolumeA-Log | 40 GB | VolumeA-Replica-Log | 40 GB |
| VolumeB | 500 GB | VolumeB-Replica | 500 GB |
| VolumeB-Log | 40 GB | VolumeB-Replica-Log | 40 GB |

See Appendix A for the correct PowerShell Cmdlets and Failover Cluster Manager steps to create the volumes shown in this table. It is recommended that you create two-way mirrors for all volumes to improve write performance and capacity efficiency.

**NOTE:** For Asynchronous Replication, the RPO can be set as low as 30 seconds.

For a planned site failure, when the volume replication direction is reversed, the disk reservations on the secondary site for Replica Volume and Replica-log volumes are removed and moved to the primary site. Source Data and Source Log volumes are given the disk reservations and become active on the secondary site. After 10 minutes, the virtual machines residing on the primary site associated with the migrated volume automatically Live Migrate to the secondary site.

# Storage efficiency

Due to high I/Os on the underlying disks, stretched clusters require an underlying infrastructure capable of delivering high I/Os with low latency. Dell Technologies recommends all-flash configurations for stretched cluster deployments.

All-flash configurations do not have a cache tier. The following table shows the difference in storage efficiency for a two-way and three-way mirror created on a single site and stretched cluster environment:

**Table 7. Storage efficiency differences**

| | Two-Way Mirror | Three-Way Mirror |
|---|---|---|
| Single site | 50% | 33% |
| Stretched cluster | 25% | 16.5% |

# Test-SRTopology

This cmdlet validates a potential replication partnership. The cmdlet:

- Measures bandwidth and round trip latency
- Estimates initial sync time
- Verifies that source and destination volumes exist
- Verifies that there is sufficient physical memory to run replication

**NOTE:** The file server feature has to be enabled on the nodes to run this cmdlet.

Keep the generated report for future reference.

For more information about this cmdlet, see Test-SRTopology.

# 5

# Virtual Machines

This chapter presents the following topics:

**Topics:**



## Introduction

Virtual Machines in a stretched cluster environment can be managed using:

- PowerShell
- Failover Cluster Manager
- Windows Admin Center

For more information, see Manage VMs on Azure Stack HCI using Windows Admin Center.

In a stretched cluster environment, volumes hosting the virtual machines may or may not be replicated, depending on business requirements. A VM that is hosted on a replicated volume runs on the site where the volume is mounted. If the volume moves from the primary site to the secondary site due to a planned downtime or site failure, the VMs follow the volumes to the secondary site after 10 minutes while the cluster service balances the VMs across the nodes on the secondary site after 30 minutes. To enable faster movement of VMs after the Virtual Disk ownership is transferred to the secondary site, you can initiate Live Migration manually.

## VM and storage affinity rules

You can use PowerShell to create affinity and anti-affinity rules for your VMs in a cluster. An affinity rule is one that establishes a relationship between two or more resource groups or roles, such as VMs, to keep them together in an Azure Stack HCI cluster. An anti-affinity rule does the opposite, keeping specified resource groups apart from each other.

You can use storage affinity rules to keep a VM and its associated Virtual Hard Disk v2 (VHDX) on a Cluster Shared Volume (CSV) on the same cluster node. This ensures CSV redirection does not occur and keeps application performance at optimal levels. For more information, see Storage affinity rules.

## Preferred sites

Configure preferred sites in a stretched cluster to define a location in which you want to run all your resources. This ensures that VMs and volumes come up on the preferred site after a cold start or after network connectivity issues. A dynamic quorum ensures that preferred sites survive after events such as asymmetric network connectivity failures.

In the event of a quorum split, if witnesses cannot be contacted, the preferred site is selected and the passive site drops out of the cluster membership.

VMs in a stretched cluster are placed based on the following site priority:

- Storage Affinity Site
- Group Preferred Site
- Cluster Preferred Site

**Host VMs and associated Virtual Disks**

VM placement can be difficult in a Hyper-V clustered environment. Always ensure that VMs that have I/O-intensive workloads are hosted on the node that owns the VM's Virtual Disk.

6

# Failure/Recovery from failure of Site/Node

This chapter presents the following topics:

**Topics:**



## Planned failover

Windows Admin Center has a Switch Direction feature that allows you to migrate workloads from one site to the other. This must be initiated on each volume. VMs hosted on the volumes follow the volumes to the migrated site after 10 minutes. This feature is helpful in scenarios such as:

- There is a planned downtime
- A potential weather event could take the site down

To use the Switch Direction feature, go to Windows Admin Center and select **Storage Replica** on the left pane. Then select the SR Partnership for which you would like to change the Replication Direction. Select **More** and click on **Switch Direction**.

In the event of a site failure, if a volume is replicating synchronously then the data and the log volume automatically come online on the surviving site, along with VMs associated with this volume because the RPO is 0. For asynchronous replication the data and the log volume do not come online automatically because the RPO is not equal to 0.

When the failed site comes back online, the Replica and Replica-Log volume are moved to the primary site with persistent disk reservations, and replication begins again. For a synchronous replicated volume, the replication direction cannot be changed until replication is 100 percent complete.

## Operation steps

The following sections describe the steps to take in the event of different failure types.

### Node failure

Handling a node failure on either site in a stretched cluster environment is no different than managing one in a traditional or standalone Azure Stack HCI cluster. A complete node failure would result in operating system or HBA corruption or complete hardware failure on the node. In either case, restoring system functionality is the priority.

The high level steps to do this are:

1. Replace the hardware as needed.
2. Re-install the operating system on the operating system drives (if needed).
3. Join the system to the domain.
4. Ensure you assign the new node IPs specific to the site where the node is hosted.
5. Add the node to the existing stretched cluster.
6. Based on the IP subnets used or the Cluster Fault Domains added, the cluster adds the drives to the correct pool.
7. Wait for the storage jobs to complete.
8. During this process the workloads on the affected site would still be running and there should be no interruption of replication.

## Site failure

A site failure in a stretched cluster topology requires rebuilding all of the nodes of the affected site. If the failure happens at the primary site, the following scenarios occur:

- All volumes hosted on the affected site and associated VMs become inaccessible.
- After a brief period, the volumes move to the secondary site.
- The VMs restart on the secondary site.
- Depending on whether synchronous or asynchronous replication is being used, you either have zero data loss or data loss within the limits of the defined RPO:
  - For the replica volumes configured with synchronous replication, the VMs are crash consistent. Application recovery depends on the available backup/recovery of the application.
  - For the replica volumes configured with asynchronous replication, the VMs are not crash consistent. The default RPO is 30 seconds. It can be configured using PowerShell or Windows Admin Center. Application recovery still depends on the available backup/recovery of the application.

## Site recovery

Follow these steps to recover the nodes on the failed site:

1. Remove the failed nodes from the cluster and remove the computer names from the Active Directory.
2. Remove SRPartnership and SRGroups using PowerShell cmdlets. Replication can also be disabled from the Failover Cluster Manager.
3. Bring up all the nodes on the affected site. The node names and IPs used should be the same as those used before the crash.
4. Join the nodes to the domain.
5. Add all the nodes to the existing stretched cluster at the same time.
6. All drives in the new site will be added to a new pool.
7. Re-create and enable replication for replica volumes and associated log volumes using Failover Cluster Manager or PowerShell cmdlets.

# A

# Appendices

These appendices present the following topics:

**Topics:**



# Appendix A: Sample PowerShell cmdlets for end-to-end deployment

## Install required Windows features

```
Install-WindowsFeature -Name Fs-Fileserver,Storage-Replica ,Hyper-V, Failover-
Clustering, Data-Center-Bridging -IncludeAllSubFeature -IncludeManagementTools -
Verbose
```

## Create VM switches and configure host networking

```
#Create VMSwitch for Management Network
#Mention the correct network adapters based on your environment. This
configuration has to be done on all #nodes in the stretch cluster
New-VMSwitch -Name S2DSwitch -AllowManagementOS 0 -NetAdapterName "NIC1","NIC2"
-MinimumBandwidthMode Weight -Verbose

#Host Management Adapter
Add-VMNetworkAdapter -ManagementOS -Name "Management" -SwitchName S2DSwitch -
Passthru | Set-VMNetworkAdapterVlan -Access -VlanId 202 -Verbose
New-NetIPAddress -InterfaceAlias "vEthernet (Management)" -IPAddress
192.168.100.11 -DefaultGateway 192.168.100.1 -PrefixLength 24 -AddressFamily
IPv4 -Verbose

#DNS server address
Set-DnsClientServerAddress -InterfaceAlias "vEthernet (Management)" -
ServerAddresses 192.168.100.100,192.168.100.101

#Storage 1 Adapter
New-NetIPAddress -InterfaceAlias "SLOT 3 Port 1" -IPAddress 192.168.101.11 -
PrefixLength 24 -AddressFamily IPv4 -Verbose

#Storage 2 Adapter
New-NetIPAddress -InterfaceAlias "SLOT 3 Port 2" -IPAddress 192.168.102.11 -
PrefixLength 24 -AddressFamily IPv4 -Verbose

#Enable RDMA on Storage Ports & set NetworkDirect Technology to iWarp
Enable-NetAdapterRDMA -Name "SLOT 3 Port 1", "SLOT 3 Port 2"

Set-NetAdapterAdvancedProperty -Name 'SLOT 3 PORT 1' -DisplayName 'NetworkDirect
Technology' -DisplayValue 'iWarp'
Set-NetAdapterAdvancedProperty -Name 'SLOT 3 PORT 2' -DisplayName 'NetworkDirect
Technology' -DisplayValue 'iWarp'

#Set VM Migration to SMB
Set-VMHost -VirtualMachineMigrationPerformanceOption SMB
```

```
#Configure Replica Network as applicable
#Replica 1
New-NetIPAddress -InterfaceAlias "SLOT 2 Port 1" -IPAddress 192.168.111.11 -
PrefixLength 24 -AddressFamily IPv4 -Verbose

#Replica 2
New-NetIPAddress -InterfaceAlias "SLOT 2 Port 1" -IPAddress 192.168.112.11 -
PrefixLength 24 -AddressFamily IPv4 -Verbose
```

## Test-Cluster

```
Test-Cluster -Node SiteANode1,SiteANode2,SiteBNode1,SiteBNode2 -Include 'Storage
Spaces Direct', Inventory', 'Network', 'System Configuration'
```

## New-Cluster

```
New-Cluster -Name R740StretchCluster -Node
SiteANode1,SiteANode2,SiteBNode1,SiteBNode2 -NoStorage -StaticAddress
192.168.100.100,192.168.200.100 -IgnoreNetwork
192.168.101.0/24,192.168.201.0/24, 192.168.102.0/24,192.168.202.0/24
```

## Witness

Configure a highly available file share witness at a tertiary site or on Azure cloud.

```
Set-ClusterQuorum -FileShareWitness \\tertiaryShare\witness
```

## Active Directory sites

Ensure that you have created two sites on your Active Directory based on IPs subnets. This will help with assigning the correct names of sites on the cluster.

If you do not have sites configured on Active Directory, create cluster fault domains as required. The following cmdlet overrides the site names specified on the Active Directory:

```
#This is needed if you do not have sites configured on AD
#Create Sites
New-ClusterFaultDomain -Name 'Bangalore' -Type Site
New-ClusterFaultDomain -Name 'Chennai' -Type Site

#Site membership for nodes
Set-ClusterFaultDomain -Name SiteANode1 -Parent 'Bangalore'
Set-ClusterFaultDomain -Name SiteANode2 -Parent 'Bangalore'
Set-ClusterFaultDomain -Name SiteBNode1 -Parent 'Chennai'
Set-ClusterFaultDomain -Name SiteBNode2 -Parent 'Chennai'
```

## Preferred Sites

The Preferred Site will be your primary datacenter site.

**NOTE:** Do not configure a Cluster fault domain as a Preferred Site if you plan to run active VMs on both sites, otherwise choosing to Live Migrate VMs using the "Best Possible Node" option results in all VMs moving to the Preferred Site.

```
#Preferred Site
(get-cluster).PreferredSite = 'Bangalore'
```

Preferred Sites can also be configured at cluster role and group level.

```
(Get-ClusterGroup -Name SQLServer1).PreferredSite = 'Bangalore'
```

If there is an Active-Active stretched cluster where Preferred Sites are not configured, it is highly recommended that you configure Preferred Sites for each volume. This will ensure that the volumes stay at the same site if there is a single node failure on either site.

```
(Get-ClusterSharedVolume "Cluster Virtual Disk (ax740xds2N2)"  | Get-
ClusterGroup).PreferredSite = "Chennai"

Get-ClusterSharedVolume "*ax740xds2N2)"  | Get-ClusterGroup |fl *
```

## Enable Storage Spaces

```
Enable-ClusterS2D -confirm:$false
```

This step will ensure that Storage Spaces Direct is enabled on the stretched cluster. Two storage pools are created, one for each site.

```
PS C:\Users\Administrator.TEST> Get-StoragePool
FriendlyName   Operational  HealthStatus IsPrimordial IsReadOnly  Size      AllocatedSize
               Status
------------ -------------- ------------ ------------ ---------- ---------  -------------
Primordial     OK           Healthy      True         False      15.94 TB   15.71 TB
Primordial     OK           Healthy      True         False      15.94 TB   15.71 TB
Pool for Site  OK           Healthy      False        False      15.7 TB    4.28 TB
Chennai
Pool for Site  OK           Healthy      False        False      15.7 TB    4.28 TB
Bangalore
```

## New Volume

You can create a new replicated volume by using Windows Admin Center or a mixture of PowerShell and Failover Cluster Manager.

**Using PowerShell and Failover Cluster Manager**

For each volume that you want to replicate across two sites, you will have to create its associated Replica volume and Log volume on both sites.

**NOTE:** Customers can choose to enable replication on the volumes based on their business needs.

```
#Primary Volume
New-Volume -StoragePoolFriendlyName "Pool for Site Bangalore" -FriendlyName
VolumeA -FileSystem CSVFS_ReFS -Size 1TB

#Log Volume
New-Volume -StoragePoolFriendlyName "Pool for Site Bangalore" -FriendlyName
'VolumeA-Log' -FileSystem ReFS -Size 50GB

#Replica Volume
New-Volume -StoragePoolFriendlyName "Pool for Site Chennai" -FriendlyName
'VolumeA-Replica' -FileSystem ReFS -Size 1TB

#Replica Log Volume
New-Volume -StoragePoolFriendlyName "Pool for Site Chennai" -FriendlyName
'VolumeA-Replica-Log' -FileSystem ReFS -Size 50GB
```

**NOTE:** Ensure that Replica volumes and all log volumes are ReFS (not CSVFS) before enabling replication using PowerShell or Failover Cluster Manager.

**Using Failover Cluster Manager to enable Volume Replica**

To enable replication on volumes, go to **Storage >> Disks** and right-click on the primary volume on which you want to enable replication. Then follow these steps:

- Select **Replication** and click **Enable**
- Select the log volume for the primary site
- Select the Replica volume and associated log volume for the secondary site
- Overwrite the destination volume unless you have a seeded disk
- Select the mode of replication
- Complete the wizard

This enables replication on the volume after the initial block copy. The initial block copy process can take a few minutes to a few hours, depending on the size of the volume.

## Test-SRTopology

This cmdlet validates a potential replication partnership between source and destination computers. Follow these steps:

- Create a local CSVFS volume, for example 1 TB.
- Create a local log volume (ReFS), for example 50 GB.
- Create a Replica volume (ReFS), for example 1 TB (Ensure that the local and replica volumes are the same size).
- Create a Replica-log volume (ReFS), for example 50 GB.

```
###Step 1###
New-Volume -StoragePoolFriendlyName "Pool for Site Bangalore" -FriendlyName
'VolumeA' -FileSystem CSVFS_ReFS -Size 1TB
#Log Volume
New-Volume -StoragePoolFriendlyName "Pool for Site Bangalore" -FriendlyName
'VolumeA-Log' -FileSystem ReFS -Size 50GB
## Move 'Available Storage to Site B' ##
Get-ClusterGroup -Name 'Available Storage' | Get-ClusterResource | Stop-
ClusterResource
Move-ClusterGroup -Name 'Available Storage' -Node ax740xds2n1

###Step 2###
#Replica Volume
New-Volume -StoragePoolFriendlyName "Pool for Site Chennai" -FriendlyName
'VolumeA-Replica' -FileSystem ReFS -Size 1TB
#Replica Log Volume
New-Volume -StoragePoolFriendlyName "Pool for Site Chennai" -FriendlyName
'VolumeA-Replica-Log' -FileSystem ReFS -Size 50GB

###Step 3
# Create Replication Group for secondary volumes
$PathVolARep =Get-Volume -FriendlyName VolumeA-Replica | Select -
ExpandProperty Path
$PathVolARepLog=Get-Volume -FriendlyName VolumeA-Replica-Log | Select -
ExpandProperty Path

New-SRGroup -ComputerName ax740xds2n1 -Name Group108 -VolumeName $PathVolARep -
LogVolumeName $PathVolARepLog -LogSizeInBytes 2GB

Get-ClusterGroup -Name 'Available Storage' | Get-ClusterResource | Stop-
ClusterResource
Move-ClusterGroup -Name 'Available Storage' -Node ax740xds1n2

# Assign drive letters
Get-Volume -FriendlyName VolumeB-Log | Get-Partition | Set-Partition -
NewDriveLetter H
Get-Volume -FriendlyName VolumeB-Replica | Get-Partition | Set-Partition -
NewDriveLetter I
Get-Volume -FriendlyName VolumeB-Replica-Log | Get-Partition | Set-Partition -
NewDriveLetter J

Test-SRTopology -SourceComputerName ax740xds1n2 -SourceVolumeName
C:\ClusterStorage\VolumeB -SourceLogVolumeName H -DestinationComputerName
ax740xds2n1 -DestinationVolumeName I -DestinationLogVolumeName J -
DurationInMinutes 30 -ResultPath .\TopologyResults
```

The preceding cmdlet completes in 30 minutes and displays the results in HTML format.

**NOTE:** Run `Test-SRTopology` only for a single volume.

**NOTE:** If you choose asynchronous replication, ensure that you choose a log volume size of at least 30 GB.

**NOTE:** Use either PowerShell or Windows Admin Center to create volumes, do not mix the tools. There is a minor difference in volume sizes created using PowerShell and Windows Admin Center that results in a failure if you try to enable replication.

## Set-SRNetworkConstraint

In a network topology which has multiple routes to the secondary site, it is imperative to provide a correct path for the Replica network. `Set-SRNetworkConstraint` is a cmdlet that is useful for specifying an array of network interfaces to be used for replica traffic. This cmdlet has to be run once for each volume.

```
Set-SRNetworkConstraint -SourceRGName "Replication 2" -SourceNWInterface "SR -
Site B" -DestinationRGName "Replication 1" -DestinationNWInterface "SR - Site A"
-SourceComputerName SiteANode1 -DestinationComputerName SiteBNode1 -Verbose
```

`Set-SRNetworkConstraint` ensures that the Management network does not become bottlenecked because of Replica traffic.

# Appendix B: Supported hardware

See the Dell support matrix for hardware and configurations validated with stretched clusters:

- Support Matrix for Dell EMC Solutions for Microsoft Azure Stack HCI