

SQLskills Immersion Event

IE3: High Availability & Disaster Recovery

Module 7: Failover Clustering

Jonathan Kehayias
Jonathan@SQLskills.com



Overview

- Clustering concepts
- Planning for failover clustering
- How failover works
- Installation and patching SQL Server
- Troubleshooting cluster Issues



2

© SQLskills, All rights reserved.
<http://www.SQLskills.com>



Clustering Concepts

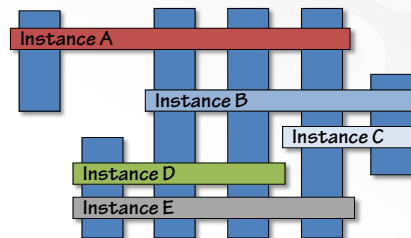
What is Failover Clustering?

- High-availability solution built on top of Windows Clustering that provides automatic detection and failover at the instance level
- Only one node can own the instance or service and its dependent resources such as shared disks, IP address, and MSDTC (if configured as a cluster resource)
- Standard Edition supports 2 nodes, Enterprise and Datacenter support the OS-Edition maximum for nodes
 - Up to 16 nodes can participate in a failover cluster
- Only maintains a single copy of the databases on the shared storage
 - Single point of failure

Clustering Concepts

Multi-Instance and Multi-Node Clusters

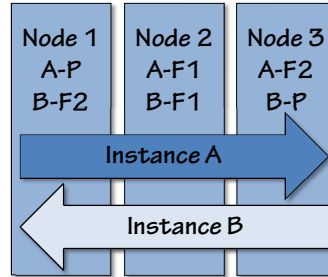
- Multi-instance clusters have multiple instances of SQL Server installed
 - Each instance requires its own cluster resource group
 - Each instance should be assigned a dedicated TCP port to avoid firewall and client connectivity issues from dynamic ports if the Browser service is disabled
 - Nodes must be able to handle the workload of all instances that can failover to the node
- Multi-node clusters provide added flexibility to the configuration
 - N+1: N active nodes with 1 passive standby server
 - N+K: N active nodes with K passive standby servers



Clustering Concepts

Multi-Instance Preferred Ownership

- Multi-instance clusters have should have preferred owners and failure paths configured for the instances to ensure that SLAs continue to be met in the event of a failure
- Example three-node cluster with two instances of SQL Server installed
 - Instance A has Node 1 as its preferred owner and a defined failure path to Node 2 and then Node 3
 - Instance B has Node 3 as its preferred owner and a defined failure path to Node 2 and then Node 1



Clustering Concepts

Understanding Quorum

- Quorum determines the number of failures that a cluster can sustain and still remain online
- Quorum exists to handle scenarios where communication between cluster nodes has failed to prevent multiple nodes from trying to host the same resource group(s) simultaneously resulting in a 'split-brain'
- Voting towards quorum
 - The cluster has quorum if more than half of the voters are online and communicating with each other
 - Each cluster node has 1 vote
 - A disk or file share witness can be configured for 1 vote
 - Always configure quorum so that an odd number of votes exists in the cluster
 - 5-node cluster needs 3 nodes online to run, so does a 4-node cluster without a disk or file share witness

Clustering Concepts

Voting Towards Quorum Windows 2008R2 and earlier

- The cluster has quorum if more than half of the voters are online and communicating with each other
- Each cluster node has 1 vote
 - Based on its NodeWeight configuration in Windows Server 2008 onwards
 - <http://support.microsoft.com/kb/2494036>
- A disk or file share witness can be configured for 1 vote
- Always configure quorum so that an odd number of votes exists in the cluster
 - 5-node cluster needs 3 nodes online to run, so does a 4-node cluster without a disk or file share witness
 - Plan for servers participating in the cluster that are in a remote data center when working with geo-clusters

Clustering Concepts

Voting Towards Quorum Windows 2012

- **Dynamic Quorum**
 - Provides the ability of the cluster to recalculate quorum on the fly to maintain a working cluster
 - Provides the ability to continue to run a cluster even if less than 50% of the nodes remain available
 - Allows the cluster to be reduced down to the last node (known as last man standing) and still maintain quorum
- **Dynamic Quorum requires:**
 - The cluster has already achieved quorum
 - Sequential failures of nodes occurs
- **If multiple nodes in a cluster go down simultaneously, dynamic quorum will not recalculate the number of votes to maintain quorum**
 - When this occurs, a regroup must occur to determine quorum can be maintained, and then dynamic quorum will resume when a subsequent node failure occurs

Clustering Concepts

Quorum Types

- **Node Majority**
 - Default for odd number of nodes
- **Node and Disk Majority**
 - Recommended for even number of nodes
 - Requires an additional small clustered disk that can failover between the nodes as a witness and additional voter for quorum
- **Node and File Share Majority**
 - Required for multi-site or geo-clustering
 - Requires a file share in the same AD forest as the cluster nodes
- **Legacy: Disk Only**
 - Is a single point of failure for the cluster quorum and is not recommended for use
 - Maintained only for backwards-compatibility reasons

Clustering Concepts

Geo-Clustering

- Eliminates the SAN as a single point of failure
- Requires specific hardware from the Geographic Cluster Hardware Compatibility List for the solution to be supported (KB280743)
- Requires a stretched VLAN in SQL Server 2008R2 and prior
 - Public and private networks must appear non-routed
 - Round trip latency must be under 500ms
- SQL Server 2012 allows OR dependency on IP Address for the cluster name resource allowing multiple subnets to be used
- Requires third party SAN replication technology
- SQL install must be started from the command line to skip Cluster_VerifyForErrors rule check
 - Setup /SkipRules=Cluster_VerifyForErrors /Action=InstallFailoverCluster

Planning for Failover Clustering

Server Hardware

- Traditionally failover clustering required identical servers from the Hardware Compatibility List (HCL)
- Windows Server 2008 introduced the Cluster Validation Tool (CVT)
 - Any hardware that passes the configuration tests run by the CVT can be used for failover clustering
 - Heterogeneous cluster nodes are supported as long as the hardware and OS architecture (x86/x64/IA64) matches
- Separate public and private networks are no longer required, but still recommended as a best practice
 - The cluster heartbeat can now communicate across any network available to the cluster

Planning for Failover Clustering

Multi-Instance Memory Management

- When multiple instances can failover to a single cluster node memory resources will be constrained during the failure
 - Instance memory = (memory for thread stack) + (memory for buffer pool) + (memory for plan cache) + (memory for multipage allocators)
 - Memory for thread stack = (max worker threads) *(stack size)
 - x86 = 512K; x64 = 2MB; ia64 = 4MB
- Buffer pool memory usage is limited by the 'max server memory' sp_configure option and the buffer pool will grow/shrink based on OS memory availability

Planning for Failover Clustering

Multi-Instance Memory Management

- The 'min server memory' sp_configure option can be used to guarantee a minimum amount of memory to the buffer pool after memory ramp up
- For multi-instance configurations 'min server memory' should be set for each of the instances based on their priority/importance to guarantee a minimum amount of buffer pool memory
 - The aggregated total of 'min server memory' values should always be lower than the total server memory installed to prevent memory pressure and paging problems especially when using Lock Pages in Memory
 - Memory\Available Mbytes > 150-300MB
- Aaron Bertrand's blog post on automating reconfiguration:
 - <http://sqlskills.com/help/AutoMemConfig>

Planning for Failover Clustering

Multi-Instance Memory: Lock Pages in Memory

- Lock Pages in Memory allows SQL Server to allocate memory using AllocateUserPhysicalPages() instead of VirtualAlloc()
- Required for AWE on 32-bit instances
- Prevents paging of the buffer pool on 64-bit instances
 - Does not require configuration of "AWE enabled" in SQL Server
 - Trace flag -T845 required for Standard Edition instances
- Detected at startup only

Planning for Failover Clustering

Multi-Instance Memory: Large Pages

- For 32-bit the OS page size is 4KB
- 64-bit can use large pages (2-16 MB)
 - Helps avoid translation look-aside buffer (TLB) misses
- Requires
 - Enterprise Edition
 - Locked Pages In Memory
 - Trace flag -T834
 - <http://support.microsoft.com/kb/920093>
- All memory allocated at startup, and must be contiguous
 - For this reason, large pages should not be enabled for multi-instance configurations where two instances can run on a single node at the same time

Planning for Failover Clustering

Storage

- The shared disk arrays must be presented to each of the cluster nodes and pass the CVT Disk tests before creating failover cluster
- Determines the disks available for clustering, validates arbitration, failover ability, and multipath configuration
- SCSI-3 Persistent Reservation Requirement
 - Avoid SCSI bus resets, much less disruptive than the older reserve/release commands
- All tests must pass in order for the cluster configuration to be supported for SQL Server failover clustering

Planning for Failover Clustering

SMB Storage in SQL Server 2012

- Skip Cluster Validation Disk checks
- Use Node and File Share quorum configuration if an even number of nodes are participating in the cluster
 - Requires that the Virtual Computer Object (VCO) for the cluster have Full Control over the share
- During SQL Server cluster installation specify UNC paths for the file location configuration
 - Requires that the service account for the Database Engine has Full Control permission over the share locations
- Blog post:
 - <http://sqlskills.com/help/2012onSMB>

Planning for Failover Clustering

Storage: Mount Points or Drive Letters

- Mount points provide a way to exceed the 26 volume limit imposed by using drive letters
 - Additional volumes can be mounted as folders under a drive letter
 - May be problematic for space monitoring
- Each instance requires at least one disk with a drive letter assigned to it
- SMB storage allows up to 50 instances to be configured in a cluster in SQL Server 2012
- SQL Server 2012 supports tempdb on local SSD storage

Planning for Failover Clustering

Storage: Virtual Machines

- Failover clustering is supported using VM guests in SQL Server 2005+ on Windows Server 2008+
 - Cluster must pass the Cluster Validation Tests in Windows Server 2008 to be supported
 - VMware requires fiber channel SAN connection for RDM shared disks
 - Limited to a single path currently, which imposes performance limitations that have to be considered
 - Support for in-guest iSCSI initiator with MPIO
- Reference:
 - <http://sqlskills.com/help/VMClusterRegs>

Planning for Failover Clustering

Active Directory/DNS

- Active Directory is required to configure failover clustering
- For best security a separate domain service account should be used for each service with the minimum permissions required
- DNS name resolution must be working properly using fully-qualified domain names (FQDN)
- If using multiple NICs, ensure that the domain-name suffix is configured appropriately

Planning for Failover Clustering

Networking

- Separate private network is not required for heartbeat using Windows Server 2008+ but is still recommended
- Redundant NICs connected to separate switches should be used to remove single point of failure
 - For 2 node clusters a x-over cable can be used for the private network
- NIC Teaming can be used to overcome network bottleneck if necessary
 - Allows multiple network cards to share a common IP address and balance load across their connections
- Configure all adapters for full-duplex operation

Planning for Failover Clustering

MSDTC

- MSDTC is not required in failover clustering
- If you need MSDTC for distributed transactions, it must be configured as a cluster resource using the Cluster Administrator
- MSDTC can be installed after SQL Server as a resource in the instance cluster group
 - Allows multi-instance clusters to have separate DTCs per instance
 - Ensures that the instance and DTC resource run on the same cluster node and failover together

Blog by Cindy Gross:

- <http://sqlskills.com/help/msdtdc>

SQL Server Installation

Adding a Node

- Setup must be run on each node that an instance is able to failover to
- Install the SQL Server prerequisites on the node
- SQL Server 2008 (only, not 2008 R2 or 2012) requires slipstream installation on Windows Server 2008 R2
 - Choose “Add node to SQL Server failover cluster” as the setup option
 - On the Cluster Node Configuration page select the name of the SQL Server failover cluster instance to add in the drop down
 - Configure the appropriate service accounts
 - Restart the server if required when installation completes

SQL Server Installation

Removing a Node

- Prior to evicting a node from the cluster using the Cluster Administrator run SQL Server setup to remove the SQL instance(s) from the node
 - If the node is being formatted or repurposed this is not required
 - Run SQL Server setup
 - On the Cluster Node Configuration page select the name of the SQL Server failover cluster instance to remove in the drop down
 - Restart the server if required when installation completes
- Evict the node from MSCS using the Cluster Administrator

How Failover Works

Failure Detection

- Node failure is detected by MSCS when a node misses 6 consecutive heartbeats
- Resource failure is detected through IsAlive and LooksAlive checks provided by the Resource DLL
 - SQL Server performs the LooksAlive check every 5 seconds and validates that the SQL Server service status matches the status reported by the Service Control Manager (SCM)
 - SQL Server performs the IsAlive check every 60 seconds by executing the command `SELECT @@SERVERNAME` against the instance
- Intervals for the IsAlive and LooksAlive checks and the number of missed heartbeats to trigger failure can be configured using the Failover Cluster Manager

How Failover Works

Instance Failover

- When instance failover occurs the resource group for the instance is taken offline on the current owning node
- The Failover Manager performs arbitration to locate a new owner for the group and brings the resources online on the new node in dependency order
- When the SQL Server service starts up instance recovery begins
 - The SQL Resource DLL is up once the master database completes recovery
 - User databases become available when REDO completes for Enterprise Edition or after UNDO completes for Standard Edition
 - Called 'Fast Recovery'

How Failover Works

Instance Failover and SQL Agent

- By default SQL Agent failures will cause a failover of the SQL Server Cluster Resource Group
- If SQL Agent is not required in the environment, or failover should not occur if the SQL Agent service goes offline under normal operation, this should be changed for the environment

How Failover Works

Failback

- Cluster resource groups can be assigned a 'preferred owner'
- When the preferred owner node comes back online, the cluster resource group can automatically failback to the node based on its configuration
- Failback is not required and should only be configured in multi-node/multi-instance clusters where specific instances should ideally be running on specific nodes
- Cluster resource groups should be configured to perform automatic failbacks outside of normal operating hours only
- Failback is the same as a failover and all client application connections are broken impacting end users

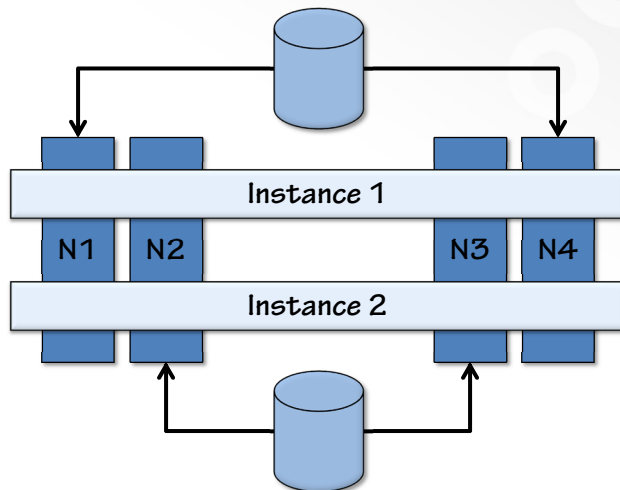
SQL Server Installation

Non-Cluster Aware Components

- **Reporting Services**
 - Should be setup in a scale-out solution with network load balancing instead of failover clustering
- **Integration Services**
 - Possible to cluster manually
<http://msdn.microsoft.com/en-us/library/ms345193.aspx>
- **Notification Services (2005)**

Patching a Clustered Instance

1. Remove N3 & N4 from possible owners
2. Patch N3 & N4
3. Add N3 & N4 to possible owners, remove N1 and N2 from possible owners
4. Failover to N3 & N4
5. Patch N1 & N2
6. Add N1 & N2 back to possible owners



Troubleshooting Cluster Issues

- Windows event logs
- Cluster log
 - cluster.exe /Cluster:ClusterName log /gen /copy "c:\\"
- SQL Server error log
- SQL Server setup log
- Connection failures
 - Ensure that each instance is configured to use a fixed TCP/IP port in the cluster
 - Ensure that the appropriate firewall rules have been implemented in each of the cluster nodes for the fixed ports
 - Ensure there are no IP conflicts in the network

Troubleshooting Clusters Issues

Recovering from Cluster Hardware Failure

- Evict the failed node from the cluster using the Cluster Administrator
- Verify that the node has been removed from the cluster
- Repair or replace the failed server hardware and reinstall Windows Server 2008
- Add the node to the cluster using the Cluster Administrator
 - Plan a downtime to re-run Cluster Validation Tests
- Run SQL Server setup on the node

SQL Server 2012 Flexible Failover

- User sets new cluster properties for HealthCheckTimeout and FailureConditionLevel
- IsAlive/LooksAlive result based on output from sp_server_diagnostics and FailureConditionLevel

Level	Failover Type	Description	Example
0	No automatic failover or restart	Never Fail, Log Only	
1	Failover or restart on SQL Server down	Server Down	Instance Not Starting
2	Failover or restart on SQL Server unresponsive	Server Hang	17884 Deadlock
3	Failover or restart on critical SQL Server errors	System Unhealthy	Stack Dumps Occurring
4	Failover or restart on moderate SQL Server errors	Resource Unhealthy	Low on Memory
5	Failover or restart on any qualified failure	QP Unhealthy	Queries Stalled

Review

- Clustering concepts
- Planning for failover clustering
- How failover works
- Installation and patching SQL Server
- Troubleshooting cluster issues

Resources

- List of known issues when you install SQL Server on Windows 7 or on Windows Server 2008 R2
 - <http://support.microsoft.com/kb/955725>
- How to update or slipstream an installation of SQL Server 2008
 - <http://support.microsoft.com/kb/955392>



35

© SQLskills, All rights reserved.
<http://www.SQLskills.com>

Questions?

