



# Mellanox DPDK

## Quick Start Guide

---

Rev 16.11\_3.0

## 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 NONINFRINGEMENT 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  
350 Oakmead Parkway Suite 100  
Sunnyvale, CA 94085  
U.S.A.  
[www.mellanox.com](http://www.mellanox.com)  
Tel: (408) 970-3400  
Fax: (408) 970-3403

© Copyright 2017. Mellanox Technologies Ltd. All Rights Reserved.

Mellanox®, Mellanox logo, Accelio®, BridgeX®, CloudX logo, CompustorX®, Connect-IB®, ConnectX®, CoolBox®, CORE-Direct®, EZchip®, EZchip logo, EZappliance®, EZdesign®, EZdriver®, EZsystem®, GPUDirect®, InfiniHost®, InfiniBridge®, InfiniScale®, Kotura®, Kotura logo, Mellanox CloudRack®, Mellanox CloudXMellanox®, Mellanox Federal Systems®, Mellanox HostDirect®, Mellanox Multi-Host®, Mellanox Open Ethernet®, Mellanox OpenCloud®, Mellanox OpenCloud Logo®, Mellanox PeerDirect®, Mellanox ScalableHPC®, Mellanox StorageX®, Mellanox TuneX®, Mellanox Connect Accelerate Outperform logo, Mellanox Virtual Modular Switch®, MetroDX®, MetroX®, MLNX-OS®, NP-1c®, NP-2®, NP-3®, Open Ethernet logo, PhyX®, PlatformX®, PSIPHY®, SiPhy®, StoreX®, SwitchX®, Tiler®, Tiler logo, TestX®, TuneX®, The Generation of Open Ethernet logo, UFM®, Unbreakable Link®, Virtual Protocol Interconnect®, Voltaire® and Voltaire logo are registered trademarks of Mellanox Technologies, Ltd.

All other trademarks are property of their respective owners.

For the most updated list of Mellanox trademarks, visit <http://www.mellanox.com/page/trademarks>

# Table of Contents

- Document Revision History..... 5**
- 1 Overview ..... 8**
- 2 MLNX\_DPDK Installation and Configuration ..... 9**
  - 2.1 Installing ConnectX-3 Pro ..... 9
    - 2.1.1 MLNX\_OFED/Firmware Versions..... 9
    - 2.1.2 ConnectX-3 Pro Bare Metal..... 9
    - 2.1.3 ConnectX-3 Pro KVM ..... 10
  - 2.2 Installing ConnectX-4/ConnectX-4 Lx /ConnectX-5 ..... 12
    - 2.2.1 MLNX\_OFED/Firmware Versions..... 12
    - 2.2.2 ConnectX-4/ConnectX-4 Lx /ConnectX-5 Bare Metal ..... 12
    - 2.2.3 ConnectX-4/ConnectX-4 Lx / ConnectX-5 KVM..... 13
    - 2.2.4 ConnectX-4/ConnectX-4 Lx ESX6.0 UP1..... 14
  - 2.3 Configuring PMD Debug Mode..... 16
  - 2.4 mlx5 PMD Command Line Parameters ..... 17
  - 2.5 Checksum and TSO offload for VXLAN Packets without the Inner Ethernet Header on ConnectX-4/ConnectX-4 Lx ..... 19
    - 2.5.1 System Requirements ..... 19
    - 2.5.2 Command Line Requirements ..... 19
  - 2.6 Sending and Receiving Jumbo Frames on ConnectX-3 Pro..... 19
  - 2.7 Setting RX VLAN Filter on ConnectX-3 Pro ..... 19
- 3 System Performance Configuration..... 20**
  - 3.1 General Settings ..... 20
    - 3.1.1 ConnectX-3 Pro Setting..... 22
    - 3.1.2 ConnectX-4 / ConnectX-4 Lx /ConnectX-5 Settings..... 23
  - 3.2 KVM Settings ..... 23
- 4 Running DPDK Application with Mellanox Poll-Mode Driver ..... 25**
  - 4.1 ConnectX®-3 Pro ..... 25
  - 4.2 ConnectX®-4 / ConnectX®-4 Lx / ConnectX®-5..... 25
- 5 Sanity Check..... 27**

## List of Tables

|  |    |
|--|----|
| Table 1: Document Revision History .....           | 5  |
| Table 2: Reference Documents .....                 | 7  |
| Table 3: MLNX_OFED/Firmware Versions.....          | 9  |
| Table 4: MLNX_OFED/Firmware Versions.....          | 10 |
| Table 5: MLNX_OFED/Firmware Versions.....          | 12 |
| Table 6: MLNX_OFED/MLNX_EN/Firmware Versions ..... | 13 |
| Table 7: Firmware Versions .....                   | 14 |

## Document Revision History

**Table 1: Document Revision History**

| Revision  | Date                            | Description   |
|-----------|---------------------------------|---|
| 16.11_3.0 | June 30 <sup>th</sup> , 2017    | <ul style="list-style-type: none"> <li>Updated MLX_OFED and Firmware versions to the latest across the document</li> <li>Updated section <a href="#">mlx5 PMD Command Line Parameters</a>, added the following parameters:               <ul style="list-style-type: none"> <li>delay_drop_rxq</li> <li>swp_en</li> </ul> </li> <li>Updated sections:               <ul style="list-style-type: none"> <li><a href="#">System Requirements</a></li> <li><a href="#">Command Line Requirements</a></li> </ul> </li> </ul>  |
| 16.11_2.3 | April 06 <sup>th</sup> , 2017   | <ul style="list-style-type: none"> <li>Updated the supported MLNX-OFED/ESXi/Firmware versions across the document.</li> <li>Updated section <a href="#">mlx5 PMD Command Line Parameters</a>; added the follow parameters:               <ul style="list-style-type: none"> <li>txq_mpw_hdr_dseg_en</li> <li>txq_max_inline_len</li> <li>tso</li> <li>udp_port</li> </ul> </li> </ul>   |
| 16.11_1.5 | January 31 <sup>st</sup> , 2017 | <ul style="list-style-type: none"> <li>Updated the supported MLNX-OFED/ESXi/Firmware versions across the document.</li> <li>Updated the following sections:               <ul style="list-style-type: none"> <li><a href="#">Configuring PMD Debug Mode</a></li> <li><a href="#">ConnectX®-4 / ConnectX®-4 Lx / ConnectX®-5</a></li> </ul> </li> </ul>  |
| 2.2_4.2   | October 09 <sup>th</sup> , 2016 | <ul style="list-style-type: none"> <li>Added MLNX_EN support and updated the following sections:               <ul style="list-style-type: none"> <li><a href="#">MLNX_OFED/Firmware Versions</a></li> <li><a href="#">ConnectX-4/ConnectX-4 Lx /ConnectX-5 Bare Metal</a></li> <li><a href="#">MLNX_OFED/Firmware Versions</a></li> <li><a href="#">ConnectX-4/ConnectX-4 Lx ESXi: VM Virtual Machine</a></li> </ul> </li> <li>Added the following sections:               <ul style="list-style-type: none"> <li><a href="#">ConnectX-3 Pro Setting</a></li> <li><a href="#">ConnectX-4 / ConnectX-4 Lx /ConnectX-5 Settings</a></li> </ul> </li> <li>Updated MLNX_OFED/ESXi and ConnectX-4/Lx versions across the document</li> <li>Updated the following sections:               <ul style="list-style-type: none"> <li><a href="#">Configuring PMD Debug Mode</a></li> </ul> </li> </ul> |

| Revision | Date                              | Description  |
|----------|-----------------------------------|--|
|          |                                   | <ul style="list-style-type: none"> <li>• <a href="#">mlx5 PMD Command Line Parameters</a></li> </ul>   |
| 2.2_3.9  | September 09 <sup>th</sup> , 2016 | <ul style="list-style-type: none"> <li>• Updated MLNX_OFED version across the document</li> <li>• Updated section <a href="#">mlx5 PMD Command Line Parameters</a>; added parameter txq_inline_auto_en</li> </ul>  |
| 2.2_2.7  | June 09 <sup>th</sup> , 2016      | <ul style="list-style-type: none"> <li>• Added the following sections:               <ul style="list-style-type: none"> <li>• <a href="#">Error! Not a valid result for table.</a></li> <li>• <a href="#">ConnectX-4/ConnectX-4 Lx ESXi:</a></li> <li>• <a href="#">ConnectX-4/ConnectX-4 Lx ESXi: VM Virtual Machine</a></li> </ul> </li> <li>• Updated the following sections:               <ul style="list-style-type: none"> <li>• <a href="#">ConnectX-3 Pro Bare Metal</a></li> <li>• <a href="#">ConnectX-3 Pro KVM Hypervisor</a></li> <li>• <a href="#">ConnectX-3 Pro KVM Virtual Machine</a></li> </ul> </li> <li>• Removed the following sections:               <ul style="list-style-type: none"> <li>• ConnectX-3/ConnectX-3 Pro ESX: VMWare</li> <li>• ConnectX-3/ConnectX-3 Pro ESX: VM Virtual Machine</li> </ul> </li> </ul> |
| 2.2_2.3  | May 16 <sup>th</sup> , 2016       | <ul style="list-style-type: none"> <li>• Updated the following sections:               <ul style="list-style-type: none"> <li>• <a href="#">ConnectX-4/ConnectX-4 Lx /ConnectX-5 Bare Metal</a></li> <li>• <a href="#">ConnectX-3 Pro KVM HypervisorConnectX-3 Pro KVM Virtual Machine</a></li> <li>• <a href="#">ConnectX-3/ConnectX-3 Pro ESX: VM Virtual Machine</a></li> <li>• <a href="#">ConnectX-4/ConnectX-4 Lx /ConnectX-5 Bare Metal</a></li> <li>• <a href="#">General Settings</a></li> <li>• <a href="#">ConnectX®-4 / ConnectX®-4 Lx</a></li> </ul> </li> <li>• Removed the following section:               <ul style="list-style-type: none"> <li>• PowerKVM Settings</li> </ul> </li> </ul>   |

## Related Documents

The following table lists the documents referenced in this User Manual.

**Table 2: Reference Documents**

| Document Name                     | Description   |
|-----------------------------------|---|
| Mellanox OFED Linux Release Notes | Describes the new features and changes of the latest MLNX_OFED release.   |
| Mellanox OFED Linux User Manual   | Provides general information on the installation, configuration, management and maintenance of the software and hardware of VPI (InfiniBand, Ethernet) adapter cards. |
| Mellanox DPDK Release Notes       | Describes the new features and changes of the latest Mellanox DPDK  |

# 1 Overview

Mellanox Poll Mode Driver (PMD) is designed for fast packet processing and low latency by providing kernel bypass for receive, send, and by avoiding the interrupt processing performance overhead.

This Quick Start Guide provides information of how to install and configure mlx4 and mlx5 DPDK Poll-Mode Driver (PMD) for Mellanox ConnectX®-3 Pro and ConnectX®-4/ConnectX®-4 Lx /ConnectX®-5/ConnectX®-5 Ex Ethernet adapters.



## 2 MLNX\_DPDK Installation and Configuration

### 2.1 Installing ConnectX-3 Pro

#### 2.1.1 MLNX\_OFED/Firmware Versions

**Table 3: MLNX\_OFED/Firmware Versions**

| Driver/Firmware         | Version             | Download Link                          |
|-------------------------|---------------------|--|
| MLNX_OFED               | 4.1-1.x.x.x         | <a href="#">MLNX_OFED website</a>      |
| ConnectX-3 Pro Firmware | 2.36.5000 and above | <a href="#">Firmware download page</a> |

#### 2.1.2 ConnectX-3 Pro Bare Metal

1. Install MLNX\_OFED. (See [Table 3: MLNX\\_OFED/Firmware Versions](#) for the Driver/Firmware versions)
2. Verify that ConnectX-3 Pro firmware is **2.36.5000 or above** (Use `ibstat` command).

Set all the ports to Ethernet, and follow the instructions on the screen.

```
connectx_port_config
```

For further instructions on how to run the script, please refer to the MLNX\_OFED User Manual.

3. Add the following line to `/etc/modprobe.d/mlx4_core.conf`.

```
options mlx4_core log_num_mgm_entry_size=-7
```



**NOTE:** If VLAN filtering is used, set `log_num_mgm_entry_size=-1`.



**NOTE:** Please be aware, performance penalty can occur in this case.

4. Restart MLNX\_OFED.

```
/etc/init.d/openib restart
```

5. Extract the package `MLNX_DPDK_16.11_3.0.tar.gz`

The default `mlx4` configuration in `config/common_linuxapp` is the following:

```
#
# Compile burst-oriented Mellanox ConnectX-3 (MLX4) PMD
#
CONFIG_RTE_LIBRTE_MLX4_PMD=y
CONFIG_RTE_LIBRTE_MLX4_DEBUG=n
CONFIG_RTE_LIBRTE_MLX4_SGE_WR_N=1
CONFIG_RTE_LIBRTE_MLX4_MAX_INLINE=0
CONFIG_RTE_LIBRTE_MLX4_TX_MP_CACHE=8
CONFIG_RTE_LIBRTE_MLX4_SOFT_COUNTERS=1
```

6. Compile DPDK.

```
On x86_64:
make install T=x86_64-native-linuxapp-gcc
```

```
on Power8:
make install T=ppc_64-power8-linuxapp-gcc
```

For more advanced DPDK compilation options, please refer to [dpdk.org](http://dpdk.org) documentation:

[http://www.dpdk.org/doc/guides/linux\\_gsg/index.html](http://www.dpdk.org/doc/guides/linux_gsg/index.html)

## 7. Configure huge pages according to the NUMA the card is connected to

```
echo $PAGE_NUM >
/sys/devices/system/node/node$NUMA_NODE/hugepages/hugepages-
"$PAGE_SIZE"kB/nr_hugepages

mkdir -p /mnt/huge

HTLB_MOUNTED=$( mount | grep "type hugetlbfs" | wc -l)
if [ $HTLB_MOUNTED -eq 0 ]; then
    mount -t hugetlbfs hugetlb /mnt/huge
fi
```

### 2.1.3 ConnectX-3 Pro KVM

#### 2.1.3.1 MLNX\_OFED/Firmware Versions

**Table 4: MLNX\_OFED/Firmware Versions**

| Driver/Firmware         | Version             | Download Link                          |
|-------------------------|---------------------|--|
| MLNX_OFED               | 4.1-1.x.x.x         | <a href="#">MLNX_OFED website</a>      |
| ConnectX-3 Firmware     | 2.36.5000 and above | <a href="#">Firmware download page</a> |
| ConnectX-3 Pro Firmware | 2.36.5000 and above | <a href="#">Firmware download page</a> |

#### 2.1.3.2 ConnectX-3 Pro KVM Hypervisor

1. Download MLNX\_OFED. (See [Table 4: MLNX\\_OFED/Firmware Versions](#) for the Driver/Firmware versions)
2. Install MLNX\_OFED enable SR-IOV.
3. Please follow MLNX\_OFED User Manual instructions how to enable SR-IOV.
4. Verify that ConnectX-3 Pro firmware is **2.36.5000 or above** (Use `ibstat` command).
5. Add the following line to `/etc/modprobe.d/mlx4_core.conf`.

```
options mlx4_core log_num_mgm_entry_size=-7
```



**NOTE:** If VLAN filtering is used, set `log_num_mgm_entry_size=-1`.

Please be aware, performance penalty can occur in this case.

6. Restart MLNX\_OFED.

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

#### 2.1.3.3 ConnectX-3 Pro KVM Virtual Machine

1. Download MLNX\_OFED. (See [Table 4: MLNX\\_OFED/Firmware Versions](#) for the Driver/Firmware versions)
2. Install MLNX\_OFED and enable the guest.

```
./mlnxofedinstall --guest
```

### 3. Extract the package MLNX\_DPDK\_16.11\_3.0.tar.gz

The default mlx4 configuration in `config/common_linuxapp` is the following:

```
#
# Compile burst-oriented Mellanox ConnectX-3 (MLX4) PMD
#
CONFIG_RTE_LIBRTE_MLX4_PMD=y
CONFIG_RTE_LIBRTE_MLX4_DEBUG=n
CONFIG_RTE_LIBRTE_MLX4_SGE_WR_N=1
CONFIG_RTE_LIBRTE_MLX4_MAX_INLINE=0
CONFIG_RTE_LIBRTE_MLX4_TX_MP_CACHE=8
CONFIG_RTE_LIBRTE_MLX4_SOFT_COUNTERS=1
```

### 4. Compile DPDK.

```
On x86_64:
make install T=x86_64-native-linuxapp-gcc
on Power8:
make install T=ppc_64-power8-linuxapp-gcc
```

For more advanced DPDK compilation options, please refer to [dpdk.org](http://dpdk.org) documentation:

[http://www.dpdk.org/doc/guides/linux\\_gsg/index.html](http://www.dpdk.org/doc/guides/linux_gsg/index.html)

### 5. Configure huge pages.

```
echo $PAGE_NUM > /sys/devices/system/node/node0/hugepages/hugepages-
"$PAGE_SIZE"kB/nr_hugepages

mkdir -p /mnt/huge

HTLB_MOUNTED=$( mount | grep "type hugetlbfs" | wc -l)
if [ $HTLB_MOUNTED -eq 0 ]; then
    mount -t hugetlbfs hugetlb /mnt/huge
fi
```

## 2.2 Installing ConnectX-4/ConnectX-4 Lx /ConnectX-5

### 2.2.1 MLNX\_OFED/Firmware Versions

**Table 5: MLNX\_OFED/Firmware Versions**

| Driver/Firmware        | Version     | Download Link                          |
|------------------------|-------------|--|
| MLNX_OFED              | 4.1-1.x.x.x | <a href="#">MLNX_OFED website</a>      |
| ConnectX-4 Firmware    | 12.20.1000  | <a href="#">Firmware download page</a> |
| ConnectX-4 Lx Firmware | 14.20.1000  | <a href="#">Firmware download page</a> |
| ConnectX-5 Firmware    | 16.20.1000  | <a href="#">Firmware download page</a> |

### 2.2.2 ConnectX-4/ConnectX-4 Lx /ConnectX-5 Bare Metal

➤ *To clean installation or upgrade from any previous MLNX\_DPDK GA version:*

1. Download MLNX\_OFED. (See [Table 5: MLNX\\_OFED/Firmware Versions](#) for the Drivers/Firmware versions)

2. Install the MLNX\_OFED driver.

```
./mlnxofedinstall
```

3. Verify that ConnectX-4 firmware is 12.20.1000, ConnectX-4 Lx firmware is 14.18.2000 and ConnectX-5 firmware is 16.20.1000

```
ibv_devinfo
```

4. Set all the ports to Ethernet if MLNX\_OFED has been installed

```
mst start
mlxconfig -d <device> set LINK_TYPE_P1/2=1/2/3
* LINK_TYPE_P1=<1|2|3>, 1=Infiniband 2=Ethernet 3=VPI(auto-sense)
```

For example:

```
mlxconfig -d <device> set LINK_TYPE_P1=2
mlxfwreset -d <device> reset
```

For further instructions on how to run the script, please refer to the MLNX\_OFED User Manual.

5. Restart the driver:

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

6. Extract the package MLNX\_DPDK\_16.11\_3.0.tar.gz

The default mlx5 configuration in `config/common_base` is the following:

```
# Compile burst-oriented Mellanox ConnectX-4 (MLX5) PMD
#
CONFIG_RTE_LIBRTE_MLX5_PMD=y
CONFIG_RTE_LIBRTE_MLX5_DEBUG=n
CONFIG_RTE_LIBRTE_MLX5_TX_MP_CACHE=8
```

7. Compile DPDK.

```
On x86_64:
make install T=x86_64-native-linuxapp-gcc
on Power8:
make install T=ppc_64-power8-linuxapp-gcc
```

For more advanced DPDK compilation options, please refer to [dpdk.org](http://dpdk.org) documentation:

<http://www.dpdk.org/doc/guides-16.11/>

- Configure huge pages according to the NUMA the card is connected to.

```
echo $PAGE_NUM >
/sys/devices/system/node/node$NUMA_NODE/hugepages/hugepages-
"$PAGE_SIZE"kB/nr_hugepages

mkdir -p /mnt/huge

HTLB_MOUNTED=$( mount | grep "type hugetlbfs" | wc -l)
if [ $HTLB_MOUNTED -eq 0 ]; then
    mount -t hugetlbfs hugetlb /mnt/huge
fi
```

## 2.2.3 ConnectX-4/ConnectX-4 Lx / ConnectX-5 KVM

### 2.2.3.1 MLNX\_OFED/Firmware Versions

**Table 6: MLNX\_OFED/MLNX\_EN/Firmware Versions**

| Driver/Firmware        | Version     | Download Link                          |
|------------------------|-------------|--|
| MLNX_OFED              | 4.1-1.x.x.x | <a href="#">MLNX_OFED website</a>      |
| ConnectX-4 Firmware    | 12.20.1000  | <a href="#">Firmware download page</a> |
| ConnectX-4 Lx Firmware | 14.20.1000  | <a href="#">Firmware download page</a> |
| ConnectX-5 Firmware    | 16.20.1000  | <a href="#">Firmware download page</a> |

### 2.2.3.2 ConnectX-4/ConnectX-4 Lx /ConnectX-5 KVM Hypervisor

- Download MLNX\_OFED. (See [Table 5: MLNX\\_OFED/Firmware Versions](#) for the Drivers/Firmware versions)

- Install the MLNX\_OFED driver.

```
./mlnxofedinstall
```

- Restart the driver.

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

- Verify that ConnectX-4 firmware is 12.20.1000, ConnectX-4 Lx firmware is 14.18.2000 and ConnectX-5 firmware is 16.20.1000

```
ibv_devinfo
```

- Check if SR-IOV is enabled in the firmware.

```
mlxconfig -d <device> q
Device #1: -----
Device type: Connect4
PCI device: /dev/mst/mt4115_pciconf0
Configurations: Current
SRIOV_EN 1
NUM_OF_VFS 8
```

- If needed, use mlxconfig to set the relevant fields:

```
mlxconfig -d <device> set SRIOV_EN=1 NUM_OF_VFS=16
mlxfwreset -d <device> reset
```

- Set all the ports to Ethernet.

```
mst start
mlxconfig -d <device> set LINK_TYPE_P1/2=1/2/3
* LINK_TYPE_P1=<1|2|3> , 1=Infiniband 2=Ethernet 3=VPI (auto-sense)
For example:
mlxconfig -d <device> set LINK_TYPE_P1=2
```

```
mlxfwreset -d <device> reset
```

For further instructions on how to run the script, please refer to the MLNX\_OFED User Manual.

- Write to the sysfs file the number of Virtual Functions you need to create for the PF.

The following is an example of a standard Linux kernel generated file that is available in the new kernels.

```
echo [num_vfs] > /sys/class/infiniband/mlx5_0/device/sriov_numvfs
```

For further information, please refer to the MLNX\_OFED User Manual section “Configuring SR-IOV for ConnectX-4/Connect-IB”.

### 2.2.3.3 ConnectX-4/ConnectX-4 Lx/ConnectX-5 KVM Virtual Machine

- Download MLNX\_OFED. (See [Table 6: MLNX\\_OFED/MLNX\\_EN/Firmware Versions](#) for the Drivers/Firmware versions)
- Install the MLNX\_OFED driver.

```
./mlnxofedinstall
```

- Extract the package MLNX\_DPK-16.11\_3.0.tar.gz

The default mlx5 configuration in config/common\_base is the following:

```
#
# Compile burst-oriented Mellanox ConnectX-4 (MLX5) PMD
#
CONFIG_RTE_LIBRTE_MLX5_PMD=y
CONFIG_RTE_LIBRTE_MLX5_DEBUG=n
CONFIG_RTE_LIBRTE_MLX5_TX_MP_CACHE=8
```

- Compile DPDK.

```
On x86_64:
make install T=x86_64-native-linuxapp-gcc
on Power8:
make install T=ppc_64-power8-linuxapp-gcc
```

For more advanced DPDK compilation options, please refer to [dpdk.org](#) documentation:

<http://www.dpdk.org/doc/guides-16.11/>

- Configure huge pages.

```
echo $PAGE_NUM > /sys/devices/system/node/node0/hugepages/hugepages-
"$PAGE_SIZE"kB/nr_hugepages

mkdir -p /mnt/huge

HTLB_MOUNTED=$( mount | grep "type hugetlbfs" | wc -l)
if [ $HTLB_MOUNTED -eq 0 ]; then
    mount -t hugetlbfs hugetlb /mnt/huge
fi
```

## 2.2.4 ConnectX-4/ConnectX-4 Lx ESX6.0 UP1

### 2.2.4.1 Firmware Versions

**Table 7: Firmware Versions**

| Driver/Firmware     | Version    | Download Link                          |
|---------------------|------------|--|
| ConnectX-4 Firmware | 12.20.1000 | <a href="#">Firmware download page</a> |

| Driver/Firmware        | Version    | Download Link                          |
|------------------------|------------|--|
| ConnectX-4 Lx Firmware | 14.20.1010 | <a href="#">Firmware download page</a> |

#### 2.2.4.2 ConnectX-4/ConnectX-4 Lx ESXi: VMware

1. Download MLNX-NATIVE-ESX-ConnectX-4.

[http://www.mellanox.com/page/products\\_dyn?product\\_family=29&mtag=vmware\\_driver](http://www.mellanox.com/page/products_dyn?product_family=29&mtag=vmware_driver)

2. Install MLNX-NATIVE-ESX-ConnectX-4 v4.15.10.2 for ESX 6.0.

```
esxcli software vib install -d /FULL_PATH/<VERSION>.zip
```

3. Update the firmware on ESXi.

4. Verify that ConnectX-4 firmware is 12.20.1000 and ConnectX-4 Lx firmware is 14.20.1000.

##### ConnectX-4:

```
/opt/mellanox/bin/mlxburn -d <device> -fw_dir /FULL_PATH/fw-4115-rel-12_20_1000
```

##### ConnectX-4 Lx:

```
/opt/mellanox/bin/mlxburn -d mt4115_pciconf0 -fw_dir /FULL_PATH/fw-4117-rel-14_20_1000
```

5. Verify that NIC firmware is burned as expected:

```
/opt/mellanox/bin/flint -d <device> q
```

6. Reboot the ESXi machine.

7. Enable SR-IOV on the VMware.

```
esxcli system module parameters set -m nmlx5_core -p max_vfs=2
esxcfg-module -g nmlx5_core
```

8. Restart the driver.

```
/opt/mellanox/bin/openibd restart
```

#### 2.2.4.3 ConnectX-4/ConnectX-4 Lx ESXi: VM Virtual Machine

1. Download MLNX\_OFED/. (See [Table 6: MLNX\\_OFED/MLNX\\_EN/Firmware Versions for the Driver/Firmware versions](#))

2. Install the MLNX\_OFED driver.

```
./mlnxofedinstall
```

3. Extract the package MLNX\_DPDK-16.11\_3.0.tar.gz

The default mlx5 configuration in config/common\_base is the following:

```
#
# Compile burst-oriented Mellanox ConnectX-4 (MLX5) PMD
#
CONFIG_RTE_LIBRTE_MLX5_PMD=y
CONFIG_RTE_LIBRTE_MLX5_DEBUG=n
CONFIG_RTE_LIBRTE_MLX5_TX_MP_CACHE=8
```

4. Compile DPDK.

```
make install T=x86_64-native-linuxapp-gcc
```

For more advanced DPDK compilation options, please refer to [dpdk.org](http://dpdk.org) documentation:

<http://www.dpdk.org/doc/guides-16.11/>

5. Configure huge pages.

```
echo $PAGE_NUM > /sys/devices/system/node/node0/hugepages/hugepages-
"$PAGE_SIZE"kB/nr_hugepages

mkdir -p /mnt/huge

HTLB_MOUNTED=$( mount | grep "type hugetlbfs" | wc -l)
if [ $HTLB_MOUNTED -eq 0 ]; then
    mount -t hugetlbfs hugetlb /mnt/huge
fi
```

## 2.3 Configuring PMD Debug Mode

➤ *To enable Debug mode:*

1. Modify the `config/common_base` file.

- `CONFIG_RTE_LOG_LEVEL=RTE_LOG_DEBUG`
- For `mlx4` PMD: `CONFIG_RTE_LIBRTE_MLX4_DEBUG=y`
- For `mlx5` PMD: `CONFIG_RTE_LIBRTE_MLX5_DEBUG=y`

2. Compile DPDK.

```
On x86_64:
rm -rf x86_64-native-linuxapp-gcc
make install T=x86_64-native-linuxapp-gcc
On Power8:
rm -rf ppc_64-power8-linuxapp-gcc
make install T=ppc_64-power8-linuxapp-gcc
```

3. Add `--log-level=9` to EAL command line parameters when you run your application.



## 2.4 mlx5 PMD Command Line Parameters

| Parameter          | Description   |
|--------------------|---|
| txq_inline_auto_en | <p><b>[Beta]</b> This mode is disabled by default and supported only on ConnectX-4 single port card. It should not be used to measure 0 packet loss performance.</p> <p>When enabled, the running algorithm for auto inline is:</p> <ul style="list-style-type: none"> <li>The PPS is measured with and without inline, and the decision is taken according to the highest PPC.</li> <li>After the decision was taken, every 100 ms of the packet rate is tested and if there was 10% change in the packet rate, the measurements with and without inline are done again.</li> </ul> <p>It is expected that at the beginning of the test the performance will not be at maximum, but after 1 sec, the performance should be stabilized.</p> <p>When enabled, txq_inline option is ignored.</p>  |
| txq_inline         | <p>Amount of data to be inlined during TX operations. This improves latency. It can improve PPS performance when PCI back pressure is detected and may be useful for scenarios involving heavy traffic on many queues.</p> <p>On ConnectX-4: It is recommended to set it to 128 for messages <math>\leq 128B</math> and to set it to 64 for messages <math>&gt; 128B</math></p> <p>On ConnectX-4 Lx: When is used with txq_mpw_en option (see below), in case of PCI back pressure, it is recommended to set to 200</p> <p>On ConnectX-5: It is set to 896 by default if Enhanced MPW is enabled. If not, it is disabled by default (set to 0) since the additional software logic necessary to handle this mode can lower performance when back pressure is not expected. (see below explanation regarding Enhanced MPW)</p>   |
| txqs_min_inline    | <p>Enable inline Send only when the number of TX queues is higher or equal to this value. This option should be used in combination with txq_inline above.</p> <p>It is set to 8 by default only if Enhanced MPW is enabled. If not, it is set to zero by default.</p>  |
| txq_mpw_en         | <p>When set to 1 MPW/Enhanced MPW is enabled<br/>When set to 0 MPW is disabled</p> <p>This option is to enable multi-packet send. This feature allows the TX burst function to pack up to multiple packets in a single descriptor session in order to save PCI bandwidth.</p> <p>MPW is enabled for ConnectX-4 Lx and Enhanced MPW is enabled for ConnectX-5.</p> <p>A MPW session can carry only identical size of packets and the number of packets in a session is also limited while Enhanced MPW does not have such limitations.</p> <p>When txq_inline is set along with this option, TX burst function tries to copy entire packet data on to TX descriptor instead of including pointer of packet only if there is enough room remained in the descriptor. txq_inline sets per-descriptor space for either pointers or inlined packets. Enhanced MPW supports hybrid mode - mixing inlined packets and pointers in a descriptor.</p> <p>This option cannot be used in conjunction with tso. When tso is set, txq_mpw_en is disabled.</p> <p>It is set to 1 by default.</p> <p>To disable please set to 0.</p> |

| Parameter           | Description  |
|---------------------|--|
|                     | When multi-packet send is enabled, there is no support for TX VLAN HW insertion. To use TX VLAN HW insertion please disable it.  |
| txq_mpw_hdr_dseg_en | Enables including two pointers in the first block of TX descriptor. This can be used to lessen CPU load for memory copy.<br>Effective only when Enhanced MPW is supported. Disabled by default.  |
| txq_max_inline_len  | Maximum size of packet to be inlined. This option limits the size of packet to be inlined. If the size of a packet is larger than configured value, the packet isn't inlined even though there's enough space remained in the descriptor. Instead, the packet is included with pointer.<br>Effective only when Enhanced MPW is supported. The default value is 256.  |
| tso                 | Enables hardware TSO.<br>When hardware TSO is enabled, packets marked with TCP segmentation offload will be divided into segments by the hardware.<br>It is disabled by default.   |
| rxq_cqe_comp_en     | A nonzero value that enables optimized received packets' completion algorithm when PCI back pressure is detected. This improves performance at the cost of a slightly higher CPU usage.<br>It is enabled by default. To disable this please set to 0   |
| udp_port            | A nonzero value enables traffic dispatching between the kernel and DPDK according to the packet UDP port.<br>UDP packets with destination port equals to udp_port will be directed to DPDK.<br>RSS will be support for those packets, the RSS configuration must be ETH_RSS_UDP.   |
| delay_drop_rxq      | A nonzero value enables tolerance mode for rxq tail drop.<br>In case of no available descriptors on the rxq, the NIC will tail drop packets. The options forces the NIC to check the descriptors count multiple times, before the drop decision. Thus providing the PMD the option to repost descriptions without packet loss.<br>The feature is valuable for cases where the DPDK application wants to keep the Rx descriptors count small without experiencing packet lost.<br>The feature should be used with care, as delaying the drop may cause backpressure to propagate toward the link.<br>It is disabled by default. |
| swp_en              | A nonzero value enables the SW parser feature.<br>When SW parser is enabled, the packet headers location is determined according to the length fields provided in the mbuf.<br>The feature is useful in case application wants to use checksum and TSO Tx offload for non-standard packets.<br>Software parser is supported only for PFs. In case of VF, option is ignored.<br>See section <a href="#">2.5</a> for more details<br>It is disabled by default.  |

## 2.5 Checksum and TSO offload for VXLAN Packets without the Inner Ethernet Header on ConnectX-4/ConnectX-4 Lx

The feature provides the ability for HW to compute the checksums and segment non-standard VXLAN packets which do not contain the inner Ethernet header.

### 2.5.1 System Requirements

- Must run on PF

### 2.5.2 Command Line Requirements

- For ConnectX-4 Lx, force `txq_mpw_en=0`
- Enable software parser `swp_en=1`
- If TSO is required, set `tso=1`

#### 2.5.2.1 Using Checksum and TSO offload for VXLAN Packets without the Inner Ethernet Header on ConnectX-4/ConnectX-4 Lx

➤ *To use the feature, the application should:*

1. Prepare mbuf to send with the non-standard VXLAN packet
2. Set `PKT_TX_TUNNEL_VXLAN` on the mbuf
3. Fill all mbuf header length fields correctly
4. Set the checksum and TSO flags according to the application needs.

## 2.6 Sending and Receiving Jumbo Frames on ConnectX-3 Pro

If the `mbuf` size is smaller than the MTU size and you need to use scattered mbuf.

1. Modify the `config/common_base` file.
  - For ConnectX-3 PMD: `CONFIG_RTE_LIBRTE_MLX4_SGE_WR_N=4`
2. Compile DPDK.

```
On x86_64:
rm -rf x86_64-native-linuxapp-gcc
make install T=x86_64-native-linuxapp-gcc
On Power8:
rm -rf ppc_64-power8-linuxapp-gcc
make install T=ppc_64-power8-linuxapp-gcc
```

3. Set the appropriate MTU using the `rte_eth_dev_set_mtu` API.

## 2.7 Setting RX VLAN Filter on ConnectX-3 Pro

1. Make sure that regular steering mode is configured.

```
cat /sys/module/mlx4_core/parameters/log_num_mgm_entry_size -1
```

2. Modify the `/etc/modprob.d/mlnx.conf` file if required and restart `MLNX_OFED`.

Configure VLAN interface on the port using standard Linux tools.

3. Add or remove VLAN using the `rte_eth_dev_vlan_filter()` DPDK API.

## 3 System Performance Configuration

### 3.1 General Settings

- Use server with at least 4 memory channels, for better performance. In order to check the number of memory channels use:

```
# lshw -C memory
```

The output will show the server banks with their utilization

```
*-bank:1
  description: DIMM Synchronous 2400 MHz (0.4 ns)
  product: 18ASF2G72PDZ-2G3B1
  vendor: 002C00B3002C
  physical id: 1
  serial: 15B8ED61
  slot: A2
  size: 16GiB
  width: 64 bits
  clock: 2400MHz (0.4ns)
*-bank:2
  description: [empty]
  physical id: 2
  slot: A3
```

- Use the CPU near local NUMA node to which the PCIe adapter is connected, for better performance.

For Virtual Machines (VM), verify that the right CPU and NUMA node are pinned for the VM according to the above. If possible, connect you NIC near NUMA node 0. Run `mst status -v` to identify the NUMA node to which the PCIe adapter is connected

```
mst status -v
MST modules:
-----
MST PCI module loaded
MST PCI configuration module loaded
PCI devices:
-----
DEVICE_TYPE          MST          PCI          RDMA
NET
ConnectX3Pro(rev:0)  /dev/mst/mt4103_pciconf0
ConnectX3Pro(rev:0)  /dev/mst/mt4103_pci_cr0    04:00.0    mlx4_0
net-eth4,net-eth5
```

- If more than one adapter is used, verify that both adapters are located on the same PCI bus (as each CPU socket on a Crown Pass platform manages its own PCI bus) in order to forward packets from one to the other without NUMA performance penalty.

- **Check the Core Frequency**

Check that the output CPU frequency for each core is equal to the maximum supported and that all core frequencies are consistent.

- Check the maximum supported CPU frequency:

```
#cat /sys/devices/system/cpu/cpu*/cpufreq/cpuinfo_max_freq
```

- Check that the core frequencies are consistent

```
#cat /proc/cpuinfo | grep "cpu MHz"
```

- Check that the output frequencies are the same as the maximum supported

- Check the current CPU frequency to check whether it is configured to max available frequency:

```
#cat /sys/devices/system/cpu/cpu*/cpufreq/cpuinfo_cur_freq
```

When the following CPU frequency modules are loaded, CPU scaling is enabled, and you can improve performance by setting the scaling mode to performance:

- **Set the scaling mode to performance for every CPU**

```
# echo performance >
/sys/devices/system/cpu/cpu<cpunumber>/cpufreq/scaling_governor
```

For further information, please refer to Performance Tuning Guide for Mellanox Adapters

([http://www.mellanox.com/related-docs/prod\\_software/Performance\\_Tuning\\_Guide\\_for\\_Mellanox\\_Network\\_Adapters.pdf](http://www.mellanox.com/related-docs/prod_software/Performance_Tuning_Guide_for_Mellanox_Network_Adapters.pdf))

- **Verify that Max\_Read\_Req BIOS parameter is set to 1K**

To obtain the current setting for the Max\_Read\_Req BIOS parameter:

```
setpci -s <NIC BIOS address> 68.w
```

```
example:
setpci -s 21:00.0 68.w
3026
```

If the output is different than 3XXX, set it by:

```
setpci -s <NIC BIOS address> 68.w=3XXX
```



**NOTE:** The XXX can be different on different systems. Make sure to configure according to the setpci output..

For example:

```
setpci -s 84:00.0 68.w
2026
Run: setpci -s 84:00.0 68.w=3026
```

- **Disable pause frames on all network interfaces managed by mlx4\_en/mlx5\_en**

```
lldpad stop
```

```
ethtool -A eth16 rx off tx off
```



**NOTE:** In certain systems, pause frames are used to increase performance.

- **Use 1Gb huge pages**

- **Hyper threading**

In certain DPDK application, enabling hyper threading results in better performance. For benchmarking purposes, it is recommended to disable hyper threading

- **Make sure that unnecessary System Management Interrupts (SMIs) are disabled**  
SMI that are used for Power Monitoring and for Memory PreFailure Notification are recommended to be disabled. Please refer to your server provider guides for recommended platform tuning.

- **Isolate used cores**

Use `isolcpus` command for boot configuration.

For example, add the following to kernel boot parameters:

```
isolcpus=2,3
```

- **Disable kernel memory compaction**

```
echo never > /sys/kernel/mm/transparent_hugepage/defrag
echo never > /sys/kernel/mm/transparent_hugepage/enabled
echo 0 > /sys/kernel/mm/transparent_hugepage/khugepaged/defrag
sysctl -w vm.swappiness=0
sysctl -w vm.zone_reclaim_mode=0
```

- **Interrupts configuration**



**NOTE:** Interrupts configuration should be performed only if the needed performance was not achieved.

- Stop irqbalancer

```
service irqbalance stop
```

- Set all possible interrupts to different NUMA:

```
Example: echo '6-9' | sudo tee /proc/irq/*/smp_affinity_list
```

- Set NIC interrupts to same NUMA:

Example:

```
set_irq_affinity_cpulist.sh 0-1 ethX
```

- Set other NIC interrupts to different NUMA:

Example:

```
set_irq_affinity_cpulist.sh 6-9 ethY
```

- **IO non-posted prefetch settings**

Verify IO non-posted prefetch is disabled by default. This can be checked via the BIOS configuration.

Please contact your server provider for more information about the settings.

### 3.1.1 ConnectX-3 Pro Setting

- Verify the optimized steering mode is configured

```
cat /sys/module/mlx4_core/parameters/log_num_mgm_entry_size -7
```

If not, modify `/etc/modprobe.d/mlx4_core.conf` and restart `MLNX_OFED`

- Use environment variable `MLX4_INLINE_RECV_SIZE=64` to get maximum performance for 64B messages

Example:

```
MLX4_INLINE_RECV_SIZE=64 ./testpmd -c 0xe000 -n 4 --socket-mem=0,2048 --
--port-numa-config=0,1,1,1 --socket-num=1 --burst=64 --txd=256 --rxd=256
--mbcache=512 --rxq=1 --txq=1 --nb-cores=2 --i
```

### 3.1.2 ConnectX-4 / ConnectX-4 Lx / ConnectX-5 Settings

- Configure aggressive CQE Zipping for maximum performance

```
mlxconfig -d mlx5_0 s CQE_COMPRESSION=1
```

- To set it back to the default CQE Zipping mode

```
mlxconfig -d mlx5_0 s CQE_COMPRESSION=0
```

- x86\_64

Use command line parameter *txq\_inline* and *txqs\_min\_inline* to get maximum performance:

- Messages sizes <=128B: *txq\_inline*=128, *txqs\_min\_inline*=4
- Message sizes > 128: *txq\_inline*=64, *txqs\_min\_inline*=4
- ConnectX-4 LX: *txq\_inline*=200, *txqs\_min\_inline*=4
- ConnectX-5: use default values

Example:

```
./testpmd -c 0x1ff -n 4 -w
0000:08:00.0,txq_inline=128,txqs_min_inline=4 -w
0000:08:00.1,txq_inline=128,txqs_min_inline=4 --socket-mem=2048,0 --
--port-numa-config=0,0,1,0 --socket-num=0 --burst=64 --txd=1024 --
rxd=256 --mbcache=512 --rxq=4 --txq=4 --nb-cores=8 -i
```

## 3.2 KVM Settings

1. Make sure that Hypervisor kernel is 3.16 or newer (For example Ubuntu 14.10 or Fedora 20/21 can be used).
2. Configure boot with "iommu=pt".
3. Use 1G huge pages.
4. Make sure to allocate a VM on huge pages

Example:

qemu is started with the following commands:

```
umount /mnt/huge 2> /dev/null
mount -t hugetlbfs none /mnt/huge &&
echo 8192 > /sys/kernel/mm/hugepages/hugepages-2048kB/nr_hugepages &&

numactl --cpunodebind 1 --membind 1 -- \
qemu-system-x86_64 \
-smp 24 \
-m 4G \
-mem-path /mnt/huge \
-mem-prealloc \
-enable-kvm \
-cpu host \
-serial tcp::999,server,nowait \
-nographic \
-vga none \
- -device pci-assign,host=83:00.1 \
-device pci-assign,host=84:00.1 \
-drive snapshot=on,file=/opt/vm/ubuntu-14.04-template.qcow2 \
-drive file=/data/data.img
```

Since both adapters are installed on NUMA node 1 PCI slots (CPU socket 1), \*numactl is used to bind qemu to CPU threads and memory from that node only, which makes a Virtual Machine without NUMA internally.

All its memory is allocated from huge pages in /mnt/huge.

- After loading VM, verify huge pages on your Hypervisor is used by VM:

```
cat /sys/devices/system/node/node<NUM>/hugepages/hugepages-<PAGE-  
SIZE>/free_hugepages
```

- Make sure to set CPU pining

For example if you run qemu:

```
(qemu) info cpus  
CPU #0: pc=0xffffffff81056306 (halted) thread_id=2719  
CPU #1: pc=0xffffffff81056306 (halted) thread_id=2720  
CPU #2: pc=0xffffffff81056306 (halted) thread_id=2723  
CPU #3: pc=0xffffffff81056306 (halted) thread_id=2724  
taskset -p 0x1 2719  
taskset -p 0x2 2720  
taskset -p 0x4 2723  
taskset -p 0x8 2724
```

- It is recommended to use small TX and RX queues, the number of descriptors should be <= 512



## 4 Running DPDK Application with Mellanox Poll-Mode Driver

Mellanox NICs must not be unbound from the kernel. Although the control path is not through uio/vfio, it still uses the existing kernel drivers while the data path bypasses kernel.

### 4.1 ConnectX®-3 Pro

For example:

```
./testpmd -c 0xe000 -n 4 --socket-mem=0,2048 -- --port-numa-config=0,1,1,1
--socket-num=1 --burst=64 --txd=256 --rxd=256 --mbcache=512 --rxq=1 --
txq=1 --nb-cores=2 --i
```

Specific PCI address can be used to specify the NIC:

```
./testpmd -c 0xe000 -n 4 -w 0000:08:00.0 --socket-mem=0,2048 -- --port-numa-
config=0,1,1,1 --socket-num=1 --burst=64 --txd=256 --rxd=256 --mbcache=512
--rxq=1 --txq=1 --nb-cores=2 --i
```



**NOTE:** In ConnectX-3 NICs, a single PCI address represents 2 ports.

When running bi-directional traffic, for better performance, use the receive-inline feature that can be enabled by the env variable `MLX4_INLINE_RECV_SIZE`.

Example: for 64B messages

```
MLX4_INLINE_RECV_SIZE=64 ./testpmd -c 0xe000 -n 4 --socket-mem=0,2048 -- --
port-numa-config=0,1,1,1 --socket-num=1 --burst=64 --txd=256 --rxd=256 --
mbcache=512 --rxq=1 --txq=1 --nb-cores=2 --i
```

### 4.2 ConnectX®-4 / ConnectX®-4 Lx / ConnectX®-5

For example:

Specific PCI address can be used to specify NIC's ports:

```
./testpmd -c 0x1fff -n 4 -w 0000:08:00.0
-w 0000:08:00.1 --socket-mem=2048,0 -- --port-numa-config=0,0,1,0 --socket-
num=0 --burst=64 --txd=1024 --rxd=256 --mbcache=512 --rxq=4 --txq=4 --nb-
cores=8 --i
```



**NOTE:** In ConnectX-4 NICs, a PCI address represents each port.

For best performance with ConnectX-4 NIC:

```
./testpmd -c 0x1fff -n 4 -w 0000:08:00.0,txq_inline=128
-w 0000:08:00.1,txq_inline=128 --socket-mem=2048,0 -- --port-numa-
config=0,0,1,0 --socket-num=0 --burst=64 --txd=1024 --rxd=256 --mbcache=512
--rxq=4 --txq=4 --nb-cores=8 --rss-udp --i
```



**NOTE:** ConnectX-4: It is recommended to use **txq\_inline=128** parameter for best performance for 64B and 128B messages with more multiple cores to achieve max performance.



**NOTE:** ConnectX-4: It is recommended to use **txq\_inline=64** parameter for best performance for messages  $\geq 256B$  with multiple cores to achieve max performance.



**NOTE:** `--rss-udp testpmd` option should be used to achieve best spreading of UDP flows.

Some ConnectX-4 Lx cards are single port cards. To run testpmd fwd test on one port:

```
./testpmd -c 0xe000 -n 4 -w 0000:08:00.0,txq_inline=200 --socket-  
mem=2048,0 -- --port-numa-config=0,0,1,0 --socket-num=0 --burst=64 --  
txd=4096 --rxd=1024 --mbcache=512 --rxq=4 --txq=4 --nb-cores=4 --rss-udp --i
```



**NOTE:** ConnectX-4 LX: It is recommended to use **txq\_inline=200** parameter for best performance for 64B messages with multiple cores to achieve maximum performance.



**NOTE:** `--rss-udp testpmd` option should be used to achieve best spreading of UDP flows.

For ConnectX-5 cards, default values will bring the best performance:

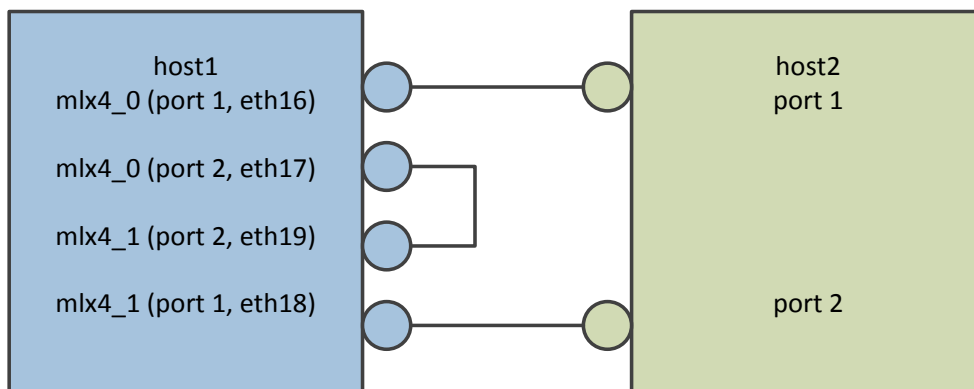
```
./testpmd -c 0xff00ff00 -n 4 -w 84:00.0 --socket-mem=0,8192 -- --port-numa-  
config=0,1 --socket-num=1 --burst=128 --txd=512 --rxd=512 --mbcache=512 -  
rxq=15 --txq=15 --nb-cores=15 -i --rss-udp -a
```

## 5 Sanity Check

Provided that all software components have been successfully installed and at least one ConnectX® adapter is present in the host system, run testpmd to test PMD.

These examples assume that there is a host with two dual port adapters that that:

- First port of each NIC is linked to another similar host
- Second port of each NIC is linked with each other



Run `*testpmd*` interactively from the DPDK build tree (for more information about its command-line options, please refer to its documentation:

[http://www.dpdk.org/doc/guides/testpmd\\_app\\_ug/index.html](http://www.dpdk.org/doc/guides/testpmd_app_ug/index.html)):

```

root# ~/DPDK/x86_64-native-linuxapp-gcc/app/test-pmd/testpmd -c
0xf000f000 -n 4 -d -- -i
EAL: Detected lcore 0 as core 0 on socket 0
EAL: Detected lcore 1 as core 1 on socket 0
EAL: Detected lcore 2 as core 2 on socket 0
EAL: Detected lcore 3 as core 3 on socket 0
EAL: Detected lcore 4 as core 4 on socket 0
[...]
EAL: Detected lcore 27 as core 3 on socket 1
EAL: Detected lcore 28 as core 4 on socket 1
EAL: Detected lcore 29 as core 5 on socket 1
EAL: Detected lcore 30 as core 6 on socket 1
EAL: Detected lcore 31 as core 7 on socket 1
EAL: Support maximum 64 logical core(s) by configuration.
EAL: Detected 32 lcore(s)
EAL: cannot open VFIO container, error 2 (No such file or directory)
EAL: VFIO support could not be initialized
EAL: Setting up memory...
EAL: Ask a virtual area of 0x6400000 bytes
EAL: Virtual area found at 0x7f15fd600000 (size = 0x6400000)
EAL: Ask a virtual area of 0x200000 bytes
[...]
EAL: PCI device 0000:83:00.0 on NUMA socket 1
EAL: probe driver: 15b3:1007 librte_pmd_mlx4
PMD: librte_pmd_mlx4: PCI information matches, using device "mlx4_0"
(VF: false)
PMD: librte_pmd_mlx4: 2 port(s) detected
PMD: librte_pmd_mlx4: bad state for port 1: "down" (1)
PMD: librte_pmd_mlx4: port 1 MAC address is 00:02:c9:b5:b7:50
PMD: librte_pmd_mlx4: bad state for port 2: "down" (1)
PMD: librte_pmd_mlx4: port 2 MAC address is 00:02:c9:b5:b7:51
EAL: PCI device 0000:84:00.0 on NUMA socket 1
EAL: probe driver: 15b3:1007 librte_pmd_mlx4
    
```

```

PMD: librte_pmd_mlx4: PCI information matches, using device "mlx4_1"
(VF: false)
PMD: librte_pmd_mlx4: 2 port(s) detected
PMD: librte_pmd_mlx4: bad state for port 1: "down" (1)
PMD: librte_pmd_mlx4: port 1 MAC address is 00:02:c9:b5:ba:b0
PMD: librte_pmd_mlx4: bad state for port 2: "down" (1)
PMD: librte_pmd_mlx4: port 2 MAC address is 00:02:c9:b5:ba:b1
Interactive-mode selected
Configuring Port 0 (socket 0)
PMD: librte_pmd_mlx4: 0x7f35e0: TX queues number update: 0 -> 1
PMD: librte_pmd_mlx4: 0x7f35e0: RX queues number update: 0 -> 1
Port 0: 00:02:C9:B5:B7:50
Configuring Port 1 (socket 0)
PMD: librte_pmd_mlx4: 0x7f3620: TX queues number update: 0 -> 1
PMD: librte_pmd_mlx4: 0x7f3620: RX queues number update: 0 -> 1
Port 1: 00:02:C9:B5:B7:51
Configuring Port 2 (socket 0)
PMD: librte_pmd_mlx4: 0x7f3660: TX queues number update: 0 -> 1
PMD: librte_pmd_mlx4: 0x7f3660: RX queues number update: 0 -> 1
Port 2: 00:02:C9:B5:BA:B0
Configuring Port 3 (socket 0)
PMD: librte_pmd_mlx4: 0x7f36a0: TX queues number update: 0 -> 1
PMD: librte_pmd_mlx4: 0x7f36a0: RX queues number update: 0 -> 1
Port 3: 00:02:C9:B5:BA:B1
Checking link statuses...
Port 0 Link Up - speed 10000 Mbps - full-duplex
Port 1 Link Up - speed 40000 Mbps - full-duplex
Port 2 Link Up - speed 10000 Mbps - full-duplex
Port 3 Link Up - speed 40000 Mbps - full-duplex
Done
testpmd>

```

The following commands are typed from the \*testpmd\* interactive prompt.

### 1. Check port status:

```

testpmd> show port info all
***** Infos for port 0 *****
MAC address: 00:02:C9:B5:B7:50
Connect to socket: 0
memory allocation on the socket: 0
Link status: up
Link speed: 10000 Mbps
Link duplex: full-duplex
Promiscuous mode: enabled
Allmulticast mode: disabled
Maximum number of MAC addresses: 128
Maximum number of MAC addresses of hash filtering: 0
VLAN offload:
  strip on
  filter on
  qinq(extend) off

***** Infos for port 1 *****
MAC address: 00:02:C9:B5:B7:51
Connect to socket: 0
memory allocation on the socket: 0
Link status: up
Link speed: 40000 Mbps
Link duplex: full-duplex
Promiscuous mode: enabled
Allmulticast mode: disabled
Maximum number of MAC addresses: 128
Maximum number of MAC addresses of hash filtering: 0
VLAN offload:
  strip on
  filter on

```

```

qinq(extend) off

***** Infos for port 2 *****
MAC address: 00:02:C9:B5:BA:B0
Connect to socket: 0
memory allocation on the socket: 0
Link status: up
Link speed: 10000 Mbps
Link duplex: full-duplex
Promiscuous mode: enabled
Allmulticast mode: disabled
Maximum number of MAC addresses: 128
Maximum number of MAC addresses of hash filtering: 0
VLAN offload:
  strip on
  filter on
  qinq(extend) off

***** Infos for port 3 *****
MAC address: 00:02:C9:B5:BA:B1
Connect to socket: 0
memory allocation on the socket: 0
Link status: up
Link speed: 40000 Mbps
Link duplex: full-duplex
Promiscuous mode: enabled
Allmulticast mode: disabled
Maximum number of MAC addresses: 128
Maximum number of MAC addresses of hash filtering: 0
VLAN offload:
  strip on
  filter on
  qinq(extend) off
testpmd>

```

2. Start IO forwarding between ports 1 and 3. The `*tx_first*` argument tells `*testpmd*` to send a single packet burst which will be always forwarded by both ports:

```

testpmd> set fwd io
Set io packet forwarding mode
testpmd> set portlist 1,3
previous number of forwarding ports 4 - changed to number of
configured ports 2
testpmd> start tx_first
io packet forwarding - CRC stripping disabled - packets/burst=32
nb forwarding cores=1 - nb forwarding ports=2
RX queues=1 - RX desc=128 - RX free threshold=0
RX threshold registers: pthresh=8 hthresh=8 wthresh=0
TX queues=1 - TX desc=512 - TX free threshold=0
TX threshold registers: pthresh=32 hthresh=0 wthresh=0
TX RS bit threshold=0 - TXQ flags=0x0
testpmd>

```

3. Display `*testpmd*` port statistics:

```

testpmd> show port stats all

##### NIC statistics for port 0 #####
RX-packets: 0          RX-missed: 0          RX-bytes: 0
RX-badcrc: 0          RX-badlen: 0          RX-errors: 0
RX-nombuf: 0
TX-packets: 0          TX-errors: 0          TX-bytes: 0
#####

##### NIC statistics for port 1 #####
RX-packets: 60800584  RX-missed: 0          RX-bytes: 3891239534
RX-badcrc: 0          RX-badlen: 0          RX-errors: 0
RX-nombuf: 0
TX-packets: 61146609  TX-errors: 0          TX-bytes: 3913382976

```

```
#####
##### NIC statistics for port 2 #####
RX-packets: 0          RX-missed: 0          RX-bytes: 0
RX-badcrc: 0          RX-badlen: 0          RX-errors: 0
RX-nombuf: 0
TX-packets: 0          TX-errors: 0          TX-bytes: 0
#####
##### NIC statistics for port 3 #####
RX-packets: 61146920  RX-missed: 0          RX-bytes: 3913402990
RX-badcrc: 0          RX-badlen: 0          RX-errors: 0
RX-nombuf: 0
TX-packets: 60800953  TX-errors: 0          TX-bytes: 3891262080
#####
testpmd>
```

4. Stop forwarding:

```
testpmd> stop
Telling cores to stop...
Waiting for lcores to finish...

----- Forward statistics for port 1 -----
RX-packets: 78238689      RX-dropped: 0          RX-total: 78238689
TX-packets: 78681769      TX-dropped: 0          TX-total: 78681769
-----

----- Forward statistics for port 3 -----
RX-packets: 78681737      RX-dropped: 0          RX-total: 78681737
TX-packets: 78238721      TX-dropped: 0          TX-total: 78238721
-----

+++++++ Accumulated forward statistics for all ports+++++++
RX-packets: 156920426      RX-dropped: 0          RX-total: 156920426
TX-packets: 156920490      TX-dropped: 0          TX-total: 156920490
+++++++

Done.
testpmd>
```

5. Exit testpmd.

```
testpmd> quit
Stopping port 0...done
Stopping port 1...done
Stopping port 2...done
Stopping port 3...done
bye...
root#
```