

Dell PS Series DCB Configuration Best Practices

Dell Storage Engineering
May 2017

Revisions

Date	Description
February 2013	Initial publication
February 2013	Changed PC81xx configuration steps to external SGC link
May 2013	Added Force10 S4820T information
May 2014	Updated links to new Switch Configuration Guides
May 2017	Updated document to include consolidation of two other documents (BP1017 & BP1044)

The information in this publication is provided "as is." Dell Inc. makes no representations or warranties of any kind with respect to the information in this publication, and specifically disclaims implied warranties of merchantability or fitness for a particular purpose.

Use, copying, and distribution of any software described in this publication requires an applicable software license.

Copyright © 2013 - 2017 Dell Inc. or its subsidiaries. All Rights Reserved. Dell, EMC, and other trademarks are trademarks of Dell Inc. or its subsidiaries. Other trademarks may be the property of their respective owners. Published in the USA [5/8/2017] [Best Practices] [BP1058]

Dell believes the information in this document is accurate as of its publication date. The information is subject to change without notice.

Table of Contents

1	Introduction	5
1.1	Objective	5
1.2	Audience	5
2	Data Center Bridging (DCB) Explained	6
2.1	DCB requirements	6
2.2	DCB Terminology	7
2.2.1	Priority-based Flow Control	7
2.2.2	Enhanced Transmission Selection	8
2.2.3	Application priority configuration	8
2.2.4	Data Center Bridging Capability Exchange	9
2.3	DCB parameter configuration	10
2.3.1	End-device auto configuration	10
2.3.2	Switch to switch configuration	11
2.4	Converged network example	12
2.5	DCB configuration process overview	13
2.5.1	Enabling DCB and configuring the VLAN	14
2.5.2	Switch configuration	15
2.5.3	Disabling DCB on the SAN	15
2.5.4	DCB configuration verification	15
2.5.5	Priority Group, ETS, and PFC verification	15
2.5.6	Invalid DCB configuration detection	16
3	Network configuration and deployment topologies	18
3.1	General configuration process flow	18
3.2	Configuration for switches with DCB support	21
3.2.1	Dell Networking S4810	21
3.2.2	Dell Networking S4820T	21
3.2.3	Dell Networking MXL	21
3.2.4	PowerConnect 81XX Series	21
3.2.5	Configuration for switches that do not support DCB for PS Series	22
3.2.6	PowerConnect 8024, 8024F, and M8024	22
3.3	Deployment topologies	22
3.3.1	Single layer switching with blade servers	22

3.3.2	Single layer switching with rack servers	23
3.3.3	Dual layer switching	24
4	Host adapter configuration for DCB	27
4.1	QLogic 57810 (previously Broadcom) configuration	27
4.2	QLogic 57810 DCBX willing mode verification	27
4.3	QLogic 57810 iSCSI VLAN configuration	28
4.4	QLogic 57810 DCB configuration verification	28
5	DCB testing	30
5.1	Workload definitions	30
5.2	Topologies	32
6	Results and analysis	33
6.1	DCB versus non-DCB in a dedicated SAN network	33
6.2	DCB versus non-DCB in a shared network	34
A	PS Series command line reference for DCB configuration verification	37
B	Technical support and resources	40
B.1	Additional resources	40

1 Introduction

Data Center Bridging (DCB) is a networking standard that first appeared on Fibre Channel over Ethernet (FCoE) SANs as a method to ensure lossless delivery of Fibre Channel (FC) traffic from Ethernet hosts to FC storage targets. Over recent years, this technology has been extended to deliver similar benefits for iSCSI storage solutions and continues to gain acceptance with storage networking device manufacturers and customers. Beginning with Dell™ PS Series Firmware release 5.1 in 2010, Dell introduced DCB for iSCSI on all 10 GbE PS Series products. Given that all network elements in a DCB-enabled iSCSI SAN need to support DCB, Dell offers PowerEdge™ converged network adapters (CNAs) and Ethernet switches in the Dell Networking product line that support DCB for iSCSI. A complete listing of all Ethernet switches and CNAs validated for DCB compliance with Dell PS Series can be found in the [Dell Storage Compatibility Matrix](#).

1.1 Objective

This document provides an overview of the DCB requirements for PS Series arrays and assists with identifying and resolving basic DCB configuration issues in the Dell PS Series SAN. This information may be used to ensure that DCB is properly enabled across all devices in the SAN or to ensure that DCB is effectively disabled if it is not required.

Note: For an in-depth tutorial of DCB, refer to the Dell white paper, *Data Center Bridging: Standards, Behavioral Requirements, and Configuration Guidelines with Dell EqualLogic iSCSI SANs* at <http://en.community.dell.com/dell-groups/dtcmedia/m/mediagallery/20283700>

1.2 Audience

This document is primarily targeted at storage and networking administrators who are responsible for the design, deployment, and maintenance of Dell PS Series SANs and related system components. It is assumed that the reader has operational and administrative knowledge of PS Series storage, iSCSI SAN network design, and related PS Series networking best practices. For a comprehensive listing of Dell PS Series networking best practices, please refer to the [Dell PS Series Configuration Guide](#)

2 Data Center Bridging (DCB) Explained

A key data center resource is the network infrastructure that interconnects various devices. These devices include server hardware systems that host enterprise applications and storage systems that host application data. Data Center Bridging (DCB) enables sharing the same network infrastructure between multiple traffic types such as server application traffic and storage data traffic. DCB provides network resource sharing mechanisms including bandwidth allocation and lossless traffic. This allows converged network implementations to carry multiple traffic types on the same network infrastructure with fair resource sharing. This white paper covers I/O convergence using DCB for Dell PS Series iSCSI storage.

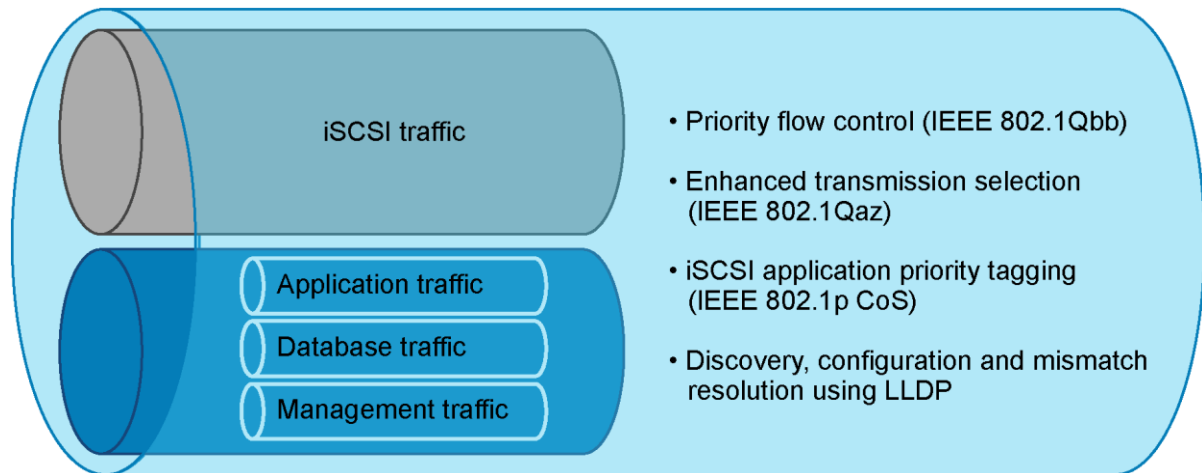


Figure 1 Converged traffic using DCB through single 10GB pipe

2.1 DCB requirements

It is required that all devices in the Dell PS Series SAN support DCB for iSCSI when this functionality is enabled. If any device in the SAN does not support DCB, then DCB must be disabled at the switches for the entire SAN. Once all devices in the SAN are DCB compliant, then DCB can be re-enabled. Devices that are designated as DCB Supported in the *Dell Storage Compatibility Matrix* have been fully validated by Dell to ensure compatibility for PS Series SANs. In deployments that contain devices not listed in the *Dell Storage Compatibility Matrix*, ensure that all Ethernet switches and CNAs in the SAN adhere to the following DCB standards:

- **VLAN tagging** within an iSCSI SAN where DCB configuration values are embedded within the Ethernet header (IEEE 802.1Q).
- **Data Center Bridging Exchange (DCBX):** DCB protocol performs discovery, configuration and mismatch resolution using the Link Layer Discovery Protocol (IEEE 802.1Qaz).
- **iSCSI application priority:** Support for the iSCSI protocol in the application priority DCBX Type Length Value (TLV). Advertises the priority value (IEEE 802.1p CoS, PCP field in VLAN tag) for iSCSI protocol. End devices identify and tag Ethernet frames containing iSCSI data with this priority value.
- **Enhanced Transmission Selection (ETS):** (IEEE 802.1Qaz) Provides minimum, guaranteed bandwidth allocation per traffic class/priority group during congestion and permits additional bandwidth allocation during non-congestion.

- **Priority-based Flow Control (PFC):** (IEEE 802.1Qbb) Independent traffic priority pausing and enablement of lossless packet buffers/queueing for iSCSI.

2.2 DCB Terminology

The following sub-sections explain the various DCB standards in more detail and why they are needed within a DCB enabled SAN.

2.2.1 Priority-based Flow Control

Priority-based Flow control (PFC) is an evolution of the concept of Flow Control originally implemented in the MAC Pause feature of Ethernet (IEEE 802.3x). The MAC Pause feature is a simplistic control of traffic made by requesting that the sender stop transmitting for a specific period of time. With no granularity applied to this request, all Ethernet frames are stopped.

PFC also asks the sender to stop transmitting, but it in addition leverages classes of traffic to apply granularity to the process. In a DCB environment, all traffic is tagged with a Class of Service (CoS) using the VLAN Q-tag. PFC can then request that a specific CoS be paused for a time, while other classes can continue unhindered as shown in Figure 2. In a storage environment, this may mean that other TCP/IP traffic is paused, while storage traffic tagged with a higher priority can be managed using PFC.

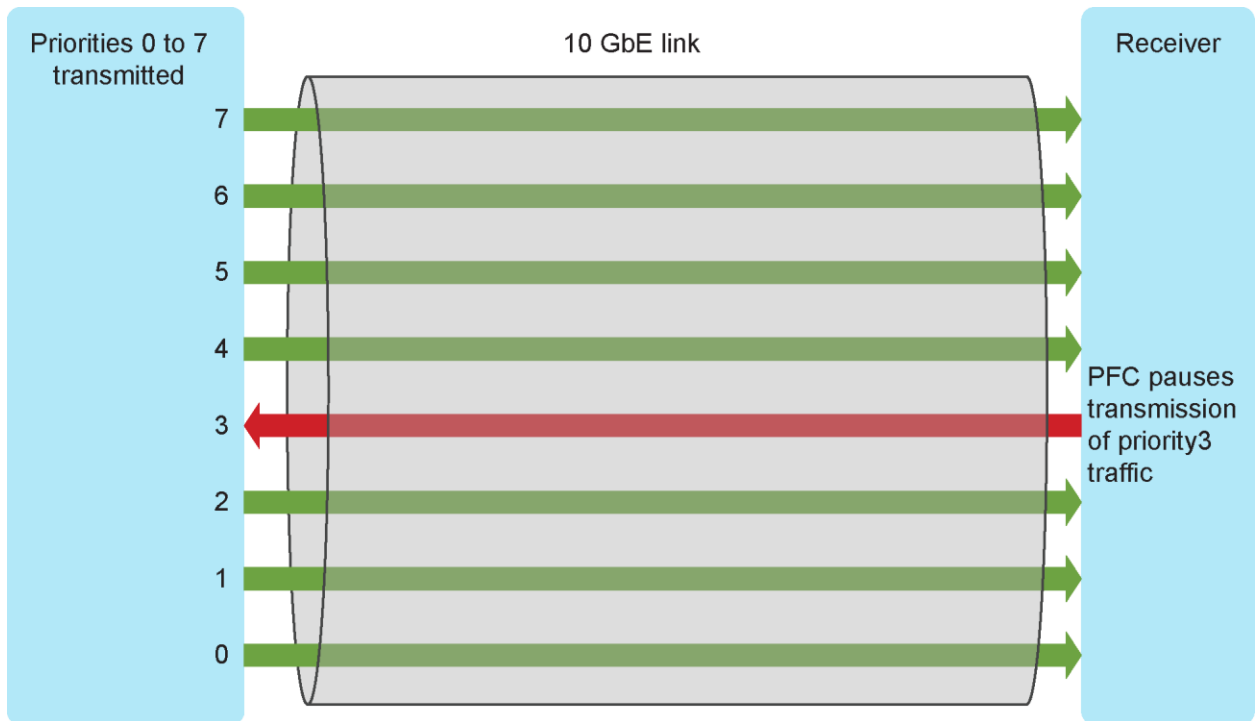


Figure 2 Example PFC diagram

2.2.2 Enhanced Transmission Selection

Enhanced Transmission Selection (ETS) is a mechanism for guaranteeing a percentage of bandwidth to a traffic class (TC). A traffic class contains one or more Classes of Service from the VLAN Q-tag. Each traffic class is then assigned a percentage of bandwidth with a granularity of 1%. All traffic class bandwidths must add up to 100%; oversubscription is not allowed.

The bandwidth percentage defined is a minimum guaranteed bandwidth for that traffic class. If a traffic class is not using its entire minimum amount, it can be utilized by other traffic classes that may need it. However, as soon as the original traffic class requires its bandwidth again, the other traffic flow must be throttled to allow the bandwidth to be recovered. This is accomplished through the use of PFC discussed earlier. PFC will issue a pause for the overreaching traffic classes in a manner to allow the bandwidth to be regained with a minimum number of dropped frames for the throttled traffic class.

Note: An important note to consider when deciding on bandwidth percentages for each traffic class is that the required accuracy of the ETS algorithm is only plus or minus 10%.

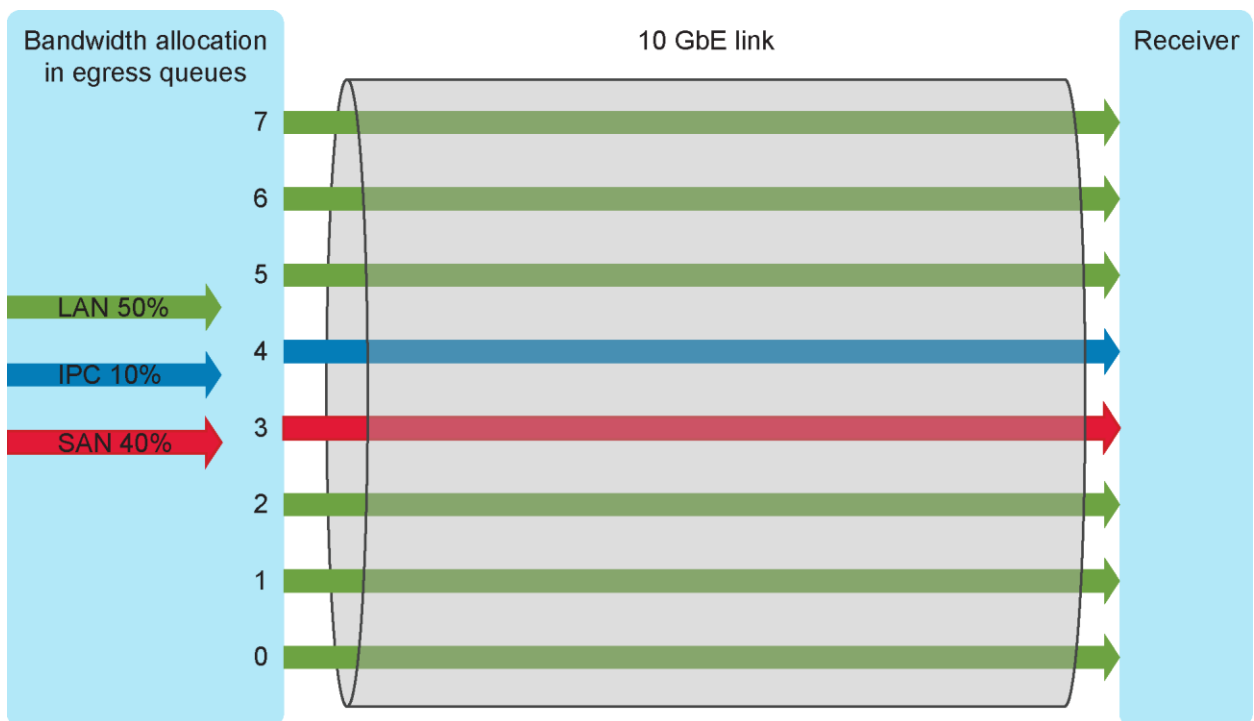


Figure 3 Example ETS diagram

2.2.3 Application priority configuration

Application priority is a DCB configuration construct which defines the mapping of an application or protocol to a priority (PCP) value. All frames belonging to a particular application or protocol are tagged with a priority value (0 to 7) which in turn gets mapped to a traffic class or priority group (PG). This mapping provides dedicated bandwidth for a protocol and also facilitates PFC with lossless behavior for a protocol across the network. The application or protocol is identified by TCP/UDP port numbers or by an Ethertype number. For example, iSCSI protocol is identified by TCP port 3260 and FCoE protocol is identified by Ethertype 0x8906.

Protocols identified by this method can be mapped to a priority value using the application priority feature. For example, iSCSI identified by TCP port number 3260 can be mapped to priority 4. An end device which is the source of a frame carrying iSCSI payload marks the frame with priority 4 in the VLAN tag. All devices in the network forward the frame with this priority value and process it with the same configured PFC lossless behavior and PG/TC bandwidth settings for this priority. The default priority value is 0 and any network traffic not assigned a priority value using the application priority feature will have priority 0 marked in the frame.

2.2.4 Data Center Bridging Capability Exchange

Datacenter Bridging Capability Exchange (DCBx) is an extension of the IEEE standard 802.1AB for Link Layer Discovery Protocol (LLDP). It uses the existing LLDP framework for network devices to advertise their identity and capabilities. LLDP relies on the use of Type-Length-Values (TLV) to advertise the device capabilities for a multitude of Ethernet functions, as well as its identity. DCBx defines new TLVs specific to the DCB functionalities.

PFC and ETS have specific TLVs defining items such as:

- Whether PFC is to be used
- Priorities that will be managed using PFC
- Which priorities belong to a specific traffic class
- Bandwidth minimums for each defined traffic class

The standard also defines Application TLVs. These TLVs allow for the definition of which protocols will be managed, as well as the priorities to be assigned to each. Currently FCoE and iSCSI have Application TLVs.

DCBX allows mismatch detection and remote configuration of DCB parameters. Each peer port on a link initially advertises its locally configured DCB parameters. It then verifies the advertised configuration parameters from the remote peer and detects mismatches. DCBX also allows remote configuration of a peer port with the willing mode feature. PS Series arrays are configured, by default, to always be in willing mode. A port may advertise that it is willing to apply DCB configuration parameters from a peer port. The advertised configuration of the peer port is then accepted as the local configuration for operation. For example an end device such as a server CNA or a storage port in willing mode accepts and operates with the switch recommended DCB parameters. Typically most server CNAs and storage ports are in willing mode by default. This enables the network switches to be the source of DCB configuration. This model of DCB operation with end devices in willing mode is recommended and prevalent due to ease of configuration and operation by minimizing mismatches.

For Dell PS Series environments using DCB, support for the iSCSI TLV is required. For iSCSI, the end station needs to know how to identify iSCSI frames from the other TCP/IP traffic. The iSCSI TLV identifies what TCP port iSCSI traffic is using, so that the end station can properly assign the class of service to it. Once this has been outlined, PFC and ETS can manage the iSCSI traffic as its own class.

Note: PS Series arrays always use port 3260 for iSCSI traffic.

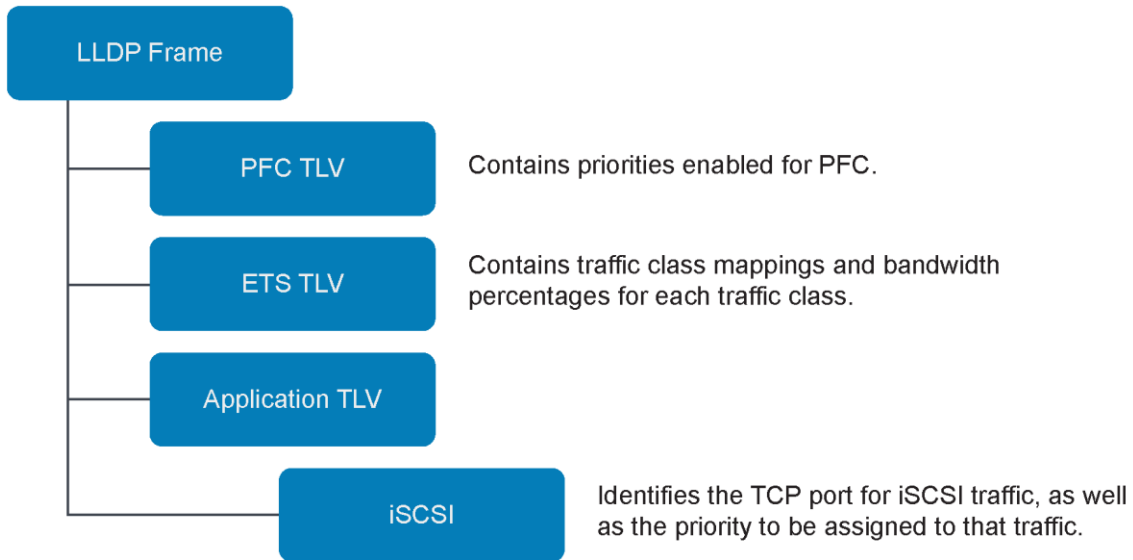


Figure 4 LLDP hierarchy supporting DCBx

2.3 DCB parameter configuration

The key DCB technologies discussed in section 2.2 are applicable peer to peer on an Ethernet link between two ports. A link comprised of peer ports may be between a NIC/CNA and a switch, between two switches, or between a switch and a storage device. A successful converged network deployment based on DCB includes proper configuration and operation of DCB parameters on all peer device ports that carry converged traffic. DCB parameters include PFC, ETS, and application priority mapping as described in section 2.2.

2.3.1 End-device auto configuration

The DCBX willing mode on end devices enables automatic configuration of end devices and minimizes mismatches on DCB parameters. In this model of configuration, network switches are configured with the required DCB parameters and they advertise the configuration through DCBX to attached end devices. The end devices operating in willing mode learn the DCB configuration information from DCBX and operate with those parameters. The TLV fields within the LLDP protocol are used by DCBX for communicating the DCB parameters. The DCBX TLV exchange between peer ports is illustrated in Figure 5, Figure 6, and Figure 7 below. It is a simple example showing only one port per device to illustrate the DCBX exchange. The server CNA and storage ports are in willing mode (willing = 1) to accept DCB parameters from switch port. The switch ports are configured with DCB parameters as per deployment requirements and operate with the willing mode off (willing = 0) as shown in Figure 5. In the figures, DCBX TLVs include PFC TLV, PG/ETS TLV, and application priority TLV.

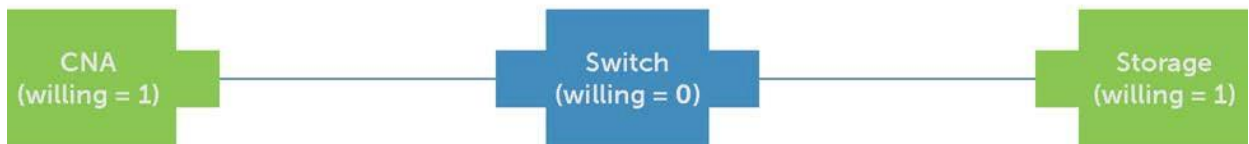


Figure 5 Step 1 – Willing mode state of switch, CNA, and storage ports

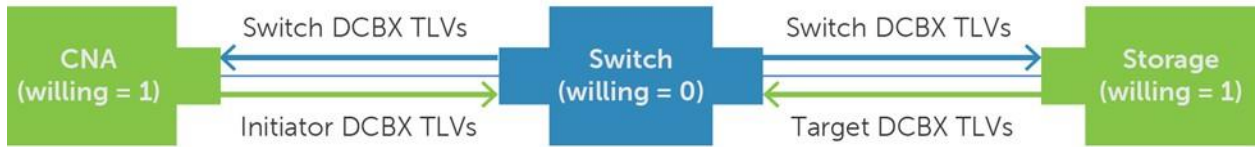
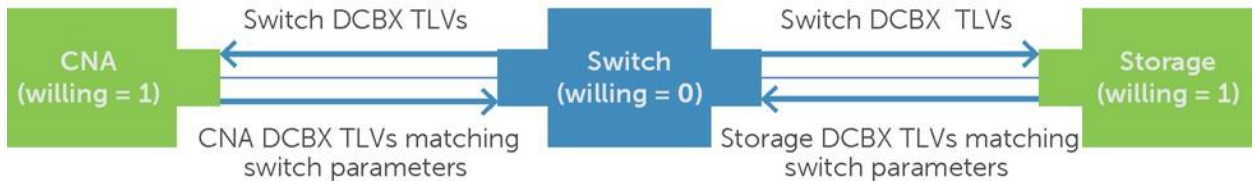


Figure 6 Step 2 – DCBX TLVs exchange between peer ports on the operational link



DCBX TLVs = PFC TLV, PG/ETS TLV, Application priority TLV

Figure 7 Step 3 – CNA and storage port accept and confirm switch DCB parameters

In step 2 depicted in Figure 6, peer ports communicate their local DCB configuration through DCBX and their willing mode. The server CNA and storage ports initially have a default local DCB configuration that they communicate. In step 3 depicted in Figure 7, they apply the switch advertised configuration locally and communicate the same parameters back to the switch.

Mismatches do not occur when the server NIC/CNAs and the storage devices support all the requisite parameters configured on the switch. This can be ensured by deploying only components that meet the converged network design requirements.

2.3.2 Switch to switch configuration

A DCB enabled network typically consists of multiple switches for redundancy and scalability. There are two possible ways to configure the network switches with the required DCB parameters. The methods are:

- Configuration propagation: DCB parameters are manually configured on two or more switches (source switches) and they advertise the configuration through DCBX to other switches in the network. Other switches have a certain number of ports facing the source switches called upstream ports. These ports operate in DCBX willing mode, learn the DCB parameters from an upstream device, and internally propagate the DCB configuration to other ports (downstream ports) on the switch. The downstream ports apply the internally propagated configuration and advertise the same to peer devices.
- Manual configuration: All switches on the network are manually configured with the same DCB parameters. The switch ports are configured with willing mode turned off and they only advertise DCBX configuration to peer devices.

2.4 Converged network example

This section discusses a sample converged network implementation using DCB. The various DCB parameters (PFC, PG/TC, and application priority) are illustrated in the context of a deployment scenario. The scenario uses the hypothetical requirements given below:

- A virtualized server environment that has requirements to converge both LAN and SAN traffic on the same physical Ethernet infrastructure (Server adapters and Ethernet switches).
- Server NICs/CNAs and Ethernet Switches are 10GbE.
- LAN traffic segregation requirements on virtualized servers:
 - Separate logical network for hypervisor LAN traffic with bandwidth allocated for all LAN based traffic.
 - All LAN traffic can be lossy from a Layer 2 perspective. Lossy implies that the switches and adapters can discard frames under congestion. Upper layer protocols such as TCP will perform recovery of lost frames by re-sending data.
- SAN traffic requirements on virtualized servers:
 - A separate logical network for iSCSI SAN traffic with dedicated bandwidth allocation.
 - All SAN traffic must be lossless from a Layer 2 perspective (PFC enabled).

Figure 8 illustrates the configuration values for various DCB parameters (PFC, PG/TC, and Application Priority) for this deployment scenario and their relationship. The table columns in Figure 8 that are highlighted in blue are descriptive text that is not part of the actual DCB parameters. The network switch ports are configured with the DCB parameters to support the requirements listed above and advertised to peer device ports, as follows:

- The first step is to assign the priority value for each traffic type by using the application priority mapping. This information is advertised as part of the DCBX Application TLV. The Application priority mapping table in the figure shows the priority mapping for iSCSI protocol. Priority 4 is the industry accepted value for iSCSI protocol and implies that all frames tagged with priority 4 carry iSCSI as payload. To ensure this, the CNA and storage must support iSCSI protocol in the application priority TLV and tag outgoing iSCSI frames with the configured priority. All other un-configured traffic types are mapped to default priority 0. In this example, LAN traffic is mapped to default priority 0 since it is not explicitly configured.
- The next step is to enable PFC (to enable lossless behavior) for iSCSI traffic on priority 4 and disable PFC for LAN traffic on priority 0 since it can be lossy. An example of priority values along with the PFC requirement is shown in the PFC table. The PFC enabled/disabled information for the priorities is advertised as part of the DCBX PFC TLV.

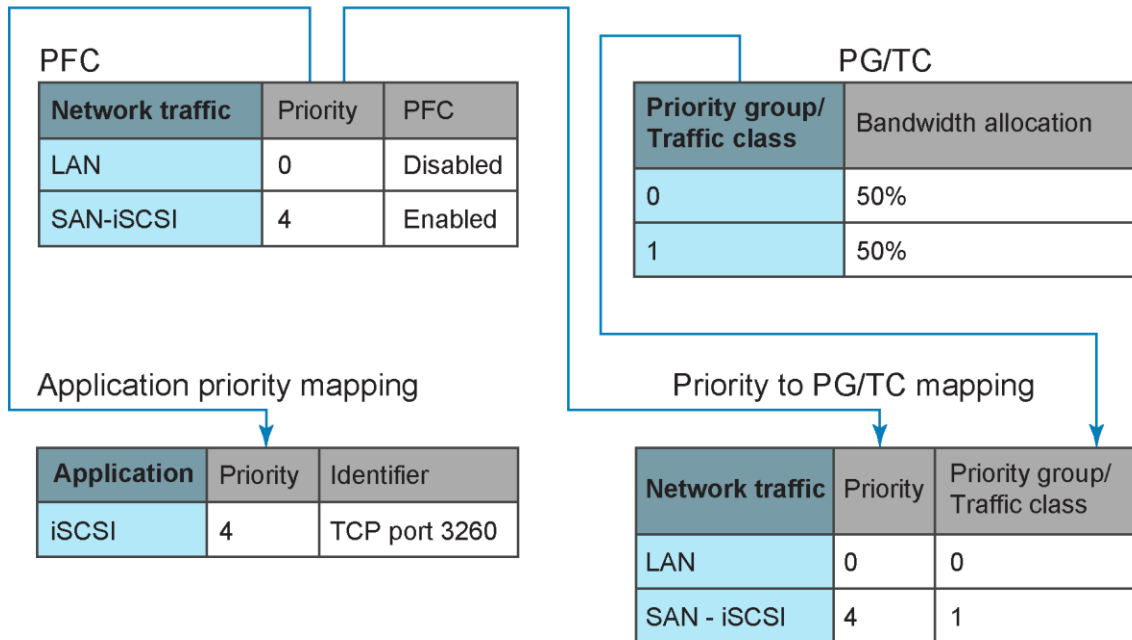


Figure 8 DCB parameters

- The priority groups or traffic classes are the final step. The PG/TC table in the figure shows the priority group or traffic classes and the respective bandwidth percentage assignment. The Priority to PG/TC mapping table shows the priorities that are mapped to priority groups or traffic classes. This information is advertised as part of the DCBX ETS or PG TLV. It must be noted that multiple priorities can be mapped to the same PG or TC. In that case, all those priorities will share the same bandwidth allocation and will not have dedicated bandwidth. In this example, we mapped iSCSI priority to its own PG/TC to provide dedicated bandwidth.

2.5 DCB configuration process overview

Important: With all network configuration changes, there is the possibility of service interruptions. Network changes required to properly configure DCB for PS Series Storage will result in a temporary loss of connectivity. Therefore, Dell strongly recommends that all network environment changes are performed during a planned maintenance window.

This section provides general steps to configure DCB for PS Series and section 3.2 contains detailed steps for specific Dell switch models. These steps apply to both new deployments and those deployments that have DCB improperly configured. It is important to verify that all components in the SAN are listed in the *Dell Storage Compatibility Matrix* as DCB Supported or that the components support all PS Series requirements for DCB. If multiple switches are in the path between iSCSI Hosts and PS Series arrays, then all configuration steps below must be applied to these switches. For SAN deployments with two or more switch hops, please refer to *Data Center Bridging: Standards, Behavioral Requirements, and Configuration Guidelines with Dell EqualLogic iSCSI SANs* for additional configuration guidance and considerations at:

<http://en.community.dell.com/dell-groups/dtcmmedia/m/mediagallery/20283700>

When configuring DCB on a new PS Series array member, ensure the switch is setup properly first. The PS Series CLI setup wizard or the PS Series Remote Setup Wizard will detect whether the switch is DCB capable and prompt for the required configuration information, including the VLAN ID, during setup.

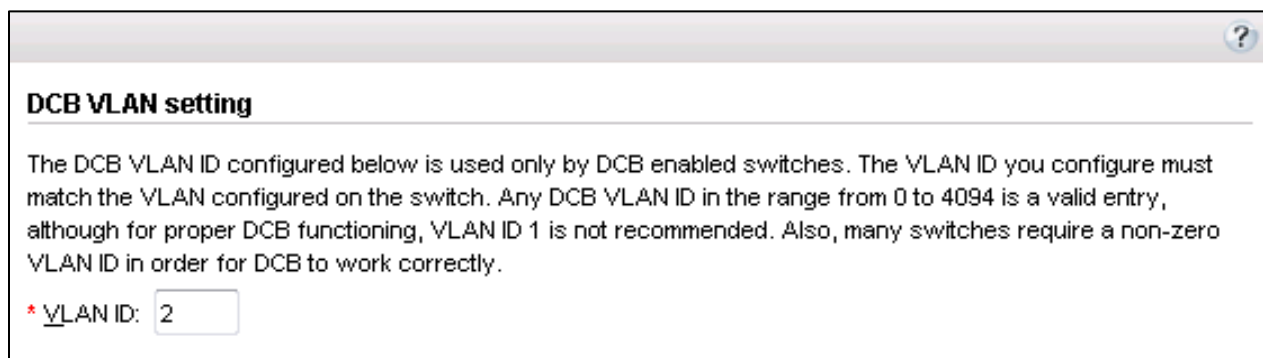
To configure DCB on an existing PS Series array member:

1. Disable DCB on the switch, if it is enabled.
2. Verify that **all** CNAs are enabled and configured in DCB willing mode.
3. Verify that the PS Series array has DCB enabled (Note: This is enabled by default and should never be disabled).
4. Enter a VLAN ID from 2-4096 in Group Manager. With firmware v7 and later, the VLAN ID will default to 2 when DCB is detected on the switch.
5. On the switch, create a VLAN in trunk mode for all switch ports connected to arrays and iSCSI hosts with the same VLAN ID configured on the array.
6. Configure all CNAs with same VLAN ID configured on the array.
7. On the switch, configure Priority Groups/ETS, PFC, and iSCSI application priority on all ports connected to PS Series arrays and iSCSI host CNAs. The same should be configured for any switch interconnects (LAGs, VLTi, vPC, and others)
8. Re-enable DCB on the switch.
9. Verify all DCB configuration and operational parameters on array ports through Group Manager.
10. Verify DCB operational parameters on CNAs using the respective management application.

Steps 1, 3, 4, 7, 8, and 9 are outlined in the following section. Section 4.1 provides additional guidance for steps 2, 6, and 10 using the QLogic 57810S CNA.

2.5.1 Enabling DCB and configuring the VLAN

The DCB VLAN setting can be found in the Group Manager **Summary** page, under the **Advanced** tab (as shown below).



DCB VLAN setting

The DCB VLAN ID configured below is used only by DCB enabled switches. The VLAN ID you configure must match the VLAN configured on the switch. Any DCB VLAN ID in the range from 0 to 4094 is a valid entry, although for proper DCB functioning, VLAN ID 1 is not recommended. Also, many switches require a non-zero VLAN ID in order for DCB to work correctly.

* VLAN ID:

Figure 9 DCB VLAN setting from EqualLogic Group Manager

Note: With PS Series firmware version 7.x and later, DCB is always enabled and in willing mode.

Next, a valid VLAN ID must be entered. The range of recommended values is 2-4096. However, verify that the VLAN ID choice does not conflict with any default or reserved IDs for your specific switch make and model. With firmware v7 and later, the VLAN ID will default to 2 when DCB is detected on the switch.

Important: When DCB is enabled on the switch, it is necessary to configure a non-default VLAN on the array, switch, and all host ports that are part of the PS Series SAN. VLAN IDs 0 and 1 should not be used as these may be the default or reserved VLAN for some switches, and as such, may forward frames untagged (with no VLAN tagging). VLAN tagging is required to fully support DCB.

2.5.2 Switch configuration

For switch configuration steps, refer to Section 3.2 and find your specific switch model. If the specific switch is not listed, consult with your switch vendor to confirm the DCB support requirements for PS Series and obtain instructions for configuring (or disabling) DCB on the switch as needed.

2.5.3 Disabling DCB on the SAN

When DCB functionality is not required for the SAN deployment, DCB must be disabled on all switches in the SAN. Once DCB is disabled on all switches, the attached PS Series arrays and CNAs will no longer receive DCBx messages from the attached switches, effectively disabling DCB functionality on the SAN completely. Configuration steps for disabling DCB can also be found in Section 3.2.

Important: If DCB is disabled on the switch, ensure that 802.3x flow control is enabled on all switch ports attached to the PS Series arrays and iSCSI hosts. Dell PS Series best practices recommend a minimum of RX flow control on the switch ports. Also it is assumed that the switches are dedicated for SAN traffic alone when DCB is disabled.

Important: If DCB is disabled on a switch, then all DCB functionality will be disabled, including support for other protocols such as FCoE.

Note: VLAN tagging is only supported when DCB mode is fully enabled and configured. If DCB is disabled on the switch, switch ports must be configured as untagged or as access ports (or left in the native VLAN).

2.5.4 DCB configuration verification

The following sections show how to verify the DCB configuration of a switch through the PS Series Group Manager. Equivalent methods for verifying the configuration via the command line interface (CLI) are provided in Appendix A.

2.5.5 Priority Group, ETS, and PFC verification

To determine what configuration steps in Section 2.5 are needed to properly configure switches and correct the error condition, verify the following information in Group Manager.

1. In the PS Series Group Manager, click a member array and go to the **Network** tab.
2. Right-click on each 10 Gb interface (for example **eth0**) and click **DCB Details** to display the dialog box shown below.

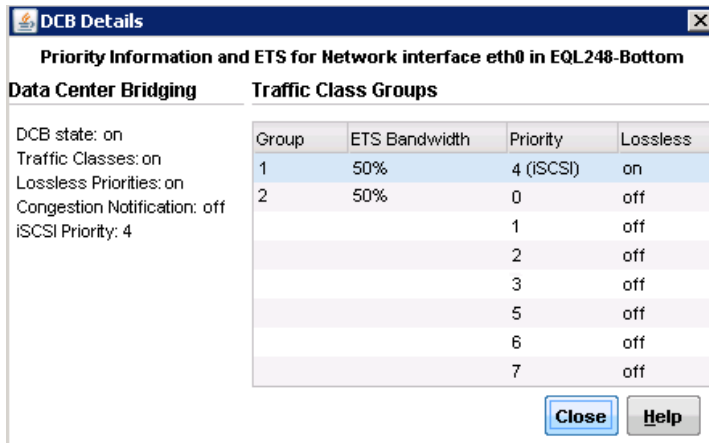


Figure 10 EqualLogic Group Manager NIC DCB Details

3. From these details, verify the following:
 - a. Column 1: For the iSCSI Traffic Class Group (Group 1, in this example), only iSCSI is present (no other traffic classes should appear, such as FCoE).
 - b. Column 2: ETS % from 1 to 100% (recommend 50% as an initial setting).
 - c. Column 3: Priority level for iSCSI is 4 (recommended).
 - d. Column 4: Lossless state for iSCSI equals PFC ON for priority 4.
4. Repeat this verification for all 10 Gb array interfaces in the group.

2.5.6 Invalid DCB configuration detection

PS Series firmware automatically detects if there is an invalid configuration state on the Ethernet switch attached to the PS Series array due to the following conditions:

- PFC is configured (PFC TLV received)
- iSCSI is using a priority group which is not PFC enabled (lossless)

This state will result in a warning message that will be displayed at the bottom of the EqualLogic Group Manager interface as shown in Figure 11.



Figure 11 Invalid DCB configuration detection warning message in Group Manager v7

If a switch does not support the iSCSI TLV, then iSCSI DCB is not possible as the switch would only allow prioritizing TCP traffic. In this case, the array will place all iSCSI frames in default priority 0. Typically, priority

0 will not have PFC enabled on the switch, which means that iSCSI traffic would be treated with the same priority as all other traffic or in some cases, with a lower priority. It is also possible that even when the iSCSI TLV is supported by the switch, the iSCSI priority may not have PFC enabled for that priority. Upon detection of either of these conditions, the array will issue a warning message that Ethernet flow control has been disabled on one or more interfaces. Operation with no flow control may cause performance degradation and service interruptions.

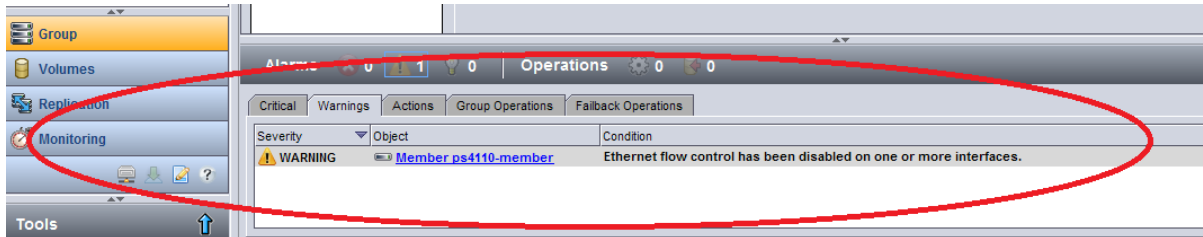


Figure 12 Invalid DCB configuration detection warning message in Group Manager v6

In addition to this warning message, more detailed warning messages will appear in the event log and CLI.

3 Network configuration and deployment topologies

This section discusses detailed DCB parameter configuration for network deployments with PS Series storage. It includes the configuration process for multi-layered switching environments and discusses sample topologies.

Note: The term initiator in this document refers to an iSCSI initiator which is one of the following:

- A NIC in conjunction with an OS software initiator
- A dependent hardware initiator (CNA) that offloads iSCSI in conjunction with OS software initiator support for configuration and session management
- An independent hardware initiator (CNA) that offloads iSCSI completely with no OS support

3.1 General configuration process flow

The general configuration steps for setting up a DCB network infrastructure with PS Series storage are given below. The assumptions for the discussion in this section are:

- iSCSI traffic is prioritized in its own traffic class or priority group
- iSCSI traffic is PFC enabled and all other traffic is not
- iSCSI traffic is in a tagged VLAN ID and prioritized with priority 4
- All other traffic flow is in the default traffic class or priority group
- All other traffic is untagged with default priority 0

The DCB configuration sequence involves the following steps:

1. Configure at least two top layer switches within the Layer 2 DCB network domain administratively with the required DCB parameters and DCBX willing mode turned off on all ports. These become the source switches. At least two switches are required since if one switch is shut down or fails, another is available as a configuration source. Only the ports on these switches that are part of the Layer 2 DCB domain need to be configured with DCB.
2. Configure intermediate layer switches and other peer switches:
 - a. Configuration propagation: If this mechanism is supported by switch model, then peer switches, downstream switches, and I/O Aggregator modules should accept the DCBX configuration from the source switches through certain Inter Switch Link (ISL) ports set as upstream ports capable of receiving DCBX configuration from peers (enabled with willing mode turned on the ports). In this scenario, a port is a single switch interface or a LAG port-channel. These upstream ports internally propagate DCBX configuration to other ports and ISLs on the same switch. Other ports and ISLs are set as downstream ports capable of recommending DCBX configuration to their peer ports (enabled with willing mode turned off). The downstream ports apply internally propagated DCBX configuration from upstream ports and recommend them to connected edge devices and other switches. The upstream and downstream ports on a switch are administratively identified based on network topology and are then configured manually on each switch based on the topology. This identification and configuration is automatic for certain switching modules such as I/O aggregators that are on the edge of the network connecting to only edge devices on the downstream ports. In addition, when multiple upstream ports are configured, an internal

- arbitration process is defined for one of them to propagate DCBX configuration internally and the other ports accept that configuration.
- b. Manual configuration: If this internal propagation method is not supported by your switch model, then peer and intermediate layer switches must be manually configured with the same DCB parameters as the source switches with willing mode off across all ports.
- 3. Configure and verify edge device DCBX parameters.
 - a. Configure edge device ports with the VLAN ID configured for iSCSI in the switches.
 - b. Verify operational DCBX parameters on the initiator/target storage ports and peer switch ports.
 - c. If mismatches occur, ensure edge device ports are in willing mode and they are capable of supporting the DCB parameter requirements.

The configuration propagation model and DCBX exchanges using the willing bit is illustrated in Figure 13 below. The target storage is PS Series storage and the initiator hosts are Dell PowerEdge servers installed with DCB capable NICs/CNAs. In the figure, SI links stands for Switch Interconnect links.

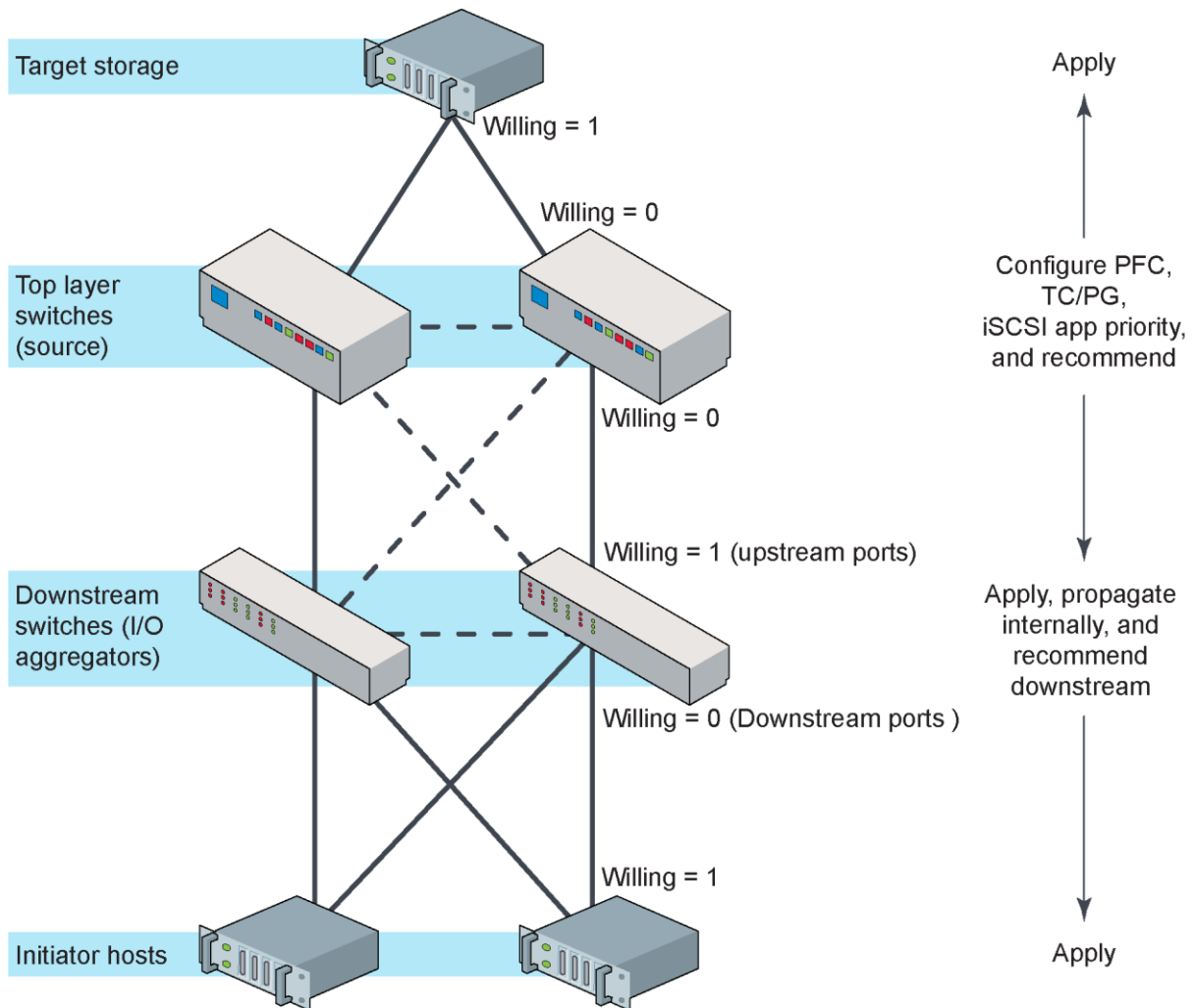


Figure 13 DCBX configuration propagation in intermediate switch layer

Note: The upstream port method can be used to configure other peer switches in the same top layer. A pair of source switches can be configured with DCBX parameters and other switches in the same layer can accept configuration from the source switches through the inter switch links (Willing mode on for ISLs in the same layer). All switches in this layer then pass on configuration to downstream links and edge devices like initiators and targets (willing mode off for these ports and downlinks).

The manual DCB configuration across all switches is illustrated in Figure 14 below.

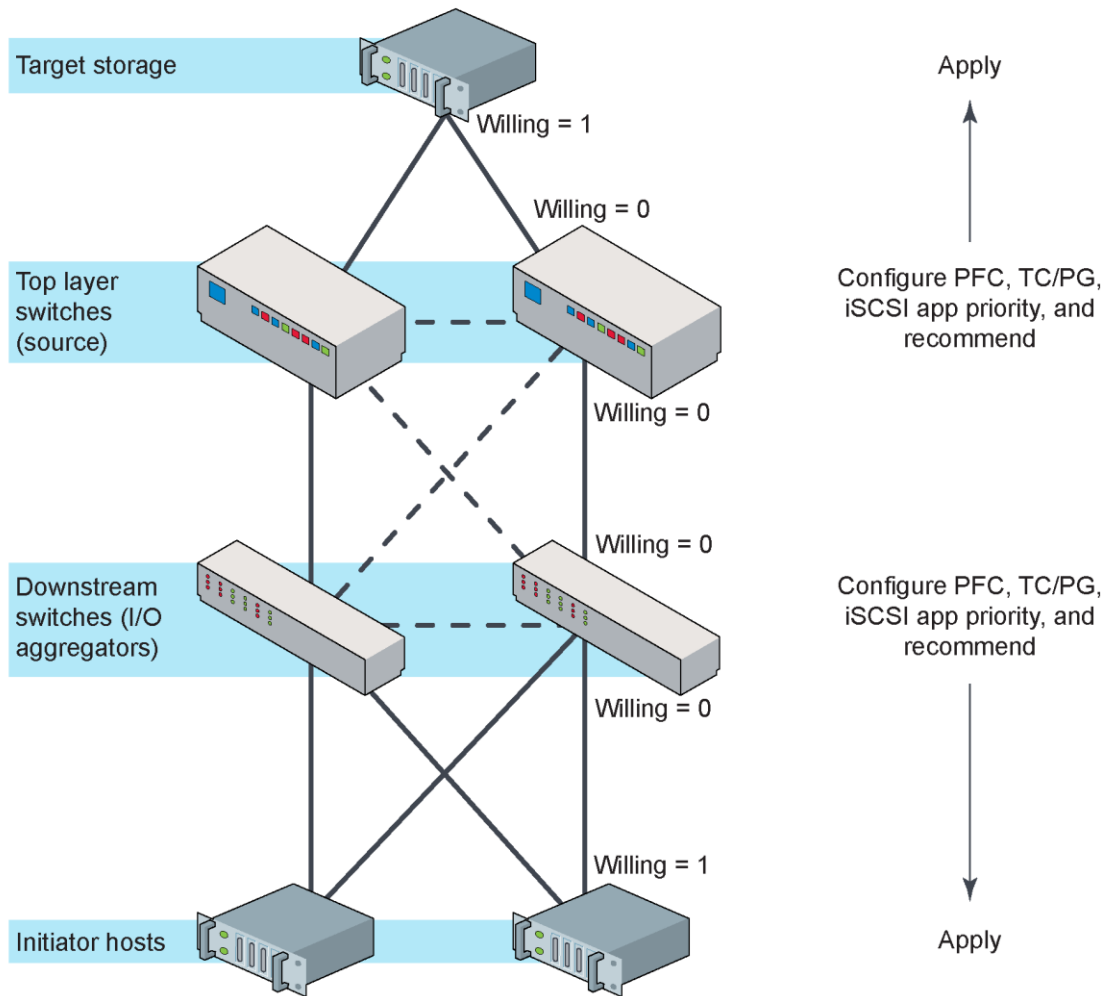


Figure 14 Manual DCB configuration across all switch layers

Network configuration assumptions:

- All switches in the Layer 2 network carrying iSCSI traffic must be configured to be either DCB enabled or DCB disabled. Mixed configuration with a partial set of switches in non-DCB mode should not exist in the Layer 2 DCB network domain. The administrator must ensure that all switches are either in DCB enabled state or DCB disabled state on the network during the configuration process.

- For switches that support the DCB propagation mechanism: Switch inter-connectivity in the Layer 2 network must be designed so that there is no single point of failure that would cause a disjoint network. When designed properly, the failure of a single switch device, link, or port interface should not cause a disjoint network. The administrator must ensure that network redundancy design best practices are followed to avoid disjoint networks.

Important: With all network configuration changes, there is a possibility of service interruptions. Network changes required to properly configure DCB will result in a temporary loss of connectivity. Therefore, Dell strongly recommends that all network environment changes are performed during a planned maintenance window. Further, when re-configuring an existing PS Series SAN to function in DCB mode, it is recommended to first configure DCB VLAN ID for iSCSI on the arrays and then configure DCB along with VLAN ID on the switch ports. This will ensure that inter-array connectivity is not lost during the process.

3.2 Configuration for switches with DCB support

The following switches feature DCB with the requirements necessary to fully support PS Series storage systems. These switches can be configured for DCB mode or as a dedicated SAN switch with standard link-level (802.3x) flow control.

3.2.1 Dell Networking S4810

Refer to the *Dell Networking S4810 Switch Configuration Guide* for instructions related to DCB or non-DCB configuration at

<http://en.community.dell.com/dell-groups/dtcmedia/m/mediagallery/20220824>

3.2.2 Dell Networking S4820T

Refer to the *Dell Networking S4820T Switch Configuration Guide* for instructions related to DCB or non-DCB configuration at

<http://en.community.dell.com/dell-groups/dtcmedia/m/mediagallery/20293376>

3.2.3 Dell Networking MXL

Refer to the *Dell Networking MXL 10/40 GbE Blade Switch Configuration Guide* for instructions related to DCB or non-DCB configuration at

<http://en.community.dell.com/dell-groups/dtcmedia/m/mediagallery/20279157>

3.2.4 PowerConnect 81XX Series

Refer to the *Dell PowerConnect 8100 Series Switch Configuration Guide* for instructions related to DCB or non-DCB configuration at

<http://en.community.dell.com/dell-groups/dtcmedia/m/mediagallery/20308559>

3.2.5 Configuration for switches that do not support DCB for PS Series

The following switches do not fully support DCB requirements for PS Series, so DCB must be disabled on these SAN switches.

Important: If DCB is disabled on a switch, then all DCB functionality will be disabled, including support for other protocols such as FCoE. Mixing of FCoE and iSCSI on the same converged fabric is not recommended and not supported by Dell.

3.2.6 PowerConnect 8024, 8024F, and M8024

The first step in configuring this line of switches is to first verify the firmware version installed is 5.1.10.1,A27 or higher.

Note: If an earlier firmware version is installed, it will not be possible to fully disable DCB on the switch and correct the invalid flow control condition on the arrays and hosts.

Upgrade instructions can be found under the corresponding product firmware download pages at <http://support.dell.com>.

To disable DCB, refer to the *Dell PowerConnect 8024, 8024F, M8024, and M8024-k Switch Configuration Guide for EqualLogic SANs* at: http://en.community.dell.com/techcenter/extras/m/white_papers/20437136

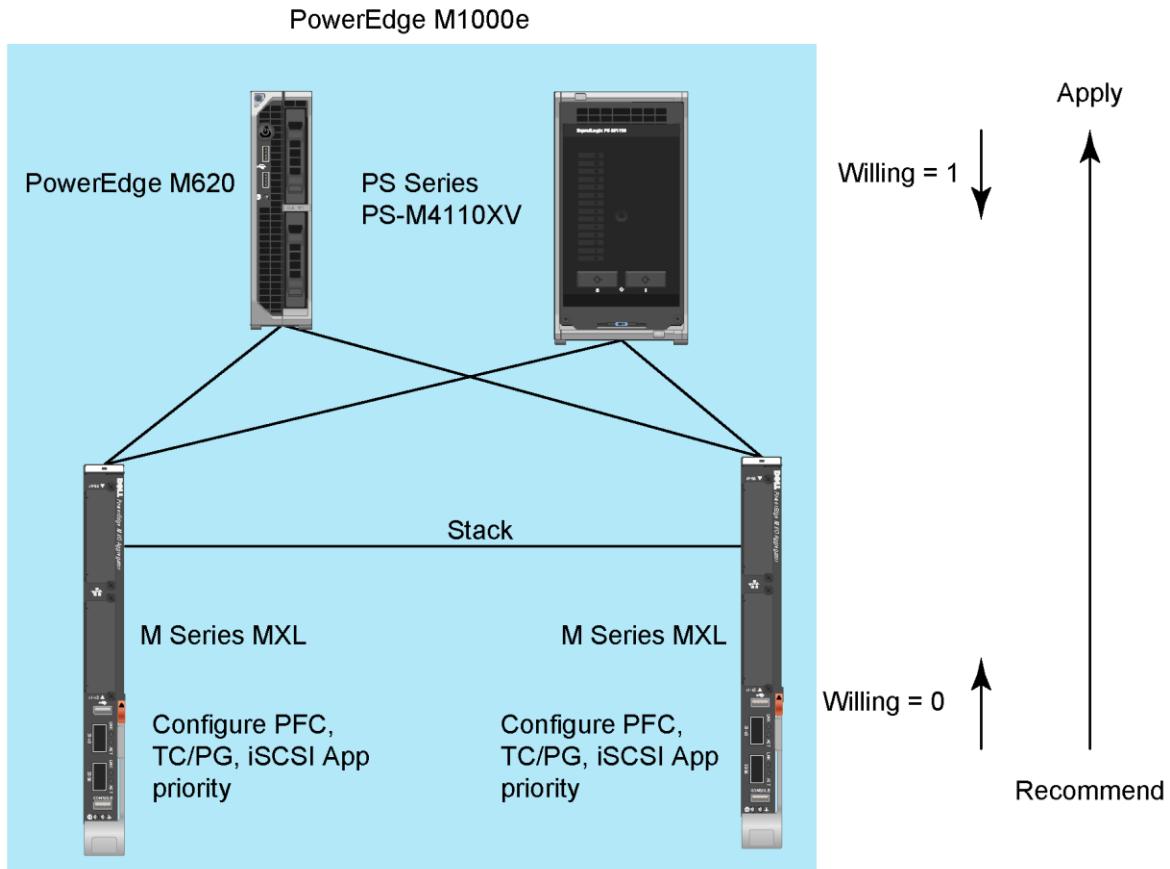
3.3 Deployment topologies

In this section, sample deployment topologies with PS Series iSCSI storage, Dell Networking switches, and Dell PowerEdge servers are demonstrated. For these topologies, it is assumed that there are no prior infrastructure constraints from the network, initiators, or target storage to be considered. The assumption is that the DCB infrastructure is a green field deployment.

3.3.1 Single layer switching with blade servers

Switch network configurations that are not hierarchical and include a single switching layer are covered in this section. This section demonstrates a deployment configuration where all the components are within a blade server chassis.

The sample configuration for this discussion includes PowerEdge M620 blade servers attached to PS-M4110XV Series storage arrays through the Dell PowerEdge MXL blade switches within the PowerEdge M1000e chassis. The switches are inter-connected through a stack and the configuration is illustrated in Figure 15.



Connectivity to LAN core not depicted

Figure 15 Single switching layer – Blade Servers

3.3.2 Single layer switching with rack servers

This section demonstrates deployment configurations with rack servers and a single layer of top of rack (ToR) switches. The sample configuration for this discussion includes PowerEdge R620 servers attached to PS6110XV Series storage arrays through the PowerConnect 8132F ToR switches. The switches are interconnected through a LAG (Link Aggregation Group) and the configuration is illustrated in Figure 16.

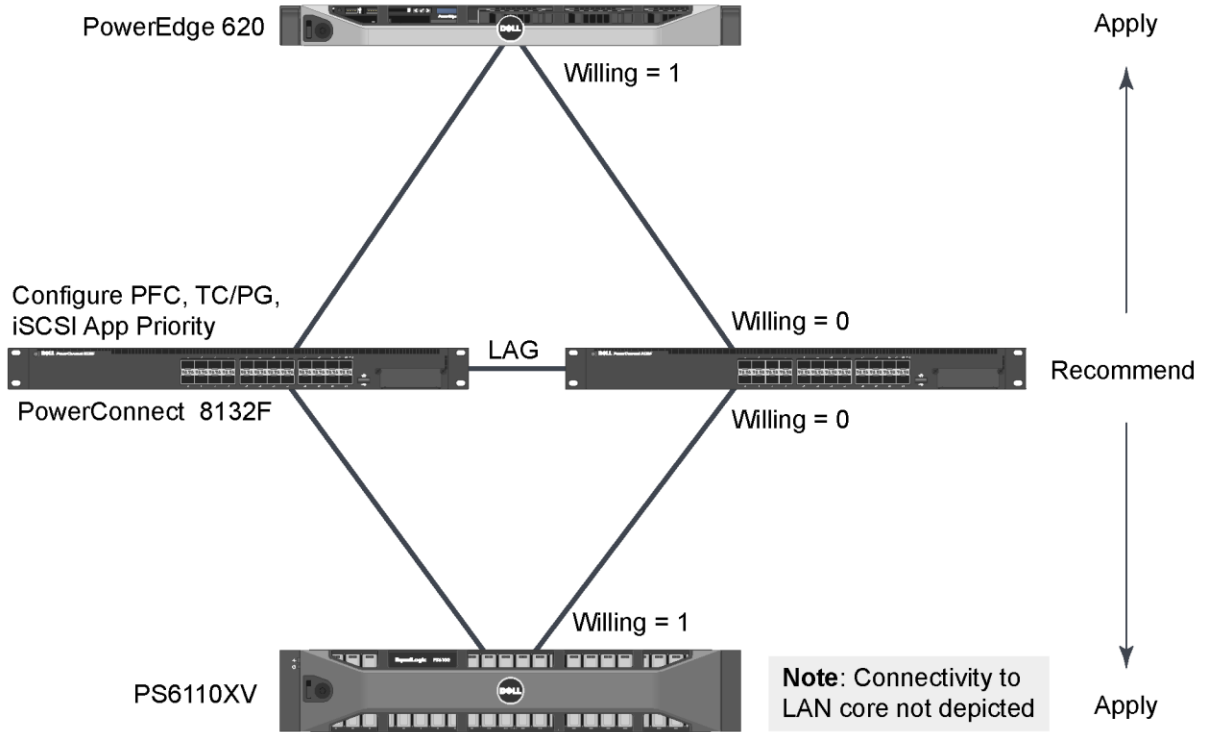


Figure 16 Single switching layer – Rack servers

In both the Blade server and Rack server configurations, the first step is to configure the switches with DCB parameters.

3.3.3 Dual layer switching

A hierarchical switching configuration with two layers of switching, one for servers and another for storage, is covered in this section. An illustration of this configuration is given in Figure 17. The configuration includes PowerEdge M620 blade servers connecting to Dell PowerEdge M I/O Aggregator modules within the PowerEdge M1000e chassis and connecting externally to S4810 Series switches. PS6110XV Series arrays are connected to the external S4810 Series switches. The S4810 Series switches are connected in VLT mode. The I/O Aggregator modules connect to the VLT switches through LACP.

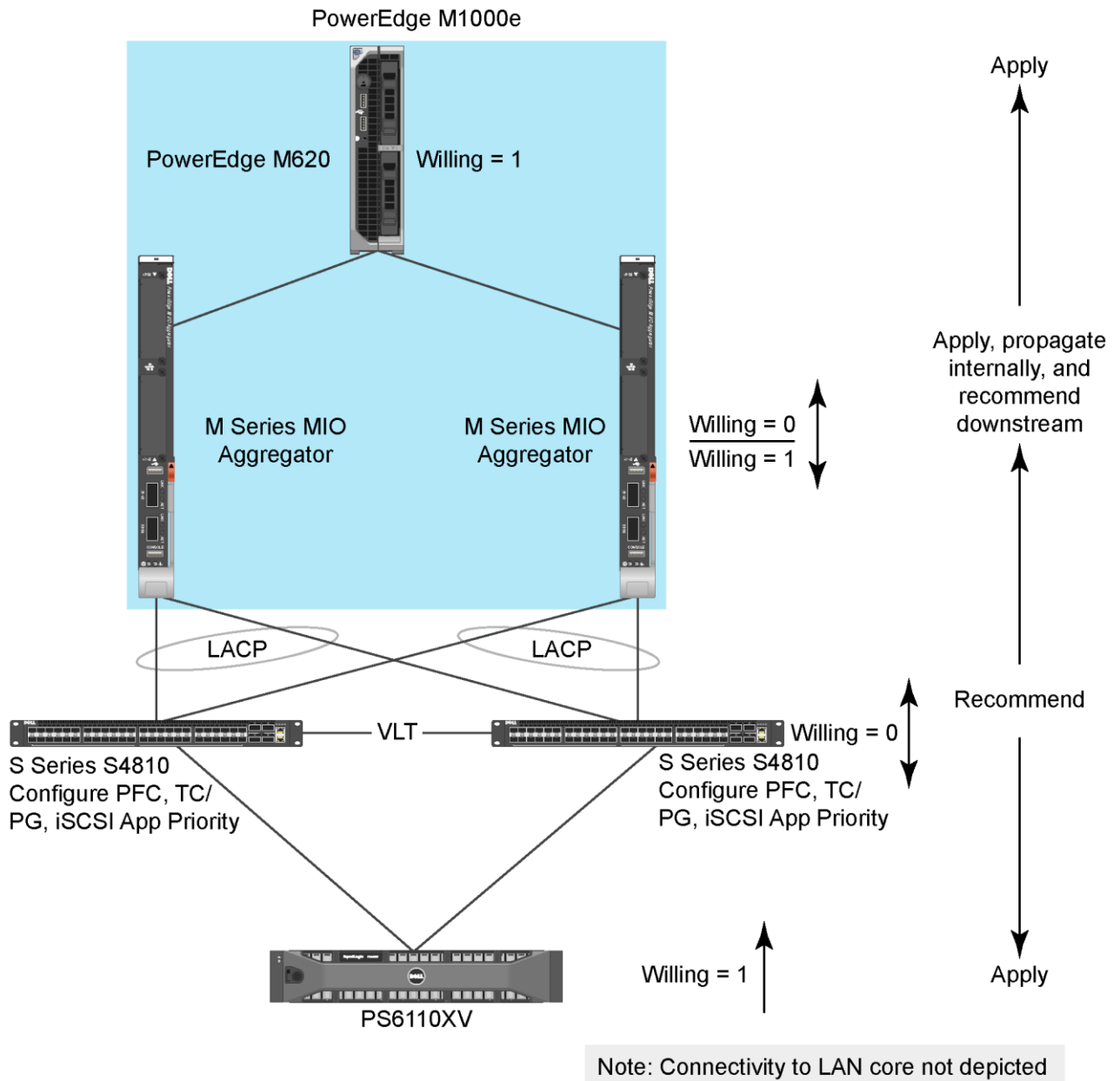


Figure 17 Dual Layer Switching

In this scenario, the first step is to configure the external S4810 switches with DCB parameters.

The Dell PowerEdge M I/O aggregator modules (FTOS version 9.11.0.P8) are by default configured with DCB enabled and all ports belonging to all VLANs. The external uplink ports (33 to 56) are auto configured to be in a single LACP LAG (LAG 128). They are also auto configured to be in the DCB auto-upstream mode. The auto-upstream ports are set with willing bit turned on. One of the auto-upstream uplink ports (configuration source) will accept the recommended DCB configuration from the uplink peer and internally propagate to remaining auto-upstream and all auto-downstream ports. The internal server facing ports (1 to 32) are auto configured to operate in auto-downstream mode. This means that they have the willing bit turned off and

accept the internally propagated DCB configuration from the auto- upstream port. These ports advertise the DCB configuration information to their server peers.

The second step is to setup the server initiators and targets with correct iSCSI VLAN ID. DCB is enabled in willing mode by default. The last step is the verification of DCB parameters on the initiator, target, and corresponding switch ports.

4 Host adapter configuration for DCB

4.1 QLogic 57810 (previously Broadcom) configuration

Configuring the QLogic 57810 adapter for iSCSI and DCB requires the use of the QLogic Control Suite (QCS) utility. For this paper, the following software versions were used: QCS version 17.0.14.0, QLogic driver version 7.12.32.0 and QLogic firmware version 7.12.19. QCS software and CNA drivers can be found at the Dell support site: <https://support.dell.com>.

4.2 QLogic 57810 DCBX willing mode verification

Figure 18 below shows a screen shot from the QLogic Control Suite (QCS) software. This is applicable to the QLogic 57810 Dual-Port 10 GbE SFP+ adapter and the QLogic 57810 Dual-Port 10 GbE KR Blade Converged Mezzanine Card on Windows Server 2008 R2 SP1. The parameter *local machine willing* indicates the willing mode state and is enabled by default.

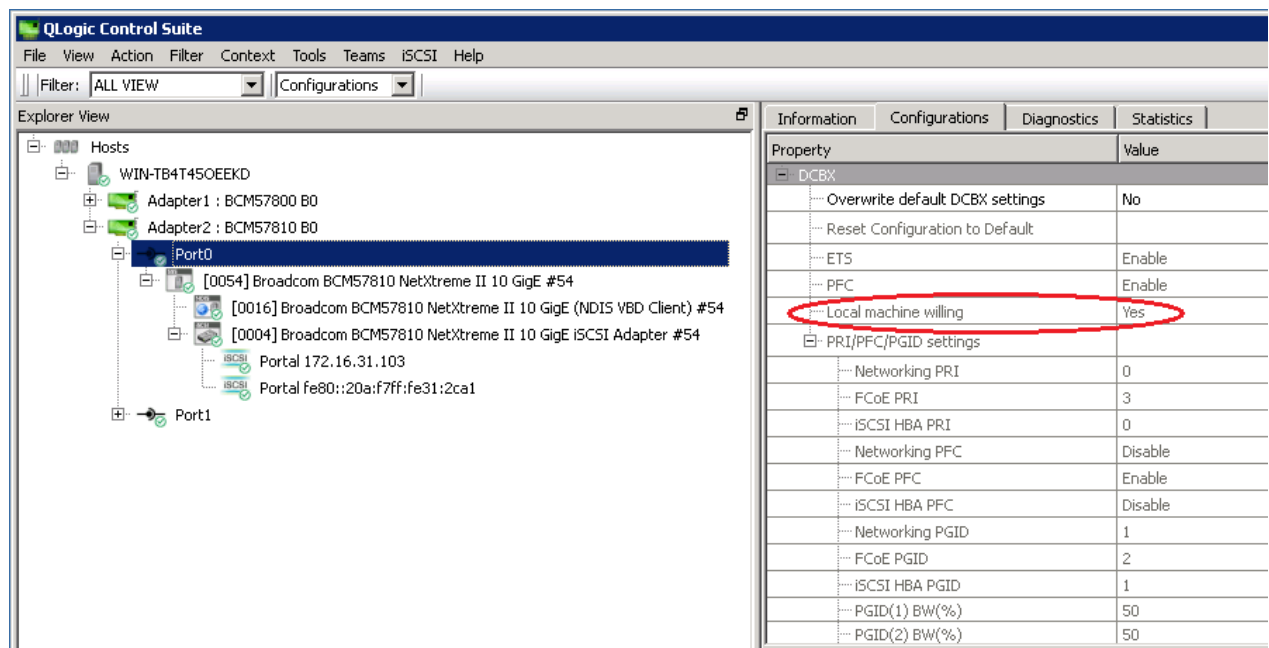


Figure 18 Initiator willing mode configuration

In this figure, the adapter is configured with willing mode enabled by default. Also the local DCBX settings are shown in the default configuration. These settings include the priority and PFC values for FCoE and iSCSI. This also includes the priority groups with their bandwidth allocation and priority mapping. Once the adapter port establishes an operational link with a peer switch port and receives the switch DCBX settings, it will use those settings for DCB operations. The local settings will not be operational at that point. The default local DCBX setting can be changed by overriding them.

4.3 QLogic 57810 iSCSI VLAN configuration

Figure 19 below illustrates the VLAN configuration for iSCSI traffic on the adapter port.

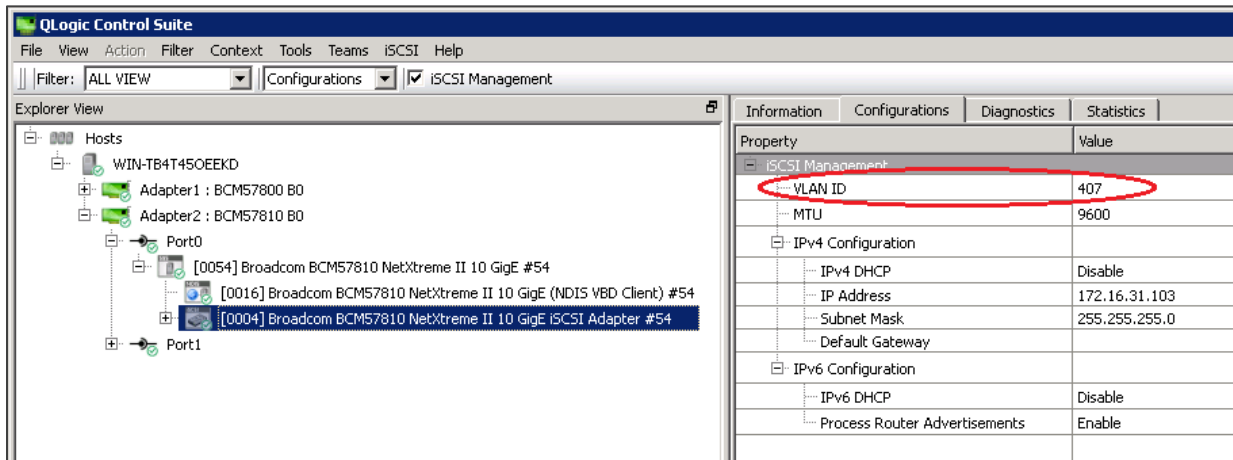


Figure 19 Initiator VLAN ID configuration

4.4 QLogic 57810 DCB configuration verification

Figure 20 below is a screen shot from the QCS software. This is applicable to the QLogic 57810 Dual-Port 10 GbE SFP+ adapter and the QLogic 57810 Dual-Port 10 GbE KR Blade Converged Mezzanine Card on Windows Server 2008 R2 SP1. The DCB parameters shown are the values advertised by the switch peer and accepted by the NIC port. It also indicates the DCB operational state for each feature.

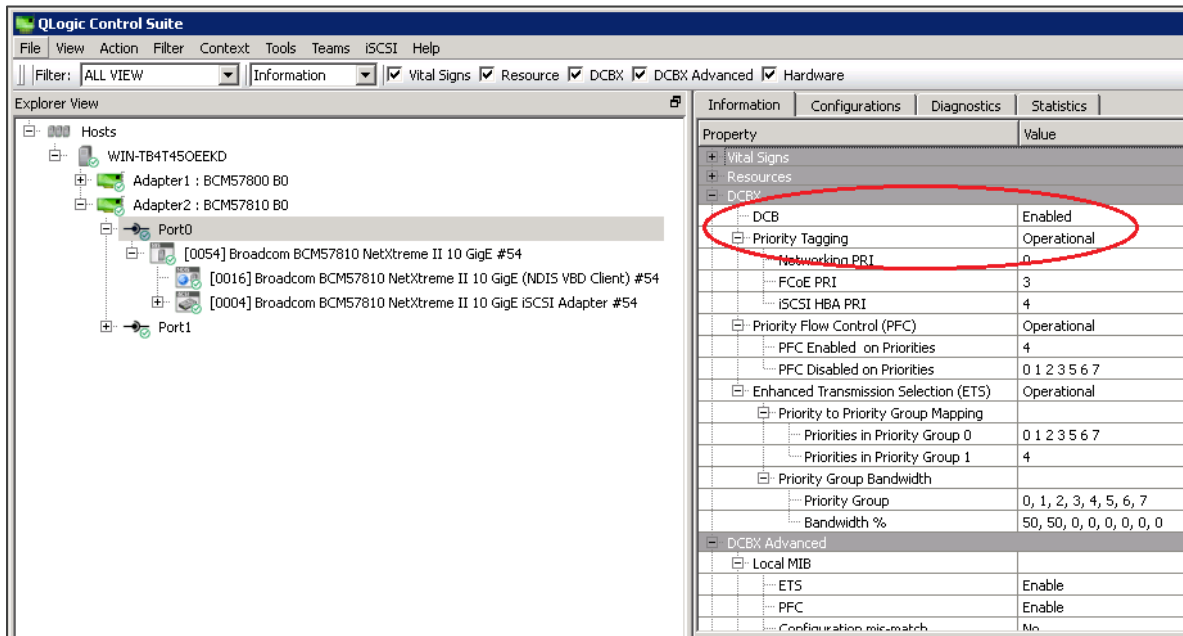


Figure 20 DCBX operation state and configuration verification

The local and remote DCBX settings are displayed in the **DCBX Advanced** section of the QCS, see Figure 21.

The screenshot shows the QLogic Control Suite (QCS) interface. The Explorer View on the left displays the hardware configuration for a host named WIN-TB4T450EEKD, including two BCM57800 B0 adapters and their respective ports. The Properties window on the right is set to the 'DCBX Advanced' section, showing a table of Local and Remote MIB settings.

Property	Value
DCBX Advanced	
Local MIB	
ETS	Enable
PFC	Enable
Configuration mis-match	No
Networking PRI	0
FCoE PRI	3
iSCSI HBA PRI	4
PFC Enabled on Priorities	4
PFC Disabled on Priorities	0 1 2 3 5 6 7
Networking PGID	0
FCoE PGID	0
iSCSI HBA PGID	1
PGID(0) BW(%)	50
PGID(1) BW(%)	50
PGID(2) BW(%)	0
PGID(3) BW(%)	0
PGID(4) BW(%)	0
PGID(5) BW(%)	0
PGID(6) BW(%)	0
PGID(7) BW(%)	0
Remote MIB	
Remote application priority willing	No
Remote PFC willing	No
Remote ETS willing	No
Remote ETS recommendation valid	No
Remote FCoE PRI	
Remote iSCSI PRI	4
Remote PFC Enabled on Priorities	4
Remote PFC Disabled on Priorities	0 1 2 3 5 6 7
Remote Networking PGID	0
Remote FCoE PGID	
Remote iSCSI PGID	1
Remote PGID(0) BW(%)	50
Remote PGID(1) BW(%)	50
Remote PGID(2) BW(%)	0
Remote PGID(3) BW(%)	0
Remote PGID(4) BW(%)	0
Remote PGID(5) BW(%)	0
Remote PGID(6) BW(%)	0
Remote PGID(7) BW(%)	0

Figure 21 Advanced DCB operation state and configuration verification

5 DCB testing

The following workloads and tests were developed to provide empirical evidence of the benefits of DCB in two common scenarios: Dedicated iSCSI networks and converged traffic networks. In the dedicated iSCSI tests, traffic was only running from the iSCSI host to the target. No other traffic was being injected. This simulates environments where the customer is implementing DCB, but continuing to separate storage and LAN traffic through the use of dedicated, separate physical networks. For the converged traffic tests, iSCSI traffic was still run, and non-iSCSI TCP traffic was also streamed from a TCP source host to the same ports on the iSCSI host. This allowed the impacts of DCB to be directly measured on the initiator links. This scenario provides a high throughput test of both storage and LAN traffic on the same network segment.

5.1 Workload definitions

During each test throughput (MB/s) and I/Os per second (IOPS) data at the host initiator were gathered. The TCP retransmission rate was also monitored to ensure that it never went above 0.5% during the test runs. The Tool used to generate iSCSI traffic during testing was VDBench. The tool used to generate raw IP traffic to simulate other LAN traffic was iPerf.

The three I/O workload types tested were: sequential read, sequential write, and a random read/write mix. These workloads were selected to imitate three common real-world scenarios.

Sequential read: Designed to imitate large-block, high throughput scenarios such as video streaming or backups.

Sequential write: Designed to imitate medium block sequential write scenarios such as file transfers or day-to-day OS processes.

Random read/write: Designed to imitate small block, high IOPS scenarios such as databases and OLTP transactions

Each test consisted of a single workload type running for 15 minutes. We completed three separate test runs for each workload type, and then computed an average of the results of all three runs. Details on the workload types are shown in Table 1.

Table 1 Workloads

Workload type	Block size	Read/Write ratio
Sequential read	256 K	100%/0%
Sequential write	64 K	0%/100%
Random read/write	8 K	67%/33%

After comparison runs early in the testing process, it was determined that the sequential write and random read/write workloads did not provide sufficient throughput to cause a noticeable effect. As seen from the charts below, virtually no difference was seen from a DCB versus non-DCB standpoint.

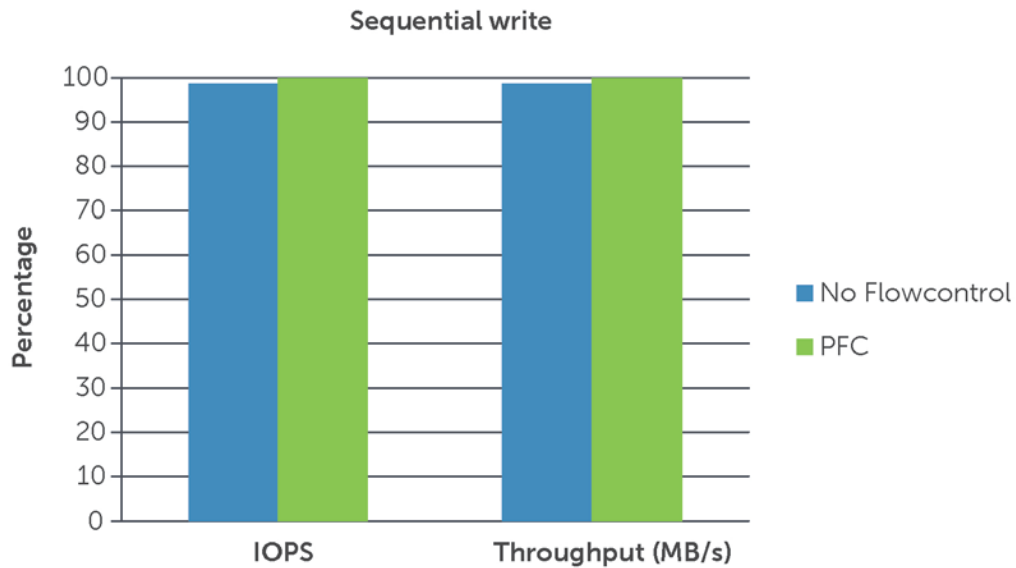


Figure 22 Sequential write I/O and throughput comparisons



Figure 23 Random read/write I/O and throughput comparisons

Note: For the remainder of this paper, only the results derived from the sequential read workload will be discussed.

5.2 Topologies

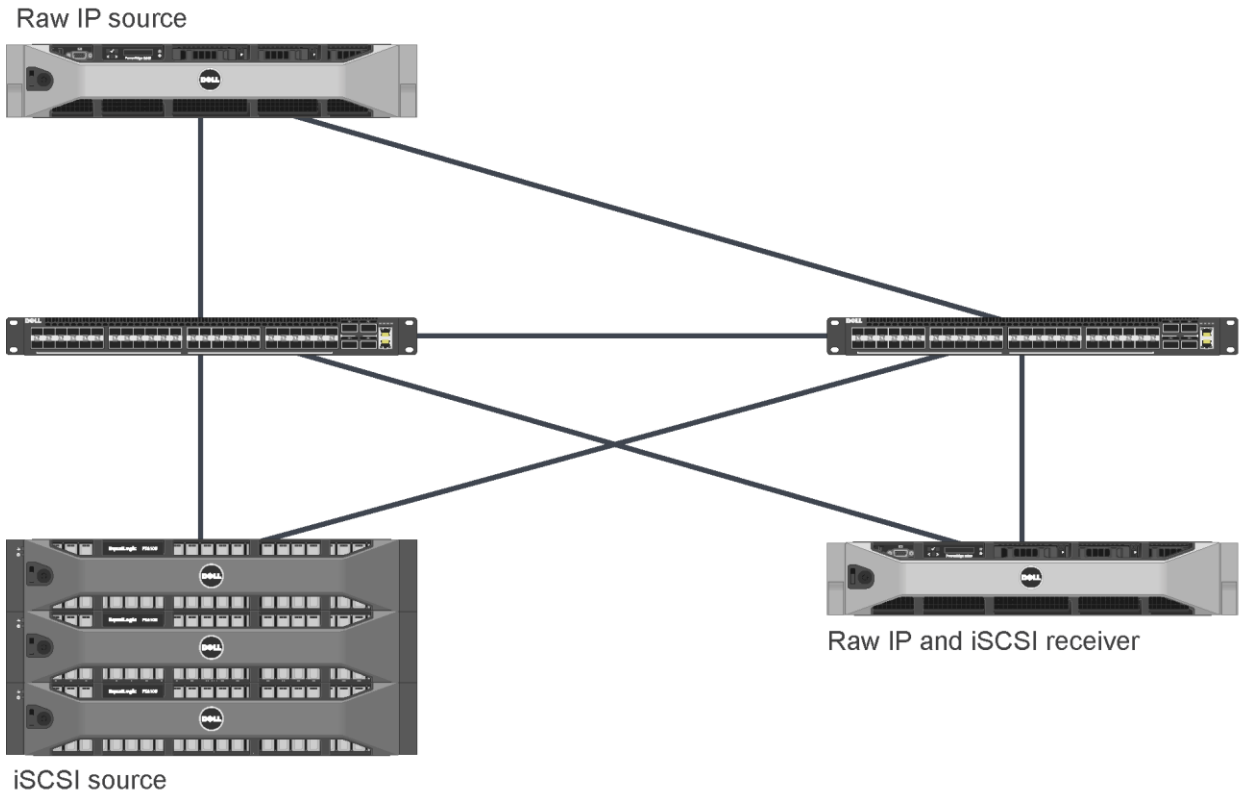


Figure 24 Test topology

In this topology, three PS6110 Series arrays were used to ensure that the throughput for the host ports could be maximized. All physical connections were made in a redundant manner, with ports connected across separate switches. These switches were then connected through a high speed link aggregation group (LAG) of 2 x 40 Gb ports, providing sufficient bandwidth for any traffic traversing the inter-switch link. A separate server was used to provide TCP traffic to the test host, ensuring that the converged traffic would occur on the proper links being measured for the test.

6 Results and analysis

6.1 DCB versus non-DCB in a dedicated SAN network

The following charts show the difference in retransmitted iSCSI frames on the network given each flowcontrol method. Retransmitted network frames waste available network resources by transmitting data a second time resulting in lower throughput and I/O performance on the storage array. Notice how in a pure iSCSI environment MAC PAUSE and PFC both are able to reduce the retransmitted traffic to basically zero percent, the same does not hold true for mixed traffic scenario.

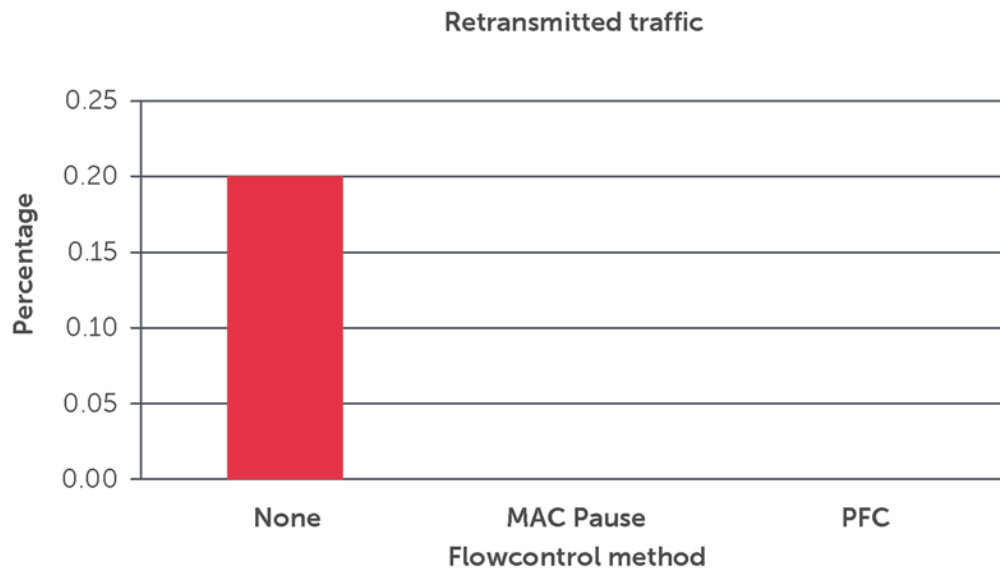


Figure 25 TCP retransmission comparison for iSCSI only traffic

For pure iSCSI environments, tests were run to determine if there was any benefit to running DCB protocols. It represents an opportunity to upgrade existing infrastructure without changing the logical layout of the network. For throughput there is an approximate 15% improvement with PFC over no flowcontrol and a 5% improvement over MAC PAUSE.

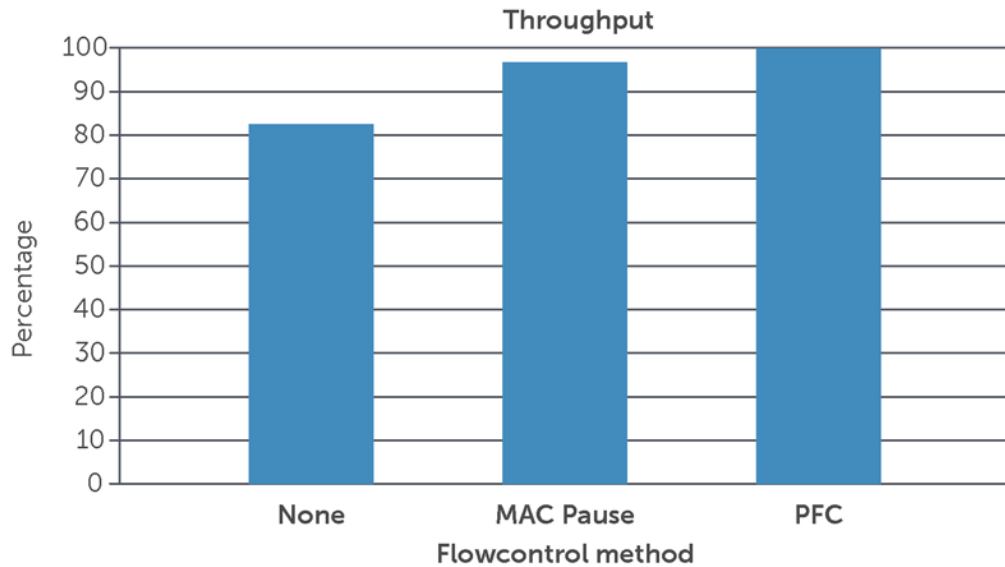


Figure 26 Throughput comparison for iSCSI only traffic

For I/Os per second, the result was consistent with a difference of approximately 15% improvement with PFC over no flowcontrol and a 5% improvement over MAC PAUSE.

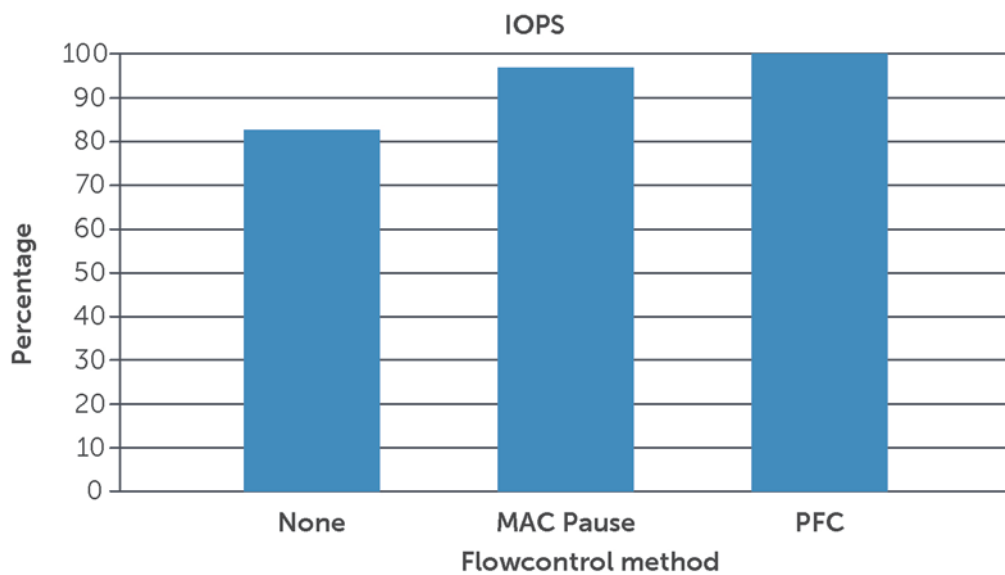


Figure 27 I/Os per second comparison for iSCSI only traffic

6.2 DCB versus non-DCB in a shared network

Finally, the throughput for both DCB and non-DCB environments was compared for fully converged traffic. In this series of tests, the same iSCSI load was placed on the network while simultaneously injecting TCP traffic to another server. The ETS settings on the switch were configured so that iSCSI should receive a minimum of 50% of the available bandwidth at any time.

In a shared network, controlling the amount of bandwidth given to different resources has been tough even in the best scenario. Traditional flowcontrol as implemented with MAC PAUSE provided a minimal method to stop some bursting traffic by pausing all traffic on a given link. With PFC it is now possible to stop specific traffic classes over a given link allowing a much more controlled environment for shared network traffic, in this case iSCSI traffic and other LAN data. The chart below shows how MAC PAUSE greatly reduces retransmissions over no form of flow control but how PFC is able to eliminate all retransmission, even in a shared network environment.

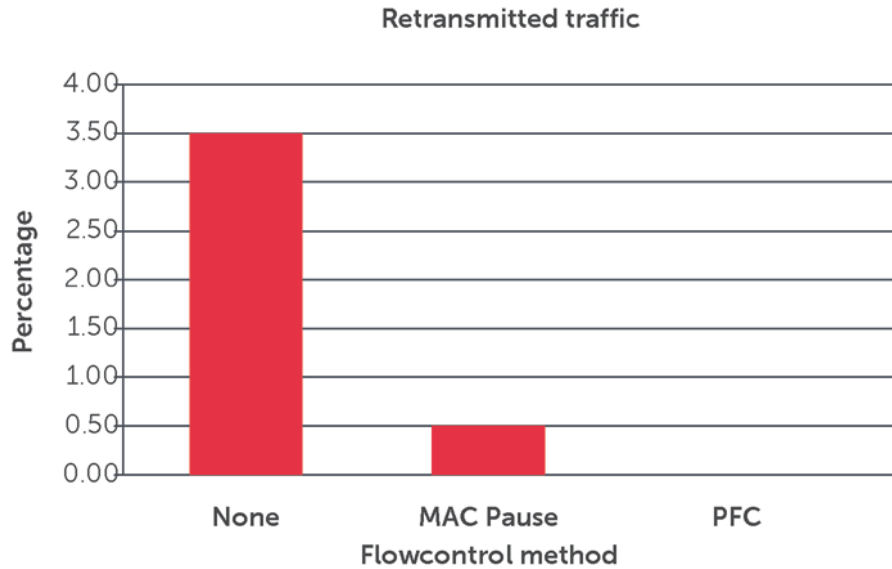


Figure 28 TCP (iSCSI) retransmission comparison mixed traffic environment

Prior to the introduction of PFC, running multiple traffic types in a network would yield unpredictable results. Figure 29 shows that without flowcontrol, iSCSI traffic performance was degraded and overrun by all the other LAN traffic on the network. The basic flow control ability of MAC PAUSE was able to improve on the control of network bandwidth allocation. However, PFC did a great job of controlling the network and giving iSCSI traffic its expected bandwidth.

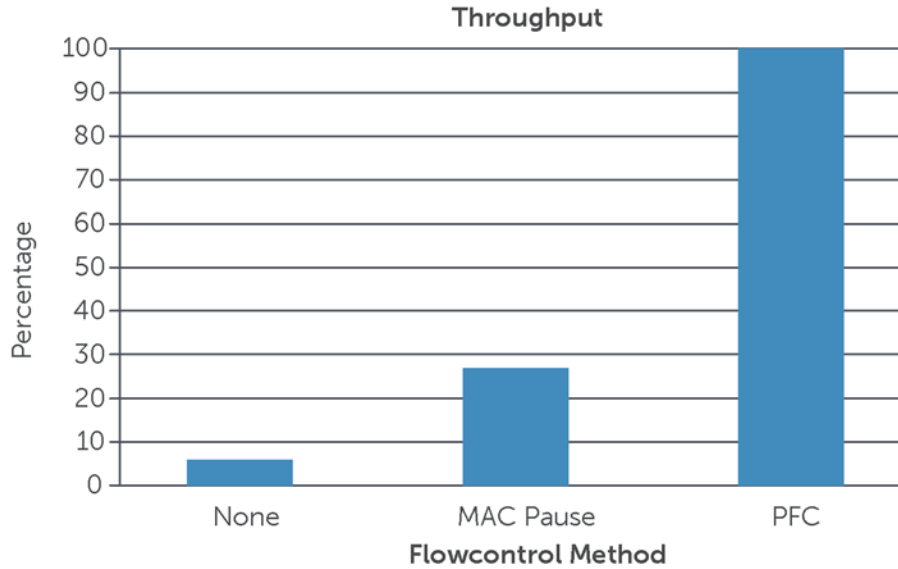


Figure 29 iSCSI throughput comparison for mixed traffic environment

In Figure 30, we see the same benefit in IOPS performance that was noted in the previous throughput section. PFC is able to create a controlled environment providing consistent results.

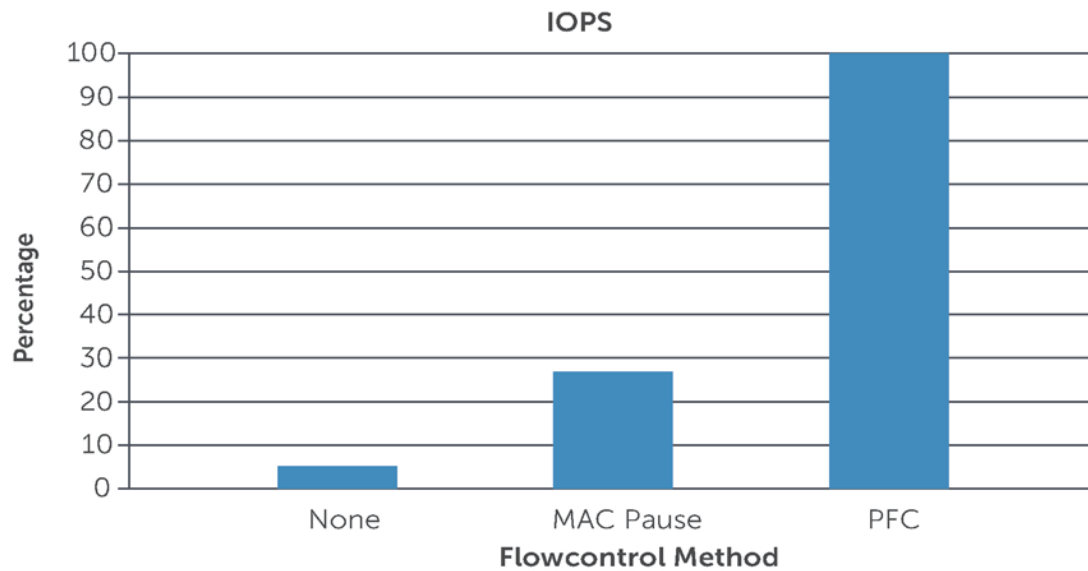


Figure 30 iSCSI IOPS comparison for mixed traffic environment

A PS Series command line reference for DCB configuration verification

From the CLI (GroupAdmin login), the following commands can be executed to obtain the same information that is presented in the Group Manager GUI:

- `show grpparams` (Display Global DCB and VLAN status):

```
PG1> show grpparams
```

```
_____ Group Information _____
Name: PG1
Group-Mgmt-Gateway: 192.168.140.1
Def-Snap-Warn: 10%
Def-Thin-Growth-Warn: 60%
DateAndTime: Fri Nov 9 13:32:41 2012
Description:
Def-Iscsi-Alias: yes
Info-Messages: enabled
Webaccess-noencrypt: enabled
Cliaccess-Telnet: enabled
Syslog-Notify: disabled
Email-List:
Sntp-Server-List:
DNS-Suffix-List:
Target-Auth-UserName:
ZpHr8HHZj2PJX22H
Email-From:
Conn-Balancing: enabled
Email-Contact:
Disallow-Downgrades: yes
SSH-V1-Protocol: enabled
Standby-Button: disabled
Def-DCB-VLAN-Id: 100
Crypto-Legacy-Protocols: enabled
Session-Idle-Timeout: 0
Session-Banner-Enable: disabled

Group-IpAddress: 10.10.10.150
Def-Snap-Reserve: 100%
Def-Snap-Depletion: delete-oldest
Def-Thin-Growth-Max: 100%
TimeZone: America/New_York
Def-Iscsi-Prefix:
iqn.2001-05.com.equallogic
Webaccess: enabled
Cliaccess-SSH: enabled
Email-Notify: disabled
iSNS-Server-List:
NtpServers:
DNS-Server-List:
Syslog-Server-List:
Target-Auth-Password:
BbH822Bx2xXpBRJZ
Location: default
Discovery-Use-Chap: disabled
Perf-balancing: enabled
Management-IpAddress:
192.168.140.170
FTP-service: enabled
DCB: enabled
Thermal-Shutdown: enabled
Volume-Recovery: enabled
Session-Idle-Timeout-Enable:
disabled
Def-Snapshot-Borrow: disabled
```

- member select <array-name> eth show (Display DCB ON/OFF):

The example below is from a PS6110 array, so eth0 is the iSCSI port and eth1 is the management port which shows the DCB as off for the management port.

```
PG1> member select P1 eth show
Name      ifType          ifSpeed  Mtu    Ipaddress      Status  Errors  DCB
eth0      ethernet-      10 Gbps  9000   10.10.10.141   up      0       on
eth1      ethernet-      100 Mbps 1500   192.168.140.171 up      0       off
```

- member select <array-name> eth select 0 dcb show (Display DCB details):

```
PG1> member select P1 eth select 0 dcb show
_____ Eth DCB Information _____
Name: eth0                      DCB: on
Lossless-Priorities: on         Traffic-Classes: on
Congestion Notification: off    iSCSI-Priority: 4
```

- member select <array-name> eth select 0 dcb lossless-priority show (Display Lossless Priority configuration):

```
PG1> member select P1 eth select 0 dcb lossless-priority show
Priority  Status
0         off
1         off
2         off
3         off
4         on
5         off
6         off
7         off
```

- member select <array-name> eth select 0 dcb traffic-class show (Display Traffic Classes):

```
PG1> member select P1 eth select 0 dcb traffic-class show
Traffic      Bandwidth  Priorities
```

Class		
1	50%	4
2	50%	0, 1, 2, 3, 5, 6, 7

B Technical support and resources

Dell.com/support is focused on meeting customer needs with proven services and support.

Dell TechCenter is an online technical community where IT professionals have access to numerous resources for Dell EMC software, hardware and services.

Dell.com/StorageResources on Dell TechCenter provide expertise that helps to ensure customer success on Dell EMC Storage platforms.

B.1 Additional resources

See the following referenced or recommended Dell publications:

- Dell Storage Compatibility Matrix:

<http://en.community.dell.com/dell-groups/dtcmmedia/m/mediagallery/19856862.aspx>

- Data Center Bridging: Standards, Behavioral Requirements, and Configuration Guidelines with Dell EqualLogic iSCSI SANs:

<http://en.community.dell.com/dell-groups/dtcmmedia/m/mediagallery/20283700/download.aspx>

- Switch Configuration Guides for EqualLogic or Compellent:

<http://en.community.dell.com/techcenter/storage/w/wiki/4250.switch-configuration-guides-for-EqualLogic-or-compellent-sans.aspx>