Sun™ Dual Port 4x DDR IB Host Channel Adapter PCIe® ExpressModule™ User Guide

Sun Microsystems, Inc.
www.sun.com

Part No. 820-3522-10
November 2007, Revision A

Submit comments about this document at: `http://www.sun.com/hwdocs/feedback`

Please
Recycle

Adobe PostScript

Contents

Figures

Tables

Preface

The *Sun Dual Port 4x DDR IB Host Channel Adapter PCIe ExpressModule User's Guide* provides installation instructions for the Sun™ Dual Port 4x DDR IB Host Channel Adapter PCIe ExpressModule. The guide provides an overview of the adapter, installation instructions and adapter specifications. The instructions in this guide are designed for system administrators with experience installing network hardware and software.

Note – The Sun Dual Port 4x DDR IB Host Channel Adapter PCIe ExpressModule is based on Mellanox Technologies' MT25408 ConnectX™ IB adapter (HCA) device.

How This Document Is Organized

Chapter 1 provides an overview of the ExpressModule.

Chapter 2 provides installation instructions for the ExpressModule.

Chapter 3 provides InfiniBand software overview.

Chapter 4 provides Internet Protocol Over InfiniBand configuration.

Chapter 5 provides information about updating firmware on a Linux system.

Chapter 6 provides interface, power, and memory information.

Appendix A provides specifications for the ExpressModule.

Appendix B provides an overview of diagnostics and troubleshooting.

Using UNIX Commands

This document might not contain information about basic UNIX® commands and procedures such as shutting down the system, booting the system, and configuring devices. Refer to the following for this information:

- Software documentation that you received with your system
- Solaris™ Operating System documentation, which is at:

  `http://docs.sun.com`

Shell Prompts

Shell	Prompt
C shell	*machine-name*`%`
C shell superuser	*machine-name*`#`
Bourne shell and Korn shell	`$`
Bourne shell and Korn shell superuser	`#`

Typographic Conventions

Typeface	Meaning	Examples
`AaBbCc123`	The names of commands, files, and directories; on-screen computer output	Edit your `.login` file. Use `ls -a` to list all files. `% You have mail.`
`AaBbCc123`	What you type, when contrasted with on-screen computer output	`%` **`su`** `Password:`
AaBbCc123	Book titles, new words or terms, words to be emphasized. Replace command-line variables with real names or values.	Read Chapter 6 in the *User's Guide*. These are called *class* options. You *must* be superuser to do this. To delete a file, type `rm` *filename*.

Note – Characters display differently depending on browser settings. If characters do not display correctly, change the character encoding in your browser to Unicode UTF-8.

Documentation, Support, and Training

Sun Function	URL
Documentation	http://www.sun.com/documentation/
Support	http://www.sun.com/support/
Training	http://www.sun.com/training/

Third-Party Web Sites

Sun is not responsible for the availability of third-party web sites mentioned in this document. Sun does not endorse and is not responsible or liable for any content, advertising, products, or other materials that are available on or through such sites or resources. Sun will not be responsible or liable for any actual or alleged damage or loss caused by or in connection with the use of or reliance on any such content, goods, or services that are available on or through such sites or resources.

Information	URL
PCI Express specifications	http://www.pcisig.com
Mellanox Technologies HCA information	http://www.mellanox.com
OFED software downloads	http://www.openfabrics.org/downloads.htm

Sun Welcomes Your Comments

Sun is interested in improving its documentation and welcomes your comments and suggestions. You can submit your comments by going to:

http://www.sun.com/documentation/feedback

Please include the title and part number of your document with your feedback:

Sun Dual Port 4x DDR IB Host Channel Adapter PCIe ExpressModule User's Guide, part number 820-3522-10.

CHAPTER 1

Overview

The main features of the Sun Dual Port 4x DDR IB Host Channel Adapter PCIe ExpressModule are:

- Compliant to the *PCI ExpressModule Electromechanical Specification, Revision 1.1*
- Compliant to the InfiniBand Architecture Specification, Release 1.2
- Two 20 Gb/s 4X InfiniBand (IB) copper ports for connecting InfiniBand traffic (4X IB connectors)
- Media detect circuit supporting external InfiniBand active copper and fiber solutions
- EU Restriction of Hazardous Substances (RoHS) compliant

Note – The PCI Express (or PCIe) interface is sometimes called PCI-E.

FIGURE 1-1 ExpressModule

Figure Legend Parts of the ExpressModule

1	Metal case
2	InfiniBand Port 1
3	InfiniBand Port 2
4	PCI Express connector

FIGURE 1-2 ExpressModule Without Metal Case

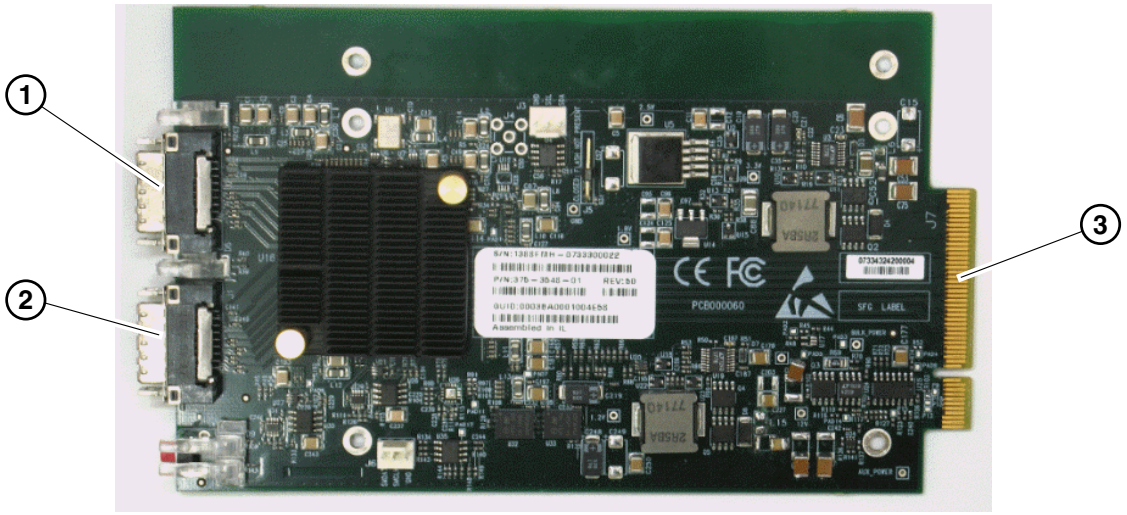

Figure Legend Connectors on the ExpressModule Without Metal Case

1	InfiniBand Port 1
2	InfiniBand Port 2
3	PCI Express connector

CHAPTER 2

Installing the Adapter

This chapter describes how to install the Sun Dual Port 4x DDR IB Host Channel Adapter PCIe ExpressModule and verify that it has been installed correctly. It contains the following sections:

Installing the Hardware

The following instructions describe the basic tasks required to install the ExpressModule. Refer to your system installation or service manual for detailed ExpressModule installation instructions.

▼ To Install the ExpressModule

1. **Halt and power off your system.**
2. **Install the ExpressModule into the ExpressModule slot, pushing the module's edge connector into the connector on the chassis.**

   Ensure that the front plate on the ExpressModule card mounts flush with the chassis panel opening.
3. **Connect the 4x end of the InfiniBand I/O cable(s) to the ExpressModule port connector(s).**

   Ensure that the connectors are properly engaged.

Caution – Avoid putting unnecessary stress on the connection. Do not bend or twist the cable near the connectors and avoid cable bends of more than 90 degrees.

4. **If not already connected, connect the InfiniBand I/O cable(s) to the appropriate ports on the InfiniBand switch.**
5. **Turn power back on to the system and allow the server to reboot.**

This completes the hardware installation. Proceed to the verification instructions in "To Verify the Installation" on page 6.

Verifying the Installation

Before you can verify the installation, you must install the ExpressModule in the chassis, power the server, and cable it to an operational InfiniBand switch. The InfiniBand switch should automatically recognize InfiniBand servers when they are connected to the fabric.

▼ To Verify the Installation

1. **Ensure that the cables are connected to the ExpressModule and switches.**
2. **Use the `lspci` command to verify that the IB Subnet Manager is running on the IB switch:**

```
> lspci
...
80:0e.0 PCI bridge: nVidia Corporation CK804 PCIE Bridge (rev a3)
81:00.0 InfiniBand: Mellanox Technologies MT25204 [InfiniHost III
Lx HCA] (rev 20)
```

   In this example, the IB ExpressModule (Mellanox InfiniHost III Lx HCA) is configured as PCI bus number 81, device 00, function 0 (`81:00.0`), which is EM slot 1 in a Sun Blade™ 8000 P Series Modular System. On your system, you might see a different designation for the EM.

3. **Check that the green LED is illuminated for each port that is connected to the switch.**

   If the green LED is not on, check the cable connections at the ExpressModule and at the switch.

4. **Check that the amber LED is illuminated for each port that is connected to the switch.**

   If the amber LED is not on, refer to Appendix B, "Diagnostic Software and Troubleshooting" on page 41 for more information.

5. **Verify that the IB EM ports are up and the driver is attached, by typing:**

```
> ibstat
```

   The state of each port is displayed:

State	Comments
`Active`	The port is working.
`Down`	The port is not working or the cable is not connected. If the cable is connected, see Appendix B for troubleshooting information.
`Initializing`	If the initialization does not finish, the Subnet Manager might not be running for the subnet. If restarting the Subnet Manager has no effect, see Appendix B for troubleshooting information.

CHAPTER 3

InfiniBand Software Overview

InfiniBand is a network architecture that is designed for the large-scale interconnection of computing and I/O nodes through a high-speed switched fabric. To operate InfiniBand on a Sun Blade 8000 Series Modular System, you need an InfiniBand HCA (the ExpressModule) and an InfiniBand software stack.

This chapter provides an overview and installation instructions for the InfiniBand software stack for the Linux operating system.

Consult the *Sun Blade 8000 Series Product Notes* for the most recent information about supported operating systems, firmware and software updates, and other issues not covered in the main product documentation.

InfiniBand Software for Linux

If you have installed current releases of Red Hat Enterprise Linux Advanced Server (RHEL AS 4-U3 or later) or SUSE Linux Enterprise Server (SLES9 SP3 or later, SLES10) on a Sun Blade Server Module and you have installed the bundled drivers and OFED Release 1.2.5 or later, you do not need to install or configure additional drivers to support the IB ExpressModule (IB EM).

Specifically, RHEL AS 4-U4 contains support in the kernel for HCA hardware produced by Mellanox (`mthca` driver). The kernel also includes core InfiniBand modules, which provide the interface between the lower-level hardware driver and the upper-layer InfiniBand protocol drivers. The InfiniBand modules provide user space access to InfiniBand hardware.

The kernel also includes the Sockets Direct Protocol (SDP) driver, IP over Infiniband (IPoIB) and the SCSI RDMA Protocol (SRP) driver.

RHEL AS 4-U4 includes the following user space packages:

- `kernel-ib`. Base package that is required to support all other packages. Includes the files necessary to configure the kernel portion of the `openib` stack, create the proper `udev` rules, add the `init` script that allows the kernel modules to be selectively loaded at boot, and so on.
- `dapl`. RDMA API that supports the DAT 1.2 specification.
- `ibibcm`. InfiniBand Connection Management API.
- `libibcommon`. Common utility functions for the IB diagnostic and management tools.
- `libibmad`. Low-layer IB functions for use by the IB diagnostic and management programs, including MAD, SA, SMP, and other basic IB functions.
- `libibumad`. User MAD library functions that sit on top of the user MAD modules in the kernel. Used by the IB diagnostic and management tools, including OpenSM.
- `libibverbs`. Library that allows user space processes to use InfiniBand "verbs" as described in the InfiniBand Architecture Specification.
- `libibverbs-utils`. Useful subnet and device diagnostic utilities.
- `libmthca`. Device-specific user space driver for Mellanox HCAs (MT23108 InfiniHost and MT25208 InfiniHost III Ex) for use with the `libibverbs` library.
- `libipathverbs`. Device-specific driver for Pathscale HCAs for use with `libibverbs` (only available on x86_64 and ia64 systems).
- `librdmacm`. RDMA Connection Management (`cm`) library.
- `libsdp`. Driver that enables a sockets application to use InfiniBand Sockets Direct Protocol (SDP) instead of TCP transparently and without recompiling the application.
- `openib-diags`. Diagnostic programs and scripts that diagnose the IB subnet.
- `opensm`. Subnet manager software for InfiniBand networks.
- `opensm-libs`. Shared libraries for InfiniBand user space access.
- `perftest`. InfiniBand performance tests.
- `srptools`. In conjunction with the kernel `ib_srp` driver, allows discovery and and use of SCSI class devices via the SCSI RDMA Protocol over InfiniBand.
- `mstflint`. Tool to query and update firmware flash memory attached to Mellanox InfiniBand HCAs.

Note – These package names can change, depending on the Linux OS.

The packages selected to support any given configuration will vary. TABLE 3-1 lists the packages considered the absolute minimum needed to support the environment described in this guide.

TABLE 3-1 Required Packages for InfiniBand Support

Package	Command Enabled	Description
`kernel-ib`	`openibd`	IB master control script
`openib-diags`	`ibstat`	IB utility to display HCAs
`openib-diags`	`ibnetdiscover`	IB utility to probe and show the fabric
`mstflint`	`mstflint`	Mellanox utility to update HCA FLASHRAM
`libibcommon`	NA	IB support package
`libibmad`	NA	IB support package
`libibumad`	NA	IB support package
OFED Release 1.2.5 or later	NA	IB support package

If you elected not to install these packages when installing the Linux OS or if you want to upgrade your drivers, you can install these packages at any time from the OS distribution source or by downloading the required files from OpenFabrics.org. For information on both of these procedures, see "Installing the InfiniBand Drivers on Linux" on page 12.

OpenFabrics Enterprise Distribution for Linux

As the popularity of InfiniBand technology increases, the number of Linux distributions and open source organizations producing drivers and tools will increase. For up-to-date information, check with open source organizations and your current vendors.

The OpenFabrics organization is the Open Software solution in the InfiniBand software space and OpenFabrics Enterprise Distribution (OFED) is the InfiniBand suite of software produced by this organization. Various vendors contribute their drivers (and other software components) to OFED.

TABLE 3-2 lists the tested Linux platforms and the corresponding OFED release.

TABLE 3-2 Linux Platforms and OFED Release

Linux Platform	OFED Release
RHEL AS 4-U3 or later	For RHEL AS 4-U3, Sun has tested OFED Release 1.2.5 of the OpenFabrics stack. Note: RHEL AS 4-U4 includes an older version of OFED, so you must install OFED Release 1.2.5 or a later version.
SLES9 SP3 or later, SLES10	Sun has tested OFED Release 1.2.5 for the SLES10 platform. Note: You must have OFED Release 1.2.5 or a later version.

OFED contains the following components:

- OpenFabrics core and Upper Layer Protocols (ULPs):
  - HCA drivers
  - Core
  - ULPs, including: IPoIB, SDP, SRP Initiator, iSER Host, RDS and uDAPL
- OpenFabrics utilities:
  - OpenSM (InfiniBand subnet manager)
  - Diagnostic tools
  - Performance tests
- Message Passing Interface (MPI):
  - OSU MPI stack supporting the InfiniBand interface
  - Open MPI stack supporting the InfiniBand interface
  - MPI benchmark tests (OSU BW/LAT, Pallas, Presta)
- Sources of all software modules (under conditions mentioned in the modules' LICENSE files)
- Documentation

Installing the InfiniBand Drivers on Linux

If you did not install the InfiniBand drivers when installing the Linux OS, you can install them at any time from the OS distribution source or by downloading the necessary files from OpenFabrics.org.

To do so, choose one of the following procedures:

- "To Install IB Drivers From Linux Distribution Source" on page 13
- "To Install the OFED Package" on page 14

If you need to determine whether or not the drivers are already installed, see "To Verify Driver Installation on Linux" on page 20.

▼ To Install IB Drivers From Linux Distribution Source

To install the InfiniBand drivers, you need access to the Red Hat Package Manager (RPM) files. Access to these files is dependent on your individual installation configuration (net boot, CD/DVD boot, .iso files, and so on). When you decide on the appropriate access method and package selection, you can add the packages to the KickStart configuration file (on RHEL) for automatic inclusion in future installations.

Note – On a 32-bit RHEL4 system, all packages have a `.i386.rpm` extension (as shown in the following procedure). On a 64-bit RHEL4 system, all packages have a `.x86_64.rpm` extension instead.

---

1. **Enter the `rpm -ivh` command for each InfiniBand package that you need to install.**

   Packages must be installed in the following order:

   - `libibcommon`
   - `libibumad`
   - `libibmad`
   - `openib-diags`
   - `mstflint`
   - `perftest`

   The following example shows the installation of one package (`libibcommon`) and the resulting dialog on an RHEL AS 4-U4 32-bit system:

```
 > rpm -ivh libibcommon-1.0-1.i386.rpm
 warning: libibcommon-1.0-1.i386.rpm: V3 DSA signature: NOKEY, key
ID db42a60e
 Preparing...      ########################################## [100%]
1:libibcommon  ########################################### [100%]
> rpm -ivh libibumad-1.0-1.i386.rpm
.
.
.
```

2. **If you are running the CSH or TCSH shell, enter the `rehash` command to rebuild the shell's view of available executables.**

3. **Enter the `ibstat` command to verify that the OS sees the IB em.**

```
> ibstat
CA 'mthca0'
     CA type: MT25204
     Number of ports: 1
     Firmware version: 1.1.0
     Hardware version: a0
     Node GUID: 0x001b00000ca72640
     System image GUID: 0x001b00000ca72643
     Port 1
         State: Active
         Physical state: LinkUp
         Rate: 20
         Base lid: 71
         LMC: 0
         SM lid: 2
         Capability mask: 0x02510a68
         Port GUID: 0x001b00000ca72641
```

4. **(Optional) You can enter the `ibnetdiscover` command to verify the presence of an operational IB fabric.**

   For an example of the output of this command, see "To Verify Driver Installation on Linux" on page 20.

5. **(Optional) You can check the status of the `ib0` network interface to determine whether the `ib_ipoib` driver is installed.**

   For details on this step, see "To Install IPoIB Driver" on page 24.

▼ To Install the OFED Package

Note – The Sun Dual Port 4x DDR IB Host Channel Adapter PCIe ExpressModule requires OFED Release 1.2.5 or later.

---

1. **On the Sun Blade Server Module, log in as `root` and copy the required files from the following location:**

   `http://www.openfabrics.org/downloads.htm`

   In the following example, `OFED-1.2.5.tar` is used only as an example.

Note – You need Write access to the files to execute the install script.

2. **From root, extract the files by typing:**

```
> tar -zxvf OFED-1.2.5.tar
```

3. **From the OFED-1.2.5 directory, initiate the installation process by typing:**

```
> ./install.sh
```

4. **When the InfiniBand OFED Distribution Software Installation Menu appears, select option 2 (Install OFED Software).**
5. **When the Select OFED Software menu appears, select option 3 (All packages).**
6. **When you are asked if you wish to create/install an MPI RPM with gcc, enter** n**.**

```
The following compiler(s) on your system can be used to
build/install MPI:  gcc
Do you wish to create/install an MPI RPM with gcc? [Y/n]: n
```

7. **Next, you are asked if you wish to create/install an openmpi RPM with gcc. Again, type** n**.**

```
The following compiler(s) on your system can be used to
build/install openmpi:  gcc
Do you wish to create/install an openmpi RPM with gcc? [Y/n]: n
```

The installation script lists the OFED packages that it will build. See the following sample output.

```
Following is the list of OFED packages that you have chosen (some
may have been added by the installation program due to package
dependencies):
ib_ipath
ib_ipoib
...
mpitests
ibutils

WARNING: This installation program will remove any previously
installed IB packages on your machine.

Do you want to continue? [Y/n]: Y
```

8. **Type `Y` to continue, as shown above.**

   Next, you are prompted to configure InfiniBand IP support.

9. **Type `Y` when asked if you want to include IPoIB configuration files.**

```
Do you want to include IPoIB configuration files (ifcfg-ib*)? [Y/n]: Y
```

10. **Press Enter to accept the default when prompted to enter a temporary directory for OFED.**

```
RPM build process requires a temporary directory.
Please enter the temporary directory [/var/tmp/OFED]:
```

11. **Press Enter to accept the default when prompted for the OFED installation directory.**

```
Please enter the OFED installation directory [/usr/local/ofed]:
```

At this point, the installer begins compiling InfiniBand packages. The process of building packages takes approximately 15–20 minutes.

The system displays output like the following:

```
The MPI_COMPILER_openmpi variable is not defined. Trying the
default compiler: gcc

The following compiler(s) will be used to build the openmpi RPM(s):
gcc

Checking dependencies. Please wait ...

Building InfiniBand Software RPMs. Please wait...

Building openib RPMs. Please wait...
.
.
.
33 packages were built

Build process finished ...
```

Installation begins. See the following message.

```
Removing previous InfiniBand Software installation
Running /bin/rpm -e libibverbs libibverbs-devel libibverbs-
utils...
```

The actual installation takes about one minute.

Assuming the IB EM hardware is installed (and, therefore, an InfiniBand HCA is present), you are prompted to configure InfiniBand IP support.

12. **Enter `Y` in response to the following prompt:**

```
Do you want to configure IPoIB interfaces [Y/n]? Y
```

The default IPoIB interface configuration is based on DHCP. A special patch for DHCP is required for supporting IPoIB. The patch is available under:

OFED-1.2.5/docs/dhcp

If you do not have DHCP, you must change this configuration in the following steps.

The system next displays the current configuration.

13. **When asked if you want to change the configuration as displayed, type** `y`**.**

```
The current IPOIB configuration for ib0 is:
DEVICE=ib0
BOOTPROTO=dhcp
ONBOOT=yes
Do you want to change this configuration? [y/N]: Y
```

The configuration script guides you through the changes one at a time. See the following as an example.

```
Enter an IP Address:10.0.0.52
Enter the Netmask: 255.255.255.0
Enter the Network:10.0.0.0
Enter the Broadcast Address:10.0.0.255
Start Device On Boot? [Y/n]:Y

Selected configuration:

IPADDR=10.0.0.52
NETMASK=255.255.255.0
NETWORK=10.0.0.0
BROADCAST=10.0.0.255
ONBOOT=yes

Do you want to save the selected configuration? [Y/n]: Y
```

14. **Type** `Y` **to save the configuration.**

If you have entered a valid IP configuration for `ib0`, you are now properly configured for IPoIB operations.

15. **Iterate the InfiniBand configuration over all InfiniBand interfaces.**

Enter a valid IP configuration for each network interface.

Once all IPoIB interfaces have been configured, you are prompted as follows to configure OpenSM for the blade.

```
Do you want to configure OpenSM [Y/n]? n
```

16. **Enter `n` to complete this part of the installation.**

    You should see a message like the following.

```
Installation finished successfully...
Press Enter to continue...
```

17. **Press Enter.**

    The InfiniBand OFED Distribution Software Installation Menu is displayed.

18. **Type `Q` to exit.**

The Sun Blade Server Module is configured now to start up the InfiniBand software on reboot (`ONBOOT=yes`).

If this is not the desired behavior, you can edit the `/etc/infiniband/openib.conf` file, changing `ONBOOT` to equal `no`. You can also manually control basic InfiniBand behavior by entering the following command:

```
/etc/init.d/openibd option
```

where *option* can be `start`, `stop`, or `status`.

19. **After a successful installation, reboot the Server Module.**

    After the reboot, the Server Module should come up as a functional member of the InfiniBand fabric.

▼ To Verify Driver Installation on Linux

1. **Verify that the Linux software driver is installed and attached to the IB EM by typing the** `openibd status` **command.**

Note – When using the `openibd` command, type the entire path as shown in the example.

The following example shows the IB driver installed, running and presenting one IB HCA channel or network device (`ib`*n*) to the OS. In the example, the Linux network device appears as `ib0`.

```
> /etc/init.d/openibd status
     HCA driver loaded
Configured devices:
ib0
Currently active devices:
ib0
     The following modules are also loaded:
ib_cm
ip_ipoib
.
.
.
```

2. **To view details of operational status, type the `ibstat` command.**

   The following example shows one operational IB channel into the IB fabric (or network). The `LinkUp` state indicates active participation in an IB fabric. It is present as `lid 69` and it is being managed by `lid 2`.

```
> ibstat
CA 'mthca0'
     CA type: MT25204
     Number of ports: 1
     Firmware version: 1.1.0
     Hardware version: a0
     Node GUID: 0x001b00000ca72620
     System image GUID: 0x001b00000ca72623
     Port 1
         State: Active
         Physical state: LinkUp
         Rate: 20
         Base lid: 69
         LMC: 0
         SM lid: 2
         Capability mask: 0x02510a68
         Port GUID: 0x001b00000ca72621
```

   You can also verify that the InfiniBand fabric is operational by entering the `ibnetdiscover` command. The output from this command will list all the nodes, as shown in the following sample output.

```
> ibnetdiscover
#
# Topology file: generated on Thu Jan 11 15:19:59 2007
#
# Max of 4 hops discovered
# Initiated from node 001b00000ca72620 port 001b00000ca72621

vendid=0x8f1
devid=0x5a31
sysimgguid=0x8f10400411ef9
switchguid=0x8f10400411ef8

Switch  24 "S-0008f10400411ef8"    # Switch port 0 lid 9
[21]       "H-0002c90109761ea0"[2]
[12]       "S-0005ad00000161ba"[5]
[7]        "H-001b00000ca72630"[1]
[6]        "H-001b00000ca72620"[1]
vendid=0x5ad
devid=0xa87c
sysimgguid=0x5ad01010161b6
switchguid=0x5ad00000161ba
```

```
Switch  8 "S-0005ad00000161ba"    # Switch - U3 port 0 lid 3
[4]"       H-0005ad0000011310"[1]
[3]       "S-0005ad00000161b6"[1]
[2]       "S-0005ad00000161b6"[2]
[1]       "S-0005ad00000161b8"[3]
[5]       "S-0008f10400411ef8"[12]
.
.
.
vendid=0x2c9
devid=0x6274
sysimgguid=0x1b00000ca72633
caguid=0x1b00000ca72630
Ca  1 "H-001b00000ca72630"  # 4x DDR IB 10-Port PCIe Network
Express Module
[1]     "S-0008f10400411ef8"[7]     # lid 68 lmc 0n
```

CHAPTER 4

Internet Protocol Over InfiniBand

This chapter describes configuration aspects of running the Internet Protocol over InfiniBand (IPoIB).

Configuring IPoIB on Linux

You might decide to change your IPoIB configuration for a variety of reasons, including the installation of an additional InfiniBand ExpressModule (IB EM).

▼ To Install IPoIB Driver

1. **Detemine whether the IPoIB driver is already installed by typing the** `lsmod | grep ib` **command.**

   The output from this command shows all the IB drivers.

   In the following sample output, note that the driver, `ib_ipoib`, is not listed.

```
> lsmod | grep ib
ib_sdp        45340 0
rdma_cm       26760 1 ib_sdp
ib_addr       10504 1 rdma_cm
ib_local_sa   14232 1 rdma_cm
findex        6528 1 ib_local_sa
ib_ipath      70552 0
ipath_core    179652 1 ib_ipath
ib_mthca      139184 0
ib_uverbs     47536 0
ib_umad       19888 0
ib_ucm        21512 0
ib_sa         18196 2 rdma_cm,ib_local_sa
ib_cm         39952 2 rdma_cm,ib_ucm
ib_mad        43176 5 ib_local_sa,ib_mthca,ib_umad,ib_sa,ib_cm
ib_core       59520 11
ib_sdp,rdma_cm,ib_local_sa,ib_ipath,ib_mthca,ib_uverbs,ib_umad,
ib_ucm,ib_sa,ib_cm,ib_mad
```

2. **To install the IPoIB driver, enter the** `modprobe` **command:**

```
> modprobe ib_ipoib
```

3. **Enter the** `lsmod | grep ib` **command again and note that** `ib_ipoib` **is now listed.**

```
> lsmod | grep ib
ib_ipoib      59800 0
ib_sdp        45340 0
rdma_cm       26760 1 ib_sdp
ib_addr       10504 1 rdma_cm
ib_local_sa   14232 1 rdma_cm
.
.
.
ib_core       59520 11
ib_sdp,rdma_cm,ib_local_sa,ib_ipath,ib_mthca,ib_uverbs,ib_umad,
ib_ucm,ib_sa,ib_cm,ib_mad
```

4. **Type the `ifconfig` command to check for network interface `ib0`.**

```
> ifconfig ib0
ib0     Link encap:UNSPEC  HWaddr 00-00-00-00-00-00-00-00-00-00-
00-00-00-00-00-00
        BROADCAST MULTICAST  MTU:2044  Metric:1
        RX packets:0 errors:0 dropped:0 overruns:0 frame:0
        TX packets:0 errors:0 dropped:0 overruns:0 carrier:0
        collisions:0 txqueuelen:128
        RX bytes:0 (0.0 b)  TX bytes:0 (0.0 b)
```

Note that network interface `ib0` is present but has no valid IP address.

To assign an address, see "To Change IPoIB Configuration Without Rebooting" on page 25.

▼ To Change IPoIB Configuration Without Rebooting

Caution – The changes made by this procedure are only temporary. These changes are lost during a reboot. To make permanent changes, see "To Change IB Startup Behavior on Linux" on page 27.

1. **Start the InfiniBand IP network by typing the `ifconfig` command and assigning a valid IP address for `ib0`.**

```
> ifconfig ib0 10.0.0.50/24
ib0     Link encap:UNSPEC  HWaddr 00-00-00-00-00-00-00-00-00-00-
00-00-00-00-00-00
        inet
addr:10.0.0.50  Bcast: 10.0.0.255  Mask:255.255.255.0
        UP BROADCAST RUNNING MULTICAST  MTU:2044  Metric:1
        RX packets:0 errors:0 dropped:0 overruns:0 frame:0
        TX packets:0 errors:0 dropped:0 overruns:0 carrier:0
        collisions:0 txqueuelen:128
        RX bytes:0 (0.0 b)  TX bytes:0 (0.0 b)
```

If ib0 is unconfigured, it appears without an IP address as shown in the following output.

```
> ifconfig ib0
ib0     Link encap:UNSPEC  HWaddr 00-00-00-00-00-00-00-00-00-00-
00-00-00-00-00-00
        BROADCAST MULTICAST  MTU:2044  Metric:1
        RX packets:0 errors:0 dropped:0 overruns:0 frame:0
        TX packets:0 errors:0 dropped:0 overruns:0 carrier:0
        collisions:0 txqueuelen:128
        RX bytes:0 (0.0 b)  TX bytes:0 ( 0.0 b)
```

2. **Type the** route **command to verify that the** 10.0.0 **subnet is available.**

   The following output shows subnet 10.0.0 present and routed through ib0.

```
> route
Kernel IP routing table
Destination Gateway    Genmask         Flags  Metric Ref  Use  Ifa
ce
10.0.0.0    *          255.255.255.0   U      0      0    0    ib0
10.8.134.0  *          255.255.255.0          0      0    0    eth
0
169.254.0.0 *          255.255.0.0     U      0      0    0    eth
0
default     ban3rtr0d0 0.0.0.0         UG     0      0    0    eth
0
```

3. **As shown in the following example, you can enter the** ping **command to see another IPoIB node on the** 10.0.0 **subnet:**

```
> ping 10.0.0.79
PING 10.0.0.79 (10.0.0.79) 56(84) bytes of data.
64 bytes from 10.0.0.79: icmp_seq=0 ttl=255 time=1.82 ms
64 bytes from 10.0.0.79: icmp_seq=1 ttl=255 time=0.082 ms
64 bytes from 10.0.0.79: icmp_seq=2 ttl=255 time= 0.062 ms

--- 10.0.0.79 ping statistics ---
3 packets transmitted, 3 received, 0% packet loss, time 2001ms
rtt min/avg/max/mdev = 0.062/0.655/1.823/0.825 ms, pipe 2
```

At this point, the IPoIB network is active and properly configured without rebooting.

▼ To Change IB Startup Behavior on Linux

You can change the InfiniBand startup behavior by editing the `openib.conf` configuration file. (The full path is `/etc/infiniband/openib.conf`).

In the following example, `openib.conf` specifies that whenever the system boots, the InfiniBand services, IPoIB, and the SDP IP service are to start up automatically (`ONBOOT=yes`, `IPOIB_LOAD=yes`, `SDP_LOAD=yes`). However, `openib.conf` specifies that the SRP service is NOT to start up automatically (`SRP_LOAD=no`). You can alter any and all of these parameters.

1. **Edit `ipoib.conf`.**

   The following is an example of the `ipoib.conf` file.

```
# Start HCA driver upon boot
ONBOOT=yes
# Load UCM module
UCM_LOAD=no
# Load RDMA_CM module
RDMA_CM_LOAD=no
# Load RDMA_UCM module
RDMA_UCM_LOAD=no
# Load MTHCA
MTHCA_LOAD=yes
# Load IPATH
IPATH_LOAD=yes
# Load IPoIB
IPOIB_LOAD=yes
# Load SDP module
SDP_LOAD=yes
# Load SRP module
SRP_LOAD=no
# Load RDS module
RDS_LOAD=no
```

2. **Create (or edit) the `ifcfg-ib`*n* file to configure an individual network interface.**

   For each InfiniBand network interface, you will need a corresponding startup file (`ifcfg-ib`*n*) in your startup scripts directory.

   The directory for startup scripts might be `/etc/sysconfig/network/`, `/etc/sysconfig/network-scripts/`, or similar, depending on your version of Linux distribution.

   As an example, the startup file for `ib0`, might look something like the following.

```
more /etc/sysconfig/network-scripts/ifcfg-ib0
 DEVICE=ib0
 BOOTPROTO=static
 IPADDR=10.0.0.50
 NETMASK=255.255.255.0
 NETWORK=10.0.0.0
 BROADCAST=10.0.0.255
 ONBOOT=yes
```

   The `ONBOOT=yes` parameter indicates that the corresponding IP network interface is to automatically start up when the system boots. Specifying `ONBOOT=no` will configure the interface but not start it.

   Generally, if you have enabled IPoIB services, the IB stack installation scripts automatically create the `ifcfg-ibn` configuration files for all IB network interfaces present. If you install an IB EM ***after*** you have installed the IB stack, you might need to manually create the `ifcfg-ib`*n* files for the newly installed network interfaces.

CHAPTER 5

Updating IB EM Firmware

This chapter provides information on updating the InfiniBand ExpressModule (IB EM) firmware on Linux.

Consult the *Sun Blade 8000 Series Product Notes* for the most recent information about the availability of firmware updates.

Updating IB EM Firmware for Linux

The firmware version on your IB EM should be ready to use and should not require updating. However, if you want to update the firmware, you must use vendor-specific and (generally) OS-specific firmware updating tools.

For Linux (RHEL AS 4-U3 or later, SLES9 SP3 or later, and SLES10), use the OFED `mstflint` tool to load new IB EM firmware. The tool, `mstflint`, is available both as part of the bundled software and from the standard OFED stack.

Installed by default, `mstflint` is similar to the Mellanox `flint` tool with the following exception: you must identify the IB EM in the PCI *bus*:*dev*.*fun* format to satisfy the `mstflint` command `-d` *device* syntax requirement.

▼ To Update IB EM Firmware for Linux

1. **Enter the** `lspci` **command to identify the IB EM.**

```
> lspci
...
80:0e.0 PCI bridge: nVidia Corporation CK804 PCIE Bridge (rev a3)
81:00.0 InfiniBand: Mellanox Technologies MT25204 [InfiniHost III
Lx HCA] (rev 20)
```

In this example, the IB EM (Mellanox InfiniHost III Lx HCA) is configured as PCI bus number 81, device 00, function 0 (`81:00.0`), which is EM slot 1 in a Sun Blade 8000 P Series Modular System. On your system, you might see a different designation for the EM.

2. **Enter the** `mstflint -d` **command in the** *bus*`:`*dev*`.`*fun* **format.**

```
> mstflint -d 81:00.0 -i ibem.bin  burn
Current FW version on flash:  1.1.0
New FW version:               1.2.100

Burn image with the following GUIDs:
  Current GUIDs are:
       Node:      001b00000ca72680
       Port1:     001b00000ca72681
       Sys.Image: 01b00000ca72683

Read and verify Invariant Sector - OK
Read and verify PPS/SPS on flash - OK
Burning second FW image without signatures - OK
Restoring second signature -OK

Do you want to continue ? (y/n) [n] : y
```

Note – The GUIDs that you will see (Node, Port1, and Sys. Image) during the burn process will differ from those shown in the example.

---

3. **To burn the image, type** `y`**.**

   As with any IB EM FLASHRAM update, you must reset the Server Module (or at least the IB EM) to load and execute the new firmware image.

4. **After resetting the Server Module (or the EM), enter the `ibstat` command to verify the new firmware version.**

```
> ibstat
CA 'mthca0'
CA type: MT25204
     Number of ports: 1
     Firmware version: 1.1.0
     Hardware version: a0
     Node GUID: 0x001b00000ca72600
     System image GUID: 0x001b00000ca72603
     Port 1:
          State: Active
          Physical state: LinkUp
          Rate: 20
          Base lid: 70
          LMC: 0
          SM lid: 2
          Capability mask: 0x02510a68
          Port GUID: 0x001b00000ca72601
```

CHAPTER 6

Interfaces

This chapter describes the following ExpressModule interfaces:

I/O Interfaces

The ExpressModule includes the following interfaces:

- Two 4x InfiniBand copper connectors
- PCI ExpressModule to Server/Workstation connector
- I/O panel LEDs

InfiniBand Interface

The Sun Dual Port 4x DDR IB Host Channel Adapter PCIe ExpressModule is compliant with the *InfiniBand Architecture Specification, Release 1.2.* It has two compliant 4x InfiniBand ports (Port 1 and Port 2); each having four Tx/Rx pairs of SerDes. The ExpressModule provides access to these ports by means of two 4x InfiniBand connectors for external InfiniBand copper cables also compliant with the IBTA specification 1.2. Connector 1 connects to Port 1 of the device, while connector 2 connects to Port 2.

Furthermore, the ExpressModule is embedded with a *media detect circuit* that supports external InfiniBand active copper and fiber solutions.

PCI ExpressModule Interface

The ExpressModule is compatible with the *PCI ExpressModule Electromechanical Specification, Revision 1.1*. The device can be either a master initiating the PCI Express bus operations, or a slave responding to PCI bus operations.

LED Assignment

The ExpressModule has four LEDs located on the InfiniBand ports panel, with two LEDs for each 4x port – see FIGURE 6-1.

- The Physical Link (green) illuminates once VAPI (InfiniBand Verbs API) is started and a physical connection is made between two nodes.
- The Data Activity Link (yellow) illuminates once the InfiniBand network is discovered over the physical link.
- The Activity Link is a steady yellow when the network is discovered but no data is being passed. The activity link blinks when data is being passed.
- If the LEDs are not active, either the physical or the logical (or both) connections have not been established.

FIGURE 6-1 InfiniBand Ports and LEDs

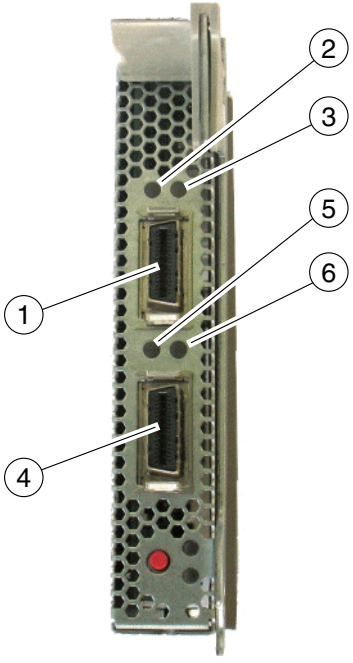

Figure Legend Ports and LEDs on the Infiniband Ports Panel

1	Port 1
2	Physical Link - Green
3	Data Activity - Yellow
4	Port2
5	Physical Link - Green
6	Data Activity - Yellow

Power

The ExpressModule receives power from the Server/Workstation slot connector. All other required power is generated by ExpressModule switch-mode regulators. For power consumption see Appendix A.

Memory

The ExpressModule supports multiple memory devices through the PCI Express, Flash, and I^2C-compatible interfaces.

The ExpressModule utilizes the PCI Express interface to store and access IB fabric connection information on the system memory.

APPENDIX **A**

Specifications

This appendix contains the following sections:

ExpressModule Specifications

TABLE A-1 Board Specifications

Specification	Description	
Physical	Size	198mm X 112mm (7.8in. x 4.4in.)
	Air Flow	200LFM @ 55C
	4x 20Gb/s Connector	InfiniBand (Copper, current rating: 0.5A max) with active media adapter support
Power and environmental	Voltage	12V, 3.3V
	Max Power	12W
	Temperature	0 to 55 Celsius
Protocol support	IBTA v1.2	Auto-Negotiation* (20Gb/s, 5Gb/s) or (10Gb/s, 2.5Gb/s)
	QoS	8 InfiniBand Virtual Lanes for each port
	RDMA Support	Yes, all ports

* The auto-negotiation protocol is proprietary of Mellanox Technologies and compliant with the *InfiniBand Architecture Specification, Release 1.2*.

Mechanical Drawing and Dimensions

FIGURE A-1 shows miniature size photographs of the ExpressModule with and without the metal case.

FIGURE A-1 ExpressModule Miniature Photos

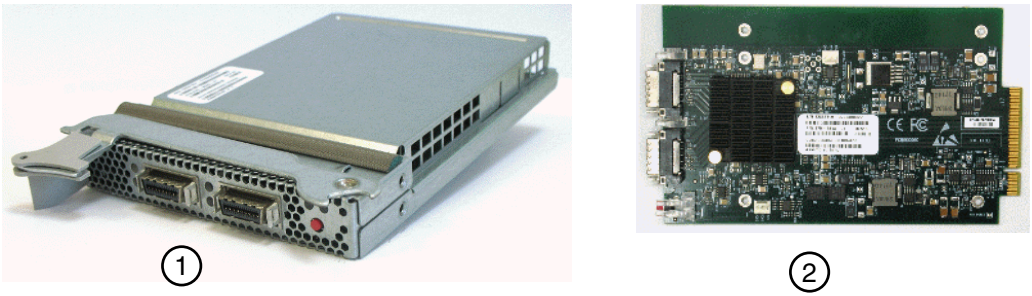

Figure Legend Express Module

1	With Metal Case
2	Without Metal Case

FIGURE A-2 shows the mechanical drawing and dimensions of the ExpressModule board without the metal case.

FIGURE A-2 Mechanical Drawing of ExpressModule Board

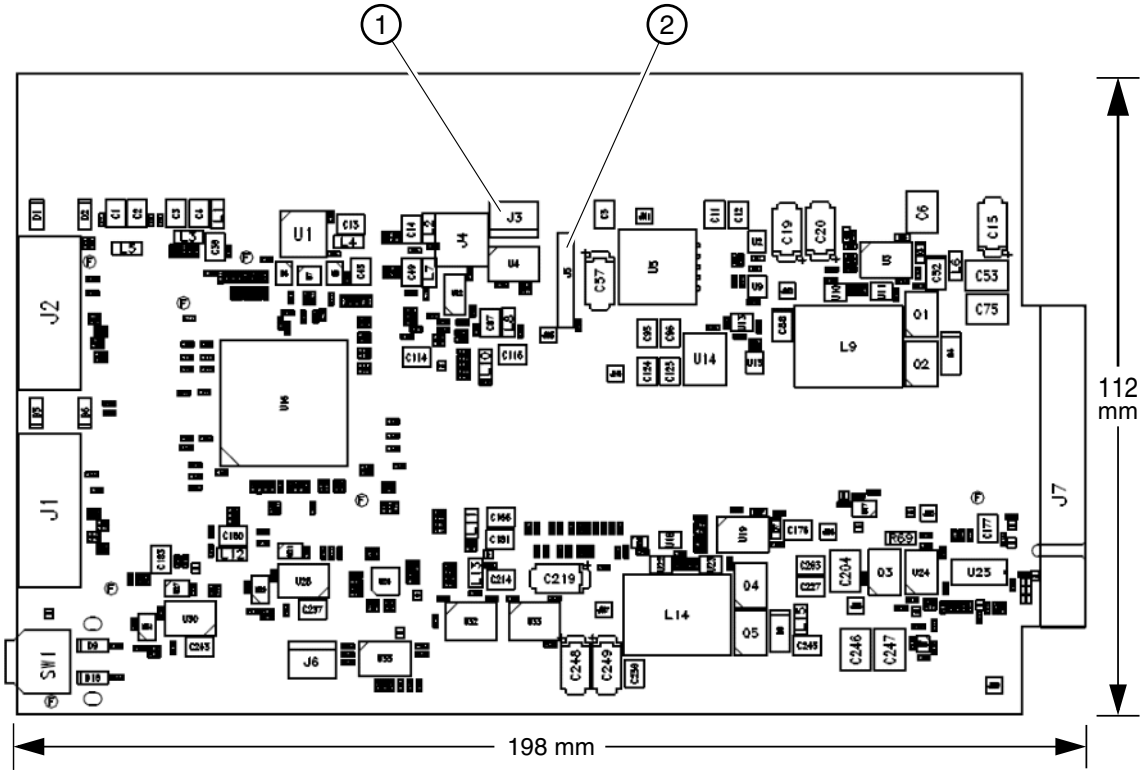

Figure Legend

1	Jumper J3 - I^2C Header
2	Jumper J5 - Flash Present/Not-Present

APPENDIX **B**

Diagnostic Software and Troubleshooting

This appendix provides an overview of the SunVTS™ diagnostic application and troubleshooting tips for checking the adapter. There is also a section outlining some common troubleshooting issues. This appendix contains the following sections:

SunVTS Diagnostic Testing

The SunVTS software executes multiple diagnostic hardware tests from a single user interface, and is used to verify the configuration and functionality of most hardware controllers and devices. The SunVTS software operates primarily from a graphical user interface, enabling test parameters to be set quickly and easily while a diagnostic test operation is being performed.

Note – SunVTS diagnostic software is not currently available for Solaris x86 Operating Systems.

Refer to the SunVTS documents (listed in TABLE B-1) for instructions on how to run and monitor the `nettest` diagnostic. These SunVTS documents are available online at the following URL:

`http://docs.sun.com/app/docs/prod/test.validate/sunvts/index.html`

Select the document for the Solaris release on your system

TABLE B-1 SunVTS Documentation

Title	Descriptions
SunVTS 6.0 PS1 Documentation Supplement (819-1804)	Describes the new SunVTS features and tests, including the `ibhcatest`.
SunVTS 6.0 User's Guide (817-7664)	Describes the SunVTS diagnostic environment.
SunVTS 6.0 Test Reference Manual (817-7665)	Describes each SunVTS test and describes the various test options and command-line arguments.
SunVTS 6.0 Quick Reference Card (817-7686)	Provides an overview of the user interface.

Using the SunVTS `ibhcatest`

The `ibhcatest` diagnostic test checks the functionality of Sun Dual Port 4x DDR IB Host Channel Adapter PCIe ExpressModule card. This test can be run from the SunVTS user interface, or it can be run from the command line. See the *SunVTS 6.0 Test Reference Manual* (817-7665) for more information about the `ibhcatest` test.

The `ibhcatest` diagnostic test is included in the SunVTS 6.0 Patch Set 1 and subsequent SunVTS software releases. SunVTS 6.0 Patch Set 1 is available for downloading from the SunSolve℠ web site `http://sunsolve.sun.com` using the following patch numbers:

- SPARC: patch 118962-01, or higher
- X86: patch 118961-01, or higher

The adapter and Tavor device driver must be installed, and the IB port interface must be configured `offline` for the `ibhcatest` to run. A loopback cable is not needed because `ibhcatest` includes an internal loopback test. Use the following procedure when running the `ibhcatest` command.

▼ To Use the `ibhcatest` Command

1. **Ensure that the SunVTS software and the Tavor driver are installed on your system, by typing:**

```
# pkginfo SUNWvts SUNWvtsx SUNWtavor
```

   If a SunVTS software package is not installed, refer to the *SunVTS User's Guide* for installation instruction. If the `SUNWtavor` package is not installed, check your Solaris Operating System documentation for software package information.

2. **Unplumb the interface from the system, using the `ifconfig` command:**

```
# ifconfig ibdn down unplumb
```

   where *n* is the instance number of the interface.

3. **Refer to** *SunVTS 6.0 PS1 Documentation Supplement* (819-1804) **for instructions on how to run the `ibhcatest` command.**

Troubleshooting Tasks

The following tasks can be useful when troubleshooting the IB-HCA and the link.

- **Verify that the InfiniBand software packages are installed on the host.** To do this, type the command:

```
pkginfo -l | egrep "InfiniBand | Tavor | udapl"
```

  Check that the following packages are installed:

  - `SUNWib` – Sun InfiniBand Framework
  - `SUNWtavor` – Sun Tavor HCA Driver
  - `SUNWipoib` – Sun IP over InfiniBand
  - `SUNWudaplr` – Direct Access Transport (DAT) registry package (`root`)
  - `SUNWudaplu` – Direct Access Transport (DAT) registry packages (`usr`)
  - `SUNWudapltr` – Service Provider for Tavor packages (`root`)
  - `SUNWudapltu` – Service Provider for Tavor packages

  If an InfiniBand software package is not installed, check your Solaris Operating System documentation for software package information.

- **Check the system log or console for Tavor driver error messages.**

  See `tavor`(7D) for error messages and descriptions. When the driver is attached to a port on the adapter, the following message is sent.

```
tavorn: port m up (link width 4x).
```

  In the message, *n* is the instance of the Tavor device number and *m* is the port number on the adapter.

  One way to check Tavor messages is by typing the following command:

```
# dmesg | grep tavor
```

Other Useful Utilities

These utilities can display status and other information about InfiniBand devices:

- `cfgadm`
- `snoop`
- `netstat`
- `kstat`

`cfgadm`

The `cfgadm` utility displays status and other information about the IB-HCA and IB fabric. See `cfgadm_ib`(1M) for details. For example:

```
# cfgadm -al
  Ap_Id                     Type       Receptacle   Occupant      Condition
  hca:21346543210a987         IB-HCA     connected    configured    ok
  ib                          IB-FABRIC  connected    configured    ok
  ib::80020123456789a         IB-IOC     connected    configured    ok
  ib::802abc9876543           IB-IOC     connected    unconfigured  unknown
  ib::80245678,ffff,ipib      IB-VPPA    connected    configured    ok
  ib::12245678,0,nfs          IB-PORT    connected    configured    ok
  ib::21346543,0,hnfs         IB-HCA_SVC connected    configured    ok
  ib::sdp,0                   IB-PSEUDO  connected    configured    ok
```

snoop

The snoop program captures and inspects network packets. See the snoop(1M) man page for details. For example:

```
# snoop -d ibd1
Using device /dev/ibd1 (promiscuous mode)
    ib-1-167 -> *            ARP C Who is 199.1.1.168, ib-1-168 ?
    ib-1-168 -> ib-1-167     ARP R 199.1.1.168, ib-1-168 is
0:2:4:7:0:0:0:0:a:4:7c:4f:0:2:c9:2:0:0:55:91
    ib-1-167 -> ib-1-168     ICMP Echo request (ID: 35608 Sequence number: 0)
    ib-1-168 -> ib-1-167     ICMP Echo reply (ID: 35608 Sequence number: 0)
```

netstat

netstat shows network status. See the netstat(1M) man page for details. For example:

```
# netstat -I ibd 4
    input   ibd1      output       input   (Total)    output
packets errs  packets errs  colls  packets errs  packets errs  colls
2458394 0     2458268 0     0      2467288 0     2465951 0     0
92233   0     92237   0     0      92247   0     92238   0     0
92703   0     92702   0     0      92709   0     92704   0     0
```

kstat

kstat displays kernel statistics. See the kstat(1M) man page for details. For example:

```
# kstat ibd:1
module: ibd                             instance: 1
name:   ibd1                            class:    net

                                                   0
        opackets                        27381595
        opackets64                      27381595
        promisc                         off        xmt_badinterp
              0
        xmtretry                        4
```

Declaration of Conformity

Compliance Model Number: **375-3548-01**
Product Family Name: **Sun IB-HCA Dual port 4x DDR PCI-E ExpressModule**

EMC

USA—FCC Class A

This equipment complies with Part 15 of the FCC Rules. Operation is subject to the following two conditions:

1. This equipment may not cause harmful interference.
2. This equipment must accept any interference that may cause undesired operation.

Canada

This Class A digital apparatus complies with Canadian ICES-003.

European Union

This equipment complies with the following requirements of the EMC Directive 2004/108/EC:

As Information Technology Equipment (ITE) Class A per (as applicable):

EN 55022:2006	Class A
EN 61000-3-2:2000 +A2:2005	Pass
EN 61000-3-3:1995 +A1:2001	Pass

EN 55024:1998 +A1:2001 +A2:2003 Required Limits:

IEC61000-4-2	4 kV (Direct), 8 kV (Air)
IEC61000-4-3	3 V/m
IEC61000-4-4	1 kV AC Power Lines, 0.5 kV Signal and DC Power Lines
IEC61000-4-5	1 kV AC Line-Line and Outdoor Signal Lines, 2 kV AC Line-Gnd, 0.5 kV DC Power Lines
IEC61000-4-6	3 V
IEC61000-4-8	1 A/m
IEC61000-4-11	Pass

Safety

This equipment complies with the following requirements of the Low Voltage Directive 2006/95/EC:

EC Type Examination Certificates:

EN 60950-1:2001, 1st Edition
IEC 60950-1:2001, 1st Edition
Evaluated to all CB Countries
UL 60950-1:2003, CSA C22.2 No. 60950-03 File: Harmon Laboratories MELSAF_EN.18203

Supplementary Information: This product was tested and complies with all the requirements for the CE Mark. This equipment complies with the Restriction of Hazardous Substances (RoHS) directive 2002/95/EC.

_____/S/ _____

Dennis P. Symanski DATE
Worldwide Compliance Office
Sun Microsystems, Inc.
4150 Network Circle, MPK15-102
Santa Clara, CA 95054 U.S.A.
Tel: 650-786-3255
Fax: 650-786-3723

Index