



RoCE with Priority Flow Control Application Guide

Rev. 1.1

www.mellanox.com

NOTE:

THIS HARDWARE, SOFTWARE OR TEST SUITE PRODUCT (“PRODUCT(S)”) AND ITS RELATED DOCUMENTATION ARE PROVIDED BY MELLANOX TECHNOLOGIES “AS-IS” WITH ALL FAULTS OF ANY KIND AND SOLELY FOR THE PURPOSE OF AIDING THE CUSTOMER IN TESTING APPLICATIONS THAT USE THE PRODUCTS IN DESIGNATED SOLUTIONS. THE CUSTOMER’S MANUFACTURING TEST ENVIRONMENT HAS NOT MET THE STANDARDS SET BY MELLANOX TECHNOLOGIES TO FULLY QUALIFY THE PRODUCT(S) AND/OR THE SYSTEM USING IT. THEREFORE, MELLANOX TECHNOLOGIES CANNOT AND DOES NOT GUARANTEE OR WARRANT THAT THE PRODUCTS WILL OPERATE WITH THE HIGHEST QUALITY. ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NON INFRINGEMENT ARE DISCLAIMED. IN NO EVENT SHALL MELLANOX BE LIABLE TO CUSTOMER OR ANY THIRD PARTIES FOR ANY DIRECT, INDIRECT, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES OF ANY KIND (INCLUDING, BUT NOT LIMITED TO, PAYMENT FOR PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY FROM THE USE OF THE PRODUCT(S) AND RELATED DOCUMENTATION EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.



Mellanox Technologies, Inc.
350 Oakmead Parkway Suite 100
Sunnyvale, CA 94085
U.S.A.
www.mellanox.com
Tel: (408) 970-3400
Fax: (408) 970-3403

Mellanox Technologies Ltd
PO Box 586 Hermon Building
Yokneam 20692
Israel
Tel: +972-4-909-7200
Fax: +972-4-959-3245

© Copyright 2011. Mellanox Technologies. All rights reserved.

Mellanox®, BridgeX®, ConnectX®, CORE-Direct®, InfiniBridge®, InfiniHost®, InfiniScale®, PhyX®, Virtual Protocol Interconnect and Voltaire are registered trademarks of Mellanox Technologies, Ltd.

FabricIT, MLNX-OS and SwitchX are trademarks of Mellanox Technologies, Ltd.

All other trademarks are property of their respective owners.

Contents

1 Overview	4
1.1 Software Dependencies	4
1.2 Firmware Dependencies.....	4
1.3 General Guidelines	4
1.4 Ported Applications.....	5
1.5 Using VLANs	5
1.6 Priority Pause Frames	5
1.6.1 Using Priority Flow Control	6
1.6.2 Using Global Pause	6
1.7 Sanity Tests	7
1.8 Reading Port Counters Statistics.....	7
2 Various Switch Configuration.....	9
2.1 Vantage 6048	9
2.1.1 Vantage 6048 PFC Configuration.....	9
2.1.2 Vantage 6048 Global Pause Configuration	9
2.2 Vantage 6024	10
2.2.1 Vantage 6024 PFC Configuration.....	10
2.2.2 Vantage 6024 Global Pause Configuration	10
2.3 Cisco Nexus 5020.....	11
2.3.1 Cisco Nexus 5020 PFC Configuration.....	11
2.3.2 Cisco Nexus 5020 Global Pause Configuration	12

1 Overview

RDMA over Converged Ethernet (RoCE) enables InfiniBand (IB) transport over Ethernet networks. It encapsulates IB transport and GRH headers in Ethernet packets using an IEEE assigned ethtype.

Classic Ethernet is a best-effort protocol in the event of congestion. Ethernet discards packets and relies on higher level protocols to provide retransmission and other reliability mechanisms. IEEE 802.3x pause allows a congested receiver to signal the other side of the link to pause the data transmission for a short period of time. The pause functionality is applied to all the traffic on the link.

The recently introduced priority flow control (PFC) 802.1 Qbb feature applies pause functionality to specific classes of traffic on the Ethernet link. For example, PFC can provide lossless service for the RoCE traffic and best-effort service for the standard Ethernet traffic. PFC can provide different levels of service to specific classes of Ethernet traffic (using IEEE 802.1p traffic classes).

This document focuses on the configuration of RoCE with lossless transport layer.

1.1 Software Dependencies

To use RoCE over Mellanox ConnectX[®] hardware, the `mlx4_en` driver must be loaded. For further information, please refer to [MLNX_EN Linux README](#) file.

1.2 Firmware Dependencies

To use RoCE over Mellanox ConnectX[®] hardware, RoCE requires ConnectX[®] firmware version 2.7.700 or higher. Features such as loopback require higher firmware versions.

1.3 General Guidelines

Since RoCE encapsulates InfiniBand traffic in Ethernet frames, the corresponding net device must be up and running. In case of Mellanox hardware, `MLNX_OFED` (version 1.5.2 and above) must be installed and `mlx4_en` must be loaded and the corresponding interface configured.

1. Make sure that `mlx4_en.ko` is loaded. Run:

```
lsmod | grep mlx4_en
```

If the module is loaded, the `mlx4_en` should be displayed as shown in the example below.

```
# lsmod | grep mlx4_en
*mlx4_en 75276 0
```

2. Run `ibv_devinfo`.

There is a new field, `link_layer` which can be either "Ethernet" or "IB". If the value is IB, then you need to use `connectx_port_config` to change the ConnectX/ConnectX-2 ports designation to eth. For further information, see `mlx4_release_notes`.

- Configure the IP address of the interface so that the link will become active. For further information, see section [Using VLANs](#) (on page 5)

All IB verbs applications which run over IB verbs should work on RoCE links if they use GRH headers (if the use of GRH is specified in the address vector)

1.4 Ported Applications

The following applications are ported with RoCE:

- `ibv_*_pingpong` examples

You must specify the GID of the remote peer using the new '-g' option. The GID has the same format as that in

```
/sys/class/infiniband/mlx4_0/ports/1/gids/0
```



Use `ibv_ud_pingpong` cautiously. The default message size is 2K, which is likely to exceed the MTU of the RoCE link. Use `ibv_devinfo` to inspect the link MTU and specify an appropriate message size.

- All `rdma_cm` applications
- `libsdp`
- Performance tests

1.5 Using VLANs

For RoCE traffic to use VLAN tagged frames, you need to specify the GID table entries that are derived from the VLAN devices when creating address vectors.

To do so, perform the following:

- Verify VLAN support is enabled by the kernel. Usually this requires loading the `802.1q` module.

```
modprobe 8021q
```

- Add a VLAN device. Run:

```
vconfig add [interface name] [vlan id]
```

- Assign an IP address to the VLAN interface. This should create a new entry in the GID table (as index 1).

```
ifconfig [interface name].[vlan id] [ip] [netmask id]
```

- For `rdma_cm` applications, specify only the IP address of a VLAN device in order for the traffic to go with the VLAN tagged frames.



PFC cannot be used when using an interface without VLANs.

1.6 Priority Pause Frames

Tagged Ethernet frames carry a 3-bit priority field. The value of this field is derived from the InfiniBand Service Level (SL) field by taking the 3 least significant bits of the SL field.

1.6.1 Using Priority Flow Control

Priority based Flow Control policy on TX and RX [7:0]. `mlx4_en` parameters:

- `pfctx` - Priority based Flow Control policy on TX[7:0]. Per priority bit mask (default is 0)
- `pfcrx` - Priority based Flow Control policy on RX[7:0]. Per priority bit mask (default is 0)

When using MLNX_OFED 1.5.2.

➤ ***To configure the `mlx4_en` Ethernet driver to support PFC:***

Add the following line to the file `/etc/modprobe.conf`, and restart the network driver.

```
options mlx4_en pfctx=0xff pfcrx=0xff
```

When using MLNX_OFED 1.5.3 and above.

➤ ***To configure the `mlx4_en` Ethernet driver to support PFC:***

Change the values of `pfctx` and `pfcrx` in the line below in the file `/etc/modprobe.conf`, and restart the network driver.

```
options mlx4_core pfctx=0xff pfcrx=0xff
```



The values of `pfctx` and `pfcrx` should be set according to the priority you need the flow control to have.



If the file `/etc/modprobe.conf` does not exist, please create an `mlx4_en.conf` under `/etc/modprobe.d/` and place the options line in there.

1.6.2 Using Global Pause

When using MLNX_OFED 1.5.2.

Verify that both `/sys/module/mlx4_en/parameters/pfctx` and `pfcrx` are set to 0 to enable global pauses (as opposed to PFC). The `mlx4_en` Ethernet driver supports link pause by default (priority bit mask is 0).

When using MLNX_OFED 1.5.3 and above.

Verify that both `/sys/module/mlx4_core/parameters/pfctx` and `pfcrx` are set to 0 to enable global pauses (as opposed to PFC).

➤ ***Global pause support can be turned ON and OFF using the following command:***

```
#> ethtool -A eth<x> [rx on|off] [tx on|off]
```

1.7 Sanity Tests

- *To verify RoCE is configured correctly, check the entries in a port's GID table:*

The first entry always contains the IPv6 link's local address of the corresponding Ethernet interface. The link's local address is formed in the following way:

```
gid[0..7] = fe80000000000000
gid[8] = mac[0] ^ 2
gid[9] = mac[1]
gid[10] = mac[2]
gid[11] = ff
gid[12] = fe
gid[13] = mac[3]
gid[14] = mac[4]
gid[15] = mac[5]
```

If VLAN is supported by the kernel and there are VLAN interfaces on the main Ethernet interface (the interface that the IB port is bind with), then each such VLAN will appear as a new GID in the port's GID table. The format of the GID entry will be identical to the one described above, except for the following change:

```
gid[11] = VLAN ID high byte (4 MS bits).
gid[12] = VLAN ID low byte
```

where the VLAN ID is 12 bits wide.

To simulate a packet drop scenario and verify the pause prevent drops, the following are required:

- at least 4 machines connected to the switch where one will act as server and the rest as clients.
- for `ibv_rc_pingpong` test, the amount of iteration must be increased to 100000 and packet size to 1000000.
- Run Verbs test:

- On server:

```
> ibv_rc_pingpong -g 1
```

- On client1:

```
> ibv_rc_pingpongs -g 1 server
```

- On client2:

```
> ibv_rc_pingpongs -g 1 server
```

- On client3:

```
> ibv_rc_pingpongs -g 1 server
```

1.8 Reading Port Counters Statistics

It is possible to read port statistics in the same manner as regular InfiniBand ports. The information is available from the sysfs at `/sys/class/infiniband/<device>/ports/<port number>/counters`, and the supported counters are `port_rcv_packets`, `port_xmit_packets`, `port_rcv_data` and `port_xmit_data`. These counters count only InfiniBand data and are not account for Ethernet traffic.

➤ ***To read the number of transmitted packets, run:***

```
> cat /sys/class/infiniband/<device>/ports/<port  
number>/counters/port_xmit_packets
```



RoCE traffic is not shown in the associated Ethernet device's counters since it is offloaded by the hardware and does not go through Ethernet network driver.

2 Various Switch Configuration

2.1 Vantage 6048

2.1.1 Vantage 6048 PFC Configuration

Table 1: Vantage 6048 PFC Configuration

	Command	Purpose
Step 1	<code>console(config)# {no} dce priority-flow-control enable</code>	Globally enables the Priority Flow Control feature on the switch
Step 2	<code>console(config)# {no} dce priority-flow-control priority {priority number} enable</code>	Enables priority flow control for a priority. The priority range is 0-7.
Step 3	<code>console(config)# interface range tengigabitethernet {type slot/port type slot/port range}</code>	Specifies the interface to configure, and enters the interface configuration mode.
Step 4	<code>console(config-if-range)# {no} dce priority-flow-control enable</code>	Enable priority flow control for a priority
Step 5	<code>console(config-if-range)# no dce dcbx advertise application-protocol</code>	
Step 6	<code>console(config-if-range)# no dce dcbx advertise priority-groups</code>	
Step 7	<code>console(config-if-range)# no dce dcbx advertise priority-flow-control</code>	
Step 8	<code>console(config-if-range)# exit</code>	Exits interface configuration mode

PFC full configuration example:

```
console# configure
console(config)# dce priority-flow-control enable
console(config)# dce priority-flow-control priority 3 enable
console(config)# interface range tengigabitethernet 0/1-2
console(config-if-range)# dce priority-flow-control enable
console(config-if-range)# no dce dcbx advertise application-protocol
console(config-if-range)# no dce dcbx advertise priority-groups
console(config-if-range)# no dce dcbx advertise priority-flow-control
```

2.1.2 Vantage 6048 Global Pause Configuration

Table 2: Vantage 6048 Global Pause Configuration

	Command	Purpose
Step 1	<code>console(config)# interface tengigabitethernet {type slot/port}</code>	Specifies the interface to configure, and enters the interface configuration mode.
Step 2	<code>console(config)# flowcontrol {on off}</code>	<ul style="list-style-type: none"> on – Force flow-control enable off – Force flow-control disable

This example will tag all traffic as lossless:

```
console# configure
console(config)# interface tengigabitethernet 0/1
console(config-if-range)# flowcontrol on
```



You must disable PFC to configure Global Pause.

2.2 Vantage 6024

2.2.1 Vantage 6024 PFC Configuration

Table 3: Vantage 6024 PFC Configuration

	Command	Purpose
Step 1	Console (config)# cee enable	Enables cee globally on the switch
Step 2	Console (config)# cee global pfc priority {priority} enable	Enables PFC on 802.1P priority [0-7]
Step 3	Console (config)# cee global pfc priority {priority} description "optional description"	Gives a description to the pfc priority [0-7]
Step 4	Console (config)# interface port { port }	Enters the interface configuration mode
Step 5	Console (config-if)# dot1p {priority}	Enters 802.1p value [0-7] on the interface

PFC full configuration example:

```
admin@voltaire#configure terminal
admin@voltaire(config)# cee enable
admin@voltaire(config)# cee global pfc priority 3 enable
admin@voltaire(config)# cee global pfc priority 3 description "Rocee Traffic"
admin@voltaire(config)# interface port 1-4
admin@voltaire(config-if)# dot1p 3
```

2.2.2 Vantage 6024 Global Pause Configuration

Table 4: Vantage 6024 Global Pause Configuration

	Command	Purpose
Step 1	console(config)# interface port {port#}	Specifies the interface to configure, and enters the interface configuration mode.
Step 2	console(config)# flowcontrol {both receive send}	<ul style="list-style-type: none"> • both – Both side flow control • receive – Receive flow control • send – send flow control

This example will tag all traffic as lossless:

```
admin@voltaire#configure terminal
admin@voltaire(config)# interface port 1
admin@voltaire(config-if)# flowcontrol both
```



You must disable PFC to configure Global Pause.



Turning CEE ON will automatically change some 802.1p QoS and 802.3x standard flow control settings on the Vantage 6024.

It is recommended that you backup your configuration prior to turning CEE on. Viewing the file will allow you to manually re-create the equivalent configuration once CEE is turned on, and will also allow you to recover your prior configuration if you need to turn CEE off.

Please refer to the Vantage 6024 user manual for all CEE effects on the switch operation.

2.3 Cisco Nexus 5020

2.3.1 Cisco Nexus 5020 PFC Configuration

Table 5: Cisco Nexus 5020 PCF Configuration

	Command	Purpose
Step 1	<code>switch# configure terminal</code>	Enters configuration mode.
Step 2	<code>switch(config)# vlan {vlan-id vlan-range}</code>	Enters VLAN configuration submode. If the VLAN does not exist, the system first creates the specified VLAN.
Step 3	<code>switch(config)# vlan {vlan-id vlan-range}</code>	Names the VLAN. You can enter up to 32 alphanumeric characters to name the VLAN. You cannot change the name of VLAN1 or the internally allocated VLANs. The default value is VLANxxxx where xxxx represent four numeric digits (including leading zeroes) equal to the VLAN ID number.
Step 4	<code>switch(config)# interface {type slot/port port-channel number}</code>	Specifies the interface to configure, and enters the interface configuration mode. The interface can be a physical Ethernet port or a port channel.
Step 5	<code>switch(config-if)# switchport mode trunk</code>	Configures the interface as a trunk port.
Step 6	<code>switch(config-if)# switchport trunk {allowed vlan vlan-id native vlan vlan-id}</code>	(Optional) Configures necessary parameters for a trunk port.
Step 7	<code>switch(config-if)# priority-flow-control mode {auto on }</code>	Sets PFC mode for the selected interface. Specify auto to negotiate PFC capability. Specify on to force-enable PFC.

	Command	Purpose
Step 8	<code>switch(config-if)# flowcontrol [receive {on off}] [transmit {on off}]</code>	(Optional) Enables IEEE 802.3x link-level flow control for the selected interface. Set receive and/or transmit on or off.
Step 9	<code>switch(config-vlan)# no shutdown</code>	Enables the VLAN. The default value is no shutdown (or enabled). You cannot shut down the default VLAN, VLAN1, or VLANs 1006 to 4094.

The example below shows how to configure two ports:

```
switch# configure terminal
switch(config)# vlan 50
switch(config-vlan)# name roce
switch(config)# interface ethernet 1/3
switch(config-if)# switchport mode trunk
switch(config-if)# switchport trunk allowed vlan 50
switch(config-if)# priority-flow-control mode on
switch(config-if)# flowcontrol receive on transmit on
switch(config-vlan)# state active
switch(config-vlan)# no shutdown
switch(config)# interface ethernet 1/11
switch(config-if)# switchport mode trunk
switch(config-if)# switchport trunk allowed vlan 50
switch(config-if)# priority-flow-control mode on
switch(config-if)# flowcontrol receive on transmit on
switch(config-vlan)# state active
switch(config-vlan)# no shutdown
```

2.3.2 Cisco Nexus 5020 Global Pause Configuration

Table 6: Cisco Nexus 5020 Global Pause Configuration

	Command	Purpose
Step 1	<code>switch# configure terminal</code>	Enters configuration mode.
Step 2	<code>switch(config)# mac access-list name</code>	Creates the MAC ACL and enters ACL configuration mode.
Step 3	<code>switch(config-mac-acl)# [sequence-number] {permit deny} source destination protocol</code>	Creates a rule in the MAC ACL. The permit and deny options support many ways of identifying traffic.
Step 4	<code>switch(config)# class-map type qos class-name</code>	Creates a named object that represents a class of traffic. Class-map names can contain alphabetic, hyphen, or underscore characters, are case sensitive, and can be up to 40 characters.
Step 5	<code>switch(config-cmap-qos)# match access-group name acl-name</code>	Configures a traffic class by matching packets based on the <i>acl-name</i> . The permit and deny ACL keywords are ignored in the matching.

	Command	Purpose
Step 6	<code>switch(config-cmap-qos)# policy-map type qos policy-name</code>	Creates a named object that represents a set of policies that are to be applied to a set of traffic classes. Policy-map names can contain alphabetic, hyphen, or underscore characters, are case sensitive, and can be up to 40 characters.
Step 7	<code>switch(config-cmap-qos)# class class-name</code>	
Step 8	<code>switch(config-pmap-c-qos)# set qos-group qos-group-value</code>	Configures one or more qos-group values to match on for classification of traffic into this class map. The range of qos-group-value is from 2 to 5. There is no default value.
Step 9	<code>switch(config)# class [type {network-qos}] class-name</code>	Associates a class map with the policy map, and enters configuration mode for the specified system class. The three class-map configuration modes are as follows: <ul style="list-style-type: none"> • network-qos—Network-wide (global) mode.
Step 10	<code>switch(config-cmap-nq)# match qos-group qos-group-value</code>	Configures the traffic class by matching packets based on a list of QoS group values. Values can range from 0 to 5. QoS group 0 is equivalent to class-default and QoS group 1 is equivalent to class-foe. <p>Note: qos-groups 0 and 1 are reserved for default classes and cannot be configured.</p>
Step 11	<code>switch(config-cmap-nq)# policy-map [type {network-qos}] policy-name</code>	See: Step 6
Step 12	<code>switch(config-cmap-nq)# class type network-qos class-name</code>	Associates a class map with the policy map, and enters configuration mode for the specified system class.
Step 13	<code>switch (config-pmap-nq-c)# pause no-drop [pfc-cos pfc-cos-value]</code>	Configures a no-drop class. If you do not specify this command, the default policy is drop. The pfc-cos-value range is from 0 to 7. This option is supported only for for a ACL-based system class. <p>Note: The operation for the drop policy is a simple tail drop, where arriving packets will be dropped if the queue increases to its allocated size.</p>

	Command	Purpose
Step 14	switch (config)# system qos	Enters system class configuration mode.
Step 15	switch (config-sys-qos)# service-policy type qos input policy-name	Specifies the policy map to use as the service policy for the system. There are three policy-map configuration modes: <ul style="list-style-type: none"> • network-qos—Network-wide (system qos) mode • qos—Classification mode (system qos input or interface input only) • queuing—Queuing mode (input and output at system qos and interface)
Step 17	switch (config-sys-qos)# service-policy type network-qos policy-name	
Step 18	switch (config)# interface type slot/port	Specifies the interface to be changed.
Step 19	flowcontrol [receive {on off}] [transmit {on off}]	Enables LLC for the selected interface. Set receive and/or transmit on or off .

This example tags all traffic as lossless:

```
switch# configure terminal
switch(config)# mac access-list test
switch(config-mac-acl)# 10 permit any any
!
switch(config)# class-map type qos test1
switch(config-cmap-qos)# match access-group name test
switch(config-cmap-qos)# policy-map type qos test1
switch(config-cmap-qos)# class test1
switch(config-pmap-c-qos)# set qos-group 4
!
switch(config)# class-map type network-qos test1
switch(config-cmap-nq)# match qos-group 4
switch(config-cmap-nq)# policy-map type network-qos test1
switch(config-cmap-nq)# class type network-qos test1
switch(config-pmap-nq-c)# pause no-drop
!
switch(config)# system qos
switch(config-sys-qos)# service-policy type qos input test1
switch(config-sys-qos)# service-policy type network-qos test1
!
switch(config)# interface ethernet 1/2
switch(config-if)# flowcontrol receive on transmit on
```