



COMWARE SUPPORT FOR ROCE

CONTENTS

Introduction.....	2
Versions	2
Benefits.....	2
Deployment.....	2
Build a lossless Ethernet for RoCE	3
Priority-based Flow Control.....	3
PFC PAUSE generation mechanism	4
Mapping between message priority and queue	4
PFC extended functions	4
Explicit congestion notification	5
Data Center Bridging Exchange Protocol.....	5
Enhanced Transmission Selection	6
Configuration example of lossless Ethernet	6
Device configuration.....	7
Resources.....	9

INTRODUCTION

RoCE technology supports carrying RDMA protocol on Ethernet to realize RDMA over Ethernet. RoCE and IB technologies have the same application layer and transmission control layer; only the network layer and the Ethernet link layer are different.

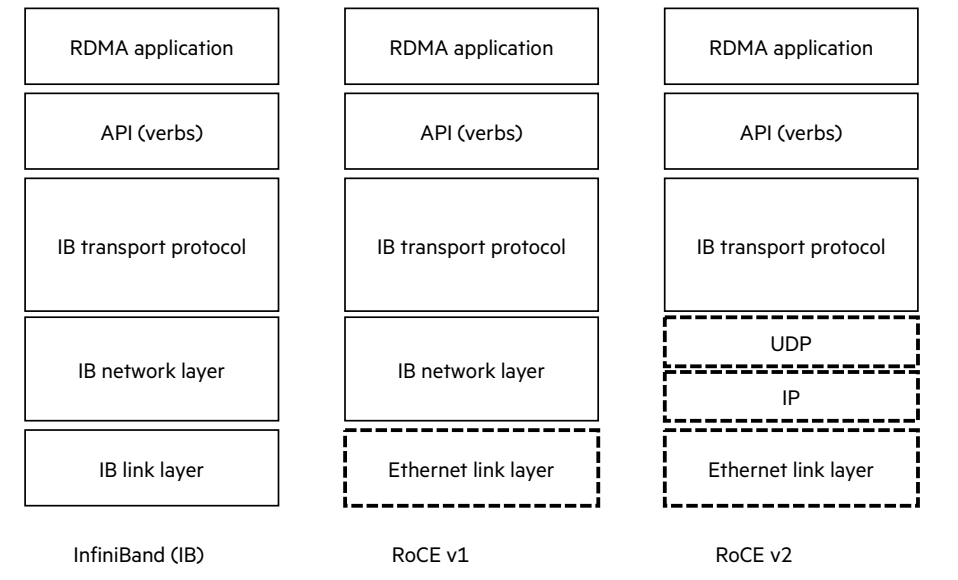


FIGURE 1. InfiniBand and RoCE architecture

VERSIONS

The RoCE protocol has two versions:

- RoCE v1 protocol is based on Ethernet bearer RDMA. It can only be deployed in a Layer 2 network. RoCE v1 message adds a Layer 2 Ethernet header to the original IB architecture message and identifies the RoCE message through EtherType 0x8915.
- RoCE v2 protocol is based on the UDP/IP protocol to carry RDMA and can be deployed in a 3-layer network. RoCE v2 message adds a UDP header, an IP header, and a Layer 2 Ethernet message header to the original IB architecture message. The RoCE message is identified by the UDP destination port number 4791. RoCE v2 supports hash based on the source port number and uses ECMP to achieve load sharing, which improves network utilization.

BENEFITS

RoCE enables Ethernet-based data transmission to:

- Improve data transmission throughput
- Reduce network latency
- Reduce CPU load

DEPLOYMENT

RoCE technology can be realized by Ethernet switches. To deploy RoCE technology, the server-side is required to support RoCE network cards, and the network side must support lossless Ethernet. This is because the loss of any packet in the IB packet loss processing mechanism will cause a large number of retransmissions, which severely reduces data transmission performance.



BUILD A LOSSLESS ETHERNET FOR ROCE

In the RoCE network, building a lossless Ethernet to ensure no packet loss during network transmission requires four key features:

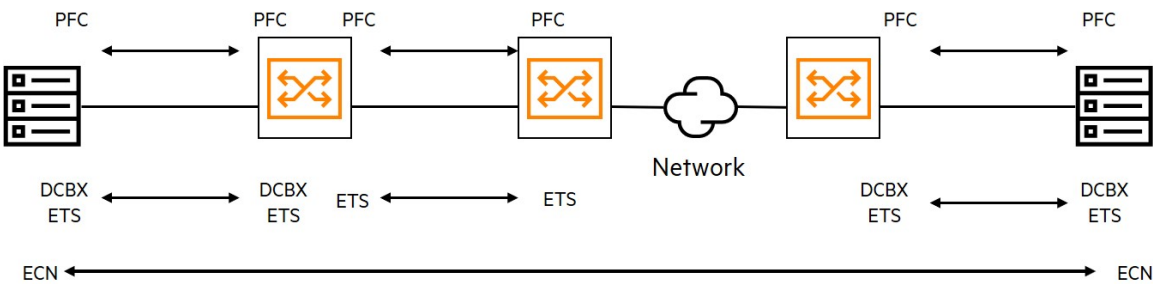


FIGURE 2. Schematic diagram of building a lossless Ethernet network with key features

PRIORITY-BASED FLOW CONTROL

Priority-based Flow Control (PFC) provides priority-based flow control hop by hop, enabling multiple types of traffic to run on the Ethernet link without affecting each other.



FIGURE 3. Schematic diagram of PFC PAUSE frame generation

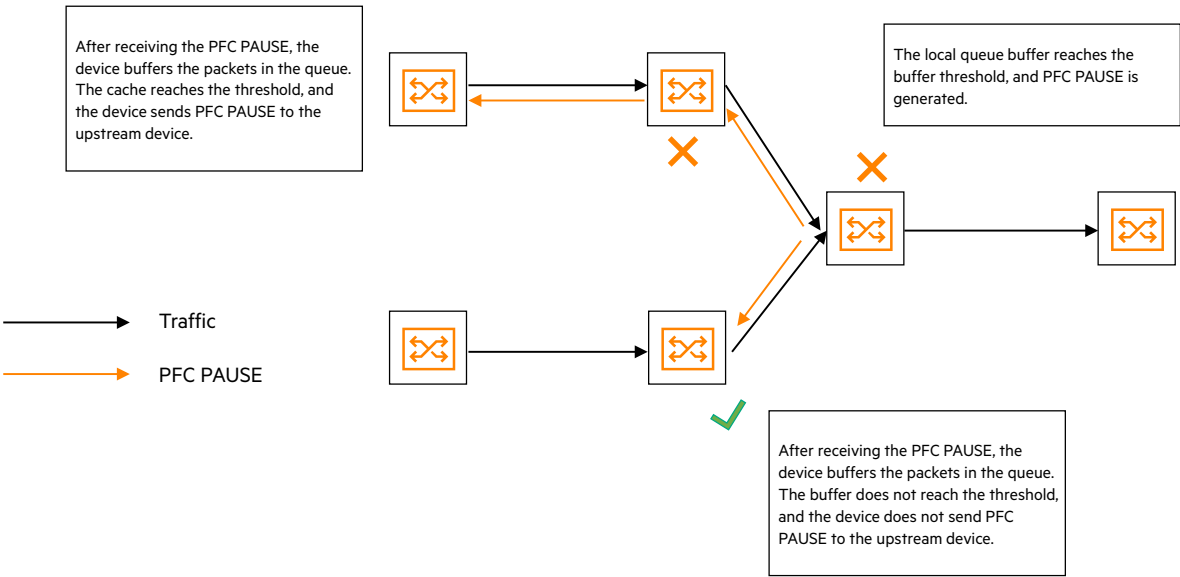


FIGURE 4. PFC PAUSE frame processing between multihop devices



PFC PAUSE GENERATION MECHANISM

- After port 1 of Device B receives a message from Device A, the memory management unit (MMU) will allocate cell resources for the message. PFC counts the cell resources occupied based on IEEE 802.1P priority.
- When the statistical count of cell resources occupied by packets of a certain priority on Port 1 of Device B reaches the set threshold and Port 1 continues to receive new packets of that priority, Port 1 will send the corresponding priority PFC PAUSE frame to Device A.
- After Device A receives the PFC PAUSE frame of this priority, it stops sending the packets of the priority and buffers the packets of this priority. If the buffer threshold is triggered, it also sends the PFC PAUSE frame to its upstream device.

MAPPING BETWEEN MESSAGE PRIORITY AND QUEUE

• Priority trust mode

- IEEE 802.1P: Trust the 802.1p priority of the packet and perform priority mapping with this priority
- dscp: Trust the DSCP priority of the IP packet and perform priority mapping based on this priority

• Port priority mode

When the priority trust mode of the port is not configured, the device will use the port priority as the priority of the packet itself.

PFC EXTENDED FUNCTIONS

PFC threshold configuration—PFC provides multiple modes of threshold setting, which can effectively solve problems such as buffer being discarded caused by insufficient buffer space and excessively large incoming traffic queues.

PFC deadlock detection—When the packets of the specified priority form a loop, the packets in the data buffer cannot be forwarded, and PFC frames are repeatedly sent and received between devices. As a result, the buffer cell resources of the device interface are always occupied and cannot be released. Currently, the device enters PFC deadlock status. By configuring the PFC deadlock detection function, the device periodically detects whether it is in the PFC deadlock state. When the device detects the PFC deadlock state, it will automatically release the deadlock state during the recovery period.

PFC one-click escape—When the PFC function of the device has an emergency failure, the user can turn off the PFC function of all interfaces with a simple command line instead of turning off the PFC function of interfaces one by one.

PFC message alarm threshold—Users can configure the alarm threshold of PFC packets in ingress or egress direction of the interface based on the networking.

Statistics and alarm reports by gRPC—PFC and gRPC can automatically report alarms for packet loss and over limit. At the same time, it provides various statistical inquiries for packet loss and dynamic usage.

Rich diagnostic and maintenance functions—The display priority-flow-control command displays the configuration of the PFC function on the port, and the total number of PFC frames sent and received, and the sending and receiving rate of each port and each queue.

The `display packet-drop` command diagnoses and queries the total packet loss information on the receiving and sending ends and the packet loss information on each port.

The `display qos queue-statistics interface outbound` command is used to display statistics in the egress port queue.



EXPLICIT CONGESTION NOTIFICATION

Explicit congestion notification (ECN) realizes end-to-end congestion management and slows down the deterioration of congestion proliferation. ECN defines a flow control and end-to-end congestion notification mechanism based on the IP layer and the transport layer. The ECN function uses the DS field in the IP message header to mark the congestion state on the message transmission path. Terminal devices that support this function can determine that congestion has occurred on the transmission path based on the content of the message, thereby adjusting the way of sending the message to avoid the aggravation of congestion.

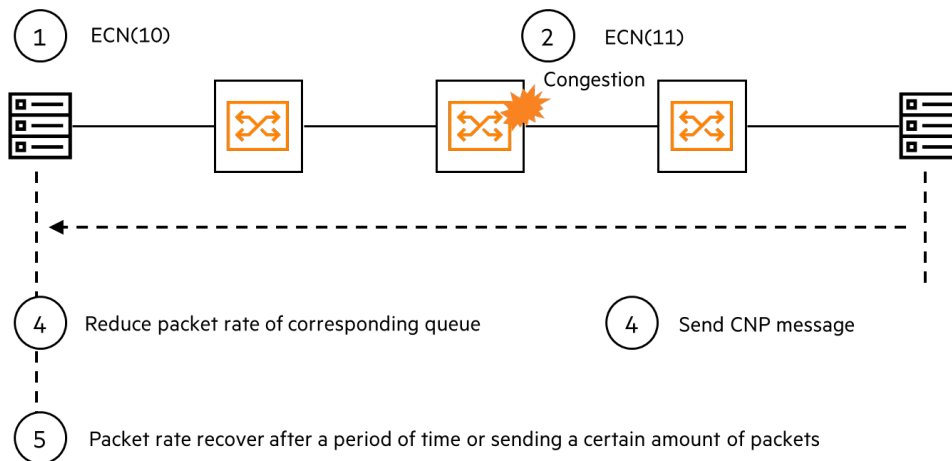


FIGURE 5. ECN mechanism

DATA CENTER BRIDGING EXCHANGE PROTOCOL

Data Center Bridging Exchange Protocol (DCBX) is used in the port of the access switch to connect to the server to negotiate the capabilities with the server network card. Through DCBX, DCB parameters can be negotiated and automatically configured between switches or between switches and server network cards to simplify configuration and ensure configuration consistency.

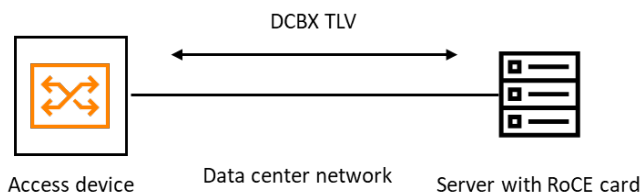


FIGURE 6. DCBX configuration

DCBX processes the information exchange through LLDP and supports configuration information such as ETS, PFC, and application priority of both sides.

Users can enable the DCBX function by LLDP function globally or on the interface and allow the interface to publish the DCBX TLV, and then configure the device to publish Application Protocol (APP), ETS, and PFC parameters through the interface according to application requirements.

ENHANCED TRANSMISSION SELECTION

Enhanced Transmission Selection (ETS) provides minimum bandwidth guarantees for different traffic while increasing link utilization to ensure the bandwidth percentage of important traffic. The ETS mechanism divides the traffic priority in the network into different priority groups and allocates a certain amount of bandwidth to each priority group. If a priority group does not consume its allocated bandwidth, other priority groups can use these unused bandwidths. Ensure that important traffic has the committed bandwidth during transmission.

Configuration example of lossless Ethernet

Device list

No.	Device	Quantity	Software version
1	5945 48SFP28 8QSFP28	2	5945-CMW710-R6616P01
2	Server (support RoCE)	3	VMware vSphere® 6.7 or above

Device networking

As shown in Figure 7, RoCE network cards are installed on Server 1, Server 2, and Server 3. Server 1 and Server 2 are connected to Server 3 through Ethernet switches Device A and Device B.

To support the RoCE technology, it is now required to build the entire network as a lossless Ethernet. The specific requirements are as follows:

All ports of the packet forwarding path enable the PFC function. In this example, the lossless transmission of packets with 802.1p priority 5.

The switch connects to the server port to enable the DCBX function so that the device and the server network card can negotiate ETS and PFC parameters.

Configure the ETS function on the Twenty-FiveGigE1/0/3 of Device A and Twenty-FiveGigE1/0/2 of Device B to ensure the transmission bandwidth of packets with 802.1p priority 5.

The Twenty-FiveGigE1/0/3 port of Device A is configured with the ECN function so that the device can notify the sender to adjust the sending rate when congestion occurs.

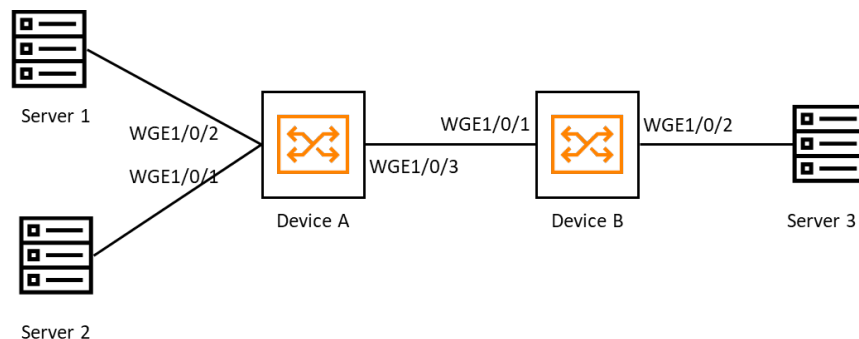


FIGURE 7. Device networking example



Device configuration

Device A

1. **Configure interfaces to trust the 802.1p priority of the packet on Twenty-FiveGigE1/0/1, Twenty-FiveGigE1/0/2, and Twenty-FiveGigE1/0/3; enable the PFC function of interfaces; and enable PFC function for 802.1p priority 5.**

```
<DeviceA> system-view
[DeviceA] interface range twenty-fivegige 1/0/1 to twenty-fivegige 1/0/3
[DeviceA-if-range] qos trust dot1p
[DeviceA-if-range] priority-flow-control enable
[DeviceA-if-range] priority-flow-control no-drop dot1p 5
[DeviceA-if-range] quit
```

2. **Enable LLDP globally.**

```
[DeviceA] lldp global enable
```

3. **Enable the LLDP function on Twenty-FiveGigE1/0/1 and Twenty-FiveGigE1/0/2 and allow the publication of DCBX TLVs. Configure the DCBX on Twenty-FiveGigE1/0/1 and Twenty-FiveGigE1/0/2 as version v1.01.**

```
[DeviceA] interface range twenty-fivegige 1/0/1 to twenty-fivegige 1/0/2
[DeviceA-if-range] lldp enable
[DeviceA-if-range] lldp tlv-enable dot1-tlv dcbx
[DeviceA-if-range] dcbx version rev101
[DeviceA-if-range] quit
```

4. **Open the WRR queue on the interface Twenty-FiveGigE1/0/3, calculate based on the number of bytes that can be sent in each poll, and configure a strict priority scheduling algorithm for port queue 5 (802.1p priority 5 to local priority 5 is the default mapping).**

```
[DeviceA] interface twenty-fivegige 1/0/3
[DeviceA-Twenty-FiveGigE1/0/3] qos wrr byte-count
[DeviceA-Twenty-FiveGigE1/0/3] qos wrr 5 group sp
[DeviceA-Twenty-FiveGigE1/0/3] quit
```

5. **Create WRED table queue-table5 and enter the WRED table view. Configure the average queue length index of queue 5 and WRED table parameters, and enable the ECN function. Apply WRED table queue-table5 on interface Twenty-FiveGigE1/0/3.**

```
[DeviceA] qos wred queue table queue-table5
[DeviceA-wred-table-queue-table5] queue 5 weighting-constant 12
[DeviceA-wred-table-queue-table5] queue 5 drop-level 0 low-limit 10 high-limit 20
discard-probability 30
[DeviceA-wred-table-queue-table5] queue 5 ecn
[DeviceA-wred-table-queue-table5] quit
[DeviceA] interface twenty-fivegige 1/0/3
[DeviceA-Twenty-FiveGigE1/0/3] qos wred apply queue-table5
```



Device B

1. **Configure interfaces to trust the 802.1p priority of packets on Twenty-FiveGigE1/0/1 and Twenty-FiveGigE1/0/2; enable the PFC function of the interface; and enable the PFC function for 802.1p priority 5.**

```
<DeviceB> system-view
[DeviceB] interface range twenty-fivegige 1/0/1 to twenty-fivegige 1/0/2
[DeviceB-if-range] qos trust dot1p
[DeviceB-if-range] priority-flow-control enable
[DeviceB-if-range] priority-flow-control no-drop dot1p 5
[DeviceB-if-range] quit
```

2. **Enable LLDP globally.**

```
[DeviceB] lldp global enable
```

3. **Enable the LLDP function on Twenty-FiveGigE1/0/2 and allow the publication of DCBX TLVs. Configure the DCBX on Twenty-FiveGigE1/0/2 as version v1.01.**

```
[DeviceB] interface twenty-fivegige 1/0/2
[DeviceB-Twenty-FiveGigE1/0/2] lldp enable
[DeviceB-Twenty-FiveGigE1/0/2] lldp tlv-enable dot1-tlv dcbx
[DeviceB-Twenty-FiveGigE1/0/2] dcbx version rev101
[DeviceB-Twenty-FiveGigE1/0/2] quit
```

4. **Open the WRR queue on the interface Twenty-FiveGigE1/0/2, calculate based on the number of bytes that can be sent in each poll, and configure a strict priority scheduling algorithm for port queue 5 (802.1p priority 5 to local priority 5 is the default mapping).**

```
[DeviceB] interface twenty-fivegige 1/0/2
[DeviceB-Twenty-FiveGigE1/0/2] qos wrr byte-count
[DeviceB-Twenty-FiveGigE1/0/2] qos wrr 5 group sp
```

Configuration verification

1. **Display the number of discarded packets on Device B.**

```
<DeviceB> display packet-drop summary
```

All interfaces:

```
Packets dropped due to Fast Filter Processor (FFP): 0
Packets dropped due to STP non-forwarding state: 0
Packets dropped due to insufficient data buffer. Input dropped: 0 Output dropped: 0
Packets of ECN marked: 1622267130
Packets of WRED dropped: 0
```

2. **Display the bandwidth utilization of port Twenty-FiveGigE1/0/2 on Device B.**

```
<DeviceB> display counters rate outbound interface Twenty-FiveGigE 1/0/2
```

Usage: Bandwidth utilization in percentage

Interface	Usage (%)	Total [pps]	Broadcast [pps]	Multicast [pps]
WGE1/0/2	100	2825427	--	--

Additional information

QoS setting could be mapped to the VXLAN EVPN underlay in the case of DCI. When IP QoS setting is leveraged for use cases such as RoCE v2, the payload DSCP setting can be mapped to the underlay network. Underlay network must honor such setting from overlay in support of the QoS requirements.



By default, on inbound ports with XConnect configured, QoS markings are trusted. The following configuration sample can remark the incoming traffic when needed.

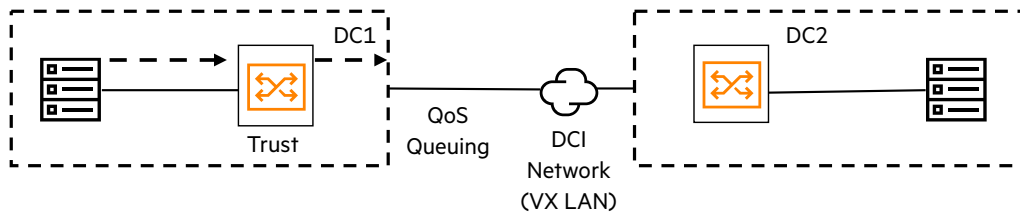


FIGURE 8. DCI networking

```
acl advanced 3000
  rule 0 permit icmp
#
traffic classifier IP operator and
  if-match acl 3000
#
traffic behavior REMARK_EF
  remark dscp ef
#
qos policy REMARK
  classifier IP behavior REMARK_EF
#
interface ten-gigabitethernet 1/1/15
  qos apply policy REMARK inbound
```

RESOURCES

[HPE FlexFabric 5945 Switch Series Fundamentals Configuration Guide](#)

[HPE FlexFabric 5945 Switch Series documentation](#)

LEARN MORE AT

hpe.com/us/en/networking/comware.html

Make the right purchase decision.
Contact our presales specialists.



Chat now (sales)



Call now



Get updates