Best Practices for Dell Networking N4000 Series Switch with DCB Configured for iSCSI

A Dell EMC Best Practices
DCB features

Dell Engineering
April 2014

A Dell EMC Deployment and Configuration Guide

Revisions

Date	Description	Author/Editor
April 2014	Rev 2.0, updates to branding for Dell Networking N Series.	Kevin Locklear, Mike Matthews
August 2012	Initial release of PC8100 DCB deployment guide	Kili Land, Kevin Locklear

Table of contents

Executive summary

The Dell Networking N4000 series switches (Figure 1) are the newest 10 and 40 GB Ethernet multilayer switches with a feature set targeted for deployment in campus networks.

One of these features is Data Center Bridging (DCB). DCB is an overarching architecture that includes several protocols, including Priority Flow Control (PFC), Enhanced Transmission Selection (ETS), and Data Center Bridging Exchange (DCBx). These protocols help to improve the reliability of congested networks by allowing Class of Service (CoS) queues to be paused instead of dropped and bandwidth to be allocated for specific types of traffic.

The goal of this document is to explain and provide best practices for the various settings needed to enable DCB and DCB iSCSI configuration.

For specific steps to configure these switches for DCB/Non-DCB ISCSI with Equallogic Arrays please refer to this document:

- Dell Networking N4000 and Dell PowerConnect 8100 Series configuration guide for Equalogic SANs
- Best Practices for Configuring DCB with Windows Server and EqualLogic Arrays
- Best Practices for Configuring DCB with VMware ESXi 5.1 and Dell EqualLogic Storage
- Best Practices for a DCB-enabled Dell M-Series Blade Solution with EqualLogic PS-M4110
- Creating a DCB Compliant EqualLogic iSCSI SAN with Mixed Traffic (revised)
- EqualLogic DCB Configuration Best Practices
- Data Center Bridging: Standards, Behavioral Requirements, and Configuration Guidelines

Figure 1 Dell Networking N4000 Series

1 Introduction

Data Center Bridging (DCB) is a collection of enhancements that combines several different features from the IEEE802.1 standards that help to improve the reliability of Ethernet-based networks in the Data Center environment.

DCB contains several protocols, which include:

- **Priority Flow Control (PFC):** PFC allows you to specify the CoS queues that should be paused (due to greater loss sensitivity) instead of dropped when congestion occurs on a link.
- **Enhanced Transmission Selection (ETS):** ETS allows bandwidth allocation to be applied on groups of Class of Service (CoS) queues.
- **Data Center Bridging Exchange (DCBx):** DCBx allows DCB devices to exchange configuration information.

Overview of Class of Service and Data Center Bridging Operation

The default operation for an uncongested Dell Networking switch is for packets to be scheduled for output in the order in which they are received.

When the switch is congested, CoS allows the administrator to prioritize high priority packets ahead of other packets, choose which packets to drop, and assign a minimum bandwidth guarantee to ensure scheduling fairness.

DCB includes the PFC and ETS protocols, which provide additional options for controlling the packets while the switch is congested.

Using PFC, CoS queues can be configured to be no-drop queues. These queues may be paused when congestion occurs but will not be dropped.

ETS provides a second level of scheduling the packets that are selected for transmission by the CoS scheduler. Each CoS queue is mapped to one of the three Traffic Class Groups (TCGs). These TCGs can be configured with weights, scheduling and both minimum and maximum bandwidths. DCB also provides DCBx that allows DCB configurations to be exchanged between DCB peers.

For additional details, refer to the ETS Operation section of the Dell Networking N2000, N3000, and N4000 Series Switches User's Configuration Guide.

The following sections provide an overview of CoS and DCB, followed by configuration and troubleshooting information.

2 Class of Service

The Class of Service (CoS) helps to streamline traffic in a congested network by allowing priority traffic to receive preferential treatment. This includes mapping traffic to specific CoS queues (Queue Mapping). Determining if packets are queued or dropped (Queue Management), and deciding the order the queued packets are serviced (Scheduling). By

using queue mapping, queue management and scheduling, traffic can be better managed to help avoid crippling congestion on oversubscribed network connections.

2.1 Queue Mapping

Queue mapping (Figure 2 and Table 1) takes the priority assigned to the packet through either VLAN Priority Tagging (VPT) or Internet Protocol - Differentiated Services Code Point (IP-DSCP) and maps it to a specific Traffic Class or CoS Queue.

There are seven queues available in the Dell Networking N series. These queues are identified as Traffic Class by the switches and are numbered 0-6. To help facilitate congestion control, the network administrator can assign the user priorities (VPT) to different traffic classes.

User Priority	Traffic Class
0	1
1	0
2	0
3	1
4	2
5	2
6	3
7	3

Figure 2 Default Mapping for the VPT (Dot1p) in the Dell Networking N series

Table 1 Recommended User Priority to Traffic Mapping

		Number of Available Traffic Classes							
		1	2	3	4	5	6	7	8
Priority	0 (Default)	0	0	0	0	0	1	1	1
	1	0	0	0	0	0	0	0	0
	2	0	0	0	1	1	2	2	2
	3	0	0	0	1	1	2	3	3
	4	0	1	1	2	2	3	4	4
	5	0	1	1	2	2	3	4	5
	6	0	1	2	3	3	4	5	6
	7	0	1	2	3	4	5	6	7

2.2 Queue Management

Queue management helps control traffic flow within a queue. It is used to determine if packets are queued or dropped when the queue is full. The queue management for CoS in the Dell Networking N series of switches is tail drop and weighted random early detection.

Tail drop

Tail drop is a very basic congestion mechanism that drops packets when the queue is full. No priority is given to any one packet. No matter what packets are incoming, if the queue is full, the packets are dropped. This is the default congestion queue management implemented by Dell Networking switches and is the required queue management for no-drop/lossless Ethernet queues, such as PFC.

Weighted Random Early Detection

Weighted Random Early Detection (WRED) functions by assigning various thresholds within a particular queue that trigger certain traffic classes to be dropped or queued. As congestion increases, the coloring assignment determines how priorities will be serviced and which packets will be dropped.

At levels of extreme congestion, WRED is no longer effective and traffic is handled in a tail drop manner until congestion is reduced to manageable levels. Because of the various thresholds of congestion handling within a WRED queue management mechanism, no-drop queues cannot be assigned to this type of queue management.

In the Dell Networking N series of switches, the congestion thresholds are all set to equitable distribution unless coloring is assigned to a particular user priority.

Packet Coloring

The Dell Networking N series switches use Single Rate mode of traffic policing. By default, all packets are colored with the same conform-color and are handled equitably.

Note: This document does not cover Packet Coloring information in any detail.

2.3 Scheduling

Scheduling is used to determine how packets are handled once they hit the queue. CoS uses strict or weighted scheduling.

Strict priority

Strict priority scheduling gives an absolute priority based on CoS queue number. The traffic in the highest numbered queue is sent first, then the next lowest numbered queue, and so on. Weighted queues are serviced after all strict priority queues have been serviced.

Weighted

Weighted scheduling selects packets for transmission based on a fixed weighting equal to the CoS queue number plus one.

Note: Round Robin is the default algorithm used by Dell Networking N series of switches when there is no congestion. The Round Robin algorithm sends packets across each queue in a sequential order that allocates equitable distribution for all traffic.

2.3.1 Minimum Bandwidth Guarantee

Along with the scheduler type, the minimum bandwidth can be specified. This is a percentage of the port's maximum negotiated bandwidth reserved for the queue. This allows unreserved bandwidth to be utilized by lower-priority queues. If the sum of the minimum bandwidth is 100%, there is no unreserved bandwidth and no sharing of bandwidth is possible.

3 Data Center Bridging

DCB increases the reliability of Ethernet-based networks in the data center by allowing CoS queues to be configured as no-drop queues (PFC), as well as having bandwidth, weights and scheduling allocated (ETS). DCB also provides the ability to exchange configuration information between connected peers (DCBx).

3.1 Data Center Bridging Exchange

Data Center Bridging Exchange Protocol (DCBx) allows DCB configuration information to be automatically exchanged between DCB compliant peers. This configuration information includes peer discovery, configuration mismatch detection, and automatic configuration of willing peers. The automatic configuration includes ETS and PFC settings.

In DCBX, a willing peer is a device that actively takes the DCBX configuration from an upstream peer. The configuration source is the upstream peer that sets the DCBX for the connected willing peers.

DCBx uses type-length-value (TLV) information elements over Link Layer Discovery Protocol (LLDP) to exchange information, so LLDP must be enabled on the port to enable the information exchange. By default, LLDP is enabled on all of the ports on the Dell Networking N4000 switches.

3.2 Priority Flow Control

Priority Flow Control (PFC) is used to pause traffic on a particular queue without having to pause traffic on all queues (as in classic Flow Control). By specifying a particular CoS queue to be a PFC queue, it is ensured that the traffic is not dropped when congestion is present on the line. Instead, a pause frame is sent to the queue to halt the traffic until the traffic can be reliably passed. While this causes a decrease in throughput rates, it increases data integrity and ensures there are no dropped packets in the PFC queue.

3.3 Enhanced Transmission Selection

ETS allows bandwidth allocation to be applied on groups of queues. Each of the CoS queues can be assigned to a Traffic Class Group (TCG). ETS allows weights, scheduling and both minimum and maximum bandwidths.

Note: N4000 switches support three TCGs internally, up to two of which may be configured as lossless.

3.3.1 Assigning Weights to TCGs

Weighting is used to assign the basic amount of bandwidth that is allocated to the assigned TCGs.

The weights are only strictly applied when the TCGs require their fully allocated bandwidth. As an example, consider a scenario where there are three Traffic Class Groups; TCG0 is assigned a weight of 20, TCG1 a weight of 70, and TCG2 a weight of 10. Essentially the bandwidth has been allocated as 20% to TCG0, 70% to TCG1 and 10% to TCG2.

The bandwidth allocation fluctuates as demand for the bandwidth shifts. Therefore, if TCG1 traffic is only using 50% of its allocated 70% of bandwidth and TCG0 traffic needs 40%, ETS allows the TCG0 traffic to use 40%. Then when the TCG1 traffic increases and needs its full 70% of bandwidth allocation, ETS will minimize the TCG0 traffic back to 20%.

3.3.2 Scheduling

Scheduling is used to determine how the packets are handled once they hit the queue. ETS uses strict or Weighted Deficit Round Robin (WDRR) scheduling.

Strict priority

Strict priority TCGs are scheduled first, but have their bandwidth reduced by the bandwidth guarantees configured on other TCGs. The strict priority TCGs are scheduled from the highest numbered TCG to the lowest.

When all TCGs have met their bandwidth limits (or the queues are empty), TCGs that have not met their bandwidth limit are scheduled. Once the limits for a TCG are satisfied (maximum bandwidth, no frames available for transmission, etc.), the scheduler moves to the next TCG.

Weighted Deficit Round Robin

WDRR is an evolution of the basic Round Robin algorithm. It is based on the weights assigned to the TCG. When WDRR is used, the packets at the head of every non-empty queue are checked against a deficit counter. If the deficit counter is greater than the packet size at the head of the queue, the queue is served. If the deficit counter is not greater, then the queue is skipped and the queue is given a credit. This credit, which

increases the deficit counter, is added each time the queue is skipped. Thus, eventually, the deficit counter increases to the point that allows the queue to be served.

As mentioned above, each TCG is handled according to its weight. If all the TCGs are weighted and scheduled the same, the highest numbered TCG is served first, followed by the next highest, and so on until the lowest numbered TCG is served.

3.3.3 Minimum and Maximum Bandwidth

Minimum Bandwidth

By setting minimum bandwidth allocations, it can be ensured that the TCGs will always be guaranteed a certain amount of bandwidth, regardless of how congested the line becomes. If the strict queue wants to take the bandwidth, it will not be allowed.

Maximum Bandwidth

By setting maximum bandwidth allocations, the maximum bandwidth allowed for a TCG can be capped. The TCG will never exceed the allocated maximum bandwidth, even if configured for strict priority.

4 Configuration

The following sections include an overview of configuring CoS and DCB.

4.1 Class of Service Configuration

Queue Mapping

The first step in deploying CoS is determining which traffic gets priority. For the purpose of this paper, the iSCSI traffic is given priority.

Next, the priorities must be mapped to the appropriate CoS queues. The default mapping in the N4000 series of switches is shown in Figure 3. iSCSI is generally assigned to VPT 4 and VPT 4 is assigned to traffic class 2. For this scenario, the VPT 4 traffic is moved to traffic class 2.

User Priority	Traffic Class
0	1
1	0
2	0
3	1
4	2
5	2
6	3
7	3

Figure 3 Default VPT and Traffic Class Mapping

Move the VPT 4 traffic from traffic class (CoS queue) 2 to traffic class 4 by entering the following command at the config prompt:

```
classofservice dot1p-mapping 4 4
```

Figure 4, shows the dot1p mapping table, which is displayed using the **`show classofservice dot1p-mapping`** command. Notice that for VPT (user priority) 4, the traffic class has moved from 2 to 4.

```
console#show classofservice dot1p-mapping

 User Priority    Traffic Class
 -------------    -------------
       0                1
       1                0
       2                0
       3                1
       4                4
       5                2
       6                3
       7                3
```

Figure 4 Traffic Class for VPT 4 Moved from 2 to 4

Note: These settings can be done at both the global and interface level. When done at the interface level, the settings will preempt the global settings. Commands in this guide are performed at the global level.

4.2 iSCSI Configuration

iSCSI Optimization

By default iSCSI optimization is enabled. For the purpose of this guide, the importance of iSCSI optimization is that once all the settings are configured, it will force the DCBx Application Priority TLV to be sent. These TLVs provide attached devices with information that includes (but is not limited to) the iSCSI protocol (TCP port 3260) and PFC settings. The `iSCSI Enable` command enables iSCSI optimization.

Note: This document does not cover iSCSI optimization in any detail.

iSCSI CoS

By default iSCSI CoS is not enabled. Enabling iSCSI CoS is required if there is not a configuration source (such as a Top of Rack (ToR) switch) that is sending the iSCSi application priority TLVs.

For this scenario, there is no configuration source, so iSCSI CoS must be enabled in order for the DCBx Application Priority TLV to be sent. The default iSCSI VPT assignment of 4 will be kept.

To enable iSCSI CoS enter the following command from the config prompt:

```
iscsi cos enable
```

Since queue mapping has been configured with the assumption that iSCSI CoS will be set to the default of 4, no changes to the VPT are necessary.

In the event the iSCSI VPT needed to be changed from the default VPT of 4, the following command would be used.

```
iscsi cos vpt <#>
```

Where **<#>** is the VPT assignment. This number can range from 0 to 7.

Note: At this point, a strict priority CoS queue has not been configured. Since the iSCSI traffic will belong to an allocated Traffic Class Group (TCG), a strict priority CoS queue is not needed.

Checking the iSCSI Settings

The iSCSI settings can be viewed with the following command:

```
show iscsi
```

Figure 5, shows the output of the `show iSCSI` command.

```
console#show iscsi

iSCSI enabled
iSCSI CoS disabled
iSCSI vpt is 4
Session aging time: 10 min
Maximum number of sessions is 192
-------------------------------------------------
iSCSI Targets and TCP Ports:
-------------------------------------------------
TCP Port    Target IP Address    Name
860               -                    -
3260              -                    -
-----------------------------------------------------
iSCSI Static Rule Table
-----------------------------------------------------
Index TCP Port    IP Address       IP Address Mask
```

Figure 5 `show iSCSI` displaying the iSCSI settings

Note that in the `show iscsi` output (Figure 5) iSCSI target ports are listed. These ports can be changed, based on the network and which ports the iSCSI traffic is traversing.

The ability to change the iSCSI target ports is useful when connecting classic flow control iSCSI devices to the network. Since these devices do not have the ability to allocate PFCs, the ports that are sending and receiving iSCSI traffic can be added as target ports, which add the traffic on those ports to the iSCSI queue.

Ports can be added with the following command at the config prompt:

```
iscsi target port <TCP port>
```

Ports can be removed by adding **no** in front of the command. Figure 6 shows the iSCSI target port 908 being added, target port 860 being removed and the output of the `show iscsi` command.

```
console(config)#iscsi target port 908

console(config)#no iscsi target port 860

console(config)#show iscsi

iSCSI enabled
iSCSI CoS disabled
iSCSI vpt is 4
Session aging time: 10 min
Maximum number of sessions is 192
-------------------------------------------------
iSCSI Targets and TCP Ports:
-------------------------------------------------
TCP Port    Target IP Address    Name
908               -                    -
3260              -                    -
-----------------------------------------------------
iSCSI Static Rule Table
-----------------------------------------------------
Index TCP Port    IP Address       IP Address Mask
```

Figure 6 Adding and Removing iSCSI Target Ports

4.3 Configuring Data Center Bridging

So far, the majority of the configuration has involved CoS. These CoS settings provide the foundation for the DCB configuration.

The following sections provide instructions on manually configuring ETS and PFC, if the device being configured will be a willing device, skip the following section and go to the as a willing device section.

4.3.1 Configuring Traffic Class Groups and Priority Flow Control

Assigning iSCSI Traffic to a Traffic Class Group

Since all the CoS queues are assigned to TCG0 by default, the iSCSI traffic needs to be moved to its own TCG. In this example, the iSCSI traffic is assigned to TCG1. The remaining CoS queues are left in TGC0 (the default TCG).

To move the iSCSI traffic from TCG0 to TCG1, enter the following command at the config prompt:

```
classofservice traffic-class-group 4 1
```

Setting Priority Flow Control to No-Drop

The CoS queues have been assigned TCGs, but there is still the potential for dropped packets as soon as congestion occurs. This is where the PFC no-drop option of is used.

Priority Flow Control can only be set on the individual port level. This means that each port that needs PFC set to no drop needs to be manually configured. The `interface range` command can be used to configure PFC on multiple ports at once.

Figure 7 displays the commands to configure the iSCSI CoS queue (Traffic Class 4) as no-drop.

```
console(config)#interface range te1/0/2-9,te1/0/48

console(config-if)#datacenter-bridging

console(config-if-dcb)#priority-flow-control mode on

console(config-if-dcb)#priority-flow-control priority 4 no-drop
```

Figure 7 Configuring the PFC No-Drop Queue

Setting TCG as strict

Strict scheduling can be used to allow the iSCSI traffic to take all the available bandwidth, regardless of the impact on other network traffic.

In the scenario used in this guide, only the iSCSI TCG (TCG1) is configured as strict. To configure the iSCSI TCG as strict, enter the following command:

```
(config)#traffic-class-group strict 1
```

Assigning Weights to TCGs

There are a couple of ways to configure the TCGs, depending on the network requirements. The TCGs can be configured to allocate a certain percentage of bandwidth using weights, or TCG1 (iSCSI traffic) can be given the right to take all the bandwidth it needs.

This example uses the first option. With this option, the TCGs are assigned a set of weights and the traffic is allowed to fluctuate as necessary. This option ensures the iSCSI traffic will never be able to choke out the other network traffic.

In this example, the default weights of **100 0 0** are used. If the default weights have been changed, enter the following command to set the TCG weights:

```
traffic-class-group weight 100 0 0
```

Note: When configuring with strict scheduling, if the command **`traffic-class-group weight 100 0 0`** is used, it might look as if the weights have been configured so all the bandwidth is allocated to TCG0. However, since the strict TCG is serviced first, this is not the case. Consider an example where there are three TCGs being configured: one as strict (TCG2) and two as not strict (TCG0, TCG1). The following command is used to allocate bandwidth: **`traffic-class-group weight 50 50 0`** In this case, the strict TCG (TCG2) takes all the bandwidth and leaves TCG 0 and TCG 1 with no bandwidth at all. However, if TCG2 does not need all the bandwidth, then the bandwidth will be equally divided betweenTCG0 and TCG1.

Figure 8 shows the output of the show **interfaces traffic-class-group** command**.**

```
console#show interfaces traffic-class-group

Global Configuration

Traffic Class    Min.       Max.                    Scheduler Type
Group           Bandwidth   Bandwidth    Weight
-------------   ---------   ---------    -------  --------------
0               10          50           100      Weighted Round Robin
1               0           0            0        Strict
2               0           0            0        Weighted Round Robin
```

Figure 8 show interfaces traffic-class-group

Setting minimum bandwidth

The TCGs can always be guaranteed a certain amount of bandwidth by setting minimum bandwidth. The minimum bandwidth allocation can be set at the global or interface level, as mentioned before the interface level settings preempt the global settings.
To set minimum bandwidth allocation, enter the following command:

```
(config)#traffic-class-group min-bandwidth 10 20 0
```

In this example, the minimum bandwidth is allocated to two TCGs. The total value of the minimums set for both equals 30 percent. This leaves more than enough room for bursty traffic. Remember, this is the *minimum* bandwidth. This is not the weight, or the maximum allowed. This is simply the amount of bandwidth that is guaranteed to the TCGs.

Note: Do not set the minimum bandwidth where the amount of the bandwidth allocated equals 100 percent. This hard sets the bandwidth, and is counter- productive and detrimental to a production environment.
The preferred behavior is to configure the minimum bandwidths so they total 80 percent or less. This ensures that the system can respond to bursts in traffic.
Setting the minimum bandwidth totals to 100 percent effectively sets the scheduler so that no queue can exceed its minimum bandwidth percentage when congestion occurs.

Setting maximum bandwidth
The following command can be run at the config prompt to make sure that the TCG never exceeds *X* amount of bandwidth by using the maximum bandwidth settings:

```
traffic-class-group max-bandwidth 50 0 0
```

In this example, the configuration has been set so that TCG0 traffic can never exceed 50 percent of the bandwidth, thus allowing TCG 1 and TCG2 to have the remaining 50 percent at all times. This is not a bandwidth guarantee for TCG0; it is the maximum it is allowed to use. If the TCG0 wants 51 percent, it cannot have it, even if TCG1 and TCG2 are not transmitting traffic.

4.3.2 Configuring Data Center Bridging on a Willing Device

Manual configuration of ETS and PCF is not needed, if the switch is going to be set as a willing device and connected into a network where an upstream configuration source already exists.

Configuring the DCBx Port Role
To configure a device as willing, set the ports to connect to the upstream device as auto-up or configuration source. Then configure the downstream ports as auto-down.

The configuration source setting can only be applied to a single port. Therefore, if the configuration is a port-channel to the upstream device, all the ports must be configured to use auto-up.

To set a port as configuration source or auto-up, enter one of the following commands:

```
(config-if-Te1/0/23)#lldp dcbx port-role configuration-source
                     or
(config-if-Te1/0/23)#lldp dcbx port-role auto-up
```

Note: Port Modes can only be applied on the physical interface. When applying port modes in a port-channel, be absolutely certain that all ports are set to the same mode.

Auto-down is used on the ports connecting to the downstream willing devices. This mode takes the information obtained from the ports connected to the configuration source and pushes the configuration down to the willing devices. This command must be entered on all ports connected to the downstream, willing devices.

To set a port as auto-down, enter the following command:

```
(config-if-Te1/0/23)#lldp dcbx port-role auto-down
```

Setting the Default DCBx Version

Generally, the default setting of **Auto** is fine for DCBx version. However, if equipment is being used that requires a specific, hard coded version be set, this can be done at the global or interface level with the following command at the config prompt:

```
lldp dcbx version <option>
```

The options available are **Auto**, **CEE** (1.06), **CIN** (1.0) or **IEEE** (802.1Qaz).

Note: It is strongly recommended that the global level DCBx version setting remain set to **auto** and only the ports connected to the individual equipment that requires a hard-set DCBx version be modified.

The DCBx version can be checked with the show command:

```
show lldp dcbx interface te1/0/23 status
```

Figure 9 shows the output of the `show lldp dcbx interface te1/0/23 status` command.

```
console#show lldp dcbx interface te1/0/23 status

DCBX operational status:....................... Enabled
Configured DCBX version:....................... Auto
Peer DCBX version:............................. IEEE
Peer MAC:.....................................
Peer Description:.............................
Auto-configuration Port Role:.................. Manual
Peer Is configuration Source:.................. False

Error counters:
ETS incompatible configuration................. 0
PFC incompatible configuration................. 0
Disappearing neighbor.......................... 0
Multiple neighbors detected.................... 0
```

Figure 9 Show lldp dcbx interface te1/0/23 status

Note: By default, the DCBx port mode on the switch is **Manual**. This means that it will push this configuration to any willing device that is connected.

5 Troubleshooting

The following sections provide information to aid in troubleshooting and verifying the settings of a DCB and/or DCB iSCSI deployment.

When troubleshooting keep the following in mind:

- Congestion control is only used when there is actual congestion on the line. If there is no congestion, then ETS and PFC do not affect the network.
- Do not set minimum bandwidth allocations to a total of 100 percent as this prevents the ability to adapt to bursts of traffic and causes problems on the network.
- Do not apply WRED queue management to the PFC queue as this causes PFC to pause the queue before the queue is truly congested.
- DCBx uses type-length-value (TLV) information elements over Link Layer Discovery Protocol (LLDP) to exchange information, so LLDP must be enabled on the port.

- DCBx is set on the physical ports, not on the port channel. So, be sure to check the DCBx settings on each of the physical connections.

- Verify the connection. Make sure the connection between the two ports is up. Ensure spanning tree is not blocking these ports. Also, keep an eye on link speed.

- In order for things to work as expected, the ETS and PFC information should match between the devices. This does not apply for an asymmetrical network.

5.1 Class of Service Settings

For CoS settings it is necessary to check the user priority and traffic class mappings. CoS settings can be confirmed using the following commands.

To show the dot10 mappings of the CoS, enter the following command:

```
show classofservice dot1p-mapping
```

To show the queue management, scheduling and minimum bandwidth for each class of service queue, enter the following command:

```
show interfaces cos-queue
```

Figure 10 shows the output of the `show interfaces cos queue` command.

```
console#show interfaces cos-queue

Global Configuration
Interface Shaping Rate..........................  0 kbps
WRED Decay Exponent............................  9

Queue Id   Min. Bandwidth   Scheduler Type  Queue Management
Type
--------   --------------   --------------   -------------------
0          0                Weighted         Tail Drop
1          0                Weighted         Tail Drop
2          0                Weighted         Tail Drop
3          0                Weighted         Tail Drop
4          0                Weighted         Tail Drop
5          0                Weighted         Tail Drop
6          0                Weighted         Tail Drop
```

Figure 10 show interfaces cos-queue

If the individual ports have been set with CoS queue settings, the command can be further detailed:

```
show interfaces cos-queue te1/0/48
```

Figure 11 shows the output of the `show interfaces cos queue te1/0/48` command.

```
console#show interfaces cos-queue te1/0/48

Interface..................................... Te1/0/48
Interface Shaping Rate........................ 0 kbps
WRED Decay Exponent........................... 9

Queue Id   Min. Bandwidth   Scheduler Type   Queue Management
Type
--------   --------------   --------------   ------------------
--
0          0                Weighted         Tail Drop
1          0                Weighted         Tail Drop
2          0                Weighted         Tail Drop
3          0                Weighted         Tail Drop
4          0                Weighted         Tail Drop
5          0                Weighted         Tail Drop
6          0                Weighted         Tail Drop
```

Figure 11 Interface level cos-queue settings

To display mapping of the dot1p queue to the TCG, enter the following command:

```
show classofservice traffic-class-group
```

Figure 12 shows the output of the `show classofservice traffic class group` command.

```
console#show classofservice traffic-class-group

 Traffic Class    Traffic Class Group
---------------- -------------------
0                0
1                0
2                0
3                0
4                0
5                0
6                0
```

Figure 12 show classofservice traffic-class-group

5.2 Traffic Class Group Settings

To view the traffic class group bandwidth allocations, weights and scheduling; enter the following command:

```
show interfaces traffic-class-group
```

Figure 13 shows the output of the `show interfaces traffic class group` command.

```
console#show interfaces traffic-class-group

Global Configuration

Traffic Class    Min.        Max.                    Scheduler Type
Group            Bandwidth   Bandwidth    Weight
-------------    ---------   ---------    -------   --------------
0                10                50      100       Weighted Round Robin
1                0                 0       0         Strict
2                0                 0       0         Weighted Round Robin
```

Figure 13 `show interfaces traffic class group`

Just as the CoS queues can be set on an individual port level, so can the TCGs. To view the settings at the individual port level, enter the following command:

```
show interfaces traffic-class-group te1/0/48
```

Figure 14 shows the output of the `show interfaces traffic-class-group te1/0/48` command.

```
console#show interfaces traffic-class-group te1/0/48
Interface..................................... Te1/0/48

Traffic Class      Min.            Max.                        Scheduler
Group              Bandwidth       Bandwidth       Weight      Type
-------------      ---------       ---------       ------      ------------
0                  10              50              100         Weighted Round Robin
1                  0               0               0           Strict
2                  0               0               0           Weighted Round Robin
```

Figure 14 show interfaces traffic-class-group te1/0/48

5.3 show lldp dcbx interface *port* detail

To view the configuration for both the local and peer ports, the **`show lldb dcbx interface port detail`** command (where *port* is the port i.e. te1/0/48) can be used (Figure 15).

```
console#show lldp dcbx interface te1/0/48 detail

DCBX operational status:....................... Enabled
Configured DCBX version:....................... Auto
Peer DCBX version:............................. IEEE
Peer MAC:...................................... 00:1E:C9:DD:BB:09
Peer Description:..............................
Auto-configuration Port Role:.................. Manual
Peer Is configuration Source:.................. False

Error counters:
ETS incompatible configuration................. 0
PFC incompatible configuration................. 0
Disappearing neighbor.......................... 0
Multiple neighbors detected.................... 0

Local configuration:
PFC configuration (Tx enabled)
Willing: False  MBC:  False   Max PFC classes supported:  2
PFC enable vector:    0:0   1:0   2:0   3:0   4:1   5:0   6:0   7:0

ETS configuration (Tx enabled)
Willing: False  Credit shaper: False   Number of TCs supported: 3
Priority assignment:        0:0   1:0   2:0   3:0   4:1   5:0   6:0   7:0
Traffic class bandwidth (%):0:100 1:0   2:0   3:0   4:0   5:0   6:0   7:0
Traffic selection algorithm:0:2   1:2   2:2   3:2   4:2   5:2   6:2   7:2

Application priority (Tx enabled)
Type         Application    Priority    Status
------------ -------------- ----------- --------
TCP/UDP      0xcbc          4           Enabled

Peer configuration:
Willing: True  MBC:  False   Max PFC classes supported:  2
PFC enable vector:    0:0   1:0   2:0   3:0   4:1   5:0   6:0   7:0
ETS configuration
Willing: True Credit shaper: False   Number of TCs supported: 3
Priority assignment:        0:0   1:0   2:0   3:0   4:1   5:0   6:0   7:0
Traffic class bandwidth (%):0:100 1:0   2:0   3:0   4:0   5:0   6:0   7:0
Traffic selection algorithm:0:2   1:2   2:2   3:2   4:2   5:2   6:2   7:2

Application priority (Tx enabled)
Type         Application    Priority    Status
------------ -------------- ----------- --------
TCP/UDP      0xcbc          4           Enabled
```

Figure 15 show lldp dcbx interface te1/0/48 detail output

5.4 Priority Flow Control Settings

The PFC information should match between the devices. If the local device has PFC enabled for traffic class 4, but the peer device has PFC for traffic class 3 enabled, iSCSI traffic will not flow correctly. The highlighted sections in Figure 16 show PFC set up for traffic class 4 on both the local and peer device.

```
Local configuration:
PFC configuration (Tx enabled)
Willing: False  MBC:  False   Max PFC classes supported:  2
PFC enable vector:    0:0   1:0   2:0   3:0   4:1   5:0   6:0   7:0

ETS configuration (Tx enabled)
Willing: False  Credit shaper: False   Number of TCs supported: 3
Priority assignment:        0:0   1:0   2:0   3:0   4:1   5:0   6:0   7:0
Traffic class bandwidth (%):0:100 1:0   2:0   3:0   4:0   5:0   6:0   7:0
Traffic selection algorithm:0:2   1:2   2:2   3:2   4:2   5:2   6:2   7:2

Application priority (Tx enabled)
Type         Application    Priority    Status
------------ -------------- ----------- --------
TCP/UDP      0xcbc          4           Enabled

Peer configuration:
Willing: True  MBC:  False   Max PFC classes supported:  2
PFC enable vector:    0:0   1:0   2:0   3:0   4:1   5:0   6:0   7:0
ETS configuration
Willing: True Credit shaper: False   Number of TCs supported: 3
Priority assignment:        0:0   1:0   2:0   3:0   4:1   5:0   6:0   7:0
Traffic class bandwidth (%):0:100 1:0   2:0   3:0   4:0   5:0   6:0   7:0
Traffic selection algorithm:0:2   1:2   2:2   3:2   4:2   5:2   6:2   7:2

Application priority (Tx enabled)
Type         Application    Priority    Status
------------ -------------- ----------- --------
TCP/UDP      0xcbc          4           Enabled
```

Figure 16 Local and Peer PFC Settings - show lldp dcbx interface te1/0/48 detail output

Priority Flow Control No-Drop Settings

To confirm that the no drop queues were properly allocated enter the following command:

```
show interfaces priority-flow-control
```

Figure 17 shows the output of the `show interfaces priority flow control` command. These results are expected since the only configured no-drop queue is PFC 4.

```
console#show interfaces priority-flow-control

Port       Drop              No-Drop           Operational
           Priorities        Priorities        Status
--------   ---------------   ---------------   -----------
Te1/0/1    0-7                                 Inactive
Te1/0/2    0-3,5-7           4                 Active
Te1/0/3    0-3,5-7           4                 Active
Te1/0/4    0-3,5-7           4                 Active
Te1/0/5    0-3,5-7           4                 Active
Te1/0/6    0-3,5-7           4                 Active
Te1/0/7    0-3,5-7           4                 Active
Te1/0/8    0-3,5-7           4                 Active
Te1/0/9    0-7                                 Inactive
......
Te1/0/45   0-7                                 Inactive
Te1/0/46   0-7                                 Inactive
Te1/0/47   0-7                                 Inactive
Te1/0/48   0-3,5-7           4                 Active
Fo1/0/1    0-7                                 Inactive
```

Figure 17 show interfaces priority-flow-control

5.5 Enhanced Transmission Selection Settings

ETS is purposefully flexible and needs to be modified from device to device depending on network. However, the connected ports should match up (or at the very least not be reporting incompatible errors). This is where individual port configuration makes it possible to customize a network for maximum efficiency. Figure 18 shows the ETS and PFC incompatible configuration Error Counters in the show lldp dcbx interface te1/0/48 detail output.

```
Error counters:
ETS incompatible configuration................. 0
PFC incompatible configuration................. 0
Disappearing neighbor.......................... 0
Multiple neighbors detected.................... 0
```

Figure 18 ETS and PFC Error Counters - show lldp dcbx interface te1/0/48 detail output

5.6 DCBx settings- Willing vs. Non-Willing

Whether the system is configured to be willing or non-willing will cause different results to be seen.

Non-Willing

If a device is showing up as non-willing, the particular port (local or peer) is most likely set to be the configuration source or is set to be manually configured (Figure 19). Either of these will cause the port not to accept the values that are passed to it.

```
console#show lldp dcbx interface te1/0/48 detail

DCBX operational status:....................... Enabled
Configured DCBX version:....................... Auto
Peer DCBX version:............................. IEEE
Peer MAC:.....................................
00:1E:C9:DD:BB:09
Peer Description:...............................
Auto-configuration Port Role:................... Manual
Peer Is configuration Source:................... False
```

Figure 19 Non-Willing Auto-configuration Port Role

Willing

If a local device is configured as willing, then the local device's settings must match the peer's settings and the peer should be shown as the configuration source (Figure 20).

```
console#show lldp dcbx interface te1/0/23 detail

DCBX operational status:....................... Enabled
Configured DCBX version:....................... Auto
Peer DCBX version:............................. IEEE
Peer MAC:.....................................
00:1E:C9:DD:BD:0D
Peer Description:...............................
Auto-configuration Port Role:................... Auto-up
Peer Is configuration Source:................... True
```

Figure 20 Non-Willing Auto-configuration Port Role and Peer Is configuration Source

Again, the settings should match, with the exception of the local device being willing and the peer being non-willing.

show lldp dcbx interface all

The show lldp dcbx interface all command can be used to determine if a configuration source is selected, and to view the DCBx status of the ports.

```
show lldp dcbx interface all
```

The output from this command (Figure 21) provides the status of the LLDP on the port, and the port role assigned. It also provides what DCBx version has been allocated for the port (if it has been hard-set) and provides an overview of the DCBx TX and RX.

```
console#show lldp dcbx interface all

Is configuration source selected............... True
Configuration source port...................... Te1/0/23
                                              DCBX  DCBX  DCBX   unknown
Interface  Status   Role          Version     Tx    Rx    Errors TLV
---------- -------- ------------- --------- ----- ----- ------ -------
Te1/0/1    Enabled  Auto-down     Auto      86    0     0      0
Te1/0/2    Enabled  Auto-down     Auto      0     0     0      0
Te1/0/3    Enabled  Auto-down     Auto      0     0     0      0
Te1/0/4    Enabled  Auto-down     Auto      0     0     0      0
......
......
Te1/0/21   Enabled  Auto-down     Auto      0     0     0      0
Te1/0/22   Enabled  Auto-down     Auto      0     0     0      0
Te1/0/23   Enabled  Auto-up       Auto      86    86    0      0
Te1/0/24   Enabled  Auto-down     Auto      0     0     0      0
```

Figure 21 LLDP DCBx status for all switch ports

Clearing the Port Counters

When troubleshooting, it is helpful to start with a clean slate and observe the changes that occur. To do this, use the `clear counters` command at either the global level or interface specific level.

To clear all counters across the entire switch, enter the command:

```
clear counters
```

To clear the counters on a specific interface, enter the command:

```
clear counters <interface>
```

Note: This command will clear all the counters on the port, not just the LLDP/DCBx counters.

6 Example topology

The topology presented in Figure 22 is an example of one type of environment in which ETS for iSCSI would be useful. This is not a suggested topology, nor is it the only topology supported.

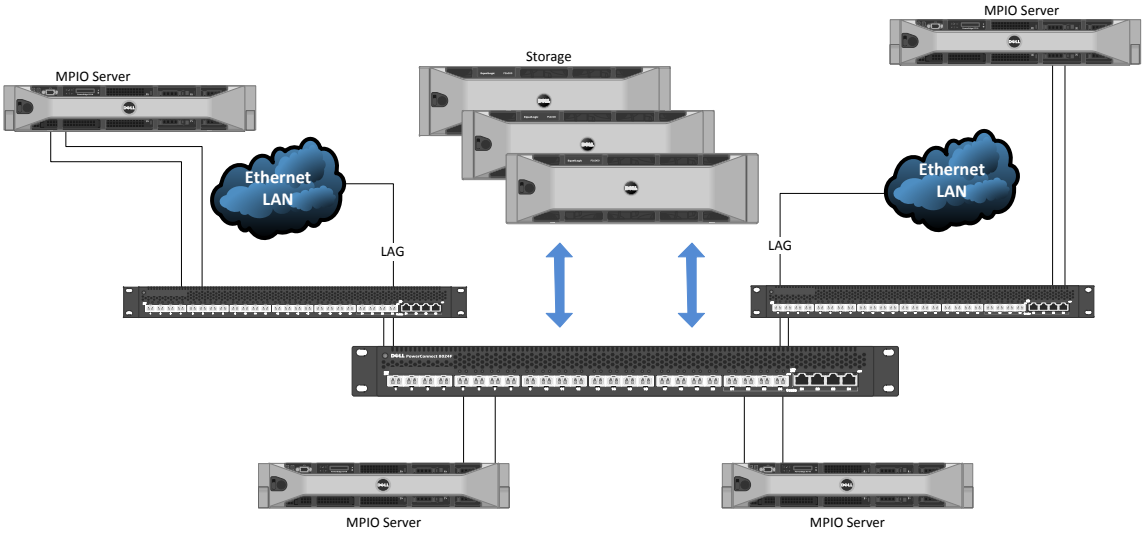

Figure 22 Example Topology

A Glossary

Class of Service (CoS): CoS allows preferential treatment to be given to certain types of traffic over others. CoS helps to streamline network traffic by taking the priority assigned to the packet and mapping it to a specific queue (queue mapping).

Configuration Source: The configuration source is the upstream peer that provides the DCBX configuration for the connected willing peers.

Data Center Bridging (DCB): An architecture type that helps to improve the reliability of Ethernet-based networks in the Data Center environment. DCB contains several protocols, including DCBx, PFC, and ETS. DCB combines several features from the IEEE 802.1Q standards.

Data Center Bridging Exchange (DCBx): Allows DCB devices to exchange configuration information. DCBx is addressed in the IEEE 802.1Qaz standard.

Dot1p: See VLAN priority.

Enhanced Transmission Selection (ETS): Allows bandwidth allocation to be applied on groups of queues. ETS is addressed in the IEEE 802.1Qaz standard.

Internet Protocol - Differentiated Services Code Point (IP DSCP): IP DSCP enables specific levels of service to be assigned to traffic on a network. It is based on RFC 2474 and 2475. IP DSCP can be backward compatible with IP Type of Service if configured carefully, but they are not interchangeable. Within the Dell Networking N series of switches, the DSCP values are mapped to various Class of Service traffic classes.

Minimum Bandwidth Guarantee: A percentage of the port's maximum negotiated bandwidth reserved for the queue.

Priority Flow Control (PFC): Provides a way to distinguish which traffic on a physical link is paused when congestion occurs based on the priority of the traffic. PFC is addressed in the IEEE 802.1Qbb standard.

Queue Management: Queue management is used to determine if packets are queued or dropped when the queue is full.

Queues: Queues are groups of data packets that are waiting to be transmitted. Queues are addressed in the IEEE 802.1p standard.

Round Robin: The Round Robin algorithm sends packets across each queue in a sequential order that allocates equitable distribution for all traffic. Round Robin is the default algorithm used by Dell Networking N series of switches when there is no congestion.

Scheduling: Used to determine how the packets are handled once they hit the queue.

Strict Priority: Strict Priority scheduling gives an absolute priority based on CoS queue or Traffic Class Group number.

Tail Drop: A very basic congestion mechanism that drops packets when the queue is full. No priority is given to any one packet. This is the default congestion queue management implemented by Dell Networking switches.

User Priority: See VLAN priority.

VLAN Priority Tagging: A VLAN priority tag is 32-bit field located between the source MAC and Ether Type/Size fields in the Ethernet Frame (Figure 23). By using VLAN priority tagging, it is possible to assign VLANs to appropriate queues. These queues can be further controlled by using additional Class of Service settings and Data Center Bridging.

Within this 32-bit field, the tag protocol identifier (TPID) and the tag control information (TCI) are present. The TCI consists of the Priority Code Point (PCP) (from the IEEE 802.1p standard), the Drop Eligibility Indicator (DEI) and the VLAN ID (VLAN ID).

The TPID simply makes sure that the equipment knows that the packet is a VLAN tagged frame. The priority code specifies the priority, which can range from 0 to 7 with the lowest being best effort and increasing in priority to level 7, which is considered the highest priority available. The Drop Eligible Indicator (DEI)is a one-bit field that is used along with the priority code to determine how the packet will be handled during congestion. The VLAN ID is simply the VLAN tag marking the packet for its specific VLAN.

The specifics on VLAN priority tagging can be found in the 802.1Q IEEE standard.

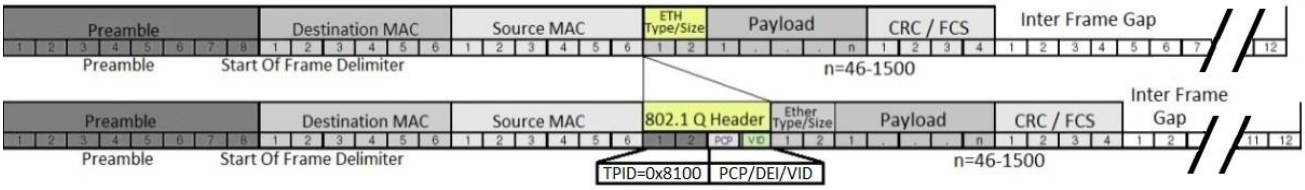

Figure 23 VLAN priority tag placement

Weighted Round Robin (WRR): Weighted Round Robin is built on top of the basic Round Robin algorithm. As congestion begins to occur, the allocated weights are brought in to the equation. The higher the weight, the more packets are allowed from a queue before moving on to the next queue. WRR serves every non-empty queue.

Weighted Deficit Round Robin (WDRR): Weighted Deficit Round Robin is another step in the evolution of Round Robin. Similar to WRR, it is based on weights that have been assigned. When congestion occurs, the packets at the head of every non-empty queue are checked against a deficit counter. If the deficit counter is greater than the packet size at the head of the queue, the queue is served. Otherwise, the queue is skipped and a

credit is given to the queue, this credit increases the deficit counter. Eventually, the deficit counter increases to the point that the queue is allowed to be services.

Weighted Random Early Detection (WRED): WRED functions by assigning various thresholds within a particular queue that trigger certain traffic classes to be dropped or queued.

Willing Peer: A willing peer is a device that actively takes the DCBX configuration from an upstream peer.

B Feedback

Readers are encouraged to provide feedback on the quality and usefulness of this publication by sending an email to to Dell_Networking_Solutions@Dell.com.