

What's new in Failover Clustering in Windows Server 2016

Table of Contents

Cluster OS Rolling upgrade.....	3
Cloud Witness	4
Workgroup and Multi-Domain Clusters	7
Virtual Machine Load Balancing / Node Fairness	7
Virtual machine start ordering	9
VM Compute/Storage Resiliency	11
Site-Aware/Stretched Clusters and Storage Replica	13
Miscellaneous	15

Finally, I'd like to review what's new in failover clustering in Windows Server 2016. Actually, I wrote this article a couple of months ago for Russian official Microsoft blog so if you are Russian you can go to [this resource](#) to read it in your native language.

Also, I described some of the new features before RTM-version (when only TPs were available) and almost all of them can be applied to Windows Server 2016 as well. It means there are no significant changes in RTM for them. I'll provide a short description of such features and links to my previous posts with a detailed information. And yes, of course, completely new functionality (Load Balancing, for instance) will also be described here.

Cluster OS Rolling upgrade

Cluster migration is usually a headache for administrators. It could be a reason of huge downtime (because we need to evict some nodes from old cluster, build the new one based on these nodes or new hardware and migrate roles from source cluster. So, in the case of overcommitment we won't have enough resources to run migrated VMs). It's critical for CSPs and other customers that have implemented SLA policy.

Windows Server 2016 fixes this by adding possibility to place Windows Server 2012 R2 and Windows Server 2016 nodes in the same cluster during upgrade/migration phase.

The new feature named as Cluster Rolling Upgrade (CRU) significantly simplifies overall process and allows us to successively upgrade existed nodes without destroying cluster. It helps to reduce downtime and any required costs (hardware, staff time and etc.)



The full list of CRU benefits is listed below:

- Hyper-V virtual machine and Scale-out File Server workloads can be upgraded ONLY from Windows Server 2012 R2 to Windows Server 2016 without any downtime. Other cluster workloads will be unavailable during the time it takes to failover (for example, SQL Server with AlwaysOn FCI ~ 5 minutes of downtime)

- It does not require any additional hardware (for example, you evicted 1 node of 4. The rest 3 nodes are online and they must have resources for workloads live migrated from evicted node. In this case zero-downtime is predicted)
- The cluster does not need to be stopped or restarted.
- In-Place OS upgrading is supported BUT Clean OS install is highly recommended. Use In-Place upgrading carefully and always check logs/services before adding node back to cluster.
- A new cluster is not required. In addition, existing cluster objects stored in Active Directory are used.
- The upgrade process is reversible until the customer crosses the “point-of-no-return”, when all cluster nodes are running Windows Server Technical Preview, and when the Update-ClusterFunctionalLevel PowerShell cmdlet is run.
- The cluster can support patching and maintenance operations while running in the mixed-OS mode.
- CRU is supported by VMM 2016 and can be automated through PowerShell/WMI

To get more details [read my previous post](#) that shows CRU in action (it’s been written for Technical Preview but can still be used with RTM)

Hint: get list of supported VM’s version by host (*Get-VMHostSupportedVersion*).

```

PS C:\Windows\system32> Get-VMHostSupportedVersion

Name                                     Version IsDefault
----
Microsoft Windows 8.1/Server 2012 R2    5.0     False
Microsoft Windows 10 1507/Server 2016 Technical Preview 3 6.2     False
Microsoft Windows 10 1511/Server 2016 Technical Preview 4 7.0     False
Microsoft Windows Server 2016 Technical Preview 5 7.1     False
Microsoft Windows 10 Anniversary Update/Server 2016 8.0     True
Prerelease                               254.0   False
Experimental                             255.0   False

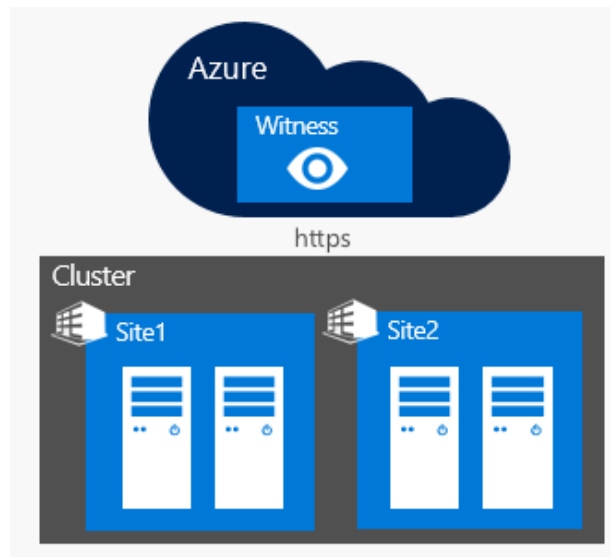
```

Cloud Witness

Failover cluster in Windows Server 2012 R2 can be deployed with an external disk or file share witness which must be available for each cluster nodes and it’s needed as a source of extra vote. As you may know, witness is highly recommended (I’d say it’s required!) for Windows Server 2012 R2 cluster regardless of a number nodes in it (dynamic quorum automatically decides when to use witness).

In Windows Server 2016 a new witness type has been introduced – Cloud Witness. Yes, it’s Azure-based and it’s specially created for DR-scenarios, Workgroup/Multi-Domain cluster (will be described later), guest clusters and clusters without shared storage between nodes.

Cloud Witness uses Azure Storage resources (Azure Blob Storage through HTTPS protocol. HTTPS port should be opened on all cluster nodes) for read/write operations. Same storage account can be used for different clusters because Azure creates a blob-file generated for each cluster with unique IDs. These blob-files are kept in msft-cloud-witness container and require just KBs of storage. So, costs are minimal and Cloud Witness can be simply used as a third site (“arbitration”) in stretched clusters and DR solutions.

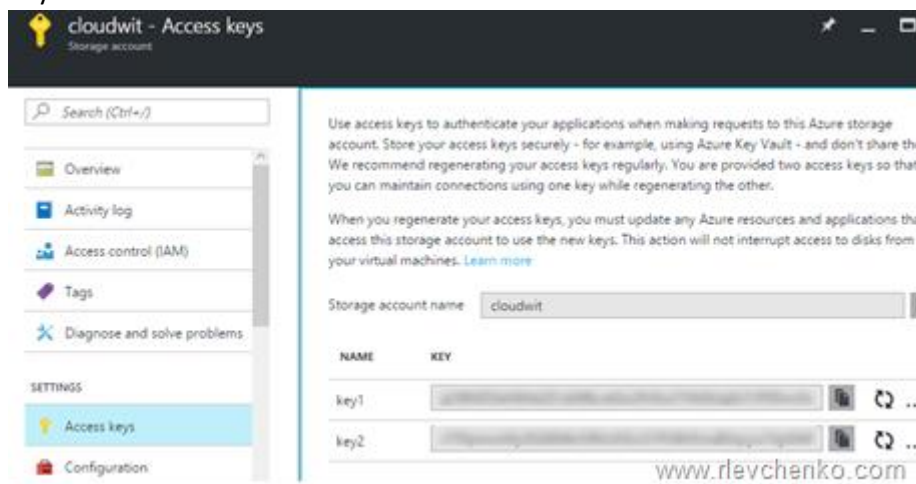


Cloud Witness scenarios:

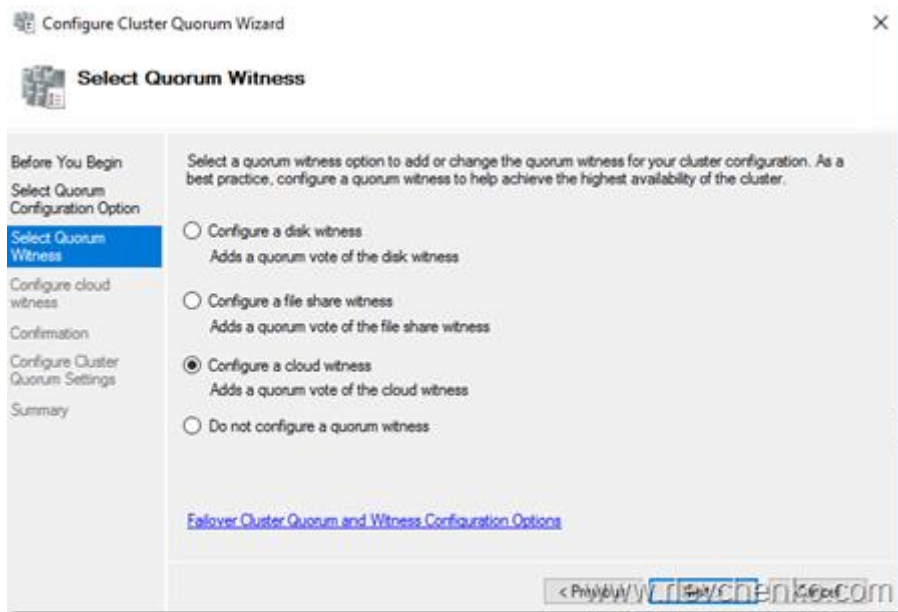
- Multi-Site clusters
- Clusters without shared storage (Exchange DAG, SQL Always-On and etc.)
- Guests clusters running on Azure and On-Premises
- Storage Cluster with or without shared storage (SOFs)
- WorkGroup and Multi-Domain Clusters (new in WS2016. It'll be described later)

How to create and add cloud witness to cluster

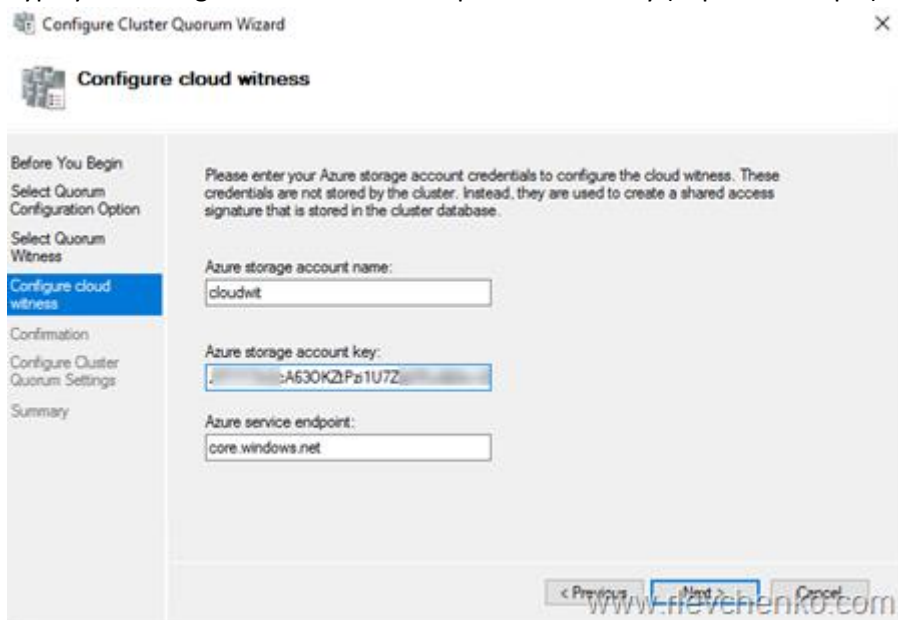
- 1) Create a new Azure Storage Account (Locally-redundant storage) and copy one of the shared key



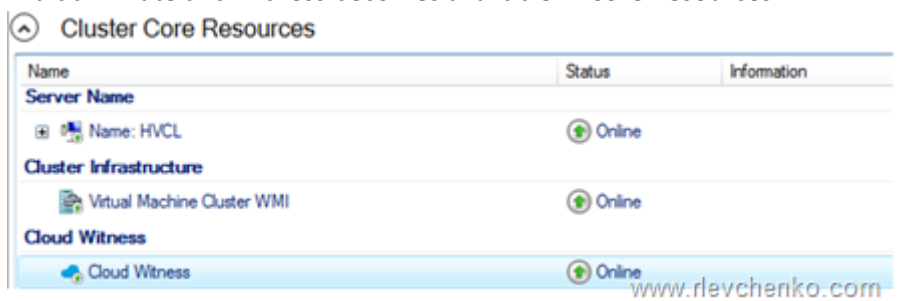
- 2) Run Quorum Configuration on your cluster and choose “Select the Quorum Witness – Configure a Cloud Witness”



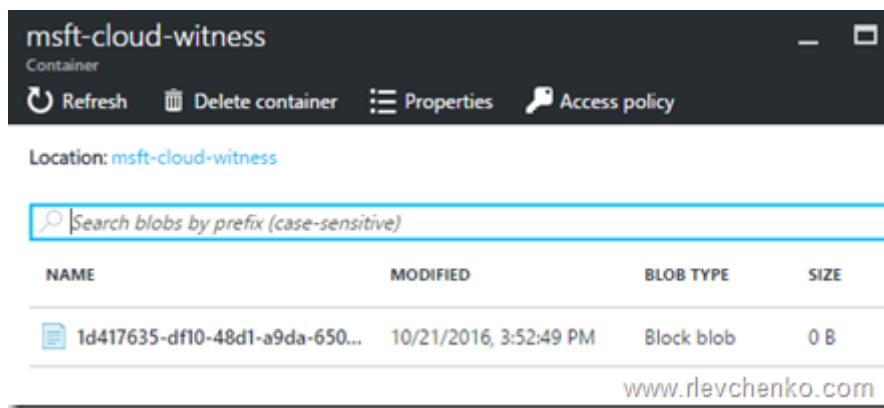
- 3) Type your storage account name and paste shared key (copied on step 1)



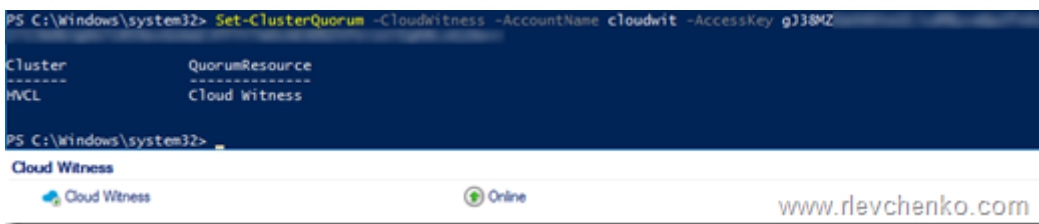
- 4) Wait a minute and witness becomes available in Core Resources



- 5) Here is a blob-file in Azure created for added cluster



Powershell oneliner:



Workgroup and Multi-Domain Clusters

In Windows Server 2012/2012 R2 and previous versions, there is one global requirement for cluster: single-domain joined nodes. [Active Directory Detached cluster](#), which was introduced in 2012 R2, has the same requirement and does not provide advanced flexibility either. Beginning from Windows Server 2016 you have additional options: create cluster with nodes in **Workgroup** and create cluster in **multi-domain** environment.

More details are in my [previous post](#)

Virtual Machine Load Balancing / Node Fairness

VMM historically provides advanced options to efficiently manage cluster resources. Dynamic optimization available in VMM automatically load balances resources between nodes. The trimmed-down version of that has been introduced in Windows Server 2016. VM Load Balancing or Node Fairness moves resources (live migration) every 30 minutes to other nodes in cluster based on configured heuristics:

- Current % of memory usage
- Average CPU load (last 5 mins)

WSFC uses AutoBalancerLevel and AutoBalancerMode for making a decision when to move resources:

```
get-cluster /fl *autobalancer*
```

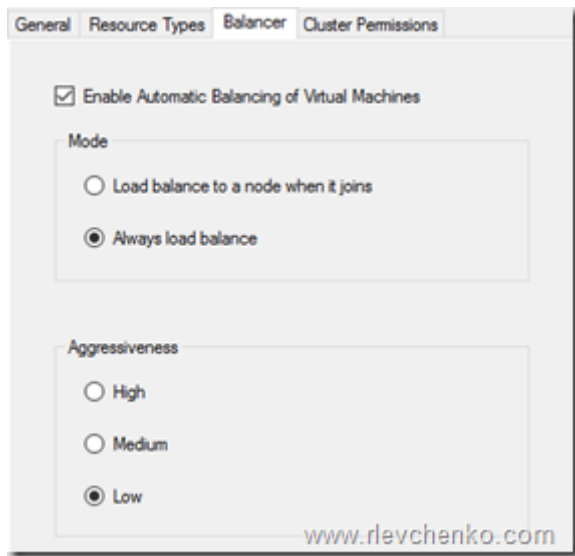
```
AutoBalancerMode : 2
```

What's new in Failover Clustering | Roman Levchenko | rlevchenko.com

AutoBalancerLevel : 1

AutoBalancerLevel	Aggressiveness	When to move?
1 (default)	Low	When host load is more than 80%
2	Medium	When host load is more than 70%
3	High	When host load is more than 60%

You can also set these settings in GUI (cluadmin.msc). By default, WSFC (cluster) uses Low level of aggressiveness and tries to always load balance.



I'm using the following values for demo:

AutoBalancerLevel: 2

```
(Get-Cluster).AutoBalancerLevel = 2
```

AutoBalancerMode:2

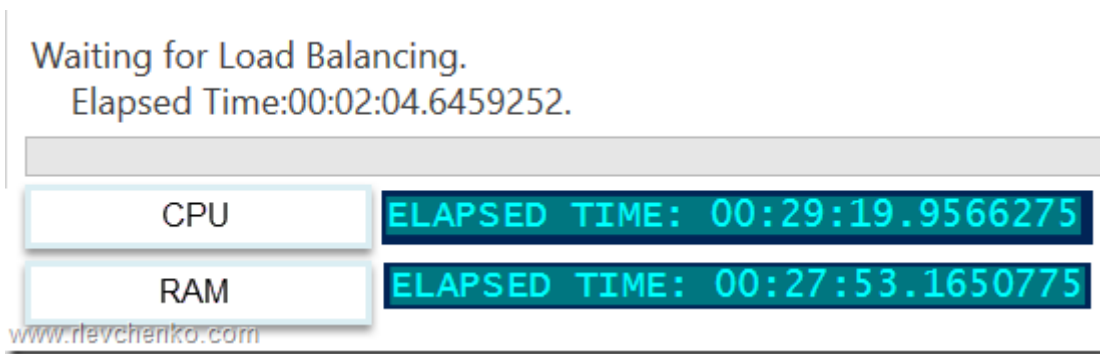
```
(Get-Cluster).AutoBalancerMode = 2
```

And verifying two scenarios:

- 1) High CPU load on my host (about 88%)
- 2) High RAM usage (about 77%).

As we use medium aggressiveness (70%) virtual machines will be moved from that busy host to another. My script waits while live migration starts and then outputs the elapsed time

When my host had a CPU load (~88%), VM balancer moved more than 1 VM. In case of RAM usage (~77%) - 1 VM was moved. All live migrations started in 30-minute interval. Load Balancing works perfectly.

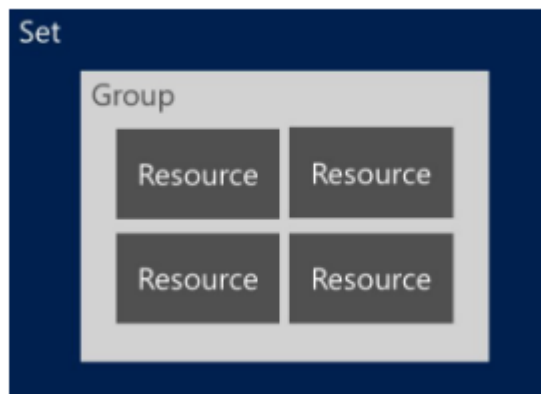


If you manage your nodes from VMM, Dynamic optimization is a preferred method for load balancing (it has more advanced settings and more powerful). When it's enabled, **VMM automatically disables VM Load Balancing feature in Windows Server.**

Virtual machine start ordering

In the previous versions start ordering is addressed by configuring VM's priority. We have Low, Medium, High priority levels and this ability helps to identify which resources should be started before other dependent (for example, Active Directory or SQL Server). Unfortunately, there is a one big limitation: no cross-node orchestration and VMs are considered to be running once it reaches online state.

Windows Server 2016 changes this behavior by adding **VM Start Ordering** which allows you to define dependencies between VMs and group VMs using Set (see thi pic below). Originally added to WS to orchestrate VMs start ordering but can be useful for any applications that represented as cluster groups.



Let's review some examples:

1 VM (Clu-VM02) runs application that dependent from Active Directory, running on the VM called as Clu-VM-01. But virtual machine Clu-VM03 depends from application on Clu-VM02.

To solve this, I'll create the new Set using PowerShell:

```
PS C:\> Get-Command *clustergroup* -CommandType Function
```

CommandType	Name	Version	Source
Function	Add-ClusterGroupSetDependency	2.0.0.0	FailoverClusters
Function	Add-ClusterGroupToSet	2.0.0.0	FailoverClusters
Function	Get-ClusterGroupSet	2.0.0.0	FailoverClusters
Function	Get-ClusterGroupSetDependency	2.0.0.0	FailoverClusters
Function	New-ClusterGroupSet	2.0.0.0	FailoverClusters
Function	Remove-ClusterGroupFromSet	2.0.0.0	FailoverClusters
Function	Remove-ClusterGroupSet	2.0.0.0	FailoverClusters
Function	Remove-ClusterGroupSetDependency	2.0.0.0	FailoverClusters
Function	Set-ClusterGroupSet	2.0.0.0	FailoverClusters

For VM with Active Directory:

```
PS C:\> New-ClusterGroupSet -Name AD -Group Clu-VM01
```

Name : AD

GroupNames : {Clu-VM01}

ProviderNames : {}

StartupDelayTrigger : Delay

StartupCount : 4294967295

IsGlobal : False

StartupDelay : 20

For VM with Application:

```
New-ClusterGroupSet -Name Application -Group Clu-VM02
```

For the service that depends from Application:

```
New-ClusterGroupSet -Name SubApp -Group Clu-VM03
```

Dependencies between groups:

```
Add-ClusterGroupSetDependency -Name Application -Provider AD
```

```
Add-ClusterGroupSetDependency -Name SubApp -Provider Application
```

To change already created Set use the cmdlet Set-ClusterGroupSet:

Example: Set-ClusterGroupSet Application -StartupDelayTrigger Delay -StartupDelay 30

StartupDelayTrigger defines what action should trigger the start and can have one of two values:

- Delay – waits 20 second (by default). Uses StartupDelay value.
- Online – waits until the group has reached an online state

StartupDelay – delay time in seconds. 20 seconds by default

isGlobal – defines if the set should start before all other sets (for example, set with Active Directory VMs must be globally available and, therefore, has to be started before all other sets)

Let's start VM Clu-VM03 (service that depends from Application):

What's new in Failover Clustering | Roman Levchenko | rlevchenko.com

It waits while Active Directory on Clu-VM01 becomes available (StartupDelayTrigger – Delay , StartupDelay – 20 seconds)

Name	Status	Type	Owner Node	Priority	Info
Clu-VM01	Running	Virtual Machine	HV01	Medium	
Clu-VM02	Off	Virtual Machine	HV02	Medium	
Clu-VM03	Starting	Virtual Machine	HV01	Medium	

When Active Directory is online, Clu-VM02 VM starts (StartupDelay is also used here)

Name	Status	Type	Owner Node	Priority	Info
Clu-VM01	Running	Virtual Machine	HV01	Medium	
Clu-VM02	Running	Virtual Machine	HV02	Medium	
Clu-VM03	Starting	Virtual Machine	HV01	Medium	

Clu-VM02 is available and running -> signal to start Clu-VM03 VM

Name	Status	Type	Owner Node	Priority	Info
Clu-VM01	Running	Virtual Machine	HV01	Medium	
Clu-VM02	Running	Virtual Machine	HV02	Medium	
Clu-VM03	Running	Virtual Machine	HV01	Medium	

VM Compute/Storage Resiliency

Now we have the new states of nodes and VMs for their better resiliency in scenarios with network or storage issues. **Storage and compute resiliencies** have been added to achieve proactive action, react on “small” problems and predict the most critical problems. Let’s review some examples.

Isolated Mode

Cluster service on node HV01 is unavailable so there is an issue with intra-cluster communication. In this case HV01 node changes state to Isolated (*ResiliencyLevel* parameter) and being removed from active cluster membership

Name	Status	Assigned Vote	Current Vote
HV01	Isolated	1	1
HV02	Up	1	1

HV01 continues to host all VMs* but state of all VMs becomes “Unmonitored” (it means that cluster service does not manage them)

Clu-VM02	Running	Virtual Machine	HV02	Medium
Clu-VM03	Unmonitored	Virtual Machine	HV01	Medium

*VMs continues to run if they sit on SMB storage. VMs placed to “Paused Critical” state if they are on block storage (FC/iSCSI and etc) ‘cause isolated node no longer has access to CSV

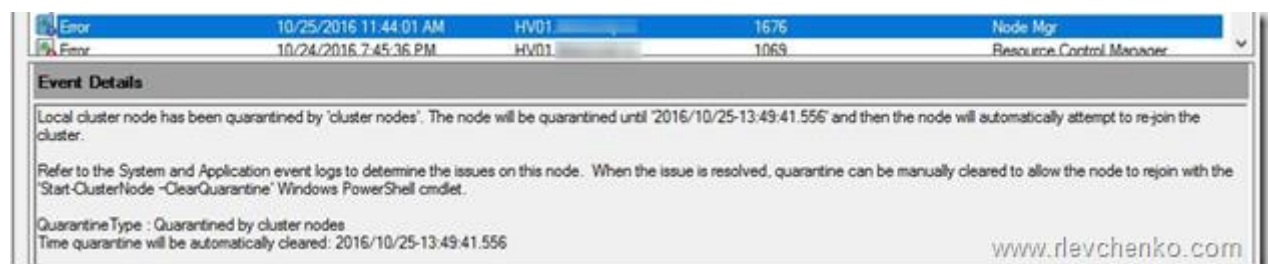
It's allowed to HV01 be in *Isolated* state within *ResiliencyDefaultPeriod* (240 seconds by default). But if the cluster service (in my case) is not in online state after 240 seconds, HV01 host will go into a down state and VMS will be migrated to the "health" node.

Quarantined

Let's say that HV01 recovered from *Isolated* state, cluster service became online and it seems all nodes are in a good condition. Unexpectedly, cluster service on HV01 becomes unavailable again and this issue is repeated once and more times during a last hour. In this case *QuarantineThreshold* (number of failures before a node is Quarantined. By default, it's 3) will be reached and node will go into **Quarantine** state for 2 hours (*QuarantineDuration* parameter). All VMs will be moved from HV01 to the health HV02 node.



Name	Status	Assigned Vote	Current Vote
HV01	Quarantined	1	1
HV02	Up	1	1



We fixed all issues in HV01 and want to bring it out of quarantine. In this case, we need to run the following command:

```
PS C:\> Start-ClusterNode HV01 -ClearQuarantine
```

Name	ID	State
----	--	----
HV01	1	Joining

Please note that **no more than 25% of nodes can be quarantined** at any given time

To customize settings:

(Get-Cluster). *QuarantineDuration* = 1800

Storage Resiliency

Do you know what happens when shared storage becomes unavailable? Yes, you are right. VMs go to Offline state and then require cold boot on the next start. It was...Now Windows Server 2016 takes these VMs to **Paused-Critical** (*AutomaticCriticalErrorAction* parameter) and freezes their state (r/w operations stopped, VM is unavailable but it's not turned off)

If storage comes back during 30 minutes (*AutomaticCriticalErrorActionTimeout*, 30 minutes is default), VM goes out of Paused-Critical state and becomes available (analogy – pause/play in your audio player). If storage is still unavailable after configured timeout, VMs will be turned off.

```
AutomaticStartAction : StartIfRunning
AutomaticStartDelay : 0
AutomaticStopAction : Save
AutomaticCriticalErrorAction : Pause
AutomaticCriticalErrorActionTimeout : 30
```

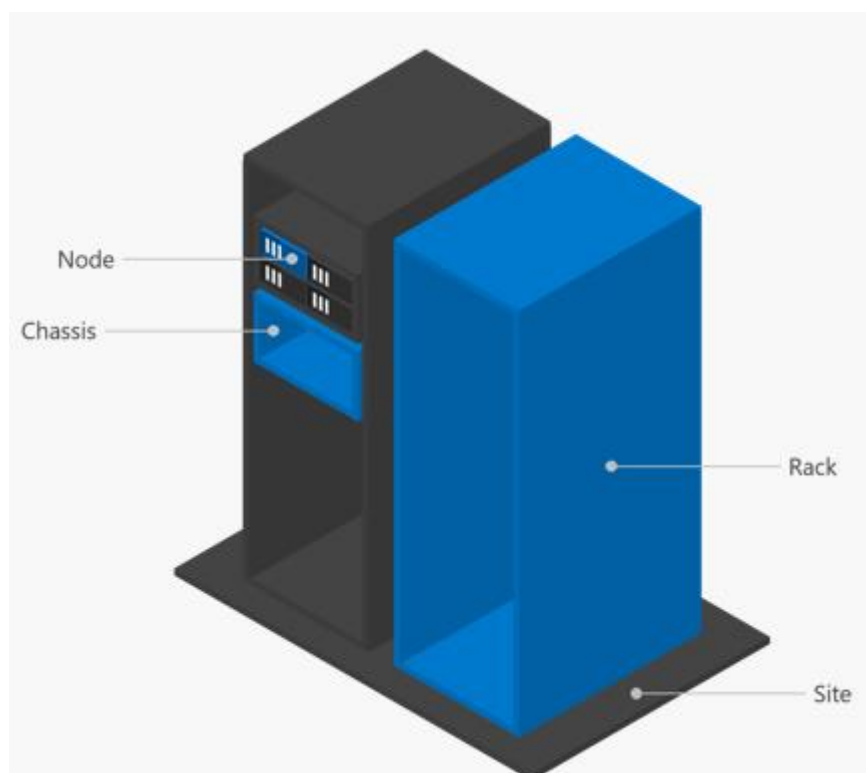
Site-Aware/Stretched Clusters and Storage Replica

Previously, we needed to find 3rd party solutions for SAN-to-SAN replication or etc. And building stretched clusters required a huge amount of money. Windows Server 2016 can help to significantly reduce costs and enhance unification in such scenarios.

Storage Replica is the main component of multi-site clusters or DR-solution and it supports both asynchronous and synchronous (!) replication between any storage devices (including Storage Spaces Direct). Storage Replica is only available in Datacenter Edition and can be used in the following configurations:



Storage Replica supports automatic failover in stretched clusters and works side-by-side with the other newest feature in Windows Server 2016 – site-awareness. **Site-Awareness** allows you to define groups of cluster nodes and link them to physical locations (site fault domain/sites) in order to form custom failover policies, VMs and S2D data placement. And in addition, we can link these groups on node, rack or chassis level. See the examples below.



New-ClusterFaultDomain -Name Voronezh -Type Site -Description "Primary" -Location "Voronezh DC"

New-ClusterFaultDomain -Name Voronezh2 -Type Site -Description "Secondary" -Location "Voronezh DC2"

New-ClusterFaultDomain -Name Rack1 -Type Rack

New-ClusterFaultDomain -Name Rack2 -Type Rack

New-ClusterFaultDomain -Name HPc7000 -type Chassis

New-ClusterFaultDomain -Name HPc3000 -type Chassis

Set-ClusterFaultDomain -Name HV01 -Parent Rack1

Set-ClusterFaultDomain -Name HV02 -Parent Rack2





Set-ClusterFaultDomain Rack1,HPc7000 -parent Voronezh

Set-ClusterFaultDomain Rack2,HPc3000 -parent Voronezh2

```
PS C:\Windows\system32> Get-ClusterFaultDomain

Name      Type ParentName ChildrenNames
-----
Voronezh   Site
Voronezh2  Site
Rack1      Rack Voronezh  HV01
Rack2      Rack Voronezh2  HV02
HPc3000    Chassis Voronezh2
HPc7000    Chassis Voronezh
HV01       Node Rack1
HV02       Node Rack2
```

Final result:

Name	Status	Assigned Vote	Current Vote	Site	Rack	Chassis
 HV01	 Up	1	1	Voronezh	Rack 1	
 HV02	 Up	1	1	Voronezh2	Rack 2	

Site-Awareness benefits:

- Groups failover to a node within the same site, before failing to a node in a different site
- During Node Drain VMs are moved first to a node within the same site before being moved cross site
- The CSV load balancer will distribute within the same site
- Virtual Machines (VMs) follow storage and are placed in same site where their associated storage resides. VMs will begin live migrating to the same site as their associated CSV after 1 minute of the storage being moved.

Using site-awareness we can define the parent site for all new created VMs:

(Get-Cluster).PreferredSite = <site name>

Or set it for the specific cluster group:

(Get-ClusterGroup -Name GroupName).PreferredSite = <preferred site name>

Miscellaneous

- Storage Spaces Direct and Storage QoS support
- **Online shared VHDX** resizing for guest clusters, Hyper-V replica and host-level backup support
- Enhanced scalability and performance of **CSV Cache** with added support for tiered spaces, storage spaces direct and deduplication (it becomes usual to give a tens Gbs to CSV Cache)
- **Cluster Log Changes** (time zones information, active memory dumps) to simplify overall diagnostic
- WSFC automatically recognizes and configures multiple NICs on the same subnet for **SMB MultiChannel**. No configuration is necessary.

Thank you very much for reading!