Study Planner

Video Training

Flash Cards

Practice tests

Review Exercises

Official Cert Guide

Advance your IT career with hands-on learning

CCNP and CCIE Data Center Core

DCCOR 350-601

SOMIT MALOO, CCIE NO. 28603, CCDE NO. 20170002
FIRAS AHMED, CCIE NO.14967

ciscopress.com

CCNP and CCIE Data Center Core

DCCOR 350-601

Official Cert Guide

SOMIT MALOO, CCIE No. 28603, CCDE No. 20170002

FIRAS AHMED, CCIE No. 14967

Cisco Press

CCNP and CCIE Data Center Core DCCOR 350-601 Official Cert Guide

Somit Maloo & Firas Ahmed

Copyright © 2020 Cisco Systems, Inc.

Published by:
Cisco Press

All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without written permission from the publisher, except for the inclusion of brief quotations in a review.

03 22

Library of Congress Control Number: 2020930073

ISBN-13: 978-0-13-644962-1

ISBN-10: 0-13-644962-X

Warning and Disclaimer

This book discusses the content and skills needed to pass the 350-601 CCNP Data Center Core certification exam, which is the prerequisite for CCNP as well as CCIE certification. Every effort has been made to make this book as complete and as accurate as possible, but no warranty or fitness is implied.

The information is provided on an “as is” basis. The authors, Cisco Press, and Cisco Systems, Inc. shall have neither liability nor responsibility to any person or entity with respect to any loss or damages arising from the information contained in this book or from the use of the discs or programs that may accompany it.

The opinions expressed in this book belong to the authors and are not necessarily those of Cisco Systems, Inc.

Trademark Acknowledgments

All terms mentioned in this book that are known to be trademarks or service marks have been appropriately capitalized. Cisco Press or Cisco Systems, Inc., cannot attest to the accuracy of this information. Use of a term in this book should not be regarded as affecting the validity of any trademark or service mark.

Special Sales

For information about buying this title in bulk quantities, or for special sales opportunities (which may include electronic versions; custom cover designs; and content particular to your business, training goals, marketing focus, or branding interests), please contact our corporate sales department at corpsales@pearsoned.com or (800) 382-3419.

For government sales inquiries, please contact governmentsales@pearsoned.com.

For questions about sales outside the U.S., please contact intlcs@pearson.com.

Feedback Information

At Cisco Press, our goal is to create in-depth technical books of the highest quality and value. Each book is crafted with care and precision, undergoing rigorous development that involves the unique expertise of members from the professional technical community.

Readers' feedback is a natural continuation of this process. If you have any comments regarding how we could improve the quality of this book, or otherwise alter it to better suit your needs, you can contact us through email at feedback@ciscopress.com. Please make sure to include the book title and ISBN in your message.

We greatly appreciate your assistance.

Editor-in-Chief: Mark Taub

Alliances Manager, Cisco Press: Arezou Gol

Director, ITP Product Management: Brett Bartow

Senior Editor: James Manly

Managing Editor: Sandra Schroeder

Development Editor: Ellie Bru

Senior Project Editor: Tonya Simpson

Copy Editor: Chuck Hutchinson

Technical Editor: Ozden Karakok

Editorial Assistant: Cindy Teeters

Cover Designer: Chuti Prasertsith

Composition: codeMantra

Indexer: Ken Johnson

Proofreader: Charlotte Kughen

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **www.cisco.com/go/offices.**

Cisco and the Cisco logo are trademarks or registered trademarks of Cisco and/or its affiliates in the U.S. and other countries. To view a list of Cisco trademarks, go to this URL: www.cisco.com/go/trademarks. Third party trademarks mentioned 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. (1110R)

About the Authors

Somit Maloo, CCIE No. 28603, CCDE No. 20170002, is a content lead from the data center team in the Learning@Cisco organization. He holds a master's degree in telecommunication networks and a bachelor's degree in electronics and telecommunication engineering. He is also a penta CCIE in routing and switching, service provider, wireless, security, and data center technologies. Somit holds various industry-leading certifications, including CCDE, PMP, RHCSA, and VMware VCIX6 in Data Center and Network Virtualization. Somit has extensive experience in designing and developing various data center courses for the official Cisco curriculum. He started his career as a Cisco TAC engineer. Somit has more than 10 years of experience in the networking industry, working mostly with data center networks. You can reach Somit on Twitter: @somitmaloo.

Firas Ahmed, CCIE No. 14967, is a solution architect from the enterprise data center team at Cisco Customer Experience (CX). He completed a master's degree in systems and control engineering following a bachelor's degree in computer engineering. Firas holds CCIE certificates in routing and switching, collaboration, wireless, security, and data center technologies in addition to industry-based certifications, including CISSP, PMP, VMware VCP6.5-DCV, ITIL, and GICSP. Firas has more than 15 years of experience in designing, developing, and supporting various data centers for enterprise and IoT customers. Firas has additional experience as a seasonal instructor in a number of community colleges in Toronto, where he taught various computer networking courses. You can reach Firas on Twitter: @dccor_firas.

About the Technical Reviewer

Ozden Karakok, CCIE No. 6331, is a technical consultant on data center technologies and solutions at Flint Consulting. She worked at Cisco for 19 years as a technical leader supporting data center solutions. Prior to joining Cisco, Ozden spent five years working for a number of Cisco's large customers in various telecommunication roles. She is a Cisco Certified Internetwork Expert in routing and switching, SNA/IP, and storage. She co-authored three Cisco Press books: *CCNA Data Center DCICN 200-150*, *CCNA Data Center DCICT 200-155*, and *Data Center Technologies DCICT 640-916*. Ozden holds a degree in computer engineering from Istanbul Bogazici University. You can reach Ozden on Twitter: @okarakok.

Dedications

Somit:

To my loving wife, Renuka, for her unending love and support.

To my wonderful parents, who supported me in every phase of my life.

To Navya and Namit, who agreed not to fight while Papa was working on the book.

To my aunt, Tara, for being the guiding angel in my life.

Firas:

To my amazing wife, Nora, who has been extremely supportive throughout this process. Thanks for letting me spend long hours on my computer once again!

To Ibrahim and Maryam, you are growing so fast. Never give up on what you want. If at first you don't succeed, try and try again. I love you more than anything!

To my parents, you are still the guiding light that keeps me on the right path.

Acknowledgments

Somit Maloo:

I would like to thank my co-author, Firas Ahmed, for working as a team to complete this book. Without your support, this book would not have been possible. I am thankful to all our professional editors, especially Mary Beth Ray, James Manly, and Ellie Bru, for their patience and guidance at every step of the book process. I would also like to thank our technical editor, Ozden Karakok, for her keen attention to detail and for agreeing to review the book, taking time out of her busy schedule.

Firas Ahmed:

I would like to thank my co-author, Somit Maloo, for taking the initiative to form this partnership and for his dedication in putting together the outline of this book. Thank you for your valuable input and continuous support throughout the process.

A special recognition to the technical editor, Ozden Karakok, for her valuable contributions in reviewing the book. Her suggestions and feedback helped improve the quality of the content in multiple areas.

Thanks to the Cisco Press team, especially Mary Beth Ray and James Manly, for believing in us, and Ellie Bru, for her guidance and extreme patience while editing and amending the chapters of the book.

A special credit to Hazim Dahir, distinguished engineer at Cisco Systems, for his help and support with the technical review of the book.

In addition, I want to thank my colleague Naveen Chapa for reviewing and providing constructive feedback that helped enhance the ACI chapter.

Contents at a Glance

Online Elements

Reader Services

Other Features

In addition to the features in each of the core chapters, this book has additional study resources on the companion website, including the following:

Practice exams: The companion website contains an exam engine that enables you to review practice exam questions. Use these to prepare with a sample exam and to pinpoint topics where you need more study.

An online interactive Flash Cards application to help you drill on Key Terms by chapter.

Glossary quizzes: The companion website contains interactive quizzes that enable you to test yourself on every glossary term in the book.

More than 2 hours of video training: The companion website contains multiple hours of unique test-prep videos.

To access this additional content, simply register your product. To start the registration process, go to www.ciscopress.com/register and log in or create an account*. Enter the product ISBN 9780136449621 and click Submit. After the process is complete, you will find any available bonus content under Registered Products.

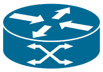
*Be sure to check the box that you would like to hear from us to receive exclusive discounts on future editions of this product.

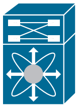
Contents

Online Elements

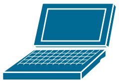
Icons Used in This Book

Cisco Nexus 9500 Series

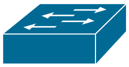

ATM Router

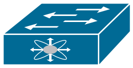

Cisco Nexus 7000

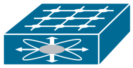
File Server


Laptop

Server


Switch


Cisco Nexus 5000

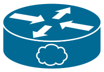

Cisco Nexus 2000

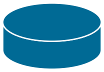
Terminal

Cloud

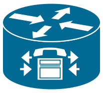
Cisco Nexus 9300 Series

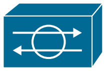
API Controller

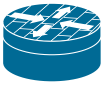

Generic/ Unknown


Database

Storage Array


Telephony Router


Net Ranger


Router with Firewall

IP Phone

Command Syntax Conventions

The conventions used to present command syntax in this book are the same conventions used in the IOS Command Reference. The Command Reference describes these conventions as follows:

- **Boldface** indicates commands and keywords that are entered literally as shown. In actual configuration examples and output (not general command syntax), boldface indicates commands that are manually input by the user (such as a **show** command).
- *Italic* indicates arguments for which you supply actual values.
- Vertical bars (|) separate alternative, mutually exclusive elements.
- Square brackets ([ ]) indicate an optional element.
- Braces ({ }) indicate a required choice.
- Braces within brackets ([{ }]) indicate a required choice within an optional element.

Introduction

Professional certifications have been an important part of the computing industry for many years and will continue to become more important. Many reasons exist for these certifications, but the most popularly cited reason is that of credibility. All other considerations held equal, the certified employee/consultant/job candidate is considered more valuable than one who is not.

Goals and Methods

The most important and somewhat obvious goal of this book is to help you pass the 350-601 CCNP Data Center Core Exam. In fact, if the primary objective of this book were different, the book's title would be misleading; however, the methods used in this book to help you pass the 350-601 CCNP Data Center Core Exam are designed to also make you much more knowledgeable about how to do your job. Although this book and the companion website together have more than enough questions to help you prepare for the actual exam, the method in which they are used is not simply to make you memorize as many questions and answers as you possibly can.

One key methodology used in this book is to help you discover the exam topics that you need to review in more depth, to help you fully understand and remember those details, and to help you prove to yourself that you have retained your knowledge of those topics. So, this book does not try to help you pass by memorization, but helps you truly learn and understand the topics. The Data Center Core Exam is just one of the foundation topics in the CCNP and CCIE certification, and the knowledge contained within is vitally important to consider yourself a truly skilled data center engineer or specialist. This book would do you a disservice if it didn't attempt to help you learn the material. To that end, the book will help you pass the Data Center Core Exam by using the following methods:

- Helping you discover which test topics you have not mastered
- Providing explanations and information to fill in your knowledge gaps
- Supplying exercises and scenarios that enhance your ability to recall and deduce the answers to test questions
- Providing practice exercises on the topics and the testing process via test questions through the companion website

Who Should Read This Book?

This book is not designed to be a general networking topics book, although it can be used for that purpose. This book is intended to tremendously increase your chances of passing the CCNP Data Center Core Exam. Although other objectives can be achieved from using this book, the book is written with one goal in mind: to help you pass the exam.

So why should you want to pass the CCNP Data Center Core Exam? Because it's one of the milestones toward getting the CCNP and CCIE certification—no small feat in itself. What would getting the CCNP or CCIE mean to you? A raise, a promotion, recognition?

How about to enhance your resume? To demonstrate that you are serious about continuing the learning process and that you're not content to rest on your laurels. To please your reseller-employer, who needs more certified employees for a higher discount from Cisco. Or one of many other reasons.

Strategies for Exam Preparation

The strategy you use for the CCNP Data Center Core Exam might be slightly different from strategies used by other readers, mainly based on the skills, knowledge, and experience you already have obtained. For instance, if you have attended the DCICN and DCICT course, you might take a different approach than someone who learned data center technologies via on-the-job training.

Regardless of the strategy you use or the background you have, the book is designed to help you get to the point where you can pass the exam with the least amount of time required. For instance, there is no need for you to practice or read about OSPF or BGP if you fully understand it already. However, many people like to make sure that they truly know a topic and thus read over material that they already know. Several book features will help you gain the confidence that you need to be convinced that you know some material already and to also help you know what topics you need to study more.

The Companion Website for Online Content Review

All the electronic review elements, as well as other electronic components of the book, exist on this book's companion website.

How to Access the Companion Website

To access the companion website, which gives you access to the electronic content with this book, start by establishing a login at www.ciscopress.com and register your book. To do so, simply go to www.ciscopress.com/register and enter the ISBN of the print book: 9780136449621. After you have registered your book, go to your account page and click the **Registered Products** tab. From there, click the **Access Bonus Content** link to get access to the book's companion website.

Note that if you buy the Premium Edition eBook and Practice Test version of this book from Cisco Press, your book will automatically be registered on your account page.

Simply go to your account page, click the Registered Products tab, and select Access Bonus Content to access the book's companion website.

How to Access the Pearson Test Prep Practice Test Software

You have two options for installing and using the Pearson Test Prep practice test software: a web app and a desktop app. To use the Pearson Test Prep practice test software, start by finding the registration code that comes with the book. You can find the code in these ways:

- **Print book:** Look in the cardboard sleeve in the back of the book for a piece of paper with your book's unique PTP code.

- **Premium Edition:** If you purchase the Premium Edition eBook and Practice Test directly from the Cisco Press website, the code will be populated on your account page after purchase. Just log in at www.ciscopress.com, click **Account** to see details of your account, and click the **Digital Purchases** tab.
- **Amazon Kindle:** For those who purchase a Kindle edition from Amazon, the access code will be supplied directly from Amazon.
- **Other Bookseller e-books:** Note that if you purchase an e-book version from any other source, the practice test is not included because other vendors to date have not chosen to vend the required unique access code.

NOTE Do not lose the activation code because it is the only means with which you can access the QA content with the book.

When you have the access code, to find instructions about both the PTP web app and the desktop app, follow these steps:

Step 1. Open this book's companion website, as shown earlier in this Introduction under the heading "How to Access the Companion Website."

Step 2. Click the Practice Exams button.

Step 3. Follow the instructions listed there both for installing the desktop app and for using the web app.

Note that if you want to use the web app only at this point, just navigate to www.pearsontestprep.com, establish a free login if you do not already have one, and register this book's practice tests using the registration code you just found. The process should take only a couple of minutes.

NOTE Amazon e-book (Kindle) customers: It is easy to miss Amazon's email that lists your PTP access code. Soon after you purchase the Kindle eBook, Amazon should send an email. However, the email uses very generic text and makes no specific mention of PTP or practice exams. To find your code, read every email from Amazon after you purchase the book. Also do the usual checks for ensuring your email arrives, like checking your spam folder.

NOTE Other e-book customers: As of the time of publication, only the publisher and Amazon supply PTP access codes when you purchase their e-book editions of this book.

How This Book Is Organized

Although this book could be read cover-to-cover, it is designed to be flexible and allow you to easily move between chapters and sections of chapters to cover just the material that you need more work with.

The core chapters, Chapters 1 through 20, cover the following topics:

- **Chapter 1, "Implementing Routing in the Data Center":** This chapter discusses data center Layer 3 routing protocols, focusing on OSPF and BGP routing protocols. It also discusses multicast and First Hop Redundancy Protocols such as HSRP and VRRP.
- **Chapter 2, "Implementing Data Center Switching Protocols":** This chapter discusses data center Layer 2 switching protocols, focusing on spanning tree and multiport aggregation. It also discusses virtual port channels (multichassis port channels).
- **Chapter 3, "Implementing Data Center Overlay Protocols":** This chapter discusses various data center Overlay protocols, including Overlay Transport Virtualization (OTV) and Virtual Extensible LAN (VXLAN).
- **Chapter 4, "Describe Cisco Application Centric Infrastructure":** This chapter discusses various aspects of Cisco ACI, including but not limited to fabric discovery, fabric access policies, fabric packet flow, tenants, and VMM domains.
- **Chapter 5, "Cisco Cloud Services and Deployment Models":** This chapter discusses an overview of what cloud computing is along with cloud service models per the NIST 800-145 definition, such as Infrastructure as a Service (IaaS), Software as a Service (SaaS), and Platform as a Service (PaaS). It also discusses various cloud deployment models per the NIST 800-145 definition, such as public, private, community, and hybrid cloud.
- **Chapter 6, "Data Center Network Management and Monitoring":** This chapter discusses data center network disruptive/nondisruptive upgrade procedures, network configurations, and infrastructure monitoring aspects in detail. It also discusses data center network assurance and data telemetry.
- **Chapter 7, "Implement Fibre Channel":** This chapter discusses the Fibre Channel protocol in detail. It discusses Fibre Channel topologies, port types, switched fabric initialization, CFS distribution, VSAN, zoning, device alias, FLOGI, and FCNS databases. It also discusses NPV and NPIV features in detail.
- **Chapter 8, "Implement FCoE Unified Fabric":** This chapter discusses the FCoE Unified Fabric Protocol in detail. It discusses various Ethernet enhancements that enable FCoE support on Ethernet interfaces. It also discusses FCoE topology options and various FCoE implementations—for example, FCoE over FEX and FCoE NPV.
- **Chapter 9, "Describe NFS and NAS Concepts":** This chapter discusses NFS basics along with various NFS versions. It also discusses NAS basics with an overview of the Cisco NSS 3000 Series NAS product.
- **Chapter 10, "Describe Software Management and Infrastructure Monitoring":** This chapter discusses how the Cisco MDS NX-OS Setup Utility helps to build an initial configuration file using the System Configuration dialog. It also discusses Cisco MDS NX-OS software upgrade and downgrade procedures, along with infrastructure monitoring features such as SPAN, RSPAN, RMON, and Call Home.

- **Chapter 11, "Cisco Unified Computing Systems Overview":** This chapter discusses the Cisco Unified Computing System (UCS) architecture. It also discusses in detail UCS initial setup, along with network management aspects of Cisco UCS such as identity pools, policies, QoS, and templates.

- **Chapter 12, "Cisco Unified Computing Infrastructure Monitoring":** This chapter discusses Cisco Unified Compute traffic monitoring and Intersight cloud management.

- **Chapter 13, "Cisco Unified Compute Software and Configuration Management":** This chapter discusses Cisco UCS configuration management such as backup and restore. It also discusses aspects of firmware and software updates on Cisco UCS.

- **Chapter 14, "Cisco HyperFlex Overview":** This chapter discusses the Cisco Hyperflex solution and benefits. It also discusses edge solutions that enable any application to be deployed, monitored, and managed anywhere.

- **Chapter 15, "Automation and Scripting Tools":** This chapter discusses various automation and scripting tools. It discusses the Embedded Event Manager (EEM), Scheduler, Bash Shell, and Guest Shell for Cisco NX-OS software, and various data formats such as XML and JSON. It also discusses how the REST API can be used to configure Cisco NX-OS devices.

- **Chapter 16, "Evaluate Automation and Orchestration Technologies":** This chapter discusses various automation and orchestration technologies. It discusses how Ansible, Puppet, and Python can be used to automate Cisco Data Center products. It also discusses the PowerOn Auto Provisioning (POAP) process, Cisco Data Center Network Manager (DCNM) tool, Cisco UCS Director (UCSD) tool, along with how the PowerShell Agent executes various tasks on UCSD.

- **Chapter 17, "Network Security":** This chapter discusses network authentication, authorization, and accounting (AAA) and user role-based access control (RBAC). It also discusses various network security protocols in detail, including control plan policing, dynamic ARP inspection, DHCP snooping, and port security.

- **Chapter 18, "Compute Security":** This chapter discusses Cisco UCS authentication and user role-based access control. It also discusses the keychain authentication method.

- **Chapter 19, "Storage Security":** This chapter discusses various storage security features in detail. It discusses authentication, authorization, and accounting (AAA), user accounts, and RBAC. It also discusses configuration and verification of port security and fabric binding features on the Cisco MDS 9000 Series switches.

- **Chapter 20, "Final Preparation":** This chapter suggests a plan for final preparation after you have finished the core parts of the book, in particular explaining the many study options available in the book.

Certification Exam Topics and This Book

The questions for each certification exam are a closely guarded secret. However, we do know which topics you must know to *successfully* complete this exam. Cisco publishes them as an exam blueprint for the Implementing Cisco Data Center Core Technologies (DCCOR 350-601) Exam. Table I-1 lists each exam topic listed in the blueprint along with a reference to the book chapter that covers the topic. These are the same topics you should be proficient in when working with Cisco data center technologies in the real world.

Table I-1 DCCOR Exam 350-601 Topics and Chapter References

DCCOR 350-601 Exam Topic	Chapter(s) in Which Topic Is Covered
1.0 Network	
1.1 Apply routing protocols	1
1.1.a OSPFv2, OSPFv3	1
1.1.b MP-BGP	1
1.1.c PIM	1
1.1.d FHRP	1
1.2 Apply switching protocols such as RSTP+, LACP and vPC	2
1.3 Apply overlay protocols such as VXLAN EVPN and OTV	3
1.4 Apply ACI concepts	4
1.4.a Fabric setup	4
1.4.b Access policies	4
1.4.c VMM	4
1.4.d Tenant policies	4
1.5 Analyze packet flow (unicast, multicast, and broadcast)	4
1.6 Analyze Cloud service and deployment models (NIST 800-145)	5
1.7 Describe software updates and their impacts	6
1.7.a Disruptive/nondisruptive	6
1.7.b EPLD	6
1.7.c Patches	6
1.8 Implement network configuration management	6
1.9 Implement infrastructure monitoring such as NetFlow and SPAN	6
1.10 Explain network assurance concepts such as streaming telemetry	6
2.0 Compute	
2.1 Implement Cisco Unified Compute System Rack Servers	11
2.2 Implement Cisco Unified Compute System Blade Chassis	11
2.2.a Initial setup	11
2.2.b Infrastructure management	11
2.2.c Network management (VLANs, pools and policies, templates, QoS)	11

DCCOR 350-601 Exam Topic	Chapter(s) in Which Topic Is Covered
2.2.d Storage management (SAN connectivity, Fibre Channel zoning, VSANs, WWN pools, SAN policies, templates)	11
2.2.e Server management (Server pools and boot policies)	11
2.3 Explain HyperFlex Infrastructure Concepts and benefits (Edge and Hybrid Architecture versus all-flash)	14
2.4 Describe firmware and software updates and their impacts on B-Series and C-Series servers	13
2.5 Implement compute configuration management (Backup and restore)	13
2.6 Implement infrastructure monitoring such as SPAN and Intersight	12
3.0 Storage Network	
3.1 Implement Fibre Channel	7
3.1.a Switch fabric initialization	7
3.1.b Port channels	7
3.1.c FCID	7
3.1.d CFS	7
3.1.e Zoning	7
3.1.f FCNS	7
3.1.g Device alias	7
3.1.h NPV and NPIV	7
3.1.i VSAN	7
3.2 Implement FCoE Unified Fabric (FIP and DCB)	8
3.3 Describe NFS and NAS concepts	9
3.4 Describe software updates and their impacts (Disruptive/ nondisruptive and EPLD)	10
3.5 Implement infrastructure monitoring	10
4.0 Automation	
4.1 Implement automation and scripting tools	15
4.1.a EEM	15
4.1.b Scheduler	15
4.1.c Bash Shell and Guest Shell for NX-OS	15
4.1.d REST API	15
4.1.e JSON and XML encodings	15
4.2 Evaluate automation and orchestration technologies	16
4.2.a Ansible	16
4.2.b Puppet	16
4.2.c Python	16
4.2.d POAP	16

DCCOR 350-601 Exam Topic	Chapter(s) in Which Topic Is Covered
4.2.e DCNM	16
4.2.f UCSD	16
4.2.g PowerShell	16
5.0 Security	
5.1 Apply network security	17
5.1.a AAA and RBAC	17
5.1.b ACI contracts and microsegmentation	17
5.1.c First-hop security features such as dynamic ARP inspection (DAI), DHCP snooping, and port security	17
5.1.d CoPP	17
5.2 Apply compute security	18
5.2.a AAA and RBAC	18
5.2.b Keychain authentication	18
5.3 Apply storage security	19
5.3.a AAA and RBAC	19
5.3.b Port security	19
5.3.c Fabric binding	19

Each version of the exam can have topics that emphasize different functions or features, and some topics can be rather broad and generalized. The goal of this book is to provide the most comprehensive coverage to ensure that you are well prepared for the exam. Although some chapters might not address specific exam topics, they provide a foundation that is necessary for a clear understanding of important topics. Your short-term goal might be to pass this exam, but your long-term goal should be to become a qualified data center professional.

It is also important to understand that this book is a “static” reference, whereas the exam topics are dynamic. Cisco can and does change the topics covered on certification exams often.

This exam guide should not be your only reference when preparing for the certification exam. You can find a wealth of information available at Cisco.com that covers each topic in great detail. If you think that you need more detailed information on a specific topic, read the Cisco documentation that focuses on that topic.

Note that as data center technologies continue to develop, Cisco reserves the right to change the exam topics without notice. Although you can refer to the list of exam topics in Table I-1, always check Cisco.com to verify the actual list of topics to ensure that you are prepared before taking the exam. You can view the current exam topics on any current Cisco certification exam by visiting the Cisco.com website, choosing **Menu**, and **Training & Events**, then selecting from the Certifications list. Note also that, if needed, Cisco Press might post additional preparatory content on the web page associated with this book at http://www.ciscopress.com/title/9780136449621. It's a good idea to check the website a couple of weeks before taking your exam to be sure that you have up-to-date content.

Taking the CCNP Data Center Core Exam

As with any Cisco certification exam, you should strive to be thoroughly prepared before taking the exam. There is no way to determine exactly what questions are on the exam, so the best way to prepare is to have a good working knowledge of all subjects covered on the exam. Schedule yourself for the exam and be sure to be rested and ready to focus when taking the exam.

The best place to find out the latest available Cisco training and certifications is under the Training & Events section at Cisco.com.

Tracking Your Status

You can track your certification progress by checking http://www.cisco.com/go/certifications/login. You must create an account the first time you log in to the site.

How to Prepare for an Exam

The best way to prepare for any certification exam is to use a combination of the preparation resources, labs, and practice tests. This guide has integrated some practice questions and sample scenarios to help you better prepare. If possible, get some hands-on experience with ACI, Nexus, and UCS equipment. There is no substitute for real-world experience; it is much easier to understand the designs, configurations, and concepts when you can actually work with a live data center network.

Cisco.com provides a wealth of information about Application Centric Infrastructure (ACI), Nexus switches, and Unified Computing System—Blade and Rack servers, and data center LAN technologies and features.

Assessing Exam Readiness

Exam candidates never really know whether they are adequately prepared for the exam until they have completed about 30 percent of the questions. At that point, if you are not prepared, it is too late. The best way to determine your readiness is to work through the "Do I Know This Already?" quizzes at the beginning of each chapter and review the foundation and key topics presented in each chapter. It is best to work your way through the entire book unless you can complete each subject without having to do any research or look up any answers.

Cisco Data Center Certifications in the Real World

Cisco is one of the most recognized names on the Internet. Cisco Certified data center specialists can bring quite a bit of knowledge to the table because of their deep understanding of data center technologies, standards, and networking devices. This is why the Cisco certification carries such high respect in the marketplace. Cisco certifications demonstrate to potential employers and contract holders a certain professionalism, expertise, and dedication required to complete a difficult goal. If Cisco certifications were easy to obtain, everyone would have them.

Exam Registration

The 350-601 CCNP Data Center Core Exam is a computer-based exam, with around 100 to 110 multiple-choice, fill-in-the-blank, list-in-order, and simulation-based questions. You can take the exam at any Pearson VUE (http://www.pearsonvue.com) testing center. According to Cisco, the exam should last about 120 minutes. Be aware that when you register for the exam, you might be told to allow a certain amount of time to take the exam that is longer than the testing time indicated by the testing software when you begin. This discrepancy is because the testing center will want you to allow for some time to get settled and take the tutorial about the test engine.

Book Content Updates

Because Cisco occasionally updates exam topics without notice, Cisco Press might post additional preparatory content on the web page associated with this book at http://www.ciscopress.com/title/9780136449621. It is a good idea to check the website a couple of weeks before taking your exam to review any updated content that might be posted online. We also recommend that you periodically check back to this page on the Cisco Press website to view any errata or supporting book files that may be available.

CHAPTER 1

Implementing Routing in the Data Center

Data centers are an essential element of the Internet and cloud infrastructure. It is the data networks that deliver data services around the world. This task would be impossible without routing. Even in the new generation of facilities like edge data centers, routers play an important role in connecting network services to end users.

This chapter covers the following key topics:

OSPF: This section discusses the NX-OS OSPFv2 and OSPFv3 routing protocols and includes OSPF area types, OSPF routing device functions, and NX-OS configuration commands plus an example.

Border Gateway Protocol (BGP): This section covers the NX-OS BGP external routing protocols, including Multiprotocol BGP (MBGP or MP-BGP), along with configuration commands and an example.

Bidirectional Forwarding Detection (BFD): This section covers NX-OS routing with BFD failure detection and configuration commands and an example.

Multicast: This section discusses the NX-OS Layer 2 and Layer 3 multicast protocols, which include IGMP, MLD, MDT, PIM, and multicast forwarding, along with configuration commands and examples.

Hot Standby Router Protocol (HSRP): This section discusses NX-OS HSRP as a First Hop Redundancy Protocol (FHRP) on the Ethernet network, including HSRP object tracking and load sharing along with configuration commands and an example.

Virtual Router Redundancy Protocol (VRRP): This section discusses the NX-OS VRRP operation, groups, and object tracking. In addition this section covers IPv6 first hop redundancy and configuration commands and includes an example.

“Do I Know This Already?” Quiz

The “Do I Know This Already?” quiz allows you to assess whether you should read this entire chapter thoroughly or jump to the “Exam Preparation Tasks” section. If you are in doubt about your answers to these questions or your own assessment of your knowledge of the topics, read the entire chapter. Table 1-1 lists the major headings in this chapter and their corresponding “Do I Know This Already?” quiz questions. You can find the answers in Appendix A, “Answers to the ‘Do I Know This Already?’ Quizzes.”

Table 1-1 “Do I Know This Already?” Section-to-Question Mapping

Foundation Topics Section	Questions
OSPF	1–5
Border Gateway Protocol (BGP)	6–7
Bidirectional Forwarding Detection (BFD)	9
Multicast	10–11
Hot Standby Router Protocol (HSRP)	12–14
Virtual Router Redundancy Protocol (VRRP)	15

CAUTION The goal of self-assessment is to gauge your mastery of the topics in this chapter. If you do not know the answer to a question or are only partially sure of the answer, you should mark that question as wrong for purposes of the self-assessment. Giving yourself credit for an answer you correctly guess skews your self-assessment results and might provide you with a false sense of security.

1. When Open Shortest Path First (OSPF) starts neighbor negotiations, what setting must match between OSPF neighbors for them to be able to establish adjacencies?

- **a.** Router ID
- **b.** Hello intervals
- **c.** Link cost
- **d.** IP address

2. Which two parameters will assist in designated router (DR) elections?

- **a.** IP address
- **b.** Router ID
- **c.** Priority
- **d.** Code version

3. Designated router elections occur on which type of network? (Choose two answers.)

- **a.** Point-to-Point
- **b.** Broadcast (Ethernet)
- **c.** NBMA mode
- **d.** Point-to-Multipoint

4. What are two enhancements that OSPFv3 supports over OSPFv2?

- **a.** It requires the use of ARP.
- **b.** It can support multiple IPv6 subnets on a single link.
- **c.** It supports up to two instances of OSPFv3 over a common link.
- **d.** It routes over links rather than over networks.

5. Which statements about IPv6 and routing protocols are true?
   a. Link-local addresses are used to form routing adjacencies.
   b. OSPFv3 is the only routing protocol that supports IPv6.
   c. Loopback addresses are used to form routing adjacencies.
   d. MBGP does not support the IPv6 protocol.
6. When selecting the best path, the BGP protocol takes into account the following information in the stated order:
   a. AS_Path, origin type, multi-exit discriminator, local preference
   b. AS_Path, origin type, local preference, multi-exit discriminator
   c. Local preference, AS_Path, origin type, multi-exit discriminator
   d. Local preference, AS_Path, multi-exit discriminator, origin type
7. Which command displays the iBGP and eBGP neighbors that are configured?
   a. **show ip bgp**
   b. **show ip bgp paths**
   c. **show ip bgp peers**
   d. **show ip bgp summary**
8. What kind of BGP session is established between two routers that are adjacent but in two different autonomous systems?
   a. eBGP
   b. iBGP
   c. dBGP
   d. mBGP
9. What is a BFD detect multiplier?
   a. The interval at which this device wants to send BFD hello messages
   b. The minimum interval at which this device can accept BFD hello messages from another BFD device
   c. The number of missing BFD hello messages from another BFD device before this local device detects a fault in the forwarding path
   d. The time between BFD hello packets
10. In Ethernet LANs, what is the functional equivalent to IGMPv3 in IPv6?
    a. IGMPv3 includes IPv6 multicast support; that's why it is v3.
    b. MLDv2.
    c. MLDv1.
    d. IPv6's native support for multicast routing deprecates this need.
11. How is multicast RPF checking important when running PIM sparse mode? (Choose two answers.)
    a. To prevent multicast source spoofing
    b. To prevent receiver spoofing
    c. To prevent multicast forwarding loops by validating that the receiving interface is the reverse path to the S address
    d. To prevent multicast forwarding loops by validating that the receiving interface is the reverse path of the G address

12. Which statements about HSRP operation are true? (Choose three answers.)
    - a. The HSRP default timers are a 3-second hello interval and a 10-second dead interval.
    - b. HSRP supports only cleartext authentication.
    - c. The HSRP virtual IP address must be on a different subnet than the router's interface IP address.
    - d. The HSRP virtual IP address must be on the same subnet as the router's interface address.
    - e. HSRP V1 supports up to 256 groups.
13. Which HSRP feature was new in HSRPv2?
    - a. Group numbers that are greater than 255
    - b. Virtual MAC addresses
    - c. Tracking
    - d. Preemption
14. When a router with the highest HSRP priority recovers from failure, which option will ensure that the router immediately becomes the active router?
    - a. **standby preempt**
    - b. **standby priority**
    - c. **standby tracker**
    - d. **standby delay**
15. Which statement describes Virtual Router Redundancy Protocol (VRRP) object tracking?
    - a. It monitors traffic flow and link utilization.
    - b. It ensures the best VRRP router is the virtual router master for the group.
    - c. It causes traffic to dynamically move to higher bandwidth links.
    - d. It thwarts man-in-the-middle attacks.

Foundation Topics

OSPF

Open Shortest Path First (OSPF) is an IETF link-state routing protocol. OSPF has two versions: IPv4 OSPFv2 (version 2), which is described in RFC 2328, and IPv6 OSPFv3 (version 3), which is described in RFC 2740.

OSPF utilizes *hello packets* (multicast IPv4 224.0.0.5 or IPv6 FF02::5) for neighbor discovery. These hello packets are sent out on each OSPF-enabled interface to discover other OSPF neighbor routers. OSPF sends hello packets every 10 seconds (the OSPF default *hello_interval* is set to 10 seconds). In addition, OSPF uses hello packets for keepalive and bidirectional traffic. For keepalive, OSPF determines whether a neighbor is still communicating. If a router does not receive a hello packet within 40 seconds (the OSPF *dead-interval* is usually a multiple of the hello interval; the default is four times the hello interval), the neighbor is removed from the local neighbor table.

When a neighbor is discovered, the two routers compare information in the hello packet to determine whether the routers have compatible configurations. The neighbor routers attempt to establish *adjacency*, which means that the routers synchronize their link-state databases to ensure that they have identical OSPF routing information. Adjacent routers share link-state advertisements (LSAs) that include information about the operational state of each link, the cost of the link, and any other neighbor information. The routers then flood these received LSAs out every OSPF-enabled interface so that all OSPF routers eventually have identical link-state databases. When all OSPF routers have identical link-state databases, the network is *converged*. Each router then uses Dijkstra's Shortest Path First (SPF) algorithm to build its route table.

The key differences between the OSPFv3 and OSPFv2 protocols are as follows:

- OSPFv3 expands on OSPFv2 to provide support for IPv6 routing prefixes and the larger size IPv6 addresses, OSPF Hello address FF02::5.
- LSAs in OSPFv3 are expressed as prefix and prefix length instead of address and mask.
- The router ID and area ID are 32-bit numbers with no relationship to IPv6 addresses.
- OSPFv3 uses link-local IPv6 addresses for neighbor discovery and other features.
- OSPFv3 can use the IPv6 authentication trailer (RFC 6506) or IPSec (RFC 4552) for authentication. However, neither of these options is supported on Cisco NX-OS.
- OSPFv3 redefines LSA types.

Key Topic

OSPF Link-State Advertisements

OSPF uses link-state advertisements (LSAs) to build its routing table. When an OSPF router receives an LSA, it forwards that LSA out every OSPF-enabled interface, flooding the OSPF area with this information. This LSA flooding guarantees that all routers in the network have identical routing information. LSA flooding depends on the OSPF area configuration. The LSAs are flooded based on the link-state refresh time (every 30 minutes by default). Each LSA has its own link-state refresh time.

OSPFv3 redefines LSA types. OSPFv2 LSAs have seven different types (LSA type 1 to 7) and extensions (LSA 9 to 11) called Opaque, as shown in Table 1-2.

Opaque LSAs consist of a standard LSA header followed by application-specific information. This information might be used by OSPFv2 or by other applications. OSPFv2 uses Opaque LSAs to support the OSPFv2 graceful restart capability. The three Opaque LSA types are defined as follows:

- **LSA type 9:** Flooded to the local network.
- **LSA type 10:** Flooded to the local area.
- **LSA type 11:** Flooded to the local autonomous system.

In OSPFv3, LSA changed by creating a separation between prefixes and the SPF tree. There is no prefix information in LSA types 1 and 2. You find only topology adjacencies in these LSAs; you don't find any IPv6 prefixes in them. Prefixes are now advertised in type 9 LSAs, and the link-local addresses that are used for next hops are advertised in type 8 LSAs. Type 8 LSAs are flooded only on the local link, whereas type 9 LSAs are flooded within the area. The designers of OSPFv3 could have included link-local addresses in type 9 LSAs, but because these are only required on the local link, it would be a waste of resources.

Key Topic

Table 1-2 OSPFv2 and OSPFv3 LSAs Supported by Cisco NX-OS

Type	OSPFv2 Name	Description	OSPFv3 Name	Description
1	Router LSA	LSA sent by every router. This LSA includes the state and the cost of all links and a list of all OSPFv2 neighbors on the link. Router LSAs trigger an SPF recalculation. Router LSAs are flooded to the local OSPFv2 area.	Router LSA	LSA sent by every router. This LSA includes the state and cost of all links but does not include prefix information. Router LSAs trigger an SPF recalculation. Router LSAs are flooded to the local OSPFv3 area.
2	Network LSA	LSA sent by the DR. This LSA lists all routers in the multi-access network. Network LSAs trigger an SPF recalculation.	Network LSA	LSA sent by the DR. This LSA lists all routers in the multi-access network but does not include prefix information. Network LSAs trigger an SPF recalculation.
3	Network Summary LSA	LSA sent by the area border router to an external area for each destination in the local area. This LSA includes the link cost from the area border router to the local destination.	Inter-Area Prefix LSA	Same as OSPFv2; just the name changed.
4	ASBR Summary LSA	LSA sent by the area border router to an external area. This LSA advertises the link cost to the ASBR only.	Inter-Area Router LSA	Same as OSPFv2; just the name changed.
5	AS External LSA	LSA generated by the ASBR. This LSA includes the link cost to an external autonomous system destination. AS External LSAs are flooded throughout the autonomous system.	AS External LSA	Same as OSPFv2.

Type	OSPFv2 Name	Description	OSPFv3 Name	Description
7	NSSA External LSA	LSA generated by the ASBR within a not-so-stubby area (NSSA). This LSA includes the link cost to an external autonomous system destination. NSSA External LSAs are flooded only within the local NSSA.	NSSA External LSA	Same as OSPFv2.
8	N/A		Link LSA (New OSPFv3 LSA)	LSA sent by every router, using a link-local flooding scope. This LSA includes the link-local address and IPv6 prefixes for this link.
9	Opaque LSAs	LSA used to extend OSPF.	Intra-Area Prefix LSA	LSA sent by every router. This LSA includes any prefix or link state changes. Intra-Area Prefix LSAs are flooded to the local OSPFv3 area. This LSA does not trigger an SPF recalculation.
10	Opaque LSAs	LSA used to extend OSPF.	N/A	
11	Opaque LSAs	LSA used to extend OSPF.	Grace LSAs	LSA sent by a restarting router, using a link-local flooding scope. This LSA is used for a graceful restart of OSPFv3.

To control the flooding rate of LSA updates in your network, you can use the *LSA group pacing* feature. LSA group pacing can reduce high CPU or buffer usage. This feature groups LSAs with similar link-state refresh times to allow OSPF to pack multiple LSAs into an OSPF update message.

Each router maintains a link-state database for the OSPF network. This database contains all the collected LSAs and includes information on all the routes through the network. OSPF uses this information to calculate the best path to each destination and populates the routing table with these best paths.

LSAs are removed from the link-state database if no LSA update has been received within a set interval, called the MaxAge. Routers flood a repeat of the LSA every 30 minutes to prevent accurate link-state information from being aged out. The Cisco NX-OS operating system supports the LSA grouping feature to prevent all LSAs from refreshing at the same time.

OSPF Areas

Key Topic

An *area* is a logical division of routers and links within an OSPF domain that creates separate subdomains. *LSA flooding is contained within an area*, and the link-state database is limited to links within the area, which reduces the CPU and memory requirements for an OSPF-enabled router. You can assign an area ID to the interfaces within the defined area.

The area ID is a 32-bit value that you can enter as a number or in dotted-decimal notation, such as 3.3.3.3 or 0.0.0.3. Cisco NX-OS always displays the area in dotted-decimal notation. If you define more than one area in an OSPF network, you must also define the *backbone area*, which has the reserved area ID of 0.0.0.0. If you have more than one area, one or more routers become *area border routers* (ABRs). An ABR connects to both the backbone area and at least one other defined area (see Figure 1-1).

Key Topic

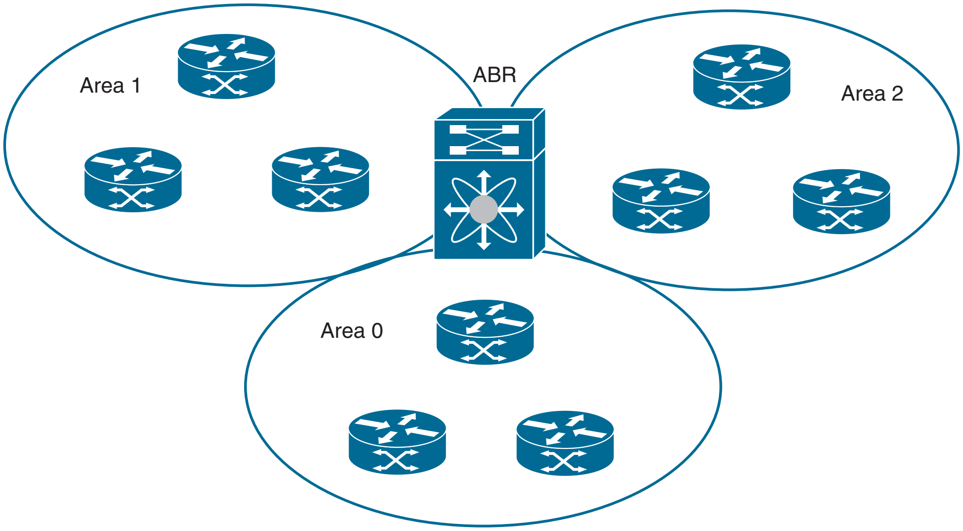


Figure 1-1 *OSPF Areas*

The ABR has a separate link-state database for each area to which it connects. The ABR sends *Network Summary (type 3) LSAs* from one connected area to the backbone area. The backbone area sends summarized information about one area to another area.

OSPF defines another router type as an *autonomous system boundary router* (ASBR). This router connects an OSPF area to another autonomous system. An autonomous system is a network controlled by a single technical administration entity. OSPF can redistribute its routing information into another autonomous system or receive redistributed routes from another autonomous system.

You can limit the amount of external routing information that floods an area by making it a *stub area*. A stub area is an area that does not allow AS External (type 5) LSAs. These LSAs are usually flooded throughout the local autonomous system to propagate external route information. Stub areas have the following requirements:

- All routers in the stub area are stub routers.
- No ASBR routers exist in the stub area.
- You cannot configure virtual links in the stub area.

Figure 1-2 shows an example of an OSPF autonomous system where all routers in area 0.0.0.5 have to go through the ABR to reach external autonomous systems. Area 0.0.0.5 can be configured as a stub area.

Key Topic

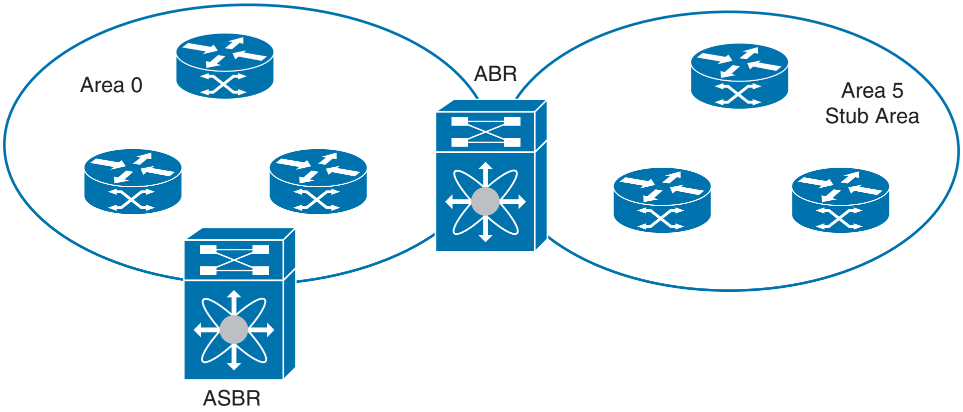


Figure 1-2 *OSPF Stub Area*

Stub areas use a default route for all traffic that needs to go through the backbone area to the external autonomous system.

There is an option to allow OSPF to import autonomous system external routes within a stub area; this is a *not-so-stubby area* (NSSA). An NSSA is similar to a stub area, except that an NSSA allows you to import autonomous system (AS) external routes within an NSSA using redistribution. The NSSA ASBR redistributes these routes and generates NSSA External (type 7) LSAs that it floods throughout the NSSA. You can optionally configure the ABR that connects the NSSA to other areas to translate this NSSA External LSA to AS External (type 5) LSAs. The ABR then floods these AS External LSAs throughout the OSPF autonomous system. Summarization and filtering are supported during the translation.

You can, for example, use NSSA to simplify administration if you are connecting a central site using OSPF to a remote site that is using a different routing protocol. Before NSSA, the connection between the corporate site border router and a remote router could not be run as an OSPF stub area because routes for the remote site could not be redistributed into a stub area. With NSSA, you can extend OSPF to cover the remote connection by defining the area between the corporate router and remote router as an NSSA.

NOTE The backbone area 0 cannot be an NSSA.

All OSPF areas must physically connect to area 0 (backbone area). If one area cannot connect directly to area 0, you need a virtual link. Virtual links allow you to connect an OSPF area ABR to a backbone area ABR when a direct physical connection is not available. Figure 1-3 shows a virtual link that connects area 5 to the backbone area 0 through area 3.

You can also use virtual links to temporarily recover from a partitioned area, which occurs when a link within the area fails, isolating part of the area from reaching the designated ABR to the backbone area.

Key Topic

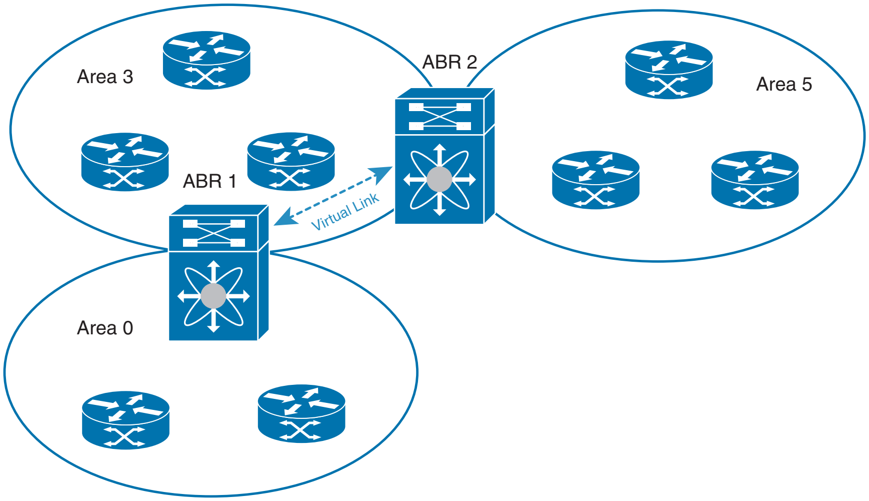


Figure 1-3 *OSPF Virtual Links*

Key Topic

Designated Routers and Backup Designated Routers

OSPF routers with the broadcast network type will flood the network with LSAs. The same link-state information needs to be sent from multiple sources. For this type, OSPF uses a single router, the *designated router* (DR), to control the LSA floods and represent the network to the rest of the OSPF area. OSPF selects a *backup designated router* (BDR). If the DR fails, the BDR will take the DR role of redistributing routing information.

Network types are as follows:

- **Point-to-point:** A network that exists only between two routers. All neighbors on a point-to-point network establish adjacency, and *there is no DR required*.
- **Broadcast:** A network with multiple routers that can communicate over a shared medium that allows broadcast traffic, such as Ethernet. OSPF routers establish a DR and BDR that control LSA flooding on the network. In OSPFv2, DR uses the well-known IPv4 multicast address 224.0.0.5 and the MAC address 0100.5e00.0005 to communicate with neighbors, and in OSPFv3, it uses the well-known IPv6 multicast address FF02::5 and the MAC address 3333.0000.0005 to communicate with neighbors. Likewise, in OSPFv2, each non-DR or non-BDR router uses the well-known IPv4 multicast address 224.0.0.6 and the MAC address 0100.5e00.0006 to send routing information to a DR or BDR, and in OSPFv3, it uses the well-known IPv6 multicast address FF02::6 and the MAC address 3333.0000.0006 to send routing information to a DR or BDR.

OSPF Authentication

OSPFv2 supports authentication to prevent unauthorized or invalid routing updates in the network. Cisco NX-OS supports two authentication methods:

- Simple password authentication
- MD5 authentication digest

Simple password authentication uses a simple cleartext password that is sent as part of the OSPFv2 message. The receiving OSPFv2 router must be configured with the same cleartext password to accept the OSPFv2 message as a valid route update. Because the password is in clear text, anyone who can watch traffic on the network can learn the password.

Cisco recommends that you use MD5 authentication to authenticate OSPFv2 messages. You can configure a password that is shared at the local router and all remote OSPFv2 neighbors. For each OSPFv2 message, Cisco NX-OS creates an MD5 one-way message digest based on the message itself and the encrypted password. The interface sends this digest with the OSPFv2 message. The receiving OSPFv2 neighbor validates the digest using the same encrypted password. If the message has not changed, the digest calculation is identical, and the OSPFv2 message is considered valid.

MD5 authentication includes a sequence number with each OSPFv2 message to ensure that no message is replayed in the network.

OSPFv3 doesn't have an authentication field in its header like OSPFv2; instead, OSPFv3 relies on IPsec.

OSPF Configurations and Verifications

Table 1-3 lists the OSPFv2/v3 default parameters. You can alter OSPF parameters as necessary. You are not required to alter any of these parameters, but the following parameters must be consistent across all routers in an attached network: ospf hello-interval and ospf dead-interval. If you configure any of these parameters, be sure that the configurations for all routers on your network have compatible values.

Key Topic

Table 1-3 Default OSPFv2/OSPFv3 Parameters

Parameters	Default
Hello interval	10 seconds
Dead interval	40 seconds
Graceful restart grace period	60 seconds
OSPFv2/OSPFv3 feature	Disabled
Stub router advertisement announce time	600 seconds
Reference bandwidth for link cost calculation	40 Gbps
LSA minimal arrival time	1000 milliseconds
LSA group pacing	240 seconds
SPF calculation initial delay time	200 milliseconds
SPF calculation maximum wait time	5000 milliseconds
SPF minimum hold time	1000 milliseconds

Cisco NX-OS is a modular system and requires a specific license to enable specific features. Table 1-4 covers the NX-OS feature licenses required for OSPFv2/OSPFv3. For more information, visit the Cisco NX-OS Licensing Guide.

Table 1-4 Feature-Based Licenses for Cisco NX-OS OSPFv2 and OSPFv3

Platform	Feature License	Feature Name
Cisco Nexus 9000 Series	Enterprise Services Package	OSPF
Cisco Nexus 7000 Series	LAN_ENTERPRISE_SERVICES_PKG	OSPFv3

Platform	Feature License	Feature Name
Cisco Nexus 6000 Series Cisco Nexus 5600 Series Cisco Nexus 5500 Series Cisco Nexus 5000 Series	Layer 3 Base Services Package LAN_BASE_SERVICES_PKG	OSPF OSPFv3
Cisco Nexus 3600 Series	Layer 3 Enterprise Services Package LAN_ENTERPRISE_SERVICES_PK	OSPF OSPFv3
Cisco Nexus 3000 Series	Layer 3 Base Services Package LAN_BASE_SERVICES_PK	OSPF (limited routes)

OSPFv2 and OSPFv3 have the following configuration limitations:

- Cisco NX-OS displays areas in dotted-decimal notation regardless of whether you enter the area in decimal or dotted-decimal notation.
- The OSPFv3 router ID and area ID are 32-bit numbers with no relationship to IPv6 addresses.

Tables 1-5 through 1-8 describe the most-used OSPFv2/v3 configuration commands. For a full list of the commands, refer to the Nexus Unicast Routing Configuration Guide links shown in the reference list at the end of this chapter.

Table 1-5 OSPF Global-Level Commands

Command	Purpose
feature ospf	Enables the OSPFv2 feature.
feature ospfv3	Enables the OSPFv3 feature.
router ospf *ospf-instance-tag*	Creates a new OSPFv2 routing instance.
router ospfv3 *ospf-instance-tag*	Creates a new OSPFv3 routing instance.

Table 1-6 OSPF Routing-Level Commands

Command	Purpose
router-id *ip-address*	(Optional) Configures a unique OSPFv2 or OSPFv3 router ID. *ip-address* must exist on a configured interface in the system.
area *area-id* **authentication** [**message-digest**]	Configures the authentication mode for an area.
area *area-id* **stub**	Creates this area as a stub area.
area *area-id* **nssa** [**no-redistribution**] [**default-information-originate**]**originate** [**route-map** *map-name*]] [**no-summary**] [**translate type7** {**always** | **never** } [**suppress-fa**]]	Creates this area as an NSSA.
address-family ipv6 unicast	Enters IPv6 unicast address family mode.

Table 1-7 OSPF Interface-Level Commands

Command	Purpose
ip ospf *cost number*	(Optional) Configures the OSPFv2 cost metric for this interface. The default is to calculate the cost metric, based on reference bandwidth and interface bandwidth. The range is from 1 to 65,535.
ip ospf dead-interval *seconds*	(Optional) Configures the OSPFv2 dead interval in seconds. The range is from 1 to 65,535. The default is four times the hello interval in seconds.
ip ospf hello-interval *seconds*	(Optional) Configures the OSPFv2 hello interval in seconds. The range is from 1 to 65,535. The default is 10 seconds.
ip ospf mtu-ignore	(Optional) Configures OSPFv2 to ignore any IP maximum transmission unit (MTU) mismatch with a neighbor. The default is not to establish adjacency if the neighbor MTU does not match the local interface MTU.
ospf network {**broadcast** | **point-point** }	(Optional) Sets the OSPFv2 network type.
[**default** | **no**] **ip ospf passive-interface**	(Optional) Suppresses routing updates on the interface. This command overrides the router or VRF command mode configuration. The default option removes this interface mode command and reverts to the router or VRF configuration if present.
ip ospf priority *number*	(Optional) Configures the OSPFv2 priority used to determine the DR for an area. The range is from 0 to 255. The default is 1.
ip ospf shutdown	(Optional) Shuts down the OSPFv2 instance on this interface.
ip ospf message-digest-key *key-id* **md5** [**0** | **3**] *key*	Configures message digest authentication for this interface. Use this command if the authentication is set to message-digest. The *key-id* range is from 1 to 255. The MD5 option 0 configures the password in clear text and 3 configures the pass key as 3DES encrypted.
ip router ospf *instance-tag* area *area-id* **[secondaries none]**	Adds the interface to the OSPFv2 instance and area.
ipv6 router ospfv3 *instance-tag* area *area-id* **[secondaries none]**	Adds the interface to the OSPFv3 instance and area.

Table 1-8 OSPF Global-Level Verification and Process Clear Commands

Command	Purpose
show ip ospf [*instance-tag*] [**vrf** *vrf-name*]	Displays the OSPFv2 configuration.
show ip ospf interface [*instance-tag*] [*interface-type interface-number*] [**brief**] [**vrf** *vrf-name*]	Displays the OSPFv2 interface configuration.
show ip ospf route [*ospf-route*] [**summary**] [**vrf** { *vrf-name* | **all** | **default** | **management** }]	Displays the internal OSPFv2 routes.

Command	Purpose
show ip ospf virtual-links [**brief**] [**vrf** { *vrf-name* | **all** | **default** | **management** }]	Displays information about OSPFv2 virtual links.
show running-configuration ospf	Displays the current running OSPFv2 configuration.
show ip ospf statistics [**vrf** { *vrf-name* | **all** | **default** | **management** }]	Displays the OSPFv2 event counters.
show ip ospf traffic [*interface - type number*] [**vrf** { *vrf-name* | **all** | **default** | **management** }]	Displays the OSPFv2 packet counters.
clear ip ospf [instance-tag] neighbor {* | *neighbor-id* | *interface-type number* | *loopback number* | *port-channel number*} [**vrf** *vrf-name*]	Clears neighbor statistics and resets adjacencies for Open Shortest Path First (OSPFv2). **NOTE:** Clearing the OSPF **neighbor** command will reload the OSPF process, so take extra precaution before executing the command in a production environment.
show [ipv6] ospfv3 [*instance-tag*] [**vrf** *vrf-name*]	Displays the OSPFv3 configuration.
show [ipv6] ospfv3 interface [*instance-tag*] [*interface-type interface-number*] [**brief**] [**vrf** *vrf-name*]	Displays the OSPFv3 interface configuration.
clear ospfv3 [*instance-tag*] **neighbor** {* | *neighbor-id* | **interface-type** *number* | **loopback number** | **port-channel** *number*} [**vrf** *vrf-name*]	Clears neighbor statistics and resets adjacencies for Open Shortest Path First (OSPFv3). **NOTE:** Clearing the OSPF **neighbor** command will reload the OSPF process, so take extra precaution before executing the command in a production environment.

Figure 1-4 shows the network topology for the configuration that follows, which demonstrates how to configure Nexus OSPF for IPv4 and IPv6.

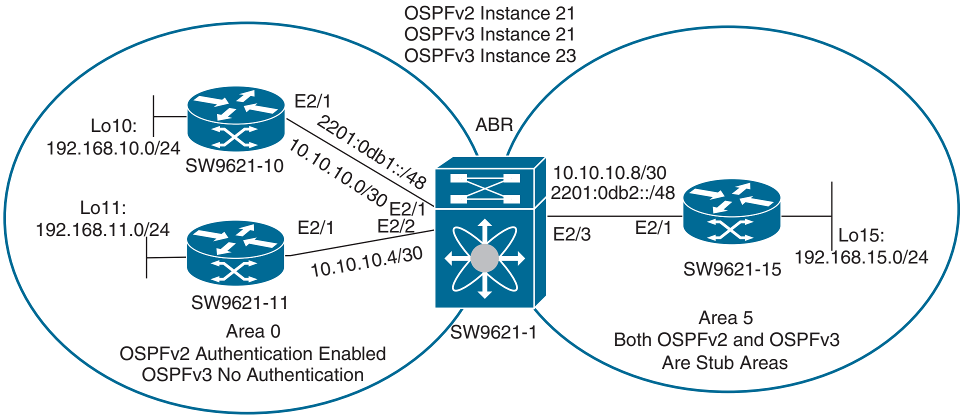


Figure 1-4 *OSPF Network Topology*

NOTE To practice Nexus configuration, you can use the Cisco Virtual Internet Routing Lab (VIRL) (http://virl.cisco.com/). If you don't have VIRL in your lab, you can use Cisco dCloud (https://dcloud.cisco.com); search for "VIRL 1.6.55 Sandbox."

Example 1-1 shows SW9621-1 OSPFv2 feature enabling and router configurations.

Example 1-1 *OSPF Instance 21*

```
SW9621-1(config)# feature ospf
SW9621-1(config)# router ospf 21
SW9621-1(config-router)# router-id 1.1.1.1
SW9621-1(config-router)# area 0.0.0.0 authentication message-digest
SW9621-1(config-router)# area 0.0.0.5 stub
```

Example 1-2 shows SW9621-1 OSPFv3 feature enabling and router configurations.

Example 1-2 *OSPF Instance 21 and 23*

```
SW9621-1(config)# feature ospfv3
SW9621-1(config)# router ospfv3 21
SW9621-1(config-router)# router-id 1.1.1.1
SW9621-1(config)# router ospfv3 23
SW9621-1(config-router)# area 0.0.0.5 stub
```

NOTE We didn't configure the router ID for OSPFv3 23; it is recommended that you configure the router ID.

Examples 1-3 and 1-4 show SW9621-1 OSFP interface and authentication configurations.

Example 1-3 *OSPF Interface Configurations*

```
SW9621-1(config)# interface loopback0
SW9621-1(config-if)# ip address 1.1.1.1/32
SW9621-1(config-if)# ip router ospf 21 area 0.0.0.0
SW9621-1(config)# interface Ethernet2/1
SW9621-1(config-if)# ip address 10.10.10.1/30
SW9621-1(config-if)# ip ospf authentication message-digest
SW9621-1(config-if)# ip ospf authentication key-chain mypass
SW9621-1(config-if)# ip router ospf 21 area 0.0.0.0
SW9621-1(config-if)# ipv6 address 2201:db1::1/48
SW9621-1(config-if)# ipv6 router ospf 21 area 0.0.0.0
SW9621-1(config-if)# no shutdown
SW9621-1(config)# interface Ethernet2/2
SW9621-1(config-if)# ip address 10.10.10.5/30
SW9621-1(config-if)# mtu 9216
SW9621-1(config-if)# ip ospf authentication message-digest
SW9621-1(config-if)# ip ospf authentication key-chain mypass
```

```
SW9621-1(config-if)# ip ospf network point-to-point
SW9621-1(config-if)# ip router ospf 21 area 0.0.0.0
SW9621-1(config-if)# no shutdown
SW9621-1(config)# interface Ethernet2/3
SW9621-1(config-if)# ip address 10.10.10.9/30
SW9621-1(config-if)# ip ospf hello-interval 25
SW9621-1(config-if)# ip router ospf 21 area 0.0.0.5
SW9621-1(config-if)# ipv6 address 2201:db2::1/48
SW9621-1(config-if)# ipv6 router ospf 23 area 0.0.0.5
SW9621-1(config-if)# no shutdown
SW9621-1(config-if)# exit
```

NOTE Use the **ip ospf mtu-ignore** command for OSPFv2 or **ipv6 ospf mtu-ignore** command for OSPFv3 to disable MTU mismatch detection on an interface. By default, OSPF checks whether neighbors use the same MTU on a common interface. If the receiving MTU is higher than the IP MTU configured on the incoming interface, OSPF does not establish adjacencies. The **mtu-ignore** command will disable this check and allow adjacencies when the MTU value differs between OSPF neighbors. This command will help only if the OSPF LSA database packet size is less than the lowest interface MTU. If the OSPF LSA packet is greater than the interface MTU, the physical interface will drop the packet as defragment disabled, and the OSPF neighbor will continuously change state between Down and Full.

NOTE The dead interval default will be 4xhello-interval. In this example, it is set to 100 seconds. You can set the dead interval using the OSPF interface command **ospf dead-interval** *seconds*.

Example 1-4 *OSPF Authentication Shared Key Configuration*

```
SW9621-1(config)# key chain mypass
SW9621-1(config-keychain)# key 0
SW9621-1(config-keychain-key)# key-string cisco
SW9621-1(config-keychain-key)# exit
SW9621-1(config-keychain)# exit
```

Examples 1-5 and 1-6 show the SW9621-1 OSPF status.

Example 1-5 *OSFPv2 ABR Verification (SW9621-1)*

```
SW9621-1# show ip ospf neighbors

OSPF Process ID 21 VRF default
 Total number of neighbors: 3
 Neighbor ID     Pri State            Up Time  Address         Interface
 1.1.1.10          10 FULL/BDR         00:20:19 10.10.10.2      Eth2/1
 1.1.1.11          1 FULL/ -           00:20:48 10.10.10.6      Eth2/2
 1.1.1.15          1 FULL/BDR          00:04:39 10.10.10.10     Eth2/3
```

```
SW9621-1# show ip ospf
 Routing Process 21 with ID 1.1.1.1 VRF default
 Routing Process Instance Number 1
 Stateful High Availability enabled
 Graceful-restart is configured
   Grace period: 60 state: Inactive
   Last graceful restart exit status: None
 Supports only single TOS(TOS0) routes
 Supports opaque LSA
 This router is an area border
 Administrative distance 110
 Reference Bandwidth is 40000 Mbps
 SPF throttling delay time of 200.000 msecs,
   SPF throttling hold time of 1000.000 msecs,
   SPF throttling maximum wait time of 5000.000 msecs
 LSA throttling start time of 0.000 msecs,
   LSA throttling hold interval of 5000.000 msecs,
   LSA throttling maximum wait time of 5000.000 msecs
 Minimum LSA arrival 1000.000 msec
 LSA group pacing timer 10 secs
 Maximum paths to destination 8
 Number of external LSAs 0, checksum sum 0
 Number of opaque AS LSAs 0, checksum sum 0
 Number of areas is 2, 1 normal, 1 stub, 0 nssa
 Number of active areas is 2, 1 normal, 1 stub, 0 nssa
 Install discard route for summarized external routes.
 Install discard route for summarized internal routes.
   Area BACKBONE(0.0.0.0)
        Area has existed for 00:37:14
        Interfaces in this area: 3 Active interfaces: 3
        Passive interfaces: 0  Loopback interfaces: 1
        Message-digest authentication
        SPF calculation has run 20 times
         Last SPF ran for 0.000420s
        Area ranges are
        Number of LSAs: 7, checksum sum 0x363a9
   Area (0.0.0.5)
        Area has existed for 00:37:14
        Interfaces in this area: 1 Active interfaces: 1
        Passive interfaces: 0  Loopback interfaces: 0
        This area is a STUB area
        Generates stub default route with cost 1
        No authentication available
        SPF calculation has run 20 times
         Last SPF ran for 0.000078s
        Area ranges are
        Number of LSAs: 11, checksum sum 0x5306a

SW9621-1# show ip ospf interface brief
 OSPF Process ID 21 VRF default
 Total number of interface: 4
 Interface               ID     Area            Cost   State    Neighbors Status
 Lo0                     1      0.0.0.0         1      LOOPBACK 0         up
 Eth2/1                  4      0.0.0.0         40     DR       1         up
 Eth2/2                  3      0.0.0.0         40     P2P      1         up
 Eth2/3                  2      0.0.0.5         40     DR       1         up


SW9621-1# show ip ospf interface
 loopback0 is up, line protocol is up
    IP address 1.1.1.1/32
    Process ID 21 VRF default, area 0.0.0.0
    Enabled by interface configuration
    State LOOPBACK, Network type LOOPBACK, cost 1
    Index 1
 Ethernet2/1 is up, line protocol is up
    IP address 10.10.10.1/30
    Process ID 21 VRF default, area 0.0.0.0
    Enabled by interface configuration
    State DR, Network type BROADCAST, cost 40
    Index 4, Transmit delay 1 sec, Router Priority 1
    Designated Router ID: 1.1.1.1, address: 10.10.10.1
    Backup Designated Router ID: 1.1.1.10, address: 10.10.10.2
    1 Neighbors, flooding to 1, adjacent with 1
    Timer intervals: Hello 10, Dead 40, Wait 40, Retransmit 5
      Hello timer due in 00:00:05
    Message-digest authentication, using keychain mypass (ready)
    Number of opaque link LSAs: 0, checksum sum 0
 Ethernet2/2 is up, line protocol is up
    IP address 10.10.10.5/30
    Process ID 21 VRF default, area 0.0.0.0
    Enabled by interface configuration
    State P2P, Network type P2P, cost 40
    Index 3, Transmit delay 1 sec
    1 Neighbors, flooding to 1, adjacent with 1
    Timer intervals: Hello 10, Dead 40, Wait 40, Retransmit 5
      Hello timer due in 00:00:02
    Message-digest authentication, using keychain mypass (ready)
    Number of opaque link LSAs: 0, checksum sum 0
 Ethernet2/3 is up, line protocol is up
    IP address 10.10.10.9/30
    Process ID 21 VRF default, area 0.0.0.5
    Enabled by interface configuration
    State DR, Network type BROADCAST, cost 40
    Index 2, Transmit delay 1 sec, Router Priority 1
    Designated Router ID: 1.1.1.1, address: 10.10.10.9
    Backup Designated Router ID: 1.1.1.15, address: 10.10.10.10
    1 Neighbors, flooding to 1, adjacent with 1
    Timer intervals: Hello 25, Dead 100, Wait 100, Retransmit 5
      Hello timer due in 00:00:06
    No authentication
    Number of opaque link LSAs: 0, checksum sum 0
```

Example 1-6 *OSFPv3 ABR Verification (SW9621-1)*

```
SW9621-1# show ospfv3 neighbors

 OSPFv3 Process ID 23 VRF default
 Total number of neighbors: 1
 Neighbor ID     Pri State            Up Time  Interface ID    Interface
 1.1.1.15        1 FULL/DR            00:00:54 36              Eth2/3
   Neighbor address fe80::200:ff:feff:1ff
 OSPFv3 Process ID 21 VRF default
 Total number of neighbors: 1
 Neighbor ID     Pri State            Up Time  Interface ID    Interface
 1.1.1.10        1 FULL/DR            01:02:49 38              Eth2/1
   Neighbor address fe80::200:ff:fe00:2f

SW9621-1# show ipv6 ospfv3 interface

 Ethernet2/3 is up, line protocol is up
    IPv6 address 2201:db2::1/48
    Process ID 23 VRF default, Instance ID 0, area 0.0.0.5
    Enabled by interface configuration
    State BDR, Network type BROADCAST, cost 40
    Index 1, Transmit delay 1 sec, Router Priority 1
    Designated Router ID: 1.1.1.15, address: fe80::200:ff:feff:1ff
    Backup Designated Router ID: 1.1.1.1, address: fe80::200:ff:feff:ddfe
    1 Neighbors, flooding to 1, adjacent with 1
    Timer intervals: Hello 10, Dead 40, Wait 40, Retransmit 5
      Hello timer due in 00:00:05
    Number of link LSAs: 2, checksum sum 0xe4da
```

```
Ethernet2/1 is up, line protocol is up
   IPv6 address 2201:db1::1/48
   Process ID 21 VRF default, Instance ID 0, area 0.0.0.0
   Enabled by interface configuration
   State BDR, Network type BROADCAST, cost 40
   Index 2, Transmit delay 1 sec, Router Priority 1
   Designated Router ID: 1.1.1.10, address: fe80::200:ff:fe00:2f
   Backup Designated Router ID: 1.1.1.1, address: fe80::200:ff:feff:ef22
   1 Neighbors, flooding to 1, adjacent with 1
   Timer intervals: Hello 10, Dead 40, Wait 40, Retransmit 5
     Hello timer due in 00:00:00
   Number of link LSAs: 2, checksum sum 0xbaac
```

Example 1-7 shows the SW9621-10 OSPFv2/v3 configurations and status.

Example 1-7 *Router 1 Configuration and Verification (SW9621-10)*

```
SW9621-10(config)# feature ospf
SW9621-10(config)# feature ospfv3
SW9621-10(config)# router ospf 21
SW9621-10(config-router)# router-id 1.1.1.10
SW9621-10(config-router)# area 0.0.0.0 authentication message-digest
SW9621-10(config)# router ospfv3 21
SW9621-10(config-router)# router-id 1.1.1.10
SW9621-10(config-router)# key chain mypass
SW9621-10(config-keychain)# key 0
SW9621-10(config-keychain-key)# key-string cisco

SW9621-10(config-keychain-key)# interface loopback0
SW9621-10(config-if)# ip address 1.1.1.10/32
SW9621-10(config-if)# ip router ospf 21 area 0.0.0.0

SW9621-10(config-if)# interface loopback10
SW9621-10(config-if)# ip address 192.168.10.1/24
SW9621-10(config-if)# ip router ospf 21 area 0.0.0.0

SW9621-10(config-if)# interface Ethernet2/1
SW9621-10(config-if)# no switchport
SW9621-10(config-if)# ip address 10.10.10.2/30
SW9621-10(config-if)# ip ospf authentication key-chain mypass
SW9621-10(config-if)# ip ospf priority 10        !# Note higher priorities win the
DR election
SW9621-10(config-if)# ipv6 address 2201:db1::2/48
SW9621-10(config-if)# ip router ospf 21 area 0.0.0.0
SW9621-10(config-if)# ipv6 router ospfv3 21 area 0.0.0.0
SW9621-10(config-if)# no shutdown
```

```
SW9621-10# show ip ospf neighbors
 OSPF Process ID 21 VRF default
 Total number of neighbors: 1
 Neighbor ID     Pri State            Up Time  Address         Interface
 1.1.1.1           1 FULL/DR          00:00:26 10.10.10.1      Eth2/1

SW9621-10# show ipv6 ospfv3 neighbors
 OSPFv3 Process ID 21 VRF default
 Total number of neighbors: 1
 Neighbor ID     Pri State            Up Time  Interface ID    Interface
 1.1.1.1           1 FULL/DR          00:01:54 37              Eth2/1
   Neighbor address fe80::200:ff:feff:ef22
```

Example 1-8 shows the SW9621-11 OSPFv2 configurations and status.

Example 1-8 *Router 2 Configuration and Verification (SW9621-11)*

```
SW9621-11(config)# feature ospf
SW9621-11(config)# router ospf 21
SW9621-11(config-router)# router-id 1.1.1.11
SW9621-11(config-router)# area 0.0.0.0 authentication message-digest
SW9621-11(config-router)# key chain mypass
SW9621-11(config-keychain)# key 0
SW9621-11(config-keychain-key)# key-string cisco
SW9621-11(config-keychain-key)# interface loopback0
SW9621-11(config-if)# ip address 1.1.1.11/32
SW9621-11(config-if)# ip router ospf 21 area 0.0.0.0
SW9621-11(config-if)# interface loopback11
SW9621-11(config-if)# ip address 192.168.11.1/24
SW9621-11(config-if)# ip router ospf 21 area 0.0.0.0
SW9621-11(config-if)# interface Ethernet2/1
SW9621-11(config-if)# mtu 9216
SW9621-11(config-if)# ip address 10.10.10.6/30
SW9621-11(config-if)# ip ospf authentication key-chain mypass
SW9621-11(config-if)# ip ospf network point-to-point
SW9621-11(config-if)# ip router ospf 21 area 0.0.0.0
SW9621-11(config-if)# no shutdown

SW9261-11# show ip ospf neighbors
 OSPF Process ID 21 VRF default
 Total number of neighbors: 1
 Neighbor ID     Pri State            Up Time  Address         Interface
 1.1.1.1           1 FULL/ -          01:23:31 10.10.10.5      Eth2/1

SW9261-11# show ipv6 osp?                  <-== Note: ospfv3 feature not enabled
                               ^
% Invalid command at '^' marker.
```

Example 1-9 shows the SW9621-15 OSPF configuration and verification status.

Example 1-9 *Router 3 Configuration and Verification (SW9621-15)*

```
SW9621-15(config)# feature ospf
SW9621-15(config)# feature ospfv3

SW9621-15(config)# router ospf 21
SW9621-15(config-router)# router-id 1.1.1.15
SW9621-15(config-router)# area 0.0.0.5 stub
SW9621-15(config)# router ospfv3 23
SW9621-15(config-router)# area 0.0.0.5 stub
SW9621-15(config-keychain-key)# interface loopback0
SW9621-15(config-if)# ip address 1.1.1.15/32
SW9621-15(config-if)# ip router ospf 21 area 0.0.0.5
SW9621-15(config-if)# interface loopback15
SW9621-15(config-if)# ip address 192.168.15.1/24
SW9621-15(config-if)# ip router ospf 21 area 0.0.0.5
SW9621-15(config-if)# interface Ethernet2/1
SW9621-15(config-if)# no switchport
SW9621-15(config-if)# ip address 10.10.10.10/30
SW9621-15(config-if)# ip ospf hello-interval 25
SW9621-15(config-if)# ip router ospf 21 area 0.0.0.5
SW9621-15(config-if)# ipv6 address 2201:db2::2/48
SW9621-15(config-if)# ipv6 router ospfv3 23 area 0.0.0.5
SW9621-15(config-if)# no shutdown

SW9621-15# show ip ospf neighbors
 OSPF Process ID 21 VRF default
 Total number of neighbors: 1
 Neighbor ID     Pri State            Up Time  Address         Interface
 1.1.1.1           1 FULL/DR          01:08:08 10.10.10.9      Eth2/1

SW9621-15# show ipv6 ospfv3 neighbors
 OSPFv3 Process ID 23 VRF default
 Total number of neighbors: 1
 Neighbor ID     Pri State            Up Time  Interface ID    Interface
 1.1.1.1           1 FULL/BDR         00:07:55 39              Eth2/1
   Neighbor address fe80::200:ff:feff:ddfe
```

Key Topic

Border Gateway Protocol

The Border Gateway Protocol (BGP) uses a path-vector routing algorithm to exchange routing information between BGP speakers. Based on this information, each BGP speaker determines a path to reach a particular destination while detecting and avoiding paths with routing loops. The routing information includes the actual route prefix for a destination, the path of autonomous systems to the destination, and additional path attributes.

BGP selects a single path, by default, as the best path to a destination host or network. Each path carries *well-known mandatory*, *well-known discretionary*, and *optional transitive attributes* that are used in BGP best-path analysis. You can influence BGP path selection by altering some of these attributes by configuring BGP policies.

BGP also supports load balancing or equal-cost multipath (ECMP), where next-hop packet forwarding to a single destination can occur over multiple "best paths" that tie for top place in routing metric calculations. It potentially offers substantial increases in bandwidth by load-balancing traffic over multiple paths.

Cisco NX-OS supports BGP version 4, which includes multiprotocol extensions that allow BGP to carry routing information for IP multicast routes and multiple Layer 3 protocol address families. BGP uses TCP (Port 179) as a reliable transport protocol to create TCP sessions with other BGP-enabled devices.

The BGP autonomous system (AS) is a network controlled by a single administration entity. An autonomous system forms a routing domain with one or more Interior Gateway Protocols (IGPs) and a consistent set of routing policies. BGP supports 16-bit and 32-bit autonomous system numbers.

External BGP autonomous systems dynamically exchange routing information through external BGP (eBGP) peering sessions. BGP speakers within the same autonomous system can exchange routing information through internal BGP (iBGP) peering sessions.

BGP supports 2-byte or 4-byte AS numbers. Cisco NX-OS displays 4-byte AS numbers in plain-text notation (that is, as 32-bit integers). You can configure 4-byte AS numbers as either plain-text notation (for example, 1 to 42,94,967,295) or AS.dot notation (for example, 1.0).

BGP Peering

A BGP speaker does not discover and peer with another BGP speaker automatically. You must configure the relationships between BGP speakers. A *BGP peer* is a BGP speaker that has an active TCP connection to another BGP speaker.

BGP uses TCP port 179 to create a TCP session with a peer. When a TCP connection is established between peers, each BGP peer initially exchanges all of its routes—the complete BGP routing table—with the other peer. After this initial exchange, the BGP peers send only *incremental* updates when a topology change occurs in the network or when a routing policy change occurs. In the periods of inactivity between these updates, peers exchange special messages called *keepalives*. The *hold time* is the maximum time limit that can elapse between receiving consecutive BGP update or keepalive messages. Cisco NX-OS supports the following peer configuration options:

Key Topic

- **Individual IPv4 or IPv4 address:** BGP establishes a session with the BGP speaker that matches the remote address and AS number.
- **IPv4 or IPv6 prefix peers for a single AS number:** BGP establishes sessions with BGP speakers that match the prefix and the AS number.
- **Dynamic AS number prefix peers:** BGP establishes sessions with BGP speakers that match the prefix and an AS number from a list of configured AS numbers.

Cisco NX-OS accepts a range or list of AS numbers to establish BGP sessions and does not associate prefix peers with dynamic AS numbers as either interior BGP (iBGP) or external BGP (eBGP) sessions until after the session is established.

For example, if you configure BGP to use IPv4 prefix 172.16.2.0/8 and AS numbers 10, 30, and 100, BGP establishes a session with 172.16.2.1 with AS number 30 but rejects a session from 172.16.2.2 with AS number 20.

NOTE The dynamic AS number prefix peer configuration overrides the individual AS number configuration that is inherited from a BGP template.

To establish BGP sessions between peers, BGP must have a router ID, which is sent to BGP peers in the OPEN message when a BGP session is established. The BGP router ID is a 32-bit value that is often represented by an IPv4 address. You can configure the router ID. By default, Cisco NX-OS sets the router ID to the IPv4 address of a loopback interface on the router. If no loopback interface is configured on the router, the software chooses the highest IPv4 address configured to a physical interface on the router to represent the BGP router ID. The BGP router ID must be unique to the BGP peers in a network.

NOTE If BGP does not have a router ID, it cannot establish any peering sessions with BGP peers.

Key Topic

BGP Path Selection

The best-path algorithm runs each time a path is added or withdrawn for a given network. The best-path algorithm also runs if you change the BGP configuration. BGP selects the best path from the set of valid paths available for a given network.

Cisco NX-OS implements the BGP best-path algorithm in the following steps.

Step 1: Comparing Pairs of Paths

This first step in the BGP best-path algorithm compares two paths to determine which path is better. The following sequence describes the basic steps that Cisco NX-OS uses to compare two paths to determine the better path:

1. Cisco NX-OS chooses a valid path for comparison. (For example, a path that has an unreachable next-hop is not valid.)
2. Cisco NX-OS chooses the path with the highest weight.
3. Cisco NX-OS chooses the path with the highest local preference.
4. If one of the paths is locally originated, Cisco NX-OS chooses that path.
5. Cisco NX-OS chooses the path with the shorter AS-path.

NOTE When calculating the length of the AS-path, Cisco NX-OS ignores confederation segments and counts AS sets as 1.

6. Cisco NX-OS chooses the path with the lower origin. Interior Gateway Protocol (IGP) is considered lower than EGP.

7. Cisco NX-OS chooses the path with the lower multi-exit discriminator (MED).

   You can configure a number of options that affect whether this step is performed. In general, Cisco NX-OS compares the MED of both paths if the paths were received from peers in the same autonomous system; otherwise, Cisco NX-OS skips the MED comparison.

   You can configure Cisco NX-OS to always perform the best-path algorithm MED comparison, regardless of the peer autonomous system in the paths. Otherwise, Cisco NX-OS will perform a MED comparison that depends on the AS-path attributes of the two paths being compared:

   a. If a path has no AS-path or the AS-path starts with an AS_SET, the path is internal, and Cisco NX-OS compares the MED to other internal paths.

   b. If the AS-path starts with an AS_SEQUENCE, the peer autonomous system is the first AS number in the sequence, and Cisco NX-OS compares the MED to other paths that have the same peer autonomous system.

   c. If the AS-path contains only confederation segments or starts with confederation segments followed by an AS_SET, the path is internal and Cisco NX-OS compares the MED to other internal paths.

   d. If the AS-path starts with confederation segments followed by an AS_SEQUENCE, the peer autonomous system is the first AS number in the AS_SEQUENCE, and Cisco NX-OS compares the MED to other paths that have the same peer autonomous system.

NOTE If Cisco NX-OS receives no MED attribute with the path, Cisco NX-OS considers the MED to be 0 unless you configure the best-path algorithm to set a missing MED to the highest possible value.

   e. If the nondeterministic MED comparison feature is enabled, the best-path algorithm uses the Cisco IOS style of MED comparison.

8. If one path is from an internal peer and the other path is from an external peer, Cisco NX-OS chooses the path from the external peer.

9. If the paths have different IGP metrics to their next-hop addresses, Cisco NX-OS chooses the path with the lower IGP metric.

10. Cisco NX-OS uses the path that was selected by the best-path algorithm the last time that it was run.

    If all path parameters in step 1 through step 9 are the same, you can configure the best-path algorithm to compare the router IDs. If the path includes an originator attribute, Cisco NX-OS uses that attribute as the router ID to compare to; otherwise, Cisco NX-OS uses the router ID of the peer that sent the path. If the paths have different router IDs, Cisco NX-OS chooses the path with the lower router ID.

NOTE When the attribute originator is used as the router ID, it is possible that two paths have the same router ID. It is also possible to have two BGP sessions with the same peer router, and therefore, you can receive two paths with the same router ID.

11. Cisco NX-OS selects the path with the shorter cluster length. If a path was not received with a cluster list attribute, the cluster length is 0.
12. Cisco NX-OS chooses the path received from the peer with the lower IP address. Locally generated paths (for example, redistributed paths) have a peer IP address of 0.

NOTE Paths that are equal after step 9 can be used for multipath if you configure it.

Step 2: Determining the Order of Comparisons

The second step of the BGP best-path algorithm implementation is to determine the order in which Cisco NX-OS compares the paths:

1. Cisco NX-OS partitions the paths into groups. Within each group, Cisco NX-OS compares the MED among all paths. Cisco NX-OS uses the same rules as in step 1 to determine whether MED can be compared between any two paths. Typically, this comparison results in one group being chosen for each neighbor autonomous system. If you configure the **bgp bestpath med always** command, Cisco NX-OS chooses just one group that contains all the paths.
2. Cisco NX-OS determines the best path in each group by iterating through all paths in the group and keeping track of the best one so far. Cisco NX-OS compares each path with the temporary best path found so far, and if the new path is better, it becomes the new temporary best path, and Cisco NX-OS compares it with the next path in the group.
3. Cisco NX-OS forms a set of paths that contain the best path selected from each group in step 2. Cisco NX-OS selects the overall best path from this set of paths by going through them as in step 2.

Step 3: Determining the Best-Path Change Suppression

The next part of the implementation is to determine whether Cisco NX-OS will use the new best path or suppress it. The router can continue to use the existing best path if the new one is identical to the old path (if the router ID is the same). Cisco NX-OS continues to use the existing best path to avoid route changes in the network.

You can turn off the suppression feature by configuring the best-path algorithm to compare the router IDs. If you configure this feature, the new best path is always preferred to the existing one.

You cannot suppress the best-path change if any of the following conditions occur:

- The existing best path is no longer valid.
- Either the existing or new best paths were received from internal (or confederation) peers or were locally generated (for example, by redistribution).
- The paths were received from the same peer (the paths have the same router ID).
- The paths have different weights, local preferences, origins, or IGP metrics to their next-hop addresses.
- The paths have different MEDs.

NOTE The order of comparison determined in Step 2 is important. Consider the case where you have three paths—A, B, and C. When Cisco NX-OS compares A and B, it chooses A. When Cisco NX-OS compares B and C, it chooses B. But when Cisco NX-OS compares A and C, it might not choose A because some BGP metrics apply only among paths from the same neighboring autonomous system and not among all paths.

The path selection uses the BGP AS-path attribute. The AS-path attribute includes the list of autonomous system numbers (AS numbers) traversed in the advertised path. If you subdivide your BGP autonomous system into a collection or confederation of autonomous systems, the AS-path contains confederation segments that list these locally defined autonomous systems.

Key Topic

Multiprotocol BGP

Cisco NX-OS supports multiple address families. Multiprotocol BGP (MBGP) can carry different sets of routes depending on the address family. For example, BGP can carry one set of routes for IPv4 unicast routing, one set of routes for IPv4 multicast routing, and one set of routes for IPv6 multicast routing. You can use MBGP for Reverse Path Forwarding (RPF) checks in IP multicast networks.

NOTE Because Multicast BGP does not propagate multicast state information, you need a multicast protocol, such as Protocol Independent Multicast (PIM).

You need to use the router address-family and neighbor address-family configuration modes to support Multiprotocol BGP configurations. MBGP maintains separate Routing Information Bases (RIBs) for each configured address family, such as a unicast RIB and a multicast RIB for BGP.

A Multiprotocol BGP network is backward compatible, but BGP peers that do not support multiprotocol extensions cannot forward routing information, such as address family identifier information, that the multiprotocol extensions carry.

BGP Configurations and Verifications

Table 1-9 lists the BGP default parameters; you can alter BGP default parameters as necessary.

Table 1-9 Default BGP Parameters

Parameters	Default
BGP feature	Disabled
Keepalive interval	60 seconds
Hold timer	180 seconds
BGP PIC core	Enabled
Auto-summary	Always disabled
Synchronization	Always disabled

Table 1-10 shows NX-OS feature license required for BGP. For more information, visit the Cisco NX-OS Licensing Guide.

Table 1-10 Feature-Based Licenses for Cisco NX-OS

Platform	Feature License	Feature Name
Cisco Nexus 9000 Series Cisco Nexus 7000 Series	Enterprise Services Package LAN_ENTERPRISE_SERVICES_PKG	BGP
Cisco Nexus 6000 Series Cisco Nexus 5600 Series Cisco Nexus 5500 Series Cisco Nexus 5000 Series	Layer 3 Enterprise Services Package LAN_ENTERPRISE_SERVICES_PKG	BGP
Cisco Nexus 3600 Series	Layer 3 Enterprise Services Package LAN_ENTERPRISE_SERVICES_PKG	BGP
Cisco Nexus 3000 Series	Layer 3 Enterprise Services Package LAN_ENTERPRISE_SERVICES_PKG	BGP

BGP has the following configuration limitations:

- The dynamic AS number prefix peer configuration overrides the individual AS number configuration inherited from a BGP template.
- If you configure a dynamic AS number for prefix peers in an AS confederation, BGP establishes sessions with only the AS numbers in the local confederation.
- BGP sessions created through a dynamic AS number prefix peer ignore any configured eBGP multihop time-to-live (TTL) value or a disabled check for directly connected peers.
- Configure a router ID for BGP to avoid automatic router ID changes and session flaps.
- Use the maximum-prefix configuration option per peer to restrict the number of routes received and system resources used.
- Configure the update source to establish a session with BGP/eBGP multihop sessions.
- Specify a BGP policy if you configure redistribution.
- Define the BGP router ID within a VRF.
- If you decrease the keepalive and hold timer values, you might experience BGP session flaps.
- The BGP minimum route advertisement interval (MRAI) value for all iBGP and eBGP sessions is zero and is not configurable.

Tables 1-11 through 1-13 describe the most-used BGP configuration commands. For a full list of the commands, refer to the Nexus Unicast Routing Configuration Guide listed in the reference section at the end of the chapter.

Table 1-11 BGP Global-Level Configurations

Command	Purpose
feature bgp	Enables the BGP feature.
router bgp *autonomous-system-number*	Enables BGP and assigns the AS number to the local BGP speaker. The AS number can be a 16-bit integer or a 32-bit integer in the form of a higher 16-bit decimal number and a lower 16-bit decimal number in *xx.xx* format.

Table 1-12 BGP Routing-Level Configurations

Command	Purpose
router-id *ip-address*	(Optional) Configures a unique BGP router ID. This IP address identifies this BGP speaker.
description *text*	(Optional) Adds a description for the neighbor. The description is an alphanumeric string up to 80 characters.
neighbor { *ip-address* | *ipv6-address* } **remote-as** *as-number*	Configures the IPv4 or IPv6 address and AS number for a remote BGP peer. The *ip-address* format is *x.x.x.x*. The *ipv6-address* format is A:B::C:D.
address-family { **ipv4** | **ipv6** | **vpnv4** | **vpnv6** }{ **unicast** | **multicast** }	(Optional) Enters global address family configuration mode for the IP or VPN address family.
network *ip-prefix* [**route-map** *map-name*]	(Optional) Specifies a network as local to this autonomous system and adds it to the BGP routing table. For exterior protocols, the **network** command controls which networks are advertised. Interior protocols use the **network** command to determine where to send updates.
timers *keepalive-time hold-time*	(Optional) Adds the keepalive and hold time BGP timer values for the neighbor. The range is from 0 to 3600 seconds. The default is 60 seconds for the keepalive time and 180 seconds for the hold time.
shutdown	(Optional) Administratively shuts down this BGP neighbor. This command triggers an automatic notification and session reset for the BGP neighbor sessions.
neighbor *prefix* **remote-as route-map** *map-name*	Configures the IPv4 or IPv6 prefix and a route map for the list of accepted AS numbers for the remote BGP peers. The *prefix* format for IPv4 is x.x.x.x/length. The length range is from 1 to 32. The *prefix* format for IPv6 is A:B::C:D/length. The length range is from 1 to 128. The *map-name* can be any case-sensitive, alphanumeric string up to 63 characters.

Table 1-13 BGP Verification and BGP Clear Commands

Command	Purpose
show bgp all [**summary**] [**vrf** *vrf-name*]	Displays the BGP information for all address families.
show {**ipv** | **ipv6**} **bgp** *options*	Displays the BGP status and configuration information. This command has multiple options. One important option is summary (**show ip bgp summary**).
show bgp convergence [**vrf** *vrf-name*]	Displays the BGP information for all address families.
show bgp {**ip** | **ipv6**} {**unicast** | **multicast**} [*ip-address* | *ipv6-prefix*] **community** {**regexp expression** | [**community**] [**no-advertise**] [**no-export**] [**no-export-subconfed**]} [**vrf** *vrf-name*]	Displays the BGP routes that match a BGP community.
show bgp *process*	Displays the BGP process information.
show running-configuration bgp	Displays the current running BGP configuration.
show bgp sessions [**vrf** *vrf-name*]	Displays the BGP sessions for all peers. You can use the **clear bgp sessions** command to clear these statistics.
show bgp statistics	Shows BGP statistics.
clear bgp all { **neighbor** | * | *as-number* | **peer-template** *name* | *prefix* } [**vrf** *vrf-name*]	Clears one or more neighbors from all address families. * clears all neighbors in all address families. The arguments are as follows: *neighbor*: IPv4 or IPv6 address of a neighbor. *as-number*: Autonomous system number. The AS number can be a 16-bit integer or a 32-bit integer in the form of higher 16-bit decimal number and a lower 16-bit decimal number in xx.xx format. *name*: Peer template name. The name can be any case-sensitive, alphanumeric string up to 64 characters. *prefix*: IPv4 or IPv6 prefix. All neighbors within that prefix are cleared. *vrf-name*: VRF name. All neighbors in that VRF are cleared. The name can be any case-sensitive, alphanumeric string up to 64 characters.
clear bgp all dampening [**vrf** *vrf-name*]	Clears route flap dampening networks in all address families. The *vrf-name* can be any case-sensitive, alphanumeric string up to 64 characters.
clear bgp all flap-statistics [**vrf** *vrf-name*]	Clears route flap statistics in all address families. The *vrf-name* can be any case-sensitive, alphanumeric string up to 64 characters.

Figure 1-5 shows the network topology for the configuration that follows, which demonstrates how to configure Nexus BGP for IPv4 and IPv6.

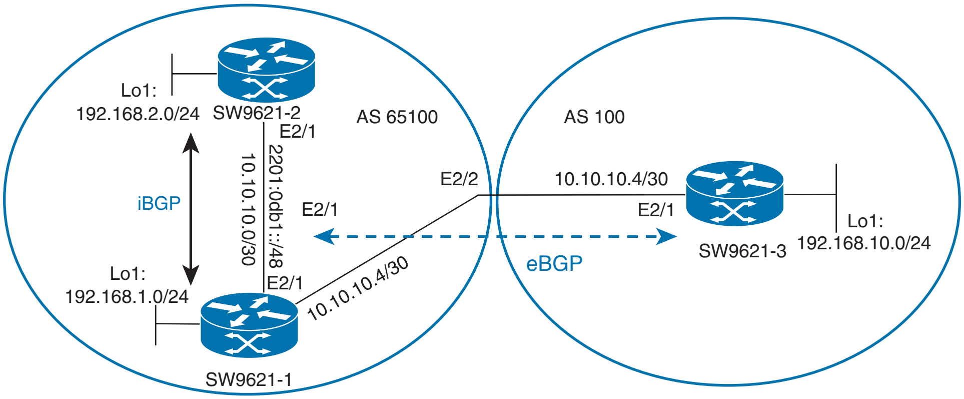


Figure 1-5 *BGP Network Topology*

Example 1-10 shows SW9621-1 BGP feature enabling and BGP router configurations.

Example 1-10 *BGP AS 65100 Creation Configuration*

```
SW9621-1(config)# feature bgp
SW9621-1(config)# router bgp 65100
SW9621-1(config-router)# router-id 1.1.1.1
SW9621-1(config-router)# address-family ipv4 unicast
SW9621-1(config-router-af)# network 192.168.1.0/24
SW9621-1(config-router)# neighbor 2201:db1::2 remote-as 65100
SW9621-1(config-router-neighbor)# address-family ipv6 unicast
SW9621-1(config-router)# neighbor 10.10.10.2 remote-as 65100
SW9621-1(config-router-neighbor)# address-family ipv4 unicast
SW9621-1(config-router-neighbor)# address-family ipv4 multicast
SW9621-1(config-router)# neighbor 10.10.10.6 remote-as 100
SW9621-1(config-router-neighbor)# address-family ipv4 unicast
```

Example 1-11 shows SW9621-1 interface configurations.

Example 1-11 *Interface Configurations*

```
SW9621-1(config)# interface loopback1
SW9621-1(config-if)# ip address 192.168.1.1/24
SW9621-1(config)# interface Ethernet2/1
SW9621-1(config-if)# ip address 10.10.10.1/30
SW9621-1(config-if)# no shutdown
SW9621-1(config)# interface Ethernet2/2
SW9621-1(config-if)# ip address 10.10.10.5/30
SW9621-1(config-if)# no shutdown
```

Example 1-12 shows the SW9621-1 BGP neighbors and summary.

Example 1-12 *BGP Verification*

```
SW9621-1(config)# show ip bgp summary

BGP summary information for VRF default, address family IPv4 Unicast
BGP router identifier 1.1.1.1, local AS number 65100
BGP table version is 14, IPv4 Unicast config peers 2, capable peers 2
3 network entries and 3 paths using 432 bytes of memory
BGP attribute entries [3/432], BGP AS path entries [1/6]
BGP community entries [0/0], BGP clusterlist entries [0/0]

Neighbor        V    AS    MsgRcvd MsgSent   TblVer          InQ OutQ             Up/Down
State/PfxRcd
10.10.10.2      4   65100    154      152            14          0     0              01:46:14
1
10.10.10.6      4    100     23       25             14          0     0              00:10:52
1
SW9621-1(config)# show ip bgp
BGP routing table information for VRF default, address family IPv4 Unicast
BGP table version is 14, local router ID is 1.1.1.1
Status: s-suppressed, x-deleted, S-stale, d-dampened, h-history, *-valid, >-best
Path type: i-internal, e-external, c-confed, l-local, a-aggregate, r-redist, I-injected
Origin codes: i - IGP, e - EGP, ? - incomplete, | - multipath, & - backup

   Network            Next Hop            Metric     LocPrf          Weight     Path
*>l192.168.1.0/24     0.0.0.0                        100             32768      i
*>i192.168.2.0/24     10.10.10.2                     100             0          i
*>e192.168.3.0/24     10.10.10.6                                     0          100 i

SW9621-1(config)# show bgp all

BGP routing table information for VRF default, address family IPv4 Unicast
BGP table version is 14, local router ID is 1.1.1.1
Status: s-suppressed, x-deleted, S-stale, d-dampened, h-history, *-valid, >-best
Path type: i-internal, e-external, c-confed, l-local, a-aggregate, r-redist, I-injected
Origin codes: i - IGP, e - EGP, ? - incomplete, | - multipath, & - backup

   Network            Next Hop            Metric     LocPrf          Weight     Path
*>l192.168.1.0/24     0.0.0.0                        100             32768      i
*>i192.168.2.0/24     10.10.10.2                     100             0          i
*>e192.168.3.0/24     10.10.10.6                                     0          100 i

SW9621-1(config)# show bgp sessions

Total peers 3, established peers 3
ASN 65100
```

```
VRF default, local ASN 65100
peers 3, established peers 3, local router-id 1.1.1.1
State: I-Idle, A-Active, O-Open, E-Established, C-Closing, S-Shutdown

Neighbor        ASN    Flaps    LastUpDn|LastRead|LastWrit St Port(L/R)     Notif(S/R)
10.10.10.2    65100     3        01:48:28|00:00:23|00:00:20 E   14994/179        0/3
10.10.10.6      100     1        00:13:05|00:00:20|00:00:04 E   16596/179        0/1
2201:db1::2   65100     0        00:27:27|00:00:25|00:00:25 E   179/46276        0/0

SW9621-1(config)# show ipv6 bgp sum

BGP summary information for VRF default, address family IPv6 Unicast
BGP router identifier 1.1.1.1, local AS number 65100
BGP table version is 3, IPv6 Unicast config peers 1, capable peers 1
0 network entries and 0 paths using 0 bytes of memory
BGP attribute entries [0/0], BGP AS path entries [0/0]
BGP community entries [0/0], BGP clusterlist entries [0/0]

Neighbor    V    AS     MsgRcvd MsgSent    Tbl   Ver   InQ OutQ      Up/Down  State/PfxRcd
2201:db1::2  4   65100     33        34       3    0     0             00:27:59      0
```

Example 1-13 shows SW9621-2 full configurations and BGP neighbors.

Example 1-13 *Router 2 Configuration and Verification (SW9621-2)*

```
SW9621-2(config-router-neighbor)# show run bgp

feature bgp
router bgp 65100
  router-id 2.2.2.2
  address-family ipv4 unicast
    network 192.168.2.0/24
  neighbor 2201:db1::1 remote-as 65100
    address-family ipv6 unicast
  neighbor 10.10.10.1 remote-as 65100
    address-family ipv4 unicast
    address-family ipv4 multicast


SW9621-2(config-router-neighbor-af)# show run int lo1
interface loopback1
  ip address 192.168.2.1/24

SW9621-2(config-router-neighbor)# show run int e2/1

interface Ethernet2/1
  no switchport
  mac-address 0000.0000.002f
  ip address 10.10.10.2/30
  ipv6 address 2201:db1::2/48
  no shutdown

SW9621-2(config-router-neighbor)# show bgp all
BGP routing table information for VRF default, address family IPv4 Unicast
BGP table version is 15, local router ID is 2.2.2.2
Status: s-suppressed, x-deleted, S-stale, d-dampened, h-history, *-valid, >-best
Path type: i-internal, e-external, c-confed, l-local, a-aggregate, r-redist,
I-injected
Origin codes: i - IGP, e - EGP, ? - incomplete, | - multipath, & - backup

   Network            Next Hop            Metric     LocPrf        Weight   Path
*>i192.168.1.0/24     10.10.10.1                      100          0          i
*>l192.168.2.0/24     0.0.0.0                         100          32768      i
  i192.168.3.0/24     10.10.10.6                      100          0          100 i
```

Example 1-14 shows SW9621-3 full configurations and BGP neighbors.

Example 1-14 *Router 3 Configuration (SW9621-3)*

```
SW9621-3(config-router-neighbor)# show run bgp
feature bgp
router bgp 100
  router-id 3.3.3.3
  address-family ipv4 unicast
    network 192.168.3.0/24
  neighbor 10.10.10.5 remote-as 65100
    address-family ipv4 unicast
      weight 100

SW9621-3(config-router-neighbor-af)# show run int lo
interface loopback10
  ip address 192.168.3.1/24

SW9621-3(config-router-neighbor-af)# show run int e2/1
interface Ethernet2/1
  no switchport
  mac-address 0000.0000.003f
  ip address 10.10.10.6/30
  no shutdown

SW9621-3# show bgp all
BGP routing table information for VRF default, address family IPv4 Unicast
BGP table version is 11, local router ID is 3.3.3.3
Status: s-suppressed, x-deleted, S-stale, d-dampened, h-history, *-valid, >-best
Path type: i-internal, e-external, c-confed, l-local, a-aggregate, r-redist,
I-injected
```

```
Origin codes: i - IGP, e - EGP, ? - incomplete, | - multipath, & - backup

   Network            Next Hop            Metric     LocPrf     Weight  Path
*>e192.168.1.0/24     10.10.10.5                                  100   65100 i
*>e192.168.2.0/24     10.10.10.5                                  100   65100 i
*>l192.168.3.0/24      0.0.0.0                           100      32768       i
```

Key Topic

Bidirectional Forwarding Detection

Bidirectional Forwarding Detection (BFD) is a detection protocol designed to provide fast forwarding–path failure detection times for media types, encapsulations, topologies, and routing protocols. BFD provides subsecond failure detection between two adjacent devices and can be less CPU-intensive than protocol hello messages because some of the BFD load can be distributed onto the data plane on supported modules.

Cisco NX-OS supports the BFD asynchronous mode, which sends BFD control packets between two adjacent devices to activate and maintain BFD neighbor sessions between the devices. You configure BFD on both devices (or BFD neighbors). Once BFD has been enabled on the interfaces and on the appropriate protocols, Cisco NX-OS creates a BFD session, negotiates BFD session parameters, and begins to send BFD control packets to each BFD neighbor at the negotiated interval. The BFD session parameters include the following:

- **Desired minimum transmit interval:** The interval at which this device wants to send BFD hello messages.
- **Required minimum receive interval:** The minimum interval at which this device can accept BFD hello messages from another BFD device.
- **Detect multiplier:** The number of missing BFD hello messages from another BFD device before this local device detects a fault in the forwarding path.

Figure 1-6 shows how a BFD session is established. The figure shows a simple network with two routers running OSPF and BFD. When OSPF discovers a neighbor (1), it sends a request to the local BFD process to initiate a BFD neighbor session with the OSPF neighbor router (2). The BFD neighbor session with the OSPF neighbor router is now established (3).

Key Topic

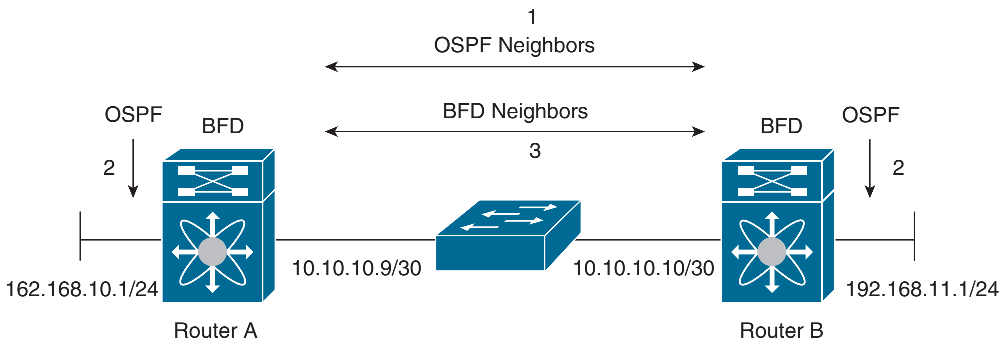


Figure 1-6 *Establishing a BFD Neighbor Relationship*

Key Topic

Rapid Detection of Failures

After a BFD session has been established and timer negotiations are complete, BFD neighbors send BFD control packets that act in the same manner as an IGP hello protocol to detect liveliness, except at a more accelerated rate. BFD detects a failure, but the protocol must take action to bypass a failed peer.

BFD sends a failure detection notice to the BFD-enabled protocols when it detects a failure in the forwarding path. The local device can then initiate the protocol recalculation process and reduce the overall network convergence time.

Figure 1-7 shows what happens when a failure occurs in the network (1). The BFD neighbor session with the OSPF neighbor router is torn down (2). BFD notifies the local OSPF process that the BFD neighbor is no longer reachable (3). The local OSPF process tears down the OSPF neighbor relationship (4). If an alternative path is available, the routers immediately start converging on it.

NOTE The BFD failure detection occurs in less than a second, which is much faster than OSPF hello messages could detect the same failure.

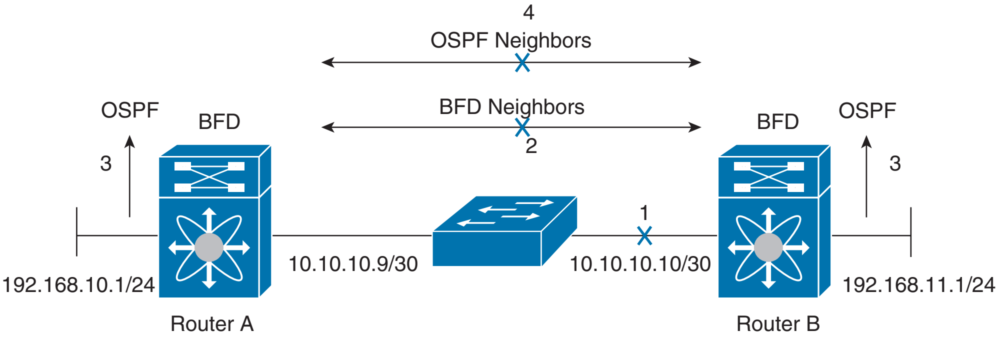


Figure 1-7 *Tearing Down an OSPF Neighbor Relationship*

BFD Configurations and Verifications

Table 1-14 lists BFD default parameters; you can alter BFD parameters as necessary.

Key Topic

Table 1-14 Default Settings for BFD Parameters

Parameters	Default
BFD feature	Disabled
Required minimum receive interval	50 milliseconds
Desired minimum transmit interval	50 milliseconds
Detect multiplier	3
Echo function	Enabled
Mode	Asynchronous
Port channel	Logical mode (one session per source-destination pair address)
Slow timer	2000 milliseconds
Subinterface optimization	Disabled

NOTE No license is required for BFD features.

BFD has the following configuration limitations:

- NX-OS supports BFD version 1.
- BFD supports single-hop BFD; BFD for BGP supports single-hop EBGP and iBGP peers.
- BFD depends on Layer 3 adjacency information to discover topology changes, including Layer 2 topology changes. A BFD session on a VLAN interface (SVI) may not be up after the convergence of the Layer 2 topology if no Layer 3 adjacency information is available.
- For port channels used by BFD, you must enable the Link Aggregation Control Protocol (LACP) on the port channel.
- HSRP for IPv4 is supported with BFD. HSRP for IPv6 is not supported with BFD.

Tables 1-15 through 1-18 described the most-used BFD configuration commands. For a full list of the commands, refer to the Nexus Unicast Routing Configuration Guide in the reference section at the end of the chapter.

Table 1-15 BFD Global-Level Configurations

Command	Purpose
feature bfd	Enables the BFD feature.
bfd interval *mintx* **min_rx** *msec* **multiplier** *value*	Configures the BFD session parameters for all BFD sessions on the device. You can override these values by configuring the BFD session parameters on an interface. The *mintx* and *msec* range is from 50 to 999 milliseconds, and the default is 50. The multiplier range is from 1 to 50. The multiplier default is 3.
bfd slow-timer *msec*	Configures the slow timer used in the echo function. This value determines how fast BFD starts up a new session and at what speed the asynchronous sessions use for BFD control packets when the echo function is enabled. The slow-timer value is used as the new control packet interval, while the echo packets use the configured BFD intervals. The echo packets are used for link failure detection, while the control packets at the slower rate maintain the BFD session. The range is from 1000 to 30,000 milliseconds. The default is 2000.

Table 1-16 BFD Routing-Level Configurations

Command	Purpose
router ospf *instance-tag* **bfd**	(Optional) Enables BFD for all OSPFv2 interfaces.
router bgp *as-number* **neighbor** { *ip-address* | *ipv6-address* } **remote-as** *as-number* **bfd**	Enables BFD for this BGP peer.

Table 1-17 BFD Interface-Level Command

Command	Purpose
bfd echo	Enables the echo function. The default is enabled.
bfd optimize subinterface	Optimizes subinterfaces on a BFD-enabled interface. The default is disabled.
ip ospf bfd	(Optional) Enables or disables BFD on an OSPFv2 interface. The default is disabled.

Table 1-18 BFD Verification Commands

Command	Purpose
show bfd neighbors [*application name*] [**details**]	Displays information about BFD for a supported application, such as BGP or OSPFv2.
show bfd neighbors [**interface** *int-if*] [**details**]	Displays information about BGP sessions on an interface.

Figure 1-8 shows the network topology for the configuration that follows, which demonstrates how to configure Nexus BFD for OSPFv2.

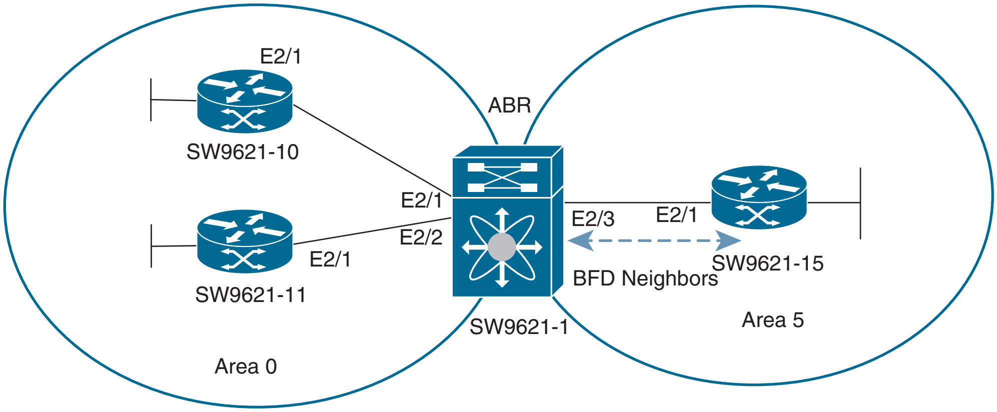


Figure 1-8 *BFD Network Topology*

Example 1-15 shows the SW9621-1 BFD feature enabling and its configuration.

Example 1-15 *SW9621-1 BFD Configuration*

```
SW9621-1(config)# feature bfd
SW9621-1(config)# interface Ethernet2/3
SW9621-1(config-if)# ip ospf bfd
SW9621-1(config-if)# no ip redirects
SW9621-1(config-if)# no ipv6 redirects
SW9621-1(config-if)# bfd interval 100 min_rx 100 multiplier 5
```

Example 1-16 shows SW9621-1 BFD neighbors.

Example 1-16 *SW9621-1 OSPFv2/BFD Verification*

```
SW9621-1# show bfd neighbors
OurAddr      NeighAddr    LD/RD                 RH/RS  Holdown(mult)  State   Int      Vrf
10.10.10.9   10.10.10.10  1090519042/1090519042  Up     9193(5)        Up      Eth2/3  default

SW9621-1# show ip ospf interface
 Ethernet2/3 is up, line protocol is up
    IP address 10.10.10.9/30
    Process ID 21 VRF default, area 0.0.0.5
    Enabled by interface configuration
    State DR, Network type BROADCAST, cost 40
    BFD is enabled
    Index 2, Transmit delay 1 sec, Router Priority 1
    Designated Router ID: 1.1.1.1, address: 10.10.10.9
    Backup Designated Router ID: 1.1.1.15, address: 10.10.10.10
    1 Neighbors, flooding to 1, adjacent with 1
    Timer intervals: Hello 25, Dead 100, Wait 100, Retransmit 5
      Hello timer due in 00:00:04
    No authentication
    Number of opaque link LSAs: 0, checksum sum 0
```

Example 1-17 shows SW9621-15 BFD configurations and neighbors.

Example 1-17 *Router 4 Config and Verification (SW9621-15)*

```
SW9621-15(config)# feature bfd
SW9621-15(config)# interface Ethernet2/1
SW9621-15(config-if)# ip ospf bfd
SW9621-15(config-if)# no ip redirects
SW9621-15(config-if)# no ipv6 redirects
SW9621-15(config-if)# bfd interval 100 min _ rx 100 multiplier 5

SW9621-15# show bfd neighbors
OurAddr      NeighAddr    LD/RD                 RH/RS   Holdown(mult)  State  Int     Vrf
10.10.10.10  10.10.10.9   109051902/109051902   Up       9190(5)         Up    Eth2/1  default
```

Multicast

Multicast IP routing protocols used to distribute data to multiple recipients in a single session (for example, audio/video streaming broadcasts). Multicast IP can send IP data to a group of interested receivers in a single transmission. Multicast IP addresses are called *groups*. A multicast address that includes a group and source IP address is often referred to as a *channel*. You can use multicast in both IPv4 and IPv6 networks to provide efficient delivery of data to multiple destinations.

The Internet Assigned Numbers Authority (IANA) has assigned 224.0.0.0 through 239.255.255.255 as IPv4 multicast addresses, and IPv6 multicast addresses begin with 0xFF.

Figure 1-9 shows one source transmitting multicast data that is delivered to two receivers. In the figure, because the center host is on a LAN segment where no receiver requested multicast data, no data is delivered to that receiver.

Key Topic

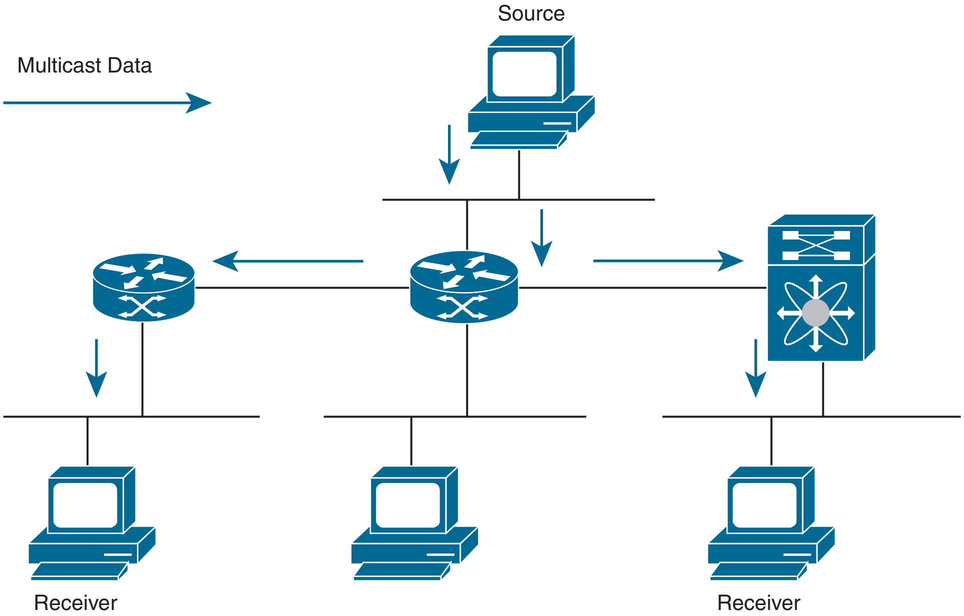


Figure 1-9 *Multicast Traffic from One Source to Two Receivers*

Key Topic

Internet Group Management Protocol

The Internet Group Management Protocol (IGMP) is used by hosts that want to receive multicast data to request membership in multicast groups. Once the group membership is established, multicast data for the group is directed to the LAN segment of the requesting host.

IGMP is an IPv4 protocol that a host uses to request multicast data for a particular group. Using the information obtained through the IGMP, the software maintains a list of multicast group or channel memberships on a per-interface basis. The systems that receive these IGMP packets send multicast data that they receive for requested groups or channels out the network segment of the known receivers.

NX-OS supports IGMPv2 and IGMPv3. By default, NX-OS enables IGMPv2 when it starts the IGMP process. You can enable IGMPv3 on interfaces where you want its capabilities.

IGMPv3 includes the following key changes from IGMPv2:

- IGMPv3 supports source-specific multicast (SSM), which builds shortest path trees from each receiver to the source, through the following features:
    - Host messages that can specify both the group and the source.
    - The multicast state that is maintained for groups and sources, not just for groups as in IGMPv2.
- Hosts no longer perform report suppression, which means that hosts always send IGMP membership reports when an IGMP query message is received.

For detailed information about IGMPv2, see RFC 2236.

The basic IGMP process of a router that discovers multicast hosts is shown in Figure 1-10. Hosts 1, 2, and 3 send unsolicited IGMP membership report messages to initiate receiving multicast data for a group or channel.

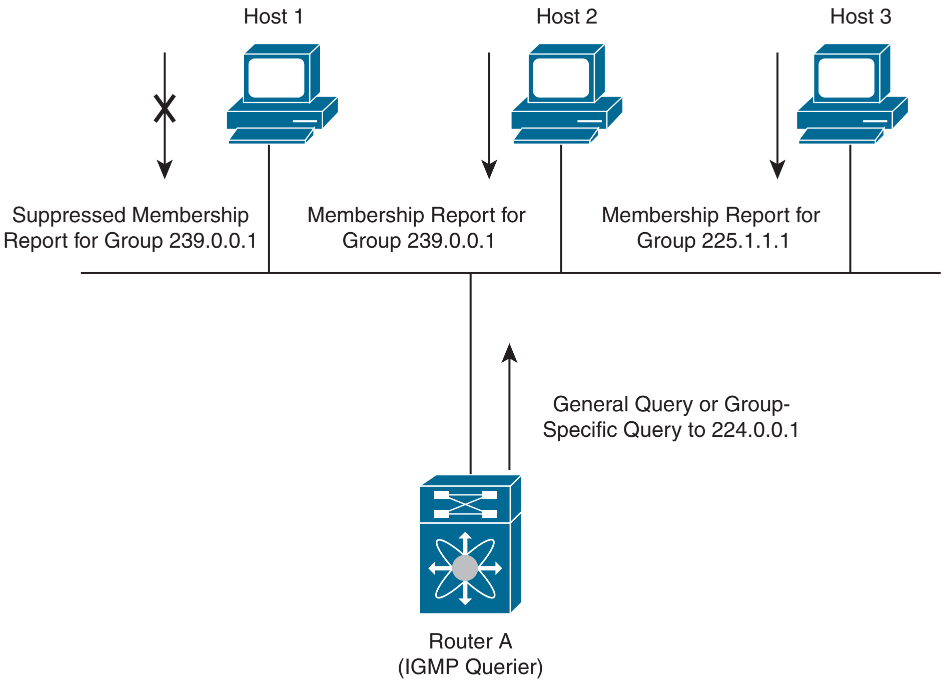

Host 1
Host 2
Host 3
Suppressed Membership Report for Group 239.0.0.1
Membership Report for Group 239.0.0.1
Membership Report for Group 225.1.1.1
General Query or Group-Specific Query to 224.0.0.1
Router A (IGMP Querier)

Figure 1-10 *IGMPv1 and IGMPv2 Query-Response Process*

In Figure 1-10, router A, which is the IGMP designated querier on the subnet, sends query messages to the all-hosts multicast group at 224.0.0.1 periodically to discover whether any hosts want to receive multicast data. You can configure the group membership timeout value that the router uses to determine that no members of a group or source exist on the subnet.

The software elects a router as the IGMP querier on a subnet if it has the lowest IP address. As long as a router continues to receive query messages from a router with a lower IP address, it resets a timer that is based on its querier timeout value. If the querier timer of a router expires, it becomes the designated querier. If that router later receives a host query message from a router with a lower IP address, it drops its role as the designated querier and sets its querier timer again.

In Figure 1-10, host 1's membership report is suppressed, and host 2 sends its membership report for group 239.0.0.1 first. Host 1 receives the report from host 2. Because only one membership report per group needs to be sent to the router, other hosts suppress their reports to reduce network traffic. Each host waits for a random time interval to avoid sending reports at the same time. You can configure the query maximum response time parameter to control the interval in which hosts randomize their responses.

NOTE IGMPv2 membership report suppression occurs only on hosts that are connected to the same port.

In Figure 1-11, router A sends the IGMPv3 group-and-source-specific query to the LAN. Hosts 2 and 3 respond to the query with membership reports that indicate that they want to receive data from the advertised group and source. This IGMPv3 feature supports SSM.

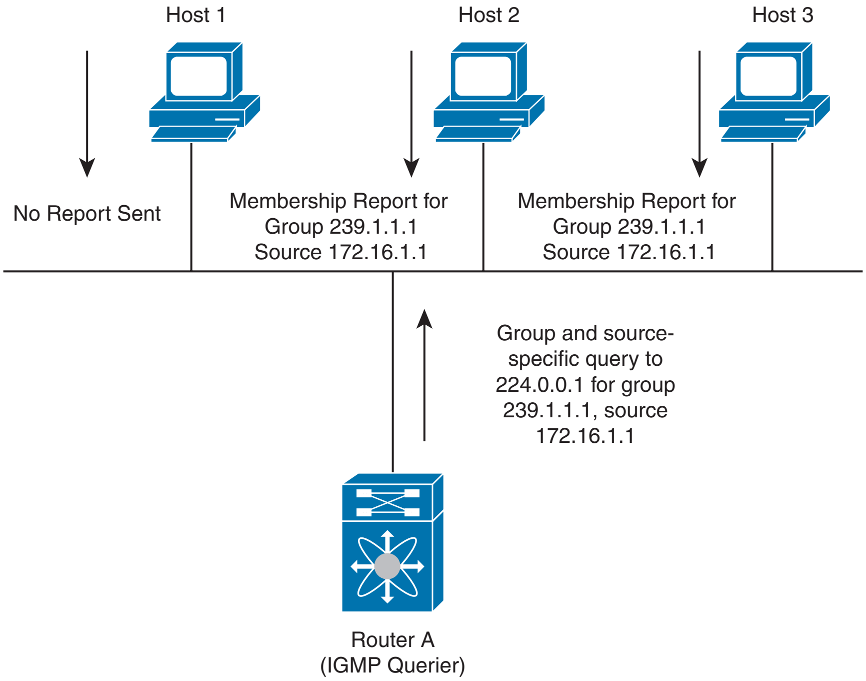


Figure 1-11 *IGMPv3 Group-and-Source-Specific Query*

NOTE IGMPv3 hosts do not perform IGMP membership report suppression.

IGMP messages sent by the designated querier have a time-to-live (TTL) value of 1, which means that the messages are not forwarded by the directly connected routers on the subnet. You can configure the frequency and number of query messages sent specifically for IGMP start-up, and you can configure a short query interval at start-up so that the group state is established as quickly as possible. Although usually unnecessary, you can tune the query interval used after start-up to a value that balances the responsiveness to host group membership messages and the traffic created on the network.

NOTE Changing the query interval can severely impact multicast forwarding.

When a multicast host leaves a group, a host that runs IGMPv2 or later sends an IGMP *leave message*. To check if this host is the last host to leave the group, the software sends an IGMP query message and starts a timer that you can configure, called the *last member query response interval*. If no reports are received before the timer expires, the software removes the group state. The router continues to send multicast traffic for a group until its state is removed.

You can configure a robustness value to compensate for packet loss on a congested network. The robustness value is used by the IGMP software to determine the number of times to send messages.

Link-local addresses in the range 224.0.0.0/24 are reserved by the Internet Assigned Numbers Authority (IANA). Network protocols on a local network segment use these addresses; routers do not forward these addresses because they have a TTL of 1. By default, the IGMP process sends membership reports only for nonlink-local addresses, but you can configure the software to send reports for link-local addresses.

Switch IGMP Snooping

IGMP snooping is a feature that limits multicast traffic on VLANs to the subset of ports that have known receivers. The IGMP snooping software examines IGMP protocol messages within a VLAN to discover which interfaces are connected to hosts or other devices interested in receiving this traffic. Using the interface information, IGMP snooping can reduce bandwidth consumption in a multi-access LAN environment to avoid flooding the entire VLAN. The IGMP snooping feature tracks which ports are attached to multicast-capable routers to help it manage the forwarding of IGMP membership reports. Multicast traffic is sent only to VLAN ports on which interested hosts reside. The IGMP snooping software responds to topology change notifications.

By default, IGMP snooping is enabled on the Cisco NX-OS system.

Multicast Listener Discovery

Multicast Listener Discovery (MLD) is an IPv6 protocol that a host uses to request multicast data for a particular group. Using the information obtained through MLD, the software maintains a list of multicast group or channel memberships on a per-interface basis.

The devices that receive MLD packets send the multicast data that they receive for requested groups or channels out the network segment of the known receivers.

MLDv1 is derived from IGMPv2, and MLDv2 is derived from IGMPv3. IGMP uses IP Protocol 2 message types, whereas MLD uses IP Protocol 58 message types, which is a subset of the ICMPv6 messages.

The MLD process is started automatically on the device. You cannot enable MLD manually on an interface. MLD is enabled automatically when you perform one of the following configuration tasks on an interface:

- Enable PIM6.
- Statically bind a local multicast group.
- Enable link-local group reports.

Cisco NX-OS supports MLDv1 and MLDv2. MLDv2 supports MLDv1 listener reports.

By default, the software enables MLDv2 when it starts the MLD process. You can enable MLDv1 on interfaces where you want only its capabilities.

MLDv2 includes the following key changes from MLDv1:

- MLDv2 supports source-specific multicast (SSM), which builds shortest path trees from each receiver to the source, through the following features:
  - Host messages that can specify both the group and the source.
  - The multicast state that is maintained for groups and sources, not just for groups as in MLDv1.
- Hosts no longer perform report suppression, which means that hosts always send MLD listener reports when an MLD query message is received.

For detailed information about MLDv1, see RFC 2710. For detailed information about MLDv2, see RFC 3810.

The MLD process is similar to IGMP: MLD utilizes link-local addresses in the range FF02::0/16, as defined by the IANA. Network protocols on a local network segment use these addresses; routers do not forward these addresses because they have a TTL of 1. By default, the MLD process sends listener reports only for nonlink-local addresses, but you can configure the software to send reports for link-local addresses.

Multicast Distribution Trees

A multicast distribution tree (MDT) represents the path that multicast data takes between the routers that connect sources and receivers. The multicast software builds different types of trees to support different multicast methods.

A *source* tree represents the shortest path that the multicast traffic takes through the network from the sources that transmit to a particular multicast group to receivers that requested traffic from that same group. Because of the shortest path characteristic of a source tree, this tree is often referred to as a *shortest path tree* (SPT). Figure 1-12 shows a source tree for group 239.1.1.1 that begins at host A and connects to hosts B and C.

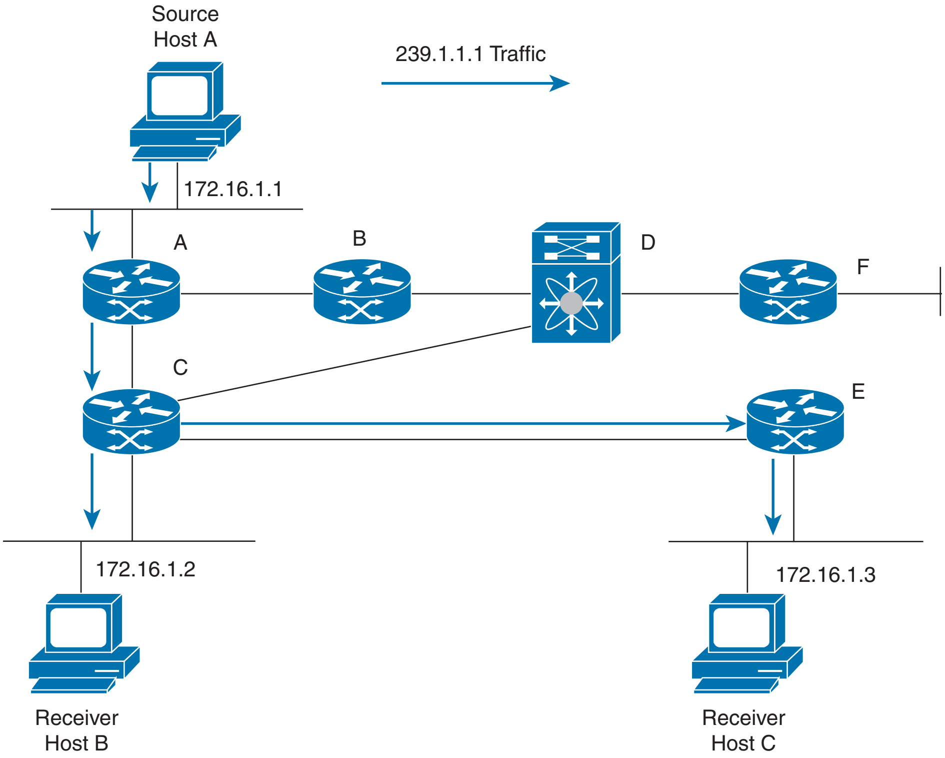


Figure 1-12 *Source Tree*

The notation (S, G) represents the multicast traffic from source S on group G. The SPT in Figure 1-12 is written (172.16.1.1, 239.1.1.1) and called "S comma G." Multiple sources can be transmitting on the same group.

A *shared tree* represents the shared distribution path that the multicast traffic takes through the network from a shared root or rendezvous point (RP) to each receiver. (The RP creates an SPT to each source.) A shared tree is also called an *RP tree* (RPT). Figure 1-13 shows a shared tree for group 239.1.1.1 with the RP at router D. Source hosts A and D send their data to router D, the RP, which then forwards the traffic to receiver hosts B and C.

The notation (*, G) ("star comma G") represents the multicast traffic from any source on group G. The shared tree in Figure 1-13 is written (*, 239.1.1.1).

A *bidirectional shared tree* represents the shared distribution path that the multicast traffic takes through the network from a shared root, or rendezvous point, to each receiver. Multicast data is forwarded to receivers encountered on the way to the RP. The advantage of the bidirectional shared tree is shown in Figure 1-14. Multicast traffic flows directly from host A to host B through routers B and C. In a shared tree, the data from source host A is first sent to the RP (router D) and then forwarded to router B for delivery to host B.

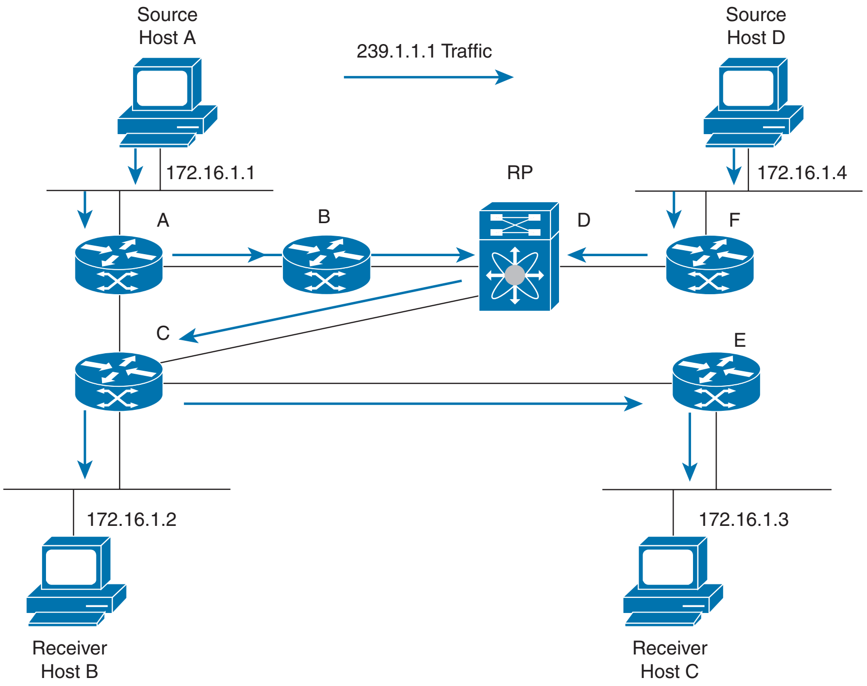


Figure 1-13 *Shared Tree*

The notation (*, G) represents the multicast traffic from any source on group G. The bidirectional tree in Figure 1-14 is written (*, 239.0.0.1).

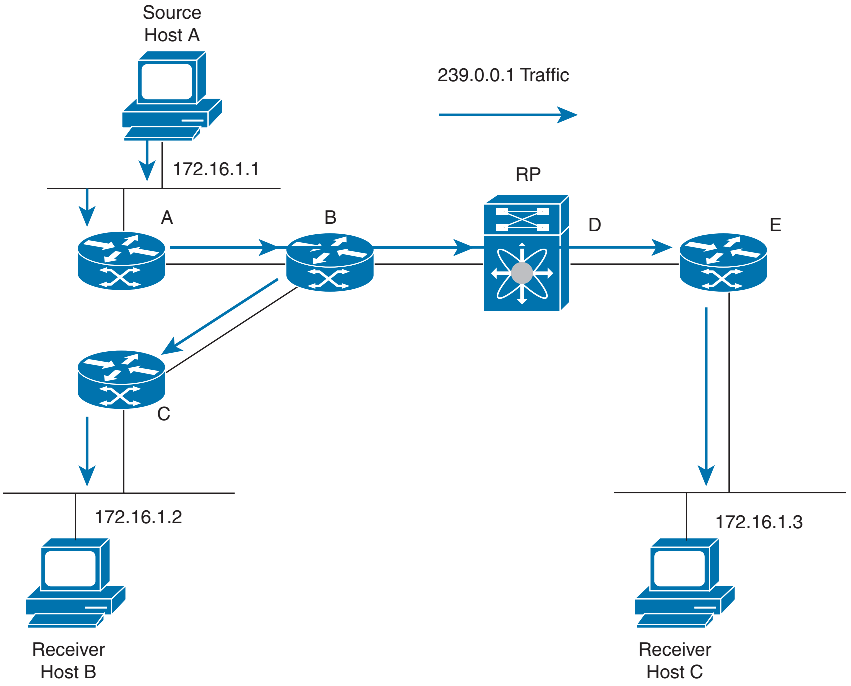


Figure 1-14 *Bidirectional Shared Tree*

Key Topic

Protocol Independent Multicast

Cisco NX-OS supports multicasting with Protocol Independent Multicast (PIM) sparse mode. PIM (PIMv2) is an independent IPv4 routing protocol, and PIM6 is an independent IPv6 routing protocol. In PIM sparse mode, multicast traffic is sent only to locations of the network that specifically request it. PIM dense mode is not supported by Cisco NX-OS.

You need to enable the PIM or PIM6 feature before configuring multicast. Multicast is enabled only after you enable PIM or PIM6 on an interface of each router in a domain. You configure PIM for an IPv4 network and PIM6 for an IPv6 network. By default, IGMP and MLD are enabled on the NX-OS system.

PIM, which is used between multicast-capable routers, advertises group membership across a routing domain by constructing multicast distribution trees. PIM builds shared distribution trees on which packets from multiple sources are forwarded, as well as source distribution trees, on which packets from a single source are forwarded.

The distribution trees change automatically to reflect the topology changes due to link or router failures. PIM dynamically tracks both multicast-capable sources and receivers although the source state is not created in Bidir mode.

The router uses the unicast routing table and RPF routes for multicast to create multicast routing information. In Bidir mode, additional routing information is created.

NOTE In this book, we use *PIMv2 for IPv4* and *PIM6 for IPv6* to refer to the Cisco NX-OS implementation of PIM sparse mode. A PIM domain can include both an IPv4 and an IPv6 network. Figure 1-15 shows two PIM domains in an IPv4 network.

Key Topic

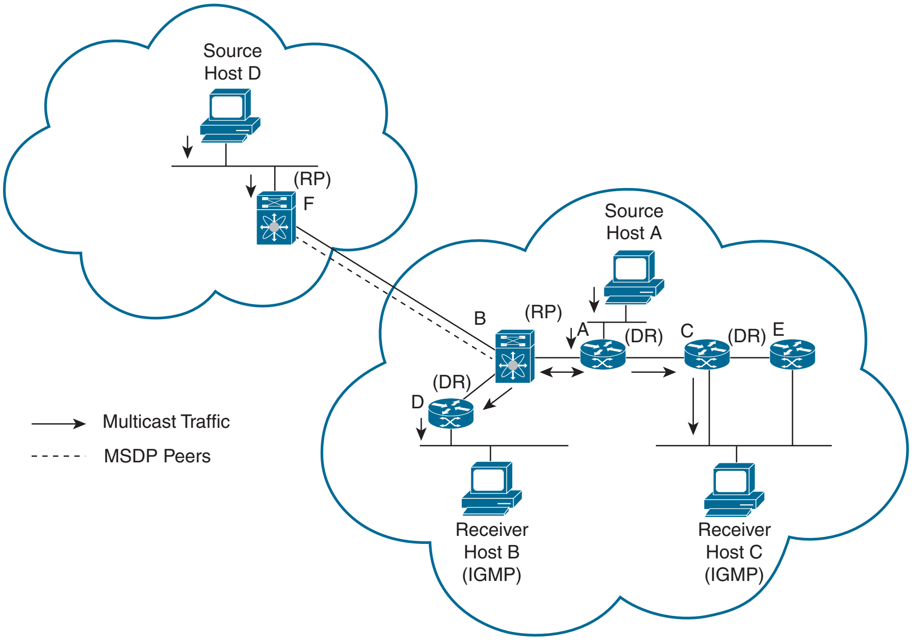


Figure 1-15 *PIM Domains in an IPv4 Network*

Figure 1-15 shows the following PIM elements:

- The lines with arrows show the path of the multicast data through the network. The multicast data originates from the sources at hosts A and D.
- The dashed line connects routers B and F, which are Multicast Source Discovery Protocol (MSDP) peers. MSDP supports the discovery of multicast sources in other PIM domains.
- Hosts B and C receive multicast data by using Internet Group Management Protocol (IGMP) to advertise requests to join a multicast group.
- Routers A, C, and D are designated routers (DRs). When more than one router is connected to a LAN segment, such as C and E, the PIM software chooses one router to be the DR so that only one router is responsible for putting multicast data on the segment.

Router B is the rendezvous point for one PIM domain, and router F is the RP for the other PIM domain. The RP provides a common point for connecting sources and receivers within a PIM domain.

Figure 1-16 shows two PIM6 domains in an IPv6 network. In an IPv6 network, receivers that want to receive multicast data use the Multicast Listener Discovery (MLD) protocol to advertise requests to join a multicast group. MSDP, which allows for discovery of multicast sources in other PIM domains, is not supported for IPv6. You can configure IPv6 peers and use source-specific multicast (SSM) and Multiprotocol BGP (MBGP) to forward multicast data between PIM6 domains.

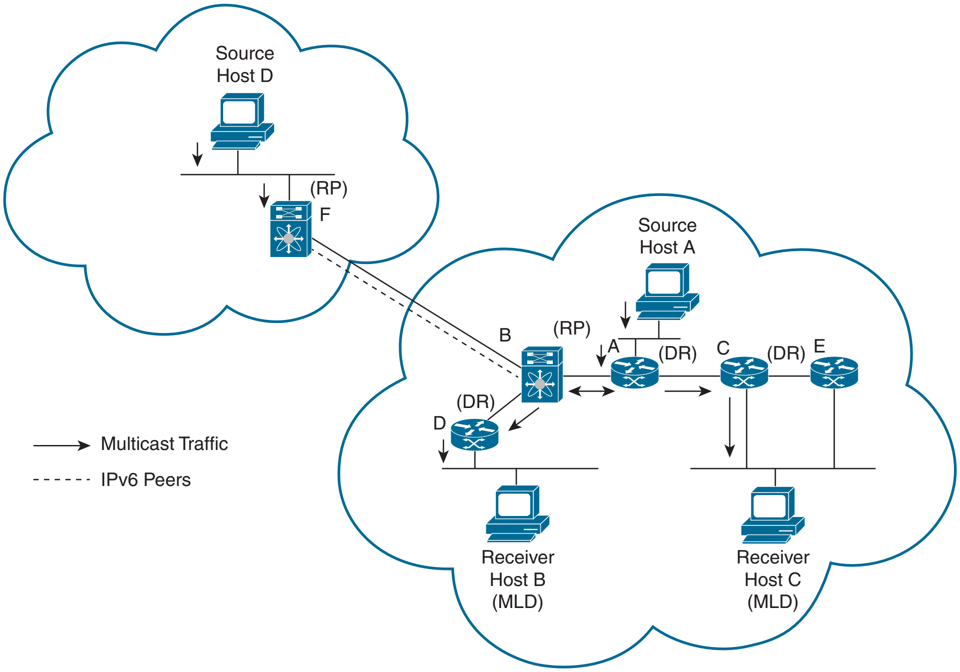


Figure 1-16 *PIM6 Domains in an IPv6 Network*

Cisco NX-OS supports a combination of three modes for different ranges of multicast groups. You can also define RPF routes for multicast.

Any-source multicast (ASM) is a PIM tree building mode that uses shared trees to discover new sources and receivers as well as source trees to form shortest paths from receivers to sources. The shared tree uses a network node as the root, called the rendezvous point. The source tree is rooted at first hop routers, directly attached to each source that is an active sender. The ASM mode requires an RP for a group range. An RP can be configured statically or learned dynamically by the Auto-RP or BSR group-to-RP discovery protocols. If an RP is learned and is not known to be a Bidir RP, the group operates in ASM mode.

The ASM mode is the default mode when you configure RPs.

Bidirectional shared trees (Bidir) is a PIM mode that, like the ASM mode, builds a shared tree between receivers and the RP but does not support switching over to a source tree when a new receiver is added to a group. In the Bidir mode, the router that is connected to a receiver is called the designated forwarder (DF) because multicast data can be forwarded directly from the designated router to the receiver without first going to the RP. The Bidir mode requires that you configure an RP.

The Bidir mode can reduce the amount of resources required on a router when there are many multicast sources and can continue to operate whether or not the RP is operational or connected.

Source-specific multicast (SSM) is a PIM mode that builds a source tree that originates at the designated router on the LAN segment that receives a request to join a multicast source. Source trees are built by sending PIM join messages in the direction of the source. The SSM mode does not require you to configure RPs.

The SSM mode allows receivers to connect to sources outside the PIM domain. PIM messages include the following:

- **Hello:** The PIM process begins when the router establishes PIM neighbor adjacencies by sending PIM hello messages to the multicast address 224.0.0.13. Hello messages are sent periodically at an interval of 30 seconds. After all neighbors have replied, the PIM software chooses the router with the highest priority in each LAN segment as the designated router. The DR priority is based on a DR priority value in the PIM hello message. If the DR priority value is not supplied by all routers, or the priorities match, the highest IP address is used to elect the DR.

  The hello message also contains a hold-time value, which is typically 3.5 times the hello interval. If this hold time expires without a subsequent hello message from its neighbor, the device detects a PIM failure on that link.

  For security, you can configure an MD5 hash value that the PIM software uses to authenticate PIM hello messages with PIM neighbors.

- **Join-Prune:** When the DR receives an IGMP membership report message from a receiver for a new group or source, the DR creates a tree to connect the receiver to the source by sending a PIM join message out the interface toward the rendezvous point (ASM or Bidir mode) or source (SSM mode). The rendezvous point is the root of a shared tree, which is used by all sources and hosts in the PIM domain in the ASM or the Bidir mode. SSM does not use an RP but builds a shortest path tree (SPT) that is the lowest cost path between the source and the receiver.

  When the DR determines that the last host has left a group or source, it sends a PIM prune message to remove the path from the distribution tree.

  The routers forward the join or prune action hop by hop up the multicast distribution tree to create (join) or tear down (prune) the path.

NOTE In this book, we use the terms *PIM join message* and *PIM prune message* to simplify the action taken when referring to the PIM join-prune message with only a join or prune action.

  The software sends join-prune messages as quickly as possible. You can filter the join-prune messages by defining a routing policy. For information about configuring the join-prune message policy please refer to Table 1-24.

- **PIM register:** PIM register messages are unicast to the RP by designated routers that are directly connected to multicast sources. The PIM register message has the following functions:

  - To notify the RP that a source is actively sending to a multicast group.

  - To deliver multicast packets sent by the source to the RP for delivery down the shared tree.

  The DR continues to send PIM register messages to the RP until it receives a register-stop message from the RP. The RP sends a register-stop message in either of the following cases:

  - The RP has no receivers for the multicast group being transmitted.

  - The RP has joined the SPT to the source but has not started receiving traffic from the source.

  You can use the **ip pim register-source** command to configure the IP source address of register messages when the IP source address of a register message is not a uniquely routed address to which the RP can send packets. This situation might occur if the source address is filtered so that the packets sent to it are not forwarded or if the source address is not unique to the network. In these cases, the replies sent from the RP to the source address will fail to reach the DR, resulting in Protocol Independent Multicast sparse mode (PIM-SM) protocol failures.

PIM requires that multicast entries are refreshed within a 3.5-minute timeout interval. The state refresh ensures that traffic is delivered only to active listeners, and it keeps routers from using unnecessary resources.

To maintain the PIM state, the last-hop DR sends join-prune messages once per minute. State creation applies to both (*, G) and (S, G) states as follows:

- **(*, G) state creation example:** An IGMP (*, G) report triggers the DR to send a (*, G) PIM join message toward the RP.
- **(S, G) state creation example:** An IGMP (S, G) report triggers the DR to send an (S, G) PIM join message toward the source.

If the state is not refreshed, the PIM software tears down the distribution tree by removing the forwarding paths in the multicast outgoing interface list of the upstream routers.

Key Topic

PIM Rendezvous Points

A rendezvous point (RP) is a router that you select in a multicast network domain that acts as a shared root for a multicast shared tree. You can configure as many RPs as you like, and you can configure them to cover different group ranges:

- **Static RP:** You can statically configure an RP for a multicast group range. You must configure the address of the RP on every router in the domain. You can define static RPs for the following reasons:
  - To configure routers with the Anycast RP address.
  - To manually configure an RP on a device.
- **BSRs:** The bootstrap router ensures that all routers in the PIM domain have the same RP cache as the BSR. You can configure the BSR to help you select an RP set from BSR candidate RPs. The function of the BSR is to broadcast the RP set to all routers in the domain. You select one or more candidate BSRs to manage the RPs in the domain. Only one candidate BSR is elected as the BSR for the domain.

CAUTION You should not configure both Auto-RP and BSR protocols in the same network.

- **Auto-RP:** Auto-RP is a Cisco protocol that was used prior to the Internet standard bootstrap router mechanism. You configure Auto-RP by selecting candidate mapping agents and RPs. Candidate RPs send their supported group range in RP-Announce messages to the Cisco RP-Announce multicast group 224.0.1.39. An Auto-RP mapping agent listens for RP-Announce messages from candidate RPs and forms a Group-to-RP mapping table. The mapping agent multicasts the Group-to-RP mapping table in RP-Discovery messages to the Cisco RP-Discovery multicast group 224.0.1.40.
- **Anycast RP:** An Anycast RP is used to define redundant and load-balanced RPs. An Anycast RP has two implementations:
  - Using Multicast Source Discovery Protocol (MSDP)
  - Using Protocol Independent Multicast (PIM)

  An Anycast RP allows two or more RPs to share the load for source registration and to act as hot backup routers for each other. MSDP is the protocol that RPs use to share information about active sources. With an Anycast RP, the RPs are configured to establish MSDP peering sessions using a TCP connection. Group participants use the closest RP that is favored by the IP unicast route table.

You can use the PIM Anycast RP to assign a group of routers to a single RP address that is configured on multiple routers. The set of routers that you configure as Anycast RPs is called the Anycast RP set. This RP method is the only one that supports more than one RP per multicast group, which allows you to load balance across all RPs in the set. The Anycast RP supports all multicast groups.

PIM register messages are sent to the closest RP, and PIM join-prune messages are sent in the direction of the closest RP as determined by the unicast routing protocols. If one of the RPs goes down, unicast routing ensures that these message will be sent in the direction of the next-closest RP.

You must configure PIM on the loopback interface that is used for the PIM Anycast RP.

Key Topic

PIM Designated Routers/Forwarders

In PIM ASM and SSM modes, the software chooses a designated router from the routers on each network segment. The DR is responsible for forwarding multicast data for specified groups and sources on that segment. The DR for each LAN segment is determined as described in the hello messages.

In ASM mode, the DR is responsible for unicasting PIM register packets to the RP. When a DR receives an IGMP membership report from a directly connected receiver, the shortest path is formed to the RP, which may or may not go through the DR. The result is a shared tree that connects all sources transmitting on the same multicast group to all receivers of that group.

In SSM mode, the DR triggers (S, G) PIM join or prune messages toward the source. The path from the receiver to the source is determined hop by hop. The source must be known to the receiver or the DR.

In PIM Bidir mode, the software chooses a designated forwarder (DF) at RP discovery time from the routers on each network segment. The DF is responsible for forwarding multicast data for specified groups on that segment. The DF is elected based on the best metric from the network segment to the RP.

If the router receives a packet on the RPF interface toward the RP, the router forwards the packet out all interfaces in the OIF list. If a router receives a packet on an interface on which the router is the elected DF for that LAN segment, the packet is forwarded out all interfaces in the OIF list except the interface that it was received on and also out the RPF interface toward the RP.

NOTE Cisco NX-OS puts the RPF interface into the OIF list of the MRIB but not in the OIF list of the MFIB.

Key Topic

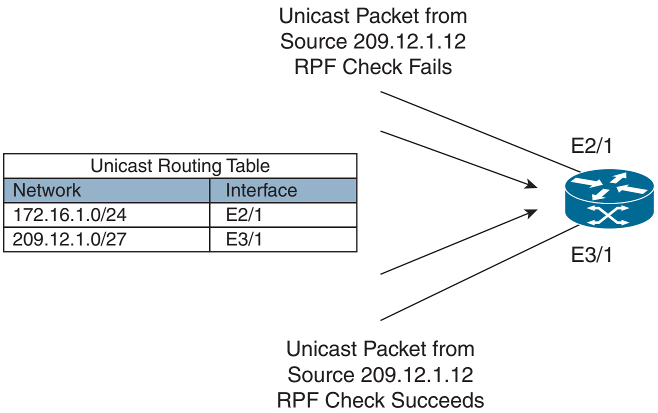
Multicast Forwarding

Because multicast traffic is destined for an arbitrary group of hosts, the router uses *Reverse Path Forwarding* (RPF) to route data to active receivers for the group. When receivers join a group, a path is formed either toward the source (SSM mode) or the RP (ASM or Bidir mode). The path from a source to a receiver flows in the reverse direction from the path that was created when the receiver joined the group.

For each incoming multicast packet, the router performs an RPF check. If the packet arrives on the interface leading to the source, the packet is forwarded out each interface in the *outgoing interface* (OIF) list for the group. Otherwise, the router drops the packet.

NOTE In Bidir mode, if a packet arrives on a non-RPF interface, and the interface was elected as the designated forwarder, the packet is also forwarded in the upstream direction toward the RP.

Figure 1-17 shows an example of RPF checks on packets coming in from different interfaces. The packet that arrives on E2/1 fails the RPF check because the unicast route table lists the source of the network on interface E3/1. The packet that arrives on E3/1 passes the RPF check because the unicast route table lists the source of that network on interface E3/1.



Unicast Routing Table

Network	Interface
172.16.1.0/24	E2/1
209.12.1.0/27	E3/1

Figure 1-17 *RPF Check Example*

Multicast Configurations and Verifications

Table 1-19 lists IGMP/MLD default parameters; you can alter these parameters as necessary.

Key Topic

Table 1-19 Default IGMP/MLD Parameters

Parameters	Default
IGMP version	2
MLD version	2
Startup query interval	30 seconds
Startup query count	2
Robustness value	2
Querier timeout	255 seconds
Query timeout	255 seconds
Query max response time	10 seconds

Parameters	Default
Query interval	125 seconds
Last member query response interval	1 second
Last member query count	2
Group membership timeout	260 seconds
Report link-local multicast groups	Disabled
Enforce router alert	Disabled
Immediate leave	Disabled

Table 1-20 lists the default PIM/PIM6 parameters.

Table 1-20 Default PIM/PIM6 Parameters

Parameters	Default
Use shared trees only	Disabled
Flush routes on restart	Disabled
Log Neighbor changes	Disabled
Auto-RP message action	Disabled
BSR message action	Disabled
SSM multicast group range or policy	232.0.0.0/8 for IPv4 and FF3x::/96 for IPv6
PIM sparse mode	Disabled
Designated router priority	0
Hello authentication mode	Disabled
Domain border	Disabled
RP address policy	No message filtering
PIM register message policy	No message filtering
BSR candidate RP policy	No message filtering
BSR policy	No message filtering
Auto-RP mapping agent policy	No message filtering
Auto-RP RP candidate policy	No message filtering
Join-prune policy	No message filtering
Neighbor adjacency policy	Become adjacent with all PIM neighbors
BFD	Disabled

NOTE No license is required for IGMP/MLD.

Because PIMv2 required a license, Table 1-21 shows the required NX-OS feature licenses. For more information, visit the Cisco NX-OS Licensing Guide.

Table 1-21 Feature-Based Licenses for Cisco NX-OS

Platform	Feature License	Feature Name
Cisco Nexus 9000 Series Cisco Nexus 7000 Series	Enterprise Services Package LAN_ENTERPRISE_SERVICES_PKG	PIMv2 (all modes) MSDP
Cisco Nexus 6000 Series Cisco Nexus 5600 Series Cisco Nexus 5500 Series Cisco Nexus 5000 Series	Layer 3 Base Services Package LAN_BASE_SERVICES_PKG	PIMv2 (sparse mode)
Cisco Nexus 6000 Series Cisco Nexus 5600 Series Cisco Nexus 5500 Series Cisco Nexus 5000 Series	Layer 3 Enterprise Services Package LAN_ENTERPRISE_SERVICES_PKG	PIMv2 (all modes) MSDP
Cisco Nexus 3600 Series	Layer 3 Enterprise Services Package LAN_ENTERPRISE_SERVICES_PKG	PIMv2
Cisco Nexus 3000 Series	Layer 3 Enterprise Services Package LAN_ENTERPRISE_SERVICES_PKG	PIMv2 (spares mode)

PIM and PIM6 have the following configuration guidelines and limitations:

- Cisco NX-OS PIM and PIM6 do not interoperate with any version of PIM dense mode or PIM sparse mode version 1.
- Do not configure both Auto-RP and BSR protocols in the same network.
- Configure candidate RP intervals to a minimum of 15 seconds.
- If a device is configured with a BSR policy that should prevent it from being elected as the BSR, the device ignores the policy. This behavior results in the following undesirable conditions:
    - If a device receives a BSM that is permitted by the policy, the device, which incorrectly elected itself as the BSR, drops that BSM so that routers downstream fail to receive it. Downstream devices correctly filter the BSM from the incorrect BSR so that these devices do not receive RP information.
    - A BSM received by a BSR from a different device sends a new BSM but ensures that downstream devices do not receive the correct BSM.

You can configure separate ranges of addresses in the PIM or PIM6 domain using the multicast distribution modes described in Table 1-22.

Table 1-22 PIM and PIM6 Multicast Distribution Modes

Multicast Distribution Mode	Requires RP Configuration	Purpose
ASM	Yes	Any source multicast
Bidir	Yes	Bidirectional shared trees
SSM	No	Single-source multicast
RPF routes for multicast	No	RPF routes for multicast

Tables 1-23 through 1-25 describe the most-used multicast configuration commands. For the full list of commands, refer to the Nexus Multicast Routing Configuration Guide in the reference section at the end of this chapter.

Table 1-23 Multicast Global-Level Commands

Command	Purpose
feature pim	Enables PIM. By default, PIM is disabled.
feature pim6	Enables PIM6. By default, PIM6 is disabled.
[**ip** | **ipv6**] [**pim**|**pim6**] **rp-address** *rp-address* [**group-list** *ip-prefix* | **prefix-list** *name* | **route-map** *policy-name*] [**bidir**]	Configures a PIM static RP address for a multicast group range. You can specify a route-map policy name that lists the group prefixes to use with the **match ip multicast** command. The default mode is ASM unless you specify the **bidir** keyword. The default group range for IPv4 is 224.0.0.0 through 239.255.255.255; for IPv6, it is ff00::0/8. Example 1 configures PIM ASM mode for the specified group range. Example 2 configures PIM Bidir mode for the specified group range.
ip pim auto-rp {**listen** [**forward**] | **forward** [**listen**]}	(Optional) Enables listening or forwarding of Auto-RP messages. The default is disabled, which means that the software does not listen to or forward Auto-RP messages.
ip pim bsr {**listen** [**forward**] | **forward** [**listen**]}	(Optional) Enables listening or forwarding of BSR messages. The default is disabled, which means that the software does not listen or forward BSR messages.
[**ip** | **ipv6**] [**pim**|**pim6**] **rp-address** *anycast-rp-address* [**group-list** *ip-address*]	Configures a PIM Anycast RP peer address for the specified Anycast RP address. Each command with the same Anycast RP address forms an Anycast RP set. The IP addresses of RPs are used for communication with RPs in the set.
ip pim bidir-rp-limit *limit*	(Optional) Specifies the number of Bidir RPs that you can configure for IPv4. The maximum number of Bidir RPs supported per VRF for PIM and PIM6 combined cannot exceed 8. Values range from 0 to 8. The default is 6.
[**ip** | **ipv6**] [**pim**|**pim6**] **register-rate-limit** *rate*	(Optional) Configures the rate limit in packets per second. The range is from 1 to 65,535. The default is no limit.
ip pim spt-threshold infinity group-list *route-map-name*	(Optional) Configures the initial hold-down period in seconds. The range is from 90 to 210. Specify 0 to disable the hold-down period. The default is 210.

Command	Purpose
[**ip** | **ipv6**] **routing multicast holddown** *holddown-period*	(Optional) Configures the initial hold-down period in seconds. The range is from 90 to 210. Specify 0 to disable the hold-down period. The default is 210.
ip igmp ssm-translate	Configures the translation of IGMPv1 or IGMPv2 membership reports by the IGMP process to create the (S,G) state as if the router had received an IGMPv3 membership report.

Table 1-24 Multicast Interface-Level Commands

Command	Purpose
[**ip** | **ipv6**] [**pim**|**pim6**] **sparse-mode**	Enables sparse mode on this interface. The default is disabled.
[**ip** | **ipv6**] [**pim**|**pim6**] **dr-priority** *priority*	(Optional) Sets the designated router priority that is advertised in PIM/PIM6 hello messages. Values range from 1 to 4,294,967,295. The default is 1.
[**ip** | **ipv6**] [**pim**|**pim6**] **hello-authentication ah-md5** *auth-key*	(Optional) Enables an MD5 hash authentication key in PIM/PIM6 hello messages. You can enter an unencrypted (cleartext) key or one of these values followed by a space and the MD5 authentication key: **0:** Specifies an unencrypted (cleartext) key. **3:** Specifies a 3-DES encrypted key. **7:** Specifies a Cisco Type 7 encrypted key. The key can be up to 16 characters. The default is disabled.
[**ip** | **ipv6**] [**pim**|**pim6**] **hello-interval** *interval*	(Optional) Configures the interval at which hello messages are sent in milliseconds. The range is from 1000 to 18,724,286. The default is 30,000.
[**ip** | **ipv6**] [**pim**|**pim6**] **neighbor-policy prefix-list** *prefix-list*	(Optional) Configures which PIM/PIM6 neighbors to become adjacent to based on a route-map policy with the **match ipv6 address** command. The policy name can be up to 63 characters. The default is to become adjacent with all PIM6 neighbors. **NOTE:** We recommend that you configure this feature only if you are an experienced network administrator.
[**ip igmp** | **ipv6 mld**] **version** *value*	Sets the IGMP/MLD version to the value specified. The default is 2.
[**ip igmp** | **ipv6 mld**] **join-group** { *group* [**source** *source*] | **route-map** *policy-name* }	Statically binds a multicast group to the interface. If you specify only the group address, the (*, G) state is created. If you specify the source address, the (S, G) state is created. You can specify a route-map policy name that lists the group prefixes, group ranges, and source prefixes to use with the **match ip multicast** command.

Command	Purpose
	NOTE: A source tree is built for the (S, G) state only if you enable IGMPv3. **CAUTION:** The device CPU must be able to handle the traffic generated by using this command. Because of CPU load constraints, using this command, especially in any form of scale, is not recommended. Consider using the **ip igmp static-oif** command instead.
[**ip igmp** | **ipv6 mld**] **startup-query-interval** *seconds*	Sets the query interval used when the software starts up. Values can range from 1 to 18,000 seconds. The default is 31 seconds.
[**ip igmp** | **ipv6 mld**] **query-timeout** *seconds*	Sets the querier timeout that the software uses when deciding to take over as the querier. Values can range from 1 to 65,535 seconds. The default is 255 seconds.
[**ip igmp** | **ipv6 mld**] **query-interval** *interval*	Sets the frequency at which the software sends IGMP host query messages. Values can range from 1 to 18,000 seconds. The default is 125 seconds.
[**ip igmp** | **ipv6 mld**] **access-group** *policy*	Configures a route-map policy to control the multicast groups that hosts on the subnet serviced by an interface can join.
[**ip igmp** | **ipv6 mld**] **immediate-leave**	Enables the device to remove the group entry from the multicast routing table immediately upon receiving a leave message for the group. Use this command to minimize the leave latency of IGMPv2/MLD group memberships on a given IGMP/MLD interface because the device does not send group-specific queries. The default is disabled. **NOTE:** Use this command only when there is one receiver behind the interface for a given group.
[ip|ipv6] **pim ip-policy** *policy-name* [**in** | **out**]	(Optional) Enables join-prune messages to be filtered based on a route-map policy where you can specify group, group and source, or group and RP addresses with the **match ip multicast** command. The default is no filtering of join-prune messages. This command filters messages in both incoming and outgoing directions.

Table 1-25 Multicast Global-Level BGP Verification Commands

Command	Purpose
show [ip|ipv6] **pim rp**	Displays rendezvous points known to the software, how they were learned, and their group ranges. For similar information, see also the **show ip pim group-range** command.
show [ip|ipv6] **mroute ip-address** {*source group* | *group*[*source*]} [**vrf** *vrf-name* | **all**]	Displays the IP or IPv6 multicast routing table.
show [ip|ipv6] **pim group-range** [*ip-prefix* | **vrf** *vrf-name*]	Displays the learned or configured group ranges and modes. For similar information, see also the **show ip pim rp** command.

Command	Purpose
show running-configuration pim [**6**]	Displays the running-configuration information.
show [**ip igmp** | **ipv6 mld**] **interface** [**interface**] [**vrf** *vrf-name* | **all**] [**brief**]	Displays IGMP information about all interfaces or a selected interface, the default VRF, a selected VRF, or all VRFs.
show [**ip igmp** | **ipv6 mld**] **groups** [**group** | **interface**] [**vrf** *vrf-name* | **all**]	Displays the IGMP attached group membership for a group or interface, the default VRF, a selected VRF, or all VRFs.
show [**ip igmp** | **ipv6 mld**] **route** [**group** | **interface**] [**vrf** *vrf-name* | **all**]	Displays the IGMP attached group membership for a group or interface, the default VRF, a selected VRF, or all VRFs.
show [**ip igmp** | **ipv6 mld**] **local - groups**	Displays the IGMP local group membership.
show running-configuration [**igmp**|**mld**]	Displays the IGMP/MLD running-configuration information.

Figure 1-18 shows the network topology for the configuration that follows, which demonstrates how to configure Nexus multicast routing.

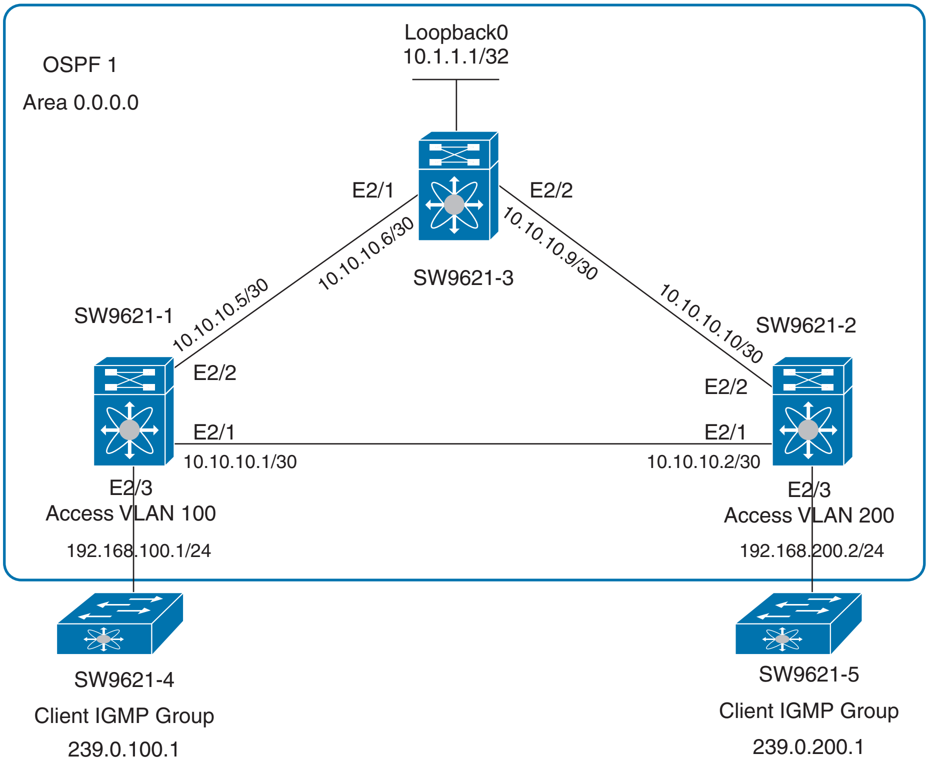


Figure 1-18 *Multicast Network Topology*

Example 1-18 shows the SW9621-1 Multicast PIM feature enabling and IGMP configurations.

Example 1-18 *Multicast IGMP SW9621-1 Configuration and Verifications*

```
SW9621-1(config)# feature interface-vlan
SW9621-1(config)# feature pim
SW9621-1(config)# vlan 100
SW9621-1(config)# interface vlan 100
SW9621-1(config-if)# ip pim sparse-mode
SW9621-1(config-if)# ip address 192.168.100.1/24
SW9621-1(config-if)# no shut
SW9621-1(config)# interface e2/3
SW9621-1(config-if)# switchport mode access
SW9621-1(config-if)# switchport access vlan 100

W9621-1# show ip igmp int vlan 100
IGMP Interfaces for VRF "default"
Vlan100, Interface status: protocol-up/link-up/admin-up
  IP address: 192.168.100.1, IP subnet: 192.168.100.0/24
  Active querier: 192.168.100.1, version: 2, next query sent in: 00:00:15
  Membership count: 0
  Old Membership count 0
  IGMP version: 2, host version: 2
  IGMP query interval: 125 secs, configured value: 125 secs
  IGMP max response time: 10 secs, configured value: 10 secs
  IGMP startup query interval: 31 secs, configured value: 31 secs
  IGMP startup query count: 2
  IGMP last member mrt: 1 secs
  IGMP last member query count: 2
  IGMP group timeout: 260 secs, configured value: 260 secs
  IGMP querier timeout: 255 secs, configured value: 255 secs
  IGMP unsolicited report interval: 10 secs
  IGMP robustness variable: 2, configured value: 2
  IGMP reporting for link-local groups: disabled
  IGMP interface enable refcount: 1
  IGMP interface immediate leave: disabled
  IGMP VRF name default (id 1)
  IGMP Report Policy: None
  IGMP State Limit: None
  IGMP interface statistics: (only non-zero values displayed)
    General (sent/received):
      v2-queries: 33/33, v2-reports: 0/0, v2-leaves: 0/0
    Errors:
  Interface PIM DR: Yes
Interface vPC SVI: No
  Interface vPC CFS statistics:

SW9621-1# show ip igmp snooping vlan 100
Global IGMP Snooping Information:
  IGMP Snooping enabled
  Optimised Multicast Flood (OMF) enabled
  IGMPv1/v2 Report Suppression enabled
  IGMPv3 Report Suppression disabled
  Link Local Groups Suppression enabled

IGMP Snooping information for vlan 100
  IGMP snooping enabled
  Lookup mode: IP
  Optimised Multicast Flood (OMF) enabled
  IGMP querier present, address: 192.168.100.1, version: 2, i/f Vlan100
  Querier interval: 125 secs
  Querier last member query interval: 1 secs
  Querier robustness: 2
  Switch-querier disabled
  IGMPv3 Explicit tracking enabled
  IGMPv2 Fast leave disabled
  IGMPv1/v2 Report suppression enabled
  IGMPv3 Report suppression disabled
  Link Local Groups suppression enabled
  Router port detection using PIM Hellos, IGMP Queries
  Number of router-ports: 1
  Number of groups: 0
  VLAN vPC function disabled
  Active ports:
    Eth2/3
```

Example 1-19 shows SW9621-2 Multicast PIM configurations and IGMP status.

Example 1-19 *Multicast IGMP SW9621-2 Configuration and Verifications*

```
SW9621-2(config)# feature interface-vlan
SW9621-2(config)# feature pim
SW9621-2(config)# vlan 200
SW9621-2(config)# int vlan 200
SW9621-2(config-if)# ip address 192.168.200.1/24
SW9621-2(config-if)# ip igmp ver 3
SW9621-2(config-if)# ip pim sparse-mode
SW9621-2(config-if)# no shut
SW9621-2(config)# interface e2/3
```

```
SW9621-2(config-if)# switchport mode access
SW9621-2(config-if)# switchport access vlan 200
SW9621(config-if)# show ip igmp snooping vlan 200
Global IGMP Snooping Information:
  IGMP Snooping enabled
  Optimised Multicast Flood (OMF) enabled
  IGMPv1/v2 Report Suppression enabled
  IGMPv3 Report Suppression disabled
  Link Local Groups Suppression enabled

IGMP Snooping information for vlan 200
  IGMP snooping enabled
  Lookup mode: IP
  Optimised Multicast Flood (OMF) enabled
  IGMP querier present, address: 192.168.200.1, version: 3, i/f Vlan200
  Querier interval: 125 secs
  Querier last member query interval: 1 secs
  Querier robustness: 2
  Switch-querier disabled
  IGMPv3 Explicit tracking enabled
  IGMPv2 Fast leave disabled
  IGMPv1/v2 Report suppression enabled
  IGMPv3 Report suppression disabled
  Link Local Groups suppression enabled
  Router port detection using PIM Hellos, IGMP Queries
  Number of router-ports: 1
  Number of groups: 0
  VLAN vPC function disabled
  Active ports:
    Eth2/3
```

Example 1-20 shows SW9621-1, SW9621-2, and SW9621-3 Multicast PIM configurations and IP PIM status.

Example 1-20 *Enabling PIM on Routers 1 to 3 (SW9621-1 to SW9621-3) and Verifications*

```
SW9621-1(config)# feature pim
SW9621-1(config)# interface e2/1
SW9621-1(config-if)# ip pim sparse-mode
SW9621-1(config)# interface e2/2
SW9621-1(config-if)# ip pim sparse-mode

SW9621-2(config)# feature pim
SW9621-2(config)# interface e2/1
```

```
SW9621-2(config-if)# ip pim sparse-mode
SW9621-2(config)# interface e2/2
SW9621-2(config-if)# ip pim sparse-mode

SW9621-3(config)# feature pim
SW9621-3(config)# interface e2/1
SW9621-3(config-if)# ip pim sparse-mode
SW9621-3(config)# interface e2/2
SW9621-1(config-if)# ip pim sparse-mode

SW9621-1# show ip pim neighbor
PIM Neighbor Status for VRF "default"
Neighbor        Interface .          Uptime    Expires   DR       Bidir-  BFD
                                                         Priority Capable State
10.10.10.2      Ethernet2/1          01:23:13  00:01:38  1        yes     n/a
10.10.10.6      Ethernet2/2          01:22:01  00:01:36  1        yes     n/a

SW9621-2(config-if)# show ip pim neighbor
PIM Neighbor Status for VRF "default"
Neighbor        Interface            Uptime    Expires   DR       Bidir-  BFD
                                                         Priority Capable State
10.10.10.1      Ethernet2/1          01:25:09  00:01:34  1        yes     n/a
10.10.10.9      Ethernet2/2          01:23:45  00:01:41  1        yes     n/a


SW9621-3# show ip pim neighbor
PIM Neighbor Status for VRF "default"
Neighbor        Interface            Uptime    Expires   DR       Bidir-  BFD
                                                         Priority Capable State
10.10.10.5      Ethernet2/1          01:24:15  00:01:37  1        yes     n/a
10.10.10.10     Ethernet2/2          01:24:04  00:01:18  1        yes     n/a
```

Example 1-21 shows SW9621-1 Multicast Static RP configurations and status.

Example 1-21 *Multicast Static RP Configurations and Verification*

```
SW9621-1(config)# ip pim rp 10.10.10.2

SW9621-1(config)# show ip pim rp
PIM RP Status Information for VRF "default"
BSR disabled
Auto-RP disabled
BSR RP Candidate policy: None
BSR RP policy: None
Auto-RP Announce policy: None
Auto-RP Discovery policy: None
```

```
RP: 10.10.10.2, (0),
 uptime: 00:00:14   priority: 0,
 RP-source: (local),
 group ranges:
 224.0.0.0/4

SW9621-1(config)# ip pim rp 10.10.10.6 group-list 239.0.200.0/24

SW9621-1(config)# show ip pim rp
PIM RP Status Information for VRF "default"
BSR disabled
Auto-RP disabled
BSR RP Candidate policy: None
BSR RP policy: None
Auto-RP Announce policy: None
Auto-RP Discovery policy: None

RP: 10.10.10.6, (0),
 uptime: 00:02:16   priority: 0,
 RP-source: (local),
 group ranges:
 239.0.200.0/24
```

Example 1-22 shows SW9621-1 multicast BSR RP configurations and status.

Example 1-22 *Multicast BSRs RP Configurations and Verification*

```
SW9621-1(config)# ip pim bsr bsr-candidate e2/1

SW9621-1(config)# show ip pim rp
PIM RP Status Information for VRF "default"
BSR: 10.10.10.1*, next Bootstrap message in: 00:00:53,
     priority: 64, hash-length: 30
Auto-RP disabled
BSR RP Candidate policy: None
BSR RP policy: None
Auto-RP Announce policy: None
Auto-RP Discovery policy: None

RP: 10.10.10.2, (0),
 uptime: 00:04:24   priority: 0,
 RP-source: (local),
 group ranges:
 224.0.0.0/4

SW9621-2(config)# ip pim bsr listen
```

```
SW9621-2(config)# show ip pim rp
PIM RP Status Information for VRF "default"
BSR listen-only mode
BSR: Not Operational
Auto-RP disabled
BSR RP Candidate policy: None
BSR RP policy: None
Auto-RP Announce policy: None
Auto-RP Discovery policy: None

RP: 10.10.10.6, (0),
 uptime: 00:03:39   priority: 0,
 RP-source: (local),
 group ranges:
 239.0.200.0/24
```

Example 1-23 shows SW9621-2 Multicast Auto RP configurations and status.

Example 1-23 *Multicast Auto RP Configurations and Verification*

```
SW9621-2(config)# ip pim auto-rp rp-candidate e2/1 group-list 239.0.0.0/24 bidir
SW9621-2(config)# ip pim auto-rp mapping-agent e2/1
SW9621-2(config)# ip pim auto-rp forward listen

SW9621-2(config)# show ip pim rp
PIM RP Status Information for VRF "default"
BSR listen-only mode
BSR: 10.10.10.1, uptime: 00:04:37, expires: 00:01:21,
     priority: 64, hash-length: 30
Auto-RP RPA: 10.10.10.2*, next Discovery message in: 00:00:05
BSR RP Candidate policy: None
BSR RP policy: None
Auto-RP Announce policy: None
Auto-RP Discovery policy: None

RP: 10.10.10.2*, (1),
 uptime: 00:00:16   priority: 0,
 RP-source: 10.10.10.2 (A),
 group ranges:
 239.0.0.0/24  (bidir)  , expires: 00:02:43 (A)
RP: 10.10.10.6, (0),
 uptime: 00:08:54   priority: 0,
 RP-source: (local),
 group ranges:
 239.0.200.0/24
```

Example 1-24 shows SW9621-3 multicast anycast RP configurations and status.

Example 1-24 *Multicast Anycast RP Configurations and Verification*

```
SW9621-3(config)# int loopback 0
SW9621-3(config-if)# ip add 10.1.1.1/32
SW9621-3(config-if)# ip router ospf 1 area 0
SW9621-3(config-if)# no shut

SW9621-3(config)# ip pim anycast-rp 10.1.1.1 10.10.10.6
SW9621-3(config)# ip pim anycast-rp 10.1.1.1 10.10.10.9

SW9621-2(config)# ip pim auto-rp mapping-agent e2/1
SW9621-2(config)# ip pim auto-rp forward listen

SW9621-3(config)# show ip pim rp
PIM RP Status Information for VRF "default"
BSR disabled
Auto-RP disabled
BSR RP Candidate policy: None
BSR RP policy: None
Auto-RP Announce policy: None
Auto-RP Discovery policy: None

Anycast-RP 10.1.1.1 members:
  10.10.10.6  10.10.10.9*
```

Example 1-25 shows SW9621-1 multicast SSM range configurations and status.

Example 1-25 *Multicast SSM Configurations and Verification*

```
SW9621-1(config)# ip ssm range 239.0.100.0/24

SW9621-1(config)# show ip pim group-range
PIM Group-Range Configuration for VRF "default"
Group-range        Action Mode  RP-address      Shrd-tree-range   Origin

239.0.100.0/24        Accept SSM   -               -                 Local

224.0.0.0/4     -      ASM   10.10.10.2           -                   Static
```

Example 1-26 shows the IP multicast routing table.

Example 1-26 *Multicast IP Route Example*

```
SW9621-1(config)# show ip mroute

IP Multicast Routing Table for VRF "default"
```

```
(*, 239.0.100.0/24), uptime: 00:03:07, pim ip
  Incoming interface: Null, RPF nbr: 0.0.0.0
  Outgoing interface list: (count: 0)

SW9621-2(config)# show ip mroute
IP Multicast Routing Table for VRF "default"

(*, 232.0.0.0/8), uptime: 00:30:21, pim ip
  Incoming interface: Null, RPF nbr: 0.0.0.0
  Outgoing interface list: (count: 0)

(*, 239.0.0.0/24), bidir, uptime: 00:10:46, pim ip
  Incoming interface: Ethernet2/1, RPF nbr: 10.10.10.2
  Outgoing interface list: (count: 0)

SW9621-3(config)# show ip mroute
IP Multicast Routing Table for VRF "default"

(*, 232.0.0.0/8), uptime: 00:29:36, pim ip
  Incoming interface: Null, RPF nbr: 0.0.0.0
  Outgoing interface list: (count: 0)
```

Key Topic

Hot Standby Router Protocol

Hot Standby Router Protocol (HSRP) is a First Hop Redundancy Protocol (FHRP) that allows a transparent failover of the first hop gateway router. HSRP provides first hop routing redundancy for IP hosts on Ethernet networks configured with a gateway or default route. You can use HSRP in a group of routers for selecting an active router and a standby router. In a group of two routers, the active router is the router that routes packets; the standby router is the router that takes over when the active router fails or when preset conditions are met.

Many host implementations do not support any dynamic router discovery mechanisms but can be configured with a default router. Running a dynamic router discovery mechanism on every host is not feasible for a number of reasons, including administrative overhead, processing overhead, and security issues. HSRP provides failover services to these hosts.

When you use HSRP, you need an HSRP virtual IP address as the host's default router (instead of the IP address of the actual router). The virtual IP address is an IP address that is shared among a group of routers that run HSRP. Configuring HSRP on a network segment will provide a virtual MAC address and a virtual IP address for the HSRP group. You need to configure the same virtual address on each HSRP-enabled interface in the group. You also configure a unique IP address and MAC address on each interface that acts as the real address. HSRP selects one of these interfaces to be the active router. The active router receives and routes packets destined for the virtual MAC address of the group.

HSRP detects when the active router fails. At that point, a selected standby router assumes control of the virtual MAC and IP addresses of the HSRP group. HSRP also selects a new standby router at that time.

HSRP uses a priority mechanism to determine which HSRP-configured interface becomes the default active router. To configure an interface as the active router, you assign it with a priority that is higher than the priority of all the other HSRP-configured interfaces in the group. The default priority is 100, so if you configure just one interface with a higher priority, that interface becomes the default active router.

Interfaces that run HSRP send and receive multicast User Datagram Protocol (UDP)-based hello messages to detect a failure and to designate active and standby routers. When the active router fails to send a hello message within a configurable period of time, the standby router with the highest priority becomes the active router. The transition of packet forwarding functions between the active and standby router is completely transparent to all hosts on the network.

NOTE You can configure multiple HSRP groups on an interface.

Figure 1-19 shows a network configured for HSRP. By sharing a virtual MAC address and a virtual IP address, two or more interfaces can act as a single virtual router.

Key Topic

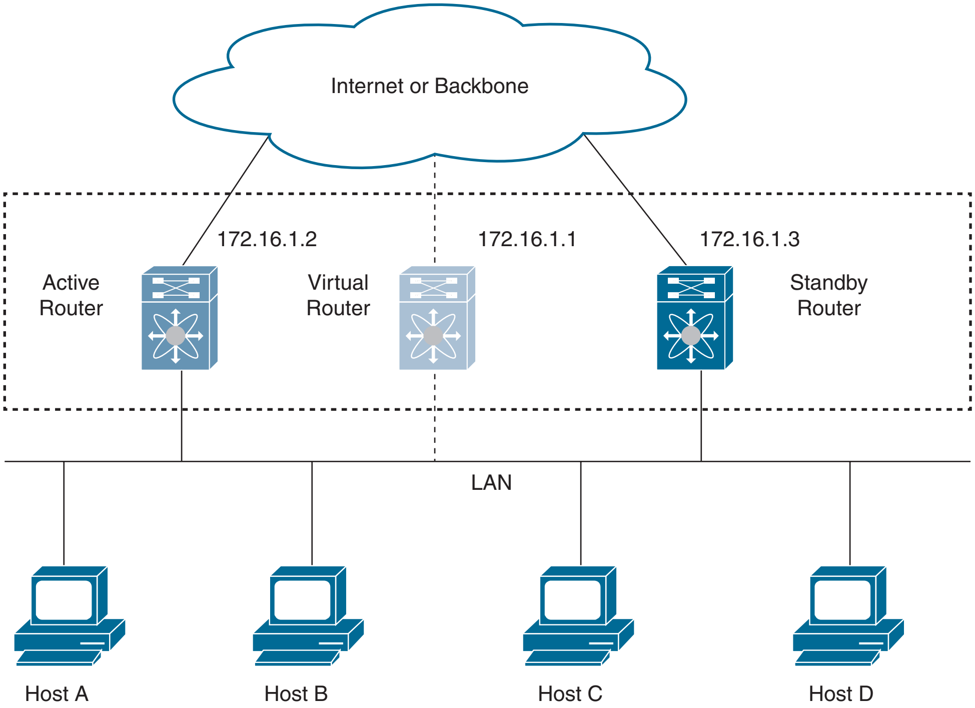


Figure 1-19 *HSRP Topology with Two Enabled Routers*

The virtual router does not physically exist but represents the common default router for interfaces that are configured to provide backup to each other. You do not need to configure the hosts on the LAN with the IP address of the active router. Instead, you configure them with the IP address (virtual IP address) of the virtual router as their default router.

If the active router fails to send a hello message within the configurable period of time, the standby router takes over, responds to the virtual addresses, and becomes the active router, assuming the active router duties. From the host perspective, the virtual router remains the same.

Cisco NX-OS supports HSRP version 1 by default. You can, instead, configure an interface to use HSRP version 2.

HSRP version 2 has the following enhancements to HSRP version 1:

- Expands the group number range. HSRP version 1 supports group numbers from 0 to 255. HSRP version 2 supports group numbers from 0 to 4095.
- Uses the new IP multicast address 224.0.0.102 to send hello packets instead of the multicast address of 224.0.0.2, which is used by HSRP version 1.
- Uses the MAC address range from 0000.0C9F.F000 to 0000.0C9F.FFFF. HSRP version 1 uses the MAC address range 0000.0C07.AC00 to 0000.0C07.ACFF.
- Adds support for MD5 authentication.

When you change the HSRP version, Cisco NX-OS reinitializes the group because it now has a new virtual MAC address.

HSRP version 2 has a different packet format than HSRP version 1. The packet format uses a type-length-value (TLV) format. HSRP version 2 packets received by an HSRP version 1 router are ignored.

HSRP message digest 5 (MD5) algorithm authentication protects against HSRP-spoofing software and uses the industry-standard MD5 algorithm for improved reliability and security.

HSRP routers communicate with each other by exchanging HSRP hello packets. These packets are sent to the destination IP multicast address 224.0.0.2 (reserved multicast address used to communicate to all routers) on UDP port 1985. The active router sources hello packets from its configured IP address and the HSRP virtual MAC address while the standby router sources hellos from its configured IP address and the interface MAC address, which may or may not be the burned-in address (BIA). The BIA is the last 6 bytes of the MAC address that is assigned by the manufacturer of the network interface card (NIC).

Because hosts are configured with their default router as the HSRP virtual IP address, they must communicate with the MAC address associated with the HSRP virtual IP address. This MAC address is a virtual MAC address, 0000.0C07.ACxy, where xy is the HSRP group number in hexadecimal based on the respective interface. For example, HSRP group 1 will use the HSRP virtual MAC address 0000.0C07.AC01. Hosts on the adjoining LAN segment use the normal Address Resolution Protocol (ARP) process to resolve the associated MAC addresses.

HSRP version 2 uses the new IP multicast address 224.0.0.102 to send hello packets instead of the multicast address 224.0.0.2, which is used by version 1. HSRP version 2 permits an expanded group number range of 0 to 4095 and uses a new MAC address range of 0000.0C9F.F000 to 0000.0C9F.FFFF.