PathScale™ InfiniPath™ Interconnect Installation Guide

PathScale, Inc.

Version 1.0

CHAPTER 1

About This Guide

The *PathScale InfiniPath Interconnect Installation Guide* contains complete instructions for installing the InfiniPath HT-460 hardware and the InfiniPath software. For instructions on administering and using the InfiniPath HT-460 product, see the *PathScale InfiniPath Interconnect User Guide.*

This chapter describes the objectives, intended audience, and organization of the *PathScale InfiniPath Interconnect Installation Guide*, and defines the conventions used to convey instructions and noteworthy information.

Who should read this Guide

Much of this Guide is intended for cluster administrators responsible for installing the InfiniPath HT-460 product. Users installing the InfiniPath HT-460 on an AMD Opteron™ machine cluster and administering the InfiniPath cluster should read this entire Guide.

This Guide assumes you are familiar with cluster networking and with the specific hardware to which your processors are connected. Prior to installing the InfiniPath

Adapter, you should have basic knowledge of your host and target operating systems, and working knowledge of message passing concepts.

How this Guide is organized

The *PathScale InfiniPath Interconnect Installation Guide* is organized into these sections:

- Chapter 2, “Hardware Installation”, provides instructions for installing the PathScale InfiniPath HT-460 Adapter hardware.
- Chapter 3, “Software Installation”, includes instructions for installing the PathScale InfiniPath software.
- Appendix A, “Installation Troubleshooting”, lists installation error messages that sometimes occur during installation, and provides recommendations for fixing them.
- Appendix B, “Regulatory information”.

Conventions used in this Guide

This Guide uses these typographical conventions:

Convention	Meaning
`command`	Fixed-space font is used for literal items such as commands, functions, programs, files and pathnames.
`variable`	Italic fixed-space font is used for variable names in programs and command lines, indicating parameter values that you supply.
concept	Italic font is used for emphasis, concepts, and publication titles.
user input, system output	Fixed-space font is used for code output, and for commands or constructs you type in.
$	Indicates a command line prompt.
#	Indicates a command line prompt as root.
[ ]	Brackets enclose optional elements of a command or program construct.
...	Ellipses indicate that a preceding element can be repeated.
>	Right caret identifies the path of menu commands used in a procedure.
NOTE:	Indicates important information.

CHAPTER 2

Hardware Installation

This chapter provides the requirements and instructions for installing the InfiniPath HT-460 adapter hardware. The HT-460 adapter hardware consists of the HT-460 card and the HTX™ riser card. These components will be referred to as the *adapter* and the *riser card* in the remainder of this Guide.

System requirements

This section lists system requirements and identifies areas of flexibility in your system configuration when installing the InfiniPath HT-460.

The InfiniPath HT-460 is installed in the only HTX (HyperTransport™) slot on the motherboard of your system. The HTX slot is indicated in Figure 4 for a typical Opteron motherboard.

Currently, the only supported motherboard is Iwill™ DK8S2-HT, a dual slot Opteron motherboard.

Installation of the InfiniPath adapter in a 1U or 2U chassis requires the use of the riser card. This is similar to the problem of installing a full size PCI card in these chassis. When the adapter is installed with the riser card, it is not possible to install any PCI cards, because the adapter will cover the PCI slots.

Currently, the InfiniPath adapter installation is supported only for the version 2.6.11 Linux kernel. See “Linux environment” for information on kernel and other software requirements.

Before installing any software, however, complete the hardware installation as described in this chapter.

Safety with electricity

Observe these guidelines and safety precautions when working around computer hardware and electrical equipment:

1. Locate the power source shutoff for the computer room or lab where you are working. This is where you will turn OFF the power in the event of an emergency or accident. Never assume that power has been disconnected for a circuit. Always check first.
2. Don’t wear loose clothing. Fasten your tie or scarf, remove jewelry, and roll up your sleeves. Wear safety glasses when working under any conditions that might be hazardous to your eyes.
3. Shut down the power supply to your system before you begin work, but do not disconnect the power cord from the system power or wall socket. This connection provides the ground path necessary to safely install your InfiniPath HT-460.
4. Use normal precautions to prevent electrostatic discharge.

Unpacking information

This section provides instructions for safely unpacking and handling the InfiniPath HT-460. To avoid damaging the adapter card, always take normal precautions to avoid electrostatic discharge.

Verify package contents

The InfiniPath HT-460 system should arrive in good condition. Before unpacking, check for any obvious damage to the packaging. If you find any obvious damage to the packaging or to the contents, please notify your reseller immediately.

List of the package contents

The package contents are:

- InfiniPath HT-460 adapter.
- HTX riser card for use in 1U or 2U chassis.

Unpacking the PathScale InfiniPath HT-460

When unpacking, ground yourself before removing the InfiniPath adapter from the anti-static bag.

- Grasping the InfiniPath adapter by its InfiniBand connector, pull the adapter out of the anti-static bag. Handle the adapter only by its edges or the IB connector. Do not allow the InfiniPath adapter or any adapter card components to touch any metal parts.
- After checking for visual damage, store the InfiniPath adapter and the riser card in the anti-static bag until you are ready to install them.

Hardware installation

This section contains hardware installation instructions. Please refer to the photographs in Figures 1-5 on the following pages, when performing the described installation steps. The HT-400, indicated in Figures 1 and 2, is the PathScale InfiniPath interconnect ASIC, which is the central component of the interconnect.

FIGURE 1. InfiniPath HT-460 card, top view

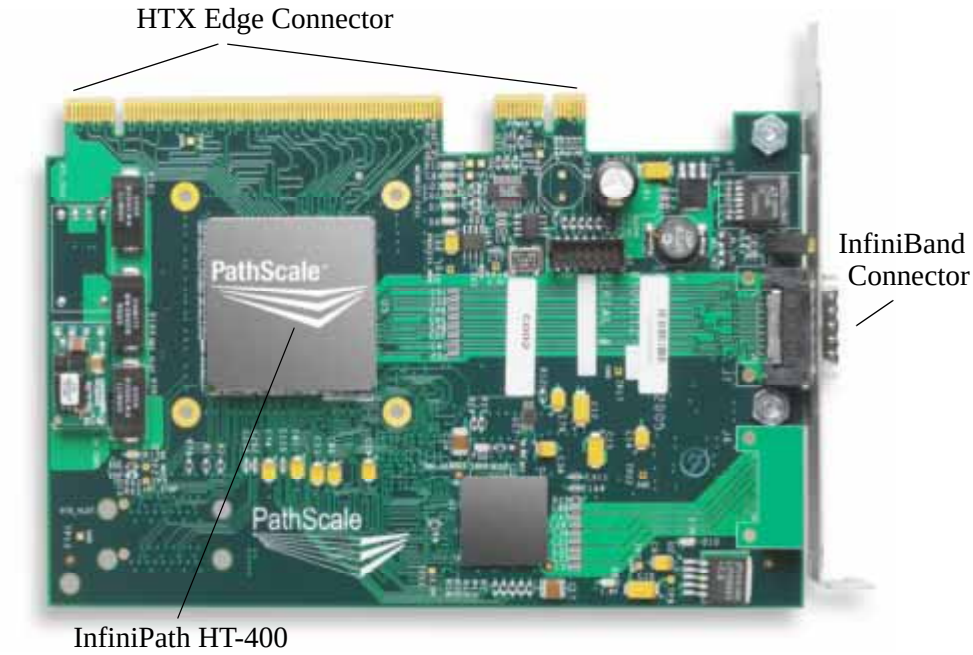

FIGURE 2. HTX riser card

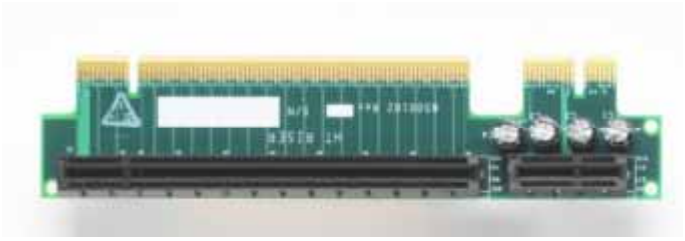

FIGURE 3. InfiniPath HT-460 with riser card attached and heat sink

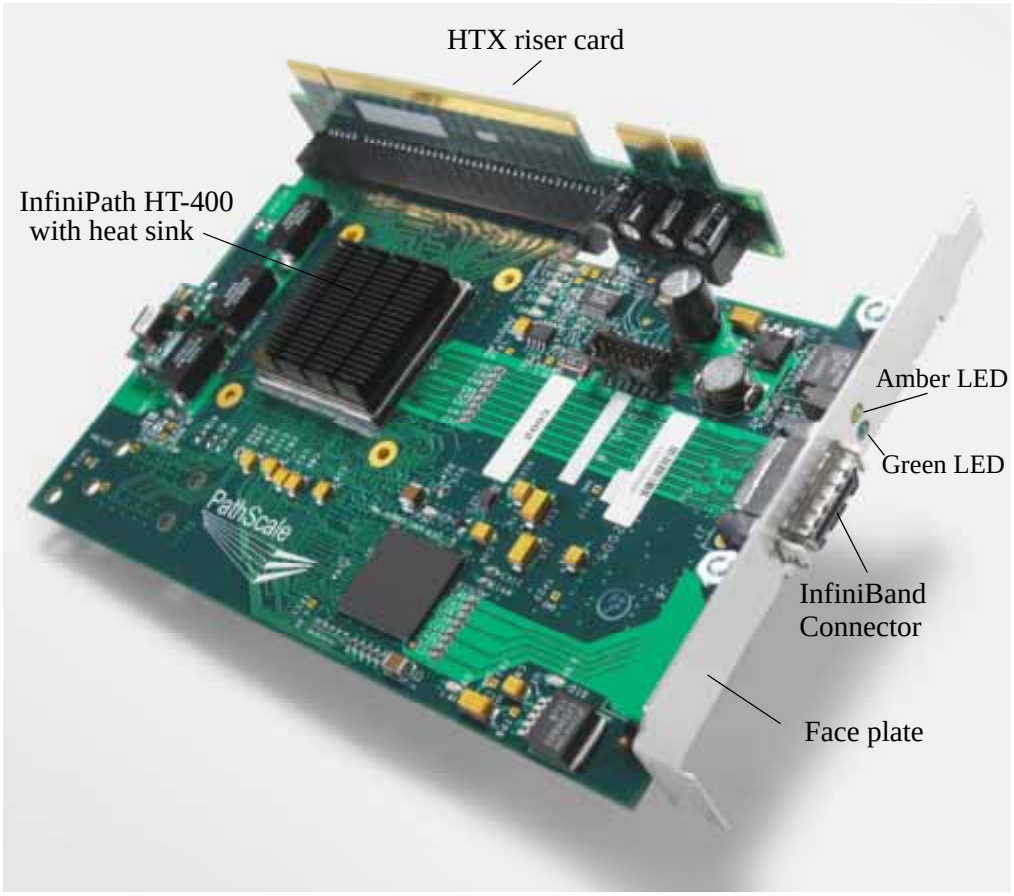

FIGURE 4. Typical dual-Opteron motherboard with HTX slot

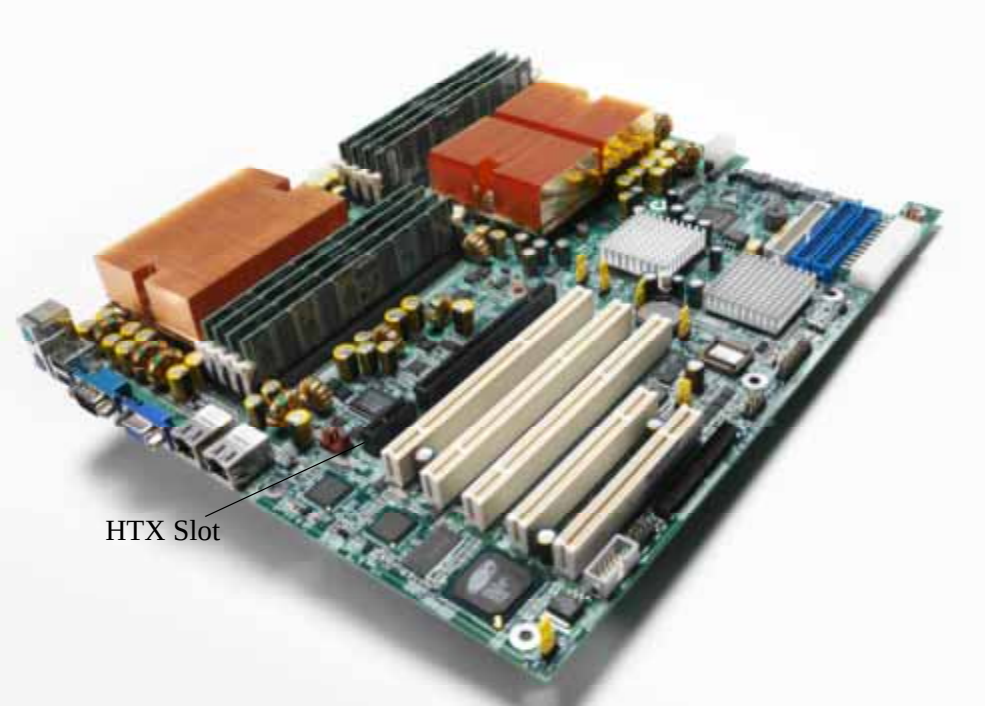

FIGURE 5. Motherboard with InfiniPath HT-460 and riser installed

Hardware installation With HTX riser

Installation of InfiniPath HT-460 in 1U or 2U chassis requires installation with an HTX riser card.

To install PathScale's InfiniPath adapter with an HTX riser card:

1. If any BIOS configuration is required, it will usually need to be done before installing the InfiniPath adapter. See "Configuring the BIOS".
2. Shut down the power supply to the system into which you'll be installing the InfiniPath adapter.
3. Take precautions to avoid damage to the cards by grounding yourself or touching the metal chassis to discharge static electricity before handling them.

4. Remove the cover screws and cover plate to expose the system's motherboard. For specific instructions on how to do this, follow the hardware documentation that came with your system.
5. If you are installing the InfiniPath adapter directly onto a motherboard, the cover plate will most likely already be removed. If not, refer to the separate instructions from your system vendor and remove the cover plate and back panel.
6. Locate the HTX slot on your motherboard. See Figure 4.
7. Determine if a blanking panel is installed in your chassis. If it is, remove it. Refer to your system vendor instructions for how to remove the blanking panel.
8. Remove the InfiniPath HT-460 from the anti-static bag.
9. Locate the face plate on the connector edge of the card.
10. Connect the InfiniPath adapter and HTX riser card (Figure 3) together, forming the assembly that you'll insert into your motherboard. To do this, first visually line up the card slot connector edge with the edge connector of the HTX riser card.
11. Holding the InfiniPath adapter by its edges, carefully insert the card slot connector into the HTX riser card edge connector. The result is a combined L-shaped assembly of the HTX riser card and InfiniPath adapter. This assembly is what you'll insert into the HTX slot on the motherboard in the next step.
12. Holding this HT assembly above the motherboard at about a 45 degree angle, slowly lower it so that the connector edge of the InfiniPath adapter clears the blanking panel opening of the chassis from the inside. Slowly align the connector edge of the HTX riser card with the motherboard's HTX slot. The HT riser and HTX slot should line up perfectly at this point. If they don't, don't force them. You probably just need to remove the blanking panel. Go back to Step 7 and start again there.
13. Insert the HT riser assembly into the motherboard HTX slot, ensuring good contact. This results in the InfiniPath adapter positioned over the motherboard, parallel to the motherboard about one inch above it. See Figure 5.
14. Secure the face plate to the chassis with a retention screw.

The InfiniPath HT-460 with HTX riser card is now installed. Next, install the cables. as described in "Cabling the HT-460 to the InfiniBand switch". Then test your installation by powering up and verifying link status. See "Completing the installation".

Hardware installation without HTX riser

The installation of InfiniPath HT-460 without an HTX riser card requires a 3U or larger chassis.

To install PathScale's InfiniPath adapter with no HTX riser card:

1. If any BIOS configuration is required, it will usually need to be done before installing the InfiniPath HT-460. See "Configuring the BIOS".
2. Shut down the power supply to the system into which you'll be installing the InfiniPath adapter.
3. Take precautions to avoid damage to the cards by grounding yourself or touching the metal chassis to discharge static electricity before handling them.
4. If you are installing the InfiniPath adapter into a covered system, you will first need to remove the cover screws and cover plate to expose the system's motherboard. For specific instructions on how to do this, follow the hardware documentation that came with your system.
5. If you are installing the InfiniPath adapter directly onto a motherboard, the cover plate will most likely already be removed. If not, refer to the separate instructions from your system vendor and remove the cover plate and back panel.
6. Locate the HTX slot on your motherboard. See Figure 4.
7. Remove the InfiniPath adapter from the anti-static bag.
8. Visually line up the card slot connector with the motherboard's HTX slot.
9. Holding the InfiniPath adapter by its edges, carefully insert the card slot connector into the motherboard HTX slot ensuring good contact. This results in the InfiniPath adapter positioned vertically in the motherboard, perpendicular to it.
10. Secure the face plate to the chassis with a retention screw.

Next, install the cables. as described in "Cabling the HT-460 to the InfiniBand switch". Then test your installation by powering up and verifying link status. See "Completing the installation".

Cabling the HT-460 to the InfiniBand switch

Standard InfiniBand copper cabling

The cable installation uses a standard InfiniBand cable. While any standard cable should work, PathScale has explicitly qualified cables from two vendors: Volex, Inc. and Leoni High Speed Cables. Cable part numbers are shown in Table 1. The longest IB cable we support is ten meters.

TABLE 1. IB Cables Part Numbers from two Qualified Vendors

IB cable length	Volex, Inc.	Leoni High Speed Cables
1 meter	VIB04UA-28-01M	L45590-A499-A30
2 meter	VIB04UA-28-02M	L45590-A499-A31
3 meter	VIB04UA-28-03M	L45590-A499-A32
5 meter	VIB04UA-26-05M	L45590-A499-A33
10 meter	VIB04UA-24-10M	L45590-A499-A44

To install the InfiniBand cables:

1. Check that you have removed the protector plugs from the cable connector ends.
2. Different vendor cables might have different latch mechanisms. Determine if your cable has a spring-loaded latch mechanism. If your cable is spring-loaded, grasp the metal shell and pull on the plastic latch to release the cable. To insert, push and the cable snaps into place. You will hear a short “click” sound from the cable connector when it snaps in.
3. If your cable latch mechanism is not spring-loaded, simply push on the metal case, then push the plastic latch to lock the cable in place.
4. The InfiniBand cables are symmetric; either end can be plugged into the switch. Connect the InfiniBand cable to the connector on the InfiniPath HT-460. Depress the side latches of the cable when connecting. (On some cables this latch is located at the top of the cable connector.) Make sure the lanyard handle on the cable connector is slid forward toward the card connector until fully engaged.
5. Connect the other end of the cable to the InfiniBand switch.

6. Two InfiniPath HT-460s can be connected directly with a standard IB cable, without the use of a switch. This can be useful for testing and benchmarking.
7. You can also do limited testing with a loopback connector. If you want to test sending and receiving data and link packets only on the InfiniPath adapter itself, connect it using a loopback connector. You cannot run MPI programs with a loopback connector. There is an internal test flag that allows for some testing.

Optical fibre option

PathScale InfiniPath adapter also supports connection to the switch by means of optical fibres through the Emcore QT2400 optical media converter. Not all switches support this convertor. Contact Emcore via www.emcore.com for details.

Completing the installation

To complete the hardware installation:

1. Complete any other installation steps for other components.
2. Replace the cover plate and back panel.
3. Verify that the power cable is properly connected.
4. Turn on the power supply, and boot the system normally.
5. Watch for LED indicators. The amber LED will normally illuminate first. See “LED link and data indicators”.

LED link and data indicators

There are two LED indicators next to the InfiniBand connector on the InfiniPath HT-460 to monitor link and error status. These are visible with the card installed and the system running. For 1U and 2U chassis installations, the amber LED is directly below the green LED. See Figure 3. For installations in 4U chassis systems, the LEDs will be side by side.

Note that the LEDs are functional only after the InfiniPath software has been installed and the system is connected to one or more other systems through the InfiniBand switch. Otherwise, the LEDs will flash once at power-on.

The green LED will normally illuminate first. Normal state is Green On, Amber On.

TABLE 2.

LED	On	Off
Amber	Link configured. Properly connected and ready to receive data packets and link packets.	Link not configured. Check the connection.
Green	Signal detected. Can only send and receive link and SMA packets.	Green Off = Link not up. Probable loss of signal, cabling problems, switch not powered up, etc. Check the connection and switch, and be sure that the software is installed and running.

Configuring the BIOS

The BIOS, stored in non-volatile memory, contains code needed to start the bootstrap process when the system is powered up. It also contains certain parameters characterizing the system.

You can check the BIOS settings for the HyperTransport configuration using the BIOS Setup Utility. HyperTransport link speed and width are adjustable. For specific instructions on how to do this, follow the hardware documentation that came with your system.

CHAPTER 3 *Software Installation*

This chapter provides instructions for installing the PathScale InfiniPath software. Software requirements and components are listed and described.

The InfiniPath software includes drivers, protocol libraries, PathScale's implementation of the MPI message passing standard, and example programs, including benchmark programs. This chapter describes what is needed for software installation. Further information, particularly on the use of MPI, is found in the companion document, *PathScale InfiniPath Interconnect User's Guide*. For convenience, some information appears in both documents.

Software installation involves the following steps.

1. Obtain and install the Linux kernel software on each node in your cluster. The required kernel and supported Linux distributions are defined below under "Linux environment"
2. Download the InfiniPath software from the PathScale web site www.pathscale.com and install the appropriate packages on each cluster node as described under "Software packaging" and "Installing InfiniPath software".
3. Configure the `infinipath` and `ipath` drivers as described under ""Loading and unloading the driver""

4. Perform the recommended health checks as described under “Customer acceptance utility”.

Linux environment

The current version of the InfiniPath software requires Linux kernel version 2.6.11. The supported distributions are Fedora Core 3 and SuSE 9.3. Each distribution requires a different version of the InfiniPath software distribution, as described below under “Software packaging”.

Among the many optional packages that each distribution offers, the InfiniPath software requires, on every node, `openssh`, `openssh-server`, and, if the MPD job launcher is to be used, `python`.

Software packaging

Linux distributions of InfiniPath software are installed from binary RPMs. There are multiple interdependent RPM packages that make up InfiniPath software. The RPMs can be downloaded from
http://www.pathscale.com/infinipath_support/downloads.html

Some RPMs are required only on compute nodes. Some RPMs are required only on nodes that launch jobs, that is, frontend nodes. Some RPMs are required only on development machines. Of course, any machine can serve any combination of these three purposes, but a typical cluster has many compute nodes and only one or just a few frontends. Some RPMs are optional.

InfiniPath software RPMs

The InfiniPath software consists of the following RPMs. In each case `xxx` is a build identifier and `yyy` is either `fc3` (for the Fedora Core 3) distribution or `suse9.3` (for the SuSE 9.3 distribution).

- `infinipath-1.0-xxx_yyy_psc.x86_64.rpm`
  InfiniPath drivers, binaries and source code.
  InfiniPath configuration files.
  Needed on compute nodes and frontend nodes.

- `mpi-benchmark-1.0-xxx_yyy_psc.x86_64.rpm`
  MPI benchmark binaries.
  Needed on job launch nodes, but only if you want to run the benchmarks.

- `mpi-devel-1.0-xxx_yyy_psc.noarch.rpm`
  Source code for MPI examples and benchmarks.
  Optional. Needed on build computers if you want to build the examples or rebuild the benchmarks.

- `mpi-libs-1.0-xxx_yyy_psc.i386.rpm`
  Shared libraries for MPI.

Needed on all compute nodes.

- `mpi-frontend-1.0-xxx_yyy_psc.i386.rpm`
  MPI job launch scripts and binaries.
  Needed on job launch nodes.

- `infinipath-libs-1.0-xxx_yyy_psc.i386.rpm`
  InfiniPath protocol shared libraries for 32-bit and 64-bit systems.
  Needed on all compute nodes.

Two additional RPMs include man pages and would generally be installed only on frontend nodes. Of course, they are not required:

- `infinipath-doc-1.0-xxx_yyy_psc.noarch.rpm`
  InfiniPath man pages and other documents.

- `mpi-doc-1.0-xxx_yyy_psc.noarch.rpm`
  Man pages for MPI functions and other MPI documents.

To generate a list of InfiniPath software package contents, run:

`rpm -qlp` *rpm_file_name*

on each RPM.

Installing InfiniPath software

In this section we assume that the correct Linux kernel and a supported distribution have been installed on every node. The previous section specifies which RPMS are required or optional for each type of node, according to its function as a compute node, frontend node, or development machine.

Copy the InfiniPath RPMs to a directory accessible (e.g., via NFS) to every node, or at least, copy to a directory on each node the RPMs needed on that node. Then for each node, login as root and, for each relevant RPM, run the command

```
rpm -Uvh rpmdirectorypath/rpm_name.rpm
```

After having installed the software on a frontend node, you can install on the compute nodes in parallel with the help of the `mpirun` command, as in

```
mpirun -np n -m hostfile -nonmpi \
rpm -Uvh rpmdirectorypath/rpm_name.rpm
```

See the *User's Guide* for description of the use of `mpirun`.

You can verify the installation of all RPMs with

```
rpm --verify 'mpi*' 'infinipath*'
```

Once the RPMs have been installed, you will need to modify a few files on each node to allow for proper cluster operation.

Configuring the infinipath driver

The `infinipath` driver is responsible for InfiniPath HT-460 initialization, handling interrupts (mostly for errors) mediating access to the interconnect for user programs, and handling memory mapping. It also provides support to layered drivers such as the `ipath` driver, and support programs such as the Subnet Management Agent (SMA).

Debug messages are printed with the function name preceding the message.

The primary configuration file for the `infinipath` driver, `ipath` Ethernet driver, the Subnet Management Agent (SMA) and associated daemons is:

```
/etc/sysconfig/infinipath
```

This is where options to either the driver or the SMA are provided. Normally this configuration file is set up correctly at installation and the cluster administrator does not need to change it.

See the `infinipath(4)`, `ipath(4)`, `ipath_sma(8)`, and `ipath_mux(8)` man pages for more information.

Refer also to the *User Guide.*The device files are `/dev/ipath` and `/dev/ipath_sma`.

Configuring the ipath Ethernet driver

You will need to create a network device configuration file for the layered Ethernet device on InfiniPath HT-460. This configuration file will resemble the configuration files for the other Ethernet devices on the nodes.

ipath configuration on Fedora

The Ethernet driver is loaded at startup by a line in

```
/etc/modprobe.conf
```

such as

```
alias eth2 ipath
```

and by corresponding lines in

```
/etc/sysconfig/network-scripts/ifcfg-eth2
```

namely,

```
# PathScale Interconnect Ethernet
DEVICE=eth2
ONBOOT=yes
BOOTPROTO=dhcp
```

You can check whether the Ethernet driver has been loaded with

```
lsmod | grep ipath
```

ipath configuration on SuSE

1. Check to see that the ipath module is loaded:

```
lsmod | grep -w ipath
```

If it is not listed, load it with the command:

```
modprobe ipath
```

2. Determine the MAC address to be used with the following command:

```
sed -n -e 's/.*GUID=[0-9a-f]*:[0-9a-f]*//' \
-e 's/:\(.\):/:0\1:/g' \
-e 's/://' -e 's/,.*//p' /proc/driver/infinipath/status
```

The output should appear similar to this (6 hex digit pairs, separated by colons):

```
00:11:75:04:e0:11
```

This is referred to as $MAC in the rest of the instructions, where $MAC must be replaced by the actual MAC address output by the `sed` command.

2. Create the file

```
/etc/sysconfig/hardware/hwcfg-eth-id-$MAC
```

with the following lines:

```
MODULE=ipath
STARTMODE=auto
```

3. Create the file

```
/etc/sysconfig/network/ifcfg-eth-eth#
```

where # is the number of the next Ethernet interface. If `ifconfig -a` showed `eth0` and `eth1`, then # would typically be `2` and the file name would be: `/etc/sysconfig/network/ifcfg-eth-eth2`. It should contain the following lines if you are using DHCP:

```
BOOTPROTO=dhcp
NAME=ipath
STARTMODE=auto
_nm_name=eth-id-$MAC
```

If you are using static IP addresses, then it will instead appear similar to the lines below, with the actual network and IP address that you are using (this example uses a private IP address of 192.168.1.2, with the normal matching netmask and broadcast addresses):

```
BOOTPROTO=static
IPADDR=192.168.1.2
NETMASK=255.255.255.0
NETWORK=192.168.1.0
NAME=ipath
STARTMODE=auto
_nm_name=eth-id-$MAC
```

It is important that the name in `_nm_name` exactly match the name of the file created in step 2, with `hwcfg-` prepended.

4. When the system is next rebooted, the `ipath` Ethernet device should be correctly configured. Verify this with the command:

```
ifconfig -a
```

If you do not see a device listed with a HWaddr matching $MAC, check for errors in steps 1, 2, or 3. If it is listed, but does not have a valid "inet addr" listed, check for errors with the IP address or DHCP server configuration.

Loading and unloading the driver

Normally, the driver(s) are loaded during system boot, unless they have been configured off. To check the configuration state, use the command:

```
chkconfig --list infinipath
```

To enable the driver, use the command (as root):

```
chkconfig infinipath on 2345
```

To disable the driver, use the command (as root):

```
chkconfig infinipath off
```

To stop the driver, use the command (as root):

```
/etc/rc.d/init.d/infinipath [start|stop|restart]
```

To manually unload the driver, use the command (as root):

```
modprobe -r ipath infinipath
```

To manually load the driver, use the command (as root):

```
modprobe infinipath
```

This is done only as a way to test loading problems, since it will not complete the configuration, nor will it run the daemon(s) needed for correct operation of InfiniPath HT-460.

The `INFINIPATH_NOSMA`, `INFINIPATH_NOMUX`, and `INFINIPATH_NODUMP` variables in `/etc/sysconfig/infinipath` are used to modify SMA behavior.

Recompiling the driver

If you upgrade the kernel then you either have to reinstall the InfiniPath software or recompile the driver. You can recompile the driver by

```
cd /usr/src/pathscale/drivers
./make-install.sh
```

Switch configuration and monitoring

Follow the vendor documentation for installing and configuring your switches.

Customer acceptance utility

`ipath_checkout` is a `bash` script to verify that the installation is correct and that all the nodes of the network are functioning and mutually connected by the InfiniPath fabric. It is to be run on a frontend node, and requires specification of a hosts file:

```
ipath_checkout [options] hostsfile
```

where *`hostsfile`* designates a file listing the hostnames of the nodes of the cluster, one hostname per line, just as the hosts file for the `mpirun` command. (See the discussion of hosts file in the *PathScale InfiniPath Interconnect Users Guide*).

`ipath_checkout` performs the following seven tests on the cluster:

1. `ping` all nodes to verify all are reachable from the frontend.
2. `ssh` to each node to verify correct configuration of ssh.
3. Gather and analyze system configuration from nodes.
4. Gather and analyze RPMs installed on nodes
5. Verify InfiniPath hardware and software status and configuration.
6. Verify ability to run MPI jobs on nodes by use of `mpirun`.
7. Run bandwidth and latency test on every pair of nodes and analyze results.

The possible options to `ipath_checkout` are:

`-h, --help`
Display help message giving defined usage.

`-v, --verbose`
`-vv, --vverbose`
`-vvv, --vvverbose`
These specify three successively higher levels of detail in reporting results of tests. So,

there are four levels of detail in all, including the case of where none these options are given.

`-c, --continue`
When not specified, the test terminates when any test fails. When specified, the tests continue after a failure, with failing nodes excluded from subsequent tests.

`--workdir=DIR`
Use DIR to hold intermediate files created while running tests. DIR must not already exist.

`-k, --keep`
Keep intermediate files that were created while performing tests and compiling reports. Results will be saved in a directory named pathscale_* or the directory name given to `--workdir`.

`--skip=`*`LIST`*
Skip the tests in *`LIST`* (e.g. --skip=2457 will skip tests 2, 4, 5, and 7.)

In most cases of failure, the script suggests recommended actions. Please see the `ipath_checkout` man page for further information and updates.

Removing software packages

To uninstall the InfiniPath software packages on any node, using a `bash` shell, type the command (as root):

```
rpm -e $(rpm -qa 'mpi*' 'infinipath')
```

This will uninstall the InfiniPath software RPMs.

APPENDIX A

Installation Troubleshooting

This appendix describes anomalies that you may encounter during the installation process. It describes possible causes and provides recommended actions.

Troubleshooting InfiniPath adapter installation

This section lists conditions you may encounter while installing the PathScale InfiniPath HT-460, and offers suggestions for working around them.

Mechanical and Electrical Considerations

If the InfiniPath interconnect boards are all installed correctly and solidly connected to an appropriate InfiniBand switch, then, on powering up, each board should show both LEDs lit.

If a node repeatedly and spontaneously reboots when attempting to load the InfiniPath driver, it may be a symptom that its InfiniPath interconnect board is not well seated in the HTX slot.

Driver load fails

Symptom: If you try to load the infinipath driver on a kernel that InfiniPath software does not support (which for now is anything besides version 2.6.11), the load fails. Error messages similar to this appear:

```
modprobe: error inserting '/lib/modules/2.6.3-1.1659-smp/
kernel/drivers/net/infinipath.ko': -1 Invalid module
format
```

Suggestion: Install Linux kernel version 2.6.11, then reload the driver.

Initiation message

Symptom: When I start the program, it fails with messages similar to:

```
userinit: userinit ioctl failed: Network is down [1]:
device init failed
userinit: userinit ioctl failed: Fatal Error in
keypriv.c(520): device init failed
```

Suggestion: One or more nodes do not have the Interconnect in a usable state. A cable is not connected, the switch is down, SMA is not running, or a hardware error has occurred.

You can check the file `/proc/driver/infinipath/status` to verify that the Infinipath software is loaded and functioning Normally, it should contain node's LID, MLID, GUID, and serial number, plus the following items

```
Initted
InfiniPath_found
IB_link_up
IB_configured
```

plus, if `ipath` is in use, the following items

```
ipath_loaded
ipath_up.
```

In case of trouble the entry `Fatal_Hardware_Error` might appear.

If, on any node, the driver status appears abnormal, you can try restarting it with

```
/etc/rc.d/init.d/infinipath restart
```

or

```
/etc/rc.d/init.d/infinipath stop

/etc/rc.d/init.d/infinipath start
```

Broken intermediate link

Symptom: Some message traffic passes through the fabric while other traffic appears to be blocked. MPI jobs fail to run.

Explanation: In large cluster configurations, switches may be attached to other switches in order to supply the necessary inter-node connectivity. Problems with these inter-switch (or intermediate) links are sometime more difficult to diagnose than failure of the final link between a switch and a node. The failure of an intermediate link may allow some traffic to pass through the fabric while other traffic is blocked or degraded.

Suggestion: If you encounter such behavior in a multi-layer fabric, check that all switch cable connections are correct

.

SMA syslog messages

This section lists SMA syslog messages you may encounter, and offers suggestions for working around them.

Most of these are written once at SMA startup, or once for each unit. A few are printed at state changes after startup.

`Failed to take IB link to DOWN state: errno`
Not fatal, but likely a sign of other troubles. Printed when the SMA is bouncing the link in order to get the attention of the Subnet Manager and the link won't go down.

The following messages are normally written just once when `infinipath` is started.

`unit u devstatus = hex_str desc_str`
The `hex_str` holds a status code, encoding a set of binary status variables. The *desc_str* is a concatenation of descriptive strings from the following list, each indicating the corresponding bit being on:

```
IB_NOCABLE
IB_CONF
IB_READY
CHIP_PRESENT
SMA
PLACEHOLDER
LAYERUP
LAYERLOADED
INITTED
```

or, if the determination of status failed:

`unable to get status for unit u`
That occurrence is not fatal, but indicates something anomalous.

`connected to mux, id=mux_id`
Indicates normal operation.

`looking for n units`
Indicates normal operation and shows the maximum number of units supported by the installed driver and SMA.

```
unit u found
    or
unit u not found
```

Written once at startup for each unit. “Found” is normal. “Not found” is not fatal, but usually means that not all possible units are configured..

`expected 1 port on unit u, got n`

A unit claimed to support more than one port. Not fatal, but it almost certainly means a driver problem or library mismatch.

`setting unit u LID to hx(n) from command line`

Only printed if someone sets the LID on the command line.

`setting unit u GUID to xx:xx:xx:xx:xx:xx:xx:xx from command line` Only printed if someone set the GUID on the command line.

`calloc() failure in mcast_register()`

Printed if someone (likely an MPI job) tries to set up a multicast group and there isn't enough memory to `malloc` the struct, which is tiny. Not fatal, but it means the system is in bad trouble.

`unexpected state transition s1-->s2 in mcast_mp_state()`
`unexpected state transition s1-->s2 in mcast_state()`

Not fatal but they shouldn't happen. They might indicate a problems with a particular SM implementation in setting up a multicast group or handling link state transitions.

`no subnet management activity detected on unit u, bouncing link`

At startup, if we haven't made sufficient progress in setting up our link, we start taking the link up and down periodically to try to get the attention of the SM. This message may indicate that the SM is down, that the fabric is partitioned and the SM is unreachable, or that there is a compatibility problem with a particular implementation of an SM.

`SM does not support Multicast`

Self explanatory. The SM managing the fabric does not support multicast groups. This implies that neither the layered Ethernet driver nor MPI will work. Otherwise, not fatal.

`SM set MLID to 0xxxxx`

Printed once per multicast group creation. One multicast group is created at startup for `ipath`, so there is always one.

or

```
Multicast join failed, status = 0xxxxx
```

The multicast join request was refused by the SM for some reason.

```
SM set unit u LID to 0xxxxx(n)
SM set unit %d link state to 0xxxxx
```

Printed whenever the SM wants this node to change its link state. Normally, there may be a few such transitions at startup, and then whenever there is an unplug/replug event.

```
SM set unit u MTU to 0xxxxx
```

Printed whenever the SM assigns this node an MTU. Normally only once, at startup.

The following seven are all non fatal diagnostics indicating something abnormal.

```
handle_sigchld: waitpid() returned error, err_no
handle_sigchld: waitpid() returned 0
handle_sigchld: waitpid() returned unexpected pid pid
```

```
clock_sanity: fstat(\"tmpname\") failed -- err_no
clock_sanity: mkstemp(\"tmplat\") failed -- err_no
clock_sanity: gmtime(long) failed -- err_no
clock_sanity: year yyyy is likely bogus -- check hwclock
```

The next is normally written at shutdown

```
child process exited, exiting
```

The following may be written by `ipath_sma` or by `ipath_dumpmads`. They are all diagnostics indicating non-fatal problems opening, writing, or rotating the log file.

```
backup log file ipath_mux_logfile is not a regular file"
unable to rename log file ipath_mux_logfile, deleting
unable to delete log file ipath_mux_logfile
unable to open log file ipath_mux_logfile
unable to stat() log file ipath_mux_logfile
unable to open log file ipath_mux_logfile
unable to stat() log file ipath_mux_logfile
unable to open log file ipath_mux_logfile
```

Here is an example of the syslog record for a normal startup.

```
May 16 20:45:26 iqa-16 /usr/bin/ipath_sma[5301]: connected to
mux, id=4
May 16 20:45:26 iqa-16 /usr/bin/ipath_sma[5301]: looking for 1
units
May 16 20:45:26 iqa-16 /usr/bin/ipath_sma[5301]: unit 0 found
May 16 20:45:26 iqa-16 /usr/bin/ipath_sma[5301]: unit 0 devsta-
tus = 0x61 IB_READY CHIP_PRESENT INITTED
May 16 20:45:26 iqa-16 /usr/bin/ipath_sma[5301]: SM set unit 0
LID to 0x12(18)
May 16 20:45:26 iqa-16 /usr/bin/ipath_sma[5301]: SM set unit 0
MTU to 0x4
May 16 20:45:26 iqa-16 /usr/bin/ipath_sma[5301]: SM set unit 0
link state to 0x4
May 16 20:45:36 iqa-16 /usr/bin/ipath_sma[5301]: SM set MLID to
0xc010
```

/var/log/ipath_sma

If verbose logging has been enabled, the SMA creates a log of Subnet Management MADs it sends and receives. The default location is `/var/log/ipath_sma`. It is cleared when the SMA is started so the entries are for this session only. Verbose logging is enabled by use of the `-v` option to `ipath_sma`.

Looking at the SMA log can sometimes help diagnose problems. If you have enabled verbose logging and there are no entries in the sma logfile even after several minutes, make sure that:

- the software is loaded (use `rpm`). See the `infinipath` service is enabled (use `chkconfig`). See “Loading and unloading the driver”.
- the infinipath driver is loaded (use `lsmod`). See “Loading and unloading the driver”.
- the sma has been started (use `ps`). See “Switch configuration and monitoring” and “Configuring the ipath Ethernet driver”.
- the fabric is correctly cabled. See “Cabling the HT-460 to the InfiniBand switch”.
- all devices on the fabric are powered up. See “Completing the installation”.

- that a Subnet Manager is up and functioning. See vendor documentation.

Symptom: You are convinced all the above are correct and still aren't getting any SMA log output.

Suggestion: Try using the `nodeinfo` program to check the low-level InfiniBand connectivity and probe the low-level topology of the fabric. See `nodeinfo` man page for details.

Symptom: You see that the SMA is receiving the same message over and over from the SM. It is likely that the SM is not seeing the replies from the SMA.

Suggestion: Try unloading /reloading the driver and the problem usually goes away. See "Loading and unloading the driver".

ipathbug-helper

The InfiniPath software includes a shell script `ipathbug-helper`, which can gather status and history information for use in analyzing InfiniPath adapter problems. When seeking assistance from PathScale technical support, you should run this script on the head node of your cluster and perhaps on some of the compute nodes which are suspected to have problems, and send its `stdout` output to your reseller.

It is best to run `ipathbug-helper` with root privilege, since some of the queries it makes requires it. There is also a `--verbose` which greatly increases the amount of gathered information.

APPENDIX B

Regulatory Information

Warning

This is a Class A product. In a domestic environment this product may cause radio interference in which case the user may be required to take adequate measures.

NOTE: This equipment has been tested and found to comply with the limits for a Class A digital device, pursuant to Part 15 of the FCC Rules. These limits are designed to provide reasonable protection against harmful interference when the equipment is operated in a commercial environment. This equipment generates, uses, and can radiate radio frequency energy and, if not installed and used in accordance with the instruction manual, may cause harmful interference to radio communications. Operation of this equipment in a residential area is likely to cause harmful interfer-

ence in which case the user will be required to correct the interference at his own expense.

Any modification to this unit not expressly approved by PathScale could void the user's authority to operate the equipment.