vSAN Stretched Cluster Bandwidth Sizing

Recommendations for vSAN 8 U3 and VMware Cloud Foundation 5.2

January 2, 2025

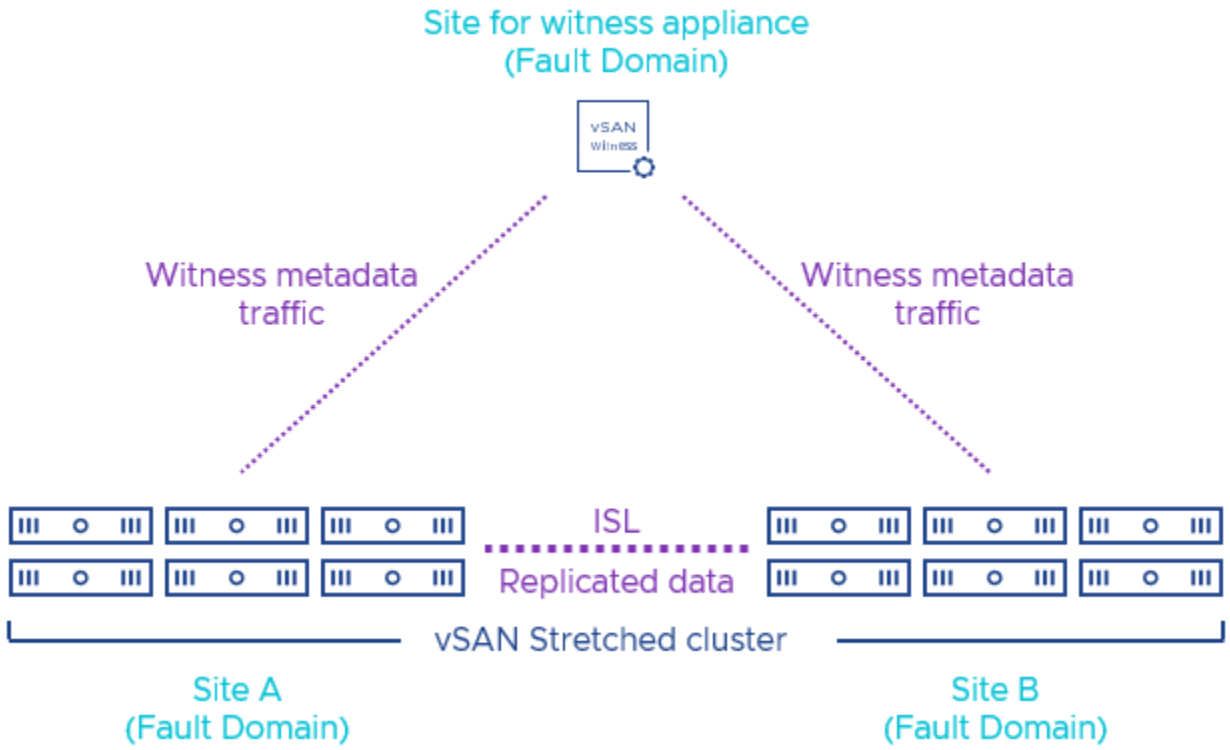
vmware by Broadcom

Table of Contents

Introduction

This document explains how to size network bandwidth between sites when using vSAN HCI clusters and vSAN Max clusters in a stretched cluster configuration.

In stretched cluster configurations, two data fault domains have one or more hosts, and the third fault domain contains a virtual witness appliance to help determine availability of the data sites. In this document each data fault domain will be referred to as a site. vSAN stretched cluster configurations can be spread across distances, provided bandwidth and latency requirements are met.



Scope of Topics

The information provided in this document will assume the use of vSAN 8 U3, and/or VMware Cloud Foundation (VCF) 5.2. VCF deployments may have additional requirements and support limitations that fall outside of the scope of this document.

Minimum Supported Bandwidth and Latency Between Sites

The bandwidth needed between sites will depend largely on the amount of I/O that the workloads are generating, but will be influenced by other factors, such as the architecture used (vSAN ESA or OSA), cluster size, and failure handling scenarios. The minimum supported values are noted below, but your calculated estimates may result in bandwidth requirements that exceed the stated minimums.

Connectivity	Bandwidth	Latency
Data Site to Data Site	Minimum of 10Gbps	<5ms latency RTT
Data Site to Witness Site	2Mbps per 1,000 vSAN components	<200ms latency RTT (up to 10 hosts per site) <100ms latency RTT (11 to 15 hosts per site) <500ms latency RTT (1 host per site)

Recommendation. Please use the vSAN Sizer paired with the vSAN Design Guide to come up with an optimal configuration for your stretched cluster environment.

Understanding I/O Activity, Read & Write Ratios, and I/O sizes

Workloads are comprised of several characteristics. Not only does the amount of I/O activity vary greatly from workload to workload, but reads and writes are often occurring at different rates, with different I/O sizes. In the interest of providing a simple formula for calculation, we will use the following variables to calculate the estimated bandwidth for the stretched cluster inter-site link (ISL).

- I/O rate of all read and write commands, measured in IOPS
- Read/Write ratios, measured as a ratio (e.g. 70/30)
- Average I/O size, measured in KB

Bandwidth is a rate-based measurement, meaning that it expresses how much data is moved for a given period of time. Unlike measuring data at rest, where it takes the form of Bytes, KiloBytes, etc, the measurement of data in transit across a network is in bits (b), Kilobits (Kb), Megabits (Mb), or Gigabits (Gb), and the period of time is "per second" (ps). When discussing rates, we must remember to convert bytes to bits.

Lets use a simple example where a total I/O profile requires 100,000 I/Os per second (IOPS) averaging 8KB in size, where 70% of the IOPS are read, and 30% are write, in a stretched configuration. In this scenario, the write I/O (30,000 IOPS averaging 8KB in size). would calculate to 240,000 Kbps, or 240 Mbps. This would be the estimate of ISL bandwidth needed for this profile.

The simplest, but potentially less accurate way of estimating bandwidth requirements is to simply use inputs that represent the total needs of the cluster. If one is using the formula to calculate discrete workloads in a cluster, the sum of those calculations would provide the resulting bandwidth requirements. One may wish to allocate much more than the minimum calculated amount, as workloads inevitably become more resource intensive.

The formulas to calculate bandwidth estimates are noted in the sections below, distinguished by their respective architecture used: vSAN OSA or vSAN ESA. With the introduction of the vSAN Express Storage Architecture (ESA) comes all new capabilities in delivering storage performance at all new levels. Therefore, when using the ESA, it is advised to monitor the utilization of the ISL closely to ensure there is sufficient bandwidth necessary so that the workloads are not adversely constrained by the ISL. For more information, see the post: "Using the vSAN ESA in a Stretched Cluster Topology."

Bandwidth Requirements Between Data Sites

The calculation formulas below help estimate the network bandwidth required between data sites. It is assumed that the two data sites are close enough in geographic location to meet the latency requirements.

Bandwidth Calculation Formulas for vSAN ESA

Replica traffic from guest VM I/O is the dominate traffic type that we must account for when estimating the bandwidth needed for the inter-site link (ISL) traffic. With vSAN ESA stretched clusters, read traffic is, by default, serviced by the site that the VM resides on. The required bandwidth between two data sites **(B)** is equal to **Write bandwidth (Wb) * data multiplier (md) * resynchronization multiplier (mr) * compression ratio (CR),** or:

The calculation of bandwidth will be for the ISL serving a vSAN stretched cluster using the ESA is different than the formula used for the OSA. The **vSAN ESA will compress the data prior to it being replicated across the sites.** As a result, the ESA will reduce the bandwidth usage of an ISL, effectively increasing its capabilities to transmit more data across the link. To account for this in the calculation in a topology using the ESA, the formula factors in the estimated compression ratio of the data stored. Lets use the following example, where we are estimating a 2x savings by using compression in vSAN ESA. We are using a simple "2x" (also referred to as 2:1, or 50%) for simplicity. Actual compression ratios will depend on the data stored in your environment. The 2:1 example is not a suggestion of what you may see in your environment.

1. Convert this to a percentage of the original size by dividing the ending unit size (1) by the starting unit size (2). This will result in ".50"
2. Multiply the final calculated amount from the formula described in this document by ".50"

As a result, the formula would look like:

B = Wb * md * CR * mr * CR

As noted above, compression ratios can often be shown in different ways to express capacity savings. For example:

- Compression **expressed as a ratio**. [starting unit size]:[ending unit size]. A compression ratio of 2:1 indicates that the data would be compressed to half its original size.
- Compression **expressed as a multiplier of savings**. A compression ratio of 2x indicates that the data would be compressed to half its original size. This is what is rendered in the vSAN cluster capacity view.
- Compression **expressed as a percentage reduced from its original size**. A compression value of 50% (or, .50) indicates that the data would be compressed to half its original size.

Site to Site Examples (for vSAN ESA)

Workload 1

- With an example workload of 10,000 writes per second to a workload on vSAN with an average of a 8KB size write, that would require 80MB/s, or 640 Mbps bandwidth. Lets assume the data has a 2x compression ratio.
- B = 640 Mbps * 1.4 * 1.25 * .50 = 560 Mbps.
- Including the vSAN network requirements, the required bandwidth would be 560 Mbps.

Workload 2

- In another example, 30,000 writes per second, 8KB writes, would require 240MB/s, or 1,920Mbps bandwidth. Lets assume the data has a 2x compression ratio.
- B = 1,920 Mbps * 1.4 * 1.25 * .50 = 1,680 Mbps or ~1.7Gbps
- The required bandwidth would be approximately 1.7Gbps.

Workloads are seldom all reads or writes, and normally include a general read to write ratio for each use case. Using the general situation where a total I/O profile requires 100,000 IOPS, of which 70% are write, and 30% are read, in a Stretched configuration, the write IO is what is sized against for inter-site bandwidth requirements. With vSAN stretched clusters, read traffic is, by default, serviced by the site that the VM resides on.

Bandwidth Calculation Formulas for vSAN OSA

Much like stretched clusters running the ESA, replica traffic from guest VM I/O in a stretched cluster using the OSA is the dominate traffic type that we must account for when estimating the bandwidth needed for the inter-site link (ISL) traffic. With vSAN OSA stretched clusters, read traffic is, by default, serviced by the site that the VM resides on. The required bandwidth between two data sites **(B)** is equal to **Write bandwidth (Wb) * data multiplier (md) * resynchronization multiplier (mr),** or:

B = Wb * md * mr

The data multiplier is comprised of overhead for vSAN metadata traffic and miscellaneous related operations. VMware recommends a data multiplier of 1.4. The resynchronization multiplier is included to account for resynchronizing events. It is recommended to allocate bandwidth capacity on top of required bandwidth capacity for resynchronization events. Making room for resynchronization traffic, an additional 25% is recommended.

Recommendation: Plan on using vSAN ESA for all new vSAN clusters. It is faster, and more efficient than the vSAN OSA, and often times reduce the number of hosts required for the same amount of workloads. The formula for vSAN OSA remains available for those with existing installations of vSAN OSA.

Bandwidth Requirements Between Witness and Data Sites

In a stretched cluster topology, the witness host appliance simply stores components that are comprised of a small amount of metadata that help determine availability of the objects stored at the data sites. As a result, network traffic sent to the site for the witness host appliance is quite small when compared to the traffic between the two data sites via the ISL.

Therefore, the sizing of bandwidth for a connection to a virtual witness host appliance from a data site will result in much smaller bandwidth requirements. One of the most significant variables is the amount of data stored in each site. vSAN stores data in the form of objects and components. It determines how many components are needed for a given object, and where they should be placed across the cluster to maintain data resilience as prescribed by the assigned storage policy.

The vSAN Express Storage Architecture (ESA) typically uses more components than the Original Storage Architecture (OSA). One should factor this in when calculating the potential bandwidth needed for the witness host appliance.

Recommendation: Make sure to choose the correct size of a vSAN Witness host deployment. The deployment offers four witness host appliance sizes ("Tiny," "Medium," "Large," & "Extra Large") each one geared to accommodate different sizes of environments. See "Deploying a vSAN Witness Appliance" for more information.

Witness Bandwidth Calculation Formulas for vSAN ESA and OSA

Network communication from the data sites to the witness host appliance is comprised entirely of metadata. This makes for a much lighter demand on networking to and from the witness host appliance, which is reflected in the supported minimum bandwidth and latency requirements for the witness site.

Since the formula for estimating bandwidth required for the witness site is based on the number of components, it can be used for calculating stretched clusters that are using the OSA, or the ESA. Since the ESA typically uses more components per object (potentially 2-3x) than the OSA, one should factor in a higher component count per VM when using a design that is running the ESA.

Basic Formula

The required bandwidth between the Witness and each site is equal to ~1138 B x Number of components / 5s

1138 B x NumComp / 5 seconds

The 1138 B value comes from operations that occur when the Preferred Site goes offline, and the Secondary Site takes ownership of all of the components. When the primary site goes offline, the secondary site becomes the leader. The Witness sends updates to the new leader, followed by the new leader replying to the Witness as ownership is updated. The 1138 B requirement for each component comes from a combination of a payload from the Witness to the backup agent, followed by metadata indicating that the Preferred Site has failed. In the event of a Preferred Site failure, the link must be large enough to allow for the cluster ownership to change, as well as ownership of all of the components within 5 seconds.

Witness to Site Examples (OSA)

Workload 1

With a VM being comprised of

- 3 objects
- Failure to Tolerate of 1 (FTT=1)

Approximately 166 VMs with the above configuration would require the Witness to contain 996 components (166 VMs * 3 components/VM * 2 (FTT+1) * 1 (Stripe Width)). To successfully satisfy the Witness bandwidth requirements for a total of 1,000 components on vSAN, the following calculation can be used:

Converting Bytes (B) to Bits (b), multiply by 8

B = 1138 B * 8 * 1,000 / 5s = 1,820,800 Bits per second = 1.82 Mbps

VMware recommends adding a 10% safety margin and round up.

B + 10% = 1.82 Mbps + 182 Kbps = 2.00 Mbps

Therefore, with the 10% margin of safety included, a rule of thumb can be stated that for every 1,000 components, 2 Mbps is appropriate.

Workload 2

With a VM being comprised of

- 3 objects
- Failure to Tolerate of 1 (FTT=1)
- Stripe Width of 2

Approximately 1,500 VMs with the above configuration would require 18,000 components to be stored on the Witness. To successfully satisfy the Witness bandwidth requirements for 18,000 components on vSAN the resulting calculation is:

B = 1138 B * 8 * 18,000 / 5s = 32,774,400 Bits per second = 32.78 Mbps

B + 10% = 32.78 Mbps + 3.28 Mbps = 36.05 Mbps

Using the general equation of 2Mbps for every 1,000 components, (NumComp/1000) X 2Mbps, it can be seen that 18,000 components does in fact require 36Mbps.

Witness Bandwidth for 2 Node Configurations (OSA)

Remote Site Example 1

Take the example of 25 VMs in a 2 Node configuration, each with a 1TB virtual disk protected at FTT=1 and a Stripe Width=1. Each vmdk would be comprised of 8 components (vmdk and replica) and 2 components each for the VM namespace and swap file. The total number of components is 300 (12/VMx25VMs). With 300 components, using the rule of thumb (300/1000 x 2Mbps), 600kbps of bandwidth is required.

Remote Site Example 2

Take another example of 100 VMs on each host, of the same VM above, with 1TB virtual disk, FTT=1 & SW=1. The total number of components would be 2,400. Using the rule of thumb (2,400/1000 x 2Mbps), 4.8Mbps of bandwidth is required.

Summary

The calculations in this document will provide a reasonably accurate estimate of bandwidth needed for a stretched cluster environment. While the numbers may produce a result that are well below the bandwidth you currently have between sites, having high bandwidth, low latency connectivity will provide a much better experience for your workloads and the consumers who use them.

Additional Resources

The following are a collection of useful links that relate to bandwidth sizing for vSAN stretched clusters.

Performance Recommendations for vSAN ESA. This is a collection of recommendations to help achieve the highest levels of performance in a vSAN ESA cluster. Many of these same recommendations apply to vSAN storage clusters.

vSAN Proof of Concept (PoC) Performance Testing. This is a collection of recommendations that will guide users to test the performance of a vSAN cluster. While it is currently written for the OSA, many of the testing methods used are also applicable to the ESA.

Design and Sizing for vSAN ESA clusters. This post offers some nice guidance on using the vSAN Sizer for the ESA that summarizes some key points that can be found in the VMware vSAN Design Guide.

vSAN Network Design Guide. This network design guide applies to environments running vSAN 8 and later.

vSAN technical blogs. Stay up to date on the most recently published technical information about vSAN. These posts are created by the vSAN Technical Marketing team.

VMware Resource Center. The location for design guides, operations guides and other technical white papers on vSAN. These assets are created by the vSAN Technical Marketing and Product Enablement teams.

Official vSAN documentation. The location for all "how to" documentation on vSAN.

About the Author

Pete Koehler is a Product Marketing Engineer in the VCF division at Broadcom. With a primary focus on vSAN, Pete covers topics such as design and sizing, operations, performance, troubleshooting, and integration with other products and platforms.