Microsoft SQL Server Failover Cluster Instance on VMware vSAN Stretched Cluster

Table of contents

Microsoft SQL Server Failover Cluster Instance on VMware vSAN Stretched Cluster 3
Business Case 3
Solution Overview 4
Solution Configuration 5
- Architecture 5
- Hardware Resources 6
- Software Resources 7
- Network Configuration 8
- SQL Server 2017 Virtual Machine Configuration 8

Performance Results 11
Application Role Failover and Site Failure Validation 13
Recommendations 14
Conclusion 16
References 17
About the Author 18

Microsoft SQL Server Failover Cluster Instance on VMware vSAN Stretched Cluster

Business Case

Clustered SQL Server solution with shared disks provides the high availability capability in case of physical or operating system failures. The most used active-passive model can be extended to cross data centers with advanced storage capability such as Dell EMC's VPLEX and IBM's SVC.

vSAN Stretched clusters extend the vSAN cluster from a single data site to two sites for a higher level of availability and inter-site load balancing. Stretched clusters can be used to manage planned maintenance and avoid disaster scenarios, because maintenance or loss of one site does not affect the overall operation of the cluster. In a stretched cluster configuration, both data sites are active sites. If either site fails, data is still available on the alternate site. When configured with vSphere HA, VM's are automatically restarted on the remaining active site.

With the release of vSAN 6.7 Update 3, vSAN Stretched Clusters provide a solution for clustered applications like Microsoft SQL Server (further referenced as SQL Server) to use shared disk across sites. This allows data center administrators to run workloads using legacy clustering technologies on vSAN across two data centers which can fully leverage compute resource on the data centers while having the capability to sustain one site failure.

Solution Overview

This reference architecture validates the solution of a SQL Server Failover Clustering Instance (FCI) using shared disks backed by vSAN stretched cluster. OLTP Performances of SQL Server 2017 on Windows Server Failover Clustering (WSFC) at different inter-site latency are demonstrated. We showcased the FCI role failover across sites using shared disks, and the capability of vSAN stretched cluster to support the FCI without application outage during a site failure.

Solution Configuration

- Architecture
- Hardware resources
- Software resources
- Network configuration
- SQL Server configuration including:
- Database space usage
- Node placement and vSAN policy
- SQL Server virtual machine and disk layout

Architecture

This solution designed two SQL Server Failover Clusters on vSAN stretched cluster. We emulated a 300GB tier-1 application database with cross site RAID-1 for the data and log disks and the quorum disk.

We used four DELL PowerEdge R630 servers, a 1U platform for density, performance and scalability, and with optimized application performance, to form the vSAN all-flash stretched cluster. A vSAN Witness Appliance was used as the vSAN Witness, was located on another management cluster. A virtual application located on the management cluster was used for Layer 3 routing, bridging the two sites for vSAN traffic.

The configuration consisted of six Windows 2016 virtual machines on a vSAN stretched cluster as shown in Figure 1. Two SQL Server failover clusters were configured with identical disk layouts and vSAN Storage Policy assignments. Their operating system, database, log, and quorum disks all reside on the vSAN stretched cluster. We put one active node on one site and another active node on another site for both WSFCs. Two domain controllers were created for the WSFC with each site having one to provide the authentication and DNS service with the capability to serve when one site was down.

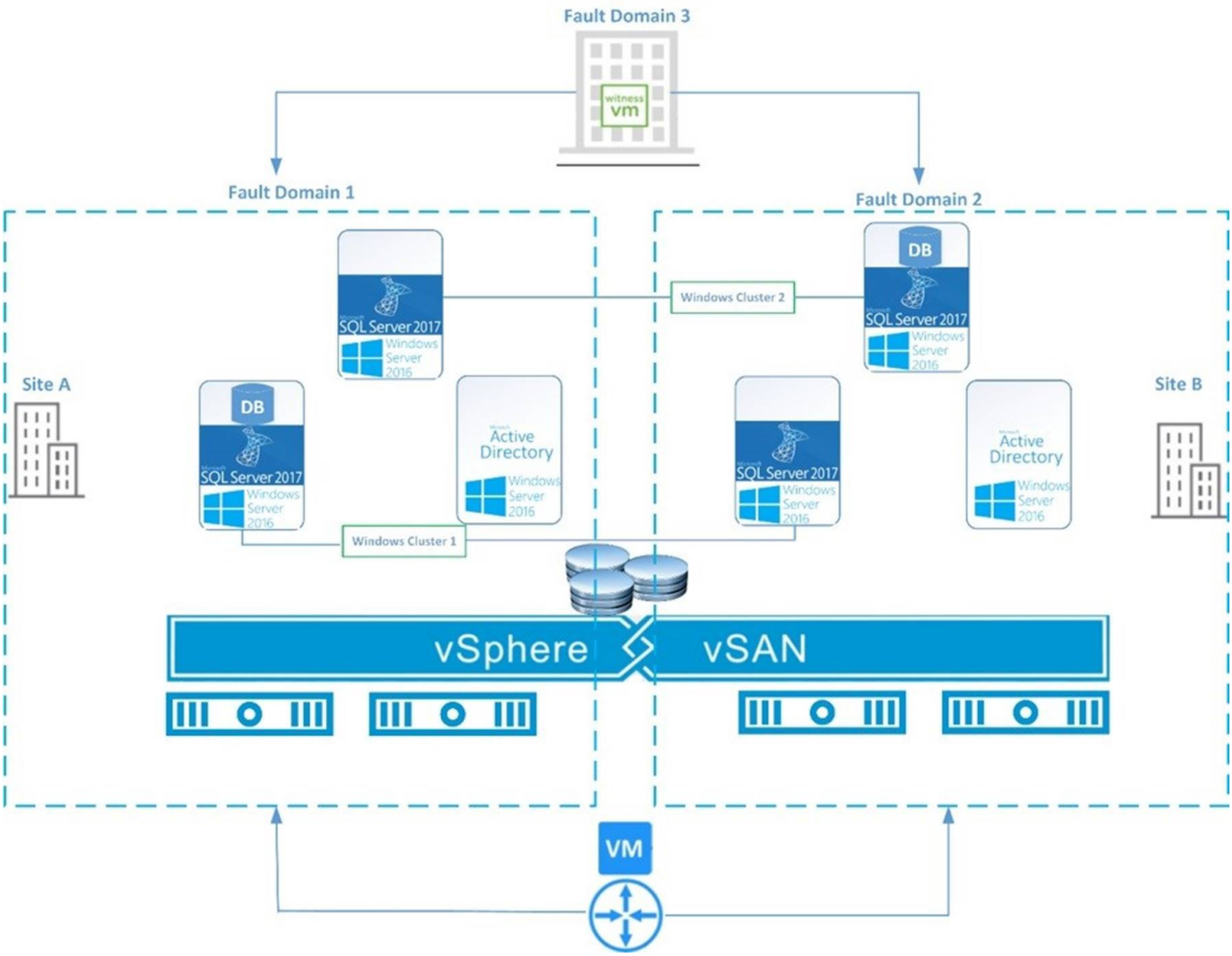

Figure 1. Solution Architecture

Hardware Resources

Each VMware ESXi™ host contains two disk groups, with each disk group consisting of one cache-tier NVMe SSD and four capacity-tier SAS SSDs. We configured pass-through mode for the capacity-tier storage controller.

The ESXi Server in the vSAN Cluster has the following configuration as shown in Table 1.

Table 1. Hardware Resources

Property	SPECIFICATION
Server	Dell PowerEdge R630
CPU	2 sockets, 24 cores each of 2.6GHz
RAM	256 GB DDR4 RDIMM
Network adapter	2 x Intel 10 Gigabit X540-AT2, + I350 1Gb Ethernet
Storage adapter	2 x 12 Gbps SAS PCIExpress (Dell PowerEdge RAID H730 mini)
Disks	NVMe: Samsung 2 x 1.6 TB drives as cache SSD SSD: 6 x 400 GB drives as capacity SSD

Software Resources

Table 2 shows the software resources used in this solution.

Table 2. Software Resources

Software	Version	Purpose
VMware vCenter ® Server and VMware ESXi™/Witness Virtual Appliance	6.7 update 03 ESXi build 14320388 vCenter build 14016707	ESXi Cluster to host virtual machines and provide the vSAN Cluster. VMware vCenter Server provides a centralized platform for managing VMware vSphere environments
VMware vSAN	6.7 update 03	Software-defined storage solution for hyper-converged infrastructure
Microsoft SQL Server	2017 Enterprise Edition, RTM-CU17	Database software

Software	Version	Purpose
Microsoft Windows Server	2016, x64, Standard Edition	Operating System for the VMS:
Benchmark Factory for Databases	8.1	OLTP database and workload generator
Ubuntu Linux	14.04.2	L3 router for intersite data flow and latency emulation

Network Configuration

A VMware vSphere Distributed Switch™ acts as a single virtual switch across all associated hosts in the data cluster. This setup allows virtual machines to maintain a consistent network configuration as they migrate across multiple hosts. The vSphere Distributed Switch uses two 10GbE adapters for the teaming and failover purposes. A port group defines properties regarding security, traffic shaping, and NIC teaming. We used default port group setting except the uplink failover order as shown in Table 3. It also shows the distributed switch port groups created for different functions and the respective active and standby uplink to balance traffic across the available uplinks.

Table 3. Uplink and VLAN settings of the Distributed Switch Port Groups

Distributed Switch Port Group Name	VLAN	Active Uplink	Standby Uplink
vSAN Witness/Manage ment/vMotion	1284	Uplink 1	Uplink2
vSAN Stretched Cluster (Site A)	4040	Uplink 2	Uplink1
vSAN Stretched Cluster (Site B)	4041	Uplink 2	Uplink1

An Ubuntu Linux virtual machine was configured as a Layer 3 router. Static routes were added to all vSAN hosts to ensure proper communication over the Layer 3 network.

SQL Server 2017 Virtual Machine Configuration

We followed the Architecting Microsoft SQL Server on VMware vSphere Best Practice Guide to ensure an optimal SQL Server configuration.

Database Space Usage

The Scale Factor determines the database size. We configured three sized DB servers for the performance tests. We set SF=32 for medium-sized database which created 300GB database, or a TPC-E like OLTP user database with customer number 32,000, which generated around 327GB data in the data files.

Node Placement and vSAN Policy

A vSAN stretched cluster was formed by configuring two nodes in one fault domain, two nodes in another fault domain, and designating the vSAN Witness Host, which resided on the management cluster in a different site. See the figure below for the fault domain configuration for the vSAN stretched cluster.

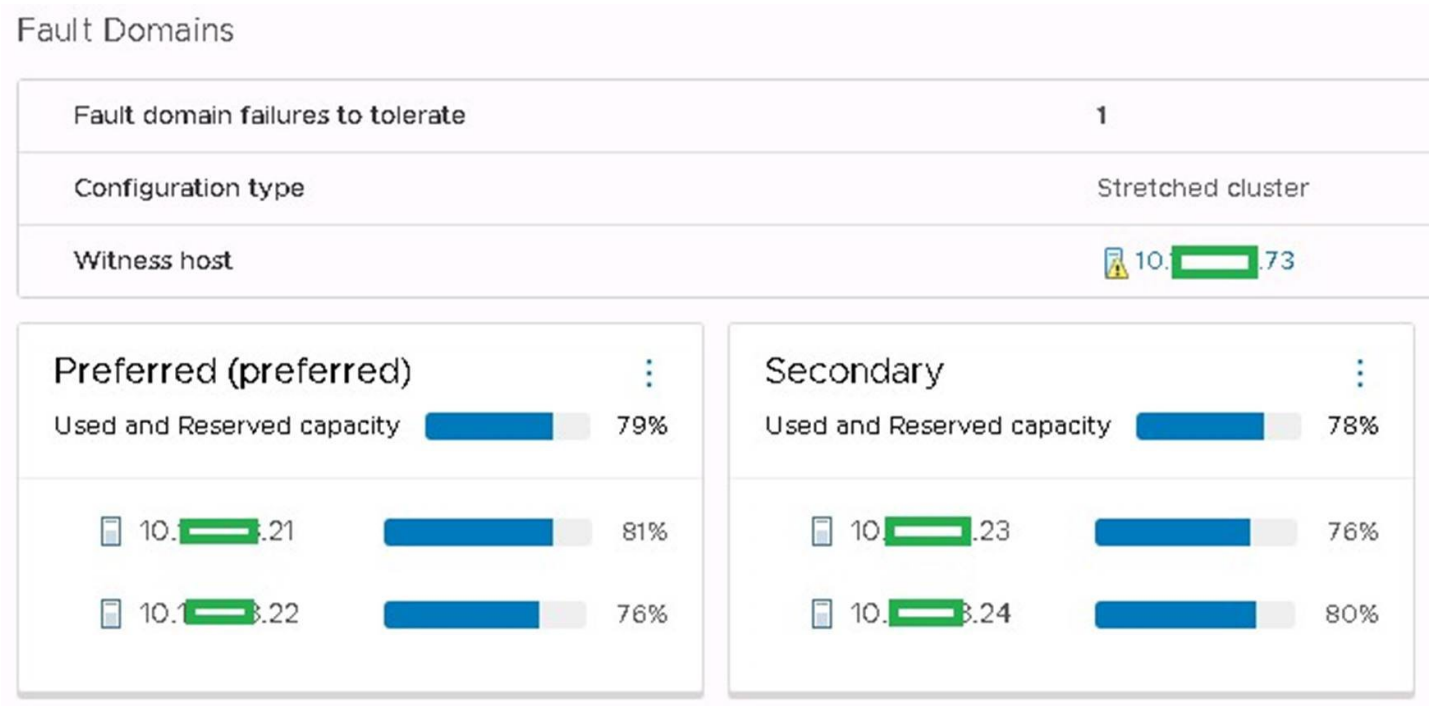

Figure 2. Fault Domain Setting of the vSAN Stretched Cluster

As for WSFC cluster 1, the Active node was placed on site A, and the standby node was place on site B, and WSFC cluster 2 nodes were placed with the reversed order. We set the "Site disaster tolerance" rule of the VM Storage Policy to "Dual site mirroring (stretched cluster)" for VM home, OS, DB/Log and TempDB VMDK, and the quorum VMDK. See the figure below for the cross-site RAID-1 setting in the VM Storage policy.

Note: The test configuration of two nodes per site only allowed data protection across sites. Data protection within sites requires three or more hosts to satisfy local site protection policies.

Figure 3. vSAN Storage Policy Configuration

Advanced vSAN data services, including deduplication & compression, encryption, and the iSCSI Target Service were not enabled. The vSAN Performance Service was enabled.

SQL Server Virtual Machine and Disk Layout

We designed two SQL Server failover clusters with each supporting one 300GB database. We configured four SQL Server virtual machines for the performance tests. Databases were created by the Benchmark Factory for Databases. The database and index files consumed approximately 300GB space. The four virtual machines were assigned 24 vCPUs, 64GB RAM, and their memory was reserved.. We set maximum and minimum memory of SQL Server instance to 51GB. Each SQL Server virtual machine in the cluster had VMware Tools installed, and each virtual hard disk for data and log was connected to separate VMware Paravirtual SCSI (PVSCSI) adapter to ensure the optimal throughput and lower CPU utilization. The virtual network adapters used the VMXNET 3 adapter, which was designed for the optimal performance. For the TPC-E-like OLTP workload, the database size is based on the actual disk space requirement and additional space for the database growth. Table 4 is the disk layout of the two 300GB

databases.

Table 4. SQL Server Disk Layout

Purpose	Number x Size (GB)	SCSI Controller
Operating system	1 x 100	LSI Logic SAS 0
Log disk	1 x 80	PVSCSI 1
Data disks	2 x 250	PVSCSI 2
Tempdb	1 x 60	PVSCSI 3
Quorum disk	1 x 20	PVSCSI 1

Performance Results

We measured the performance of two databases on the four virtual machines that formed two SQL Server FCI with one cluster hosted one TPC-E-like OLTP database. The purpose of this performance test aimed to evaluate the impact of the inter-site latency to the OLTP workloads. And to seek for the optimal inter-site latency for tier-1 application running on SQL Server FCI on vSAN stretched cluster.

The number of users were adjusted as needed to generate enough OLTP activity necessary to saturate the host. Repeatability was ensured by restoring a backup of the database before each run. We kicked off the test concurrently from the test clients on the management cluster remotely, to avoid the side effect of the clients. The test duration was one hour with a 15-minutes warm-up duration. Table 5 shows the test results. From the table we found that the two databases performance results were quite similar on both TPS and Average Transaction Time in milliseconds. As inter-site latency increased the Average Transaction Time increased, while the TPS value dropped.

Table 5. TPS and Average Transaction Time in Milliseconds of the 2 x 300GB Databases

2 x 300 GB TPC-E like Performance	No Intersite Latecy	2ms Intersite Latency	4ms Intersite Latency	8ms Intersite Latency
TPS	4764	4413	4278	3196
Average Transaction Time(ms)	33.14	35.67	36.83	49.72

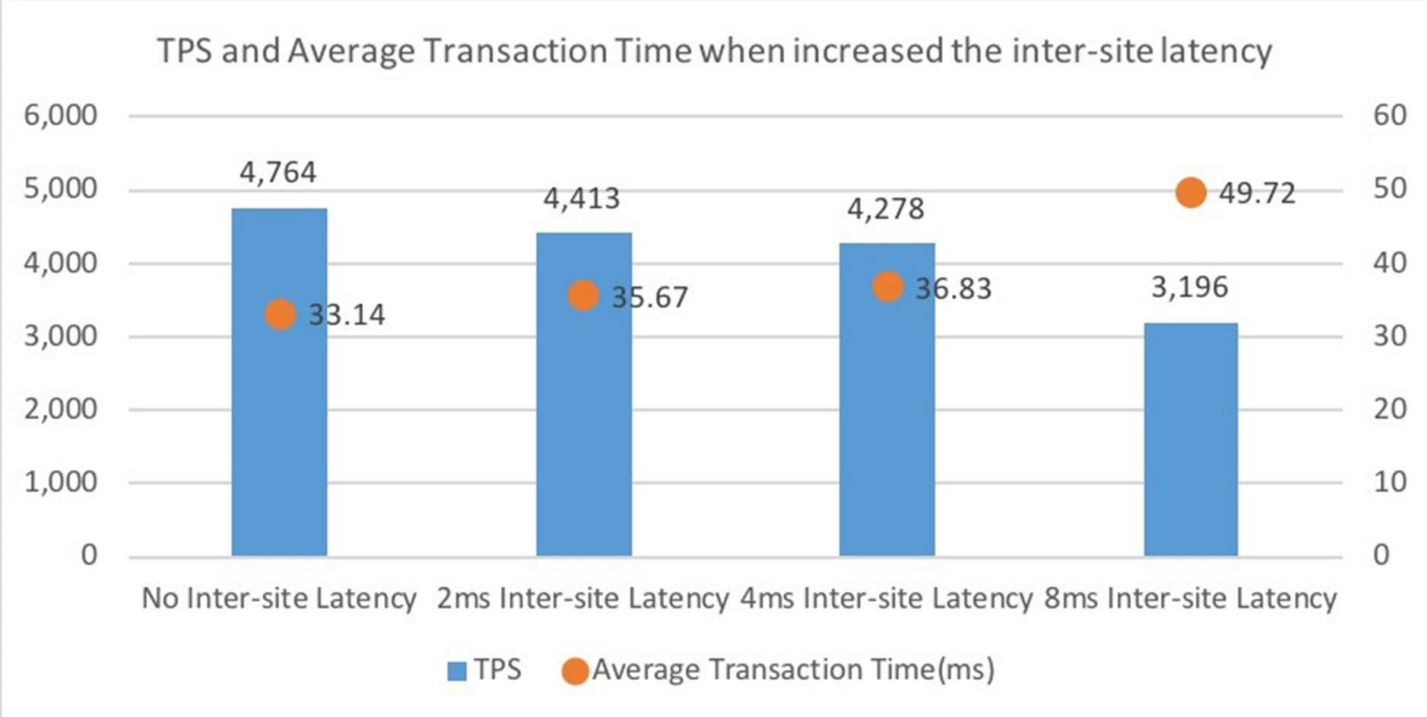

Figure 4. SQL Server FCI OLTP Performance on Stretched vSAN

We measured the vSAN VM performance and vSAN backend performance. Table 6 is the vSAN VM Performance of the two 300GB databases. While the average VM latency increased from 0.83-0.96ms to 2.45-2.82ms on Read, and the write latency increased from 1.9-2.5ms to 20.66-25.45ms on Write when increasing the inter-site latency up to 8ms, the average vSAN backend latency was kept ad 0.2ms on read and 0.16ms on write. That indicated that the latency was on the on-flight channel like network instead of on the backend. Also from SQL Server DBA point of view, more than 4ms inter-site would introduce 9-10ms average write latency, that means running SQL Server OLTP workload on a stretched cluster with more than 4ms inter-site latency for tier-1 application is not recommended and will have the performance impact.

Table 6. vSAN VM Performance of the 2 x 300GB Databases

vSAN VM Performance	No Intersite Latency	2ms Intersite Latency	4ms Intersite Latency	8ms Intersite Latency
Read IOPS	9300-17100	10300-16070	9170-16500	7700-11600
Write IOPS	1510-1960	890-1530	870-1160	680-1030
Read Latency(ms)	0.83ms-0.96ms	2.36ms-3.14ms	1.48ms-1.69ms	2.45ms-2.82ms
Write Latency(ms)	1.9ms-2.5ms	6.9ms-8.21ms	9.96ms-11.20 ms	20.66 ms25.45 ms

Application Role Failover and Site Failure Validation

Application Roles are clustered services like SQL Server FCI on WSFC. When a FCI is built on vSAN stretched cluster using shared disks for databases, role failover can happen from one site to another, when the active node is on one site and passive node(s) on another. The duration of a SQL Server instance failover from the active node to the standby node(s) in a WSFC only requires a few seconds. When configured with a Dual Site Mirroring Storage Policy, a vSAN stretched cluster can guarantee data access even when one site has failed or is isolated.

In this reference architecture, we verified the following scenarios and demonstrated the capability of Failover Cluster and advantages of vSAN stretched cluster.

- Scenario 1- Application Role Failover: we manually moved the SQL Server FCI from one node to another node.
- Scenario 2- Host Shutdown (non-primary node of WSFC): we shut down the host which hosted the non-primary node of the WSFC using iDRAC console of the Dell R630 server, to emulate the unplanned host failure of the physical host. Before this failure validation, we initiated the workload on one database.
- Scenario 3- Host Shutdown (primary node of WSFC): following scenario 2, we shut down the host which hosted the primary node of one cluster after the failure scenario 2, by shutting down the host using iDRAC console of the Dell R630 server, to emulate the unplanned site A failure. This scenario should cause VMs restart on another node in different site.

The test results were shown in Table 7.

Table 7. Failover and Failure Impact to Clustered Service

	Application Role Failover	Host Shutdown (non-primary node of WSFC)	Host Shutdown (primary node of WSFC) /Site A down
Failure observation and impact to running workload	The failover duration was less than 35 seconds		

The validation results showed:

- Application role failover worked as expected moving the service from one node to another.
- Failure of the non-primary node did not cause the workload interruption, and HA restarted the impacted VM on the remaining hosts.
- Failure of the primary node (simulating a site failure scenario) causes the restarting of the two VMs on site A (they were running on the only host on site A before it was turned off) and no cluster service down was monitored, when using disk witness of vSAN stretched cluster as the quorum.

Recommendations

- Less than four milliseconds inter-site (round trip) latency is recommended for tier-1 SQL Server databases running on vSAN stretched cluster.
- Enable DRS VM/Hosts Rule and create rules to separate the VMs of one WSFC on different ESXi hosts. And enable VM/Hosts Rule to separate the VMs of different WSFC nodes on different ESXi hosts for performance consideration terms. See the Figure 5 which is an example to create a rule to separate the two VMs in one WSFC on different ESXi hosts.

Figure 5. Create VM/Hosts Rule to Separate the VMs of a WSFC

- Use quorum disk witness as the cluster service quorum setting and vSAN stretched cluster can ensure the witness disk accessibility in a site failure without tearing down the cluster service. See the Figure 6 and Figure 7 which are an example to configure a quorum disk witness for the WSFC with two nodes.

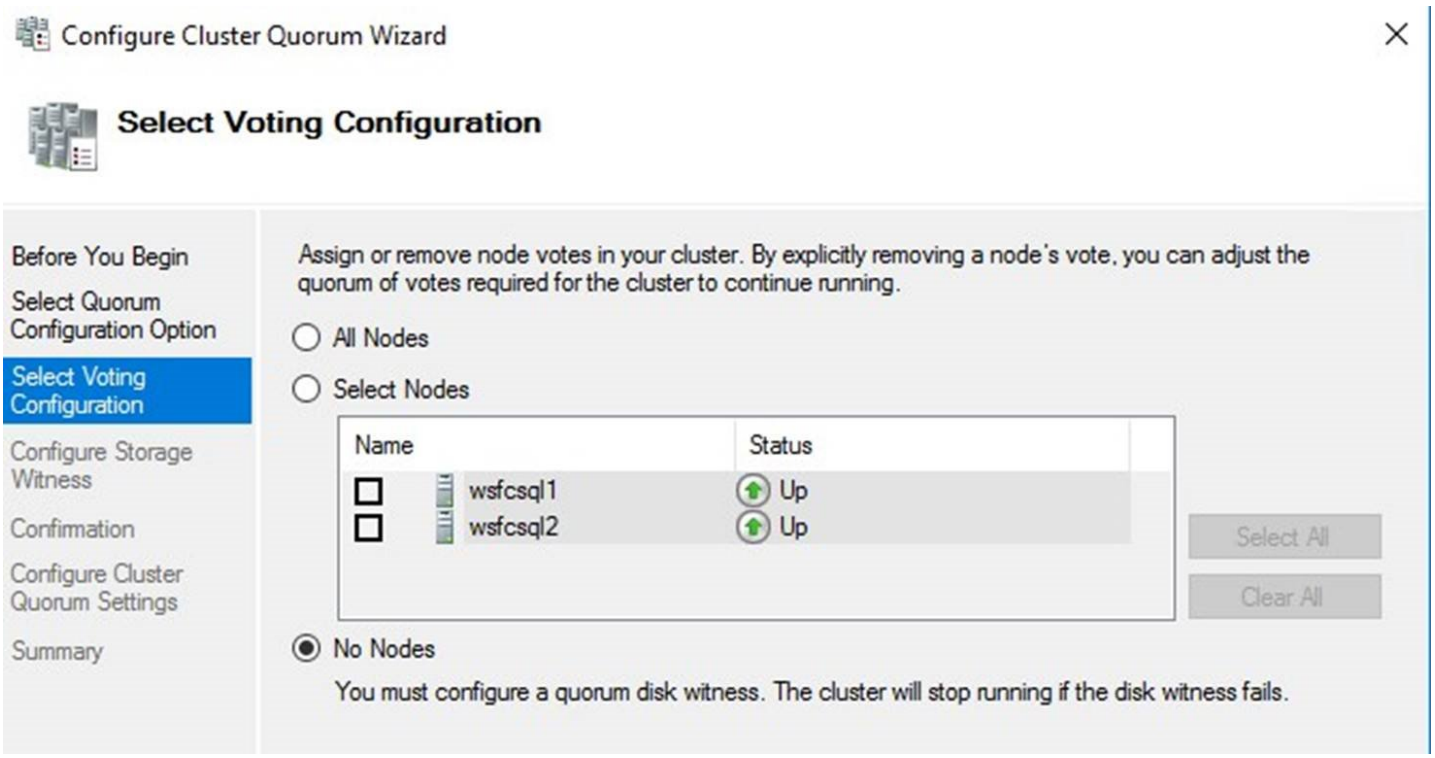

Figure 6. Choose No Nodes (quorum disk witness) for WSFC

Figure 7. Choose Shared Disk for the Cluster

Conclusion

VMware vSAN is optimized for modern All-Flash storage with efficient near-line deduplication, compression, and erasure coding capabilities that lower TCO while delivering incredible performance.

vSAN 6.7 Update 3 and later releases support SCSI-3 Persistent Reservations (SCSI3-PRs) on a virtual disk level required by WSFC to arbitrate an access to a shared disk between nodes, for both standard and stretched cluster. The generic support for WSFC without limitation provides maximum flexibility for you to deploy Windows Failover Clustering on vSAN stretched cluster.

References

- Microsoft Windows Server Failover Clustering on VMware vSphere 6.x: Guidelines for supported configurations
- Architecting Microsoft SQL Server on VMware vSphere
- Configuring a shared disk resource for Windows Server Failover Cluster (WSFC) and migrating SQL Server Failover Cluster Instance (FCI) from SAN (RDMs) to vSAN

About the Author

Tony Wu is a Senior Solutions Architect at VMware in the HCI Business Unit, Solutions Architecture Team with a focus on storage solutions.

The following reviewers also contribute to the paper contents:

- Oleg Ulyanov, Senior Solution Architect in the Cloud Platform Business Unit
- Jase McCarty, Staff Technical Marketing Architect in the HCI Business Unit