



Ubuntu 20.10 Linux Inbox Driver User Manual

20.10

Document History

Version	Date	Description of Change
Ubuntu 20.10	January 2021	Initial release of this document

Table of Contents

- 1 **Firmware Burning4**
- 2 **Port Type Management.....6**
 - 2.1 Port Type Management/VPI Cards Configuration 6
- 3 **Modules Loading and Unloading11**
- 4 **Important Packages and Their Installation12**
- 5 **SR-IOV Configuration.....13**
 - 5.1 Setting up SR-IOV 13
- Default RoCE Mode Setting for RDMA_CM Application15**

1 Firmware Burning

1. Check the device's PCI address.

```
lspci | grep Mellanox
```

Example:

```
04:00.0 Ethernet controller: Mellanox Technologies MT27700 Family
[ConnectX-4]
04:00.1 InfiniBand controller: Mellanox Technologies MT27700 Family
[ConnectX-4]
07:00.0 Ethernet controller: Mellanox Technologies MT27710 Family
[ConnectX-4 Lx]
07:00.1 Ethernet controller: Mellanox Technologies MT27710 Family
[ConnectX-4 Lx]
0a:00.0 Network controller: Mellanox Technologies MT27520 Family
[ConnectX-3 Pro]
21:00.0 InfiniBand controller: Mellanox Technologies MT27600 [Connect-IB]
24:00.0 Ethernet controller: Mellanox Technologies MT28800 Family
[ConnectX-5 Ex]
24:00.1 InfiniBand controller: Mellanox Technologies MT28800 Family
[ConnectX-5 Ex]
```

2. Identify the adapter card's PSID.

```
# mstflint -d 81:00.0 q
Image type:          FS4
FW Version:          16.26.4012
FW Release Date:     10.12.2019
Product Version:     16.26.4012
Rom Info:            type=UEFI version=14.19.17 cpu=AMD64
                    type=PXE version=3.5.805 cpu=AMD64
Description:         UID                               GuidNumber
Base GUID:           ec0d9a0300d42de4                  8
Base MAC:             ec0d9ad42de4                     8
Image VSD:           N/A
Device VSD:          N/A
PSID:                MT_0000000009
Security Attributes: N/A
```

3. Download the firmware BIN file from the Mellanox website that matches your card's PSID:
www.mellanox.com → Support → Support → Firmware Download
4. Burn the firmware.
5. Reboot your machine after the firmware burning is completed.
6. Validate new firmware burned successfully:

```
# ethtool -i ens3
driver: mlx5_core
```

```
version: 5.0-0
firmware-version: 16.26.4012 (MT_00000000009)
expansion-rom-version:
bus-info: 0000:24:00.0
supports-statistics: yes
supports-test: yes
supports-eeprom-access: no
supports-register-dump: no
supports-priv-flags: yes
```

2 Port Type Management

2.1 Port Type Management/VPI Cards Configuration

ConnectX®-3/ConnectX®-3 Pro/ConnectX®-4 ports can be individually configured to work as InfiniBand or Ethernet ports. By default, both ConnectX®-5 VPI ports are initialized as InfiniBand ports. If you wish to change the port type use the `mstconfig` after the driver is loaded.

1. Install mstflint tools.

```
apt install mstflint
```

2. Check the PCI address.

```
lspci | grep Mellanox
```

Example:

```
24:00.0 Ethernet controller: Mellanox Technologies MT28800 Family
[ConnectX-5 Ex]
```

3. Use mstconfig to change the link type as desired IB - for InfiniBand, ETH - for Ethernet.

```
mstconfig -d <device pci> s LINK_TYPE_P1/2=<ETH|IB|VPI>
```

Example:

```
# mstconfig -d 00:06.0 s LINK_TYPE_P1=ETH
```

Device #1:

```
Device type:    ConnectX5
Name:           MCX556A-EDA_Ax
Description:    ConnectX-5 Ex VPI adapter card; EDR IB (100Gb/s)
and 100GbE; dual-port QSFP28; PCIe4.0 x16; tall bracket; ROHS R6
Device:         24:00.0
```

Configurations:	Next Boot	New
LINK_TYPE_P1	IB (1)	
ETH (2)		

```
Apply new Configuration? (y/n) [n] : y
```

```
Applying... Done!
```

```
-I- Please reboot machine to load new configurations.
```

4. Reboot your machine.

5. Query the device's parameters to validate the new configuration.

```
# mstconfig -d 00:06.0 q

Device #1:
-----

Device type:      ConnectX5
Name:             MCX556A-EDA_Ax
Description:      ConnectX-5 Ex VPI adapter card; EDR IB (100Gb/s)
and 100GbE; dual-port QSFP28; PCIe4.0 x16; tall bracket; ROHS R6
Device:           24:00.0

Configurations:                                     Next Boot
MEMIC_BAR_SIZE                                     0
MEMIC_SIZE_LIMIT                                  _256KB(1)
HOST_CHAINING_MODE                                DISABLED(0)
HOST_CHAINING_DESCRIPTORs                         Array[0..7]
HOST_CHAINING_TOTAL_BUFFER_SIZE                   Array[0..7]
FLEX_PARSER_PROFILE_ENABLE                        0
FLEX_IPV4_OVER_VXLAN_PORT                         0
ROCE_NEXT_PROTOCOL                                254
ESWITCH_HAIRPIN_DESCRIPTORs                       Array[0..7]
ESWITCH_HAIRPIN_TOT_BUFFER_SIZE                   Array[0..7]
NON_PREFETCHABLE_PF_BAR                          False(0)
NUM_OF_VFS                                        4
SRIOV_EN                                           True(1)
PF_LOG_BAR_SIZE                                   5
VF_LOG_BAR_SIZE                                   1
NUM_PF_MSIX                                       63
NUM_VF_MSIX                                       11
INT_LOG_MAX_PAYLOAD_SIZE                         AUTOMATIC(0)
SW_RECOVERY_ON_ERRORS                            False(0)
RESET_WITH_HOST_ON_ERRORS                       False(0)
ADVANCED_POWER_SETTINGS                          False(0)
CQE_COMPRESSION                                  BALANCED(0)
IP_OVER_VXLAN_EN                                 False(0)
PCI_ATOMIC_MODE                                   PCI_ATOMIC_MODE
PCI_ATOMIC_DISABLED_EXT_ATOMIC_ENABLED(0)
LRO_LOG_TIMEOUT0                                  6
LRO_LOG_TIMEOUT1                                  7
LRO_LOG_TIMEOUT2                                  8
LRO_LOG_TIMEOUT3                                  13
LOG_DCR_HASH_TABLE_SIZE                          11
DCR_LIFO_SIZE                                     16384
LINK_TYPE_P1                                     ETH(2)
LINK_TYPE_P2                                     IB(1)
```

ROCE_CC_PRIO_MASK_P1	255
ROCE_CC_ALGORITHM_P1	ECN(0)
ROCE_CC_PRIO_MASK_P2	255
ROCE_CC_ALGORITHM_P2	ECN(0)
CLAMP_TGT_RATE_AFTER_TIME_INC_P1	True(1)
CLAMP_TGT_RATE_P1	False(0)
RPG_TIME_RESET_P1	300
RPG_BYTE_RESET_P1	32767
RPG_THRESHOLD_P1	1
RPG_MAX_RATE_P1	0
RPG_AI_RATE_P1	5
RPG_HAI_RATE_P1	50
RPG_GD_P1	11
RPG_MIN_DEC_FAC_P1	50
RPG_MIN_RATE_P1	1
RATE_TO_SET_ON_FIRST_CNP_P1	0
DCE_TCP_G_P1	1019
DCE_TCP_RTT_P1	1
RATE_REDUCE_MONITOR_PERIOD_P1	4
INITIAL_ALPHA_VALUE_P1	1023
MIN_TIME_BETWEEN_CNPS_P1	2
CNP_802P_PRIO_P1	6
CNP_DSCP_P1	48
CLAMP_TGT_RATE_AFTER_TIME_INC_P2	True(1)
CLAMP_TGT_RATE_P2	False(0)
RPG_TIME_RESET_P2	300
RPG_BYTE_RESET_P2	32767
RPG_THRESHOLD_P2	1
RPG_MAX_RATE_P2	0
RPG_AI_RATE_P2	5
RPG_HAI_RATE_P2	50
RPG_GD_P2	11
RPG_MIN_DEC_FAC_P2	50
RPG_MIN_RATE_P2	1
RATE_TO_SET_ON_FIRST_CNP_P2	0
DCE_TCP_G_P2	1019
DCE_TCP_RTT_P2	1
RATE_REDUCE_MONITOR_PERIOD_P2	4
INITIAL_ALPHA_VALUE_P2	1023
MIN_TIME_BETWEEN_CNPS_P2	2
CNP_802P_PRIO_P2	6
CNP_DSCP_P2	48
LLDP_NB_DCBX_P1	False(0)
LLDP_NB_RX_MODE_P1	OFF(0)
LLDP_NB_TX_MODE_P1	OFF(0)
LLDP_NB_DCBX_P2	False(0)

LLDP_NB_RX_MODE_P2	OFF(0)
LLDP_NB_TX_MODE_P2	OFF(0)
DCBX_IEEE_P1	True(1)
DCBX_CEE_P1	True(1)
DCBX_WILLING_P1	True(1)
DCBX_IEEE_P2	True(1)
DCBX_CEE_P2	True(1)
DCBX_WILLING_P2	True(1)
KEEP_ETH_LINK_UP_P1	True(1)
KEEP_IB_LINK_UP_P1	False(0)
KEEP_LINK_UP_ON_BOOT_P1	False(0)
KEEP_LINK_UP_ON_STANDBY_P1	False(0)
KEEP_ETH_LINK_UP_P2	True(1)
KEEP_IB_LINK_UP_P2	False(0)
KEEP_LINK_UP_ON_BOOT_P2	False(0)
KEEP_LINK_UP_ON_STANDBY_P2	False(0)
NUM_OF_VL_P1	_4_VLs(3)
NUM_OF_TC_P1	_8_TCs(0)
NUM_OF_PFC_P1	8
NUM_OF_VL_P2	_4_VLs(3)
NUM_OF_TC_P2	_8_TCs(0)
NUM_OF_PFC_P2	8
DUP_MAC_ACTION_P1	LAST_CFG(0)
SRIOV_IB_ROUTING_MODE_P1	LID(1)
IB_ROUTING_MODE_P1	LID(1)
DUP_MAC_ACTION_P2	LAST_CFG(0)
SRIOV_IB_ROUTING_MODE_P2	LID(1)
IB_ROUTING_MODE_P2	LID(1)
PCI_WR_ORDERING	per_mkey(0)
MULTI_PORT_VHCA_EN	False(0)
PORT_OWNER	True(1)
ALLOW_RD_COUNTERS	True(1)
RENEG_ON_CHANGE	True(1)
TRACER_ENABLE	True(1)
IP_VER	IPv4(0)
BOOT_UNDI_NETWORK_WAIT	0
UEFI_HII_EN	False(0)
BOOT_DBG_LOG	False(0)
UEFI_LOGS	DISABLED(0)
BOOT_VLAN	1
LEGACY_BOOT_PROTOCOL	PXE(1)
BOOT_RETRY_CNT1	NONE(0)
BOOT_LACP_DIS	True(1)
BOOT_VLAN_EN	False(0)
BOOT_PKEY	0
EXP_ROM_UEFI_x86_ENABLE	False(0)

EXP_ROM_PXE_ENABLE	True(1)
IBM_TUNNELED_ATOMIC_EN	False(0)
IBM_AS_NOTIFY_EN	False(0)
ADVANCED_PCI_SETTINGS	False(0)
SAFE_MODE_THRESHOLD	10
SAFE_MODE_ENABLE	True(1)

3 Modules Loading and Unloading

Mellanox modules for ConnectX®-2/ConnectX®-3/ConnectX®-3 Pro are:

- ▶ mlx4_en, mlx4_core, mlx4_ib

Mellanox modules for ConnectX®-4/ConnectX®-4 Lx/ConnectX®-5 are:

- ▶ mlx5_core, mlx5_ib

In order to unload the driver, you need to first unload mlx*_en/ mlx*_ib and then the mlx*_core module.

- ▶ To load and unload the modules, use the commands below:

- **Loading the driver:** `modprobe <module name>`

```
modprobe mlx5_ib
```

- **Unloading the driver:** `modprobe -r <module name>`

```
modprobe -r mlx5_ib
```

4 Important Packages and Their Installation

rdma-core

rdma-core RDMA core userspace libraries and daemons

libibmad5: Low layer InfiniBand diagnostic and management programs

libibmad5 OpenFabrics Alliance InfiniBand MAD library

opensm: InfiniBand Subnet Manager

opensm OpenIB InfiniBand Subnet Manager and management utilities

ibutils: OpenIB Mellanox InfiniBand Diagnostic Tools

ibutils OpenIB Mellanox InfiniBand Diagnostic Tools

infiniband-diags: OpenFabrics Alliance InfiniBand Diagnostic Tools

infiniband-diags OpenFabrics Alliance InfiniBand Diagnostic Tools

perftest: IB Performance tests

perftest IB Performance Tests

mstflint: Mellanox Firmware Burning and Diagnostics Tools

mstflint Mellanox firmware burning tool

- To install the packages above, run:

```
# apt-get install <packages names>
```

5 SR-IOV Configuration

5.1 Setting up SR-IOV

1. Download mstflint tools.

```
# apt install mstflint
```

2. Check the device's PCI.

```
lspci | grep Mellanox
```

Example:

```
24:00.0 Ethernet controller: Mellanox Technologies MT28800 Family
[ConnectX-5 Ex]
```

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

```
mstconfig -d <device pci> q
```

Example:

```
# mstconfig -d 00:06.0 q
```

Device #1:

```
Device type:    ConnectX3Pro
PCI device:     00:06.0
```

Configurations:	Current
SRIOV_EN	True (1)
NUM_OF_VFS	8
LINK_TYPE_P1	ETH (2)
LINK_TYPE_P2	IB (1)
LOG_BAR_SIZE	3
BOOT_PKEY_P1	0
BOOT_PKEY_P2	0
BOOT_OPTION_ROM_EN_P1	True (1)
BOOT_VLAN_EN_P1	False (0)
BOOT_RETRY_CNT_P1	0
LEGACY_BOOT_PROTOCOL_P1	PXE (1)
BOOT_VLAN_P1	1
BOOT_OPTION_ROM_EN_P2	True (1)
BOOT_VLAN_EN_P2	False (0)
BOOT_RETRY_CNT_P2	0
LEGACY_BOOT_PROTOCOL_P2	PXE (1)
BOOT_VLAN_P2	1
IP_VER_P1	IPv4 (0)

4. Enable SR-IOV:

```
mstconfig -d <device pci> s SRIOV_EN=<False|True>
```

5. Configure the needed number of VFs.

```
mstconfig -d <device pci> s NUM_OF_VFS=<NUM>
```

Note: This file will be generated only if IOMMU is set in the grub.conf file (by adding "intel_iommu=on" to /boot/grub/grub.conf file).

6. [mlx4 devices only] Edit the file /etc/modprobe.d/mlx4.conf:

```
options mlx4_core num_vfs=[needed num of VFs] port_type_array=[1/2  
for IB/ETH],[ 1/2 for IB/ETH]
```

Example:

```
options mlx4_core num_vfs=8 port_type_array=1,1
```

7. [mlx5 devices only] Write to the sysfs file the number of needed VFs.

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

8. Reboot the driver.

9. Load the driver and verify that the VFs were created.

```
lspci | grep mellanox
```

Example:

```
24:00.0 Ethernet controller: Mellanox Technologies MT28800 Family  
[ConnectX-5 Ex]  
24:00.1 Infiniband controller: Mellanox Technologies MT28800  
Family [ConnectX-5 Ex]  
24:00.2 Ethernet controller: Mellanox Technologies MT28800 Family  
[ConnectX-5 Ex Virtual Function]  
24:00.3 Ethernet controller: Mellanox Technologies MT28800 Family  
[ConnectX-5 Ex Virtual Function]  
24:00.4 Ethernet controller: Mellanox Technologies MT28800 Family  
[ConnectX-5 Ex Virtual Function]  
24:00.5 Ethernet controller: Mellanox Technologies MT28800 Family  
[ConnectX-5 Ex Virtual Function]
```

For further information, refer to section Setting Up SR-IOV MLNX_OFED User Manual.

Default RoCE Mode Setting for RDMA_CM Application

1. Mount the configs file.

```
# mount -t configfs none /sys/kernel/config
```

2. Create a directory for the mlx4/mlx5 device.

```
mkdir -p /sys/kernel/config/rdma_cm/mlx4_0/
```

3. Validate what is the used RoCE mode in the default_roce_mode configs file.

```
# cat /sys/kernel/config/rdma_cm/mlx4_0/ports/1/default_roce_mode  
IB/RoCE v1
```

4. Change the default RoCE mode,

- For RoCE v1: IB/RoCE v1
- For RoCE v2: RoCE v2

```
# echo "RoCE v2" >  
/sys/kernel/config/rdma_cm/mlx4_0/ports/1/default_roce_mode  
# cat /sys/kernel/config/rdma_cm/mlx4_0/ports/1/default_roce_mode  
RoCE v2
```

```
# echo "IB/RoCE v1" >  
/sys/kernel/config/rdma_cm/mlx4_0/ports/1/default_roce_mode  
# cat /sys/kernel/config/rdma_cm/mlx4_0/ports/1/default_roce_mode  
IB/RoCE v1
```

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.

Trademarks

NVIDIA, the NVIDIA logo, and Mellanox 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.

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

Copyright

© 2020 NVIDIA Corporation. All rights reserved.

