SGI® 10-Gigabit Ethernet PCI Express Network Adapter User's Guide

007–4983–002

FCC WARNING
This equipment has been tested and found compliant 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 interference, in which case the user will be required to correct the interference at personal expense.

VDE 0871/6.78
This equipment has been tested to and is in compliance with the Level A limits per VDE 0871.

EUROPEAN UNION STATEMENT
This device complies with the European Directives listed on the “Declaration of Conformity” which is included with each product. The CE mark insignia displayed on the device is an indication of conformity to the aforementioned European requirements.

International Special Committee on Radio Interference (CISPR)
This equipment has been tested to and is in compliance with the Class A limits per CISPR publication 22.

Canadian Department of Communications Statement
This digital apparatus does not exceed the Class A limits for radio noise emissions from digital apparatus as set out in the Radio Interference Regulations of the Canadian Department of Communications.

Attention

Cet appareil numérique n'émet pas de perturbations radioélectriques dépassant les normes applicables aux appareils numériques de Classe A préscrites dans le Règlement sur les interferences radioélectriques etabli par le Ministère des Communications du Canada.

Japanese Compliance Statement

この装置は、情報処理装置等電波障害自主規制協議会 (VCCI) の基準に基づくクラス A 情報技術装置です。この装置を家庭環境で使用すると電波妨害を引き起こすことがあります。この場合には使用者が適切な対策を講ずるよう要求されることがあります。

Compliance Statement in Chinese

警告使用者：

這是甲類的資訊產品，在居住的環境中使用時，可能會造成射頻干擾，在這種情況下，使用者會被要求採取某些適當的對策．

New Features in this Guide

This revision includes additional SGI performance tuning recommendations. See Chapter 4, "Performance Tuning" on page 13.

Record of Revision

Version	Description
001	April 2007 Original publication
002	October 2007 Revision

Contents

Figures

Tables

About this Guide

This guide describes the SGI 10-Gigabit (Gbit) Ethernet PCI Express network adapter. It requires one of the following SGI ProPack for Linux releases:

- SGI ProPack 5 Service Pack 1 or later
- SGI ProPack 4 Service Pack 3 or later

You can use the SGI 10-Gbit Ethernet PCI Express network adapter in addition to your current adapter.

This guide shows you how to connect the adapter to an Ethernet network and explains how to operate the adapter.

Audience

This guide assumes that you have general knowledge of Ethernet networks and the system in which the adapter is installed.

Important Information

Warning: Never look into the end of a fiber optic cable to confirm that light is being emitted (or for any other reason).

Do not use any type of magnifying device, such as a microscope, eye loupe, or magnifying glass. Such activity causes a permanent burn on the retina of the eye. Optical signal cannot be determined by looking into the fiber end.

Most fiber optic laser wavelengths (1300 nm and 1550 nm) are invisible to the eye and cause permanent eye damage. Shorter wavelength lasers (for example, 780 nm) are visible and can cause significant eye damage.

Use only an optical power meter to verify light output.

Scope of this Guide

This guide is written to facilitate installation of the adapter and does not cover detailed points of network configuration. It contains the following chapters:

- Chapter 1, "Features and Capabilities", summarizes features, cabling, configuration limits, and tools.
- Chapter 2, "Connecting the Adapter to a Network", shows you how to connect the adapter to your network.
- Chapter 3, "Operating the Adapter", explains how to verify installation of the adapter and software, how to reset the adapter, and how to set configuration parameters.
- Chapter 4, "Performance Tuning", discusses performance tuning topics.

Related Publications

This guide is part of a document set that fully supports the installation, operation, and service of the adapter. For more information about installing and servicing the adapter, see the user's guide for the system in which the adapter is installed.

Also see the following:

- *Linux Configuration and Operations Guide*
- *The Network Administrators' Guide*
- The following Myricom webpages:
  - *Myri-10G 10-Gigabit Ethernet Solutions*:

    http://www.myri.com/Myri-10G/10gbe_solutions.html
  - *README - Myricom 10GbE driver for Linux*:

    http://www.myri.com/scs/READMES/README.myri10ge-linux
  - *Myri-10G PCI-Express NIC with a 10GBase-R port*:

    http://www.myri.com/Myri-10G/NIC/10G-PCIE-8A-R.html

- Standard Linux man pages that are useful for any Ethernet device:

  `ethtool`(8)
  `ifconfig`(8)
  `ip`(8)

Obtaining Publications

You can obtain SGI documentation as follows:

- See the SGI Technical Publications Library at http://docs.sgi.com. Various formats are available. This library contains the most recent and most comprehensive set of online books, release notes, man pages, and other information.
- You can view release notes on your system by accessing the `README.txt` file for the product. This is usually located in the `/usr/share/doc/`*productname* directory, although file locations may vary.
- You can view man pages by typing `man` *title* at a command line.

Conventions

The following conventions are used throughout this document:

Convention	Meaning
`command`	This fixed-space font denotes literal items such as commands, files, routines, path names, signals, messages, and programming language structures.
variable	Italic typeface denotes variable entries and words or concepts being defined.
`user input`	This bold, fixed-space font denotes literal items that the user enters in interactive sessions. (Output is shown in nonbold, fixed-space font.)

[ ]	Brackets enclose optional portions of a command or directive line.
...	Ellipses indicate that a preceding element can be repeated.

Product Support

SGI provides a comprehensive product support and maintenance program for its products:

- If you are in North America, contact the Technical Assistance Center at +1 800 800 4SGI or contact your authorized service provider.
- If you are outside North America, contact the SGI subsidiary or authorized distributor in your country.

Reader Comments

If you have comments about the technical accuracy, content, or organization of this publication, contact SGI. Be sure to include the title and document number of the publication with your comments. (Online, the document number is located in the front matter of the publication. In printed publications, the document number is located at the bottom of each page.)

You can contact SGI in any of the following ways:

- Send e-mail to the following address:

  techpubs@sgi.com

- Contact your customer service representative and ask that an incident be filed in the SGI incident tracking system.
- Send mail to the following address:

  SGI
  Technical Publications
  1140 East Arques Avenue
  Sunnyvale, CA 94085–4602

SGI values your comments and will respond to them promptly.

Features and Capabilities

This chapter discusses the following:

SGI Systems Supported

The SGI 10-Gigabit (Gbit) Ethernet PCI Express network adapter is supported in the following systems:

- SGI Altix 4700
- SGI Altix 450
- SGI Altix XE210
- SGI Altix XE240
- SGI Altix XE310
- SGI Altix XE1200
- SGI Altix XE1300

Key Features

The adapter includes the following key features:

- Low-profile PCI Express x8 add-in card.
- 10-Gigabit Ethernet.

Note: SGI does not support dual-protocol or 10–Gigabit Myrinet.

- Wire-speed performance.
- Firmware-controlled offload engine. The driver and NIC firmware implement zero-copy on the send side with all supported operating systems, and, depending on the operating system, use a variety of stateless offloads, including:
  - Interrupt coalescing
  - IP and TCP checksum offload, send and receive
  - TCP segmentation offload (TSO), also known as large send offload (LSO)
  - Receive-side scaling (RSS)
  - Large receive offload (LRO)
  - Multicast filtering

For additional details, see:

http://www.myri.com/Myri-10G/NIC/10G-PCIE-8A-R.html

10-Gbit Ethernet Technology

The 10-Gbit Ethernet technology is an extension of Gigabit Ethernet (1000-Base-T) technology that allows over-the-wire speeds of up to 10 Gbits per second (Gbps), which is theoretically ten times the rate of existing technology.

The 10-Gbit Ethernet technology is targeted at backbone networks and interserver connectivity. It provides an upgrade path for high-end workstations that require more bandwidth than Gigabit Ethernet can provide.

Cabling

The adapter has an LC connector and uses a `10GBASE-SR` transceiver at 850 nm. It is connected to the network using a multimode fiber (MMF) 50-micron cable. The cable (which is not included in the shipment) must have a modal bandwidth in the range

from 400-MHz * km to 2000-MHz * km, depending on its length, as shown in Table 1-1.

Table 1-1 10-Gbit Cable Standards

Diameter (Microns)	Modal Bandwidth (MHz * km)	Range (Meters)
50	400	2 to 66
50	500	2 to 82
50	2000	2 to 300

Configuration Limits

The number of the 10-Gbit Ethernet PCI Express network adapters supported varies by system. Consult with your SGI representative to determine the currently supported maximum for your configuration.

Tools

The following standard Linux commands are useful with any Ethernet device:

- `ethtool`(8)
- `ifconfig`(8)
- `ip`(8)

For more information, see the man page associated with each tool.

Connecting the Adapter to a Network

To install the SGI 10-Gbit Ethernet PCI Express network adapter, refer to the instructions for installing a PCI card in the user's or owner's guide that came with the SGI system.

To connect the adapter to a network, do the following:

1. Remove the protective end caps and **save them**.

Caution: 10-Gbit optics are very sensitive. If you plan on leaving them disconnected for any length of time, you must replace the end caps. The optics on the SGI 10-Gbit Ethernet PCI Express network adapter cannot be cleaned.

2. Insert the LC connector on one end of the fiber-optic cable into the adapter, as shown in Figure 2-1 Ensure that the connector is inserted completely into the jack.

Note: If the network connects to an Ethernet switch, consult the operating manual for the switch to ensure that the switch port is enabled and configured correctly and has the correct adapter type (`10GBASE-SR`).

3. Insert the connector on the other end of the fiber-optic cable into the connector on the Ethernet switch, or another computer system (as appropriate).

Figure 2-1 shows the fiber optic connectors for the card.

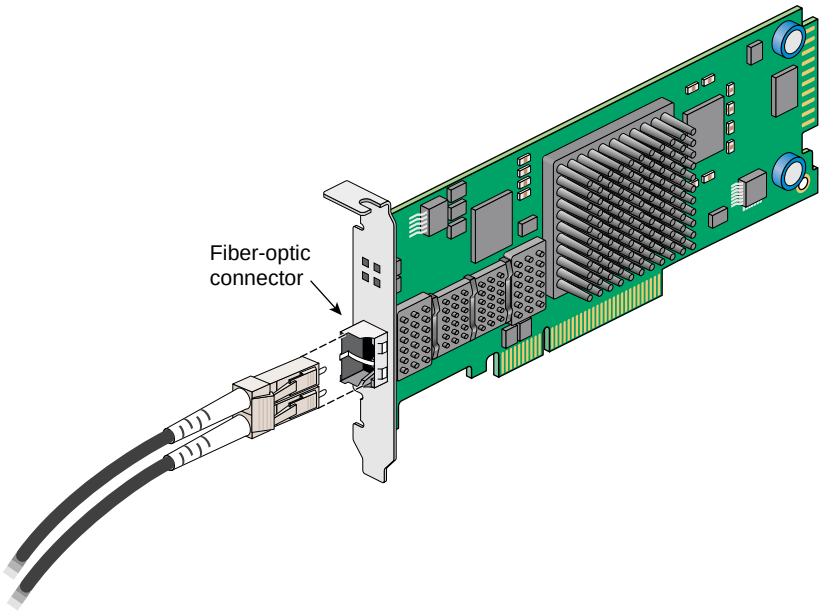

Figure 2-1 Fiber Optic Connections

For information about configuring the interfaces, see *The Network Administrators' Guide*.

Operating the Adapter

This chapter describes various issues that may occur when using the SGI 10-Gbit Ethernet PCI Express network adapter in a 10-Gbit Ethernet network. It includes the following sections:

Verifying Functionality

This section explains the following:

Using LEDs to Determine Functionality

The SGI 10-Gbit Ethernet PCI Express network adapter has light-emitting diodes (LEDs) that indicate whether the adapter is configured correctly and connected to an active Ethernet.

Table 3-1 Faceplate LEDs

Label	Color	Meaning
S	Yellow	Controlled by the adapter's firmware
L	Green	Link connectivity
R	Green	Receive (RX) traffic
T	Green	Transmit (TX) traffic

Verifying Adapter Recognition

To verify that the adapter has been recognized, do the following:

1. Use the `/sbin/lspci` command to ensure that the device has been recognized. For example, you might see one of the following:

```
[root@linux root]# /sbin/lspci |grep -i myricom
Ethernet controller: MYRICOM Inc.: Unknown device 0008
     Subsystem: MYRICOM Inc.: Unknown device 0008

[root@linux root]# /sbin/lspci |grep -i myricom
Ethernet controller: MYRICOM Inc. Myri-10G Dual-Protocol NIC (10G-PCIE-8A)
     Subsystem: MYRICOM Inc. Myri-10G Dual-Protocol NIC (10G-PCIE-8A)
```

   If `lspci` shows `Unknown device` as in the previous examples, you can use the `update-pciids` utility to update the `pciid` database. For more information, see the `update-pciids`(8) man page.

   Note: If the driver module is not yet loaded, files and commands such as `/proc/net/dev` and `/sbin/ifconfig` will not display the device.

2. If `/sbin/lsmod` does not show the `myri10ge` module, use the following command to load it:

```
[root@linux root]# /sbin/modprobe myri10ge
```

Enabling the Adapter

To enable the adapter, enter the following:

```
[root@linux root]# ifconfig IPaddress netmask netmaskvalue broadcast broadcastaddress mtu 1500|9000
```

For example:

```
[root@linux root]# ifconfig eth2 10.0.0.1 netmask 0xffffff00 broadcast 10.0.0.255 mtu 9000
```

Note: Ethernet interfaces are named `eth0`, `eth1`, and so on, always with a common prefix of `eth`.

For other systems to see the new address, you must enter the new hosts' addresses in DNS or in `host` files or NIS as required for your system.

For details, see the operating-system specific documentation about networking.

Verifying that the Adapter is Properly Configured and Enabled

To verify that the network interface is configured properly and is enabled on, enter the following on a Linux system:

```
[root@linux root]# ifconfig ethN
```

For example, for `eth2`:

```
[root@linux root]# ifconfig eth2
eth2      Link encap:Ethernet  HWaddr 00:60:DD:47:81:24
          inet addr:10.0.0.1  Bcast:10.0.0.255  Mask:255.255.255.0
          inet6 addr: fe80::260:ddff:fe47:8124/64 Scope:Link
          UP BROADCAST RUNNING MULTICAST  MTU:9000  Metric:1
          RX packets:0 errors:0 dropped:0 overruns:0 frame:0
          TX packets:6 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:0 (0.0 b)  TX bytes:468 (468.0 b)
          Interrupt:65
```

Resetting the Adapter

In the unlikely event that you need to reset the adapter, enter the following, where *N* is the adapter number:

```
# ifconfig ethN down
# ifconfig ethN up
```

Changing the Configuration

To change the configuration of your adapter, use the `ethtool`(8) command. For more information, see the `ethtool`(8) man page.

Setting MTU Sizes

The maximum transmission unit (MTU) size is controlled by the `mtu` *mtu_size* switch of the `ifconfig` command. The most common MTU sizes are 1500 bytes (standard-size Ethernet frames) and 9000 (jumbo Ethernet frames). The adapter supports an MTU size of up to 9000 bytes. Configuring the adapter to use jumbo frames can increase network throughput and reduce CPU load, but only if the network supports jumbo frames.

By default, the Myri10GE driver configures the Myri-10G NIC with a 9000-byte MTU. The 10-Gigabit Ethernet switch to which the NICs are connected must support jumbo frames; otherwise, the NICs must be configured with a 1500-byte MTU.

To configure the MTU size, follow these steps:

1. To display information about the network adapters currently installed in the system, enter the following command:

   ```
   [root@linux root]# netstat -i
   ```

   For example:

```
[root@linux root]# netstat -i
Kernel Interface table
Iface   MTU Met   RX-OK RX-ERR RX-DRP RX-OVR   TX-OK TX-ERR TX-DRP TX-OVR Flg
... other interfaces removed from output
eth2   9000   053042986      0      0      073338167      0      0      0 BMRU
```

2. To change the MTU size of the 10-Gbit Ethernet adapter, enter the following, where *N* is the number of the adapter:

   ```
   [root@linux root]# ifconfig ethN mtu mtu_size
   ```

   For example:

   ```
   [root@linux root]# ifconfig eth2 mtu 1500
   ```

3. To verify that the MTU size has been changed, enter the following:

   ```
   [root@linux root]# netstat -i
   ```

   For example:

```
[root@linux root]# netstat -i
Kernel Interface table
Iface   MTU Met   RX-OK RX-ERR RX-DRP RX-OVR   TX-OK TX-ERR TX-DRP TX-OVR Flg
... other interfaces removed from output
eth2   1500   053042986      0      0      073338167      0      0      0 BMRU
```

Troubleshooting

For information about troubleshooting, see:

http://www.myri.com/scs/README.myri10ge-linux

Performance Tuning

The default settings have been carefully chosen to maximize throughput for the SGI 10-Gigabit (Gbit) Ethernet PCI Express network adapter so that no further tuning is required. However, if you wish to experiment with the settings found on Myricom's web site at http://www.myri.com/serve/cache/511.html, it is important that you follow the information in this chapter.

This chapter discusses the following:

Jumbo Frames

Using a large maximum transmission unit (MTU) is necessary for the best 10–Gbit Ethernet performance. Generally, the bigger the MTU, the better. The driver supports MTUs as large as 9000 bytes.

Read/Write Size

Applications should read large buffers from and write large buffers to the network for the best throughput and to reduce CPU utilization.

For example, an application that uses `recv`(2) calls with 32–KB buffers will generally have better throughput than if the application were to use twice as many `recv` calls with 16-KB buffers.

Network Buffer Sizes

Normally, larger network buffers are called for with 10-Gbit Ethernet than when lower-bandwidth network interface cards are used. The following network buffer

sizes were chosen to be compatible with those recommended by the card manufacturer and required by SGI's software. To set the network buffer sizes:

1. Add or change the following entries in the `/etc/sysctl.conf` file:

```
net.core.rmem_default = 524287
net.core.rmem_max = 524287
net.core.wmem_default = 524287
net.core.wmem_max = 524287
net.ipv4.tcp_rmem = 10000000 10000000 16777216
net.ipv4.tcp_wmem = 10000000 10000000 16777216
net.ipv4.tcp_mem = 10000000 10000000 16777216
net.core.netdev_max_backlog = 300000
net.core.optmem_max = 524287
net.ipv4.tcp_timestamps = 1
net.ipv4.tcp_sack = 1
net.ipv4.tcp_tw_recycle = 1
net.ipv4.tcp_tw_reuse = 1
```

2. Make the sizes take effect:

```
[root@linux root]# sysctl -p /etc/sysctl.conf
```

TCP Time Stamps

TCP time stamps are turned on by default to greatly reduce the chance of data corruption at high throughput rates. Therefore, SGI strongly recommends their use.

Glossary

Ethernet
A communication network used to connect computers.

gigabit (Gbit)
A communication rate of 10^9 bits per second.

host
Any system connected to the network.

hostname
The name that uniquely identifies each host (system) on the network.

IP address
A number that uniquely identifies each host (system) on a TCP/IP network.

LED
Light-emitting diode, a light on a piece of hardware that indicates status or error conditions.

MAC
Medium access control, also called the physical layer.

MAC address
The physical address of the SGI 10-Gbit Ethernet Network adapter, which is distinct from the IP address.

MTU
Maximum Transmission Unit is a configuration parameter that controls the size of the Ethernet frames that the SGI 10-Gigabit Ethernet network adapter can transmit and receive.

man (manual) page

An online document that describes how to use a particular command. Also called reference page.

multiclient configuration

A TCP/IP configuration in which the system is connected via 10-Gbit Ethernet to a switch that fans out to multiple clients via 1-Gbit Ethernet.

NIS

Network Information Service is a distributed database mechanism for user accounts, host names, mail aliases, and so on.

PCI Express

Peripheral Component Interconnect Express is a high-performance I/O interconnect. Traditional PCI attributes are maintained from a usage model, but the parallel PCI bus interconnect is replaced by a highly scalable serial interface.

reference page

See *man (manual)* page.

TCP/IP

Transmission Control Protocol/Internet Protocol is a standard networking protocol

Index