Chapter 11

Ensuring High Availability and Business Continuity

After all your servers are installed, all storage is provisioned, all virtual networking is pinging, and all virtual machines are running, it is time to define the strategies or the methods to put into place in a virtual infrastructure that will provide high availability and business continuity. The deployment of a virtual infrastructure opens many new doors for disaster-recovery planning. The virtual infrastructure administrator will lead the charge into a new era of ideologies and methodologies for ensuring that business continues as efficiently as possible in the face of corrupted data, failed servers, or even lost datacenters.

With the release of VMware vSphere, we have been given more tools at our disposal to reach our goal of increased uptime and recoverability of the infrastructure. You'll learn about the methods and new features available to reach this goal.

In this chapter, you will learn to:

- Understand Windows clustering and the types of clusters
- Understand built-in high availability options
- Understand the differences between VCB and VCDR
- Understand data replication options

Clustering Virtual Machines

Let's start with the most well-known technique for helping administrators achieve high availability: Microsoft Clustering Service (MSCS), or *failover clustering* as it is called in Windows 2008. Failover clustering in Windows 2008 is used when critical services and applications call for the highest levels of availability. Microsoft Windows Server 2003 and 2008 both support network load balancing (NLB) clusters as well as server clusters depending on the version of the Windows Server operating system that is installed on the server. Moving forward, I'll just use the term Microsoft Clustering Service or MSCS to describe any forms or versions of Windows clustering.

Microsoft Clustering

The NLB configuration involves an aggregation of servers that balances the requests for applications or services. In a typical NLB cluster, all nodes are active participants in the cluster and are consistently responding to requests for services. NLB clusters are most commonly deployed as a means of providing enhanced performance and availability. NLB clusters are best suited for scenarios involving Internet Information Services (IIS), virtual private networking (VPN), and Internet Security and Acceleration (ISA) server, to name a few. Figure 11.1 details the architecture of an NLB cluster.

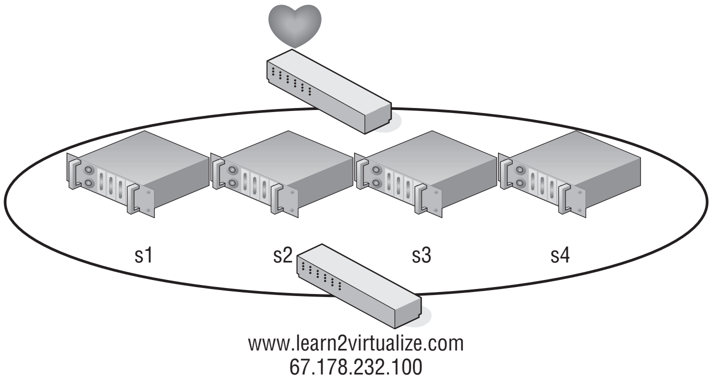


Figure 11.1
An NLB cluster can contain up to 32 active nodes that distribute traffic equally across each node. The NLB software allows the nodes to share a common name and IP address that is referenced by clients.

NLB Support from VMware

As of this writing, VMware supports NLB, but you will need to run NLB in multicast mode to support VMotion and virtual machines on different physical hosts. You will also need to configure static Address Resolution Protocol (ARP) entries on the physical switch to achieve this. If NLB is running in unicast mode, then the virtual machines will all need to be running on the same host. Another option to consider would be the use of third-party load balancers to achieve the same results.

Unlike NLB clusters, server clusters are used solely for the sake of availability. Server clusters do not provide performance enhancements outside of high availability. In a typical server cluster, multiple nodes are configured to be able to own a service or application resource, but only one node owns the resource at a given time. Server clusters are most often used for applications like Microsoft Exchange, Microsoft SQL Server, and DHCP services, which each share a need for a common datastore. The common datastore houses the information accessible by the node that is online and currently owns the resource, as well as the other possible owners that could assume ownership in the event of failure. Each node requires at least two network connections: one for the production network and one for the cluster service heartbeat between nodes. Figure 11.2 details the structure of a server cluster.

The different versions of Windows Server 2003 and 2008 offer various levels of support for NLB and server clusters. Table 11.1 outlines the cluster support available in each version of Windows Server 2003. The only difference in Windows 2008 is that a server cluster can have up to 16 nodes.

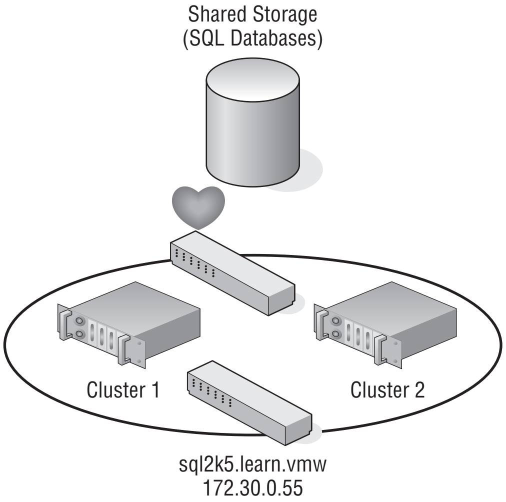
TABLE 11.1: Windows Server 2003 Clustering Support

OPERATING SYSTEM	NETWORK LOAD BALANCING	SERVER CLUSTER
Windows Server 2003/2008 Web Edition	Yes (up to 32 nodes)	No
Windows Server 2003/2008 Standard Edition	Yes (up to 32 nodes)	No
Windows Server 2003/2008 Enterprise Edition	Yes (up to 32 nodes)	Yes (up to 8 nodes in 2003 and 16 nodes in 2008)
Windows Server 2003/2008 Datacenter Edition	Yes (up to 32 nodes)	Yes (up to 8 nodes in 2003 and 16 nodes in 2008)

WINDOWS CLUSTERING STORAGE ARCHITECTURES

Server clusters built on Windows Server 2003 can support only up to eight nodes, and Windows 2008 can support up to 16 nodes when using a Fibre Channel–switched fabric. Storage architectures that use SCSI disks as direct attached storage or that use a Fibre Channel–arbitrated loop result in a maximum of only two nodes in a server cluster. Clustering virtual machines in an ESX/ESXi host utilizes a simulated SCSI shared storage connection and is therefore limited to only two-node clustering. In addition, in ESX 3.*x*, the clustered virtual machine solution uses only SCSI 2 reservations, not SCSI 3 reservations, and supports only the SCSI miniport drivers, not the Storport drivers. This has been changed in VMware vSphere, which now allows SCSI 3 reservations and the use of the Storport drivers.

FIGURE 11.2
Server clusters are best suited for applications and services like SQL Server, Exchange Server, DHCP, and so on, that use a common data set.



MSCS, when constructed properly, provides automatic failover of services and applications hosted across multiple cluster nodes. When multiple nodes are configured as a cluster for a service or application resource, as I said previously, only one node owns the resource at any given time. When the current resource owner experiences failure, causing a loss in the heartbeat between the cluster nodes, another node assumes ownership of the resource to allow continued access with minimal data loss. To configure multiple Windows Server nodes into a Microsoft cluster, the following requirements must be met:

- Nodes must be running either Windows Server Enterprise Edition or Datacenter Edition
- All nodes should have access to the same storage device(s)
- All nodes should have two similarly connected and configured network adapters: one for the production network and one for the heartbeat network
- All nodes should have Microsoft Cluster Services for the version of Windows that you are using

Virtual Machine Clustering Scenarios

The clustering of Windows Server virtual machines using Microsoft Cluster Services can be done in one of three different configurations. The following gives you a quick peek now, and I will get into more details in a minute:

Cluster in a box The clustering of two virtual machines on the same ESX/ESXi host is also known as a *cluster in a box*. This is the easiest of the three configurations to set up. No special configuration needs to be applied to make this configuration work.

Cluster across boxes The clustering of two virtual machines that are running on different ESX/ESXi hosts is known as a *cluster across boxes*. VMware had restrictions in place for this configuration in earlier versions: the cluster node's `C:` drive must be stored on the host's local storage or local VMFS datastore, the cluster shared storage must be stored on Fibre Channel external disks, and you must use raw device mappings on the storage. This has been changed and updated to allow `.vmdk` files on the SAN and to allow the cluster VM boot drive or `C:` drive on the SAN, but VMotion and Distributed Resource Scheduling (DRS) are not supported using Microsoft-clustered virtual machines. The exact warning from VMware is ''Clustered virtual machines cannot be part of VMware clusters (DRS or HA).''

Physical to virtual clustering The clustering of a physical server and a virtual machine together is often referred as a *physical to virtual cluster*. This configuration of using both physical and virtual servers together gives you the best of both worlds, and the only other added restriction is that you cannot use virtual compatibility mode with the RDMs. I'll cover these options in more detail and show how to set them up in a virtual environment later in this chapter.

Clustering has long been considered an advanced technology implemented only by those with high technical skills in implementing and managing high-availability environments. Although this might be more rumor than truth, it is certainly a more complex solution to set up and maintain.

Although you might achieve results setting up clustered virtual machines, you may not receive support for your clustered solution if you violate any of the clustering restrictions put forth by VMware. The following list summarizes and reviews the do's and don'ts of clustering virtual machines as published by VMware:

- 32-bit and 64-bit virtual machines can be configured as nodes in a server cluster.
- Majority Node Set clusters with application-level replication (for example, Microsoft Exchange 2007 Cluster Continuous Replication) is now supported.

- Only two-node clustering is allowed.
- Clustering is not supported on iSCSI or NFS disks.
- Clustering does not support NIC teaming in the virtual machines.
- Virtual machines configured as cluster nodes must use the LSI Logic SCSI adapter and the vmxnet network adapter.
- Virtual machines in a clustered configuration are not valid candidates for VMotion, and they can’t be part of a DRS or HA cluster.
- ESX/ESXi hosts that run virtual machines that are part of a server cluster can now be configured to perform a boot from SAN.
- ESX/ESXi hosts that run virtual machines that are part of a server cluster cannot have both QLogic and Emulex HBAs.

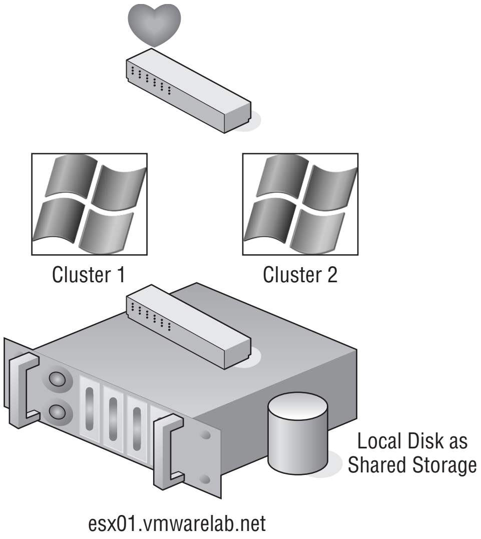
There is something else that you need to do. You must set the I/O timeout to 60 seconds or more by modifying HKLM\System\CurrentControlSet\Services\Disk\TimeOutValue, and if you re-create a cluster, you need to reset the value again.

So, let’s get into some more details on clustering and look at the specific clustering options available in the virtual environment. I will start with the most basic design configuration, the cluster in a box.

Examining Cluster-in-a-Box Scenarios

The cluster-in-a-box scenario involves configuring two virtual machines hosted by the same ESX/ESXi host as nodes in a server cluster. The shared disks of the server cluster can exist as `.vmdk` files stored on local VMFS volumes or on a shared VMFS volume. Figure 11.3 details the configuration of a cluster in a box.

FIGURE 11.3
A cluster-in-a-box configuration does not provide protection against a single point of failure. Therefore, it is not a common or suggested form of deploying Microsoft server clusters in virtual machines.



After reviewing the diagram of a cluster-in-a-box configuration, you might wonder why you would want to deploy such a thing. The truth is, you wouldn’t want to deploy cluster-in-a-box configuration because it still maintains a single point of failure. With both virtual machines running on the same host, if that host fails, both virtual machines fail. This architecture contradicts the very reason for creating failover clusters. A cluster-in-a-box configuration still contains a single point of failure that can result in downtime of the clustered application. If the ESX/ESXi host hosting the two-node cluster-in-a-box configuration fails, then both nodes are lost, and a failover does not occur. This setup might, and I use *might* loosely, be used only to ''play'' with clustering services or to test clustering services and configurations. But ultimately, even for testing, it is best to use the cluster-across-box configurations to get a better understanding of how this might be deployed in a production scenario.

Configuration Options for Virtual Clustering

As suggested in the first part of this chapter, server clusters are deployed for high availability. High availability is not achieved by using a cluster-in-a-box configuration, and therefore this configuration should be avoided for any type of critical production applications and services.

Examining Cluster-Across-Boxes Configurations

Although the cluster-in-a-box scenario is more of an experimental or education tool for clustering, the cluster-across-boxes configuration provides a solid solution for critical virtual machines with stringent uptime requirements—for example, the enterprise-level servers and services like SQL Server and Exchange Server that are heavily relied on by the bulk of end users. The cluster-across-boxes scenario, as the name applies, draws its high availability from the fact that the two nodes in the cluster are managed on different ESX/ESXi hosts. In the event that one of the hosts fails, the second node of the cluster will assume ownership of the cluster group, and its resources and the service or application will continue responding to client requests.

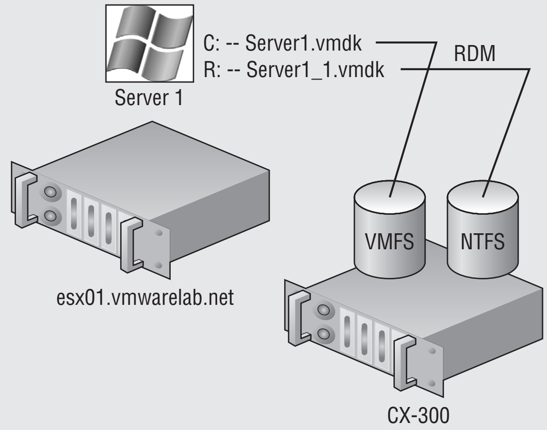
The cluster-across-boxes configuration requires that virtual machines have access to the same shared storage, which must reside on a Fibre Channel storage device external to the ESX/ESXi hosts where the virtual machines run. The virtual hard drives that make up the operating system volume of the cluster nodes can be a standard VMDK implementation; however, the drives used as the shared storage must be set up as a special kind of drive called a *raw device mapping* (RDM). An RDM is a feature that allows a virtual machine to establish direct access to a LUN on a SAN device.

Using Raw Device Mappings in your Virtual Clusters

An RDM is not a direct access to a LUN, and it is not a normal virtual hard disk file. An RDM is a blend between the two. When adding a new disk to a virtual machine, as you will soon see, the Add Hardware Wizard presents the RDMs as an option on the Select a Disk page. This page defines the RDM as having the ability to give a virtual machine direct access to the SAN, thereby allowing SAN management. I know this seems like a contradiction to the opening statement of this sidebar; however, I'm getting to the part that, oddly enough, makes both statements true.

By selecting an RDM for a new disk, you're forced to select a compatibility mode for the RDM. An RDM can be configured in either Physical Compatibility mode or Virtual Compatibility mode. The Physical Compatibility mode option allows the virtual machine to have direct raw LUN access. The

Virtual Compatibility mode, however, is the hybrid configuration that allows raw LUN access but only through a VMDK file acting as a proxy. The following image details the architecture of using an RDM in Virtual Compatibility mode.



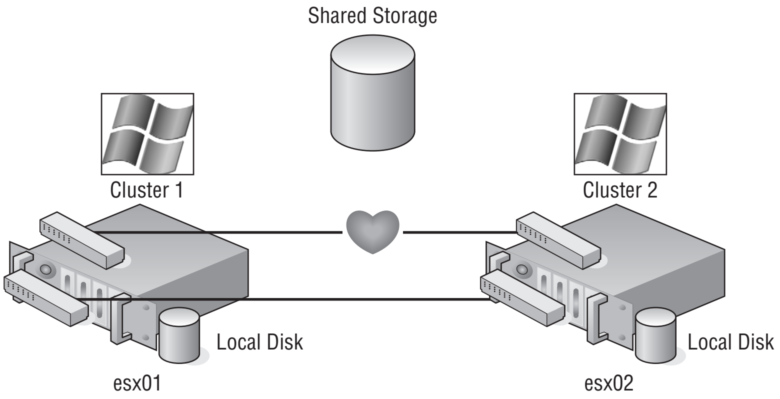
So, why choose one over the other if both are ultimately providing raw LUN access? Because the RDM in Virtual Compatibility mode uses a VMDK proxy file, it offers the advantage of allowing snapshots to be taken. By using the Virtual Compatibility mode, you will gain the ability to use snapshots on top of the raw LUN access in addition to any SAN-level snapshot or mirroring software. Or, of course, in the absence of SAN-level software, the VMware snapshot feature can certainly be a valuable tool. The decision to use Physical Compatibility or Virtual Compatibility is predicated solely on the opportunity and/or need to use VMware snapshot technology or when using physical to virtual clustering.

A cluster-across-box configuration requires a more complex setup than a cluster-in-a-box configuration. When clustering across boxes, all proper communication between virtual machines and all proper communication from virtual machines and storage devices must be configured properly. Figure 11.4 provides details on the setup of a two-node virtual machine cluster-across-box configurations using Windows Server guest operating systems.

Make sure you document things well when you start using RDMs. Any storage that is presented to ESX and is not formatted with VMFS will show up as available storage. If all the administrators are not on the same page, it is easy to take a LUN that was used for an RDM and reprovision that LUN as a VMFS datastore, effectively blowing away the RDM data in the process. I have seen this mistake happen firsthand, and let me tell you, the process is very quick to erase any data that is there. I have gone so far as to create a separate column in vCenter Server to list any RDM LUNs that are configured to make sure everyone has a reference point to refer to.

Let's keep moving and perform the following steps to configure Microsoft Cluster Services on Windows 2003 across virtual machines on separate ESX/ESXi hosts.

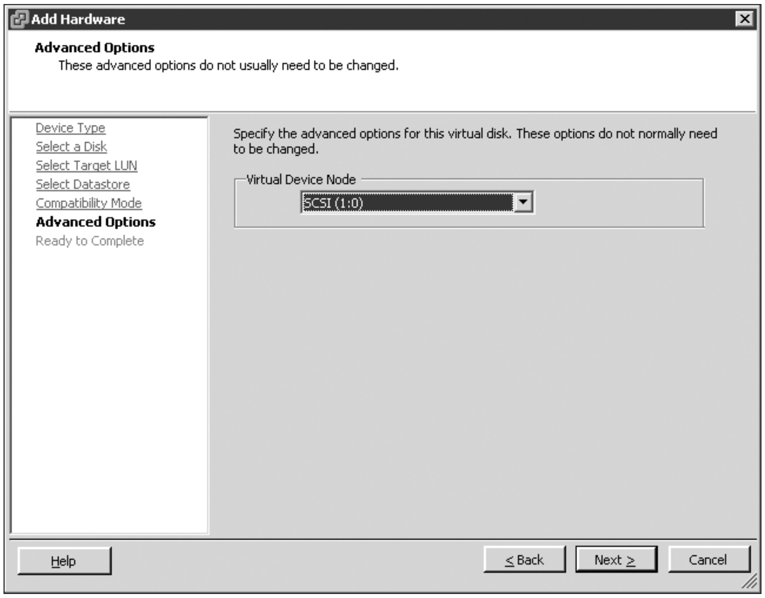
FIGURE 11.4
A Microsoft cluster built on virtual machines residing on separate ESX hosts requires shared storage access from each virtual machine using an RDM.



CREATING THE FIRST CLUSTER NODE IN WINDOWS 2003

Perform the following steps to create the first cluster node:

1. Inside the vSphere client, create a virtual machine that is a member of a Windows Active Directory domain.
2. Right-click the new virtual machine, and select the Edit Settings option.
3. Click the Add button, and select the Hard Disk option.
4. Select the Raw Device Mappings radio button, and then click the Next button.
5. Select the appropriate target LUN from the list of available targets.
6. Select the datastore location where the VMDK proxy file should be stored, and then click Next.
7. Select the Virtual radio button to allow VMware snapshot functionality for the RDM, and then click Next.
8. Select the virtual device node to which the RDM should be connected, as shown in Figure 11.5, and then click Next.
9. Click the Finish button.
10. Right-click the virtual machine, and select the Edit Settings option.
11. Select the new SCSI controller that was added as a result of adding the RDMs on a separate SCSI controller.
12. Select the Virtual radio button under the SCSI Bus Sharing options, as shown in Figure 11.6.
13. Repeat steps 2 through 9 to configure additional RDMs for shared storage locations needed by nodes of a Microsoft server cluster.



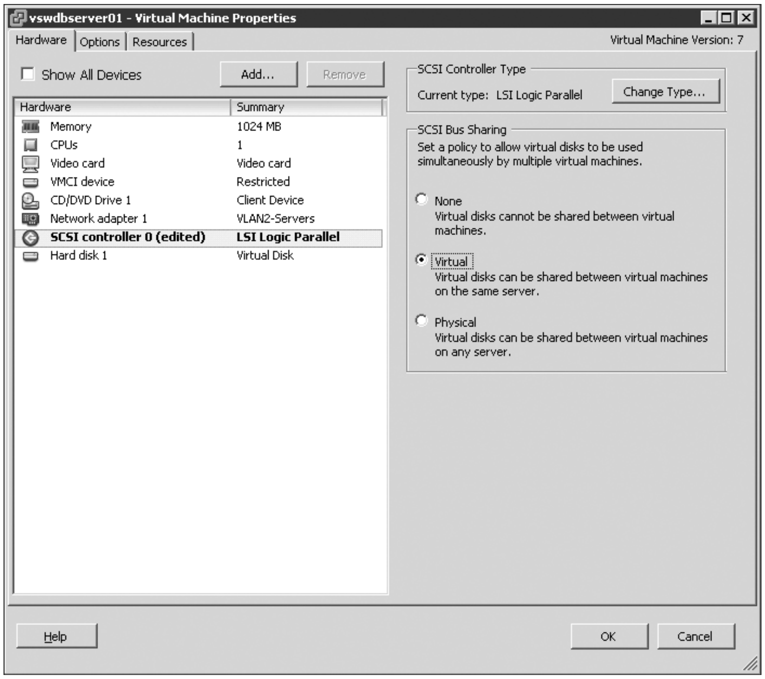
Advanced Options

These advanced options do not usually need to be changed.

Specify the advanced options for this virtual disk. These options do not normally need to be changed.

Virtual Device Node: SCSI (1:0)

FIGURE 11.5
The virtual device node for the additional RDMs in a cluster node must be on a different SCSI node.



Virtual Machine Version: 7

□ Show All Devices

Hardware	Summary
Memory	1024 MB
CPUs	1
Video card	Video card
VMCI device	Restricted
CD/DVD Drive 1	Client Device
Network adapter 1	VLAN2-Servers
SCSI controller 0 (edited)	**LSI Logic Parallel**
Hard disk 1	Virtual Disk

SCSI Controller Type

Current type: LSI Logic Parallel

SCSI Bus Sharing

Set a policy to allow virtual disks to be used simultaneously by multiple virtual machines.

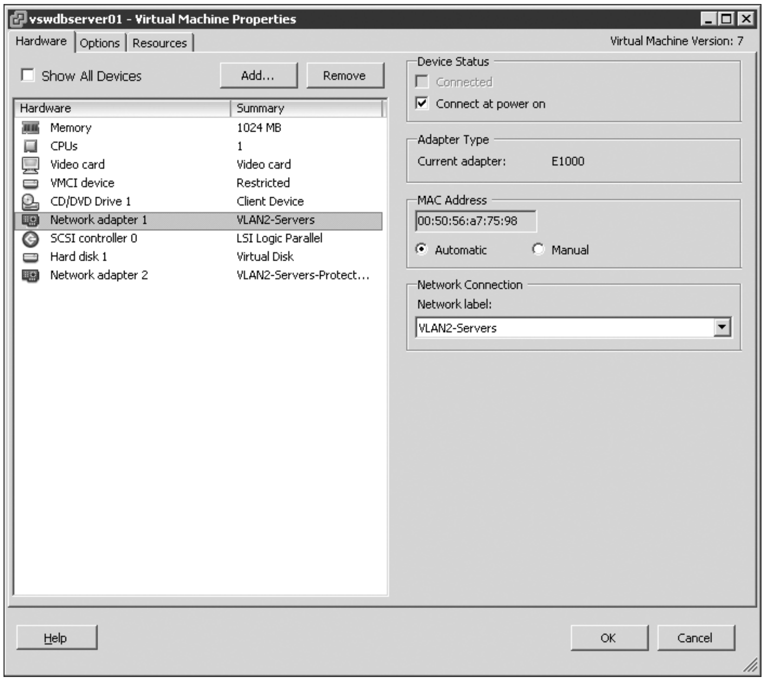
○ None
Virtual disks cannot be shared between virtual machines.

● Virtual
Virtual disks can be shared between virtual machines on the same server.

○ Physical
Virtual disks can be shared between virtual machines on any server.

FIGURE 11.6
The SCSI bus sharing for the new SCSI adapter must be set to Virtual to support running a virtual machine as a node in a Microsoft server cluster.

14. Configure the virtual machine with two network adapters. Connect one network adapter to the production network, and connect the other network adapter to the network used for heartbeat communications between nodes. Figure 11.7 shows a cluster node with two network adapters configured.



vswdbserver01 - Virtual Machine Properties

Virtual Machine Version: 7

☐ Show All Devices

Hardware	Summary
Memory	1024 MB
CPUs	1
Video card	Video card
VMCI device	Restricted
CD/DVD Drive 1	Client Device
Network adapter 1	VLAN2-Servers
SCSI controller 0	LSI Logic Parallel
Hard disk 1	Virtual Disk
Network adapter 2	VLAN2-Servers-Protect…

Device Status
☐ Connected
☑ Connect at power on

Adapter Type
Current adapter: E1000

MAC Address
00:50:56:a7:75:98
● Automatic ○ Manual

Network Connection
Network label:
VLAN2-Servers

FIGURE 11.7
A node in a Microsoft server cluster requires at least two network adapters. One adapter must be able to communicate on the production network, and the second adapter is configured for internal cluster heartbeat communication.

15. Power on the first node of the cluster, and assign valid IP addresses to the network adapters configured for the production and heartbeat networks. Then format the additional drives, and assign drive letters, as shown in Figure 11.8.

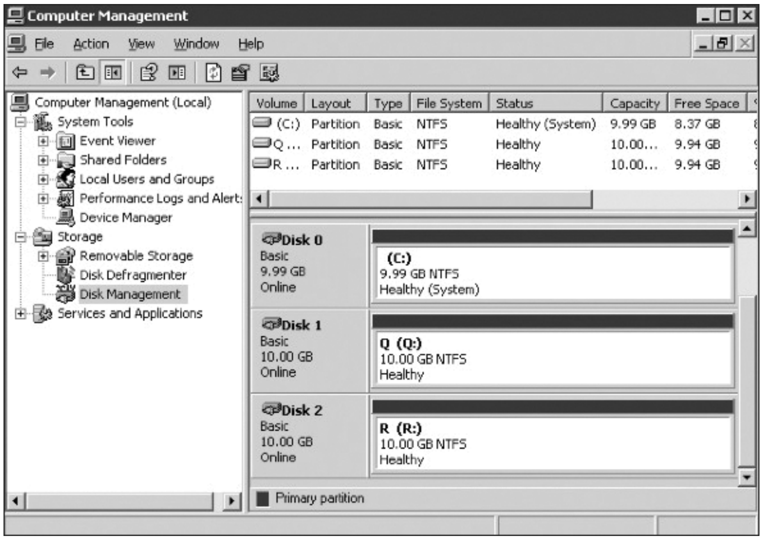
16. Shut down the first cluster node.

17. In the VCenter Server inventory, select the ESX/ESXi host where the first cluster node is configured, and then select the Configuration tab.

18. Select Advanced Settings from the Software menu.

19. In the Advanced Settings dialog box, configure the following options:

    - Set the Disk.ResetOnFailure option to 1.
    - Set the Disk.UseLunReset option to 1.
    - Set the Disk.UseDeviceReset option to 0.



Volume	Layout	Type	File System	Status	Capacity	Free Space
(C:)	Partition	Basic	NTFS	Healthy (System)	9.99 GB	8.37 GB
Q ...	Partition	Basic	NTFS	Healthy	10.00...	9.94 GB
R ...	Partition	Basic	NTFS	Healthy	10.00...	9.94 GB

Figure 11.8
The RDMs presented to the first cluster node are formatted and assigned drive letters.

20. Proceed to the next section to configure the second cluster node and the respective ESX/ESXi host.

SCSI Nodes for RDMs

RDMs used for shared storage in a Microsoft server cluster must be configured on a SCSI node that is different from the SCSI to which the hard disk is connected that holds the operating system. For example, if the operating system's virtual hard drive is configured to use the SCSI0 node, then the RDM should use the SCSI1 node. This rule applies to both virtual and physical clustering.

Because of PCI addressing issues, all RDMs should be added prior to configuring the additional network adapters. If the NICs are configured first, you may be required to revisit the network adapter configuration after the RDMs are added to the cluster node.

Creating the Second Cluster Node in Windows 2003

Perform the following steps to create the second cluster node:

1. Starting from inside the vSphere client, create a second virtual machine that is a member of the same Active Directory domain as the first cluster node.
2. Add the same RDMs to the second cluster node using the same SCSI node values.

   For example, if the first node used SCSI 1:0 for the first RDM and SCSI 1:1 for the second RDM, then configure the second node to use the same configuration. As in the first cluster node configuration, add all RDMs to the virtual machine before moving on to step 3 to configure the network adapters. Don't forget to edit the SCSI bus sharing configuration for the new SCSI adapter.

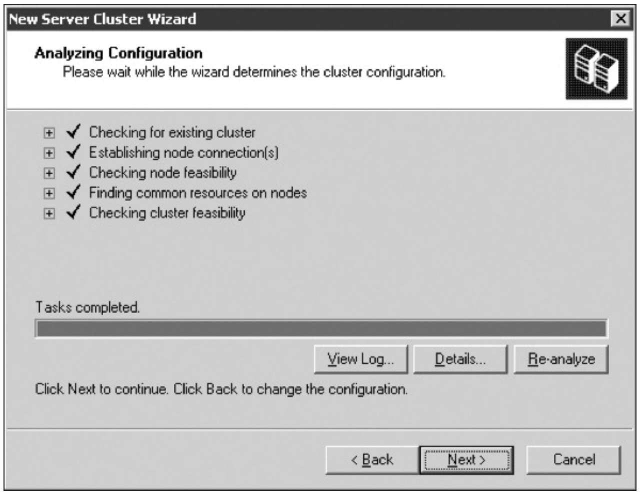
3. Configure the second node with an identical network adapter configuration.
4. Verify that the hard drives corresponding to the RDMs can be seen in Disk Manager. At this point, the drives will show as a status of ‘‘Healthy,’’ but drive letters will not be assigned.
5. Power off the second node.
6. Edit the advanced disk settings for the ESX/ESXi host with the second cluster node.

Creating the Management Cluster in Windows 2003

Perform the following steps to create the management cluster:

1. Starting from Active Directory Users and Computers, if you have the authority, create a new user account that belongs to the same Windows Active Directory domain as the two cluster nodes. The account does not need to be granted any special group memberships at this time.
2. Power on the first node of the cluster, and log in as a user with administrative credentials.
3. Click Start ➢ Programs ➢ Administrative Tools, and select the Cluster Administrator console.
4. Select the Create New Cluster option from the Open Connection To Cluster dialog box. Then click OK.
5. Provide a unique name for the name of the cluster. Ensure that it does not match the name of any existing computers on the network.
6. The next step is to execute the cluster feasibility analysis to check for all cluster-capable resources, as shown in Figure 11.9. Then click Next.

Figure 11.9
The cluster analysis portion of the cluster configuration wizard identifies that all cluster-capable resources are available.



New Server Cluster Wizard

Analyzing Configuration
Please wait while the wizard determines the cluster configuration.

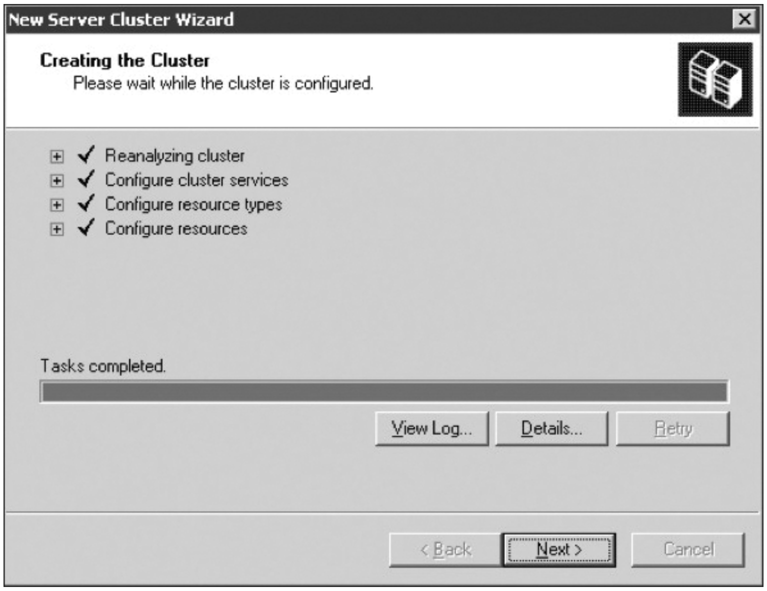
⊞ ✓ Checking for existing cluster
⊞ ✓ Establishing node connection(s)
⊞ ✓ Checking node feasibility
⊞ ✓ Finding common resources on nodes
⊞ ✓ Checking cluster feasibility

Tasks completed.

Click Next to continue. Click Back to change the configuration.

7. Provide an IP address for cluster management. The IP address configured for cluster management should be an IP address that is accessible from the network adapters configured on the production network. Click Next.
8. Provide the account information for the cluster service user account created in step 1. The Cluster Service Account page of the New Server Cluster Wizard acknowledges that the account specified will be granted membership in the local administrators group on each cluster node. Therefore, do not share the cluster service password with users who should not have administrative capabilities. Click Next.
9. At the completion of creating the cluster timeline, shown in Figure 11.10, click Next.

FIGURE 11.10
The cluster installation timeline provides a running report of the items configured as part of the installation process.



New Server Cluster Wizard

Creating the Cluster
Please wait while the cluster is configured.

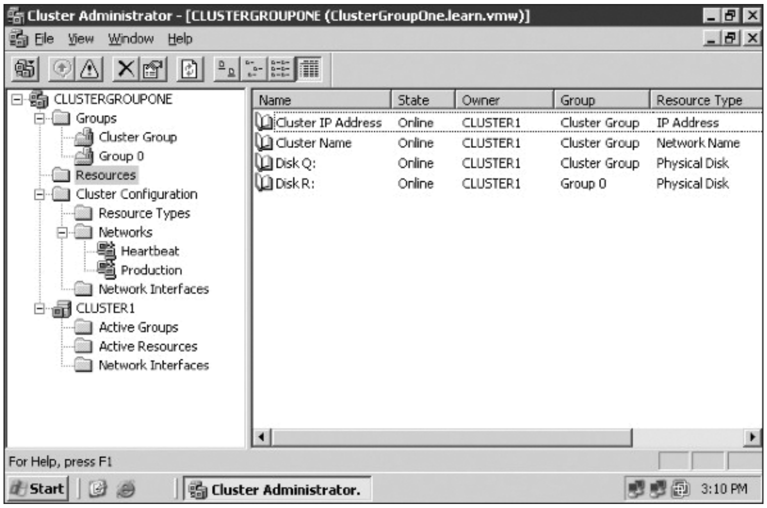
- ✓ Reanalyzing cluster
- ✓ Configure cluster services
- ✓ Configure resource types
- ✓ Configure resources

Tasks completed.

10. Continue to review the Cluster Administrator snap-in, and review the new management cluster that was created, shown in Figure 11.11.

CLUSTER MANAGEMENT

To access and manage a Microsoft cluster, create a Host (A) record in the zone that corresponds to the domain to which the cluster nodes belong.



Name	State	Owner	Group	Resource Type
Cluster IP Address	Online	CLUSTER1	Cluster Group	IP Address
Cluster Name	Online	CLUSTER1	Cluster Group	Network Name
Disk Q:	Online	CLUSTER1	Cluster Group	Physical Disk
Disk R:	Online	CLUSTER1	Group 0	Physical Disk

FIGURE 11.11
The completion of the initial cluster management creation wizard results in a cluster group and all associated cluster resources.

ADDING THE SECOND NODE TO THE MANAGEMENT CLUSTER IN WINDOWS 2003

Perform the following steps to add the second node to the management cluster:

1. Leave the first node powered on, and power on the second node.
2. Starting from the Cluster Administrator, right-click the name of the cluster, select the New option, and then click the Node option, as shown in Figure 11.12.

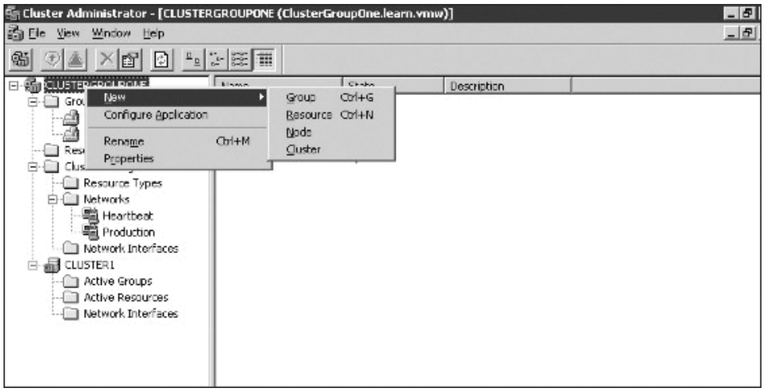
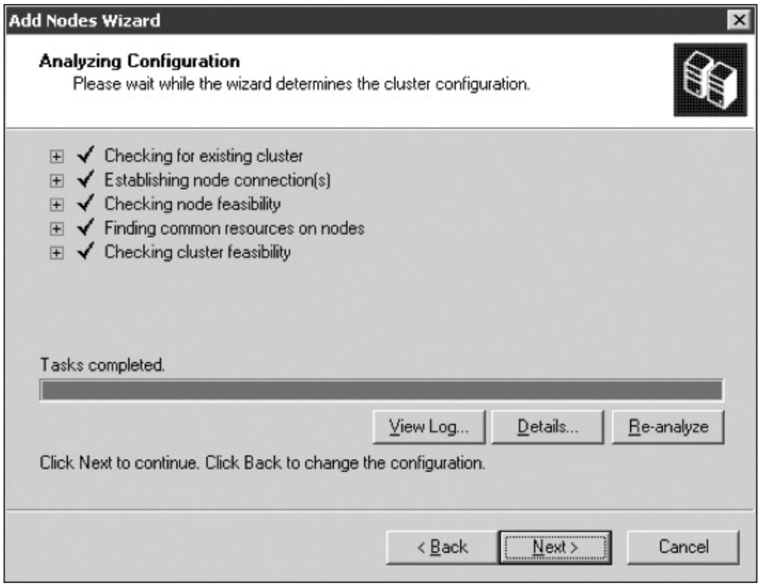


FIGURE 11.12
After the management cluster is complete, you can add a node.

3. Specify the name of the node to be added to the cluster, and then click Next.
4. After the cluster feasibility check has completed (see Figure 11.13), click the Next button.



Add Nodes Wizard

Analyzing Configuration

Please wait while the wizard determines the cluster configuration.

- ✓ Checking for existing cluster
- ✓ Establishing node connection(s)
- ✓ Checking node feasibility
- ✓ Finding common resources on nodes
- ✓ Checking cluster feasibility

Tasks completed.

Click Next to continue. Click Back to change the configuration.

FIGURE 11.13
A feasibility check is executed against each potential node to validate the hardware configuration that supports the appropriate shared resources and network configuration parameters.

5. Proceed to review the Cluster Administrator, identifying that two nodes now exist within the new cluster.

Feasibility Stall

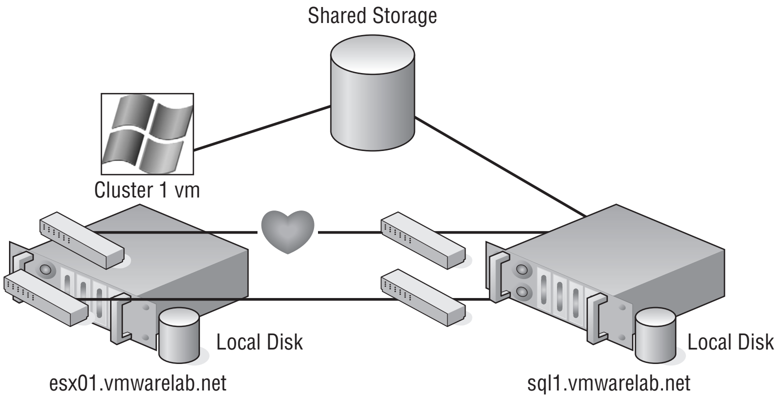
If the feasibility check stalls and reports a 0x00138f error stating that a cluster resource cannot be found, the installation will continue to run. This is a known issue with the Windows Server 2003 cluster configuration. If you allow the installation to continue, it will eventually complete and function as expected. For more information, visit `http://support.microsoft.com/kb/909968`.

At this point, the management cluster is complete; from here, application and service clusters can be configured. Some applications, such as Microsoft SQL Server 2005 and Microsoft Exchange Server 2007, are not only cluster-aware applications but also allow for the creation of a server cluster as part of the standard installation wizard. Other cluster-aware applications and services can be configured into a cluster using the cluster administrator.

Examining Physical to Virtual Clustering

The last type of clustering scenario to discuss is physical to virtual clustering. As you might have guessed, this involves building a cluster with two nodes where one node is a physical machine and the other node is a virtual machine. Figure 11.14 details the setup of a two-node physical to virtual cluster.

FIGURE 11.14
Clustering physical machines with virtual machine counterparts can be a cost-effective way of providing high availability.



The constraints surrounding the construction of a physical to virtual cluster are identical to those noted in the previous configuration. Likewise, the steps to configure the virtual machine acting as a node in the physical to virtual cluster are identical to the steps outlined in the previous section, with one addition: you must set the RDMs up in Physical Compatibility mode. The virtual machine must have access to all the same storage locations as the physical machine. The virtual machine must also have access to the same pair of networks used by the physical machine for production and heartbeat communication, respectively.

The advantage to implementing a physical to virtual cluster is the resulting high availability with reduced financial outlay. Physical to virtual clustering, because of the two-node limitation of virtual machine clustering, ends up as an N+1 clustered solution, where N is the number of physical servers in the environment plus one additional physical server to host the virtual machines. In each case, each physical virtual machine cluster creates a failover pair. With the scope of the cluster design limited to a failover pair, the most important design aspect in a physical to virtual cluster is the scale of the host running ESX/ESXi host. As you may have figured, the more powerful the ESX/ESXi host, the more failover incidents it can handle. A more powerful ESX/ESXi host will scale better to handle multiple physical host failures, whereas a less powerful ESX/ESXi host might handle only a single physical host failure before performance levels experience a noticeable decline.

Now that I've covered clustering, let's take a look at VMware's version of high availability. VMware has a built-in option called VMware High Availability that is just what the name implies.

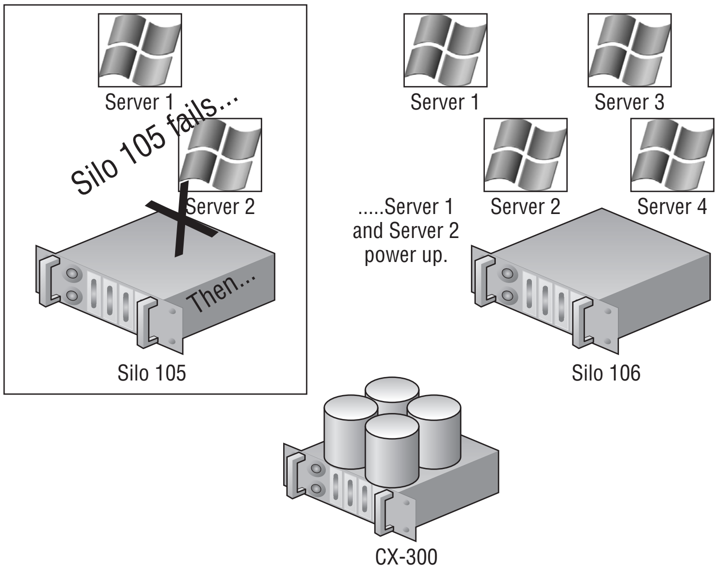
Implementing VMware High Availability

High availability has been an industry buzzword that has stood the test of time. The need and/or desire for high availability is often a significant component to any infrastructure design. Within the scope of an ESX/ESXi host, VMware High Availability (HA) is a component of the vSphere 4 product that provides for the automatic failover of virtual machines. But—and it's a big *but* at this point in time—HA does not provide high availability in the traditional sense of the term. Commonly, HA means the automatic failover of a service or application to another server.

Understanding HA

The VMware HA feature provides an automatic restart of the virtual machines that were running on an ESX/ESXi host at the time it became unavailable, shown in Figure 11.15.

Figure 11.15
VMware HA provides an automatic restart of virtual machines that were running on an ESX/ESXi host when it failed.



In the case of VMware HA, there is still a period of downtime when a server fails. Unfortunately, the duration of the downtime is not a value that can be calculated because it is unknown ahead of time how long it will take to boot a series of virtual machines. From this you can gather that, at this point in time, VMware HA does not provide the same level of high availability as found in a Microsoft server cluster solution. When a failover occurs between ESX/ESXi hosts as a result of the HA feature, there is potential for data loss as a result of the virtual machine that was immediately powered off when the server failed and then brought back up minutes later on another server.

Real World Scenario

HA Experience in the Field

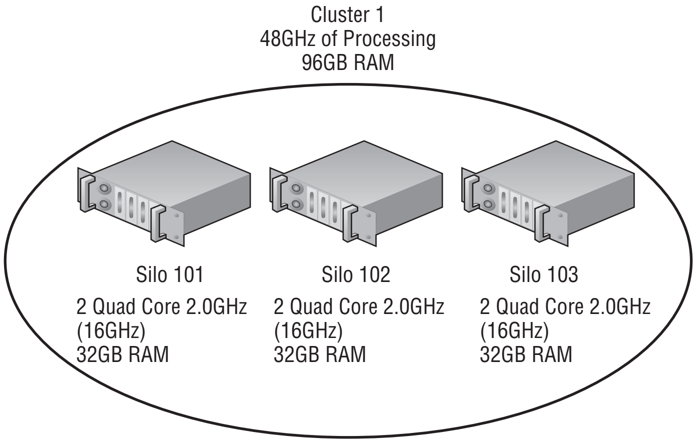
With that said, I want to mention my own personal experience with HA and the results I encountered. Your mileage might vary but should give you a reasonable expectation of what to expect. I had a VMware ESX/ESXi host that was a member of a five-node cluster. This node crashed sometime during the night, and when the host went down, it took anywhere from 15 to 20 virtual machines with it. HA kicked in and restarted all the virtual machines as expected.

What made this an interesting experience is that the crash must have happened right after the polling of the monitoring and alerting server. All the virtual machines that were on the general alerting schedule were restarted without triggering any alerts. We did have some of those virtual machines with a more aggressive monitoring that did trip off alerts that were recovered before anyone was able to log on to the system and investigate. I tried to argue the point that if an alert never fired, did the downtime really happen? I did not get too far with that argument but was pleased with the results.

In another case, during testing I had a virtual machine running on a two-node cluster. I pulled the power cords on the host that the virtual machine was running to create the failure. My time to recovery from pull to ping was between five and six minutes. That's not too bad for general use but not good enough for everything. VMware Fault Tolerance can now fill that gap for even the most important and critical servers in your environment. I'll talk more about FT in a bit.

In the VMware HA scenario, two or more ESX/ESXi hosts are configured in a cluster. Remember, a VMware cluster represents a logical aggregation of CPU and memory resources, as shown in Figure 11.16. By editing the cluster settings, you can enable the VMware HA feature for a cluster. The HA cluster then determines the number of hosts failures it must support.

Figure 11.16
A VMware ESX/ESXi host cluster logically aggregates the CPU and memory resources from all nodes in the cluster.



HA: Within, but Not Between, Sites

A requisite of HA is that each node in the HA cluster must have access to the same SAN LUNs. This requirement prevents HA from being able to failover between ESX/ESXi hosts in different locations unless both locations have been configured to have access to the *same* storage devices. It is not acceptable just to have the data in LUNs the same because of SAN replication software. Mirroring data from a LUN on a SAN in one location to a LUN on a SAN in a hot site is not conducive to allowing HA (VMotion or DRS).

When ESX/ESXi hosts are configured into a VMware HA cluster, they receive all the cluster information. vCenter Server informs each node in the HA cluster about the cluster configuration.

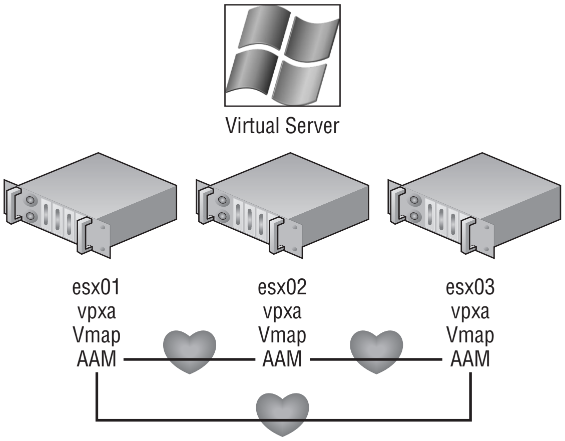
HA and vCenter Server

Although vCenter Server is most certainly required to enable and manage VMware HA, it is not required to execute HA. vCenter Server is a tool that notifies each VMware HA cluster node about the HA configuration. After the nodes have been updated with the information about the cluster, vCenter Server no longer maintains a persistent connection with each node. Each node continues to function as a member of the HA cluster independent of its communication status with vCenter Server.

When an ESX/ESXi host is added to a VMware HA cluster, a set of HA-specific components are installed on the ESX/ESXi host. These components, shown in Figure 11.17, include the following:

- Automatic Availability Manager (AAM)
- Vmap
- vpxa

Figure 11.17 Adding an ESX/ESXi host to an HA cluster automatically installs the AAM, Vmap, and possibly the vpxa components on the host.



The AAM, effectively the engine or service for HA, is a Legato-based component that keeps an internal database of the other nodes in the cluster. The AAM is responsible for the intracluster heartbeat used to identify available and unavailable nodes. Each node in the cluster establishes a heartbeat with each of the other nodes over the Service Console network, or you can use or define another VMkernel port group for the HA heartbeat. As a best practice, you should provide redundancy to the AAM heartbeat by establishing the Service Console port group on a virtual switch with an underlying NIC team. Though the Service Console could be multihomed and have an AAM heartbeat over two different networks, this configuration is not as reliable as the NIC team. The AAM is extremely sensitive to hostname resolution; the inability to resolve names will most certainly result in an inability to execute HA. When problems arise with HA functionality, look first at hostname resolution. Having said that, during HA troubleshooting, you should identify the answers to questions such as these:

- Is the DNS server configuration correct?
- Is the DNS server available?
- If DNS is on a remote subnet, is the default gateway correct and functional?
- Does the `/etc/hosts` file have bad entries in it?
- Does the `/etc/resolv.conf` have the right search suffix?
- Does the `/etc/resolv.conf` have the right DNS server?

Adding a Host to vCenter Server

When a new host is added into the vCenter Server inventory, the host must be added by its hostname, or the HA will not function properly. As just noted, HA is heavily reliant on successful name resolution. ESX/ESXi hosts should not be added to the vCenter Server inventory using IP addresses.

The AAM on each ESX/ESXi host keeps an internal database of the other hosts belonging to the cluster. All hosts in a cluster are considered either a primary host or a secondary host. However, only one ESX/ESXi host in the cluster is considered the primary host at a given time, with all others considered secondary hosts. The primary host functions as the source of information for all new hosts and defaults to the first host added to the cluster. If the primary host experiences failure, the HA cluster will continue to function. In fact, in the event of primary host failure, one of the secondary hosts will move up to the status of primary host. The process of promoting secondary hosts to primary host is limited to four other hosts. Only five hosts could assume the role of primary host in an HA cluster.

While AAM is busy managing the intranode communications, the vpxa service (or vCenter Server agent) manages the HA components. The vpxa service communicates to the AAM through a third component called the Vmap.

Name Resolution Tip

If DNS is set up and configured correctly, then you should not need anything else for name resolution. However, as a method of redundancy, consider adding the other VMware ESX and vCenter Server information to the local host file (`/etc/hosts`). If there is a failure and the ESX/ESXi host is unable to talk to DNS, this setup will ensure that HA would still work as designed.

Configuring HA

Before I detail how to set up and configure the HA feature, let's review the requirements of HA. To implement HA, all of the following requirements should be met:

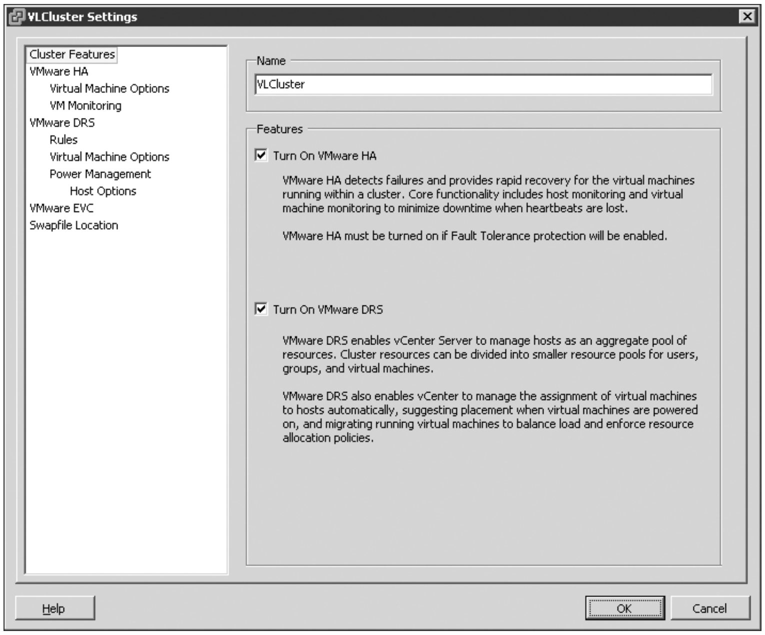
- All hosts in an HA cluster must have access to the same shared storage locations used by all virtual machines on the cluster. This includes any Fibre Channel, iSCSI, and NFS datastores used by virtual machines.
- All hosts in an HA cluster should have an identical virtual networking configuration. If a new switch is added to one host, the same new switch should be added to all hosts in the cluster.
- All hosts in an HA cluster must resolve the other hosts using DNS names.

A Test for HA

An easy and simple test for identifying HA capability for a virtual machine is to perform a VMotion. The requirements of VMotion are actually more stringent than those for performing an HA failover, though some of the requirements are identical. In short, if a virtual machine can successfully perform a VMotion across the hosts in a cluster, then it is safe to assume that HA will be able to power on that virtual machine from any of the hosts. To perform a full test of a virtual machine on a cluster with four nodes, perform a VMotion from node 1 to node 2, node 2 to node 3, node 3 to node 4, and finally, node 4 back to node 1. If it works, then you have passed the test!

First and foremost, to configure HA, you must create a cluster. After you create the cluster, you can enable and configure HA. Figure 11.18 shows a cluster enabled for HA.

Figure 11.18
A cluster of ESX/ESXi hosts can be configured with HA and DRS. The features are not mutually exclusive and can work together to provide availability and performance optimization.



Name: VLCluster

☑ Turn On VMware HA

VMware HA detects failures and provides rapid recovery for the virtual machines running within a cluster. Core functionality includes host monitoring and virtual machine monitoring to minimize downtime when heartbeats are lost.

VMware HA must be turned on if Fault Tolerance protection will be enabled.

☑ Turn On VMware DRS

VMware DRS enables vCenter Server to manage hosts as an aggregate pool of resources. Cluster resources can be divided into smaller resource pools for users, groups, and virtual machines.

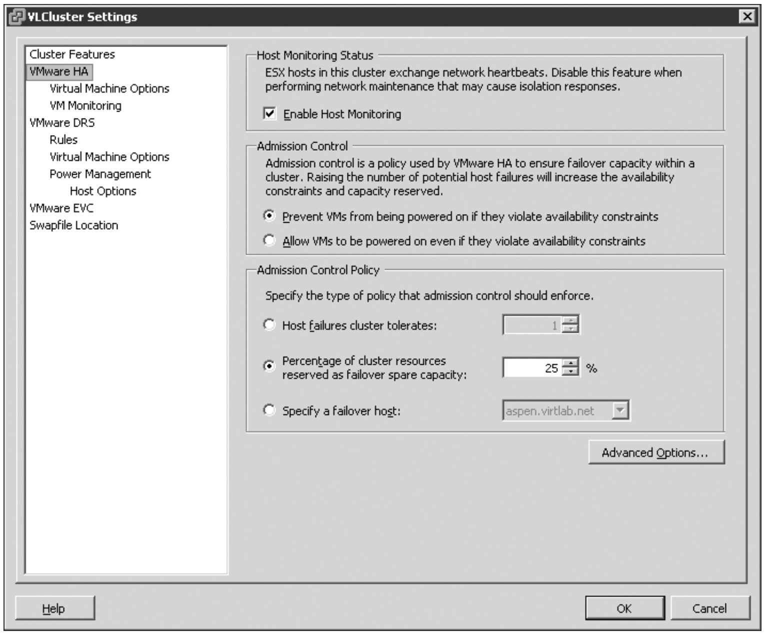
VMware DRS also enables vCenter to manage the assignment of virtual machines to hosts automatically, suggesting placement when virtual machines are powered on, and migrating running virtual machines to balance load and enforce resource allocation policies.

Configuring an HA cluster revolves around three different settings:

- Host failures allowed
- Admission control
- Virtual machine options

The configuration option for the number of host failures to allow, shown in Figure 11.19, is a critical setting. It directly influences the number of virtual machines that can run in the cluster before the cluster is in jeopardy of being unable to support an unexpected host failure. vSphere now gives up the capability to set a percentage for the failover spare capacity or specify a specific node for failover.

FIGURE 11.19
The number of host failures allowed dictates the amount of spare capacity that must be retained for use in recovering virtual machines after failure.



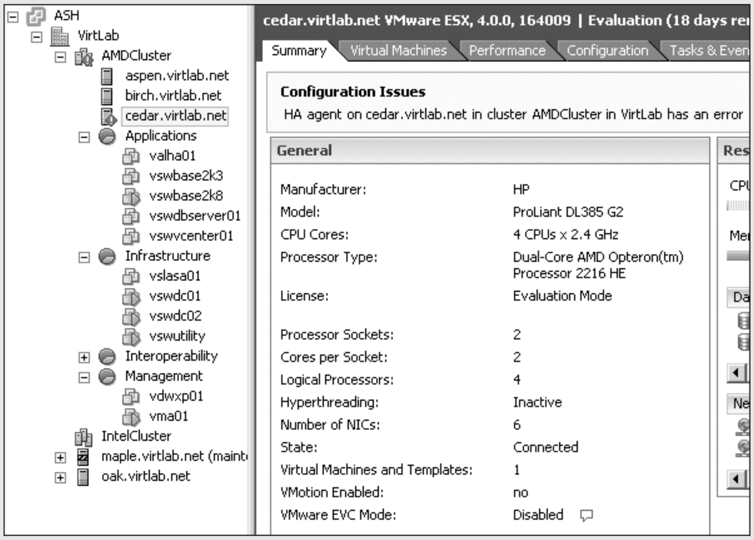
Host Monitoring Status
ESX hosts in this cluster exchange network heartbeats. Disable this feature when performing network maintenance that may cause isolation responses.
☑ Enable Host Monitoring

Admission Control
Admission control is a policy used by VMware HA to ensure failover capacity within a cluster. Raising the number of potential host failures will increase the availability constraints and capacity reserved.
● Prevent VMs from being powered on if they violate availability constraints
○ Allow VMs to be powered on even if they violate availability constraints

Admission Control Policy
Specify the type of policy that admission control should enforce.
○ Host failures cluster tolerates: 1
● Percentage of cluster resources reserved as failover spare capacity: 25 %
○ Specify a failover host: aspen.virtlab.net

HA Configuration Failure

It is not uncommon for a host in a cluster to fail during the configuration of HA. Remember the stress we put on DNS, or name resolution in general, earlier in this chapter? Well, if DNS is not set correctly, you will find that the host cannot be configured for HA. Take, for example, a cluster with three nodes being configured as an HA cluster to support two-node failure. Enabling HA forces a configuration of each node in the cluster. The image here shows an HA cluster where one of the nodes, Silo 104, has thrown an error related to the HA agent and is unable to complete the HA configuration.



```
ASH
  VirtLab
    AMDCluster
      aspen.virtlab.net
      birch.virtlab.net
      cedar.virtlab.net
      Applications
        valha01
        vswbase2k3
        vswbase2k8
        vswdbserver01
        vswvcenter01
      Infrastructure
        vslasa01
        vswdc01
        vswdc02
        vswutility
      Interoperability
      Management
        vdwxp01
        vma01
    IntelCluster
    maple.virtlab.net (maint
    oak.virtlab.net
```

cedar.virtlab.net VMware ESX, 4.0.0, 164009 | Evaluation (18 days re

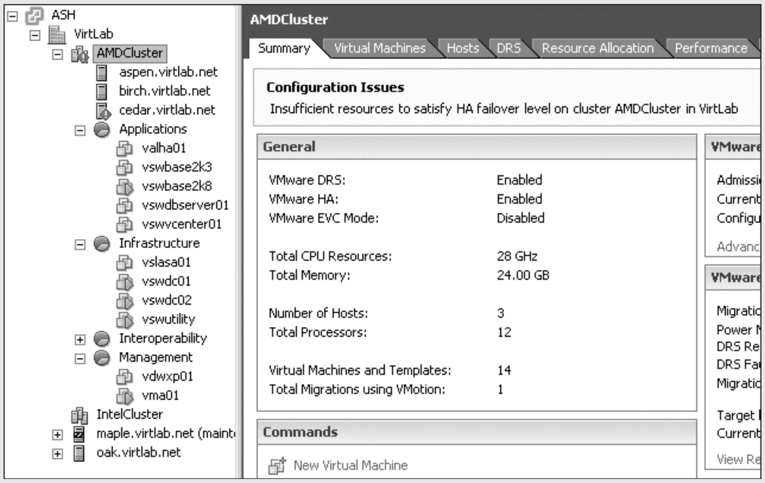
Configuration Issues

HA agent on cedar.virtlab.net in cluster AMDCluster in VirtLab has an error

General

Manufacturer:	HP
Model:	ProLiant DL385 G2
CPU Cores:	4 CPUs x 2.4 GHz
Processor Type:	Dual-Core AMD Opteron(tm) Processor 2216 HE
License:	Evaluation Mode
Processor Sockets:	2
Cores per Socket:	2
Logical Processors:	4
Hyperthreading:	Inactive
Number of NICs:	6
State:	Connected
Virtual Machines and Templates:	1
VMotion Enabled:	no
VMware EVC Mode:	Disabled

In this example, because the cluster was attempting to allow for two-node failure and there are only two nodes successfully configured, this would be impossible. The cluster in this case is now warning that there are insufficient resources to satisfy the HA failover level. Naturally, with only two nodes, we cannot cover a two-node failure. The following image shows an error on the cluster because of the failure in Silo 104.



```
ASH
  VirtLab
    AMDCluster
      aspen.virtlab.net
      birch.virtlab.net
      cedar.virtlab.net
      Applications
        valha01
        vswbase2k3
        vswbase2k8
        vswdbserver01
        vswvcenter01
      Infrastructure
        vslasa01
        vswdc01
        vswdc02
        vswutility
      Interoperability
      Management
        vdwxp01
        vma01
    IntelCluster
    maple.virtlab.net (maint
    oak.virtlab.net
```

AMDCluster

Configuration Issues

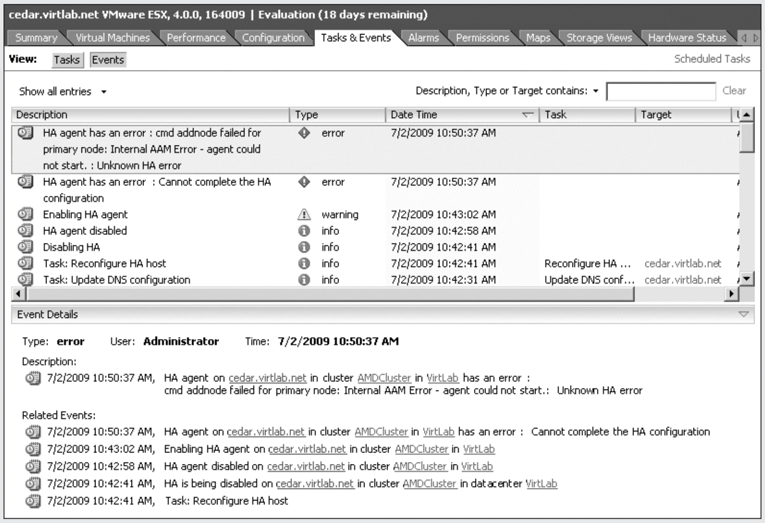
Insufficient resources to satisfy HA failover level on cluster AMDCluster in VirtLab

General

VMware DRS:	Enabled
VMware HA:	Enabled
VMware EVC Mode:	Disabled
Total CPU Resources:	28 GHz
Total Memory:	24.00 GB
Number of Hosts:	3
Total Processors:	12
Virtual Machines and Templates:	14
Total Migrations using VMotion:	1

Commands

In the Tasks pane of the graphic, you might have noticed that Silo 105 and Silo 106 both completed the HA configuration successfully. This provides evidence that the problem is probably isolated to Silo 104. Reviewing the Tasks & Events tab to get more detail on the error reveals exactly that. The following image shows that the error was caused by an inability to resolve a name. This confirms the suspicion that the error is with DNS.



Description	Type	Date Time	Task	Target
HA agent has an error : cmd addnode failed for primary node: Internal AAM Error - agent could not start. : Unknown HA error	error	7/2/2009 10:50:37 AM		
HA agent has an error : Cannot complete the HA configuration	error	7/2/2009 10:50:37 AM		
Enabling HA agent	warning	7/2/2009 10:43:02 AM		
HA agent disabled	info	7/2/2009 10:42:58 AM		
Disabling HA	info	7/2/2009 10:42:41 AM		
Task: Reconfigure HA host	info	7/2/2009 10:42:41 AM	Reconfigure HA ...	cedar.virtlab.net
Task: Update DNS configuration	info	7/2/2009 10:42:31 AM	Update DNS conf...	cedar.virtlab.net

Type: **error** User: **Administrator** Time: **7/2/2009 10:50:37 AM**

Description:

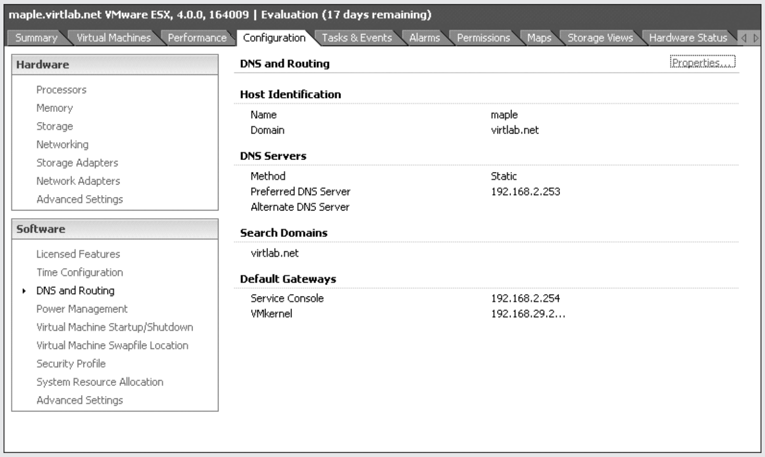
7/2/2009 10:50:37 AM, HA agent on cedar.virtlab.net in cluster AMDCluster in VirtLab has an error :
cmd addnode failed for primary node: Internal AAM Error - agent could not start.: Unknown HA error

Related Events:

7/2/2009 10:50:37 AM, HA agent on cedar.virtlab.net in cluster AMDCluster in VirtLab has an error : Cannot complete the HA configuration
7/2/2009 10:43:02 AM, Enabling HA agent on cedar.virtlab.net in cluster AMDCluster in VirtLab
7/2/2009 10:42:58 AM, HA agent disabled on cedar.virtlab.net in cluster AMDCluster in VirtLab
7/2/2009 10:42:41 AM, HA is being disabled on cedar.virtlab.net in cluster AMDCluster in datacenter VirtLab
7/2/2009 10:42:41 AM, Task: Reconfigure HA host

Perform the following steps to review or edit the DNS server for an ESX/ESXi host:

1. Use the vSphere Client to connect to a vCenter Server.
2. Click the Hosts And Clusters button on the Home page.
3. Click the name of the host in the inventory tree on the left.
4. Click the Configuration tab in the details pane on the right.
5. Select DNS And Routing in the Advanced menu.
6. If needed, edit the DNS server, as shown in the following image, to a server that can resolve the other nodes in the HA cluster.



maple.virtlab.net VMware ESX, 4.0.0, 164009 | Evaluation (17 days remaining)

DNS and Routing

Host Identification

Name	maple
Domain	virtlab.net

DNS Servers

Method	Static
Preferred DNS Server	192.168.2.253
Alternate DNS Server	

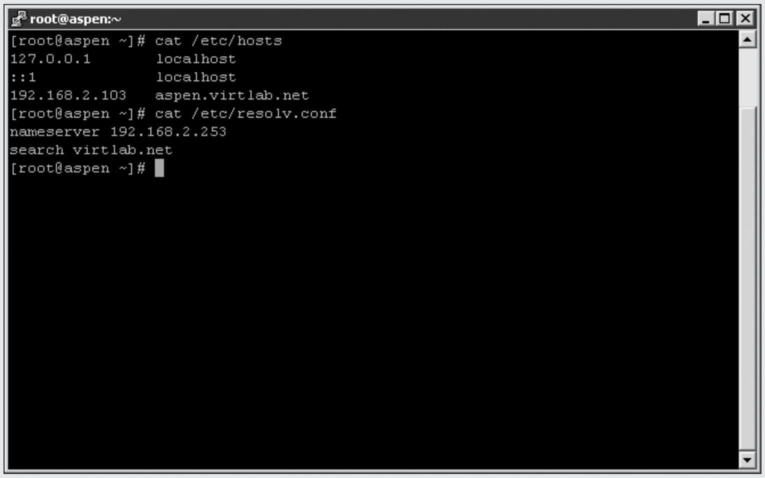
Search Domains

virtlab.net

Default Gateways

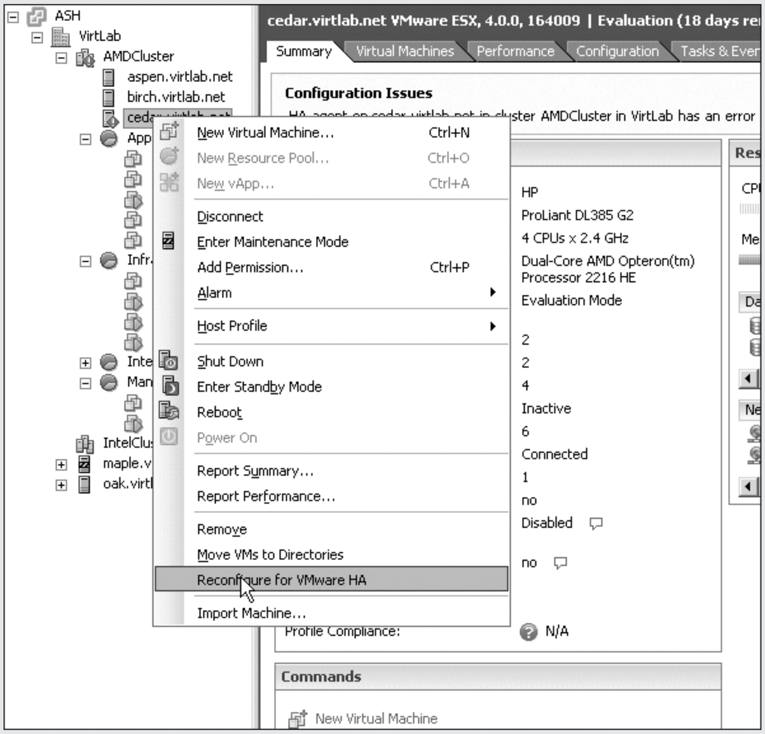
Service Console	192.168.2.254
VMkernel	192.168.29.2...

Although they should not be edited on a regular basis, you can also check the /etc/hosts and /etc/resolv.conf files, which should contain static lists of hostnames to IP addresses or the DNS search domain and name servers, respectively. The following image offers a quick look at the information inside the /etc/hosts and /etc/resolv.conf files. These tools can be valuable for troubleshooting name resolution.


root@aspen:~

```
[root@aspen ~]# cat /etc/hosts
127.0.0.1       localhost
::1             localhost
192.168.2.103   aspen.virtlab.net
[root@aspen ~]# cat /etc/resolv.conf
nameserver 192.168.2.253
search virtlab.net
[root@aspen ~]#
```

After the DNS server, `/etc/hosts`, or `/etc/resolv.conf` has been corrected, the host with the failure can be reconfigured for HA. It's not necessary to remove the HA configuration from the cluster and then reenable it. The following image shows the right-click context menu of Silo 104, where it can be reconfigured for HA now that the name resolution problem has been fixed.



cedar.virtlab.net VMware ESX, 4.0.0, 164009 | Evaluation (18 days re

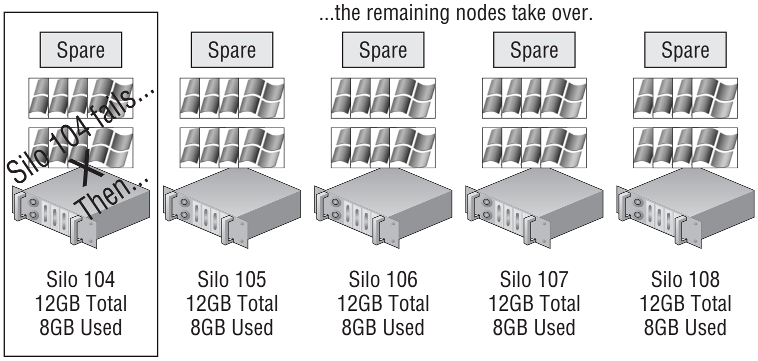
Configuration Issues

HA agent on cedar.virtlab.net in cluster AMDCluster in VirtLab has an error

Upon completion of the configuration of the final node, the errors at the host and cluster levels will be removed, the cluster will be configured as desired, and the error regarding the inability to satisfy the failover level will disappear.

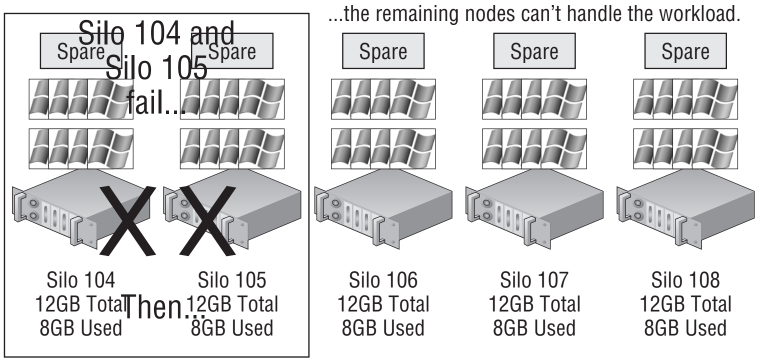
To explain the workings of HA and the differences in the configuration settings, let's look at some implementation scenarios. For example, consider five ESX/ESXi hosts named Silo 101 through Silo 105. All five hosts belong to an HA cluster configured to support single-host failure. Each node in the cluster is equally configured with 12GB of RAM. If each node runs eight virtual machines with 1GB of memory allocated to each virtual machine, then 8GB of unused memory across four hosts is needed to support a single-host failure. The 12GB of memory on each host minus 8GB for virtual machines leaves 4GB of memory per host. Figure 11.20 shows our five-node cluster in normal operating mode.

Figure 11.20
A five-node cluster configured to allow single-host failure



Let’s assume that Service Console and virtual machine overhead consume 1GB of memory, leaving 3GB of memory per host. If Silo 101 fails, the remaining four hosts will each have 3GB of memory to contribute to running the virtual machines orphaned by the failure. The 8GB of virtual machines will then be powered on across the remaining four hosts that collectively have 12GB of memory to spare. In this case, the configuration supported the failover. Figure 11.21 shows our five-node cluster down to four after the failure of Silo 101. Assume in this same scenario that Silo 101 and Silo 102 both experience failure. That leaves 16GB of virtual machines to cover across only three hosts with 3GB of memory to spare. In this case, the cluster is deficient, and not all of the orphaned virtual machines will be restarted.

Figure 11.21
A five-node cluster configured to allow single-host failure is deficient in resources to support a second failed node.



Primary Host Limit

In the previous section introducing the HA feature, I mentioned that the AAM caps the number of primary hosts at five. This limitation translates into a maximum of four host failures allowed in a cluster.

The admission control setting goes hand in hand with the Number Of Host Failures Allowed setting. There are two possible settings for admission control:

- Do not power on virtual machines if they violate availability constraints (known as *strict admission control*).
- Allow virtual machines to be powered on even if they violate availability constraints (known as *guaranteed admission control*).

In the previous example, virtual machines would not power on when Silo 102 experienced failure because by default an HA cluster is configured to use strict admission control. Figure 11.22 shows an HA cluster configured to use the default setting of strict admission control.

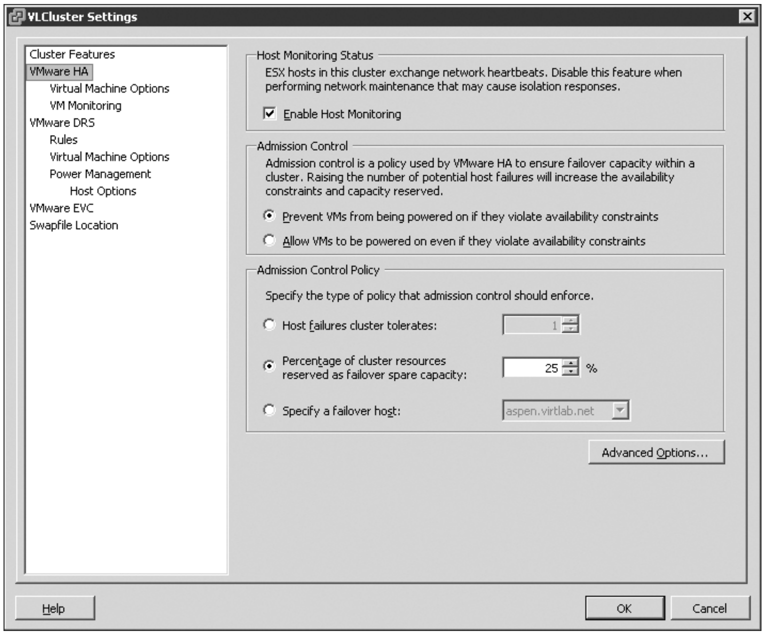


FIGURE 11.22 Strict admission control for an HA cluster prioritizes resource balance and fairness over resource availability.

Host Monitoring Status

ESX hosts in this cluster exchange network heartbeats. Disable this feature when performing network maintenance that may cause isolation responses.

☑ Enable Host Monitoring

Admission Control

Admission control is a policy used by VMware HA to ensure failover capacity within a cluster. Raising the number of potential host failures will increase the availability constraints and capacity reserved.

● Prevent VMs from being powered on if they violate availability constraints

○ Allow VMs to be powered on even if they violate availability constraints

Admission Control Policy

Specify the type of policy that admission control should enforce.

○ Host failures cluster tolerates: 1

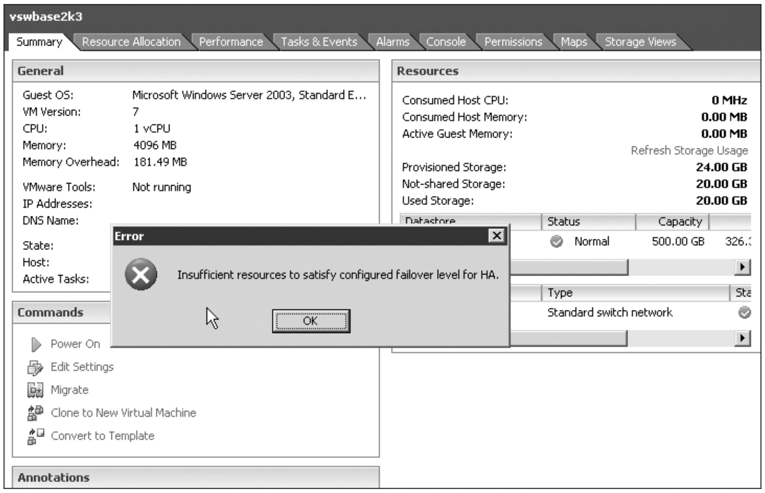
● Percentage of cluster resources reserved as failover spare capacity: 25 %

○ Specify a failover host: aspen.virtlab.net

With strict admission control, the cluster will reach a point at which it will no longer start virtual machines. Figure 11.23 shows a cluster configured for two-node failover. A virtual machine with more than 3GB of memory reserved is powering on, and the resulting error is posted, stating that insufficient resources are available to satisfy the configured HA level.

If the admission control setting of the cluster is changed from strict admission control to guaranteed admission control, then virtual machines will power on even in the event that the HA failover level is jeopardized.

Figure 11.24 shows a cluster reconfigured to use guaranteed admission control.



Guest OS: Microsoft Windows Server 2003, Standard E...
VM Version: 7
CPU: 1 vCPU
Memory: 4096 MB
Memory Overhead: 181.49 MB
VMware Tools: Not running
IP Addresses:
DNS Name:
State:
Host:
Active Tasks:

Consumed Host CPU: 0 MHz
Consumed Host Memory: 0.00 MB
Active Guest Memory: 0.00 MB
Provisioned Storage: 24.00 GB
Not-shared Storage: 20.00 GB
Used Storage: 20.00 GB

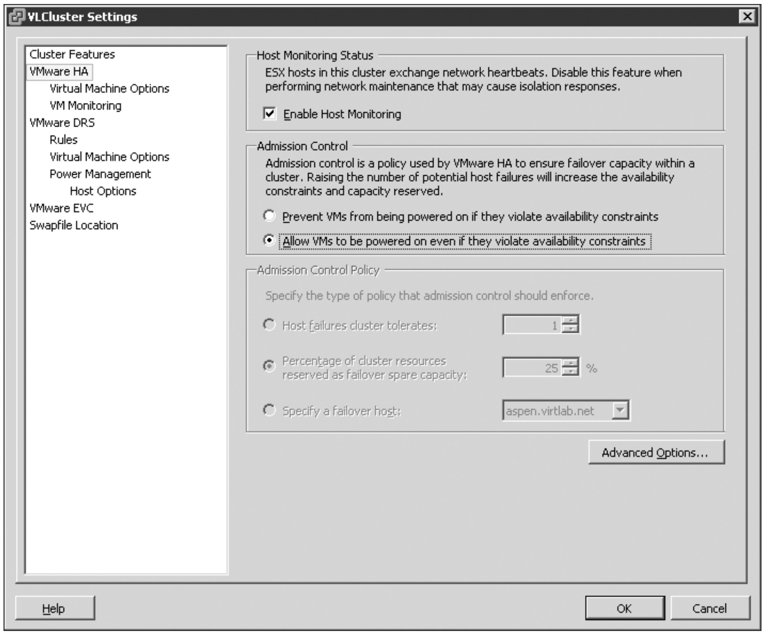
Datastore	Status	Capacity	
	Normal	500.00 GB	326.:

	Type	Sta
	Standard switch network	

Error

Insufficient resources to satisfy configured failover level for HA.

FIGURE 11.23
Strict admission control imposes a limit at which no more virtual machines can be powered on because the HA level would be jeopardized.



Host Monitoring Status

ESX hosts in this cluster exchange network heartbeats. Disable this feature when performing network maintenance that may cause isolation responses.

☑ Enable Host Monitoring

Admission Control

Admission control is a policy used by VMware HA to ensure failover capacity within a cluster. Raising the number of potential host failures will increase the availability constraints and capacity reserved.

○ Prevent VMs from being powered on if they violate availability constraints

● Allow VMs to be powered on even if they violate availability constraints

Admission Control Policy

Specify the type of policy that admission control should enforce.

○ Host failures cluster tolerates: 1

● Percentage of cluster resources reserved as failover spare capacity: 25 %

○ Specify a failover host: aspen.virtlab.net

FIGURE 11.24
Guaranteed admission control reflects the idea that when failure occurs, availability is more important than resource fairness and balance.

With that same cluster now configured with guaranteed admission control, the virtual machine with more than 3GB of memory can now successfully power on.

Overcommitment in an HA Cluster

When the admission control setting is set to allow virtual machines to be powered on even if they violate availability constraints, you could find yourself in a position where there is more physical memory allocated to virtual machines than actually exists.

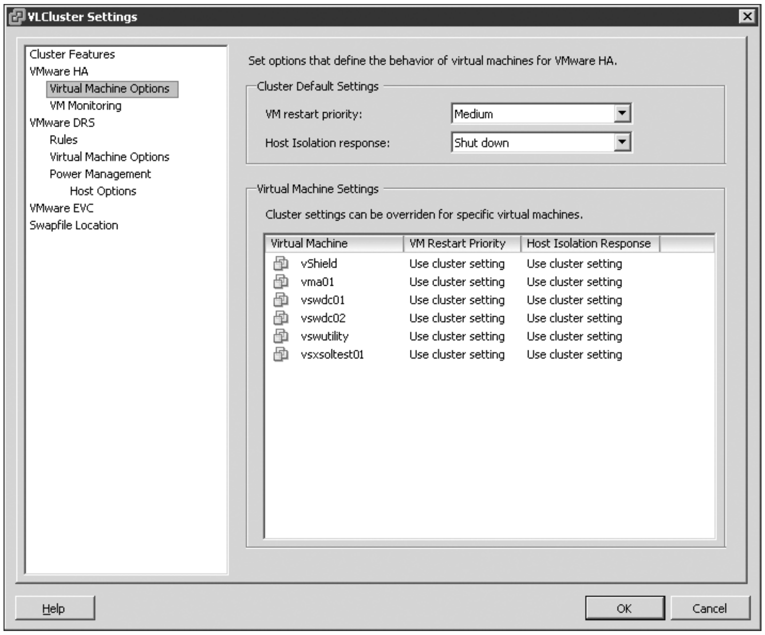
This situation, called *overcommitment*, can lead to poor performance on virtual machines that become forced to page information from fast RAM out to the slower disk-based swap file. Yes, your virtual machines will start, but after the host gets maxed out, the whole system and all virtual machines will slow down dramatically. This will increase the amount of time that HA will need to recover the virtual machines. What should have been a 20- to 30-minute recovery could end up being an hour or even more.

HA Restart Priority

Not all virtual machines are equal. There are some that are more important or more critical and that require higher priority when ensuring availability. When an ESX/ESXi host experiences failure and the remaining cluster nodes are tasked with bringing virtual machines back online, they have a finite amount of resources to fill before there are no more resources to allocate to virtual machines that need to be powered on. Rather than leave the important virtual machines to chance, an HA cluster allows for the prioritization of virtual machines. The restart priority options for virtual machines in an HA cluster include Low, Medium, High, and Disabled. For those virtual machines that should be brought up first, the restart priority should be set to High. For those virtual machines that should be brought up if resources are available, the restart priority can be set to Medium and/or Low. For those virtual machines that will not be missed for a period of time and should not be brought online during the period of reduced resource availability, the restart priority should be set to Disabled. Figure 11.25 shows virtual machines with various restart priorities configured to reflect their importance. Figure 11.25 details a configuration where virtual machines such as domain controllers, database servers, and cluster nodes are assigned higher restart priority.

The restart priority is put into place only for the virtual machines running on the ESX/ESXi hosts that experience an unexpected failure. Virtual machines running on hosts that have not failed are not affected by the restart priority. It is possible then that virtual machines configured with a restart priority of High might not be powered on by the HA feature because of limited resources, which are in part because of lower-priority virtual machines that continue to run. For example, as shown in Figure 11.26, Silo 101 hosts five virtual machines with a priority of High and five other virtual machines with priority values of Medium and Low. Meanwhile, Silo 102 and Silo 103 each hold 10 virtual machines, but of the 20 virtual machines between them, only four are considered of high priority. When Silo 101 fails, Silo 102 and Silo 103 will begin powering the virtual machines with a high priority. However, assume there were only enough resources to power on four of the five virtual machines with high priority. That leaves a high-priority virtual machine powered off while all other virtual machines of medium and low priorities continue to run on Silo 102 and Silo 103.

FIGURE 11.25
Restart priorities help minimize the downtime for more important virtual machines.



VLCluster Settings

- Cluster Features
- VMware HA
  - Virtual Machine Options
  - VM Monitoring
- VMware DRS
  - Rules
  - Virtual Machine Options
  - Power Management
    - Host Options
- VMware EVC
- Swapfile Location

Set options that define the behavior of virtual machines for VMware HA.

Cluster Default Settings

VM restart priority: Medium

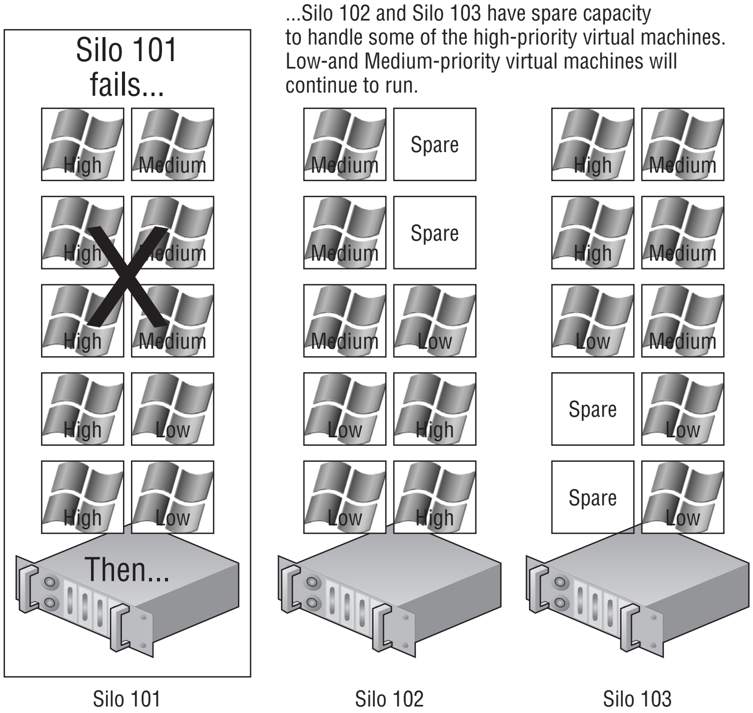
Host Isolation response: Shut down

Virtual Machine Settings

Cluster settings can be overriden for specific virtual machines.

Virtual Machine	VM Restart Priority	Host Isolation Response
vShield	Use cluster setting	Use cluster setting
vma01	Use cluster setting	Use cluster setting
vswdc01	Use cluster setting	Use cluster setting
vswdc02	Use cluster setting	Use cluster setting
vswutility	Use cluster setting	Use cluster setting
vsxsoltest01	Use cluster setting	Use cluster setting

FIGURE 11.26
High-priority virtual machines from a failed ESX/ESXi host might not be powered on because of a lack of resources—resources consumed by virtual machines with a lower priority that are running on the other hosts in an HA cluster.



At this point in the vSphere product suite, you can still manually remedy this imbalance. Any disaster recovery plan in a virtual environment built on vSphere should include a contingency plan that identifies virtual machines to be powered off to make resources available for those virtual machines with higher priority as a result of the network services they provide. If the budget allows, construct the HA cluster to ensure that there are ample resources to cover the needs of the critical virtual machines, even in times of reduced computing capacity.

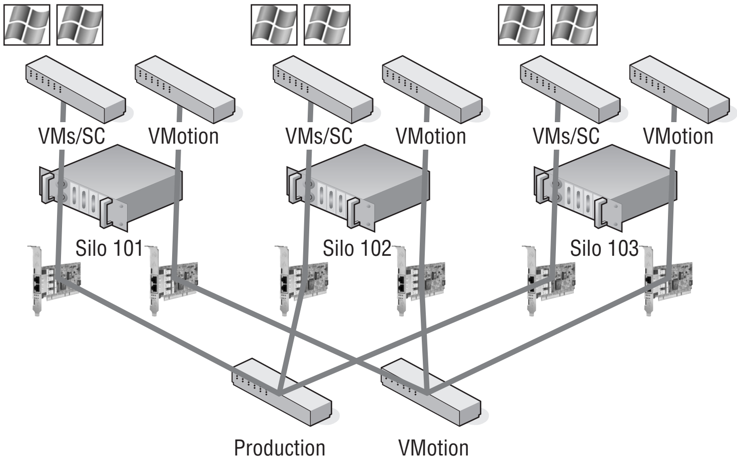
HA Isolation Response

Previously, we introduced the AAM and its role in conducting the heartbeat that occurs among all the nodes in the HA cluster. The heartbeat among the nodes in the cluster identifies the presence of each node to the other nodes in the cluster. When a heartbeat is no longer presented from a node in the HA cluster, the other cluster nodes spring into action to power on all the virtual machines that the missing node was running.

But what if the node with the missing heartbeat was not really missing? What if the heartbeat was missing, but the node was still running? And what if the node with the missing heartbeat is still locking the virtual machine files on a SAN LUN, thereby preventing the other nodes from powering on the virtual machines?

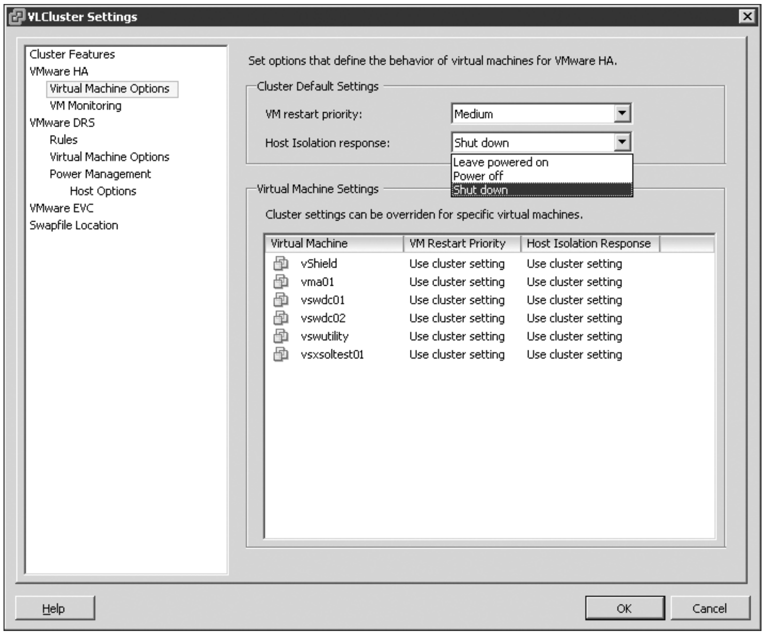
Let's look at two particular examples of a situation VMware refers to as a *split-brained* HA cluster. Let's assume there are three nodes in an HA cluster: Silo 101, Silo 102, and Silo 103. Each node is configured with a single virtual switch for VMotion and with a second virtual switch consisting of a Service Console port and a virtual machines port group, as shown in Figure 11.27.

Figure 11.27
ESX/ESXi hosts in an HA cluster using a single virtual switch for Service Console and virtual machine communication



To continue with the example, suppose that an administrator mistakenly unplugs the Silo 101 Service Console network cable. When each of the nodes identifies a missing heartbeat from another node, the discovery process begins. After 15 seconds of missing heartbeats, each node then pings an address called the *isolation response address*. By default this address is the default gateway IP address configured for the Service Console. If the ping attempt receives a reply, the node considers itself valid and continues as normal. If a host does not receive a response, as Silo 101 wouldn't, it considers itself in isolation mode. At this point, the node will identify the cluster's isolation response configuration, which will guide the host to either power off the existing virtual machines it is hosting or leave them powered on. This isolation response value, shown in Figure 11.28, is set on a per-virtual machine basis. So, what should you do? Power off the existing virtual machine? Or leave it powered on?

FIGURE 11.28
The isolation response identifies the action to occur when an ESX/ESXi host determines it is offline but powered on.



Set options that define the behavior of virtual machines for VMware HA.

Cluster Default Settings

VM restart priority: Medium

Host Isolation response: Shut down

Leave powered on
Power off
Shut down

Virtual Machine Settings

Cluster settings can be overriden for specific virtual machines.

Virtual Machine	VM Restart Priority	Host Isolation Response
vShield	Use cluster setting	Use cluster setting
vma01	Use cluster setting	Use cluster setting
vswdc01	Use cluster setting	Use cluster setting
vswdc02	Use cluster setting	Use cluster setting
vswutility	Use cluster setting	Use cluster setting
vsxsoltest01	Use cluster setting	Use cluster setting

The answer to this question is highly dependent on the virtual and physical network infrastructures in place. In our example, the Service Console and virtual machines are connected to the same virtual switch bound to a single network adapter. In this case, when the cable for the Service Console was unplugged, communication to the Service Console and every virtual machine on that computer was lost. The solution, then, should be to power off the virtual machines. By doing so, the other nodes in the cluster will identify the releases on the locks and begin to power on the virtual machines that were not otherwise included.

In the next example, we have the same scenario but a different infrastructure, so we don't need to worry about powering off virtual machines in a split-brain situation. Figure 11.29 diagrams a virtual networking architecture in which the Service Console, VMotion, and virtual machines all communicate through individual virtual switches bound to different physical network adapters. In this case, if the network cable connecting the Service Console is removed, the heartbeat will once again be missing; however, the virtual machines will be unaffected because they reside on a different network that is still passing communications between the virtual and physical networks.

Figure 11.30 shows the isolation response setting of Leave Powered On that would accompany an infrastructure built with redundancy at the virtual and physical network levels.

FIGURE 11.29
Redundancy in the physical infrastructure with isolation of virtual machines from the Service Console in the virtual infrastructure provides greater flexibility for isolation response.

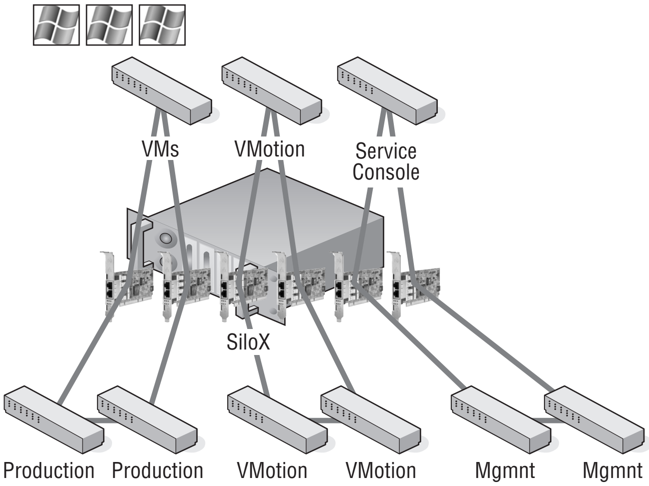
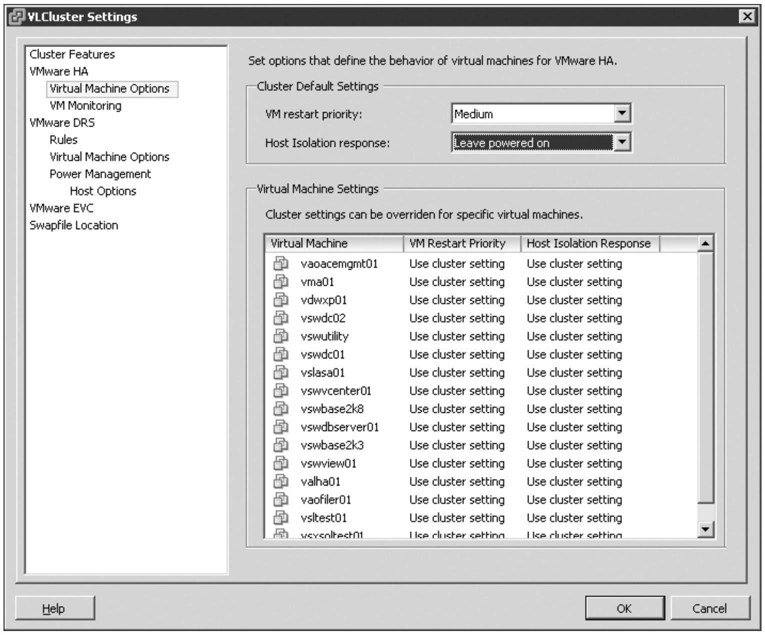


FIGURE 11.30
The option to leave virtual machines running when a host is isolated should be set only when the virtual and the physical networking infrastructures support high availability.



Set options that define the behavior of virtual machines for VMware HA.

Cluster Default Settings

VM restart priority: Medium

Host Isolation response: Leave powered on

Virtual Machine Settings

Cluster settings can be overriden for specific virtual machines.

Virtual Machine	VM Restart Priority	Host Isolation Response
vaoacemgmt01	Use cluster setting	Use cluster setting
vma01	Use cluster setting	Use cluster setting
vdwxp01	Use cluster setting	Use cluster setting
vswdc02	Use cluster setting	Use cluster setting
vswutility	Use cluster setting	Use cluster setting
vswdc01	Use cluster setting	Use cluster setting
vslasa01	Use cluster setting	Use cluster setting
vswvcenter01	Use cluster setting	Use cluster setting
vswbase2k8	Use cluster setting	Use cluster setting
vswdbserver01	Use cluster setting	Use cluster setting
vswbase2k3	Use cluster setting	Use cluster setting
vswview01	Use cluster setting	Use cluster setting
valha01	Use cluster setting	Use cluster setting
vaofiler01	Use cluster setting	Use cluster setting
vsltest01	Use cluster setting	Use cluster setting
vsxsqltest01	Use cluster setting	Use cluster setting

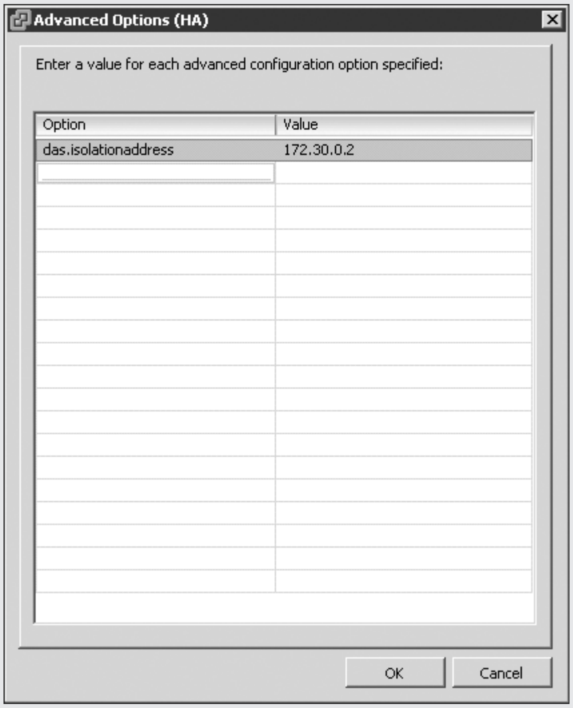
CONFIGURING THE ISOLATION RESPONSE ADDRESS

In some highly secure virtual environments, Service Console access is limited to a single, nonrouted management network. In some cases, the security plan calls for the elimination of a default gateway on the Service Console port configuration. The idea is to lock the Service Console onto the local subnet, thus preventing any type of remote network access. The disadvantage, as you might have guessed, is that without a default gateway IP address configured for the Service Console, there is no isolation address to ping as a determination of isolation status.

It is possible, however, to customize the isolation response address for scenarios just like this. The IP address can be any IP address but should be an IP address that is not going to be unavailable or taken from the network at any time.

Perform the following steps to define a custom isolation response address:

1. Use the vSphere Client to connect to a vCenter server.
2. Open the Hosts And Clusters View, right-click an existing cluster, and select the Edit Settings option.
3. Click the VMware HA node.
4. Click the Advanced Options button.
5. Enter **das.isolationaddress** in the Option column in the Advanced Options (HA) dialog box.
6. Enter the IP address to be used as the isolation response address for ESX/ESXi hosts that miss the AAM heartbeat. The following image shows a sample configuration in which the servers will ping the IP address 172.30.0.2.



Advanced Options (HA)

Enter a value for each advanced configuration option specified:

Option	Value
das.isolationaddress	172.30.0.2

7. Click the OK button twice.

This interface can also be configured with the following options:

- das.isolationaddress1: To specify the first address to try
- das.isolationaddress2: To specify the second address to try
- das.defaultfailoverhost: To identify the preferred host to failover to
- das.failuredetectiontime: To change the amount of time required for failover detection
- das.AllowNetwork: To specify a different port group to use for HA heartbeat

To support a redundant HA architecture, it is best to ensure that the Service Console port is sitting atop a NIC team where each physical NIC bound to the virtual switch is connected to a different physical switch.

Clustering is configured to give you, the administrator of an environment, a form of fault tolerance, and VMware has taken this concept to a whole other level. Although VMware does not call FT clustering, it functions the same in that FT will failover the primary virtual machine to a secondary virtual machine. VMware Fault Tolerance (FT) is based on vLockstep technology and provides zero downtime, zero data loss, and continuous availability for your applications.

That sounds pretty impressive, doesn't it? But how does it work?

Implementing VMware Fault Tolerance

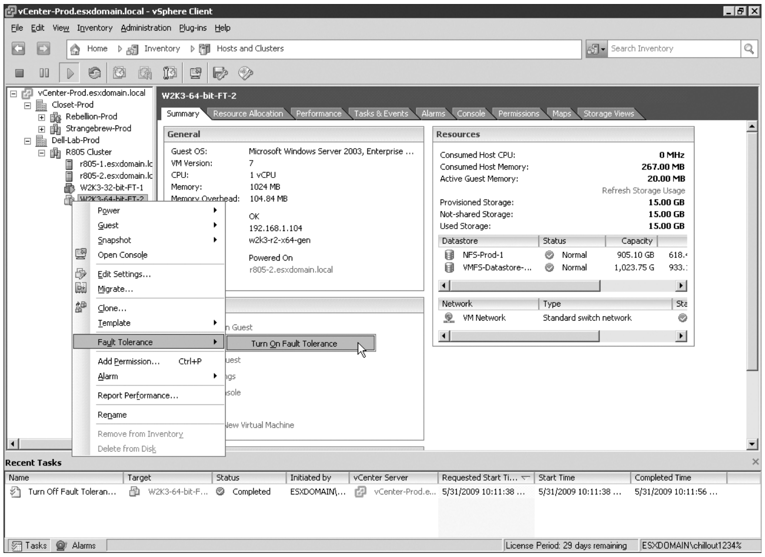
VMware FT is the evolution of ''continuous availability'' that works by utilizing VMware vLockstep technology to keep a primary machine and a secondary machine in a virtual lockstep. This virtual lockstep is actually the record/playback technology that VMware introduced in VMware Workstation in 2006. VMware FT will stream data that will be recorded, only nondeterministic events are recorded, and the replay will occur deterministically. By doing it this way, VMware has created a process that matches instruction-for-instruction and memory-for-memory to get identical results.

Deterministic means that the computer processor will execute the same instruction stream on the secondary virtual machine as to end up in the same state as the primary virtual machine. On the other hand, nondeterministic events are functions, such as network/disk/keyboard I/O as well as hardware interrupts. So, the record process will take the data stream, and the playback will perform all the keyboards and mouse clicks. It is pretty slick to move the mouse on the primary virtual machine and see it also move on the secondary virtual machine.

Perform the following steps to enable FT:

1. Starting in the vSphere client, right-click a running virtual machine, and then select Turn Fault Tolerance On, as shown in Figure 11.31.
2. A pop-up message appears to warn you that any thin disks will be converted to eagerthick, and DRS will be disabled for this virtual machine, as shown in Figure 11.32.

FIGURE 11.31
Turning on fault tolerance



Guest OS: Microsoft Windows Server 2003, Enterprise ...
VM Version: 7
CPU: 1 vCPU
Memory: 1024 MB
Memory Overhead: 104.84 MB
OK
192.168.1.104
w2k3-r2-x64-gen
Powered On
r805-2.esxdomain.local

Consumed Host CPU: 0 MHz
Consumed Host Memory: 267.00 MB
Active Guest Memory: 20.00 MB
Provisioned Storage: 15.00 GB
Not-shared Storage: 15.00 GB
Used Storage: 15.00 GB

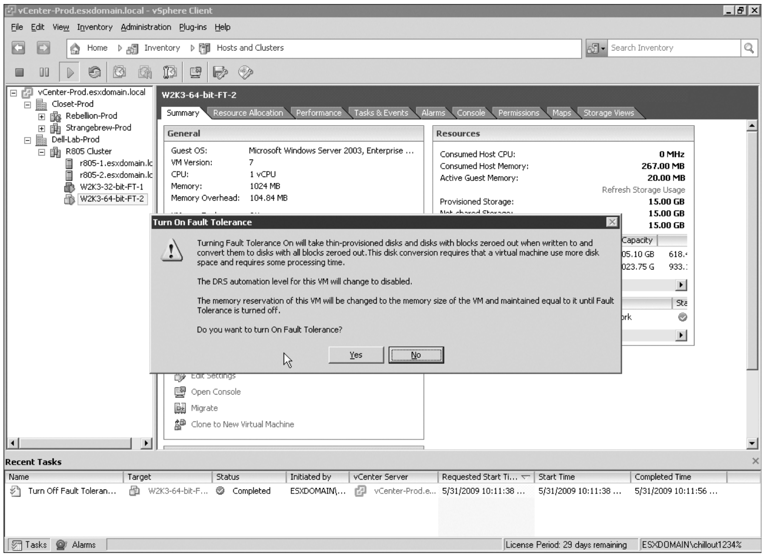
Datastore	Status	Capacity	
NFS-Prod-1	Normal	905.10 GB	618.·
VMFS-Datastore-...	Normal	1,023.75 G	933.:

Network	Type	Sta
VM Network	Standard switch network	

Name	Target	Status	Initiated by	vCenter Server	Requested Start Ti...	Start Time	Completed Time
Turn Off Fault Toleran...	W2K3-64-bit-F...	Completed	ESXDOMAIN\...	vCenter-Prod.e...	5/31/2009 10:11:38 ...	5/31/2009 10:11:38 ...	5/31/2009 10:11:56 ...

License Period: 29 days remaining ESXDOMAIN\chillout1234%

FIGURE 11.32
Warning about disks being converted to eagerthick and that DRS will be disabled on that virtual machine



Guest OS: Microsoft Windows Server 2003, Enterprise ...
VM Version: 7
CPU: 1 vCPU
Memory: 1024 MB
Memory Overhead: 104.84 MB

Consumed Host CPU: 0 MHz
Consumed Host Memory: 267.00 MB
Active Guest Memory: 20.00 MB
Provisioned Storage: 15.00 GB

Turn On Fault Tolerance

Turning Fault Tolerance On will take thin-provisioned disks and disks with blocks zeroed out when written to and convert them to disks with all blocks zeroed out.This disk conversion requires that a virtual machine use more disk space and requires some processing time.

The DRS automation level for this VM will change to disabled.

The memory reservation of this VM will be changed to the memory size of the VM and maintained equal to it until Fault Tolerance is turned off.

Do you want to turn On Fault Tolerance?

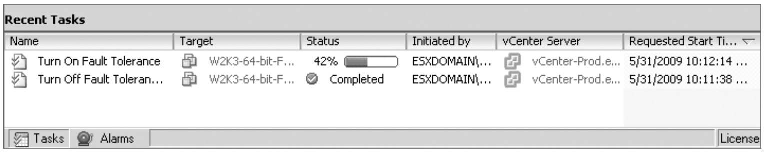
Yes No

Name	Target	Status	Initiated by	vCenter Server	Requested Start Ti...	Start Time	Completed Time
Turn Off Fault Toleran...	W2K3-64-bit-F...	Completed	ESXDOMAIN\...	vCenter-Prod.e...	5/31/2009 10:11:38 ...	5/31/2009 10:11:38 ...	5/31/2009 10:11:56 ...

License Period: 29 days remaining ESXDOMAIN\chillout1234%

3. After you have selected to enable FT, the creation task begins, as shown in Figure 11.33.

FIGURE 11.33
FT task in progress

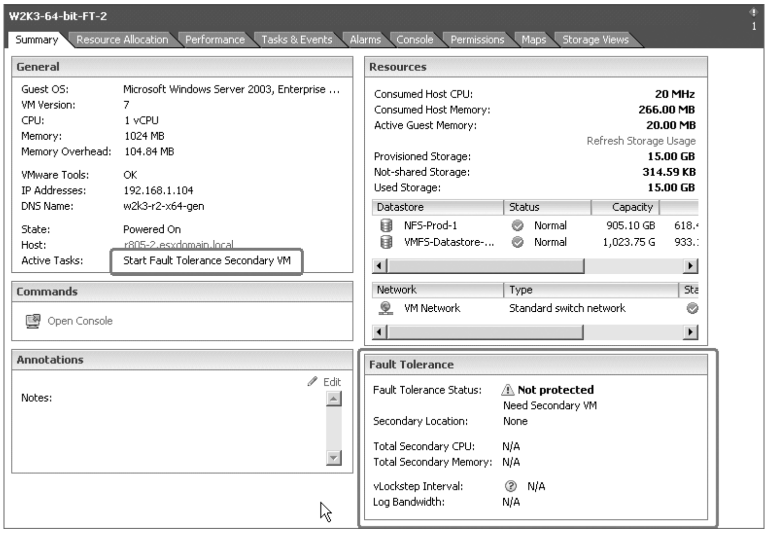


Recent Tasks

Name	Target	Status	Initiated by	vCenter Server	Requested Start Ti...
Turn On Fault Tolerance	W2K3-64-bit-F...	42%	ESXDOMAIN\...	vCenter-Prod.e...	5/31/2009 10:12:14 ...
Turn Off Fault Toleran...	W2K3-64-bit-F...	Completed	ESXDOMAIN\...	vCenter-Prod.e...	5/31/2009 10:11:38 ...

It is literally that simple. After VMware FT is turned on, vCenter Server will then initiate the creation of the secondary virtual machine, as shown in Figure 11.34, by using a special type of VMotion. Both the primary and secondary virtual machines will share a common disk between them, and using record/replay, VMware FT will then be able to keep the virtual machines in sync. Only the primary virtual machine will respond across the network, which leaves the secondary virtual machine a silent partner. You can almost compare this to active/passive cluster configuration in that only one node owns the shared network at a time. When the primary VM fails, the secondary VM takes over immediately with no break in network connection. A reverse ARP is sent to the physical switch to notify the network of the new location of the VM. Does that sound familiar? It is exactly what VMotion does when the VM switches to a new host. Once the secondary VM then becomes the primary, the creation of the new secondary VM is repeated until the sync is locked. After the sync is locked, as shown in Figure 11.35, you'll see green icons.

FIGURE 11.34
The vCenter client showing the start FT process on secondary virtual machine



General

Guest OS: Microsoft Windows Server 2003, Enterprise ...
VM Version: 7
CPU: 1 vCPU
Memory: 1024 MB
Memory Overhead: 104.84 MB

VMware Tools: OK
IP Addresses: 192.168.1.104
DNS Name: w2k3-r2-x64-gen

State: Powered On
Host: r805-2.esxdomain.local
Active Tasks: Start Fault Tolerance Secondary VM

Commands

Open Console

Annotations

Edit
Notes:

Resources

Consumed Host CPU: 20 MHz
Consumed Host Memory: 266.00 MB
Active Guest Memory: 20.00 MB
Refresh Storage Usage
Provisioned Storage: 15.00 GB
Not-shared Storage: 314.59 KB
Used Storage: 15.00 GB

Datastore	Status	Capacity	
NFS-Prod-1	Normal	905.10 GB	618.‹
VMFS-Datastore-...	Normal	1,023.75 G	933.:

Network	Type	Sta
VM Network	Standard switch network	✓

Fault Tolerance

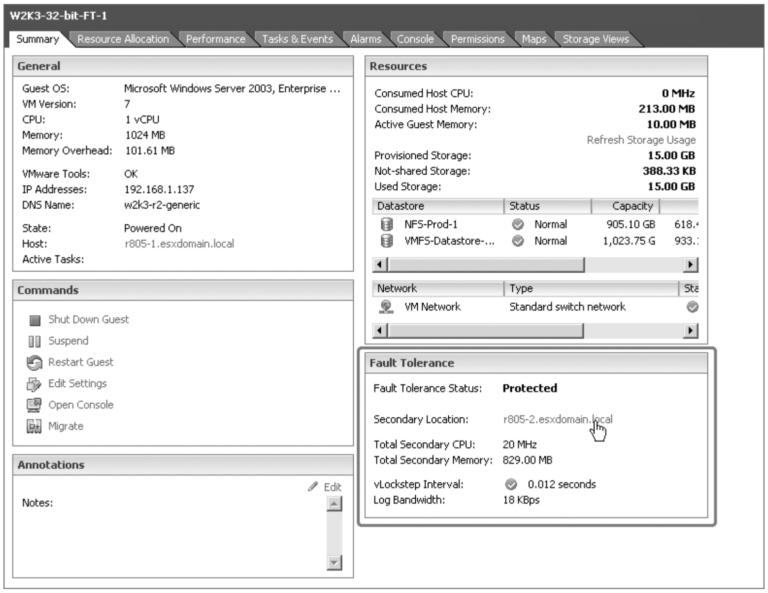
Fault Tolerance Status: ⚠ **Not protected**
Need Secondary VM
Secondary Location: None

Total Secondary CPU: N/A
Total Secondary Memory: N/A

vLockstep Interval: ? N/A
Log Bandwidth: N/A

For VMware FT to work, the virtual machines must be a member of an HA-enabled cluster. In this scenario, you will have VMotion, DRS, and HA as available services to use with VMware FT. You will need another VMkernel port group to use for FT logging, so you will need to plan your design to take that addition into account. Just like all the other physical NICs used in your host, the FT logging NIC requires Gigabit Ethernet. VMware recommends using a ''dedicated'' or ''separate'' NIC for FT logging, but in my lab I was able to set this up without a dedicated NIC.

I mention this so you know it can work, but I repeat the recommendation is for FT logging to use its own NIC.



General

Guest OS: Microsoft Windows Server 2003, Enterprise ...
VM Version: 7
CPU: 1 vCPU
Memory: 1024 MB
Memory Overhead: 101.61 MB

VMware Tools: OK
IP Addresses: 192.168.1.137
DNS Name: w2k3-r2-generic

State: Powered On
Host: r805-1.esxdomain.local
Active Tasks:

Annotations

Notes:

Resources

Consumed Host CPU: **0 MHz**
Consumed Host Memory: **213.00 MB**
Active Guest Memory: **10.00 MB**

Provisioned Storage: **15.00 GB**
Not-shared Storage: **388.33 KB**
Used Storage: **15.00 GB**

Datastore	Status	Capacity	
NFS-Prod-1	✓ Normal	905.10 GB	618.4
VMFS-Datastore-...	✓ Normal	1,023.75 G	933.1

Network	Type	Sta
VM Network	Standard switch network	✓

Fault Tolerance

Fault Tolerance Status: **Protected**

Secondary Location: r805-2.esxdomain.local

Total Secondary CPU: 20 MHz
Total Secondary Memory: 829.00 MB

vLockstep Interval: ✓ 0.012 seconds
Log Bandwidth: 18 KBps

FIGURE 11.35
The vCenter client showing the FT process is in sync

In the event of multiple host failure, VMware HA will do the following:

1. Restart the primary VM.
2. VMware FT will re-create the secondary process again until the sync is locked.

   In the case of a guest operating system failure, VMware FT will take no action because, as far as FT is concerned, the VMs are in sync. Both VMs will fail at the same time and place. VMware HA's guest failure monitor can restart the primary, and the secondary creation process will start again. Have you noticed a pattern about the secondary VMs? After the sync has failed, the secondary machine is always re-created.

There is no special hardware as such, but there are some requirements or restrictions for running FT. Let's take a look at some of these requirements/restrictions. You can also call this your checklist for VMware FT success.

- Starting with the hardware, you will need to make sure that the processors are supported. At the time of this writing, VMware FT requires Intel 31xx, 33xx, 52xx, 54xx, or 74xx or AMD 13xx, 23xx, or 83xx series processors.
- The processors listed in the previous bulleted item all support hardware virtualization (AMD-V, Intel VT), but this setting is usually disabled by default; therefore, you will need to enable this setting in the BIOS of the VMware ESX/ESXi host.
- For better performance, it is recommended that hyperthreading and power management (also known as *power-capping*) is turned off in the BIOS.

- VMware FT–protected VMs are required to be on shared storage (Fibre Channel, iSCSI, or NFS).
- You cannot use physical RDM, but virtual RDM is a supported configuration.
- You can use VMotion with an FT-protected VM, but you cannot use Storage VMotion.
- N_Port ID Virtualization (NPIV) is not supported with VMware FT.
- A virtual disk used with VMware FT must be eagerthick, but during the creation process a thin or sparsely allocated disk will be automatically converted to eagerthick.
- Gigabit Ethernet is required, and 10 Gigabit Ethernet is supported with jumbo frames enabled.
- All the hosts used for VMware FT must be in an HA-enabled cluster.
- DRS cannot be used with VMware FT–enabled virtual machines, but manual VMotion is allowed and supported on one at a time.
- The ESX hosts must be running the same version and build of VMware ESX or VMware ESXi.
- VMware FT currently supports only those VMs with one vCPU. SMP and multiprocessor VMs are not supported.
- Nested page tables/extended page tables (NPT/EPT) are not supported. VMware FT will disable NPT/EPT on the VMware ESX host.
- Hot-plugging devices are not supported, so no hardware changes when the VMs are powered on.
- USB (must be disabled) and sound devices (cannot be configured) are not supported with VMware FT–enabled VMs.
- Snapshots are not supported for VMware FT. Make sure any snapshots are applied or deleted from the VM before protecting the VM with FT.
- Virtual machines must be running at a hardware level of 7. Make sure your virtual hardware is upgraded before you begin.
- VMware FT–protected guests cannot be running a paravirtualized operating system.
- Because VMware FT is, in my opinion, the next generation of clustering, you cannot protect MSCS clusters. Remove MSCS first, or build another VM.

I also make it a habit to store any ISOs on a shared LUN that all ESX/ESXi hosts can see. This way, VMotion will still happen if an ISO gets left connected to a VM. VMware FT should be done the same way, or you will get an error message that will get reported that there is no media on the secondary VM.

You need another VMkernel port group for FT logging, and VMware recommends that you have a pair of NICs for VMotion and a pair of NICs for FT logging. That is a total of four NICs for the VMkernel.

VMware FT is not designed or meant to be run on all your virtual machines. You should use this service sparingly and take this form of fault tolerance only for your most important virtual machines. Suggested general guidelines recommend that there be no more than four to eight VMware FT–protected virtual machines—primary or secondary—on any single ESX/ESXi host. Your mileage will vary, so be cautious in your own environment. Remember, once you have primary and secondary virtual machines locked and in sync, you will be using double the resources for a protected virtual machine.

Now that you have taken a look a couple of different high availability options, let's move on to the next important thing to plan and design for. No configuration or design is complete without considering disaster recovery.

Recovering from Disasters

High availability is only one half of the ability to keep your application/systems up in day-to-day operation. The other half is disaster recovery, which is the ability to recover from a catastrophic failure. Hurricane Andrew and Hurricane Katrina demonstrated the importance of having a well-thought-out and well-designed plan in place. They also showed the importance of being able to execute that plan. Datacenters disappeared from the power of these storms, and the datacenters that remained standing and functioning did not stay functioning long when the generators ran out of gas. I believe when Hurricane Katrina came to visit New Orleans, the aftermath drove the point home that businesses need to be prepared.

I can remember what life was like before virtualization. The disaster recovery (DR) team would show up, and the remote recovery site was slated with the task of recovering the enterprise in a timely manner. A timely manner back then was at least a few days to build and install the recovery servers and then restore the enterprise from the backup media.

Sounds simple, right? Well, in theory, it was supposed to be, but there are always problems that occur during the process. First, during the recovery process, you almost never get to restore your environment at the remote datacenter location to the same make and model that you have running in your current environment. After you restore your data from your backup media, one of the joys is the pretty blue screen that you get because the drivers are different. For the most part, after the restore is finished, you can rerun the installation of the drivers for the recovery servers, but Murphy tends to show up and lay down his law.

Second, the restore process itself is another form of literal contention. If your backup strategy is not designed to consider which servers you want to recover first, then during a disaster, when you try to restore and bring up systems based on importance, you will have a lot a time wasted waiting for tape machines to become available. This contention becomes even worse if your backups span more than one tape. Speaking of tapes, it was not uncommon for tapes to become corrupt and not have the ability to be read. It was common for backups to be done and the tapes to be sent off-site, but the tapes were hardly tested until they were needed. If all goes well, in a few days you might be done, but to be honest, success was sometimes a hard thing to find.

That old-school methodology has advanced and has changed the future with it. Now, a majority of data is kept on the SAN, and the data is replicated to another SAN at your remote disaster recovery co-location site. So, your data is waiting for you when it becomes time to recover, which really speeds up the recovery process in general. At first this was an expensive undertaking because only the high-dollar enterprise SANs had this capability. Over the years, though, this is becoming more the standard and is now a must-have in any SAN environment you work with. I'll cover SAN to SAN replication options a little later.

Now let's take some time and look at the different methods and tools available to use for recovering your environments by using backups.

With the release of VMware vSphere, VMware has also enhanced the ability to work within the storage layer and has set the groundwork for third-party vendors to use the new vStorage APIs for data protection. The vStorage APIs are a framework, not a backup application, and will enable backup vendors to run backups of the virtual machines without having the VMware ESX servers perform any of the processing of these backup tasks. At the time of this writing, most third-party backup vendors are still anywhere from six months to a year from releasing or updating their own backup software to take advantage of this new programming interface. So, in the meantime, we have VMware Consolidated Backup (VCB) for backups to tape and VMware Data Recovery (VCDR) for backups to disk. I will give you more information about VCDR a little later in the chapter. VMware VCB takes the most configuration by the administrators in the field, so it is worth taking a good look at this technology, even with the understanding that VCB is no longer on the VMware road map and will be phased out as the third-party backup vendors move forward using the vStorage APIs.

Backing Up with VMware Consolidated Backup

Virtual machines are no less likely to lose data, become corrupted, or fail than their physical counterparts. And though some might argue against that point, it is most certainly the best way for you to look at virtual machines. With the opposite school of thought, you might be jeopardizing the infrastructure with overconfidence in virtual machine stability. It's better to be safe than sorry. When it comes to virtual machine backup planning, VMware suggests three different methods you can use to support your disaster recovery plan:

- Using backup agents inside the virtual machine
- Using VCB to perform virtual machine backups
- Using VCB to perform file-level backups (Windows guests only)

VCB is VMware's first entry into the backup space. (For those of you who have worked with ESX 2, I am not counting `vmsnap.pl` and `vmres.pl` as attempts to provide a backup product). VCB is a framework for backing up that integrates easily into a handful of third-party products. Although VCB can be used on its own, it lacks some of the nice features that third-party backup products bring to the table. These include features such as cataloging backups, scheduling capability, and media management backups. For this reason, I recommend that you use the VCB framework in conjunction with third-party products that have been tested.

More than likely, none of the three methods listed will suffice if used alone. As this chapter provides more details about each of the methods, you'll see how a solid backup strategy is based on using several of these methods in a complementary fashion.

Using Backup Agents in a Virtual Machine

Oh so many years ago when virtualization was not even a spot on your IT road map, you were backing up your physical servers according to some kind of business need. For most organizations, the solution involved the purchase, installation, configuration, and execution of third-party backup agents on the operating systems running on physical hardware. Now that you have jumped onto the cutting edge of technology by leading the server consolidation charge into a virtual IT infrastructure, you can still back up using the same traditional methods. Virtual machines like physical machines are targets for third-party backup tools. The downside to this time-tested model is the need to continue paying for the licenses needed to perform backups across all servers. As shown in Figure 11.36, you'll need a license for every virtual machine you want to back up:

100 virtual machines = 100 licenses. Some vendors allow for a single ESX/ESXi host license that permits an unlimited number of agent licenses to be installed on virtual machines on that host.

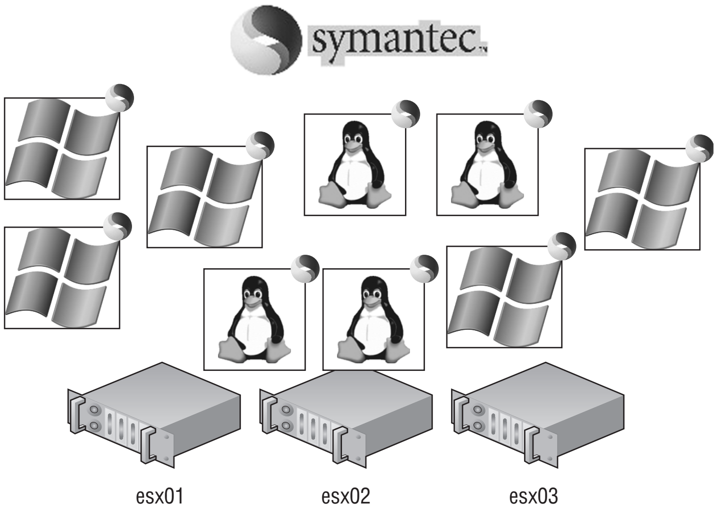


Figure 11.36
Using third-party backup agents inside a virtual machine does not take advantage of virtualization. Virtual machines are treated just like their physical counterparts for the sake of a disaster recovery plan.

In this case, virtualization has not lowered total ownership costs, and the return on investment has not changed with regard to the fiscal accountability to the third-party backup company. So, perhaps this is not the best avenue that you should travel down. With that being said, let's look at other options that rely heavily on the virtualized aspect of the guest operating system. These options include the following:

- Using VCB for full virtual machine backups
- Using VCB for single VMDK backups
- Using VCB for file-level backups

Let's take a look at each of these, starting with using VCB to capture full backups of the virtual machines.

Using VCB for Full Virtual Machine Backups

As we mentioned briefly in the opening section, VCB is a framework for backup that integrates with third-party backup software. It is a series of scripts that perform online, LAN-free backups of virtual machines or virtual machine files.

VCB for Fibre Channel...and iSCSI Too!

When first released, VCB was offered as a Fibre Channel—only solution; VMware did not support VCB used over an iSCSI storage network. The latest release of VCB offers support for use with iSCSI storage.

The requirements for VCB 1.5 include the following:

- A physical or virtual server running Windows Server 2003. (Windows 2008 is now supported as a VCB proxy. If using Windows Server Standard Edition, the VCB server must be configured not to automatically assign drive letters using Diskpart to execute automount disable and automount scrub.)
- Network connectivity for access to vCenter Server.
- Fibre Channel HBA with access to all SAN LUNs where virtual machine files are stored.
- Installation of the third-party software prior to installing and configuring VCB.

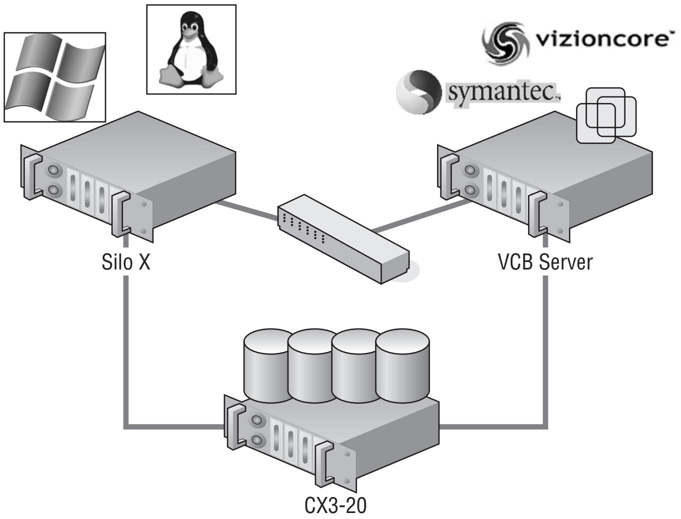
VCB on Fibre Channel Without Multipathing

The VCB proxy requires a Fibre Channel HBA to communicate with Fibre Channel SAN LUNs regarding backup processes. VCB does not, however, support multiple HBAs or multipathing software like EMC PowerPath. Insert only one Fibre Channel HBA into a VCB proxy.

Figure 11.37 looks at the VCB components and architecture.

If you do not have a SAN in your environment, you still have the ability to run VCB in your environment. Without a SAN, consolidated backup can run in what is called *LAN mode*. This mode will allow you to run VCB on a physical or virtual machine connected to your ESX/ESXi host system over the regular TCP/IP network.

Figure 11.37
A VCB deployment relies on several communication mediums, including network access to vCenter Server and Fibre Channel access to all necessary SAN LUNs.



Although considered a framework for backup, VCB can actually be used as a backup product. However, it lacks the nice scheduling and graphical interface features of third-party products like Vizioncore vRanger Pro. Two of the more common VCB commands are the following:

- `vcbVmName`: This command enumerates the various ways a virtual machine can be referenced in the `vcbMounter` command. Here's an example:

  ```
   vcbVmName -h 172.30.0.120 -u administrator -p Sybex!!
  -s ipaddr:172.30.0.24
  ```

  The following are the options:

  - `-h <vCenter Server name or IP address>`
  - `-u <vCenter Server username>`
  - `-p <vCenter Server user password>`
  - `-s ipaddr: <IP address of virtual machine to backup>`

- `vcbMounter`:

  - `-h <vCenter Server name or IP address>`
  - `-u <vCenter Server username>`
  - `-p <vCenter Server user password>`
  - `-a <name | ipaddr | moref | uuid>: <attribute value>`
  - `-t [fullvm | file]`
  - `-r <Backup directory on VCB proxy>`

VCB Proxy Backup Directory

When specifying the VCB backup directory using the `-r` parameter, do not specify an existing folder. For example, if the backup directory `E:\VCBBackups` already exists and a new backup should be stored in a subdirectory named `Server1`, then specify the subdirectory *without* creating it first. In this case, the `-r` parameter might read as follows:

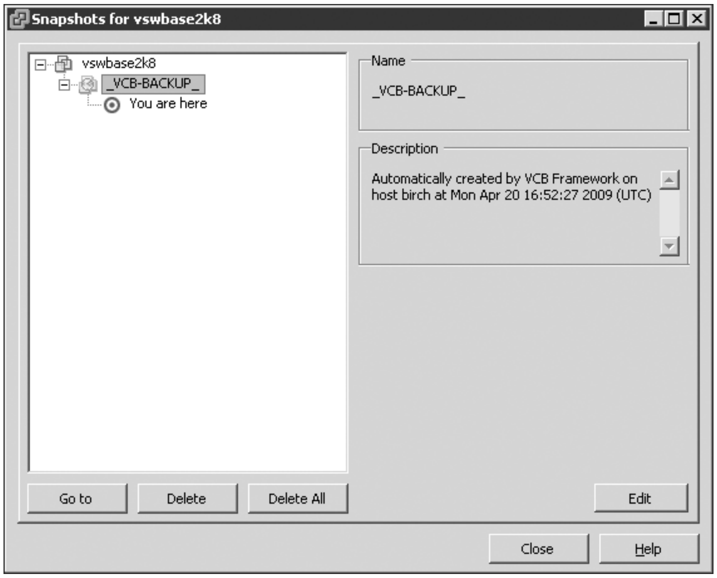
```
-r E:\VCBBackups\Server1
```

The `vcbMounter` command will create the new directory as needed. If the directory is created first, an error will be thrown at the beginning of the backup process. The error will state that the directory already exists.

When VCB performs a full backup of a virtual machine, it engages the VMware snapshot functionality to quiesce the file system and perform the backup. Remember that snapshots are not complete copies of data. Instead, a snapshot is the creation of a differencing file (or redo log) to which all changes are written. When the `vcbMounter` command is used, a snapshot is taken of the virtual machine, as shown in Figure 11.38.

Any writes that occur during the backup are done to the differencing file. Meanwhile, VCB is busy making a copy of the VMDK, which is now read-only for the duration of the backup job. Figure 11.39 details the full virtual machine backup process. After the backup job is completed, the snapshot is deleted, forcing all writes that occurred to the differencing file to be written to the virtual machine disk file.

FIGURE 11.38
You can view the snapshots that VCB created in the snapshot manager of a virtual machine.



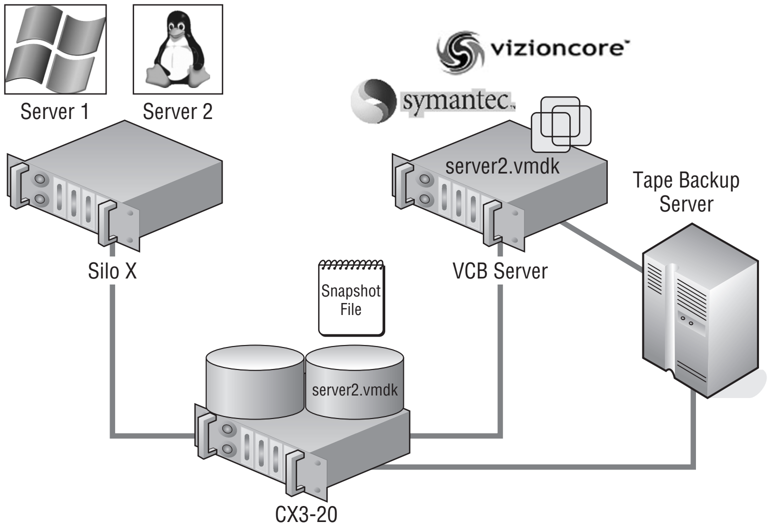
Snapshots for vswbase2k8

vswbase2k8
VCB-BACKUP
You are here

Name: _VCB-BACKUP_

Description: Automatically created by VCB Framework on host birch at Mon Apr 20 16:52:27 2009 (UTC)

FIGURE 11.39
Performing a full virtual machine backup utilizes the VMware snapshot functionality, which ensures that an online backup is correct as of that point in time.



Snapshots and VMFS Locking

Snapshots grow by default in 16MB increments, and for the duration of time it takes to grow a snapshot, a lock is held on the VMFS volume so the respective metadata can be updated to reflect the change in the snapshot. For this reason, do not instantiate a snapshot for many virtual machines at once. Although the lock is held only for the update to the metadata, the more virtual machines trying at the same time, the greater the chance of contention on the VMFS metadata. From an IT standpoint, this factor should drive your backup strategy to perform backups of many virtual machines only if the virtual machine files have been located on separate VMFS volumes.

Perform the following steps to perform a full virtual machine backup using VCB:

1. Log in to the backup proxy where VCB is installed.
2. Open a command prompt, and change directories to the `C:\Program Files\VMware\VMware Consolidated Backup Framework` directory.
3. Use the `vcbVmName` tool to enumerate virtual machine identifiers. At the command prompt, enter the following:

```
vcbVmName <IP or name of VCenter Server> -u <username>
    -p <password> -s ipaddr:<IP address of virtual
    machine to backup>
```

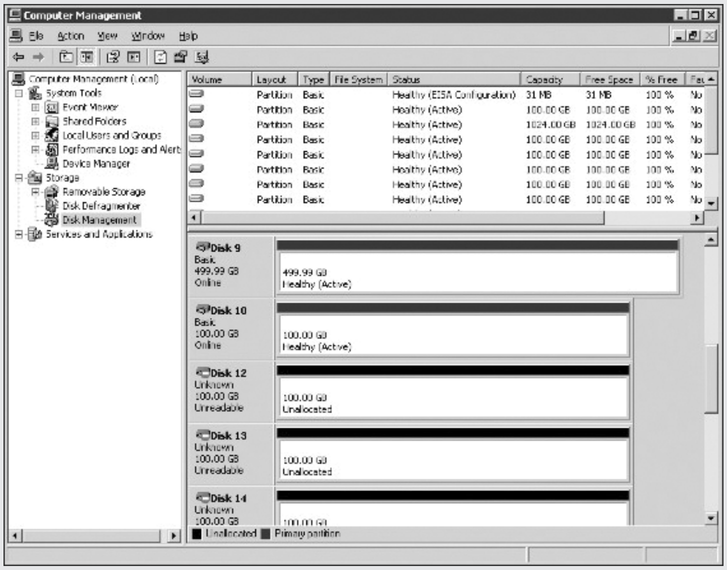
4. From the results of running the vcbVmName tool, select which identifier to use (`moref`, `name`, `uuid`, or `ipaddr`) in the `vcbMounter` command.
5. At the command prompt, enter the following:

```
vcbMounter -h <IP or name of VCenter Server> -u
    <username> -p <password> -a ipaddr:<IP address of
    virtual machine to backup> -t fullvm -r <VCB proxy
    backup directory>
```

After the backup is complete, you can review a list of the files in the directory provided in the backup script. Figure 11.40 shows the files created as part of the completed full backup of a virtual machine named Server 1.

Redundant Paths

Let’s look at an example of a VCB backup proxy with a single QLogic Fibre Channel HBA that is connected to a single Fibre Channel switch connected to two storage processors on the storage device. This configuration results in two different paths being available to the VCB server. The following image shows that a VCB server with a single HBA will find LUNs twice because of the redundancy at the storage-processor level.

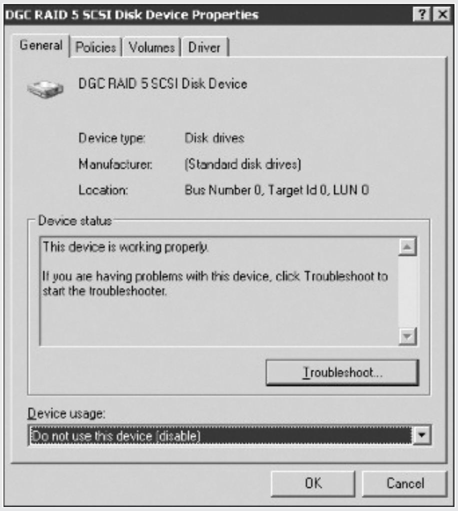


Volume	Layout	Type	File System	Status	Capacity	Free Space	% Free	Fa
	Partition	Basic		Healthy (EISA Configuration)	31 MB	31 MB	100 %	No
	Partition	Basic		Healthy (Active)	100.00 GB	100.00 GB	100 %	No
	Partition	Basic		Healthy (Active)	1024.00 GB	1024.00 GB	100 %	No
	Partition	Basic		Healthy (Active)	100.00 GB	100.00 GB	100 %	No
	Partition	Basic		Healthy (Active)	100.00 GB	100.00 GB	100 %	No
	Partition	Basic		Healthy (Active)	100.00 GB	100.00 GB	100 %	No
	Partition	Basic		Healthy (Active)	100.00 GB	100.00 GB	100 %	No
	Partition	Basic		Healthy (Active)	100.00 GB	100.00 GB	100 %	No

Disk	Partition
Disk 9 Basic 499.99 GB Online	499.99 GB Healthy (Active)
Disk 10 Basic 100.00 GB Online	100.00 GB Healthy (Active)
Disk 12 Unknown 100.00 GB Unreadable	100.00 GB Unallocated
Disk 13 Unknown 100.00 GB Unreadable	100.00 GB Unallocated
Disk 14 Unknown 100.00 GB	100.00 GB

■ Unallocated ■ Primary partition

When Disk Management shows this configuration for the older versions of VCB, it presents a problem that causes every backup attempt to fail. For the pre-VCB 1.0.3 versions, the LUNs identified as Unknown and Unreadable must be disabled in Disk Management. The option to disable is located on the properties of a LUN. The following image displays the General tab of LUN properties from Disk Management where a path to a LUN can be disabled. To remove the redundant unused paths from Computer Management, the Device Usage drop-down list should be set to Do Not Use This Device (Disable).



DGC RAID 5 SCSI Disk Device

Device type: Disk drives

Manufacturer: (Standard disk drives)

Location: Bus Number 0, Target Id 0, LUN 0

Device status

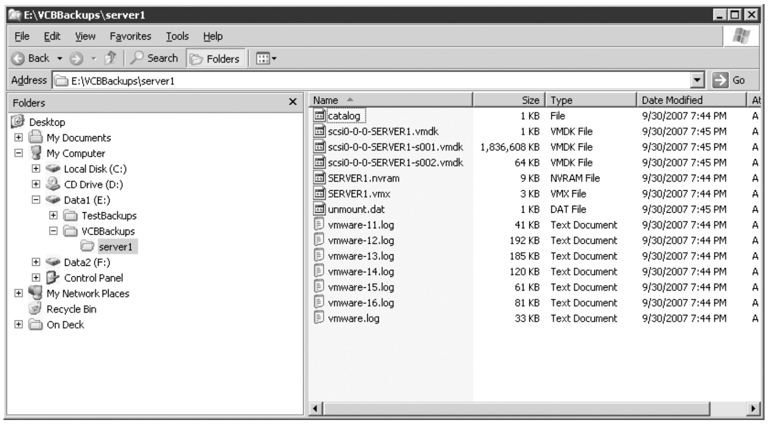
This device is working properly.

If you are having problems with this device, click Troubleshoot to start the troubleshooter.

Device usage: Do not use this device (disable)

With redundant paths disabled, this will, of course, present a problem when a LUN trespasses to another storage processor. This requires a path to the LUN that is likely disabled.

FIGURE 11.40
A full virtual machine backup that uses VCB creates a directory of files that include a configuration file (VMX), log files, and virtual machine hard drives (VMDK).



Name	Size	Type	Date Modified	At
catalog	1 KB	File	9/30/2007 7:44 PM	A
scsi0-0-0-SERVER1.vmdk	1 KB	VMDK File	9/30/2007 7:45 PM	A
scsi0-0-0-SERVER1-s001.vmdk	1,836,608 KB	VMDK File	9/30/2007 7:45 PM	A
scsi0-0-0-SERVER1-s002.vmdk	64 KB	VMDK File	9/30/2007 7:45 PM	A
SERVER1.nvram	9 KB	NVRAM File	9/30/2007 7:44 PM	A
SERVER1.vmx	3 KB	VMX File	9/30/2007 7:44 PM	A
unmount.dat	1 KB	DAT File	9/30/2007 7:45 PM	A
vmware-11.log	41 KB	Text Document	9/30/2007 7:44 PM	A
vmware-12.log	192 KB	Text Document	9/30/2007 7:44 PM	A
vmware-13.log	185 KB	Text Document	9/30/2007 7:44 PM	A
vmware-14.log	120 KB	Text Document	9/30/2007 7:44 PM	A
vmware-15.log	61 KB	Text Document	9/30/2007 7:44 PM	A
vmware-16.log	81 KB	Text Document	9/30/2007 7:44 PM	A
vmware.log	33 KB	Text Document	9/30/2007 7:44 PM	A

Using VCB for Single VMDK Backups

Sometimes a full backup is just too much: too much data that hasn't changed or too much data that is backed up more regularly and isn't needed again. For example, what if just the operating system drive needs to be backed up and not all the user data stored on other virtual machine disk files? A full backup would back up everything. For those situations, VCB provides a means of performing single virtual machine disk backups.

Perform the following steps for a single VMDK backup:

1. Log in to the backup proxy where VCB is installed.
2. Open a command prompt, and change directories to the `C:\Program Files\VMware\VMware Consolidated Backup Framework` directory.
3. Use the `vcbVmName` tool to enumerate virtual machine identifiers. At the command prompt, enter the following:

```
vcbVmName <IP or name of VCenter Server> -u <username>
    -p <password> -s ipaddr:<IP address of virtual
    machine to backup>
```

4. From the results of running the `vcbVmName` command, note the `moref` value of the virtual machine.
5. Use the following command to create a snapshot of the virtual machine:

```
vcbSnapshot -h <IP or name of VCenter Server> -u
    <username> -p <password> -c <moref value of
    virtual machine> <name of snapshot>
```

6. Note the snapshot ID (SSID) from the results of step 5.
7. Enumerate the disks within the snapshot using the `vcbSnapshot` command:

```
vcbSnapshot -h <IP or name of VCenter Server> -u
    <username> -p <password> -l <moref value of
    virtual machine> <snapshot ID>
```

8. Change to the backup directory of the virtual machine, and export the desired VMDK using the `vcbExport` command:

```
vcbExport -d <name of new VMDK copy> -s <name of
    existing VMDK>
```

9. Remove the snapshot by once again using the `vcbSnapshot` command:

```
vcbSnapshot -h <IP or name of VCenter Server> -u
    <username> -p <password> -d <moref value of
    virtual machine> <snapshot ID>
```

Using VCB for File-Level Backups

For Windows virtual machines and *only* for Windows virtual machines, VCB offers file-level backups. A file-level backup is an excellent complement to the full virtual machine or the single VMDK backup discussed in the previous sections. For example, suppose you built a virtual machine using two virtual disks: one for the operating system and one for the custom user data. The operating system's virtual disk will not change often with the exception of the second Tuesday of each month when new patches are released. So, that virtual disk does not need consistent and regular backups. On the other hand, the virtual disk that stores user data might be updated quite frequently. To get the best of both worlds and implement an efficient backup strategy, you need to do a single VMDK backup (for the OS) and file-level backup (for the data).

Perform the following steps to conduct a file-level backup using VCB:

1. Log in to the backup proxy where VCB is installed.
2. Open a command prompt, and change directories to `C:\Program Files\VMware\VMware Consolidated Backup Framework`.
3. Use the `vcbVmName` tool to enumerate virtual machine identifiers. At the command prompt, enter the following:

```
vcbVmName <IP or name of VCenter Server> -u <username>
    -p <password> -s ipaddr:<IP address of virtual
    machine to backup>
```

4. From the results of running the `vcbVmName` tool, select which identifier to use (`moref`, `name`, `uuid`, or `ipaddr`) in the `vcbMounter` command.
5. As shown in Figure 11.41, enter the following at the command prompt:

```
vcbMounter -h <IP or name of VCenter Server> -u
    <username> -p <password> -s ipaddr:<IP address of
    virtual machine to backup> -t file -r <VCB proxy
    backup directory>
```

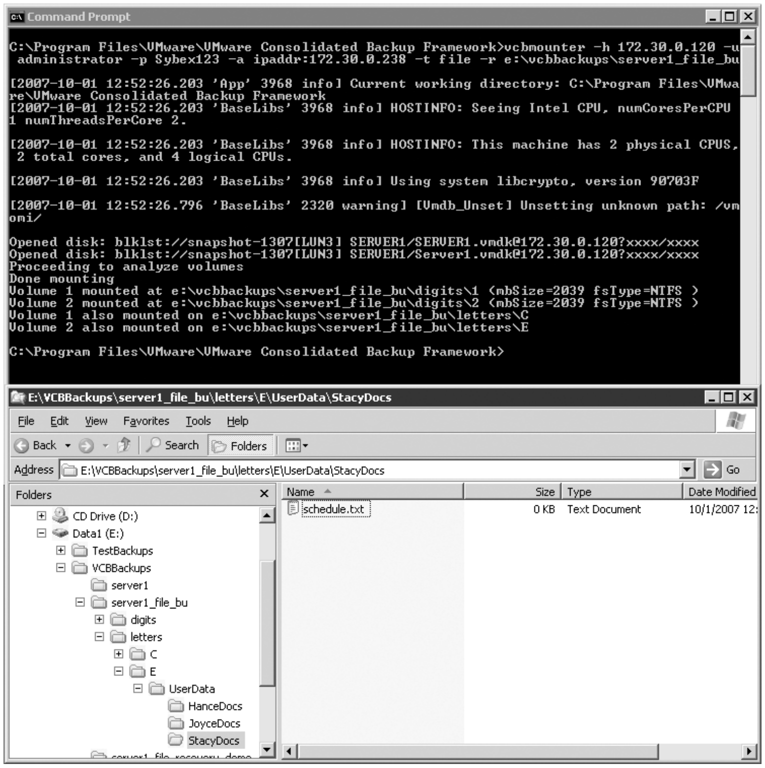
6. Browse to the mounted directory to back up the required files and folders.

7. After the file- or folder-level backup is complete, use the following command, shown in Figure 11.42, to remove the mount point:

```
mountvm -u <path to mount point>
```

8. Exit the command prompt.

FIGURE 11.41
A file- or folder-level backup begins with mounting the virtual machine drives as directories under a mount point on the VCB server.



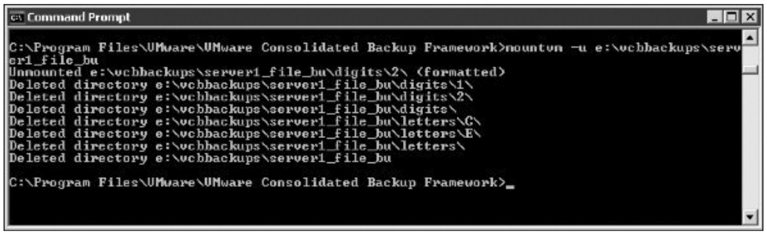
```
C:\Program Files\VMware\VMware Consolidated Backup Framework>vcbmounter -h 172.30.0.120 -u
 administrator -p Sybex123 -a ipaddr:172.30.0.238 -t file -r e:\vcbbackups\server1_file_bu

[2007-10-01 12:52:26.203 'App' 3968 info] Current working directory: C:\Program Files\VMwa
re\VMware Consolidated Backup Framework
[2007-10-01 12:52:26.203 'BaseLibs' 3968 info] HOSTINFO: Seeing Intel CPU, numCoresPerCPU
1 numThreadsPerCore 2.

[2007-10-01 12:52:26.203 'BaseLibs' 3968 info] HOSTINFO: This machine has 2 physical CPUS,
 2 total cores, and 4 logical CPUs.

[2007-10-01 12:52:26.203 'BaseLibs' 3968 info] Using system libcrypto, version 90703F

[2007-10-01 12:52:26.796 'BaseLibs' 2320 warning] [Vmdb_Unset] Unsetting unknown path: /vm
omi/

Opened disk: blklst://snapshot-1307[LUN3] SERVER1/SERVER1.vmdk@172.30.0.120?xxxx/xxxx
Opened disk: blklst://snapshot-1307[LUN3] SERVER1/Server1.vmdk@172.30.0.120?xxxx/xxxx
Proceeding to analyze volumes
Done mounting
Volume 1 mounted at e:\vcbbackups\server1_file_bu\digits\1 (mbSize=2039 fsType=NTFS )
Volume 2 mounted at e:\vcbbackups\server1_file_bu\digits\2 (mbSize=2039 fsType=NTFS )
Volume 1 also mounted on e:\vcbbackups\server1_file_bu\letters\C
Volume 2 also mounted on e:\vcbbackups\server1_file_bu\letters\E

C:\Program Files\VMware\VMware Consolidated Backup Framework>
```

E:\VCBBackups\server1_file_bu\letters\E\UserData\StacyDocs

Address E:\VCBBackups\server1_file_bu\letters\E\UserData\StacyDocs

Name	Size	Type	Date Modified
schedule.txt	0 KB	Text Document	10/1/2007 12:

FIGURE 11.42
After performing a file- or folder-level backup using the vcbMounter command, the mount point must be removed using the mountvm command.



```
C:\Program Files\VMware\VMware Consolidated Backup Framework>mountvm -u e:\vcbbackups\serv
er1_file_bu
Unmounted e:\vcbbackups\server1_file_bu\digits\2\ (formatted)
Deleted directory e:\vcbbackups\server1_file_bu\digits\1\
Deleted directory e:\vcbbackups\server1_file_bu\digits\2\
Deleted directory e:\vcbbackups\server1_file_bu\digits\
Deleted directory e:\vcbbackups\server1_file_bu\letters\C\
Deleted directory e:\vcbbackups\server1_file_bu\letters\E\
Deleted directory e:\vcbbackups\server1_file_bu\letters\
Deleted directory e:\vcbbackups\server1_file_bu

C:\Program Files\VMware\VMware Consolidated Backup Framework>_
```

If you come across a situation where a snapshot refuses to delete when you issue the mountvm -u command, you can always delete it from the snapshot manager user interface, which is accessible through the VI client.

VCB with Third-Party Products

After you master VCB framework by understanding the VCB mounter commands and the way that VCB works, then working with VCB and third-party products is an easy transition. The third-party products simply call upon the VCB framework to perform the vcbMounter command. All the while, the process is wrapped up nicely inside the GUI of the third-party product. This allows for scheduling the backups through backup jobs.

Let's look at an example with Symantec Backup Exec 11d. After the 11d product is installed, followed by the installation of VCB, a set of integration scripts can be extracted from VCB to support the Backup Exec installation. When a backup job is created in Backup Exec 11d, a pre-backup script runs (which calls vcbMounter to create the snapshot and mount the VMDKs), and after the backup job completes, a post-backup script runs to unmount the VMDKs. During the period of time that the VMDKs are mounted into the file system, the Backup Exec product has access to the mounted VMDKs in order to back them up to disk or tape, as specified in the backup job. See the following sample scripts, which perform a full virtual machine backup of a virtual machine with the IP address 192.168.4.1.

First, here's the pre-backup script example:

```
"C:\Program Files\VMware\VMware Consolidated Backup
   Framework\backupexec\pre-backup.bat" Server1_FullVM 192.168.4.1-fullvm
```

Now, here's the post-backup script example:

```
"C:\Program Files\VMware\VMware Consolidated Backup
   Framework\backupexec\post-backup.bat" Server1_FullVM 192.168.1.10-fullVM
```

There is no reference to vcbMounter or the parameters required to run the command. Behind the scenes, the pre-backup.bat and post-backup.bat files are reading a configuration file named config.js to pull defaults for some of the parameters for vcbMounter and then using the information given in the lines shown previously. When vcbMounter extracts the virtual machine files to the file system of the VCB proxy, the files will be found in a folder named 192.168.4.1-fullvm in a directory specified in the configuration file. A portion of the configuration file is shown here. The file identifies the directory to mount the backups to (F:\\mnt), as well as the vCenter server to connect to (vc01.vlearn.vmw) and the credentials to be used (administrator/Password1).

```
/*
 * Generic configuration file for VMware Consolidated Backup (VCB).
 */
 /*
 * Directory where all the VM backup jobs are supposed to reside in.
 * For each backup job, a directory with a unique name derived from the
 * backup type and the VM name will be created here.
 * If omitted, BACKUPROOT defaults to c:\\mnt.
 *  * Make sure this directory exists before attempting any VM backups.
 */ BACKUPROOT="F:\\mnt";/*
 * URL that is used by "mountvm" to obtain the block list for a
* disk image that is to be mounted on the backup proxy.
*
* Specifying this option is mandatory. There is no default
* value.  */ HOST="vc01.learn.vmw"; /*
* Port for communicating with all the VC SDK services.
* Defaults to 902
*/ // PORT="902";
/*
* Username/password used for authentication against the mountvm server.
* Specifying these options is mandatory.
*/ USERNAME="administrator"; PASSWORD="Password1";
```

It is this combination of the configuration file with the parameters passed at the time of execution that results in a successful mount and copy of the virtual machine disk files followed by an unmount.

Depending on the size of the virtual machines to be backed up, it might be more feasible to back up to disk and then create a second backup job to take the virtual machine backups to a tape device.

Now that you have looked at the backup process, the next logical step is to take a look at the restoration or restore process of the virtual machines.

Restoring with VMware Consolidated Backup

Restoring data in a virtual environment can take many forms. If you are using VCB in combination with an approved third-party backup application, there are three specific types of restores:

- Centralized restore: One backup agent on the VCB proxy.
- Decentralized restore: Several backup agents installed around the network, but not every system has one.
- Self-service restore: Each virtual machine contains a backup agent.

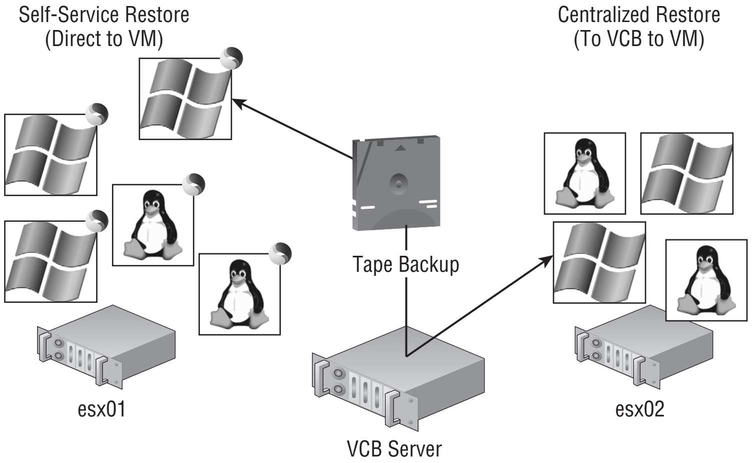
Why am I discussing backup agents in the restore section? Remember, the number of backup agents purchased directly influences the virtual machines that can also be restored directly. No matter how you implant the whole backup/restore process, you must understand that it’s ‘‘either pay me now or pay me later.’’ Something that is easier to back up is often more difficult to restore. On the flip side, something that is the most difficult to back up is often easier to restore. Figure 11.43 shows the difference between the centralized restore and the self-service restore.

Self-Service Restore Is Always Quicker

If you are looking for a restore solution focused solely on the speed of the restore and administrative effort, then the self-service restore method is ideal. Of course, the price is a bit heftier than its counterparts because an agent is required in the virtual machine. A centralized restore methodology would require two touches on the data to be restored. The first touch gets the data from the backup media to the VCB proxy server, and the second touch gets the data from the VCB proxy server back into the virtual machine. The latter happens via standard Server Message Block (SMB) or Common Internet File System (CIFS) traffic in a Windows environment. This is a literal `\\`*`servername`*`\`*`sharename`* copy of the data back to the virtual machine where the data exists.

Perhaps the best solution is to find a happy, solid relationship between the self-service restore and the centralized restore methods. This way you can reduce (not necessarily minimize) the number of backup agents while still allowing critical virtual machines to have data restored immediately.

FIGURE 11.43
Backup agents are not just for backup. They also allow restore capability. The number of backup agents purchased and installed directly affects the recovery plan.



To demonstrate a restore of a full virtual machine backup, let's continue with the earlier examples. Server 1 at this point has a full backup created. Figure 11.44 shows that Server 1 has now been deleted and is gone.

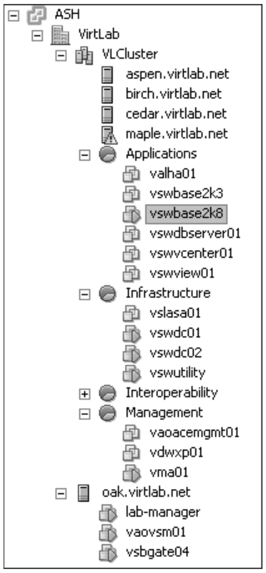
Restoring a Full Virtual Machine Backup

When bad things happen, such as the deletion or corruption of a virtual machine, a restore from a full virtual machine backup will return the environment to the point in time when the backup was taken.

Perform the following steps on a virtual machine from a full virtual machine backup:

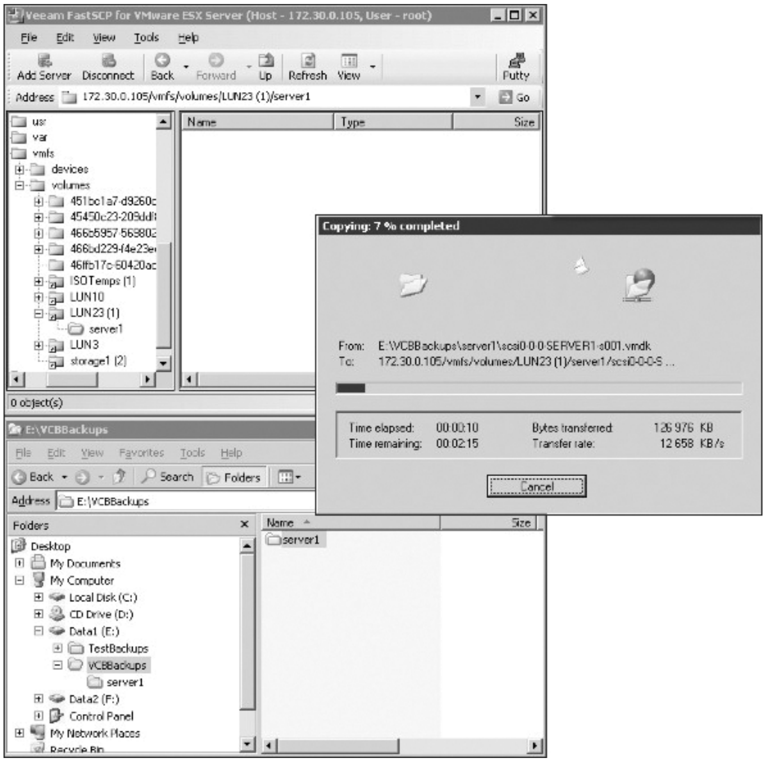
1. Connect to the VCB proxy, and use FastSCP or WinSCP to establish a secure copy protocol session with the remote host. Shown in Figure 11.45, the data from the `E:\VCBBackups\Server1` folder can be copied into a temporary directory in the Service Console. The temporary directory houses all the virtual machine files from the backup of the original virtual machine.

FIGURE 11.44
A server from the inventory is missing, and a search through the datastores does not locate the virtual machine disk files.



```
ASH
  VirtLab
    VLCluster
      aspen.virtlab.net
      birch.virtlab.net
      cedar.virtlab.net
      maple.virtlab.net
      Applications
        valha01
        vswbase2k3
        vswbase2k8
        vswdbserver01
        vswvcenter01
        vswview01
      Infrastructure
        vslasa01
        vswdc01
        vswdc02
        vswutility
      Interoperability
      Management
        vaoacemgmt01
        vdwxp01
        vma01
    oak.virtlab.net
      lab-manager
      vaovsm01
      vsbgate04
```

FIGURE 11.45
The FastSCP utility, as the name proclaims, offers a fast, secure copy protocol application to move files back and forth between Windows and ESX.



Veeam FastSCP for VMware ESX Server (Host - 172.30.0.105, User - root)

Address 172.30.0.105/vmfs/volumes/LUN23 (1)/server1

0 object(s)

Copying: 7 % completed

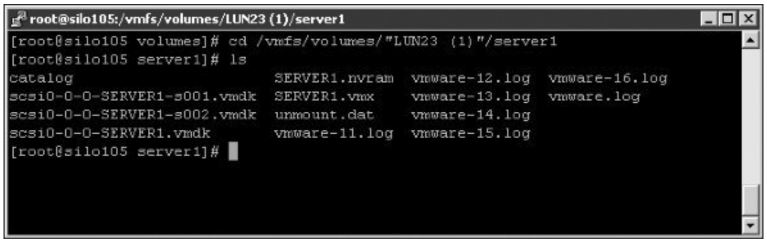
From: E:\VCBBackups\server1\scsi0-0-0-SERVER1-s001.vmdk
To: 172.30.0.105/vmfs/volumes/LUN23 (1)/server1/scsi0-0-0-S ...

Time elapsed: 00:00:10 Bytes transferred: 126 976 KB
Time remaining: 00:02:15 Transfer rate: 12 658 KB/s

E:\VCBBackups

Address E:\VCBBackups

2. Upon completion of the restore to a temporary location process, verify the existence of the files by navigating to the shared site, as shown in Figure 11.46. Use `Putty.exe` to connect to the Service Console, and navigate to the temporary directory where the backup files are stored. Then use the `ls` command to list all the files in the temporary directory.



```
[root@silo105 volumes]# cd /vmfs/volumes/"LUN23 (1)"/server1
[root@silo105 server1]# ls
catalog                     SERVER1.nvram  vmware-12.log  vmware-16.log
scsi0-0-0-SERVER1-s001.vmdk  SERVER1.vmx    vmware-13.log  vmware.log
scsi0-0-0-SERVER1-s002.vmdk  unmount.dat    vmware-14.log
scsi0-0-0-SERVER1.vmdk      vmware-11.log  vmware-15.log
[root@silo105 server1]# 
```

FIGURE 11.46
Virtual machine files needed for the restore are located in the temporary directory specified in the command.

3. From a command prompt, enter the following, all on one line:

```
vcbRestore -h 172.30.0.120 -u administrator -p
Sybex123 -s <path to temp directory>
```

4. Upon completion of the restore from the temporary location process, verify the existence of the files by navigating to datastore or by quickly glancing at the tree view of vCenter Server.

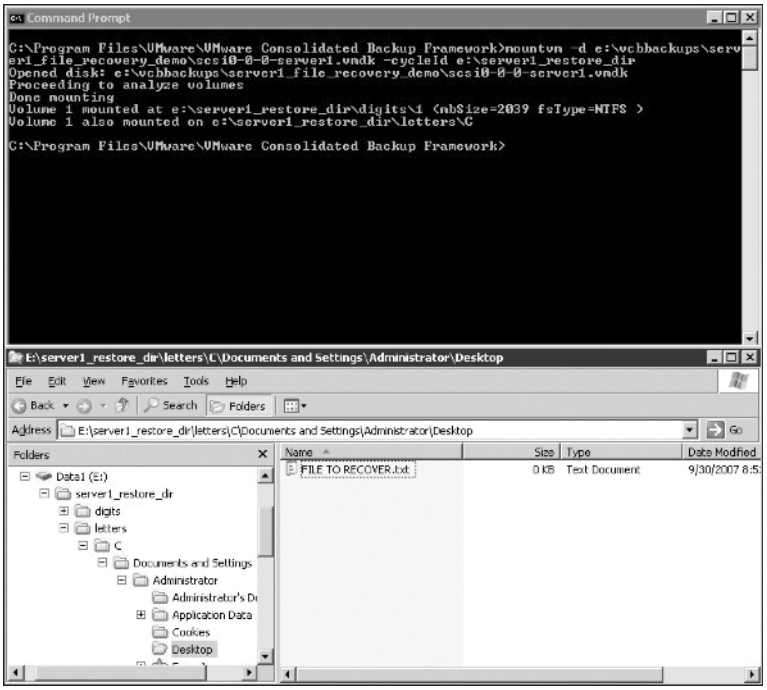
Restoring a Single File from a Full Virtual Machine Backup

Problems in the datacenter are not always as catastrophic as losing an entire virtual machine because of corruption or deletion. In fact, it is probably more common to experience minor issues like corrupted or deleted files. A full virtual machine backup does not have to be restored as a full virtual machine. Using the `mountvm` tool, it is possible to mount a virtual machine hard drive into the file system of the backup proxy (VCB) server. After the hard drive is mounted, it can be browsed the same as any other directory on the server.

PUTTING FILES INTO A VMDK

Files cannot be put directly into a VMDK. Restoring files directly to a virtual machine requires a backup agent installed on the virtual machine.

Let's say that a virtual machine named Server 1 has a full backup that has been completed. An administrator deletes a file named `FILE TO RECOVER.txt` that was on the desktop of his or her profile on Server 1 and now needs to recover the file. (No, it's not in the Recycle Bin anymore.) Using the `mountvm` command, the VMDK backup of Server 1 can be mounted into the file system of the VCB proxy server, and the file can be recovered. Figure 11.42 shows the `mountvm` command used to mount a backup VMDK named `scsi0-0-0-server1.vmdk` into a mount point named `server1_restore_dir` off the root of the `E:` drive. Figure 11.47 also shows the Windows Explorer application drilled into the mounted VMDK.



```
C:\Program Files\VMware\VMware Consolidated Backup Framework>mountvm -d e:\vcbbackups\serv
er1_file_recovery_demo\scsi0-0-0-server1.vmdk -cycleId e:\server1_restore_dir
Opened disk: e:\vcbbackups\server1_file_recovery_demo\scsi0-0-0-server1.vmdk
Proceeding to analyze volumes
Done mounting
Volume 1 mounted at e:\server1_restore_dir\digits\1 (nbSize=2039 fsType=NTFS )
Volume 1 also mounted on e:\server1_restore_dir\letters\C

C:\Program Files\VMware\VMware Consolidated Backup Framework>
```

E:\server1_restore_dir\letters\C\Documents and Settings\Administrator\Desktop

Address: E:\server1_restore_dir\letters\C\Documents and Settings\Administrator\Desktop

Name	Size	Type	Date Modified
FILE TO RECOVER.txt	0 KB	Text Document	9/30/2007 8:5

FIGURE 11.47
The mountvm command allows a VMDK backup file to be mounted into the VCB proxy server file system, where it can be browsed in search of files or folders to be recovered.

Perform the following steps to conduct a single file restore from a full virtual machine backup:

1. Log in to the VCB proxy, and navigate to the directory holding the backup files for the virtual machine that includes the missing file.
2. Browse the backup directory, and note the name of the VMDK to mount to the VCB file system.
3. Open a command prompt, and change to the `C:\Program Files\VMware\VMware Consolidated Backup Framework` directory.
4. Enter the following command:

   ```
   mountvm -d <name of VMDK to mount> -cycleId <name of    mount point>
   ```

5. Browse the file system of the VCB proxy server to find the new mount point. The new mount point will contain a subdirectory named `Letters` followed by a directory for the drive letter of the VMDK that has been mounted. These directories can now be browsed and manipulated as needed to recover the missing file.
6. After the file or folder recovery is complete, enter the following command:

   ```
   mountvm -u <mount point name>
   ```

7. Close the command prompt window.

User Data Backups in Windows

Although VCB offers the functionality to mount the virtual machine hard drives for file- or folder-level recovery, I recommend that you back up your custom user data drives and directories on a more regular basis than a full virtual machine backup. In addition to the methods discussed here like the file- and folder-level backups with VCB or third-party backup software like Vizioncore vRanger Pro, there are also tools like Shadow Copies of Shared Folders that are native to the Windows operating system.

Shadow Copies of Shared Folders builds off the Volume Shadow Services available in Windows Server 2003 and newer. It offers scheduled online backups to changes in files that reside in shared folders. The frequency of the schedule determines the number of previous versions that will exist up to the maximum of 63. The value in complementing a VCB and Vizioncore backup strategy with shadow copies is in the restore ease. Ideally, with shadow copies enabled, users can be trained on how to recover deletions and corruptions without involving the IT staff. Only when the previous version is no longer in the list of available restores will the IT staff need to get involved with a single file restore. And for the enterprise-level shadow copy deployment, Microsoft has recently released the System Center Data Protection Manager (SCDPM). SCDPM is a shadow copy on steroids, which is used to provide online frequent backups of files and folders across the entire network.

For more information on Shadow Copies of Shared Folders, visit Microsoft's website at `http://www.microsoft.com/windowsserver2003/techinfo/overview/scr.mspx`.

For more information on System Center Data Protection Manager, visit Microsoft's website at `http://www.microsoft.com/systemcenter/dpm/default.mspx`.

There is another option to use for the restore process, and that is the VMware Converter tools that VMware has built in to the vSphere Server.

Restoring VCB Backups with VMware Converter Enterprise

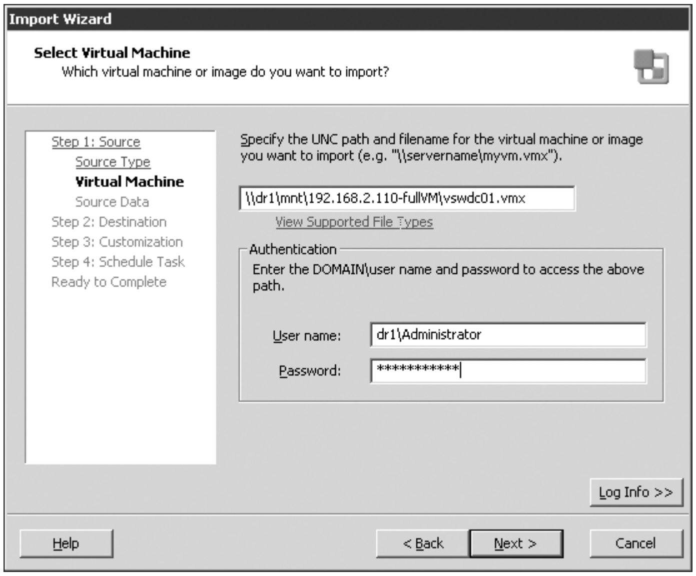
Perhaps one of the best new features of vCenter Server is the integration of the VMware Converter Enterprise. But to add to its benefit, VMware extended the functionality of the VMware Converter to allow it to perform restores of backups that were made using VMware Consolidated Backup.

During the import process, shown in Figure 11.48, you will need to provide the UNC path to the VMX file for the virtual machine to be restored.

The examples in the previous two figures show the configuration for a backup server named DR1 with a folder that has been shared as MNT. Therefore, the appropriate path for the VMX file of the virtual machine to be restored would be `\\DR1\MNT\192.168.168.8-fullVM\VAC-DC3.vmx`. The remaining steps of the Import Wizard are identical to those outlined in Chapter 7.

This particular feature alone makes vCenter an invaluable tool for building a responsive disaster recovery plan.

With the release of VMware vSphere and specifically the release of the vStorage APIs, VMware has come out with vCenter Disaster Recovery (VCDR).



Select Virtual Machine
Which virtual machine or image do you want to import?

Specify the UNC path and filename for the virtual machine or image you want to import (e.g. "\\servername\myvm.vmx").

\\dr1\mnt\192.168.2.110-fullVM\vswdc01.vmx

Authentication
Enter the DOMAIN\user name and password to access the above path.

User name: dr1\Administrator
Password: ***********

FIGURE 11.48
The VCB backup files include a VMX file with all the data about the virtual machine.

Implementing VMware Data Recovery

VCDR is a disk-based backup and recovery solution. This solution fully integrates with VMware vCenter Server to enable centralized and efficient management backup jobs, and it also includes data deduplication. It is my understanding, at the time of this writing, that VMware Data Recovery will not be a replacement for VCB but rather an enhancement and add-on to vSphere's vCenter Server.

So, how does VCDR work? VMware Data Recovery is composed of three main components. The first component is the VCDR virtual appliance that will manage the backup and recovery process. The second component is the user interface plug-in for VMware vCenter Server. The third and last component is the deduplicated destination storage.

Using the VMware vCenter Server interface, you can pick the virtual machines that you want to protect. You can then schedule the backup job, configure the data retention policy, and select the destination disk that the backup will go to. VMware vCenter Server will then send the job information to the VMware Data Recovery virtual appliance to start the backup process by initiating the point-in-time snapshots of the protected virtual machine. Like its predecessor, VMware Data Recovery frees up network traffic on the LAN by mounting the snapshot directly to the VMware Data Recovery virtual appliance. After the snapshot is mounted, the virtual appliance begins streaming the block-level data directly to the destination storage. It is during this streaming process before the data gets to the destination disks that the VMware Data Recovery appliance will deduplicate the data to ensure the redundant data is eliminated. After all the data has been written to the destination disk, the VMware Data Recovery appliance will then dismount the snapshot and then apply the snapshot to the virtual machine.

The recovery process is a point-in-time file-level or complete system restoration. The VMware Date Recovery virtual appliance will retrieve and stream the specific blocks of data that are needed for the restore. The virtual appliance will efficiently transfer only that data that has changed. This speeds up and streamlines the process. When restoring a single file, or file-level restore, the process is initiated from inside the virtual machine console.

This process in itself is a great improvement from the processes that were needed using VCB. One of my biggest complaints about VCB was the lack of central management, and VCDR really addresses those needs and brings ease of use to the table.

What makes VCDR so much different from VCB is that VCDR plugs into the vStorage APIs to accomplish its tasks. Currently, VCDR is a backup-to-disk product for VMs only, and the appliance does not have an interface to allow for backing up to tape. When using VCB, you can have the backup provider you use with your physical boxes send your backup data to tape, but there is no ability to do the same with the VCDR virtual appliance. Until this restriction has been addressed, VCB will be around for the given future.

VCB allows the use of the same backup software product that you have been running to back up the rest of your physical environment. VCB's life cycle is limited, and VMware has actually taken VCB off its road map for future development. Moving forward, third-party backup software vendors will be updating their solutions to use the new vStorage framework APIs for data protection. These APIs are the framework of the future and not a specific application.

Sometimes when you need to find a solution to a specific problem, you might find the ability to reuse a solution that was designed for something else. Let's take a look at what I call the *office in a box*.

Implementing an Office in a Box

Administrators working in the IT field are expected to find and deploy solutions for problems that they encounter in the field. I do not think we come up with answers to problems but rather solution to problems. These solutions can sometimes be applied to a completely different problem, and the office in a box was one of those solutions.

I have worked in companies that made it a habit to buy or open new offices around the country and the world. After the acquisition was made, the race was on to order equipment, get it shipped to the location, and then get the systems up and running. The typical order cycle would sometime fall way short of what was wanted for the ones leading the charge. This leaves a simple question to ask. What is the quickest way to design, order, and build this new environment? Most of the time, it was an office that was acquired, and the final end result was to fully migrate the original systems to the new environment quickly.

A solution that I designed for this was to create what I call the *office in a box*. I would take a stand-alone server that was fully configured with a VMware ESX/ESXi host and have all the virtual machines that would be needed in an office up and going. There was a domain controller, Microsoft Exchange Server, Microsoft SQL Server, and so on. The idea was to have this confirmation running as a remote site that would be receiving updates from the domain. When the time came that the system was needed, then it would be shipped overnight to the location where it was needed and brought online. Now we have an office that would be fully working in a day. We could start moving data into this site right away, and after the new permanent equipment arrived, it was a very easy process to migrate the virtual machines over and not lose a beat as far as the migration process.

So, how does this fit into disaster recovery? If you do not have a permanent remote datacenter in place, then after a location has been found, you can use this process to help speed up the recovery process. If it takes at least a little time to get your backup tapes delivered, then you can have the infrastructure in place about the same time and will have Active Directory functionality right off the bat. This should give you the appearance of being up and running much quicker.

Now let's take this design and change this just a little. We just got done talking about VCB and VCDR, so what if you took this stand-alone system and used it to receive the virtual machines that were just backed up with VCB and/or VCDR? You would then be able to store a copy of the backups on your systems. When disaster happens, then you could ship this stand-alone system and dramatically decrease your recovery time objectives. Sometimes a solution for one task can open doors as a solution for another task. It is the creativity of the administrator to build on what he already has done to give innovation for something entirely different. Remember, we create solutions, so sometimes you can use what you have and spend time thinking out of the box.

Replicating SANs

The next form of continuity is replication at the SAN level. What used to be a very expensive solution to deploy has become a much more mainstream configuration. The technology to do block-level replication at a SAN level at first was just for the high-end and most expensive SAN solutions but now is a standard configuration in just about all SAN solutions.

To set up this solution, a company would purchase two SANs that would be set up at different locations, and the data would be replicated between the two sites. The replication could occur via the SAN Fibre Channel network or over an IP connection. A snapshot of the LUN would be taken by the SAN, and then the data would be replicated at a block level. The greater the distance between the two sites would also increase the latency or the time the data takes to travel.

The two SANs were communicating with each other via Symmetrix Remote Data Facility (SRDF). SRDF logically pairs a LUN or a group of LUNs from each array and replicates data from one to the other either synchronously or asynchronously. An established pair of LUNs can be split so that separate hosts can access the same data independently and then be resynchronized. This might also be good for backups. Data can be replicated in synchronous or asynchronous mode.

In synchronous mode (SRDF/S), the primary array waits until the secondary array has acknowledged each write before the next write is accepted, ensuring that the replicated copy of the data is always as current as the primary. This is where the resultant latency comes into play and increases significantly with the distance.

Asynchronous SRDF (SRDF/A) transfers to the secondary array in what is known as *contained delta sets*. These contained delta sets are then transferred at defined intervals. Using this method, the remote copy of the data will never be as current as the primary copy, but this method can replicate data over very long distances and with reduced bandwidth requirements.

SRDF is proprietary EMC technology, but the fundamentals will be the same and apply to any SAN or storage vendor. All the vendors will use synchronous or asynchronous replication to move the data between sites. You now also have the ability to sync to more than one remote site at a time, which can give you even greater flexibility when creating your solutions and designs.

Earlier in this chapter I told you that you could not perform HA failover to another site, and as a general rule, this is true, but with SAN replication happening in synchronous mode, you will now be able to do this as long as the distance between the two sites is not too great. If you have two datacenters in different buildings on the same campus, you should be able to use HA to failover to the other building. Your mileage will vary from each of the different storage vendors, but the technology keeps getting better, and the distances keep getting farther.

If you had a real budget for your design and could afford a big enough and private network connection between datacenters, then you have another option. I have seen a WAN accelerator used between two different datacenters across the country and was able to watch a VMotion happen between two different VMware vSphere servers that were 1,000 miles apart. The purpose of mentioning this is that technology keeps changing at such a fast pace. What is not possible this very moment might be just around the corner.

The next big thing to come along in the SAN infrastructure is the ability to do point-in-time restores. The way this works is you have a dedicated disk to be used as the cache that will need to be big enough to hold information for a predetermined amount of time. This time frame could be hours, days, weeks, or months depending on the amount of disk you have provided for this. You know when something major changed in your environment at a specific time. Let's say a virtual machine was deleted by accident. With point-in-time restore, you can dial back in time to right before the virtual machine was deleted. Mount the LUN from that specific moment in time, and restore your virtual machine. This is only one example, and the sky is the limit as far as the practical application of this technology goes.

In this chapter, I discussed that high availability is for increasing uptime, and disaster recovery is for recovery from a problem. The bottom line, to be blunt, is that you better have both in place in your environment. High availability is an important part of any IT shop, and proper thought should be used when creating or designing a solution. You cannot stop there and absolutely must test, test, and test again any solution to make sure that it is working as designed and, most importantly, that it will work when you need it.

The Bottom Line

Understand Windows clustering and the types of clusters. Windows clustering plays a central role in the design of any high availability solution for both virtual and physical servers. Microsoft Windows clustering gives us the ability to have application failover to the secondary server when the primary server fails.

Master It Specifically with regard to Windows clustering in a virtual environment, what are three different types of cluster configurations that you can have?

Understand built-in high-availability options. VMware Virtual Infrastructure has high-availability options built in and available to you out of the box. These options help you provide better uptime for our critical applications.

Master It What are the two types of high-availability options that VMware provides in vSphere?

Understand the differences between VCB and VCDR. VMware has provided some disaster recovery options via backup and restore capability. First there was VMware Consolidated Backup (VCB), and now VMware has introduced VMware Data Center Recovery (VCDR), giving you different options for backing up and restoring your virtual machines.

Master It What are the main differences between VCB and VCDR?

Understand data replication options. There are other methods to keep continuity in your businesses. Replication of your data to a secondary location is a must to address a company's needs during a real disaster.

Master It What are three methods to replicate your data to a secondary location as well as the golden rule for any continuity plan?