Red Hat Enterprise Linux (RHEL) 8.10 Driver Documentation

Table of contents

Overview

This is the documentation for Red Hat Enterprise Linux (RHEL) Inbox Driver. This document provides instructions on drivers for NVIDIA® ConnectX® adapter cards used in a RHEL Inbox Driver environment.

Included Documentation

- Release Notes
- User Manual

Release Notes

These are the release notes for Red Hat Enterprise Linux (RHEL) Inbox Driver. The release notes include the following sections.

General Support

Supported Uplinks to Servers

This version supports the following uplinks to servers.

Uplink/Adapter Card	Driver Name	Uplink Speed
BlueField-2	mlx5	• InfiniBand: SDR, FDR, EDR, HDR • Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE[2], 100GbE[2]
BlueField		• InfiniBand: SDR, QDR, FDR, FDR10, EDR • Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE, 100GbE
ConnectX-7		• InfiniBand: EDR, HDR100, HDR, NDR200, NDR • Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE[2], 100GbE[2], 200GbE[3]

Uplink/Adapter Card	Driver Name	Uplink Speed
ConnectX-6 Lx		• Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE[2]
ConnectX-6 Dx		• Ethernet: 10GbE, 25GbE, 40GbE, 50GbE[2], 100GbE[2], 200GbE[2]
ConnectX-6		• InfiniBand: SDR, FDR, EDR, HDR100, HDR • Ethernet: 10GbE, 25GbE, 40GbE, 50GbE[2], 100GbE[2], 200GbE[2]
ConnectX-5/ConnectX-5 Ex		• InfiniBand: SDR, QDR, FDR, FDR10, EDR • Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE, 100GbE
ConnectX-4 Lx		• Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE
ConnectX-4		• InfiniBand: SDR, QDR, FDR, FDR10, EDR • Ethernet: 1GbE, 10GbE, 25GbE, 40GbE, 50GbE, 56GbE[1], 100GbE
Innova™ IPsec EN		• Ethernet: 10GbE, 40GbE
Connect-IB®		• InfiniBand: SDR, QDR, FDR10, FDR
ConnectX-3/ConnectX-3 Pro	mlx4	• InfiniBand: SDR, QDR, FDR10, FDR • Ethernet: 10GbE, 40GbE and 56GbE [1]

1. 56GbE is an NVIDIA proprietary link speed and can be achieved while connecting an NVIDIA adapter card to NVIDIA SX10XX switch series or when connecting an NVIDIA adapter card to another NVIDIA adapter card.

2. Speed that supports both NRZ and PAM4 modes in Force mode and Auto-Negotiation mode.

3. Speed that supports PAM4 mode only.

Note

- BlueField is supported as a standard ConnectX-5 Ethernet NIC only.
- BlueField-2 is supported as a standard ConnectX-6 Dx Ethernet NIC. On the DPU, BlueField-2 is only supported as technical preview (i.e., the feature is not fully supported for production).
- ConnectX-7 is only supported as technical preview (i.e., the feature is not fully supported for production).

Supported Adapter Card Firmware Versions

This version of Red Hat Enterprise Linux (RHEL) driver supports the following NVIDIA network adapter card firmware versions.

HCA	Recommended Firmware Version
BlueField-3 (Technical Preview)	32.38.3056
ConnectX-7	28.39.1002
BlueField-2	24.38.1002
ConnectX-6 Lx	26.39.1002
ConnectX-6 Dx	22.39.1002
ConnectX-6	20.39.1002
BlueField	18.33.1048
ConnectX-5	16.35.3006
ConnectX-4 Lx	14.32.1010
ConnectX-4	12.28.2006
ConnectX-3/ConnectX-3 Pro	2.42.5000
Connect-IB	10.16.1002

SR-IOV Support

Driver	Support
mlx4_core, mlx4_en, mlx4_ib	Ethernet InfiniBand: Technical Preview
mlx5_core (includes Ethernet functionality), mlx5_ib	Ethernet InfiniBand: Technical Preview

Note

Running InfiniBand (IB) SR-IOV requires IB Virtualization support on the OpenSM (Session Manager).This capability is supported only on OpenSM provided by NVIDIA, that is not available Inbox.This support can be achieved by running the highest-priority OpenSM on a NVIDIA switch in an IB fabric.The switch SM can support this feature by enabling the virt flag (# ib sm virt enable).

Please note that this capability is not tested over the Inbox environment and is considered a technical preview.

RoCE Support

Driver	Support
mlx4—RoCE v1/v2	Yes
mlx5—RoCE v1/v2	Yes

VXLAN Support

Driver	Support
mlx4—VXLAN offload	Yes
mlx5—VXLAN offload	Yes (without RSS)

DPDK Support

Driver	Support
mlx4	NVIDIA PMD is enabled by default.
mlx5	NVIDIA PMD is enabled by default.

Open vSwitch Hardware Offloads Support

Driver	Support
mlx4	No
mlx5	Yes

Changes and New Features

Component	Feature/ Change	Description
ASAP 2	CT with Header Rewrite	**[BlueField-2]** Added support for offloading CT rules with header rewrite of L3.
	VxLAN Group Based Policy (GBP) Offload	**[ConnectX-6 Dx and above]** Added support for hardware offload of VxLAN with GBP configured.

Component	Feature/Change	Description
VDPA	Mergeable Buffer Support	**[ConnectX-6 Dx and above]** Added support for Enabled Mergeable Buffer feature on vdpa interfaces using vdpa tool to achieve better performance with large MTUs.
	Posted Interrupts	**[ConnectX-6 Dx and above]** Added support for posted interrupts in mlx5_vdpa, allowing direct IRQ propagation from the NIC to vCPU within the guest.

RHEL NIC Qualification

The following RHEL and NIC combinations successfully passed RHEL NIC qualification covering OVS functional, OVS non-offload, OVS-offload, and OVS-DPDK:

Adapter Cards	RHEL Versions
ConnectX-4 Lx, ConnectX-5	RHEL 8.0–8.x
ConnectX-5 Ex	RHEL 8.0–8.x
ConnectX-6, ConnectX-6 Dx	RHEL 8.0–8.x
ConnectX-6 Lx	RHEL 8.0–8.x

Note

For more details, see the Red Hat page “Network Adapter Fast Datapath Feature Support Matrix” at access.redhat.com/articles/3538141.

Known Inbox-Related Issues

The following table describes known issues in this release and possible workarounds.

Internal Ref.	Bugzilla Ref.	Description
2482177	-	**Description:** RDMA device name for VFs may change after resetting all VFs at once.
		Workaround: Either reset interfaces one by one with a delay in between, or use a network interface naming scheme with predictable interface names, such as NAME_PCI or NAME_GUID: copy /lib/udev/rules.d/60-rdma-persistent-naming.rules to /etc/udev/rules.d/ and edit the last line accordingly. Note that this will change interface names.
-	-	**Description:** RPM package **kernel-modules-extra** is required for supporting various OVS Hardware Offloads. To use OVS Hardware Offloads, make sure to install the **kernel-modules-extra** RPM package which provides various kernel modules that are required for supporting this functionality.
	1816660	**Description:** When the NUM_OF_VFS parameter configured in the Firmware (using the mstconfig tool) is higher than 64, VF LAG mode will not be supported while deploying OVS offload.
		Workaround: N/A
		Keywords: ConnectX-5, VF LAG, ASAP ^{2}, SwitchDev
	1816660	**Description:** An internal firmware error occurs either when attempting to disable single-root input/output virtualization or when unbinding PF using a function (such as ifdown and ip link) under the following condition: VF LAG mode in an OVS offload deployment, where at least one VF of any PF is still bound on the host or attached to a VM.
		Workaround: Unbind or detach VFs before you perform these actions as follows. 1. Shutdown and detach any VMs. 2. Remove VF LAG bond interface from OVS. 3. Unbind VFs, perform for each configured VF:

Internal Ref.	Bugzilla Ref.	Description

		`# echo <VF PCIe BDF> > /sys/bus/pci/drivers/mlx5_core/unbind`
		
		1. Disable SR-IOV, perform for each PF:
		
		`# echo 0 > /sys/class/net/<PF>/device/sriov_numvfs`

		Keywords: ConnectX-5, VF LAG, ASAP^2, SwitchDev
1284047	-	**Description:** Bandwidth degradations due to PTI (Page Table Isolation) in Intel's CPU security fix.
		Keywords: Performance
1610281	-	**Description:** Setting speed to 56GbE on ConnectX-4 causes firmware syndrome (0x1a303e).
		Workaround: N/A
		Keywords: ConnectX-4, syndrome
1609804	-	**Description:** Kernel panic during MTU change under stress traffic.
		Workaround: N/A
		Keywords: Panic, MTU

User Manual

This is the user manual for Red Hat Enterprise Linux (RHEL) Inbox Driver. The user manual includes the following sections.

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 e41d2d0300b3f590 e41d2d0300b3f591 e41d2d0300b3f592
GUIDs:
e41d2d0300b3f593
MACs: e41d2db3f591 e41d2db3f591
VSD:
PSID: MT_1090111019
```

3. Download the firmware BIN file from the NVIDIA website that matches your card's PSID. To download the firmware, go to NVIDIA's Firmware Downloads page.

4. Burn the firmware.

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

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

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 (InfiniBand) or ETH (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.

Modules Loading and Unloading

Adapter Cards	Modules
ConnectX-3/ConnectX-3	mlx4_en, mlx4_core, mlx4_ib
ConnectX-4 and above	mlx5_core, mlx5_ib

To unload the driver, 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
```

Important Packages and Their Installation

Package	Name	Description
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	infiniband-diags	OpenFabrics Alliance InfiniBand Diagnostic Tools

Package	Name	Description
and libibmad Low layer InfiniBand diagnostic and management programs		
perftest: IB Performance tests	perftest	IB Performance Tests
mstflint: NVIDIA Firmware Burning and Diagnostics Tools	mstflint	NVIDIA firmware burning tool

To install the packages above run the following:

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

SR-IOV Configuration

To set up SR-IOV, do the following:

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

```
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" in the MLNX_OFED User Manual.

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

```
# 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
```

- 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
```

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
```