Dell EMC VxRail vSAN Stretched Cluster Planning Guide

Abstract

This planning guide provides best practices and requirements for using stretched clusters with VxRail appliances.

January 2020

Contents

1 Executive summary

A stretched cluster is a deployment model in which two or more virtualization host servers are part of the same logical cluster but are located in separate geographical locations. The vSAN stretched cluster feature enables synchronous replication of data between sites. This feature allows for an entire site failure to be tolerated. It extends the concept of fault domains to data center awareness domains.

vCenter Server is the centralized platform for managing a VMware environment. It is the primary point of management for both server virtualization and vSAN. It is also the enabling technology for advanced capabilities such as vMotion, Distributed Resource Scheduler (DRS), and HA. vCenter scales to enterprise levels where a single vCenter can support up to 1000 hosts (VxRail nodes) and 10,000 virtual machines. vCenter supports a logical hierarchy of data centers, clusters, and hosts, which allow resources to be separated by use cases or lines of business. It also allows resources to be moved as needed dynamically. These operations all done from a single interface.

1.1 Intended Use and Audience

This guide is intended for customers, Dell EMC and business partners, and implementation professionals. It is designed to help you understand the requirements for stretched cluster support with the Dell EMC VxRail Appliance. Services from Dell EMC or an authorized VxRail Services Partner are required for implementation of stretched clusters.

This document is not intended to replace the implementation guide or to bypass the service implementation required for stretched clusters. An attempt to set up stretch clusters on your own will invalidate support.

Upgrading from VxRail 4.7.300 to a later version no longer requires a Professional services (PS) engagement. You can choose to automatically update the witness. You can also choose which data site to upgrade first.

To upgrade from a version 4.7.300 or earlier, contact your Technical Support representative.

2 Overview

This planning guide provides best practices and requirements for using stretched cluster with a VxRail Appliance. This guide assumes that the reader is familiar with the *vSAN Stretched Cluster Guide*. This guide is for use with a VxRail Appliance only.

The vSAN stretched cluster feature creates a stretched cluster between two geographically separate sites, synchronously replication data between sites. This feature allows for an entire site failure to be tolerated. It extends the concept of fault domains to data center awareness domains.

VxRail 4.5.070 introduced vSAN 6.6 which includes local site protection and site affinity for stretched clusters allowing unbalanced configurations. The following is a list of the terms that are used for vSAN stretched clusters:

- *Preferred or Primary site* – one of the two data sites that is configured as a vSAN fault domain.
- *Secondary site* – one of the two data sites that is configured as a vSAN fault domain.
- *Witness host* – a dedicated ESXi host or vSAN witness appliance that is host to the witness component which coordinates data placement between the preferred and secondary site. It also helps with the failover process. This is the third fault domain.

The vSAN storage policies that impact the VxRail Cluster configuration are:

- *Primary Failures to Tolerate (PFTT) 1/Failures to Tolerate (FTT)* – for stretched clusters this rule has two possible values: 0 ensures protection on a single site; 1 enables protection across sites.
- *Secondary Failures to Tolerate (SFTT)* – This term is used in vSAN 6.6/VxRail 4.5.070 or later. The rule that defines the number of host and device failures that a virtual machine object can tolerate in the local site. Possible values: 0,1,2,3.
- *Failure Tolerance Method-* either RAID-1 (mirroring) used when performance is important or starting with vSAN 6.6/VxRail 4.5.070, RAID-5/6 (erase coding) used when capacity is important. For stretched clusters, this policy only applies to the SFTT setting. This failure tolerance method is the local file protection mode.
- *Affinity (only applicable starting with vSAN 6.6/VxRail 4.5.070)* - this policy is applicable when PFTT is set to 0. It is set to preferred or secondary to determine which site stores the vSAN object.

2.1 vSphere and vSAN

For vSAN stretched cluster functionality on VxRail, vSphere Distributed Resource Scheduler (DRS) is required. DRS provides initial placement assistance, and automatically migrates virtual machines to the corrected site in accordance with the Host/VM affinity rules. It can also help locate virtual machines to their correct site when a site recovers after a failure.

2.2 Fault domains

Fault domains (FD) provide the core functionality of vSAN stretched cluster. The maximum number of fault domains in a vSAN stretched cluster is 3. The first Fault Domain can be referred as Preferred data site, the second Fault Domain can be referred as Secondary data site, and the third Fault Domain is the witness host site. It is important to keep utilization per data site below 50% to ensure proper availability, if either the Preferred or Secondary site goes offline.

2.3 VxRail cluster nodes

vSAN stretched clusters are deployed across two sites in an Active/Active configuration. An identical number of ESXi hosts was required in versions earlier than vSAN 6.6/VxRail 4.5.070 to ensure a balanced distribution of resources. Starting with vSAN 6.6/VxRail 4.5.070, unbalanced configurations are supported; however, we recommend having an identical number of ESXi hosts across the two sites. VM and Host Affinity rules must be set for an unbalanced configuration.

Each data site is configured as a Fault Domain. An externally available third site houses a Witness appliance, which makes up the third Fault Domain.

2.3.1 VxRail cluster deployment options

You must plan the VxRail stretched cluster deployment before installation. Depending on the number of nodes in the VxRail cluster, you can:

- Deploy up to 16 nodes, 8 per site, on initial deployment or
- Initially deploy the minimum number of nodes per site and then scale out additional nodes either at installation or during the VxRail stretched cluster life cycle.

2.4 Witness host

Each vSAN stretched-cluster configuration requires a Witness host. The Witness must reside on a third site that has independent paths to each data site. While the Witness host must be part of the same vCenter as the hosts in the data sites, it must not be on the same cluster as the data site hosts. The Witness ESXi OVA is deployed using a virtual standard switch (vSS).

A vSAN Witness Appliance, or a physical host, can be used for the Witness function. The vSAN Witness Appliance includes licensing, while a physical host must be licensed accordingly.

NOTE: The Witness host OVA file comes with a license, therefore, it does not consume a vSphere license. However, a physical host requires a vSphere license.

2.5 VxRail cluster requirements

This section describes the requirements necessary to implement vSAN stretched clusters in a VxRail Cluster.

- The VxRail Cluster must be deployed across two physical sites in an Active/Active configuration.
- The VxRail Cluster must be running VxRail version 3.5 or later.
- For VxRail 3.5, 4.0 and 4.5.0, each data site must have an identical number of nodes.
- Starting with vSAN 6.6/VxRail 4.5.070, we recommend each data site has an identical number of nodes, but it is not required.
- Failure Tolerance Method of RAID-5/6, available starting with vSAN6.6/VxRail 4.5.070, the configuration must be all-flash.
- The minimum supported configuration is 1+1+1. See the *vSAN 2-Node Cluster on VxRail Planning Guide* for more detailed information.
- The minimum number of nodes depends the VxRail version and stretched cluster configuration. See Table 1 for Best Practice configurations.

VxRail Version		Minimum Nodes Preferred Site + Secondary Site + Witness
VxRail 3.5		4 + 4 + 1
VxRail 4.0.x and 4.5.0		3 + 3 + 1
VxRail 4.5.070 and beyond **NOTE:** This configuration depends the values set for PFTT, SFTT, and Failure Tolerance Method.	PFTT = 1; SFTT=1; Failure Tolerance Method=RAID-1 (Mirroring)	3 + 3 + 1
	PFTT = 1; SFTT=2; Failure Tolerance Method=RAID-1 (Mirroring)	5 + 5 +1
	PFTT = 1; SFTT=3; Failure Tolerance Method=RAID-1 (Mirroring)	7 + 7 + 1
	PFTT = 1; SFTT=1; Failure Tolerance Method=RAID-5/6 (Erasure Coding)	4 + 4 + 1
	PFTT = 1; SFTT=2; Failure Tolerance Method=RAID-5/6 (Erasure Coding)	6 + 6 + 1

Table 1 **VxRail Version Minimum Number of Nodes per Site**

- Starting with VxRail v4.7.x, 2+2+1 configurations are supported.
- The maximum supported configuration is 15+15+1 (30 nodes+1 witness).
- A witness host must be installed on a separate site as part of the installation engagement. See Table 2 for version compatibility.

VxRail Version	Witness Host OVA Version
VxRail 3.5	OVA Version 6.2
VxRail 4.0.x	OVA Version 6.2
VxRail 4.5.x	OVA Version 6.5
VxRail 4.7.x	OVA Version 6.7

Table 2 **VxRail/Witness Host OVA Compatibility Chart**

2.6 vCenter Server requirements

Before VxRail 4.5.200, only a customer-supplied vCenter could be used for stretched clusters. Starting with VxRail 4.5.200, either a VxRail vCenter Server or a customer-supplied vCenter Server can be used for stretched clusters. See the *VxRail vCenter Server Planning Guide* for caveats of using VxRail vCenter Server.

Customer-supplied vCenter Server Appliance is the recommended choice.

2.7 Customer-supplied vCenter Server requirements

The following are the customer-supplied vCenter Server requirements:

- The customer must provide the vSphere Enterprise Plus license.
- The customer-supplied vCenter Server version must be in the VxRail and external vCenter interoperability matrix. In addition, the ESXi version hosting the vCSA should be running version 6.0 or later.
- Check the VxRail Release Notes for to determine the proper version numbers.
  - VxRail 3.5 and vSphere 6.0, version details can be found in *VxRail Appliance Software 3.5 Release Notes*.
  - VxRail 4.0.x and vSphere 6.0, version details can be found in *VxRail Appliance Software 4.0.x Release Notes.*
  - VxRail 4.5.x and vSphere 6.5, version details can be found *in VxRail Appliance Software 4.5.x Release Notes*.
  - VxRail 4.7.x and vSphere 6.7, version details can be found *in VxRail Appliance Software 4.7.Release Notes*.

To join the customer-supplied vCenter Server, you must:

- Know whether your customer-supplied vCenter Server has an embedded or non-embedded Platform Services Controller. If the PSC is non-embedded, you will need the PSC FQDN.
- Know the customer-supplied vCenter Server FQDN.
- Know the Customer Existing Single Sign-on domain (SSO) (For example vsphere.local).
- Create or select a data center on the customer-supplied vCenter Server for the VxRail Cluster to join.
- Specify the name of the cluster that will be created by VxRail in the selected data center when the cluster is built. It will also be the name of the distributed switch. This name must be unique and not used anywhere in the data center on the customer-supplied vCenter Server.
- Verify that the customer DNS server can resolve all VxRail ESXi hostnames before deployment.
- Create or reuse a VxRail management user and password for this VxRail cluster on the customer-supplied vCenter Server. The user must be:
  - Created with no permissions
  - Created with no roles assigned to it
- (Optional) Create a VxRail admin user and password for VxRail on the Customer-Supplied vCenter Server.

3 Networking and latency

3.1 Layer 2 and Layer 3 support

Starting with VxRail 4.7.300, you can use Layer 3 between the two sites, but you must use Layer 2 for the management network.

For vSAN, you can use either Layer 2 or Layer 3. However, for Layer 2 configurations. Ensure that the Witness site in L2 enclosures is using high bandwidth and low latency.

Connectivity between the data sites and the witness must be Layer 3. Figure 1 illustrates a supported configuration.

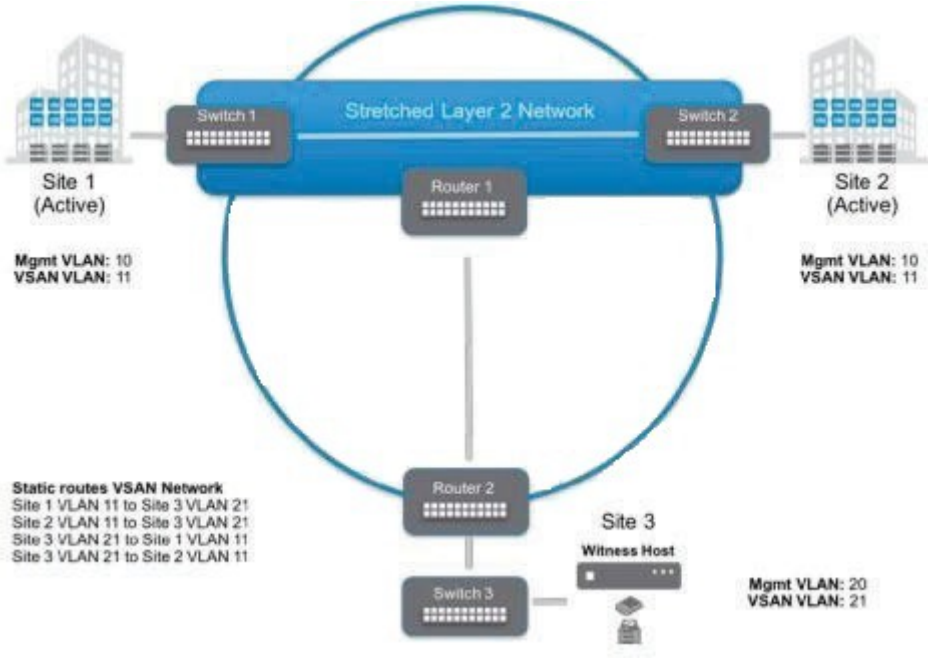

Figure 1 **VxRail Supported Topology**

3.2 Supported geographical distances

For vSAN stretched clusters, support is based on network latency and bandwidth requirements, rather than distance. The key requirement is the actual latency numbers between sites.

3.3 Data site to data site network latency

Latency or Round-Trip Time (RTT) between sites hosting virtual machine objects should not be greater than 5 msec (< 2.5 msec one-way).

3.4 Data site to data site bandwidth

Bandwidth between sites hosting virtual machine objects are workload-dependent. For most workloads, VMware recommends a minimum of 10 Gbps or greater bandwidth between sites.

3.5 Data site to witness network latency

In most vSAN stretched-cluster configurations, latency or RTT between sites hosting VM objects and the witness nodes should not be greater than 200 msec (100 msec one-way).

The latency to the witness is dependent on the number of objects in the cluster. VMware recommends that on vSAN stretched-cluster configurations up to 10+10+1, a latency of less than or equal to 200 milliseconds is acceptable. If possible, a latency of less than or equal to 100 milliseconds is preferred. For configurations that are greater than 10+10+1, VMware recommends a latency of less than or equal to 100 milliseconds is required.

3.6 Data site to witness network bandwidth

Bandwidth between sites hosting VM objects and the witness nodes are dependent on the number of objects residing on vSAN. It is important to size data site to witness bandwidth appropriately for both availability and growth. A standard guideline is 2 Mbps for every 1000 objects on vSAN.

3.7 Inter-site MTU consistency

It is important to maintain a consistent MTU (maximum transmission unit) size between data nodes and the witness in a stretched-cluster configuration. Ensuring that each VMkernel interface designated for vSAN traffic is set to the same MTU size will prevent traffic fragmentation. The vSAN Health Check checks for a uniform MTU size across the vSAN data network, and reports on any inconsistencies.

3.8 Connectivity

- Management network: Connectivity to all three sites
- VM network: Connectivity between the data sites (the witness will not run virtual machines that are deployed on the vSAN cluster).
- vMotion network: Connectivity between the data sites (virtual machines are never migrated from a data host to the witness host).
- vSAN network: Connectivity to all three sites

3.9 Witness traffic separation

vCenter Server 6.7u1 supports Witness Traffic Separation (WTS) for VxRail stretched cluster deployments. This feature allows an alternate VMkernel interface to be designated to carry traffic that is destined for the Witness rather than the vSAN tagged VMkernel interface. This feature supports more flexible network configurations by allowing separate networks for node-to-node and node-to-witness traffic. From a routing perspective, this feature allows two independent subnets/routes to be advertised from each Data Node site to the Witness site.

Appendix A: VxRail stretched cluster setup checklist

Required Reading	✓ Read the *VMware vSAN Stretched Cluster Guide.* ✓ Read the *VxRail vCenter Server Planning Guide.*
VxRail Version	✓ The minimum version is VxRail 3.5. ✓ No mixed clusters are supported (that is, VxRail 4.5 and 4.0 in the same cluster).
vSphere License	✓ vSphere Enterprise Plus license is required. ✓ You cannot reuse the VxRail vCenter Server license on any other deployments.
Number of Nodes	✓ Review Table 1 in this guide for the minimum number of nodes. ✓ The minimum supported configuration is 1+1+1. ✓ The maximum supported configuration is 15+15+1 (30 nodes+1 witness).
customer-supplied vCenter Server (Recommended choice)	✓ Required for versions earlier than VxRail 4.5.200 ✓ The customer-supplied vCenter Server version must be in the *VxRail and external vCenter interoperability matrix.*
Fault Domains	✓ Must have 3 Fault Domains (preferred, secondary, and witness host).
Network Topology	✓ vSAN traffic between the data sites can be Layer 2 or Layer 3. Ensure that the Witness site in L2 enclosures is using high bandwidth and low latency. ✓ vSAN traffic between the witness host and the data sites must be Layer 3.
Data Site to Data Site Network Latency	✓ Latency or RTT between data sites should not be greater than 5 msec. (<2.5 msec one-way)
Data Site to Data Site Bandwidth	✓ A minimum of 10 Gbps is required.
Data Site to Witness Network Latency	✓ For configurations up to 10+10+1, latency or RTT less than or equal to 200 msec is acceptable, but 100 msec is preferred. ✓ For configuration greater than 10+10+1, latency or RTT less than or equal to 100 msec is required.
Data Site to Witness Network Bandwidth	✓ The guideline is 2 Mbps for every 1000 objects on vSAN.
Inter-site MTU consistency	✓ Required to be consistent between data sites.
Network Ports	✓ Review *Appendix B* for required port connectivity.

Appendix B: VxRail stretched cluster open port requirements

The following table lists the open port requirements for a VxRail stretched cluster. See https://ports.vmware.com/home/vSphere for the latest list of port connectivity requirements.

Description	Connectivity To and From	L4 Protocol	Port
vSAN Clustering Service	vSAN Hosts	UDP	12345, 23451
vSAN Transport	vSAN Hosts	TCP	2233
vSAN VASA Vendor Provider	vSAN Hosts and vCenter	TCP	8080
vSAN Unicast Agent (to Witness Host)	vSAN Hosts and vSAN Witness Appliance	UDP	12321