NETWORKING BEST PRACTICES

FOR VMWARE® INFRASTRUCTURE 3 ON DELL™ POWEREDGE™ BLADE SERVERS

April 2009

Dell Virtualization Solutions Engineering
www.dell.com/virtualization

Information in this document is subject to change without notice.

This white paper is for informational purposes only and may contain typographical errors or technical inaccuracies. The content is provided as is, without express or implied warranties of any kind.

Contents

Table of Figures

1 Introduction

This whitepaper provides an overview of the networking architecture for VMware® Infrastructure 3 on Dell™ PowerEdge blade servers. It provides best practices for deploying and configuring your network in the VMware environment. References to other guides for step by step instructions are provided. The intended audiences for this whitepaper are systems administrators who want to deploy VMware virtualization on Dell PowerEdge blade servers and iSCSI storage.

The network architecture discussed in this white paper primarily focuses on iSCSI SAN. Best practices for Fibre Channel SAN are not covered in this document.

2 Overview

The PowerEdge M1000e is a high density and energy efficient blade chassis. It supports up to sixteen half height blade servers or eight full height blade servers and three layers of I/O fabric (A, B and C), which you can select between combinations of Ethernet, InfiniBand, and Fibre Channel modules. You can install up to six hot-swappable I/O modules in the enclosure, including Fibre Channel switch I/O modules, Fibre Channel pass-through I/O modules, InfiniBand switch I/O modules, Ethernet switch I/O modules, and Ethernet pass-through module I/O modules.
The integrated Chassis Management Controller also enables easy management of I/O Modules through a single secure interface.

2.1 Fabrics

The PowerEdge M1000e system consists of three I/O fabrics: Fabric A, B and C. Each fabric is comprised of two I/O modules which add up to a total of six I/O modules. The modules are A1, A2, B1, B2, C1 and C2.
The following figure illustrates the different I/O modules supported by the chassis.

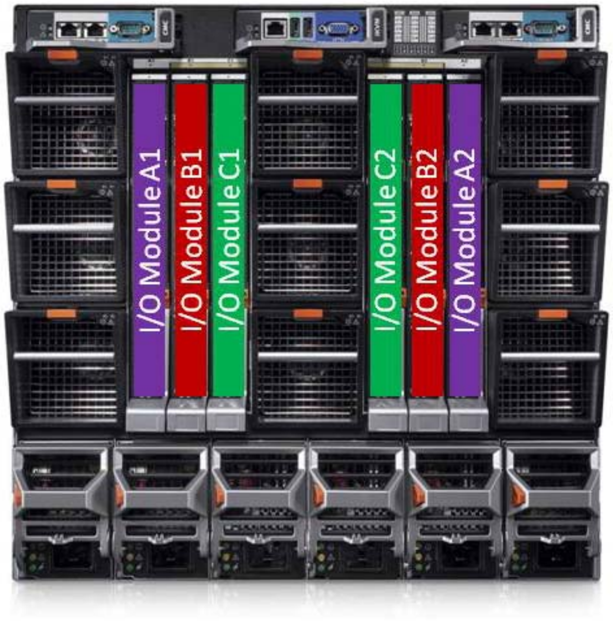

Figure 1: Blade Fabric Layout

- **Fabric A** is a redundant 1Gb Ethernet fabric that supports I/O module slots A1 and A2. The integrated Ethernet controllers in each blade dictate Fabric A as an Ethernet-only fabric.
- **Fabric B** is a 1/10 Gb/sec dual port, redundant fabric that supports I/O module slots B1 and B2. Fabric B currently supports 1/10Gb Ethernet, InfiniBand, and Fibre Channel modules. To communicate with an I/O

module in the Fabric B slots, a blade must have a matching mezzanine card installed in a Fabric B mezzanine card location. You can install modules designed for Fabric A in the Fabric B slots.

- **Fabric C** is a 1 to 10 Gb/sec dual port, redundant fabric that supports I/O module slots C1 and C2. Fabric C currently supports Gb Ethernet, Infiniband, and Fibre Channel modules. To communicate with an I/O module in the Fabric C slots, a blade must have a matching mezzanine card installed in a Fabric C mezzanine card location. You can install modules designed for Fabric A in the Fabric C slots.

2.2 I/O Modules

This subsection lists all the I/O modules that the PowerEdge M1000e chassis supports.

- **PowerConnect M6220 Ethernet Switch:** This Includes 16 internal server 1Gb Ethernet ports, 4 fixed copper 10/100/1000Mb Ethernet uplinks plus two of the following optional modules:
    - 48Gb (full duplex) Stacking module
    - 2 x 10Gb Optical (XFP-SR/LR) uplinks
    - 2 x 10Gb copper CX4 uplinks.

  The Standard Features include:
    - Layer 3 routing (OSPF, RIP, VRRP)
    - Layer 2/3 QoS
- **PowerConnect M8024 Ethernet Switch (10Gb Module)**: This includes 16 internal server 1/10Gb Ethernet ports, upto 8 external 10GbE ports via upto 2 selectable uplinks modules, 4-port SFP plus 1 10GbE module and 3-port CX-4 10GbE copper module.

  The Standard Features include:
    - Layer 3 routing (OSPF, RIP, VRRP)
    - Layer 2/3 QoS
- **Cisco® Catalyst Blade Switch M 3032:** This includes 16 internal server 1Gb Ethernet ports, 4 fixed copper 10/100/1000Mb Ethernet uplinks plus 2 optional module bays which can support either 2 x 1Gb copper or optical SFPs.

  The Standard features include:
    - Base Layer 3 routing (static routes, RIP)
    - L2/3 QoS
- **Cisco Catalyst Blade Switch M 3130G:** This **i**ncludes 16 internal server 1Gb Ethernet ports, 4 fixed copper 10/100/1000Mb Ethernet uplinks plus 2 optional module bays which can support either 2 x 1Gb copper or optical SFPs.

  The Standard Features include:
    - Base Layer 3 routing (static routes, RIP)
    - L2/3 QoS
    - Virtual Blade Switch Technology provides a high bandwidth interconnection between 8 CBS 3130 switches. You can configure and manage the switches as 1 logical switch. This radically simplifies management, allows server to server traffic to stay within the VBS domain as against congesting the core network, and can significantly help consolidate external cabling.
    - Optional software license key upgrades to IP Services (Advanced L3 protocol support) and Advanced IP Services (IPv6)
- **Cisco Catalyst Blade Switch M 3130X (supports 10G modules)**: This includes 16 internal server 1Gb Ethernet ports, 4 fixed copper 10/100/1000Mb Ethernet uplinks, 2 stacking ports and support for 2 X2 modules for up to a total of two 10G CX4 or SR/LRM uplinks.

  The Standard Features include:
    - Base Layer 3 routing (static routes, RIP)
    - L2/3 QoS
    - Virtual Blade Switch Technology provides a high bandwidth interconnect between up to 8 CBS 3130 switches enabling them to be configured and managed as 1 logical switch. This radically

simplifies management, allows server-server traffic to stay within the VBS domain vs. congesting the core network, and can help significantly consolidate external cabling.

  - Optional software license key upgrades to IP Services (Advanced L3 protocol support) and Advanced IP Services (IPv6)

- **Dell Ethernet Pass-Through Module:** This supports 16 x 10/100/1000Mb copper RJ45 connections. This is the only Ethernet Pass-through module in the market that supports the full range of 10/100/1000Mb operation.
  Note: PowerEdge M1000e also supports additional I/O modules - Brocade M5424 SAN I/O Module , Brocade M4424 SAN I/O Module, 4Gb Fibre Channel Pass-through Module and Infiniband.

  New I/O modules may have been released after this document is published. For the latest information and detailed specification refer to www.dell.com

  For more information on the fabrics, I/O modules, mezzanine cards, mapping between mezzanine cards and I/O modules refer to *Hardware Owner's Manual* of your blade server model under section *About Your System* at http://support.dell.com.

2.3 Mapping between Blade Server and I/O Modules in Chassis

This section describes how the onboard network adapter and add-in mezzanine cards map to the I/O modules in the chassis.

Each half height blade has a dual port onboard network adapter and two optional dual port mezzanine I/O cards. One mezzanine I/O card is for Fabric B and one mezzanine I/O card for Fabric C.

The following figure illustrates how these adapters are connected to the I/O modules in the chassis.

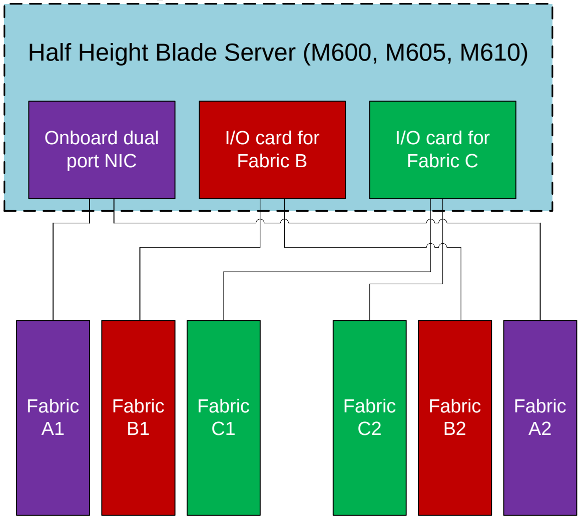

Figure 2: Adapter and I/O Modules connection in Chassis for Half Height Blades

Each full height blade has two dual port onboard network adapter and four optional dual port I/O mezzanine cards. Two I/O mezzanine cards are for Fabric B and two I/O mezzanine cards are for Fabric C. The following figure illustrates how the network adapters on a full height blade are connected to the I/O modules.

The following figure illustrates how the network adapters on a full height blade are connected to the I/O modules.

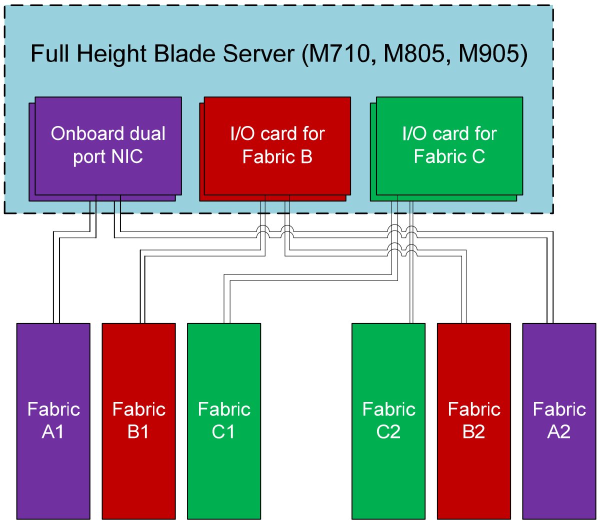

Figure 3: Adapter and I/O Modules connection in Chassis for Full Height Blades

For more information on port mapping, see the *Hardware Owner's Manual* for your blade server model at http://support.dell.com.

2.4 Mapping between ESX Physical Adapter Enumeration and I/O Modules

The following table shows how the ESX/ESXi 3.5 servers enumerate the physical adapters and the I/O modules they connect to. This enumeration only applies to blade servers that have all their I/O mezzanine cards populated with dual port network adapters.

Table 1: ESX/ESXi Physical Adapter Enumeration

ESX/ESXi Network Adapter enumeration	Full Height Blade Connection (M710, M805, M905)	Half Height Blade Connection (M600, M605, M610)
vmnic0	I/O Module A1 (port *n*)	I/O Module A1(port *n*)
vmnic1	I/O Module A2 (port *n*)	I/O Module A2(port *n*)
vmnic2	I/O Module A1 (port *n+8*)	I/O Module B1(port *n*)
vmnic3	I/O Module A2 (port *n+8*)	I/O Module B2(port *n*)
vmnic4	I/O Module C1 (port *n*)	I/O Module C1(port *n*)
vmnic5	I/O Module C2 (port *n*)	I/O Module C2(port *n*)
vmnic6	I/O Module B1 (port *n*)	N/A
vmnic7	I/O Module B2 (port *n*)	N/A
vmnic8	I/O Module C1 (port *n+8*)	N/A
vmnic9	I/O Module C2 (port *n+8*)	N/A
vmnic10	I/O Module B1 (port *n+8*)	N/A
vmnic11	I/O Module B2 (port *n+8*)	N/A

In the above table, *port n* refers to the port in the I/O modules to which the physical adapter connects, where *n* represents the slot in which the blade is installed. For example, *vmnic0* of a PowerEdge M710 blade in slot 3 is connected to I/O module A1 at port 3. *vmnic3* for the same server connects to I/O module A2 at port 11.

3 Network Architecture

Network traffic can be divided into two primary types - Local Area Network (LAN) and iSCSI Storage Area Network (SAN). LAN consists of traffic from virtual machines, ESX/ESXi management (service console for ESX), and VMotion. iSCSI SAN consists of iSCSI storage network traffic. You can replace the iSCSI network with the Fibre Channel SAN by replacing the network adapters with the Fibre Channel and network switches with Fibre Channel switches. This section discusses the best practices for the iSCSI SAN only.

3.1 Design Principles

The following design principles are used to develop the network architecture:

- **Redundancy:** Both LAN and iSCSI SAN have redundant I/O modules. Redundancy of the network adapters is achieved through NIC teaming at the virtual switch.
- **Simplified management through stacking:** You can combine switches servicing the same traffic type into logical fabrics using the high-speed stacking ports on the switches.
- **iSCSI SAN physical isolation:** You should physically separate the iSCSI SAN network from the LAN network. Typically iSCSI traffic is network intensive and may consume disproportionate share of the switch resources if sharing a switch with LAN traffic.
- **Logical isolation of VMotion using VLAN:** VMotion traffic is unencrypted. It is important to logically isolate the VMotion traffic using VLANs.
- **Optimal performance:** Load balancing is used to achieve the highest throughput possible

3.2 Recommended Configurations

Based on the bandwidth requirements of LAN and iSCSI SAN, there are different ways to configure the I/O modules. The different configurations are listed in the table below. They meet the design principles listed above.

Table 2: Bandwidth Configurations for LAN and iSCSI SAN

	Minimum Configuration	**Base - High LAN Bandwidth**	**Balanced**	**High iSCSI Bandwidth**	**Isolated Fabric**
I/O Module A1	LAN	LAN	LAN	LAN	LAN
I/O Module B1	iSCSI SAN	iSCSI SAN	iSCSI SAN	iSCSI SAN	iSCSI SAN
I/O Module C1	Blank	LAN	LAN	iSCSI SAN	Isolated Fabric
I/O Module C2	Blank	LAN	iSCSI SAN	iSCSI SAN	Isolated Fabric
I/O Module B2	iSCSI SAN	iSCSI SAN	iSCSI SAN	iSCSI SAN	iSCSI SAN
I/O Module A2	LAN	LAN	LAN	LAN	LAN

- **Minimum Configuration:** This is the simplest configuration and has the minimum number of I/O modules. Two I/O modules are dedicated for LAN and two for iSCSI SAN. Two modules are left blank and you can populate them at any time to meet any growing bandwidth demands.
- **Base - High LAN Bandwidth:** In this configuration four I/O modules are dedicated to the LAN and two I/O modules dedicated to the iSCSI SAN. This configuration is useful for environments which have high LAN bandwidth requirements. Requirements of most environments can be met with this configuration. The rest of this whitepaper uses this configuration to further illustrate best practices. You can easily apply the best practices to other configurations.
- **Balanced:** In this configuration three I/O modules are dedicated to both LAN and iSCSI SAN. Both fabrics have an equal amount of bandwidth allocated. This configuration is useful for environments which have high back end SAN requirements such as database environments.

- **High iSCSI SAN Bandwidth:** In this configuration two I/O modules are dedicated to the LAN and four I/O modules are dedicated to the iSCSI SAN. This configuration is useful for environments which have high back end SAN requirements such as database environments and low LAN bandwidth requirements.
- **Isolated Fabric:** Certain environments require physically isolated network of certain class of virtual machines (such as credit card transactions). To accommodate those virtual machines, we can dedicate two redundant I/O modules. The two additional switches are stacked together to form a third fault-tolerant logical fabric.

The following sections describe the best practices to configure the LAN and iSCSI SAN network. The *high LAN bandwidth* configuration is used as an example for illustrations.

3.3 Local Area Network (LAN)

The LAN traffic includes the traffic generated from Virtual Machines and ESX Management VMotion. This section provides the best practices to configure LAN including traffic isolation using VLANs, load balancing, and external connectivity using uplinks. Based on Table 2, the I/O modules are dedicated to the LAN.
All the I/O Modules are stacked together to create single Virtual Switch to further simplify the deployment, management, and increase the load balancing capabilities of the solution.

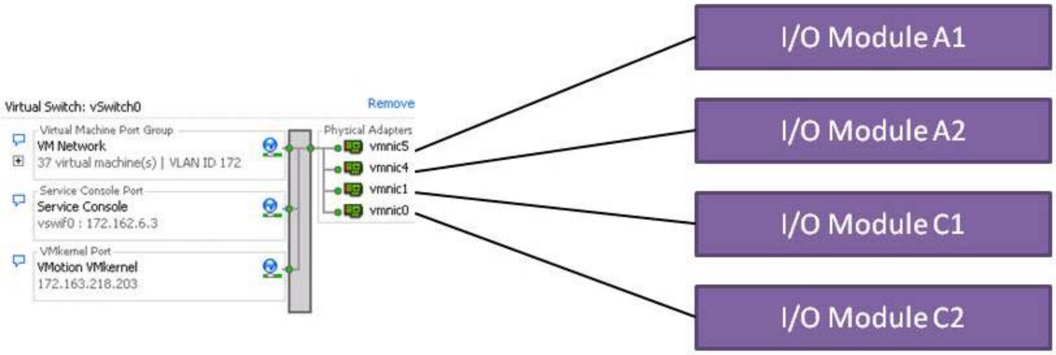

Figure 4: Virtual Switch for LAN on Half Height Blade Servers

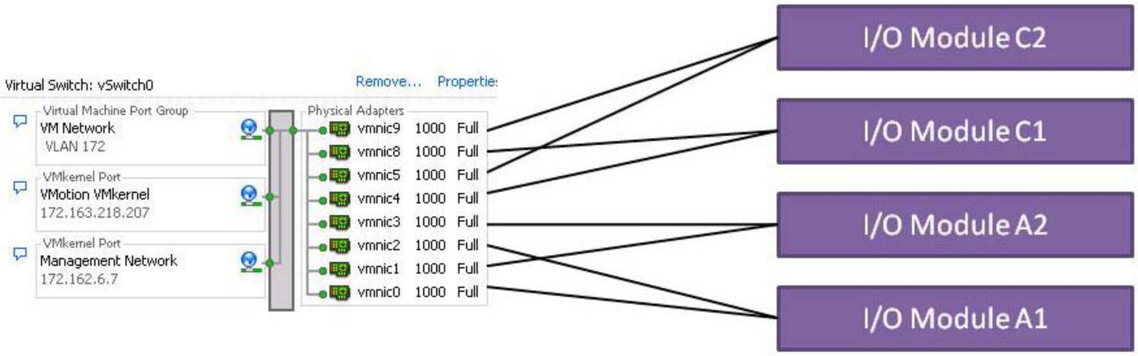

Figure 5: Virtual Switch for LAN on Full Height Blade Servers

3.3.1 Traffic Isolation using VLANs

The traffic on the LAN network is separated into three VLANs, one VLAN each for management, VMotion, and virtual machine traffic. Network traffic is tagged with respective VLAN ID for each traffic type in the virtual switch. This is achieved through the Virtual Switch Tagging (VST) mode. In this mode, a VLAN is assigned to each of the three port groups. The virtual switch port group tags all outbound frames and removes tags for all inbound frames. For example:

- Service Console (VLAN 162)
- vMotion (VLAN 163)
- General Virtual Machine Traffic (VLAN 172)
- Special Virtual Machine Traffic #1 (VLAN 173)
- Special Virtual Machine Traffic #1 (VLAN 174)

Trunking must be used so that all the VLANs can share the same physical connection. To achieve this, all the internal ports in the Cisco I/O modules are configured to be in the trunk mode.

VMotion traffic is unencrypted, and so it is highly recommended to isolate VMotion traffic . Using the above VLAN configuration, we achieve traffic isolation between various traffic types, including the VMotion traffic. The four network adapters provide sufficient bandwidth for all the traffic types.

3.3.2 Load Balancing

The virtual switch provides fault-tolerance and load balancing by allowing multiple physical Network Identification Cards (NICs) to be connected to a single switch. The stacking link between the I/O Modules (used for LAN) creates a single virtual switch which provides failover and load-balancing between the physical NICs connected to different I/O Modules.

VMware virtual switch provides three options to configure load balancing:

- **Route based on the originating virtual switch port ID (default configuration):** Here a physical adapter is selected for transmit based on the hash of the virtual port. This means that a given virtual network adapter will use only one physical adapter at any given time to transmit network packets. Packets are received on the same physical adapter.
- **Route based on source MAC hash:** Here a physical adapter is selected for transmit based on the hash on the source MAC address. This means that a given virtual network adapter will use only one physical adapter at any given time to transmit network packets. Packets are received on the same physical adapter.
- **Route based on IP hash:** Here the physical adapter is selected for transmit based on the hash on the source and destination IP address. Because you may select different adapters based on the destination IP, you need to configure both the virtual switches and the physical switches to support this method. The physical switch combines the connections to multiple NICs into a single logical connection using EtherChannel, and the load balancing algorithm selected for the switch will then determine which physical adapter receives the packets.

Note: The virtual switch and the physical switch hashing algorithms work independent of each other.

If connectivity to a physical network adapter is lost, then any virtual network adaptor that is currently using that physical channel will fail-over to a different physical adapter, and the physical switch will learn that the MAC address has moved to a different channel.

3.3.3 External Connectivity using Uplinks

There are multiple options for connecting the blade chassis to an existing LAN. You can use the pass-through module to connect each blade server directly into an existing network. This is the simplest solution to connect to an existing infrastructure, but it requires many cables.

Each Ethernet switch has four built-in 1 Gb ports, and there are various options for adding additional 1Gb and 10Gb Ethernet ports. When using multiple Ethernet ports, you must join them together into a single EtherChannel, and distribute them evenly across all the physical switches in a stack of switches to provide redundancy.

You can also connect multiple blade chassis together. If the total number of front-end switches is less than 8 for Cisco, or 12 for Dell, then you can stack all the switches together into a single Virtual Blade Switch. Multiple Virtual Blade Switches could be daisy-chained together by creating two EtherChannels.

3.4 iSCSI Storage Area Network (SAN)

The iSCSI SAN traffic includes the traffic generated between the ESX Servers and Storage Arrays. This section provides the best practices to configure iSCSI SAN including storage connectivity, load balancing, and external connectivity using uplinks.

The following figure illustrates the virtual switch configuration with port groups and how the virtual switch is connected to the physical network adapters and in turn to the I/O modules.

Figure 5 is based on ESX. If ESXi is used, you need not configure the service console port group for iSCSI.

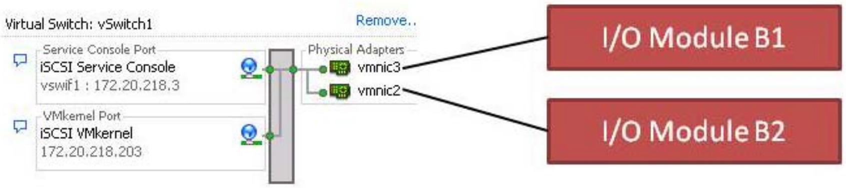

Figure 6: Virtual Switch for SAN on Half Height Blade Servers

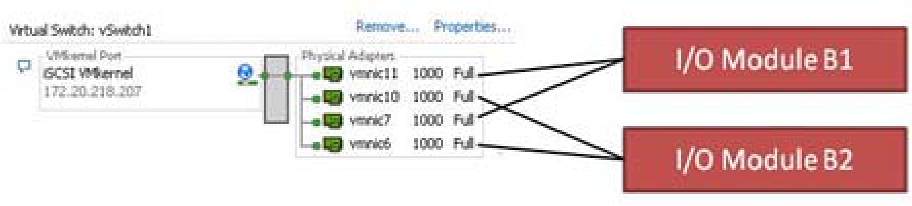

Figure 7: Virtual Switch for SAN on Full Height Blade Servers

3.4.1 Load Balancing

To understand load balancing, it is important to understand the basics of iSCSI and how it is implemented in VMware ESX/ESXi. Load balancing configuration depends on the iSCSI storage device that is being used. Following are some key fundamentals that will you help understand how load balancing works:

- ESX software iSCSI initiator establishes only one connection to each target.
- The current ESX version uses only one path for each connection.
- For Dell EqualLogic SAN:
    - Dell EqualLogic SAN expose each LUN or volume as a separate target with LUN ID 0. Dell EqualLogic SAN exposes all the LUNs in a group using one IP address and hence only one path.
    - EqualLogic SAN automatically load balances the LUNs between the physical interfaces in the storage array
    - In the virtual switch, the load balancing must be enabled with the configuration 'Route based on IP hash'
- For Dell EMC SAN:
    - Dell EMC SAN exposes each physical interface as a separate target. Hence, each LUN has multiple paths depending on the number of physical interfaces
    - Multi-pathing can be achieved by manually selecting the active paths for each LUN in the vCenter multi-pathing dialog. The paths are selected so that the LUNs are balanced across the different physical paths.

4 References

iSCSI overview - A “Multivendor Post” to help our mutual iSCSI customers using VMware
http://virtualgeek.typepad.com/virtual_geek/2009/01/a-multivendor-post-to-help-our-mutual-iscsi-customers-using-vmware.html

Integrating Blade Solutions with EqualLogic SANs
http://www.dell.com/downloads/global/partnerdirect/apj/Integrating_Blades_to_EqualLogic_SAN.pdf

Cisco Products
http://www.cisco.com/en/US/products/ps6746/Products_Sub_Category_Home.html

Cisco 3130 Product Page
http://www.cisco.com/en/US/products/ps8764/index.html

VMware Infrastructure 3 in a Cisco Network Environment
http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/vmware/VMware.html

Cisco Catalyst 3750 and 2970 Switches: Using Switches with a PS Series Group
http://www.equallogic.com/resourcecenter/assetview.aspx?id=5269