BUSINESS READY SOLUTIONS FOR VIRTUAL INFRASTRUCTURE AVAILABILITY

USING DELL™ POWEREDGE™ SERVERS, DELL POWERVAULT™ STORAGE, AND VMWARE vSPHERE™

July 2009

Dell Virtualization Solutions Engineering
www.dell.com/virtualization

Contents

Introduction

With advances in virtualization technology, both at the hardware and software layer, more and more virtual machines and hence applications are being run on a single server. While this improves consolidation ratios, it also presents new challenges, especially the risk of losing multiple critical applications in case of a fault at any layer of the infrastructure. As shown in Figure 1, where availability is a concern, one should consider architecting a solution that provides availability at different layers: the virtual machine, hypervisor, server, network fabrics, and storage layer. This white paper presents a sample Dell solution that is designed to tolerate faults at multiple infrastructure layers to provide maximum level of availability for the virtual infrastructure. We start with an example solution configuration and then discuss how the configuration helps in maximizing uptime.

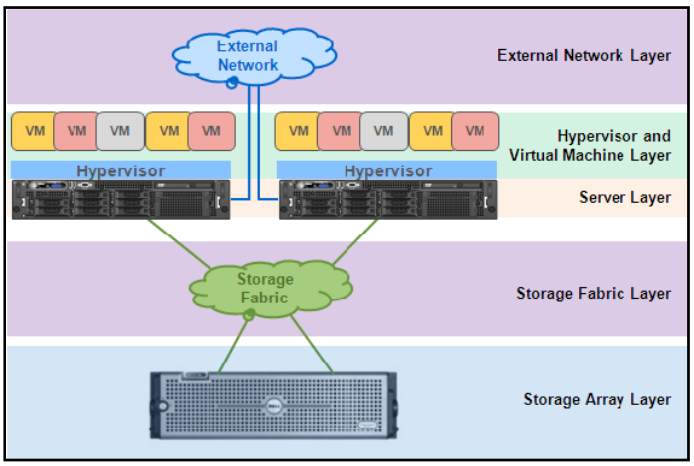

Figure 1: Sample virtual infrastructure design

Audience and Scope

The intended audience for this white paper includes IT administrators, IT managers, and channel partners who are planning to implement server virtualization to run applications for which availability is a top concern. It is expected that the reader is well versed with fundamentals of server, storage, and networking technologies and VMware vSphere 4 server virtualization software.

This white paper presents an example configuration and best practices for deploying and configuring a highly available architecture using Dell's 11th generation servers, Dell PowerVault MD3000i iSCSI array, Dell PowerConnect switches, and VMware vSphere 4 software. Within the constraints of infrastructure components, the solution discussed is designed in order to maximize availability of software and hardware infrastructure components.

Although this paper presents a sample architecture using Dell PowerEdge R710 servers, PowerVault MD3000i, PowerConnect switches, and VMware vSphere 4.0 software, the design principles and availability features presented in this paper can be used to design a highly available infrastructure using other Dell servers[1], storage products such as Dell™ EqualLogic™ and VMware vSphere software.

VMware High Availability (HA) and VMware Fault Tolerance (FT) features of vSphere 4.0 are used to provide availability at the virtual machine layer. Other solutions for providing availability may exist; however, they are outside of the scope of this paper.

Availability at the application layer is not discussed as applications may have their own built-in availability mechanisms which may vary by application type. The design approach presented in this paper may supplement such mechanisms and also help to increase availability of applications for which it is too hard to configure, manage, or sustain availability by other methods.

Based on customer requirements, further customization of the recommended architecture may be required. Consult your Dell technical sales representative for more information on how to customize this solution for your specific needs.

[1] Dell servers that are qualified to run VMware vSphere 4.0 and VMware Fault Tolerance. Refer to VMware HCL at www.vmware.com/go/hcl for a complete list of Dell hardware that is qualified to run with ESX Server software.

Solution Components

This section provides an overview of the components used for a highly available virtualized infrastructure stack solution. In the rest of this paper, we discuss how we use Dell PowerEdge R710/R610 servers, PowerVault MD3000i iSCSI storage array, PowerConnect Switches, and VMware vSphere 4 software to design a solution that provides availability at different infrastructure layers.

Dell's 11th Generation PowerEdge™ R610 and R710 servers help lower the total cost of ownership with improved performance per watt, enhanced virtualization capabilities, improved energy efficiency, and simplified systems management tools.

Availability at Memory Subsystem

The PowerEdge R610 and R710 servers support up to a maximum of 192GB DDR3 memory respectively. DDR3 memory is available as un-buffered (UDIMMs) or registered (RDIMMs). While UDIMMs offer cost and power savings benefits, RDIMMs offer higher memory capacities and additional reliability, availability, and serviceability (RAS) features that make them ideal to enhance availability for virtualized infrastructure. Availability features such as error correcting code (ECC), single device data correction (SDDC) and memory mirroring help prevent failures at the memory subsystem.

Dell supports the following memory modes and availability features:

- *Memory Optimized*: Provides maximum performance by utilizing all three available memory channels; however, the SDDC feature is supported only for x4 chips.
- *Advanced ECC Mode*: This mode uses two memory controller hubs (MCH) and ties them together to emulate a 128-bit data bus DIMM. This enables SDDC for DIMMs based on both x4 and x8 DRAM technology.
- *Mirror Mode*: This mode uses two of the three memory channels. Identical data writes are performed on each channel but the reads are alternated between the two channels. Memory mirroring provides redundancy against a failure of a DIMM on a channel. Note that since memory is mirrored across two channels, the operating system sees only half of the physical capacity installed in the server.

Refer to the System Memory section in the R610/R710 *Hardware Owner's Manual* for more details on configuring system memory.

Availability at Storage Subsystem in PowerEdge Servers

The Dell PowerEdge RAID Controller (PERC) 6 family of controllers is designed for enhanced performance, reliability, and fault tolerance. The following features enhance reliability and fault tolerance of the internal or direct attached storage connected to the PERC 6 controller:

- RAID Levels: PERC 6 controllers support RAID levels 0, 1, 5, 6, 10, 50, and 60, thus providing a range of configuration options for both performance and availability. Note that using RAID level 0 is not recommended as it provides no protection against the failure of a hard disk.
- PERC 6 RAID controllers support global hot spare, dedicated hot spare, and affinity configurations, which administrators can set up using the Dell BIOS Configuration Utility as well as Dell OpenManage Storage Management. Global hot spares can typically be used in any degraded RAID array when the hot spare has sufficient capacity to fit into the RAID container. Dedicated hot spares are reserved for a particular disk group.
- The PERC 6/I controller has 256MB of error correcting code (ECC) battery-backed cache. The PERC 6/E controller offers 256MB or 512MB of ECC battery-backed cache.
- The PERC 6 RAID controller provides advanced media-error monitoring and repair technologies, such as consistency check, background initialization, and SMART features that help prevent data loss and enable safe data retrieval.

Refer to the *Dell Open Manage Server Administrator Storage Management User's Guide* for more information on additional features and usage instructions for Dell PERC controllers.

Availability at the I/O Device layer

Dell PowerEdge R710 and R610 servers have two dual-port embedded Broadcom NetXtreme II 5709c Gigabit Ethernet NICs. Additionally, the Dell PowerEdge 710 can be configured to have:

- Two PCIe x8 and two PCIe x4 G2 or

- One PCIe x16 and two PCIe x4 G2 I/O slots.

The PowerEdge R610 has two PCIe x8 G2 I/O slots.

Multiple I/O slots and devices enable using redundant devices such as Ethernet or Fibre Channel controllers to provide multiple paths to network or storage fabrics. In combination with teaming and multi-pathing software, redundant devices provide protection against a device failure.

Other Availability Features

- Up to two hot-plug, energy-smart or high-output redundant power supplies.
- Hot-plug redundant fans (PowerEdge R710 only).

Dell PowerVault™ MD3000i is a high-performance iSCSI storage area network (SAN) designed to deliver storage consolidation, high performance, and availability for up to 16 hosts. With shared storage being a requirement for advanced virtualization features, such as live migration, the MD3000i serves as an ideal entry-level storage array for virtualized infrastructures. The following features of the MD3000i array increase availability of storage available to the hosts:

- Dual RAID controllers provide failover and redundant enclosure-management capabilities. Each controller has 512MB of battery-backed cache providing protection for up to 72 hours in case of power failure. Each RAID controller has two 1 Gb Ethernet ports, with a total of 4 ports per array. At any given time, a virtual disk or LUN on an MD3000i is controlled by one of the two RAID controllers. In case of a controller failure, all the virtual disks failover to the other controller, maintaining availability of storage to the hosts.
- The RAID controllers provide a choice of RAID levels 0, 1, 10, 5, and 6, thus providing multiple options for configuring storage for both performance and availability. Note that using RAID level 0 is not recommended as it provides no protection against the failure of a hard disk.
- Hot spare disks provide additional data protection against physical disk failures within a disk group. In case of a physical disk failure, the data on the failed disk is reconstructed to the hot spare disk. When the failed disk is replaced, the data on the hot spare is copied back to the replacement physical disk. The hosts continue to access data during the failover or failback process.
- Each RAID controller has a management port (10/100 Mb Ethernet) available for out-of-band management of the array. In case of a controller or management port/path failure, full access to the array—including all management functionality—is available through the management port on the other controller.
- Redundant power supplies and cooling.

Refer to the MD30000i documentation for more information on MD3000i features and usage.
Note: Availability against a complete catastrophic failure (e.g. failure of both RAID controllers) at the MD3000i array level is out of the scope of this paper.

Dell PowerConnect™ 5400 Series Switches deliver 24 ports (PowerConnect 5424) or 48 ports (PowerConnect 5448) of wire-speed Gigabit Ethernet with advanced security and enterprise management features to help meet the needs of organizations of all sizes. PowerConnect 5400 switches also automatically optimize for iSCSI traffic. By detecting iSCSI traffic flows and assigning iSCSI packets higher quality of service, PowerConnect switches ensure that iSCSI storage traffic is prioritized in the event of a resource contention.

To provide availability at the network layer, we use redundant Ethernet switches in combination with NIC teaming on the ESX hosts. This way we achieve protection against the failure of a switch or a network device.

Refer to PowerConnect documentation for more information on features and usage.

VMware vSphere 4 Software is the next-generation server virtualization software from VMware. Among many features in vSphere 4, VMware High Availability (HA), VMware Fault Tolerance (FT), NIC teaming and iSCSI multipathing are especially interesting with regards to enhancing availability of the virtualized infrastructure.

VMware High Availability (HA)
VMware HA provides protection for virtual machines against an ESX host failure. All hosts that are part of an HA cluster need to have access to shared storage. In case an ESX host fails, the virtual machines on the failed host are

restarted on other available hosts in the cluster. Note that depending on the application, special steps may need to be taken to resume application availability after the impacted virtual machines reboot.

VMware Fault Tolerance (FT)
VMware FT also provides protection against an ESX host failure. However, unlike VMware HA where VMs are restarted on other available hosts, if protected by VMware FT, the virtual machines continue to run in the event of a host failure. Leveraging VMware *vLockstep* technology, VMware FT protects a VM by creating and maintaining an active Secondary VM in lockstep with the protected or Primary VM. The Secondary VM executes same sequence of instructions as the Primary VM. In case of failure of the Primary due to ESX host failure, the Secondary VM takes over without any loss of data or interruption of service. VMware FT also protects a workload when all HBAs (Fibre Channel HBAs only) providing access to shared storage fail.

It is worth noting that VMware FT does not protect a virtual machine against failures such as:

- Software failure inside the protected virtual machine.
- A component failure on the host. For example, failure of a network card used for virtual machine traffic will not cause the Secondary VM to take over from the Primary VM.
- Failure at the storage array level, etc.

Note that VMware FT has additional requirements for virtual machine configuration: a single vCPU for the VMs, no support for snapshots, only thick provisioned virtual disks, no automated DRS for FT-protected VMs, that all virtual disks have to be on shared iSCSI/Fibre Channel/NFS storage,etc. Refer to the VMware *vSphere Availability Guide* for a complete list of requirements for VMware FT and VMware HA.

NIC Teaming
The NIC Teaming feature allows the creation of a team or a bond of two or more physical network ports on a server. NIC teaming provides capabilities for both failover at network device layer as well as load balancing of the network traffic. While creating a network team, administrators should consider choosing network ports that reside on different PCI buses/slots/adapters. This provides additional protection against the failure of a PCI bus, slot, or adapter port.

iSCSI Multipathing
vSphere 4 includes enhanced software iSCSI initiator that provides multipathing capabilities at the iSCSI session layer. This enables both path redundancy and load balancing of iSCSI traffic. Refer to the *vSphere iSCSI SAN Configuration Guide* for more information on configuration and best practices for iSCSI storage.

VMware vCenter Server
vCenter server provides central management of the VMware ESX virtual infrastructure environment. vCenter also provides a control point for advanced features such as VMotion, Distributed Resource Scheduling (DRS), High Availability, and Fault Tolerance. vCenter availability is not considered in this paper. Once features such as HA and FT are configured, vCenter is not required to maintain availability of the virtual machines. If availability of vCenter is desired, the following approaches can be taken:

- Use VMware vCenter HeartBeat to protect vCenter server.
- Use Microsoft Cluster services to protect vCenter server application.
- Run vCenter server inside a virtual machine running on a VMware HA cluster.

Sample Three-Node Architecture

Figure 2 shows a sample three-node architecture designed for availability at different infrastructure layers. In the case of a failure or fault at a pre-defined hardware or software component, the virtual machines suffer minimal downtime. In the subsequent sections, we discuss architectural details and how this solution provides availability during typical failure scenarios.

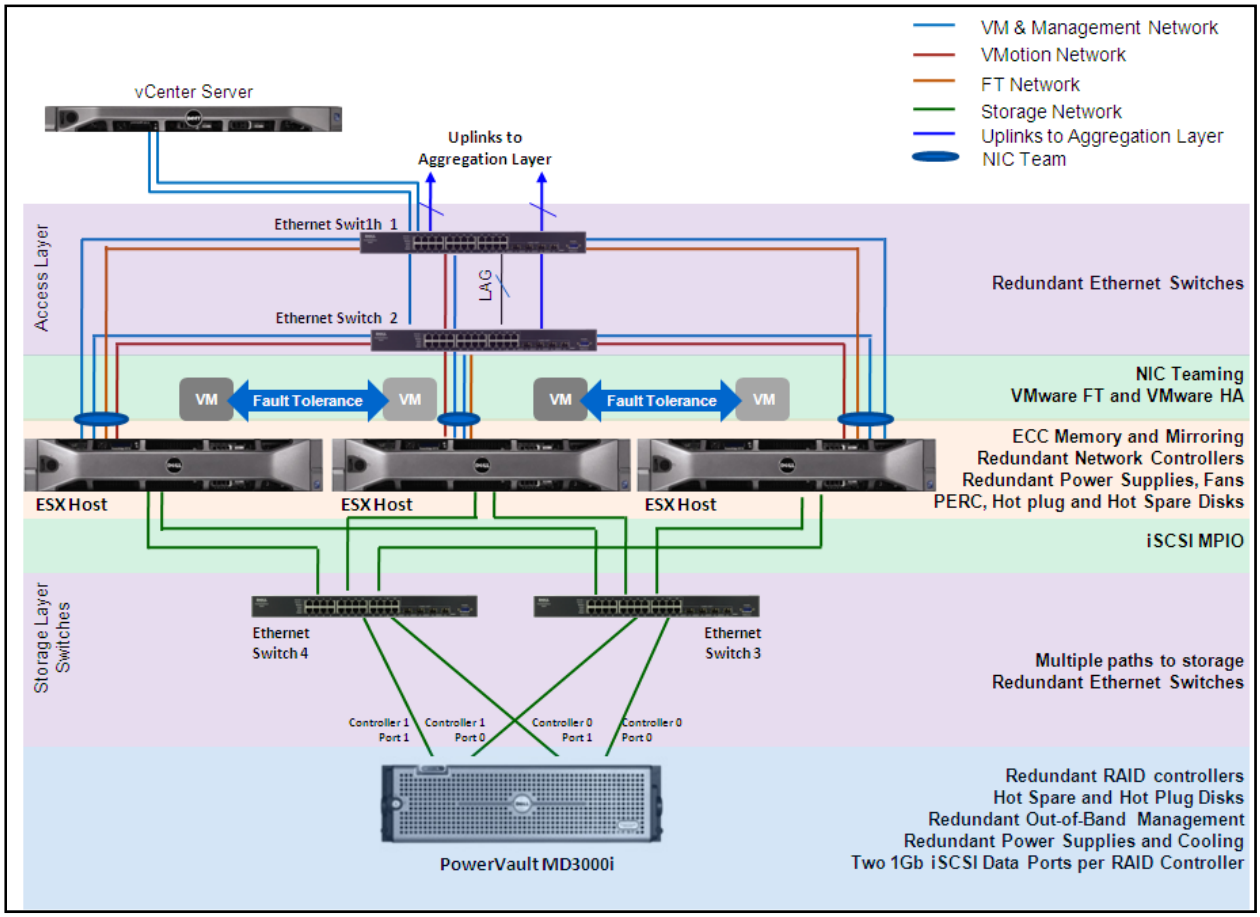

Figure 2: Sample three-node highly available architecture

As shown in Figure 2, three R710 servers run VMware ESXi Server 4.0 and are managed by vCenter Server 4.0 server running on PowerEdge R610. Each layer of the solution stack is configured to tolerate faults at different levels mentioned in the "Solution Components" section. As illustrated, the three PowerEdge R710 servers are connected to Dell PowerVault MD3000i iSCSI storage. Ethernet switches 3 and 4 enable multiple ESX hosts to connect to the MD3000i enclosure. Ethernet switches 1 and 2 form the Access Layer (Layer 2) for the ESX servers and provide Layer 2 adjacency for teamed Ethernet adapters on ESX hosts. vCenter Server has redundant connection to the access layer switches, and in turn to the ESX server hosts.

Configuration Guidelines

This section lists some general configuration guidelines pertaining to the hardware and software used for highly available architecture shown in Figure 2.

MD3000i Configuration

- The MD3000i array in HA configuration has two redundant dual-port RAID controllers.
- Controller 0, Port 0, and Controller 1 Port 0 are connected to Ethernet switch 3; Controller 0, Port 1, and Controller 1 Port 1 are connected to Ethernet switch 4.

For detailed information on configuring MD3000i with VMware ESX Server, refer to the *Dell PowerVault Configuration Guide for VMware ESX Server*.

Dell PowerEdge R710 Configuration

- Number of processors: Two Quad Core Intel® Xeon® 5500 series processors. Note that if using VMware FT to protect virtual machines, additional CPU resources are needed for the Secondary virtual machine. As a best practice, size your CPU requirements by taking into account both Primary and Secondary virtual machines. By default, CPU resources for the Secondary VM are not reserved on the ESX host running the Secondary VM. This can, however, be changed by setting explicit CPU reservation for the Primary VM. This reservation is also applied to the Secondary VM and hence ensures that sufficient resources are available on hosts running Primary and Secondary VMs.
- Memory: Use RDIMMs instead of UDIMMs, as RDIMMs have better RAS features compared to UDIMMs. Note that when using VMware FT to protect a VM, memory equal to the Primary VM's memory is reserved for the Secondary VM on the host on which the secondary VM resides. Since the memory requirement for each protected VM doubles, size your memory requirements accordingly.
    - To provide additional availability at the memory subsystem, memory mirroring can be used. However, only one half of the installed physical memory capacity is available to the ESX host. Hence, if using mirroring, size the physical memory to be double of what the ESX host and VMs would need.
- Internal RAID controller: Dell PowerEdge RAID Controller (PERC) 6/I
- Hard disks: At least three hard disks. Two hard disks that are configured as a RAID-1 volume for ESX and one hard disk that is configured as a hot spare for the RAID-1 volume.
- Network adapters: At least six Gigabit Ethernet network adapters.

vSphere 4 Configuration

- VMware vSphere 4 software components:

    —VMware vSphere software: vSphere 4.0 Patch 1 Advanced, Enterprise or Enterprise Plus edition.

    —VMware vCenter Server 4.0

Note: The ESX/ESXi 4.0 Patch 1 needs to be installed in order to deploy Dell-supported iSCSI configurations. This patch carries critical fixes related to iSCSI failover functionality.

Validation of Availability Features

As described in the previous section, a fault tolerant infrastructure should have multiple levels of redundancy to accommodate faults at different infrastructure tiers. To test the behavior of virtual machines in the architecture shown in Figure 2, an HA cluster was set up consisting of three R710 physical servers and one MD3000i storage enclosure. To test the impact of various failures on the application availability, we used the Dell DVDStore application inside Microsoft® Windows® Server 2003 virtual machines. The following sections describe details on software and hardware configuration for the setup shown in Figure 2.

Infrastructure Configuration

ESX Host Configuration

Three Dell PowerEdge R710 servers were used with identical physical configuration, with configuration details listed in Table 1. VMware ESXi 4.0 Installable was installed on local RAID 1 volume. A PowerEdge R610 server running Microsoft Windows Server 2003 Standard Edition was used to run vCenter.

ESX Host Configuration	
Server	Dell PowerEdge R710
Processors	2 X Intel Xeon Processor X5560 (2.8GHz, 8MB cache)
Memory	48GB (6 x 8 GB 1066 MHz RDIMMs)
Hypervisor	VMware ESXi Server 4.0 Patch 1 Installable
Network Adapters	4x1Gb on-board Broadcom NetXtreme II 5709c network adapters Intel PRO/1000 PT Dual Port Server Adapter
RAID Controller	Dell PERC 6/I
Hard Disks	3 x 73GB 15K SAS Drives

RAID Configuration	2 Drives in a RAID 1 volume for ESX installation 1 drive configured as hot spare for the RAID 1 volume
BIOS Version	1.1.4 or later
BIOS Option-Virtualization Technology	Enabled
BIOS Option-Power Management	OS Control

Table 1: ESX Host Configuration

Note: BIOS option *'Virtualization Technology'* needs to be set to *'Enabled'* for VMware FT to work.
Note: As a best practice for running VMware FT, both the Primary and Secondary VMs should run on processors with the same speed. Hence, the "Power Management" BIOS option should be set to "OS Control" to let ESX manage the processor power states.

ESX Host Network Configuration

The ESX network was configured with the following objectives in mind:

Availability: Provide redundancy for all different kinds of network traffic: ESX management, virtual machines, iSCSI storage, VMotion, and fault tolerance logging.
Traffic Isolation: Isolate different kinds of network traffic using physical or virtual segregation. In this scenario, we physically isolate the iSCSI storage traffic from the rest. The rest of the traffic is isolated using VLANs.
Performance: To achieve optimal performance, we use iSCSI multipathing capabilities of the software iSCSI initiator and network layer load balancing for the virtual machine traffic.

Figure 3 illustrates what the network configuration looks like for any ESX host in the cluster. The location of network ports on the PCI buses and their respective network controllers was considered while choosing the adapters to team. For each PowerEdge 710 used in this setup, the physical network adapters are enumerated as follows:

vmnic0, *vmnic1*: First and second LOM ports (marked as Gb1 and Gb2 on chassis) respectively on first dual-port Broadcom NetXtreme II 5709 Gigabit Ethernet controller integrated on the system motherboard.
vmnic2, *vmnci3*: First and second LOM ports (marked as Gb3 and Gb4 on chassis) respectively on second dual-port Broadcom NetXtreme II 5709 Gigabit Ethernet controller integrated on the system motherboard.
vmnic4, *vmic5:* First and second ports on the add-in Intel PRO/1000PT dual-port server adapter.

Below are configuration details about various kinds of network traffic:

All Traffic Other Than iSCSI Storage

As shown in Figure 3, four network adapters namely *vmnic0*, *vmnic2*, *vmnic3*, and *vmnic5* are connected as uplinks to *vSwitch0*. Although all four ports are teamed at the vSwich level, we modified teaming configuration at each port group level. Out of four NICs, *vmnic0* and *vmnic2* are primarily used for the management and virtual machine traffic, *vmnic3* is primarily used for VMotion and *vmnic5* for FT logging traffic. This provides both dedicated bandwidth and failover capabilities for all kinds of network traffic.

Management Network (Port Group Management Network)

NIC Teaming: *vmnic0* as active. *vmnic2*, *vmnic3* and *vmnic5* (in the order listed) as stand-by adapters.
Load Balancing: Originating virtual port ID.
VLAN Configuration: *Management Network* port group is configured to tag ESX host management packets with VLAN ID 162. Note that the VLAN IDs used may be different as per your setup.
Failback: No. Disabling failback avoids any loss of management traffic during the time that it takes for the port to be ready for serving traffic. This avoids the triggering of any false host isolation due to a temporary loss of management network.

Virtual Machine Network (Port Group VM Network)

NIC Teaming: *vmnic0* and *vmnic2* as active/active. *vmnic3* and *vmnic5* (in the order listed) as stand-by adapters.
Load Balancing: Originating virtual port ID.
VLAN Configuration: *VM Network* port group is configured to tag virtual machine network packets with VLAN ID 164. Note that the VLAN IDs used may be different, depending on your setup.

Failback: Yes

This setup enables both high availability and load balancing for a virtual machine network. Note that enabling failback may cause temporary loss of network access to virtual machines which fail back after an adapter recovers from failure. If this is not a tolerable situation, disable failback.

VMotion Network (Port Group VMkernel-VMotion)

NIC Teaming: *vmnic3* as active; *vmnic0*, *vmnic2* and *vmnic5* (in the order listed) as stand-by. In order to keep VMotion and FT traffic separate in the case of a failure of VMotion NIC, *vmnic0* and *vmnic2* are kept first in the standby list.

Load Balancing: Default virtual switch policy.

VLAN Configuration: *VMkernel-VMotion* port group is configured to tag VMotion packets with VLAN ID 163. Note that the VLAN IDs used may be different, depending on your setup.

Failback: Yes

This setup provides full Gigabit bandwidth and high availability for VMotion traffic.

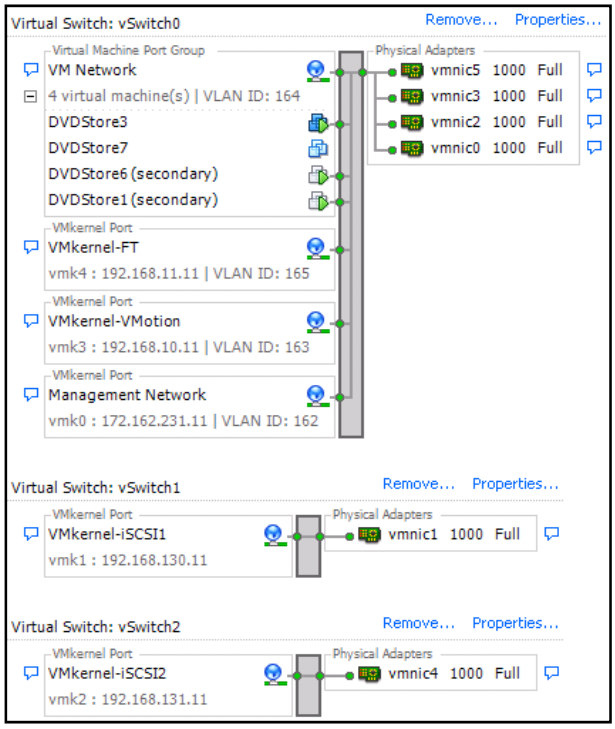

Figure 3: ESX Host Network Configuration

FT logging Network (Port Group VMkernel-FT)

NIC Teaming: *vmnic5* as active; *vmnic2*, *vmnic0* and *vmnic3* (in the order listed) as stand-by. In order to keep FT and VMotion traffic separate in the case of a failure of FT NIC, *vmnic2* and *vmnic0* are kept first in the standby list.

Load Balancing: Default virtual switch policy.

VLAN Configuration: *VMkernel-FT* port group is configured to tag FT logging packets with VLAN ID 165. Note that the VLAN IDs used may be different, depending on your setup.

Failback: Yes

This setup provides full Gigabit bandwidth and high availability for FT logging traffic.

Note that enabling failback may cause a temporary loss of connectivity of the VMotion and FT logging NIC after a network adapter recovers from failure. This may cause the VMs protected by FT to be temporarily in a not-protected state. Once the adapter (to which VMotion/FT logging network fails back) is fully available to serve traffic and FT/VMotion links are available, the VMs will automatically be protected. Since both VMotion and FT logging have high bandwidth requirements, failback was enabled so that traffic can utilize the dedicated NIC after the faulty adapter recovers from a failure.

Notes:

- To reduce time taken for ports to come online, configure the speed/duplex settings for network adapter *vmnic0, vmnic2, vmnic3* and *vmnic5* to be 1000Mb/Full Duplex.
- The sample configuration described in this paper uses a Gigabit network adapter for FT logging traffic. If high utilization of FT logging link is observed, it is recommended to use a 10Gb Ethernet adapter.
- In the case of a failure of the VMotion or FT link, one of *vmnic0* or *vmnic2* is used to maintain respective traffic. Similarly if Ethernet switch 1 or 2 fails, only two network adapters are available for all types of network traffic other than iSCSI storage traffic. Since both VMotion and FT logging links have high bandwidth requirements, depending on the bandwidth requirement for virtual machine traffic, it may not be desirable to share the physical network interfaces after a failure. For such scenarios, two additional dedicated network adapters (a total of four dedicated network adapters for VMotion and FT logging) may be used to provide failover support for VMotion and FT logging links. Failback may be disabled for this configuration.

iSCSI Storage Network Configuration

Multi-Pathing: To provide path redundancy and traffic load balancing for the iSCSI storage traffic, two network adapters (*vmnic1* and *vmnic4*) are used. As shown in Figure 3, *vmnic1* is uplinked to *vSwitch1* and *vmnic4* is uplinked to *vSwitch2*. Using the *esxcli* command, VMkernel interfaces *vmk1* (corresponding to *vmnic1*) and *vmk2* (corresponding to *vmnic4*) are attached to the software iSCSI initiator. Refer to the vSphere 4 *iSCSI SAN Configuration Guide* for detailed information about configuring software iSCSI initiator for multi-pathing.

VMkernel-iSCSI1 interface is configured to be in the same IP subnet as MD3000i Controller 0 Port 0 and Controller 1 Port 0. *VMkernel-iSCSI2* interface is configured to be in the same IP subnet as MD3000i Controller 0 Port 1 and Controller 1 Port 1. This enables two active paths (one through *VMkernel-iSCSI1* and other through *VMkernel-iSCSI2* interface) to a LUN owned by any MD3000i controller.

Load Balancing: The configuration described above provides two active paths for any LUN on any MD3000i controller. However, to load balance traffic through both available active paths, for each LUN exposed to the ESX hosts, the path selection policy is set to *'Round Robin'*. The other path selection policies namely *Most Recently Used (MRU)* or *Fixed* offer only failover and no load balancing capability. *Round Robin* provides both path redundancy and load balancing for both transmit and receive traffic.

VLANs: No VLANs are used as traffic is physically isolated using dedicated switches.

Jumbo Frames: These can be optionally enabled for improved performance. Jumbo frames must be enabled for both the vSwitch and the VMkernel port. Refer to the vSphere *iSCSI Storage Configuration* guide for details about how to configure VMkernel interfaces to use jumbo frames. Note that if using jumbo frames, jumbo frames need to be enabled end to end on the ESX host virtual switch, VMkernel interfaces, Ethernet switches and MD3000i data ports.

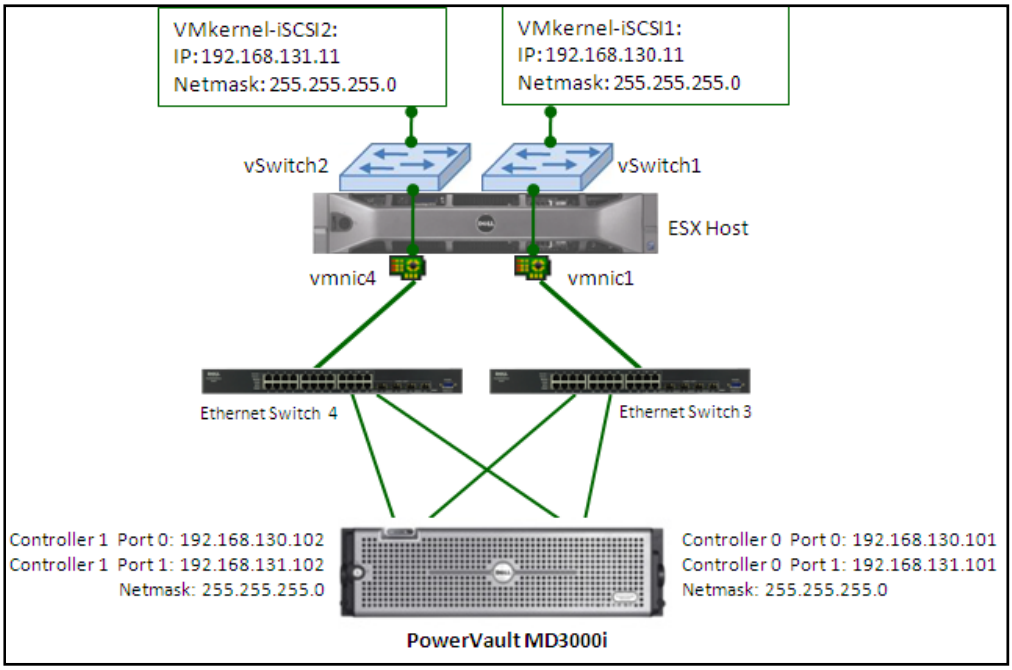

Figure 4: ESX-MD3000i Storage Network Configuration

ESX Cluster Configuration

The ESX cluster was enabled for both VMware HA and VMware DRS. The cluster HA setting was configured to tolerate a single node failure. DRS policy was set to fully automated; however, for VMs protected by VMware FT, the DRS settings are automatically set to manual.

MD3000i Configuration

MD3000i Network Configuration

MD3000i array comprises of two active RAID controllers. However, at any given time, only one RAID controller owns a virtual disk or a LUN. Each RAID controller has two Gigabit Ethernet ports for iSCSI data traffic. By configuring Controller 0 Port 0 and Controller 1 Port 0 in one IP subnet and Controller 0 Port 1 and Controller 1 Port 1 in another IP subnet, both traffic isolation across physical Ethernet segments (using redundant Ethernet switches) and path redundancy are achieved.

For the purpose of this paper, the MD3000i iSCSI data ports were configured with the following IP configuration as shown in Figure 4:

Controller 0, Port 0: 192.168.130.101/255.255.255.0

Controller 0, Port 1: 192.168.131.101/255.255.255.0

Controller 1, Port 0: 192.168.130.102/255.255.255.0

Controller 1, Port 1: 192.168.131.102/255.255.255.0

The two controllers are connected to redundant PowerConnect 5424 switches.

MD3000i Storage Configuration

For the purpose of validating the configuration shown in Figure 2, a single MD3000i array with two controllers and 15 146GB 15K SAS drives was used. 14 drives were used to configure the disk groups with different RAID levels, and one disk was used as a global hot spare. Each disk group is protected by its individual RAID configuration, at the same time the global hot spare provides protection against a degraded RAID set due to a physical disk failure. Separate LUNs were created to store the database, log and OS virtual disks (for DVDStore virtual machines). Two identical virtual disks were created on each disk group and assigned to the two controllers in order to balance load across two RAID controllers. Each virtual disk was mapped to each host in the cluster.

Table 2 describes the MD3000i storage configuration.

Physical Disks	Disk Group	RAID Level	Virtual Disk/Owning Controller
0,1,2,3,4,5,6,7	DB	RAID 10	DB1 (Controller 0) and DB2 (Controller 1)
8,9,10, 11	Log	RAID 10	Log1 (Controller 0) and Log2 (Controller 1)
12, 13	OS	RAID 1	OS1 (Controller 0) and OS2 (Controller 1)
14 (Hot Spare)	--	--	--

Table 2: MD3000i Storage Configuration

External Network layer Configuration

Ethernet Switches 1 and 2: As shown in Figure 5, Ethernet switches 1 and 2 handle the following types of network traffic groups: virtual machines, management, VMotion and FT logging. This networking configuration along with the vSwitch configuration on each ESX host described earlier achieves following key objectives:

- Segregating VMotion and FT logging traffic on separate network adapters and switches ensures that there is no overloading of network links within the Gigabit network bandwidth limit.
- Using two Ethernet switches and teaming four network adapters on the ESX hosts provides redundancy against both network adapter and switch failures.

Since each of the traffic groups is in its own VLAN, the network ports connected to *vmnic0* and *vmnic3* on Ethernet switch 1 and *vmnic2* and *vmnic5* on Ethernet switch 2 are configured as trunk ports. Ethernet switches 1 and 2 are connected to each other using a 4- port LAG also configured as a trunk. A 4-port LAG provides both availability and enough bandwidth for setup described in Figure 2. A layer 2 connection (the 4-port trunked LAG in this case) is required between the two switches in order to provide path adjacency among the team members. If Ethernet switches 1 and 2 are connected to Aggregation Layer (layer 2/3) switches, care must be taken to configure layer 2/3 aggregation layer switches so as to avoid layer 2 loops. If the spanning tree protocol (STP) is used, then the STP should be adjusted so that there is always a physical layer 2 path (either through access or aggregation layer switches) between the NIC team members. It is recommended to have Ethernet switches 1 and 2 connected using a LAG so as to keep VMotion, FT logging and ESX host heart beat traffic (over management network) local to access layer switches.

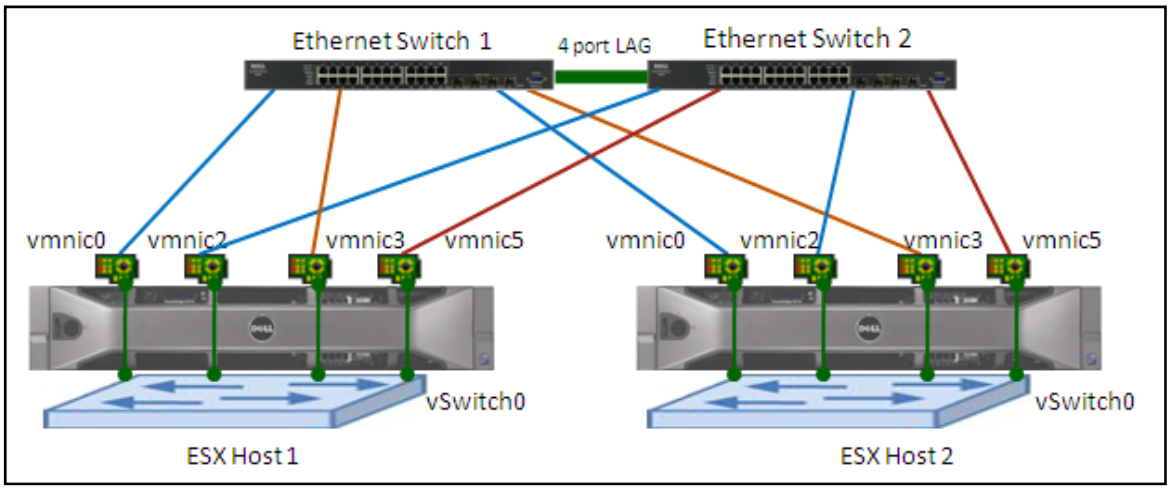

Figure 5: Detailed External Network Configuration

To understand why a physical Layer 2 path is required among teamed network adapters, consider the Management network (using *vmnic0*) configuration in Figure 4. Management network is configured to use *vmnic0* as the primary and *vmnic2* as the first stand-by. On failure of *vmnic0* on say ESX Host 2, *vmnic2* becomes the primary management NIC on Host 2. On ESX Host 1, *vmnic0* is still the primary interface for management traffic. It is a property of ESX virtual switch on ESX Host 1 to block any traffic for *vmnic0* on *vmnic2*. Hence, the only way the management interfaces between ESX Host 1 and Host 2 can connect is through the physical layer 2 path provided by the trunked LAG. Layer 2 path is also required for virtual machines using different network adapters (for example virtual machines using *vmnic0* to communicate with virtual machines using *vmnic2*) on the team to communicate with each other.

Additional configuration notes for Ethernet switches 1 and 2:

- In an HA cluster, it is important to avoid any false host isolation events due to the temporary loss of management traffic. Hence as a best practice, configure the edge ports (ports connecting the ESX servers) on the Ethernet switches 1 and 2 to come online as fast as possible and avoid temporary blocking of management traffic by configuring following edge port settings:
  - Enable Fast Link (or Port Fast). This enables the ports connected to ESX hosts to go directly in forwarding state rather than waiting in blocking state before STP converges.
  - Disable auto negotiation. Make sure the speed/duplex is set to 1000Mb Full Duplex.
- Use rapid spanning tree protocol (RSTP) instead of spanning tree protocol (STP).
- It is important to configure connections of Ethernet switches 1 and 2 to any uplink switches/routers in such a manner that in case Ethernet switch 1 or 2 fails, the ESX hosts still can reach the management network gateway.

Ethernet Switches 3 and 4 are used for iSCSI storage traffic. No special configuration is required on these switches. Enabling jumbo frames on the ports connecting to ESX hosts is optional for improved performance. Note that if using jumbo frames, jumbo frames need to be enabled end to end on ESX host virtual switch, VMkernel, Ethernet switches, and MD3000i data ports.

Power Supply Considerations

To ensure that the loss of a power bus does not impact the whole infrastructure, it is always recommended to connect the server/storage/switch power supplies to separate power buses. Specifically for this configuration, connect:

- The ESX/vCenter host and MD3000i power supplies from two different power buses.
- Ethernet switch 1 and 3 from one power bus and Ethernet switch 2 and 4 from a different power bus.

Alert Notification

Each layer of the infrastructure has fault alerting capabilities:

1. Dell OpenManage Server Administrator (OMSA) agent running on ESX hosts captures all hardware faults and has an ability to send alerts as Email or SNMP messages. Refer to the *Dell OpenManage Server Administrator User's Guide* for more information on how to configure and manage server alerts.

2. Alerts from MD3000i array are reported in the MD Storage Manager application and can be sent as e-mail or SNMP messages. Refer to the *Dell PowerVault Modular Disk Storage Manager User's Guide* for more information on how to configure and manage MD3000i alerts.
3. Dell PowerConnect switches can be configured to send alerts as SNMP messages. Refer to the *Dell PowerConnect User Guide* for more information on how to configure and manage alerts for PowerConnect switches.
4. vCenter and ESX server software can be configured to send SNMP or e-mail messages for alarms generated at the ESX or vCenter layer. Refer to the *vSphere Basic System Administration Guide* for more information on how to configure and manage SNMP alerts for ESX and vCenter.

Workload Configuration

Application Workload

DVD Store is an online e-commerce test application developed and maintained by Dell. It simulates a real-world database operation and contains scripts for database creation and indexing and driver programs for client access. The driver program simulates a real user access of an online DVD store application by logging on, browsing, and purchasing DVDs. The load can be varied by changing the number of active threads where each thread simulates the complete cycle of one user access to the database. The primary performance metric is orders per minute which is the cumulative average of the total number of orders executed and is reported every 10 seconds during a typical run of the benchmark.

The default sizes of the small, medium and large databases in a DVD Store setup are 10MB, 1GB and 100GB respectively. To approximate a typical use case, a 10GB custom database was created using the dataset generation scripts available with the application. The client driver program was run on a separate physical machine and connected to each of the VMs remotely. To simulate an appropriate load, five threads were run simultaneously on each VM. To accommodate storage failover scenarios, the timeout for each database operation was set to 200s. Upon failure of connectivity to the VMs, the driver program on the remote host reports errors after this timeout. The connectivity of the driver program was used as a benchmark to monitor application availability under various failure scenarios.
For more information on Dell DVDStore application, refer to http://www.delltechcenter.com/page/DVD+Store

Virtual Machine Configuration

An initial VM was created with 1 vCPU and 1 GB of RAM. Microsoft Windows Server 2003 Enterprise edition was installed as the guest operating system, and Microsoft SQL Server 2005 was installed as the database application. Three virtual disks were attached to this VM—two to store SQL Server database files and one to store SQL Server log files. A custom DVDStore database of size 10GB was created using the scripts provided with the DVDStore application. Table 3 summarizes the virtual machine configuration:

Operating System	Windows Server 2003 R2 Enterprise Edition
Number of virtual CPUs	1
Memory	1GB
Virtual Disks	OS: 8GB | Database1: 15GB | Database2:15GB | Log:15GB
Database Application	Microsoft SQL Server 2005
Workload	Dell DVDStore
DVD Store Database Size	10GB
SCSI Time Out[2]	200 seconds

Table 3: DVDStore Virtual Machine Configuration

The initial VM was then cloned to create 12 VMs and added to the cluster of R710 servers. DRS was used to provision the VMs; however, care was taken to load-balance the VMs across all physical servers. Following VMware's best practice to run up to a total of four Primary or Secondary VMs per host, VMware FT was enabled for six virtual machines in the three-node cluster.

[2] To accommodate the MD3000i controller failover scenarios, the SCSI timeout in Windows VMs was set to 200s.

Failover Scenarios and Test Results

This section describes some failure scenarios that we simulated in the lab and the effect of failures on the DVDStore application. To test availability of the DVDStore application, we used the SQL Server driver program included with the DVDStore application. The driver was run for each DVDStore VM from an external client station. No application-level availability software/mechanism was configured for DVDStore. Client connectivity to the DVDStore application was observed during failures simulated as described in Table 4. As shown in Table 4, the configuration was able to withstand different kinds of failures and maintained availability of the DVDStore application.

Failure Area	Failure Description	Protection Provider for VMs	Impact to application availability	Behavior Following Failure
MD3000i Array	Hard disk failure simulated by pulling out a hard disk.	RAID set and Hot Spare	None	The Hot Spare disk takes over the failed disk and starts rebuilding the RAID set.
	RAID controller failure simulated by pulling out an active controller.	Dual RAID controllers	None	The LUNs owned by affected controller are failover to the other controller. SCSI commands to impacted LUNs will be delayed during the failover. ESX hosts still have two active paths for I/O to all LUNs.
	Single port failure on a RAID controller simulated by pulling out the network cable of an active port.	Dual ports per controller in combination with ESX iSCSI multipathing	None	Out of two, only one path is available for I/O. The path using the failed port is marked as dead and paths through the other controller remain on standby.
	Dual port failure on a RAID controller simulated by pulling out both network cables on an active controller.	Dual controllers in combination with ESX iSCSI multipathing	None	The LUNs owned by affected controller failover to the other controller. SCSI commands to impacted LUNs will be delayed during the failover. ESX hosts still have two active paths for I/O to all volumes.
	Power supply failure simulated by pulling out the power supply from an active RAID controller.	Dual Power Supplies	None	The second power supply takes over for the failed power supply.
	Management port failure simulated by pulling out the network cable from the management port on a controller.	Dual management ports per array	None	MD3000i out-of-band management still available through the management port on the other controller.
Network Fabric	Storage switch failure (Ethernet switch 3 or 4) simulated by pulling out the power supply to a switch.	Redundant Ethernet switches	None	For each volume, only one path is available for I/O and one remains on standby. The paths using the failed switch are marked as dead.
	Access layer Ethernet switch 1 failure[3] simulated by pulling out the power cable to the switch.	Redundant Ethernet switches	None[3]	*vmnic0* and *vmnic3* links go down. Management traffic fails over to *vmnic2* and virtual machine traffic fails over to *vmnic2* and *vmnic5*.

[3] If failback is enabled for virtual machine port group, virtual machines that fail back to the adapter that recovers from failure (either due to an adapter or switch failure) may observe temporary loss of access to their network which fail back after an adapter recovers from failure.

	Access layer Ethernet switch 2 failure[4] simulated by pulling out the power cable to the switch.	Redundant Ethernet switches	None[4]	*vmnic2* and *vmnic5* links go down. VMotion traffic fails over to *vmnic0* and FT logging traffic fails over to *vmnic3*.
ESX Host	Local hard disk failure simulated by pulling out a hard disk.	RAID set provided by PERC 6/I	None	Hot spare takes over for the failed disk and RAID set starts rebuilding.
	Single iSCSI host port failure (*vmnic1* or *vmnic4*) simulated by pulling out the network cable for the iSCSI network adapter.	iSCSI multipathing	None	Out of two, only one path is available for I/O. The path using the failed port is marked as dead and paths through the other RAID controller remain on standby.
	VMotion link (*vmnic3*) failure[4] simulated by pulling out the network cable for *vmnic3*.	NIC Teaming	None[4]	VMotion traffic fails over to *vmnic0*.
	FT logging NIC (*vmnic5*) failure[4] simulated by pulling out the network cable for *vmnic5*.	NIC Teaming	None[4]	FT logging traffic fails over to *vmnic2*. FT protected VMs remain protected.
	Management link (*vmnic0*) failure simulated by pulling out network cable for *vmnic0*.	NIC Teaming	None	Management traffic fails over to *vmnic2*.
	Virtual machine link (*vmnic0* or *vmnic2*) failure[5] simulated by pulling out network cable for *vmnic0* or *vmnic2*.	NIC Teaming	None[5]	Virtual machine traffic fails over to the other adapter and no loss of network connectivity was observed during failover.
	DIMM failure simulated by pulling out a DIMM from a running system.	Memory Mirroring[6]	None	ESX and virtual machines continue to run without any impact.
	Single bit memory error simulated using a fault injecting DIMM.	ECC Memory	None	ESX and virtual machines continue to run without any impact.
	Multi-bit memory error simulated using a fault injecting DIMM.	Memory Mirroring[6]/VMware FT	None if using Memory Mirroring. If not, no impact for VMs protected by VMware FT.	If using memory mirroring, ESX and virtual machines continue to run without any impact. If no mirroring is used, ESX server software crashes. For VMs protected by FT, the Secondary VMs take over the Primary VMs, thus maintain access to

[4] If failback is enabled for VMotion/FT logging ports groups, a temporary loss of connectivity of the VMotion and FT logging NIC may occur after a network adapter/switch recovers from failure. This may cause the VMs protected by FT to be temporarily in 'not protected' state. Once the adapter (to which VMotion/FT logging network fails back) is fully available to serve traffic and FT/VMotion links are available, the VMs will automatically be protected.

[5] If failback is enabled for virtual machine port group, virtual machines that fail back to the adapter that recovers from failure (either due to an adapter or switch failure) may observe temporary loss of access to their network which fail back after an adapter recovers from failure.

[6] Requires the *Memory Operating Mode* server BIOS option to be set to *Mirror Mode*.

			For VMs protected by HA, application is not available until VMs restart on other hosts.	DVDStore application. VMware FT spawns Secondary VMs for the new Primary VMs. VMs protected by HA are restarted on remaining two hosts in the cluster. DVDStore application is available after the VMs are back online.
	ESX host failure simulated by powering off the host.	VMware FT/VMware HA	None for VMs protected by FT. For VMs protected by HA, application is not available until VMs restart on other hosts.	For VMs protected by FT, the Secondary VMs take over the Primary VMs running on failed host. VMware FT spawns Secondary VMs for the new Primary VMs. VMs protected by HA are restarted on remaining two hosts in the cluster. DVDStore application is available after the VMs are back online.
	Power supply failure simulated by pulling out a power supply.	Dual power supplies	None	The second power supply maintains the supply of power to the server.
vCenter Server	vCenter failure simulated by powering off the vCenter server.	ESX host cluster	None	FT and HA protected VMs keep on running. Management of cluster including any change in HA/FT configuration is not available until vCenter Server becomes available.
	vCenter failure simulated by powering off the vCenter server and ESX host failure simulated by powering off a host.	VMware FT/VMware HA	None for VMs protected by FT. For VMs protected by HA, application is not available until VMs restart on other hosts.	For VMs protected by FT, the Secondary VMs take over the Primary VMs running on failed host. VMware FT spawns Secondary VMs for the new Primary VMs. VMs protected by HA are restarted on remaining two hosts in the cluster. DVDStore application is available after the VMs are back online.

Table 4: Failover Tests and Observations

Conclusion

Planning for availability is critical for virtualized infrastructure due to the fact that a small failure could lead to downtime for many applications. This paper presented an approach toward building a highly available architecture for virtualized infrastructure. It is important to consider availability at every layer of the whole infrastructure, and using sample architecture, we highlighted resilience of the architecture against different kind of faults.

References/Additional Links

1. Dell PowerEdge R710 Documentation: http://support.dell.com/support/edocs/systems/per710
2. Dell PowerEdge R610 Documentation: http://support.dell.com/support/edocs/systems/per610
3. Dell PowerVault MD3000i Documentation: http://support.dell.com/support/edocs/systems/md3000i/en/2ndGen/index.htm
4. Dell PowerVault MD3000i configuration with vSphere 4.0: http://www.delltechcenter.com/page/VMware+ESX+4.0+and+PowerVault+MD3000i
5. Dell PowerConnect Documentation: http://support.dell.com/support/edocs/network/54xx/en/index.htm
6. Dell OpenManage Server Administrator Storage Management User's Guide: http://support.dell.com/support/edocs/software/svradmin/6.0.3/omss/index.htm
7. Dell vSphere 4.0 Documentation: http://support.dell.com/support/edocs/software/eslvmwre/VS/VS.htm
8. VMware vSphere 4.0 iSCSI SAN Configuration Guide: http://vmware.com/pdf/vsphere4/r40/vsp_40_iscsi_san_cfg.pdf
9. VMware vSphere 4.0 Availability Guide: http://vmware.com/pdf/vsphere4/r40/vsp_40_availability.pdf
10. VMware vSphere 4.0 Basic System Administration Guide: http://vmware.com/pdf/vsphere4/r40/vsp_40_admin_guide.pdf
11. Dell Virtualization Home Page: http://www.dell.com/virtualization
12. Dell Business Ready Configurations for Virtualization: http://content.dell.com/us/en/business/virtualization-business-ready-configurations.aspx
13. Dell Enterprise Technology Center: http://www.delltechcenter.com