About the Ethernet Gateway

This chapter describes the Ethernet gateway and includes the following sections:

- Introduction, page 10-1
- Interface Gateway Ports, page 10-1
- Configuration Options, page 10-3
- IP Fragmentation and Jumbo Packets, page 10-8
- Managing the Ethernet Gateway, page 10-8

Introduction

The Ethernet gateway is an optional expansion module for the Cisco SFS 3504 Server Switch that supports seamless connectivity between IB-enabled servers and Ethernet networks. The Ethernet gateway is the intermediary between IB-enabled servers and Ethernet devices.

The Cisco SFS 3504 Server Switch supports the 6-port 1G Ethernet gateway card, (PID number SFS-3500-ENGW-1G).

The Ethernet gateway supports flexible configuration of Layer 2 bridging. In all cases, hosts on one or more IB subnets run the IPoIB protocol stack and forward IP packets to the Ethernet gateway, which then transparently translate to IP over Ethernet. On the Ethernet side, IP over Ethernet packets are translated into IPoIB packets and then forwarded to the IB subnet.

Interface Gateway Ports

The gateway ports are the two internal ports that connect the gateway to the IB network. The gateway ports are often called internal ports. The two internal ports are 10 Gbps IB ports.

The Cisco SFS 3504 Server Switch supports a single InfiniBand switch card. Both ports from the gateway are connected to the same IB switch within the chassis. Therefore, the user is not required to specify a default port setting because the ports are both connected to the same switch.

This section describes the interface gateway ports and includes the following topics:

- Internal Gateway Ports, page 10-2
- Configuring the Optional Gateway Port for the Cisco SFS 3504 Server Switch, page 10-2
- DHCP Addressing with Ethernet Gateway, page 10-3
Internal Gateway Ports

The Cisco SFS 3504 Server Switch uses IB as the underlying fabric switch that creates a scalable server area network. The systems also seamlessly interconnect with the Ethernet resources, extending the value of IB to the rest of the network. For information regarding the supported protocols, see the Cisco SFS InfiniBand Host Drivers User Guide for Linux or the Cisco SFS InfiniBand Host Drivers User Guide for Windows.

While the external gateway ports are Ethernet ports, the internal ports are connected to the IB fabric. This is particularly important when configuring bridging. In addition to assigning a bridge-group to an external gateway and port, you can optionally choose to assign a specific internal port to a bridging group.

If you do not select a specific internal port, the second gateway port is selected for you automatically. Figure 10-1 illustrates a view of the internal ports.

For example, you may want to utilize both internal ports by assigning different bridge-groups to separate internal ports of the same gateway. Assigning bridge-group A to interface gateway 1 and bridge-group B to interface gateway 2 provides a kind of manual load balancing across two IB links within a single gateway.

Configuring the Optional Gateway Port for the Cisco SFS 3504 Server Switch

This section describes how to configure gateway ports.
The following example shows how to configure the gateway (internal) ports for the Cisco SFS 3504 Server Switch:

```
SFS-3504(config)# interface gateway 1
SFS-3504(config-if-gw-1/2)# bridge-group 3
```

**DHCP Addressing with Ethernet Gateway**

The Ethernet gateway forwards DHCP packets transparently, but for DHCP to work across the gateway, all clients and servers must follow the requirements in RFC-4390 (DHCP over IB).

DHCP must be configured correctly to work with IB. The requirements to use DHCP with IB are as follows:

- The DHCP server must use the client ID as the identifier (instead of MAC). This is a DHCP standard.
- The DHCP server must use broadcast reply.

For more information see the standards for RFC 4390 (DHCP over IB).

**InfiniBand Unique Identifier and MAC Addressing Scheme**

The Ethernet gateway uses a range of sequential unique identifiers to allocate Infiniband GUIDs, Ethernet ports, and link aggregation group MAC addresses. The range is specified as a base and the number of identifiers. The Cisco SFS 3504 gateways derive their identifiers from the chassis. They do not have any assigned unique identifiers. Each Cisco SFS 3504 gateway slot has a range of 130 identifiers assigned during manufacturing. As a result, no configuration changes are required if a gateway is replaced with a different one in the same chassis slot. However, it is possible that configuration changes are required if the gateway is moved to a different slot or to another chassis.

Under certain conditions, gateway GUID changes could require a change in configuration. For example, if there is a bridge group with a P-Key that is different from the default, the Subnet Manager configuration must change to include the new gateway in the corresponding partition. The MAC address change may also require a configuration change in some Ethernet nodes. For example, if there are some access control lists that have been programmed based on the gateway MAC, a configuration change in certain Ethernet nodes is required.

**Configuration Options**

This section describes configuration options and includes the following topics:

- Link Aggregation Trunking, page 10-4
- Layer 2 Bridging, page 10-4
- VLANs, page 10-5
- Redundancy Groups, page 10-5
- Load Balancing, page 10-5
- Layer 3 Protocols, page 10-6
- Multicast Support, page 10-6
Link Aggregation Trunking

Link aggregation, sometimes referred to as trunking, is an optional feature supported by the Ethernet gateway. It is used with Layer 2 bridging. Link aggregation allows multiple Ethernet gateway ports to merge logically into a single link. Because the full bandwidth of each physical link is available, bandwidth is not wasted by inefficient routing of traffic. As a result, the entire cluster is utilized more efficiently.

Link aggregation offers higher aggregate bandwidth to traffic-heavy servers as well as reroute capability in case of a port/cable failure.

For information on configuring link aggregation, see Chapter 13, “Configuring Link Aggregation.”

Layer 2 Bridging

This section describes Layer 2 bridging and includes the following topics:

- Bridging Multiple Subnets, page 10-4
- Subnet Forwarding, page 10-4

The Ethernet gateway acts like a Layer 2 bridge between the IB and the Ethernet. Configuring Layer 2 bridging enables the system to learn everything it needs to know about the location of nodes on the network with minimal input from the administrator. The Layer 2 bridge supports only IP traffic and forwards only IP protocol.

The administrator creates a bridge-group to bridge an IB partition to an Ethernet VLAN and to one or more IP subnets based on the IP address. Thus, IB hosts appear to be attached to an Ethernet switch.

Hosts on the IB subnet run the IPoIB protocol stack and forward IP packets to the Ethernet gateway, which then transparently translates to IP over Ethernet. On the Ethernet side, IP over Ethernet packets are translated into IPoIB packets and then forwarded to the IB subnet.

Bridging Multiple Subnets

Each bridge-group can be configured to bridge up to eight IP subnets. Additional IP subnets are added to a bridge-group one at a time; however all subnets for a bridge-group can be viewed simultaneously.

For more information, see the “Configuring Subnet Forwarding for Multiple IP Subnets” section on page 15-12.

Subnet Forwarding

Bridge-groups can contain multiple routes (forwarding rules) that point to different IP routers. The routes can belong to all IP subnets, a specific IP subnet, or a set of IP subnets.

For more information, see the “Configuring Subnet Forwarding for a Single IP Subnet” section on page 15-5.
VLANs

Each Ethernet gateway supports up to 32 VLANs. A VLAN is a way of grouping a set of switch ports together so that they form a logical network. The following list includes some key features about VLANs:

- Ethernet bridge ports can be tagged or untagged
- Standard 802.1Q VLANs are supported
- Up to 32 VLANs are supported per gateway
- Static port based VLANs are supported
- A full range of VLAN IDs are supported
- One VLAN is mapped to one IB partition. See the “Partitions P_Key” section on page 8.

Redundancy Groups

Multiple bridge-groups are assigned into redundancy groups. A redundancy group connects an IB partition to an Ethernet VLAN. Redundancy managers run on the controller cards of each chassis in the fabric and coordinate load balancing and rerouting of traffic. The managers control the behavior of redundancy groups and ensure correct operation. In active/passive mode, only one bridge-group passes traffic and the rest of them are in standby state. In active/active mode all of the bridge-groups are forwarding.

Redundant groups can be created across multiple gateways and across multiple chassis. See Chapter 18, “Configuring Redundancy” for more information.

Load Balancing

Load balancing can be enabled on a redundancy group to enter the active/active mode. With load balancing enabled, redundancy groups pass traffic on all bridge-groups. (See Figure 10-2.) Load balance distribution is IB host based.
Layer 3 Protocols

The following list provides key features about Layer 3 protocols:

- IPv4 and ARP are supported
- IPv4 based upper layer protocols work transparently
- Other Layer 3 protocols are dropped

Multicast Support

Multicast support is implemented in hardware and functions at wire speed. This feature provides the automatic discovery of multicast groups.

IPv4 multicast is the only type that is supported by the Cisco SFS 3504 Server Switch. See the “Multicast Forwarding with the Ethernet Gateway” section on page 17-3. Multicast properties can be configured for individual bridge-groups or on multiple bridge-groups by configuring them into redundancy groups.
Chapter 10      About the Ethernet Gateway

Configuration Options

ARP

ARP maps the IP network addresses to the corresponding link layer addresses. This ensures proper host-IP address routing. For Ethernet, the corresponding link layer address is MAC. For IB, the equivalent is GID and Queue Pair QP. IP ARP relies on broadcast mechanisms. ARP over IB leverages a group multicast that all IPoIB members join. Separate multicast groups are created for each IB partition (for multiple VLANs).

This section describes areas related to ARP in a redundant configuration and includes the following topics:

- Updating ARP Entries in the Event of a Bridge-Group Failover, page 10-7
- Updating ARP Entries in the Event of a Failover, page 10-7
- Setting Your OS and Layer 3 Switch Time-Out, page 10-7

Updating ARP Entries in the Event of a Bridge-Group Failover

In a redundant Ethernet gateway configuration, (where two or more bridge-groups are added to a redundancy group), the MAC address and GID:QP of the Ethernet gateway changes when a primary bridge-group fails. A change in MAC and GID:QP leads to invalid ARP entries on the Ethernet and IB IPoIB nodes, including on any Layer 3 switches.

The Ethernet gateway keeps the ARP entries fresh by keeping track of the nodes, sending gratuitous ARPs, and updating the host and Layer 3 switch ARP entries in the event that a gateway fails.

Updating ARP Entries in the Event of a Failover

Although the Ethernet gateway updates the ARP entries on hosts and Layer 3 switches by design, circumstances may prevent the gateway from updating all ARP entries. It is important that you set appropriate time-outs for your hosts and Layer 3 switches.

If the ARP entry on the host or the Layer 3 switch cannot be updated, the entry points to the failed gateway until the OS or Layer 3 switch time-out takes effect.

Setting Your OS and Layer 3 Switch Time-Out

The expiration time (time-out) is set differently for various OS and Layer 3 switches. Verify that the time-out values are reasonable in the event that an ARP entry is not updated by the gateway.

If the time-out is unacceptably long, you must change it to a more reasonable value. Use the equation that follows to determine the ARP time-out; setting it to a very short interval may overwhelm the network with ARP requests if the subnet has many nodes.

Use the following equation to calculate the appropriate ARP rate:

- Number of nodes (both Ethernet and IB IPoIB) = \( n \)
- Divide \( n \) by the time-out value in seconds.

The result is the average number of ARP requests per second. This average ARP value should not be greater than 100.
Loop Protection

The Ethernet gateway provides loop protection to prevent broadcast loops in Layer 2 switching configurations.

There are several options that provide protection against broadcast loops. For more details, see Chapter 16, “Enabling Loop Protection.”

Partitions P_Key

Partitions enforce isolation among systems that share an IB fabric, similar to the way that VLANs or zoning create isolation. Ports or servers associated with a P_Key are partitioned from each other.

For more details regarding creating partitions, see Chapter 12, “Configuring Partitions.”

IP Fragmentation and Jumbo Packets

IP fragmentation is performed in the Ethernet gateway on IP frames forwarded from Ethernet to IB. This feature, which is always active, is implemented in hardware and functions at wire speed.

IP fragmentation is activated when an Ethernet port receives a frame with a payload between 2044 bytes and 9 Kbytes. Frames larger than 9 Kbytes are dropped.

Note

Fragmentation is performed even if the Don't Fragment bit is set in the IP header. If the Don't Fragment bit is set in the IP header, there is no ICMP message generated to the sender; this may prevent some Path MTU discovery mechanisms to discover MTU differences between the Ethernet and IB network, thus making bridging transparent. However, the TCP protocol discovers the path MTU correctly and avoids fragmentation.

Fragmentation is not performed for IB frames bridged to the Ethernet, even if the frame size is between 1500-2044 bytes. These frames are considered oversized on the Ethernet fabric. If an IB host MTU is greater than 1500 bytes, the Ethernet switches must have jumbo frames enabled to be delivered to the Ethernet hosts.

Managing the Ethernet Gateway

The Ethernet gateway module is managed as part of the integrated IB management framework. It can be managed using one of the following methods:

- Command-line interface
- Element Manager Java-based GUI (see the Cisco SFS Product Family Element Manager User Guide).
- Chassis Manager web-based GUI (see the Cisco SFS Product Family Chassis Manager User Guide).
- Standard and proprietary SNMP MIBs.