Soft-RoCE

README

Rev 1.0 (Alpha)

www.mellanox.com

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 PRODUCTO(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

List of Tables

Document Revision History

Table 1: Document Revision History

Revision	Date	Description
1.0	December 04, 2014	Initial Release

1 Soft-RoCE

1.1 Overview

RoCE can be implemented in the hardware as well as in the software. Mellanox adapters include hardware implementation of RoCE to deliver the highest performance and efficiency. Soft-RoCE is the software implementation of the RoCE standard and compatible with any standard Ethernet networks.

1.2 Supported Operating Systems

The current version of the Soft-RoCE is supported in SLES11 SP3 only.

2 Soft-RoCE Installation and Configuration

NOTE: This version of Soft-RoCE supports RoCE v1 only.

2.1 Downloading Soft-RoCE

The Soft-RoCE distribution is available at:

- Kernel: git://flatbed.openfabrics.org/~amirv/rxe.git

  branch: rxe-3.0
- User space: git://flatbed.openfabrics.org/~amirv/librxe.git

  branch: librxe-1.0.0

2.2 Installing Soft-RoCE

➢ ***To install Soft-RoCE perform the following:***

1. Install SLES11 SP3.
2. Install the InfiniBand user space packages.

```
# zypper install libibverbs
# zypper install libibverbs-utils
```

3. Fetch the RXE driver and the user space sources.

   See Downloading Soft-RoCE for their location.

   a. Fetch the RXE driver.

```
# cd <workspace>/
# git clone git://flatbed.openfabrics.org/~amirv/rxe.git
```

   b. Fetch the librxe user space.

```
# git clone git://flatbed.openfabrics.org/~amirv/librxe.git
```

4. Fetch the vanilla kernel v3.0 and the RXE driver.

   a. Compile the kernel.

```
# cd <workspace>/rxe
# cp /boot/config-`uname -r` .config
# make olddefconfig
# make -j 32
# make modules_install
# make install
```

   b. Add an entry in the boot loader configuration file (grub/lilo) for the new kernel.

   c. Compile the user space.

```
# cd <workspace>/librxe
# ./configure --libdir=/usr/lib64/ --prefix=
# make
# make install
```

2.3 Configuring Soft-RoCE

Once the Soft-RoCE (aka rxe) is installed, the configuration tasks are handled via the "rxe_cfg" program.

The following are some basic Soft-RoCE functions.

- Loading kernel module and configuring previously added persistent instances

```
rxe_cfg start
```

- Adding an instance on Ethernet interface "eth4"

```
rxe_cfg add eth4
```

- Displaying the status of all rxe instances

```
rxe_cfg status
```

(or just "rxe_cfg")

- Removing an RXE instance

```
rxe_cfg remove rxe0
```

- Set the RXE MTU
  - To set the RXE MTU for all RXE instances to 4K:

```
rxe_cfg mtu 4096
```

  - To set the MTU for a single RXE instance:

```
rxe_cfg mtu rxe0 2048
```

2.3.1 Important Notes

- **MTU Settings**

  The Soft-RoCE specification allows power-of-two MTUs from 512 to 4096, although the currently available Hard RoCE implementation only supports up to 2K. Setting the MTU as described above will only take effect if the underlying Ethernet interface has a large enough MTU. For example, to support an MTU of 4096, the Ethernet MTU must be at least 4176 (80 bytes larger). If the Soft-RoCE MTU is too large to fit within the Ethernet MTU, the Soft-RoCE software will use the largest MTU that does fit. For example, if the Soft-RoCE MTU is 4096, and the Ethernet MTU is 4096, Soft-RoCE will use a 2048 byte MTU. If you subsequently increase the Ethernet MTU to 9000 (or any value greater than or equal to 4176), the Soft-RoCE MTU will automatically switch to 4096. If you use the "-f" option, as in "rxe_cfg -f mtu rxe0 4096", the relevant Ethernet MTU will be changed, if necessary, to 4176 (but will not be altered if it is already larger than 4176).

- **Persistent Instances**

  The rxe driver, as presently shipped, is not automatically loaded after boot. However, when you start up the subsystem (rxe_cfg start), the driver will be loaded and any previously configured instances will be recreated.

  By default, "rxe_cfg add" is persistent. The file /var/rxe/rxe is created, and contains the list of Ethernet interfaces for which rxe instances should be created. You can remove this file if you want to delete existing records of persistent rxe instances.

You can also add the "-n" option to the rxe_cfg add and remove subcommands. "rxe_cfg –n add eth3" will add a rxe instance on eth3 without adding a persistence entry in /var/rxe/rxe. "rxe_cfg –n remove rxe0" will remove the rxe0 instance without removing it from the /var/rxe/rxe persistence file.

For more detailed information, please refer to System Fabric Works:
http://www.systemfabricworks.com/downloads/roce

2.4 Testing Soft-RoCE over Mellanox Devices

Mellanox devices use standard RoCE with hardware offloading by default. It delivers the highest performance and efficiency. However because the default mode of the NIC is hardware offload of RoCE traffic, all traffic with ethertype 0x8915 is consumed by the NIC.

If you want to test the Soft-RoCE over such NIC, you are required to change the default ethertype value used by the RXE driver.

- ***To change the default ethertype value used by the RXE driver:***

1. Stop the RXE driver.

```
# ./rxe_cfg stop
```

2. Restart the RXE driver while assigning to it a different ethertype value.

```
# ./rxe_cfg start -p 0x8916
```