OUTDATED

StarWind Virtual SAN®

Windows Geo-Clustering: SQL Server

FEBRUARY 2016

TECHNICAL PAPER

EDWIN SARMIENTO,
Microsoft SQL Server MVP,
Microsoft Certified Master

Trademarks

Changes

The material in this document is for information only and is subject to change without notice. While reasonable efforts have been made in the preparation of this document to assure its accuracy, StarWind Software assumes no liability resulting from errors or omissions in this document, or from the use of the information contained herein. StarWind Software reserves the right to make changes in the product design without reservation and without notification to its users.

Technical Support and Services

If you have questions about installing or using this software, check this and other documents first - you will find answers to most of your questions on the Technical Papers webpage or in StarWind Forum. If you need further assistance, please contact us.

Copyright ©2009-2016 StarWind Software Inc.

In 2016, Gartner named StarWind “Cool Vendor for Compute Platforms”.

Gartner does not endorse any vendor, product or service depicted in its research publications, and does not advise technology users to select only those vendors with the highest ratings or other designation. Gartner research publications consist of the opinions of Gartner's research organization and should not be construed as statements of fact. Gartner disclaims all warranties, expressed or implied, with respect to this research, including any warranties of merchantability or fitness for a particular purpose.

About StarWind

StarWind is a pioneer in virtualization and a company that participated in the development of this technology from its earliest days. Now the company is among the leading vendors of software and hardware hyper-converged solutions. The company’s core product is the years-proven StarWind Virtual SAN, which allows SMB and ROBO to benefit from cost-efficient hyperconverged IT infrastructure. Having earned a reputation of reliability, StarWind created a hardware product line and is actively tapping into hyperconverged and storage appliances market. In 2016, Gartner named StarWind “Cool Vendor for Compute Platforms” following the success and popularity of StarWind HyperConverged Appliance. StarWind partners with world-known companies: Microsoft, VMware, Veeam, Intel, Dell, Mellanox, Citrix, Western Digital, etc.

Contents

Introduction

The goal of this documentation is to provide a high-level overview of geo-clustering using Windows Server 2012 for both high availability and disaster recovery. An understanding of what geo-clustering is and what needs to be considered form the foundation of successful implementations.

This document is intended for technology architects and system administrators who are responsible for architecting, creating and managing IT environments that utilizes Microsoft Windows Server Technologies and familiar with the concepts of Windows Server Failover Cluster (WSFC.)

The Business Need For Geo-Clustering

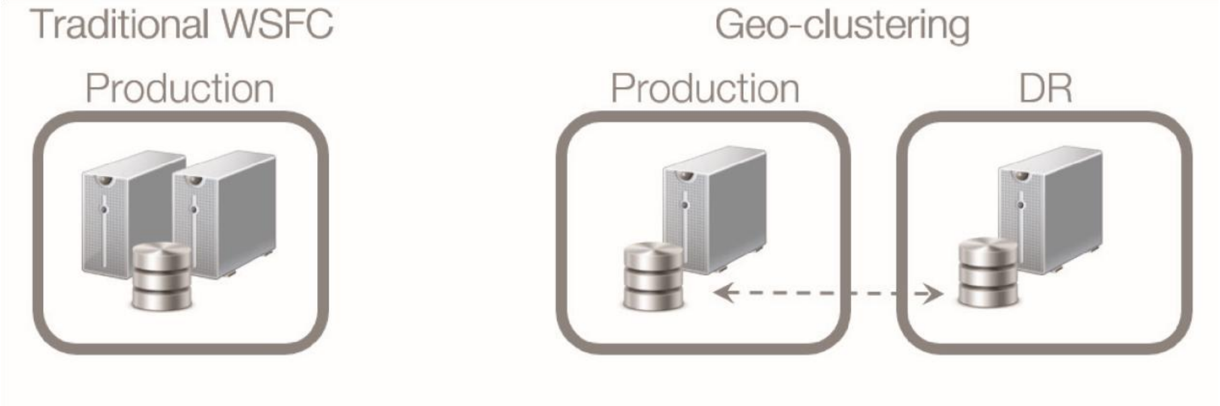
Businesses have demanded high availability in mission-critical systems to make sure business continuity is achieved. In the past, this has been provided for by traditional Microsoft Windows Server Failover Clusters that consists of two or more cluster nodes in the same physical location that have access to shared storage. This is typically referred to as local high availability and provides redundancy and availability should a component in an application architecture fail or experiences degradation in the level of service provided.

Local high availability addresses the business concern regarding recovery time objective (RTO) – the maximum amount of time it takes to restore a business process or an application system after a service disruption to avoid unacceptable loss as defined by the business. While local high availability would be sufficient for most mission-critical business applications, it does not address the risk of severe catastrophic failures like the loss of an entire data center or natural calamities like earthquakes and hurricanes. In the past, this was typically addressed by taking backups from the production data center and restoring them to a remote disaster recovery (DR) data center. While this strategy addresses the issue, it does not meet an organization's applicable recovery point objective (RPO) – the maximum point in time to which a business process or an application system can afford data loss – which also ties back in to the RTO in order to restore a business process or application system in a remote data center.

Geo-clustering - also called multi-subnet clustering, wide area network (WAN) clustering and stretched clustering – addresses both RPO and RTO when a severe catastrophic failure occurred and affected the production data center. It involves having two or more cluster nodes but, unlike the traditional Microsoft Windows Server Failover Clusters that have the nodes in the same physical locations, the nodes are located in different geographical locations forming a single highly available system. Implementing geo-clustering as both a high availability and disaster recovery solution can be a viable strategy to meet your organization's recovery objectives and service level agreements (SLA.) Depending on the desired architecture, several factors need to be considered when planning and implementing a geo-clustering solution. While the concepts described in this whitepaper apply to any cluster-aware applications, the main focus will be on Microsoft SQL Server 2012 and higher versions running as a failover clustered instance on the geo-cluster WSFC.

Single-Location Clustering Versus Geo-Clustering

To better understand geo-clustering concepts, we can compare it to the traditional WSFC that is typically implemented as a high availability solution for applications running on the Microsoft Windows Server platform. A graphic is shown below to demonstrate their similarities and differences.



In a traditional WSFC, all of the nodes in the cluster reside in the same physical location and connected to the same shared storage. Whereas in geo-clustering, the nodes can be in different geographical locations and are connected to different storage subsystem. However, the storage subsystem used by one node is an exact replica of the storage subsystem being used by the other nodes.

This is referred to as **asymmetric storage configuration** in WSFC and was introduced as a deployment option in Windows Server 2008 via a hotfix (for more information about this hotfix, see the Knowledge Base article Hotfix to add support for asymmetric storages to the Failover Cluster Management MMC snap-in for a failover cluster that is running Windows Server 2008 or Windows Server 2008 R2.) Storage or block-level replication can be handled by a third-party application like StarWind Virtual SAN and is outside of the Windows Server operating system. Because of the differences between the two implementations, several factors need to be considered for proper design of the architecture, keeping in mind meeting recovery objectives and SLAs.

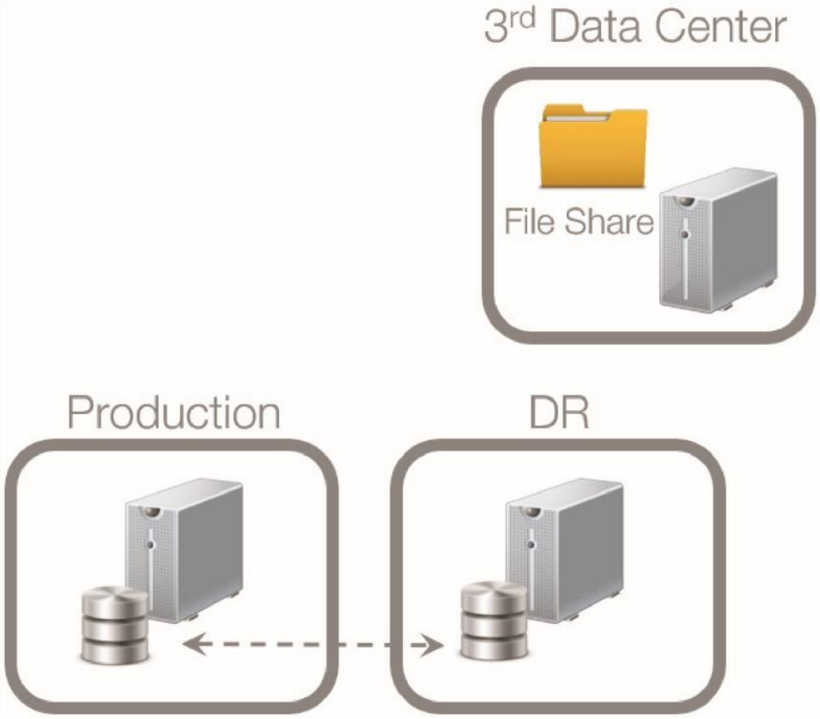
Planning and Considerations

This section highlights planning considerations, requirements, and prerequisites to consider before implementing a geo-cluster solution for high availability and disaster recovery using Windows Server 2012.

Quorum Considerations

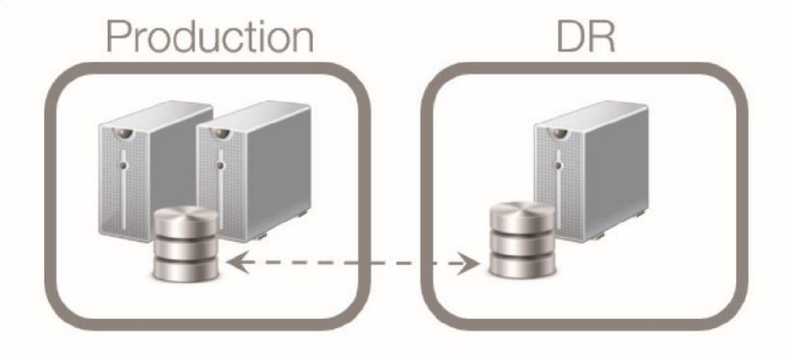
In a traditional WSFC, the concept of a quorum is typically achieved by introducing a shared storage resource in an even number of nodes to achieve majority of votes in order for the cluster to function. Windows Server 2008 introduced four different quorum models: Node Majority, Node and File Share Majority, Node and Disk Majority, No Majority:
Disk Only (for more information about quorum models, see Failover Cluster Step-by-Step Guide: Configuring the Quorum in a Failover Cluster.) Since the nodes in a geo-cluster are not connected to the same storage subsystem, the quorum model used by the traditional WSFC no longer makes sense.

By default, all of the nodes in a cluster has a vote. In a 2-node cluster (or any even number of nodes,) the shared storage resource provides a tie-breaker to achieve majority of votes. For a geo-cluster implementation, the Node and File Share Majority quorum model provides an alternative to the shared storage resource required in a Node and Disk Majority quorum model, where, ideally, the file share resides in a data center that is independent of both the production and DR data centers. A sample diagram is shown below to describe this architecture.



This architecture and quorum model allows loss of network connectivity between the two nodes, and even the file share, without the risk of the entire cluster shutting down. However, it is not enough to have only 2 nodes in the WSFC because it does not provide for local high availability. A false positive may trigger the cluster to failover to the remote data center, making the application connecting to the cluster cross over the wide area network which may introduce latency and performance degradation. Also, it is not cost effective to have an independent data center specifically just for the file share. We want to make sure that the cluster will only failover to the DR data center in the event of a catastrophic failure or when manually initiated as part of a DR exercise.

To achieve both high availability and disaster recovery with a geo-cluster implementation, a traditional 2-node cluster can be implemented in the production data center and a 3rd node with asymmetric storage on the DR data center, as shown in the sample diagram below.



With the recommended architecture, we need to ensure that the quorum of the nodes in the production data center is not compromised by outages in the DR data center or loss of connectivity between the two data centers. This can be done by disabling the voting option of the cluster node in the DR data center and implementing either a Node and Disk Majority or Node and File Share Majority quorum model to still achieve majority of votes. As a best practice, the total number of votes for the WSFC should be an odd number. If there is an even number of voting nodes, consider adjusting the quorum voting by applying the guidelines in the Recommended Adjustments to Quorum Voting section in SQL Server Books Online.

Alternatively, an additional cluster node can also be introduced in the production data center and configure the cluster to use the Node Majority quorum model. The choice of quorum model and configuration should be dictated by the recovery objectives, SLAs and within reasonable cost.

NOTE: The concept of Dynamic Witness is introduced in Windows Server 2012 R2 with the dynamic quorum enabled, by default. This means that the cluster dynamically adjusts the vote of the witness – whether it be a disk or a file share – based on the number of available voting nodes.

Network Considerations

In previous versions of the Windows Server operating system, a stretched virtual local area network or VLAN is required to implement a cluster that spans multiple geographical locations. This made it more challenging for network engineers to design the infrastructure required to support geo-clusters. Starting with Windows Server 2008, a geo-cluster can contain nodes in different network subnets. This means that a cluster resource's client access point remain the same when it fails over to a cluster node residing on a different network subnet (for more information on client access points in a WSFC, see Understanding Access Points (Names and IP Addresses) in a Failover Cluster.) In a SQL Server failover clustered instance, the client access point is the virtual server name assigned to the instance.

Because a Windows Server 2008 and higher geo-cluster can contain nodes in different network subnets, the number of virtual IP addresses for the client access point should be the number of network subnets where the cluster nodes reside. When failing over a clustered resource to a node on a different subnet, the virtual IP address assigned to the client access point will be in the IP address range of that subnet. Depending on the recovery objectives and service level agreements, this may affect client applications' connectivity to the client access point. A client will cache the DNS entry corresponding to the client access point equivalent to its Time-To-Live (TTL) property value.

The default TTL value of a DNS entry is 1200 seconds or 20 minutes. This means that a client will have, at most, 20 minutes before it queries the DNS server again for a new IP address value assigned to the client access point. If this value is beyond your recovery objectives, you can adjust the client access point's TTL value to meet your objectives. Be careful not to set the TTL value way too low that it severely impacts your network infrastructure. For example, if you set the TTL value to 1 minute and you have hundreds, or even thousands, of client applications connecting to the client access point, your network and DNS servers will get overloaded with DNS client requests every minute.

Another consideration when working with geo-cluster nodes is the frequency of the DNS servers to replicate to the other DNS servers. Ideally, each geographical location should have its own DNS server and Active Directory domain controller. In an Active Directory-integrated DNS zone, DNS server replication frequency is the same as that of Active Directory replication. When a client access point is created, a corresponding virtual computer object in Active Directory is created. This object gets replicated to the other domain controllers in the network and is the basis of the DNS entry that also gets replicated to the other DNS servers. Also, the nodes in the cluster need to be joined in the same Active Directory domain. Talk to your network administrators for the configured value of your network's Active Directory replication frequency and what other applications are dependent on it.

NOTE: Windows Server 2012 R2 introduced the concept of Active Directory-detached cluster. This allows administrators to deploy a WSFC and a SQL Server failover clustered instance without dependencies in Active Directory for the client access point. However, for a SQL Server failover clustered instance, it is recommended to have client applications authenticate using SQL Server authentication instead of Kerberos.

Still another consideration when working with geo-cluster nodes is the communication between cluster nodes or what is commonly known as "heartbeat." There are two major settings that affect heartbeat. First, the frequency at which the nodes send signals to the other nodes in the cluster (subnet delays) and, second, the number of heartbeats that a node can miss before the cluster initiates a failover (subnet threshold). In a traditional WSFC, these settings were rarely modified because the delay and threshold values are tolerable enough for the cluster to handle without initiating a false failover. However, in a geo-cluster environment, when the cluster nodes are separated by geographical locations, inter-node communication may take longer and could possibly miss heartbeats. The table below outlines the default values for cluster subnet delays and thresholds, respectively.

Heartbeat Parameter	Default value
SameSubnetDelay	1000 (in milliseconds)
SameSubnetThreshold	5 heartbeats
CrossSubnetDelay	1000 (in milliseconds)
CrossSubnetThreshold	5 heartbeats

The cluster heartbeat configuration settings are considered advanced settings and are only exposed via the Windows PowerShell Failover Cluster Modules. Adjust these cluster parameters accordingly and monitor appropriately. Discuss them with your network engineers to understand potential issues with network and how they can impact your cluster environment.

Client Connectivity

In a traditional WSFC, client applications need not worry about the multiple virtual IP addresses assigned to client access points because they only have one. In a geo-cluster, there will be multiple virtual IP addresses assigned to a client access point. This means that when a client queries the DNS for the client access point, it will try to connect to the first IP address in the list, depending on the priority of the DNS servers assigned to it. Legacy client applications do not have reconnection logic to try all of the virtual IP addresses assigned to the client access point and, therefore, will not be able to establish connectivity to the clustered resource. This translates to system unavailability from the client application's point of view even when the clustered resource is online and available. In order for client applications to be able to handle multiple virtual IP addresses for a client access point, they need to either be using at least

- the SQL Server Native Client 11.0
  the Data Provider for SQL Server in .NET Framework 4.02
- the Microsoft JDBC Driver 4.0 for SQL Server

A new connection string attribute named MultiSubnetFailover is made available to allow client applications to try all the virtual IP addresses assigned for the client access point and connects to the first one that responds. This improves a client application's connectivity after a failover and, therefore, reduce overall system downtime. Legacy client applications need to update their client libraries to support the MultiSubnetFailover attribute.

Storage Considerations

Since there is no shared storage for all of the nodes in a geo-cluster, data must be replicated between the storage subsystems used by the nodes at each of the physical locations. Data replication at the storage or block-level can be handled by a third-party application like StarWind Virtual SAN and is outside of the Windows Server operating system. While the local nodes need access to the storage subsystem in their own location, there is also connectivity between the storage units used to link them together.

This connectivity allows the cluster nodes to take exclusive ownership of the storage during failover and also improves availability and I/O performance. This concept of storage or block-level replication is the backbone of geo-clustering.

Since the storage is replicated between the different storage subsystems in the cluster, it is important to understand the capabilities of the storage subsystem to make sure that they meet your recovery objectives and service level agreement. The amount of data loss that can be occurred during a failover to a remote data center is dependent on the capabilities of the storage subsystem and the network bandwidth between the nodes of the cluster. There are two types of data replication – synchronous and asynchronous – both of which have direct impacts to the amount of data loss and I/O performance. With synchronous storage replication, data that has been changed on one storage subsystem is not considered completed unless the same change has been applied to all of the storage subsystems participating in a replication topology.

This type of storage replication holds the promise of no data loss when a cluster failover occurs on a remote data center. However, since the I/O performance of the storage subsystem is dependent on the network bandwidth between the cluster nodes, synchronous storage replication is ideal for high bandwidth, low latency network connections where the cluster nodes are only separated by shorter distances. With asynchronous storage replication, data that has been changed on one storage subsystem does not necessarily have to be applied immediately to the other storage subsystems in a replication topology in order to be considered completed.

This type of storage replication holds the promise of improved I/O performance but at the cost of potential data loss, thus, impacting the recovery objectives. Be sure to perform extensive failover testing with the appropriate workload to verify if asynchronous replication can meet your allowable data loss in the event of a disaster.

Summary

Geo-clustering on Windows Server 2012 provides customers with flexible design choices for building both a high availability and disaster recovery solution for mission critical applications like SQL Server. Given the different considerations when implementing a successful geo-cluster solution, it is essential to work not just with the database team but to collaborate closely with the other IT teams involved such as the network team and the server administration team.

The implementation of a geo-cluster solution requires additional complexity in terms of architecture design, implementation and operations and comes at an additional cost. It is important to properly evaluate the overall benefits of implementing a geo-cluster solution, keeping in mind that the ultimate goal of implementing such a solution is to meet organizational recovery objectives and service level agreements.

Contacts

US Headquarters	EMEA and APAC
☎ 1-617-449-7717 📠 1-617-507-5845	☎ +44 20 3769 1857 (UK) ☎ +49 302 1788 849 (Germany) ☎ +33 097 7197 857 (France) ☎ +7 495 975 94 39 (Russian Federation and CIS) 🗨 1-866-790-2646

Customer Support Portal:	https://www.starwind.com/support
Support Forum:	https://www.starwind.com/forums
Sales:	sales@starwind.com
General Information:	info@starwind.com

StarWind Software, Inc. 35 Village Rd., Suite 100, Middleton, MA 01949