SUSE Linux Enterprise Server (SLES) 12 SP3 Driver User Manual

SLES 12 SP3

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

Document Revision History

Table 1: Document Revision History

Revision	Date	Description
SLES 12 SP3	July 13, 2017	Initial version of this document.

1 Firmware Burning

1. Identify the adapter card's PSID.

```
# mstflint -d 81:00.0 q
Image type:          FS2
FW Version:          2.36.5000
FW Release Date:     26.1.2016
Rom Info:            type=PXE version=3.4.718 devid=4103
Device ID:           4103
Description:         Node             Port1            Port2
Sys image
GUIDs:               e41d2d0300b3f590 e41d2d0300b3f591 e41d2d0300b3f592
e41d2d0300b3f593
MACs:                                     e41d2db3f591     e41d2db3f592
VSD:
PSID:                MT_1090111019
```

2. Download the firmware BIN file from the Mellanox website that matches your card's PSID:

   www.mellanox.com → Support/Education→Support Downloader

3. Burn the firmware.

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

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

2 Port Type Management

2.1 Port Type Management/VPI Cards Configuration

NOTE: This tool is supported in the following devices:

- 4th generation devices: ConnectX-3, ConnectX-3 Pro (FW 2.31.5000 and above).
- 5th generation devices: Connect-IB, ConnectX-4, ConnectX-4 Lx, ConnectX-5.

Device ports can be individually configured to work as InfiniBand or Ethernet ports. By default, device ports are initialized as InfiniBand ports. If you wish to change the port type, use the mstflint tool after the driver is loaded.

5. Install mstflint tools: Zypper install mstflint.
6. Check the PCI address.

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

7. 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 82:00.1 s LINK_TYPE_P1=ETH
```

8. Reboot your machine.

3 Modules Loading and Unloading

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

- mlx4_en, mlx4_core, mlx4_ib

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

- mlx5_core, mlx5_ib

➢ ***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 userspace infrastructure and documentation, including initialization scripts, kernel driver-specific modprobe override configs, IPoIB network scripts, dracut rules, and the rdma-ndd utility.

- **Libibverbs - InfiniBand verbs library**

  Sub-packages: libibverbs and libmlx5-1

  libibverbs is a library that allows userspace processes to use RDMA "verbs" as described in the InfiniBand Architecture Specification and the RDMA Protocol Verbs Specification. This includes direct hardware access from userspace to InfiniBand/iWARP adapters (kernel bypass) for fast path operations.

  Device-specific plug-in ibverbs userspace drivers are included:

  - libcxgb3: Chelsio T3 iWARP HCA
  - libcxgb4: Chelsio T4 iWARP HCA
  - libhfi1: Intel Omni-Path HFI
  - libi40iw: Intel Ethernet Connection X722 RDMA
  - libipathverbs: QLogic InfiniPath HCA
  - libmlx4: Mellanox ConnectX-3 InfiniBand HCA
  - libmlx5: Mellanox Connect-IB/X-4+ InfiniBand HCA
  - libmthca: Mellanox InfiniBand HCA
  - libnes: NetEffect RNIC
  - libocrdma: Emulex OneConnect RDMA/RoCE Device
  - librxe: A software implementation of the RoCE protocol

- **librdmacm: RDMA cm library**
  - librdmacm-utils - Tools and Example test programs for the librdmacm library
  - librdmacm1 - librdmacm provides a userspace RDMA Communication Management API.

- **libibcm - Userspace InfiniBand Connection Management API**

  libibcm provides a userspace library that handles the majority of the low level work required to open an RDMA connection between two machines.

- **libibmad: Low layer InfiniBand diagnostic and management programs**

- libibmad-devel Development files for the libibmad library
- libibmad5 Libibamd runtime library

- **libibumad: InfiniBand Userspace Management Datagram library**

  libibumad provides the userspace management datagram (umad) library functions, which sit on top of the umad modules in the kernel. These are used by the IB diagnostic and management tools, including OpenSM.

- **opensm: InfiniBand Subnet Manager**
  - **opensm** InfiniBand Subnet Manager
  - **opensm-devel** Development files for OpenSM
  - **opensm-devel-static** Static libraries for OpenSM
  - **opensm-libs3** OpenSM runtime libraries

- **Ibutils: OpenIB Mellanox InfiniBand Diagnostic Tools ibutils**

  The ibutils package provides a set of diagnostic tools that check the health of an InfiniBand fabric.

- **Infiniband-diags OpenFabrics Alliance InfiniBand Diagnostic Tools**
  - infiniband-diags-devel - OpenFabrics Alliance InfiniBand Diagnostic Tools diags provides IB diagnostic programs and scripts needed to diagnose an IB subnet.
  - infiniband-diags-devel OpenIB InfiniBand Diagnostic Tools SDK

- **srp_daemon Tools for using the InfiniBand SRP protocol devices**

  In conjunction with the kernel ib_srp driver, srp_daemon allows you to discover and use SCSI devices via the SCSI RDMA Protocol over InfiniBand.

- **perftest: IB Performance tests**

  uverbs microbenchmarks

- **mstflint: Mellanox Firmware Burning and Diagnostics Tools**

  This package contains a burning tool and diagnostic tools for Mellanox manufactured HCA/NIC cards. It also provides access to the relevant source code. Please see the file LICENSE for licensing details. This package is based on a subset of the Mellanox Firmware Tools (MFT) package. For a full documentation of the MFT package, please refer to the downloads page at the Mellanox web site.

- ***To install the packages above, run:***

```
#zypper -n install  <package-name>
```

The following packages will be automatically installed:

- rdma-core
- librdmacm1
- libibmad5
- libibumad3

5 SR-IOV Configuration

5.1 Setting up SR-IOV in ConnectX-3/ConnectX-3 Pro

9. Download mstflint tools: zypper install mstflint
10. Check the device's PCI.

```
lspci | grep mellanox
```

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

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

Example:

```
# mstconfig -d 81:00.0 q

Device #1:
----------

Device type:    ConnectX3Pro
PCI device:     81:00.0

Configurations:                                 Current
         SRIOV_EN                               True(1)
         NUM_OF_VFS                             0
         LINK_TYPE_P1                           VPI(3)
         LINK_TYPE_P2                           VPI(3)
         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)…
```

12. Check SRIOV_EN and NUM_OF_VFS configurations.
13. Enable SR-IOV:

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

14. 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).

15. **[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
```

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

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

Example:

```
echo 8 > /sys/class/infiniband/mlx5_0/device/sriov_numvfs
```

17. Reboot the driver.

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

```
lspci | grep mellanox
```

Example:

```
dev-r-vrt-214:~ # lspci | grep nox
82:00.0 Ethernet controller: Mellanox Technologies MT27700 Family
[ConnectX-4]
82:00.1 Ethernet controller: Mellanox Technologies MT27700 Family
[ConnectX-4]
82:00.2 Ethernet controller: Mellanox Technologies MT27700 Family
[ConnectX-4 Virtual Function]
82:00.3 Ethernet controller: Mellanox Technologies MT27700 Family
[ConnectX-4 Virtual Function]
82:00.4 Ethernet controller: Mellanox Technologies MT27700 Family
[ConnectX-4 Virtual Function]
82:00.5 Ethernet controller: Mellanox Technologies MT27700 Family
[ConnectX-4 Virtual Function]
```

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

6 Default RoCE Mode Setting for RDMA_CM Application

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

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

20. 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
```

21. 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
```