



Cisco SFS InfiniBand Ethernet Gateway User Guide

Release 2.8.0

June 2007

Americas Headquarters

Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
<http://www.cisco.com>
Tel: 408 526-4000
800 553-NETS (6387)
Fax: 408 527-0883

Text Part Number: OL-12960-01

THE SPECIFICATIONS AND INFORMATION REGARDING THE PRODUCTS IN THIS MANUAL ARE SUBJECT TO CHANGE WITHOUT NOTICE. ALL STATEMENTS, INFORMATION, AND RECOMMENDATIONS IN THIS MANUAL ARE BELIEVED TO BE ACCURATE BUT ARE PRESENTED WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED. USERS MUST TAKE FULL RESPONSIBILITY FOR THEIR APPLICATION OF ANY PRODUCTS.

THE SOFTWARE LICENSE AND LIMITED WARRANTY FOR THE ACCOMPANYING PRODUCT ARE SET FORTH IN THE INFORMATION PACKET THAT SHIPPED WITH THE PRODUCT AND ARE INCORPORATED HEREIN BY THIS REFERENCE. IF YOU ARE UNABLE TO LOCATE THE SOFTWARE LICENSE OR LIMITED WARRANTY, CONTACT YOUR CISCO REPRESENTATIVE FOR A COPY.

The Cisco implementation of TCP header compression is an adaptation of a program developed by the University of California, Berkeley (UCB) as part of UCB's public domain version of the UNIX operating system. All rights reserved. Copyright © 1981, Regents of the University of California.

NOTWITHSTANDING ANY OTHER WARRANTY HEREIN, ALL DOCUMENT FILES AND SOFTWARE OF THESE SUPPLIERS ARE PROVIDED "AS IS" WITH ALL FAULTS. CISCO AND THE ABOVE-NAMED SUPPLIERS DISCLAIM ALL WARRANTIES, EXPRESSED OR IMPLIED, INCLUDING, WITHOUT LIMITATION, THOSE OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT OR ARISING FROM A COURSE OF DEALING, USAGE, OR TRADE PRACTICE.

IN NO EVENT SHALL CISCO OR ITS SUPPLIERS BE LIABLE FOR ANY INDIRECT, SPECIAL, CONSEQUENTIAL, OR INCIDENTAL DAMAGES, INCLUDING, WITHOUT LIMITATION, LOST PROFITS OR LOSS OR DAMAGE TO DATA ARISING OUT OF THE USE OR INABILITY TO USE THIS MANUAL, EVEN IF CISCO OR ITS SUPPLIERS HAVE BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

CCVP, the Cisco logo, and the Cisco Square Bridge logo are trademarks of Cisco Systems, Inc.; Changing the Way We Work, Live, Play, and Learn is a service mark of Cisco Systems, Inc.; and Access Registrar, Aironet, BPX, Catalyst, CCDA, CCDP, CCIE, CCIP, CCNA, CCNP, CCSP, Cisco, the Cisco Certified Internetwork Expert logo, Cisco IOS, Cisco Press, Cisco Systems, Cisco Systems Capital, the Cisco Systems logo, Cisco Unity, Enterprise/Solver, EtherChannel, EtherFast, EtherSwitch, Fast Step, Follow Me Browsing, FormShare, GigaDrive, HomeLink, Internet Quotient, IOS, iPhone, IP/TV, iQ Expertise, the iQ logo, iQ Net Readiness Scorecard, iQuick Study, LightStream, Linksys, MeetingPlace, MGX, Networking Academy, Network Registrar, *Packet*, PIX, ProConnect, ScriptShare, SMARTnet, StackWise, The Fastest Way to Increase Your Internet Quotient, and TransPath are registered trademarks of Cisco Systems, Inc. and/or its affiliates in the United States and certain other countries.

All other trademarks mentioned in this document or Website are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (0705R)

Any Internet Protocol (IP) addresses used in this document are not intended to be actual addresses. Any examples, command display output, and figures included in the document are shown for illustrative purposes only. Any use of actual IP addresses in illustrative content is unintentional and coincidental.

Cisco SFS InfiniBand Ethernet Gateway User Guide
© 2007 Cisco Systems, Inc. All rights reserved.



CONTENTS

Preface ix

Audience ix

Organization ix

Conventions x

Related Documentation xi

Obtaining Documentation, Obtaining Support, and Security Guidelines xi

CHAPTER 1

Understanding the Ethernet Gateway 1-1

Introducing the Ethernet Gateway 1-1

Understanding the Interface Gateway Ports 1-1

Internal Gateway Ports 1-2

Selecting an Interface Gateway Port 1-3

Understanding the Optional Gateway Port Configuration 1-3

Understanding DHCP Addressing with Ethernet Gateway 1-3

Understanding IB Unique Identifier and MAC Addressing Scheme 1-3

6-Port Ethernet Gateways 1-4

Understanding the Configuration Options 1-4

Layer 2 Bridging 1-5

Bridging Multiple Subnets 1-5

Subnet Forwarding 1-5

Link Aggregation Trunking 1-5

VLANs 1-5

Redundancy Groups 1-6

Load Balancing 1-6

Layer 3 Protocols 1-7

Multicast Support 1-7

ARP 1-8

ARP in a Redundant Configuration 1-8

Loop Protection 1-9

Partitions P_key 1-9

IP Fragmentation and Jumbo Packets 1-9

Managing the Ethernet Gateway 1-10

CHAPTER 2

Using the Ethernet Gateway Hardware 2-1

- About the Ethernet Gateway Expansion Module 2-1
 - Bandwidth and Speed 2-2
- Installing an Ethernet Gateway 2-2
- Removing an Ethernet Gateway 2-3
- Connecting Network Devices 2-4
 - About Connectors 2-4
 - Connecting to Other Devices 2-4
- Managing the Ethernet Gateway 2-4
 - Enabling/Disabling the Ethernet Gateway Using GUI 2-5
 - Enabling/Disabling the Ethernet Gateway Using CLI 2-5
- Interpreting Ethernet LEDs 2-6
 - Ethernet Gateway LEDs 2-6
 - Ethernet Gateway Interface LEDs 2-6
 - Ethernet Gateway Port LEDs 2-7
- Upgrading Gateways 2-7
 - Upgrading the Software for a Gateway 2-7

CHAPTER 3

Configuring Partitions 3-1

- Partition Overview 3-1
 - What is a Partition? 3-1
 - Default Partition 3-1
 - P_Keys 3-1
 - How Partitions Work 3-2
 - Partition Utility 3-2
 - Membership Types 3-2
 - Full Membership 3-2
 - Limited Membership 3-2
 - About the Default Partition 3-3
 - Selecting Valid P_Key Values 3-3
 - Use P_KEYS with Bridge-Groups 3-3
- Creating Partitions 3-4
 - Creating a P_Key Partition ID 3-4
 - Specifying Partition Members and Membership Types 3-4

CHAPTER 4

Understanding and Configuring Link Aggregation 4-1

- About Link Aggregation 4-1
 - Benefits 4-1

Supported Features	4-1
Load Distribution Types	4-2
Configuration Overview	4-2
Configuring Link Aggregation	4-3
Planning the Number of Link Aggregation Groups	4-3
Creating a Link Aggregation Interface	4-3
Setting the Distribution Type	4-4
Setting the Distribution on Your Ethernet Switch	4-4
Assigning Ports to the Link Aggregation Group	4-4
(Optional) Creating Additional Link Aggregation Groups	4-5
Managing Link Aggregation Groups	4-5
Disabling a Link Aggregation Group	4-5
Deleting a Link Aggregation Group	4-6

CHAPTER 5

Bridging InfiniBand and Ethernet Networks 5-1

About Bridge-Groups	5-1
Understanding Bridge-Group Options	5-2
Attributes of the Bridge Port	5-2
Bridge-Group Configuration Overview	5-3
Configuring Bridge-Groups with Default Pkey and No VLAN	5-4
Creating a Single Bridge-Group	5-4
Assigning Ethernet Ports to the Bridge-Group	5-5
Assigning Ethernet Ports without Link Aggregation	5-5
Assigning a Bridge-Group to a Link Aggregation Group	5-5
Assigning the Gateway Slot for the Bridge-Group	5-6
Configuring Broadcast Forwarding	5-6
Configuring Bridge-Groups with a Non-Default Pkey and a VLAN	5-6
Prerequisites	5-6
Creating a Single Bridge-Group	5-6
Assigning Ethernet Ports and VLAN ID to the Bridge-Group	5-7
Assigning Ethernet Ports and VLAN ID without Link Aggregation	5-7
Assigning a Bridge-Group and VLAN ID to a Link Aggregation Group	5-7
Assigning the Gateway Slot for the Bridge-Group	5-8
Associating the Bridge-Group to a Non-Default Partition	5-8
Configuring Broadcast Forwarding	5-8
Adding Bridge-Groups to the Larger Network Configuration	5-8

CHAPTER 6

Understanding and Configuring Subnet Forwarding 6-1

- About Subnet Forwarding 6-1
 - IB Host Subnet Forwarding 6-1
 - Ethernet Host Subnet Forwarding 6-1
 - Understanding Same Subnet Forwarding 6-1
 - Sample Scenario Description 6-2
 - Process Overview 6-2
 - Process Details 6-3
 - Understanding Remote Subnet Forwarding 6-4
 - Sample Scenario Description 6-4
 - Process Overview 6-4
 - Process Details 6-5
- Configuring Subnet Forwarding for a Single IP Subnet 6-6
 - Configuring Subnet Forwarding without Remote Access 6-6
 - Configuring Subnet Forwarding for Remote Access 6-8
 - Configuring Bridge-Groups with Multiple Remote Access Points 6-11
- Configuring Subnet Forwarding for Multiple IP Subnets 6-14
 - Bridging Multiple IP Subnets without Remote Access 6-14
 - Bridging Multiple IP Subnets with Remote Access 6-16
- (Optional) Enabling Broadcast Forwarding 6-19
- Viewing the Configurations 6-19
 - Showing the Bridge-Groups 6-19
 - Showing the Trunk Groups 6-20

CHAPTER 7

Enabling Loop Protection 7-1

- Loop Protection Overview 7-1
 - Loop Protection Options 7-1
 - Disabling Broadcast Forwarding 7-1
 - Self-Cancelling ARP Requests 7-1
 - Delayed Proxy ARP Transaction 7-1
 - ARP Packet Painting 7-2
- Configuring Loop Protection 7-2
 - Disabling Broadcast Forwarding 7-2
 - Enabling Broadcast Forwarding 7-2
 - Disabling ARP Packet Signatures 7-3

CHAPTER 8

Configuring Multicast Forwarding 8-1

- Multicast Overview 8-1

IPv4 Multicast	8-1
Understanding Multicast Forwarding and IGMP Snooping	8-1
Multicast Groups	8-1
Multicast Forwarding	8-2
Internet Group Management Protocol	8-2
IGMP Snooping	8-2
Layer 2 Multicast Addresses	8-3
Multicast Forwarding with the Ethernet Gateway	8-3
Multicast/Broadcast Configuration in Bridge-Groups	8-4
Creating a Multicast Group	8-4
Enabling Multicast Forwarding	8-4
Disabling Multicast Forwarding	8-5

CHAPTER 9
Configuring Redundancy 9-1

Introduction	9-1
Configuring Redundant Ethernet Gateways	9-2
Verify Redundancy Group Configuration	9-5
Verify Bridge Group Configuration	9-5

APPENDIX A
Acronyms and Abbreviations A-1

INDEX



Preface

This preface describes who should read the *Cisco SFS InfiniBand Ethernet Gateway User Guide*, how it is organized, and its document conventions. It contains the following sections:

- [Audience, page ix](#)
- [Organization, page ix](#)
- [Conventions, page x](#)
- [Related Documentation, page xi](#)
- [Obtaining Documentation, Obtaining Support, and Security Guidelines, page xi](#)

Audience

The intended audience for this document is the administrator responsible for configuring and managing Ethernet gateways and related software and equipment. It is expected that the user has experience working with equipment such as server switches, the Subnet Manager, Ethernet gateways, Host Channel Adapters, and software drivers.

Organization

This publication is organized as follows:

Chapter	Title	Description
Chapter 1	Understanding the Ethernet Gateway	This chapter introduces the Ethernet gateway.
Chapter 2	Using the Ethernet Gateway Hardware	This chapter describes how to use the Ethernet gateway hardware.
Chapter 3	Configuring Partitions	This chapter describes how to configure partitions in IB nodes.
Chapter 4	Understanding and Configuring Link Aggregation	This chapter describes and shows how to configure the link aggregation feature on the Ethernet gateway.
Chapter 5	Bridging InfiniBand and Ethernet Networks	This chapter describes how to bridge InfiniBand and Ethernet networks.

Chapter	Title	Description
Chapter 6	Understanding and Configuring Subnet Forwarding	This chapter describes how to configure Subnet forwarding.
Chapter 7	Enabling Loop Protection	This chapter describes loop protection options and configuring loop protection.
Chapter 8	Configuring Multicast Forwarding	This chapter describes multicast forwarding, IGMP snooping, and creating a multicast group.
Chapter 9	Configuring Redundancy	This chapter describes how to configure the Ethernet gateway for redundancy..
Appendix A	Acronyms and Abbreviations	This appendix defines the acronyms and abbreviations that are used in this publication.

Conventions

This document uses the following conventions:

Convention	Description
boldface font	Commands, command options, and keywords are in boldface . Bold text indicates Chassis Manager elements or text that you must enter as-is.
<i>italic font</i>	Arguments in commands for which you supply values are in <i>italics</i> . Italics not used in commands indicate emphasis.
Menu1 > Menu2 > Item...	Series indicate a pop-up menu sequence to open a form or execute a desired function.
[]	Elements in square brackets are optional.
{ x y z }	Alternative keywords are grouped in braces and separated by vertical bars. Braces can also be used to group keywords and/or arguments; for example, { interface <i>interface</i> type }.
[x y z]	Optional alternative keywords are grouped in brackets and separated by vertical bars.
string	A nonquoted set of characters. Do not use quotation marks around the string or the string will include the quotation marks.
screen font	Terminal sessions and information the system displays are in screen font.
boldface screen font	Information you must enter is in boldface screen font .
<i>italic screen font</i>	Arguments for which you supply values are in <i>italic screen font</i> .
^	The symbol ^ represents the key labeled Control—for example, the key combination ^D in a screen display means hold down the Control key while you press the D key.

Convention	Description
< >	Nonprinting characters, such as passwords are in angle brackets.
!, #	An exclamation point (!) or a pound sign (#) at the beginning of a line of code indicates a comment line.

Notes use the following convention:



Note

Means *reader take note*. Notes contain helpful suggestions or references to material not covered in the manual.



Caution

Cautions use the following convention:

Means *reader be careful*. In this situation, you might do something that could result in equipment damage or loss of data.

Related Documentation

For additional information related to the Ethernet Gateway, see the following documents:

- *Cisco InfiniBand Host Channel Adapter Hardware Installation Guide*
- *Cisco SFS Product Family Chassis Manager User Guide*
- *Cisco SFS Product Family Element Manager User Guide*
- *Cisco SFS InfiniBand Host Drivers User Guide for Linux*
- *Cisco SFS InfiniBand Host Drivers User Guide for Windows*

Obtaining Documentation, Obtaining Support, and Security Guidelines

For information on obtaining documentation, obtaining support, providing documentation feedback, security guidelines, and also recommended aliases and general Cisco documents, see the monthly *What's New* in Cisco Product Documentation, which also lists all new and revised Cisco technical documentation, at:

<http://www.cisco.com/en/US/docs/general/whatsnew/whatsnew.html>



CHAPTER 1

Understanding the Ethernet Gateway

The following sections appear in this chapter:

- [Introducing the Ethernet Gateway, page 1-1](#)
- [Understanding the Interface Gateway Ports, page 1-1](#)
- [Understanding the Configuration Options, page 1-4](#)
- [IP Fragmentation and Jumbo Packets, page 1-9](#)
- [Managing the Ethernet Gateway, page 1-10](#)

Introducing the Ethernet Gateway

The Ethernet Gateway is an optional expansion module for the SFS 3012R or the SFS 3001 Server Switches that supports seamless connectivity between IB-enabled servers and Ethernet networks. The Ethernet gateway is the intermediary between IB-enabled servers and Ethernet devices.



Note

TopspinOS release 2.3.0 and later versions do not support the legacy four-port Ethernet gateway. If your chassis includes a legacy gateway, replace the gateway with the six-port EX gateway.

The Ethernet gateway allows 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, the opposite is also true.

Understanding the 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.

Internal Gateway Ports

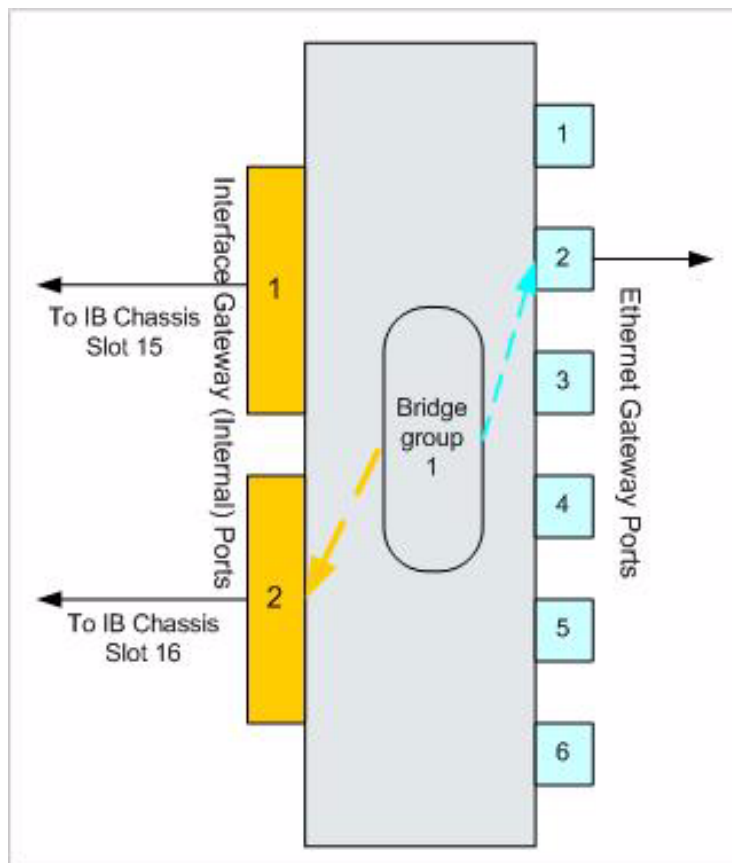
The SFS 3012R and the SFS 3001 Server Switches use IB as the underlying fabric switch that creates a scalable server area network. The systems also seamlessly interconnect with the Ethernet resources, extending the value of IB to the rest of the network. For information regarding the supported protocols, 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 to understand 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. The second gateway port is then connected to the switch card in slot 16.

The following figure illustrates a view of the internal ports.

Figure 1-1 Internal IB Ports (Interface Gateway Ports)



Selecting an Interface Gateway Port

If you decide to configure a specific internal port for a bridge-group, it is not necessarily important which internal port you assign to a bridge-group, as long as the chassis has all of the IB switch cards. For an SFS 3012R Server Switch, where it is possible that slot 15 could be vacant, you must select port 2, which is connected to the IB switch in slot 16.

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 switches within a single gateway.

Understanding the Optional Gateway Port Configuration

When configuring the gateway (internal) ports, you can use the slot number alone, or you can specify the slot number and a specific internal port number.

By default the following command would configure an Ethernet gateway in slot 6 to internal port 2 (port 2 is connected to the switch card in slot 16).

```
SFS-3012R (config)# interface gateway 6
```

You can optionally change the default of the internal port.

For example, the internal gateway configuration of an Ethernet gateway in slot 6 / internal port 1 of the chassis would be:

```
SFS-3012R (config)# interface gateway 6/1
```

Understanding DHCP Addressing with Ethernet Gateway

Ethernet gateway forwards DHCP packets transparently, but in order 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 following are the requirements to use DHCP with IB:

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

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

Understanding IB Unique Identifier and MAC Addressing Scheme

The following section describes the IB GUID and MAC addressing.

A GUID refers to an EUI-64 number. An EUI-48 number is used, and 2 bytes of 0 are inserted into the middle to create an EUI-64 number.

For example:

EUI-48 00:05:ad:01:12:34

becomes

EUI-64 00:05:ad:00:00:01:12:34

- A MAC address is an EUI-48 number
- An IB GUID is an EUI-64 number

6-Port Ethernet Gateways

32 numbers are assigned per card, by operations.

32 numbers are needed:

- 6 Gigabit Ethernet ports +
- Gateway * 4 (node + 2 ports + 1 rsv) +
- 6 Trunk Ports + 1 Debug Ethernet Port +
- 15 Reserved

The 5 least significant bits of the GUID base number should be 0.

The base number is the GUID number taken from the bar code label on the card.

Gigabit Ethernet Port 1: base number (EUI-48)

Gigabit Ethernet Port 2: base number + 1 (EUI-48)

Gigabit Ethernet Port 3: base number + 2 (EUI-48)

Gigabit Ethernet Port 4: base number + 3 (EUI-48)

Gigabit Ethernet Port 5: base number + 4 (EUI-48)

Gigabit Ethernet Port 6: base number + 5 (EUI-48)

Gateway Node: base number + 6 (EUI-64)

Gateway Port1: base number + 7 (EUI-64)

Gateway Port2: base number + 8 (EUI-64)

Gateway Rsvd: base number + 9 (EUI-64)

Debug Ethernet Port: EUI-48 MAC address: use base number (minus middle 0s) + 10 (EUI-48)

Trunk Port 1: base number + 16 (EUI-48)

Trunk Port 2: base number + 17 (EUI-48)

Trunk Port 3: base number + 18 (EUI-48)

Trunk Port 4: base number + 19 (EUI-48)

Trunk Port 5: base number + 20 (EUI-48)

Trunk Port 6: base number + 21 (EUI-48)

Understanding the Configuration Options

This section describes configuration options.

Layer 2 Bridging

The Ethernet gateway acts like a Layer 2 bridge between IB and 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 to one or more IP subnets based on IP address. 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, the opposite is also true.

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 6-14.

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 set of IP subnets.

For more information, see the [“Configuring Subnet Forwarding for a Single IP Subnet”](#) section on page 6-6.

Link Aggregation Trunking

Link aggregation, sometimes referred to as trunking, is an optional feature available on 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 logically combines multiple links into a single one, to support a larger data stream than a single link and related objects or operations.

Link aggregation offers:

- Higher aggregate bandwidth to traffic-heavy servers
- Reroute capability in case of a port/cable failure

For information on configuring link aggregation, see [Chapter 4, “Understanding and Configuring Link Aggregation”](#).

VLANs

Each Ethernet gateway supports up to 32 Virtual LANs (VLANs). A VLAN is a way of grouping a set of switch ports together so that they form a logical network.

- Ethernet bridge ports can be tagged or untagged

- Standard 802.1Q VLANs are supported
- Up to 32 VLANs can be 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 9.

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.

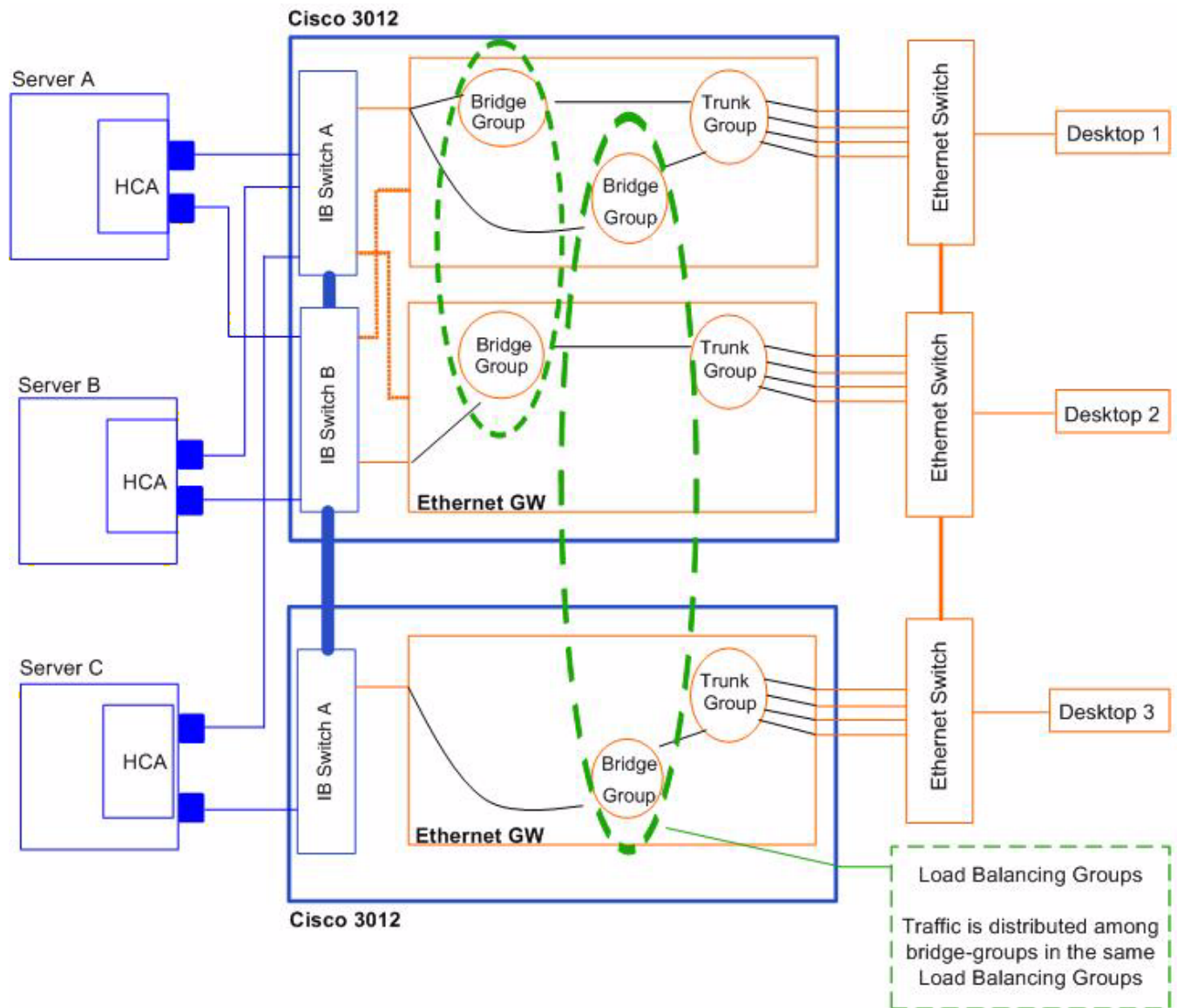
Redundant groups can be created across multiple gateways and across multiple chassis. See [Chapter 9, “Configuring Redundancy”](#) for more information.

Load Balancing

Load balancing can be enabled on a redundancy group to enter active/active mode.

Unlike redundancy groups, load balancing groups pass traffic on all bridge-groups.

Redundancy groups with load balancing enabled can be created using multiple gateways across multiple chassis. Load balance distribution is IB host based.

Figure 1-2 Bridge-Group and Redundancy Group Configuration

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. See the [“Multicast Forwarding with the Ethernet Gateway” section on page 8-3](#). Multicast properties can be configured for individual bridge-groups or on multiple bridge-groups by configuring them into redundancy groups.

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)

ARP in a Redundant Configuration

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 any Layer 3 switches.

The Ethernet gateway keeps the ARP entries fresh by:

- keeping track of the nodes
- sending gratuitous ARPs
- 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, the mechanism is a best effort and 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 below 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 value should not be greater than 50-100.

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:

- [Disabling Broadcast Forwarding, page 7-1](#)
- [Self-Cancelling ARP Requests, page 7-1](#)
- [Delayed Proxy ARP Transaction, page 7-1](#)
- [ARP Packet Painting, page 7-2](#)

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 complete information regarding creating partitions, see [Chapter 3, “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 and 9k bytes. Frames larger than 9k 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 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, the Ethernet switches must have jumbo frames enabled in order 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 through 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.



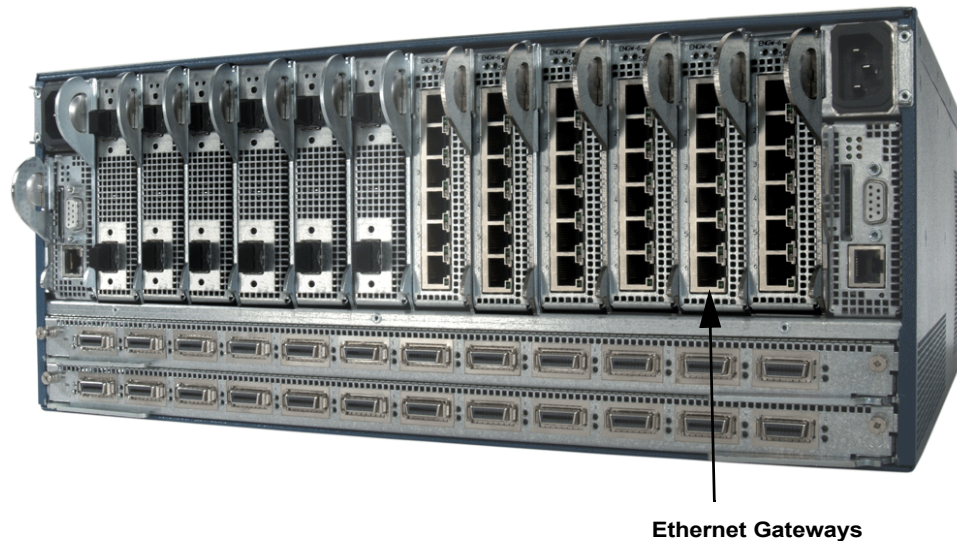
CHAPTER 2

Using the Ethernet Gateway Hardware

The following sections appear in this chapter:

- [About the Ethernet Gateway Expansion Module, page 2-1](#)
- [Installing an Ethernet Gateway, page 2-2](#)
- [Removing an Ethernet Gateway, page 2-3](#)
- [Connecting Network Devices, page 2-4](#)
- [Managing the Ethernet Gateway, page 2-4](#)
- [Interpreting Ethernet LEDs, page 2-6](#)
- [Upgrading Gateways, page 2-7](#)

Figure 2-1 *Ethernet Gateways in the SFS 3012R Server Switch*



About the Ethernet Gateway Expansion Module

The Ethernet gateway is designed for expansion and ease-of-use. The flexible design allows for the following:

- Adding a new Ethernet gateway card to expand current capacity

- Hot-swapping a card with another like card
- Replacing an older cards with a new, more efficient, and higher capacity card
- Removing a defective card

The exchange of modules in and out of the chassis may be performed without bringing the chassis down, or disrupting other network configurations.

Always ground yourself before touching any removable hardware components to avoid damage from ESD. Grounding is performed by attaching a wrist-strap to the chassis, in the manner recommended by the wrist-strap manufacturer.

Figure 2-2 **Ethernet EX Gateway with 6-ports**



Bandwidth and Speed

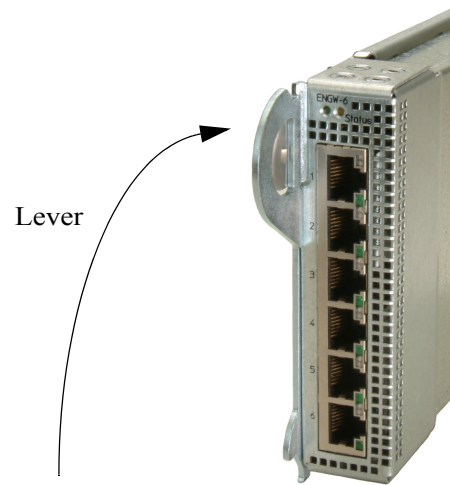
- The Ethernet EX gateway has 6 GigE ports at wire speed or 12 Gbps aggregate throughput.
- The two internal ports are 10 Gbps IB ports.

Installing an Ethernet Gateway

To install an Ethernet gateway card in a chassis, perform the following steps:

-
- | | |
|---------------|--|
| Step 1 | If a blanking panel is already in the expansion bay, remove it. |
| Step 2 | Make sure the ejector lever is completely open and insert card into the open slot until the card is firmly seated. |
| Step 3 | Close the ejector lever completely. |

Figure 2-3 *Installing an Ethernet Gateway*

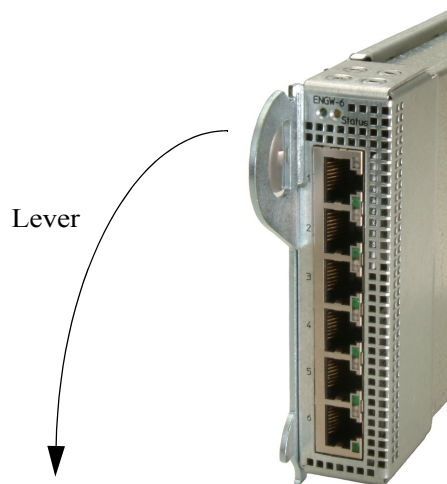


Removing an Ethernet Gateway

To remove an Ethernet gateway from the chassis, perform the following steps:

-
- Step 1** Make sure the chassis is secure.
 - Step 2** Properly ground yourself to avoid potential electrostatic discharge and damage to the card.
 - Step 3** Face the back of the chassis, place one or more fingers in the ejector lever to the left of the card you wish to remove.
 - Step 4** Pull the lever slowly out and down.
The lever pivot point is on the bottom of the card. As you pull the lever down, the card should start to come out of the chassis bay.

Figure 2-4 *Removing an Ethernet Gateway*



- Step 5** Pull the lever towards you. When the lever is all the way down, pull the card out of the chassis bay.
- Step 6** Install the blanking panel into the empty chassis bay to ensure appropriate cooling.
-

Connecting Network Devices

This section describes how to connect network devices.

About Connectors

The Ethernet gateway expansion module supports the Ethernet gateway module: RJ-45 jacks for standard unshielded twisted pair connections.

Connecting to Other Devices

To connect to other workstations, routers, or switches using Ethernet, you can use either straight-through or cross-over UTP cabling. The Ethernet gateway auto-detects the kind of cable that is connected.

Managing the Ethernet Gateway

You can manage the Ethernet gateway using one of three methods:

- Command Line Interface (CLI) - a text-based interface accessible through a direct serial connection, Telnet over IP, or SSH over IP.
- Element Manager (GUI) - A graphical interface installed on a workstation, accessible over IP.
- Chassis Manager (Web-based GUI).

Enabling/Disabling the Ethernet Gateway Using GUI

To enable or disable the Ethernet gateway using the Element Manager, perform the following steps:

-
- Step 1** Launch Element Manager.
- Step 2** Double-click an Ethernet gateway card.
- The Card tab appears.
- The type of card is automatically detected. The AdminType radio button already has the appropriate card type selected, such as en4port1G.
- Step 3** Enable or disable the radio button.
- a. Click the **up** radio button in the Enable/Disable Card section to enable the card.
 - b. Click the **down** radio button in the Enable/Disable Card section to disable the card.
- Step 4** Click the **Apply** button.
- Step 5** Click the **Close** button.
- Step 6** Double-click the gateway again. The Card tab appears. Check the Current Card Status to determine that the card has come online. It may take up to two minutes for the card to initialize.
- Step 7** Check the LEDs on the Ethernet gateway module(s) and connected devices for indications of physical connections. See the [“Interpreting Ethernet LEDs” section on page 2-6](#).
- Step 8** (Optional) If you are configuring an SFS 3012R Server Switch, repeat this process for all the Ethernet gateway cards you use.
-

Enabling/Disabling the Ethernet Gateway Using CLI

To enable the Ethernet gateway using the CLI, perform the following steps:

-
- Step 1** Start a CLI session.
- Step 2** Enter privileged EXEC mode.
- ```
SFS-3012R > enable
SFS-3012R #
```
- Step 3** Enter global configuration mode.
- ```
SFS-3012R # configure
SFS-3012R (config)#
```
- Step 4** Specify the card to be configured by entering the card number.
- ```
SFS-3012R # card 2
SFS-3012R (config-card-2)#
```
- The type of card to be enabled is automatically detected. For example, the en6port1G represents the 6-port 1G Ethernet.
- Step 5** Enter the **no shutdown** command.
- ```
SFS-3012R (config-card-2)# no shutdown
SFS-3012R (config-card-2)#
```

Step 6 Exit to privileged EXEC mode with the **Cntrl-Z** command.

```
SFS-3012R (config-card-2) # ^Z
SFS-3012R #
```

Interpreting Ethernet LEDs

This section describes Ethernet LEDs.

Ethernet Gateway LEDs

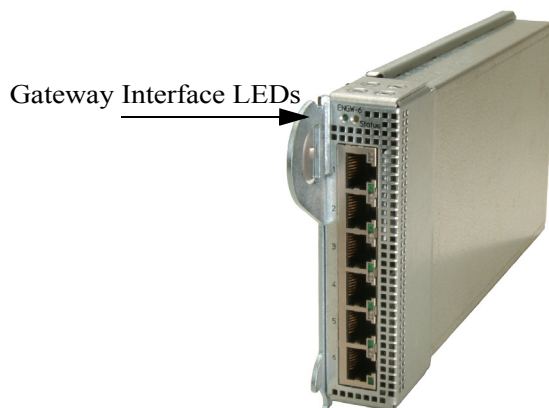
The Ethernet gateway has LEDs for the gateway and for the ports.

- [Ethernet Gateway Interface LEDs, page 2-6](#)
- [Ethernet Gateway Port LEDs, page 2-7](#)

Ethernet Gateway Interface LEDs

Ethernet gateway interface LEDs appear on the Ethernet gateway. [Figure 2-5](#) displays the Ethernet interface LEDs.

Figure 2-5 *Ethernet Gateway Interface LEDs*



[Table 2-1](#) explains the Ethernet gateway interface LED indications.

Table 2-1 *Ethernet Gateway Interface LED Indications*

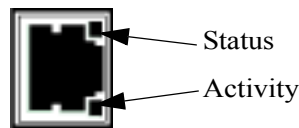
LED State	Indication
Yellow LED on/Green LED off	Auxiliary power available. Gateway is either in the process of booting up, or has been disabled.
Green LED on/Yellow LED off	Gateway is enabled and running without errors.

Table 2-1 Ethernet Gateway Interface LED Indications

LED State	Indication
Green LED on/Yellow LED on	An error has occurred.
Yellow LED off/Green LED off	Indicates an error, such as no auxiliary power to the gateway.

Ethernet Gateway Port LEDs

Ethernet Port LEDs are adjacent to each Ethernet port. They indicate the port status and activity.

Figure 2-6 Ethernet Port LEDs

LED State	Indication
Status LED off	No physical link.
Status LED solid	Successful physical link.
Activity LED off	No logical link.
Activity LED blinking	Traffic runs over the port.
Activity LED solid	Logical link established but traffic does not yet run.

Upgrading Gateways

This section describes how to upgrade gateways.

Upgrading the Software for a Gateway

To upgrade the OS for the entire chassis, see the appropriate hardware documentation for your product. However, if you do not need to update the chassis, but have a gateway with mismatched software, use the following steps:

Step 1 Verify that you have a .img file on the chassis by entering the **dir image** command.

If you do not have a .img file, you must copy an uncompressed image to the chassis controller before upgrading the gateway. See the appropriate hardware documentation for your product.

```
SFS-3012R > enable
SFS-3012R # dir image
=====
Existing Boot-Images on System
=====
slot date-created          size      file-name
-----
1   Thu Oct  7 03:11:39 2004 33065427 Topspin360-TopspinOS-2.1.0-build157.img
1   Thu Oct  7 03:31:26 2004 4359168  TopspinOS-2.1.0/build157
```

Step 2 Activate the image. Once you have verified that you have an image that can be installed on the chassis (a.img file), you must install the image before it can be used.

- a. In the privileged EXEC mode, enter **install image:image-file-name.img**

```
SFS-3012R # install image:Topspin-360-TopspinOS-2.1.0-build497.img
*****operation completed successfully
```

Step 3 Specify the new boot image to be used.

- a. Find the name of your new system image by entering the **dir image** command. Copy the name of the new image.
- b. Enter **boot-config primary-image-source**, then paste the copied image file name.

```
SFS-3012R # configure
SFS-3012R (config)# boot-config primary-image-source TopspinOS-2.1.0/build497
SFS-3012R (config)#
```

Step 4 Reboot the gateway(s) that you want to upgrade to the new image. Use the **shutdown** and **no shutdown** commands to reboot a single card instead of the whole chassis.

This example shows how to use the **shutdown** command.

```
SFS-3012R (config)# card 12
SFS-3012R (config-card-12)# shutdown
```

This example shows how to use the **no shutdown** command.

```
SFS-3012R # configure
SFS-3012R (config)# card 12
SFS-3012R (config-card-12)# no shutdown
```



CHAPTER 3

Configuring Partitions

The following sections appear in this chapter:

- [Partition Overview, page 3-1](#)
- [Creating Partitions, page 3-4](#)

Partition Overview

For more detailed information regarding Partitions, see the *Cisco SFS Product Family Element Manager User Guide*.

What is a Partition?

A Partition defines a set of IB nodes that are permitted to communicate with one another. Each node may be part of multiple partitions so that a system administrator can define overlapping partitions as the situation requires.

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.

Default Partition

All nodes in the IB fabric are automatically members of the default partition (ff:ff). When left as the default, all members can communicate with all other members. However, additional partitions can be created to isolate communication.

P_Keys

A Partition Key value (p_key) is the value that is assigned to a partition. To configure the Ethernet gateway, you need to associate the internal IB ports of the gateway with a p_key, and configure the host IB port to use that p_key.

Upon creation, the p_key value is technically a 15-bit number. However, after the p_key is created and the membership type has been established, the entire value becomes 16-bits. The MSB displays the type of membership (0 = Limited member, 1 = Full member).

How Partitions Work

Partitions work by requiring that packets contain a Partition Key (P_Key). The P_Key must match a P_Key that is stored in the receiver's P_Key table, or be discarded. A P_Key is a value that is carried in packets and stored in the channel adapters; they are used to determine the membership in a partition.

Essentially, a partition is a collection of channel adapter ports that are allowed to communicate with one another. Ports may be members of multiple partitions simultaneously. Ports in different partitions are unaware of each other's presence.

Enforcement of partitions is handled primarily by the source and destination channel adapters.

Partition Utility

Partitions are used to create isolation between ports. However, when configuring the Ethernet gateway, partitions are used to bridge to ports that are tagged with a VLAN.

Membership Types

A partition contains a group of members, and partition memberships allows greater control over communication within that group.

There are two types of partition memberships:

- Full Membership
- Limited Membership

Full Membership

Partition members with full-membership can communicate with both full and limited members.

A full-membership partition member can communicate with all other partition members, including other full members, as well as limited members. A limited-membership partition member cannot communicate with other limited-membership partition members. However, a limited partition member can communicate with a full member.

Limited Membership

By configuring a partition member to have a limited membership, a partition can contain a node that communicates with all of the clients nodes, but those nodes cannot use that partition to communicate with each other.

A limited-membership partition member cannot communicate with other limited-membership partition members. However, a limited partition member can communicate with a full member.

Full Membership	Limited Membership
A full-membership partition member can communicate with all other partition members, including other full members, as well as limited members.	A limited-membership partition member cannot communicate with other limited-membership partition members. However, a limited partition member can communicate with a full member.

About the Default Partition

The Subnet Manager automatically configures a default partition, which is always p_key ff:ff. The default partition controls all connected ports, and by default, everything is a full-member of the default partition. The default p_key cannot be altered or deleted as it is the controlling mechanism that manages the configuration of all the partitions.

Selecting Valid P_Key Values

See the *Cisco SFS Product Family Element Manager User Guide* for details on how p_key values are read.

You can choose your own p_key values if you have knowledge of how the values are used, or you can simply choose your values from the list in the following table:

00:01	00:11
00:02	00:12
00:03	00:13
00:04	00:14
00:05	00:15
00:06	00:16
00:07	00:17
00:08	00:18
00:09	00:19
00:10	00:20

Use P_KEYS with Bridge-Groups

Once you have created partitions, create bridge-groups (see the [“Creating a Single Bridge-Group” section on page 5-4](#)).

Assign bridge-groups to the partitions according to the level of isolation you want to enforce. For example, if you are using partitions and want to allow one bridge-group to fail over to another, you must assign both bridge-groups to the same partition. Bridge-groups in different partitions will not have access to each other.

Creating Partitions

This section describes how to create partitions.

Creating a P_Key Partition ID



Note

Partitions are created on the IB switch and should have been configured previously if they are to be used (see the *Cisco SFS Product Family Element Manager User Guide*).

Default partitions are configured automatically. The members of a default partition include all connected ports, and provide full membership. However, to create separation between traffic, you must configure specific partitions.

Specifying Partition Members and Membership Types



Note

Partitions are created on the IB switch, and should previously have been configured if they are going to be used (see the *Cisco SFS Product Family Element Manager User Guide*).

- Step 1** Locate the GUID of the IB node that you wish to add to the newly created partition.
- There are different ways to find the GUID of the IB node. It depends of the OS type and IB stack running on the node. For more details see your *Cisco SFS InfiniBand Host Drivers User Guide for Linux* or the *Cisco SFS InfiniBand Host Drivers User Guide for Windows*, as appropriate.
- Step 2** Locate the GUID of the Ethernet gateway internal port so you can add it to the newly created partition. See the [“Understanding the Interface Gateway Ports” section on page 1-1](#).
- a. Log into the Server Switch.
 - b. Enter the **show ib-agent summary** command.
- ```
SFS-3012R > enable
SFS-3012R # show ib-agent summary
```
- Step 3** Add the following information for partition members:
- The **ib sm subnet-prefix** command
  - The subnet-prefix that is to be partitioned.
  - The **p\_key** command
  - The current p\_key value
  - The **partition-member** command
  - The GUID of the node that you want to add to the partition.
  - The Internal IB port number that is to be added to the partition.
  - The membership type of the partition member (full-member or limited-member) see the [“Membership Types” section on page 3-2](#).
- In the following example, two gateways are being added to the same partition.

**Note**

The internal port being used is port 2, which is connected to the switch in slot 16. See the [“Internal Gateway Ports” section on page 1-2](#).

This is an example of two IB members in a single partition

```
SFS-3012R (config)# ib sm subnet-prefix 255.255.0.0 p_key 00:01 partition-member
00:05:ad:00:00:01:60:44 2 full-member
SFS-3012R (config)# ib sm subnet-prefix 255.255.0.0 p_key 00:01 partition-member
00:05:ad:00:00:00:04:a0 2 full-member
SFS-3012R (config)# exit
```

This is an example of two IB members in two separate partitions

```
SFS-3012R (config)# ib sm subnet-prefix 255.255.0.0 p_key 00:01 partition-member
00:05:ad:00:00:00:02:30 2 full-member
SFS-3012R (config)# ib sm subnet-prefix 255.255.0.0 p_key 00:02 partition-member
00:05:ad:00:00:00:04:a0 2 full-member
SFS-3012R (config)# exit
```





## CHAPTER 4

# Understanding and Configuring Link Aggregation

---

The following sections appear in this chapter:

- [About Link Aggregation, page 4-1](#)
- [Configuration Overview, page 4-2](#)
- [Configuring Link Aggregation, page 4-3](#)
- [Managing Link Aggregation Groups, page 4-5](#)

## About Link Aggregation

Link aggregation, also called trunking, is an optional feature available on the Ethernet gateway, and is used with Layer 2 Bridging. Link aggregation allows multiple ports to merge logically in 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 logically combines multiple links to support a larger data stream than a single link and related objects or operations into a single one.

## Benefits

Link aggregation offers the following benefits:

- Higher aggregate bandwidth to traffic-heavy servers
- Reroute capability in case of a single port/cable failure

## Supported Features

- Standard 802.3ad link aggregation is supported
- Static link aggregation group configuration
- One link aggregation group can be assigned to one bridge-group or to multiple bridge-groups
- Six link aggregation groups are supported for each Ethernet gateway
- You can use VLAN tagging with different VLAN IDs.
- Seven different frame distribution types are supported. See the [“Load Distribution Types” section on page 4-2](#).

- Each link aggregation group can carry up to 32 VLANs
- A link aggregation group cannot span multiple gateways

## Load Distribution Types

When configuring link aggregation on a switch that is link aggregation aware, you must select a type of distribution for the switch by using the distribution-type command.

A link aggregation Ethernet switch is required in order to use link aggregation on the gateway. It is not required that the distribution type be the same on the Ethernet switch and on the Ethernet gateway, but it is recommended to achieve the same load distribution in both directions.

| Distribution Type | Function                                                                                                                                                                                                                                                                                                                                                                                                    |
|-------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| dst-ip            | Load distribution is based on the destination IP address. Packets to the same destination are sent on the same port, but packets to different destinations are sent on different ports in the channel.                                                                                                                                                                                                      |
| dst-mac           | IB hosts do not have a MAC address, so load distribution is based on the LID address of the IB node and MAC address of Ethernet node. Packets to the same destination are sent on the same port, but packets to different destinations are sent on different ports in the channel.                                                                                                                          |
| src-dst-ip        | Load distribution is based on the source logic gate (XOR) destination IP address.                                                                                                                                                                                                                                                                                                                           |
| src-dst-mac       | IB hosts do not have a MAC address, so load distribution is based on the source logic gate (XOR) LID and MAC address.                                                                                                                                                                                                                                                                                       |
| src-ip            | Load distribution is based on the source IP address. Packets to the same destination are sent on the same port, but packets to different destinations are sent on different ports in the channel.                                                                                                                                                                                                           |
| src-mac           | IB hosts do not have a MAC address, so load distribution is based on the source-LID address of the incoming packet. Packets from different hosts use different ports in the channel, but packets from the same host use the same port in the channel.                                                                                                                                                       |
| round robin       | Round Robin is a load balancing algorithm that distributes load in a circular fashion, thereby creating an evenly distributed load.<br><br>When using redundancy groups and load balancing, selecting the Round Robin distribution can increase performance in many cases. Even with a topology that contains as few as one Ethernet host, the performance could benefit from using this distribution type. |

## Configuration Overview

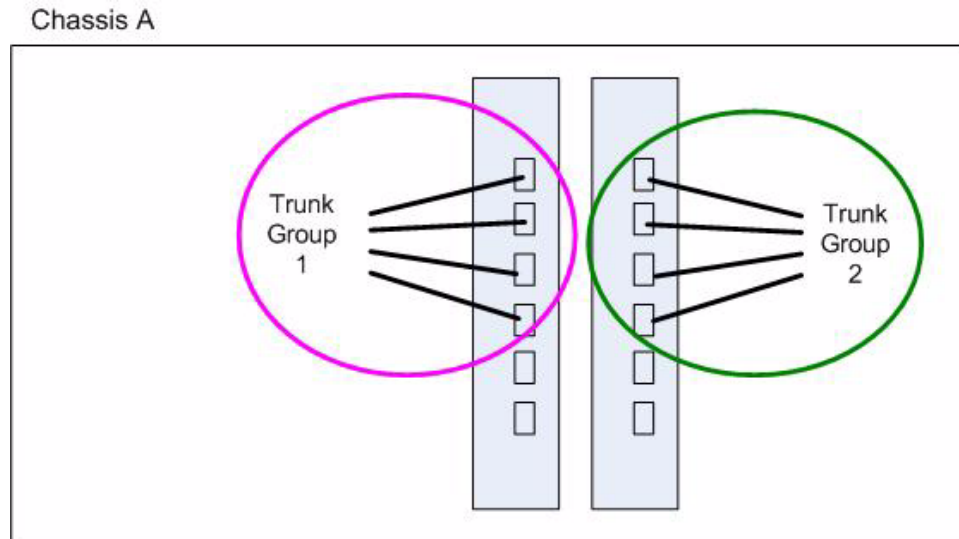
Create trunk groups to configure link aggregation between two or more ports on a single gateway. Follow the sections listed below:

- (Optional) Create a Partition ID (P\_Key). See the [“Creating Partitions” section on page 3-4](#).
- [“Creating a Link Aggregation Interface” section on page 4-3](#)
- [“Setting the Distribution Type” section on page 4-4](#)

- “Setting the Distribution on Your Ethernet Switch” section on page 4-4
- “Assigning Ports to the Link Aggregation Group” section on page 4-4
- “(Optional) Creating Additional Link Aggregation Groups” section on page 4-5
- “Disabling a Link Aggregation Group” section on page 4-5
- “Deleting a Link Aggregation Group” section on page 4-6

The following diagrams show a logical example of trunk groups.

**Figure 4-1** Create Trunk Groups to Enable Link Aggregation



## Configuring Link Aggregation

This section describes how to configure link aggregation.

### Planning the Number of Link Aggregation Groups

The number of link aggregation groups that you can create is limited by the number of gateways that you have.

### Creating a Link Aggregation Interface

Using link aggregation is not required but is recommended for maximum performance.

- Step 1** Specify the trunk interface, and assign a number to the trunk group. The trunk group ID can be any number between 1 - 256.

```
SFS-3012R # config
SFS-3012R (config#) interface trunk 1
SFS-3012R (config-if-trunk)#
```

**Step 2** Enable the trunk group.

```
SFS-3012R (config-if-trunk) # enable
```

**Step 3** (Optional) Assign a name to the trunk group.

```
SFS-3012R (config-if-trunk) # name eth-grp-1
SFS-3012R (config-if-trunk) #
```

---

## Setting the Distribution Type

If you are using link aggregation, you must set the type of distribution for the trunk group.

The default distribution is src-dist-mac.

Select a type of distribution that is appropriate for the gateway and your Ethernet switch.

For more information, see the [“Load Distribution Types” section on page 4-2](#).

```
SFS-3012R (config-if-trunk) # distribution-type src-mac
SFS-3012R (config-if-trunk) #
```

## Setting the Distribution on Your Ethernet Switch

**Step 1** Verify that your Ethernet switch that connects to the gateway supports link aggregation.

**Step 2** Configure the distribution of the corresponding ports on your Ethernet switch. For best performance, use the same distribution-type as you have set for the link aggregation group.

---

## Assigning Ports to the Link Aggregation Group

If you are using link aggregation, you must specify the ports that belong to a particular trunk group.

**Step 1** Specify the interface and the ports that belong to the first trunk.

```
SFS-3012R (config) # interface ethernet 2/1-2/4
```

**Step 2** Enter the trunk group ID or the name of the new trunk.

```
SFS-3012R (config-if-ether-2/1-2/4) # trunk-group 1
```

**Step 3** Return to global configuration mode.

```
SFS-3012R (config-if-ether-2/1-2/4) # exit
```

---



## (Optional) Creating Additional Link Aggregation Groups

**Step 1** Add an additional link aggregation group on another gateway in the same chassis.

You cannot add ports from two separate gateways to the same trunk group.

```
SFS-3012R # config
SFS-3012R (config#)interface trunk 2
SFS-3012R (config-if-trunk)# enable
SFS-3012R (config-if-trunk)# name eth-grp-2
SFS-3012R (config-if-trunk)# distribution-type src-mac
SFS-3012R (config)# interface ethernet 3/1-3/4
SFS-3012R (config-if-ether-3/1-3/4)# trunk-group 2
```

**Step 2** Return to global configuration mode.

```
SFS-3012R (config-if-ether-3/1-3/4)# exit
SFS-3012R (config)#
```

## Managing Link Aggregation Groups

This section describes how to manage link aggregation groups.

### Disabling a Link Aggregation Group

Use the following steps to disable a link aggregation group without removing it from the configuration. The link aggregation group can be re-enabled later.

**Step 1** Start a CLI session.

**Step 2** Enter privileged EXEC mode.

```
SFS-3012R > enable
SFS-3012R #
```

**Step 3** Enter global configuration mode.

```
SFS-3012R # configure
SFS-3012R (config)#
```

**Step 4** Specify the trunk interface and the number of the trunk group.

```
SFS-3012R # interface trunk 1
SFS-3012R (config-if-trunk)#
```

**Step 5** Disable the trunk group.

```
SFS-3012R (config-if-trunk)# disable
SFS-3012R (config-if-trunk)#
```

## Deleting a Link Aggregation Group

To delete or destroy a trunk:

---

**Step 1** Start a CLI session, unless you have already done so.

**Step 2** Enter privileged EXEC mode.

```
SFS-3012R > enable
SFS-3012R #
```

**Step 3** Enter global configuration mode.

```
SFS-3012R # configure
SFS-3012R (config) #
```

**Step 4** Specify the trunk interface and the number of the trunk group.

```
SFS-3012R # interface trunk 1
SFS-3012R (config-if-trunk) #
```

**Step 5** Disable the trunk group.

```
SFS-3012R (config-if-trunk) # no trunk 1
SFS-3012R (config-if-trunk) #
```

The trunk group is deleted.

**Step 6** Exit to privileged EXEC mode with the **Cntrl-Z** command.

```
SFS-3012R (config-if-trunk) # ^Z
SFS-3012R #
```

---



## CHAPTER 5

# Bridging InfiniBand and Ethernet Networks

---

This chapter describes configuring bridging between IB and Ethernet networks with the CLI.

The following sections appear in this chapter:

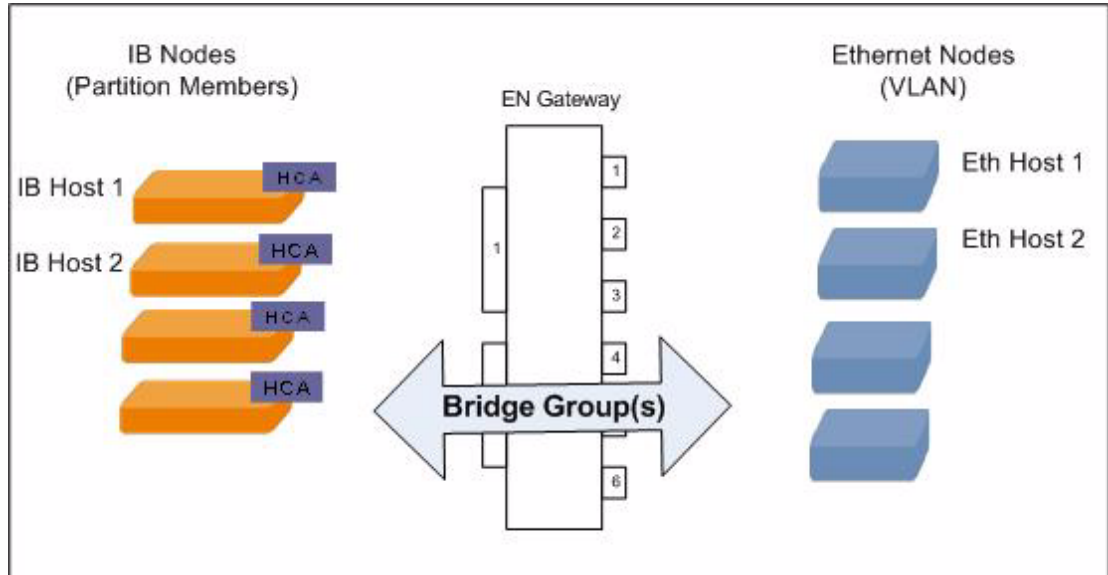
- [About Bridge-Groups, page 5-1](#)
- [Bridge-Group Configuration Overview, page 5-3](#)
- [Configuring Bridge-Groups with Default Pkey and No VLAN, page 5-4](#)
- [Configuring Bridge-Groups with a Non-Default Pkey and a VLAN, page 5-6](#)
- [Adding Bridge-Groups to the Larger Network Configuration, page 5-8](#)

## About Bridge-Groups

A bridge-group is an entity, that runs on the Ethernet gateway and enables the bridging of one IPoIB partition to one VLAN.

The Ethernet gateway acts like a Layer 2 bridge between IB and Ethernet. Each Ethernet gateway must be configured for Layer 2 bridging (with or without link aggregation and redundancy groups).

Configuring Layer 2 bridging enables the system to learn everything it needs to know about the location of nodes on the network with very little input from the administrator. The bridge-group bridges only IP protocol and drops all others.

**Figure 5-1** *Bridge-Groups Connect IB Networks to Ethernet Networks*

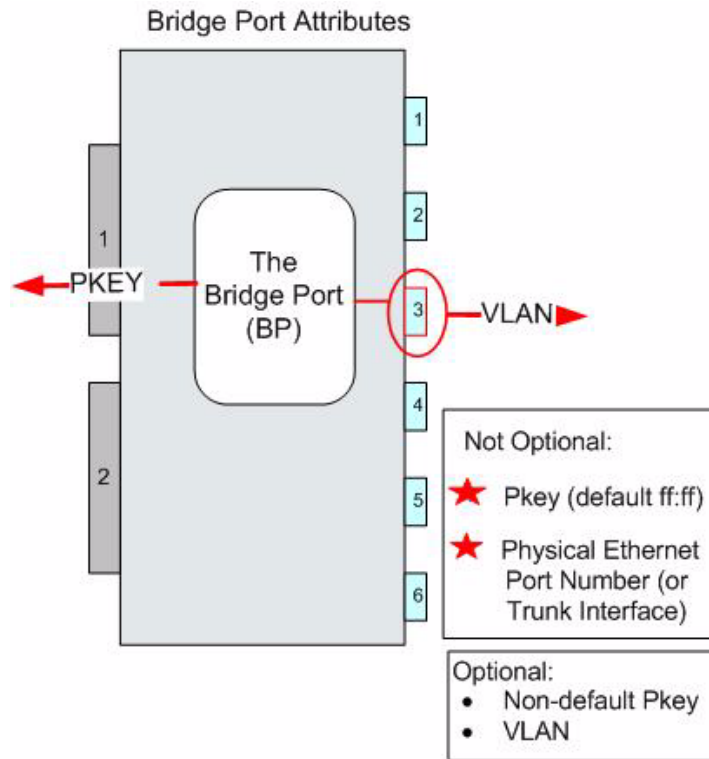
## Understanding Bridge-Group Options

The parameters that you define for the bridge-group determine the way that traffic is handled between the IB network and the Ethernet network.

### Attributes of the Bridge Port

The bridge port has certain attributes that are always attached to the bridge port entity, and certain attributes that are optional to the bridge port.

Attributes of the bridge port must be created or deleted with the bridge-group. For example, the physical port attribute is assigned at the time the bridge-group is created, and cannot be changed without deleting the bridge-group.

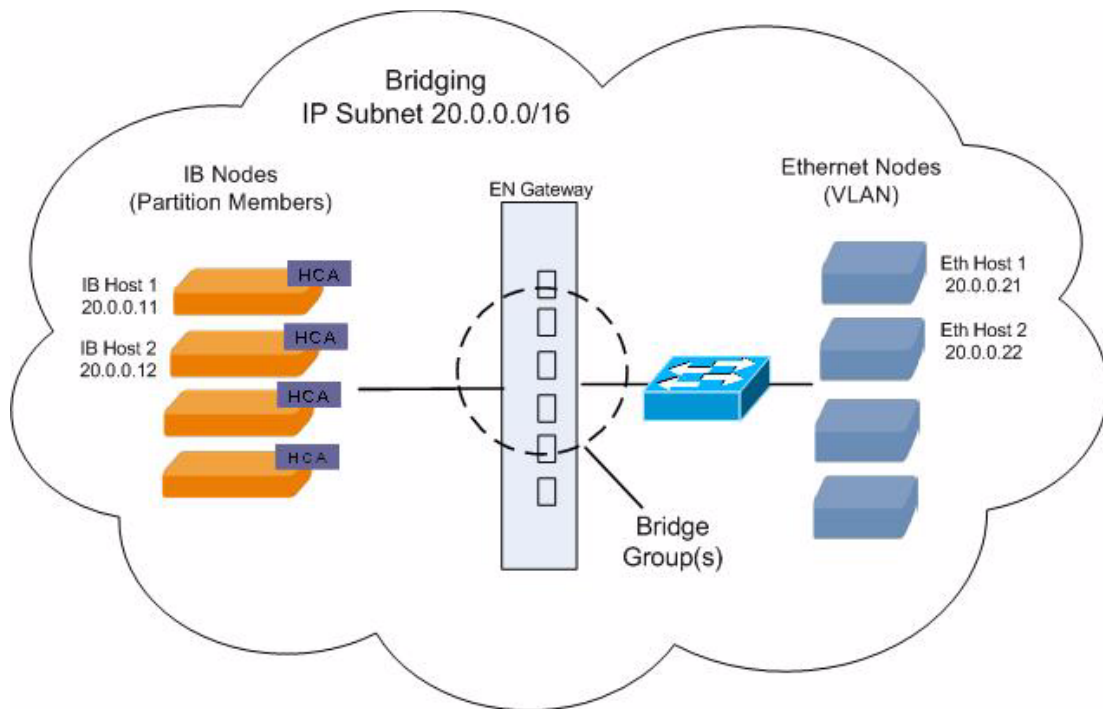
**Figure 5-2 Required and Optional Bridge Port Attributes**

## Bridge-Group Configuration Overview

See the following sections in the sequence listed for information on configuring bridging between the Ethernet gateway and an Ethernet switch:

- [“Creating Partitions” section on page 3-4](#)
- [“Configuring Link Aggregation” section on page 4-3](#)
- [“Configuring Bridge-Groups with Default Pkey and No VLAN” section on page 5-4](#)
- [“Creating a Single Bridge-Group” section on page 5-4](#)
- [“Associating the Bridge-Group to a Non-Default Partition” section on page 5-8](#)
- [“Assigning a Bridge-Group to a Link Aggregation Group” section on page 5-5](#)
- [“Assigning Ethernet Ports and VLAN ID to the Bridge-Group” section on page 5-7](#)

The diagram in [Figure 5-3](#) shows an example of two bridge-groups created on two separate gateways in a single chassis.

**Figure 5-3** Create a Bridge-Group to Bridge a Single Broadcast Domain

## Configuring Bridge-Groups with Default Pkey and No VLAN

This section describes how to configure bridge-groups with default Pkey and no VLAN.

### Creating a Single Bridge-Group

Use the following steps to create one or more bridge-groups.

- Step 1** The default partitions is created automatically and all bridge-groups are members.



**Note** Do not create a partition unless you wish you create non-default partitions.

- Step 2** (Optional) [“Configuring Link Aggregation” section on page 4-3.](#)

```
SFS-3012R > enable
SFS-3012R # configure terminal
SFS-3012R (config)# interface trunk 1
SFS-3012R (config-if-trunk)# enable
SFS-3012R (config-if-trunk)# interface ethernet 3/1-3/6
SFS-3012R (config-if-ether-3/1-3/6)# trunk-group 1
SFS-3012R (config-if-ether-3/1-3/6)# exit
```

- Step 3** Create the bridge-group by entering the **bridge-group** command, assigning a number and assigning a subnet.
- Step 4** Enter **bridge-group bridge-group# subnet-prefix address of subnet-prefix to be bridged and length**

```
SFS-3012R (config)# bridge-group 1 subnet-prefix 20.0.0.0 16
```

**Step 5** (Optional) Assign a name to the bridge-group.

```
SFS-3012R (config)# bridge-group 1 name PrimaryBG
SFS-3012R (config)#
```

---

## Assigning Ethernet Ports to the Bridge-Group

You must assign Ethernet ports 1-6 to a bridge-group. This can be done either by assigning individual Ethernet ports on the gateway, or by assigning a link aggregation group (with multiple ports) to the bridge-group.

- [Assigning Ethernet Ports without Link Aggregation, page 5-5](#)
- [Assigning a Bridge-Group to a Link Aggregation Group, page 5-5](#)

### Assigning Ethernet Ports without Link Aggregation

(Optional) If you are not using link aggregation, assign individual Ethernet ports to the bridge-group.

**Step 1** Specify the type of interface that is to be assigned to the bridge-group, which is the Ethernet interface.

```
SFS-3012R (config)# interface ethernet 3/1
```

**Step 2** Specify the bridge-group.

```
SFS-3012R (config-if-ether-3/1-3/4)# bridge-group 1
```

---

### Assigning a Bridge-Group to a Link Aggregation Group

If you are using link aggregation, you must associate a bridge-group to a trunk interface.

If you are not using trunking (link aggregation), you must assign the interface ethernet port to the bridge-group.

**Step 1** Specify the type of interface that is to be assigned to the bridge-group, which in this case is the trunk interface.

```
SFS-3012R (config)# interface trunk 1
SFS-3012R (config-if-trunk)#
```

**Step 2** Specify the bridge-group.

```
SFS-3012R (config-if-trunk)# bridge-group 1
```

---

## Assigning the Gateway Slot for the Bridge-Group

Associate the bridge-group to the slot number in the chassis. By assigning the gateway slot, the internal gateway port is associated by default. Unless you specify otherwise, internal port 2 is defined. See the [“Internal Gateway Ports” section on page 1-2](#).

```
SFS-3012R (config)# interface gateway 3
SFS-3012R (config)#
```

## Configuring Broadcast Forwarding

If you are using redundancy groups, broadcast-forwarding should be configured on the redundancy group and *not* on the individual bridge-groups.

However, if you are not using redundancy groups, broadcast forwarding parameters is configured on the individual bridge-group. Enable broadcast forwarding if you are using a DHCP server.

```
SFS-3012R # config
SFS-3012R (config#) bridge-group 1 broadcast-forwarding
SFS-3012R (config#)
```

## Configuring Bridge-Groups with a Non-Default Pkey and a VLAN

This section describes how to configure bridge-groups with a non-default Pkey and a VLAN.

### Prerequisites

1. (Optional) Create a non-default partition. See the [“Creating Partitions” section on page 3-4](#). The default partitions are created automatically.
2. Create a VLAN on your Ethernet switch, if you have not already done so.

### Creating a Single Bridge-Group

Use the following steps to create one or more bridge-groups.

- Step 1** (Optional) Configure link aggregation. For more details, see the [“Configuring Link Aggregation” section on page 4-3](#).

```
SFS-3012R > enable
SFS-3012R # configure terminal
SFS-3012R (config)# interface trunk 1
SFS-3012R (config-if-trunk)# enable
SFS-3012R (config-if-trunk)# interface ethernet 3/1-3/6
SFS-3012R (config-if-ether-3/1-3/6)# trunk-group 1
SFS-3012R (config-if-ether-3/1-3/6)# exit
```



- Step 2** Create the bridge-group by entering the **bridge-group** command, assigning a number and assigning a subnet.
- Enter **bridge-group** *bridge-group# subnet-prefix address of subnet-prefix to be bridged and length*
- ```
SFS-3012R (config)# bridge-group 1 subnet-prefix 20.0.0.0 16
```
- Step 3** Assign port or link aggregation group connected to the gateway to the VLAN.
- Enable VLAN tagging on this port or link aggregation group.
-

Assigning Ethernet Ports and VLAN ID to the Bridge-Group

You must assign a minimum of 1 and a maximum of 6 Ethernet ports to a bridge-group. This can be done either by assigning individual Ethernet ports on the gateway, or by assigning a link aggregation group (with multiple ports) to the bridge-group.

- [Assigning Ethernet Ports without Link Aggregation, page 5-5](#)
- [Assigning a Bridge-Group to a Link Aggregation Group, page 5-5](#)

Assigning Ethernet Ports and VLAN ID without Link Aggregation

(Optional) If you are not using link aggregation, assign individual Ethernet ports to the bridge-group.

- Step 1** Specify the type of interface that is to be assigned to the bridge-group, which is the Ethernet interface.
- ```
SFS-3012R (config)# interface ethernet 3/1
```
- Step 2** Specify the bridge-group.
- ```
SFS-3012R (config-if-Ether-3/1-3/4)# bridge-group 1 vlan-tag 4
```
-

Assigning a Bridge-Group and VLAN ID to a Link Aggregation Group

If you are not using trunking (link aggregation), you must assign the interface Ethernet port to the bridge-group, as described in [Assigning Ethernet Ports and VLAN ID without Link Aggregation, page 5-7](#).

- Step 1** Specify the type of interface that is to be assigned to the bridge-group, which is the trunk interface in this case.
- ```
SFS-3012R (config)# interface trunk 1
SFS-3012R (config-if-trunk)#
```
- Step 2** Specify the bridge-group and VLAN ID of the tagged Ethernet switch port.
- ```
SFS-3012R (config-if-trunk)# bridge-group 1 vlan-tag 4
```
-

Assigning the Gateway Slot for the Bridge-Group

Associate the bridge-group to the slot number in the chassis. By assigning the gateway slot, the internal gateway port is associated by default. Unless you specify otherwise, internal port 2 is defined. See the [“Internal Gateway Ports” section on page 1-2](#).

```
SFS-3012R (config)# interface gateway 3  
SFS-3012R (config-if-gw-3)#
```

Associating the Bridge-Group to a Non-Default Partition

(Optional) If you have created a partition, as described in [“Creating a P_Key Partition ID” section on page 3-4](#), create the bridge-group and assign the partition key (pkey).

```
SFS-3012R (config-if-gw-3)# bridge-group 1 pkey 80:01
```

Configuring Broadcast Forwarding

If you are using redundancy groups, broadcast-forwarding should be configured on the redundancy group.

However, if you are not using redundancy groups, broadcast-forwarding parameters would be configured on the individual bridge-group. Enable broadcast forwarding if you are using a DHCP server.

```
SFS-3012R # config  
SFS-3012R (config#) bridge-group 1 broadcast-forwarding  
SFS-3012R (config#)
```

Adding Bridge-Groups to the Larger Network Configuration

If you want your IB network to reach a remote subnet (any destination outside of your bridged subnet), you must configure the ib-next-hop. See [Chapter 6, “Understanding and Configuring Subnet Forwarding”](#) for more details.



CHAPTER 6

Understanding and Configuring Subnet Forwarding

The following sections appear in this chapter:

- [About Subnet Forwarding, page 6-1](#)
- [Configuring Subnet Forwarding for a Single IP Subnet, page 6-6](#)
- [Configuring Subnet Forwarding for Multiple IP Subnets, page 6-14](#)
- [Viewing the Configurations, page 6-19](#)

About Subnet Forwarding

If you want your IB network to reach a remote subnet (any destination outside of your bridged subnet), then you must configure the `ib-next-hop`. It is important to understand that the `ib-next-hop` refers to the next hop that the `ib` nodes will take and not the next hop on the IB fabric. In most cases this is the default gateway.

IB Host Subnet Forwarding

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.

Ethernet Host Subnet Forwarding

Hosts on the Ethernet subnet forward IP over Ethernet packets to the Ethernet gateway, which then transparently translates them to IPoIB.

Understanding Same Subnet Forwarding

The diagram in [Figure 6-1](#) shows an example of the steps that occur when an ARP is sent out over the bridge within a single subnet.

For a configuration example, see the [“Configuring Subnet Forwarding without Remote Access”](#) section on page 6-6.

The following sections describe a sequential order of events in a same subnet forwarding scenario (see Figure 6-1).

Sample Scenario Description

The IPoIB host address (referred to as IP1) must send an IP packet to the Ethernet host labelled IP3 address.

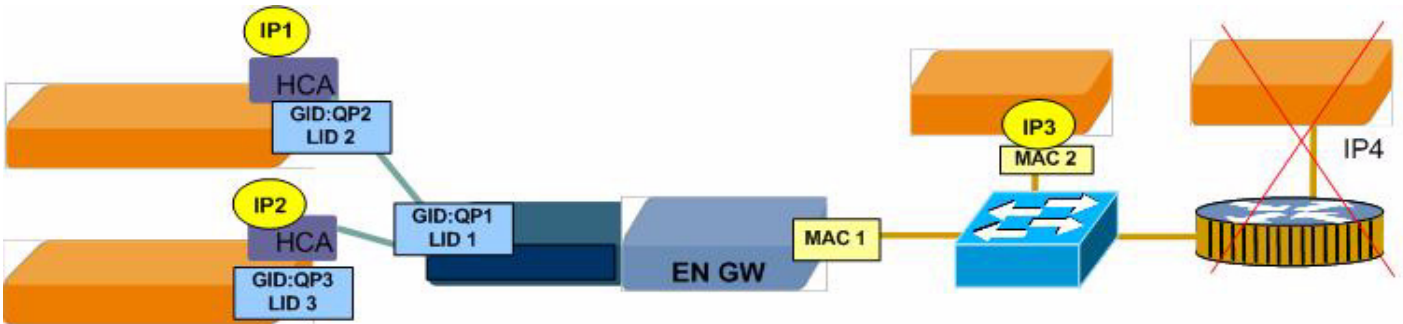

Note

The host with addresses IP1, IP2 and IP3 are on a single IP subnet.

See the following sections for additional details:

- “Process Overview” section on page 6-2
- “Process Details” section on page 6-3

Figure 6-1 Same Subnet Forwarding



Same Subnet Forwarding	Sample Trigger	Sample Action	Sample Reaction
	IP1: Arp for IP3	GW: Forward ARP replace GID with GW MAC	IP3: respond with MAC
		GW: Receive ARP resp, add to table, respond with IB GID	
	IP1: Send frame to IB GID w/ IP3	GW: Translate IB GID to IP3 MAC	IP3: receive packet

Process Overview

The high-level process of IP1 sending a packet to IP3 is as follows:

1. The IP1 host must resolve the IP3 address to map it to the link layer address.
2. Once the link layer address is resolved, the IP1 host encapsulates the IP packet into the IB packet and sends the packet.

Process Details

The single subnet forwarding events are described in sequential order (see [Figure 6-1](#)):

1. Based on the destination IP address and routing table of IP3, host IP1 determines the interface and the next hop IP address.

In this case, the interface is the IPoIB interface with IP address IP1. The next hop is the same as the final destination IP3.

- a. IP1 sends a broadcast ARP request for IP3 and waits for a reply.
 - b. All IPoIB hosts, except IP1 and the Ethernet gateway, receive the broadcast ARP request. The IPoIB hosts drop the request since the target IP address does not match their own address.
2. The Ethernet gateway creates a new ARP transaction to keep the state, then buffers the ARP request, issuing a proxy ARP request for the same IP3 target address on the Ethernet network.

The source hardware address is MAC1 on the Ethernet gateway

- a. The Ethernet host IP3 matches the target IP address, stores IP1 in its ARP cache and sends an ARP reply with its own MAC2 address.



Note The host IP3 has an ARP entry for IP1, corresponding to the MAC1 address, which is the Ethernet gateway MAC address.

- b. The ARP reply is received by the Ethernet gateway and matched with the pending ARP transaction.
3. At this point, two address entries are created in the Ethernet gateway address table:
 - IP3 -> MAC2
 - IP1 -> GID:QP2
4. To close the ARP transaction in the Ethernet gateway, an ARP reply is sent to the IP1 host with GID:QP1 as a hardware address.

IP1 host receives the ARP reply and creates an ARP entry for IP3 and link layer address GID:QP1.
5. Now the IP1 host and the Ethernet gateway have ARP entries with GID:QP link addresses; however, in order to use them, the GID needs to be resolved to a LID.
 - a. A secondary transaction sending path record request is sent to the IB subnet manager. Until the path record request is answered, the address is considered unresolved.
 - b. The Subnet Manager replies on path record and host IP1 and Ethernet gateway, which completes the address resolution.
6. IP1 now has the address of the Ethernet gateway (GID:QP1/LID1) as a link layer address for IP3, and it is ready to send the IP packet.

IP1 encapsulates the IP packet into an IB packet using Ethernet gateway LID1 and sends the packet.
7. The Ethernet gateway receives the IP packet with the IP3 destination address, recognizes that this packet is for a local host, based on subnet prefix information already configured, and queries its address table for the MAC address corresponding to IP3.

IP3 is already in the address table as a result of previous steps and is resolved to MAC.

The Ethernet gateway encapsulates the IP packet into Ethernet packet, using MAC2 as the destination address, and sends the frame.
8. The Ethernet IP3 host receives the IP packet.

Understanding Remote Subnet Forwarding

Figure 6-2 shows an example of the order of events when an ARP is sent out over the bridge to remote subnet.

Sample Scenario Description

The IPoIB host address (referred to as IP1) needs to send an IP packet to remote Ethernet host labelled IP4-3 address.

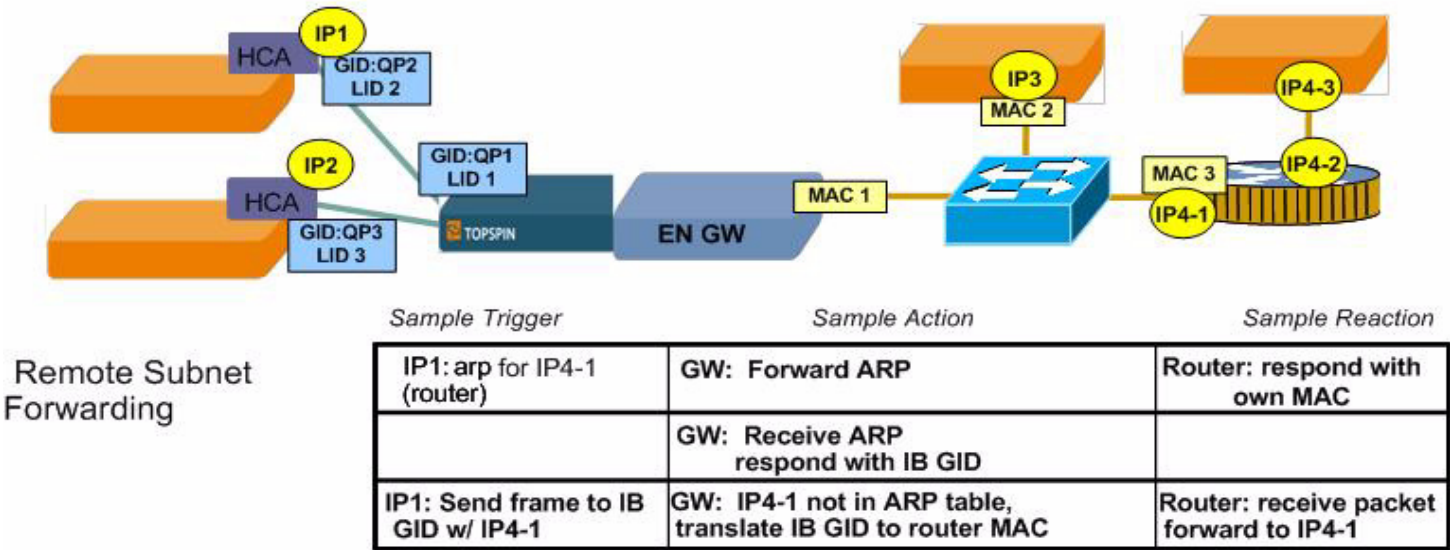


Note The hosts IP1 and IP4-3 are on different IP subnets.

See the following sections for additional details:

- “Process Overview” section on page 6-4
- “Process Details” section on page 6-5

Figure 6-2 Remote Subnet Forwarding



Process Overview

For the IB host IP1 to send an IP packet to remote Ethernet host IP4-3:

1. First the IP1 host needs to resolve IP4-1 address of the next hop such as to map it to link layer address. The IP1 host sends a packet to an L3 switch and the L3 switch forwards the packet accordingly.
2. The IP1 host encapsulates the IP packet into IB packet and sends the packet.

Process Details

The events in [Figure 6-2](#) are described in sequential order:

1. Based on destination IP address (IP4-3) and its routing table, host IP1 determines the interface and the next hop IP address.

In this case the interface is the IPoIB interface with IP address IP1. The next hop is IP4-1 address of the L3 switch, which is different from the final destination, in this example.

- a. IP1 sends a broadcast ARP request for IP4-1 and waits for a reply.
 - b. All IPoIB hosts, except IP1, and the Ethernet gateway receive the broadcast ARP request. The IPoIB hosts drop the request, since the target IP address does not match its own address.
2. The Ethernet gateway creates a new ARP transaction to keep the state, buffers the ARP request and issue, and proxy ARP request for the same IP4-1 target address on the Ethernet network.



Note

The Ethernet gateway MAC1 address is used as source hardware address.

3. L3 switch IP4-1 matches the ARP target IP address, stores IP1 in its ARP cache, and sends an ARP reply with its own MAC2 address.



Note

The L3 switch has an ARP entry for IP1 that corresponds to the MAC1 address, which is the Ethernet gateway MAC address.

4. The ARP reply is received by the Ethernet gateway and matched with the pending ARP transaction. Now two address entries are created in the Ethernet gateway address table:
 - IP4-1->MAC3
 - IP1->GID:QP2
 5. To close the ARP transaction in the Ethernet gateway, an ARP reply is sent to IP1 host with GID:QP1 as hardware address.
 - a. The IP1 host receives the ARP reply and creates an ARP entry for IP4-1 and the link layer address GID:QP1.
 - b. Now the IP1 host and Ethernet gateway have ARP entries with GID:QP link addresses, but in order to use them GID must be resolved to LID. Therefore, a secondary transaction sending path record request to the Subnet Manager. Until the path record request is answered, the address is considered unresolved.
 - c. The Subnet Manager reply on path record, host IP1 and the Ethernet gateway completes address resolution
 6. IP1 has the address of the Ethernet gateway (GID:QP1/LID1) as a link layer address for IP4-1 and is ready to send the IP packet.
 - a. IP1 encapsulates the IP packet into an IB packet using Ethernet gateway LID1 and sends the packet.
 - b. The Ethernet gateway receives the IP packet with IP4-3 destination address, recognizes that this packet is for a remote host based on subnet prefix information already configured and queries its routing table for the next hop. The result of the query is IP4-1.
 7. The Ethernet gateway looks in its address table for the IP4-1 MAC address. IP4-1 is already in the address table as a result of previous steps and is resolved to MAC.

8. The Ethernet gateway encapsulates the IP packet into Ethernet packet using MAC3 as destination address and sends the frame.
9. The L3 switch receives the IP packet and forwards it to the IP4-3.

Value	Description
Bridge-group number	Specifies the bridge-group that is being configured.
<i>next-hop-type</i>	This parameter determines through which port the route is pointing. May be either ib-next-hop or eth-next-hop .
<i>next-hop-IP</i>	Specifies the IP address of the next hop to be used in order to reach the destination subnet.
next-hop-mask	This is an optional parameter and determines the bridged subnet to which a forwarding rule belongs. If the parameter is missing, the default value is zero and the corresponding forwarding rule belongs to all bridged IP subnets. Assuming that the <i>next-hop-ip</i> address is on the same IP subnet as the bridged subnet, the <i>next-hop-mask</i> and <i>next-hop-ip</i> determines to which subnet the rule belongs to. If there are multiple matches the rule belongs to all of them.
dest <i>dest_ip</i> <i>dest_mask</i>	This parameter defines the destination IP subnet by prefix and mask. If the dest option is used, both the <i>dest_ip</i> and the <i>dest_mask</i> arguments must be provided. This is because the Ethernet gateway only supports CIDR. If this parameter is not used, the default 0/0 subnet is assumed.

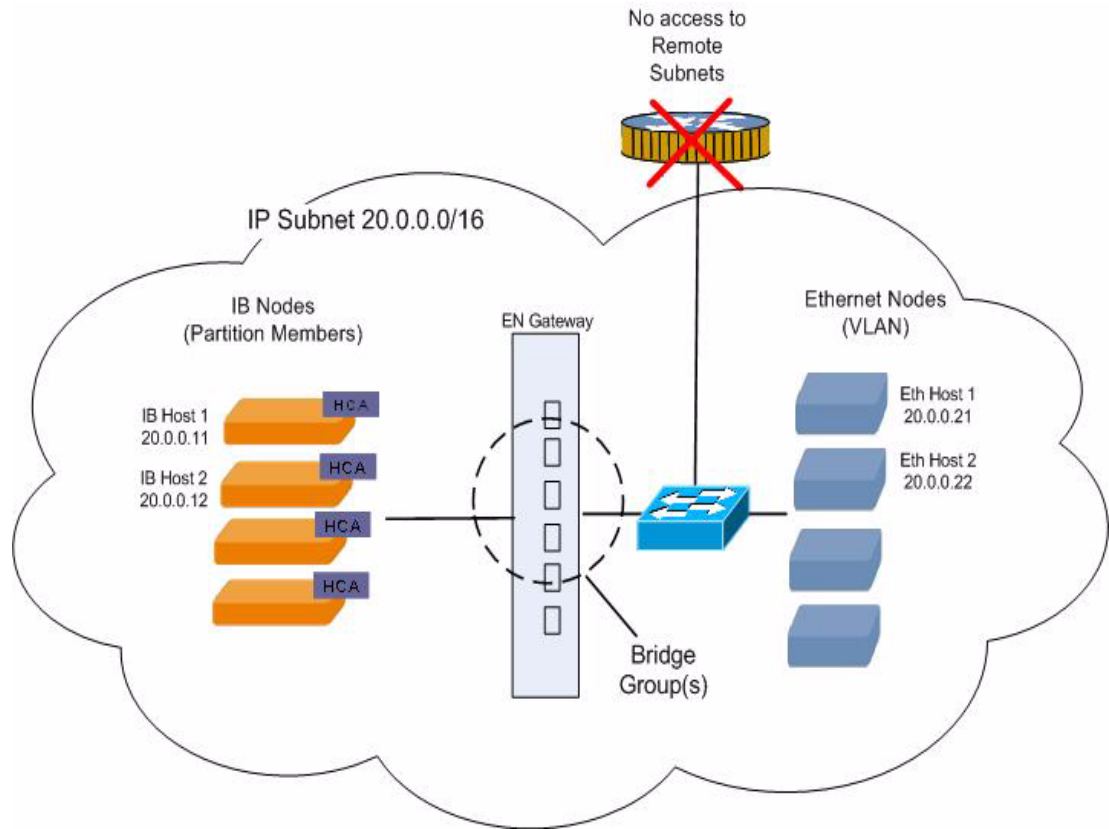
Configuring Subnet Forwarding for a Single IP Subnet

This section describes how to configure subnet forwarding for a single IP subnet.

Configuring Subnet Forwarding without Remote Access

If your IB network does not require access to remote subnets (any destination outside of your bridged subnet), then you do not need to configure a next-hop address.

Use the following sample to configure subnet forwarding for one broadcast domain. This example does not configure remote access, so an L3 switch or router is not included.

Figure 6-3 Example of No IB Next Hop Configured for Bridge Groups

- Step 1** Configure Partitions, if you are using this feature. See the [“Creating a P_Key Partition ID”](#) section on page 3-4.
- Step 2** Create a link aggregation group, if you are using this feature. See the [“Creating a Link Aggregation Interface”](#) section on page 4-3.
- Step 3** Create a bridge-group, if you have not already done so. See the [“Creating a Single Bridge-Group”](#) section on page 5-4.

The example below shows a bridge-group created without partitions or link aggregation.



Note An ib-next-hop has not been configured, as this bridge-group is not granted access outside of the broadcast domain.

```
enable
configure terminal
bridge-group 1 subnet-prefix 20.0.0.0 16
interface ethernet 3/1
bridge-group 1
interface gateway 3
bridge-group 1
exit
```

Step 4 Configure the IB Hosts.

This is an example of IB host 1

```
ifconfig ib0 20.0.0.11 netmask 255.255.0.0 broadcast 20.0.255.255
```

This is an example of IB host 2

```
ifconfig ib1 20.0.0.12 netmask 255.255.0.0 broadcast 20.0.255.255
```

Step 5 Configure the Ethernet Hosts.

This is an example of Ethernet host 1

```
ifconfig eth2 20.0.0.21 netmask 255.255.0.0 broadcast 20.0.255.255
```

This is an example of Ethernet host 2

```
ifconfig eth2 20.0.0.22 netmask 255.255.0.0 broadcast 20.0.255.255
```

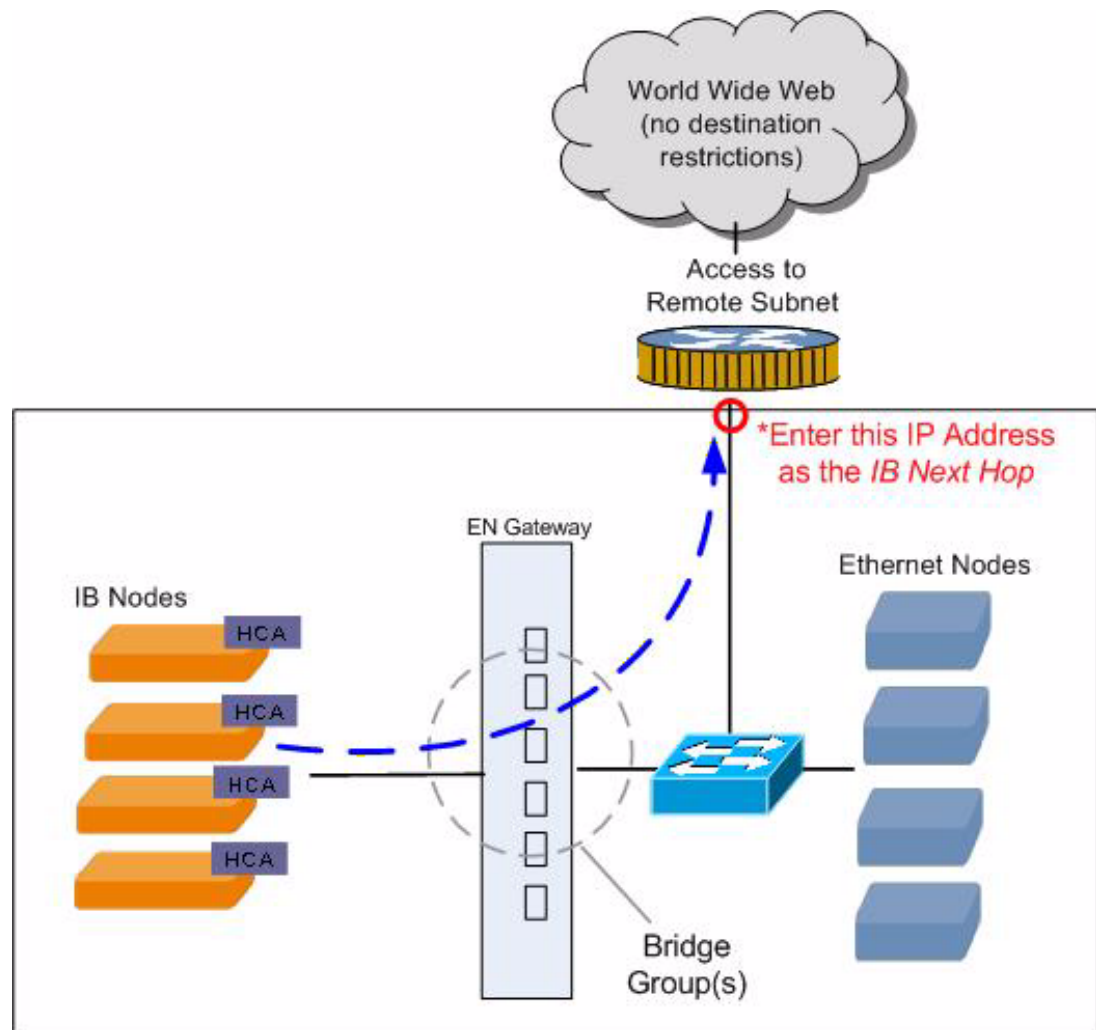
Step 6 Configure the Ethernet switch, if you have not already done so.

The ethernet switch must be configured to match the configuration of the bridge-group. For more details see [Chapter 5, “Bridging InfiniBand and Ethernet Networks”](#) that describes how to configure link aggregation and VLAN tagging if required.

Configuring Subnet Forwarding for Remote Access

If you want your IB network to reach a remote subnet (any destination outside of your bridged subnet), then you must configure the `ib-next-hop`. The `ib-next-hop` refers to the next hop that the *ib* nodes take and not the next hop on the *ib* fabric. In most cases this is the default gateway.

The configuration in [Figure 6-4](#) illustrates a bridge-group with an `ib-next-hop`. In this example, the `ib-next-hop` is the default gateway. The destination for the bridge-group is left as the default; that is, no specific destinations are configured so the default 0/0 subnet is assumed.

Figure 6-4 Example of Bridge-Groups with a Single IB Next Hop (No Destination Restriction)

- Step 1** Configure Partitions, if you are using this feature. See the [“Creating a P_Key Partition ID”](#) section on page 3-4.
- Step 2** Create a link aggregation group, if you are using this feature. See the [“Creating a Link Aggregation Interface”](#) section on page 4-3.
- Step 3** Create a bridge-group, if you have not already done so. See the [“Creating a Single Bridge-Group”](#) section on page 5-4.
- Locate the address of your ib-next-hop. This is the next hop that the IB hosts will take outside of the bridged network. In most cases this is the default gateway.
 - Assign the next hop for the IB network. The example below shows a bridge-group created without partitions or link aggregation.

**Note**

An ib-next-hop has been configured, as this bridge-group is granted access outside of the broadcast domain.

```
enable
configure terminal
bridge-group 1 subnet-prefix 20.0.0.0 16
bridge-group 1 ib-next-hop 20.0.0.1 next-hop-mask 255.255.0.0
interface ethernet 3/1
bridge-group 1
interface gateway 3
bridge-group 1
exit
```

Step 4 Configure the IB hosts.

This is an example of IB host 1

```
ifconfig ib0 20.0.0.11 netmask 255.255.0.0 broadcast 20.0.255.255
route add default gw 20.0.0.1
```

This is an example of IB host 2

```
ifconfig ib1 20.0.0.12 netmask 255.255.0.0 broadcast 20.0.255.255
route add default gw 20.0.0.1
```

Step 5 Configure the Ethernet hosts.

Example of Ethernet host 1

```
ifconfig eth2 20.0.0.21 netmask 255.255.0.0 broadcast 20.0.255.255
route add default gw 20.0.0.1
```

This is an example of Ethernet host 2

```
ifconfig eth2 20.0.0.22 netmask 255.255.0.0 broadcast 20.0.255.255
route add default gw 20.0.0.1
```

Step 6 Configure remote Ethernet hosts.

This is an example of remote Ethernet host 1

```
ifconfig eth1 100.0.0.11 netmask 255.255.0.0 broadcast 100.0.255.255
route add default gw 100.0.0.1
```

This is an example of remote Ethernet host 2

```
ifconfig eth3 100.0.0.12 netmask 255.255.0.0 broadcast 100.0.255.255
route add default gw 100.0.0.1
```

Step 7 Configure the Ethernet switch.

The Ethernet switch and the IP router must be configured to match the bridge-group configuration. This includes but is not limited to VLAN, link aggregation, and router configuration.

**Note**

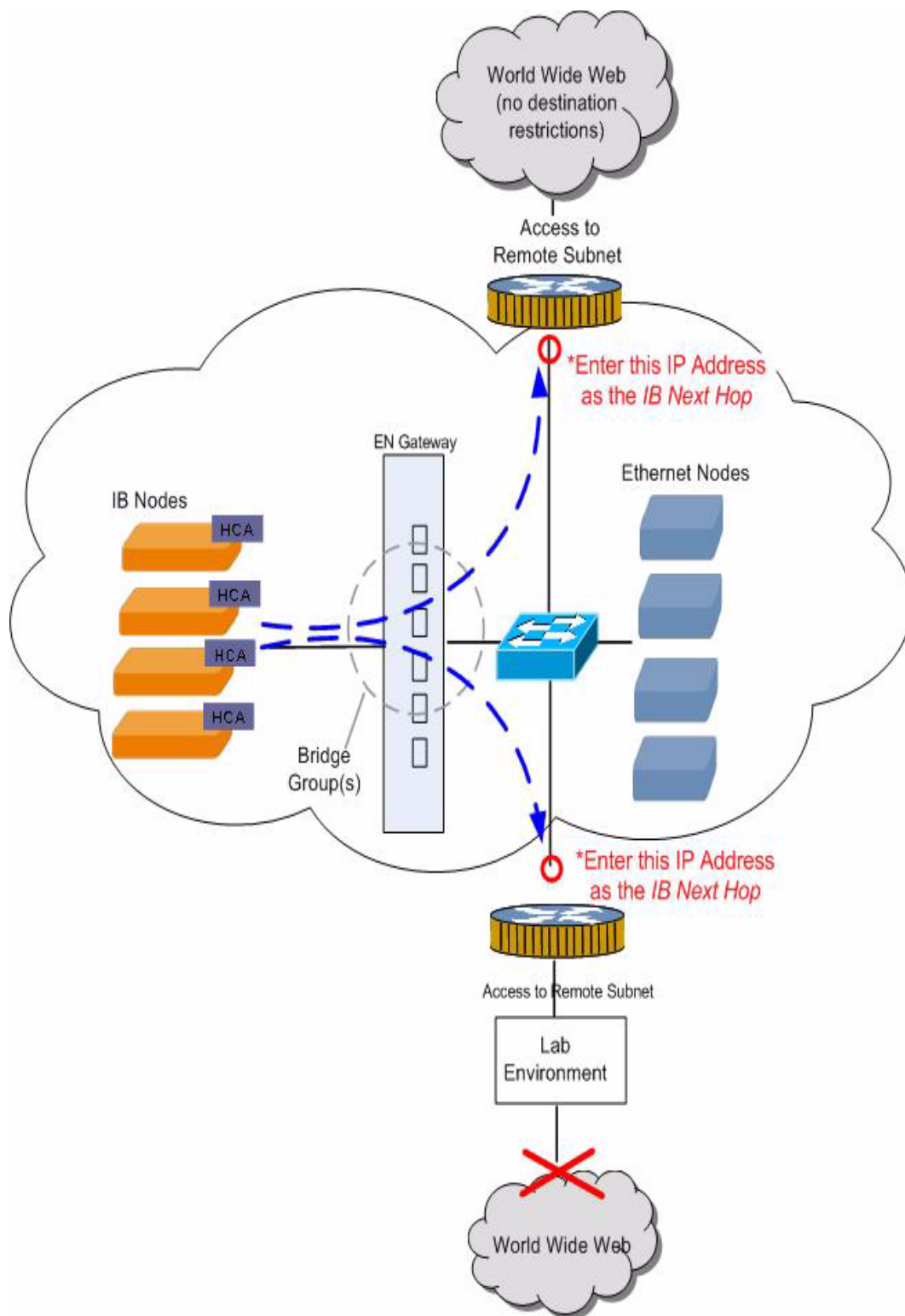
The Ethernet switch and the router may be different network elements.

Configuring Bridge-Groups with Multiple Remote Access Points

If you want your IB network to reach a remote subnet (any destination outside of your bridged subnet), then you must configure the `ib-next-hop`. It is important to understand that the `ib-next-hop` refers to the next hop that the *ib* nodes will take - not the next hop on the *ib* fabric. In most cases this is the default gateway.

The configuration in [Figure 6-5](#) illustrates a bridge-group with two `ib-next-hops`. In this example, the `ib-next-hops` are the default gateways. The destination for one of the next hops is left as the default (subnet 0/0 is assumed); however the other next hop has an explicitly configured destination.

Figure 6-5 Example of Bridge-Groups with Two IB Next Hops



- Step 1** Configure Partitions, if you are using this feature. See the [“Creating a P_Key Partition ID”](#) section on page 3-4.

Step 2 Create a link aggregation group, if you are using this feature. See the [“Creating a Link Aggregation Interface” section on page 4-3](#).

Step 3 Create a bridge-group, if you have not already done so. See the [“Creating a Single Bridge-Group” section on page 5-4](#).

- a. Locate the address of your ib-next-hop. This is the next hop that the IB hosts take outside the bridged network. In most cases this is the default gateway.

The Ethernet gateway requires that the next hop match the routing table in the ib nodes.

- b. Assign the first next hop for the IB network. The following example shows a bridge-group created without partitions or link aggregation.



Note An ib-next-hop has been configured, as this bridge-group is granted unrestricted access outside the broadcast domain.

This is an example of ib-next-hop 1

```
enable
configure terminal
bridge-group 1 subnet-prefix 20.0.0.0 16
bridge-group 1 ib-next-hop 20.0.0.1 next-hop-mask 255.255.0.0
```

Step 4 Assign the second next hop for the IB network. The example below shows a bridge-group created without partitions or link aggregation.



Note This ib-next-hop has been configured, but is restricted by a specific destination.

This is an example of ib-next-hop 2

```
bridge-group 1 ib-next-hop 20.0.0.2 dest 200.0.0.0 255.255.0.0
interface ethernet 3/1
bridge-group 1
interface gateway 3
bridge-group 1
exit
exit
```

Step 5 (Optional) Show the bridge-group configuration.

Step 6 Configure the IB hosts.

This is an example of IB host 1

```
ifconfig ib0 20.0.0.11 netmask 255.255.0.0 broadcast 20.0.255.255
route add -net 200.0.0.0 netmask 255.255.0.0 gw 20.0.0.2
route add default gw 20.0.0.1
```

This is an example of IB host 2

```
ifconfig ib1 20.0.0.12 netmask 255.255.0.0 broadcast 20.0.255.255
route add -net 200.0.0.0 netmask 255.255.0.0 gw 20.0.0.2
route add default gw 20.0.0.1
```

Step 7 Configure the Ethernet hosts.

This is an example of Ethernet host 1

```
ifconfig eth2 20.0.0.21 netmask 255.255.0.0 broadcast 20.0.255.255
route add -net 200.0.0.0 netmask 255.255.0.0 gw 20.0.0.2
route add default gw 20.0.0.1
```

Example of Ethernet host 2

```
ifconfig eth2 20.0.0.2 netmask 255.255.0.0 broadcast 20.0.255.255
ifconfig eth2:0 200.0.0.11 netmask 255.255.0.0 broadcast 200.0.255.255
```

Step 8 Configure remote Ethernet hosts.

Example of remote Ethernet host 1

```
ifconfig eth1 100.0.0.11 netmask 255.255.0.0 broadcast 100.0.255.255
route add default gw 100.0.0.1
```

Example of remote Ethernet host 2

```
ifconfig eth3 100.0.0.12 netmask 255.255.0.0 broadcast 100.0.255.255
route add default gw 100.0.0.1
```

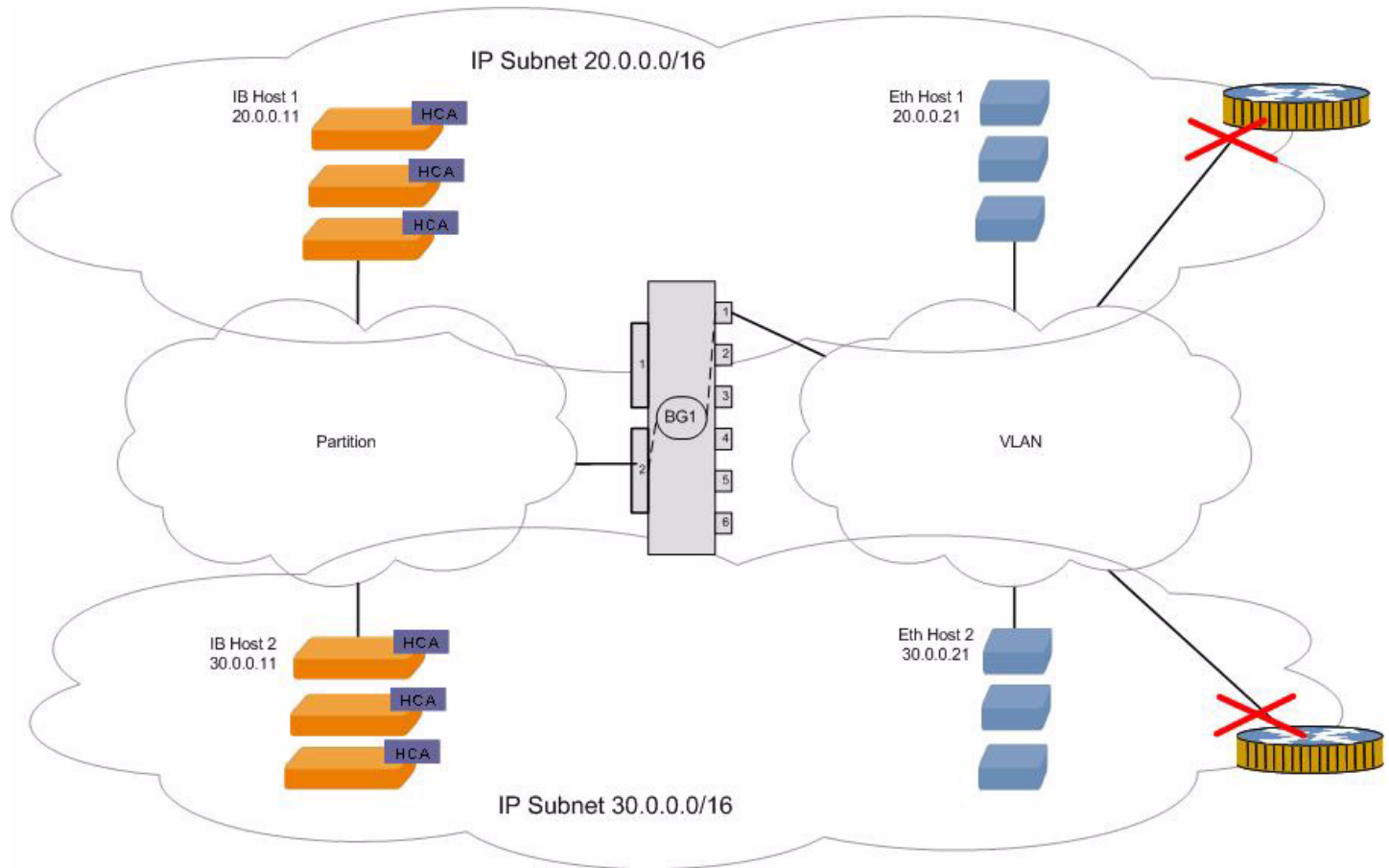
Step 9 Configure the Ethernet switch.

Configuring Subnet Forwarding for Multiple IP Subnets

Bridge-groups can be configured to bridge a single subnet or up to eight multiple IP subnets. Bridge-groups can also be configured to use one or more Layer 3 switches for remote access.

Bridging Multiple IP Subnets without Remote Access

The configuration in [Figure 6-6](#) shows one bridge-group configured for two IP subnets. This bridge-group is not configured to have remote access through a Layer 3 switch.

Figure 6-6 Example of Bridging Two IP Subnets

- Step 1** Configure partitions, if you are using this feature. See the [“Creating a P_Key Partition ID”](#) section on page 3-4.
- Step 2** Create a link aggregation group, if you are using this feature. See the [“Creating a Link Aggregation Interface”](#) section on page 4-3.
- Step 3** Create a bridge-group with the appropriate IP subnets. See the [“Creating a Single Bridge-Group”](#) section on page 5-6.

```
enable
configure terminal
bridge-group 1 subnet-prefix 20.0.0.0 16
bridge-group 1 subnet-prefix 30.0.0.0 16
interface ethernet 3/1
bridge-group 1
interface gateway 3
bridge-group 1
exit
```

- Step 4** Configure the IB hosts.

This is an example of IB host 1

```
ifconfig ib0 20.0.0.11 netmask 255.255.0.0 broadcast 20.0.255.255
```

This is an example of IB host 2

```
ifconfig ib1 30.0.0.11 netmask 255.255.0.0 broadcast 30.0.255.255
```

Step 5 Configure the Ethernet hosts.

This is an example of Ethernet host 1

```
ifconfig eth2 20.0.0.21 netmask 255.255.0.0 broadcast 20.0.255.255
```

This is an example of Ethernet host 2

```
ifconfig eth2 30.0.0.21 netmask 255.255.0.0 broadcast 30.0.255.255
```

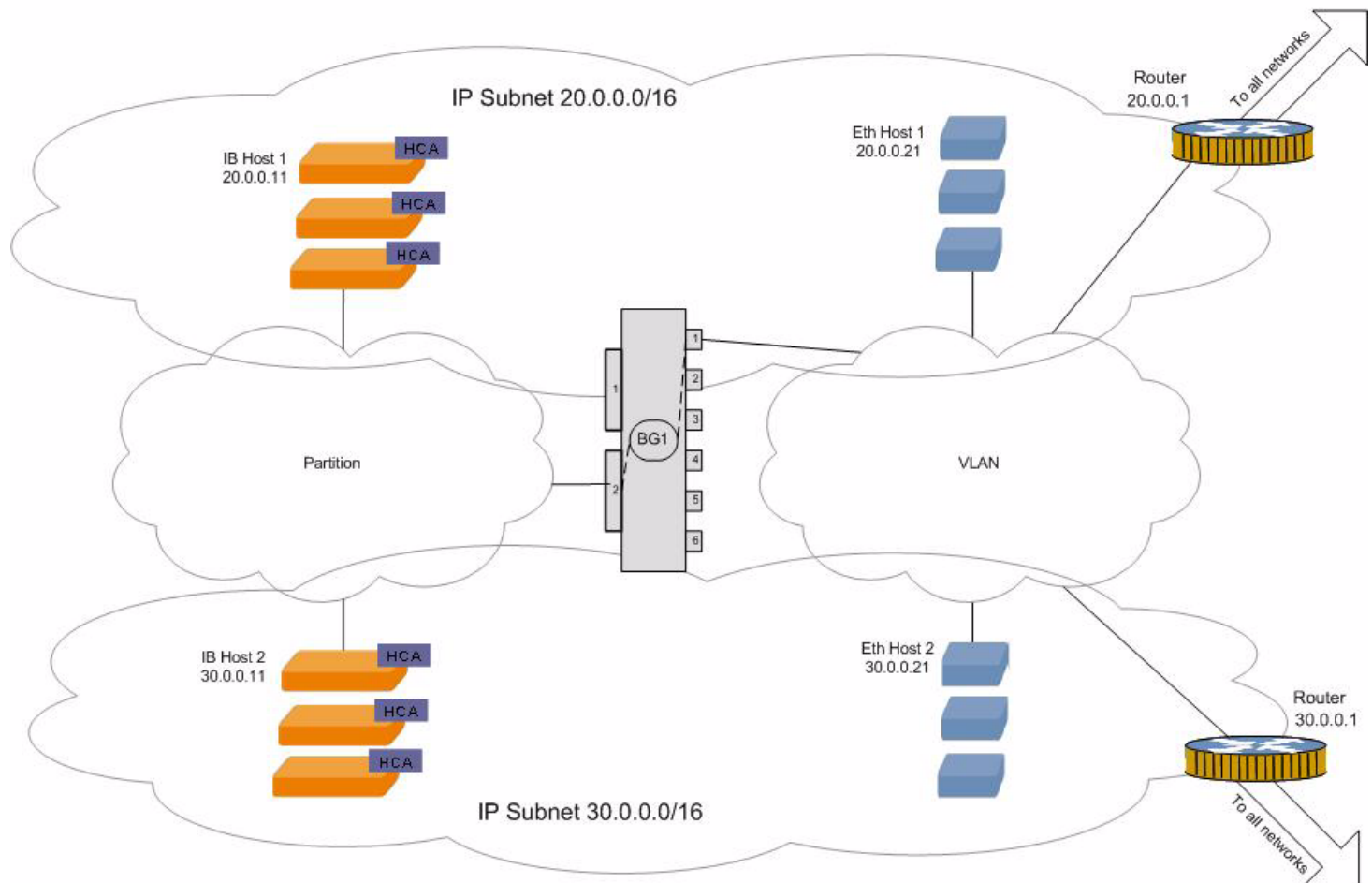
Step 6 Configure the Ethernet switch.

Bridging Multiple IP Subnets with Remote Access

The configuration in [Figure 6-7](#) shows one bridge-group configured for two IP subnets. Two Layer 3 switches are used as the default gateway for both subnets.

Figure 6-7 Example of Bridging Two IP Subnets

Documentation Setup 5



Step 1 Configure partitions, if you are using this feature. See the [“Creating a P_Key Partition ID”](#) section on page 3-4.

Step 2 Create a link aggregation group, if you are using this feature. See the [“Creating a Link Aggregation Interface”](#) section on page 4-3.

Step 3 Create a bridge-group for the first IP subnet. See the [“Creating a Single Bridge-Group”](#) section on page 5-6.

```
enable
configure terminal

bridge-group 1 subnet-prefix 20.0.0.0 16
```

Step 4 Add the second IP subnet for the bridge-group.

```
enable
configure terminal
bridge-group 1 subnet-prefix 20.0.0.0 16
```

Step 5 Assign the first next hop for the IB network.

```
bridge-group 1 ib-next-hop 20.0.0.1 next-hop-mask 255.255.0.0
```

Step 6 Assign the second next hop for the IB network.

```
bridge-group 1 ib-next-hop 30.0.0.1 next-hop-mask 255.255.0.0
```

Step 7 Complete the bridge-group configuration.

```
interface ethernet 3/1
bridge-group 1
interface gateway 3
bridge-group 1
exit
exit
```

Step 8 (Optional) Show the bridged subnets.

Example 1

```
SFS-3012R # show bridge-subnets
=====
Bridge Subnets
=====
bridge      subnet-prefix      subnet-prefix-len
-----
1           20.0.0.0           16
```

Example 2

```
SFS-3012R # show bridge-subnets
=====
Bridge Subnets
=====
bridge      subnet-prefix      subnet-prefix-len
-----
1           20.0.0.0           16
1           30.0.0.0           16
```

Step 9 (Optional) Show the bridge forwarding configuration.

```
SFS-3012R # show bridge-forwarding
=====
Bridge Forwarding
=====
bridge      port      subnet-pfx      pfx-len      next-hop-addr      dest-addr      dest-len
-----
1           ib        20.0.0.0        16           20.0.0.1           0.0.0.0        0
1           ib        30.0.0.0        16           30.0.0.1           0.0.0.0        0
```

Step 10 Configure the IB hosts.

Example of IB host 1

```
ifconfig ib0 20.0.0.11 netmask 255.255.0.0 broadcast 20.0.255.255
route add default gw 20.0.0.1
```

Example of IB host 2

```
ifconfig ib1 30.0.0.11 netmask 255.255.0.0 broadcast 30.0.255.255
route add default gw 30.0.0.1
```

Step 11 Configure the Ethernet hosts.

Example of Ethernet host 1

```
ifconfig eth2 20.0.0.21 netmask 255.255.0.0 broadcast 20.0.255.255
route add default gw 20.0.0.1
```

Example of Ethernet host 2

```
ifconfig eth2 30.0.0.1 netmask 255.255.0.0 broadcast 30.0.255.255
ifconfig eth2:0 200.0.0.11 netmask 255.255.0.0 broadcast 200.0.255.255
```

Step 12 Configure the remote Ethernet hosts.

Example of remote Ethernet Host 1

```
ifconfig eth1 100.0.0.11 netmask 255.255.0.0 broadcast 100.0.255.255
route add default gw 100.0.0.1
```

```
ifconfig eth3 100.0.0.12 netmask 255.255.0.0 broadcast 100.0.255.255
route add default gw 100.0.0.1
```

Step 13 Configure the Ethernet switch, if you have not already done so.

(Optional) Enabling Broadcast Forwarding

This section describes how to enable broadcast forwarding.

1. (Optional) Enable broadcast forwarding if you are using DHCP .
 - a. Enable broadcast forwarding on the bridge-group only if you are *not* using redundancy groups. If you are using redundancy groups, this configuration should be done at the redundancy group level.


```
SFS-3012R (config)# bridge-group 1 broadcast-forwarding
SFS-3012R (config)#
```
 - b. Enable broadcast forwarding on the redundancy groups if you are using the redundancy feature. See the [“Enabling Broadcast Forwarding” section on page 7-2](#).

Viewing the Configurations

This section describes how to view configurations.

Showing the Bridge-Groups

Display the bridge-groups by entering the **show bridge-group** command.

In the example below, no VLAN is used, and DHCP is not used (therefore the ib-next-hop address is used, and broadcast forwarding is not used).

```
SFS-3012R # show bridge-group
=====
                        Bridge Groups
=====
bridge-group-id : 2
bridge-group-name : PrimaryBrGrp
eth-bridge-port : trunk 2 (not tagged)
ib-bridge-port : 2/2 (gw) (pkey: 80:04)
```

```

broadcast-forwarding : false
broadcast-forwarding-mode : inherit-from-redundancy-group
loop-protection-method : one
multicast : false
multicast-mode : inherit-from-redundancy-group
redundancy-group : 23
status-in-redundancy-group : primary
bridge-group-id : 3
bridge-group-name : SecondaryBrGrp
eth-bridge-port : trunk 3 (not tagged)
ib-bridge-port : 3/2(gw) (pkey: 80:04)
broadcast-forwarding : false
broadcast-forwarding-mode : inherit-from-redundancy-group
loop-protection-method : one
multicast : false
multicast-mode : inherit-from-redundancy-group
redundancy-group : 23
status-in-redundancy-group : secondary

```

Showing the Trunk Groups

Display the trunk groups by entering the **show trunk** command.

```

SFS-3012R # show trunk
=====
Trunks Groups
=====
trunk-group-id : 1
trunk-group-name : eth-grp-1
distribution-type : src-dst-mac
port-members : 2/1-2/4
enable : true
mtu : 1500
mac-addr : 00:05:ad:01:37:10
    ifindex : 45057

```



CHAPTER 7

Enabling Loop Protection

The following sections appear in this chapter:

- [Loop Protection Overview, page 7-1](#)
- [Configuring Loop Protection, page 7-2](#)

Loop Protection Overview

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

Loop Protection Options

There are several options that provide protection against broadcast loops and should be used depending on your hardware and software configuration.

Disabling Broadcast Forwarding

This option disables forwarding of all IP broadcast packets except for the services for which the Ethernet gateway has special handling. Depending on the hardware and software configuration you have chosen, it may be necessary to disable broadcast forwarding to prevent the creation of loops. See the [“Disabling Broadcast Forwarding” section on page 7-2](#) for more specific information.

Self-Cancelling ARP Requests

By default, the self-cancelling ARP requests feature is always active. It prevents duplicate ARP (requests that have the same target protocol address) from creating loops. The duplicate ARP is seen by multiple gateways and discarded.

Delayed Proxy ARP Transaction

By default, the delayed proxy ARP transaction feature is always active. This feature is an extension of the self-cancelling ARP request and comes into play when a duplicate ARP request is delayed.

ARP Packet Painting

Though the loop protection mechanisms described above are very effective on most LANs, it is possible for a duplicated ARP request to be received on the other side of the bridge before the original has arrived. In this event, the duplicated packet is considered to be the original, and the original packet is dropped. To prevent this rare occurrence, the Ethernet gateway provides the option of inserting a signature at the end of every proxy ARP request. This signature allows the Ethernet gateway to filter the duplicate requests, and break the loop. This feature is enabled by default and can be disabled if your Ethernet switch does not support it. See the [“Disabling ARP Packet Signatures” section on page 7-3](#).

Configuring Loop Protection

For information regarding all the loop protection mechanisms that are available on the Ethernet gateway, see the [“Loop Protection” section on page 1-9](#).

Disabling Broadcast Forwarding

Whether or not Broadcast Forwarding should be disabled on a gateway to prevent loops depends on the physical configuration of your switches and gateways. For example, the following configuration would cause every broadcast packet to loop forever:

Two separate Ethernet gateways are connected to the same IB and Ethernet switches, and the gateways are assigned the same bridging ports with the same VLAN and P_key.

In the configuration described above, Broadcast Forwarding should be disabled. For more information about Loop Protection, see the [“Loop Protection” section on page 1-9](#).

To disable broadcast forwarding on an Ethernet gateway:

-
- | | |
|---------------|--|
| Step 1 | Enter privileged EXEC mode

SFS-3012R > enable
SFS-3012R # |
| Step 2 | Enter global configuration mode.

SFS-3012R # configure
SFS-3012R (config) # |
| Step 3 | Specify the no bridge-group number to which you want to disable broadcast forwarding, and enter the broadcast-forwarding command.

SFS-3012R (config) # no bridge-group 1 broadcast-forwarding |
-

Enabling Broadcast Forwarding

Broadcast forwarding is disabled on the Ethernet gateway by default. Whether or not broadcast forwarding should be enabled or disabled on a gateway to prevent loops depends on the configuration of your switches and gateways. For example, if you are using the DHCP communication protocol, broadcast forwarding must be enabled on the bridge-group or redundancy group.

For more information about Loop Protection, see the [“Loop Protection Overview” section on page 7-1](#).
To enable broadcast forwarding on an Ethernet gateway:

-
- Step 1** Enter privileged EXEC mode.
- ```
SFS-3012R > enable
SFS-3012R #
```
- Step 2** Enter global configuration mode.
- ```
SFS-3012R # configure
SFS-3012R (config)#
```
- Step 3** Specify the bridge-group on which you want to enable broadcast forwarding, and enter the **broadcast-forwarding** command.
- ```
SFS-3012R (config)# bridge-group 1 broadcast-forwarding
SFS-3012R (config)#
```
- 

## Disabling ARP Packet Signatures

ARP packet painting is enabled by default. This method of loop protection is used to prevent rare instances of loops that occur despite the other protection methods (see the [“ARP Packet Painting” section on page 7-2](#)).

However, if your switch does not support the insertion of these signatures on proxy ARPs, you can disable this feature.

- 
- Step 1** Enter privileged EXEC mode.
- ```
SFS-3012R > enable
SFS-3012R #
```
- Step 2** Enter global configuration mode.
- ```
SFS-3012R # configure
SFS-3012R (config)#
```
- Step 3** Specify the bridge-group on which you want to disable ARP Packet Painting, and use the **loop-protection** command.
- ```
SFS-3012R (config)# no bridge-group 1 loop-protection
```
-



CHAPTER 8

Configuring Multicast Forwarding

The following sections appear in this chapter:

- [Multicast Overview, page 8-1](#)
- [Understanding Multicast Forwarding and IGMP Snooping, page 8-1](#)
- [Creating a Multicast Group, page 8-4](#)

Multicast Overview

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

IPv4 Multicast

IPv4 Multicast is the only type of IP multicast that is supported by the Ethernet gateway.

IP multicasting enables a host to send packets to a specific subset of all hosts as a group transmission. Without the ability to multicast, a host is limited to sending to either a single host or to all hosts.

Point-to-multipoint—a single multicast stream that is replicated at branch points in a switched network to reach multiple viewers. This is useful for unidirectional, live broadcasts such as corporate communications or media events. It is also useful for content distribution from a central server to collocated servers or software distribution from a data center to multiple end station PCs or servers.

Multipoint-to-multipoint—a bidirectional hub-and-spoke architecture in which an MCU receives and redirects signals from each member of a multicast conference.

Understanding Multicast Forwarding and IGMP Snooping

Multicast Groups

Each multicast group is identified by a unique LID and GID.

Each switch is configured with routing information for the multicast traffic, which specifies all of the ports where the packet needs to travel.

Multicast Forwarding

Multicast forwarding allows the hardware engine to forward IP packets, instead of having multicast traffic sent to the line card CPU for slower path processing.

Multicast forwarding, the source is sending traffic to an arbitrary group of hosts that are represented by a multicast group address. The multicast router must determine which direction is the upstream direction (toward the source) and which one is the downstream direction (or directions). If there are multiple downstream paths, the router replicates the packet and forwards it down the appropriate downstream paths (best unicast route metric)—which is not necessarily all paths. Forwarding multicast traffic away from the source, rather than to the receiver, is called RPF.

Internet Group Management Protocol

Using Multicast Over a Layer 2 Switch—

Regarding the Ethernet gateway, Layer 2 (L2) does not see Ethernet, but instead refers to the use of IB and Ethernet.

The default behavior for a Layer 2 switch is to forward all multicast traffic to every port that belongs to the destination LAN on the switch. This behavior reduces the efficiency of the switch, whose purpose is to limit traffic to the ports that need to receive the data.

As a result of the default switch behavior, IGMP efficiently handles IP multicast in a Layer 2 switching environment.

IGMP is used to dynamically register individual hosts in a multicast group on a particular LAN. Hosts identify group memberships by sending IGMP messages to their local multicast router. Under IGMP, routers listen to IGMP messages and periodically send out queries to discover which groups are active or inactive on a particular subnet.

Used by IP routers and their immediately connected IP hosts to communicate multicast group membership states to neighboring multicast routers.

Version 3 of IGMP supports source filtering, which is the ability of a multicast receiver host to signal to a router the groups from which it wants to receive multicast traffic, and from which sources this traffic is expected. This membership information enables software to forward traffic from only those sources from which receivers requested the traffic.

IGMP Snooping

IGMP snooping manages multicast traffic in switches by allowing directed switching of IP multicast traffic.

IGMP Snooping is an IP multicast constraining mechanism that runs on a Layer 2 switch. IGMP Snooping requires that the switch examine, or snoop, some Layer 3 information (specifically IGMP join/leave messages) in the IGMP packets that are sent between the hosts and the router. When the switch hears the IGMP host report from a host for a particular multicast group, the switch adds the port number of the host to the associated multicast table entry. When the switch hears the IGMP leave group message from a host, the switch removes the table entry of the host.

Because IGMP control messages are sent as multicast packets, they are indistinguishable from multicast data at Layer 2. A switch running IGMP snooping must examine every multicast data packet to determine if it contains any pertinent IGMP control information.

Overview of the IGMP Snooping Process—

When a host wants to join an IP multicast group, it sends an IGMP join message. A join message specifies the MAC address of the host and the IP multicast group that it wants to join.

The router then builds an IGMP join message and multicasts the join message to the well-known address to which the switches listens.

Upon receipt of the join message, each switch searches its multicast table to determine if it contains the MAC address of the host that is asking to join the multicast group.

If a switch finds the MAC address of the host in its table associating the MAC address with a non-trunking port, the switch creates a multicast forwarding entry in the forwarding table.

The host associated with that port receives multicast traffic for that multicast group. In this way, the switch automatically learns the MAC addresses and port numbers of the IP multicast hosts.

Layer 2 Multicast Addresses

Historically, NICs on a LAN segment could receive only packets destined for their burned-in MAC address or the broadcast MAC address. In IP multicast, several hosts must receive a single data stream with a common destination MAC address such that multiple hosts receive the same packet and still be able to differentiate between several multicast groups.

One method to accomplish this is to map IP multicast Class D addresses directly to a MAC address. Today, using this method, NICs can receive packets destined to many different MAC addresses—their own unicast, broadcast, and a range of multicast addresses.

Multicast Forwarding with the Ethernet Gateway

The Ethernet gateway does not use IGMP snooping to discover IP multicast groups. The gateway queries or listens for IB traps from the Subnet Manager, and if there is an IPoIB multicast listener provision, the hardware to forward the corresponding IP multicast group.

This is an example of a multicast scenario:

- No multicast members are in an IP multicast group in the IP subnet (both Ethernet and IB).
 - A multicast sender is on a different IP subnet (there is an L3 switch on the Ethernet side).
1. An IPoIB host must receive this multicast and joins corresponding IP/IB multicast group.
 2. The Subnet Manager sends a notification:
 - The Ethernet gateway receives an IB trap from the Subnet Manager that a new IB multicast group is created.
 - The gateway inspects the MGID and sees that this is an IPoIB multicast group.
 - The Ethernet gateway provisions the Ethernet tables to forward the IPoIB multicast group (in less than a second). The L3 switch starts forwarding this multicast group, the Ethernet gateway is already provisioned and the multicast traffic reaches the IPoIB host.

IGMP message is sent by the IPoIB host at approximately the same time:
 3. The IPoIB host sends a notification at approximately the same time as the Subnet Manager is sending a notification:
 - The IPoIB host sends an IGMP notification to the L3 switch to inform that it is interested. The L3 switch starts forwarding the multicast group to this subnet.
 - *If* the first IGMP notification comes after the first Subnet Manager trap, the IGMP notification can be dropped.

- If the first IGMP notification is dropped, the host resends the IGMP announcement 2-3 times, and the next packets reaches L3 on the Ethernet. The L3 switch starts forwarding this multicast group, the Ethernet gateway is already provisioned, and the multicast traffic reaches the IPoIB host.

Multicast/Broadcast Configuration in Bridge-Groups

The bridge-group has a setting to enable/disable multicast/broadcast forwarding. This setting is used while the group is *not* a member of a redundancy group. If the bridging group is a member of a redundancy group the redundancy group multicast/broadcast enable/disable flag overwrites that of the bridge-group. When the bridge-group is removed from the redundancy group, the original setting is restored.

Creating a Multicast Group

This section describes how to create a multicast group.

Enabling Multicast Forwarding

Step 1 Start a CLI session, unless you have already done so.

Step 2 Enter privileged EXEC mode.

```
SFS-3001> enable
SFS-3001#
```

Step 3 Enter global configuration mode.

```
SFS-3001# configure
SFS-3001(config)#
```

Step 4 Specify the bridge-group and enable multicast forwarding:

- Enter the **bridge-group** *<bridge-group number>* command.
- Enter the **multicast** command.

```
SFS-3001(config)# bridge-group 2 multicast
SFS-3001(config)#
```

Step 5 Return to privileged EXEC mode.

```
SFS-3001(config)# exit
SFS-3001#
```

Step 6 Show the bridge-group by entering the **show bridge-group** command.

```
SFS-3001# show bridge-group
```

```
=====
Groups =====
      bridge-group-id : 2
      bridge-group-name :
      subnet-prefix : 0.0.0.0
      subnet-prefix-len : 0
      eth-bridge-port : 0/0
```

Bridge

```

        ib-bridge-port : 0/0
        eth-next-hop-addr : 0.0.0.0
        ib-next-hop-addr : 0.0.0.0
        broadcast-forwarding : false
        loop-protection-method : one
        multicast : true

```

Disabling Multicast Forwarding

- Step 1** Start a CLI session, unless you have already done so.
- Step 2** Enter privileged EXEC mode.
- ```
SFS-3001> enable
SFS-3001#
```
- Step 3** Enter global configuration mode.
- ```
SFS-3001# configure
SFS-3001(config)#
```
- Step 4** Specify the bridge-group and disable multicast forwarding:
- Enter the **no** keyword with the **bridge-group** *<bridge-group number>* command
 - Enter the **multicast** command.
- ```
SFS-3001(config)# no bridge-group 2 multicast
SFS-3001(config)#
```
- Step 5** Return to privileged EXEC mode.
- ```
SFS-3001(config)# exit
SFS-3001#
```
- Step 6** Show the bridge-group by entering the **show bridge-group** command.

```

SFS-3001# show bridge-group
=====
Groups =====
        bridge-group-id : 2
        bridge-group-name :
        subnet-prefix : 0.0.0.0
        subnet-prefix-len : 0
        eth-bridge-port : 0/0
        ib-bridge-port : 0/0
        eth-next-hop-addr : 0.0.0.0
        ib-next-hop-addr : 0.0.0.0
        broadcast-forwarding : false
        loop-protection-method : one
        multicast : false

```

Bridge



CHAPTER 9

Configuring Redundancy

This chapter contains the following sections:

- [Introduction, page 9-1](#)
- [Configuring Redundant Ethernet Gateways, page 9-2](#)

Introduction

Ethernet gateway redundancy is based on the concept of redundancy group. Redundancy group is a logical entity bridging an Ethernet VLAN to an IB partition, just like a bridge group, but in redundant fashion. A redundancy group can contain one or more bridge groups located on the same or different gateways. Gateways with bridge group members of the same redundancy group can be in the same or different chassis.

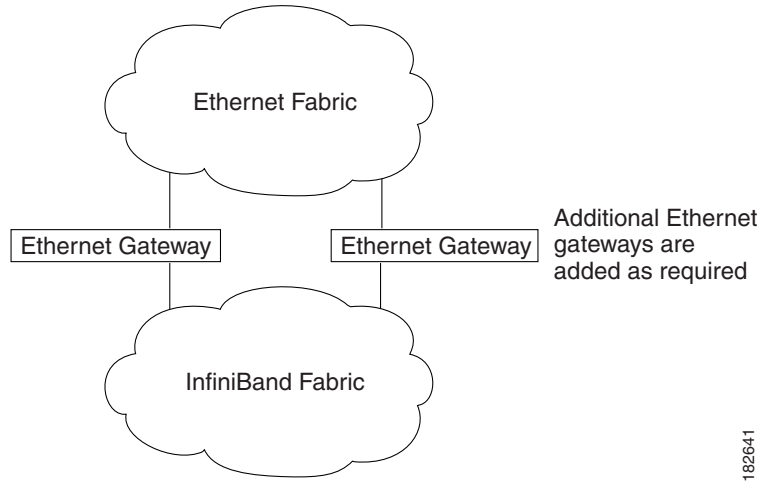
A redundancy group can operate in two modes. The first one, which is the default mode, is active-passive. In active-passive mode, only one bridge group is active and all others are in hot-standby state. In case the active bridge group fails, another member from the same redundancy group is selected and activated. In active-active mode all bridge groups are active and load balancing is enabled. Load balancing allows the user to distribute the load of IB nodes among all bridge groups of the redundancy group. In case of a bridge group failure, the load of the IB nodes is redistributed among the remaining bridge groups.



Note

Even in an active-active mode, only one bridge group forwards the broadcast and multicast traffic. In case the bridge group forwarding broadcast and multicast fails, a new bridge group member of the same redundancy group is selected to forward broadcast and multicast. The selection mechanism is identical to the one used in active-passive mode to select the active bridge group.

A logical diagram of gateways in a redundant configuration is shown in [Figure 9-1](#). Two or more Ethernet gateways are configured to bridges between the Ethernet fabric and the IB fabric. The gateways can be in the same or different chassis. The redundancy manager monitors the health of the gateways and in case of failure elects a new primary bridge group. The redundancy manager is part of the SFS OS and runs on the controller card of each chassis.

Figure 9-1 Logical Diagram of Ethernet Gateways in a Redundant Configuration

Different topologies work in similar ways but provide different levels of redundancy. For example, a redundancy group with two bridge groups in the same chassis but different gateways protects against gateway failure but does not provide chassis redundancy. On the other hand, a redundancy group with two bridge groups on gateways in different chassis provides the same level of gateway redundancy but also provides chassis redundancy. Examples are provided later in this chapter.

When a bridge group becomes a member of a redundancy group, some parameters are overwritten with the corresponding parameters from the redundancy group. These parameters are restored when the bridge group is removed from the redundancy group. Some examples of such parameters are broadcast and multicast forwarding.

The examples in the next sections show the most popular deployment of different types of I/O chassis. To simplify configuration, only one IP subnet is bridged (data IP subnet). The Ethernet ports are not VLAN tagged. The configuration on the Ethernet switch connected to the gateways determines which VLAN is bridged. This VLAN is mapped to the default IB partition using the Ethernet gateway. One IP subnet is allocated for in-band IB management (management IP subnet). The in-band IB management interface must be configured in order for the redundancy to work. This is true even in a single chassis configuration.

Each bridge group must have an IP address assigned from the data subnet in order for the redundancy with load-balancing (active-active mode) to work.

Data Subnet: 10.0.0.0/8

Default Gateway (on the Ethernet switch): 10.0.0.1

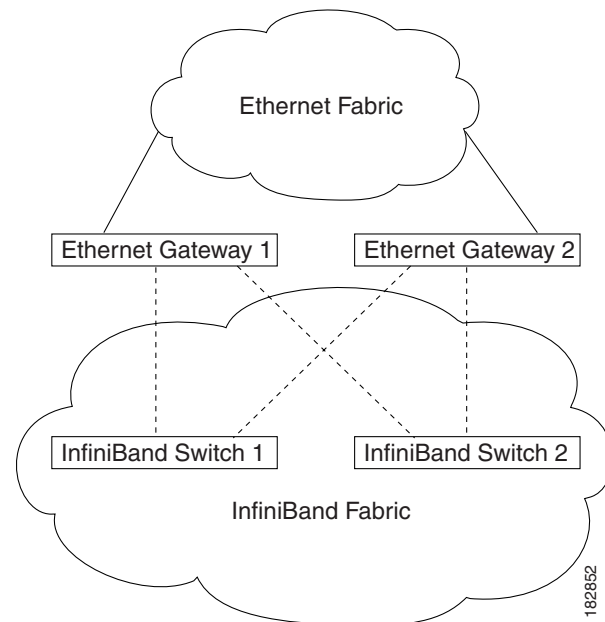
IB Management Subnet: 100.0.0.0/8

Configuring Redundant Ethernet Gateways

This example shows an SFS 3012R Server Switch setup with a total of two Ethernet gateways that are available. The gateways are configured to use different switch cards. The gateway in slot 2 uses the switch card in slot 15 and the gateway in slot 3 uses the switch card in slot 16. Thus if one of the switch cards fail, the other gateway continues to remain operational. If the Ethernet ports of the gateways are connected to two different Ethernet switches, the gateways in the same chassis must be connected to

different Ethernet switches. This topology allows one chassis and one Ethernet switch failure without interruption to the service. The topology used in this example is shown in Figure 9-2. The figure below shows two Ethernet gateways which could exist in two different chassis or a single chassis.

Figure 9-2 Ethernet Gateway Redundancy



Once the redundancy group is configured a primary bridge group is elected and forwarding is enabled. The steps to configure the Ethernet gateways are as follows:

Step 1 Enter configuration mode.

This example shows how to enter configuration mode:

```
SFS-3012R> enable
SFS-3012R# configure terminal
```

Step 2 Configure and connect the IB in-band management interface.

The IP address must be unique on each chassis.



Note If the Out-Of-Band Ethernet interface on the controller card is also configured, it must be on a different IP subnet.

This example shows how to configure and connect the IB in-band management interface:

```
SFS-3012R(config)# interface mgmt-ib
SFS-3012R(config-if-mgmt-ib)# ip address 100.0.0.1 255.0.0.0
SFS-3012R(config-if-mgmt-ib)# no shutdown
SFS-3012R(config-if-mgmt-ib)# exit
```

Step 3 Create and configure two link aggregation groups (trunks). Assign Ethernet ports to the link aggregation groups.

This example shows how to create and configure link aggregation groups:

```
SFS-3012R(config)# interface trunk 1
```

```

SFS-3012R(config-if-trunk)# enable
SFS-3012R(config)# distribution-type src-dst-ip
SFS-3012R(config)# interface ethernet 2/1-2/6
SFS-3012R(config-if-ether-2/1-2/6)# trunk-group 1

SFS-3012R(config)# interface trunk 2
SFS-3012R(config-if-trunk)# enable
SFS-3012R(config)# distribution-type src-dst-ip
SFS-3012R(config)# interface ethernet 3/1-3/6
SFS-3012R(config-if-ether-3/1-3/6)# trunk-group 2
SFS-3012R(config-if-ether-3/1-3/6)# exit

```

Step 4 Configure two bridge groups and assign ports to them.

One of the bridge groups uses the IB port 1 to connect to the switch card in slot 15 and the other bridge group uses the IB port 2 to connect to the switch card in slot 16.



Note IP addresses must be from Data IP subnet and must be unique for each bridge group.

This example shows how to configure two bridge groups and assign ports to them:

```

SFS-3012R(config)# bridge-group 1 subnet-prefix 10.0.0.0 8
SFS-3012R(config)# bridge-group 1 ip-addr 10.0.0.101
SFS-3012R(config)# bridge-group 1 ib-next-hop 10.0.0.1
SFS-3012R(config)# interface trunk 1
SFS-3012R(config-if-trunk)# bridge-group 1
SFS-3012R(config)# interface gateway 2/1
SFS-3012R(config-if-gw-2/1)# bridge-group 1
SFS-3012R(config-if-gw-2/1)# exit

SFS-3012R(config)# bridge-group 2 subnet-prefix 10.0.0.0 8
SFS-3012R(config)# bridge-group 2 ip-addr 10.0.0.102
SFS-3012R(config)# bridge-group 2 ib-next-hop 10.0.0.1
SFS-3012R(config)# interface trunk 2
SFS-3012R(config-if-trunk)# bridge-group 2
SFS-3012R(config)# interface gateway 3/2
SFS-3012R(config-if-gw-3/2)# bridge-group 2
SFS-3012R(config-if-gw-3/2)# exit

```

Step 5 Configure the redundancy group and assign both bridge groups to it.



Note The redundancy group ID must be the same in both chassis.

This example shows how to configure the redundancy group and assign bridge groups to it:

```

SFS-3012R(config)# redundancy-group 1
SFS-3012R(config)# bridge-group 1 redundancy-group 1
SFS-3012R(config)# bridge-group 2 redundancy-group 1

```

Step 6 (Optional) Enable load balancing between bridge groups.

Step 7 This example shows how to enable load balancing between bridge groups:

```

SFS-3012R(config)# redundancy-group 1 load-balancing

```

Verify Redundancy Group Configuration

To check redundancy group configuration and status use **show redundancy-group** CLI command. This command shows redundancy group properties and all members. It is important to make sure the properties match the configuration and all members are reported. Redundancy group must be checked on both chassis.

This example shows how to use the **show redundancy-group** command:

```
SFS-3012R# show redundancy-group
```

```
=====
                        Redundancy Groups
=====
rlb-id : 1
name :
group-p_key : ff:ff
load-balancing : enabled
broadcast-forwarding : false
multicast : false
gratuitous-igmp : false
igmp-version : v2
num-members : 2
new-member-force-reelection : false

=====
                        Redundancy Group Members
=====
bridge-group src-addr      last-receive
-----
1             100.0.0.1     Thu Jan  1 00:03:44 1970
2             100.0.0.1     Thu Jan  1 00:03:44 1970
```

Verify Bridge Group Configuration

To check bridge group configuration and status use **show bridge-group** CLI command. Note that when a bridge group is a member of a redundancy group most of the properties are inherited from the redundancy group. Also make sure only one bridge group, across all chassis, is primary and the rest are secondary. This is true even when load balancing is enabled (active-active mode). In active-passive mode only the primary bridge group is forwarding. All others are in hot standby state. In active-active mode all bridge groups are forwarding unicast traffic and only primary is forwarding broadcast and multicast if enabled.

Verify the bridge group configuration on the first SFS 3012R Server Switch. This example shows how to verify the bridge group configuration on the first SFS 3012R Server Switch:

```
SFS-3012R# show bridge-group
```

```
=====
                        Bridge Groups
=====
bridge-group-id : 1
bridge-group-name :
ip-addr : 10.0.0.101
eth-bridge-port : trunk 1 (not tagged)
ib-bridge-port : 2/1(gw) (pkey: ff:ff)
broadcast-forwarding : false
broadcast-forwarding-mode : inherit-from-redundancy-group
loop-protection-method : one
multicast : false
```

```
        multicast-mode : inherit-from-redundancy-group
        gratuitous-igmp : false
    gratuitous-igmp-mode : inherit-from-redundancy-group
        igmp-version : v2
        igmp-version-mode : inherit-from-redundancy-group
        redundancy-group : 1
    status-in-redundancy-group : primary

        bridge-group-id : 2
    bridge-group-name :
        ip-addr : 10.0.0.102
        eth-bridge-port : trunk 2 (not tagged)
        ib-bridge-port : 3/2(gw) (pkey: ff:ff)
    broadcast-forwarding : false
    broadcast-forwarding-mode : inherit-from-redundancy-group
    loop-protection-method : one
        multicast : false
        multicast-mode : inherit-from-redundancy-group
        gratuitous-igmp : false
    gratuitous-igmp-mode : inherit-from-redundancy-group
        igmp-version : v2
        igmp-version-mode : inherit-from-redundancy-group
        redundancy-group : 1
    status-in-redundancy-group : secondary
```



APPENDIX **A**

Acronyms and Abbreviations

[Table A-1](#) defines the acronyms and abbreviations that are used in this publication.

Table A-1 *List of Acronyms and Abbreviations*

Acronym	Expansion
ARP	Address Resolution Protocol
CIDR	Classless Inter-Domain Routing
CLI	command-line interface
DHCP	Dynamic Host Configuration Protocol
ESD	electrostatic discharge
GID	global identifier
GUID	globally unique identifier
IB	InfiniBand
ICMP	Internet Control Message Protocol
IGMP	Internet Group Management Protocol
IP	Internet Protocol
IPoIB	Internet Protocol over InfiniBand
ISL	Inter-Switch Link
LID	local identifier
MAC	Media Access Control
MCU	multipoint controller unit
MGID	Multicast Global ID
MSB	most significant bit
MTU	Maximum Transmission Unit
NIC	Network Interface Card
OOB	Out-of-Band
QP	queue pair
RPF	Reverse Path Forwarding

Table A-1 *List of Acronyms and Abbreviations (continued)*

Acronym	Expansion
SFS	Server Fabric Switching
UTP	unshielded twisted pair



INDEX

Numerics

802.1Q [1-6](#)

802.3ad [4-1](#)

A

about

- address resolution protocol [1-8](#)

- arp requests [7-1](#)

- delayed proxy ARP transaction [7-1](#)

- gateway ports [1-1](#)

- internal ports [1-2](#)

- layer 2 bridging [1-5](#)

- loop protection [1-9](#)

- MAC addressing scheme [1-3](#)

- trunking [1-5](#)

address resolution protocol [1-8](#)

ARP (address resolution protocol)

- definition [1-8](#)

- requests [7-1](#)

arp requests, about [7-1](#)

audience [ix](#)

B

bridge-group

- selecting ID [6-6](#)

bridging

- layer 2 [1-5](#)

broadcast forwarding

- enable [6-19](#)

- loop protection [1-9](#)

C

connecting to other devices [2-4](#)

connectors [2-4](#)

conventions, document [x](#)

D

delayed proxy ARP transaction

- about [7-1](#)

DF header [1-9](#)

DHCP [1-3](#)

- broadcast forwarding [6-19](#)

disable

- broadcast forwarding [7-2](#)

distribution

- round robin [4-2](#)

distribution-type [4-2](#)

document

- audience [ix](#)

- conventions [x](#)

- organization [ix](#)

- related [xi](#)

don't fragment

- IP header [1-9](#)

dst-ip [4-2](#)

dst-mac [4-2](#)

E

ethernet gateway

- LEDs [2-6, 2-7](#)

ethernet interface

LEDs [2-6, 2-7](#)

Expansion Module

installing [2-2](#)

removing [2-3](#)

G

gateway

upgrade [2-7](#)

gateway LEDs [2-6, 2-7](#)

gateway ports [1-1](#)

understanding [1-1](#)

I

image

upgrade for gateway [2-7](#)

installing

expansion module [2-2](#)

internal ports

understanding [1-3](#)

IP fragmentation [1-9](#)

IP header

don't fragment [1-9](#)

IPv4 [1-7](#)

L

layer 2 bridging [1-5](#)

how it works [1-5](#)

LEDs

ethernet gateway [2-6, 2-7](#)

ethernet meanings [2-6](#)

link aggregation

load distribution [4-2](#)

standard [4-1](#)

load distribution

types and descriptions [4-2](#)

loop protection [1-9](#)

M

MAC addressing scheme [1-3](#)

management

about [1-10](#)

MIBs [1-10](#)

N

network devices

connecting [2-4](#)

O

organization, document [ix](#)

P

partitions

default [1-9](#)

ports

gateway [1-1](#)

internal [1-2](#)

R

related documentation [xi](#)

round robin [4-2](#)

S

self-canceling ARP requests [7-1](#)

show trunk [6-20](#)

src-dst-ip [4-2](#)

src-dst-mac [4-2](#)

src-ip [4-2](#)

src-mac [4-2](#)

T

tagged

VLANs [1-5](#)

troubleshooting

disabling broadcast forwarding [7-2](#)

trunking [1-5](#)

U

untagged

VLANs [1-5](#)

upgrade

software for gateway [2-7](#)

V

VLANs

link aggregation [4-2](#)

