



DPU

Table of contents

QoS Configuration	3
Shared RQ Mode	8

This section contains the following pages:

- [QoS Configuration](#)
- [Shared RQ Mode](#)

QoS Configuration

Note

To learn more about port QoS configuration, refer to [this](#) community post.

Warning

When working in Embedded Host mode, using `mlx_qos` on both the host and Arm will result with undefined behavior. Users must only use `mlx_qos` from the Arm. After changing the QoS settings from Arm, users must restart the mlx5 driver on host.

Note

When configuring QoS using DCBX, the `lldpad` service from the NVIDIA® BlueField® networking platform's (DPU or SuperNIC) side must be disabled if the configurations are not done using tools other than `lldpad`.

This section explains how to configure QoS group and settings using devlink located under `/opt/mellanox/iproute2/sbin/`. It is applicable to host PF/VF and Arm side SFs. The following uses VF as example.

The settings of a QoS group include creating/deleting a QoS group and modifying its `tx_max` and `tx_share` values. The settings of VF QoS include modifying its `tx_max`

and `tx_share` values, assigning a VF to a QoS group, and unassigning a VF from a QoS group. This section focuses on the configuration syntax.

Please refer to section "Limit and Bandwidth Share Per VF" in the MLNX_OFED User Manual for detailed explanation on vPort QoS behaviors.

devlink port function rate add

	<code>devlink port function rate add <DEV>/<GROUP_NAME></code> Adds a QoS group.	
Syntax Description	DEV/GROUP_NAME	Specifies group name in string format
Example	<p>This command adds a new QoS group named <code>12_group</code> under device <code>pci/0000:03:00.0</code>:</p> <pre>devlink port function rate add pci/0000:03:00.0/12_group</pre>	
Notes		

devlink port function rate del

	<code>devlink port function rate del <DEV>/<GROUP_NAME></code> Deletes a QoS group.	
Syntax Description	DEV/GROUP_NAME	Specifies group name in string format
Example	<p>This command deletes QoS group <code>12_group</code> from device <code>pci/0000:03:00.0</code>:</p> <pre>devlink port function rate del pci/0000:03:00.0/12_group</pre>	
Notes		

devlink port function rate set tx_max tx_share

	<pre>devlink port function rate set {<DEV>/<GROUP_NAME> <DEV>/<PORT_INDEX>} tx_max <TX_MAX> [tx_share <TX_SHARE>]</pre> <p>Sets <code>tx_max</code> and <code>tx_share</code> for QoS group or devlink port.</p>	
Syntax Description	DEV/GROUP_NAME	Specifies the group name to operate on
	DEV/PORT_INDEX	Specifies the devlink port to operate on
	TX_MAX	<code>tx_max</code> bandwidth in MB/s
	TX_SHARE	<code>tx_share</code> bandwidth in MB/s
Example	<p>This command sets <code>tx_max</code> to 2000MB/s and <code>tx_share</code> to 500MB/s for the <code>12_group</code> QoS group:</p> <pre>devlink port function rate set pci/0000:03:00.0/12_group tx_max 2000MBps tx_share 500MBps</pre>	
	<p>This command sets <code>tx_max</code> to 2000MB/s and <code>tx_share</code> to 500MB/s for the VF represented by port index 196609:</p> <pre>devlink port function rate set pci/0000:03:00.0/196609 tx_max 200MBps tx_share 50MBps</pre>	
	<p>This command displays a mapping between VF devlink ports and netdev names:</p> <pre>\$ devlink port</pre> <p>In the output of this command, VFs are indicated by <code>flavour pcivf</code>.</p>	
Notes		

devlink port function rate set parent

	<pre>devlink port function rate set <DEV>/<PORT_INDEX> {parent <PARENT_GROUP_NAME>}</pre> <p>Assigns devlink port to a QoS group.</p>	
--	---	--

Syntax Description	DEV/PORT_INDEX	Specifies the devlink port to operate on
	PARENT_GROUP_NAME	parent group name in string format
Example	<p>This command assigns this function to the QoS group <code>12_group</code>:</p> <pre>devlink port function rate set pci/0000:03:00.0/196609 parent 12_group</pre>	
Notes		

devlink port function rate set noparent

	<pre>devlink port function rate set <DEV>/<PORT_INDEX> noparent</pre> <p>Ungroups a devlink port.</p>	
Syntax Description	DEV/PORT_INDEX	Specifies the devlink port to operate on
Example	<p>This command ungroups this function:</p> <pre>devlink port function rate set pci/0000:03:00.0/196609 noparent</pre>	
Notes		

devlink port function rate show

	<pre>devlink port function rate show [<DEV>/<GROUP_NAME> <DEV>/<PORT_INDEX>]</pre> <p>Displays QoS information QoS group or devlink port.</p>	
Syntax Description	DEV/GROUP_NAME	Specifies the group name to display
	DEV/PORT_INDEX	Specifies the devlink port to display
Example	<p>This command displays the QoS info of all QoS groups and devlink ports on the system:</p>	

```
devlink port function rate show
pci/0000:03:00.0/12_group type node tx_max 2000MBps
tx_share 500MBps
pci/0000:03:00.0/196609 type leaf tx_max 200MBps
tx_share 50MBps parent 12_group
```

This command displays QoS info of `12_group`:

```
devlink port function rate show
pci/0000:03:00.0/12_group
pci/0000:03:00.0/12_group type node tx_max 2000MBps
tx_share 500MBps
```

Notes

If a QoS group name or devlink port are not specified, all QoS groups and devlink ports are displayed.

Shared RQ Mode

When creating 1 send queue (SQ) and 1 receive queue (RQ), each representor consumes ~3MB memory per single channel. Scaling this to the desired 1024 representors (SFs and/or VFs) would require ~3GB worth of memory for single channel. A major chunk of the 3MB is contributed by RQ allocation (receive buffers and SKBs). Therefore, to make efficient use of memory, shared RQ mode is implemented so PF/VF/SF representors share receive queues owned by the uplink representor.

The feature is enabled by default. To disable it:

1. Edit the field `ALLOW_SHARED_RQ` in `/etc/mellanox/mlnx-bf.conf` as follows:

```
ALLOW_SHARED_RQ="no"
```

2. Restart the driver. Run:

```
/etc/init.d/openibd restart
```

To connect from the host to NVIDIA® BlueField® networking platform (DPU or SuperNIC) in shared RQ mode, please refer to section [Verifying Connection from Host to BlueField](#).

Note

PF/VF representor to PF/VF communication on the host is not possible.

The following behavior is observed in shared RQ mode:

- It is expected to see a 0 in the rx_bytes and rx_packets and valid vport_rx_packets and vport_rx_bytes after running traffic. Example output:

```
# ethtool -S pf0hpf
NIC statistics:
  rx_packets: 0
  rx_bytes: 0
  tx_packets: 66946
  tx_bytes: 8786869
  vport_rx_packets: 546093
  vport_rx_bytes: 321100036
  vport_tx_packets: 549449
  vport_tx_bytes: 321679548
```

- Ethtool usage – in this mode, it is not possible to change/set the ring or coalesce parameters for the RX side using ethtool. Changing channels also only affects the TX side.

Notice
This document is provided for information purposes only and shall not be regarded as a warranty of a certain functionality, condition, or quality of a product. NVIDIA Corporation (“NVIDIA”) makes no representations or warranties, expressed or implied, as to the accuracy or completeness of the information contained in this document and assumes no responsibility for any errors contained herein. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This document is not a commitment to develop, release, or deliver any Material (defined below), code, or functionality. NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and any other changes to this document, at any time without notice. Customer should obtain the latest relevant information before placing orders and should verify that such information is current and complete. NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer (“Terms of Sale”). NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this document. No contractual obligations are formed either directly or indirectly by this document. NVIDIA products are not designed, authorized, or warranted to be suitable for use in medical, military, aircraft, space, or life support equipment, nor in applications where failure or malfunction of the NVIDIA product can reasonably be expected to result in personal injury, death, or property or environmental damage. NVIDIA accepts no liability for inclusion and/or use of NVIDIA products in such equipment or applications and therefore such inclusion and/or use is at customer’s own risk. NVIDIA makes no representation or warranty that products based on this document will be suitable for any specified use. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer’s sole responsibility to evaluate and determine the applicability of any information contained in this document, ensure the product is suitable and fit for the application planned by customer, and perform the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer’s product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this document. NVIDIA accepts no liability related to any default, damage, costs, or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this document or (ii) customer product designs. No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this document. Information published by NVIDIA regarding third-party

products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA.

Reproduction of information in this document is permissible only if approved in advance by NVIDIA in writing, reproduced without alteration and in full compliance with all applicable export laws and regulations, and accompanied by all associated conditions, limitations, and notices.

THIS DOCUMENT AND ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, "MATERIALS") ARE BEING PROVIDED "AS IS." NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL NVIDIA BE LIABLE FOR ANY DAMAGES, INCLUDING WITHOUT LIMITATION ANY DIRECT, INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES, HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF ANY USE OF THIS DOCUMENT, EVEN IF NVIDIA HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA's aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the Terms of Sale for the product.

Trademarks

NVIDIA and the NVIDIA logo are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

Copyright 2025. PDF Generated on 01/02/2025