



vSphere High Availability Deployment Best Practices

TECHNICAL MARKETING DOCUMENTATION
V 1.5/UPDATED JULY 2011

Table of Contents

Introduction	3
Design Principles for High Availability	4
Host Considerations	4
Host Selection	4
Host Versioning	5
Host Placement	5
VMware vSphere Auto Deploy Hosts	5
VMware vCenter Server Availability Considerations	6
Networking Design Considerations	6
General Networking Guidelines	6
Setting Up Redundancy for VMware HA Networking	7
Network Adaptor Teaming and Management Networks	7
Management Network Changes in a VMware HA Cluster	8
Storage Design Considerations	8
Storage Heartbeats	9
Cluster Configuration Considerations	9
Host Isolation	9
Host Isolation Detection	10
Host Isolation Response	10
Leave Powered On	10
Power Off	11
Shut Down	11
Host Monitoring	12
Cluster Partitions	12
Virtual Machine and Application Health Monitoring	12
vSphere High Availability and VMware Fault Tolerance	13
Host Partitions	13
Host Isolation	13
Admission Control	13
Affinity Rules	14
Log Files	15
Configuring Log Capacity	16
General Logging Recommendations for All ESX Versions	16
Summary	17

Introduction

Downtime, whether planned or unplanned, brings with it considerable costs. Solutions to ensure higher levels of availability have traditionally been very costly, hard to implement and difficult to manage.

VMware vSphere® makes it simpler and less expensive to provide higher levels of availability for important applications. With vSphere, organizations can easily and cost-effectively increase the baseline level of availability provided for all applications.

vSphere makes it possible to reduce both planned and unplanned downtime. With the revolutionary VMware vSphere® vMotion® capabilities in vSphere, it is possible to perform planned maintenance with zero application downtime. vSphere High Availability (VMware HA) specifically reduces unplanned downtime by leveraging multiple VMware ESX®/VMware ESXi™ hosts configured as a cluster, to provide rapid recovery from outages and cost-effective high availability for applications running in virtual machines.

VMware HA provides for application availability in the following ways:

- It reacts to hardware failure and network disruptions by restarting virtual machines on active hosts within the cluster.
- It detects operating system (OS) failures by continuously monitoring a virtual machine and restarting it as required.
- It provides a mechanism to react to application failures.
- Unlike other clustering solutions, it provides the infrastructure to protect all workloads within the cluster.

At its basic level, there is no need to install additional software within the application or virtual machine. HA protects all workloads. After it is configured, no further actions are required to protect new virtual machines added to a cluster. They are automatically protected.

You can combine HA with vSphere Distributed Resource Scheduler (DRS) to protect against failures and to provide load balancing across the hosts within a cluster.

The following are among the advantages that HA has over traditional failover solutions:

- Minimal setup
- Reduced complexity (e.g., no need for quorum disks)
- Reduced hardware cost and setup
- Increased application availability without the expense of additional idle failover hosts or the complexity of maintaining identical hosts for failover pairs
- DRS and vMotion integration

Refer to the *vSphere Availability Guide* for more information on the basics of HA, including how to create HA clusters, how it works, the benefits of integrating with DRS and explanations on configuration procedures.

Design Principles for High Availability

The key to architecting a highly available computing environment is to eliminate single points of failure. With the potential of occurring anywhere in the environment, failures can affect both hardware and software. Building redundancy at vulnerable points helps reduce or eliminate downtime caused by [implied] hardware failures. These include redundancies at the following layers:

- Server components such as network adaptors and host bus adaptors (HBAs)
- Servers, including blades and blade chassis
- Networking components
- Storage arrays and storage networking

Host Considerations

Prior planning regarding hosts to be used in a cluster will provide the best results. Because it is not always possible to start with a greenfield environment, this section will discuss some of the considerations applicable to the hosts.

Host Selection

Overall vSphere availability starts with proper host selection. This includes items such as redundant power supplies, error-correcting memory, remote monitoring and notification, and so on. Consideration should also be given to removing single points of failure in host location. This includes distributing hosts across multiple racks or blade chassis to remove the ability for rack or chassis failure to impact an entire cluster.

When deploying a VMware HA cluster, it is a best practice to build the cluster out of identical server hardware. Using identical hardware provides a number of key advantages, such as the following ones:

- Simplifies configuration and management of the servers using Host Profiles
- Increases ability to handle server failures and reduces resource fragmentation. Using drastically different hardware will lead to an unbalanced cluster, as described in the “Admission Control” section. By default, HA prepares for the worst-case scenario in that the largest host in the cluster could fail. To handle the worst case, more resources across all hosts must be reserved, making them essentially unusable.

Additionally, care should be taken to remove any inconsistencies that would prevent a virtual machine from being started on any cluster host. Inconsistencies such as mounting datastores to a subset of the cluster hosts or the implementation of DRS “required” virtual machine-to-host required affinity rules are examples of items to carefully consider. Avoiding these conditions will increase the portability of the virtual machine and provide a higher level of availability. Inconsistencies can be checked for a given virtual machine by using the VMware vSphere® Client™ and selecting the migrate option to determine whether any error conditions would prevent vMotion from being able to migrate the virtual machine to other hosts in the cluster.

The overall size of a cluster is another important factor to consider. Smaller-sized clusters require a larger relative percentage of the available cluster resources to be set aside as reserve capacity to adequately handle failures. For example, for a cluster of three nodes to tolerate a single host failure, about 33 percent of the cluster resources will be reserved for failover. A 10-node cluster requires that only 10 percent be reserved. This is discussed in more detail in the “Admission Control” section of this document. In contrast, as cluster size increases, so does the HA management complexity of the cluster. This complexity is associated with general configuration factors as well as ongoing management tasks such as troubleshooting. This increase in management complexity, however, is overshadowed by the benefits a large cluster can provide. Features such as DRS and vSphere Distributed Power Management (VMware DPM) become very compelling with large clusters. In general, it is recommended that customers establish the largest clusters possible to reap the full benefits of these solutions.

Host Versioning

An ideal configuration is one where all the hosts contained within the cluster use the latest version of ESXi. When adding a host to vSphere 5.0 clusters, it is always a best practice to upgrade the host to ESXi 5.0 and to avoid using clusters with mixed-host versions.

Upgrading hosts to ESXi 5.0 enables users to utilize features that were not supported for earlier host versions and to leverage capabilities that have been improved. A good example of this is the support added in vSphere 5.0 for management network partitions. This feature is supported only if the cluster contains only ESXi 5.0 hosts.

Mixed clusters are supported but not recommended because there are some differences in VMware HA performance between host versions, and these differences can introduce operational variances in a cluster. These differences arise from earlier host versions' not offering the same capabilities as later versions. For example, ESX 3.5 hosts do not support certain properties present within ESX 4.0 and greater. These properties were added to ESX 4.0 to inform HA of conditions warranting a restart of a virtual machine. As a result, HA will not restart virtual machines that crash while running on ESX 3.5 hosts but will restart such a virtual machine if it was running on an ESX 4.0 or later host.

The following recommendations apply if using an HA-enabled cluster that includes hosts with differing versions:

- Be aware of the general limitations of using a mixed cluster, as previously mentioned.
- ESX(i) 3.5 hosts within a 5.0 cluster must include a patch to address an issue involving file locks. For ESX 3.5 hosts, you must apply the ESX350-201012401-SG patch. For ESXi 3.5, you must apply the ESXi350-201012401-I-BG patch. Prerequisite patches must be applied before applying these patches. HA will not allow an ESX(i) 3.5 host to be added to the cluster if it does not meet the patch requirements.
- Do not deploy mixed clusters if vSphere Storage vMotion or Storage DRS is required. Refer to the *vSphere Availability Guide* for more information on this topic.

Host Placement

In versions of vSphere prior to 5.0, the concept of primary and secondary hosts, with a limit of five primary hosts, was present. This construct was the focus of many best-practice recommendations. In vSphere High Availability 5.0, this construct has been eliminated. Instead, VMware HA now uses a master/slave relationship between the nodes of a cluster. Under normal operations, there will be a single host that takes on the role of the master. All other hosts are referred to as slave hosts. In the event of a failure of a host acting as a master, an election process is performed and a new master is selected.

The use of this new construct for HA eliminates concerns held previously regarding the following issues:

- Number of hosts in a cluster
- Managing the role of the host
- Number of consecutive host failures
- Placement of hosts across blade chassis and stretched clusters
- Partition scenarios likely to occur in stretched cluster environments

VMware vSphere Auto Deploy Hosts

VMware vSphere® Auto Deploy ESXi hosts provide numerous advantages within a virtualized environment. Among these are increases in flexibility and ease of management. Their use, however, brings additional considerations to bear in a highly available configuration.

Refer to the *Auto Deploy Best Practices Guide* for specific guidance in designing high availability into an environment with Auto Deploy.

VMware vCenter Server Availability Considerations

VMware™ vCenter Server is the management focal point for any VMware® environment. Although vSphere High Availability will continue to protect your environment without vCenter Server, your ability to manage the environment is severely impacted without it. It is highly recommended that you protect your vCenter Server instance as best as possible. The following are among the ways that you can accomplish this:

- Use of VMware vCenter Server Heartbeat™ – A specially designed HA solution for vCenter.
- Use of vSphere High Availability – Useful in environments where the vCenter Server instance is virtualized, such as when using the vSphere vCenter Server Appliance.

Which option you choose depends on your configuration, requirements and budget. In either case, the goal is to provide as much protection for vCenter Server as possible.

It is extremely critical when using ESXi Auto Deploy that both the Auto Deploy service and the vCenter Server instance used are highly available. In the event of a loss of the vCenter Server instance, Auto Deploy hosts might not be able to reboot successfully in certain situations.

There are several recommendations related to providing for the availability of vSphere vCenter Server as it pertains to the use of Auto Deploy. These are discussed in detail in the *Auto Deploy Best Practices Guide*. It bears repeating here, though, that if vCenter Server is made highly available through the use of HA, the vCenter Server virtual machine must be configured with a restart priority of “high.” This ensures that the vCenter Server virtual machine will be among the first virtual machines to be restarted in the event of a failure.

Additionally, this virtual machine should be configured to run on two, or more, hosts that are not managed by Auto Deploy. This can be done by using a DRS virtual machine-to-host rule or by deploying the virtual machine on a datastore accessible to only these hosts. Because Auto Deploy depends upon the availability of vCenter Server in certain circumstances, this ensures that the vCenter Server virtual machine is able to come online. This does not require that DRS be enabled if you use DRS rules, because these rules will remain in effect after DRS has been disabled.

Providing for the highest availability of the vCenter Server instance will ensure proper operation of the cluster at all times.

Networking Design Considerations

Best practices network design falls into two specific areas, increasing resiliency of “client-side” networking to ensure access from external systems to workloads running in vSphere, and increasing resiliency of communications used by HA itself.

General Networking Guidelines

The following suggestions are best practices for configuring networking in general for improved availability:

- Configuring switches. If the physical network switches that connect your servers support the PortFast (or an equivalent) setting, enable it. This setting prevents a host from incorrectly determining that a network is isolated during the execution of lengthy spanning tree algorithms on boot. For more information on this option, refer to the documentation provided by your networking switch vendor.
- Disable host monitoring when performing any network maintenance that might disable all heartbeat paths (including storage heartbeats) between the hosts within the cluster, because this might trigger an isolation response.

- With vSphere High Availability 5.0, all dependencies on DNS have been removed, making previous best practices of updating DNS for the purpose of HA irrelevant. However, it is always a best practice to ensure that the hosts and virtual machines within an environment are able to be properly resolved through DNS.
- Use consistent port names on VLANs for virtual machine networks on all ESXi hosts in the cluster. Port names are used to determine compatibility of the network by virtual machines. Before initiating a failover, HA will check whether a host is compatible. If there are no hosts with matching port group names available, no failover is possible. Use of a documented naming scheme is highly recommended. Issues with port naming can be completely mitigated by using vSphere distributed virtual switches.
- In environments where both IPv4 and IPv6 protocols are used, configure the virtual switches on all hosts to enable access to both networks. This prevents network partition issues due to the loss of a single IP networking stack or host failure.
- Ensure that TCP/UDP port 8182 is open within the network. HA will automatically open these ports when enabled and close when disabled. User action is required only if there are firewalls in place between hosts within the cluster, as in a stretched cluster configuration.
- If using an ESX host, ensure that there is no service running in the ESX console OS that uses port 8182.
- Configure redundant management networking from ESXi hosts to network switching hardware if possible. Using network adaptor teaming can enhance overall network availability as well as increase overall throughput.
- In configurations with multiple management networks, an isolation address must be configured for each network. Refer to the “Host Isolation” section for more details.

Setting Up Redundancy for VMware HA Networking

Networking redundancy between cluster hosts is absolutely critical for HA reliability. Redundant management networking enables the reliable detection of failures.

NOTE: Because this document is primarily focused on vSphere 5.0, “management network” within this document refers to the VMkernel network selected for use as a management network. Refer to the vSphere Availability Guide for information regarding the service console network if using an earlier version of vSphere.

You can implement network redundancy at the network adaptor level or at the management network level. In most implementations, network adaptor teaming provides sufficient redundancy. It is explained here. If you want to add additional redundancy, see the *ESXi Configuration Guide* for more information on setting up redundant management networks.

Network Adaptor Teaming and Management Networks

Using a team of two network adaptors connected to separate physical switches can improve the reliability of the management network. Because servers connected to each other through two network adaptors, and through separate switches, have two independent paths for cluster communication, the cluster is more resilient to failures.

To configure a network adaptor team for the management network, configure the vNICs in the vSwitch configuration for the ESXi host for active/standby configuration.

Requirements:

- Two physical network adaptors
- VLAN trunking
- Two physical switches

The vSwitch should be configured as follows:

- Load balancing = route based on the originating virtual port ID (default)
- Failback = no
- vSwitch0: Two physical network adaptors (for example: vmnic0 and vmnic2)
- Two port groups (for example, vMotion and management)

In this example, the management network runs on vSwitch0 as active on vmnic0 and as standby on vmnic2. The vMotion network runs on vSwitch0 as active on vmnic2 and as standby on vmnic0

Each port group has a VLAN ID assigned and runs dedicated on its own physical network adaptor. Only in the case of a failure is it switched over to the standby network adaptor. Failback is set to “no” because in the case of physical switch failure and restart, ESXi might falsely recognize that the switch is back online when its ports first come online. In reality, the switch might not be forwarding on any packets until it is fully online. However, when failback is set to “no” and an issue arises, both your management network and vMotion network will be running on the same network adaptor and will continue running until you manually intervene.

Management Network Changes in a VMware HA Cluster

VMware HA uses the management network as its primary communication path. As a result, it is critical that proper precautions be taken whenever a maintenance action will affect the management network.

As a general rule, whenever maintenance is to be performed on the management network, the hosts should be placed into maintenance mode. This mode will prevent HA from possibly identifying the maintenance action as a failure and taking action.

If changes that involve the management network occur without placing a host into maintenance mode, it is recommended to reconfigure HA after the maintenance action is completed. This ensures that any pertinent changes are recognized by HA. Changes that cause a loss of management network connectivity are grounds for performing a reconfiguration of HA. An example of this is the addition or deletion of networks used for management network traffic without the host's being placed into maintenance mode.

Storage Design Considerations

To maintain a constant connection between an ESXi host and its storage, ESXi supports multipathing, a technique that enables you to use more than one physical path to transfer data between the host and an external storage device.

In case of a failure of any element in the SAN network, such as an adapter, switch or cable, ESXi can move to another physical path that does not use the failed component. This process of path switching to avoid failed components is known as path failover.

In addition to path failover, multipathing provides load balancing, the process of distributing I/O loads across multiple physical paths. Load balancing reduces or removes potential bottlenecks

For Fibre Channel SAN configurations, multipathing setup is very specific to the HBA, switch and array components chosen. See the *Fibre Channel Configuration Guide* for more information.

For configurations using iSCSI, ESXi supports the creation of a second iSCSI initiator to enable multipathing configurations. See the *iSCSI SAN Configuration Guide* for more details on setting up multiple iSCSI initiators.

For all block storage configurations, VMware strongly recommends multiple paths to storage for maximum resiliency.

Storage Heartbeats

A new feature of VMware HA in vSphere 5.0 is the ability to use storage subsystems as a means of communication between the nodes of a cluster. Storage heartbeats are used when the management network is unavailable to enable a slave to communicate with a master. This provides an additional level of redundancy for internode communication. It also provides an ability to accurately distinguish between the different failure scenarios of dead, isolated or partitioned hosts. With this new ability, storage heartbeats enable detection of cluster partition scenarios that are not supported with previous versions of vSphere, providing a more coordinated failover when host isolation occurs.

By default, vCenter will automatically select two datastores to use for storage heartbeats. An algorithm designed to maximize availability and redundancy of the storage heartbeats selects these datastores. This algorithm attempts to select datastores that are connected to the highest number of hosts. It also attempts to select datastores that are hosted on different storage arrays/NFS servers. A preference is given to VMware vSphere® VMFS-formatted datastores, although NFS-hosted datastores can also be used. This weighting attempts to avoid network issues that would disrupt the storage heartbeat datastore access. vCenter selects the heartbeat datastores when HA is enabled, when a datastore is added or removed from a host and when the accessibility to a datastore changes.

Although users can manually select the datastores to be used for storage heartbeating, it is not recommended, because having vCenter automatically select and update the storage heartbeat datastores reduces management tasks. It also provides more flexibility for the system to adapt to unplanned storage outages.

Environments that provide only network-based storage must take into consideration the network architecture to fully realize the potential of the storage heartbeat feature. If the storage network traffic and the management network traffic flow through the same network, disruptions in network service potentially might disrupt both. It is recommended that these networks be separated as much as possible.

It is also recommended that all hosts within a cluster have access to the same datastores. This promotes virtual machine portability because the virtual machines will be able to fail over to any of the hosts within the cluster. It also is beneficial because it enables those datastores to be used for storage heartbeats. If network partitions or isolations are anticipated within the environment, a minimum of two datastores is recommended to be accessible to all hosts within the cluster.

Refer to the *vSphere 5.0 Availability Guide* for detailed information about the storage heartbeat feature.

Cluster Configuration Considerations

Many options are available when configuring a vSphere High Availability environment. This flexibility enables users to leverage HA in a wide range of environments while accommodating particular requirements. In this section, we will discuss some of the key cluster configuration options and the recommended best practices for them.

Host Isolation

One key mechanism within HA is the ability for a host to detect when it has become network isolated from the rest of the cluster. The host isolation feature is best suited to protecting virtual machines against loss of network connectivity in environments where the loss of connectivity in the management network likely also corresponds to a loss of connectivity in the virtual machine network(s). If failures of the management network are not likely correlated with failures of the virtual machine network, the loss of the management network simply results in the inability to manage the virtual machines on the isolated host. In such environments, it is often preferable to leave the virtual machines running while the management network connectivity is restored.

In HA, a host detects that it is isolated if attempts to ping the configured isolation address fail and no HA agent traffic appears on the management network(s). When a host detects that it is isolated, it will invoke the

configured isolation response. If it has access to its heartbeat datastores, it also writes to disk the fact that it is isolated. At the same time, the VMware HA master agent monitors the state of the host and attempts to restart any virtual machines that are powered off by the isolated host. If heartbeat datastore connectivity is not impacted by the isolation, the master waits for the isolated host to report virtual machine power-offs before attempting to restart virtual machines. If heartbeat datastore connectivity is affected, the master declares the host dead and begins attempting to fail over the virtual machines.

Although as compared to prior versions, vSphere 5.0 provides enhancements against a host isolation scenario, it is best to try to prevent this scenario from occurring at all. Therefore, as previously mentioned, it is highly recommended that care be taken in designing the network infrastructure for the management network to provide redundancy to eliminate any single points of failure.

Host Isolation Detection

The hosts within an HA cluster constantly heartbeat with the host designated as the master over the management network. The first step in determining whether a host is isolated is detecting a lack of these heartbeats.

After a host stops communicating with the master, the master attempts to determine the cause of the issue. Using heartbeat datastores, the master can distinguish whether the host is still alive by determining if the affected host is maintaining heartbeats to the heartbeat datastores. This enables the master to differentiate between a management network failure, a dead host and a partitioned/isolated situation.

The time elapsed before the host declares itself isolated varies depending on the role of the host (master or slave) at the time of the loss of heartbeats. If the host was a master, it will declare itself isolated within 5 seconds. If the host was a slave, it will declare itself isolated in 30 seconds. The difference in time is due to the fact that if the host was a slave, it then must go through an election process to identify whether any other hosts exist or if the master host simply died. This election process starts for the slave at 10 seconds after the loss of heartbeats is detected. If the host sees no response from another host during the election for 15 seconds, the HA agent on a host then elects itself as a master, checks whether it is isolated and, if so, drops into a startup state. In short, a host will begin to check to see whether it is isolated whenever it is a master in a cluster with more than one other host and has no slaves. It will continue to do so until it becomes a master with a slave or connects to a master as a slave. At this point, the host will attempt to ping its configured isolation addresses to determine the viability of the network. The default isolation address is the gateway specified for the management network. Advanced settings can be used to modify the isolation addresses used for your particular environment. The option `das.isolationaddress[X]` (where X is 1-10) is used to configure multiple isolation addresses. Additionally `das.usedefaultisolationaddress` is used to indicate whether the default isolation address (the default gateway) should be used to determine if the host is network isolated. If the default gateway is not able to receive ICMP ping packets, you must set this option to "false." It is recommended to set one isolation address for each management network used, keeping in mind that the management network links should be redundant, as previously mentioned. The isolation address used should always be reachable by the host under normal situations, because after 5 seconds have elapsed with no response from the isolation addresses, the host then declares itself isolated. After this occurs, it will attempt to inform the master of its isolated state by use of the heartbeat datastores.

Host Isolation Response

Now the isolated host must determine whether it must take any action based upon the configuration settings for the isolation response for each virtual machine that is powered on. The isolation response setting provides a means to dictate the action desired for the powered-on virtual machines maintained by a host when that host is declared isolated. There are three possible isolation response values that can be configured and applied to a cluster or individually to a specific virtual machine. These are **Leave Powered On**, **Power Off** and **Shut Down**.

Leave Powered On

With this option, virtual machines hosted on an isolated host are left powered on. In situations where a host loses all management network access, it might still have the ability to access the storage subsystem and the virtual machine network. Selecting this option enables the virtual machine to continue to function if this were to occur. This is now the default isolation response setting in vSphere High Availability 5.0.

Power Off

When this isolation response option is used, the virtual machines on the isolated host are immediately stopped. This is similar to removing the power from a physical host. This can induce inconsistency with the file system of the OS used in the virtual machine. The advantage of this action is that VMware HA will attempt to restart the virtual machine more quickly than when using the third option.

Shut Down

Through the use of the VM Tools package installed within the guest operating system of a virtual machine, this option attempts to gracefully shut down the operating system with the virtual machine before powering off the virtual machine. This is more desirable than using the **Power Off** option because it provides the OS with time to commit any outstanding I/O activity to disk. HA will wait for a default of 300 seconds (5 minutes) for this graceful shutdown to occur. If the OS is not gracefully shut down by this time, it will initiate a power-off of the virtual machine. Changing the `das.isolationshutdowntimeout` attribute will modify this timeout if it is determined that more time is required to gracefully shut down an OS. The **Shut Down** option requires that the VM Tools package be installed in the guest OS. Otherwise, it is equivalent to the **Power Off** setting.

From a best practices perspective, **Leave Powered On** is the recommended isolation response setting for the majority of environments. Isolated hosts are a rare event in a properly architected environment, given the redundancy built in. In environments that use network-based storage protocols, such as iSCSI and NFS, the recommended isolation response is **Power Off**. With these environments, it is highly likely that a network outage that causes a host to become isolated will also affect the host's ability to communicate to the datastores.

An isolated host will initiate the configured isolation response for a running virtual machine if either of the following is true: 1) the host lost access to the datastore containing the configuration (.vmx) file for the virtual machine; 2) the host still has access to the datastore and it determined that a master is responsible for the virtual machine. To determine this, the isolated host checks for the accessibility of the "home datastore" for each virtual machine and whether the virtual machines on that datastore are "owned" by a master, which is indicated by a master's having exclusively locked a key file that HA maintains on the datastore. After declaring itself as being isolated, the isolated host releases any locks it might have held on any datastores. It then checks periodically to see whether a master has obtained a lock on the datastore. After a lock is observed on the datastore by the isolated host, the HA agent on the isolated host applies the configured isolation response. Ensuring that a virtual machine is under continuous protection by a master provides an additional layer of protection. Because only one master can lock a datastore at a given time, this significantly reduces chances of "split-brain" scenarios. This also protects against situations where a complete loss of the management networks without a complete loss of access to storage would make all the hosts in a cluster determine they were isolated.

In certain environments, it is possible for a loss of the management network to also affect access to the heartbeat datastores. This is the case when the heartbeat datastores are hosted via NFS that is tied to the management network in some manner. In the event of a complete loss of connectivity to the management network and the heartbeat datastores, the isolation response activity resembles that observed in vSphere 4.x. In this configuration, the isolation response should be set to **Power Off** so another virtual machine with access to the network can attempt to power on the virtual machine.

There is a situation where the isolation response will likely take an extended period of time to transpire. This occurs when all paths to storage are disconnected, referred to as an all-paths-down (APD) state, and the APD condition does not impact all of the datastores mounted on the host. This is due to the fact that there might be outstanding write requests to the storage subsystem that must time out. Establishing redundant paths to the storage subsystem will help prevent an APD situation and this issue.

Host Monitoring

The restarting by VMware HA of virtual machines on other hosts in the cluster in the event of a host isolation or host failure is dependent on the “host monitoring” setting. If host monitoring is disabled, the restart of virtual machines on other hosts following a host failure or isolation is also disabled. Disabling host monitoring also impacts VMware Fault Tolerance because it controls whether HA will restart a Fault Tolerance (FT) secondary virtual machine after an event. Essentially a host will always perform the programmed host isolation response when it determines it is isolated. The host monitoring setting determines if virtual machines will be restarted elsewhere following this event.

Cluster Partitions

vSphere High Availability 5.0 significantly improves handling of cluster partitions. A cluster partition is a situation where a subset of hosts within the cluster loses the ability to communicate with the rest of the hosts in the cluster but can still communicate with each other. This can occur for various reasons, but the most common cause is where a stretched cluster configuration is used. A stretched cluster is defined as a cluster that spans multiple sites within a metropolitan area.

When a cluster partition occurs, one subset of hosts will still be able to communicate to a master. The other subset of hosts will not. As such, the second subset will go through an election process and elect a new master. Therefore, it is possible to have multiple masters in a cluster partition, one per partition. This situation will last only as long as the partition exists. After the network issue causing the partition is resolved, the masters will be able to communicate and discover the multiplicity of master hosts. Anytime multiple masters exist and can communicate with each other over the management network, all but one will abdicate.

In a partitioned cluster, it is possible for a master of one partition to be responsible for virtual machines that might be running on hosts in the other partition. If one such virtual machine fails, or the host on which the virtual machine is running fails, the master will attempt to restart the virtual machine.

Additionally, vCenter Server will talk to only one master at a time, so the HA cluster state seen in the vSphere Client for the hosts will be from the perspective of the master that the vCenter Server instance is communicating with. The odds are slim that vCenter Server would be able to communicate with multiple masters if the masters could not communicate with each other. However, if this were to occur and the master with which vCenter Server was communicating failed, vCenter Server might connect to the master in the other partition, causing the vSphere Client to switch perspectives.

Avoidance of cluster partition situations depends on having robust management network architecture. Refer to the networking section of this document for recommended practices on this topic.

Virtual Machine and Application Health Monitoring

The functionality of both virtual machine and application monitoring has not changed from earlier versions. These features enable the HA agent on a host to heartbeat with a virtual machine through VM Tools or to an agent running within the virtual machine that is monitoring the application health. After the loss of a defined number of VMware Tools heartbeats with the virtual machine, HA will reset the virtual machine.

Virtual machine and application monitoring are not dependent on the virtual machine protection state attribute as reported by the vSphere Client. This attribute signifies that HA detects that the desired state of the virtual machine is to be powered on. As such, it will attempt to restart it assuming that there is nothing restricting the restart. Conditions that might restrict this action include insufficient resources available and a disabled virtual machine restart priority.

This functionality is not available when the HA agent on a host is in the uninitialized state, as would occur immediately after the HA agent has been installed on the host or when hostd is not available. Additionally, the number of missed heartbeats is reset after the HA agent on the host reboots, which should occur rarely if at all, or after HA is reconfigured on the host.

Because virtual machines exist only for the purposes of hosting an application, it is highly recommended that virtual machine health monitoring be enabled. All hosts must have the VMware Tools package installed within the guest OS.

vSphere High Availability and VMware Fault Tolerance

Often vSphere High Availability is used in conjunction with VMware Fault Tolerance. Fault Tolerance provides protection for extremely critical virtual machines where any loss of service is intolerable. VMware HA detects the use of FT to ensure proper operation. This section describes some of the unique HA behavior specific to FT with HA. Additional VMware FT best practices can be found in the *vSphere Availability Guide*.

Host Partitions

HA will restart a secondary virtual machine of an FT virtual machine pair when the primary virtual machine is running in the same partition as the master HA agent that is responsible for the virtual machine. If this condition is not met, the secondary virtual machine won't be restarted until the partition is resolved.

Host Isolation

Host isolation responses are not performed on virtual machines enabled with Fault Tolerance. The rationale is that the primary and secondary FT virtual machine pairs are already communicating via the FT logging network. So they either continue to function and have network connectivity or they have lost network, they are not heartbeating over the FT logging network, and one of them will then take over as a primary FT virtual machine. Because HA does not offer better protection than that, it bypasses FT virtual machines when initiating host isolation response.

Ensure that the FT logging network that is used is implemented with redundancy to provide greater resilience to failures for FT.

Admission Control

vCenter Server uses HA admission control to ensure that sufficient resources in the cluster are reserved for virtual machine recovery in the event of host failure. Admission control will prevent the following if there is encroachment on resources reserved for virtual machines restarted due to failure:

- The power-on of new virtual machines
- Changes of virtual machine memory or CPU reservations
- A vMotion instance of a virtual machine into the cluster from another cluster

This mechanism is highly recommended to guarantee the availability of virtual machines. With vSphere 5.0, HA offers the following three configuration options for choosing your admission control strategy.

- **Host Failures Cluster Tolerates** (default): HA ensures that a specified number of hosts can fail and that sufficient resources remain in the cluster to fail over all the virtual machines from those hosts. HA uses a concept called slots to calculate available resources and required resources for a failing over of virtual machines from a failed host. Under some configurations, this policy might be too conservative in its reservations. The slot size can be controlled using a couple of advanced configuration options. The default setting was also changed in vSphere 5.0 from 256MHz to 32MHz when no reservation is used.
- **Percentage of Cluster Resources Reserved** as failover spare capacity: HA ensures that a specified percentage of memory and CPU resources are reserved for failover. This policy is recommended for situations where you must host virtual machines with significantly different CPU and memory reservations in the same cluster or have different-sized hosts in terms of CPU and memory capacity (vSphere 5.0 adds the ability to specify different percentages for memory and CPU through the vSphere Client). A key difference between this policy and the **Host Failures Cluster Tolerates** policy is that the capacity set aside for failures can be fragmented across hosts. So there is a chance that at the time of failing over a virtual machine, there might be insufficient unfragmented capacity available on a single host to power on all virtual machines. DRS, if enabled, will attempt to defragment the capacity in such situations.

- **Specify a Failover Host:** VMware HA designates a specific host or hosts as a failover host(s). When a host fails, HA attempts to restart its virtual machines on the specified failover host(s). The ability to specify more than one failover host is a new feature in vSphere High Availability 5.0. When a host is designated as a failover host, HA admission control disallows powering on virtual machines on that host, and DRS will not migrate virtual machines to the failover host. It effectively becomes a hot standby.

The best practices recommendation from VMware staff for admission control is as follows:

- Select the **Percentage of Cluster Resources Reserved** for admission control. This policy offers the most flexibility in terms of host and virtual machine sizing and is sufficient for most situations. In most cases, a simple calculation of 1/N, where N = total nodes in the cluster, will yield adequate sparing.
- If the **Host Failures Cluster Tolerates** setting is used, the following apply:
 - Ensure that all cluster hosts are sized equally. An “unbalanced” cluster results in excess capacity’s being reserved to handle failure of the largest possible node.
 - Attempt to keep virtual machine resource reservations similar across all configured virtual machines.**Host Failures Cluster Tolerates** uses a notion of “slot sizes” to calculate the amount of capacity needed to reserve for each virtual machine. The slot size is based on the largest reserved memory and CPU needed for any virtual machine. Mixing virtual machines of greatly different CPU and memory requirements will cause the slot size calculation to default to the largest possible of all virtual machines, limiting consolidation. See the vSphere Availability Guide for more information on slot size calculation and overriding slot size calculation in cases where you must configure different-sized virtual machines in the same cluster.

HA added a capability in vSphere 4.1 to balance virtual machine loading on failover, thereby reducing the issue of “resource imbalance” in a cluster after a failover. With this capability, there is less likelihood for vMotion instances after a failover. Also in vSphere 4.1, HA invokes DRS to create more contiguous capacity on hosts, to increase the chance for larger virtual machines to be restarted if some virtual machines cannot be restarted because of resource fragmentation. This does not guarantee enough contiguous resources to restart all the failed virtual machines. It simply means that vSphere will make the best effort to restart all virtual machines with the host resources remaining after a failure.

Admission control does not consider hosts that are disconnected or in maintenance mode. Hosts that are failed or in an error state according to HA are considered by the **Percentage of Cluster Resources** policy, but not for the **Host Failures Cluster Tolerates** policy. The intent is to ensure that enough capacity exists within the environment and to provide a warning to the user in certain situations.

The intent of the **Host Failures Cluster Tolerates** policy is to reserve resources on healthy hosts. This excludes from consideration hosts that are in a failed or error state. The **Percentage of Cluster Resources** policy determines the amount of resources reserved based upon how many hosts are considered. For example, 20 percent of five hosts is less than 20 percent of 10 hosts. If hosts in a failed or error state are not considered, the amount of resources being reserved is reduced. Because this is not preferable, HA considers them when using this policy.

Affinity Rules

A virtual machine–host affinity rule specifies that the members of a selected virtual machine DRS group should or must run on the members of a specific host DRS group. Unlike a virtual machine–virtual machine affinity rule, which specifies affinity (or anti-affinity) between individual virtual machines, a virtual machine–host affinity rule specifies an affinity relationship between a group of virtual machines and a group of hosts. There are “required” rules (designated by “must”) and “preferred” rules (designated by “should”). See the *vSphere Resource Management Guide* for more details on setting up virtual machine–host affinity rules.

When restarting virtual machines after a failure, HA ignores the preferential virtual machine–host rules but respects the required rules. If HA violates any preferential rule, DRS will attempt to correct it after the failover is complete by migrating virtual machines. Additionally, DRS might need to migrate other virtual machines to make space on the preferred hosts.

If required rules are specified, VMware HA will restart virtual machines only on an ESXi host in the same host DRS group. If no available hosts are in the host DRS group or the hosts are resource constrained, the restart will fail.

Any required rules defined when DRS is enabled are enforced even if DRS is subsequently disabled. So to remove the effect of such a rule, it must be explicitly disabled.

Limit the use of required virtual machine–host affinity rules to situations where they are necessary, because such rules can restrict HA target host selection when restarting a virtual machine after a failure.

Log Files

When an event occurs, it is important to determine its root cause. VMware provides excellent support services to assist in identifying and correcting any issues that might arise. Because VMware support personnel are engaged after an event has occurred, the historical information stored within the log files is a critical component for them.

In the latest version of HA, the changes in the architecture allowed for changes in how logging is performed. Previous versions of HA stored the operational logging information across several distinct log files. In vSphere High Availability 5.0, this information is consolidated into a single operational log file. This log file utilizes a circular log rotation mechanism, resulting in multiple files, with each file containing a part of the overall retained log history.

To improve the ability of the VMware support staff to diagnose problems, VMware recommends configuring logging to retain approximately one week of history. The following table provides recommended log capacities for several sample cluster configurations.

CONFIGURATION SIZE	SIZE QUALIFIER	PER-HOST MINIMUM LOG CAPACITY	DEFAULT LOG SETTINGS SUFFICIENT	
			ESXI	ESX
Small	<ul style="list-style-type: none"> • 40 total virtual machines • 8 virtual machines per host 	4MB	Yes ¹	Yes
Medium	<ul style="list-style-type: none"> • 375 total virtual machines • 25 virtual machines per host 	35MB	Yes ²	No
Large	<ul style="list-style-type: none"> • 1,280 total virtual machines • 40 virtual machines per host 	120MB	No	No
Enterprise	<ul style="list-style-type: none"> • 3,000 total virtual machines • 512 virtual machines per host 	300MB	No	No

The preceding recommendations are sufficient for most environments. If you notice that the HA log history does not span one week after implementing the recommended settings in the preceding table, consider increasing the capacity beyond what is noted.

Increasing the log capacity for HA involves specifying the number of log rotations that are preserved and the size of each log file in the rotation. For log capacities up to 30MB use a 1MB file size; for log capacities greater than 30MB, use a 5MB file size.

1. The default log settings are sufficient for ESXi hosts that are logging to persistent storage.

2. The default log setting is sufficient for ESXi 5.0 hosts if the following conditions are met: (i) they are not managed by Auto Deploy and (ii) they are configured with the default log location in a scratch directory on a VMFS partition.

Configuring Log Capacity

The mechanism used to configure the VMware HA agent logging depends on the ESX host version and the logging mechanism used. In all cases, before increasing the log capacity, you should verify that the locations where the log files are being written have sufficient space and that logging is being done to persistent storage.

When using a third-party syslog server, refer to the syslog server documentation for instructions on increasing the log capacity. ESX does not configure or control third-party syslog servers.

To configure logging for hosts predating ESXi 5.0, use the following HA options. These options are set using the HA advanced options dialog in the vSphere Client.

ADVANCED OPTION	SETTING TO USE
das.config.log.maxFileNum	The desired number of log rotations.
das.config.log.maxFileSize	The desired file size in bytes.
das.config.log.directory	The full directory path used to store the log files. Use this option if you need to change the location.

After changing the advanced options, reconfigure HA on each host in the cluster. The log values you configure in this manner will be preserved across vCenter Server updates. However, applying an update that includes a new version of the HA agent will require HA to be reconfigured on each host for the configured values to be reapplied.

Configuring logging on ESXi 5.0 hosts involves consideration of many environmental details. Refer to the following recommended information source for more in-depth information.

NOTE: The name of the VMware HA logger is Fault Domain Manager (FDM).

General Logging Recommendations for All ESX Versions

- Ensure that the location where the log files will be stored has sufficient space available.
- For ESXi hosts, ensure that logging is being done to a persistent location.
- When changing the directory path, ensure that it is present on all hosts in the cluster and is mapped to a different directory for each host.
- Configure each HA cluster separately.
- If a cluster contains 5.0 and earlier host versions, setting the **das.config.log.maxFileNum** advanced option will cause the 5.0 hosts to maintain two copies of the log files, one maintained by the 5.0 logging mechanism discussed in the ESXi 5.0 documentation (see the following) and one maintained by the pre-5.0 logging mechanism, which is configured using the advanced options previously discussed.

Multiple sources of information exist that provide additional details on the topics mentioned here. The following sources are recommended for more detailed information:

- For further information on configuring logging for ESXi 5.0 hosts, see “Providing Sufficient Space for System Logging” in the *vSphere 5.0 Installation and Setup* documentation.
- See the following VMware knowledge base articles for more information on logging:
 - 1033696 – “Creating a Persistent Scratch Location for ESXi”
 - 1016621 – “Enabling Syslog on ESXi”
 - 1021801 – “Location of ESXi Log Files”

Summary

vSphere High Availability greatly simplifies virtual machine provisioning, resource allocation, load balancing and migration while also providing an easy-to-use, cost-effective high-availability and failover solution for applications running in virtual machines. Using VMware vSphere 5.0 and VMware HA helps eliminate single points of failure in the deployment of business-critical applications in virtual machines. It also enables you to maintain other inherent virtualization benefits such as higher system utilization; closer alignment of IT resources with business goals and priorities; and more streamlined, simplified and automated administration of larger infrastructure installations and systems.

