

Reference Architecture for Implementing VMware Fault Tolerance on the Dell PowerEdge R910 with EqualLogic iSCSI Storage

April 2010

M Shabana
Dell Virtualization Solutions Engineering



THIS WHITE PAPER IS FOR INFORMATIONAL PURPOSES ONLY, AND MAY CONTAIN TYPOGRAPHICAL ERRORS AND TECHNICAL INACCURACIES. THE CONTENT IS PROVIDED AS IS, WITHOUT EXPRESS OR IMPLIED WARRANTIES OF ANY KIND.

© 2010 Dell Inc. All rights reserved. Reproduction of this material in any manner whatsoever without the express written permission of Dell Inc. is strictly forbidden. For more information, contact Dell.

Dell, the *DELL* logo, *EqualLogic*, *PowerEdge* and *PowerConnect* are trademarks of Dell Inc.; *Intel* and *Xeon* are registered trademarks of Intel Corporation; *Microsoft* and *Windows* are registered trademark of Microsoft Corporation; *VMware*, *VMware vSphere*, *vCenter*, and *VMotion* are registered trademarks or trademarks (the "Marks") of VMware, Inc. in the United States and/or other jurisdictions.

Table of Contents

1	Introduction	2
2	VMware Fault Tolerance Overview	2
3	PowerEdge R910 Overview.....	3
4	Requirements for Implementing Fault Tolerance	3
5	VMware Fault Tolerance Limitations.....	4
6	Reference Architecture	5
7	Hardware Configuration.....	6
8	Network Configuration	7
9	iSCSI Configuration.....	10
10	References	11

1 Introduction

Providing high availability to mission critical applications has always been a challenge for IT administrators. As server virtualization becomes more popular, IT organizations are moving from deploying applications in dedicated physical servers to virtual machines (VM). With the move towards virtual infrastructure environments, demand for providing high availability to mission critical applications has increased dramatically.

The VMware Fault Tolerance (FT) feature helps to provide continuous protection to virtual machines running mission critical applications from host failure. When VMware FT is enabled on a virtual machine, a copy of the Primary VM is automatically created on another host and the two VMs are synchronized using VMware's vLockstep technology. When a failure occurs in the host with the Primary VM, the Secondary VM is designated as the Primary and continues to run with little or no downtime; data, transactions, and network connections are maintained.

The Dell PowerEdge R910 supports up to four 8-core processors and 64 DIMMs of memory (up to 1 TB), allowing for increased consolidation ratios. As more virtual machines are consolidated, high availability becomes even more critical. VMware Fault Tolerance combined with the optional failsafe embedded hypervisor in the R910 provides increased availability for mission critical applications running as virtual machines. Throughout this document, references to ESX host refer to both ESX and ESXi host servers unless explicitly mentioned.

2 VMware Fault Tolerance Overview

This section provides an overview of VMware Fault Tolerance technology. For more details on how the fault tolerance features work in vSphere refer to the following white paper from VMware:

<http://www.vmware.com/resources/techresources/1094>

VMware FT allows virtual machines to run continuously, even when an ESX host server fails. When VMware FT is enabled on a virtual machine (called the Primary VM), a copy of the Primary VM is created on another host (chosen by the Distributed Resource Scheduler) and designated as the Secondary VM. If DRS is not enabled, the host for the Secondary VM is chosen manually from the list of available hosts. VMware FT runs the Primary and Secondary VMs in lockstep with each other - essentially mirroring the execution state of the Primary VM to the Secondary VM. In the event of a hardware failure that causes the Primary VM to fail, the Secondary VM immediately picks up where the Primary VM left off, and continues to run with little or no loss of network connections, transactions, or data.

VMware vLockstep Technology

VMware vLockstep technology ensures that the Primary and Secondary VMs execute the same set of instructions in an identical sequence with a small time delay. The Primary VM captures all non-deterministic events and copies them to the Secondary VM using the VMware FT logging network (a dedicated VMKernel Network). A few examples of non-deterministic events include user inputs, network packets sent to the Primary VMs, and disk I/O. The Secondary VM replays those non-deterministic events in the same sequence as in the Primary VM. The Secondary VM thus executes the

same series of instructions as the Primary VM. Only the output of the Primary virtual machine will take effect while the Secondary VM output will be suppressed.

3 PowerEdge R910 Overview

The Dell PowerEdge R910 server offers up to four Intel Xeon 8-core processors (with Nehalem-EX) and has the following key features:

Increased Memory Capacity: The R910 features integrated QuickPath memory controllers with DDR3 RDIMM modules enabling support for up to 1 TB in a four-socket configuration.

Optional Integrated 10 Gigabit Network Adapter: The R910 consists of either two dual-port embedded Broadcom 5709C Gigabit NIC controllers on a 1GbE riser or one Broadcom dual port 57710 10GbE controller and one 5709C 1GbE NIC controller.

"Failsafe" Virtualization: The PowerEdge R910 server comes with an Internal Dual SD Module (IDSDM) which consists of up to two SD flash cards containing bootable OS images for hypervisors. These cards are mirrored when set in the redundant mode for higher availability.

Embedded Management with Lifecycle Controller: The Lifecycle Controller is the engine for advanced embedded management and is delivered as part of iDRAC express or iDRAC enterprise in the new 11th generation Dell PowerEdge Servers. The Lifecycle Controller helps to simplify administrator tasks by performing a complete set of provisioning functions such as system deployment, system updates, hardware configuration and diagnostics from a single intuitive interface called the Unified Server Configurator (USC) in a pre-OS environment.

4 Requirements for Implementing Fault Tolerance

The following requirements must be fulfilled to implement VMware FT. For a detailed description of these requirements, please refer to the following link:

http://www.vmware.com/pdf/vsphere4/r40/vsp_40_availability.pdf

ESX Host Requirements

- The ESX servers must use processors which are FT capable and identical. Refer to the hardware compatibility list to check on a processor's FT status:
<http://www.vmware.com/resources/compatibility/search.php>
- The ESX hosts must have vSphere 4 or a later version for implementing VMware FT. The hosts must have the same ESX version and patch version. The VMware FT feature is available in the Advanced, Enterprise & Enterprise Plus versions of vSphere.
- The ESX hosts need to be configured in VMware High Availability (HA) mode. To ensure redundancy and maximum Fault Tolerance protection, VMware recommends that you have a minimum of three hosts in the cluster. In a failover situation, a new secondary VM will be created on the third host.

Storage Requirements

- Implementing the FT feature requires the use of shared storage between servers.
- Turning on VMware FT for a virtual machine first requires the virtual machines' virtual disk (VMDK) files to be eager-zeroed (all blocks are zeroed out) and thick-provisioned. Space allocation is done at the time of the creation of the disk. Unlike the thick and thin-provisioned disk types, an eager-zeroed disk must format the data blocks at the time of deployment. VMs with thin-provisioned and last-zeroed disks must be powered off to enable FT through vCenter.
- Backup solutions are supported in the Guest operating system for file or disk-level backups. These applications may lead to the saturation of the VMware FT logging network if heavy read access is performed. Ideally, FT enabled VMs should run on a different host.

Network Requirements

- The FT logging traffic can be high if the fault tolerant VM has excessive non-deterministic events. To calculate the amount of network bandwidth required for FT logging, the following formula can be used:

VMware FT logging bandwidth ~ =

$(\text{Avg disk reads (MB/s)} \times 8 + \text{Avg network input (Mbps)}) \times 1.2$ [20% headroom]

- Ensure that the networking latency between ESX hosts is low. Sub-millisecond latency is recommended for the FT logging network. Use *vmkping* to measure the latency.
- VMware vSwitch settings on the hosts should be uniform, such as using the same VLAN for VMware FT logging, to make these hosts available for placement of the Secondary VMs. Consider using a VMware Distributed Switch (available with Enterprise Plus Licenses) to avoid inconsistencies in the vSwitch settings.
- It is recommended to use a separate NIC for FT logging and VMotion™ to isolate network traffic. This enables VMotion traffic to continue while the Secondary VM is created. The FT logging traffic and other network traffic can be configured in a single virtual switch in an active/standby configuration. In the case of a failure of an active NIC for FT logging traffic, the standby NIC is utilized (see Section 9: Network Configuration).

5 VMware Fault Tolerance Limitations

In addition to the pre-requisites for enabling FT, there are certain limitations for implementing FT. For a detailed list of limitations, refer to the vSphere Availability Guide:

www.vmware.com/pdf/vsphere4/r40/vsp_40_availability.pdf

- Virtual machines cannot use more than 1 vCPU (vSMP is not supported).
- Hot-plug devices are not supported.
- Paravirtual SCSI adapters and legacy network drivers are not supported.
- VMXNET3 is not supported.
- VMDirectPath is not available for FT virtual machines.

- For FT enabled virtual machines, Extended Page Tables and Rapid Virtualization Indexing are automatically disabled.
- Virtual machine snapshots are not supported with FT enabled VMs.
- For backups, client agents need to be installed.
- DRS is automatically disabled for a FT enabled virtual machine.

6 Reference Architecture

This section will describe the reference architecture for implementing VMware FT on PowerEdge R910 servers using a Dell EqualLogic PS6010 iSCSI storage array and PowerConnect 8024F 10Gb Ethernet switches.

Design Principles

The following design principles were used during the design of this architecture:

- *Optimal Server Configuration for VMware Fault Tolerance:* When VMware FT is enabled on a virtual machine, both the Primary and Secondary VMs run simultaneously consuming both processor and memory resources. With increased memory capacity, the R910 proves to be a good choice for hosting FT-enabled VMs.
- *Optimal Bandwidth for FT Logging Traffic:* The FT logging network is used by the Primary VMs to send the non-deterministic events to the Secondary VMs. The logging traffic will be higher for virtual machines running I/O intensive applications. To provide enough bandwidth for the FT logging traffic, a 10 Gigabit Ethernet network is used.
- *Availability for the Network:* To achieve network availability, multiple network adapters are configured in teams at the virtual switch level. The network adapters are configured in active/standby mode.

Configuration Guidelines

The reference architecture discussed in this whitepaper consists of three PowerEdge R910 servers connected to a Dell EqualLogic PS6010 in an iSCSI SAN. The network infrastructure consists of two PowerConnect 8024F 10 Gigabit Ethernet switches (for redundancy) with iSCSI traffic on dedicated 10Gb NIC ports (as shown in Figure 1). For simplicity the figure below shows connections from only one server. The other two servers are connected in a similar fashion.

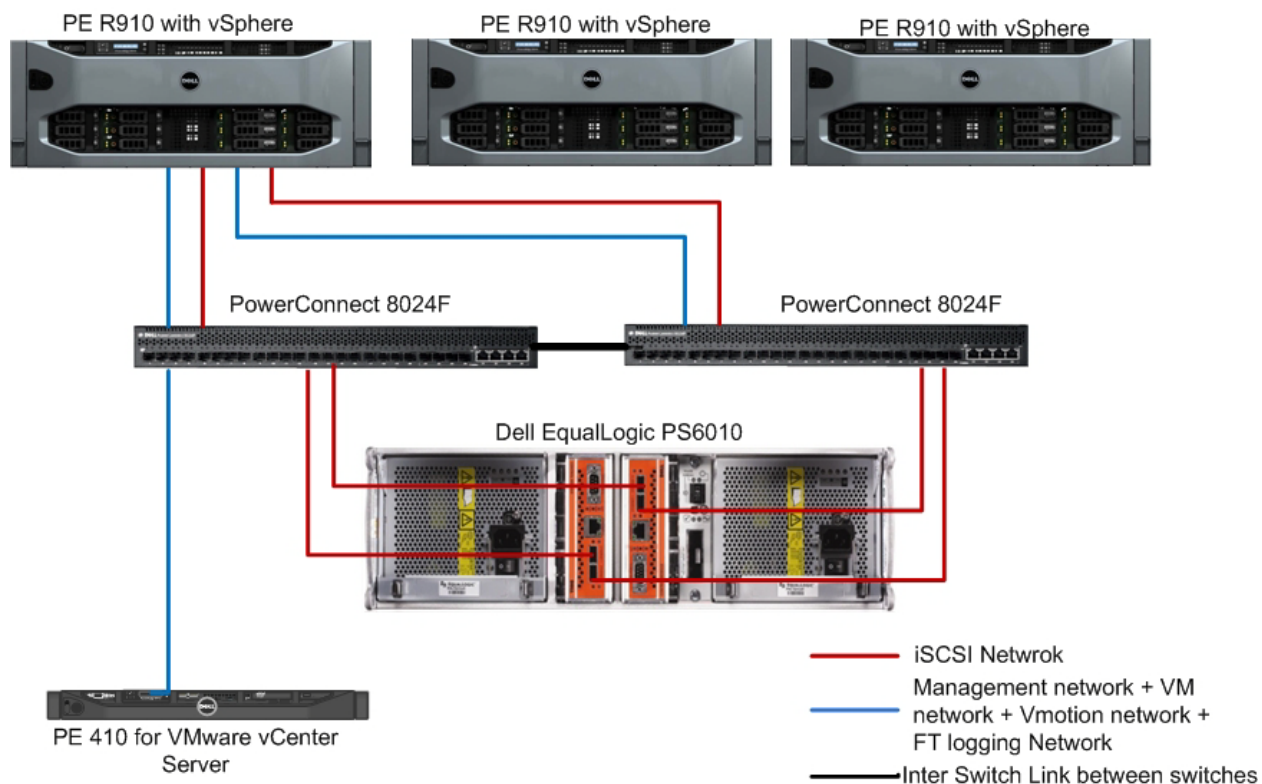


Figure 1: Reference Architecture for Implementing VMware FT

7 Hardware Configuration

Dell PowerEdge R910 Configuration

- ESX hosts that contain Secondary VMs will suffer increased CPU resources which are not reserved. The CPU requirements for these servers should include both Primary and Secondary virtual machines.
- All the ESX hosts in an FT implementation are recommended to have identical physical configurations including processor type and speed, BIOS level, peripherals, and memory configuration.
- ESXi 4.0 with the latest patch/updates is installed on all the ESX host servers. For current updates and patches, please refer to: <https://www.vmware.com/mysupport/download/>
- For increased bandwidth, 10 Gigabit Ethernet adapters are used (Broadcom BCM57111 SFP+) for the entire configuration.
- The 10GbE NIC ports are configured redundantly into two groups of network traffic. Due to high I/O demand, iSCSI has its own dedicated bandwidth. The management network, Virtual Machine network, VMotion network, and FT logging network share the remaining two ports.

Dell EqualLogic PS6010

- The Dell EqualLogic PS6010 iSCSI array has two data ports for each controller and one dedicated management port.
- For multi-pathing, both of the data ports on the controller are connected to the iSCSI network. The data ports are connected to different network switches to avoid a single point of failure at the switch level.
- For managing the PS array, the dedicated 100Mb/s management port can be configured to connect to the management network available in the infrastructure.
- Section 10 below covers the specific settings required to configure the array.
- The PS6010 array is configured with RAID 6 for better performance and reliability.

Dell PowerConnect 8024F Switches

There are five different types of networks in the virtual infrastructure:

1. Management network
2. Virtual Machine network
3. VMotion traffic network
4. Fault Tolerance network
5. iSCSI network

As shown in the reference architecture diagram, two Dell PowerConnect 8024 switches are used for all five of the different types of network traffic - isolated by VLANs. The switches are interlinked using two 10GbE links for optimal bandwidth and failover capabilities in case of a failure.

8 Network Configuration

This section describes the configuration of the physical network adapters, PowerConnect switches, virtual network adapters and virtual switches to support the various traffic types and achieve load-balancing and redundancy.

Each R910 host in the reference architecture discussed has two dual port 10GbE BCM57711 SFP+ network adapters. These physical NICs are enumerated as follows:

- vmnic4, vmnic5: First and second ports, respectively, on the first dual-port Broadcom NetXtreme II 57711 SFP+ Gigabit Ethernet controller
- vmnic6, vmnic7: First and second ports, respectively, on the second dual-port Broadcom NetXtreme II 57711 SFP+ Gigabit Ethernet controller

The PowerConnect switches are interlinked using two 10Gb Ethernet connections (Figure 1), which provides optimal bandwidth for the traffic between the switches and failover capabilities if there is a link failure. Figure 2 below shows how the host is physically connected to the network in an FT implementation.

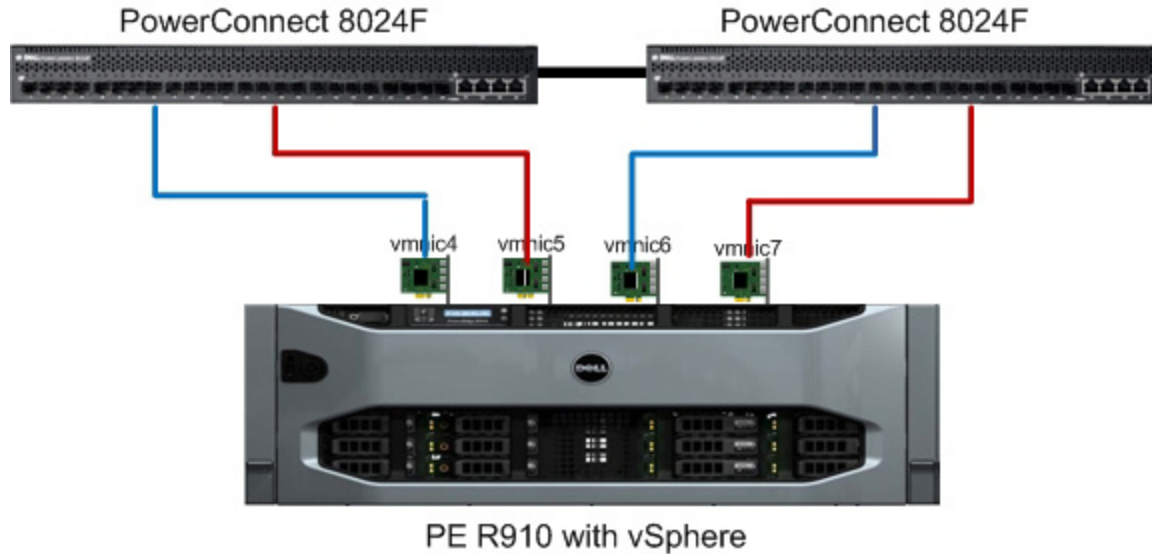


Figure 2: Configuring Network Adapters on Host

Virtual Machine, Management, VMotion and Fault Tolerance Logging Network

For the virtual machine network, management network, VMotion network and FT logging network, two 10 Gigabit Ethernet physical network ports on the ESXi host *vmnic4* and *vmnic6* are teamed at the virtual switch in an active/standby configuration. A virtual switch *vSwitch0* is created using *vmnic4* and *vmnic6* as uplinks. Four *portgroups* are created on *vSwitch0*: a *portgroup* for the VM network, a VMkernel *portgroup* for the VMotion network, a *portgroup* for ESXi host management and a VMkernel *portgroup* for the FT logging network. Figure 3 below illustrates the virtual switch configuration with *portgroups* and how the virtual switch is connected to the physical network ports on the ESXi host. Best practices recommend the enablement of jumbo frames on the VMkernel *portgroup* created for FT logging traffic.

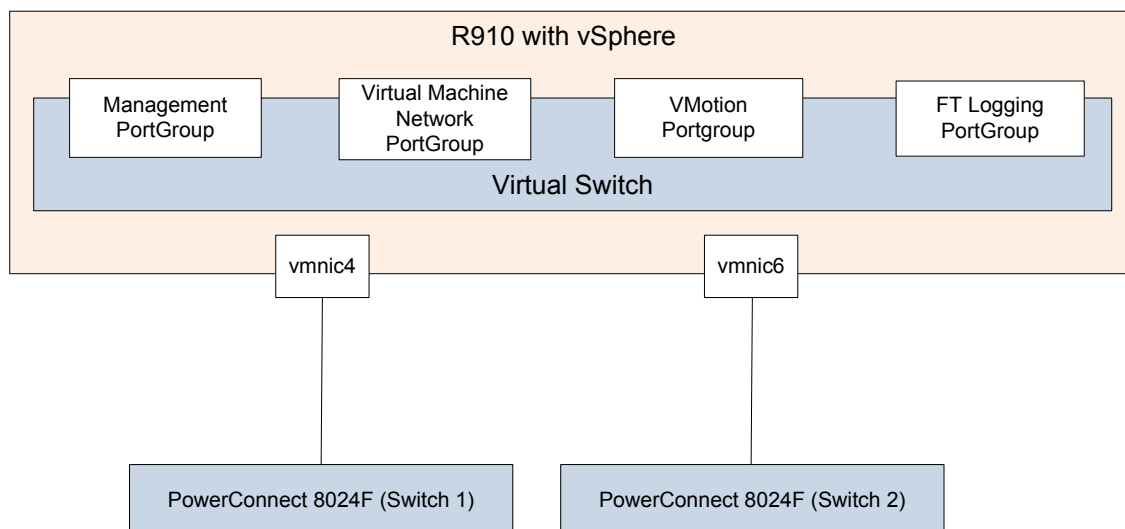


Figure 3: Virtual Switch and Portgroups

Traffic Isolation Using VLANs

All five of the different types of network traffic share the same network switches. However the iSCSI traffic is given dedicated NIC/Switch ports due to high I/O requirements. VLANs are configured to isolate the rest of the four traffic types that share common NIC adapter ports:

1. Management
2. VMotion
3. FT Logging
4. Virtual Machine

Network traffic is tagged with the respective VLAN ID for each traffic type in the virtual switch. This is achieved through Virtual Switch Tagging (VST) mode. In this mode, a VLAN is assigned to each of the four port groups. The virtual switch port group tags all outbound frames and removes tags for all inbound frames.

- Trunking must be used so that all of the VLANs can share the same physical connection. All of the ports connected to the ESXi hosts on the physical switches are configured in trunk mode.
- VMotion traffic is unencrypted; it is highly recommended that VMotion traffic is isolated.
- Routing between VLANs is dependent on specific customer requirements and is not included in this reference architecture.

NIC Teaming and Failover

The four traffic types sharing common NIC ports should be provisioned as shown in Figure 4. VMotion and virtual machine traffic are active on *vmnic4* with *vmnic6* in standby mode. FT Logging and Management traffic are active on *vmnic6* with *vmnic4* configured as the standby port.

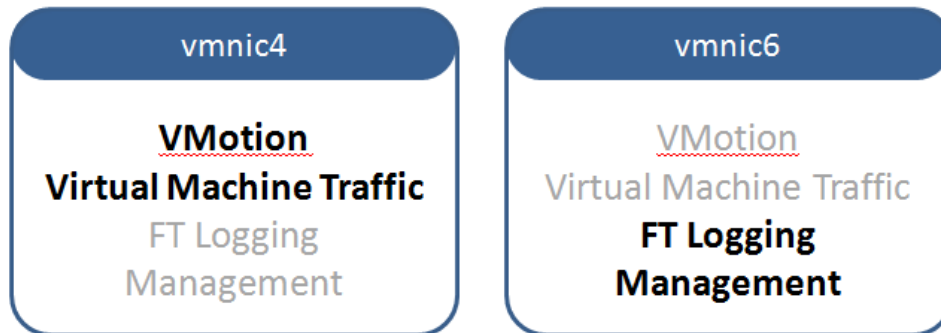


Figure 4: NIC Teaming and Failover Configuration

9 iSCSI Configuration

This section discusses the best practices for configuring the iSCSI SAN using the software iSCSI initiator in ESXi. The software initiators in guest Operating Systems are also supported, but are considered outside the scope of this document.

The two 10Gb Ethernet network ports used for iSCSI communications on each ESXi host are isolated from other network traffic on the shared physical switches with VLANs. Two virtual switches (*vSwitch1* & *vSwitch2*) are created with one network port added to each vSwitch. As shown in figure 5, two VMkernel portgroups (*vmkernel-iscsi1* and *vmkernel-iscsi2*) and a VMkernel IP interface per portgroup (*vmk0* and *vmk1*) are created. VLAN tagging is enabled at the portgroup level. These portgroups are used by the software iSCSI initiator to connect to the PS Series storage array. The software iSCSI initiator will see two paths to the target volume on the PS Series SAN. Round Robin path selection policy can be used to balance the load between the two physical NIC ports.

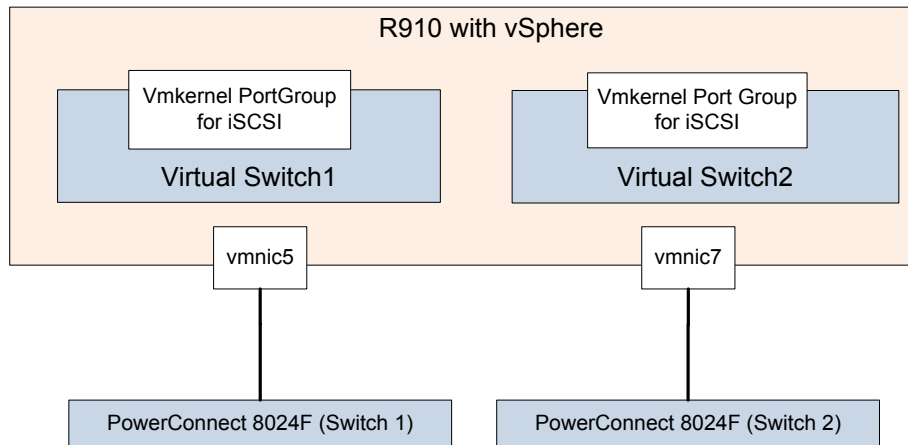


Figure 5: Virtual Switch and Portgroups for iSCSI

EqualLogic Network Requirements

In addition to the guidelines discussed previously, Dell EqualLogic storage arrays have specific recommendations for connecting to the network. A few of the more relevant recommendations are highlighted here. For more information, see the Dell EqualLogic PS Quick Start Guide at <https://www.equallogic.com/support/> (Account registration may be required).

- Do not use Spanning-Tree (STP) on switch ports that connect end-nodes (iSCSI initiators or array network interfaces). If STP or Rapid STP (preferable to STP) is required, enable the port settings (FastLink or Port Fast) on the switches to enable the port to immediately transition into the STP forwarding state upon link up. This functionality can reduce network interruptions that occur when devices restart and should only be enabled on switch ports that connect end-nodes. Note: The use of Spanning-Tree for a single-cable connection between switches is encouraged, as is the use of trunking for multi-cable connections between switches.

- Enable Flow Control on each switch port and NIC that handles iSCSI traffic. PS Series arrays will correctly respond to Flow Control.
- Disable unicast storm control on each switch that handles iSCSI traffic if the switch provides this feature. The use of broadcast and multicast storm control is encouraged on switches.
- Enable Jumbo Frames on the switches, virtual switches and VMware kernel interfaces.

10 References

- Business Ready Solution for Virtual Infrastructure Availability Using Dell PowerEdge Servers, Dell PowerVault Storage and VMware vSphere
http://content.dell.com/us/en/highered/d/business-solutions-engineering-docs-en/Documents-Dell_Virtualization_Availability_Solution.pdf.aspx
- VMware white paper on protecting mission critical workloads using FT
www.vmware.com/files/pdf/resources/ft_virtualization_wp.pdf
- VMware Fault Tolerance Recommendations and Considerations on VMware vSphere 4
www.vmware.com/files/pdf/fault_tolerance_recommendations_considerations_on_vmw_vsphere4.pdf
- VMware vSphere 4 Fault Tolerance: Architecture and Performance
http://www.vmware.com/files/pdf/perf-vsphere-fault_tolerance.pdf
- Configuring VMware vSphere Software iSCSI with Dell EqualLogic PS Series Storage
<http://www.equallogic.com/resourcecenter/assetview.aspx?id=8453>