Lossless Network Deployment with RDMA for HPE Storage and Compute

Hewlett Packard
Enterprise

Open Source Code

This product includes code licensed under certain open source licenses which require source compliance. The corresponding source for these components is available upon request. This offer is valid to anyone in receipt of this information and shall expire three years following the date of the final distribution of this product version by Hewlett Packard Enterprise Company. To obtain such source code, please check if the code is available in the HPE Software Center at https://myenterpriselicense.hpe.com/cwp-ui/software but, if not, send a written request for specific software version and product for which you want the open source code. Along with the request, please send a check or money order in the amount of US $10.00 to:

Hewlett Packard Enterprise Company
Attn: General Counsel
WW Corporate Headquarters
1701 E Mossy Oaks Rd, Spring, TX 77389
United States of America

Contents

About this Guide

This guide provides information on deploying Remote Direct Memory Access over Converged Ethernet (RoCE) solutions. It also provides information on the typical use cases of RoCE.

Intended Audience

This guide is intended for users managing Aruba-CX Data Center Bridging capable switches which are being deployed in a lossless networking solution.

Applicable Products

This document applies to the following products:

- Aruba 8100 Switch Series (R9W86A, R9W87A, R9W88A, R9W89A, R9W90A, R9W91A, R9W92A, R9W93A)
- Aruba 8325 Switch Series (JL624A, JL625A, JL626A, JL627A)
- Aruba 8360 Switch Series (JL700A, JL701A, JL702A, JL703A, JL706A, JL707A, JL708A, JL709A, JL710A, JL711A, JL700C, JL710C, JL711C, JL704C, JL705C, JL719C, JL718C, JL717C, JL720C, JL722C, JL721C)
- Aruba 8400 Switch Series (JL363A, JL365A, JL666A)
- Aruba 9300 Switch Series (R9A29A, R9A30A, R8Z96A)
- Aruba 10000 Switch Series (R8P13A, R8P14A)

As of May 2023, the Aruba CX 8325 as well as the Aruba CX 9300 Switch Series have been validated and approved by the HPE Storage Networking Stream. For more information, refer to the latest Validation Scope.

Overview of RoCE

Remote Direct Memory Access over Converged Ethernet (RoCE) is a network protocol that allows Remote Direct Memory Access (RDMA) over an Ethernet network. RDMA enables the movement of data between servers with very little CPU involvement. RoCE integrates the benefits of Ethernet and Data Center Bridging (DCB) with RDMA techniques and provides lower CPU overhead and increases enterprise data center application performance.

With increasing applications of AI, ML and High-Performance Computing (HPC), RoCE is designed to support increasingly data intensive applications ensuring higher performance and lower latency. A dependable transport is essential since database sizes are constantly growing and there is a necessity for high bandwidth for the transfer of data between processing nodes. RoCE is designed for excellent performance inside an advanced data center architecture by converging compute, network, and storage onto a single fabric. Converged solutions, such as RoCE, carry both lossless and lossy traffic and is therefore gaining popularity in modern data centers. The growth of hyper-convergence, IP storage and RoCE based solutions have grown due to solutions like Server Message Block direct.

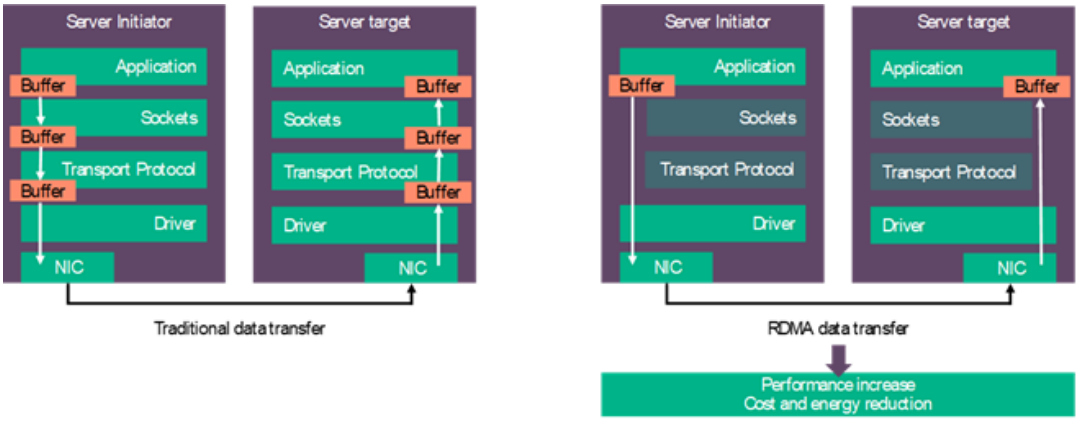
This chapter includes the following topics:

Remote Direct Memory Access

RDMA is a technology that provides direct memory access. RDMA facilitates data transfers between the working memory of two systems without taxing the CPUs of either system. When an application employs RDMA, data movement is handled by the RDMA-capable NIC, reducing CPU overhead for these operations. This allows server CPU resources to be utilized for other activities without sacrificing the I/O performance. RDMA ensures lower latency, improved resource utilization, flexible resource allocation, fabric unification, and scalability. By increasing the server productivity, RDMA reduces the need for additional servers and lowers the total cost of ownership.

The following figure illustrates the RDMA process:

Figure 1 *RDMA Data Transfer*



RoCE Topology

RoCE leverages the functions of RDMA to expedite and accelerate communications between applications hosted and running on clusters of servers and storage arrays. RDMA was first used with InfiniBand (IB) fabrics. RoCE is primarily established by replacing the IB link layer with an Ethernet link layer to transfer data. Data centers can benefit from RDMA using a converged, high performance infrastructure that supports TCP/IP.

The following table describes the structure of RoCE:

Table 1: *RoCE Topology*

	1-tier Core	2-tier Layer 2	Spine and Leaf - Underlay
Description	▪ Initiators or targets connected to DC Core (VSX Pair) ▪ **Pure L2 solution**	▪ Initiators or targets connected to access switches ▪ VSX LAGs southbound to servers, northbound to L2 Core ▪ **Pure L2 solution**	▪ Initiators or targets connected to leaf switches ▪ Leaf switches have VSX LAGs southbound to servers, northbound to L2 Core ▪ **L3 solution** ▪ **With HPE Aruba Networking CX 10000, 9300, 8360, and 8100 Switch Series, IP ECN is supported on the underlay only. There is no ECN support on these platforms in VXLAN overlay.** ▪ **With HPE Aruba Networking CX 8325 Switch Series, IP ECN is supported on the**

	1-tier Core	2-tier Layer 2	Spine and Leaf - Underlay
			underlay and the VXLAN overlay - Single hop from leaf-to-leaf.

The following are the two versions of RoCE:

- RoCE v1
- RoCE v2

RoCE v1

The RoCE v1 protocol is an Ethernet link layer protocol allowing two hosts in the same Ethernet broadcast domain to communicate.

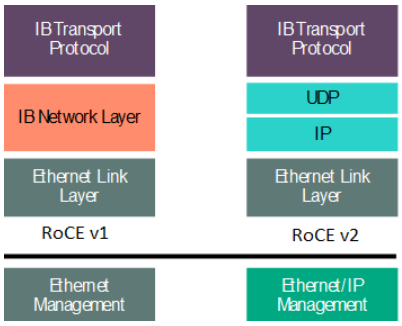
RoCE v2

The RoCE v2 protocol replaced the IB network layer with a standard IP and UDP header enabling the traffic to be routed. RoCE v2 is used on both layer-2 and layer-3 networks as packet encapsulation includes IP and UDP headers. Layer-3 routing is made possible by this, bringing RDMA to networks with multiple subnets for increased scalability.

To deploy RoCE v2, see Deploying RoCE v2 on HPE Aruba Networking CX Switches.

The following figure differentiates the structures of RoCE v1 and RoCE v2:

Figure 2 *RoCE v1 and RoCE v2*



RoCE Use Cases

RDMA and RoCE offer extensive possibilities in the networking market. Since RoCE has direct access to memory data it can facilitate low-latency and high-performance transmission. It also ensures low CPU involvement, increases productivity by improving latency and throughput, and reduces cost by using Ethernet infrastructure to handle the massive amount of data.

The following list of markets have started to leverage these solutions:

- Cloud Computing
- Data Storage
- Financial Services
- Web 2.0 Big Data

Cloud Computing

The cloud computing market has been actively leveraging the benefits of RDMA and RoCE. These environments obtain benefits, such as improved SLAs through deterministic performance and efficient clustering allowing for elasticity and scale out computing. As indicated before, the benefits of implementing RoCE include lower cost of ownership and greater return on investment that spans across traditional and modern hyper converged infrastructures.

The following vendor solutions provide a few examples of applications that leverage RDMA/RoCE.

- VMware®
  - VMware has published Ultra-Low Latency on vSphere with RDMA which compares RDMA vs regular TCP/IP stack in a cloud environment.
- Microsoft® Azure
  - Azure achieved a 40 Gbps throughput at 0% CPU utilization through a 40 GbE implementation of RoCE solution.
- Red Hat® Kernel-based Virtual Machine (KVM)
- Citrix® XenServer
- Amazon Elastic Compute Cloud
- Google™ App Engine

Data Storage

Many data storage focused applications are gaining benefits from implementing RDMA/RoCE such as SMB Direct and Azure Managed Lustre. Data storage protocols over RDMA deliver higher throughput and lower latency. Everyday applications such as Microsoft Exchange and Microsoft SharePoint are also able to leverage and benefit from RoCE.

Financial Services

The financial services market has been leveraging InfiniBand as those environments often require low latency.

The following are examples of some applications that could see high performance and I/O benefits in RoCE solutions:

- TIBCO Software Inc.
- IBM WebSphere MQ
- Red Hat® Enterprise MRG Realtime
- 29West by Informatica

Web 2.0 Big Data

Big data environment is another segment that benefits from RoCE solutions. The goal in these big data environments is to minimize the response time and increase the I/O.

The following are examples of some applications that could benefit from these RoCE solutions:

- Storage Spaces Direct (S2D):
  - Software-defined storage (SDS) stack in a Windows Server enables building highly-available (HA) storage systems with local storage.
- SMB 3 Direct:

    - SMB 3 Direct is an extension of the Microsoft Server Message Block technology used for file operations. Direct implies the use of high speed RDMA networking methods to transfer large amounts of data with little CPU intervention
- Other workload types that benefit in the Big Data enviroment using RoCE (Big Data, Databases, DFS, Cloud):
    - MicroSoft Azure Stack HCI via RoCEv2
    - Lustre® Open Source
    - Oracle RAC
    - IBM DB2 pureScale
    - Microsoft® SQL Server
    - Apache Hadoop
    - Eucalyptus
    - Apache Cassandra

Lossless Network

A lossless network is one where the devices comprising the network fabric are configured to prevent packet loss by using DCB protocols. One of the primary objectives of designing network solutions that incorporate RoCE is to deploy a lossless fabric. Even though the RoCE standards do not demand lossless networks, RoCE performance can suffer when a lossless network is not provided. With that as a premise, it is recommended that a lossless fabric be considered as requirement for RoCE implementations. Lossless fabrics can be built on Ethernet fabrics by leveraging DCB protocols.

Data Center Bridging Protocols

The following DCB protocols are leveraged to build lossless fabrics on Ethernet fabrics:

- Priority-based Flow Control
- Enhanced Transmission Selection
- Data Center Bridging Exchange
- Quantized Congestion Notification
- IP Explicit Congestion Notification

Priority-based Flow Control

IEEE standard 802.1Qbb is a link-level flow control mechanism. The flow control mechanism is similar to that used by IEEE 802.3x Ethernet PAUSE, but it operates on individual priorities. Instead of pausing all traffic on a link, Priority Flow Control (PFC) allows you to selectively pause traffic according to its class.

Enhanced Transmission Selection

Enhanced Transmission Selection (ETS) provides a configurable bandwidth guarantee for all queues. For a PFC queue, this ensures a guaranteed percentage of the link bandwidth is available, so that the lossless flow is not starved for bandwidth by flows in other queues. Through this configuration, each traffic queue gets a minimum amount of bandwidth.

Data Center Bridging Exchange

Data Center Bridging Exchange (DCBx) protocol helps to ensure that the NIC and the switch are configured correctly.Data Center Bridging Capability Exchange protocol (DCBX) is a discovery and exchange protocol for communicating configuration and capabilities among neighbors to ensure consistent configuration across the data center bridging network.The protocol allows auto exchange of Ethernet parameters and discovery functions between switches and endpoints. If the server NIC has enabled DCBx willing mode, DCBx will ensure the server NIC knows how to mark and treat traffic on that link after the switch is configured with the needed DCB and traffic marking parameters.

Quantized Congestion Notification

Quantized Congestion Notification (QCN) protocol provides a means for a switch to notify a source that there is congestion on the network. The source then reduces the flow of traffic. This helps to keep the critical traffic flowing while also reducing the need for pauses. This is only supported in pure layer-2 environments and seen very rarely now that RoCE v2 is the predominant RoCE solution.

HPE Aruba Networking CX switches with DCB-based solutions do not support QCN.

IP Explicit Congestion Notification

IP Explicit Congestion Notification (IP ECN) is not officially part of the DCB protocol suite, however, RoCE v2 supports ECN and sends Congestion Notification Packets (CNP) to an endpoint when congestion is signaled via the IP ECN bits on traffic originating from that endpoint. ECN must be enabled on both endpoints and on all the intermediate devices between endpoints for ECN to work properly. ECN notifies end nodes and connected devices about congestion with the goal of reducing packet loss and delay by signaling the sending device to decrease the transmission rate until congestion clears. When used in conjunction with PFC, ECN helps endpoints adjust their transmit rate before a PFC pause becomes necessary thereby improving performance.

The following table summarizes the DCB protocols for RoCE v2:

Table 2: *DCB Protocols for RoCE v2*

Parameter	RoCE v2	Notes
PFC	Yes	Must use always. If not used, in times of congestion the RDMA advantages will not be achieved.
ETS	Yes	Must use in converged environments. If not used, lossless traffic classes can get starved for bandwidth.
DCBx	Yes	Not mandatory, but recommended.
QCN	No	Congestion Notification (CN) helps to address pause unfairness and victim flow issues. When used with PFC, PFC acts as fast acting mechanism to address microbursts, while CN smooths out traffic flows helping to reduce pause storms under normal load. **NOTE:** (Not supported on HPE Aruba Networking CX)

Parameter	RoCE v2	Notes
IP ECN	Yes	Highly recommended for L3 RoCE v2. ECN helps to address pause unfairness and victim flow issues. When used with PFC, PFC acts as fast acting mechanism to address microbursts, while CN smooths out traffic flows helping to reduce pause storms under normal load.

HPE Storage Offerings

HPE offers several storage configurations based on modular building blocks and offers cost-optimized, balanced, and high performing options. These options allow flexibility in the types of solutions and configurations that are used. Depending on the chosen option, lossless networking through RoCEv2 may be supported.

The following storage options are offered by HPE:

- Mission Critical Storage
- Business Critical Storage
- General Purpose Storage
- Data Protection

Mission Critical Storage

Mission critical storage allows you to drive your most demanding workloads with the mission-critical reliability and unprecedented agility of HPE Alletra 9000 and HPE Primera. It consolidates traditional and modern apps with All-NVMe performance and ultra-low latency, backed by a 100% availability guarantee.

Business Critical Storage

Business critical storage allows you to harness maximum efficiency for your business-critical workloads with strict availability and performance SLAs. HPE Alletra 6000 and HPE Nimble All Flash Arrays deliver fast, consistent performance, effortless scale, industry-leading data efficiency, and 99.9999% availability guarantee.

General Purpose Storage

General purpose storage leverages a cloud operational consumption experience, guarantees 99.9999% availability, and a flexible scale to easily power your general-purpose apps. It ensures zero tuning, zero trade-offs, and zero wasted resources.

Data Protection

Data protection is employed to protect any app and meet any SLA with a suite of cloud-based and on-premises data protection solutions that deliver the industry's best ransomware protection, effortless cloud backup, instant restores, and an economical data archive.

Deploying RoCE v2 on HPE Aruba Networking CX Switches

The following components are required for deploying RoCE v2 on HPE Aruba Networking (HPE ANW) CX switches:

- DCB Ethernet and DCB capable switches to ensure a lossless networking fabric.
  - PFC to stave off loss.
  - ETS to protect Traffic Classes TC.
  - ECN (and PFC) over L3 links.
- NICs supporting RoCE.

NOTE

The deployment scenarios are applicable to HPE ANW CX 10000, 9300, 8400, 8360, 8325, and 8100 Switch Series.

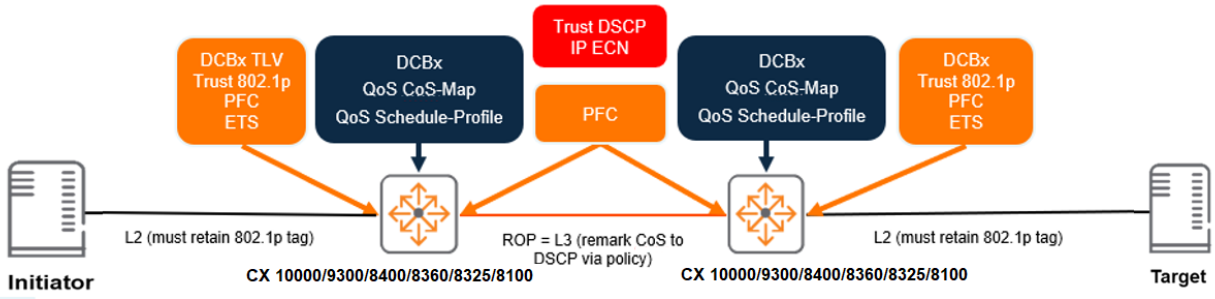
NOTE

As of May 2023, the HPE ANW CX 8325 as well as the HPE ANW CX 9300 Switch Series have been validated and approved by the HPE Storage Networking Stream. For more information, refer to the latest Validation Scope.

In RoCE-based solutions, the hosts (initiators or targets) are configured to ensure that the lossless traffic flows are sent to the attached switches with the correct 802.1p or DSCP values. The switches in the fabric leverage L2 links facing the initiators and targets as well as L3 links between the switches.

The following figure describes a RoCE v2-based solution on HPE ANW CX switches:

Figure 3 *RoCE v2 Configuration*



To deploy RoCE v2 onHPE ANW CX switches, complete the following configuration steps:

1. Enable the DCBx Link Layer Discovery Protocol (LLDP).
2. Review the default Class of Service (CoS) map and queue profile, make changes (if required) to ensure 802.1p tagged frames use the desired queue.
3. Set the initiator or target facing port to trust mode for all the interfaces carrying RoCE traffic respectively.

4. Apply a schedule profile so that it allocates the appropriate bandwidth to all queues and ensures the bandwidth allocation of the lossless RoCE queue.
5. Enable PFC for the 802.1p priority on all initiator or target facing interfaces.
6. Use the DCBx application TLV to inform the attached host to send lossless traffic marked with the correct 802.1p code point. This allows the switch to recognize the lossless traffic and treat it properly.

If the attached host does not have the **willing bit** turned on, then the administrator must manually configure the Hypervisor to mark the traffic accurately.

7. Configure ECN and trust DSCP on L3 switch-to-switch links and ensure that a remark policy (CoS to DSCP) has been applied.

The following procedures describe the configuration of RoCE v2 on HPE ANW CX switches:

1. Configuring LLDP and DCBx
2. Configuring QoS Pool
3. Configuring QoS Queue Profile
4. Configuring QoS Schedule Profile
5. Configuring Global Trust
6. Configuring PFC
7. Configuring Flow-Control Watchdog
8. Configuring DCBx Application TLVs
9. Configuring ECN Threshold Profile

DCB solutions enable lossless Ethernet. It is also important to follow the HPE recommended ratio of 7:1 for typical distributed data access.

The following table describes the I/O workload examples:

Table 3: *I/O Workload Example*

I/O Workload	Recommended Ratio
Higher I/O data intensive application requirements: (> 70 MB/s at 2 Gb/s, > 140 MB/s at 4 Gb/s, > 280 MB/s at 8 Gb/s, > 560 MB/s at 16 Gb/s, > 1120 MB/s at 32 Gb/s)	1:1 to 3:1
Lower I/O data intensive application requirements: (< 70 MB/s at 2 Gb/s, < 140 MB/s at 4 Gb/s, < 280 MB/s at 8 Gb/s, < 560 MB/s at 16 Gb/s, < 1120 MB/s at 32 Gb/s)	7:1 to 15:1

Configuring LLDP and DCBx

DCBx is a discovery and capability exchange protocol which operates between two directly-connected devices. The exchanged information is used in troubleshooting mismatches between the devices. DCBx is part of the IEEE 802.1Qaz-2011 standard.

DCBx uses LLDP as the underlying protocol for exchange of parameters with peers. The DCBx parameters are exchanged as LLDP Type Length Values (TLVs).

The following DCBx versions are supported by HPE ANW CX switches:

- IEEE DCBx
- CEE DCBx

Following are the guidelines for DCBx:

- DCBx is disabled, by default.
- LLDP must be enabled on the DCBx supporting interfaces.
- LLDP DCBx can be enabled either globally or at an interface-level.
- DCBx is supported on physical interfaces only as it relies on LLDP. It is not supported on management or logical interfaces.
- DCBx willing mode is set to 0 in all the TLVs in Aruba CX switches. This informs the peer that the switch is not willing to change its configuration to match the peer's configuration.
- Both ends of the link must be configured to use the same version of DCBx (IEEE or CEE), otherwise an error will occur.

To configure LLDP and DCBx in HPE ANW CX 10000, 9300, 8400, 8360, 8325, and 8100 Switch Series, complete the following steps:

1. Enable LLDP:

```
switch(config)# lldp
```

LLDP is enabled by default ("no" form disables).

2. Verify the LLDP status:

```
switch(config)# show lldp configuration
```

3. Enable DCBx:
   a. Globally, optionally specifying the version:

IEEE is the default version, if not specified.

```
switch(config)# lldp dcbx version cee
```

To enable IEEE at interface level:

```
switch(config)# lldp dcbx
or
switch(config)# lldp dcbx version ieee
```

b. On an interface, optionally specifying the version:

IEEE is the default version, if not specified.

```
        switch(config-if)# lldp dcbx cee
```

4. Verify the DCBx status:

```
switch(config)# show dcbx interface <IFNAME>
```

Configuring QoS Pool

The Quality of Service (QoS) pool creates a packet buffer pool for lossless traffic by configuring lossless pool size, headroom buffer associated with the pool, and the priorities that are mapped to the pool.

The lossless pool size is a percentage of the total available buffer memory on the device. The headroom pool memory is allocated from the lossless pool and is used for storing packets that arrive on a port after pause has been asserted.

The following CLI command configures the QoS pools for the HPE ANW CX 10000, 9300, and 8325 Switch Series:

```
switch(config)# qos pool {1 | 3} lossless size <factory-default | PERCENT>
percent headroom
<factory-default | KBYTES> kbytes priorities <PRIORITY>
```

The following CLI command configures QoS pool 1 and assigns a single packet priority (4):

```
switch(config)# qos pool 1 lossless size 60 percent headroom 2048 kbytes
priorities 4
```

The following CLI command configures QoS pool 1 and assigns two packet priorities (3 and 4):

```
switch(config)# qos pool 1 lossless size 60 percent headroom 2048 kbytes
priorities 3,4
```

The following table describes the QoS pool configuration parameters:

Table 4: *QoS Pool Parameters*

Parameter	Description
`{1 | 3}`	Specifies the lossless pool number. The valid values are 1-3 (HPE ANW CX 8325, 9300, and 10000 Switch Series only).

Parameter	Description
	NOTE: HPE ANW CX 8100, 8360, and 8400 Switch Series support only 1 pool.
<PERCENT>	Specifies the lossless pool size in percent. Default: 40.6. Range: 10 to 90 percent. The percentage of packet buffer memory to be allocated to this pool in integer format. or Range: 10.00 to 90.00. The percentage of packet buffer memory to be allocated to this pool in decimal format.
<factory-default>	Specifies the default lossless pool or headroom buffer sizes.
<PRIORITY>	Specifies the PFC priorities to be mapped to the pool. Range: 0 to 7. To map multiple priorities, use a comma-separated list. For example: 1,3,6.

The following CLI command displays the QoS pool usage:

```
switch# show qos pool statistics


Packet-Buffer Pool Statistics   Recent-Peak Interval: 1427 seconds
Packet Buffer Pools   Total Size     Peak Use   Recent Peak  Current Use
-------------------   ----------   ----------   -----------  -----------
Lossy Pool                19036           20            20            0
Lossless Pool 1            8679         5579          5579            0
Headroom 1                 3078            0             0            0
Lossless Pool 2               0            0             0            0
Headroom 2                    0            0             0            0
```

Configuring QoS Queue Profile

The QoS queue profile ensures that 802.1p marked traffic is placed into the correct queues.

Each PFC priority must be configured with its own queue separate from any other lossy or lossless priority. For example, if there are two PFC priorities being configured, then a minimum of 3-queue QoS queue profiles must be created.

The following CLI command configures 2-queue QoS profiles for HPE ANW CX 10000, 9300, 8400, 8360, 8325, and 8100 Switch Series, where local-priority 4 is the only traffic mapped to queue 1 and could be configured as lossless.

```
switch(config)# qos queue-profile SMB
switch(config-queue)# map queue 0 local-priority 0
switch(config-queue)# map queue 0 local-priority 1
switch(config-queue)# map queue 0 local-priority 2
```

```
switch(config-queue)# map queue 0 local-priority 3
switch(config-queue)# map queue 1 local-priority 4
switch(config-queue)# map queue 0 local-priority 5
switch(config-queue)# map queue 0 local-priority 6
switch(config-queue)# map queue 0 local-priority 7
```

The following CLI command displays the QoS queue profile:

```
switch(config)# show qos queue-profile SMB
```

Configuring QoS Schedule Profile

The QoS schedule profile determines the order in which queues transmit a packet and the amount of service defined for each queue. A QoS schedule profile is always applied on all interfaces, either from user configuration or using factory-default. The configuration may specify a global schedule profile for all ports, as well as a specific schedule profile on a per-interface basis. If both global and interface-specific schedule profiles are configured, the interface-specific schedule profile is used for the interface. A schedule profile should be configured on interfaces carrying lossless traffic so that lossless flows are allocated an appropriate bandwidth guarantee for the environment.

HPE ANW CX switches are automatically provisioned with a schedule profile named factory-default, which assigns the Weighted Fair Queueing (WFQ) and Deficit Weighted Round Robin (DWRR) scheduling algorithms to all queues with a weight of 1.

All interfaces use the factory-default scheduling profile.

HPE ANW CX switches also have a pre-defined strict schedule profile that services each queue until it is empty and in descending order of queue priority (for example, 7-0). Although the strict schedule profile is not applied by default, it can be used on any interface and with any queue profile.

The following configurations are permitted for a schedule profile:

- All queues use the same scheduling algorithm, DWRR.
- All queues use strict priority.
- The highest queue number uses strict priority, and the remaining lower queues use the same algorithm, DWRR.

The following configuration changes are permitted on an applied custom schedule profile:

- Weight of a DWRR queue.
- Bandwidth of a strict queue.
- Algorithm of the highest numbered queue can be swapped between DWRR and strict queue, and vice versa.

Only the configuration changes mentioned above are allowed to be performed on an applied custom schedule profile. Any other changes render the custom schedule profile unusable and the switch reverts to the default profile until the profile changes are corrected.

To configure the 2-queue QoS schedule profile for HPE ANW CX 10000, 9300, 8400, 8360, 8325 and 8100 Switch Series, complete the following steps:

1. Create a schedule profile (no form removes). For example, test SMB1:

```
switch(config)# qos schedule-profile SMB1
```

2. Configure each queue with appropriate bandwidth or algorithm:

```
switch(config)# dwrr queue 0 weight 15
switch(config)# dwrr queue 1 weight 15
```

3. Verify the schedule profile:

```
switch(config)# show qos schedule-profile SMB1
```

4. Apply QoS queue profile and schedule profile:

```
switch(config)# apply qos queue-profile SMB schedule-profile SMB
```

The configuration settings use a weight to set the amount of available bandwidth for each queue. The configuration settings ensure that 50% of bandwidth is applied to both queue 0 and queue 1.

For more information about schedule profiles, see AOS-CX Quality of Service Guide.

Configuring Global Trust

The appropriate trust configurations must be applied to the relevant ports. With RoCE-based solutions that largely rely on the 802.1p marking, users must ensure that the markings are being trusted.

It is therefore recommended to set the Global Trust mode to CoS. This is applied to all interfaces that do not already have an individual trust mode configured. A DSCP override can then be applied to any layer-3 interfaces that do not carry a 802.1p tag.

The following CLI command configures Global Trust for HPE ANW CX 10000, 9300, 8400, 8360, 8325, and 8100 Switch Series:

```
switch(config)# qos trust
cos   Trust 802.1p priority and preserve DSCP or IP-ToS
dscp  Trust DSCP and remark the 802.1p priority to match
```

For RoCE v2, trust CoS or trust DSCP can be used depending on the method set for the endpoints when transmitting lossless packets. If there are no VLAN tags, then trust DSCP is used. If there are VLAN tags, then users can choose between trust CoS or trust DSCP depending on which packet fields the endpoints are known to populate with the correct packet priority.

The following CLI commands display Global Trust information for HPE ANW CX 10000, 9300, 8400, 8360, 8325 and 8100 Switch Series:

```
switch(config)# show qos trust
qos trust cos

switch(config)# show qos trust
qos trust dscp
```

Configuring PFC

PFC enables flow control over a unified 802.3 Ethernet media interface, for LAN and SAN technologies. PFC is intended to prevent loss of packets of a specific priority due to congestion in a device when a link is experiencing contention for bandwidth. This allows loss-sensitive protocols, such as RoCE, to coexist with traditional loss-tolerant protocols over the same unified fabric.

While PFC is supported on theHPE ANW CX 10000, 9300, 8400, 8360, 8325 and 8100 Switch Series, the caveat for PFC (as well as for LLFC RxTx) on the HPE ANW CX 10000 Switch Series is that flows arriving on port priorities configured for PFC are not seen by Distributed Services Processor. Therefore, these flows are not visible to monitoring and any configured rules.

Important

The following points are to be noted prior to the configuration:

- To ensure lossless behavior for a given deployment, PFC must be enabled on all endpoints and switches in the flow path.
- PCP 0 should not be used for PFC lossless queue. Additionally, it is also recommended to not configure PFC for PCP 7 since some switch-generated protocol and control packets use this priority. There can be protocol-related issues if these packets are stuck in a paused queue.
- All flow-controlled priorities must first be mapped to a lossless pool using the **qos pool** command.
- PFC should be configured on both ends of a link that requires lossless networking.
- On layer-3 interfaces, PFC uses DSCP values for classification. In case VLAN tags are used, then PFC uses 802.1Q tags instead of DSCP values.

On the HPE ANW CX 10000 Switch Series and HPE ANW CX 8325 Switch Series, when an interface priority is configured for PFC, the packet field that is used for determining whether the packet is lossy or lossless may differ from the packet field used for local-priority classification (which is based on the configured trust mode). While the QoS trust mode determines local priority (and the resulting transmit queue), the local priority value is only used to determine the lossy or lossless nature of the packet for untagged frames.

The following points provide information on the configuration of PFC priorities specific to HPE ANW CX switches:

- Seven lossless PFC priorities per interface can be configured on HPE ANW CX 10000, 9300, and 8325 Switch Series (isolated resources).
- Two lossless PFC priorities per interface can be configured on HPE ANW CX 8360, 8100 Switch Series (shared resources).

- One lossless PFC priority per interface can be configured on HPE ANW CX 8400 Switch Series (shared resources).

When PFC is configured on an interface, the intent is that packets arriving on that interface marked with that priority should not be dropped.

The following CLI command enables PFC for arriving packets of priority 4 on interface 1/1/2 (HPE ANW CX 10000, 9300, 8400, 8360, 8325, and 8100 Switch Series):

```
switch(config)# interface 1/1/2
switch(config-if)# flow-control priority 4
```

For more details on PFC, see the Aruba CX Fundamentals Guide.

Configuring Flow-Control Watchdog

The Flow-Control Watchdog feature enables monitoring of PFC-enabled queues on a physical interface. When a lossless queue is paused (that is, lossless packets in the queue cannot be transmitted) for an excessive duration of time, problematic lossless buffer congestion can occur throughout the network. To prevent such a situation, egress lossless queues are monitored to detect when no transmissions have occurred for a globally specified detection timeout. When the condition is detected, the Flow-Control Watchdog triggers for the affected queue resulting in the following actions:

- The Flow-Control Watchdog timeout counter on the interface is incremented.
- All packets occupying the affected queue are discarded.
- New packet arrivals destined for the affected queue are discarded.

After the configured resume interval has elapsed since the trigger, the queue is returned to normal operation.

Flow-Control Watchdog is only supported on interfaces configured with PFC. Link-level flow control is not compatible with Flow-Control Watchdog. When Flow-Control Watchdog is enabled, it is active on all lossless queues of the port.

The Flow-Control Watchdog feature is supported only on HPE ANW CX 10000, 9300, and 8325 Switch Series.

The following command determines whether the Flow-Control Watchdog is enabled on an interface, or to view the number of times the Flow-Control Watchdog has been triggered:

```
show interface flow-control
```

To configure Flow-Control Watchdog, complete the following steps:

1. Enable Flow-Control Watchdog for an interface:

   ```
   switch(config-if)# flow-control watchdog
   ```

2. Enable Flow-Control Watchdog timeouts for an interface:

```
switch(config)# flow-control watchdog timeout <MILLISECONDS> resume
<MILLISECONDS>
```

3. Revert to factory default Flow-Control Watchdog timeout and resume values:

```
switch(config)# no flow-control watchdog timeout <MILLISECONDS> resume
<MILLISECONDS>
```

The following table provides information on the parameters used in the commands:

Table 5: *Flow-Control Watchdog Parameters*

Parameter	Description
`timeout <MILLISECONDS>`	Specifies the amount of time in milliseconds, that a queue must be paused for watchdog to trigger. ■ Range: 10 to 1500 milliseconds. ■ Default: 100 milliseconds.
`resume <MILLISECONDS>`	Specifies the duration of time in milliseconds, that a queue remains in the triggered state. ■ Range: 1 to 100000 milliseconds. ■ Default: 100 milliseconds.

Configuring DCBx Application TLVs

The DCBx application to priority map Type Length Value (TLV) gets advertised in the DCBx application priority messages sent to the attached devices. These messages tell the DCBx peer (with willing bit on) to send the application traffic with the configured priority so that the network can receive and queue traffic properly.

Multiple applications can be configured in this manner.

If the attached device does not honor the DCBx application TLVs, then the device has to be configured manually to mark traffic correctly.

The following CLI command configures an application priority for advertisement by DCBx.:

```
switch(config)# dcbx application {ISCSI | TCP-SCTP <PORT-NUM> |
TCP-SCTP-UDP <PORT-NUM> | UDP <PORT-NUM> | ether <ETHERTYPE>}
priority <PRIORITY>
```

The following CLI command configures DCBx to advertise iSCSI traffic by using priority 4:

```
switch(config)# dcbx application iscsi priority 4
```

The following CLI command configures DCBx to advertise tcp-sctp port 860 traffic by using priority 4:

```
switch(config)# dcbx application tcp-sctp 860 priority 4
```

Configuring ECN Threshold Profile

Explicit Congestion Notification (ECN) enables network congestion notification between an ECN-enabled sender and an an ECN-enabled receiver on TCP/IP-based networks. ECN threshold profiles set up individual queue utilization thresholds as triggers for taking action (that is, ECN marking) on a packet. A threshold profile is applied per-port and defines the threshold and action for each queue.

ECN configuration supports a range of values which offers users an opportunity to tune the parameters to achieve the best performance. When used in conjunction with PFC, ECN offers the ability to improve performance by giving endpoints feedback on congestion before PFC is required to avoid buffer overflow. Following are the results if ECN thresholds are configured too high or too low:

- Configuring ECN thresholds too high could result in no benefit since the thresholds won't be hit before a PFC pause occurs to avoid loss.
- Configuring ECN thresholds too low reduces or eliminates burst absorption in the network, reducing performance if not enough packets can be buffered in the network to maintain traffic at line rate when needed.

As a starting point, ECN can be configured with min-threshold of 500 kbytes and a max threshold of 1500 kbytes, providing a 1 megabyte range for increasing probability-based congestion marking of packets as egress queue utilization increases from 500 kbytes to 1500 kbytes. Depending on the link speed, round-trip latency, and amount of over-subscription of interfaces in the data path, it may be necessary to increase these values to allow the buffer use to grow larger before signaling congestion.

Higher link speeds can drain packet buffers faster, benefiting from larger buffers to maintain link utilization when multiple bursty flows are competing for bandwidth. In addition, higher round-trip latency means it takes longer for the sending device to receive the notification of network congestion and react.

Experimentally adjusting the ECN min-threshold and max-threshold and executing performance tests can result in improved performance, keeping in mind that the range between the two values is what provides probabilistic congestion marking of packets. Following are the scenarios resulting if this range is small or large:

- If this range is small, the probability of marking packets increases rapidly as buffer use grows.
- If this range is large, the probability of marking packets increases gradually as buffer use grows.

If users find that ECN congestion marking is causing an excessive back-off reaction and under utilization of the link, then the min-threshold and max-threshold values may be too low, as well as the range between the values may be too small. If users find that ECN is not signaling congestion and instead you are seeing only PFC pauses between link peers, then ECN min-threshold and max-threshold are likely too high.

Note the following:

- When the queue exceeds the minimum threshold limit, the ECN action marks ECN-Capable Transport (ECT) packets CE based on the probability of marking at the current queue size, where the probability increases linearly from 1–100% between the minimum and maximum thresholds.
- When the queue exceeds the maximum threshold limit, the ECN action marks all ECT packets CE.

NOTE

For ECN to work, all switches in path between two ECN-enabled endpoints must have ECN enabled.

- With HPE ANW CX 10000, 9300, 8360, and 8100 Switch Series, IP ECN is supported on the underlay only. There is no ECN support on these platforms in VXLAN overlay.
- With HPE ANW CX 8325 Switch Series, IP ECN is supported on the underlay and the VXLAN overlay – Single hop from leaf-to-leaf.

To configure a threshold profile named ECN, complete the following steps:

1. Create a non-probability-based threshold profile with ECN action on queue for:
   - HPE ANW CX 8360 and 8100 Switch Series:

```
config)# qos threshold-profile ECN
(config-threshold)# queue 1 action ecn all threshold 50 percent
```

   - HPE ANW CX 8400 Switch Series:

```
config)# qos threshold-profile ECN
(config-threshold)# queue 1 action ecn all threshold 50 kbytes
```

NOTE

No packets are marked when the queue length is below the threshold and all packets are marked when the queue length is above the threshold.

2. Create a dual threshold (probability-based) profile with ECN action on queue (HPE ANW CX 10000, 9300, and 8325 Switch Series):

```
config)# qos threshold-profile ECN
(config-threshold)# queue 5 action ecn all min-threshold 2000 kbytes max-
threshold 4000 kbytes
```

   - (Option 1) Apply profile globally (all ports):

```
config)# apply qos threshold-profile ECN
```

   - (Option 2) Apply profile on interface-level:

```
config)# int 1/1/3
(config-if)# apply qos threshold-profile ECN
```

- (Option 3) Apply profile to specific Ethernet or LAG interfaces:

```
(config)# int lag 10
(config-if)# apply qos threshold-profile ECN
```

3. Verify if the ECN threshold is applied (HPE ANW CX 8100, 8325, 8360, 9300, and 10000 Switch Series):

```
switch (config)# show qos threshold-profile ECN
Queue Action         Color   Minimum  Maximum  Max Probability
----- -------------- ------- -------- -------- ----------------
1       ecn            all     50       50
Port      Status
--------- -------
1/1/1     applied
1/1/10    applied
1/1/11    applied
1/1/12    applied
1/1/13    applied
```

The following table lists the QoS threshold parameters:

Table 6: *QoS Threshold Parameters*

Parameter	Description
<QUEUE-NUMBER>	Specifies the queue number. Range: 0 to 7.
`action ecn`	Apply ECN when the threshold is exceeded. Applies the action to all colors. Colors are reserved for future use.
`all`	Applies the action to all colors. Colors are reserved for future use. Specifies the threshold value in kilobytes. Range: 0 to 6000.
`threshold <AMOUNT> kbytes`	Applies the action to all colors. Colors are reserved for future use. Specifies the threshold value in kilobytes. Range: 0 to 6000.

For more details on ECN, see the AOS-CX QoS Configuration Guide .

Deploying RoCE v2 on Ethernet NIC Adapters

Ethernet NIC adapters are the ideal solution for network connectivity for high performance computing, secure data center connectivity and AI/ML applications as they are designed for cloud scale and enterprise environments. Ethernet NIC adapters support RoCE based on the adapter's supported Ethernet speeds. RoCE is a complete hardware offload feature supported on Ethernet NIC adapters, which allows RDMA functionality over an Ethernet network. RoCE helps to reduce CPU workload as it provides direct memory access for applications, bypassing the CPU.

The following table provides information on Mellanox Ethernet NIC Adapters that support RoCE:

Table 7: *Mellanox Ethernet NIC Adapters with RoCE Support*

Part Number	ASIC	Ports	I/O	PCIE/OCP
P13188-B21	MT27800 Family [ConnectX-5]	2x 25G	SFP28	PCIE

This chapter describes the following:

- Installing and Configuring Mellanox CX5 Ethernet NIC Adapters on Ubuntu
- Configuring PFC on Mellanox Ethernet NIC Adapters
- Configuring Congestion Control on Mellanox Ethernet NIC Adapters

Installing and Configuring Mellanox CX5 Ethernet NIC Adapters on Ubuntu

To install and configure Mellanox drivers, complete the following steps.

1. Create a directory for installing the configuration:

```
Mkdir mlnx_ofed
```

2. Download the Mellanox driver (MLNX_OFED).
3. Untar the downloaded package for Mellanox driver:

```
tar –xzfv MLNX_OFED_LINUX-5.9-0.5.6.0-ubuntu20.04-x86_64.tgz
```

4. Create an apt-get repository configuration file called "/etc/apt/source.list.d/mlnx_ofed.list" and point the location of the debian package from the untarred files.

a. Create the configuration file:

```
Sudo nano /etc/apt/source.list.d/mlnx_ofed.list
```

b. Add the configuration file to the debian package:

```
deb file:/<path to extracted MLNX_OFED package/DEBS ./
```

Example:

```
deb file:/home/someUser/mlx_ofed/MLNX_OFED_LINUX-5.9-0.5.6.0-ubuntu20.04-x86_64/DEBs
```

5. Download and install the Mellanox Technologies GPG-Key from RPM-GPG-KEY-Mellanox:

```
Wget –qO – http://www.mellanox.com/downloads/ofed/RPM-GPG-KEY-Mellanox | sudo apt-key add
sudo apt-key add RPM-GPG-KEY-Mellanox
```

6. Verify that the GPG key is successfully imported:

```
Apt-key list
pub   1024D/A9EE4B643 2013-08-11
uid MEllanox Technologies support@mellanox.com
sub 1024g/09FCC 2013-08-11
```

7. Update the apt-get cache:

```
Apt-get update
```

8. Install the OFED tool:

```
Apt-get install mlnx-ofed-all –y
```

9. Install the upstream libs for all the users:

```
Apt-get install mlx-ofed-dpdk-upstream-libs
```

10. Check if the driver is successfully installed. Refer to the instructions at Nvidia Support:

```
se20@se20iscsi2:~$ sudo mlnx_qos -i ens3f0np0
[sudo] password for se20:
DCBX mode: OS controlled
Priority trust state: pcp
default priority:
Receive buffer size (bytes): 262016,0,0,0,0,0,0,0,total_size=262016
Cable len: 7
PFC configuration:
priority    0   1   2   3   4   5   6   7
enabled     0   0   0   0   0   0   0   0
buffer      0   0   0   0   0   0   0   0
tc: 1 ratelimit: unlimited, tsa: vendor
priority:  0
tc: 0 ratelimit: unlimited, tsa: vendor
priority:  1
tc: 2 ratelimit: unlimited, tsa: vendor
priority:  2
tc: 3 ratelimit: unlimited, tsa: vendor
priority:  3
tc: 4 ratelimit: unlimited, tsa: vendor
priority:  4
tc: 5 ratelimit: unlimited, tsa: vendor
priority:  5
tc: 6 ratelimit: unlimited, tsa: vendor
priority:  6
tc: 7 ratelimit: unlimited, tsa: vendor
priority:  7
```

Configuring PFC on Mellanox Ethernet NIC Adapters

This section provides information on configuring and enabling PFC on Mellanox Ethernet NIC adapters.

Before initiating the configuration process, run the following command to view the default buffer and PFC settings:

```
se20@se20iscsi2:~$ sudo mlnx_qos -i ens3f0np0
[sudo] password for se20:
DCBX mode: OS controlled
Priority trust state: pcp
default priority:
Receive buffer size (bytes): 262016,0,0,0,0,0,0,0,total_size=262016
Cable len: 7
PFC configuration:
       priority    0   1   2   3   4   5   6   7
```

```
        enabled     0   0   0   0   0   0   0   0
        buffer      0   0   0   0   0   0   0   0
tc: 1 ratelimit: unlimited, tsa: vendor
        priority:  0
tc: 0 ratelimit: unlimited, tsa: vendor
        priority:  1
tc: 2 ratelimit: unlimited, tsa: vendor
        priority:  2
tc: 3 ratelimit: unlimited, tsa: vendor
        priority:  3
tc: 4 ratelimit: unlimited, tsa: vendor
        priority:  4
tc: 5 ratelimit: unlimited, tsa: vendor
        priority:  5
tc: 6 ratelimit: unlimited, tsa: vendor
        priority:  6
tc: 7 ratelimit: unlimited, tsa: vendor
        priority:  7
```

Configuring Mellanox Ethernet NIC Adapters for PFC

To manually configure Mellanox Ethernet NIC adapters for PFC using DSCP 24, PCP 3, and Local Priority 3, complete the following steps:

1. Change the buffer size:

```
se20@se20iscsi2:~$ sudo mlnx_qos -i ens3f0np0 --buffer_size
130944,130944,0,0,0,0,0,0
[sudo] password for se20:
DCBX mode: OS controlled
Priority trust state: pcp
default priority:
Receive buffer size (bytes): 130944,130944,0,0,0,0,0,0,total_size=262016
Cable len: 7
PFC configuration:
        priority    0   1   2   3   4   5   6   7
        enabled     0   0   0   0   0   0   0   0
        buffer      0   0   0   0   0   0   0   0
tc: 1 ratelimit: unlimited, tsa: vendor
        priority:  0
tc: 0 ratelimit: unlimited, tsa: vendor
        priority:  1
tc: 2 ratelimit: unlimited, tsa: vendor
        priority:  2
tc: 3 ratelimit: unlimited, tsa: vendor
        priority:  3
tc: 4 ratelimit: unlimited, tsa: vendor
        priority:  4
tc: 5 ratelimit: unlimited, tsa: vendor
        priority:  5
tc: 6 ratelimit: unlimited, tsa: vendor
        priority:  6
tc: 7 ratelimit: unlimited, tsa: vendor
        priority:  7
```

2. Enable PFC on Local Priority 3:

```
mlnx_qos –i eno5 –pfc 0,0,0,1,0,0,0
```

3. Set priority to buffer configuration:

This sets all priorities except priority 3 to buffer 0 and priority 3 to buffer 1.

```
mlnx qos –i eno5 –prio2buffer 0,0,0,1,0,0,0,0
```

```
se20@se20iscsi2:~$ sudo mlnx_qos -i ens3f0np0 --prio2buffer 0,0,0,1,0,0,0,0
DCBX mode: OS controlled
Priority trust state: dscp
default priority:
Receive buffer size (bytes): 130944,130944,0,0,0,0,0,0,total_size=262016
Cable len: 7
PFC configuration:
        priority    0   1   2   3   4   5   6   7
        enabled     0   0   0   1   0   0   0   0
        buffer      0   0   0   1   0   0   0   0
tc: 1 ratelimit: unlimited, tsa: vendor
        priority:  0
tc: 0 ratelimit: unlimited, tsa: vendor
        priority:  1
tc: 2 ratelimit: unlimited, tsa: vendor
        priority:  2
tc: 3 ratelimit: unlimited, tsa: vendor
        priority:  3
tc: 4 ratelimit: unlimited, tsa: vendor
        priority:  4
tc: 5 ratelimit: unlimited, tsa: vendor
        priority:  5
tc: 6 ratelimit: unlimited, tsa: vendor
        priority:  6
tc: 7 ratelimit: unlimited, tsa: vendor
        priority:  7
```

Enabling Mellanox Ethernet NIC Adapters CNP for DSCP 48 and PCP 6

To enable Mellanox Ethernet NIC adapter CNP for DSCP 48 and PCP 6, complete the following steps:

1. Find bus information:

```
user@host2:/sys/class/infiniband/mlx5_4/ports/1/hw_counters$ sudo lshw -c
network –businfo
Bus info          Device        Class          Description
========================================================
pci@0000:5d:00.0  eno5np0       network        MT27710 Family [ConnectX-5
Lx]
```

2. Using the bus information, modify the Mellanox Ethernet NIC adapter to transmit values to CNP

with CNP 6 and DSCP 48:

```
mlxconfig -d 0000:5d:00.0 -y s CNP_DSCP_P1=48 CNP_802P_PRIO_P1=6
```

Configuring Congestion Control on Mellanox Ethernet NIC Adapters

To configure Ethernet NIC adapter for DCBx, complete the following steps:

1. Install the Mellanox driver by following the steps provided in the Installing and Configuring Mellanox CX5 Ethernet NIC Adapters on Ubuntu section.
2. Find the Mellanox device (NIC) on which LLDP DCBx is to be enabled:

```
sudo mst start
Starting MST (Mellanox Software Tools) driver set
Loading MST PCI module - Success
Loading MST PCI configuration module - Success
Create devices
Unloading MST PCI module (unused) - Success
```

```
sudo mst status
MST modules:
------------
        MST PCI module is not loaded
        MST PCI configuration module loaded
MST devices:
------------
/dev/mst/mt4117_pciconf0         - PCI configuration cycles access.
                         domain:bus:dev.fn=0000:5d:00.0 addr.reg=88 data.reg=92
                         cr_bar.gw_offset=-1
                         Chip revision is: 00
/dev/mst/mt4119_pciconf0         - PCI configuration cycles access.
                         domain:bus:dev.fn=0000:d8:00.0 addr.reg=88 data.reg=92
                         cr_bar.gw_offset=-1
                         Chip revision is: 00
/dev/mst/mt4125_pciconf0         - PCI configuration cycles access.
                         domain:bus:dev.fn=0000:12:00.0 addr.reg=88 data.reg=92
                         cr_bar.gw_offset=-1
                         Chip revision is: 00
```

3. Configure the switch:

```
lldp dcbx
vlan 77
        description RoCE traffic vlan
qos queue-profile lossless
        map queue 0 local-priority 0,1,2,4,5
        map queue 1 local-priority 3
        map queue 2 local-priority 6,7
qos schedule-profile lossless
        dwrr queue 0 weight 1
```

```
        dwrr queue 1 weight 1
        strict queue 2
qos threshold-profile lossless
        queue 1 action ecn all min-threshold 120 kbytes max-threshold
        150 kbytes
apply qos queue-profile lossless schedule-profile lossless
apply qos threshold-profile lossless
qos trust dscp
qos pool 1 lossless size 50.00 percent headroom 5000 kbytes priorities 3
interface 1/1/43
        description Connects to SE20-Leaf1-Sec
        no shutdown
        mtu 9198
        flow-control priority <0-7>
        no routing
        vlan trunk native 1
        vlan trunk allowed 77
        qos trust dscp
interface 1/1/48
        no shutdown
        mtu 9198
        flow-control priority <0-7>
        no routing
        vlan trunk native 1
        vlan trunk allowed 77
        qos trust dscp
interface vlan 77
        description RoCE traffic SVI
        ip mtu 9198
        ip address 77.1.1.100/24
dcbx application tcp-udp 4791 priority 3
```

4. Configure the Ethernet NIC adapter to transmit ECN and CNP egress with DSCP 48 and PCP 6.
   a. Find bus information:

```
user@host2:/sys/class/infiniband/mlx5_4/ports/1/hw_counters$ sudo lshw -c
network -businfo
Bus info          Device        Class          Description
========================================================
pci@0000:5d:00.0  eno5np0       network        MT27710 Family [ConnectX-5
Lx]
```

   b. Using the bus information, modify the Ethernet NIC adapter to transmit values to CNP with CNP 6 and DSCP 48:

```
mlxconfig -d 0000:5d:00.0 -y s CNP_DSCP_P1=48 CNP_802P_PRIO_P1=6
```

5. Configure qos trust to DSCP on hosts:

```
Mlnx_qos –I eno5np0 –trust dscp
```

6. Enable and verify DCBx on the Ethernet NIC adapter:

```
mlxconfig -d /dev/mst/mt4119_pciconf0 set LLDP_NB_DCBX_P1=TRUE
LLDP_NB_TX_MODE_P1=2 LLDP_NB_RX_MODE_P1=2 LLDP_NB_DCBX_P2=TRUE
LLDP_NB_TX_MODE_P2=2 LLDP_NB_RX_MODE_P2=2
```

```
mlxconfig -d /dev/mst/mt4119_pciconf0 q
DCBX_IEEE_P1 True(1)
DCBX_CEE_P1 True(1)
DCBX_WILLING_P1 True(1)
DCBX_IEEE_P2 True(1)
DCBX_CEE_P2 True(1)
DCBX_WILLING_P2 True(1)
```

7. Reset the firmware and allow DCBx to be handled by the firmware:

```
mlxfwreset -d /dev/mst/mt4119_pciconf0 --level 3 reset
```

```
mlnx_qos -i ens3f0np0 -d fw
```

8. Verify PFC or ETS configuration:

```
se20@se20iscsi2:~$ sudo mlnx_qos -i ens3f0np0 -d fw
se20@se20iscsi2:~$ sudo mlnx_qos -i ens3f0np0
DCBX mode: Firmware controlled
Priority trust state: dscp
dscp2prio mapping:
        prio:0 dscp:07,06,05,04,03,02,01,00,
        prio:1 dscp:15,14,13,12,11,10,09,08,
        prio:2 dscp:23,22,21,20,19,18,17,16,
        prio:3 dscp:31,30,29,28,27,26,25,24,
        prio:4 dscp:39,38,37,36,35,34,33,32,
        prio:5 dscp:47,46,45,44,43,42,41,40,
        prio:6 dscp:55,54,53,52,51,50,49,48,
        prio:7 dscp:63,62,61,60,59,58,57,56,
default priority:
Receive buffer size (bytes): 130944,130944,0,0,0,0,0,0,total_size=262016
Cable len: 7
PFC configuration:
        priority    0   1   2   3   4   5   6   7
        enabled     0   0   0   1   0   0   0   0
        buffer      0   0   0   1   0   0   0   0
tc: 0 ratelimit: unlimited, tsa: ets, bw: 50%
        priority:  0
        priority:  1
        priority:  2
        priority:  4
        priority:  5
tc: 1 ratelimit: unlimited, tsa: ets, bw: 50%
        priority:  3
tc: 2 ratelimit: unlimited, tsa: vendor
        priority:  6
        priority:  7
```