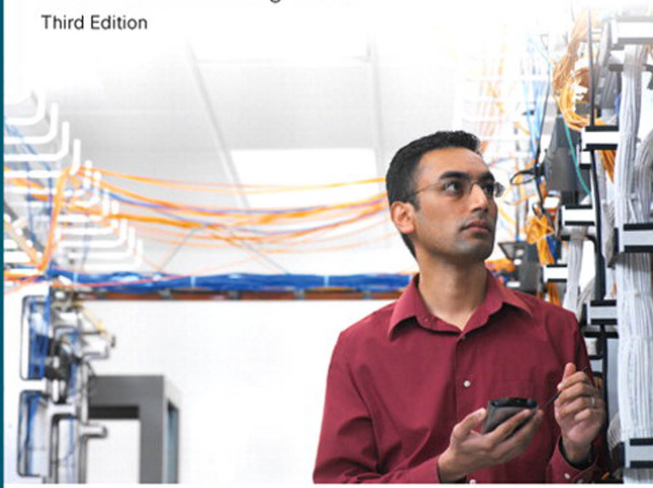
cisco

# Designing Cisco Network Service Architectures (ARCH)

## Foundation Learning Guide

Third Edition



ciscopress.com

John Tiso, CCIE® No. 5162

Foundation Learning Guide

# Designing Cisco Network Service Architectures (ARCH)

## Third Edition

John Tiso

**Cisco Press**

800 East 96th Street

Indianapolis, IN 46240

**Foundation Learning Guide**

# Designing Cisco Network Service Architectures (ARCH)

**Third Edition**

John Tiso



Published by:
Cisco Press
800 East 96th Street
Indianapolis, IN 46240 USA



Printed in the United States of America

Second Printing December 2013

Library of Congress Cataloging-in-Publication Data

Tiso, John.
Authorized self-study guide designing Cisco network service architectures (arch) / John Tiso. -- 3rd ed.
p. cm.
Rev. ed. of: Authorized self-study guide / Hutton, Keith. 2009. 2nd ed.
ISBN 978-1-58714-288-8 (hardcover)
1. Computer network architectures--Examinations--Study guides. 2. Computer networks--Design--Examinations--Study guides. 3. Internetworking (Telecommunication)--Examinations--Study guides. I. Hutton, Keith. Authorized self-study guide. II. Title.
TK5105.52.H98 2012
004.6'5--dc23

2011036250

ISBN-13: 978-1-58714-288-8

ISBN-10: 1-58714-288-0

## Warning and Disclaimer

This book is designed to provide information about designing Cisco network service architectures. 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 author, 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 author 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.

## Corporate and Government Sales

The publisher offers excellent discounts on this book when ordered in quantity for bulk purchases or special sales, which may include electronic versions and/or custom covers and content particular to your business, training goals, marketing focus, and branding interests. For more information, please contact:
**U.S. Corporate and Government Sales**
1-800-382-3419
corpsales@pearsontechgroup.com

For sales outside the United States, please contact:
**International Sales**
international@pearsoned.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 e-mail at feedback@ciscopress.com. Please make sure to include the book title and ISBN in your message.

We greatly appreciate your assistance.

**Publisher:** Paul Boger

**Associate Publisher:** Dave Dusthimer

**Executive Editor:** Brett Bartow

**Managing Editor:** Sandra Schroeder

**Project Editor:** Mandie Frank

**Editorial Assistant:** Vanessa Evans

**Designer:** Gary Adair

**Cisco Press Program Manager:** Jeff Brady

**Technical Editors:** Diane Teare; Dr. Peter J. Welcher

**Development Editor:** Marianne Bartow

**Copy Editor:** Keith Cline

**Proofreader:** Sheri Cain

**Indexer:** Tim Wright

**Composition:** Mark Shirar

## About the Author

**John Tiso**, CCIE #5162, CCDP is a Product Manager at Cisco Systems. His current responsibilities include the product management of Cisco's training and certification programs around design and architecture. Before working with Cisco, John held various engineering and architecture roles in the Cisco partner channel. In addition to his CCIE and CCDP certifications, he holds multiple industry certifications from Cisco, Microsoft, CompTIA, and Sun Microsystems. He holds a Graduate Citation in strategic management from Harvard University and a Bachelor of Science degree in computer science and mathematics from Adelphi University. John is a published author and has served as a technical editor for both McGraw-Hill and Cisco Press. He has spoken multiple times at the Cisco Live! (Networkers) conference and the national CIPTUG conference. He has served as an expert on Cisco's NetPro Forum "Ask the Expert" online events. John currently resides in Amherst, New Hampshire, with his wife, three children, and his running partner, Molly (who never complains, but sometimes barks). He is a nine-time marathon finisher, including five Boston Marathons. He can be reached at johnt@jtiso.com.

## Contributing Author

**Ed Caswell** is a Systems Engineering Manager at Cisco Systems. Ed has spoken several times at the CIPTUG national convergence event and many times at regional architectural conferences. He has served as a Subject Matter Expert (SME) on many industry panels. Ed has also edited several collaboration books for Cisco Press. Prior to Cisco, Ed held positions in system management and other individual contributor roles. He is considered a trusted collaboration expert in the industry. Ed is a military veteran and an active member of his community.

## About the Technical Reviewers

**Diane Teare**, P.Eng, CCNP, CCDP, PMP, is a professional in the networking, training, project management, and e-learning fields. She has more than 25 years of experience in designing, implementing, and troubleshooting network hardware and software, and has been involved in teaching, course design, and project management. She has extensive knowledge of network design and routing technologies, and is an instructor with one of the largest authorized Cisco Learning Partners. She was the director of e-learning for the same company, where she was responsible for planning and supporting all the company's e-learning offerings in Canada, including Cisco courses. Diane has a Bachelor's degree in applied science in electrical engineering and a Master's degree in applied science in management science. She authored or co-authored the Cisco Press titles *Implementing Cisco IP Routing (ROUTE) Foundation Learning Guide*; the previous (second) edition of *Designing Cisco Network Service Architectures (ARCH)*; *Campus Network Design Fundamentals*; the three editions of *Authorized Self-Study Guide Building Scalable Cisco Internetworks (BSCI)*; and *Building Scalable Cisco Networks*. Diane edited the *Authorized Self-Study Guide Designing for Cisco Internetwork Solutions (DESGN)* (the first two editions) and the *Designing Cisco Networks*.

**Dr. Peter J. Welcher**, CCIE #1773, works for Chesapeake NetCraftsmen, a small but highly skilled Cisco Gold Partner with a stellar reputation in the industry. Pete has developed numerous design and other courses for Cisco while teaching and consulting. His consulting includes campus and data center design and migration to *N* x 10-Gbps technology and 6500 VSS plus NAC for a fairly large federal agency, data center assessment, and other design-related work with a large international hotel chain, design review for two universities, redesign for a internationally known university and research hospital, WAN/QoS for a federal agency with more than 250,000 employees, WLAN pilot design that may impact another large federal agency, work for two New York stock market quotation firms, and so on. He has reviewed a number of book proposals and done tech edits in the past for Cisco Press and is well known in the industry. Pete is currently focusing on data center technology and enjoys teaching the Nexus class one week a month. He has also written more than 170 blog posts (a number that might be significantly larger by the time these words see print).

# Acknowledgments

I want to acknowledge and thank the following persons.

The team at Cisco Press, especially Brett Bartow, for pulling everything together and listening to my rants. Marianne Bartow for tolerating my poor formatting, whining, and general mistakes. Marianne, you made it all happen. Thank you!

The technical editors: Dr. Peter Welcher and Diane Teare. Your feedback kept me honest.

Ed Caswell for his contributions to the book.

All the friends and co-workers who have been supportive of me over the past few years.

My wife, Lauren, and my children, Danny, Nick and Katie, for tolerating me and just for being a great family!

Finally, you, the reader and certification candidate. Without you, I would have neither the opportunity to work on this nor a job. Good luck in what you seek.

# Contents at a Glance

# Contents

## Icons Used in This Book

## 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.

# Foreword

With emerging IT trends such as virtualization, data center, collaboration, and mobility, the demands placed on network professionals have become more varied and sophisticated. Knowledge of networking fundamentals is no longer sufficient. Network professionals must understand networking systems with integrated security, wireless, and voice capabilities.

The technology evolution is creating a global IT skills gap, as corporations, regardless of size, industry, or geography, seek professionals who have more than product and technology skills. Instead, the focus on today's IT professional is their ability to fulfill real-world job role requirements. In fact, new careers in IT are being established that are based more on complex services and architectures than systems.

In support of the industry's demand for "the right resources, at the right place, at the right time," Cisco remains dedicated to the development of the next-generation IT workforce. Assessment of an individual's knowledge and ability to perform the required job tasks of a particular role is measured by the Cisco Career and Specialist Certifications, which are recognized worldwide.

Cisco Press, an industry leader in networking education content, offers the only self-study books authorized by Cisco. The Cisco Press exam certification guides and preparation materials offer exceptional and flexible access to the knowledge and information required to stay current in one's field of expertise or to gain new skills. Whether used to increase internetworking skills or as a supplement to a formal certification preparation course, these materials offer networking professionals the information and knowledge required to perform on-the-job tasks proficiently.

Additional authorized Cisco instructor-led courses are available exclusively through our Authorized Cisco Learning Partners worldwide. Other self-study materials including e-learning, practice exams, labs, and simulations are available from the Cisco Learning Network, our Web 2.0 social learning community. To learn more, visit https://learningnetwork.cisco.com. I hope that you find this material to be an essential part of your education, exam preparation, and professional development and that it becomes a valuable addition to your personal library.

Tejas R. Vashi
Director, Product Management
Cisco Technical Services

August 2011

## Authors Note

This self-study work has received GOLD certification as an IPv6 certified course from the IPv6 forum for the content of this book. The related certification—CCDP—has received GOLD certification as an IPv6 certified certification as well. As a networking design professional, please keep IPv6 in the forefront of your designs:


# Introduction

*Designing Cisco Network Service Architectures (ARCH), Third Edition*, covers how to perform the conceptual, intermediate, and detailed design of a network infrastructure. This design supports network solutions over intelligent network services to achieve effective performance, scalability, and availability of the network. This book enables readers, applying solid Cisco network solution models and best design practices, to provide viable and stable enterprise internetworking solutions. In addition, the book has been written to help candidates prepare for the Designing Cisco Network Service Architectures Exam (642-874 ARCH). This exam is one of the requirements for the Cisco Certified Design Professional (CCDP) certification. This exam tests a candidate's knowledge of the latest development in network design and technologies, including network infrastructure, intelligent network services, and converged network solutions.

Since the first edition was published in 2004, the Designing Cisco Network Services Architectures (ARCH) authorized training course has been updated to keep pace with the industry. Therefore, the exam was consequently updated to match these changes. This led to the immediate need for an update to this examination preparation text. Readers of the previous edition of this work can use this text to update their knowledge and skill sets.

In certain cases, parts of this book may discuss obsolete, end-of-life, or suboptimal configurations. This ensures that the book aligns with the ARCH exam. Whenever possible, this is noted.

## Goals of This Book

Upon completing this book, you will be able to meet these objectives:

- Introduce the Cisco Borderless Networks architectural framework and explain how it addresses enterprise network needs for performance, scalability, security, unified communications, and availability.
- Describe how the Cisco enterprise architectures are used in the Borderless Networks framework for designing enterprise networks.
- Create intermediate and detailed enterprise campus network, enterprise edge, and remote infrastructure designs that offer effective functionality, performance, scalability, and availability.
- Create conceptual, intermediate, and detailed intelligent network service designs for network management, high availability, security, quality of service (QoS), and IP multicast.
- Create conceptual, intermediate, and detailed virtual private network (VPN) designs.

## Prerequisite Knowledge

Although enthusiastic readers will tackle less-familiar topics with some energy, a sound grounding in networking is advised. To gain the most from this book, you should be familiar with internetworking technologies, Cisco products, and Cisco IOS Software features. Although exams for the CCDP (or Cisco Certified Network Professional [CCNP]) do not require you to pass them in order, I strongly advise you to consider passing them in order because the program builds on itself. This book is also an excellent resource to learn about Borderless Networks (perhaps even in preparation for the Cisco Systems Borderless Networks Specialization certification).

**Note** Many Cisco specialist certifications are aligned to Channel Partner specialization requirements. Channel Partner employees should access the Partner Education Connection (Cisco.com partner level access) to find the latest roadmaps and information as it pertains to Channel Partner certifications and requirements. Specialization certifications are open to all members of the public and are not limited to Channel Partner employees.

This book covers the following topics to aid your journey toward CCDP certification:

- How to design the necessary services to extend IP addresses using variable-length subnet masking (VLSM), Network Address Translation (NAT), and route summarization
- How to implement appropriate networking routing protocols, such as Open Shortest Path First (OSPF), Enhanced Interior Gateway Routing Protocol (EIGRP), and Border Gateway Protocol (BGP) on an existing internetwork
- How to redistribute routes between different routing protocols
- The required Cisco products and services that enable connectivity and traffic transport for a multilayer campus network
- Design of data center services
- The necessary services at each layer of the network to enable all users to obtain membership in multicast groups in a working enterprise network
- How to control network traffic by implementing the necessary admission policy at each layer of the network topology
- How to identify the appropriate hardware and software solutions for a given set of WAN technology requirements, including access between a central campus, branch offices, and telecommuters
- The Cisco equipment to establish appropriate WAN connections
- How to use protocols and technologies that enable traffic flow between multiple sites while minimizing the amount of overhead traffic on each connection

- QoS capabilities to ensure that mission-critical applications receive the required bandwidth within a given WAN topology
- How to implement Cisco Unified Communications
- How to implement Cisco wireless solutions
- How to implement basic security steps and mitigation techniques

## How This Book Is Organized

Of course, you can read the chapters in this book sequentially, but the organization of this book also allows you to focus your reading on specific topics of interest. For example, if you want to focus on advanced routing design, you can skim Chapters 1 and 2 (which cover Borderless Networks and the elements of the enterprise campus network design), and then focus on the advanced IP addressing and routing topics in Chapter 3. Each chapter examines topics around a specific set of design issues. Specifically, the chapters in this book cover the following topics:

- Chapter 1, "The Cisco Enterprise Architecture," introduces the methodology configured for network engineers to design scalable, robust infrastructures to support today's complicated business applications. This includes the Cisco Borderless Networks architecture.
- Chapter 2, "Enterprise Campus Network Design," reviews high-availability designs and how to implement optimal redundancy. An in-depth look at recommended practices for Layer 2 and Layer 3 design elements follows. A discussion of the Layer 2 to Layer 3 boundary designs and issues concludes with a number of considerations for supporting infrastructure services.
- Chapter 3, "Developing an Optimum Design for Layer 3," begins by reviewing the importance of IP address planning, and then covers advanced routing elements. Discussions focus on scalable EIGRP, OSPF, and BGP designs.
- Chapter 4, "Advanced WAN Services Design Considerations," covers advanced WAN service layers. This overview goes into more detail about the common WAN optical technologies of Synchronous Optical Network (SONET), Synchronous Digital Hierarchy (SDH), dense wave-division multiplexing (DWDM), and Resilient Packet Ring. A discussion about Metro Ethernet, Virtual Private Line Service (VPLS), and Multiprotocol Label Switching virtual private network (MPLS VPN) technologies follows (and includes an examination of a number of design considerations). The discussion then turns to implementing advanced WAN services.
- Chapter 5, "Enterprise Data Center Design," focuses on the enterprise data center, and covers the data center architecture model and design consideration in the data center core, aggregation, and access layers. The discussion then turns to scaling, with a look at how to scale a three-layer data center architecture.

- Chapter 6, "SAN Design Considerations," covers storage-area networks, from components and topologies to SAN technologies. SAN design factors center on port density and topology, with some discussion about extending the SAN with various protocols.
- Chapter 7, "E-Commerce Module Design," begins with an e-commerce overview and a look at the components of high availability in this module. The chapter covers common e-commerce design components, designing an integrated e-commerce architecture, and how to fine-tune e-commerce designs.
- Chapter 8, "Security Services Design," delves into designing firewall services in various scenarios. The chapter also covers network admission control services, with a review of Cisco Network Admission Control (NAC) appliance fundamentals and NAC Appliance Server (NAS) deployment options and designs. The discussion then turns to intrusion-detection and -prevention design.
- Chapter 9, "IPsec and SSL VPN Design," examines remote-access VPN design. Site-to-site VPN designs are covered, too. This chapter also covers IPsec VPN technologies, including Cisco Easy VPN, generic routing encapsulation (GRE) over IPsec, and Dynamic Multipoint VPN (DMVPN). Recommendations for managing VPNs and considerations for scaling VPNs conclude the chapter.
- Chapter 10, "IP Multicast Design," covers IP multicast and multicast routing. Topics covered in this chapter include Protocol Independent Multicast (PIM), rendezvous points, and securing IP multicast.
- Chapter 11, "Network Management Capabilities Within Cisco IOS Software," examines Cisco network management capabilities embedded in Cisco IOS Software. This chapter also covers the syslog process, NetFlow, and Network-Based Application Recognition (NBAR), with a focus on the Cisco technologies themselves and how they enable other discovery tools, including Cisco AutoQoS. The chapter concludes with an overview of IP service-level agreement (SLA) measurements.

This book also contains an appendix and an acronym list:

- Appendix A, "Answers to Review Questions," provides the answers to all the chapter-ending review questions.
- Appendix B, "Acronyms and Abbreviations," identifies abbreviations, acronyms, and initialisms used in this book.
- Appendix C, "VoWLAN Design," introduces the Cisco Unified Wireless Network and examines requirements for voice over WLAN in the enterprise network. This appendix, which was Chapter 11 in the previous edition, also discusses VoWLAN coverage considerations and the site-survey process. It has been moved here because the matter within is not part of the ARCH exam in this version.

**Note** The website references in this book were accurate at the time of this writing. However, some might have changed since then. If a URL is unavailable, you can always search using the title as keywords in your favorite search engine.

Chapter 3

# Developing an Optimum Design for Layer 3

After completing this chapter, you will be able to

- Design IPv4 and IPv6 addressing solutions to support summarization
- Design IPv6 migration schemes
- Design routing solutions to support summarization, route filtering, and redistribution
- Design scalable EIGRP routing solutions for the enterprise
- Design scalable OSPF routing solutions for the enterprise
- Design scalable BGP routing solutions for the enterprise

This chapter examines a select number of topics on both advance IP addressing and design issues with Border Gateway Protocol (BGP), Enhanced Interior Gateway Routing Protocol (EIGRP), and Open Shortest Path First (OSPF). As one would expect, advanced IP addressing and routing protocol design encompasses a large amount of detail that has already filled a number of books on routing protocols and networking best practices.

## Designing Advanced IP Addressing

Designing IP addressing at a professional level involves several advanced considerations. This section reviews the importance of IP address planning and selection and the importance of IP address summarization. It also discusses some applications of summary addressing.

**Note** In this chapter, IP (unless specified as IPv6) refers to IPv4.

## IP Address Planning as a Foundation

Structured and modular cabling plant and network infrastructures are ideal for a good design with low maintenance and upgrade costs. In similar fashion, a well-planned IP addressing scheme is the foundation for greater efficiency in operating and maintaining a network. Without proper advanced planning, networks may not be able to benefit from route summarization features inherent to many routing protocols.

Route summarization is important in scaling any routing protocol. However, some existing IP addressing schemes may not support summarization. It takes time and effort to properly allocate IP subnets in blocks to facilitate summarization. The benefits of summarized addresses are reduced router workload and routing traffic and faster convergence. Although modern router CPUs can handle a vastly increased workload as compared to older routers, reducing load mitigates the impact of periods of intense network instability. In general, summary routes dampen out or reduce network route churn, making the network more stable. In addition, summary routes lead to faster network convergence. Summarized networks are simpler to troubleshoot because there are fewer routes in the routing table or in routing advertisements, compared to nonsummarized networks.

Just as using the right blocks of subnets enables use of more efficient routing, care with subnet assignments can also support role-based functions within the addressing scheme structure. This in turn enables efficient and easily managed access control lists (ACL) for quality of service (QoS) and security purposes.

In addition to allocating subnets in summarized blocks, it is advantageous to choose blocks of addresses within these subnets that can be easily summarized or described using wildcard masking in access control lists (ACL). With a well-chosen addressing scheme, ACLs can become much simpler to maintain in the enterprise.

### Summary Address Blocks

Summary address blocks are the key to creating and using summary routes. How do you recognize a block of addresses that can be summarized? A block of IP addresses might be able to be summarized if it contains sequential numbers in one of the octets. The sequence of numbers must fit a pattern for the binary bit pattern to be appropriate for summarization. The pattern can be described without doing binary arithmetic.

For the sequential numbers to be summarized, the block must be $x$ numbers in a row, where $x$ is a power of 2. In addition, the first number in the sequence must be a multiple of $x$. The sequence will always end before the next multiple of $x$.

For example, any address block that matches the following can be summarized:

- 128 numbers in a row, starting with a multiple of 128 (0 or 128)
- 64 numbers in a row, starting with a multiple of 64 (0, 64, 128, or 192)
- 32 numbers in a row, starting with a multiple of 32
- 16 numbers in a row, starting with a multiple of 16

If you examine 172.19.160.0 through 172.19.191.0, there are 191 – 160 + 1 = 32 numbers in a row, in sequence in the third octet. Note that 32 is $2^5$ power of 2. Note also that 160 is a multiple of 32 (5 * 32 = 160). Because the range meets the preceding conditions, the sequence 172.19.160.0 through 172.19.191.0 can be summarized.

Finding the correct octet for a subnet-style mask is fairly easy with summary address blocks. The formula is to subtract *n* from 256. For example, for 32 numbers in a row, the mask octet is 256 – 32 = 224. Because the numbers are in the third octet, you place the 224 in the third octet, to form the mask 255.255.224.0.

A summary route expressed as either 172.19.160.0, 255.255.224.0, or as 172.169.160/19 would then describe how to reach subnets starting with 172.19.160.0 through 172.19.191.0.

### Summarization for IPv6

Although the address format of IPv6 is different from IPv4, the same principles apply. Blocks of subsequent IPv6 /64 subnets can be summarized into larger blocks for decreased routing table size and increased routing table stability. To an extent, routing summarization for IPv6 is simpler than for IPv4, because you do not have to consider variable-length subnet masking (VLSM). Most IPv6 subnets have a prefix length of 64 bits, so again, you are looking for contiguous blocks of /64 subnets. The number of subnets in this block should be a power of 2, and the starting number should be a multiple of that same power of 2 for the block to be summarizable.

For example, examine the block 2001:0DB8:0:A480::/64 to 2001:0DB8:0:A4BF::/64. A quick analysis of the address block shows that the relevant part is in the last two hexadecimal characters, which are 0x80 for the first subnet in the range and 0xBF for the last subnet in the range. Conversion of these numbers to decimal yields 0x80 = 128 and 0xBF = 191. This is a block of 191 – 128 + 1 = 64 subnets. After verifying that 128 is a multiple of 64, you can conclude that the block of subnets is can be summarized.

To calculate the prefix length, you need to find the number of bits represented by the block of 64 addresses. 64 = $2^6$; therefore, 6 bits need to be subtracted from the original /64 prefix length to obtain the prefix length of the summary, which is /58 (64 – 6 = 58).

As a result, a summary route of 2001:0DB8:0:A480::/58 can be used to describe how to reach subnets 2001:0DB8:0:A480::/64 to 2001:0DB8:0:A4BF::/64.

**Note** If you have a well-chosen IPv4 addressing scheme that summarizes properly, you might consider mapping those addresses to an IPv6 addressing scheme. For example, the issued /48 prefix, append the second and third octets from your network 10.x.y.0/24, voilà! Jeff Doyle and others advise that if you're tight on subnets, if your addressing doesn't summarize well, and for other such reasons, it is usually best to do IPv6 addressing from scratch, rather than tying it to a poorly conceived legacy IPv4 addressing scheme.

### Changing IP Addressing Needs

IP address redesign is necessary to adapt to changes in how subnets are now being used. In some networks, IP subnets were initially assigned sequentially. Summary address blocks of subnets were then assigned to sites to enable route summarization.

However, newer specifications require additional subnets, as follows:

- **IP telephony:** Additional subnets or address ranges are needed to support voice services. In some cases, the number of subnets double when IP telephony is implemented in an organization.
- **Videoconferencing:** Immersive TelePresence applications are high bandwidth and sensitive to loss and latency. Generally, best practice is to segment these devices, creating the need for more subnets.
- **Layer 3 switching at the edge:** Deploying Layer 3 switching to the network edge is another trend driving the need for more subnets. Edge Layer 3 switching can create the demand for a rapid increase in the number of smaller subnets. In some cases, there can be insufficient address space, and readdressing is required.
- **Network Admission Control (NAC):** NAC is also being deployed in many organizations. Some Cisco 802.1X and NAC deployments are dynamically assigning VLANs based on user logins or user roles. In these environments, ACLs control connectivity to servers and network resources based on the source subnet, which is based on the user role.
- **Corporate requirements:** Corporate governance security initiatives are also isolating groups of servers by function, sometimes called segmentation. Describing "production" and "development" subnets in an ACL can be painful unless they have been chosen wisely. These new subnets can make managing the network more complex. Maintaining ad hoc subnets for voice security and other reasons can be time-consuming. When it is possible, describing the permitted traffic in a few ACL statements is a highly desirable. Therefore, ACL-friendly addressing which can be summarized helps network administrators to efficiently manage their networks.

### Planning Addresses

The first step in implementing ACL-friendly addressing is to recognize the need. In an environment with IP phones and NAC implemented, you need to support IP phone subnets and NAC role subnets in ACLs. In the case of IP phones, ACLs will probably be used for both QoS and voice-security rules. For NAC role-based subnets, ACLs will most likely be used for security purposes.

Servers in medium-to-large server farms should at least be grouped so that servers with different functions or levels of criticality are in different subnets. That saves listing individual IP addresses in lengthy ACLs. If the servers are in subnets attached to different access switches, it can be useful to assign the subnets so that there is a pattern suitable for wildcarding in ACLs.

If the addressing scheme allows simple wildcard rules to be written, those simple ACL rules can be used everywhere. This avoids maintaining per-location ACLs that need to define source or destination addresses to local subnets. ACL-friendly addressing supports maintaining one or a few global ACLs, which are applied identically at various control points in the network. This would typically be accomplished with a tool such as Cisco Security Manager.

The conclusion is that it is advantageous to build a pattern into role-based addressing and other addressing schemes so that ACL wildcards can match the pattern. This in turn supports implementing simpler ACLs.

**Note** For IPv6 access lists, the wildcard masks are not usually used. All source and destination addresses are notated in the form of prefixes. Therefore, it is important that subnets that are to be grouped in an access list falling within a summarized address range.

### Applications of Summary Address Blocks

Summary address blocks can be used to support several network applications:

- Separate VLANs for voice and data, and even role-based addressing
- Bit splitting for route summarization
- Addressing for virtual private network (VPN) clients
- Network Address Translation (NAT)

These features are discussed in detail in the following sections.

### Implementing Role-Based Addressing

The most obvious approach to implement role-based addressing is to use network 10. This has the virtue of simplicity. A simple scheme that can be used with Layer 3 closets is to use 10.number_for_closet.VLAN.x /24 and avoid binary arithmetic. This approach uses the second octet for closets or Layer 3 switches, the third octet for VLANs, and the fourth octet for hosts.

If you have more than 256 closets or Layer 3 switches to identify in the second octet, you might use some bits from the beginning of the third octet, because you probably do not have 256 VLANs per switch.

Another approach is to use some or all of the Class B private addressing blocks. This approach will typically involve binary arithmetic. The easiest method is to allocate bits using bit splitting. An example network is 172.0001 xxxx.xxxx xxxx.xxhh hhhh. In this case, you start out with 6 bits reserved for hosts in the fourth octet, or 62 hosts per subnet (VLAN). The *x* bits are to be split further.

This format initially uses decimal notation to the first octet and binary notation in the second, third, and fourth octets to minimize conversion back and forth.

If you do not need to use the bits in the second octet to identify additional closets, you end up with something like 172.16.cccc cccR.RRhh hhhh:

- The *c* characters indicate that 7 bits allow for $2^7$ or 128 closet or Layer 3 switches.
- The *R* characters indicate 3 bits for a role-based subnet (relative to the closet block), or 8 NAC or other roles per switch.
- The *h* characters indicate 6 bits for the 62-host subnets specified.

This addressing plan is enough to cover a reasonably large enterprise network.

Another 4 bits are available to work with in the second octet if needed.

Using a role-aware or ACL-friendly addressing scheme, you can write a small number of global **permit** or **deny** statements for each role. This greatly simplifies edge ACL maintenance. It is easier to maintain one ACL for all edge VLANs or interfaces than different ACLs for every Layer 3 access or distribution switch.

**Note** The role-based approach depends on the use of noncontiguous wildcard masks to match multiple subnets that fit a specific role. This dependency makes the method unsuitable for IPv6 and IPv4 (if devices in the path do not support discontiguous masks).

### Bit Splitting for Route Summarization

The previous bit-splitting technique has been around for a while. It can also be useful in coming up with summary address block for routing protocols if you cannot use simple octet boundaries. The basic idea is to start with a network prefix, such as 10.0.0.0, or a prefix in the range 172.16.0.0 to 172.31.0.0, 192.168.n.0, or an assigned IP address. The remaining bits can then be thought of as available for use for the area, subnet, or host part of the address. It can be useful to write the available bits as *x*, then substitute *a*, *s*, or *h* as they are assigned. The *n* in an address indicates the network prefix portion of the address, which is not subject to change or assignment.

Generally, you know how large your average subnets need to be in buildings. (A subnet with 64 bits can be summarized and will cover most LAN switches.) That allows you to convert six *x* bits to *h* for host bits.

You can then determine the number of necessary WAN links and the amount you are comfortable putting into one area to decide the number of *a* bits you need to assign. The leftover bits are *s* bits. Generally, one does not need all the bits, and the remaining bits (the *a* versus *s* boundary) can be assigned to allow some room for growth.

For example, suppose 172.16.0.0 is being used, with subnets of 62 hosts each. That commits the final 6 bits to host address in the fourth octet. If you need 16 or fewer areas, you might allocate 4 *a* bits for area number, which leaves 6 *s* bits for subnet. That would be $2^6$ or 64 subnets per area, which is many.

### Example: Bit Splitting for Area 1

This example illustrates how the bit-splitting approach would support the addresses in OSPF area 1. Writing 1 as four binary bits substitutes 0001 for the *a* bits. The area 1 addresses would be those with the bit pattern 172.16.0001 ssss.sshh hhhh. This bit pattern in the third octet supports decimal numbers 16 to 31. Addresses in the range 172.16.16.0 to 172.16.31.255 would fall into area 1. If you repeat this logic, area 0 would have addresses 172.16.0.0 to 172.16.15.255, and area 2 would have addresses 172.16.32.0 to 172.16.47.255.

Subnets would consist of an appropriate third octet value for the area they are in, together with addresses in the range 0 to 63, 64 to 127, 128 to 191, or 192 to 255 in the last octet. Thus, 172.16.16.0/26, 172.16.16.64/26, 172.16.16.128/26, 172.16.16.192/26, and 172.16.17.0/26 would be the first five subnets in area 1.

One recommendation that preserves good summarization is to take the last subnet in each area and divide it up for use as /30 or /31 subnets for WAN link addressing.

Few people enjoy working in binary. Free or inexpensive subnet calculator tools can help. For those with skill writing Microsoft Excel spreadsheet formulas, you can install Excel Toolkit functions to help with decimal-to-binary or decimal-to-hexadecimal conversion. Then, build a spreadsheet that lists all area blocks, subnets, and address assignments.

## IPv6 Address Planning

Because the IPv6 address space is much larger than the IPv4 address space, addressing plans for IPv6 are in many ways simpler to create. Subnetting an IPv4 address range is always a balancing act between getting the right number of subnets, the right number of hosts per subnet, and grouping subnets in such a way that they are easily summarizable, while also leaving room for future growth. With IPv6, creating an address plan is more straightforward.

It is strongly recommended that all IPv6 subnets use a /64 prefix. With $2^{64}$ hosts per subnet, a /64 prefix allows more hosts on each single subnet than a single broadcast domain could physically support. There is some concern that using /64 prefixes for every link, even point-to-point and loopback interfaces, unnecessarily wastes large chunks of IPv6 address space. For this reason, some organizations prefer to use /126 prefixes for point-to-point links and /128 prefixes for loopback interfaces.

**Note** When using /126 prefixes, avoid overlap with router anycast and embedded RP addresses. Another consideration in the implementation of prefix lengths that are longer than /64 is that it could cause incompatibilities with future capabilities that assume a /64 prefix length for each subnet.

Using a /64 prefix for any subnet that contains end hosts removes any considerations about the number of hosts per subnet from the addressing plan. The second consideration

in IPv4 addressing plans is to determine the right number of subnets for each site. For IPv6, this consideration is much less problematic. Local Internet Registries (LIR) commonly assign a /48 prefix from their assigned address blocks to each customer site. With 64 bits being used for the host part of the address, this leaves 128 – 64 – 48 = 16 bits to number the subnets within the site. This translates to $2^{16}$ = 65,536 possible subnets per site, which should be sufficient for all but the largest sites. If a single /48 prefix is insufficient, additional /48 prefixes can be obtained from the LIR.

Effectively, the 16 bits that are available for subnet allocation can be used freely to implement summarizable address plans or role-based addressing.

**Note** For easy configuration and renumbering, the **ipv6 general-prefix** command can be used to define a base prefix in the configuration, such as the /48 prefix that is assigned by the LIR. You can then reference this prefix by name in the interface-level IPv6 address configuration commands.

## Bit Splitting for IPv6

The 16 bits that are available for subnetting can be split in many different ways. Like IPv4, the IPv6 address plan is an integral part of the overall network design and should be synchronized with other design choices that are made. In an existing network, consider mapping the IPv6 address scheme to known numbers, such as VLANs or IPv4 addresses. This mapping eases network management and troubleshooting tasks, because network operators can relate the structure of the IPv6 addresses to existing address structures.

The following are examples of IPv6 addressing schemes that split the 16 subnet bits in different ways to support different design requirements:

- **Split by area:** If the site is split into areas, such as OSPF areas, the address structure should reflect this to support summarization between the areas. For example, the first 4 of the 16 bits could be used to represent the area, while the VLAN is coded into the last 12 bits. This scheme can support $2^4$= 16 areas and $2^{12}$ = 4096 subnets per area. A small range of VLAN numbers should be set aside to support point-to-point links and loopback interfaces within the area.
- **IPv4 mapping:** If the current IPv4 address structure is based on network 10.0.0.0/8 and all subnets are using /24 or shorter prefixes, the middle 16 bits in the IPv4 address can be mapped to the IPv6 address. For example, if a subnet has IPv4 prefix 10.123.10.0/24, the middle two octets 123.10 can be converted to hexadecimal: 123 = 0x7B and 10 = 0x0A. If the LIR-assigned prefix is 2001:0DB8:1234::/48, appending the 16 bits that are derived from the IPv4 address yields 2001:0DB8:1234:7B0A::/64 as the IPv6 prefix for the subnet. This method is convenient because it establishes a one-to-one mapping between the well known IPv4 addresses and the new IPv6 addresses. However, to use this method, the IPv4 address scheme needs to meet certain conditions, such as not using more than 16 bits for subnetting.

- **Role-based addressing:** For easier access list and firewall rule definition, it can be useful to code roles (for example, voice, office data, and guest users) into the address scheme. For example, the first 4 bits could be used to represent the role, the next 4 bits to represent the area, and the final 8 bits to represent the VLAN. This results in $2^4 = 16$ different roles that can be defined, $2^4 = 16$ areas within the site, and $2^8 = 256$ VLANs per area and per role. Using the first 4 bits for area makes it extremely easy to configure access lists or firewall rules, because all subnets for a specific role fall within a /52 address block. Summarization is slightly less efficient than in a scheme that is purely based on areas. Instead of one summarized address block per area, there is now a summarized block per role.

The methods that are shown here are just examples. When creating an address plan as part of a network design, carefully consider other address or network elements to define an address plan that matches and supports these elements.

**Note** For further information about IPv6 address planning, see RFC 5375, *IPv6 Unicast Address Assignment Considerations*.

### Addressing for VPN Clients

Focusing some attention on IP addressing for VPN clients can also provide benefits. As role-based security is deployed, there is a need for different groupings of VPN clients. These might correspond to administrators, employees, different groups of contractors or consultants, external support organizations, guests, and so on. You can use different VPN groups for different VPN client pools.

Role-based access can be controlled via the group password mechanism for the Cisco VPN client. Each group can be assigned VPN endpoint addresses from a different pool.

Traffic from the user PC has a VPN endpoint address as its source address.

The different subnets or blocks of VPN endpoint addresses can then be used in ACLs to control access across the network to resources, as discussed earlier for NAC roles. If the pools are subnets of a summary address block, routing traffic back to clients can be done in a simple way.

### NAT in the Enterprise

NAT is a powerful tool for working with IP addresses. It has the potential for being very useful in the enterprise to allow private internal addressing to map to publicly assigned addresses at the Internet connection point. However, if it is overused, it can be harmful.

NAT and Port Address Translation (PAT) are common tools for firewalls. A common approach to supporting content load-balancing devices is to perform destination NAT. A recommended approach to supporting content load-balancing devices is to perform source NAT. As long as NAT is done in a controlled, disciplined fashion, it can be useful.

Avoid using internal NAT or PAT to map private-to-private addresses internally. Internal NAT can make network troubleshooting confusing and difficult. For example, it would be difficult to determine which network 10 in an organization a user is currently connected to.

Internal NAT or PAT is sometimes required for interconnection of networks after a corporate merger or acquisition. Many organizations are now using network 10.0.0.0 internally, resulting in a "two network 10.0.0.0" problem after a merger. This is a severely suboptimal situation and can make troubleshooting and documentation very difficult. Re-addressing should be planned as soon as possible. It is also a recommended practice to isolate any servers reached through content devices using source NAT or destination NAT. These servers are typically isolated because the packets with NAT addresses are not useful elsewhere in the network. NAT can also be utilized in the data center to support small out-of-band (OOB) management VLANs on devices that cannot route or define a default gateway for the management VLAN, thereby avoiding one management VLAN that spans the entire data center.

### NAT with External Partners

NAT also proves useful when a company or organization has more than a couple of external business partners. Some companies exchange dynamic routing information with external business partners. Exchanges require trust. The drawback to this approach is that a static route from a partner to your network might somehow get advertised back to you. This advertisement, if accepted, can result in part of your network becoming unreachable. One way to control this situation is to implement two-way filtering of routes to partners: Advertise only subnets that the partner needs to reach, and only accept routes to subnets or prefixes that your staff or servers need to reach at the partner.

Some organizations prefer to use static routing to reach partners in a tightly controlled way. The next hop is sometimes a virtual Hot Standby Router Protocol (HSRP) or Gateway Load Balancing Protocol (GLBP) address on a pair of routers controlled by the partner.

When the partner is huge, such as a large bank, static routing is too labor intensive. Importing thousands of external routes into the internal routing protocol for each of several large partners causes the routing table to become bloated.

Another approach is to terminate all routing from a partner at an edge router, preferably receiving only summary routes from the partner. NAT can then be used to change all partner addresses on traffic into a range of locally assigned addresses. Different NAT blocks are used for different partners. This approach converts a wide range of partner addresses into a tightly controlled set of addresses and simplifies troubleshooting. It can also avoid potential issues when multiple organizations are using the 10.0.0.0/8 network.

If the NAT blocks are chosen out of a larger block that can be summarized, a redistributed static route for the larger block easily makes all partners reachable on the enterprise network. Internal routing then have one route that in effect says "this way to partner networks."

A partner block approach to NAT supports faster internal routing convergence by keeping partner subnets out of the enterprise routing table.

A disadvantage to this approach is that it is more difficult to trace the source of IP packets. However, if it is required, you can backtrack and get the source information through the NAT table.

# Design Considerations for IPv6 in Campus Networks

This section discusses the three different IPv6 deployment models that can be used in the enterprise campus.

## IPv6 Campus Design Considerations

As mentioned earlier, three major deployment models can be used to implement IPv6 support in the enterprise campus environment: the dual-stack model, the hybrid model, and the service block model. The choice of deployment model strongly depends on whether IPv6 switching in hardware is supported in the different areas of the network.

Dual stack is the preferred, most versatile, and highest-performance way to deploy IPv6 in existing IPv4 environments. IPv6 can be enabled wherever IPv4 is commissioned along with the associated features that are required to make IPv6 routable, highly available, and secure. In some cases, IPv6 may not be enabled on a specific interface or device because of the presence of legacy applications or hosts for which IPv6 is not supported. Inversely, IPv6 may be enabled on interfaces and devices for which IPv4 support is no longer necessary.

A key requirement for the deployment of the dual-stack model is that IPv6 switching must be performed in hardware on all switches in the campus. If some areas of the campus network do not support IPv6 switching in hardware, tunneling mechanisms are leveraged to integrate these areas into the IPv6 network. The hybrid model combines a dual-stack approach for IPv6-capable areas of the network with tunneling mechanisms such as Intra-Site Automatic Tunnel Addressing Protocol (ISATAP) and manual IPv6 tunnels where needed.

The hybrid model adapts as much as possible to the characteristics of the existing network infrastructure. Transition mechanisms are selected based on multiple criteria, such as IPv6 hardware capabilities of the network elements, number of hosts, types of applications, location of IPv6 services, and network infrastructure feature support for various transition mechanisms.

The service block model uses a different approach to IPv6 deployment. It centralizes IPv6 as a service, similar to how other services such as voice or guest access can be provided at a central location. The service block model is unique in that it can be deployed as an overlay network without any impact to the existing IPv4 network, and it is completely centralized. This overlay network can be implemented rapidly while allowing for high availability of IPv6 services, QoS capabilities, and restriction of access to IPv6

resources with little or no changes to the existing IPv4 network. As the existing campus network becomes IPv6-capable, the service block model can become decentralized. Connections into the service block are changed from tunnels (ISATAP or manually configured) to dual-stack connections. When all the campus layers are dual-stack capable, the service block can be dismantled and repurposed for other uses.

These three models are not exclusive. Elements from each of these models can be combined to support specific network requirements.

Ultimately, a dual-stack deployment is preferred. The hybrid and service block models are transitory solutions. The models can be leveraged to migrate to a dual stack design in a graceful manner, without a need for forced hardware upgrades throughout the entire campus. From an address-planning standpoint, this means that the IPv6 address plan should be designed to support a complete dual-stack design in the future.

### Dual-Stack Model

The dual-stack model deploys IPv4 and IPv6 in parallel without any tunneling or translation between the two protocols. IPv6 is enabled in the access, distribution, and core layers of the campus network. This model makes IPv6 simple to deploy, and is very scalable. No dependencies exist between the IPv4 and IPv6 design, which results in easier implementation and troubleshooting.

Deploying IPv6 in the campus using the dual-stack model offers several advantages over the hybrid and service block models. The primary advantage of the dual-stack model is that it does not require tunneling within the campus network. The dual-stack model runs the two protocols as "ships in the night," meaning that IPv4 and IPv6 run alongside one another and have no dependency on each other to function except that they share network resources. Both IPv4 and IPv6 have independent routing, high availability, QoS, security, and multicast policies. The dual-stack model also offers processing performance advantages, because packets are natively forwarded without having to account for additional encapsulation and lookup overhead.

These advantages make the dual-stack model the preferred deployment model. The stack model requires all switches in the campus to support IPv6 forwarding.

### Hybrid Model

The hybrid model strategy is to employ two or more independent transition mechanisms with the same deployment design goals. Flexibility is the key aspect of the hybrid approach. Any combination of transition mechanisms can be leveraged to best fit a given network environment. The hybrid model uses dual stack in all areas of the network where the equipment supports IPv6. Tunneling mechanisms are deployed for areas that do not currently support IPv6 in hardware. These areas can be transitioned to dual stack as hardware is upgraded later.

Various tunneling mechanisms and deployment scenarios can be part of a hybrid model deployment. This section highlights two common scenarios.

The first scenario that may require the use of a hybrid model is when the campus core is not enabled for IPv6. Common reasons why the core layer might not be enabled for IPv6 are either that the core layer does not have hardware-based IPv6 support at all, or has limited IPv6 support but with low performance.

In this scenario, manually configured tunnels are used exclusively from the distribution to aggregation layers. Two tunnels from each switch are used for redundancy and load balancing. From an IPv6 perspective, the tunnels can be viewed as virtual links between the distribution and aggregation layer switches. On the tunnels, routing and IPv6 multicast are configured in the same manner as with a dual-stack configuration.

The scalability of this model is limited, and a dual-stack model is preferred. However, this is a good model to use if the campus core is being upgraded or has plans to be upgraded, and access to IPv6 services is required before the completion of the core upgrade.

The second scenario focuses on the situation where hosts that are located in the campus access layer need to use IPv6 services, but the distribution layer is not IPv6 capable or enabled. The distribution layer switch is most commonly the first Layer 3 gateway for the access layer devices. If IPv6 capabilities are not present in the existing distribution layer switches, the hosts cannot gain access to IPv6 addressing router information (stateless autoconfiguration or Dynamic Host Configuration Protocol [DHCP] for IPv6), and then cannot access the rest of the IPv6-enabled network.

In this scenario, tunneling can be used on the IPv6-enabled hosts to provide access to IPv6 services that are located beyond the distribution layer. For example, the ISATAP tunneling mechanisms on the hosts in the access layer to provide IPv6 addressing and off-link routing. The Microsoft Windows XP and Vista hosts in the access layer must have IPv6 enabled and either a static ISATAP router definition or Domain Name System (DNS) A record entry that is configured for the ISATAP router address.

Using the ISATAP IPv4 address, the hosts establish tunnels to the IPv6-enabled core routers, obtain IPv6 addresses, and tunnel IPv6 traffic across the IPv4 distribution switches to the IPv6 enabled part of the network.

Terminating ISATAP tunnels in the core layer makes the layer appear as an access layer to the IPv6 traffic, which may be undesirable from a high-level design perspective. To avoid the blending of core and access layer functions, the ISATAP can be terminated on a different set of switches, such as the data center aggregation switches.

The main reason to choose the hybrid deployment model is to deploy IPv6 without having to go through an immediate hardware upgrade for parts of the network. It allows switches that have not reached the end of their normal life cycle to remain deployed and avoids the added cost that is associated with upgrading equipment before its time with the sole purpose of enabling IPv6.

Some drawbacks apply to the hybrid model. The use of ISATAP tunnels is not compatible with IPv6 multicast. Therefore, any access or distribution layer blocks that require the use of IPv6 multicast applications should be deployed using the dual-stack model. Manual tunnels support IPv6 multicast and can still be used to carry IPv6 across an IPv4 core. Another drawback of the hybrid model is the added complexity that is associated with

tunneling. Considerations that must be accounted for include performance, management, security, scalability, and availability.

### Service Block Model

The service block model has several similarities to the hybrid model. The underlying IPv4 network is used as the foundation for the overlay IPv6 network that is being deployed. ISATAP provides access to hosts in the access layer. Manually configured tunnels are utilized from the data center aggregation layer to provide IPv6 access to the applications and services that are located in the data center access layer. IPv4 routing is configured between the core layer and service block switches to allow visibility to the service block switches for terminating IPv6-in-IPv4 tunnels.

The biggest difference with the hybrid model is that the service block model centralizes IPv6 connectivity through a separate redundant pair of switches. The service block deployment model is based on a redundant pair of Cisco Catalyst 6500 series switches with a Cisco Supervisor Engine 32 or Supervisor Engine 720 card. The key to maintaining a highly scalable and redundant configuration in the service block is to ensure that a high-performance switch, supervisor, and modules are used to manage the load of the ISATAP, manually configured tunnels, and dual-stack connections for an entire campus network.

**Note** As the number of tunnels and required throughput increases, it might be necessary to distribute the load across an additional pair of switches in the service block.

The biggest benefit of this model compared with the hybrid model is that the centralized approach enables you to pace the IPv6 deployment in a very controlled manner.

In essence, the service block model provides control over the pace of IPv6 service introduction by leveraging the following:

- Per-user or per-VLAN tunnels, or both, can be configured via ISATAP to control the flow of connections and allow for the measurement of IPv6 traffic use.
- Access on a per-server or per-application basis can be controlled via access lists and routing policies that are implemented on the service block switches. This level of control allows for access to one, a few, or even many IPv6-enabled services, while all other services remain on IPv4 until those services can be upgraded or replaced. This enables a “per-service” deployment of IPv6.
- The use of separate dual redundant switches in the service block allows for high availability of ISATAP and manually configured tunnels as well as all dual-stack connections.
- Flexible options allow hosts access to the IPv6-enabled ISP connections, either by allowing a segregated IPv6 connection that is used only for IPv6-based Internet traffic or by providing links to the existing Internet edge connections that have both IPv4 and IPv6 ISP connections.

- Implementation of the service block model does not disrupt the existing network infrastructure and services. Because of its similarity to the hybrid model, the service block model suffers from the same drawbacks that are associated with the use of tunneling. In addition to those drawbacks, there is the cost that is associated with the service block switches.

**Note** For a detailed discussion of the three IPv6 deployment models, implementation guidelines, and examples, see the *Deploying IPv6 in Campus Networks* document at www.cisco.com/en/US/docs/solutions/Enterprise/Campus/CampIPv6.html.

# Designing Advanced Routing

This section discusses elements of advanced routing solution design using route summarization and default routing. It also discusses utilizing route filtering and redistribution in advanced routing designs. The discussion in this section

- Describes why route summarization and default routing should be used in a routing design
- Describes why route filtering should be used in a routing design
- Describes why redistribution should be used in a routing design

## Route Summarization and Default Routing

Route summarization procedures condense routing information. Without summarization, each router in a network must retain a route to every subnet in the network. With summarization, routers can reduce some sets of routes to a single advertisement, reducing both the load on the router and the perceived complexity of the network. The importance of route summarization increases with network size, as shown in Figure 3-1.

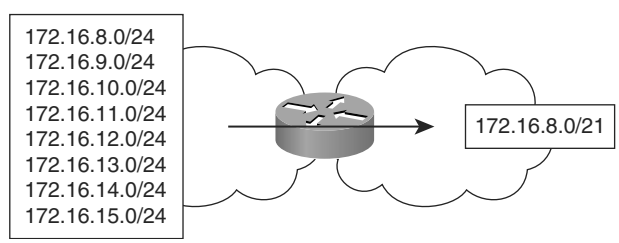



**Figure 3-1** *Route Summarization*

Medium-to-large networks often require the use of more routing protocol features than a small network. The larger the network, the more important it is to have a careful design with attention to properly scaling the routing protocol. Stability, control, predictability, and security of routing are also important. Converged networks are increasingly used to

support voice, IP telephony, storage, and other drop-sensitive traffic, and so networks must be designed for fast routing convergence.

Route summarization is one key network design element for supporting manageable and fast-converging routing. The Implementing Cisco IP Routing (ROUTE) course covers the configuration of route summarization and its benefits for routing and troubleshooting.

The design recommendations for summarizations are straightforward and include

- Using route summarization to scale routing designs.
- Designing addressing by using address blocks that can be summarized.
- Using default routing whenever possible. Route summarization is the ultimate route summarization, where all other routes are summarized in the default.

### Originating Default Routes

The concept of originating default routes is useful for summarization in routing. Most networks use some form of default routing. It is wise to have the default route (0.0.0.0 /0) advertised dynamically into the rest of the network by the router or routers that connect to Internet service providers (ISP). This route advertises the path to any route not found; more specifically in the routing table, as shown in Figure 3-2.

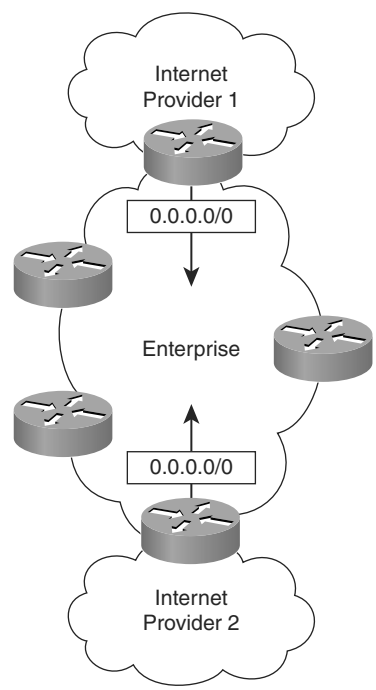



**Figure 3-2** *Originating Default Routes*

It is generally a bad idea to configure a static default route on every router, even if recursive routing is used. In recursive routing, for any route in the routing table whose next-hop IP address is not a directly connected interface of the router, the routing algorithm looks recursively into the routing table until it finds a directly connected interface to which it can forward the packets. If you configure a static default route on every router to the ISP router, the next hop is the ISP-connected router rather than a directly connected peer router. This approach can lead to black holes in the network if there is not a path to the ISP-connected router. This approach also needs to be reconfigured on every router if the exit point changes or if a second ISP connection is added.

If manually configured next hops are used, more configuration commands are needed. This approach can also lead to routing loops and is hard to change. If there are alternative paths, this static approach might fail to take advantage of them.

The recommended alternative is to configure each ISP-connected router with a static default route and redistribute it into the dynamic routing protocol. Static default route configuration needs to be done only at the network edge devices. All other routers pick up the route dynamically, and traffic out of the enterprise uses the closest exit. If the ISP-connected router loses connectivity to the ISP or fails, the default route is no longer advertised in the organization.

You might need to use the **default-information originate** command, with options, to redistribute the default route into the dynamic routing protocol.

**Note** The actual syntax of the command to inject a default route into an Interior Gateway Protocol (IGP) depends on the IGP being used. The command in the text works for RIP, OSPF, Intermediate System-to-Intermediate System (IS-IS), and BGP. For EIGRP, the **ip default-network** command is used. See the Cisco IP Command Reference for more in-depth study.

### Stub Areas and Default Route

Explicit route summarization is not the only way to achieve the benefits of summarization. The various kinds of OSPF stub areas can be thought of as a simpler form of summarization. The point of using OSPF stub areas, totally stubby areas, and not-so-stubby areas (NSSA) is to reduce the amount of routing information advertised into an area. The information that is suppressed is replaced by the default route 0.0.0.0/0 (IPv4) or ::/0 (IPv6)

OSPF cannot filter prefixes within an area. It only filters routes as they are passed between areas at an Area Border Router (ABR).

OSPF stub areas do not work to IP Security (IPsec) virtual private network (VPN) sites such as with generic routing encapsulation (GRE) over IPsec tunnels. For IPsec VPN remote sites, the 0/0 route must point to the ISP, so stub areas cannot be used. An alternative to the default route is to advertise a summary route for the organization as a "corporate default" route and filter unnecessary prefixes at the ABR. Because OSPF cannot

filter routes within an area, there still will be within-area flooding of link-state advertisements (LSA).

You can use this approach with the EIGRP, too. The **ip default-network** *network-number* command is used to configure the last-resort gateway or default route. A router configured with this command considers the network listed in the command as the candidate route for computing the gateway of last resort. This network must be in the routing table either as a static route or an IGP route before the router will announce the network as a candidate default route to other EIGRP routers. The network must be an EIGRP-derived network in the routing table or be generated by a static route that has been redistributed into EIGRP.

EIGRP networks typically configure the default route at ISP connection points. Filters can then be used so that only the default and any other critical prefixes are sent to remote sites. In many WAN designs with central Internet access, HQ just needs to advertise default to branch offices, in effect "this way to the rest of the network and to the Internet." If the offices have direct Internet access, a corporate summary can work similarly, "this way to the rest of the company."

In a site-to-site IPsec VPN network, it can be useful to also advertise a corporate summary route or corporate default route (which might be 10.0.0.0 /8) to remote sites. The advantage of doing so is that all other corporate prefixes need not be advertised to the IPsec VPN site. Even if the IPsec network uses two or three hub sites, dynamic failover occurs based on the corporate default. For the corporate default advertisement to work properly under failure conditions, all the site-specific prefixes need to be advertised between the hub sites.

Filtering the unnecessary routes out can save on the bandwidth and router CPU that is expended to provide routing information to remote sites. This increases the stability and efficiency of the network. Removing the clutter from routing tables also makes troubleshooting more effective and speeds convergence.

## Route Filtering in the Network Design

This section discusses the appropriate use of route filtering in network design. Route filtering can be used to manage traffic flows in the network, avoid inappropriate transit traffic through remote nodes, and provide a defense against inaccurate or inappropriate routing updates. You can use different techniques to apply route filtering in various routing protocols.

### Inappropriate Transit Traffic

Transit traffic is external traffic passing through a network or site, as shown in Figure 3-3.

With poorly configured topology, poorly configured filtering, or poorly configured summarization, a part of the network can be used suboptimally for transit traffic.

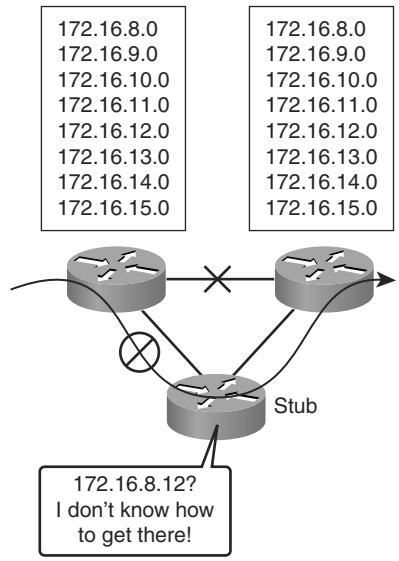



**Figure 3-3** *Avoid Inappropriate Transit Traffic*

Remote sites generally are connected with lower bandwidth than is present in the network core. Remote sites are rarely desirable as transit networks to forward network from one place to another. Remote sites typically cannot handle the traffic volume needed to be a viable routing alternative to the core network. In general, when core connectivity fails, routing should not detour via a remote site.

In OSPF, there is little control over intra-area traffic. LSAs cannot be filtered within an area. OSPF does not allow traffic to arbitrarily route into and then out of an area. The exception is area 0, which can be used for transit when another area becomes discontiguous.

With EIGRP, it can be desirable to configure EIGRP stub networks. This informs central routers that they should not use a remote site as a transit network. In addition, the use of stub networks damps unnecessary EIGRP queries, speeding network convergence. Filtering can help manage which parts of the network are available for transit in an EIGRP network.

With BGP, the most common concern about transit traffic is when a site has two Internet connections. If there is no filtering, the connections advertise routes. This advertisement can put the site at risk of becoming a transit network. This should not be a problem with two connections to the same ISP, because the autonomous system number is present in the BGP autonomous system path. Based on the autonomous system path, the ISP router ignores any routes advertised from the ISP to the site and then back to the ISP.

When two ISPs are involved, the site might inadvertently become a transit site. The best approach is to filter routes advertised outbound to the ISPs and ensure that only the company or site prefixes are advertised outward. Tagging routes with a BGP community is an

easy way to do this. All inbound routes received from the ISP should be filtered, too, so that you accept only the routes the ISP should be sending you.

### Defensive Filtering

Route filtering can also be used defensively against inaccurate or inappropriate routing traffic. This is illustrated in Figure 3-4.

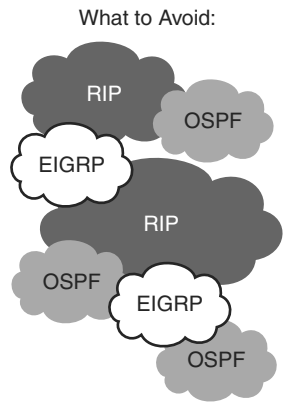



**Figure 3-4** *Defensive Filtering*

One common problem some organizations experience is that they inherit inappropriate routes from another organization, such as a business partner. Your business partner should not be advertising your routing prefixes back to your network. Those destinations are not reached through the partner, unless you have a very odd network design. The default route should not be reached via the partner, unless the partner is providing your network with Internet connectivity.

Inappropriate partner advertisements can disrupt routing without filtering. For example, a partner may define a static route to your data center. If this route leaks into your routing process, a portion of your network might think that the data center has moved to a location behind the router of the partner.

Defensive filtering protects the network from disruptions due to incorrect advertisements of others. You configure which routing updates your routers should accept from the partner and which routing updates should be ignored. For example, you would not accept routing updates about how to get to your own prefixes or about default routing.

For security reasons, you should advertise to a partner only the prefixes that you want them to be able to reach. This provides the partner with minimum information about your network and is part of a layered security approach. It also ensures that if there is an accidental leak of another partner's routes or static routes into the dynamic routing process, the inappropriate information does not also leak to others.

The approach of blocking route advertisements is also called route hiding or route starvation. Traffic cannot get to the hidden subnets from the partner unless a summary route is also present. Packet-filtering ACLs should also be used to supplement security by route starvation.

## Designing Redistribution

Redistribution is a powerful tool for manipulating and managing routing updates, particularly when two routing protocols are present in a network. This is shown in Figure 3-5.

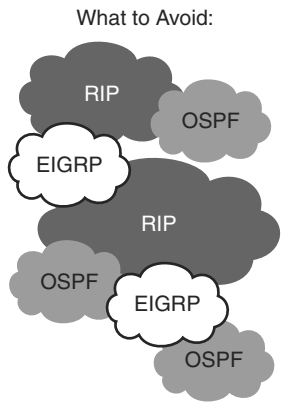



**Figure 3-5** *Designing Redistribution*

In some situations, routing redistribution is useful and even necessary. These include migration between routing protocols, corporate mergers, reorganization, and support for devices that speak only RIP or OSPF.

However, redistribution should be used with planning and some degree of caution. It is easy to create routing loops with redistribution. This is particularly true when there are multiple redistribution points, sometimes coupled with static routes, inconsistent routing summaries, or route filters.

Experience teaches that it is much better to have distinct pockets of routing protocols and redistribute than to have a random mix of routers and routing protocols with ad hoc redistribution. Therefore, running corporate EIGRP with redistribution into RIP or OSPF for a region that has routers from other vendors is viable, with due care. On the other hand, freely intermixing OSPF-speaking routers with EIGRP routers in ad hoc fashion is just asking for major problems.

When more than one interconnection point exists between two regions using different routing protocols, bidirectional redistribution is commonly considered. When running OSPF and EIGRP in two regions, it is attractive to redistribute OSPF into EIGRP, and EIGRP into OSPF.

### Filtered Redistribution

When you use bidirectional redistribution, you should prevent re-advertising information back into the routing protocol region or autonomous system that it originally came from. This is illustrated in Figure 3-6.

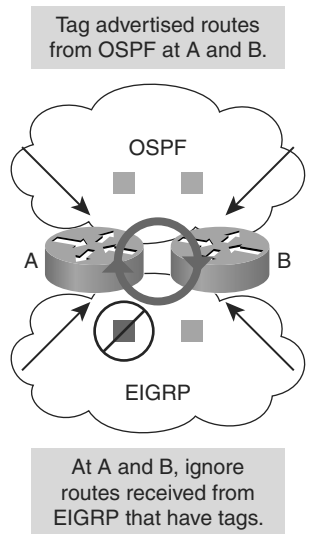



**Figure 3-6** *Filtered Redistribution*

For example, filters should be used so that OSPF information that was redistributed into EIGRP does not get re-advertised into OSPF. You also need to prevent information that came from EIGRP into OSPF from being re-advertised back into the EIGRP part of the network. This is sometimes called a manual split horizon. Split horizon is a routing protocol feature. The idea behind it is that it is counterproductive to advertise information back to the source of that information, because the information may be out of date or incorrect, and because the source of the information is presumed to be better informed.

If you do not do this filtering or use a manual split horizon, you will probably see strange convergence after an outage, you will probably see routing loops, and in general, you will experience routing problems and instability.

Both EIGRP and OSPF support the tagging of routes. A route map can be used to add the numeric tag to specific prefixes. The tag information is then passed along in routing updates. Another router may then filter out routes that match, or do not match, the tag. This is done using a route map in a distribution list.

One typical use of tags is with redistribution. In Figure 3-6, routers A and B can apply tags to routes from IGP X when they are advertised outbound into IGP Y. This in effect marks them as routes from IGP X. When routers A and B receive routes from Y, they would then filter out routes marked as from X when received from IGP Y, because both

routers learn such routes directly from IGP X. The process of filtering also applies in the opposite direction.

The point is to get routes in the most direct way, not via an indirect information path that might be passing along old information.

## Migrating Between Routing Protocols

This section discusses two common approaches for migrating between routing protocols. One approach for migrating between routing protocols is to use administrative distance (AD) to migrate the routing protocols. Another approach is to use redistribution and a moving boundary.

Migration by AD does not use redistribution. Instead, two routing protocols are run at the same time with the same routes. This assumes sufficient memory, CPU, and bandwidth are in place to support this on the routers running two routing protocols.

The first step in migration by AD is to turn on the new protocol, but make sure that it has a higher AD than the existing routing protocol so it is not preferred. This step enables the protocol and allows adjacencies or neighbors and routing databases to be checked but does not actually rely on the new routing protocol for routing decisions.

When the new protocol is fully deployed, various checks can be done with **show** commands to confirm proper deployment. Then, the cutover takes place. In cutover, the AD is shifted for one of the two protocols so that the new routing protocol will now have a lower AD.

Final steps in this process include the following:

- Check for any prefixes learned only via the old protocol.
- Check for any strange next hops (perhaps using some form of automated comparison).

With migration by redistribution, the migration is staged as a series of smaller steps. In each step, part of the network is converted from the old to the new routing protocol. In a big network, the AD approach might be used to support this conversion. In a smaller network, an overnight cutover or simpler approach might suffice.

To provide full connectivity during migration by redistribution, the boundary routers between the two parts of the network would have to bidirectionally redistribute between protocols. Filtering via tags would be one relatively simple way to manage this. The boundary routers move as more of the region is migrated.

## Designing Scalable EIGRP Designs

This section focuses on designing advanced routing solutions using Enhanced Interior Gateway Routing Protocol (EIGRP). It describes how to scale EIGRP designs and how to use multiple EIGRP autonomous systems in a large network.

## Scaling EIGRP Designs

EIGRP is tolerant of arbitrary topologies for small and medium networks. This is both a strength and a weakness. It is useful to be able to deploy EIGRP without restructuring the network. As the scale of the network increases, however, the risk of instability or long convergence times becomes greater. For example, if a network has reached the point where it includes 500 routers, EIGRP may stop working well without a structured hierarchy. As the size of the network increases, more stringent design is needed for EIGRP to work well.

**Note** This mechanism contrasts with OSPF, where structured design is imposed at an early stage. The counterpart to using EIGRP with an arbitrary topology would be an OSPF design that puts everything into OSPF area 0. That also may work for small-to-medium networks, up to around 200 or 300 OSPF routers.

To scale EIGRP, it is a good idea to use a structured hierarchical topology with route summarization.

One of the biggest stability and convergence issues with EIGRP is the propagation of EIGRP queries. When EIGRP does not have a feasible successor, it sends queries to its neighbors. The query tells the neighbor, "I do not have a route to this destination any more; do not route through me. Let me know if you hear of a viable alternative route." The router has to wait for replies to all the queries it sends. Queries can flood through many routers in a portion of the network and increase convergence time. Summarization points and filtered routes limit EIGRP query propagation and minimize convergence time.

Feasible distance is the best metric along a path to a destination network, including the metric to the neighbor advertising that path. Reported distance is the total metric along a path to a destination network as advertised by an upstream neighbor. A feasible successor is a path whose reported distance is less than the feasible distance (current best path).

### EIGRP Fast Convergence

Customers have been using EIGRP to achieve subsecond convergence for years. Lab testing by Cisco has shown that the key factor for EIGRP convergence is the presence or absence of a feasible successor. When there is no feasible successor, EIGRP uses queries to EIGRP peers and has to wait for responses. This slows convergence.

Proper network design is required for EIGRP to achieve fast convergence. Summarization helps limit the scope of EIGRP queries, indirectly speeding convergence. Summarization also shrinks the number of entries in the routing table, which speeds up various CPU operations. The effect of CPU operation on convergence is much less significant than the presence or absence of a feasible successor. A recommended way to ensure that a feasible successor is present is to use equal-cost routing.

EIGRP metrics can be tuned using the delay parameter. However, adjusting the delay on links consistently and tuning variance are next to impossible to do well at any scale.

In general, it is unwise to have a large number of EIGRP peers. Under worst-case conditions, router CPU or other limiting factors might delay routing protocol convergence. A somewhat conservative design is best to avoid nasty surprises.

## EIGRP Fast-Convergence Metrics

This section discusses EIGRP fast-convergence metrics. Cisco tested convergence of various routing protocols in the lab, as shown in Figure 3-7.

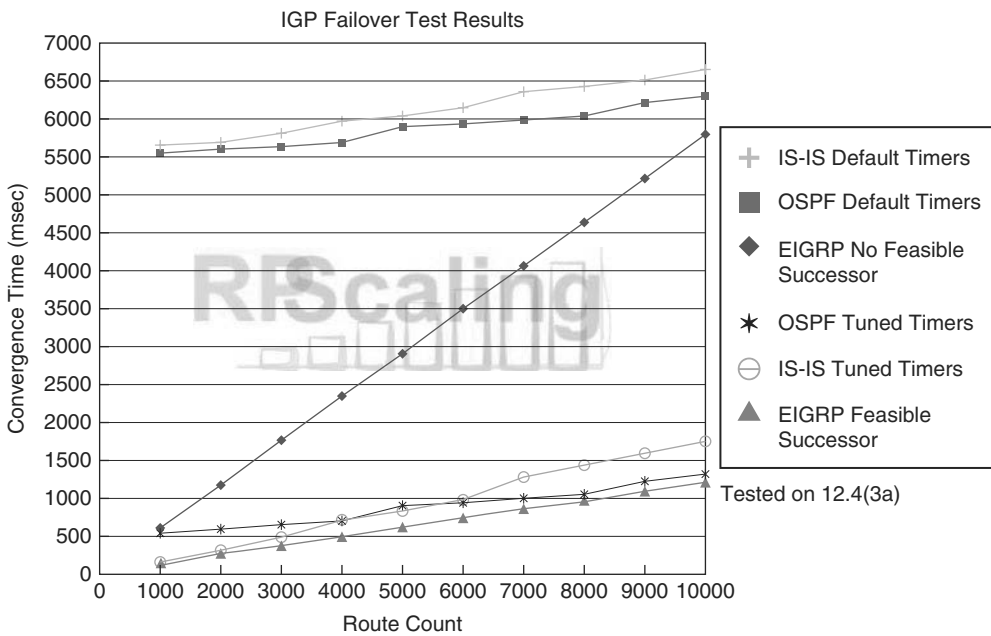



**Figure 3-7** *EIGRP Fast Convergence*

EIGRP convergence time increases as more routes need to be processed. However, there is a much bigger impact for networks without EIGRP feasible successors than for networks with no feasible successors.

With a feasible successor present, EIGRP converges in times ranging from about 1/10 second for 1000 routes to about 1.2 seconds for 10,000 routes. Without the feasible successor, convergence times increased to 1/2 to 1 second for 1000 routes and to about 6 seconds for 10,000 routes.

Subsecond timers are not available for EIGRP. One reason is that the hello timer is not the most significant factor in EIGRP convergence time. Another is that experimentation suggests that setting the EIGRP timer below two seconds can lead to instability. The recommended EIGRP minimum timer settings are two seconds for hellos and six seconds for the dead timer. Subsecond settings are not an option.

## Scaling EIGRP with Multiple Autonomous Systems

Implementing multiple EIGRP autonomous systems is sometimes used as a scaling technique. The usual rationale is to reduce the volume of EIGRP queries by limiting them to one EIGRP autonomous system. However, there can be issues with multiple EIGRP autonomous systems, as shown in Figure 3-8.

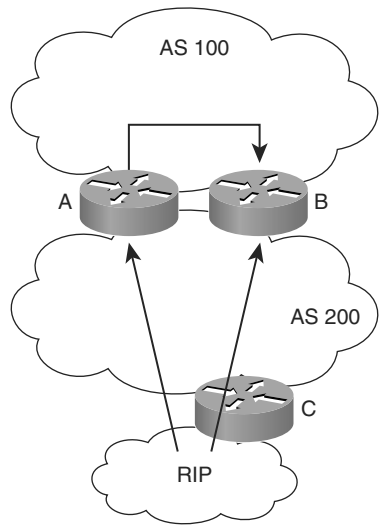



**Figure 3-8** *Scaling EIGRP with Multiple Autonomous Systems*

One potential issue is with the external route redistribution. In Figure 3-8, a route is redistributed from RIP into autonomous system 200. Router A redistributes it into autonomous system 100. Router B hears about the route prefix in advertisements from both autonomous system 200 and autonomous system 100. The AD is the same because the route is external to both autonomous systems.

The route that is installed into the EIGRP topology database first gets placed into the routing table.

### Example: External Route Redistribution Issue

If router B selects the route via autonomous system 100, it then routes to the RIP autonomous system indirectly, rather than directly via autonomous system 200, as illustrated in Figure 3-9.

Router B also advertises the route via autonomous system 100 back into autonomous system 200. Suppose B has a lower redistribution metric than router C does. If that is the case, A prefers the route learned from B over the route learned from C. In this case, A forwards traffic for this route to B in autonomous system 200, and B forwards traffic back to A in autonomous system 100. This is a routing loop!

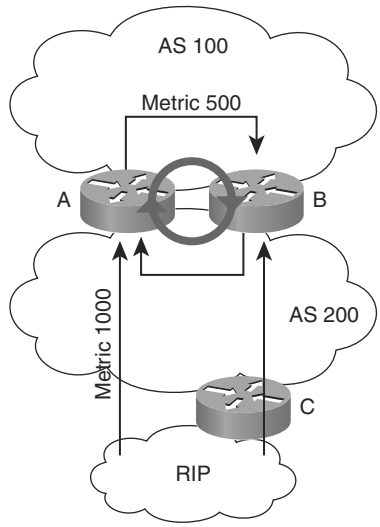



**Figure 3-9** *Example: External Route Redistribution Issue*

If two EIGRP processes run and two equal paths are learned, one by each EIGRP process, both routes do not get installed. The router installs the route that was learned through the EIGRP process with the lower autonomous system number. In Cisco IOS Software Releases earlier than 12.2(7)T, the router installed the path with the latest time stamp received from either of the EIGRP processes. The change in behavior is tracked by Cisco bug ID CSCdm47037.

The same sort of behavior may be seen with redistribution between two routing protocols, especially for routes learned from the protocol with the lower AD.

### Filtering EIGRP Redistribution with Route Tags

Outbound route tags can be used to filter redistribution and support EIGRP scaling with multiple EIGRP autonomous systems, as shown in Figure 3-10.

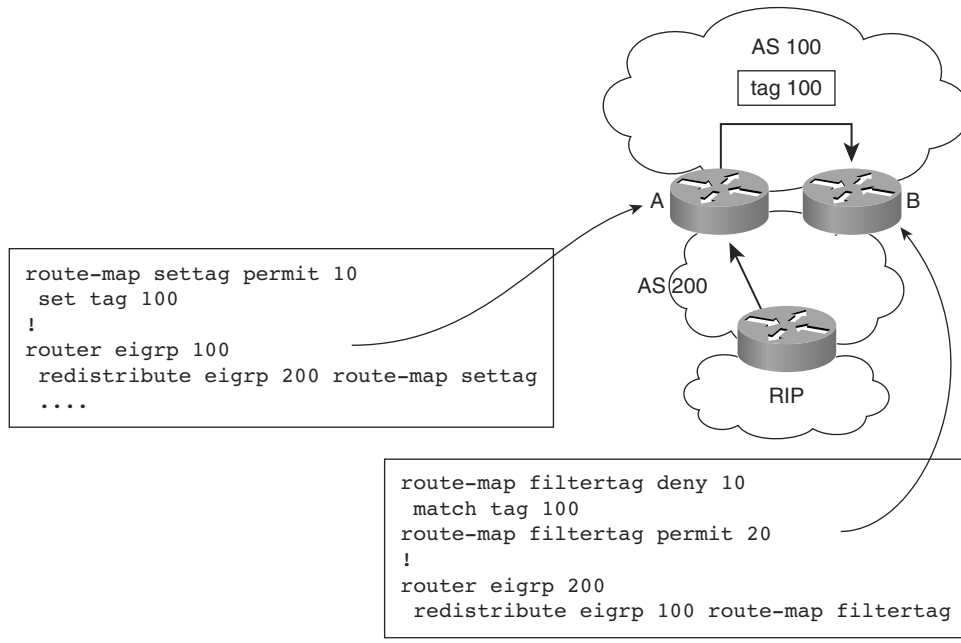



**Figure 3-10** *Filtering EIGRP Redistribution with Route Tags*

External routes can be configured to carry administrative tags. When the external route is redistributed into autonomous system 100 at router A or B, it can be tagged. This tag can then be used to filter the redistribution of the route back into autonomous system 200. This filtering blocks the formation of the loop, because router A will no longer receive the redistributed routes from router B through autonomous system 200.

In the configuration snippets, when routers A and B redistribute autonomous system 200 routes into autonomous system 100, they tag the routes with tag 100. Any routes tagged with tag 100 can then be prevented from being redistributed back into autonomous system 200. This successfully prevents a routing loop from forming.

### Filtering EIGRP Routing Updates with Inbound Route Tags

You can filter EIGRP routing updates with inbound route tags to support scaling with multiple EIGRP autonomous systems, as shown in Figure 3-11.

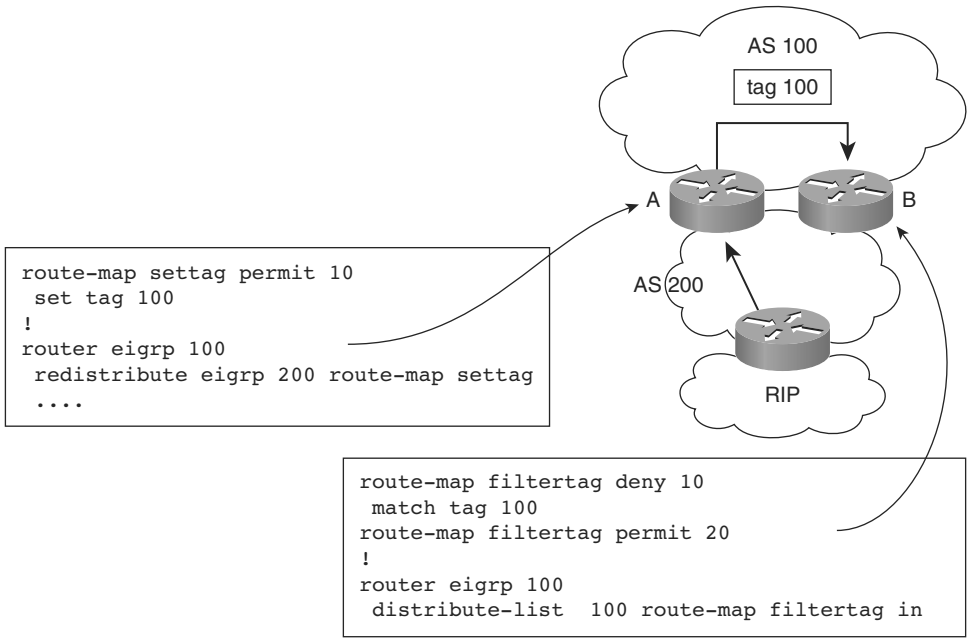



**Figure 3-11** *Filtering EIGRP Routing Updates with Inbound Route Tags*

Filtering outbound tags in the previous example does not prevent router B from learning the routes from autonomous system 100. Router B could still perform suboptimal routing by accepting the redistributed route learned from autonomous system 100.

The solution is to use inbound route tag filtering. This technique prevents routers from learning such routes, in which case they also will not be redistributed or advertised outbound. The Cisco bug fix CSCdt43016 provides support for incoming route filtering based on route maps. It allows for filtering routes based on any route map condition before acceptance into the local routing protocol database. This fix works for EIGRP and OSPF, starting with the Cisco IOS Software Releases 12.2T and 12.0S.

When routes are filtered to prevent router B from learning them, you prevent suboptimal routing by router B. The syntax shifts from using a route map with a **redistribute** command to using a route map with an inbound **distribute-list** command.

**Note** This example shows how filtering and administrative tags can help prevent routing loops with redistribution and suboptimal routing.

### Example: Queries with Multiple EIGRP Autonomous Systems

This example looks at the query behavior with multiple EIGRP autonomous systems. This is illustrated in Figure 3-12.

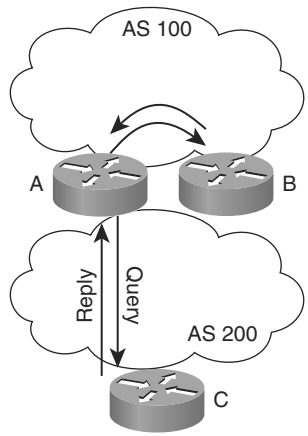



**Figure 3-12** *Example: Queries with Multiple EIGRP Autonomous Systems*

If router C sends an EIGRP query to router A, router A needs to query its neighbors. Router A sends a reply to router C, because it has no other neighbors in autonomous system 200. However, router A must also query all of its autonomous system 100 neighbors for the missing route. These routers may have to query their neighbors.

In this example, the query from router C is answered promptly by router A, but router A still needs to wait for the response to its query. Having multiple autonomous systems does not stop queries; it just delays them on the way.

**Note** The conclusion of this example is that using multiple EIGRP autonomous systems as an EIGRP query-limiting technique does not work.

What really stops a query is general scaling methods using summarization, distribution lists, and stubs.

## Reasons for Multiple EIGRP Autonomous Systems

There are several valid reasons for having multiple EIGRP autonomous systems, including the following:

- **Migration strategy after a merger or acquisition:** Although this is not a permanent solution, multiple autonomous systems are appropriate for merging two networks over time.

- **Different groups administer the different EIGRP autonomous systems:** This scenario adds complexity to the network design, but might be used for different domains of trust or administrative control.
- **Organizations with very large networks may use multiple EIGRP autonomous systems as a way to divide their networks:** Generally, this type of design approach uses summary routes at autonomous system boundaries to contain summary address blocks of prefixes in very large networks and to address the EIGRP query propagation issue.

These reasons for using multiple EIGRP autonomous systems can be appropriate, but pay careful attention to limiting queries.

# Designing Scalable OSPF Design

The ability to scale an OSPF internetwork depends on the overall network structure and addressing scheme. As outlined in the preceding sections about network topology and route summarization, adopting a hierarchical addressing environment and a structured address assignment are the most important factors in determining the scalability of your internetwork. Network scalability is affected by operational and technical considerations.

This section discusses designing advanced routing solutions using OSPF. It describes how to obtain scale OSPF designs and what factors can influence convergence in OSPF on a large network. The concepts covered are

- How to scale OSPF routing to a large network
- How to obtain fast convergence for OSPF in a routing design

## Factors Influencing OSPF Scalability

Scaling is determined by the utilization of three router resources: memory, CPU, and interface bandwidth. The workload that OSPF imposes on a router depends on these factors:

- **Number of adjacent neighbors for any one router:** OSPF floods all link-state changes to all routers in an area. Routers with many neighbors have the most work to do when link-state changes occur. In general, any one router should have no more than 60 neighbors.
- **Number of adjacent routers in an area:** OSPF uses a CPU-intensive algorithm. The number of calculations that must be performed given $n$ link-state packets is proportional to $n \log n$. As a result, the larger and more unstable the area, the greater the likelihood for performance problems associated with routing protocol recalculation. Generally, an area should have no more than 50 routers. Areas that suffer with unstable links should be smaller.
- **Number of areas supported by any one router:** A router must run the link-state algorithm for each link-state change that occurs for every area in which the router

resides. Every ABR is in at least two areas (the backbone and one adjacent area). In general, to maximize stability, one router should not be in more than three areas.

- **Designated router (DR) selection:** In general, the DR and backup designated router (BDR) on a multiaccess link (for example, Ethernet) have the most OSPF work to do. It is a good idea to select routers that are not already heavily loaded with CPU-intensive activities to be the DR and BDR. In addition, it is generally not a good idea to select the same router to be the DR on many multiaccess links simultaneously.

The first and most important decision when designing an OSPF network is to determine which routers and links are to be included in the backbone area and which are to be included in each adjacent area.

### Number of Adjacent Neighbors and DRs

One contribution to the OSPF workload on a router is the number of OSPF adjacent routers that it needs to communicate with.

Each OSPF adjacency represents another router whose resources are expended to support these activities:

- Exchanging hellos
- Synchronizing link-state databases
- Reliably flooding LSA changes
- Advertising the router and network LSA

Some design choices can reduce the impact of the OSPF adjacencies. Here are some recommendations:

- On LAN media, choose the most powerful routers or the router with the lightest load as the DR candidates. Set the priority of other routers to zero so they will not be DR candidates.
- When there are many branch or remote routers, spread the workload over enough peers. Practical experience suggests that IPsec VPN peers, for example, running OSPF over GRE tunnels are less stable than non-VPN peers. Volatility or amount of change and other workload need to be considered when determining how many peers a central hub router can support.

Any lab testing needs to consider typical operating conditions. Simultaneous restarts on all peers or flapping connections to all peers are the worst-case situations for OSPF.

### Routing Information in the Area and Domain

The workload also depends on the amount of routing information available within the area and the OSPF autonomous system. Routing information in OSPF depends on the number of routers and links to adjacent routers in an area.

There are techniques and tools to reduce this information. Stub and totally stubby areas import less information into an area about destinations outside the routing domain or the area then do normal areas. Therefore, using stub and totally stubby areas further reduces the workload on an OSPF router.

Interarea routes and costs are advertised into an area by each ABR. Totally stubby areas keep not only external routes but also this interarea information from having to be flooded into and within an area.

One way to think about Autonomous System Boundary Routers (ASBR) in OSPF is that each is in effect providing a distance vector-like list of destinations and costs. The more external prefixes and the more ASBRs there are, the more the workload for Type 5 or 7 LSAs. Stub areas keep all this information from having to be flooded within an area.

The conclusion is that area size and layout design, area types, route types, redistribution, and summarization all affect the size of the LSA database in an area.

## Designing OSPF Areas

Area design can be used to reduce routing information in an area. Area design requires considering your network topology and addressing. Ideally, the network topology and addressing should be designed initially with division of areas in mind. Whereas EIGRP will tolerate more arbitrary network topologies, OSPF requires a cleaner hierarchy with a more clear backbone and area topology.

Geographic and functional boundaries should be considered in determining OSPF area placement.

As discussed previously, to improve performance minimize the routing information advertised into and out of areas. Bear in mind that anything in the LSA database must be propagated to all routers within the area. With OSPF, note that all changes to the LSA database need to be propagated; this in turn consumes bandwidth and CPU for links and routers within the area. Rapid changes or flapping only exacerbate this effect because the routers have to repeatedly propagate changes. Stub areas, totally stubby areas, and summary routes not only reduce the size of the LSA database, but they also insulate the area from external changes.

Experience shows that you should be conservative about adding routers to the backbone area 0. The first time people configure an OSPF design, they end up with almost everything in area 0. Some organizations find that over time, too many routers ended up in area 0. A recommended practice is to put only the essential backbone and ABRs into area 0.

Some general advice about OSPF design is this:

- Keep it simple.
- Make nonbackbone areas stub areas (or totally stubby areas).
- Have the address space compressible.

## Area Size: How Many Routers in an Area?

Cisco experience suggests that the number of adjacent neighbors has more impact than the total number of routers in the area. In addition, the biggest consideration is the amount of information that has to be flooded within the area. Therefore, one network might have, for example, 200 WAN routers with one Fast Ethernet subnet in one area. Another might have fewer routers and more subnets.

It is a good idea to keep the OSPF router LSAs under the IP maximum transmission unit (MTU) size. When the MTU is exceeded, the result is IP fragmentation. IP fragmentation is, at best, a less-efficient way to transmit information and requires extra router processing. A large number of router LSAs also implies that there are many interfaces (and perhaps neighbors). This is an indirect indication that the area may have become too large. If the MTU size is exceeded, the command **ip ospf mtu ignore** must be used.

Stability and redundancy are the most important criteria for the backbone. Stability is increased by keeping the size of the backbone reasonable.

**Note** As best practice each area, including the backbone, should contain no more than 50 routers.

If link quality is high and the number of routes is small, the number of routers can be increased. Redundancy is important in the backbone to prevent partition when a link fails. Good backbones are designed so that no that single link failure can cause a partition.

Current ISP experience and Cisco testing suggest that it is unwise to have more than about 300 routers in OSPF backbone area 0, depending on all the other complexity factors that have been discussed. As mentioned in the preceding note, 50 or fewer routers is the most optimal design.

**Note** This number is intended as an appropriate indication that an OSPF design is unsatisfactory and should be reconsidered, focusing on a smaller area 0.

## OSPF Hierarchy

OSPF requires two levels of hierarchy in your network, as shown in Figure 3-13.

Route summarization is extremely desirable for a reliable and scalable OSPF network. Summarization in OSPF naturally fits at area boundaries, when there is a backbone area 0 and areas off the backbone, with one or a few routers interconnecting the other areas to area 0. If you want three levels of hierarchy for a large network, BGP can be used to interconnect different OSPF routing domains. With advanced care, two OSPF processes can be used, although it is not recommended for most networks due to complexity and the chance of inadvertent adjacencies.

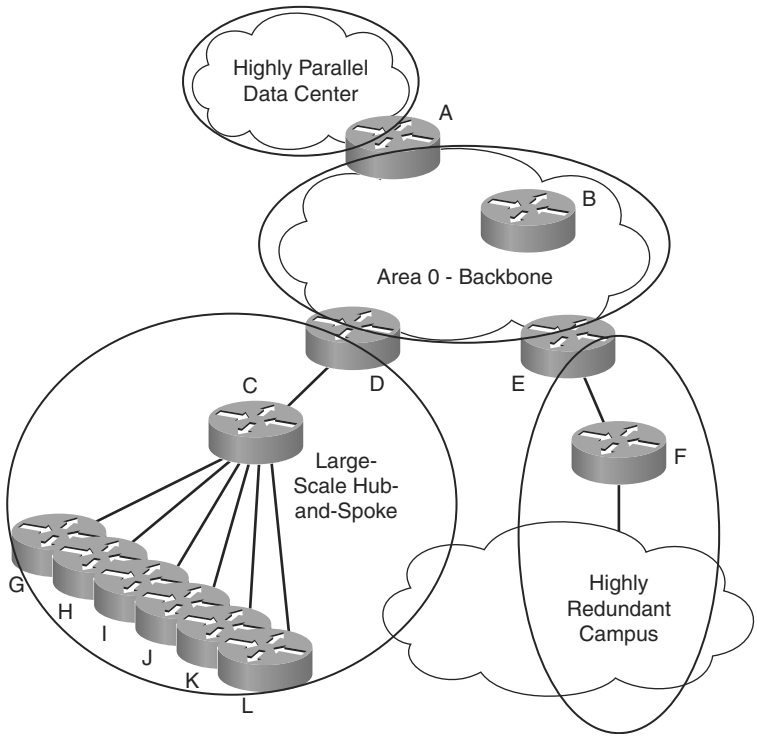



**Figure 3-13** *OSPF Hierarchy*

One difficult question in OSPF design is whether distribution or core routers should be ABRs. General design advice is to separate complexity from complexity and put complex parts of the network into separate areas. A part of the network might be considered complex when it has a lot of routing information, such as a full-mesh, a large hub-and-spoke, or a highly redundant topology such as a redundant campus or data center.

ABRs provide opportunities to support route summarization or create stub or totally stubby areas. A structured IP addressing scheme needs to align with the areas for effective route summarization. One of the simplest ways to allocate addresses in OSPF is to assign a separate network number for each area.

Stub areas cannot distinguish among ABRs for destinations external to the OSPF domain (redistributed routes). Unless the ABRs are geographically far apart, this should not matter. Totally stubby areas cannot distinguish one ABR from another, in terms of the best route to destinations outside the area. Unless the ABRs are geographically far apart, this should not matter.

## Area and Domain Summarization

There are many ways to summarize routes in OSPF. The effectiveness of route summarization mechanisms depends on the addressing scheme. Summarization should be supported into and out of areas at the ABR or ASBR. To minimize route information inserted into the area, consider the following guidelines when planning your OSPF internetwork:

- Configure the network addressing scheme so that the range of subnets assigned within an area is contiguous.
- Create an address space that will split areas easily as the network grows. If possible, assign subnets according to simple octet boundaries.
- Plan ahead for the addition of new routers to the OSPF environment. Ensure that new routers are inserted appropriately as area, backbone, or border routers.

Figure 3-14 shows some of the ways to summarize routes and otherwise reduce LSA database size and flooding in OSPF.

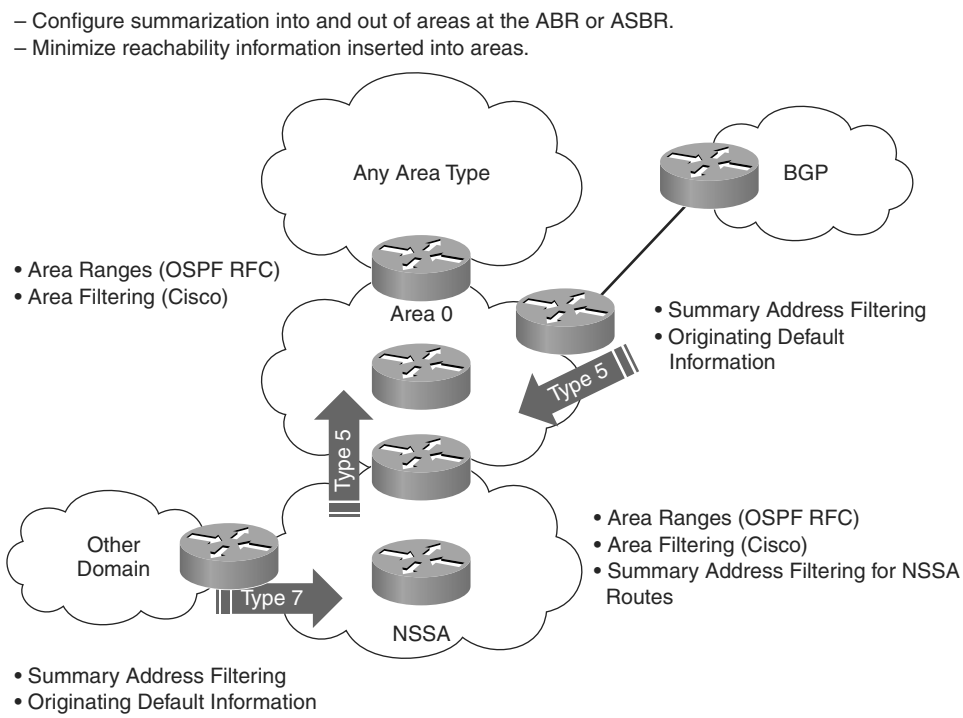



**Figure 3-14** *Area and Domain Summarization*

- **Area ranges per the OSPF RFCs:** The ability to inject only a subset of routing information back into area 0. This takes place only an ABR. It consolidates and summarizes routes at an area boundary.

- **Area filtering:** Filters prefixes advertised in type 3 LSAs between areas of an ABR.
- **Summary address filtering** Used on an ASBR to filtering on routes injected into OSPF by redistribution from other protocols.
- Originating default.
- Filtering for NSSA routes.

**Note** OSPF Version 2 (OSPFv2) for IP Version 4 (IPv4) and OSPF Version 3 (OSPFv3) for IP Version 6 (IPv6) are implemented as two entirely independent protocols. This independence means that theoretically the area structure and ABRs could be entirely different for each of these protocols. However, from a design standpoint, it is often best to align the area structure and ABRs for both protocols to reduce operational complexity and ease troubleshooting. This approach implies that the IPv6 and IPv4 address blocks that are assigned to the areas should also be aligned to support summarization for both protocols.

## OSPF Hub-and-Spoke Design

In an OSPF hub-and-spoke design, any change at one spoke site is passed up the link to the area hub and is then replicated to each of the other spoke sites. These actions can place a great burden on the hub router. Change flooding is the chief problem encountered in these designs.

Stub areas minimize the amount of information within the area. Totally stubby areas are better than stub areas in this regard. If a spoke site must redistribute routes into OSPF, make it a NSSA. Keep in mind that totally stubby NSSAs are also possible.

**Note** A Cisco proprietary extension to stub areas is what is called totally stubby areas. Cisco indicates this by adding a **no-summary** keyword to the stub area configuration. A totally stubby area is one that blocks external routes and summary routes (inter-area routes) from going into the area. This way, intra-area routes and the default of 0.0.0.0 are the only routes injected into that area.

Limiting the number of spokes per area reduces the flooding at the hub. However, smaller areas allow for less summarization into the backbone. Each spoke requires either a separate interface or a subinterface on the hub router.

## Number of Areas in an OSPF Hub-and-Spoke Design

For a hub-and-spoke topology, the number of areas and the number of sites per area need to be determined, as shown in Figure 3-15.

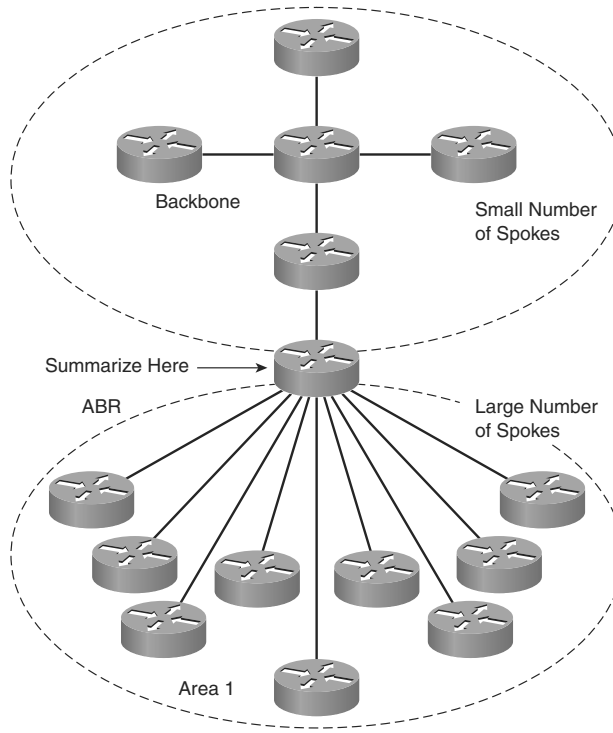



**Figure 3-15** *Number of Areas in a Hub-and-Spoke Design*

As the number of remote sites goes up, you have to start breaking the network into multiple areas. As already noted, the number of routers per area depends on a couple of factors. If the number of remote sites is low, you can place the hub and its spokes within an area. If there are multiple remote sites, you can make the hub an ABR and split off the spokes in one or more areas.

In general, the hub should be an ABR, to allow each area to be summarized into the other areas.

The backbone area is extremely important in OSPF. The best approach is to design OSPF to have a small and highly stable area 0. For example, some large Frame Relay or ATM designs have had an area 0 consisting of just the ABRs, all within a couple of racks.

### Issues with Hub-and-Spoke Design

Low-speed links and large numbers of spoke sites are the worst issues for hub-and-spoke design, as illustrated in Figure 3-16.

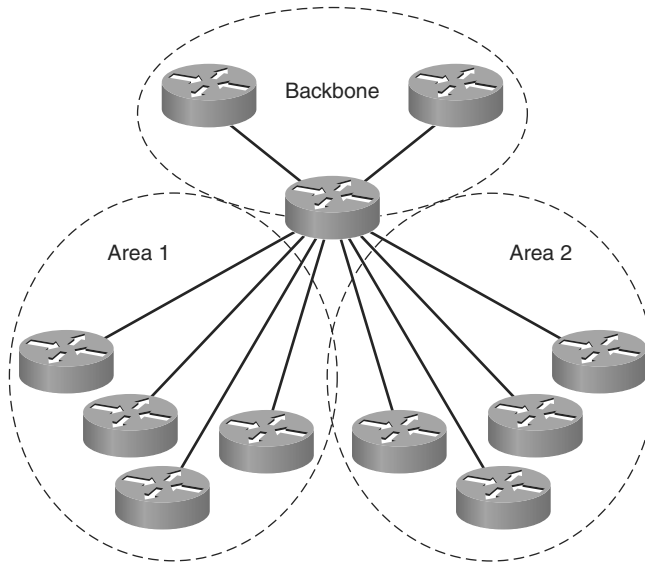



**Figure 3-16** *Issues with Hub-and-Spoke Design*

Low-speed links and large numbers of spokes may require multiple flooding domains or areas, which you must effectively support. You should balance the number of flooding domains on the hub against the number of spokes in each flooding domain. The link speeds and the amount of information being passed through the network determine the right balance.

Design for these situations must balance

- The number of areas
- The router impact of maintaining an LSA database and doing Dijkstra calculations per area
- The number of remote routers in each area

In situations with low bandwidth, the lack of bandwidth to flood LSAs when changes are occurring or OSPF is initializing becomes a driving factor. The number of routers per area must be strictly limited so that the bandwidth is adequate for LSA flooding under stress conditions (for example, simultaneous router startup or linkup conditions).

The extreme case of low-bandwidth links might be 9600-bps links. Areas for a network would consist of, at most, a couple of sites. In this case, another approach to routing might be appropriate. For example, use static routes from the hub out to the spokes, with default routes back to the hub. Flooding reduction, as discussed in the "OSPF Flooding Reduction" section later in this chapter, might help but would not improve bandwidth usage in a worst-case situation. The recommendation for this type of setting is lab testing under worst-case conditions to define the bandwidth requirements.

### OSPF Hub-and-Spoke Network Types

When using OSPF for hub-and-spoke networks, over nonbroadcast multiaccess access (that is, Frame Relay or ATM), you have several choices for the type of network you use. Figure 3-17 shows the details.

| Network Type | Advantages | Disadvantages |
|---|---|---|
| Single Interface at the Hub Treated as an OSPF Broadcast or NBMA Network | • Single IP Subnet<br>• Fewer Host Routes in Routing Table | • Manual Configuration of Each Spoke With the Correct OSPF Priority for DR/BDR<br>• No Reachability Between Spokes or Labor-Intensive Layer 2 Configuration |
| Single Interface at the Hub Treated as an OSPF Point-to-Multipoint Network<br>`ip ospf network-type point-to-multipoint` | • Single IP Subnet<br>• No Configuration Per Spoke<br>• Most Natural Solution | • Additional Host Routes Inserted in the Routing Table<br>• Longer Hello and Dead Timer Intervals |
| Individual Point-to-Point Interface at the Hub for Each Spoke<br>`ip ospf network-type point-to-point` | • Can Take Advantage of End-to-End Signaling for Down State<br>• Shorter Hello and Dead Timer Intervals | • Lost IP Address Space<br>• More Routes in the Routing Table<br>• Overhead of Subinterfaces |

Recommendation: Point-to-point or point-to-multipoint with hub-and-spoke.

**Figure 3-17** *OSPF Hub-and-Spoke Network Types*

You must use the right combination of network types for OSPF hub and spoke to work well. Generally, it is wisest to use either the point-to-multipoint OSPF network type at the hub site or configure the hub site with point-to-point subinterfaces.

Configuring point-to-multipoint is simple. The disadvantage of a point-to-multipoint design is that additional host routes are added to the routing table, and the default OPSF hello and dead-timer interval is longer. However, point-to-multipoint implementations simplify configuration as compared to broadcast or nonbroadcast multiaccess (NBMA) implementations and conserve IP address space as compared to point-to-point implementations.

Configuring point-to-point subinterfaces initially takes more work, perhaps on the order of a few hours. Each subinterface adds a route to the routing table, making this option about equal to point-to-multipoint in terms of routing table impact. More address space gets used up, even with /30 or /31 subnetting for the point-to-point links. On the other hand, after configuration, point-to-point subinterfaces may provide the most stability, with everything including management working well in this environment.

The broadcast or NBMA network types are best avoided. Although they can be made to work with some configuration effort, they lead to less stable networks or networks where certain failure modes have odd consequences.

## OSPF Area Border Connection Behavior

OSPF has strict rules for routing. They sometimes cause nonintuitive traffic patterns.

In Figure 3-18, dual-homed connections in hub-and-spoke networks illustrate a design challenge in OSPF, where connections are parallel to an area border. Traffic crossing the backbone must get into an area by the shortest path and then stay in that area.

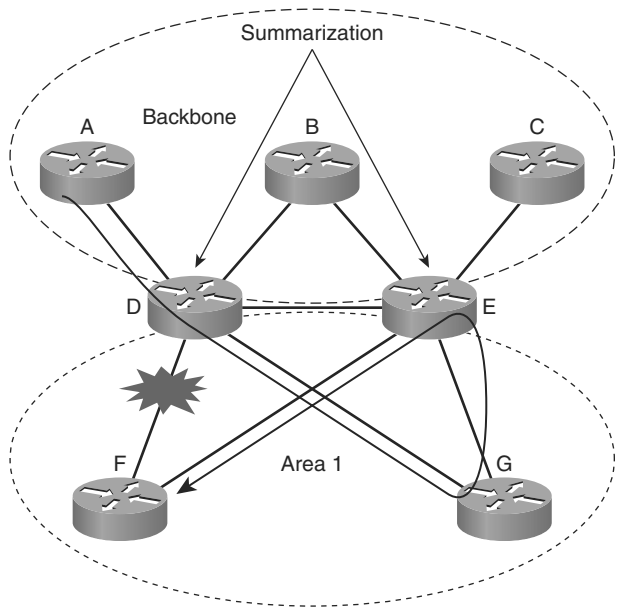



**Figure 3-18** *OSPF Area Border Connection Behavior*

In this example, the link from D to E is in area 0. If the D-to-F link fails, traffic from D to F goes from D to G to E to F. Because D is an ABR for area 1, the traffic to F is all internal to area 1 and must remain in area 1. OSPF does not support traffic going from D to E and then to F because the D-to-E link is in area 0, not in area 1. A similar scenario applies for traffic from A to F: It must get into area 1 by the shortest path through D and then stay in area 1.

In OSPF, traffic from area 1 to area 1 must stay in area 1 unless area 1 is partitioned, in which case the backbone area 0 can be used. Traffic from area 1 to area 2 must go from area 1 to area 0, and then into area 2. It cannot go into and out of any of the areas in other sequences.

OSPF area border connections must be considered in a thorough OSPF design. One solution to the odd transit situation just discussed is to connect ABRs with physical or virtual links for each area that both ABRs belong to. You can connect the ABRs within each area by either of two means:

- Adding a real link between the ABRs inside area 1
- Adding a virtual link between the ABRs inside area 0

In general, the recommendation is to avoid virtual links when you have a good alternative. OSPF virtual links depend on area robustness and therefore are less reliable than a physical link. Virtual links add complexity and fragility; if an area has a problem, the virtual link through the area has a problem. Also, if you rely too much on virtual links, you can end up with a maze of virtual links and possibly miss some virtual connections.

If the ABRs are Layer 3 switches or have some form of Ethernet connections, VLANs can be used to provide connections within each area common to both ABRs. With multiple logical links, whether physical, subinterfaces, or VLANs between a pair of ABRs, the following options are recommended:

- Consider making sure that a link exists between the ABRs within each area on those ABRs.
- Implement one physical or logical link per area.

## Fast Convergence in OSPF

Network convergence is the time that is needed for the network to respond to events. It is the time that it takes for traffic to be rerouted onto an alternative path when node or link fails or onto a more optimal path when a new link or node appears. Traffic is not rerouted until the data plane data structures such as the Forwarding Information Base (FIB) and adjacency tables of all devices have been adjusted to reflect the new state of the network. For that to occur, all network devices must go through the following procedure:

1. **Detect the event:** Loss or addition of a link or neighbor needs to be detected. This can be done through a combination of Layer 1, Layer 2, and Layer 3 detection mechanisms, such as carrier detection, routing protocol hello timers, and Bidirectional Forwarding Detection (BFD).
2. **Propagate the event:** Routing protocol update mechanisms are used to forward the information about the topology change from neighbor to neighbor.
3. **Process the event:** The information needs to be entered into the appropriate routing protocol data structures and the routing algorithm needs to be invoked to calculate updated best paths for the new topology.
4. **Update forwarding data structures:** The results of the routing algorithm calculations need to be entered into the data plane packet forwarding data structures.

At this point, the network has converged. The rest of this section focuses on the second and third steps in this procedure, because these are most specific to OSPF and tuning the associated parameters can greatly improve OSPF convergence times. The first step is dependent on the type of failure and the combination of Layer 1, Layer 2, and Layer 3 protocols that are deployed. The fourth step is not routing protocol specific, but depends on the hardware platform and the mechanisms involved in programming the data plane data structures.

## Tuning OSPF Parameters

By default, OSPF LSA propagation is controlled by three parameters:

- **OSPF_LSA_DELAY_INTERVAL:** Controls the length of time that the router should wait before generating a type 1 router LSA or type 2 network LSA. By default, this parameter is set at 500 ms.
- **MinLSInterval:** Defines the minimum time between distinct originations of any particular LSA. The value of MinLSInterval is set to 5 seconds. This value is defined in appendix B of RFC 2328.
- **MinLSArrival:** The minimum time that must elapse between reception of new LSA instances during flooding for any particular LSA. LSA instances received at higher frequencies are discarded. The value of MinLSArrival is set to 1 second. This value is defined in Appendix B of RFC 2328.

### OSPF Exponential Backoff

The default OSPF LSA propagation timers are quite conservative. Lowering the values of the timers that control OSPF LSA generation can significantly improve OSPF convergence times. However, if the value for the timeout between the generation of successive iterations of an LSA is a fixed value, lowering the values could also lead to excessive LSA flooding.

This is why Cisco has implemented an exponential backoff algorithm for LSA generation. The initial backoff timers are low, but if successive events are generated for the same LSA, the backoff timers increase. Three configurable timers control the LSA pacing:

- **LSA-Start:** The initial delay to generate an LSA. This timer can be set at a very low value, such as 1 ms or even 0 ms. Setting this timer to a low value helps improve convergence because initial LSAs for new events are sent as quickly as possible.
- **LSA-Hold:** The minimum time to elapse before flooding an updated instance of an LSA. This value is used as an incremental value. Initially, the hold time between successive LSAs is set to be equal to this configured value. Each time a new version of an LSA is generated the hold time between LSAs is doubled, until the LSA-Max-Wait value is reached, at which point that value is used until the network stabilizes.
- **LSA-Max-Wait:** The maximum time that can elapse before flooding an updated instance of an LSA. Once the exponential backoff algorithm reaches this value, it stops increasing the hold time and uses the LSA-Max-Wait timer as a fixed interval between newly generated LSAs.

What the optimal values for these values depends on the network. Tuning the timers too aggressively could result in excessive CPU load during network reconvergence, especially when the network is unstable for a period. Lower the values gradually from their defaults and observe router behavior to determine what the optimal values are for your network.

When you adjust the OSPF LSA throttling timers, it might be necessary to adjust the MinLSArrival timer. Any LSAs that are received at a higher frequency than the value of this timer are discarded. To prevent routers from dropping valid LSAs, make sure that the MinLSArrival is configured to be lower or equal to the LSA-Hold timer.

**Note** To reduce the impact of link flapping, IP event dampening can be implemented. For more information about this feature, see https://www.cisco.com/en/US/docs/ios/iproute_pi/configuration/guide/iri_ip_event_damp_ps10591_TSD_Products_Configuration_Guide_Chapter.html.

Figure 3-19 illustrates the OSPF exponential backoff algorithm. It is assumed that, every second, an event happens that causes a new version of an LSA to be generated. With the default timers, the initial LSA is generated after 500 ms. After that, a five-second wait occurs between successive LSAs.

**Result of tuning OSPF LSA throttle timers:**

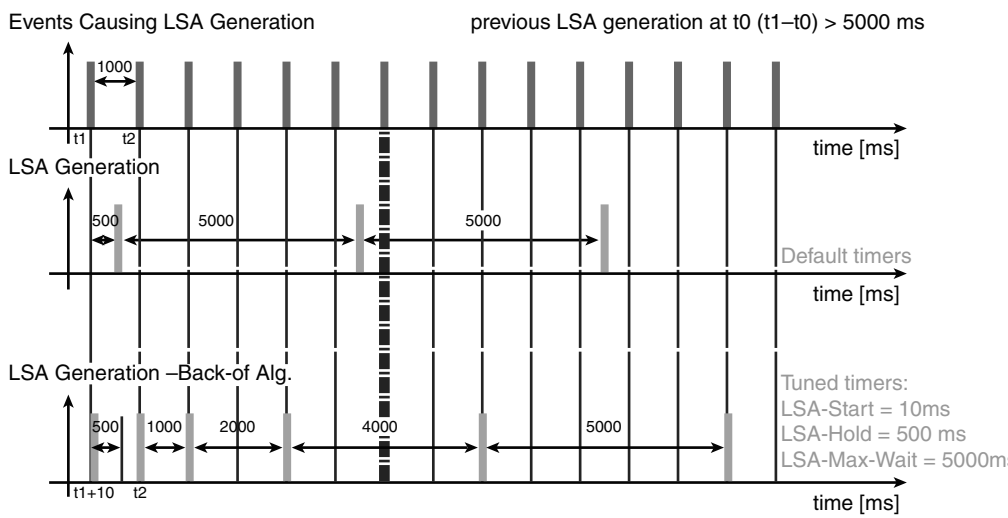



**Figure 3-19** *Tuning OSPF LSA Throttle Timers*

With the OSPF LSA throttle timers set at 10 ms for LSA-Start, 500 ms for LSA-Hold, and 5000 ms for LSA-Max-Wait, the initial LSA is generated after 10 ms. The next LSA is generated after the LSA-Hold time of 500 ms. The next LSA is generated after 2 x 500 = 1000 ms. The next LSA is generated after 4 x 500 = 2000 ms. The next LSA is generated after 8 x 500 = 4000 ms. The next one would be generated after 16 x 500 = 8000 ms, but because the LSA-Max-Wait is set at 5000 ms, the LSA is generated after 5000 ms. From this point onward, a 5000 ms wait is applied to successive LSAs, until the network stabilizes and the timers are reset.

### OSPF LSA Pacing

The LSA throttle timers control LSA generation by the originating routers. Another set of timers, the LSA pacing timers, controls the time it takes to propagate LSAs from router to router. By default, a router waits 33 ms between transmission of successive LSAs in the LSA flooding queue. There is a separate queue for LSA retransmissions, and LSAs in this queue are paced at 66 ms by default. If you adjust the LSA throttle timers to be low, you may also want to adjust these timers, because the total time for an LSA to propagate through the network is the initial LSA generation time plus the sum of the propagation delays between all routers in the path.

The intent of this timer is to ensure that you do not overwhelm neighboring routers with LSAs that cannot be processed quickly enough. However, with the increase of processing power on routers over the last decades this is not a major concern any more.

### OSPF Event Processing

The LSA throttling and pacing timers control OSPF LSA propagation. The next element in OSPF convergence is event processing. The timing of successive OSPF SPF calculations is throttled in the same manner as LSA generation, using an exponential backoff algorithm.

The timers involved in OSPF SPF throttling are very similar to the LSA throttling timers. There are three tunable timers:

- **SPF-Start:** The initial delay to schedule an SFP calculation after a change.
- **SPF-Hold:** The minimum holdtime between two consecutive SPF calculations. Similar to the LSA-Hold timer, this timer is used as an incremental value in an exponential backoff algorithm.
- **SPF-Max-Wait:** The maximum wait time between two consecutive SPF calculations.

Considerations in adjusting these timers are similar to the LSA throttling timers. An additional factor to consider is the time it takes for an SPF calculation to complete on the router platform used. You cannot schedule a new SPF run before the previous calculation has completed. Therefore, ensure that the SPF-Hold timer is higher than the time it takes to run a complete SPF. When estimating SPF run times, you should account for future network growth.

## Bidirectional Forwarding Detection

Bidirectional Forwarding Detection (BFD) is another feature that helps speed up routing convergence. One of the significant factors in routing convergence is the detection of link or node failure. In the case of link failures, there is usually an electrical signal or keepalive to detect the loss of the link. BFD is a technology that uses efficient fast Layer 2 link hellos to detect failed or one-way links, which is generally what fast routing hellos detect.

BFD requires routing-protocol support. BFD is available for OSPF, EIGRP, IS-IS, and BGP. BFD quickly notifies the routing protocol of link-down conditions. This can provide failure detection and response times down to around 50 ms, which is the typical SONET failure response time.

The CPU impact of BFD is less than that of fast hellos. This is because some of the processing is shifted to the data plane rather than the control plane. On nondistributed platforms, Cisco testing has shown a minor, 2 percent CPU increase above baseline when supporting 100 concurrent BFD sessions.

BFD provides a method for network administrators to configure subsecond Layer 2 failure detection between adjacent network nodes. Furthermore, administrators can configure their routing protocols to respond to BFD notifications and begin Layer 3 route convergence almost immediately.

**Note** BFD is currently supported only on Cisco 6500/7600 series routers, Cisco 12000 series routers, Cisco 10720 routers, Cisco Nexus 7000, and Cisco Carrier Routing System (CRS-1) routers.

# Designing Scalable BGP Designs

Border Gateway Protocol (BGP) is commonly used in sites with multiple connections to the Internet. BGP is also frequently present in medium-to large networks to provide a controlled interconnection between multiple routing domains running OSPF or EIGRP. Large-scale internal BGP networks are also becoming more prevalent as large enterprises implement internal Multiprotocol Label Switching (MPLS) VPNs for security segmentation, business unit or brand isolation, and similar purposes.

This section discusses designing advanced routing solutions using BGP. It describes how to identify scaling issues in internal BGP designs and how to use techniques to alleviate these issues.

## Scaling BGP Designs

This section discusses aspects of scaling in basic internal BGP (IBGP) design. This is illustrated in Figure 3-20.

BGP can provide a controlled interconnection between multiple routing domains running OSPF or EIGRP and support internal MPLS VPNs. IBGP requires a full mesh of BGP peers.

The full mesh of IBGP routers is needed because IBGP routers do not re-advertise routes learned via IBGP to other IBGP peers. This behavior is part of BGP protocol behavior that is used to prevent information from circulating between IBGP speaking routers in a routing information loop or cycle. External BGP (EBGP) relies on the autonomous system path to prevent loops. However, there is no way to tell whether a route advertised through

several IBGP speakers is a loop. Because IBGP peers are in the same autonomous system, they do not add anything to the autonomous system path, and they do not re-advertise routes learned via IBGP.

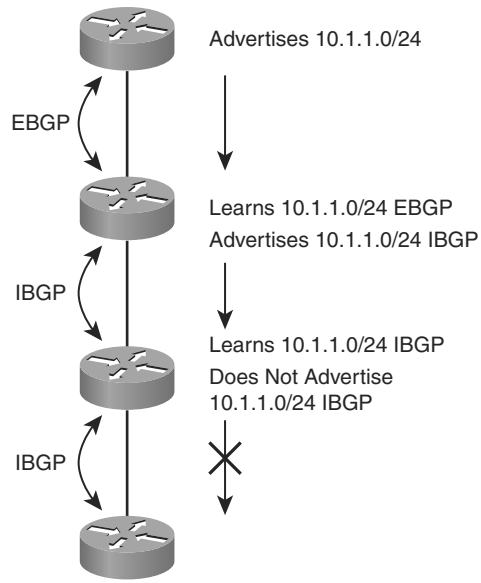



**Figure 3-20** *IBGP Full-Mesh Requirement*

**Note** BGP is commonly used in sites with multiple connections to the Internet. BGP is also common with MPLS VPNs and MPLS mixed with IP Security (IPsec) VPNs.

## Full-Mesh IBGP Scalability

Because IBGP requires a full mesh of peers, scaling the full mesh is a concern. In general, for $N$ peers in an IBGP full mesh, each would have $N - 1$ peers. There are $N (N - 1) / 2$ peering relationships. This means that each peer would need the CPU, memory, and bandwidth to handle updates and peer status for all the other routers. This is not a hierarchical design, and it would not be cost-effective to scale for large networks.

There are two IBGP alternatives to scale IBGP:

- Route reflectors
- Confederations

The following sections explore the basic design and behavior of route reflectors and confederations and demonstrate how they can be used in a routing design.

## Scaling IBGP with Route Reflectors

A BGP route reflector is an IBGP speaker that reflects or repeats routes learned from IBGP peers to some of its other IBGP peers. This is shown in Figure 3-21.

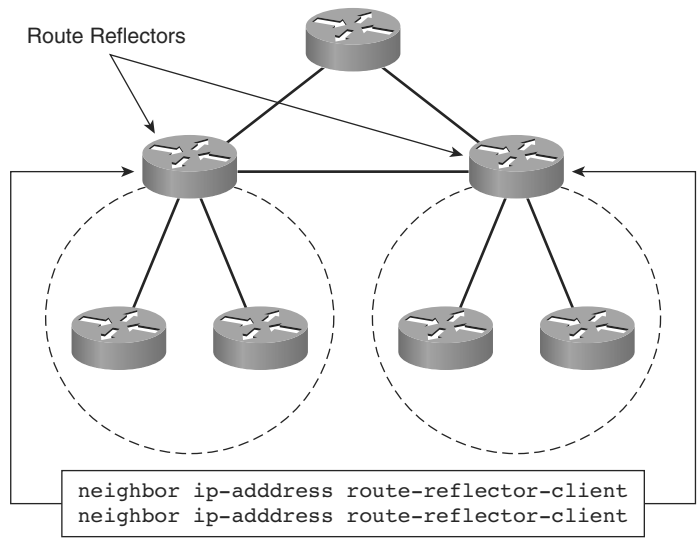



**Figure 3-21** *BGP Route Reflectors*

To prevent loops, a route reflector adds an originator ID and a cluster list to routes that it reflects between IBGP speakers. These attributes act similarly to the autonomous system path attribute to prevent routing information loops.

All configuration of the route reflector is done on the route reflector itself. The configuration identifies which IBGP peers are route reflector clients.

Implementing route reflectors is fairly simple and can be done incrementally. Each client router needs to be configured as a client on the route reflector or on multiple route reflectors. Unnecessary peers can then be removed from the configuration on the client router. Often, route reflector clients peer only with the route reflectors. In a service provider network, route reflector clients might also be provider edge (PE) devices, which also peer with customers using EBGP.

To avoid a single point of failure, redundant route reflectors are typically used.

## BGP Route Reflector Definitions

A route reflector client (shown in Figure 3-22) is an IBGP router that receives and sends routes to most other IBGP speakers via the route reflector. The route reflector client needs no special configuration, other than removing peering with some or all neighbors other than the route reflector.

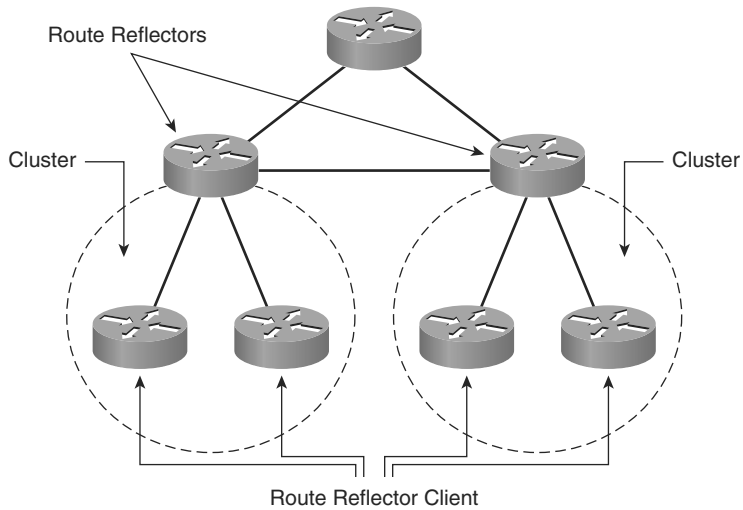



**Figure 3-22** *BGP Route Reflector Definitions*

A cluster is a route reflector together with its clients. The route reflector relieves the route reflector client routers of needing to be interconnected via an IBGP full mesh.

Route reflector clusters may overlap.

A nonclient router (shown in Figure 3-23) is any route reflector IBGP peer that is not a route reflector client of that route reflector.

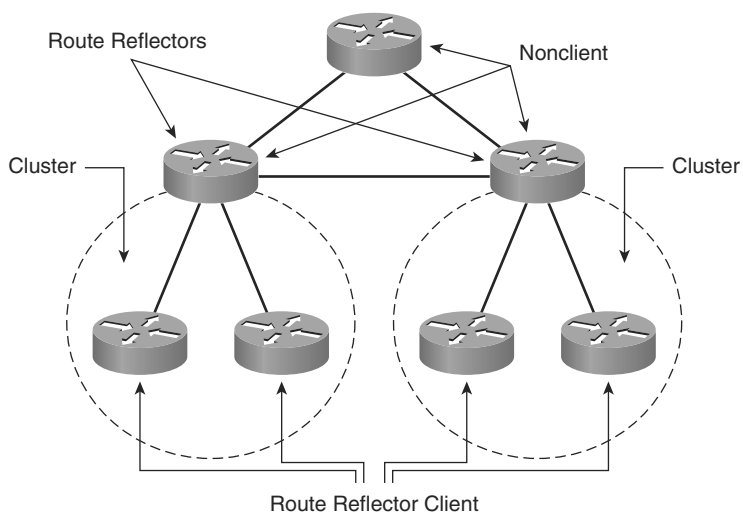



**Figure 3-23** *Additional BGP Route Reflector Definitions*

Route reflectors are typically nonclients with regard to the other route reflectors in the network.

Route reflectors must still be fully IBGP meshed with nonclients. Therefore, route reflectors reduce meshing within clusters, but all mesh links outside the cluster must be maintained on the route reflector. The route reflector clients get information from IBGP speakers outside the cluster via the route reflector.

If a route reflector receives a route from a nonclient, it reflects it to route reflector clients but not to other nonclients. The route reflector receives the routes if it has a direct peering relationship to the original nonclient. The route reflector also sends the route to EBGP peers, which is standard behavior. IBGP routes get repeated to all EBGP peers.

## Route Reflector Basics

This section briefly looks at how route advertisement works with route reflectors. This is illustrated in Figure 3-24.

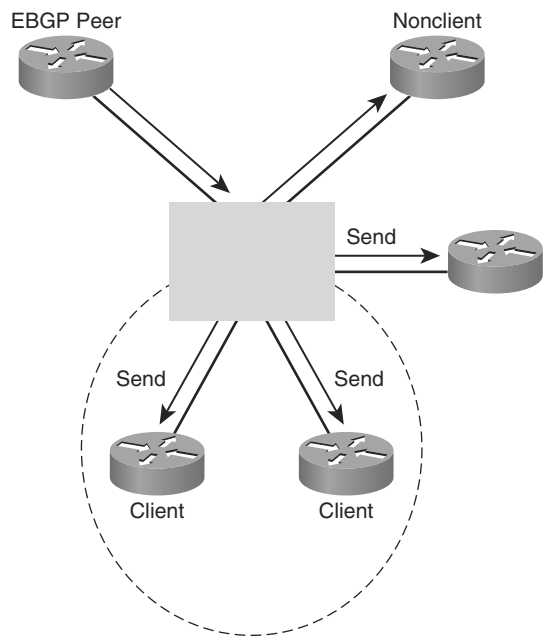



**Figure 3-24** *Route Reflector Basics*

If a route reflector receives a route from an EBGP peer, it passes that route to all route reflector clients and nonclients, just as in normal IBGP peering behavior.

If the route reflector receives a route from a route reflector client, it reflects the route to the other clients within the cluster and nonclients. It also reflects the route to EBGP peers. Here's another way to think of this: The route reflector takes over the communication

for the route reflector clients, passing along all the messages they would normally transmit directly via a peering session.

**Note** Route reflectors ignore the split-horizon design of IBGP and re-advertise routes they have received to any neighbor they have configured as a route reflector client.

## Scaling IBGP with Confederations

BGP confederations are another way of scaling IBGP. Their behavior is defined in RFC 5065. Confederations insert information using the autonomous system path into BGP routes to prevent loops within an autonomous system. The basic idea with confederations is to divide a normal BGP autonomous system into multiple sub-autonomous systems. The outer or containing autonomous system is called the confederation autonomous system. This is all that is visible to the outside world.

Each of the inner autonomous systems is a smaller sub-autonomous system that uses a different autonomous system number, typically chosen from the private autonomous system number range of 64,512 through 65,534.

## BGP Confederation Definitions

This section defines terms used with confederations (see Figure 3-25).

Peers within the same sub-autonomous system are confederation internal peers.

IBGP peers that are in a different sub-autonomous system are confederation external peers.

As IBGP information is passed around within a confederation autonomous system, the sub-autonomous system numbers are put into a confederation sequence, which works like an autonomous system path.

## Confederation Basics

Route advertisement with confederations works similarly to that of route reflectors in the following ways:

- A route learned from an EBGP peer is advertised to all confederation external and internal peers.
- A route learned from a confederation internal peer is advertised to all confederation external peers, and to EBGP peers.
- A route learned from a confederation external peer is advertised to all confederation internal peers, and to EBGP peers.

Another way to understand this is that IBGP between sub-autonomous systems acts like EBGP. Private autonomous system numbers are used internally within the confederation autonomous system and removed from updates sent outside the confederation.

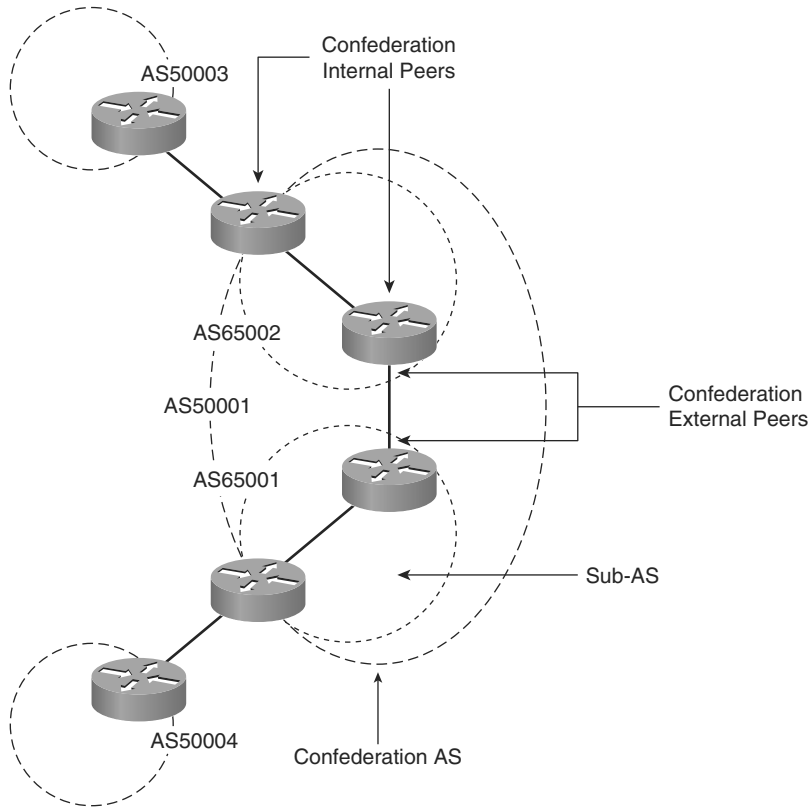



**Figure 3-25** *Confederation Definitions*

**Note** Private autonomous system numbers are typically used within the confederation.

## Confederations Reduce Meshing

Like route reflectors, confederations are used to reduce the amount of IBGP meshing needed. Without route reflectors or confederation, IBGP requires a full mesh of peering relationships, as illustrated in Figure 3-26.

**Note** The IBGP does not require peers to be directly connected.

However, confederations can reduce meshing requirements, as shown in Figure 3-27.

Routers in different sub-autonomous systems do not peer with each other, except at sub-autonomous system borders. It is generally recommended to use two or three links between sub-autonomous system borders. More links just consume CPU and memory in the border routers.

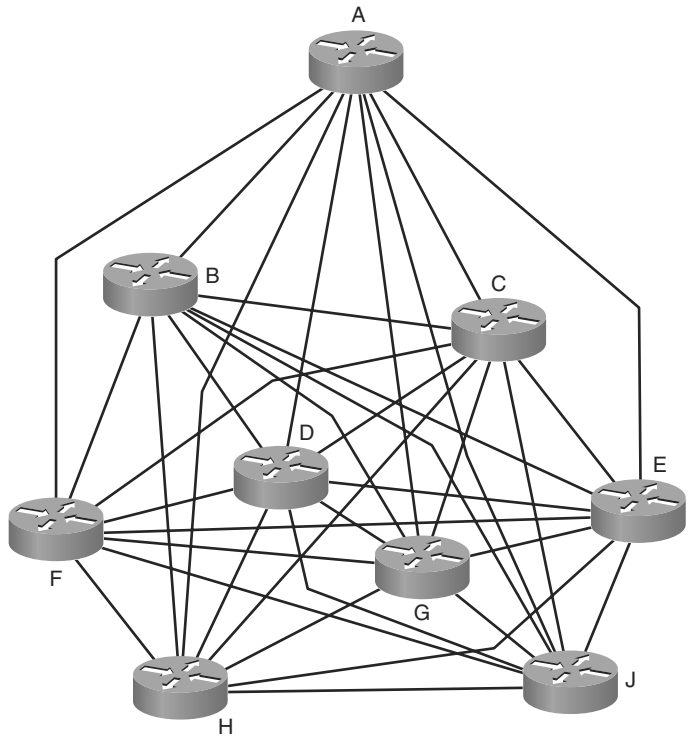



**Figure 3-26** *IBGP Full-Mesh Peering*

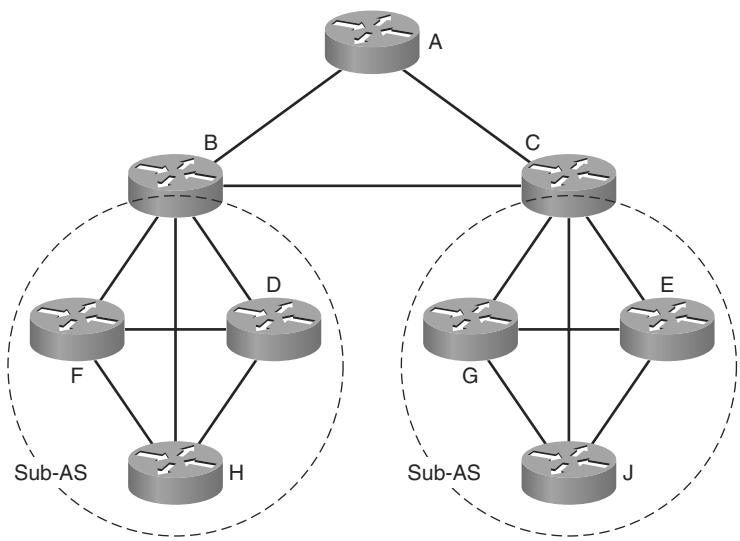



**Figure 3-27** *Confederations Reduce the Number of IBGP Peers*

When you use sub-autonomous systems for confederations, the meshing is restricted to within the sub-autonomous systems, with some additional peering between sub-autonomous system border routers.

Route reflectors can be used within confederations to further reduce network complexity. Historically, service providers have not done this, but they are now starting to. Using route reflectors alleviates the need to fully mesh within a sub-autonomous system.

## Deploying Confederations

In Figure 3-28, router B could be configured to set the BGP next hop to itself for advertisement to routers C and D. This is not normally done by IBGP routers. This would impose the constraint that routers C and D would need to have routes to the new next hop, router B.

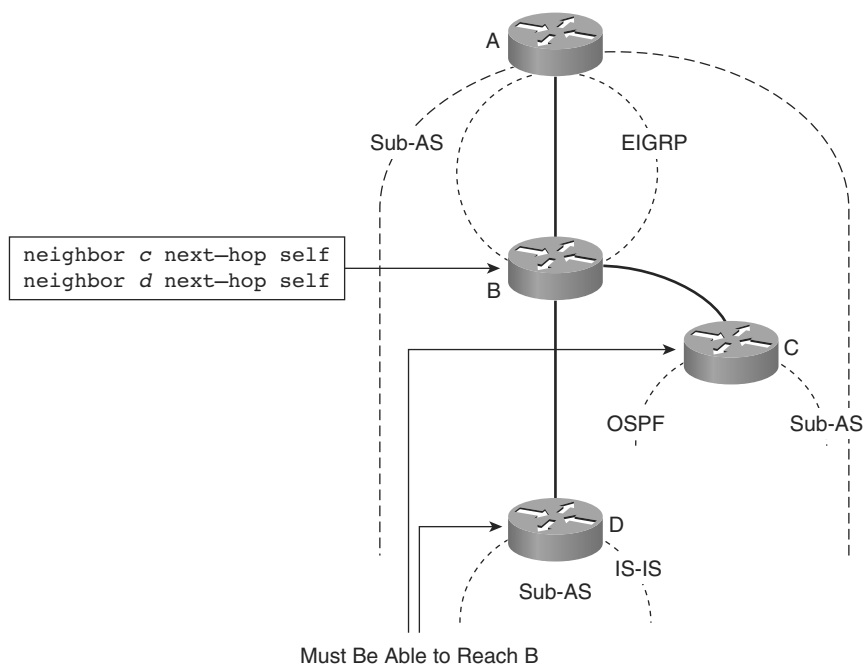



**Figure 3-28** *Deploying Confederations*

Using this configuration breaks the confederation up from a next-hop perspective from both the IGP and BGP point of view. This scenario allows for more flexibility and scaling in very large networks. This deployment might make sense for large organizations that support separate entities such as government organizations that have distinct branches or divisions.

Using confederation sub-autonomous systems has other advantages. The IBGP policies can differ internally within and between the sub-autonomous systems. In particular,

multi-exit discriminator (MED) acceptance or stripping, local preference settings, route dampening, and so on can vary between sub-autonomous systems. In addition, policy controls can be used on peerings between sub-autonomous systems.

This highlights some advantages of confederations. Confederations can ease the transition in an acquisition or merger. The new network can be treated as another sub-autonomous system and keep its IGP. It can also keep its EBGP policies with its customers.

A disadvantage of confederations is that there is no graceful way to migrate from full mesh to using confederations. The migration may well require downtime.

Table 3-1 compares how confederations and route reflectors provide various IBGP scaling features.

**Table 3-1** *Comparing Confederations to Route Reflectors*

| | **Confederation** | **Route Reflector** |
|---|---|---|
| Loop prevention | Autonomous system confederation set | Originator or set cluster ID |
| Break up a single autonomous system | Sub-autonomous systems | Clusters |
| Redundancy | Multiple connections between sub-autonomous systems | Client connects to several reflectors |
| External connections | Anywhere in the network | Anywhere in the network |
| Multilevel hierarchy | Reflectors within sub-autonomous systems | Hierarchical clusters |
| Policy control | Along outside borders and between sub-autonomous systems | Along outside borders |
| Scalability | Medium; still requires full IBGP within each sub-autonomous system | Very high |
| Migration | Very difficult (impossible in some situations) | Moderately easy (impossible in some situations) |

In general, route reflectors are simpler to migrate to and relatively simple to use, whereas confederations are more flexible as to IGP and policy.

## Summary

This chapter covered the elements of advanced routing design, and touched on the merits of a well-planned IP addressing scheme. The IP addressing scheme is the foundation for greater efficiency in operating and maintaining a network. Without proper planning in

advance, networks might not be able to benefit from route summarization features inherent to many routing protocols.

Cisco favors a transition strategy from IPv4 to IPv6 that begins from the edges of the network and moves in toward the core. This strategy allows you to control the deployment cost and focus on the needs of the applications, rather than complete a full network upgrade to a native IPv6 network at this stage. Cisco IPv6 router products offer the features for a such an integration strategy. The various deployment strategies permit the first stages of the transition to IPv6 to happen now, whether as a trial of IPv6 capabilities or as the early controlled stages of major IPv6 network implementations.IPv6 can be deployed as dual stack, hybrid, and service block.

The general advanced routing design discussion can be encapsulated in the following key points:

- Route summarization and default routing are important in scaling routing designs.
- Route filtering can be used to manage traffic flows in the network, avoiding inappropriate transit traffic and as a defense against inappropriate routing updates.
- Redistribution can be useful for manipulating and managing routing updates but needs to be designed properly to prevent routing loops or other problems.

EIGRP converges quickly as long as it has a feasible successor. With no feasible successor, EIGRP sends queries out to its neighbors. To limit the scope of these queries, use route summarization and filtering. By limiting EIGRP query scope, you can speed up EIGRP convergence and increase stability. In addition, large numbers of neighbors should be avoided for any one router. Multiple autonomous systems may be used with EIGRP providing that you understand that they do not directly limit EIGRP query scope. You would use them to support migration strategies, different administrative groups, or very large network design.

OSPF scaling depends on summarization and controlling how much LSA flooding is needed. Simple, stub, summarized designs scale most effectively. Several techniques speed up convergence for OSPF, including fast hellos, and BFD.

Finally, IBGP requires a full mesh of all IBGP routers, but full-mesh peering does not scale gracefully. Route reflectors pass along routing information to and from their clients. The route reflector clients are relieved of the burden of most IBGP peering. Confederations allow an autonomous system to be divided into sub-autonomous systems, where the sub-autonomous system border routers peer with each other and then pass along routes on behalf of the other sub-autonomous system routers. Confederation sequences are used to prevent information loops. Sub-autonomous systems can have different BGP polices from each other.

The key points to remember include the following:

- IP address design allows for route summarization that supports network scaling, stability, and fast convergence.

- Route summarization, route filtering, and appropriate redistribution help minimize routing information in the network.
- EIGRP converges quickly as long as it has a feasible successor. Multiple autonomous systems with EIGRP may be used, with care, to support special situations, including migration strategies and very large network design.
- Simple, stub, summarized OSPF designs scale most effectively. Several techniques speed up convergence for OSPF, including fast hellos and BFD.
- IBGP designs can be scaled using route reflectors to pass routing information to and from their clients and confederations to allow an autonomous system to be divided into sub-autonomous systems.

## References

Cisco Systems, Inc. *Deploying IPv6 in Campus Networks* at www.cisco.com/en/US/docs/solutions/Enterprise/Campus/CampIPv6.html

Shannon McFarland, Muninder Sambi, Nikhil Sharma, and Sanjay Hooda. *IPv6 for Enterprise Networks* (Cisco Press, 2011)

Cisco Systems, Inc. *Designing Large-Scale IP Internetworks* at www.cisco.com/en/US/docs/internetworking/design/guide/nd2003.html

*Cisco IOS IP Routing: BGP Command Reference* at www.cisco.com/en/US/docs/ios/iproute_bgp/command/reference/irg_book.html

*Cisco IOS IP Routing: EIGRP Command Reference* at www.cisco.com/en/US/docs/ios/iproute_eigrp/command/reference/ire_book.html

*Cisco IOS IP Routing: ISIS Command Reference* at www.cisco.com/en/US/docs/ios/iproute_isis/command/reference/irs_book.html

*Cisco IOS IP Routing: ODR Command Reference* at www.cisco.com/en/US/docs/ios/iproute_odr/command/reference/ird_book.html

*Cisco IOS IP Routing: OSPF Command Reference* at www.cisco.com/en/US/docs/ios/iproute_ospf/command/reference/iro_book.html

*Cisco IOS IP Routing: Protocol-Independent Command Reference* at www.cisco.com/en/US/docs/ios/iproute_pi/command/reference/iri_book.html

*Cisco IOS IP Routing: RIP Command Reference* at www.cisco.com/en/US/docs/ios/iproute_rip/command/reference/irr_book.html

The Internet Engineering Task Force. RFC 1793: *Extending OSPF to Support Demand Circuits* at www.ietf.org/rfc/rfc1793.txt

The Internet Engineering Task Force. RFC 2328: *OSPF Version 2* at www.ietf.org/rfc/rfc2328.txt

The Internet Engineering Task Force. RFC 4456: *BGP Route Reflection—An Alternative to Full Mesh IBGP* at www.ietf.org/rfc/rfc4456.txt

The Internet Engineering Task Force. RFC 5065: *Autonomous System Confederations for BGP* at www.ietf.org/rfc/rfc5065.txt

The Internet Engineering Task Force. RFC 4136: *OSPF Refresh and Flooding Reduction in Stable Topologies* at www.ietf.org/rfc/rfc4136.txt

## Review Questions

Answer the following questions, and then refer to Appendix A, "Answers to Review Questions," for the answers.

1. Which three address blocks are summarizable?
   a. 172.16.20.0/24 to 172.16.27.0/24
   b. 172.16.20.0/24 to 172.16.23.0/24
   c. 10.16.0.0/16 to 10.31.0.0/16
   d. 10.16.0.0/16 to 10.47.0.0/16
   e. 2001:0DB8:C3B7:10A0::/64 to 2001:0DB8:C3B7:10DF::/64
   f. 2001:0DB8:1234:FB40::/64 to 2001:0DB8:1234:FB5F::/64
   g. 10.96.0.0/16 to 10.159.0.0/16
2. Which two can bit-splitting techniques be used for? (Choose two.)
   a. OSPF area design
   b. Summarizable address blocks with convenient role-based subnets
   c. Access list convergence
   d. Detecting summarizable address blocks
   e. Manual route summarization
3. Which is the recommended design approach for OSPF?
   a. Configure a static default route everywhere for predictability.
   b. Configure static default routes using recursive routing for consistency.
   c. Originate the default at the edge and redistribute it into dynamic routing.
   d. Make the OSPF backbone area 0 stubby.
   e. Do not use additional parameters with the originate default command.

4. Which two statements best describe redistribution?
   a. Redistribution works poorly with an arbitrary mix of routing protocols anywhere.
   b. Redistribution seldom requires route filters.
   c. Redistribution is not useful after a merger.
   d. Redistribution works well with a limited number of redistribution points.
   e. Redistribution prevents summarization.
5. Select the best statement concerning EIGRP and OSPF routing design.
   a. Routing design needs to be done most carefully for small networks.
   b. OSPF should not be used for small networks.
   c. Routing design needs to be done most carefully for large networks.
   d. Route summarization must be used in all network designs.
   e. OSPF works best with a full mesh.
6. Which three factors are the biggest influences on OSPF scalability?
   a. Flooding paths and redundancy
   b. Amount of routing information in the OSPF area or routing domain
   c. Number of routers capable of Cisco Express Forwarding
   d. Number of adjacent neighbors
   e. Other routing protocols in use
7. Which statement best describes basic IBGP?
   a. IBGP is a link-state protocol.
   b. IBGP requires a full mesh of peers because it has no other way to prevent looping of routing information.
   c. IBGP inherently handles all full-mesh scalability issues.
   d. IBGP uses split horizoning to prevent looping of routing information.
   e. IBGP uses the autonomous system path to prevent looping of routing information.
8. A route reflector reflects routes from a route reflector client to which three types of IBGP routers?
   a. Nonclient routers
   b. Sub-autonomous system members
   c. Other route reflector client routers
   d. EBGP peers
   e. IBGP peers configured for EIGRP or OSPF routing

*This page intentionally left blank*

# Index

## Numerics



## A

# B

# C

# D

# E

# F

# G



# H



# I

# J-K-L

# M

# N

# O



# P

# Q

# R

# S

# T



# U

# V

# W-X-Y-Z