VMware vSphere® High Availability 5.0 Deployment Best Practices

TECHNICAL MARKETING DOCUMENTATION
UPDATED JANUARY 2013

vmware®

Table of Contents

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. VMware vSphere High Availability (HA) specifically reduces unplanned downtime by leveraging multiple VMware vSphere ESXi™ hosts configured as a cluster, to provide rapid recovery from outages and cost-effective high availability for applications running in virtual machines.

vSphere 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.
- It provides the infrastructure to protect all workloads within the cluster, in contrast to other clustering solutions.

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.

Users can combine HA with VMware 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 vSphere HA has over traditional failover solutions:

- Minimal setup
- Reduced complexity (in other words, 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 5.0 Availability Guide* for more information on the basics of HA, including methods for creating HA clusters, how HA works, the benefits of integrating HA with DRS and explanations on configuration procedures.

The discussions and recommendations contained in this document relate to vSphere HA 5.0 and the 5.0 update releases. Although some of the discussion applies to earlier vSphere releases, the HA guides for those releases are a better primary reference.

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 failures occurring anywhere in the environment, they can affect both hardware and software. Building redundancy at vulnerable points helps reduce or eliminate downtime caused by hardware failures. These include redundancies at the following layers:

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

Host Considerations

Proper 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 ensure that rack or chassis failure cannot impact an entire cluster.

When deploying a vSphere HA cluster, it is a best practice to build the cluster out of identical server hardware. The use of 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. The use of drastically different hardware leads to an unbalanced cluster, as described in the Admission Control section. By default, vSphere HA prepares for the worst-case scenario in which the largest host in the cluster fails. 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 the mounting of datastores to a subset of the cluster hosts or the implementation of VSphere DRS–required virtual machine–to-host affinity rules are scenarios to consider carefully. The avoidance of these conditions will increase the portability of the virtual machine and provide a higher level of availability. Users can check a given virtual machine for inconsistencies 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 handle failures adequately. For example, to ensure that a cluster of three nodes can tolerate a single host failure, about 33 percent of the cluster resources are 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 management complexity of the cluster. This complexity is associated with general configuration factors as well as ongoing management tasks such as troubleshooting. However, this increase in management complexity is overshadowed by the benefits a large cluster can provide. Features such as vSphere DRS and VMware vSphere Distributed Power Management™ (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 in which 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 employ features not supported for earlier host versions and to leverage capabilities that are improved. A good example of this is the support added in vSphere 5.0 for management network partitions. This feature is supported if the cluster contains only ESXi 5.0 hosts.

Mixed clusters are supported but not recommended because there is some differences in vSphere HA performance between host versions and these differences can introduce operational variances in a cluster. These differences arise from the fact that earlier host versions do not offer the same capabilities as later versions. For example, VMware 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 vSphere 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 apply if using a vSphere HA–enabled cluster that includes hosts with differing versions:

- Users should be aware of the general limitations of using a mixed cluster, as previously mentioned.
- Users should also know that ESXi 3.5 hosts within a 5.0 cluster must include a patch to address an issue involving file locks. For ESX 3.5 hosts, users must apply the ESX350-201012401-SG patch. For ESXi 3.5, they must apply the ESXe350-201012401-I-BG patch. Prerequisite patches must be applied before applying these patches. HA will not enable an ESX/ESXi 3.5 host to be added to the cluster if it does not meet the patch requirements.
- Users should avoid deploying mixed clusters if VMware vSphere Storage vMotion® or VMware vSphere Storage DRS™ is required. The *vSphere 5.0 Availability Guide* has more information on this topic.

Host Placement

Versions of vSphere prior to 5.0 included the use of primary and secondary hosts, with a limit of five primary hosts. This construct was the focus of many best-practice recommendations. In vSphere HA 5.0 this construct has been eliminated. Instead, vSphere 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 that a host acting as a master fails, there is an election process to select a new master.

This new construct for HA eliminates previous concerns regarding the following issues:

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

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.

The *Auto Deploy Best Practices Guide* has specific guidance for designing high availability into an environment with Auto Deploy.

VMware vCenter Server Availability Considerations

VMware vCenter Server™ is the management focal point for any vSphere environment. Although vSphere HA will continue to protect any environment without vCenter Server, the ability to manage the environment is severely impacted without it. It is highly recommended that users protect their vCenter Server instance as well as possible. The following methods can help to accomplish this:

- Use of VMware vCenter™ Server Heartbeat™—a specially designed high availability solution for vCenter Server
- Use of vSphere HA—useful in environments in which the vCenter Server instance is virtualized, such as when using the VMware vCenter Server Appliance™

Which option users choose depends on their 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 for ensuring the availability of vCenter Server as it pertains to the use of Auto Deploy. These are discussed in detail in the *Auto Deploy Best Practices Guide*. However, it bears repeating here that if vSphere HA is used to make vCenter Server highly available, 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 "must run on" 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 vSphere DRS be enabled if users employ 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 for network design fall into two specific areas. The first involves increasing resiliency of client-side networking to ensure access from external systems to workloads running in vSphere and the second involves increasing resiliency of communications used by HA itself.

General Networking Guidelines

The following suggestions are best practices for configuring networking to increase availability:

- If the physical network switches that connect the servers support the PortFast (or an equivalent) setting, this should be enabled. If this feature is not enabled, it can take a while for a host to regain network connectivity after booting due to the execution of lengthy spanning tree algorithms. While this execution is occurring, virtual machines cannot run on the host and HA will report the host as isolated or dead. Isolation will be reported if the host and an FDM master can access the host's heartbeat datastores. More information on this option is available in the documentation provided by the networking switch vendor.
- Host monitoring should be disabled 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 HA 5.0, all dependencies on DNS have been removed. This makes 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 can be resolved properly through DNS.
- Users should employ consistent port group names and network labels on VLANs for public networks. Virtual machines use port group names to reconfigure access to the network. If users employ inconsistent names for the original server and the failover server, virtual machines are disconnected from their networks after failover. Network labels are used by virtual machines to reestablish network connectivity upon restart. Use of a documented naming scheme is highly recommended. Issues with port naming can be completely mitigated by use of a VMware vSphere Distributed Switch™.
- Configure the management networks so that the vSphere HA agent on a host in the cluster can reach the agents on any of the other hosts using one of the management networks. Without such a configuration, a network partition condition can occur after a master host is elected.
- Configure the fewest possible number of hardware segments between the servers in a cluster. This limits single points of failure. Additionally, routes with too many hops can cause networking packet delays for heartbeats and increase the possible points of failure.
- In environments where both IPv4 and IPv6 protocols are used, the user should configure the distributed 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 on all network switches and firewalls that are used by the hosts for interhost communication. vSphere HA will open these ports automatically when enabled and close them when disabled. User action is required only if there are firewalls in place between hosts within the cluster, as in a stretched cluster configuration.
- Configure redundant management networking from ESXi hosts to network switching hardware if possible along with heartbeat datastores. Using network adaptor teaming will enhance overall network availability.
- Configuration of hosts with management networks on different subnets as part of the same cluster is supported. One or more isolation addresses for each subnet should be configured accordingly. Refer to the Host Isolation section for more details.
- The management network supports the use of jumbo frames as long as the MTU values and physical network switch configurations are set correctly. Ensure that the network supports jumbo frames end to end.

Setting Up Redundancy for vSphere HA Networking

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

NOTE: Because this document is primarily focused on vSphere 5.0, its use of the term "management network" 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 when using VMware ESX® 4.1, ESX 4.0, or ESX 3.5x.

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. The cluster is more resilient to failures because the hosts are connected to each other through two network adaptors and through two separate switches and thus they have two independent paths for cluster communication. To configure a network adaptor team for the management network, it is recommended to configure the vNICs in the distributed switch configuration for the ESXi host in an active/standby configuration. This is illustrated in the following example:

Requirements:

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

The distributed switch should be configured as follows:

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

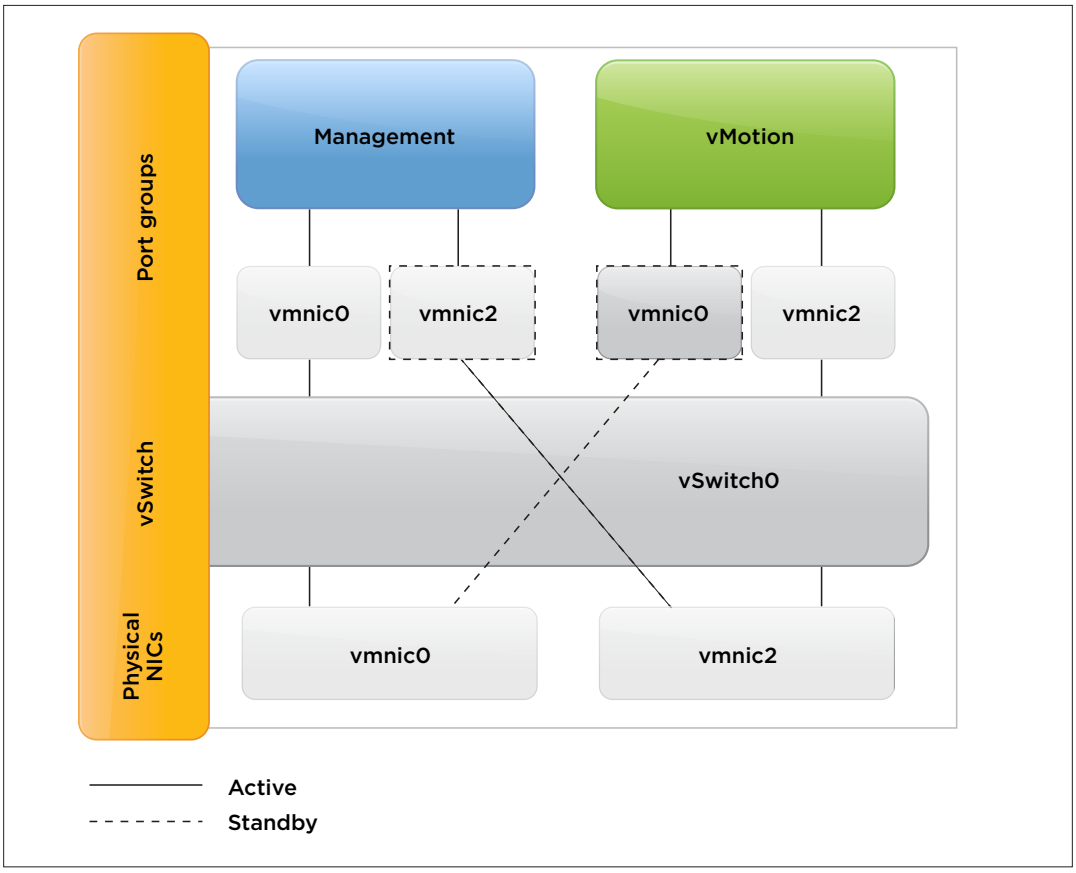

Figure 1.

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. It is recommended to use NIC ports from different physical NICs and it is preferable that the NICs are different makes and models.

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 determine that the switch is back online when its ports first come online. However, the switch itself might not be forwarding any packets until it is fully online. Therefore, when failback is set to no and an issue arises, both the management network and vMotion network will be running on the same network adaptor and will continue running until the user manually intervenes.

Management Network Changes in a vSphere HA Cluster

vSphere HA uses the management network as its primary communication path. As a result, it is critical that proper precautions are 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 host-monitoring functionality of vSphere HA should be disabled. This will prevent HA from determining that the maintenance action is a failure and from consequently triggering the isolation responses.

If there are changes involving the management network, it is advisable to reconfigure HA on all hosts in the cluster 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 when the host is not in maintenance mode.

Storage Design Considerations

Best practices for storage design reduces the likelihood of hosts losing connectivity to the storage used by the virtual machines, and that used by vSphere HA for heartbeating. To maintain a constant connection between an ESXi host and its storage, ESXi supports multipathing, a technique that enables users to employ 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, 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, which is the process of distributing I/O loads across multiple physical paths. Load balancing reduces or removes potential bottlenecks.

For Fibre Channel (FC) 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 vSphere HA in vSphere 5.0 makes it possible to use storage subsystems as a means of communication between the hosts of a cluster. Storage heartbeats are used when the management network is unavailable to enable a slave HA agent to communicate with a master HA agent. The feature also makes it possible to distinguish accurately between the different failure scenarios of dead, isolated or partitioned hosts. Storage heartbeats enable detection of cluster partition scenarios that are not supported with previous versions of vSphere. This results in a more coordinated failover when host isolation occurs.

By default, vCenter Server will select automatically two datastores to use for storage heartbeats. The algorithm is designed to maximize availability and redundancy of the storage heartbeats. It is intended to select datastores that are connected to the highest number of hosts. It is also designed to select datastores that are backed by different LUNs or NFS servers. A preference is given to VMware vSphere VMFS–formatted datastores over NFS-hosted datastores. With this weighting, it is unlikely for a host to lose both its management network and access to its heartbeat datastores after a network interruption. vCenter Server 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.

By default, vCenter Server selects the heartbeat datastores for each host from the set mounted by that host. Users can, however, configure vSphere HA to give preference to a subset of the datastores mounted by the hosts in the cluster. Alternately, they can require that HA choose only from a subset of these. VMware recommends the users employ the default setting unless there are datastores in the cluster that are more highly available than others. If there are some more highly available datastores, VMware recommends that users configure vSphere HA to give preference to these. For example, specifying a datastore preference is suggested if certain datastores are hosted by more reliable storage arrays than others, or if there are redundant paths for some but not all datastores. Another example would be if some datastores were more likely than others to remain accessible to all hosts in the cluster after a management network failure. VMware does not recommend restricting vSphere HA to using only a subset of the datastores because this setting restricts the system's ability to respond when a host loses connectivity to one of its configured heartbeat datastores.

NOTE: vSphere HA datastore heartbeating is very lightweight and will not impact in any way the use of the datastores by virtual machines.

Although users can increase to four the number of heartbeat datastores chosen for each host, increasing the number does not make the cluster significantly more tolerant of failures. (See the *vSphere Metro Storage Cluster* white paper for details about heartbeat datastore recommendations specific to stretched clusters.)

Environments that provide only network-based storage must work optimally with the network architecture to realize fully the potential of the storage heartbeat feature. If the storage network traffic and the management network traffic flow through the same network components, disruptions in network service might disrupt both. It is recommended that these networks be separated as much as possible or that datastores with a different failure domain be used for heartbeating. In cases where converged networking is used, VMware recommends that users leave heartbeating enabled. This is because even with converged networking failures can occur that disrupt only the management network traffic. For example, the VLAN tags for the management network might be incorrectly changed without impacting those used for storage traffic.

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 can then run on any of the hosts within the cluster. Such a configuration is also beneficial because it maximizes the chance that an isolated or partitioned host can communicate with a master during a network partition or isolation event. If network partitions or isolations are anticipated within the environment, users should ensure that a minimum of two shared datastores is provisioned to all hosts in the cluster.

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

vSphere 5.0 Update 1 Inaccessible Datastore Enhancements

vSphere 5.0 Update 1 introduced enhancements that enable better handling of storage connectivity failures. These include deploying an automated failover of virtual machines residing on a datastore that has a Permanent Device Loss (PDL) condition and for the automated restart of virtual machines that fail during an All Paths Down (APD) condition on their home datastore. These enhancements are detailed in KB 2015681 (*HA Implications of APD and PDL in vSphere 5.0 and 5.0 Update 1*), which discusses the enhancements and the caveats with enabling them. It is recommended to deploy these enhancements in clusters where the caveats are acceptable.

A PDL is a condition that is communicated by the array to ESXi via an SCSI sense code. This condition indicates that a device (LUN) is unavailable and likely permanently unavailable. An example would be when a LUN is set to offline. (This condition is used in specific active/active array replication technologies such as EMC VPLEX during failure scenarios to ensure that ESXi takes appropriate action when access to a LUN is revoked.)

NOTE: When a full storage failure occurs, it is impossible to reach the PDL condition because there is no communication possible between the array and the ESXi host. This state will be identified by the ESXi host as an APD condition.

To enable vSphere HA to respond to a PDL condition, two advanced settings have been introduced in vSphere 5.0 Update 1. The first setting is configured on a host level and is disk.terminateVMOnPDLDefault. This behavior is configured in /etc/vmware/settings and should be left at the default True setting. This setting ensures that a virtual machine is killed when the datastore where it resides is in a PDL state. The virtual machine is killed as soon as it initiates disk I/O on a datastore that is in a PDL condition and when all of the virtual machine files reside on this datastore.

NOTE: Virtual machines are killed only when issuing I/O to the datastore. As long as the virtual machine is not issuing I/O to the datastore, the virtual machine remains running. Virtual machines that are running memory-intensive workloads without issuing I/O to the datastore might remain active in such situations.

The second setting is a vSphere HA advanced setting called das.maskCleanShutdownEnabled. This setting is not enabled by default and it must be set to True. This setting enables vSphere HA to trigger a restart response for a virtual machine that has been killed automatically due to a PDL condition.

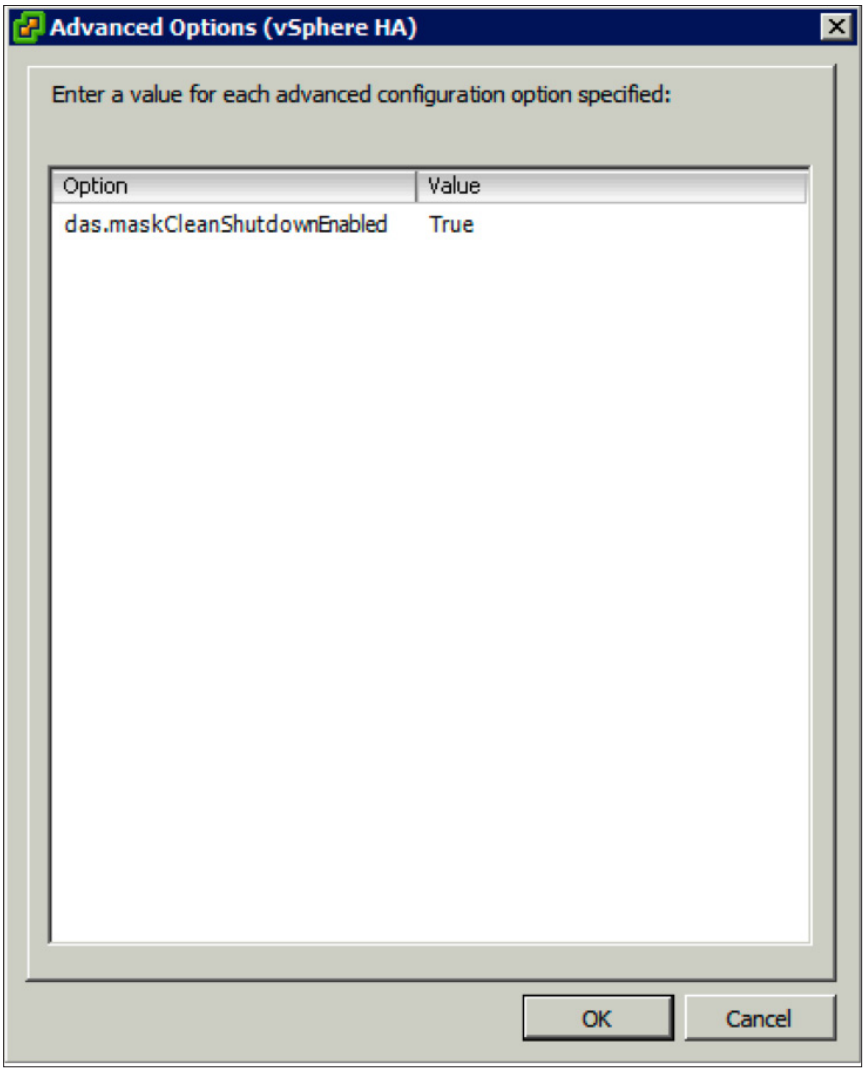

Figure 2.

Cluster Configuration Considerations

Many options are available when configuring a vSphere HA environment. This flexibility enables users to leverage vSphere HA in a wide range of environments while accommodating particular requirements. This section includes a discussion of some of the key cluster configuration options and the recommended best practices for using them.

Host Isolation

One key mechanism within vSphere HA is the ability for a host to detect when it has become network-isolated from the rest of the cluster. With this information, vSphere is able to take administrator-specified action with respect to running virtual machines on the host that has been isolated. Depending on network layout and specific business needs, the administrator might wish to tune the vSphere HA response to an isolated host to favor rapid failover or to leave the virtual machine running so clients can continue to access it.

The following section explains how a vSphere HA node detects when it has been isolated from the rest of the cluster, and the response options available to that node after that determination has been made.

Host Isolation Detection

Host isolation detection happens at the individual host level. Isolation fundamentally means a host is no longer able to communicate over the management network. To determine if it is network-isolated, the host attempts to ping its configured isolation addresses. The isolation address used should always be reachable by the host under normal situations, because after five seconds have elapsed with no response from the isolation addresses, the host then declares itself isolated.

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 0–9) 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.

Host Isolation Response

Tuning the host isolation response is typically based on whether loss of connectivity to a host via the management network would typically also indicate that clients accessing the virtual machine would also be affected. In this case it is likely that administrators would want the virtual machines shut down so other hosts with operational networks can start them up. If failures of the management network are not likely correlated with failures of the virtual machine network, where the loss of the management network simply results in the inability to manage the virtual machines on the isolated host, it is often preferable to leave the virtual machines running while the management network connectivity is restored.

The Host Isolation Response setting provides a means to set the action preferred for the powered-on virtual machines maintained by a host when that host has declared it is 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, a virtual machine might still have the ability to access the storage subsystem and the virtual machine network. By selecting this option, the user enables the virtual machine to continue to function if this were to occur. This is the default isolation response setting in vSphere HA 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 vSphere HA will attempt to restart the virtual machine more quickly than when using the Shut Down option.

Shut Down

Through the use of the VMware Tools™ package installed within the guest OS of a virtual machine, this option attempts to shut down the OS gracefully 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 time period of 300 seconds (five 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 shut down an OS gracefully. The Shut Down option requires that the VMware 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 only network-based storage protocols, such as iSCSI and NFS, and those that share physical network components between the management and storage traffic, the recommended isolation response is Power Off. With these environments, it is likely that a network outage causing a host to become isolated will also affect the host's ability to communicate to the datastores. If the host lost access to the datastores, a master HA agent would be able to power on a second instance of any virtual machine while the first is still running on the isolated host. This situation might be problematic if both instances of the virtual machine retain access to the virtual machine network. The Power Off isolation response recommendation reduces the impact of this issue by having the isolated HA agent power off the virtual machines on the isolated host.

The following table lists the recommended isolation policy for converged network configurations:

LIKELIHOOD THAT HOST WILL RETAIN ACCESS TO VIRTUAL MACHINE DATASTORES	LIKELIHOOD THAT VIRTUAL MACHINES WILL RETAIN ACCESS TO VIRTUAL MACHINE NETWORK	RECOMMENDED ISOLATION POLICY	RATIONALE
Likely	Likely	Leave Powered On	Virtual machine is running fine, so why power it off?
Likely	Unlikely	Either Leave Powered On or Shut Down	Choose Shut Down to enable vSphere HA to restart virtual machines on hosts that are not isolated and hence are likely to have access to storage.
Unlikely	Likely	Power Off	Avoid having two instances of the same virtual machine on the virtual machine network.

Table 1.

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 host with access to the network and datastores can power on the virtual machines.

Host Monitoring

The host monitoring setting determines whether vSphere HA restarts virtual machines on other hosts in the cluster after a host isolation, a host failure or after they should crash for some other reason. This setting does not impact the VM/application monitoring feature. If host monitoring is disabled, isolated hosts won't apply the configured isolation response, and vSphere HA won't restart virtual machines that fail for any reason. Disabling host monitoring also impacts VMware vSphere Fault Tolerance (FT) because it controls whether HA will restart an FT secondary virtual machine after a failure event.

Cluster Partitions

vSphere HA 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 the use of a stretched cluster configuration. 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 is still able to communicate to a master node. The other subset of hosts cannot. For this reason, the second subset will go through an election process and elect a new master node. Therefore, it is possible to have multiple master nodes in a cluster partition scenario, with one per partition. This situation will last only as long as the partition exists. After the network issue causing the partition is resolved, the master nodes will be able to communicate and discover multiple master roles. Anytime multiple master nodes exist and can communicate with each other over the management network, all but one will abdicate. The same algorithm that elects a master node when there is no master determines which master should survive after a partition ends.

Robust management network architecture helps to avoid cluster partition situations. Refer to the networking section of this document for recommended practices on this topic. Additionally, if a network partition occurs, users should ensure that each host retains access to its heartbeat datastores, and that the masters are able to access the heartbeat datastores used by the slave hosts. Refer to the earlier storage section in this document for recommended practices on this topic.

vSphere Metro Storage Cluster Considerations

VMware vSphere Metro Storage Clusters (vMSC), or stretched clusters as they are often called, are environments that span multiple sites within a metropolitan area (typically up to 100km). Storage systems in these environments typically enable a seamless failover between sites. Because this a complex environment, a paper specific to the vMSC has been produced. Download it here: http://www.vmware.com/resources/techresources/10299

Auto Deploy Considerations

Auto Deploy can provision hundreds of physical hosts with ESXi software. Users can specify the image to deploy and the hosts to provision with the image. When a physical host that is configured for Auto Deploy is powered on, Auto Deploy utilizes a PXE boot infrastructure to provision a host automatically. No host-state information is stored on the host itself. Instead, the vCenter Server manages state information for each host. For a host to boot there is a dependency on several components. When it comes to unplanned downtime, it is especially important to realize which components are critical. The best practices recommendation from VMware staff for environments using Auto Deploy is as follows:

- Deploy vCenter Server Heartbeat. vCenter Server Heartbeat delivers high availability for vCenter Server, protecting the virtual and cloud infrastructure from application-, configuration-, OS- or hardware-related outages.
- Avoid using Auto Deploy in stretched cluster environments, because this complicates the environment.
- Deploy vCenter Server in a virtual machine. Run the vCenter Server virtual machine in a vSphere HA–enabled cluster and configure the virtual machine with a vSphere HA restart priority of high. Perform one of the following actions:
  - Include two or more hosts in the cluster that are not managed by Auto Deploy and pin the vCenter Server virtual machine to these hosts by using a rule (vSphere DRS–required virtual machine–to-host rule). Users can set up the rule and then disable DRS if they do not wish to use DRS in the cluster.
  - Deploy vCenter Server and Auto Deploy in a separate management environment, that is, by hosts managed by a different vCenter server.

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 vSphere HA agent on a host to detect heartbeat information on a virtual machine through VMware Tools or an agent running within the virtual machine that is monitoring the application health. After the loss of a defined number of VMware Tools heartbeats on the virtual machine, vSphere 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 vSphere HA detects that the preferred state of the virtual machine is to be powered on. For this reason, HA will attempt to restart the virtual machine 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 vSphere HA agent on a host is in the uninitialized state, as would occur immediately after the vSphere HA agent has been installed on the host or when the host is not available. Additionally, the number of missed heartbeats is reset after the vSphere HA agent on the host reboots. This should occur rarely if at all, or after vSphere 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 virtual machines must have the VMware Tools package installed within the guest OS.

NOTE: Guest OS sleep states are not currently supported by virtual machine monitoring and can trigger an unnecessary restart of the virtual machine.

vSphere HA and vSphere FT

Often vSphere HA is used in conjunction with vSphere FT. FT provides protection for extremely critical virtual machines where any loss of service is intolerable. vSphere HA detects the use of FT to ensure proper operation. This section describes some of the unique behavior specific to vSphere FT with vSphere HA. Additional vSphere FT best practices can be found in the *vSphere 5.0 Availability Guide*.

Host Partitions

vSphere HA will restart a secondary virtual machine of a vSphere 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 in 5.0 cannot be restarted until the partition ends.

Host Isolation

Host isolation responses are not performed on virtual machines enabled with vSphere FT. 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 and they are not heartbeating over the FT logging network, in which case one of them will then take over as a primary FT virtual machine. Because vSphere 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 resiliency 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 introduced 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 configuration options for choosing users' 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 several advanced configuration options. In addition, an advanced option can be used to specify the default slot size value for CPU. This value is used when no CPU reservation has been specified for a virtual machine. The value was changed in vSphere 5.0 from 256MHz to 32MHz. When no memory reservation is specified for a virtual machine, the largest memory overhead for any virtual machine in the cluster will be used as the default slot size value for memory. See the *vSphere Availability Guide* for more information on slot-size calculation and tuning.
- Percentage of Cluster Resources Reserved as failover spare capacity: vSphere HA ensures that a specified percentage of memory and CPU resources are reserved for failover. This policy is recommended for situations where the user must have 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 with this option the capacity set aside for failures can be fragmented across hosts.
- Specify a Failover Host: vSphere 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 HA 5.0. When a host is designated as a failover host, HA admission control does not enable the powering on of virtual machines on that host, and DRS will not migrate virtual machines to the failover host. It effectively becomes a hot standby.

With each of the three admission control policies there is a chance in specific scenarios that, at the time of failing over a virtual machine, there might be insufficient contiguous capacity available on a single host to power on a given virtual machine. Although these are corner case scenarios this has been taken into account and HA will request vSphere DRS, if it is enabled, to attempt to defragment the capacity in such situations. Further, if a host had been put into standby and vSphere DPM is enabled, it will attempt to power up a host if defragmentation is not sufficient.

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

- Select the Percentage of Cluster Resources Reserved policy for admission control. This policy offers the most flexibility in terms of host and virtual machine sizing and is sufficient for most situations. When configuring this policy, the user should choose a percentage for CPU and memory that reflects the number of host failures they wish to support. For example, if the user wants vSphere HA to set aside capacity for two host failures and there are 10 hosts of equal capacity in the cluster, then they should specify 20 percent (2/10). If there are not equal capacity hosts, then the user should specify a percentage that equals the capacity of the two largest hosts as a percentage of the cluster capacity.
  - If the Host Failures Cluster Tolerates policy is used, 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 as a 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 virtual machine, limiting consolidation. See the *vSphere 5.0 Availability Guide* for more information on slot-size calculation and overriding slot-size calculation in cases where it is necessary to configure different-sized virtual machines in the same cluster.
  - If the Specify a Failover Host policy is used, decide how many host failures to support, and then specify this number of hosts as failover hosts. Ensure that all cluster hosts are sized equally. If unequally sized hosts are used with the Host Failures Cluster Tolerates policy, vSphere HA will reserve excess capacity to handle failures of the largest N hosts, where N is the number of host failures specified. With Percentage of Cluster Resources Reserved policy, unequally sized hosts will require that the user increase the percentages to reserve enough capacity for the planned number of host failures. Finally, with the Specify a Failover Host policy, users must specify failover hosts that are as large as the largest nonfailover hosts in the cluster. This ensures that there is adequate capacity in case of failures.

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 vSphere DRS to create more contiguous capacity on hosts. This increases 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.

The admission control policy is evaluated against the current state of the cluster, not the normal state of the cluster. The normal state means that all hosts are connected and healthy. Admission control does not take into account resources of hosts that are disconnected or in maintenance mode. Only healthy and connected hosts—including standby hosts, if vSphere DPM is enabled—can provide resources that are reserved for tolerating host failures. The amount of resources reserved for failure does not change based on the current state of the cluster. This results in admission control reserving the configured resources for failover across the remaining healthy hosts in the cluster. For example, if the cluster is configured with the policy Host Failures Cluster Tolerates set to tolerate one host failure, HA will effectively reserve the value of one host's resources for restarting failed virtual machines on healthy hosts in the cluster, regardless of the number of currently disconnected or failed hosts.

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 the term "must") and preferred rules (designated by the term "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 follows 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, vSphere DRS might be required to migrate other virtual machines to make space on the preferred hosts.

If required rules are specified, vSphere HA will restart virtual machines on an ESXi host in the same host DRS group only. 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 enabled changes in how logging is performed. Previous versions of HA stored the operational logging information across several distinct log files. In vSphere HA 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
Small	40 total virtual machines 8 virtual machines per host	4MB	Yes[1]
Medium	375 total virtual machines 25 virtual machines per host	35MB	Yes[2]
Large	1,280 total virtual machines 40 virtual machines per host	120MB	No
Enterprise	3,000 total virtual machines 512 virtual machines per host	300MB	No

Table 2.

The preceding recommendations are sufficient for most environments. If the user notices that the HA log history does not span one week after implementing the recommended settings in the preceding table, they should 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 vSphere VMFS partition.

Configuring Log Capacity

The mechanism used to configure the vSphere HA agent logging depends on the ESX host version and the logging mechanism used. In all cases, before increasing the log capacity, users 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.

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 vSphere 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.
- In vSphere 5.0, 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. In vSphere 5.0U1, this issue has been resolved. In this version, to maintain two sets of log files, the new HA advanced configuration option das.config.log.outputToFiles must be set to true, and das.config.log.maxFileNum must be set to a value greater than two.
- After changing the advanced options, reconfigure HA on each host in the cluster. The log values users 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.

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*

Conclusion

VMware 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 vSphere HA helps eliminate single points of failure in the deployment of business-critical applications in virtual machines. It also enables users 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.

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com

 Item No: VMW-WP-vSPHR-HA-DEPLOY-BEST-PRTC-USLET-101 Docsource: OIC - 12VM013.03