
**Release Notes**

# Linux HCA Driver for InfiniHost and InfiniHost III HCA Families

Rev 4.0.3

Linux HCA Driver for InfiniHost and InfiniHost III HCA Families Release Notes

**Document Number:**

Mellanox Technologies, Inc.
2900 Stender Way
Santa Clara, CA 95054
U.S.A.
www.Mellanox.com

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

Mellanox Technologies Ltd
PO Box 586 Hermon Building
Yokneam 20692
Israel

Tel: +972-4-909-7200
Fax: +972-4-959-3245

# 1 Overview

This is Rev 4.0.3 of the HCA driver for Linux supporting the following silicon devices:

- InfiniHost (MT23108)
- InfiniHost III Ex (MT25208)
- InfiniHost III Lx (MT25204)

This document includes the following sections:

- This Overview (describing main changes from previous release, supported platforms and operating systems, supported HCA cards)
- Driver Installation and Loading (page 7)
- Limitations and Known Bugs (page 12)
- Fixed Bugs (page 13)

## 1.1 Main Changes from version 4.0.1

- Added support for kernel 2.6.10.
- Added support for InfiniHost III Lx (MT25204).
- Bug fixes.

## 1.2 Supported Platforms and Operating Systems

Table 1 - Supported platforms and Operating Systems

| Platform | Operating System | Kernel |
|---|---|---|
| **X86** | Red Hat Enterprise Linux AS 3.0 | 2.4.21-20.ELsmp |
| | Red Hat Linux 9.0 | kernel.org: 2.4.27 (smp) |
| | Red Hat Linux 9.0 | 2.4.20-8 (smp; bigmem) |
| | SuSe SLES 9.0 | Update (2.6.5-7.111.xx-smp) |
| | SuSE Linux 9.1 Pro | 2.6.9 / 2.6.10 |
| | SuSE Linux 9.1 Pro | Update (2.6.5-7.111.xx-smp) |
| | Rocks 3.3.0 | 2.4.21-20.ELsmp |
| | Fedora Core 3 [1] | Vanilla 2.6.9 (from kernel.org) |
| **IA-64** | Red Hat Enterprise Linux AS 3.0 | 2.4.21-15.EL |
| | SuSE SLES 9.0 | 2.6.5-7.97-default |
| **AMD64 (X86_64)** | Red Hat Enterprise Linux AS 3.0 | 2.4.21-20.ELsmp |
| | SuSE SLES 9.0 | 2.6.5-7.111.xx-smp |
| | SuSE 9.1 Pro | Update (2.6.5-7.111.xx-smp) |
| | SuSE 9.1 Pro | 2.6.9/2.6.10 |
| | Rocks 3.3.0 | 2.4.21-20.ELsmp |
| | Fedora Core 3 | 2.6.9-1.667smp |
| **Intel EM64T** | Red Hat Enterprise Linux AS 3.0 | 2.4.21-20.EL |
| | SuSE SLES 9.0 RC5 | 2.6.5-7.97-smp / 2.6.5-7.111.xx-smp |
| | SuSE 9.1 Pro | 2.6.10 |
| | SuSE 9.1 Pro | Update (2.6.5-7.111.xx-smp) |
| | Rocks 3.3.0 | 2.4.21-20.ELsmp |
| | Fedora Core 3 | 2.6.9-1.667smp |

1. Fedora Core 3 originally runs kernel 2.6.9-1.667smp. However, this kernel has a small stack for the x86 architecture, therefore, the vanilla 2.6.9 kernel is used instead. For details, see Section 1.2.1 below. Furthermore, compiler gcc version 3.4.3 or later must be used.



### 1.2.1 Compiling ‘vanilla’ 2.6.9 (from kernel.org) on Fedora Core 3

Fedora Core 3 originally runs kernel 2.6.9-1.667smp. However, there are two issues regarding this kernel. The first is that it uses a small kernel stack of one page (4Kbytes) on the x86 architecture. This stack may not be sufficient for kernel applications. Therefore, the vanilla 2.6.9 is required as it allows for a larger stack.

The second issue has to do with the gcc3.4.x compiler included in Fedora Core 3. This compiler does not allow the following sequence in a source file:

1. Declare a static inline function

2. Call the function
3. Define the function

Fedora has provided a patch for the file include/compiler-gcc3.h to work around this problem by controlling the definition of the __inline__ macro.

The user needs to recompile the vanilla 2.6.9 kernel with the supplied patch before installing VAPI. To recompile, the user should follow these steps:

> Note: Use *gcc* compiler version 3.4.3 or later. The gcc version 3.4.2 that is included in Fedora Core 3 has a bug (#17581) that causes incorrect driver operation.

1. Apply the Fedora-supplied patch 'gcc34patch' (see later) as follows:
   > cd <kernel directory>
   > patch -p0 < gcc34patch  // if requested to enter a file name, enter include/linux/compiler-gcc3.h
2. Run 'make menuconfig' and make sure the following option is marked:
   General Setup | Configure standard kernel features (for small systems) | Optimize for size

### gcc34patch:

```
--- include/linux/compiler-gcc3.h.old2005-02-10 13:35:58.539171512 +0200
+++ include/linux/compiler-gcc3.h2005-02-10 13:40:32.437532632 +0200
@@ -3,10 +3,10 @@
 /* These definitions are for GCC v3.x.  */
 #include <linux/compiler-gcc.h>

-#if __GNUC_MINOR__ >= 1
-# define inline    inline                __attribute__((always_inline))
-# define __inline____inline____attribute__((always_inline))
-# define __inline__inline__attribute__((always_inline))
+#if __GNUC_MINOR__ >= 1 && __GNUC_MINOR__ < 4
+# define inline __inline__ __attribute__((always_inline))
+# define __inline__ __inline__ __attribute__((always_inline))
+# define __inline __inline__ __attribute__((always_inline))
 #endif

 #if __GNUC_MINOR__ > 0
@@ -25,7 +25,7 @@
 #if __GNUC_MINOR__ >= 1
 #define  noinline __attribute__((noinline))
 #endif
-#if __GNUC_MINOR__ >= 4
+#if __GNUC_MINOR__ > 4
 #define __must_check __attribute__((warn_unused_result))
 #endif
```

## 1.3 Supported InfiniHost Firmware Versions

This release has been QAed with:

- MT23108 InfiniHost firmware version fw-23108-rel-3.3.2
- MT25208 InfiniHost III Ex firmware version fw-25208-rel-4.6.2
- MT25208 InfiniHost III Ex firmware version fw-25218-rel-5.0.1 (MemFree)

  Note: This release will not work with fw-25208-rel-4.5.0 (MT25208 InfiniHost III Ex in InfiniHost mode)
- MT25204 InfiniHost III Lx firmware version fw-25204-rel-1.0.1

## 1.4 Supported InfiniHost Based Hardware

This release was tested with the following HCA boards:

- MHX-CEXXX-T[1] (previously MTPB23108) InfiniHost PCI-X HCA Adapter Card (Cougar)
- MHXL-CFXXX-T[1] (previously MTLP23108) Low Profile InfiniHost PCI-X HCA Adapter Card (Cougar Cub)
- MHEL-CFXXX-T[1] (previously MTLP25208) InfiniHost III Ex HCA Adapter Card (Lion)
- MHEA28-XT InfiniHost III Ex MemFree HCA Adapter Card (Lion Mini)

## 1.5 Verbs Supported

For the definition of the verbs, see the document: *Mellanox IB-Verbs API (VAPI), Rev. 1.0, Doc. #AN010601062*.

See vapi.h and evapi.h for all verbs supported.



1. XXX reflects the size of on-board memory (in MB): 128, 256, or 512.

# 2 Driver Installation and Loading

The InifiniHost HCA Driver package can be installed on Linux-based platforms.

**To install the InfiniHost HCA driver, complete the following steps:**

Note: Before installing the driver, any existing installation of Mellanox drivers (THCA or IBGD) must be unloaded first then removed using their individual un-install scripts. (The order is important to prevent the use of old modules after the new installation).

1. Download the file vapi-linux-<*version_num*>.tgz
2. Run the following as root:
   - tar xzf vapi-linux-<*version_num*>.tgz
   - cd vapi-linux-<*version_num*>
   - ./install.sh (see options “Install Script Usage”)
     Information on the installation configuration used is generated in the file: /etc/vapi-linux-release

**Notes:**

1. The kernel headers (kernel-headers-<*kernel version*>) and source (kernel-source-<*kernel version*>) packages (RPMs) must be installed in your system.
2. The file /boot/System.map-<*kernel version*> must be the same file derived from compiling the kernel image. It is used to find the function pointers of sys_mlock and sys_munlock.
3. The kernel modules are installed under: /lib/modules/<*kernel version*>/kernel/drivers/infiniband
4. The user level libraries, binaries, and include files are installed under $prefix/lib, $prefix/bin, and $prefix/include respectively. By default, $prefix is /usr.
5. The modprobe files are installed under /etc.
   The files are: kernel 2.6 - modprobe.conf.vapi; kernel 2.4 - modules.conf.vapi.
6. The driver does not use the environment variable MTHOME, and it is not defined anymore.

## 2.1 Install Script Usage

install.sh [--prefix <user-install-prefix>] [--mthome <package-install-path>] [-v <kernel version>] [-k <kernel path>] [-p <kernel patch version>] [--memtrack]

Where:

| | |
|---|---|
| --prefix | Path to user space installation (include, lib, bin; default: /usr) |
| --mthome | Path to Mellanox package installation (default: $prefix/mellanox) |
| -v, --kernel-version | Kernel version to prepare (2.4 or 2.6) (will attempt to auto-detect if not set) |
| -k, --kernel-tree | Path to top of Linux kernel tree to copy to (default is to current kernel tree) |
| --memtrack | Install with memory tracking support |
| -h, --help | For help |

The following is an example of the file content of /etc/vapi-linux-release:

BUILD_ID=”vapi-linux-4.0.2-rc<x> (TAG=vapi_4_0_2_rc<x>)” // <x> is rel. Candidate #

MTHOME=/usr/mellanox

src_path=/usr/mellanox/src

prefix=/usr

KER_PATH=/lib/modules/2.4.21-20.ELsmp/build

KERNELRELEASE=2.4.21-20.ELsmp

## 2.2 Un-Installing the Driver

To un-install the driver, run uninstall.sh. The uninstall.sh is located under MTHOME as defined in /etc/vapi-linux-release.

## 2.3 Loading the HCA Driver

### 2.3.1 Load the VAPI driver

To load the driver use the modprobe utility:

- modprobe mod_thh
  for MT23108 InfiniHost (Dev ID: 23108, firmware: fw-23108), and MT25208 InfiniHost III Ex (Dev ID: 25208, firmware: fw-25208)
- modprobe mod_rhh
  for MT25208 InfiniHost III Ex (Dev ID: 25218, firmware: fw-25218)

The utility loads the modules: mlxsys, mod_vip, mod_Xhh (were X stand for t or r depending on the module loaded). It also opens the HCA with its ports in the DOWN state.

The output on the screen should be as in the following examples:

InfiniHost: hca_id[0] = InfiniHost0 - Opened

InfiniHost III Ex: hca_id[0] = InfiniHost_III_Ex0 - Opened

#### 2.3.1.1 Module Installation Options

When loading the module mod_thh or mod_rhh, the following options can be used:

1. Common options:
   - xhh_legacy_sqp=1: Special QPs are not enabled, and the HCA is working in internal SMA mode.
   - infinite_cmd_timeout=1: Sets an infinite timeout for the execution of a command interface command. (The default timeout setting is one minute.)
   - num_cmds_outs=N: Sets the number of commands that the HCA handle in parallel. The effective number of concurrently outstanding commands is the minimum between this value and the capability reported via QUERY_FW command.
   - async_eq_size=N; Sets the size of the asynchronous events EQ. The default value is 0x4000 entries
   - cmdif_use_uar0=0; Commands posted to the HCA using HCR. Default is to use UAR0.
2. mod_thh options
   - av_in_host_mem=1: Allocates the UD AVs in host memory.

3. mod_rhh options:
   - legacy_name=0: Use “InfiniHost” as the device name instead of “InfiniHost_III_Ex”.
   - max_kernel_avs=N: Sets the number of kernel UD AVs to N. The default is 16K.

**Passing the Options:**

- Kernel 2.4: The option can be given in the command line.
  Example: modprobe mod_thh cmdif_use_uar0=0
- Kernel 2.6: The option line should be added to the modprobe.conf.vapi file.
  Example: options mod_thh cmdif_use_uar0=0

### 2.3.2 Load IB_MGT module:

modprobe mod_ib_mgt
This loads the mod_ib_mgt module and brings the ports to the initialized state, if a link is connected to another active port.

## 2.4 Unloading the Driver

To download all modules use the utility modprobe -r.

### 2.4.1 Removing the IB_MGT driver

modprobe -r mod_ib_mgt will remove the mod_ib_mgt only.

Note:
The following warning may appear in some kernels, but they can be ignored:

FATAL: Module mod_vip is in use.
WARNING: Error running remove command for mod_vip
FATAL: Module mlxsys is in use.
WARNING: Error running remove command for mlxsys
FATAL: Module mod_vip is in use.
WARNING: Error running remove command for mod_vip
FATAL: Module mlxsys is in use.
WARNING: Error running remove command for mlxsys

### 2.4.2 Removing the VAPI driver

Removing the mod_Xhh module will remove the modules: mlxsys, mod_vip, mod_Xhh and will close the HCA.

Note: mod_ib_mgt should be removed before the removal of the mod_Xhh.

- modprobe -r mod_thh
  For InfiniHost (Dev ID: 23108) and InfiniHost III Ex (Dev ID: 25208):
- modprobe -r mod_rhh
  For InfiniHost III Ex (Dev ID: 25218)

Module removal may fail if ‘close HCA’ fails (due to open HCA resources).
Example message on the screen:
InfiniHost0 - HCA is still in use. HCA resources must be released before closing it.
FATAL: Error running remove command for mod_thh

## 2.5 Directory Structure and Contents

The installation procedure installs the following under MTHOME:

- uninstall.sh -shell script to un-install the driver
- BUILD_ID - Build version of the driver
- src - the sources directory; contains two directories:
  - vapi - the HCA driver and utilities
  - ib_mgt: The special QPs modules with minism utilities

### 2.5.1 VAPI Sources Structure

Vapi directory contains two directories: kernel and user

#### 2.5.1.1 Kernel Sources

The kernel sources are partitioned into the following directories:

- include - all headers needed by the driver and other kernel modules
- mlxsys - all general and system functions
- vip - the Verbs Interface Provider layer
- mlxhh - the hardware interface layer - includes common code and two sub-directories:
  - thh - for InfiniHost (Dev ID: 23108) and InfiniHost III Ex (Dev ID: 25208)
  - rhh - for InfiniHost III Ex (Dev ID: 25218)

**Compiling kernel sources -** under the src/vapi/kernel directory run:

1. ./make_clean
2. ./make_modules
3. ./make_modules_install

#### 2.5.1.2 User Sources

Most user sources of VAPI are shared with the VAPI kernel modules. Therefore, during compilation a link to the kernel sources is created. Only the user space specific sources are located under the vapi/user directory.

- include - all user level only headers needed by the driver and other kernel modules
- mlxsys - all general and system functions
- vip - the Verbs Interface Provider layer
- mlxhh - the hardware interface layer - includes common code and two sub-directories:
  - thh - for InfiniHost MT23108 and MT25208
  - rhh - for InfiniHost III Ex MT25218
- utils - user level utilities (e.g performance test, vping, etc.)

**Compiling user level sources -** under the src/vapi/user directory run:

1. make clean

2. make VAPI_KERNEL_SRC_DIR=<*vapi/kernel full path*> install

### 2.5.2 IB_MGT Sources Structure

The ib_mgt directory contains two sub-directories:

- driver - the ib_mgt kernel module sources and user space library.
- minism - the minism utility.

**Compiling ib_mgt sources:**

1. Install the vapi driver
2. Under the src/ib_mgt directory run:
   make clean
   make install

# 3 Limitations and Known Bugs

The next table list the bugs and limitations of the VAPI driver.

Table 2 - Recent Limitations and Known Bugs

| | Description | Details |
|---|---|---|
| 1. | When using the MSI-X kernel with older FW versions the driver loading fails | Once the MSI-X is enabled, the driver will always try to activate it. |
| 2. | fw-25208-rel-4.5.0 is not supported by this driver | Due to a bug in this FW (buffer fatal error). It is possible to use it with fw-25208-rel-4.5.3 |
| 3. | Memory windows binding is not thread safe | The application should take care of the correct synchronization. |
| 4. | Virtual memory registration of IO memory is not allowed for kernel memory | This means that the following sequence is not allowed in kernel space:<br>1. Allocating memory with EVAPI_alloc_map_devmem<br>2. Calling VAPI_register_mr() with VAPI_MPR |
| 5. | Fedora Core 3 operating system support limitation | 1. Fedora Core 3 running kernel 2.6.9 or later on an X86 platform requires applying a Fedora patch. See Section 1.2.1 on page 4.<br>2. Use *gcc* compiler version 3.4.3 or later. The 3.4.2 version included with Fedora Core 3 has a bug (#17581) that prevents correct operation of the driver. |

# 4 Fixed Bugs

| | Description | Details |
|---|---|---|
| 1. | Kernel stack overflow on Fedora core 3 | Fixed |
| 2. | The VAPI version macro was not correct | Fixed |
| 3. | Removed debug print of communication established event | Fixed |
| 4. | perf_main was not optimal for mem-free InfiniHost III Ex | Reduced the number of receive scatter gather size to the needed size. (By default, this was 20) |
| 5. | MSI is not activated on Opteron machines | Though the Opteron chipset does not support MSIX, the Linux kernel recognized MSI by mistake |
| 6. | Problems in compilation of user level x86_64 systems | Did not create the 32-bit library since it was not functional |
| 7. | Resize CQ fails on a sytem with multiple HCAs | Fixed |
| 8. | The kernel stack check was wrong | Fixed |