

# Deploying an Active Fabric Network in a Small/Medium Data Center

Deployment for a PoD using virtual and physical networking

Dell Networking Solutions Engineering  
August 2013

## Revisions

| Date        | Version | Description     | Authors  |
|-------------|---------|-----------------|--|
| August 2013 | 1.1     | Initial release | Manjesh Siddamurthy, Victor Teeter, Jaiwant Virk |

Copyright © 2013 - 2016 Dell Inc. or its subsidiaries. All Rights Reserved.

Except as stated below, no part of this document may be reproduced, distributed or transmitted in any form or by any means, without express permission of Dell.

You may distribute this document within your company or organization only, without alteration of its contents.

THIS DOCUMENT IS PROVIDED “AS-IS”, AND WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED. IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE SPECIFICALLY DISCLAIMED. PRODUCT WARRANTIES APPLICABLE TO THE DELL PRODUCTS DESCRIBED IN THIS DOCUMENT MAY BE FOUND AT: <http://www.dell.com/learn/us/en/vn/terms-of-sale-commercial-and-public-sector-warranties>

Performance of network reference architectures discussed in this document may vary with differing deployment conditions, network loads, and the like. Third party products may be included in reference architectures for the convenience of the reader. Inclusion of such third party products does not necessarily constitute Dell’s recommendation of those products. Please consult your Dell representative for additional information.

Trademarks used in this text: Dell™, the Dell logo, Dell Boomi™, PowerEdge™, PowerVault™, PowerConnect™, OpenManage™, EqualLogic™, Compellent™, KACE™, FlexAddress™, Force10™ and Vostro™ are trademarks of Dell Inc. EMC VNX®, and EMC Unisphere® are registered trademarks of Dell. Other Dell trademarks may be used in this document. Cisco Nexus®, Cisco MDS®, Cisco NX-OS®, and other Cisco Catalyst® are registered trademarks of Cisco System Inc. Intel®, Pentium®, Xeon®, Core® and Celeron® are registered trademarks of Intel Corporation in the U.S. and other countries. AMD® is a registered trademark and AMD Opteron™, AMD Phenom™ and AMD Sempron™ are trademarks of Advanced Micro Devices, Inc. Microsoft®, Windows®, Windows Server®, Internet Explorer®, MS-DOS®, Windows Vista® and Active Directory® are either trademarks or registered trademarks of Microsoft Corporation in the United States and/or other countries. Red Hat® and Red Hat® Enterprise Linux® are registered trademarks of Red Hat, Inc. in the United States and/or other countries. Novell® and SUSE® are registered trademarks of Novell Inc. in the United States and other countries. Oracle® is a registered trademark of Oracle Corporation and/or its affiliates. VMware®, Virtual SMP®, vMotion®, vCenter® and vSphere® are registered trademarks or trademarks of VMware, Inc. in the United States or other countries. IBM® is a registered trademark of International Business Machines Corporation. Broadcom® and NetXtreme® are registered trademarks of QLogic is a registered trademark of QLogic Corporation. Other trademarks and trade names may be used in this document to refer to either the entities claiming the marks and/or names or their products and are the property of their respective owners. Dell disclaims proprietary interest in the marks and names of others.

# Table of Contents

|       |   |    |
|-------|---|----|
| 1     | Servers .....   | 7  |
| 1.1   | ESXi server configuration .....                                     | 7  |
| 1.1.1 | NPAR configuration .....  | 8  |
| 2     | vSwitch configuration .....   | 11 |
| 2.1.1 | Storage adapter configuration .....                                 | 12 |
| 2.1.2 | EqualLogic MEM — MPIO configuration .....                           | 14 |
| 2.2   | Windows Server configuration .....                                  | 15 |
| 2.2.1 | EqualLogic HIT kit .....  | 15 |
| 2.2.2 | MPIO configuration .....  | 15 |
| 2.2.3 | NIC Teaming - LAN load balancing .....                              | 22 |
| 3     | Networking .....  | 26 |
| 3.1   | Virtual Link Trunking (VLT) .....                                   | 26 |
| 3.1.1 | RSTP on VLT .....   | 27 |
| 3.1.2 | VLT topology .....  | 27 |
| 3.1.3 | Backup Link .....   | 32 |
| 3.1.4 | Monitoring VLT .....  | 32 |
| 3.2   | VRRP .....  | 33 |
| 3.3   | Enabling data center bridging (DCB) .....                           | 34 |
| 3.3.1 | Validating and troubleshooting the DCB settings .....               | 37 |
| 4     | Storage .....   | 40 |
| 4.1   | Configuring EqualLogic PS 6xxx/4xxx family multi-member pools ..... | 40 |
| 4.1.1 | RAID policies .....   | 41 |
| 4.1.2 | Volumes (iSCSI targets) .....                                       | 41 |
| 4.1.3 | Data Center Bridging .....  | 41 |
| 4.1.4 | Firmware Updates .....  | 44 |
| 5     | Network Management .....  | 45 |
| 5.1   | Graphing performance and traffic flow metrics .....                 | 49 |
| 5.2   | sFlow traffic .....   | 52 |
| 6     | Switch Configurations .....   | 54 |
| 7     | Observations .....  | 55 |
| A     | Configuration details .....   | 56 |

B Terminology.....57

C Additional resources.....59

    C.1 Servers .....59

    C.2 Networking.....59

    C.3 Storage .....59

## Executive summary

Active Fabric is a high-performance networking solution based on inter-connecting products, purpose-built for stitching together server, storage and workload elements in traditional and virtualized data centers. The value of this fabric is its scalability and cost-effectiveness in a new paradigm of technology where network is now the focus of virtualization and driven by software.

This deployment guide describes best practices for creating a converged and virtualized data center network. As this architecture grows, the network evolves into a fabric (physical switches that operate as a single logical device, realizing the simplicity of a single switch combined with the connectivity and resiliency of a network).

In recent years, there have been many fundamental and profound changes in the architecture of modern data centers. These data centers host the computational power, storage, networking, and applications that form the basis of any modern business. The traditional data center architecture and compute model, especially in the case of rack or blade servers use an x86 based processor. Traditionally, this approach has consisted of lightly used servers running a single operating system directly accessing all compute resources. Then came the hypervisors where multiple operating systems (VMs) run on virtualized resources. On the network side separate fabrics for storage and local traffic were the norm with multi-tier networks and complex manual provisioning. In the new dynamic data centers, there is a need for heavy server virtualization with some presence of physical servers, with high-bandwidth connections to network and a shared virtualized network.

This architecture shows a tight integration with servers and storage within the data center, as switch functionality migrates towards the network edge and the attached servers. The need for such an architecture has become apparent as computational resources, storage, and networking have become more tightly coupled in new data center designs such as point-of-delivery (PoD) systems, which are interconnected through a larger data center network (DCN). A PoD is a set of pre-defined compute, network and storage resources which often form the boundaries for workloads and their management.

This document demonstrates how to build the networking components on the servers by configuring the CNAs for network partitioning (NPAR) of physical adapter into virtual interfaces. Virtual switches (on a hypervisor) are then assigned to these virtual interfaces based on the function they perform. For example: the link that carries storage traffic has a dedicated virtual switch and interface. Similarly other applications could use the other virtual functions. Data Center Bridging (DCB) is introduced to facilitate convergence of LAN and SAN traffic over shared physical network and spans end-to-end from server to the storage. By following the design guidelines provided here, it is possible to introduce convergence and virtualization, which are a major contributor towards controlling capital and operating expenses associated with the data centers.

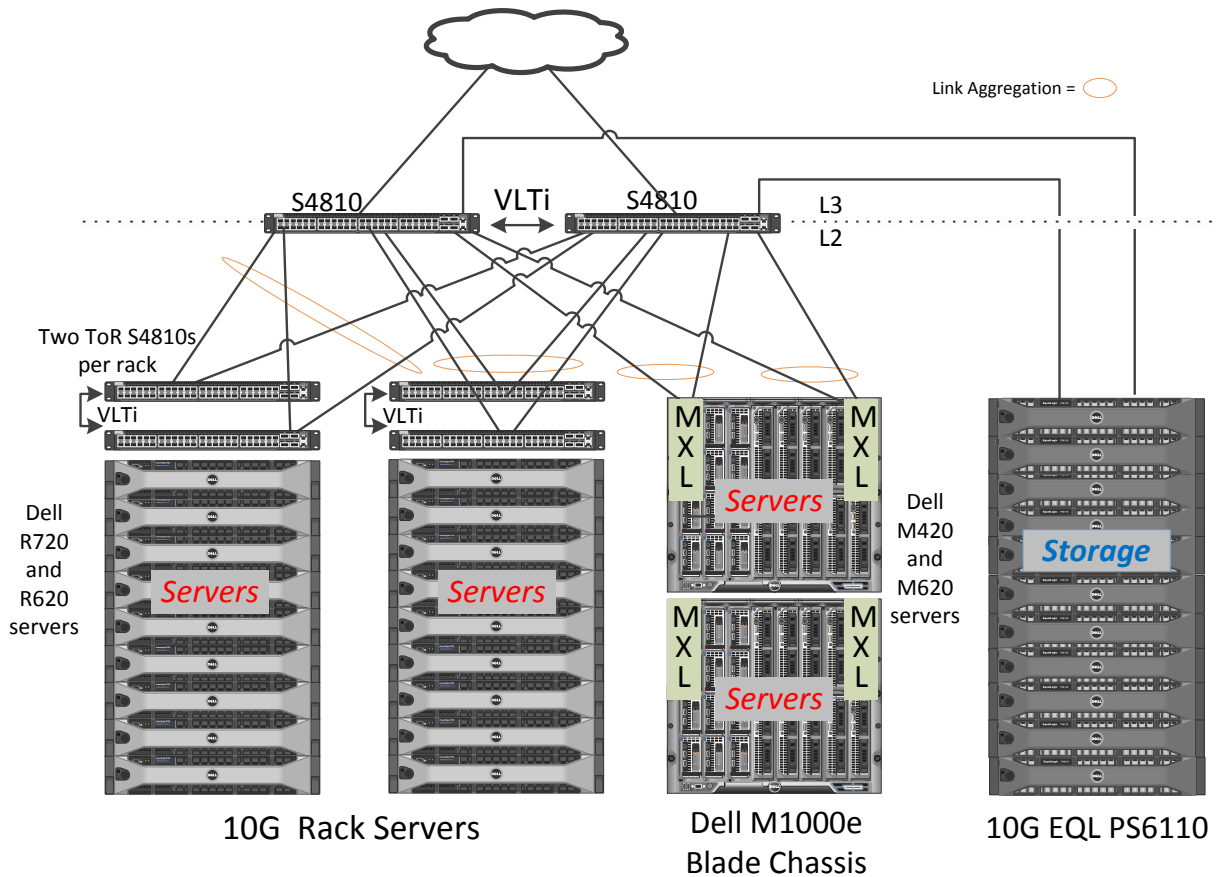


Figure 1 Active Fabric POD (point of delivery) for end-to-end iSCSI

**Note:** Not all equipment can be DCB enabled. Make sure your server, network, and storage equipment have the correct capabilities for DCB. All equipment mentioned in this document is DCB enabled.

In the topology above, each top-of-rack (ToR) switch pair needs four VLT ports at the aggregation layer. The aggregation/core layer S4810 switch pair can support up to 12 TOR VLT pairs which means the topology can be scaled up to 12 racks.

Procedures in this document are logically separated into Servers, Network and Storage for easier reference. Steps within each section are linear and can be run sequentially. You must complete all three sections somewhat in parallel.

# 1 Servers

For this iSCSI solution, the following Dell PowerEdge 12<sup>th</sup> generation servers are used:

- Dell PowerEdge R720 servers, Broadcom 10gig dual-port 57810s CNAs, VMware ESXi 5.1
- Dell PowerEdge R620 servers, Intel 10gig dual-port x520 CNAs and Broadcom 57810s, Windows Server 2012
- Dell PowerEdge M420 servers, Intel 10gig dual-port x520 CNAs, Windows Server 2012
- Dell PowerEdge M620 servers, Broadcom 10gig dual-port 57810s CNAs, VMware ESXi 5.1

**Note:** Intel and Broadcom are popular CNAs for the Dell iSCSI solution. Broadcom was chosen for NPAR with ESXi, and supports a maximum of 4 virtual NICs per physical port.

At the moment Intel does not support NIC partitioning (NPAR).

## 1.1 ESXi server configuration

Before wiring and configuring servers and switches to storage, it is a good idea to sketch how traffic segregation works with NIC partitioning and vSwitch. Figure 2 shows the blown up view of one BCM 57810 virtual ports mapping to vSwitch. The virtual NIC numbering varies based on the server PCIe slot chosen. The Dell server LOM always starts with vmnic0. After Broadcom NPAR, notice that one physical port has odd numbering virtual ports and the other physical port has even numbering virtual ports as shown in Figure 2. It is good practice to bind/map virtual ports of 2 different physical ports to a vSwitch to provide redundancy and NIC failover (See Figure 2).

**Note:** This configuration also applies to the Dell PowerEdge M620.

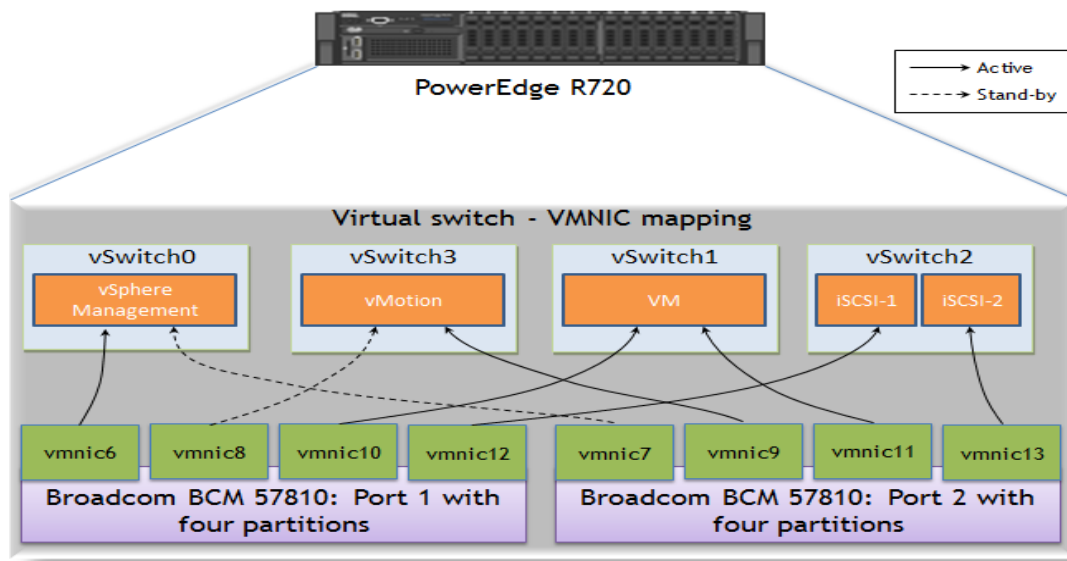


Figure 2 vSwitch and NPAR configuration on dual port BCM57810 PCIe card

### 1.1.1 NPAR configuration

NIC Partition (NPAR) – As the name says it's the technology that allows dividing physical link bandwidth to suit user requirement. Currently Broadcom & QLogic support NPAR on 10gig link. Refer appendix to learn more about NPAR.

Before installing ESXi, choose and insert the necessary PCIe cards. When the system boots, to enter Broadcom configuration screen press Ctrl+S on BIOS screen.

Figure 3 shows the Broadcom configuration page. Here, the first four NICs represent NICs on the motherboard and the highlighted PCIe NIC is used for NPAR.

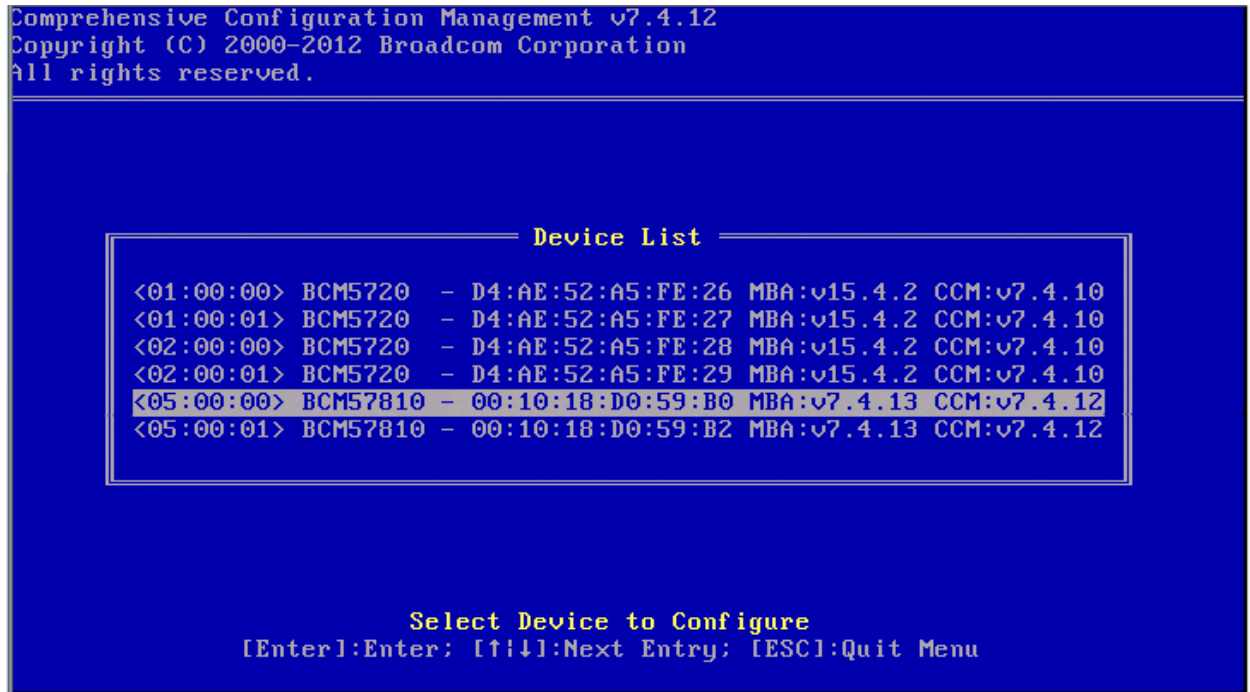


Figure 3 Landing screen for Broadcom configuration



To prepare the server for end-to-end iSCSI, the chosen physical NIC should be configured for NPAR and DCB-enabled (see Figure 4). When configured for NPAR, each Broadcom 57810s 10gig port on the Dell PowerEdge R720 is partitioned into four virtual ports.

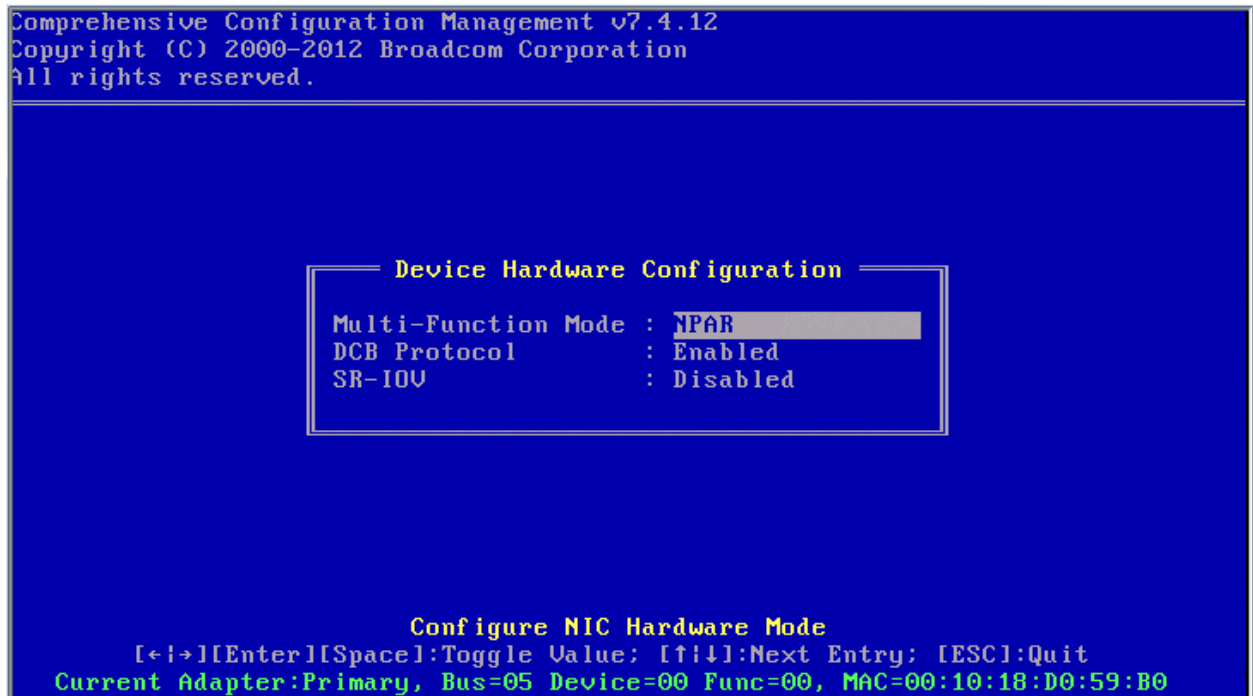


Figure 4 DCB and NPAR enabled for each selected physical NIC

As shown in Figure 5, one of four partitioned ports is enabled for iSCSI offload to carry iSCSI traffic and the remaining three ports are left as default for LAN traffic.

**Note:** No more than one of four partitioned ports can be enabled for iSCSI or FCoE.

Each partitioned port's maximum and relative bandwidth settings are left as defaults, 100 and 0 respectively. Each partition is therefore allowed to carry 10gig bandwidth, but if the total bandwidth for all four NIC partitioned ports crosses the 10gig threshold, oversubscription occurs. Using the DCB settings learned from the switch, the server then decides which packets are dropped and which packets are guaranteed bandwidth.

```
Comprehensive Configuration Management v7.4.12
Copyright (C) 2000-2012 Broadcom Corporation
All rights reserved.

----- NIC Partition Configuration -----

Flow Control : Tx: Send Pause on Rx Overflow
PF#0 L2=00:10:18:D0:59:B0(U) iSCSI=00:10:18:D0:59:B1(U) BW= 0:100 Eth
PF#2 L2=00:10:18:D0:59:B4(U) iSCSI=00:10:18:D0:59:B5(U) BW= 0:100 Eth
PF#4 L2=00:10:18:D0:59:B8(U) iSCSI=00:10:18:D0:59:B9(U) BW= 0:100 Eth
PF#6 L2=00:10:18:D0:59:BC(U) iSCSI=00:10:18:D0:59:BD(U) BW= 0:100 Eth:iSCSI
Reset Configuration to Default

Configure NIC Partition Parameters: (P):Permanent, (U):Virtual
[Enter]:Enter; [↑|↓]:Next Entry; [ESC]:Quit Menu
Current Adapter:Primary, Bus=05 Device=00 Func=00, MAC=00:10:18:D0:59:B0
```

Figure 5 Enabling iSCSI for one of four virtual NICs

## 2 vSwitch configuration

Once the NPAR settings are saved, install the server with ESXi and configured with a management address.

**Note:** Dell EMC strongly recommends that you update Broadcom firmware to the latest version. Check/update the firmware version using life cycle controller.

Use the vSphere client application to configure and manage ESXi server. As shown in Figure 6, create virtual switches and map them to virtual NICs according to Figure 2. Create the vSwitch by navigating to *Inventory->hosts->configuration->networking->add networking*.

vSwitch helps to segregate traffic using VLAN IDs. When creating vSwitch, the connection type can be either vmPort or vmKernel port based on the need. The vmKernel port is recommended for iSCSI traffic, vMotion, and TCP/IP traffic. The vmPort is recommended for VM data traffic. In Figure 6, NIC partition 1 is used for the management network, partition 2 is used for vMotion, partition 3 is used for the VM network, and partition 4 is used for iSCSI traffic. Another thing to notice is that each vSwitch has connections from both physical ports to prevent port failure. There can also be adapter/hardware level redundancy by providing more adapter ports to vSwitch.

View: vSphere Standard Switch

## Networking

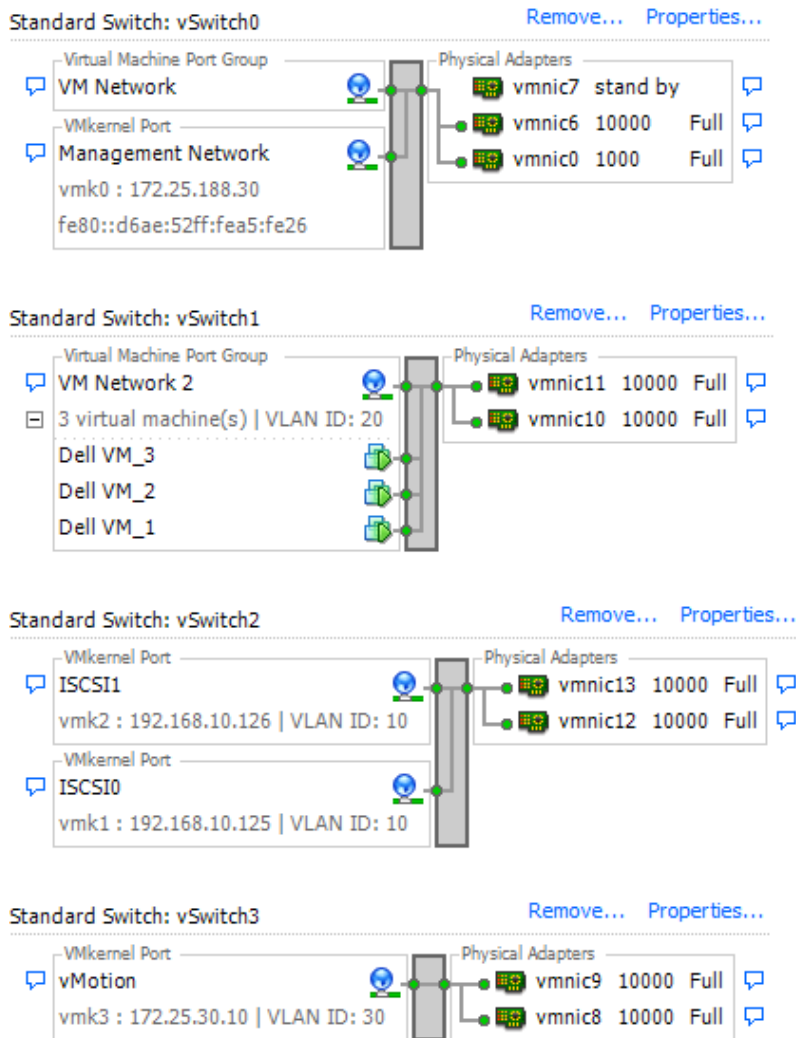


Figure 6 vSwitch view of R720 ESXi host

### 2.1.1 Storage adaptor configuration

In the virtualization world, use either hardware adapter-based iSCSI initiator or software iSCSI initiators. Figure 7 shows how to configure a hardware-based iSCSI initiator.

To bind vmnic (virtual NIC) navigate to:

*Inventory->hosts->configuration->storage adaptor->select adaptor->click properties->add vmnic/ip add*

Once the vmnic and IP address are added, the status turns active and green. For this setup, the IP address is in the same subnet as the storage device. The storage volumes can be discovered once the target (storage) IP address is entered in the Dynamic Discovery tab.

The vSwitch2 has two vmKernel ports and each vmKernel port is mapped to one vmnic. There are no standby connections. This is an active-active design ready for multipathing.

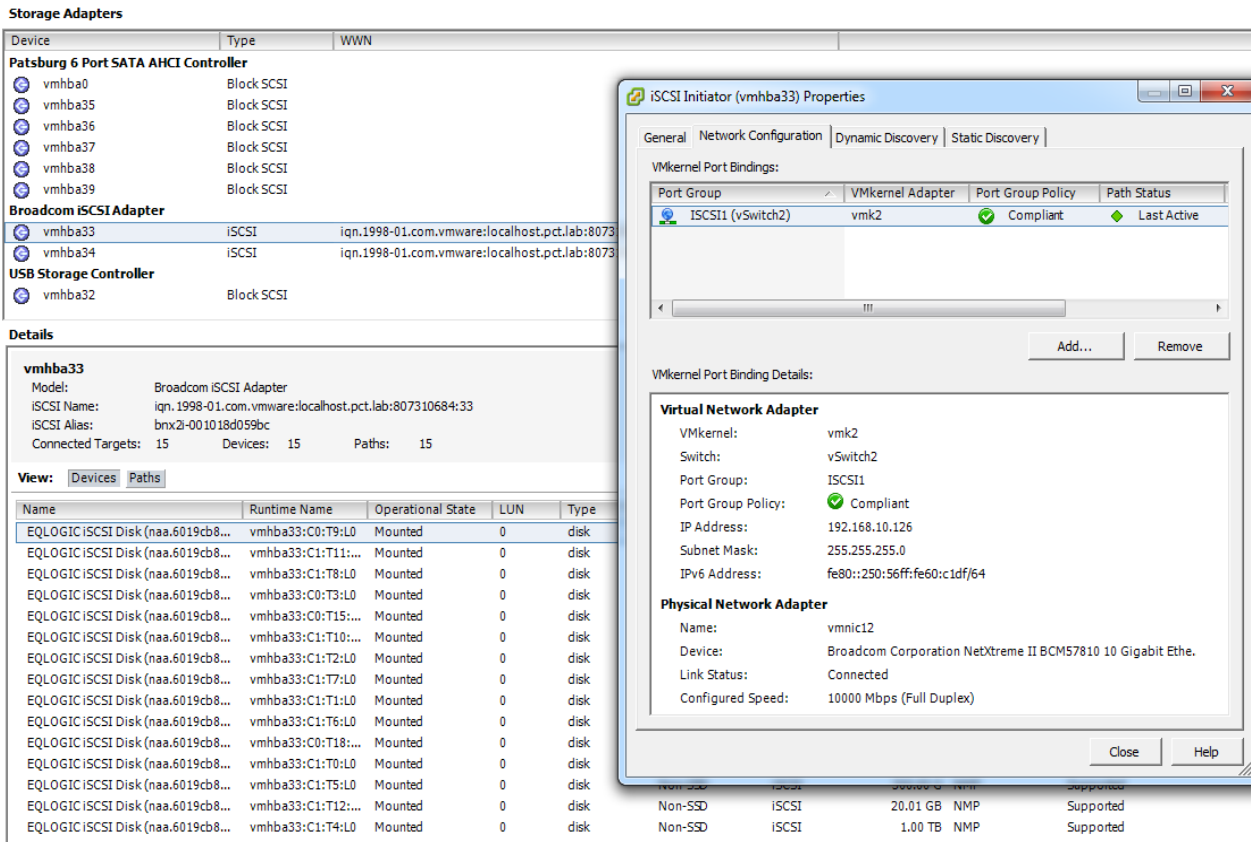


Figure 7 Binding hardware iSCSI adapter

2.1.2 EqualLogic MEM — MPIO configuration

The following method achieves the best results when using EqualLogic storage. To populate a Dell\_EQL path, install the EqualLogic multipath extension module (MEM). For installation procedures, refer to <https://support.equallogic.com/secure/login.aspx>

Figure 8 shows the MPIO setup after installing EQL MEM and taking all defaults.

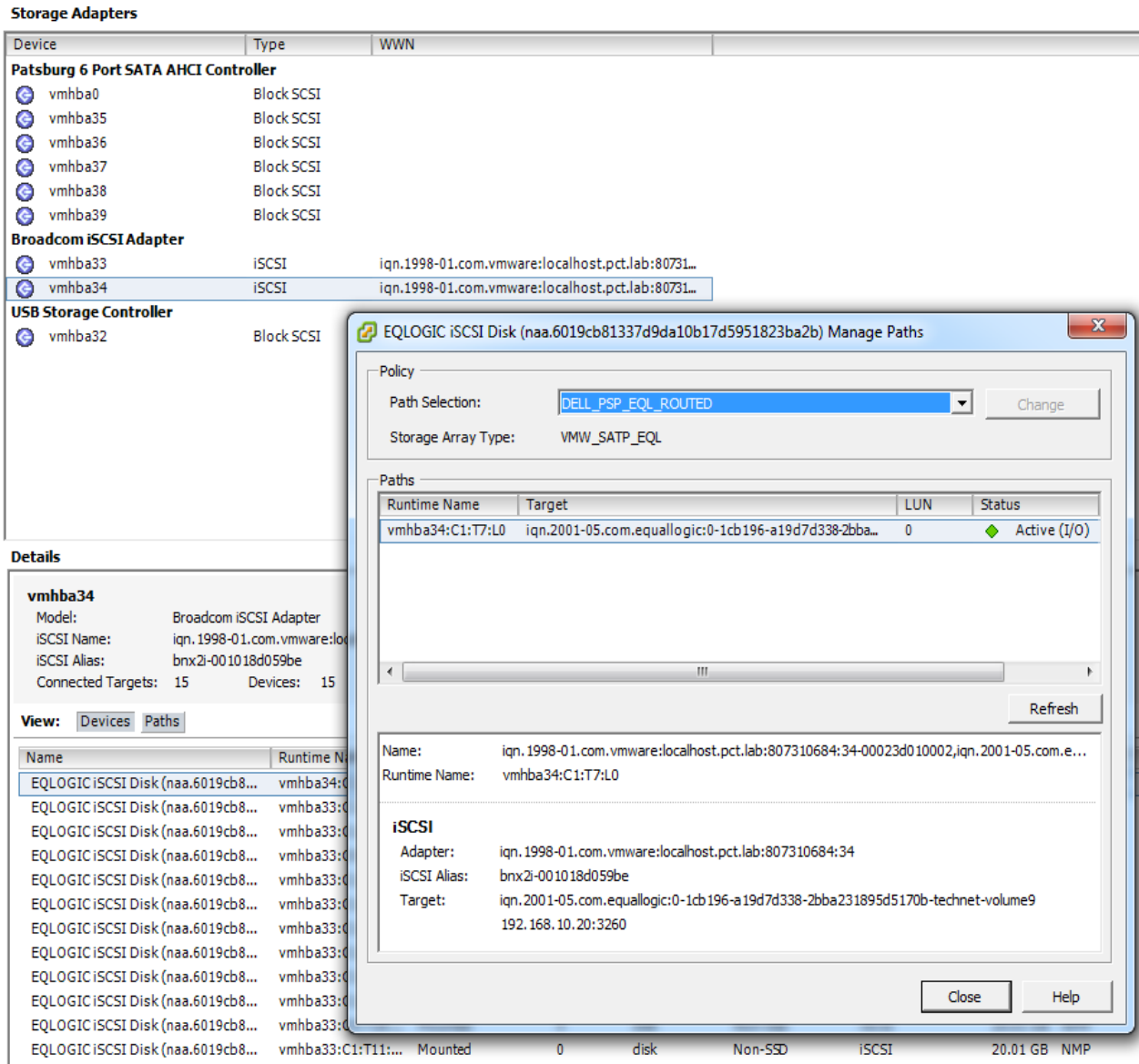


Figure 8 Enabling EQL MEM

## 2.2 Windows Server configuration

Dell PowerEdge R620 servers with Intel X520 PCIe cards or Broadcom 57810 PCIe cards are chosen to install Windows Server 2012. Once the Windows OS installation is complete, update the Intel/Broadcom drivers/firmware, and enable DCB on the CNA from Windows desktop (network properties or BASA suite) to prepare server for iSCSI solution

### 2.2.1 EqualLogic HIT kit

For best results, download and install the Host Integration Tool (HIT) kit from <https://support.equallogic.com/secure/login.aspx>. The download contains the installation and configuration guides. Upon installation, the system prompts for restart.

After reboot, the *Dell EqualLogic-MPIO* tab appears in the iSCSI initiator (see Figure 13). To open the iSCSI initiator use:

```
start menu->administrator tools->iSCSI initiator
```

### 2.2.2 MPIO configuration

Open remote setup wizard by clicking Start->Remote Setup Wizard->configure MPIO settings

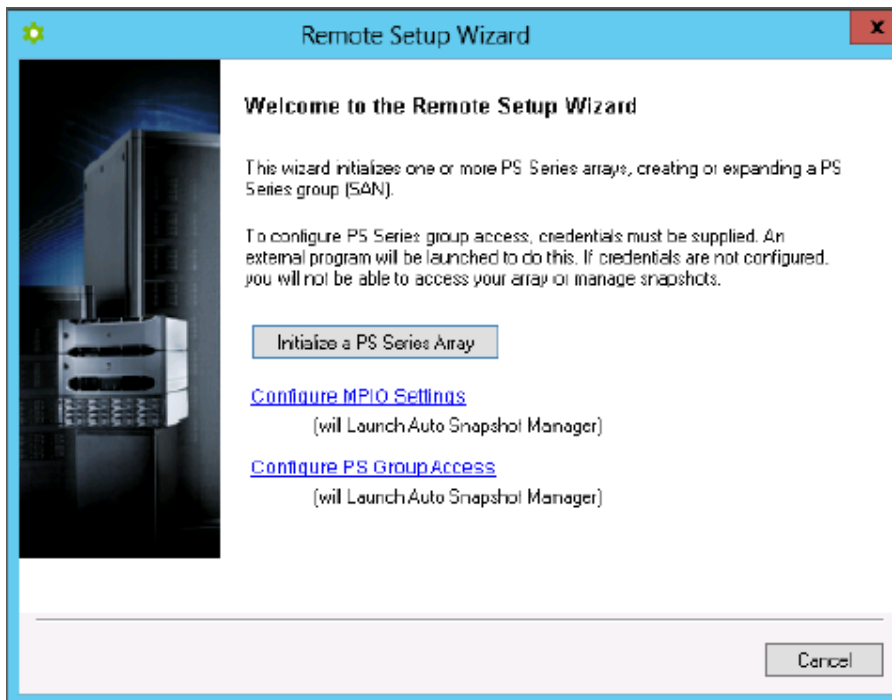


Figure 9 Remote Setup Wizard

When “*Configure MPIO settings*” is clicked in the previous screen, Auto-snapshot Manager is loaded. From the MPIO settings screen, exclude any network not used for iSCSI SAN traffic and enable the *Least Queue Depth* (see Figure 10).

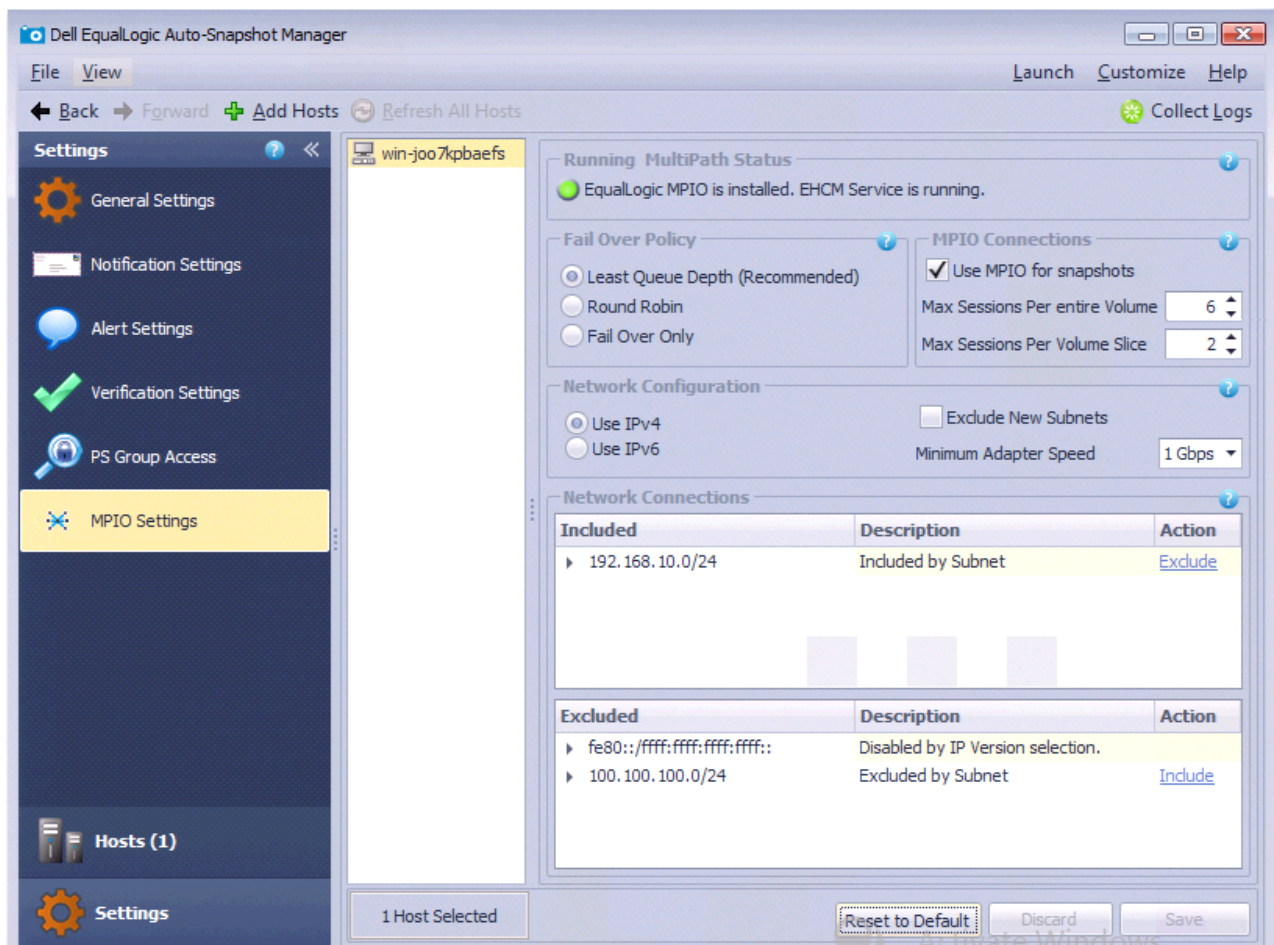


Figure 10 Enabling iSCSI for one of four virtual NICs



As Figure 11 shows, access to storage is achieved by adding a target IP address using:

*iSCSI initiator->Discovery tab->Discover portal ->target ip address*

**Note:** Leave port address (3260) & adapter address to default. EqualLogic MPIO will distribute the traffic load between the iSCSI interfaces.

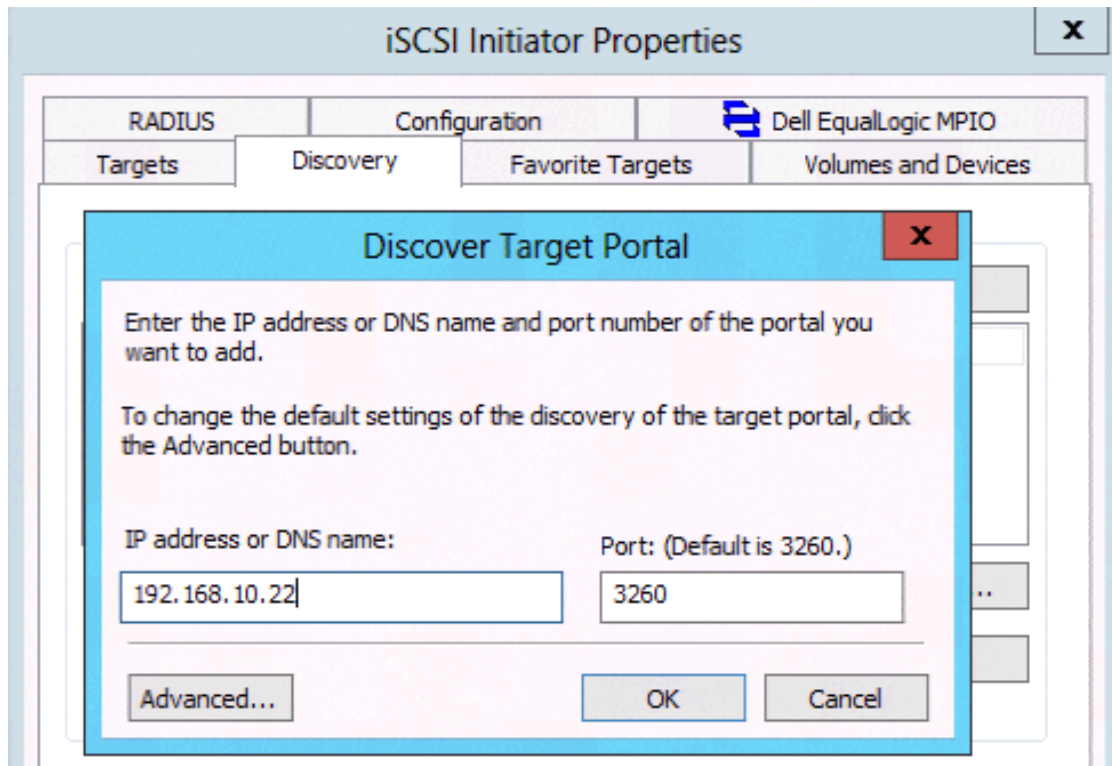


Figure 11 Connecting to storage/target

After target discovery is made, connect the volumes, as Figure 12 shows.

iSCSI initiator->Targets tab->refresh to scan iSCSI target->select volume->click connect

This enables multipath each time a connection is made to a volume.

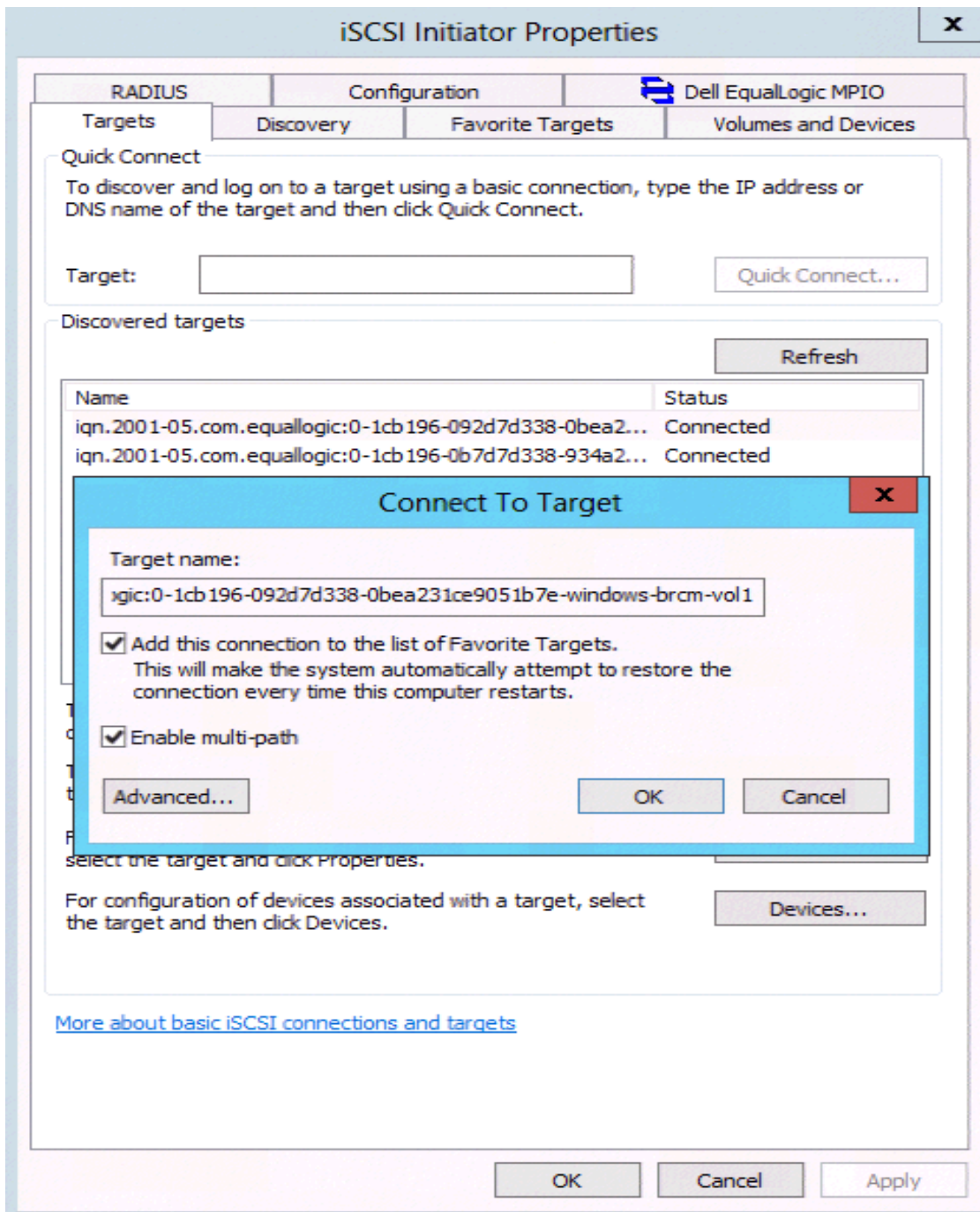


Figure 12 Connecting to volumes

After connecting to volumes, click on the DellEqualLogic MPIO tab. After a few minutes, HIT will have determined and applied the optimal connections needed.

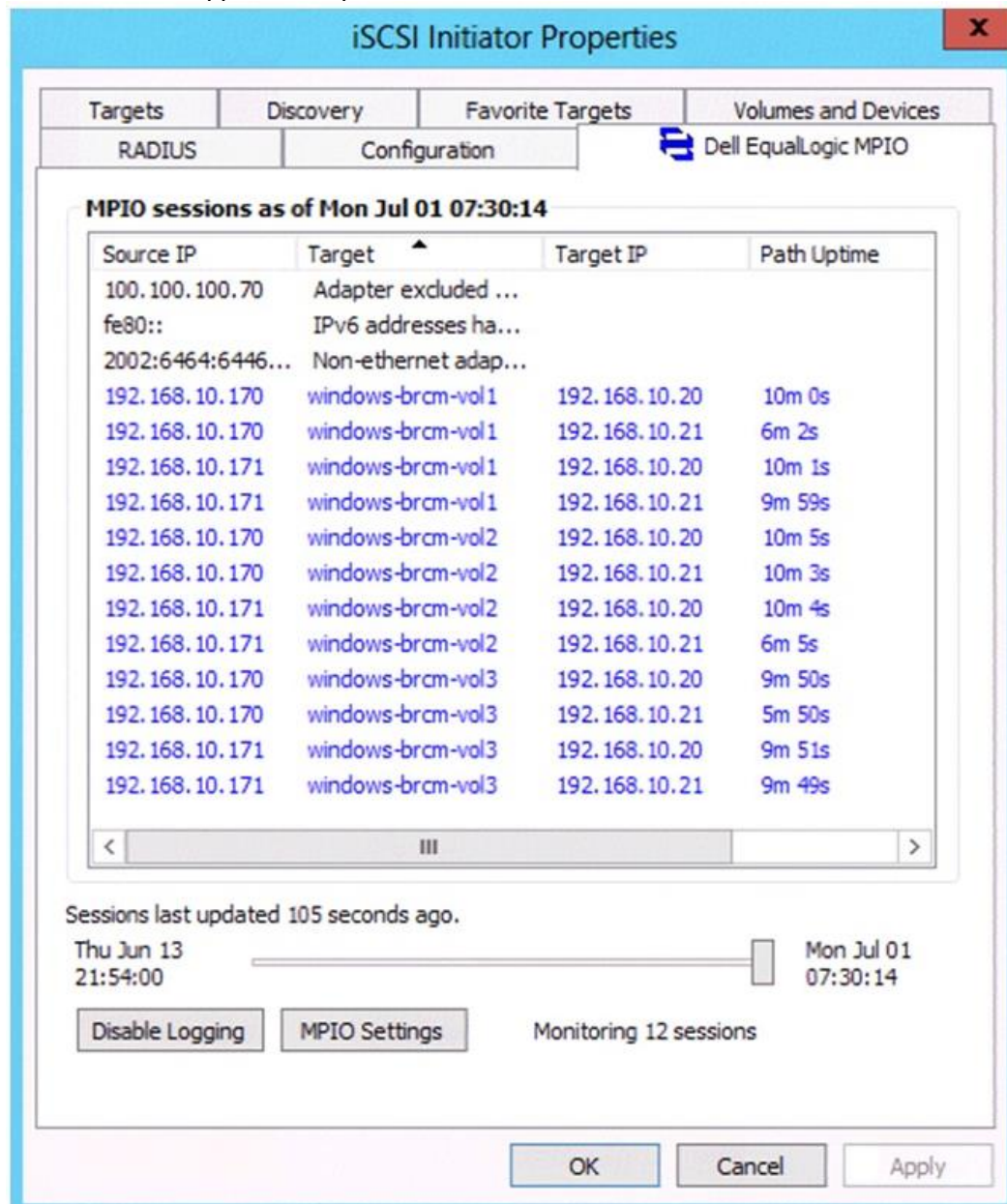


Figure 13 MPIO in action

**Note:** The *Remote Setup Wizard* (Figure 9) also provides a tool to configure access to storage by selecting *Configure PS Group access*.

Once connections to storage are achieved, LUN/Volumes show up under:

*server manager->volumes->disks*

Bring each volume online, initialize, and assign a volume name. Once this is done, it would look similar to that shown in Figure 14.

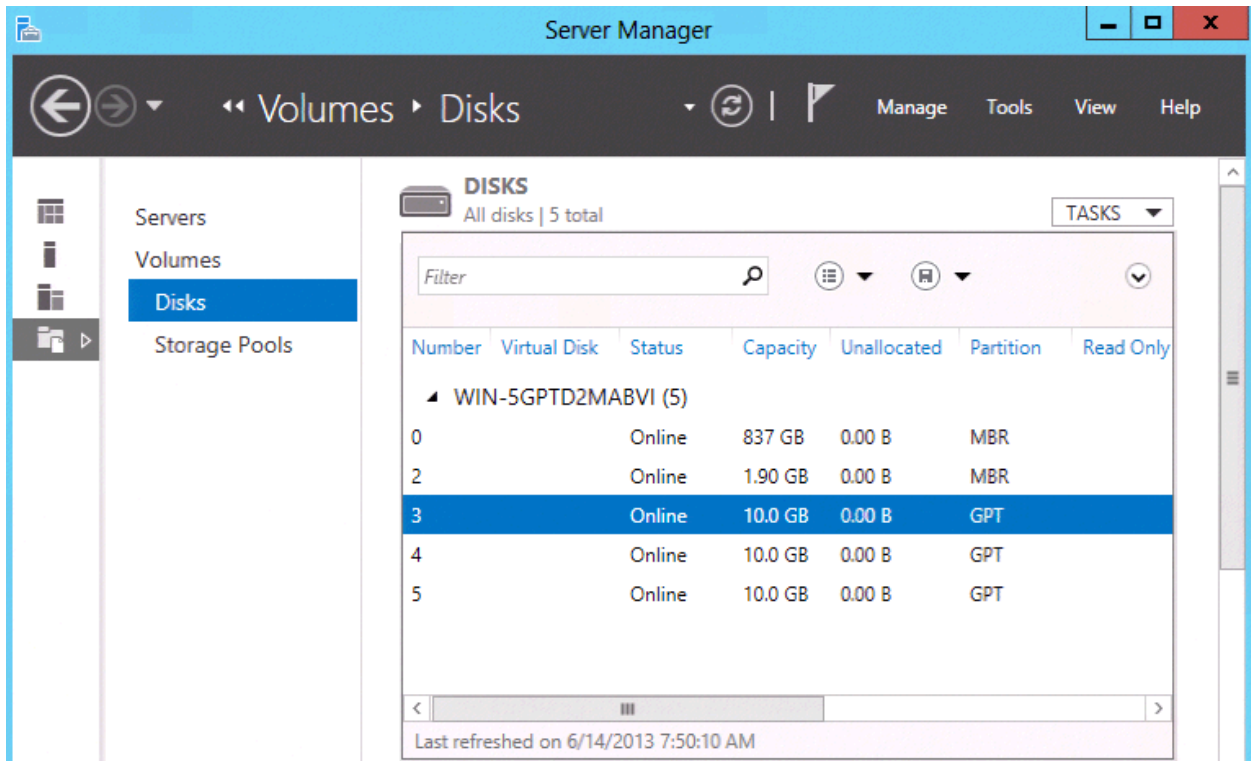


Figure 14 Bring LUN online

This may also be verified from the storage GUI, showing that connections to target are in place (Figure 15). Refer to the Storage section on how to create and access volumes.

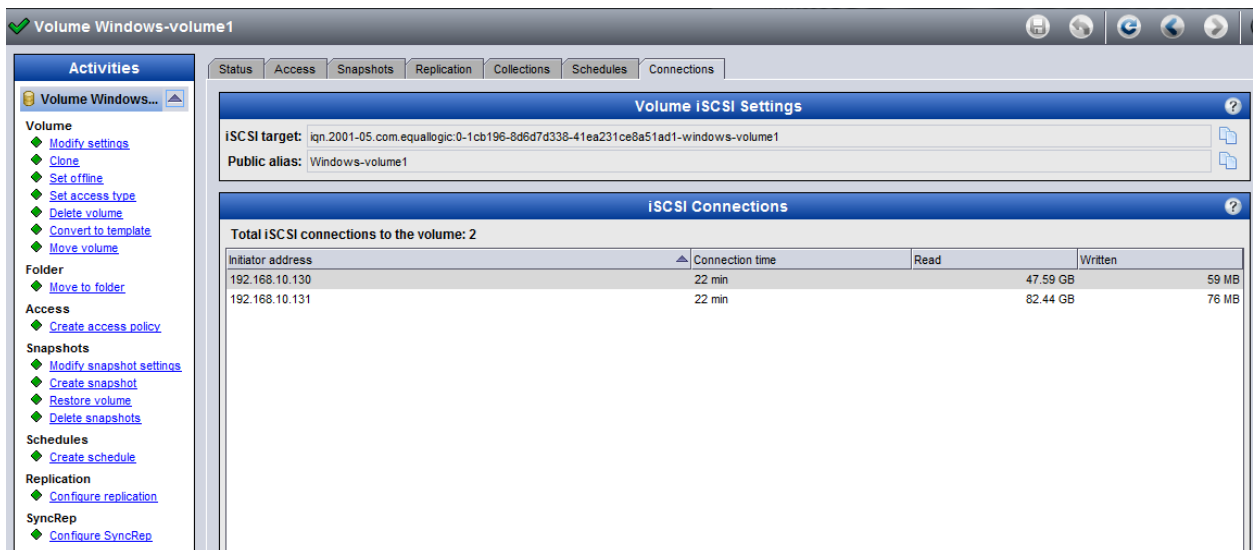


Figure 15 Verify initiator connections to target

Next step is to test end-to-end iSCSI by passing some read/write traffic.

Figure 16 shows the snapshot of live iSCSI traffic using IOMeter (available as a free download).

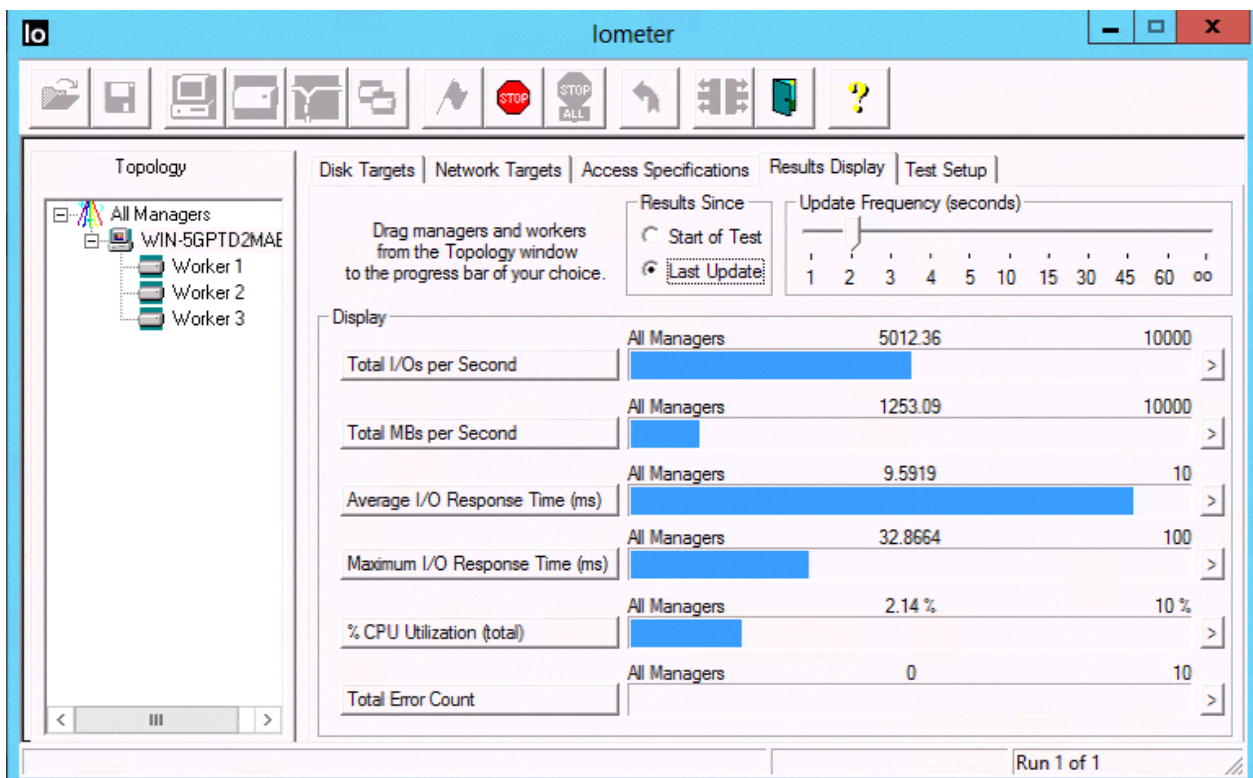


Figure 16 IOMeter MPIO in action



### 2.2.3 NIC Teaming - LAN load balancing

Similar to MPIO (multi-path IO) used for iSCSI traffic load balancing; use NIC teaming for active-active LAN traffic load balancing.

**Note:** Intel limits the use of MPIO and NIC teaming on the same single adapter. Because the experiment's prime goal is an iSCSI solution, Windows Server with Intel was chosen for MPIO only and Windows Server with Broadcom 57810s for both MPIO and Teaming.

As shown in Figure 17, BACS (Broadcom Advanced Control Suite) is used manage Broadcom adapters. By enabling Ethernet (NDIS) and iSCSI offload, separate adapters show up under the physical adapter for Ethernet and iSCSI functions, respectively. This enabling might require a system reboot.

```
Broadcom advanced control suite-> filter->All view->select-adapter->
>configure-tab->resource
```

VLAN 10 and IP address 192.168.10.X is dedicated for iSCSI traffic. Ethernet adapters are dedicated for LAN traffic with VLAN100.

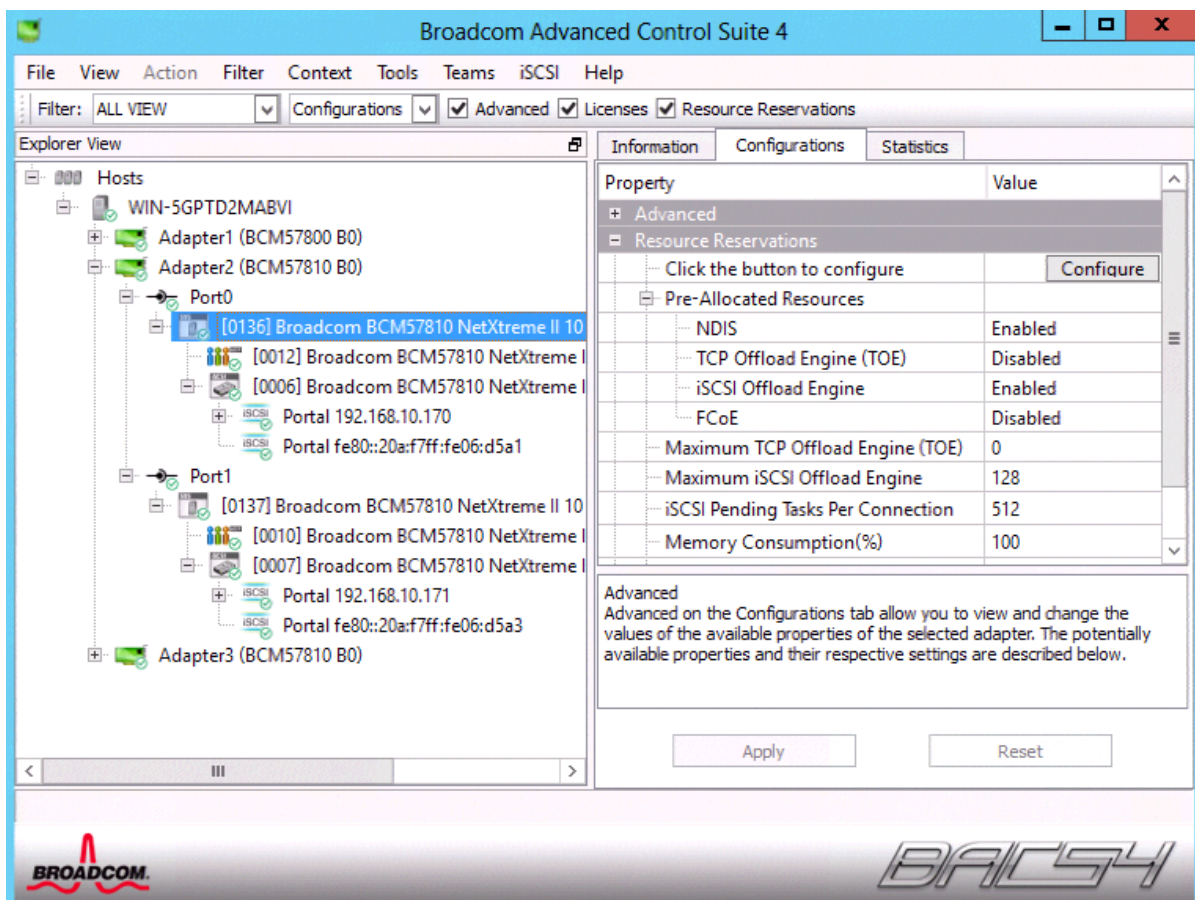


Figure 17 Enabling iSCSI and Ethernet for each port

The Ethernet adapters are combined for load balancing and for increasing bandwidth. NIC teaming helps to combine adapters of one card or multiple cards as shown below in Figure 18:

*Broadcom advanced control suite-> filter->team view->right click team->create team*

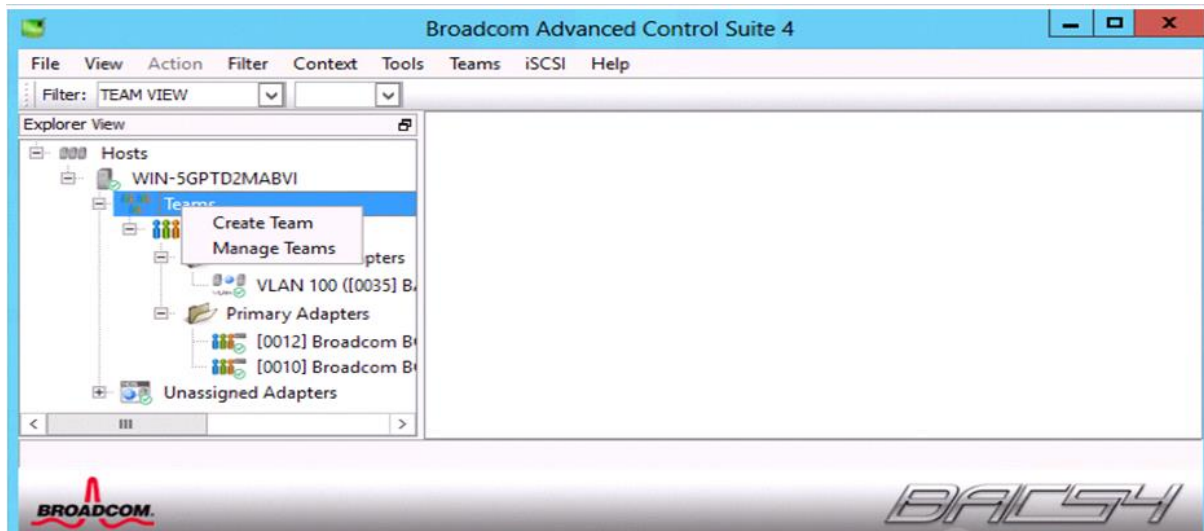


Figure 18 Create team

When creating team, choose one of the options shown in Figure 19.

**Note:** Broadcom 57810s limits the use of the LACP option when used with NPAR or iSCSI offload.

For this test, it is ideal to use LACP. If Broadcom 57810s are used LACP can't be used in conjunction with iSCSI offload, hence the next best option is to use SLB (Smart Load Balancing).

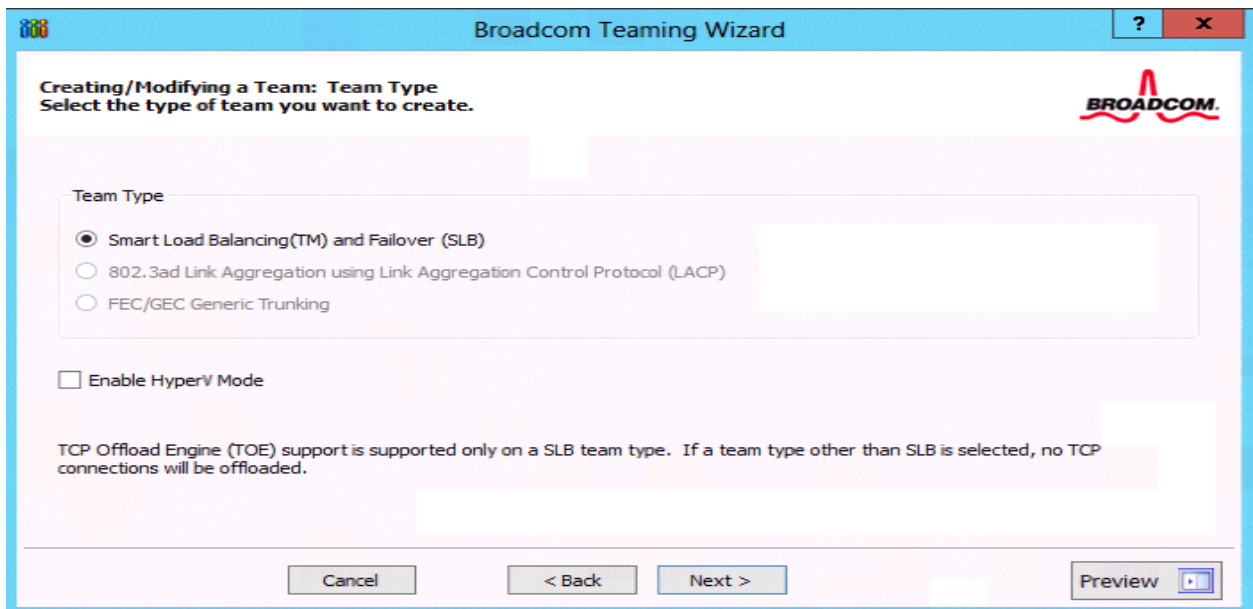


Figure 19 Types of team

The SLB option has default active-active settings. To support active-standby, choose the correct standby member setting. For our test we chose default active-active because we wanted to load balance LAN traffic.

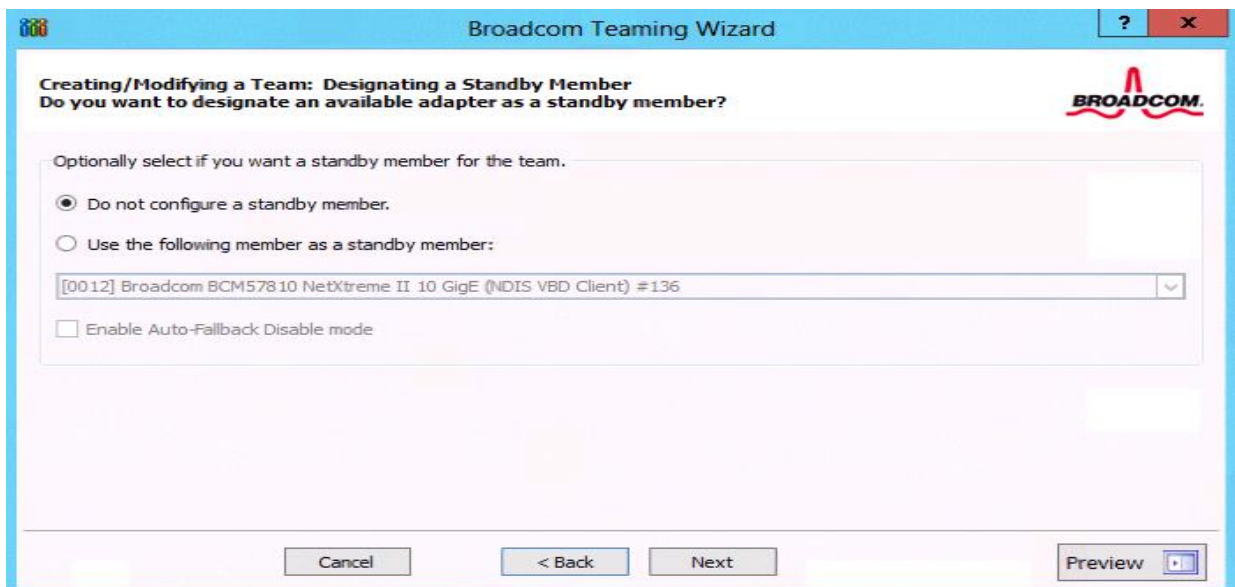


Figure 20 Active-active (no standby) vs active-standby team members



As Figure 21 shows, the VLAN 100 is the VLAN interface for team 1. Both adapters belong to VLAN 100 as well as they are primary adapters. None of adapters are in stand-by, which allows for both adapters to carry traffic and help in load balancing.

**Note:** Make sure both the VLAN interface and adapters that belong to a single team share the same VLAN ID.

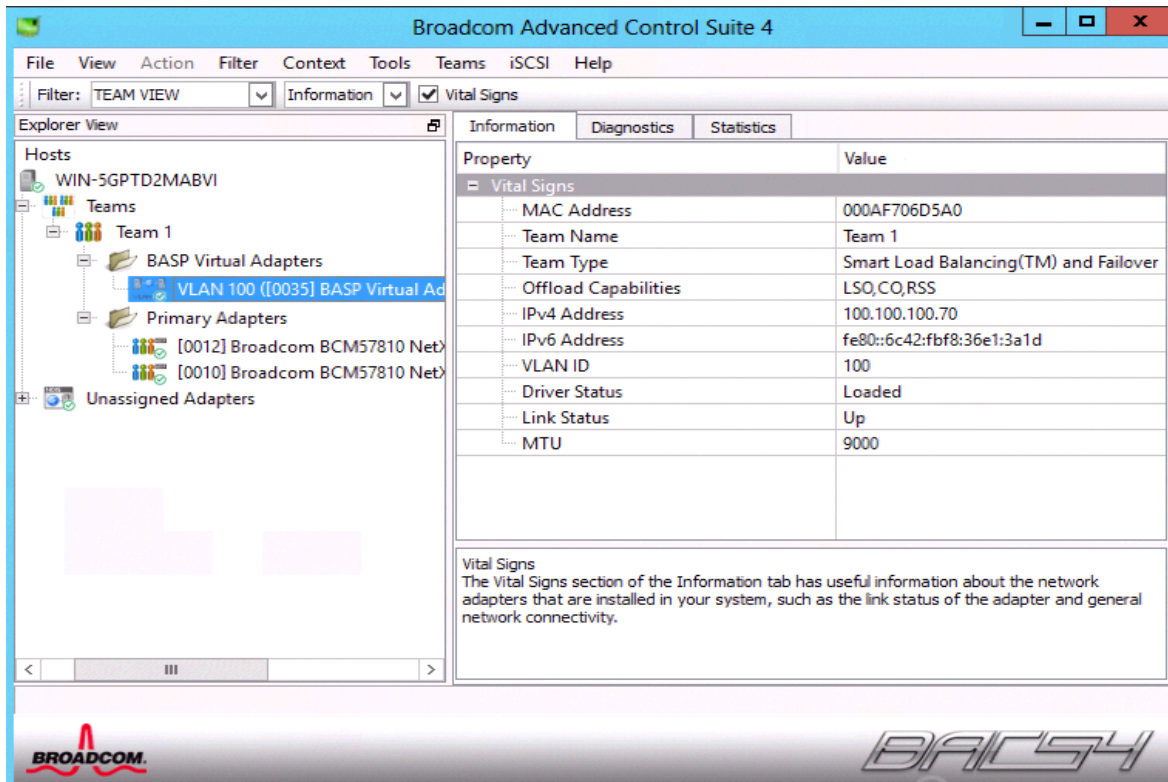


Figure 21 Team with VLAN 100 active-active members.

## 3 Networking

Use this section to setup VLT, DCB, VRRP, and Spanning Tree on the network switches. Then use the chapter on Network Management to setup OpenManage Network Manager (OMNM) in order to monitor and manage your Data Center network.

Refer to the *S4810 Configuration Guide* and the *MXL Configuration Guide* for more information on setting up this scenario and for setting up additional features not covered in this document.

- [Dell S4810 Configuration Guide - Download Now](#)
- [Dell MXL Configuration Guide - Download Now](#)

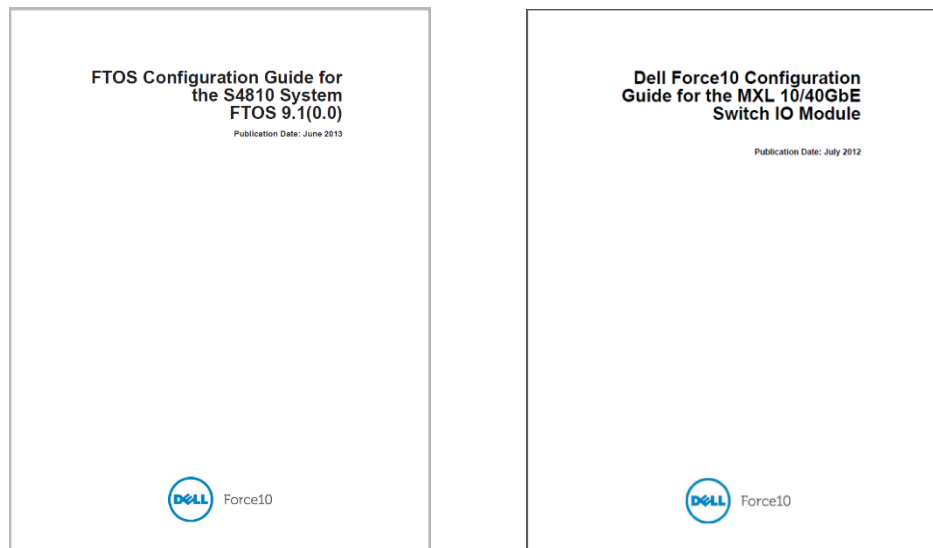


Figure 22 Dell Networking Configuration Guides

### 3.1 Virtual Link Trunking (VLT)

VLT is essential in maximizing the data center's network capacity. Implementing VLT on the S4810s at the aggregation layer can increase bandwidth of the network as much as 50 percent and provides other benefits:

- Allows a single device to use a LAG across two upstream devices
- Provides a loop-free topology
- Uses all available uplink bandwidth
- Provides fast convergence if either the link or a device fails
- Optimizes forwarding with VRRP
- Provides link-level resiliency
- Assures high availability

Check out the three minute video in the following link to see how this technology works:

<http://en.community.dell.com/dell-blogs/direct2dell/b/direct2dell/archive/2013/01/16/vlt-virtual-link-trunking-maximizing-datacenter-capacity-with-dell-force10-switches.aspx>

There are seven basic steps for setting up VLT:

1. Configure RSTP on the switch to prevent loops (if it is to be used).
2. Configure backup-link.
3. Configure a static LAG VLT interconnect (VLTi) on both peers (do not use LACP).
4. Configure the VLT domain.
5. Configure VLT port-channels (can be dynamic (LACP) or static).
6. Configure VLT traps via LLDP to monitor the VLT.
7. Enable VLT.

**Notes:**

1. Do not stack switches that use VLT. VLT is not a stack and is incompatible with stacking.
2. VLT is not supported on the MXL switch as of July, 2013. However, future versions of MXL firmware are expected to have VLT support. In this data center scenario, VLT is configured using S4810 switches.

### 3.1.1 RSTP on VLT

VLT provides loop-free redundant topologies and eliminates the need for the spanning tree protocol. If used under proper guidelines however, RSTP detects other potential issues caused by invalid configurations and minimizes topology changes after any link or node failures. If you include RSTP on the switch, configure it before VLT.

Only configure RSTP under the following guidelines when VLT is enabled:

- Configure any ports at the edge of the spanning tree's operating domain as edge ports, which are directly connected to end stations or rack servers. Ports connected directly to layer-3 only routers not running STP, must have RSTP disabled or be configured as edge ports.
- Make sure that the primary VLT peer is the root bridge and that the secondary VLT peer has the secondary bridge ID in the network. If the primary VLT peer node fails, the secondary VLT peer node becomes the root bridge, avoiding problems with spanning tree port state changes that occur when a VLT node fails or recovers.
- Even with this configuration, if the node has non-VLT ports using RSTP that are not configured as edge ports and are connected to other layer 2 switches, spanning tree topology changes can still be detected after VLT node recovery. To avoid this scenario, make sure that any non-VLT ports are configured as edge ports or have RSTP disabled.

### 3.1.2 VLT topology

In Figure 23, on the left side (rack servers/switches) of our scenario, one VLTi is configured between the ToR switch pair to let the Server NIC teaming view the pair as one logical switch. A second VLTi is then configured between the two aggregation switches to appear as a single logical link to the network core, providing additional multiple paths and load balancing.

This use of a second VLTi is an example of multi VLT. Figure 23 shows multi VLT configured in the topology.

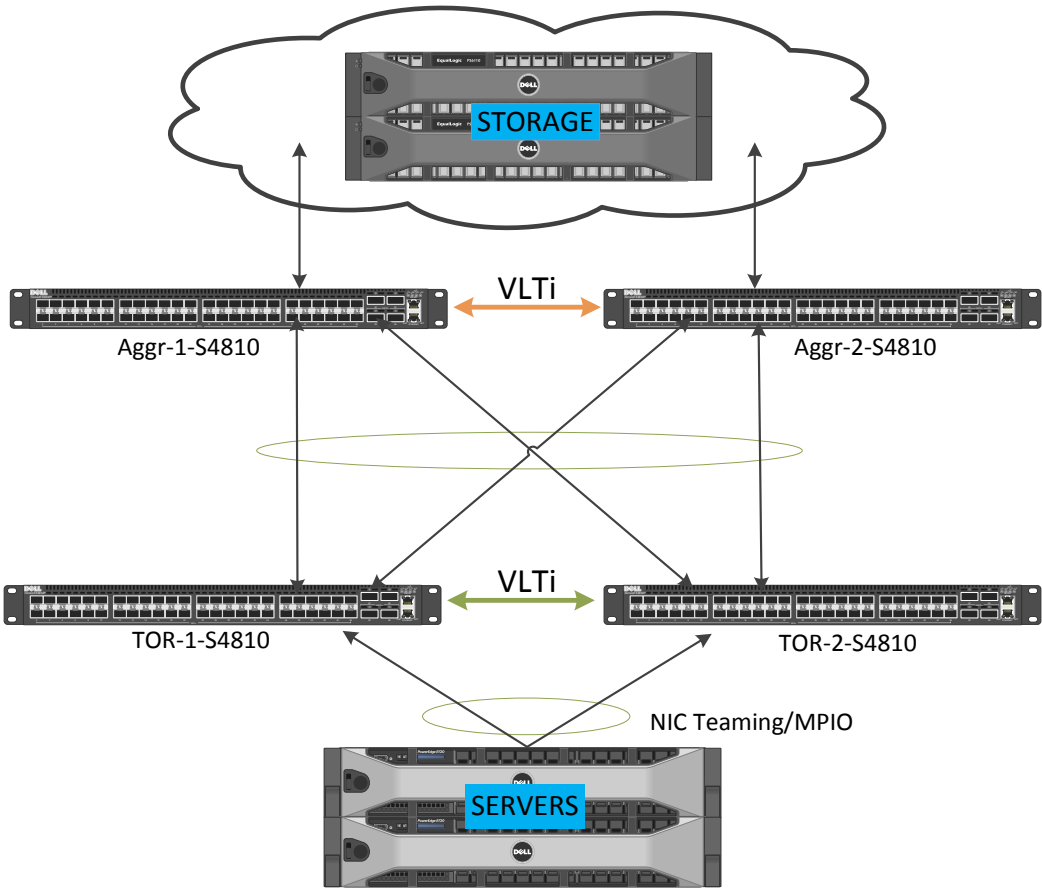


Figure 23 Multi VLT (mVLT) on aggregation and ToR (for rack-mounted servers/switches).

Below are the configuration commands used to create multi VLT (mVLT) used in this scenario.

Table 1 VLT pair configuration at the aggregation layer

| Aggr-1-S4810 (172.25.188.160)  | Aggr-2-S4810 (172.25.188.161)   |
|--|---|
| <pre>configure protocol spanning-tree rstp   no disable   bridge-priority 32768 vlt domain 200   peer-link port-channel 100   back-up destination 172.25.188.161   primary-priority 40000   system-mac mac-address 00:01:e8:11:11:11   unit-id 0 interface range tengigabitethernet 0/36 - 37 ,                   tengigabitethernet 0/40 - 41   no ip address   mtu 12000</pre> | <pre>configure protocol spanning-tree rstp   no disable   bridge-priority 4096 vlt domain 200   peer-link port-channel 100   back-up destination 172.25.188.160   primary-priority 32768   system-mac mac-address 00:01:e8:11:11:11   unit-id 1 interface range tengigabitethernet 0/36 - 37 ,                   tengigabitethernet 0/40 - 41   no ip address   mtu 12000</pre> |

| Aggr-1-S4810 (172.25.188.160)   | Aggr-2-S4810 (172.25.188.161)   |
|---|---|
| <pre> port-channel-protocol LACP   port-channel 3 mode active protocol lldp no shutdown interface fortyGigE 0/60   no ip address   mtu 12000   no shutdown interface Port-channel 3   no ip address   mtu 12000   switchport   vlt-peer-lag port-channel 4   no shutdown interface Port-channel 100   description 40Gb VLT interconnect port 60   no ip address   mtu 12000   channel-member fortyGigE 0/60   no shutdown protocol lldp </pre>                            | <pre> port-channel-protocol LACP   port-channel 4 mode active protocol lldp no shutdown interface fortyGigE 0/60   no ip address   mtu 12000   no shutdown interface Port-channel 4   no ip address   mtu 12000   switchport   vlt-peer-lag port-channel 3   no shutdown interface Port-channel 100   description 40Gb VLT interconnect port 60   no ip address   mtu 12000   channel-member fortyGigE 0/60   no shutdown protocol lldp </pre>                          |
| <p>To confirm the configuration, type <b>show vlt brief</b></p> <p>Look for all three statuses (ICL Link, HeartBeat, and VLT Peer) to show “Up” to confirm the VLT.</p>   | <p>To confirm the configuration, type <b>show vlt brief</b></p> <p>Look for all three statuses (ICL Link, HeartBeat, and VLT Peer) to show “Up” to confirm the VLT.</p>   |
| <pre> Aggr-1-S4810#show vlt brief VLT Domain Brief ----- Domain ID:           200 Role:                Secondary Role Priority:       40000 ICL Link Status:     Up HeartBeat Status:    Up VLT Peer Status:     Up Local Unit Id:       0 Version:             5(1) Local System MAC address: 00:01:e8:8b:3b:6f Remote System MAC address: 00:01:e8:8b:3b:81 Configured System MAC: 00:01:e8:11:11:11 Remote system version: 5(1) Delay-Restore timer: 90 seconds </pre> | <pre> Aggr-2-S4810#show vlt brief VLT Domain Brief ----- Domain ID:           200 Role:                Primary Role Priority:       32768 ICL Link Status:     Up HeartBeat Status:    Up VLT Peer Status:     Up Local Unit Id:       1 Version:             5(1) Local System MAC address: 00:01:e8:8b:3b:81 Remote System MAC address: 00:01:e8:8b:3b:6f Configured System MAC: 00:01:e8:11:11:11 Remote system version: 5(1) Delay-Restore timer: 90 seconds </pre> |

Table 2 VLT pair configuration at the top-of-rack

| <b>TOR-1-S4810 (172.25.188.162)</b>  | <b>TOR-2-S4810 (172.25.188.163)</b>  |
|--|--|
| <pre> configure protocol spanning-tree rstp   no disable   bridge-priority 57344 vlt domain 300   peer-link port-channel 100   back-up destination 172.25.188.163   primary-priority 32768   system-mac mac-address 00:01:e8:22:22:22   unit-id 1 stack-unit 0 port 52 portmode quad interface range TenGigabitEthernet 0/52 - 55   no ip address   mtu 12000   port-channel-protocol LACP   port-channel 1 mode active   protocol lldp   advertise management-tlv system-name   no advertise dcbx-tlv ets-reco   dcbx port-role auto-upstream   no shutdown interface fortyGigE 0/60   no ip address   mtu 12000   no shutdown interface Port-channel 1   no ip address   mtu 12000   switchport   vlt-peer-lag port-channel 2   no shutdown interface Port-channel 100   description 40Gb VLT inter port 60   no ip address   mtu 12000   channel-member fortyGigE 0/60   no shutdown protocol lldp  To confirm the configuration, type <b>show vlt brief</b> Look for all three statuses (ICL Link, HeartBeat, and VLT Peer) to show "Up" to confirm the VLT.  TOR-1-S4810#show vlt brief VLT Domain Brief ----- Domain ID:          300 Role:               Secondary Role Priority:       10000 ICL Link Status:    Up </pre> | <pre> configure protocol spanning-tree rstp   no disable   bridge-priority 61440 vlt domain 300   peer-link port-channel 100   back-up destination 172.25.188.162   primary-priority 1   system-mac mac-address 00:01:e8:22:22:22   unit-id 0 stack-unit 0 port 52 portmode quad interface range TenGigabitEthernet 0/52 - 55   no ip address   mtu 12000   port-channel-protocol LACP   port-channel 2 mode active   protocol lldp   advertise management-tlv system-name   no advertise dcbx-tlv ets-reco   dcbx port-role auto-upstream   no shutdown interface fortyGigE 0/60   no ip address   mtu 12000   no shutdown interface Port-channel 2   no ip address   mtu 12000   switchport   vlt-peer-lag port-channel 1   no shutdown interface Port-channel 100   description 40Gb VLT inter port 60   no ip address   mtu 12000   channel-member fortyGigE 0/60   no shutdown protocol lldp  To confirm the configuration, type <b>show vlt brief</b> Look for all three statuses (ICL Link, HeartBeat, and VLT Peer) to show "Up" to confirm the VLT.  TOR-1-S4810#show vlt brief VLT Domain Brief ----- Domain ID:          300 Role:               Primary Role Priority:       1 ICL Link Status:    Up </pre> |

| <b>TOR-1-S4810 (172.25.188.162)</b>          | <b>TOR-2-S4810 (172.25.188.163)</b>          |
|--|--|
| HeartBeat Status: Up                         | HeartBeat Status: Up                         |
| VLT Peer Status: Up                          | VLT Peer Status: Up                          |
| Local Unit Id: 1                             | Local Unit Id: 0                             |
| Version: 5(1)                                | Version: 5(1)                                |
| Local System MAC address: 00:01:e8:8b:36:0e  | Local System MAC address: 00:01:e8:8b:32:48  |
| Remote System MAC address: 00:01:e8:8b:32:48 | Remote System MAC address: 00:01:e8:8b:36:0e |
| Configured System MAC: 00:01:e8:22:22:22     | Configured System MAC: 00:01:e8:22:22:22     |
| Remote system version: 5(1)                  | Remote system version: 5(1)                  |
| Delay-Restore timer: 90 seconds              | Delay-Restore timer: 90 seconds              |

For the right side of our scenario (using an M1000e chassis with modular servers/switches), only one VLT interlink is configured, which is located between the aggregation switches.

The ToR switches are MXL installed in the M1000e chassis and do not support VLT. Figure 24 depicts the right side of the topology.

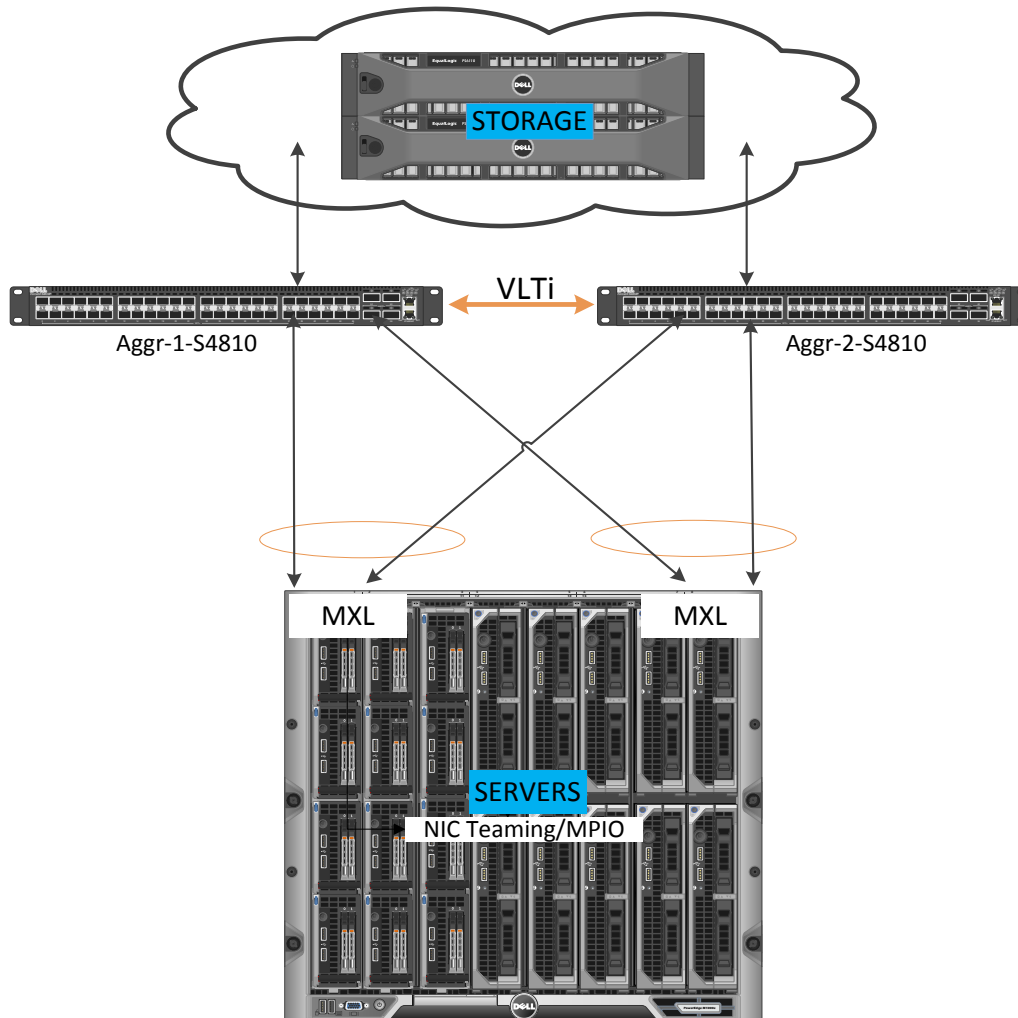


Figure 24 VLT at aggregation level only. Using M1000e modular servers and MXL switches.

### 3.1.3 Backup Link

A *backup link* should also be used to maintain heartbeat messages across an out-of-band management network. The backup link ensures that node failure conditions are correctly detected and are not confused with failures of the chassis interconnect trunk. VLT ensures that local traffic on a chassis does not traverse the chassis interconnect trunk and takes the shortest path to the destination using directly attached links.

### 3.1.4 Monitoring VLT

The ability to monitor VLT is provided using OMNM and SNMP traps. Bandwidth, for example, can be monitored by enabling VLT traps from the S4810 command line using the following command:

```
TOR-1-S4810 (conf) #snmp-server enable traps vlt
```



When bandwidth usage of the VLTi (ICL) exceeds 80 percent, the following SNMP trap (along with a syslog error message) is generated:

```

VLTi Bandwidth Usage Exceeding Threshold Value Error:  %STKUNIT0-M:CP
%VLTMGR-6-VLT-LAG-ICL: Overall Bandwidth utilization of VLT-ICL-LAG (port-
channel 25)crosses threshold. Bandwidth usage (80%)

```

When the bandwidth usage *drops below* the 80 percent threshold, the system generates another SNMP trap (and syslog message) to inform you that all is back to normal. In both cases, OMNM can capture these traps and inform the network administrator. See the OMNM section below for more information on monitoring the network.

For additional details on Virtual Link Trunking, consult the *S4810 Configuration Guide* chapter on Virtual Link Trunking.

## 3.2 VRRP

Virtual Router Redundancy Protocol (VRRP) is a protocol designed to eliminate a single point of failure in a routed network. Through the use of a redundant (second) Layer-3 switch, a backup/failover configuration is created. In this example, VRRP is configured on the aggregation layer switches (Aggr-1-S4810 and Aggr-1-S4810) to provide for L3 redundancy in the event of a single switch or link failure. Virtual configured from the previous section, the additional switch and cabling hardware required for VRRP are already installed in this network topology (see Figure 23 and Figure 24). Only the software configuration is needed using the CLI commands shown here.

Table 3 Software configuration CLI commands

| Aggr-1-S4810 (172.25.188.160)  | Aggr-2-S4810 (172.25.188.161)  |
|--|--|
| <pre> interface Vlan 10 description iSCSI-vlan ip address 192.168.10.150/24 mtu 12000 tagged TenGigabitEthernet 0/12-13 tagged Port-channel 3,20,30 vrrp-group 10 priority 100 virtual-address 192.168.10.153 no shutdown interface Vlan 20 description vm_network ip address 20.20.20.2/24 mtu 12000 tagged Port-channel 3 vrrp-group 20 priority 100 virtual-address 20.20.20.20 no shutdown  show vrrp 10 brief Interface Group Pri Pre State Master addr... ----- </pre> | <pre> interface Vlan 10 description iSCSI-vlan ip address 192.168.10.151/24 mtu 12000 tagged TenGigabitEthernet 0/12-13 tagged Port-channel 4,21,31 vrrp-group 10 priority 254 virtual-address 192.168.10.153 no shutdown interface Vlan 20 description vm_network ip address 20.20.20.3/24 mtu 12000 tagged Port-channel 4 vrrp-group 20 priority 254 virtual-address 20.20.20.20 no shutdown  show vrrp 10 brief Interface Group Pri Pre State Master addr... ----- </pre> |

| Aggr-1-S4810 (172.25.188.160)  | Aggr-2-S4810 (172.25.188.161)  |
|--|--|
| Vl 10 IPv4 10 100 Y Master 192.168.10.151...<br><br>show vrrp 10<br>-----<br>Vlan 10, IPv4 VRID: 10, Version: 2, Net: 192.168.10.150<br>State: Master, Priority: 100, Master:192.168.10.150<br>Hold Down: 0 sec, Preempt: TRUE, AdvInt: 1 sec<br>Adv rcvd: 2881570, Bad pkts rcvd: 0, Adv sent: 0,<br>Gratuitous ARP sent: 0<br>Virtual MAC address: 00:00:5e:00:01:0a<br>Virtual IP address: 192.168.10.153<br>Authentication: (none) | Vl 10 IPv4 10 254 Y Backup 192.168.10.151...<br><br>show vrrp 10<br>-----<br>Vlan 10, IPv4 VRID: 10, Version: 2, Net: 192.168.10.151<br>State: Backup, Priority: 254, Master:192.168.10.150<br>Hold Down: 0 sec, Preempt: TRUE, AdvInt: 1 sec<br>Adv rcvd: 604, Bad pkts rcvd: 0, Adv sent: 3800892,<br>Gratuitous ARP sent: 1<br>Virtual MAC address: 00:00:5e:00:01:0a<br>Virtual IP address: 192.168.10.153<br>Authentication: (none) |

### 3.3 Enabling data center bridging (DCB)

Enable DCB across both aggregation switches and both ToR switches in this scenario to accomplish end-to-end iSCSI. DCB technologies (available only on DCB-capable equipment) enable each switch-port and each network device-port in the converged network to simultaneously carry multiple traffic classes (iSCSI and non-iSCSI), while guaranteeing performance and QoS. To enable DCB, you must enable either the iSCSI Optimization configuration or the FCoE configuration. DCB is not supported if you enable link-level flow control on one or more interfaces

#### Notes:

1. Not all equipment can be DCB enabled. Make sure your equipment has the correct capabilities for DCB. All equipment mentioned in this document can be DCB enabled.
2. For switch classes used in this document, the minimum firmware version required to support DCB is version 8.3.12.x.

There are four areas in DCB to consider:

- Enable DCB to do end-to-end iSCSI on a converged network.
- Use ETS settings to assign bandwidth limits to each traffic class. This gives preference to one class over another during periods of contention between traffic types.
- Configure Priority Flow Control (PFC) for iSCSI to guarantee lossless iSCSI traffic.
- Use the Broadcom Converged Network Adapters (CNAs) to support DCB and DCBX.

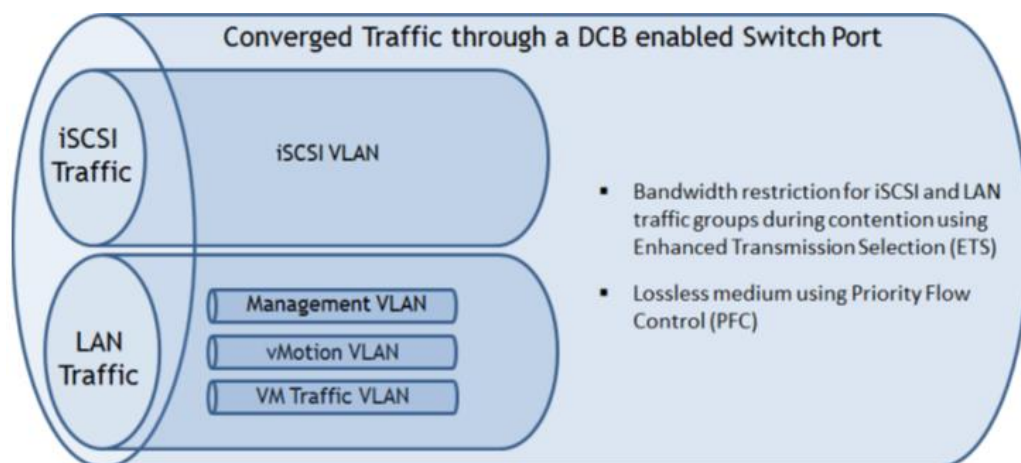


Figure 25 Conceptual view of converged traffic using DCB

The CLI commands to enable DCB on the Dell Networking switches are the same for both ToR switches, and the same for both aggregation switches. The ToR and aggregation commands differ slightly.

To enable DCB for the aggregation (Aggr) and top-of-rack (ToR) switches, enter the following:

Table 4 DCB cli commands for Aggregation and ToR switches

| Aggregation switches (S4810)   | ToR switches (S4810 and MXL)   |
|--|--|
| <pre> configure dcb enable interface TenGigabitEthernet 0/24   no ip address   mtu 12000   dcb-policy input pfc   dcb-policy output ets   port-channel-protocol LACP     port-channel 30 mode active   no shutdown interface TenGigabitEthernet 0/25   no ip address   mtu 12000   dcb-policy input pfc   dcb-policy output ets   port-channel-protocol LACP     port-channel 30 mode active   protocol lldp   no shutdown interface TenGigabitEthernet 0/36   no ip address   mtu 12000   dcb-policy input pfc   dcb-policy output ets   port-channel-protocol LACP     port-channel 3 mode active   protocol lldp   no shutdown </pre> | <pre> configure dcb enable interface TenGigabitEthernet 0/0   no ip address   mtu 12000   switchport   spanning-tree rstp edge-port   protocol lldp     dcbx port-role auto-downstream   no shutdown interface TenGigabitEthernet 0/4   no ip address   mtu 12000   switchport   spanning-tree rstp edge-port   rate-interval 5   protocol lldp     dcbx port-role auto-downstream   no shutdown interface TenGigabitEthernet 0/52   no ip address   mtu 12000   port-channel-protocol LACP     port-channel 1 mode active   protocol lldp     advertise management-tlv system-name     no advertise dcbx-tlv ets-reco     dcbx port-role auto-upstream </pre> |

| Aggregation switches (S4810)  | TOR switches (S4810 and MXL)  |
|---|---|
| <pre> interface TenGigabitEthernet 0/37   no ip address   mtu 12000   dcb-policy input pfc   dcb-policy output ets   port-channel-protocol LACP     port-channel 3 mode active   protocol lldp   no shutdown interface TenGigabitEthernet 0/40   no ip address   mtu 12000   dcb-policy input pfc   dcb-policy output ets   port-channel-protocol LACP     port-channel 3 mode active   protocol lldp   no shutdown interface TenGigabitEthernet 0/41   no ip address   mtu 12000   dcb-policy input pfc   dcb-policy output ets   port-channel-protocol LACP     port-channel 3 mode active   protocol lldp   no shutdown service-class dynamic dot1p qos-policy-output iscsi-policy ets   bandwidth-percentage 30 qos-policy-output other-policy ets   bandwidth-percentage 70 dcb-input pfc   pfc priority 4 priority-group iSCSI   priority-list 4   set-pgid 1 priority-group other   priority-list 0-3,5-7   set-pgid 2 dcb-output ets   priority-group iSCSI qos-policy iscsi-policy   priority-group other qos-policy other-policy </pre> | <pre> no shutdown interface TenGigabitEthernet 0/53   no ip address   mtu 12000   port-channel-protocol LACP     port-channel 1 mode active   protocol lldp     advertise management-tlv system-name     no advertise dcbx-tlv ets-reco     dcbx port-role auto-upstream   no shutdown interface TenGigabitEthernet 0/54   no ip address   mtu 12000   port-channel-protocol LACP     port-channel 1 mode active   protocol lldp     advertise management-tlv system-name     no advertise dcbx-tlv ets-reco     dcbx port-role auto-upstream   no shutdown interface TenGigabitEthernet 0/55   no ip address   mtu 12000   port-channel-protocol LACP     port-channel 1 mode active   protocol lldp     advertise management-tlv system-name     no advertise dcbx-tlv ets-reco     dcbx port-role auto-upstream   no shutdown service-class dynamic dot1p </pre> |

QoS settings should be carefully planned and adjusted to balance the bandwidth and performance needs of each traffic class. It is often recommended to start each traffic class with equal shares of bandwidth and then adjust as needed. In the example above, bandwidth is set to 30% and 70%, considered appropriate for this network.

Check to see if PFC buffers are configured on the switch using the following command:

```
show dcb stack-unit 0
```

If not enabled, the following command will be required:

```
dcb stack-unit all pfc-buffering pfc-ports 64 pfc-queues 2
```

Save the configuration and reboot the system to allow the changes to take effect.

### 3.3.1 Validating and troubleshooting the DCB settings

Run the following four commands on *each interface where DCB is enabled* to validate the setup,

```
show interface tengigabitethernet 0/24 dcbx detail
show interface tengigabitethernet 0/24 ets detail
show interface tengigabitethernet 0/24 pfc detail
show interface tengigabitethernet 0/24 pfc statistics
```

The DCBx parameter in the show command displays information on application priority, shows which TLVs are enabled on the port, and also displays packet counters for TLV. The *ets* parameter displays each Traffic Class priority number and associated bandwidth. The *pfc* parameters display Tx and Rx pause packets as well as total Tx/Rx frames broken down by priority. Table 5 shows an example of the output seen with the show commands.

Table 5 DCB Validation (can be used on all interfaces with DCB enabled)

| Validating the DCB setup using "show interface"                                    |   |  |
|--|---|--|
| Aggr-1-S4810#show interface tengigabitethernet 0/24 dcbx detail                    |   |  |
| E-ETS Configuration TLV enabled  | e-ETS Configuration TLV disabled  |  |
| R-ETS Recommendation TLV enabled   | r-ETS Recommendation TLV disabled   |  |
| P-PFC Configuration TLV enabled  | p-PFC Configuration TLV disabled  |  |
| F-Application priority for FCOE enabled  | f-Application Priority for FCOE disabled  |  |
| I-Application priority for iSCSI enabled   | i-Application Priority for iSCSI disabled   |  |
| -----  |   |  |
| Interface TenGigabitEthernet 0/24  |   |  |
| Remote Mac Address d0:67:e5:b5:3c:83   |   |  |
| Port Role is Manual  |   |  |
| DCBX Operational Status is Enabled   |   |  |
| Is Configuration Source? FALSE   | ← Only one port per LAG should show "TRUE" on both ToR switches. All other ports should show "FALSE". If more than one port is "TRUE" the setup is incorrect. |  |
| Local DCBX Compatibility mode is IEEEv2.5  |   |  |
| Local DCBX Configured mode is AUTO   |   |  |
| Peer Operating version is IEEEv2.5   |   |  |
| Local DCBX TLVs Transmitted: ERPfl   |   |  |
| 13 Input PFC TLV pkts, 12 Output PFC TLV pkts, 0 Error PFC pkts                    |   |  |
| 0 PFC Pause Tx pkts, 2293799792 Pause Rx pkts                                      |   |  |
| 13 Input ETS Conf TLV Pkts, 12 Output ETS Conf TLV Pkts, 0 Error ETS Conf TLV Pkts |   |  |
| 0 Input ETS Reco TLV pkts, 12 Output ETS Reco TLV pkts, 0 Error ETS Reco TLV Pkts  |   |  |
| Aggr-1-S4810#show interface tengigabitethernet 0/24 ets detail                     |   |  |

### Validating the DCB setup using “show interface”

Interface TenGigabitEthernet 0/24  
Max Supported PG is 4  
Number of Traffic Classes is 8  
Admin mode is on

Admin Parameters :

-----

Admin is enabled

| TC-grp | Priority#     | Bandwidth | TSA |
|--------|---------------|-----------|-----|
| 0      | -             | -         |     |
| 1      | 4             | 30 %      | ETS |
| 2      | 0,1,2,3,5,6,7 | 70 %      | ETS |
| 3      | -             | -         |     |
| 4      | -             | -         |     |
| 5      | -             | -         |     |
| 6      | -             | -         |     |
| 7      | -             | -         |     |

Remote Parameters :

-----

Remote is enabled

| TC-grp | Priority#     | Bandwidth | TSA |
|--------|---------------|-----------|-----|
| 0      | -             | -         |     |
| 1      | 4             | 30 %      | ETS |
| 2      | 0,1,2,3,5,6,7 | 70 %      | ETS |
| 3      | -             | -         |     |
| 4      | -             | -         |     |
| 5      | -             | -         |     |
| 6      | -             | -         |     |
| 7      | -             | -         |     |

Remote Willing Status is disabled

Local Parameters :

-----

Local is enabled

| TC-grp | Priority#     | Bandwidth | TSA |
|--------|---------------|-----------|-----|
| 0      | -             | -         |     |
| 1      | 4             | 30 %      | ETS |
| 2      | 0,1,2,3,5,6,7 | 70 %      | ETS |
| 3      | -             | -         |     |
| 4      | -             | -         |     |
| 5      | -             | -         |     |
| 6      | -             | -         |     |
| 7      | -             | -         |     |

Oper status is init

### Validating the DCB setup using “show interface”

ETS DCBX Oper status is Up  
State Machine Type is Asymmetric  
Conf TLV Tx Status is enabled  
Reco TLV Tx Status is enabled

13 Input Conf TLV Pkts, 12 Output Conf TLV Pkts, 0 Error Conf TLV Pkts  
0 Input Reco TLV Pkts, 12 Output Reco TLV Pkts, 0 Error Reco TLV Pkts

#### Aggr-1-S4810#show interface tengigabitethernet 0/24 pfc detail

Interface TenGigabitEthernet 0/24  
Admin mode is on  
Admin is enabled, Priority list is 4  
Remote is enabled, Priority list is 4  
Remote Willing Status is disabled  
Local is enabled, Priority list is 4  
Oper status is init  
PFC DCBX Oper status is Up  
State Machine Type is Symmetric  
TLV Tx Status is enabled  
PFC Link Delay 45556 pause quntams  
Application Priority TLV Parameters :  
-----  
FCOE TLV Tx Status is disabled  
ISCSI TLV Tx Status is enabled  
Local FCOE PriorityMap is 0x0  
Local ISCSI PriorityMap is 0x10  
Remote ISCSI PriorityMap is 0x10

13 Input TLV pkts, 12 Output TLV pkts, 0 Error pkts, 0 Pause Tx pkts, 2293915468 Pause Rx pkts

#### Aggr-1-S4810#show interface tengigabitethernet 0/24 pfc statistics

Interface TenGigabitEthernet 0/24

| Priority | Rx XOFF Frames |   | Rx Total Frames | Tx Total Frames |
|----------|----------------|---|-----------------|-----------------|
| 0        | 0              | 0 | 0               |                 |
| 1        | 0              | 0 | 0               |                 |
| 2        | 0              | 0 | 0               |                 |
| 3        | 0              | 0 | 0               |                 |
| 4        | 11884412057    |   | 23768824112     | 0               |
| 5        | 0              | 0 | 0               |                 |
| 6        | 0              | 0 | 0               |                 |
| 7        | 0              | 0 | 0               |                 |

## 4 Storage

Configure iSCSI storage pools with various storage array products including Dell EqualLogic, Dell Compellent, and Dell PowerVault. For this scenario, we use a commonly used EqualLogic array using an EqualLogic PS 6xxx/4xxx as an example.

### 4.1 Configuring EqualLogic PS 6xxx/4xxx family multi-member pools

It is highly recommended to use the *EqualLogic PS SERIES STORAGE ARRAYS Installation Guide* that comes in the box with each member of the EqualLogic PS 6xxx/4xxx family. This manual provides excellent guidance to quickly set up any *EqualLogic* PS storage array.

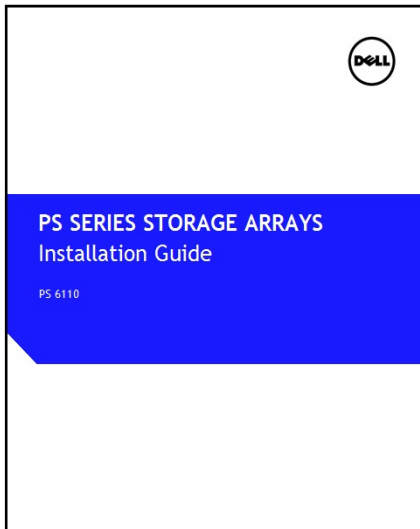


Figure 26 Installation Guide for PS Series Storage Arrays

You may also refer to the Array Configuration guide at <http://en.community.dell.com/dell-groups/dtcmedia/m/mediagallery/20094957/download.aspx> which contains step by step instructions for deploying arrays.

**Note:** It is important to cable both the active and the standby controllers of your EqualLogic arrays to achieve high availability in case of controller or cable failure. An amber ACT light on the EqualLogic controller signifies it is the standby controller. A green ACT light signifies it is the active controller.

**Caution:** On each EqualLogic PS 6110 control module, you can only use one of the two 10 Gb Ethernet ports. If connecting both the SFP+ and the 10Gbase-T ports at the same time, only the SFP+ will be active.

Enter the IP address created in the instructions above into an internet browser to connect to the *EqualLogic Group Manager*. All administrative tasks including creating volumes and pools are performed from this web management tool.



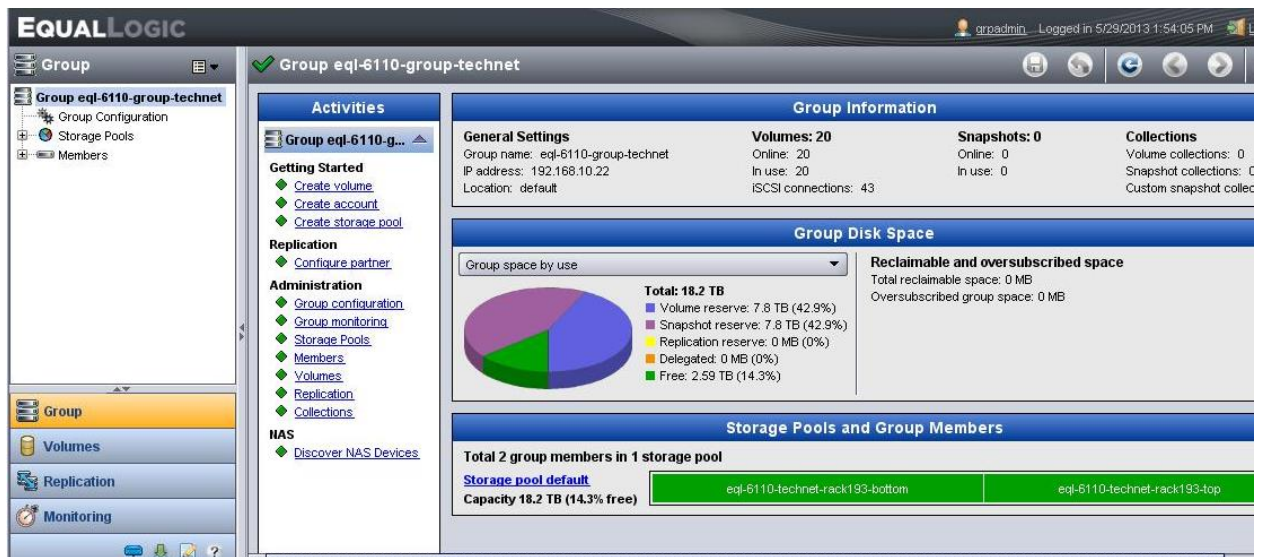


Figure 27 EqualLogic PS Series Group Manager

### 4.1.1 RAID policies

Download the technical report, *EqualLogic PS Series Storage Arrays: Choosing a Member RAID Policy*, by visiting <http://www.dellstorage.com>, and entering “choosing a member RAID policy” in the search field.

### 4.1.2 Volumes (iSCSI targets)

Use the web management GUI to allocate group storage space to users and applications. Each separate space is known as a *volume*, and appears on the network as an *iSCSI target*. Setup is simple, and the steps are provided for both Web UI and the CLI in the *PS Series Storage Arrays Installation Guide* (Figure 26) that comes with your EQL device.

You may also refer to the Array Configuration guide at <http://en.community.dell.com/dell-groups/dtcmmedia/m/mediagallery/20094957/download.aspx> which contains step by step instructions for deploying arrays.

### 4.1.3 Data Center Bridging

Once you create an EQL Group, enable Data Center Bridging (DCB) on the group so it knows to use the DCB settings propagated from the switch. You have the option of using either the CLI or the web management portal to enable/disable DCB.

Using the CLI, enter the following command:

```
grpparams dcb enable
```

Or, in the Group Manager web tool, click **Group Configuration > Advanced** tab and select the **Enable DCB** check box under Network Management > Data Center Bridging.

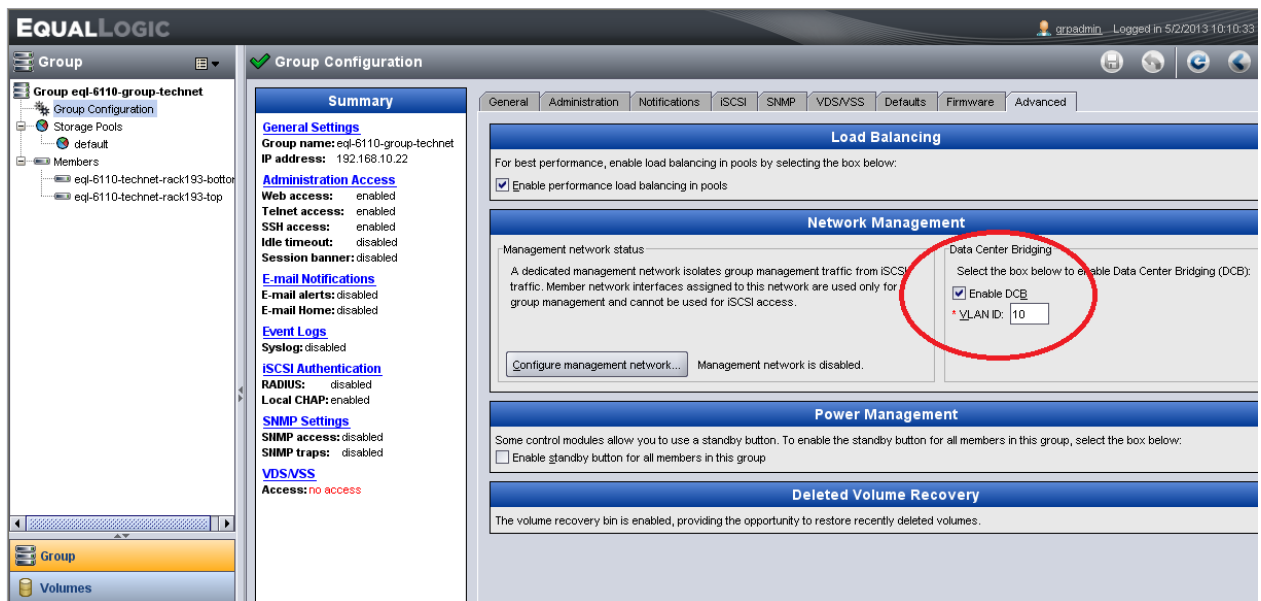


Figure 28 Enabling DCB on an EqualLogic group

Use the DCB Details page to see that DCB parameters are being properly received and used in the storage. To open the DCB Details, select the Group navigation tab, select a member under Members, select the Network tab, then click DCB Details (see Figure 29).

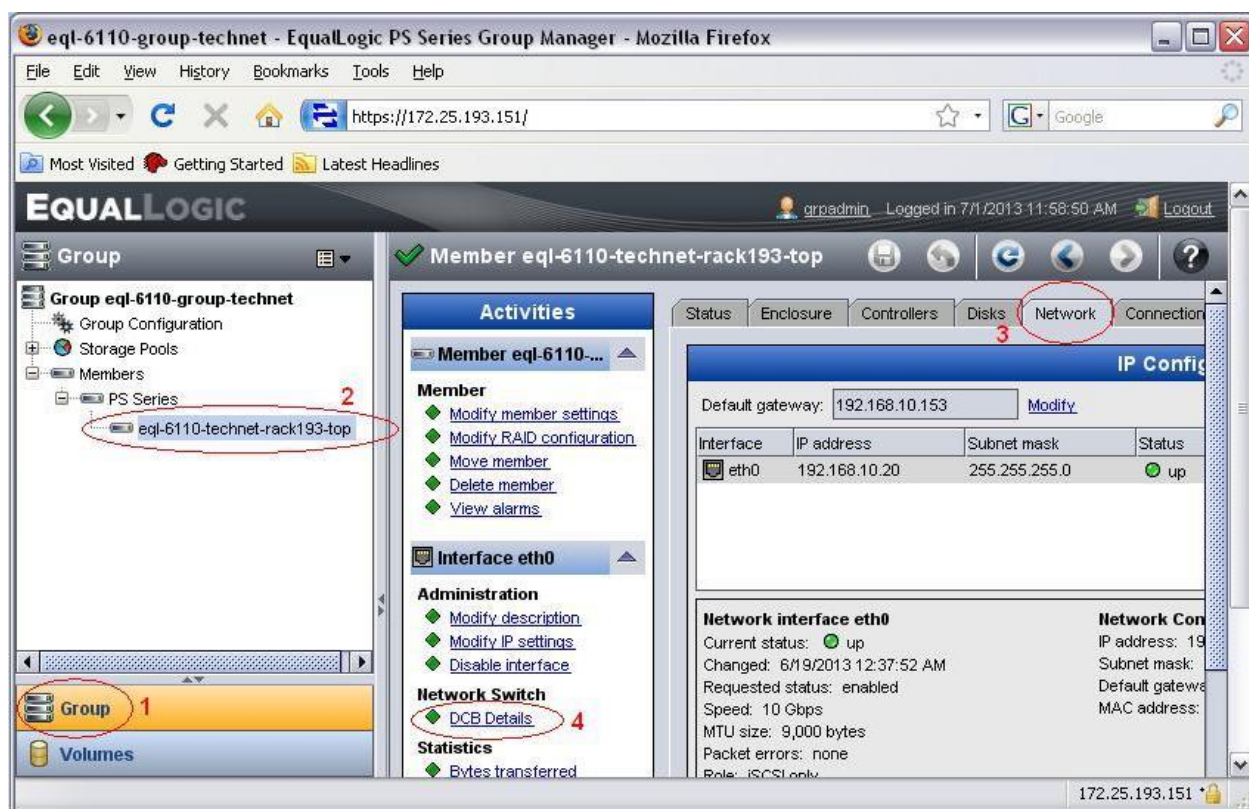


Figure 29 Launching DCB Details in EqualLogic

Click on *DCB Details* to launch the page shown in Figure 30. The information shown on this page should be identical to the information found in the **show interface** commands (see Table 4 starting on page 35).

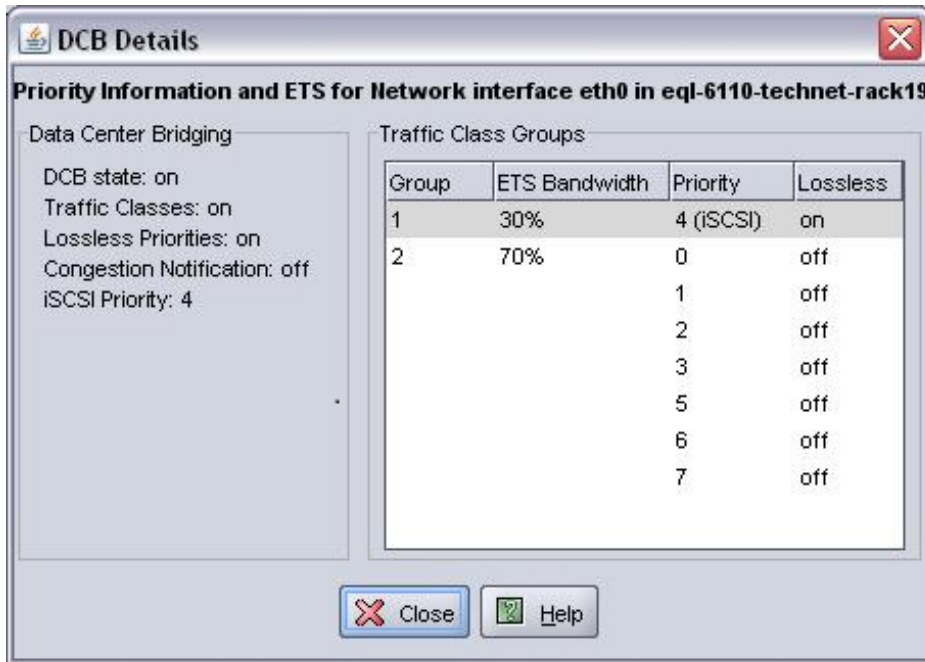


Figure 30 DCB Details

#### 4.1.4 Firmware Updates

Dell recommends using the latest firmware for all EqualLogic PS storage devices. When updating firmware on multiple members in the same pool, do not rely on what is seen in the web management tool until you update and restart *all* members. Once all members are back up, make sure to refresh the web management information.

**Note:** When updating firmware on a pool using the web management interface, by default it will only allow restarting pool members one at a time, requiring the member to fully restore before restarting the next.

## 5 Network Management

Use Dell OpenManage Network Manager (OMNM) to manage the data center network. OMNM is a one-to-many Network Management System that uses the SNMP protocol and built-in Telnet scripting to manage and monitor all Dell switches in the data center including the S4810, providing the user an easy-to-use graphical interface to simplify a wide arrange of network management tasks. To learn more about OMNM and download a free version, go to <http://www.dell.com/OMNM>. The free version provides full support for up to 10 Dell networking switches or 10 stacks of Dell networking switches.

OMNM can notify an administrator the instant that a problem occurs on the network. For example, if the VLT bandwidth exceeds a given usage threshold or drops below a certain threshold, a trap is automatically generated. Other examples include monitoring if an unauthorized user is attempting to log in or if a cable is removed. In these, and several other scenarios, traps are generated. OMNM can receive these traps and notify an administrator. Use the following commands on each S4810 to allow OMNM to monitor and manage them and to receive such traps.

```
Aggr-1-S4810(conf)#snmp-server community private rw
Aggr-1-S4810(conf)#snmp-server enable traps
Aggr-1-S4810(conf)#snmp-server host xxx.xxx.xxx.xxx traps version 2c
private
```

The host IP address is the address of the OMNM server, where traps are sent. Use the **show snmp** and **show snmp community** commands to view the SNMP configuration.

Inside OMNM, set the IP address range and the authentication required to discover the S4810s on the network. A minimum setup of SNMP and Telnet authentication is required and can be done using Quick Discovery or a Discovery Profile setup. Select **Resource Discovery** from the Quick Navigation menu on the home page. Figure 31 shows setting up the IP addresses for discovery and assigning authentication to be used.

The screenshot shows the OMNM Network Management interface. At the top, there are tabs for 'Network' and 'Results'. Below this, the interface is divided into two main sections: '1. Select Network Type and Address(es)' and '2. Select Authentication'.

Section 1: 'Select Network Type and Address(es)' contains a dropdown menu for 'IP Address(es)' and a text input field containing the IP addresses '172.25.188.160, 172.25.188.161, 172.25.188.162, 172.25.188.163'. Below the input field is a label 'Enter IP Address(es), IP Range(s) and/or Network(s)'.

Section 2: 'Select Authentication' contains a table with columns 'Name', 'Type', 'Parameters', and 'Actions'. There are two rows in the table, both with checkboxes in the 'Name' column.

| Name   | Type   | Parameters                        | Actions |
|--|--------|-----------------------------------|---------|
| <input checked="" type="checkbox"/> Telnet-1 | TELNET | Timeout: 10, Retries: 1, Port: 23 |         |
| <input checked="" type="checkbox"/> SNMP-1   | SNMPV2 | Timeout: 5, Retries: 2, Port: 161 |         |

At the top right of Section 2, there are two buttons: 'Create New' and 'Choose Existing'.

Figure 31 Selecting IP addresses and authentication to use in discovering network switches

Create appropriate authentications to match that setup on the switches by clicking the **Create New** button. Figure 32 shows the resulting screen for setting up SNMP v2c authentication using Quick Discovery. Consult the online user manual for more information on setting up Resource Discovery.

**Add New Authentication To Resource Discovery**

Authentication Name:

Protocol Type:

**Authentication** | Management Interface

Read Community:

Write Community:

Trap Community:

Figure 32 Adding Authentication to Resource Discovery

Press **Apply**, **Execute**, then **Discover**. Once discovered, devices are listed in the home page.

**Managed Resources**

| Network Status | Name                 | IP Address     | Vendor    | Model         |
|----------------|----------------------|----------------|-----------|---------------|
| ✓ Responding   | TOR-2-S4810.172.2... | 172.25.188.163 | Dell Inc. | Force10 S4810 |
| ✓ Responding   | TOR-1-S4810.172.2... | 172.25.188.162 | Dell Inc. | Force10 S4810 |
| ✓ Responding   | Aggr-2-S4810.172...  | 172.25.188.161 | Dell Inc. | Force10 S4810 |
| ✓ Responding   | Aggr-1-S4810.172...  | 172.25.188.160 | Dell Inc. | Force10 S4810 |

**Alarms**

| Severity                        | Date Opened | Entity Name | Device IP | Event Name |
|---------------------------------|-------------|-------------|-----------|------------|
| No data is available to display |             |             |           |            |

Figure 33 Discovered devices in OMNM

Choose **Alarms** from the main navigation menu to monitor important events on the network.



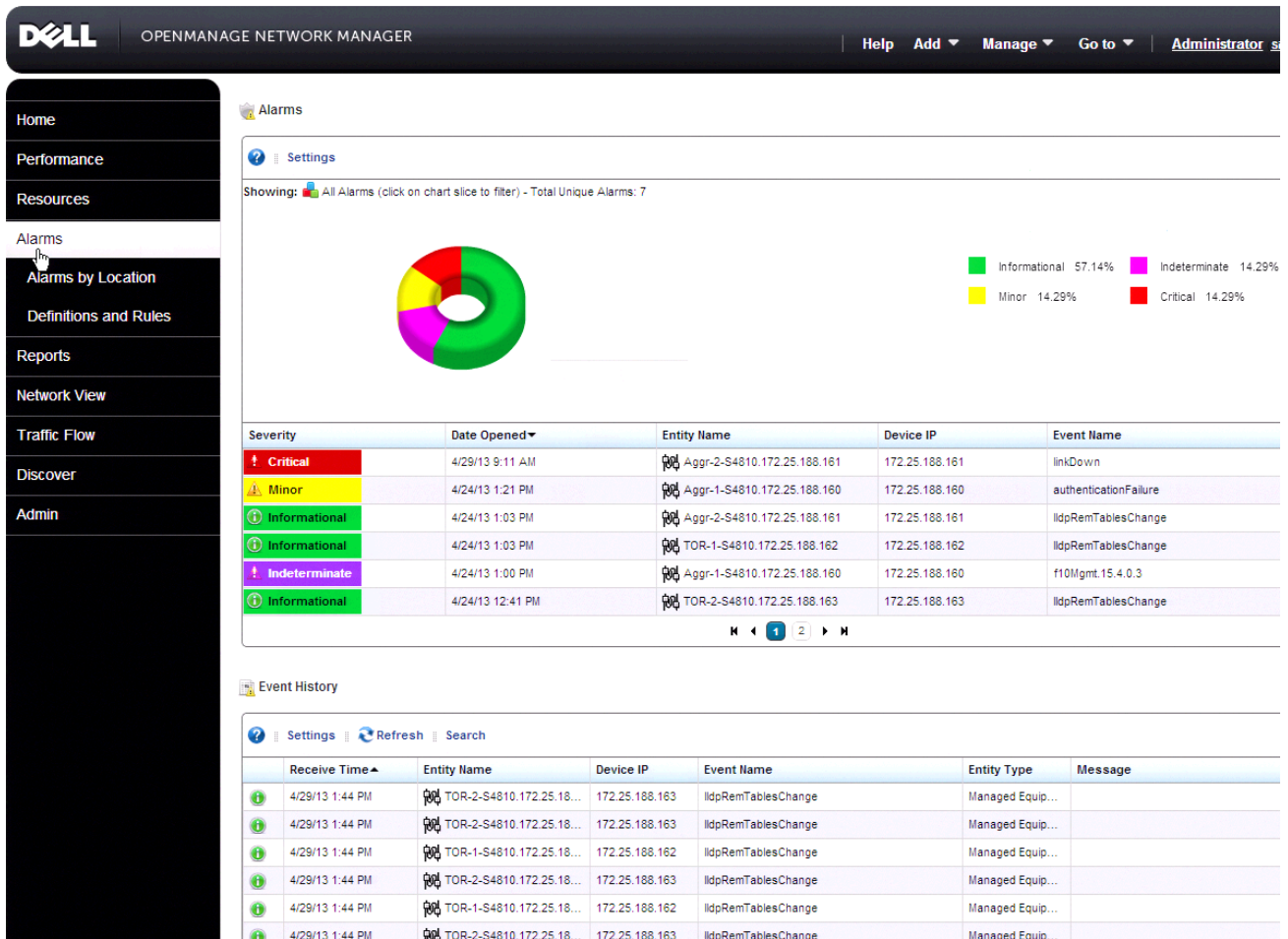


Figure 34 Network alarms displayed in OMNM

OMNM also lets you setup automatic notifications to the administrator through email (SMTP) and/or SMS. Consult the OMNM user guide for more information on these features.

When discovering S4810 switches, OMNM automatically discovers links between each device. LLDP is enabled by default on S4810 switches and must remain enabled for links to be discovered. By selecting the Network View in OMNM, a topology map is created similar to the one in Figure 35.

If any links are missing between devices, check LLDP settings on the switches as well as physical cabling.

**Note:** If the Network View does not open to show a topology map (progress indicator circles indefinitely), this is caused by one of three issues. First, try replacing localhost in the URL with the actual IP address. For example: `http://<ip-address>:8080`, and log in again. Second, try removing any proxy settings for the local network and log in again. Finally, make sure to use a Java version supported by OMNM.

As seen in Figure 35, hovering on a switch icon shows information about that switch along with the status and severity of any errors seen.

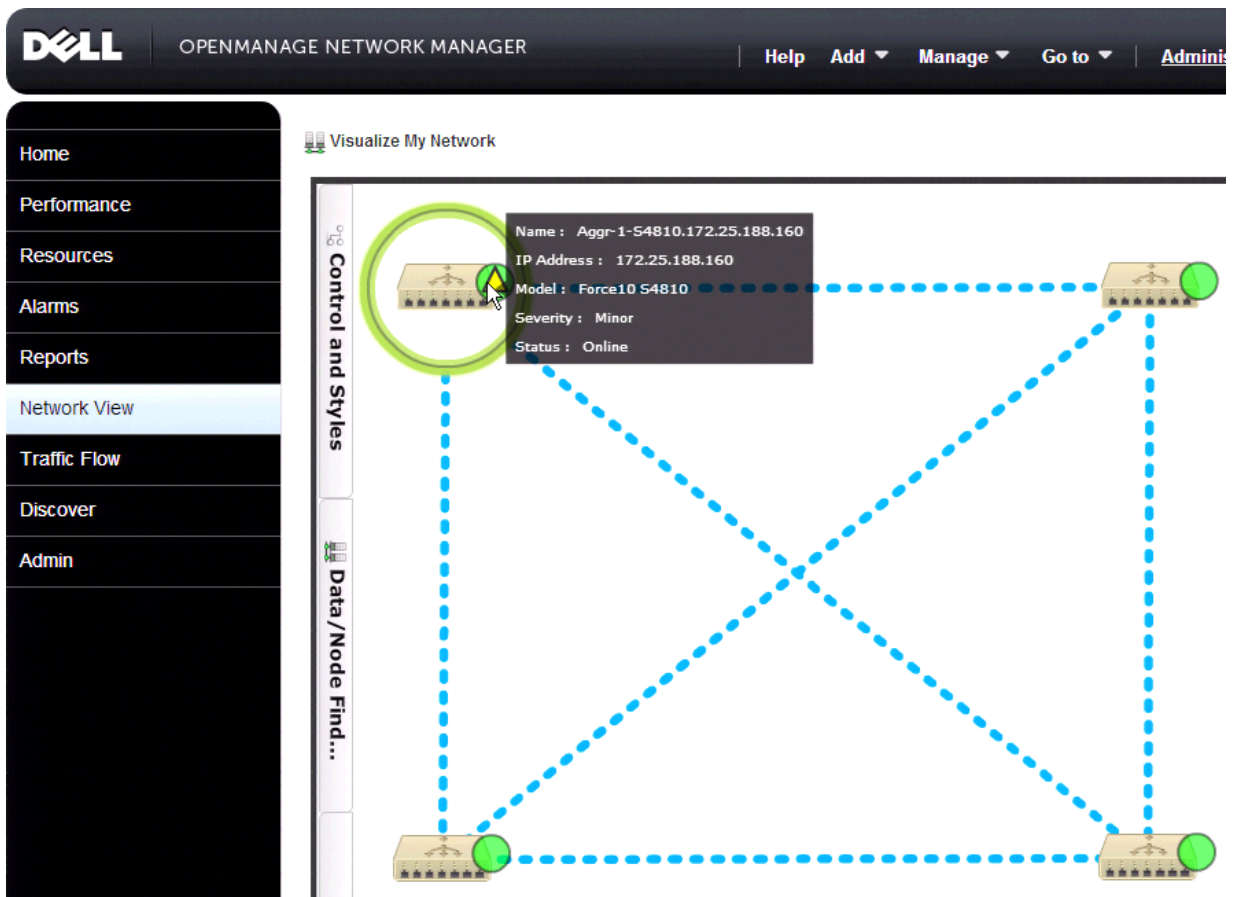


Figure 35 Identifying an alarm in an OMNM topology map

Notice the yellow warning on the top left switch in Figure 35. OMNM offers the ability to view alarms from the topology map by right clicking the switch, then selecting **Actions>Event Management>View Alarms>Execute**. Figure 36 shows the events related to that switch:

| Show Alarm(s) |                  |                    |                |                       |
|---------------|------------------|--------------------|----------------|-----------------------|
| Severity      | Date Opened▼     | Entity Name        | Device IP      | Event Name            |
| Minor         | 4/24/13 1:21 PM  | Aggr-1-S4810.17... | 172.25.188.160 | authenticationFailure |
| Indeterminate | 4/24/13 1:00 PM  | Aggr-1-S4810.17... | 172.25.188.160 | f10Mgmt.15.4.0.3      |
| Informational | 4/24/13 12:41 PM | Aggr-1-S4810.17... | 172.25.188.160 | IldpRemTablesChang    |

Figure 36 Alarm list

The topology map also offers link identification by hovering the mouse over the link as shown in Figure 37.



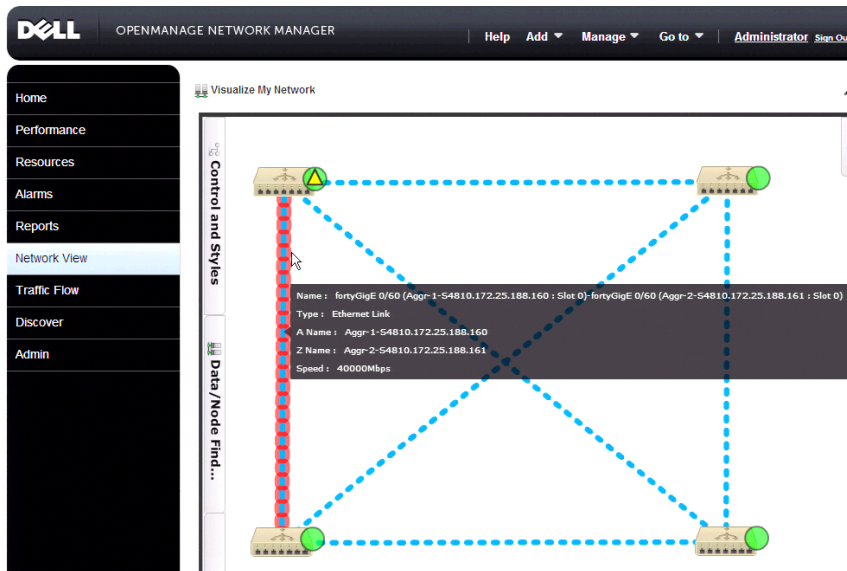


Figure 37 Link identification in OMNM

## 5.1 Graphing performance and traffic flow metrics

OMNM lets you track and graph key metrics in data flow and performance statistics. By default, only select metrics are enabled. To enable more metrics, click **Performance > Resource Monitors**. Left-click to select each *Monitor Type* then press Enable to enable each one.

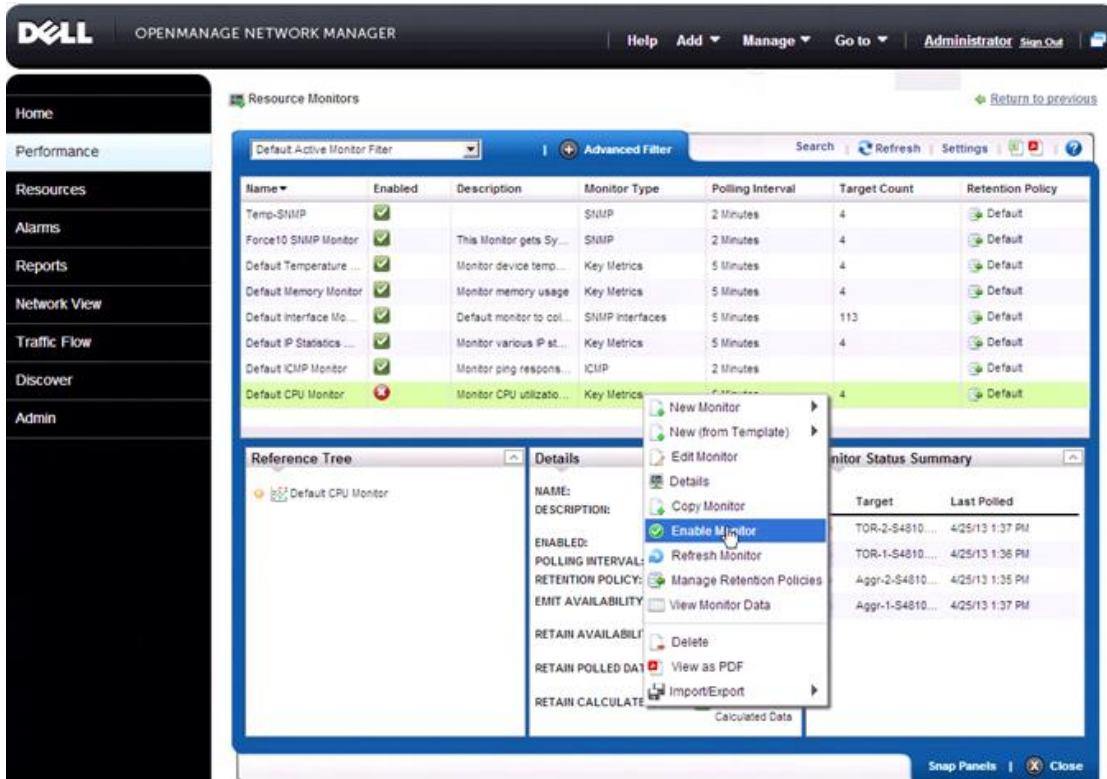


Figure 38 Enabling data collection in OMNM

Once the first polling interval has passed, graphs begin to populate under the Performance tab > Network Dashboard as shown in Figure 39.

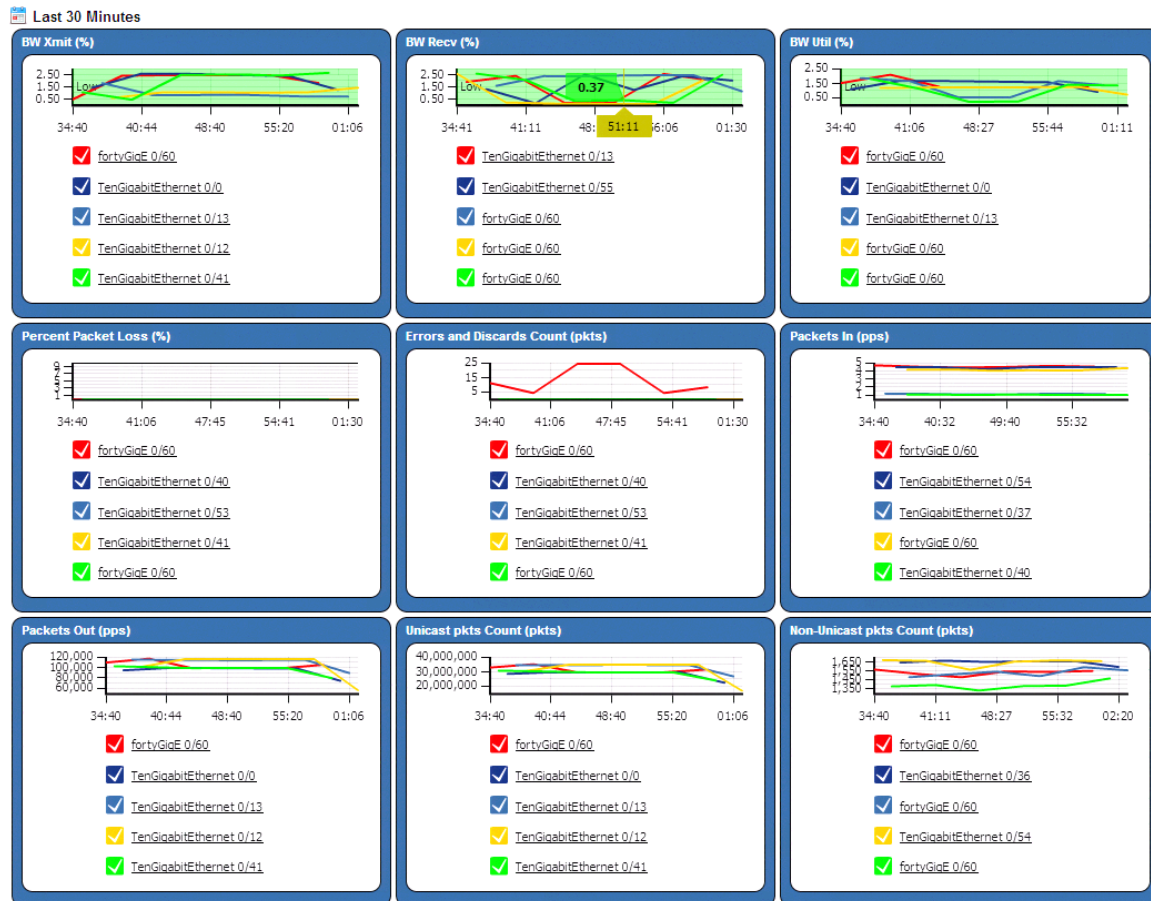


Figure 39 Network dashboard in OMNM

Hovering over any point on a line in the chart shows the exact time and value for that entity. The value is also color-coded to allow for easier viewing as shown in Figure 40.

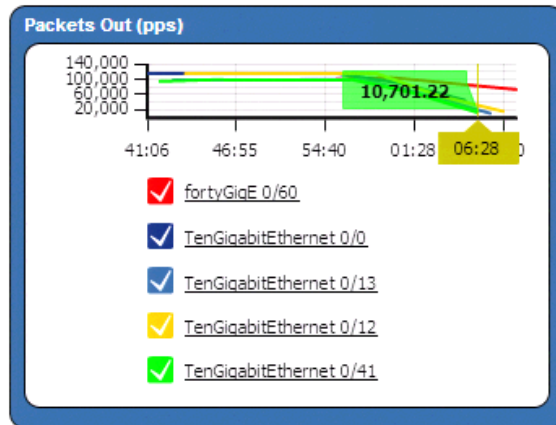


Figure 40 Color-coded for easy viewing

To ensure proper SNMP configuration and functionality from the switch side, view SNMP statistics can through the S4810 CLI by using the **show snmp** command.

```
Aggr-1-S4810#show snmp
1215  SNMP packets input
      0  Bad SNMP version errors
      0  Unknown community name
      0  Illegal operation for community name supplied
      0  Encoding errors
2916  Number of requested variables
      0  Number of altered variables
708   Get-request PDUs
506   Get-next PDUs
      0  Set-request PDUs
57399 SNMP packets output
      0  Too big errors (Maximum packet size 1500)
      0  No such name errors
      0  Bad values errors
      0  General errors
1214  Response PDUs
56185 Trap PDUs
```

## 5.2 sFlow traffic

The S4810 and OMNM also support sFlow. To enable the sFlow collector feature in OMNM, a license must be purchased and applied. Contact your Dell sales representative to purchase this license.

To receive sFlow samples from *all* front traffic ports on the S4810 switch, enter the following in *Global configuration mode*:

```
sflow enable
sflow collector 172.25.188.31 agent-addr 172.25.188.160
sflow extended-switch enable12
sflow polling-interval 20
sflow sample-rate 512
```

To receive sFlow samples from a limited number of ports (one or more), configure each port separately using the same commands and parameters from the *interface configuration* prompt. If this is implemented, do not use the global command. Enter the commands from the physical interface configuration mode (for example, TenGigabitEthernet 0/12), or from an Aggregation Link configuration mode (for example Port-channel 3). Make sure to enter the **no shutdown** command on an S4810 interface to bring up the interface.

sFlow does not work on OOB (out of band) management interfaces. sFlow traffic is only exported to the collector on in-band management interfaces. Therefore, connect the OMNM server to the network through an in-band interface to receive sFlow.

In the OMNM home page, under Managed Resources, right-click each switch that you want to view with sFlow, and select Traffic Analyzer > Register.

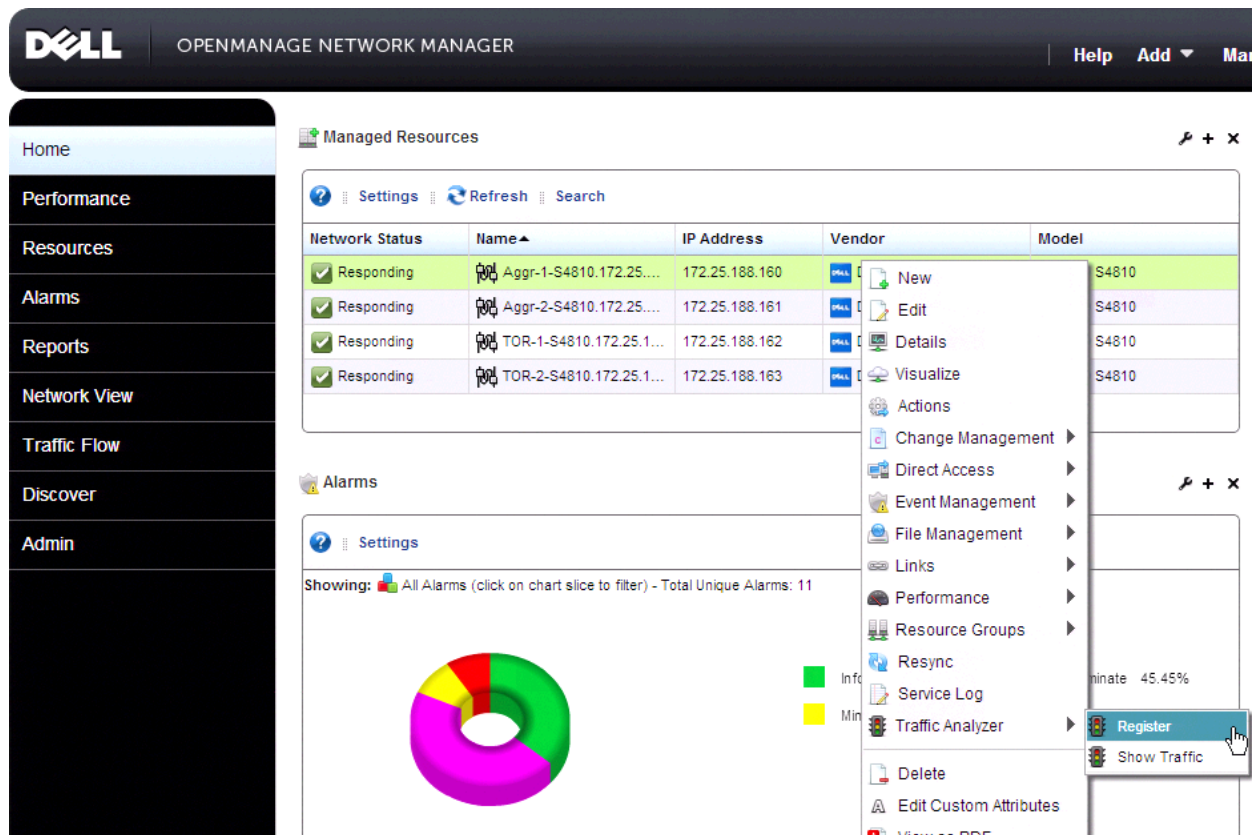


Figure 41 Registering switches for sFLOW

After the first polling interval, traffic flow statistics is seen under the Traffic Flow tool on the main navigation menu. More information is found on configuring sFlow in the *S4810 CLI guide* and the *OMNM User Guide*.

## 6 Switch Configurations

Final configuration files for each network switch in this deployment guide are attached to this document and reflect a culmination of all CLI commands previously mentioned. The topology map below shows connectivity for each.

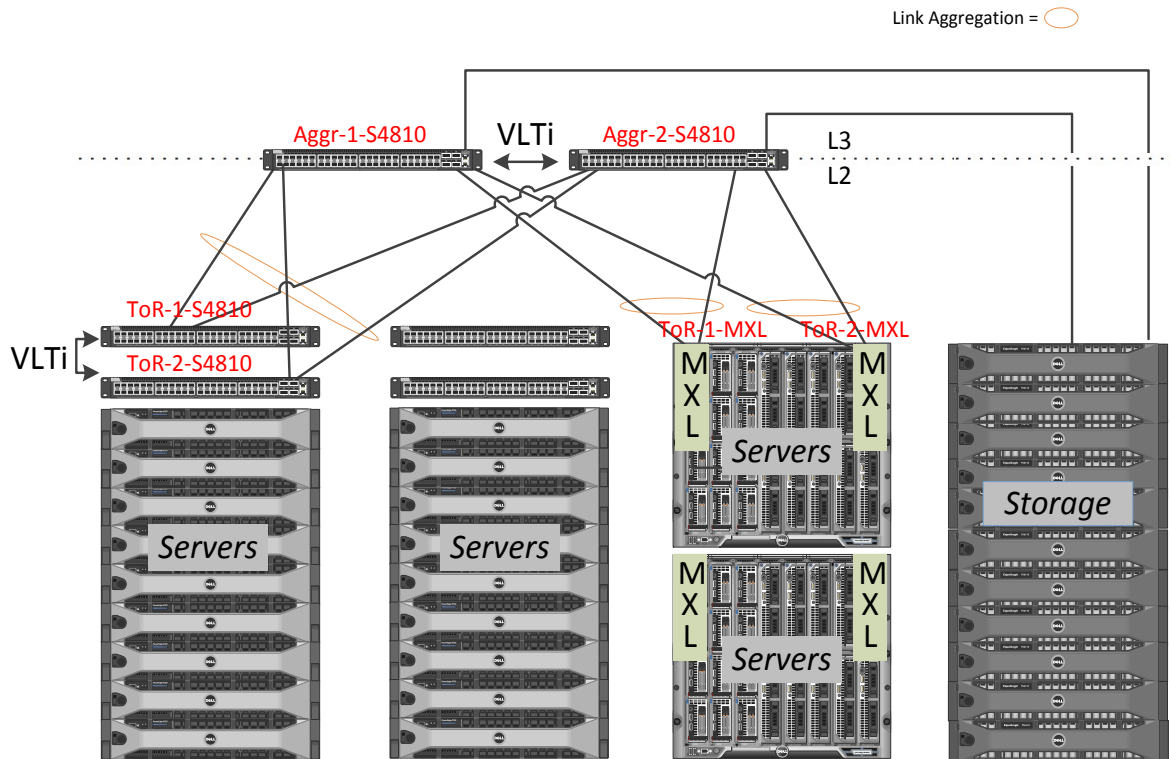



Figure 42 Switches used in the data center Active Fabric POD example

Click the paper clip  icon on the left to select one or more configuration files for viewing.

As shown in Figure 42, it is easy to scale this configuration by using the settings in this document to configure additional ToR switches. The configuration files provide an example of setting up two ToR S4810s (VLT pair) and redundant ToR MXLs. Use these same configurations to scale the data center as needed. The servers and switches shown without cables in Figure 42 are ready to be added to the network using the same configurations.

## Observations

In every setup of this size and scope there are always unexpected behaviors. Such were witnessed during the testing and creation of this deployment guide. Below are some of the observations. Your experience may or may not be the same.

- **Intel Driver** – If Lifecycle Controller is used to update the Intel driver, the update fails. There are two work-arounds for this issue:
  1. During OS installations avoid using Lifecycle Controller to update drivers. Instead, load the drivers manually.
  2. If the OS is already installed using Lifecycle Controller, uninstall video driver before installing/updating Intel driver.
- **Broadcom versus Intel CNA** – Intel currently does not support NPAR and hence it limits to achieve MPIO and NIC teaming at the same time on single adapter. If you are looking for load balancing iSCSI and LAN traffic, Broadcom fulfills this requirement.

However, Broadcom limits one to use NPAR or iSCSI offload with LACP (802.3ad) for active-active NIC-teaming, so the second best option is to use SLB (smart load balancing and failover) with iSCSI offload enabled.

- **NIC teaming** – make sure to have same VLAN-ID for Team VLAN interface and the interfaces that belong to the team. Do not assume that VLAN interface VLAN-ID overrides VLAN on interface.
- **Windows versus VMware** – when both Windows and VMware systems were running IOmeter at the same time, VMware system performance was affected by at least 20 percent.
- When multiple VMs were running IOmeter at same time on single host, All VMs suffer the hit instead of one or two VMs showing more of a load balancing nature.
- **OMNM** version 5.1.0.17 (with SP1) may give an error when the user clicks the “Click Here” button from the index.html file. To get around this issue, run the win\_install.exe found in the Dell\_OMNM directory. Also, after installing, to run OMNM 5.1 web client on a system running VMware requires the user first change the default VMware HTTP port. Attempting to open OMNM while VMware is using the default port results in the VMware client being loaded instead. Currently, OMNM provides no way to change the default HTTP port used.



## A Configuration details

| Component           | Description  |
|---------------------|--|
| Operating systems   | <ul style="list-style-type: none"> <li>• Microsoft® Windows® Server 2012 Standard</li> <li>• VMware ESXi 5.1 (5.1.0 build-799733)</li> </ul>   |
| NIC driver versions | <ul style="list-style-type: none"> <li>• Intel 10GbE/FCoE Dual KR X-520, driver version 14.0.0_W2W8_64_W2K12_A00</li> <li>• BRCM 10GbE 2P 57810S-k, driver version 17.4.0.9 (ESXi 5.1.0 native driver)</li> <li>• BRCM firmware version 7.4.8</li> <li>• Intel firmware version 14.0.12_A00</li> </ul>   |
| Applications        | <ul style="list-style-type: none"> <li>• OMNM (OpenManage Network Manager) version 5.2.0.7 (SP1)</li> <li>• EqualLogic PS Series Group Manager, version 6.0</li> <li>• EqualLogic MEM for EXSi 5.1, version 5-1.1.2.292203</li> <li>• EqualLogic HIT Kit for Windows 2012, 4.5.0</li> <li>• Latest free downloadable IOMeter, iperf</li> </ul>   |
| Servers             | <p><u>Rack Servers:</u></p> <ul style="list-style-type: none"> <li>• R720 Server, Broadcom 10gig dual-port CNAs, VMware ESXi 5.1 Intel® Xeon® CPU E5-2643 0 @ 3.30GHz, 64 GB memory</li> <li>• R620 Server, Intel 10gig dual-port CNAs, Windows Server 2012 Standard 64-bit 2xIntel® Xeon® CPU E5-26650 0 @ 2.40GHz , 192 GB RAM</li> </ul> <p><u>Blade Servers:</u></p> <ul style="list-style-type: none"> <li>• M420 Server, Slot 6B in M1000e, BIOS version 1.2.4, Windows Server 2012 Standard 64-bit, 2x Intel® Xeon® CPU E5- 2450 0 @ 2.10GHz , Total System Memory 24 GB. Fabric B: 10 GbE XAUI KR Intel 10GbE/FCoE Dual KR X-520</li> <li>• M620 Server, Slot 10 in M1000e, BIOS version 1.6.1, VMware ESXi 5.1.0, 2x Intel® Xeon® CPU E5-2670 0 @ 2.60GHz GB, Total System Memory 96 GB. Fabric B: BRCM 10GbE 2P 57810S-k Mezz</li> </ul> |
| Storage             | EQL PS6110 XS (two), RAID 6 (accelerated), 2 Pools, 27 Volumes, 51 iSCSI connections, 24 disks per enclosure, 48 disks total, SSD/HDD hybrid mix of disks. 9.1 TB per enclosure, 18.2 TB total, PS 6110 Firmware version 6.0.2 (R287892)   |
| Switches            | <ul style="list-style-type: none"> <li>• Dell MXL, located in B1 and B2 slots in M1000e chassis, firmware version 8.3.16.4, hardware version A01</li> <li>• Dell S4810, firmware version 9.1.(0.0). hardware version 3.0</li> </ul>  |
| Cables              | Split/breakout cables 40G to 4x10G   |



## B Terminology

### **CLI**

Command Line Interface (CLI) is the text-based console interface that is used for entering management and configuration commands into devices like the Dell Force10 MXL switch. You can access the MXL's CLI using telnet, SSH, an externally accessible serial connection, and also from the CMC's CLI.

### **ETS**

Enhanced Transmission Selection (ETS) is defined in the IEEE 802.1Qaz standard (IEEE, 2011). ETS supports allocation of bandwidth among traffic classes. It then allows for sharing of bandwidth when a particular traffic class does not fully use the allocated bandwidth. The management of the bandwidth allocations is done with bandwidth-allocation priorities, which coexist with strict priorities (IEEE, 2011).

### **LACP**

Link Aggregation Control Protocol (LACP) is the protocol used to make sure that the multiple links in a LAG do not form loops due to misconfiguration or device misbehavior. It is recommended practice to always use LACP on configured LAGs.

### **LAG**

Link Aggregation Group (LAG) is a configured bundle of Ethernet links that are treated as the same logical Ethernet link. There are multiple terms that apply to LAGs including channel group, port channel, trunk, and even some server Ethernet interface teaming involves a collection of links that would be considered a LAG. However while channel group and port channel always apply to LAG use, trunk and teaming do not.

### **Link**

Link is a term in networking that refers to a connection made between two nodes in a network. In Ethernet networking it is used to refer to a direct connection between two ports.

### **Out-of-Band**

An out-of-band interface provides management connectivity to a device without participating in or relying on a device's in-band (normal-use) data interfaces. On a switch, this means that an out-of-band interface does not send or receive traffic from the switched links — neither bridged nor routed. Common out-of-band interface types are Ethernet and serial console — often both are presented with RJ-45 (8P8C) connectors although on IO modules in the Dell PowerEdge M1000e chassis the serial connector is sometimes a physical USB type-A port requiring a special cable.

### **PFC**

Priority Flow Control (PFC), or Per-Priority Pause is defined in the IEEE 802.1Qbb standard. PFC is flow control based on priority settings and adds more information to the standard pause frame. The additional fields added to the pause frame allow devices to pause traffic on a specific priority instead of pausing all traffic. (IEEE, 2009) Pause frames are initiated by the FCF in most cases when its receive buffers are starting to reach a congested point. With PFC traffic is paused instead of dropped and retransmitted. This provides the lossless network behavior necessary for FC (and iSCSI) packets to be encapsulated and passed along the Ethernet paths.

### **Port Channel**

See LAG.

## **RSTP**

Rapid Spanning-Tree Protocol (RSTP) is a standards-based modified version of the basic spanning tree protocol that allows for much faster convergence times of spanning tree instances and provides for special administratively assigned port states that improve behavior in certain circumstances. RSTP is originally defined in the IEEE 802.1w standard and is included in 802.1d IEEE Ethernet bridging standard.

## **Spanning Tree / STP**

Spanning Tree refers to a family of layer-2 management protocols used by Ethernet bridges to establish a loop-free forwarding topology. At layer-2, Ethernet is a very simple technology that without intervening protocols or configuration can easily forward traffic in endless loops.

## **Switchport**

Switchport is a configuration term used to denote an Ethernet switch's link interface that is configured for layer-2 bridging (participating in one or more VLANs).

## **ToR**

Top-of-Rack (ToR) is a term for a switch that is actually positioned at the top of a server rack in a data center.

## **VLAN**

Virtual Local Area Network (VLAN) is a single layer-2 network (also called a broadcast domain as broadcast traffic does not escape a VLAN on its own). Multiple VLANs can be passed between switches using switchport trunk interfaces. When passed across trunk links, frames in a VLAN are prefixed with the number of the VLAN that they belong to—a twelve-bit value that allows just over 4000 differently numbered VLANs.

## **Virtual link trunk (VLT)**

Combined port channel between an attached device and the VLT peer switches.

## **VLT backup link**

The backup link monitors the vitality of a VLT peer switches. The backup link sends configurable, periodic keep alive messages between VLT peer switches.

## **VLT interconnect (VLTi)**

The link used to synchronize states between the VLT peer switches. Both ends must be on 10G or 40G interfaces.

## **VLT domain**

This domain includes both VLT peer devices, the VLT interconnect, and all of the port channels in the VLT connected to the attached devices. It is also associated to the configuration mode that must be used to assign VLT global parameters.

## **Trunk**

Trunk is an ambiguous term in Ethernet networking that can apply to a LAG—a group of multiple links acting as one or to a switchport interface of an Ethernet switch configured in trunk mode to pass multiple VLANs across the one link.

## C Additional resources

### C.1 Servers

<http://support.dell.com> focuses on meeting your needs with proven services and support.

<http://www.DellTechCenter.com> connects you with other Dell Customers and Dell employees to share knowledge, best practices, and information about Dell products and installations.

<http://en.community.dell.com/techcenter/b/techcenter/archive/2012/08/21/some-thoughts-about-nic-partitioning-npar.aspx> to learn more about NPAR

### C.2 Networking

<http://en.community.dell.com/dell-blogs/direct2dell/b/direct2dell/archive/2013/01/16/vlt-virtual-link-trunking-maximizing-datacenter-capacity-with-dell-force10-switches.aspx> provides a 3 minute video explaining how VLT technology works.

<https://www.force10networks.com/CSPortal20/KnowledgeBase/Documentation.aspx> provides comprehensive configuration guides for both the Dell S4810 and the Dell MXL.

### C.3 Storage

<https://eqsupport.dell.com/support/resources.aspx?id=2495> for firmware downloads and additional resources.

<https://support.equallogic.com/secure/login.aspx> to download and install the EqualLogic Host Integration Tool (HIT) kit.

<http://www.equallogic.com/WorkArea/DownloadAsset.aspx?id=5231> to download the report: *PS Series Storage Arrays: Choosing a Member RAID Policy*.

<http://www.dell.com/OMNM> to learn more about OMNM and download a free version.

<http://kb.vmware.com/selfservice/microsites/microsite.do> - VMware Knowledge base

[http://www.vmware.com/files/pdf/virtual\\_networking\\_concepts.pdf](http://www.vmware.com/files/pdf/virtual_networking_concepts.pdf) VMware Virtual networking.

[http://kb.vmware.com/selfservice/microsites/search.do?language=en\\_US&cmd=displayKC&externalId=2044431](http://kb.vmware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=2044431) - DCB solution for ESXi 5.1

<http://www.dellstorage.com/WorkArea/DownloadAsset.aspx?id=3064> provides installation procedures for the EqualLogic multipath extension module (MEM).

<http://en.community.dell.com/techcenter/storage/w/wiki/3615.rapid-equallogic-configuration-portal-by-sis.aspx> takes you to the Rapid EqualLogic Deployment site which contains a step by step guide for deploying arrays, and other information.