SOFTWARE MANUAL

Red Hat Enterprise Linux (RHEL) 8.x Driver User Manual

RHEL 8.x

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
Tel: (408) 970-3400
Fax: (408) 970-3403

Table of Contents

1 Firmware Burning

1. Check the device's PCI address.

```
lspci | grep Mellanox
```

   Example:

```
00:06.0 Infiniband controller: Mellanox Technologies MT27520 Family
[ConnectX-3 Pro]
```

2. Identify the adapter card's PSID.

```
# mstflint -d 81:00.0 q
Image type:          FS2
FW Version:          2.42.5000
FW Release Date:     26.7.2017
Rom Info:            type=PXE version=3.4.752 devid=4103
Device ID:           4103
Description:         Node             Port1            Port2
Sys image
GUIDs:               e41d2d0300b3f590 e41d2d0300b3f591 e41d2d0300b3f592
e41d2d0300b3f593
MACs:                                     e41d2db3f591     e41d2db3f592
VSD:
PSID:                MT_1090111019
```

3. Download the firmware BIN file from the Mellanox website that matches your card's PSID. To download the firmware, go to www.mellanox.com (Firmware Downloads).

4. Burn the firmware.

```
# mstflint -d <lspci-device-id> -i <image-file> b
```

5. Reboot your machine after the firmware burning is completed.

2 Port Type Management

ConnectX®-3 onwards adapter cards' ports can be individually configured to work as InfiniBand or Ethernet ports. By default, ConnectX® family adapter cards 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.

```
yum install mstflint
```

2. Check the device's PCI address.

```
lspci | grep Mellanox
```

Example:

```
00:06.0 Infiniband controller: Mellanox Technologies MT27520 Family
[ConnectX-3 Pro]
```

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:    ConnectX3Pro
PCI device:     00:06.0

Configurations:                                 Current         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.

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 onwards 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

opensm: InfiniBand Subnet Manager

`opensm-libs`	Libraries used by OpenSM and included utilities
`opensm`	OpenIB InfiniBand Subnet Manager and management utilities

infiniband-diags: OpenFabrics Alliance InfiniBand Diagnostic Tools and libibmad Low layer InfiniBand diagnostic and management programs

`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:***

```
# sudo yum install rdma-core libibverbs libibverbs-utils librdmacm libibumad
opensm infiniband-diags srptools perftest mstflint librdmacm-utils -y
```

5 SR-IOV Configuration

5.1 Setting up SR-IOV

1. Install the mstflint tools.

```
# yum install mstflint
```

2. Check the device's PCI.

```
# lspci | grep Mellanox
```

Example:

```
00:06.0 Infiniband controller: Mellanox Technologies MT27520 Family
[ConnectX-3 Pro]
```

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)
         IP_VER_P2                              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]** Create/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/net/ib2/device/sriov_numvfs
```

Example:

```
# echo 8 > /sys/class/net/ib2/device/sriov_numvfs
```

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

```
# lspci | grep mellanox
```

Example:

```
00:06.0 Network controller: Mellanox Technologies MT27520 Family
[ConnectX-3 Pro]
00:06.1 Network controller: Mellanox Technologies MT27500/MT27520 Family
[ConnectX-3/ConnectX-3 Pro Virtual Function]
00:06.2 Network controller: Mellanox Technologies MT27500/MT27520 Family
[ConnectX-3/ConnectX-3 Pro Virtual Function]
00:06.3 Network controller: Mellanox Technologies MT27500/MT27520 Family
[ConnectX-3/ConnectX-3 Pro Virtual Function]
00:06.4 Network controller: Mellanox Technologies MT27500/MT27520 Family
[ConnectX-3/ConnectX-3 Pro Virtual Function]
00:06.5 Network controller: Mellanox Technologies MT27500/MT27520 Family
[ConnectX-3/ConnectX-3 Pro Virtual Function]
00:06.6 Network controller: Mellanox Technologies MT27500/MT27520 Family
[ConnectX-3/ConnectX-3 Pro Virtual Function]
00:06.7 Network controller: Mellanox Technologies MT27500/MT27520 Family
[ConnectX-3/ConnectX-3 Pro Virtual Function]
00:06.0 Network controller: Mellanox Technologies MT27500/MT27520 Family
[ConnectX-3/ConnectX-3 Pro Virtual Function]
```

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

6 Default RoCE Mode Setting

1. Mount the configfs 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 configfs 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
```

7 PXE over InfiniBand Installation

PXE over InfiniBand infrastructure has additional parameter in the Boot Loader file for loading the necessary modules and interfaces and for allowing sufficient time to get the link.

To install RHEL from PXE using the IPoIB interfaces, add the following parameters to the Boot Loader file, located in the `var/lib/tftpboot/pxelinux.cfg` directory, at the PXE server:

```
bootdev=ib0 ksdevice=ib0 net.ifnames=0 biosdevname=0 rd.neednet=1
rd.bootif=0 rd.driver.pre=mlx5_ib,mlx4_ib,ib_ipoib ip=ib0:dhcp
rd.net.dhcp.retry=10 rd.net.timeout.iflink=60 rd.net.timeout.ifup=80
rd.net.timeout.carrier=80
```

Example:

```
default RH7.5
prompt 1
timeout 600
label RH7.5
kernel
append bootdev=ib0 ksdevice=ib0 net.ifnames=0 biosdevname=0 rd.neednet=1
rd.bootif=0 rd.driver.pre=mlx5_ib,mlx4_ib,ib_ipoib ip=ib0:dhcp
rd.net.dhcp.retry=10 rd.net.timeout.iflink=60 rd.net.timeout.ifup=80
rd.net.timeout.carrier=80
```