Building Data Centers with VXLAN BGP EVPN

A Cisco NX-OS Perspective

Lukas Krattiger, CCIE No. 21921
Shyam Kapadia
David Jansen, CCIE No. 5952

ciscopress.com

Building Data Centers with VXLAN BGP EVPN

A Cisco NX-OS Perspective

Lukas Krattiger, *CCIE No. 21921*
Shyam Kapadia
David Jansen, *CCIE No. 5952*

Cisco Press
800 East 96th Street
Indianapolis, IN 46240

Building Data Centers with VXLAN BGP EVPN

A Cisco NX-OS Perspective

Lukas Krattiger
Shyam Kapadia
David Jansen

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

Printed in the United States of America 1 2 3 4 5 6 7 8 9 0

First Printing April 2017

Library of Congress Cataloging-in-Publication Number: 2017931984

ISBN-10: 1-58714-467-0

ISBN-13: 978-1-58714-467-7

Warning and Disclaimer

This book is designed to provide information about data center network design. Every effort has been made to make this book as complete and as accurate as possible, but no warranty or fitness is implied.

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

The opinions expressed in this book belong to the 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.

Special Sales

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

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

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

Feedback Information

At Cisco Press, our goal is to create in-depth technical books of the highest quality and value. Each book is crafted with care and precision, undergoing rigorous development that involves the unique expertise of members of 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 by e-mail, at feedback@ciscopress.com. Please make sure to include the book title and ISBN in your message. We greatly appreciate your assistance.

Editor-in-Chief: Mark Taub

Alliances Manager, Cisco Press: Ron Fligge

Product Line Manager: Brett Bartow

Executive Editor: Mary Beth Ray

Managing Editor: Sandra Schroeder

Development Editor: Christopher Cleveland

Project Editor: Mandie Frank

Copy Editor: Kitty Wilson

Technical Editors: Scott Morris, Jeff Tantsura

Editorial Assistant: Vanessa Evans

Cover Designer: Okomon Haus

Composition: codeMantra

Indexer: Erika Millen

Proofreader: Abigail Manheim

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

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

Europe Headquarters
Cisco Systems International BV
Amsterdam, The Netherlands

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

About the Authors

Lukas Krattiger, CCIE No. 21921 *(Routing/Switching and Data Center)*, is principal engineer, Technical Marketing, with more than 15 years of experience in data center, Internet, and application networks. Within Cisco, he specializes in data center switching, overlay architectures, and solutions across platforms. Lukas is a double-CCIE (R&S and Data Center) with several other industry certifications and has participated in various technology leadership and advisory groups. Prior to joining Cisco, Lukas was a senior network engineer with System Integrators and Service Providers, where he was responsible for data center and Internet networks. Since joining Cisco, he has covered various technologies within the data center as well as enterprise networks portfolio, and he has built foundational solutions for customers and partners. He is from Switzerland and currently lives in California with his wife and one wonderful daughter. He can be found on Twitter at @ccie21921.

Shyam Kapadia is a principal engineer in the Data Center Group at Cisco Systems. With more than a decade of experience in the networking industry, Shyam holds more than 30 patents and has coauthored the book *Using TRILL, FabricPath, and VXLAN: Designing MSDC with Overlays*. In his 10 years at Cisco, Shyam has worked on a number of products, including the Catalyst and Nexus families of switches, with special emphasis on end-to-end data center solutions, including automation and orchestration. He holds a Ph.D. and master's degree from the University of Southern California in the field of computer science. Over the past 15 years, Shyam has been the Program Chair for the Southern California Linux Exposition (SCALE). He lives in California with his wife, enjoys watching international movies, and is passionate about sports including cricket, basketball, and football.

David Jansen, CCIE No. 5952 *(Routing/Switching)*, is a distinguished systems engineer (DSE) for Cisco, specializing in data center, campus, branch/WAN, and cloud architectures. He has 20 years of experience in the industry and has earned certifications from Novell, VMware, Microsoft, TOGAF, and Cisco. His focus is working with global enterprise customers to address their challenges with comprehensive end-to-end data center, enterprise, WAN/Internet, and cloud architectures. David has been with Cisco for more than 19 years; for the last 4 years or so as a DSE, he has gained unique experiences in building next generation data center solutions. David has a bachelor's degree in computer science engineering from the University of Michigan and a master's degree in adult education from Central Michigan University.

About the Technical Reviewers

Scott Morris, the world traveling Über-Geek has four CCIE certifications (Routing & Switching, ISP/Dial, Security and Service Provider) as well as the coveted CCDE. He also has several expert-level certifications from other major vendors, making him "multi-lingual" in the networking world.

Working on large-scale network designs, troubleshooting, and some very interesting CyberSecurity projects, has kept Scott occupied. Outside of challenging work, Scott can be found spending time with his family or finding new things to learn. Having more than 30 years of experience in just about all aspects of the industry has provided both an in-depth and an entertaining approach to disseminating knowledge. Whether involved in large-scale designs, interesting implementations, or expert-level training, you can often find Scott willing to share information.

Jeff Tantsura has been in the networking space for 20+ years and has authored/contributed to many RFCs and patents. He is the chair of the IETF Routing Working Group, chartered to work on new network architectures and technologies, including protocol independent YANG models and working on YANG modeling as the working group chair and contributor.

Jeff is a coauthor of a recently published book, *Navigating Network Complexity*, talking, among other existing topics, about why networking has become so complex and the urgent need for automation and programmable, model-driven networking.

Dedications

From Lukas Krattiger:

I want to dedicate this book to my family, especially my wife, Snjezi, and daughter, Nadalina. They have shown immense patience during nights, weekends, vacations, and other inconvenient times while this book project was being completed. I love you both!

From Shyam Kapadia:

I dedicate this book to my family, especially my wife, Rakhee, and my mother, for their constant love and support.

From David Jansen:

This book is dedicated to my loving wife, Jenise, and my three children, Kaitlyn, Joshua, and Jacob. You are the inspiration that gave me the determination to complete this project. To my three amazing children, you are learning the skills to be the best at what you do and to accomplish anything in life; keep up the great work. Thank you for all your love and support. I could not have completed yet another book without your help, support, and understanding. I would also like to further dedicate this book to my parents, Michael and Dolores. You have given me the proper tools, guidance, attitude, drive, and education to allow me to do what I do. I'm likewise grateful to God, who gives endurance, encouragement, and motivation to complete such a large project like this. In my last book dedication, I mentioned that I would not take on any additional projects like that one. As you can see, I had a hard time saying no when my good friend and colleague Lukas Krattiger convinced me to take on this project. Thank you, Lukas; it is always a pleasure, my friend. I truly enjoy working with you.

Acknowledgments

From Lukas Krattiger:

First, I'd like to thank my coauthors, Shyam Kapadia and David Jansen. Shyam, thank you for being such a great coworker and a true technical leader in our organization. I could not imagine anyone better. It has been truly wonderful sharing ideas with you, and I look forward to addressing additional challenges and innovations with you in the near future. David, thank you for stepping in to help tackle this project. You are an exceptional colleague, and it was a true pleasure working with you these many engagements, especially the video series. Each of us has unique insights and gifts to contribute, and both you and Shyam highlight the benefits the diversity in our community provides.

Likewise, I would like to send a special acknowledgment to the other team members with whom I am working. In particular, I would like to recognize Carl Solder as well as Yousuf Khan for all the support and timely guidance. Special thanks to Victor Moreno for all the discussions and the groundbreaking work with overlays.

I would also like to thank some individuals who are intimately involved with VXLAN EVPN. In particular, Ali Sajassi, Samir Thoria, Dhananjaya Rao, Senthil Kenchiah (and team), Neeraj Malhota, Rajesh Sharma, and Bala Ramaraj deserve special recognition. Similarly, all my engineering and marketing colleagues who support this innovative technology and have helped contribute to the completion of this book deserve special recognition.

A special shout-out goes to all my friends in Switzerland, Europe, Australia, the United States, and the rest of the globe. Mentioning all of you here would create an additional 11 chapters.

Finally, writing this book provided me with the opportunity to get to know some new acquaintances. I would like to thank Doug Childress for his continuous edits and reviews on the manuscript, and I would also like to thank our technical editors, Scott Morris and Jeff Tantsura, for all the feedback they provided. Finally, I would like to give a special thanks to Cisco Press for all the support on this project.

From Shyam Kapadia:

I would like to especially thank my coauthors, Lukas and David, for their collaboration and support. Lukas did the lion's share in putting together this publication, and he deserves substantial credit for that. It's hard to imagine this book coming together without his tremendous contribution. Our collaboration over the past several years has been extremely fruitful, and I look forward to additional joint innovations and deliverables in the future. In addition, I would like to thank David, who has been a good role model for many individuals at Cisco, including me.

I'd like to give a special acknowledgment to the engineering leadership team in the Data Center group at Cisco for their constant support and encouragement in pursuing this endeavor. This team includes Nilesh Shah, Ravi Amanaganti, Venkat Krishnamurthy, Dinesh Khurana, Naoshad Mehta, and Mahesh Chellappa.

Like Lukas, I want to recognize individuals at Cisco involved in taking VXLAN BGP EVPN to the summit where it is today. I would also like to acknowledge the contributions of the DCNM team for providing the management and controller aspects to the programmable fabric solution with VXLAN BGP EVPN. I would like to also thank Doug Childress for helping review and edit the book chapters, and I offer a special thanks to the reviewers and editors for their tremendous help and support in developing this book. This is my second collaboration with Cisco Press, and the experience has been even better than the first one.

From David Jansen:

This is my fourth book, and it has been a tremendous honor to work with the great people at Cisco Press. There are so many people to thank, I'm not sure where to begin. First, I would like to thank my friends and coauthors, Lukas Krattiger and Shyam Kapadia. Both of you are great friends as well as exceptional coworkers. I can't think of two better people with whom to work or complete such a project. Cisco is one of the most amazing places I've ever worked, and people like you who are exceptionally intelligent and an extreme pleasure to work with make it such a great place. I look forward to working with you on other projects in the future and growing our friendship further into as well.

I would also like to acknowledge Chris Cleveland, with whom it is always a pleasure to work. His expertise, professionalism, and follow-up as a development editor are unsurpassed. I would like to specifically thank him for all his hard work and quick turnaround times in meeting the deadlines.

To our technical editors, Jeff Tantsura and Scott Morris, I would like to offer a thank you for your time, sharp eyes, and excellent comments/feedback provided during this project. It was a pleasure having you both as part of the team.

I would like to also thank the heavy metal music world out there. It allowed me to stay focused when burning the midnight oil. I would not have been able to complete this without loud rock and roll music, air guitar, and air drums as well! So thank you.

I want to thank my family for their support and understanding while I was working on this project late at night. They were patient with me when my lack of rest may have made me a little less than pleasant to be around. I know it is also hard to sleep when Dad is downstairs writing and fails to realize how the decibel level of the music is interfering with the rest of the family's ability to sleep.

Most importantly, I would like to thank God for giving me the ability to complete such a task with the required dedication and determination and for providing me the skills, knowledge, and health needed to be successful in such a demanding profession.

Contents at a Glance

Contents

Introduction

Building Data Centers with VXLAN BGP EVPN is intended to provide a solid understanding of how data center network fabrics with VXLAN BGP EVPN function. It serves as both a technology compendium and a deployment guide.

Cisco's NX-OS-based data center switching portfolio provides a collection of networking protocols and features that are foundational to building data center networks as traditional networks evolve into fabric-based architectures, like VXLAN with the BGP EVPN control plane.

This book's goal is to explain how to understand and deploy this technology, and it begins with an introduction to the current data center challenges, before going into the technology building blocks and related semantics. It also provides an overview of the evolution of the data center fabric. The book takes a deep dive into the various fabric semantics, including the underlay, multitenancy, control and data plane interaction, unicast and multicast forwarding flows, and external, data center interconnect, and service appliance deployments.

Goals and Methods

The goal of this book is to provide a resource for readers who want to get familiar with data center overlay technologies, especially VXLAN with a control plane like BGP EVPN. This book describes a methodology that network architects and administrators can use to plan, design, and implement scalable data centers. You do not have to be a networking professional or data center administrator to benefit from this book. The book is geared toward understanding the functionality of VXLAN with BGP EVPN in data center fabric deployments. Our hope is that all readers, from university students to professors to networking experts, will benefit from this book.

Who Should Read This Book?

This book has been written with a broad audience in mind, while specifically targeting network architects, engineers, and operators. Additional audiences who will benefit from reading this book include help desk analysts, network administrators, and certification candidates. This book provides information on VXLAN with BGP EVPN for today's data centers.

For a network professional with in-depth understanding of various networking areas, this book serves as an authoritative guide, explaining detailed control and data plane concepts, with VXLAN and BGP EVPN being the primary focus. Detailed packet flows are presented, covering numerous functions, features, and deployments.

Regardless of your level of expertise or role in the IT industry, this book offers significant benefits. It presents VXLAN and BGP EVPN concepts in a consumable manner. It also describes design considerations for various fabric semantics and identifies the key benefits of adopting this technology.

How This Book Is Organized

Although this book slowly progresses conceptually from Chapter 1 to Chapter 11, you could also read individual chapters that cover only the material of interest. The first chapter provides a brief introduction to the evolution of data center networks, with an emphasis on the need for network overlays. Chapters 2 and 3 form the foundation for VXLAN BGP EVPN. The subsequent chapters describe underlying or adjacent building blocks to VXLAN BGP EVPN, with an emphasis on Layer 2 and Layer 3 services and the associated multitenancy. Chapter 10 describes the integration of Layer 4–7 services into a VXLAN network with BGP EVPN, while Chapter 11 concludes the book with an overview of fabric management and operations.

The chapter breakdown is as follows:

- **Chapter 1**, **"Introduction to Programmable Fabric."** This chapter provides a brief introduction to the Cisco VXLAN BGP EVPN fabric. It begins with a description of the requirements of today's data centers. It also gives an overview of how data centers evolved over the years, leading to a VXLAN BGP EVPN-based spine–leaf fabric. This chapter introduces common fabric-based terminology and describes what makes the fabric extremely scalable, resilient, and elastic.
- **Chapter 2**, **"VXLAN BGP EVPN Basics."** This chapter describes why overlays have become a prime design choice for next-generation data centers, with a special emphasis on VXLAN, which has become the de facto choice. The chapter describes the need for a control plane–based solution for distribution of host reachability between various edge devices and provides a comprehensive introduction to BGP EVPN. It describes the important message formats in BGP EVPN for supporting network virtualization overlays and presents representative use cases. The subsequent chapters build on this background and provide further details on the underlay, multitenancy, and single-destination and multidestination data packet flows in a VXLAN BGP EVPN–based data center network.
- **Chapter 3**, **"VXLAN/EVPN Forwarding Characteristics."** This chapter provides an in-depth discussion on the core forwarding capabilities offered by a VXLAN BGP EVPN fabric. For carrying broadcast, unknown unicast, and multicast (BUM) traffic, this chapter describes both multicast and ingress replication. It also discusses enhanced forwarding features that reduce flooding in the fabric on account of ARP and unknown unicast traffic. This chapter describes one of the key benefits of a BGP EVPN fabric: the realization of a Distributed Anycast Gateway at the ToR or leaf layer.
- **Chapter 4**, **"The Underlay."** This chapter describes the BGP EVPN VXLAN fabric underlay that needs to be able to transport both single-destination and multidestination overlay traffic. The primary objective of the underlay is to provide reachability among the various switches in the fabric. This chapter presents IP address allocation options for the underlay, using both point-to-point IP numbered options and the

rather attractive IP unnumbered option. It also discusses choices of popular IGP routing protocols, such as OSPF, IS-IS, and BGP for unicast routing. The chapter also describes the two primary choices for multidestination traffic replication in the underlay: the unicast and multicast mode.

- **Chapter 5**, **"Multitenancy."** This chapter describes how multitenancy has become a prime feature for next-generation data centers and how it is realized in the data center network with VXLAN BGP EVPN. In addition to discussing multitenancy when using VLANs or Bridge Domains (BDs) in VXLAN this chapter covers modes of operation for both Layer 2 and Layer 3 multitenancy. Overall, this chapter provides a basic introduction to the main aspects of multitenancy when using data center networks with VXLAN BGP EVPN.
- **Chapter 6**, **"Unicast Forwarding."** This chapter provides a set of sample packet flows that indicate how bridging and routing operations occur in a VXLAN BGP EVPN network. Critical concepts related to IRB functionality, symmetric IRB, and distributed anycast gateway are described in action for real-world traffic flows. This chapter pays special attention to scenarios with silent hosts as well as dual-homed hosts.
- **Chapter 7**, **"Multicast Forwarding."** This chapter provides details about forwarding multicast data traffic in a VXLAN BGP EVPN network. It discusses vanilla Layer 2 multicast traffic forwarding over VXLAN, as well as topics related to its evolution with enhancements in IGMP snooping. It also presents special considerations for dual-homed and orphan multicast endpoints behind a vPC domain.
- **Chapter 8**, **"External Connectivity."** This chapter presents external connectivity options with a VXLAN BGP EVPN fabric. After introducing the border leaf and border spine variants, it provides details on options for external Layer 3 connectivity using VRF Lite, LISP, and MPLS L3 VPN. It also details Layer 2 external connectivity options, with an emphasis on vPC.
- **Chapter 9**, **"Multi-pod, Multifabric, and Data Center Interconnect (DCI)."** This chapter describes various concepts related to multi-pod and multifabric options with VXLAN BGP EVPN deployments. It provides a brief primer on the salient distinctions between OTV and VXLAN. Most practical deployments require some form of interconnection between different pods or fabrics. This chapter discusses various considerations that need to be taken into account when making a decision on when to use the multi-pod option versus the multifabric option.
- **Chapter 10**, **"Layer 4–7 Services Integration."** This chapter provides details on how Layer 4–7 services can be integrated into a VXLAN BGP EVPN network. It covers deployments with intra-tenant and inter-tenant firewalls, which can be deployed in both transparent and routed modes. In addition, this chapter presents a common deployment scenario with load balancers, with emphasis on the nuances associated with its integration into a VXLAN BGP EVPN network. The chapter concludes with a common load balancer and firewall service chain deployment example.

- **Chapter 11, "Introduction to Fabric Management."** This chapter introduces the basic elements of fabric management, including POAP-based day-0 provisioning (using DCNM, NFM, and so on), incremental configuration using day-0.5 configuration, overlay configuration using day-1 provisioning (using DCNM, VTS, and NFM), and day-2 provisioning, which involves provisions for continuous monitoring, visibility, and troubleshooting capabilities in a VXLAN BGP EVPN fabric. It presents a brief primer on VXLAN OAM, which is an extremely efficient tool for debugging in overlay-based fabrics.

Command Syntax Conventions

The conventions used to present command syntax in this book are the same conventions used in the *NX-OS Command Reference*:

- **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).
- *Italics* indicate arguments for which you supply actual values.
- Vertical bars (|) separate alternative, mutually exclusive elements.
- Square brackets [ ] indicate optional elements.
- Braces { } indicate a required choice.
- Braces within brackets [{ }] indicate a required choice within an optional element.

Chapter 3

VXLAN/EVPN Forwarding Characteristics

In this chapter, the following topics will be covered:

- Enhanced BGP EVPN features such as ARP suppression, unknown unicast suppression, and optimized IGMP snooping
- Distributed IP anycast gateway in the VXLAN EVPN fabric
- Anycast VTEP implementation with dual-homed deployments

VXLAN BGP EVPN has been extensively documented in various standardized references, including IETF drafts[1] and RFCs.[2] While that information is useful for implementing the protocol and related encapsulation, some characteristics regarding forwarding require additional attention and discussion. Unfortunately, a common misunderstanding is that VXLAN with BGP EVPN does not require any special treatment for multidestination traffic. This chapter describes how multidestination replication works using VXLAN with BGP EVPN for forwarding broadcast, unknown unicast, and multicast (BUM) traffic. In addition, methods to reduce BUM traffic are covered. This chapter also includes a discussion on enhanced features such as early Address Resolution Protocol (ARP) suppression, unknown unicast suppression, and optimized Internet Group Management Protocol (IGMP) snooping in VXLAN BGP EVPN networks. While the functions and features of Cisco's VXLAN BGP EVPN implementation support reduction of BUM traffic, the use case of silent host detection still requires special handling, which is also discussed.

In addition to discussing Layer 2 BUM traffic, this chapter talks about Layer 3 traffic forwarding in VXLAN BGP EVPN networks. VXLAN BGP EVPN provides Layer 2 overlay services as well as Layer 3 services. For Layer 3 forwarding or routing, the presence of a first-hop default gateway is necessary. The distributed IP anycast gateway enhances the first-hop gateway function by distributing the endpoints' default gateway across all available edge devices (or Virtual Tunnel Endpoints [VTEPs]). The distributed IP anycast gateway is implemented using the Integrated Routing and Bridging (IRB) functionality. This ensures that both bridged and routed traffic—to and from endpoints—is always optimally forwarded within the network with predictable latency, based on the BGP EVPN advertised reachability information. Because of this distributed approach for the Layer 2/Layer 3 boundary, virtual machine mobility is seamlessly handled from the network point of view by employing special mobility-related functionality in BGP EVPN. The distributed anycast gateway ensures that the IP-to-MAC binding for the default gateway does not change, regardless of where an end host resides or moves within the network. In addition, there is no hair-pinning of routed traffic flows to/from the endpoint after a move event. Host route granularity in the BGP EVPN control plane protocol facilitates efficient routing to the VTEP behind which an endpoint resides—at all times.

Data centers require high availability throughout all layers and components. Similarly, the data center fabric must provide redundancy as well as dynamic route distribution. VXLAN BGP EVPN provides redundancy from multiple angles. The Layer 3 routed underlay between the VXLAN edge devices or VTEPs provides resiliency as well as multipathing. Connecting classic Ethernet endpoints via multichassis link aggregation or virtual PortChannel (vPC) provides dual-homing functionality that ensures fault tolerance even in case of a VTEP failure. In addition, typically, Dynamic Host Configuration Protocol (DHCP) services are required for dynamic allocation of IP addresses to endpoints. In the context of DHCP handling, the semantics of centralized gateways are slightly different from those of the distributed anycast gateway. In a VXLAN BGP EVPN fabric, the DHCP relay needs to be configured on each distributed anycast gateway point, and the DHCP server configuration must support DHCP Option 82.[3] This allows a seamless IP configuration service for the endpoints in the VXLAN BGP EVPN fabric.

The aforementioned standards are specific to the interworking of the data plane and the control plane. Some components and functionalities are not inherently part of the standards, and this chapter provides coverage for them. The scope of the discussion in this chapter is specific to VXLAN BGP EVPN implementation on Cisco NX-OS-based platforms. Later chapters provide the necessary details, including detailed packet flows, for forwarding traffic in a VXLAN BGP EVPN network, along with the corresponding configuration requirements.

Multidestination Traffic

The VXLAN with BGP EVPN control plane has two different options for handling BUM or multidestination traffic. The first approach is to leverage multicast replication in the underlay. The second approach is to use a multicast-less approach called *ingress replication*, in which multiple unicast streams are used to forward multidestination traffic to the appropriate recipients. The following sections discuss these two approaches.

Leveraging Multicast Replication in the Underlying Network

The first approach to handling multidestination traffic requires the configuration of IP multicast in the underlay and leverages a network-based replication mechanism. With multicast, a single copy of the BUM traffic is sent from the ingress/source VTEP toward the underlay transport network. The network itself forwards this single copy along the multicast tree (that is, a shared tree or source tree) so that it reaches all egress/destination VTEPs participating in the given multicast group. As the single copy travels along the multicast tree, the copy is replicated at appropriate branch points only if receivers have joined the multicast group associated with the VNI. With this approach, a single-copy per-wire/link is kept within the network, thereby providing the most efficient way to forward BUM traffic.

A Layer 2 VNI is mapped to a multicast group. This mapping must be consistently configured on all VTEPs where this VNI is present, typically signifying the presence of some interested endpoint below that VTEP. Once it is configured, the VTEP sends out a corresponding multicast join expressing interest in the tree associated with the corresponding multicast group. When mapping a Layer 2 VNI to a multicast group, various options are available in the mapping provision for the VNI and multicast group. The simplest approach is to employ a single multicast group and map all Layer 2 VNIs to that group. An obvious benefit of this mapping is the reduction of the multicast state in the underlying network; however, it is also inefficient in the replication of BUM traffic. When a VTEP joins a given multicast group, it receives all traffic being forwarded to that group. The VTEP does not do anything with traffic in which no interest exists (for example, VNIs for which the VTEP is not responsible). In other words, it silently drops that traffic. The VTEP continues to receive this unnecessary traffic as an active receiver for the overall group. Unfortunately, there is no suboption for a VNI to prune back multicast traffic based on both group and VNI. Thus, in a scenario where all VTEPs participate in the same multicast group or groups, the scalability of the number of multicast outgoing interfaces (OIFs) needs to be considered.

At one extreme, each Layer 2 VNI can be mapped to a unique multicast group. At the other extreme, all Layer 2 VNIs are mapped to the same multicast group. Clearly, the optimal behavior lies somewhere in the middle. While theoretically there are 2^{24} = 16 million possible VNI values and more than enough multicast addresses (in the multicast address block 224.0.0.0–239.255.255.255), practical hardware and software factors limit the number of multicast groups used in most practical deployments to a few hundred. Setting up and maintaining multicast trees requires a fair amount of state maintenance and protocol exchange (PIM, IGMP, and so on) in the underlay. To better explain this concept, this chapter presents a couple of simple multicast group usage scenarios for BUM traffic in a VXLAN network. Consider the BUM flows shown in Figure 3-1.

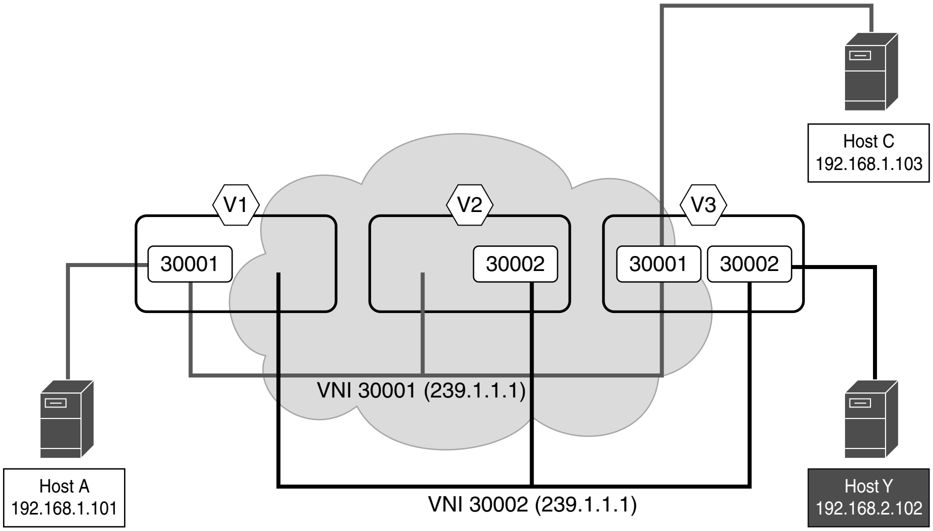


Figure 3-1 *Single Multicast Group for All VNIs*

Three edge devices participate as VTEPs in a given VXLAN-based network. These VTEPs are denoted V1 (10.200.200.1), V2 (10.200.200.2), and V3 (10.200.200.3). IP subnet 192.168.1.0/24 is associated with Layer 2 VNI 30001, and subnet 192.168.2.0/24 is associated with Layer 2 VNI 30002. In addition, VNIs 30001 and 30002 share a common multicast group (239.1.1.1), but the VNIs are not spread across all three VTEPs. As is evident from Figure 3-1, VNI 30001 spans VTEPs V1 and V3, while VNI 30002 spans VTEPs V2 and V3. However, the shared multicast group causes all VTEPs to join the same shared multicast tree. In the figure, when Host A sends out a broadcast packet, VTEP V1 receives this packet and encapsulates it with a VXLAN header. Because the VTEP is able to recognize this as broadcast traffic (DMAC is FFFF.FFFF.FFFF), it employs a destination IP in the outer IP header as the multicast group address. As a result, the broadcast traffic is mapped to the respective VNI's multicast group (239.1.1.1).

The broadcast packet is then forwarded to all VTEPs that have joined the multicast tree for the group 239.1.1.1. VTEP V2 receives the packet, but it silently drops it because it has no interest in the given VNI 30001. Similarly, VTEP V3 receives the same packet that is replicated through the multicast underlay. Because VNI 30001 is configured for VTEP V3, after decapsulation, the broadcast packet is sent out to all local Ethernet interfaces participating in the VLAN mapped to VNI 30001. In this way, the broadcast packet initiated from Host A reaches Host C.

For VNI 30002, the operations are the same because the same multicast group (239.1.1.1) is used across all VTEPs. The BUM traffic is seen on all the VTEPs, but the traffic is dropped if VNI 30002 is not locally configured at that VTEP. Otherwise, the traffic is appropriately forwarded along all the Ethernet interfaces in the VLAN, which is mapped to VNI 30002.

As it is clear from this example, unnecessary BUM traffic can be avoided by employing a different multicast group for VNIs 30001 and 30002. Traffic in VNI 30001 can be scoped to VTEPs V1 and V3, and traffic in VNI 30002 can be scoped to VTEPs V2 and V3. Figure 3-2 provides an example of such a topology to illustrate the concept of a scoped multicast group.

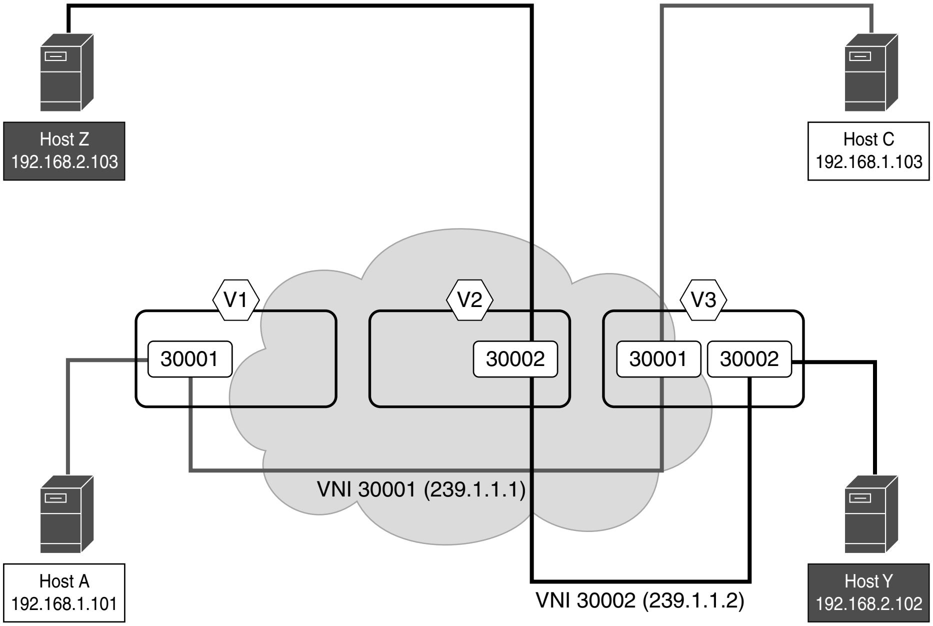


Figure 3-2 *Scoped Multicast Group for VNI*

As before, the three edge devices participating as VTEPs in a given VXLAN-based network are denoted V1 (10.200.200.1), V2 (102.200.200.2), and V3 (10.200.200.3). The Layer 2 VNI 30001 is using the multicast group 239.1.1.1, while the VNI 30002 is using a different multicast group, 239.1.1.2. The VNIs are spread across the various VTEPs, with VNI 30001 on VTEPs V1 and V3 and VNI 30002 on VTEPs V2 and V3. The VTEPs are required to join only the multicast group and the related multicast tree for the locally configured VNIs. When Host A sends out a broadcast, VTEP V1 receives this broadcast packet and encapsulates it with an appropriate VXLAN header. Because the VTEP understands that this is a broadcast packet (DMAC is FFFF.FFFF.FFFF), it uses a destination IP address in the outer header with the multicast group address mapped to the respective VNI (239.1.1.1). This packet is forwarded to VTEP V3 only, which previously joined the multicast tree for the Group 239.1.1.1, where VNI 30001 is configured. VTEP V2 does not receive the packet, as it does not belong to the multicast group because it is not configured with VNI 30001. VTEP V3 receives this packet because it belongs to the multicast group mapped to VNI 30001. It forwards it locally to all member ports that are part of the VLAN that is mapped to VNI 30001. In this way, the broadcast traffic is optimally replicated through the multicast network configured in the underlay.

The same operation occurs in relation to VNI 30002. VNI 30002 is mapped to a different multicast group (239.1.1.2), which is used between VTEPs V2 and V3. The BUM traffic is seen only on the VTEPs participating in VNI 30002—specifically multicast group 239.1.1.2. When broadcast traffic is sent from Host Z, the broadcast is replicated only to VTEPs V2 and V3, resulting in the broadcast traffic being sent to Host Y. VTEP V1 does not see this broadcast traffic as it does not belong to that multicast group. As a result, the number of multicast outgoing interfaces (OIFs) in the underlying network is reduced.

While using a multicast group per VNI seems to work fine in the simple example with three VTEPs and two VNIs, in most practical deployments, a suitable mechanism is required to simplify the assignment of multicast groups to VNIs. There are two popular ways of achieving this:

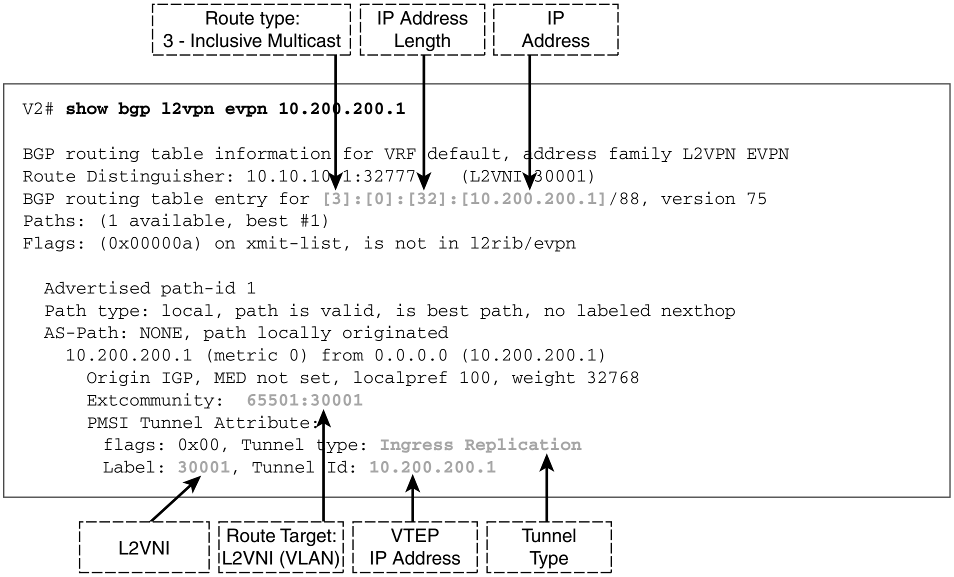
- VNIs are randomly selected and assigned to multicast groups, thereby resulting in some implicit sharing.
- A multicast group is localized to a set of VTEPs, thereby ensuring that they share the same set of Layer 2 VNIs.

The second option to handling multidestination traffic may sound more efficient than the first, but in practice it may lack the desired flexibility because it limits workload assignments to a set of servers below a set of VTEPs, which is clearly undesirable.

Using Ingress Replication

While multicast configuration in the underlay is straightforward, not everyone is familiar with or willing to use it. Depending on the platform capabilities, a second approach for multidestination traffic is available: leveraging ingress or head-end replication, which is a unicast approach. The terms *ingress replication* (*IR*) and *head-end replication* (*HER*) can be used interchangeably. IR/HER is a unicast-based mode where network-based replication is not employed. The ingress, or source, VTEP makes *N*–1 copies of every BUM packet and sends them as individual unicasts toward the respective *N*–1 VTEPs that have the associated VNI membership. With IR/HER, the replication list is either statically configured or dynamically determined, leveraging the BGP EVPN control plane. Section 7.3 of RFC 7432 (https://tools.ietf.org/html/rfc7432) defines inclusive multicast Ethernet tag (IMET) routing, or Route type 3 (RT-3). To achieve optimal efficiency with IR/HER, best practice is to use the dynamic distribution with BGP EVPN. BGP EVPN provides a Route type 3 (inclusive multicast) option that allows for building a dynamic replication list because IP addresses of every VTEP in a given VNI would be advertised over BGP EVPN. The dynamic replication list has the egress/destination VTEPs that are participants in the same Layer 2 VNI.

Each VTEP advertises a specific route that includes the VNI as well as the next-hop IP address corresponding to its own address. As a result, a dynamic replication list is built. The list is updated when configuration of a VNI at a VTEP occurs. Once the replication list is built, the packet is multiplied at the VTEP whenever BUM traffic reaches the ingress/source VTEP. This results in individual copies being sent toward every VTEP in a VNI across the network. Because network-integrated multicast replication is not used, ingress replication is employed, generating additional network traffic. Figure 3-3 illustrates sample BGP EVPN output with relevant fields highlighted, for a Route type 3 advertisement associated with a given VTEP.



```
V2# show bgp l2vpn evpn 10.200.200.1

BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 10.10.10.1:32777    (L2VNI 30001)
BGP routing table entry for [3]:[0]:[32]:[10.200.200.1]/88, version 75
Paths: (1 available, best #1)
Flags: (0x00000a) on xmit-list, is not in l2rib/evpn

  Advertised path-id 1
  Path type: local, path is valid, is best path, no labeled nexthop
  AS-Path: NONE, path locally originated
    10.200.200.1 (metric 0) from 0.0.0.0 (10.200.200.1)
      Origin IGP, MED not set, localpref 100, weight 32768
      Extcommunity:  65501:30001
      PMSI Tunnel Attribute:
        flags: 0x00, Tunnel type: Ingress Replication
        Label: 30001, Tunnel Id: 10.200.200.1
```

Figure 3-3 *Route type 3: Inclusive Multicast*

When comparing the multicast and unicast modes of operation for BUM traffic, the load for all the multidestination traffic replication needs to be considered. For example, consider a single Layer 2 VNI present on all 256 edge devices hosting a VTEP. Further, assume that a single VLAN exists on all 48 Ethernet interfaces of the edge devices where BUM traffic replication is occurring. Each Ethernet interface of the local edge device receives 0.001% of the nominal interface speed of BUM traffic. For 10G interfaces, this generates 0.0048 Gbps (4.8 Mbps) of BUM traffic at the edge device from the host-facing Ethernet interfaces. In multicast mode, the theoretical speed of BUM traffic remains at this rate. In unicast mode, the replication depends on the amount of egress/destination activity on the VTEP. In the example of 255 neighboring VTEPs, this results in 255 × 0.0048 = ~1.2 Gbps of BUM traffic on the fabric. Even though the example is theoretical, it clearly demonstrates the huge overhead that the unicast mode incurs as the amount of multidestination traffic in the network increases. The efficiency of multicast mode in these cases cannot be understated.

It is important to note that when scoping a multicast group to VNIs, the same multicast group for a given Layer 2 VNI must be configured across the entire VXLAN-based fabric. In addition, all VTEPs have to follow the same multidestination configuration mode for a given Layer 2 domain represented through the Layer 2 VNI. In other words, for a given Layer 2 VNI, all VTEPs have to follow the same unicast or multicast configuration mode. In addition, for the multicast mode, it is important to follow the same multicast protocol (for example, PIM ASM,[4] PIM BiDir[5]) for the corresponding multicast group. Failure to adhere to these requirements results in broken forwarding of multidestination traffic.

VXLAN BGP EVPN Enhancements

The following sections describe some of the feature enhancements that ride on top of the BGP EVPN control plane, further enhancing the forwarding of Layer 2 and Layer 3 traffic in a VXLAN fabric.

ARP Suppression

Address Resolution Protocol (ARP) is responsible for IPv4 address–to–MAC address mapping in IP networks. ARP facilitates obtaining endpoint MAC address information, leveraging the IP address information sent out in an ARP broadcast request. The broadcast request is treated as multidestination traffic that is typically VXLAN encapsulated and sent to every VTEP or edge device that is a member of the corresponding Layer 2 VNI. The response to the ARP request is typically sent as a unicast packet to the requestor. ARP traffic is scoped within the bounds of the broadcast domain represented by the Layer 2 VNI. Correspondingly, IPv6 networks rely on the Neighbor Discovery Protocol (ND) for IPv6 address–to–MAC address resolution. With IPv6, this resolution is triggered via an initial neighbor solicitation (NS) that is multicast through the Layer 2 broadcast domain. A neighbor advertisement (NA) sent in response from the destination completes the neighbor resolution.[6]

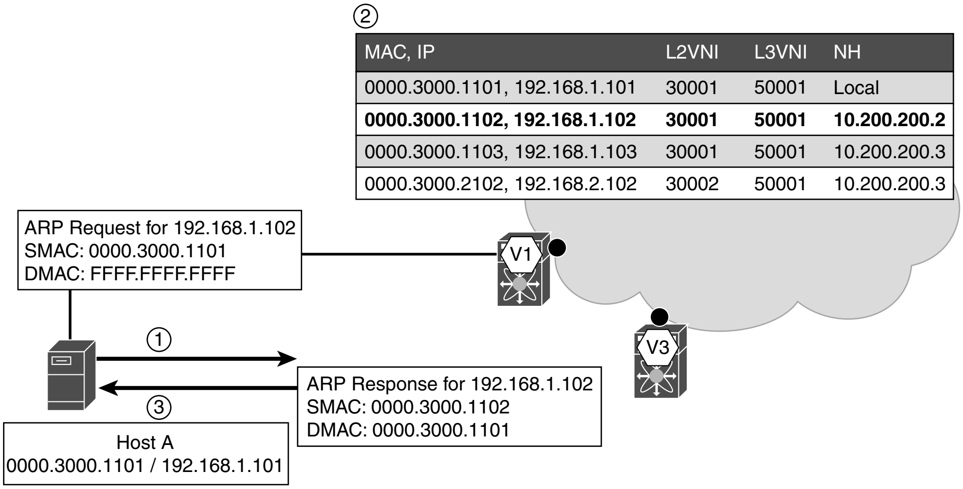
When an endpoint needs to resolve the default gateway, which is the exit point of the local IP subnet, it sends out an ARP request to the configured default gateway. The ARP operation allows the local edge device or VTEP to populate IP/MAC mappings of the locally attached endpoints. In addition to the MAC-to-IP table being populated on the edge device, all the MAC information is populated to the BGP EVPN control plane protocol. In addition, for Layer 2, Layer 2 VNI, Route Distinguisher (RD), Route Target (RT), hosting VTEP, and associated IP information is populated, and for Layer 3, Layer 3 VNI, RD, RT, hosting VTEP, and RMAC information is populated. Specifically, this information is populated in a Route type 2 advertisement and distributed to all remote VTEPs, which now have learned about that endpoint.

Typically, all ARP broadcast requests sent out from an endpoint are subject to copy-to-router and forward behavior. The local edge device learns about an endpoint as long as the endpoint sends out some type of ARP request—not necessarily an ARP request for the resolution of the default gateway hosted on the local edge device.

Typically, when an endpoint wants to talk to another endpoint in the same subnet, it sends out an ARP request for determining the IP-to-MAC binding of the destination endpoint. The ARP request is flooded to all the endpoints that are part of that Layer 2 VNI. ARP snooping coupled with the BGP EVPN control plane information can help avoid flooding for known endpoints. By using ARP snooping, all ARP requests from an endpoint are redirected to the locally attached edge device. The edge device then extracts the destination IP address in the ARP payload and determines whether it is a known endpoint. Specifically, a query is done against the known endpoint information from the BGP EVPN control plane.

If the destination is known, the IP-to-MAC binding information is returned. The local edge device then performs an ARP proxy on behalf of the destination endpoint. In other words, it sends out a unicast ARP response toward the requestor with the resolved MAC address of the known destination endpoint. In this way, all ARP requests to known endpoints are terminated at the earliest possible point, which is the locally attached edge device or VTEP or leaf. This is known as *ARP suppression*. ARP suppression is possible because all the information is known on all the VXLAN VTEPs.[7]

Note that ARP suppression is different from the Proxy ARP,[8] in which the edge device or router may serve as a proxy on behalf of the destination endpoint, using its own Router MAC. ARP suppression reduces ARP broadcast traffic by leveraging the BGP EVPN control plane information. ARP suppression is enabled on a per-Layer 2 VNI basis. In this way, for all known endpoints, ARP requests are sent only between the endpoint and the local edge device/VTEP. Figure 3-4 illustrates a scenario where ARP suppression at VTEP V1 results in early ARP termination of a request for a host 192.168.1.102 from a host 192.168.1.101 in Layer 2 VNI 30001. It is important to note that the ARP suppression feature works based on the knob enabled under the Layer 2 VNI, regardless of whether the default gateway is configured on the leafs.



MAC, IP	L2VNI	L3VNI	NH
0000.3000.1101, 192.168.1.101	30001	50001	Local
0000.3000.1102, 192.168.1.102	**30001**	**50001**	**10.200.200.2**
0000.3000.1103, 192.168.1.103	30001	50001	10.200.200.3
0000.3000.2102, 192.168.2.102	30002	50001	10.200.200.3

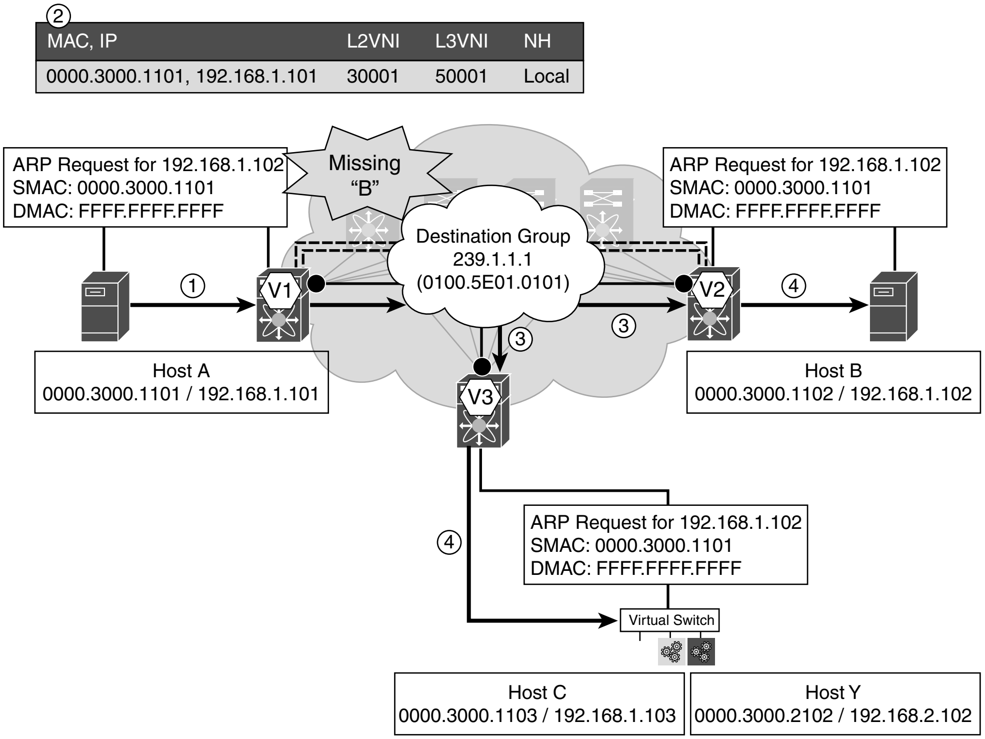
Figure 3-4 *ARP Suppression*

When the destination endpoint is not known to the BGP EVPN control plane (that is, a silent or undiscovered endpoint), the ARP broadcast needs to be sent across the VXLAN network. Recall that ARP snooping intercepts the ARP broadcast request to the locally attached edge device. Because the lookup for the destination endpoint results in a “miss,” the edge device re-injects the ARP broadcast request back into the network with appropriate source filtering to ensure that it does not return to the source port from which the ARP request was originally received. The ARP broadcast leverages the multidestination traffic forwarding implementation of the VXLAN fabric.

When the queried endpoint responds to the ARP request, the endpoint generates an ARP response. The unicast ARP response is also subjected to ARP snooping behavior so that it is sent to the remote edge device to which the destination is directly attached.

Consequently, the IP/MAC binding of the destination endpoint is learned by the remote VTEP, resulting in that information being populated into the BGP EVPN control plane. The remote edge device also sends out the ARP response to the original requesting endpoint, thereby completing the end-to-end ARP discovery process. The ARP response is sent via the data plane to ensure that there are no delays in the control plane updates. It is important to note that after the initial miss, all the subsequent ARP requests for the discovered endpoint are handled with local ARP termination, as described earlier.

Figure 3-5 illustrates a scenario involving ARP termination or ARP suppression where a miss occurs in the control plane. Host A and Host B belong to the same IP subnet, corresponding to the Layer 2 VNI 30001. Host A wants to communicate with Host B, but the respective MAC-to-IP mapping is not known in Host A’s ARP cache. Therefore, Host A sends an ARP request to determine the MAC-to-IP binding of Host B. The ARP request is snooped by VTEP V1. VTEP V1 uses the target IP address information gleaned from the ARP request payload to look up information about Host B in the BGP EVPN control plane. Since Host B is not yet known in the network, the broadcast ARP request is encapsulated into VXLAN with the destination IP address of a multicast group or a unicast destination IP address. The multicast group is leveraged when multicast is enabled in the underlay. Otherwise, using head-end replication, multiple unicast copies are generated to each remote interested VTEP, as mentioned earlier.



MAC, IP	L2VNI	L3VNI	NH
0000.3000.1101, 192.168.1.101	30001	50001	Local

Figure 3-5 *ARP lookup miss with ARP Suppression*

The ARP broadcast sent to VTEP V2 and VTEP V3 is appropriately decapsulated, and the inner broadcast payload is forwarded to all Ethernet ports that are members of VNI 30001. In this way, the ARP request reaches Host B, and Host B answers with a unicast ARP response back to Host A. Once the unicast ARP response hits VTEP V2, the information is snooped, and the control plane is then updated with Host B's IP/MAC address information. At the same time, the ARP response is encapsulated and forwarded across the VXLAN network to VTEP V1. On receiving the ARP response, VTEP V1 decapsulates the VXLAN header and forwards the decapsulated ARP response, based on a Layer 2 lookup, in VNI 30001. Consequently, Host A receives the ARP response and populates its local ARP cache. From this point forward, bidirectional communication between Host A and Host B is enabled because all MAC and IP address information of both hosts is known in the network.

The ARP request and response process between endpoints in a VXLAN BGP EVPN fabric is similar to that in classic Ethernet or any other Flood and Learn (F&L) network. However, the advantage of the control plane and ARP suppression is realized with the BGP EVPN information. A single ARP resolution of an endpoint triggers the control plane to populate that endpoint information across the entire network, at all the VTEPs. Consequently, any subsequent ARP requests to any known endpoint are locally answered instead of being flooded across the entire network.

Because ARP snooping proactively learns the endpoint information to populate the control plane, it certainly helps in reducing unknown unicast traffic. ARP requests really need to be flooded only for the initial period, when an endpoint has yet to be discovered. However, there are other sources of unknown unicast traffic (for example, vanilla Layer 2 non-IP traffic or even regular Layer 2 IP traffic) that may suffer a DMAC lookup miss and the traffic being flooded all across the VXLAN network.

To completely disable any kind of flooding due to unknown unicast traffic toward the VXLAN network, a new feature called *unknown unicast suppression* has been introduced. This independent feature is enabled on a per-Layer 2 VNI basis. With this knob enabled, when any Layer 2 traffic suffers a DMAC lookup miss, traffic is locally flooded, but there is no flooding over VXLAN. Therefore, if no silent or unknown hosts are present, flooding can potentially be minimized in a BGP EVPN VXLAN network.

Next, we consider the third piece of BUM or multidestination traffic, which is called *multicast traffic*. Layer 2 multicast in a VXLAN network is treated as broadcast or unknown unicast traffic because the multicast traffic is flooded across the underlay. Layer 2 multicast flooding leverages the configured multidestination traffic-handling method, whether it exists in unicast mode or multicast mode. Not all platforms can differentiate between the different types of Layer 2 floods, such as multicast or unknown unicast. The IGMP snooping feature presents an optimal way of forwarding Layer 2 multicast traffic if supported by the platform. The Cisco NX-OS default configuration results in behavior that floods Layer 2 multicast traffic both locally and across the VTEP interfaces (see Figure 3-6). This means that all participating VTEPs with the same VNI mapping and all endpoints in that Layer 2 VNI receive this Layer 2 multicast traffic. The multidestination traffic is forwarded to every endpoint in the VNI, regardless of whether that endpoint is an interested receiver.

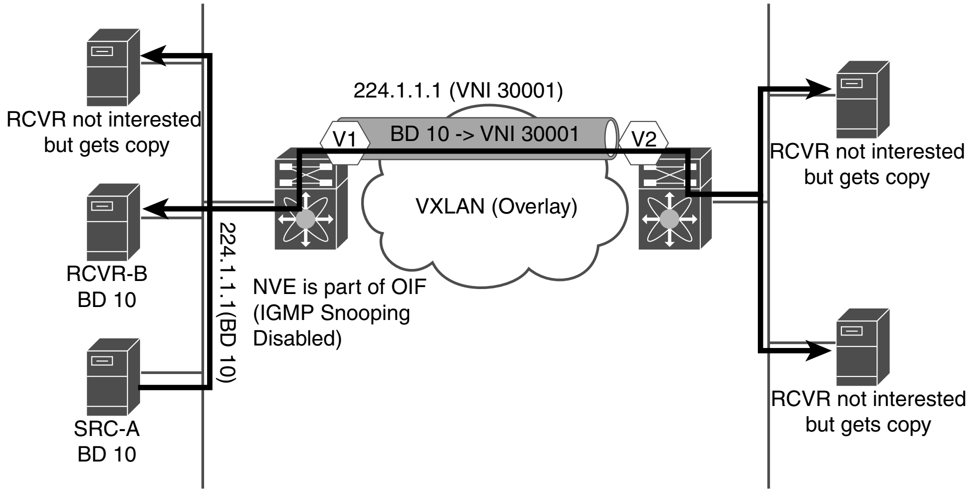


Figure 3-6 *No IGMP Snooping in VLAN/VXLAN*

With VXLAN, the option to implement IGMP snooping in the same way it is implemented in traditional Ethernet-based networks. The main difference is that the VTEP interface selectively allows Layer 2 multicast forwarding for interested receivers that are present behind a remote VTEP. If there are no interested receivers behind remote VTEPs, Layer 2 multicast traffic received from a source endpoint is not sent toward the VXLAN network. However, as long as there is at least one interested receiver behind some remote VTEP, Layer 2 multicast traffic is forwarded to all remote VTEPs since they are all part of the same Layer 2 VNI. The VTEP interface is a multicast router port that participates in IGMP snooping, thereby allowing Layer 2 multicast forwarding. This is dependent on the receipt of the IGMP “join” message from an endpoint for a given multicast group in a Layer 2 VNI.

The received IGMP join report is VXLAN encapsulated and transported to all remote VTEPs that are members of the corresponding Layer 2 VNI. Note that the multicast group(s) employed for overlay multicast traffic between endpoints should not be confused with the underlay multicast group that may be associated with the Layer 2 VNI. Recall that the latter is employed to provide the destination IP address in the outer IP header for forwarding multidestination traffic in the underlay when using multicast mode.

With IGMP snooping enabled for VXLAN VNIs, the ability to realize the true benefits of multicast occurs: Traffic is forwarded only to interested receiver endpoints (see Figure 3-7). In other words, if no interested receivers exist for a given multicast group behind a VTEP in the same VNI, the Layer 2 multicast traffic is silently dropped and not flooded. After decapsulation at the remote VTEP, Layer 2 multicast traffic is forwarded based on the local IGMP snooping state. IGMP join messages are used to selectively enable the forwarding of certain Layer 2 multicast traffic.

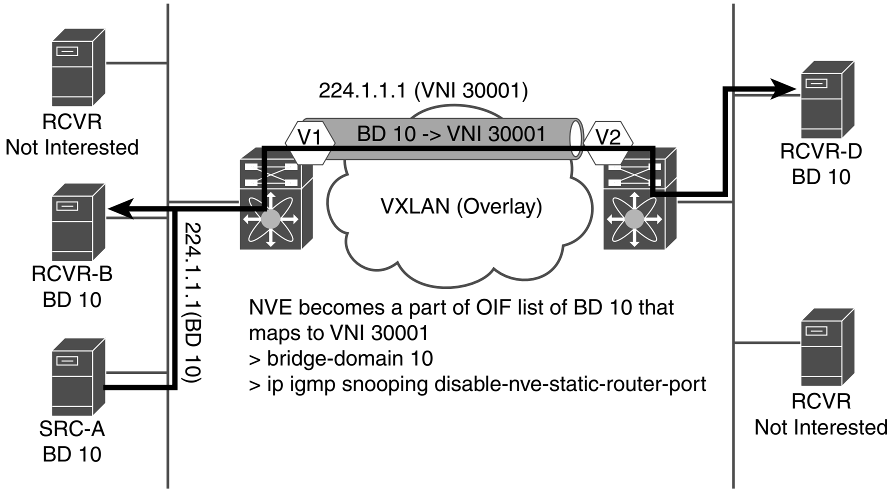


Figure 3-7 *IGMP Snooping in VLAN/VXLAN*

Depending on the platform software support, IGMP snooping for VXLAN-enabled VLANs may not be supported. IGMP snooping is a software feature that does not have any dependencies on the underlying hardware. While this section provided a brief overview, detailed IP multicast flows in a VXLAN BGP EVPN network are presented in Chapter 7 “Multicast Forwarding”.

Distributed IP Anycast Gateway

In order for two endpoints in different IP subnets to communicate with each other, a default gateway is required. Traditionally, the default gateway has been implemented centrally at the data center aggregation layer, in a redundant configuration. For an endpoint to reach the centralized default gateway, it has to first traverse a Layer 2 network. The Layer 2 network options include Ethernet, vPC, FabricPath, and even VXLAN F&L. Communication within the same IP subnet is typically bridged without any communication with the centralized default gateway. For routing communication between different IP networks/subnets, the centralized default gateway is reachable over the same Layer 2 network path.

Typically, the network is designed to be redundant as well as highly available. Likewise, the centralized gateway is highly available. First-hop redundancy protocols (FHRP) were designed to support the centralized default gateway with redundancy. Protocols such as HSRP,[9] VRRP,[10] and GLBP[11] are examples of FHRPs. HSRP and VRRP have a single node responsible for responding to ARP requests as well as routing traffic to a different IP network/subnet. When the primary node fails, the FHRP changes the operational state on the backup node to master. A certain amount of time is needed to fail over from one node to another.

FHRPs became more versatile when implemented with virtual PortChannels (vPCs), which allow both nodes to forward routing traffic and permit only the master node to respond to ARP requests. Combining vPC and FHRP significantly enhances resiliency as well as convergence time in case of failures. Enabling ARP synchronization for the FHRP in vPC environments enables ARP synchronization between the vPC primary and the vPC secondary. FabricPath with anycast HSRP[12] increases the number of active gateways from two to four nodes. The FHRP protocol exchange and operational state changes are still present with anycast HSRP. With FHRPs, the default gateway, implemented at the data center aggregation layer, provides a centralized default gateway as well as the Layer 2/Layer 3 network boundary.

The increasing importance of Layer 2 and Layer 3 operations, especially in a large data center fabric, has demanded additional resiliency in comparison to what has been provided by traditional FHRP protocols. Moving the Layer 2-Layer 3 network boundary to the fabric leaf/ToR switch or the access layer reduces the failure domain (see Figure 3-8). The scale-out approach of the distributed IP anycast gateway dramatically reduces the network and protocol state. The distributed IP anycast gateway implementation at each fabric leaf no longer requires each endpoint to traverse a large Layer 2 domain to reach the default gateway.

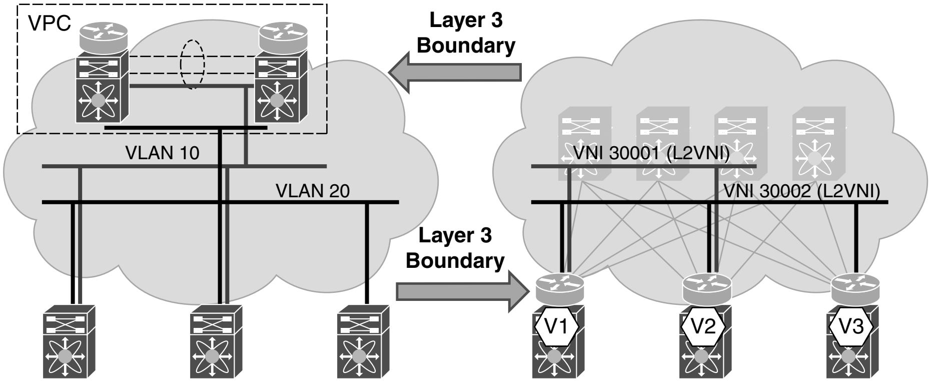


Figure 3-8 *Gateway Placement*

The distributed IP anycast gateway applies the anycast[13] network concept "one to the nearest association." Anycast is a network addressing and routing methodology in which the data traffic from an endpoint is routed topologically to the closest node in a group of gateways that are all identified by the same destination IP address. With the distributed IP anycast gateway (see Figure 3-9), the default gateway is moved closer to the endpoint—specifically to the leaf where each endpoint is physically attached. The anycast gateway is active on each edge device/VTEP across the network fabric, eliminating the requirement to have traditional hello protocols/packets across the network fabric. Consequently, the same gateway for a subnet can exist concurrently at multiple leafs, as needed, without the requirement for any FHRP-like protocols.

Redundant ToRs are reached over a Multi-Chassis Link Aggregation bundle (MC-LAG) technology such as vPC. Port channel hashing selects only one of the two available default gateways, and the "one to the nearest association" rule applies here. The VXLAN BGP EVPN network provides Layer 2 and Layer 3 services, and the default gateway association exists between the local edge device and the endpoint. When the endpoint tries to resolve the default gateway, the locally attached edge device is the only one that traps and resolves that ARP request. In this way, every edge device is responsible for performing default gateway functionality for its directly attached endpoints. The edge device also takes care of tracking the liveliness of its locally attached discovered endpoints by performing periodic ARP refreshes.

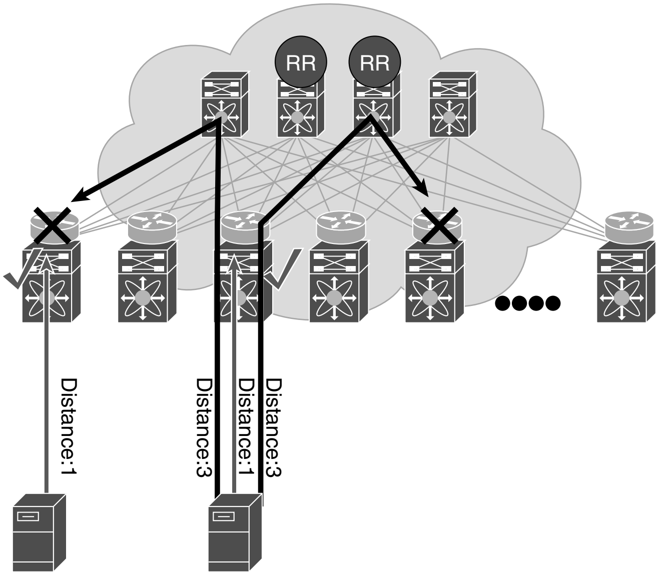


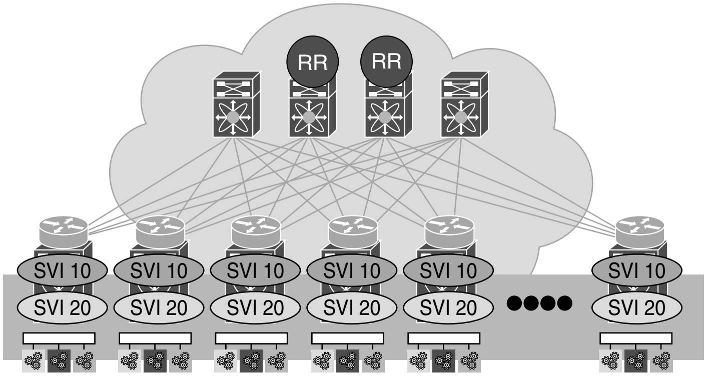
Figure 3-9 *Anycast Gateway Concept*

The scale-out implementation of the distributed anycast gateway provides the default gateway closest to each endpoint. The IP address for the default gateway is shared among all the edge devices, and each edge device is responsible for its respective IP subnet. In addition to the IP address of the default gateway being important, the associated MAC address is equally important. The MAC address is important as each endpoint has a local ARP cache that contains the specific IP-to-MAC binding for the default gateway.

For the host mobility use case, undesirable "black-holing" of traffic may occur if the gateway MAC address changes when a host moves to a server below a different ToR in a different rack even if the default gateway remains the same. To prevent this, the distributed anycast gateways on all the edge devices of the VXLAN EVPN fabric share the same MAC address for the gateway service. This shared MAC address, called the *anycast gateway MAC addresses* (*AGM*), is configured to be the same on all the edge devices. In fact, the same AGM is shared across all the different IP subnets, and each subnet has its own unique default gateway IP. The anycast gateway not only provides the most efficient egress routing but also provides direct routing to the VTEP where a given endpoint is attached, thereby eliminating hair-pinning of traffic.

In a situation where an endpoint is silent or undiscovered, no host route information for that endpoint in known to the BGP EVPN control plane. Without the host route information, the next-best route in the routing table should be used so that the packets destined to that endpoint are not dropped. Having each distributed IP anycast gateway advertise a subnet route from each VTEP where that subnet is locally instantiated allows a "next-best" route to be available in the routing table. A remote VTEP, which does not have that subnet instantiated locally, finds that the subnet prefix is reachable over multiple paths via ECMP. When traffic destined to an unknown/silent endpoint in this subnet is received by this remote VTEP, traffic is forwarded to one of the chosen VTEPs that serve the destination subnet based on the calculated ECMP hash. Once the traffic reaches that VTEP, the subnet prefix route is hit, which in turn points to a glean adjacency. This triggers the VTEP to send out an ARP request to the local-connected Layer 2 network, based on the destination subnet information. The ARP request is also sent to the Layer 3 core encapsulated with the Layer 2 VXLAN VNI associated with the destination subnet. As a result, the ARP request reaches the silent endpoint in the specific subnet. The endpoint responds to the ARP request, which in turn gets consumed at the VTEP that is directly attached to that endpoint. This is because all VTEPs share the same anycast gateway MAC. As a result, the endpoint is discovered, and the endpoint information is distributed by this VTEP into the BGP EVPN control plane.

An alternative to the subnet route advertisement approach is to allow the default route (0.0.0.0/0) to achieve similar results with the disadvantage of centralizing the discovery. However, a distributed approach based on subnet prefix routes scales much better to discover the silent endpoints. Figure 3-10 illustrates the logical realization of a distributed IP anycast gateway in a VXLAN BGP EVPN fabric.



SVI 10, Gateway IP: 192.168.1.1, Gateway MAC: 2020.0000.00AA
SVI 20, Gateway IP: 192.168.2.1, Gateway MAC: 2020.0000.00AA

Figure 3-10 *Subnet Prefix Advertisement for Distributed Anycast Gateway*

With the distributed IP anycast gateway implementation, any VTEP can service any subnet across the entire VXLAN BGP EVPN fabric. Integrated Routing and Bridging (IRB) from the VXLAN BGP EVPN construct provides the capability to efficiently route and bridge traffic. Regardless of the traffic patterns (east–west or north–south), the hairpinning of traffic is completely eliminated. The BGP EVPN control plane knows the identity of an endpoint along with the next-hop (VTEP) information that indicates its location. This leads to optimal forwarding in the network without requiring a large Layer 2 domain.

In summary, the AGM across all edge devices and subnets provides seamless host mobility as no changes to the endpoint ARP cache need to occur. This allows for "hot" workload mobility across the entire VXLAN fabric. The distributed anycast gateway moves the Layer 3 gateway to the leafs, thereby providing reduced failure domains, simplified configuration, optimized routing, and transparent endpoint/workload mobility.

Integrated Route and Bridge (IRB)

VXLAN BGP EVPN follows two different semantics for IRB that are documented and published in the IETF's draft-ietf-bess-evpn-inter-subnet-forwarding (https://tools.ietf.org/html/draft-ietf-bess-evpn-inter-subnet-forwarding).

Asymmetric IRB is the first of the documented first-hop routing operations. It follows a bridge–route–bridge pattern within the local edge device or VTEP. As the name implies, when asymmetric IRB is employed for routing, traffic egressing toward a remote VTEP uses a different VNI than the return traffic from the remote VTEP.

As noted in Figure 3-11, Host A, connected to VTEP V1, wants to communicate to Host X, connected to VTEP V2. Because Host A and Host X belong to different subnets, the asymmetric IRB routing process is used. Post default gateway ARP resolution, Host A sends data traffic toward the default gateway in VLAN 10. From VLAN 10, a routing operation is performed toward VLAN 20, which is mapped to VXLAN VNI 30002. The data traffic is encapsulated into VXLAN with VNI 30002. When the encapsulated traffic arrives on VTEP V2, it is decapsulated and then bridged over toward VLAN 20 because VLAN 20 is also mapped to VNI 30002 on VTEP V2.

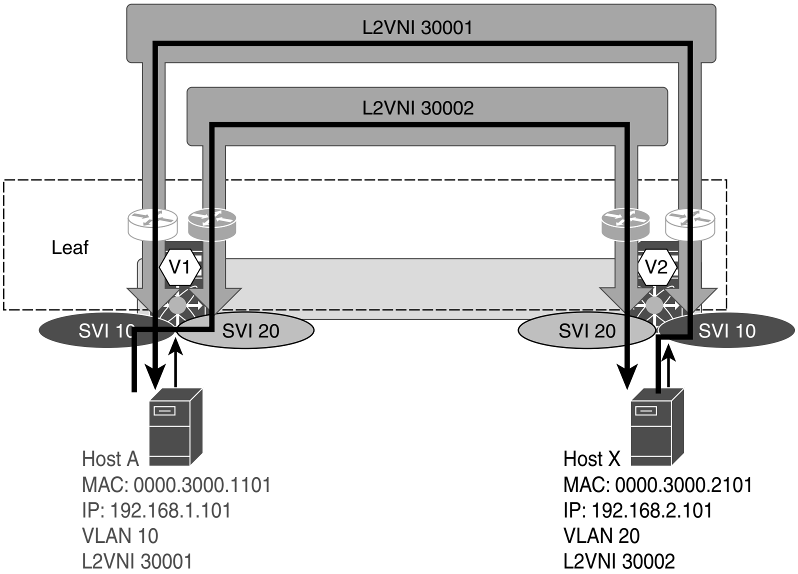


Figure 3-11 *Asymmetric IRB*

Host A–to–Host X communication results in a bridge–route–bridge pattern, with the encapsulated traffic traveling with VNI 30002. For the return traffic, Host X sends data traffic to the default gateway for the local subnet corresponding to VLAN 20. After the routing operation from VLAN 20 to VLAN 10 is performed, the traffic is encapsulated with VXLAN VNI 30001 and bridged toward VTEP V1. Once the traffic arrives at VTEP V1, the data traffic is decapsulated and bridged toward VLAN 10 because VLAN 10 is mapped to VNI 30001. Consequently, for return traffic from Host X to Host A, a bridge–route–bridge operation is performed, with encapsulated traffic traveling with VNI 30001. The end-to-end traffic flow from Host A to Host X uses VNI 30002, and the return traffic from Host X to Host A uses VNI 30001.

The preceding example demonstrates traffic asymmetry with different VNIs used for communication between Host A and Host X with asymmetric IRB. Asymmetric IRB requires consistent VNI configuration across all VXLAN VTEP(s) to prevent traffic from being black-holed. Configuration consistency is required for the second bridge operation in the bridge–route–bridge sequence because the sequence fails if there is a missing bridge-domain/VNI configuration corresponding to the network in which the destination resides. Figure 3-12 illustrates the fact that traffic from Host A and Host Y flows fine, but reverse traffic from Host Y to Host A does not work, due to the absence of the configuration for VNI 30001 at the VTEP attached to Host Y. Traffic between Host A and Host X will be correctly forwarded as both IRB interfaces (SVI 10 and SVI 20) are present on the attached VTEP.

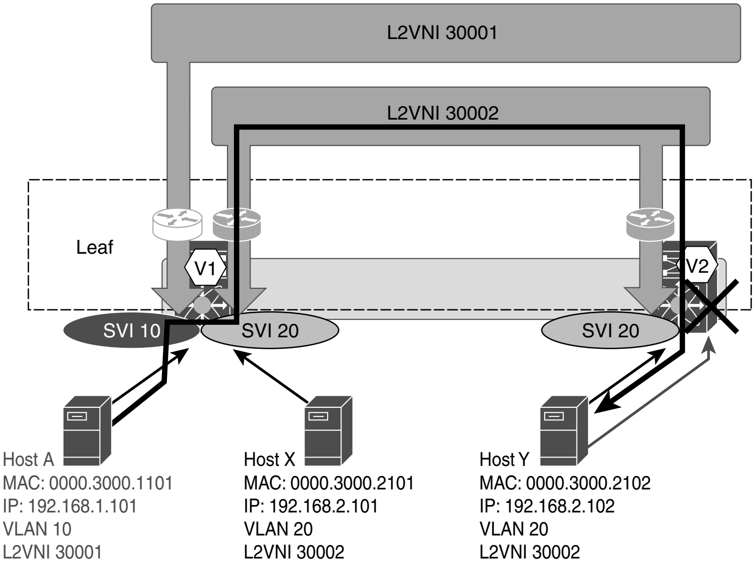


Figure 3-12 *Asymmetric IRB and Inconsistent Provisioning*

In addition to asymmetric IRB, another option for IRB operations is symmetric IRB. Whereas asymmetric IRB follows the bridge–route–bridge mode of operation, symmetric IRB follows the bridge–route–route–bridge mode of operation. Symmetric IRB provides additional use cases that are not possible with asymmetric IRB.

Symmetric IRB uses the same forwarding semantics when routing between IP subnets with VRF Lite or MPLS L3VPNs. With symmetric IRB, all traffic egressing and returning from a VTEP uses the same VNI. Specifically, the same Layer 3 VNI (L3VNI) associated with the VRF is used for all routed traffic. The distinctions between the L2VNI and L3VNI are identified in the BGP EVPN control plane specific to the VXLAN header 24-bit VNI field.

As noted in Figure 3-13, Host A is connected to VTEP V1 and wants to communicate with Host Y, attached to VTEP V2. Host A sends data traffic to the default gateway associated with the local subnet in VLAN 10. From VLAN 10, traffic is routed based on the destination IP lookup. The lookup result indicates which traffic needs to be VXLAN encapsulated and sends traffic toward VTEP V2, below which Host Y resides. The encapsulated VXLAN traffic is sent from VTEP V1 to VTEP V2 in VNI 50001, where 50001 is the Layer 3 VNI associated with the VRF in which Host A and Host Y reside. Once the encapsulated VXLAN traffic arrives at VTEP V2, the traffic is decapsulated and routed within the VRF toward VLAN 20 where Host Y resides. In this way, for traffic from Host A to Host Y, a bridge–route–route–bridge symmetric sequence is performed. Note that the Layer 2 VNIs associated with the networks in which Host A and Host Y reside are not used for the routing operation with the symmetric IRB option.

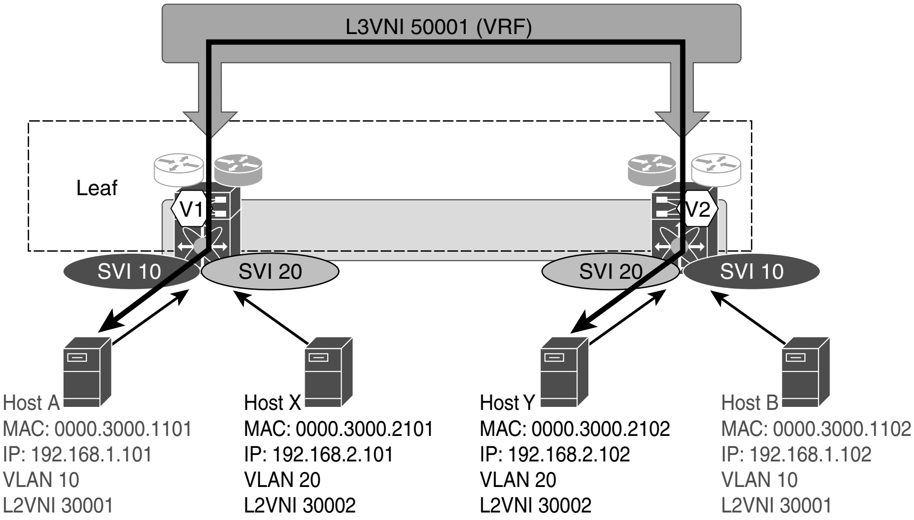


Figure 3-13 *Symmetric IRB*

For the return traffic from Host Y to Host A, the return flow is symmetric, using the same VNI 50001 associated with the VRF. Host Y sends the return traffic to the default gateway associated with its local subnet in VLAN 20. VLAN 20 makes a routing decision toward the VRF and VNI 50001, resulting in the VXLAN-encapsulated traffic being sent to VTEP V1. Once the encapsulated VXLAN traffic arrives at VTEP V1, the traffic is decapsulated and routed toward the VRF and VLAN 10. The return traffic flow from Host Y to Host A follows the same bridge–route–route–bridge sequence, following the same VNI 50001. The traffic from Host A to Host Y and traffic from Host Y to Host A leverage the same VRF VNI 50001, resulting in symmetric traffic flows. In fact, all routed traffic between hosts belonging to different networks within the same VRF employs the same VRF VNI 50001.

Symmetric IRB does not have the requirement of maintaining a consistent configuration across all the edge devices for all the VXLAN networks. In other words, scoped configuration can be implemented when it is not necessary to have all VNIs configured on all the edge devices or VTEPs (see Figure 3-14). However, for a given VRF, the same Layer 3 VNI needs to be configured on all the VTEPs since the VRF enables the sequence of bridge–route–route–bridge operation. For communication between hosts belonging to different VRFs, route leaking is required. For VRF route leaking, an external router or firewall is required to perform VRF-to-VRF communication. It should be noted that support for VRF route leaking requires software support to normalize the control protocol information with the data plane encapsulation. Next to local VRF route leaking, downstream VNI assignment can provide this function for the extranet use cases. These options are discussed further in Chapter 8, "External Connectivity."

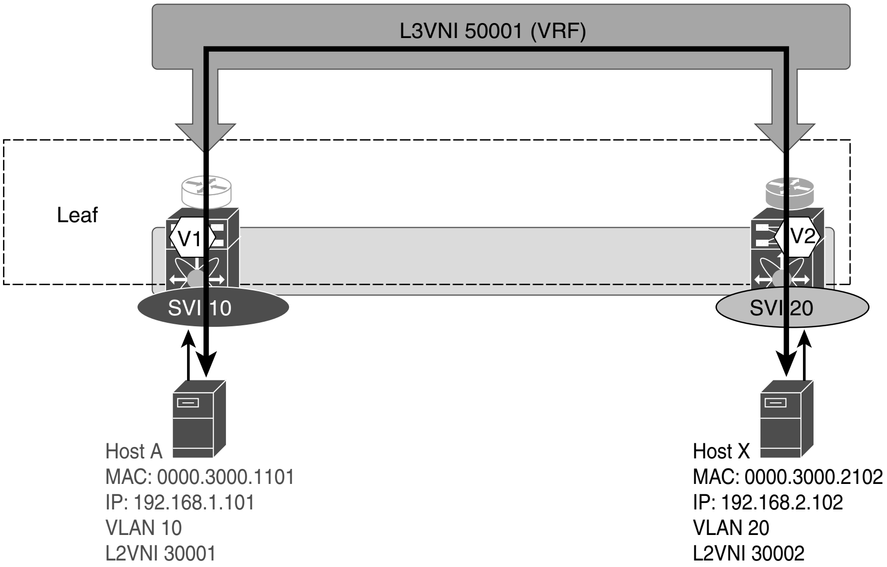


Figure 3-14 *Symmetric IRB and Inconsistent Provisioning*

Both symmetric and asymmetric modes are documented in the IETF draft. The Cisco implementation with NX-OS follows the symmetric IRB mode. The symmetric IRB bridge–route–route–bridge sequence offers flexibility for large-scale multitenant deployments. And the bridge–route–route–bridge sequence follows similar semantics to classic routing across a transit segment. The transit routing segment with VXLAN BGP EVPN is reflected by the Layer 3 VNI associated with the VRF.

Endpoint Mobility

BGP EVPN provides a mechanism to provide endpoint mobility within the fabric. When an endpoint moves, the associated host route prefix is advertised in the control plane through an updated sequence number. The sequence number avoids the need to withdraw and relearn a given prefix during an endpoint move. Instead, an update in the control plane of the new location is performed. Since forwarding in the fabric is always dictated by what is present in the BGP EVPN control plane, traffic is quickly redirected to the updated location of the endpoint that moved, thereby providing a smooth traffic convergence. When an endpoint moves, two host route prefixes for the same endpoint are present in the BGP EVPN control plane. The initial prefix is identified by the original VTEP location, and after the move, the prefix is identified by the new VTEP location.

With two prefixes in the control plane present for the same endpoint, a tiebreaker determines the winner. For this purpose, a BGP extended community called the MAC mobility sequence number is added to the endpoint host route advertisement (specifically Route type 2). The sequence number is updated with every endpoint move. In the event that the sequence number reaches its maximum possible value, the sequence number wraps around and starts over. The MAC mobility sequence number is documented in RFC 7432, which is specific to BGP EVPN.

The endpoint MAC address before the move is learned as a Route type 2 advertisement, and the BGP extended community MAC mobility sequence is set to 0. The value 0 indicates that the MAC address has not had a mobility event, and the endpoint is still at the original location. If a MAC mobility event has been detected, a new Route type 2 (MAC/IP advertisement) is added to the BGP EVPN control plane by the “new” VTEP below which the endpoint moved (its new location). The control plane then sets the MAC mobility sequence number to 1 in the BGP extended community as well as in the new location. There are now two identical MAC/IP advertisements in the BGP EVPN control plane. But only the new advertisement has the MAC mobility sequence number set to a value of 1. All the VTEPs honor this advertisement, thereby sending traffic toward the endpoint at its new location. Figure 3-15 illustrates a sample endpoint mobility use case with the relevant BGP EVPN state where endpoint Host A has moved from VTEP V1 to VTEP V3.

Route type	MAC, IP	L2VNI ("VLAN")	L3VNI ("VRF")	NH	Encap	Seq
2	0000.3000.1101, 192.168.1.101	30001	50001	10.200.200.1	8:VXLAN	0
2	0000.3000.1101, 192.168.1.101	30001	50001	10.200.200.3	8:VXLAN	1

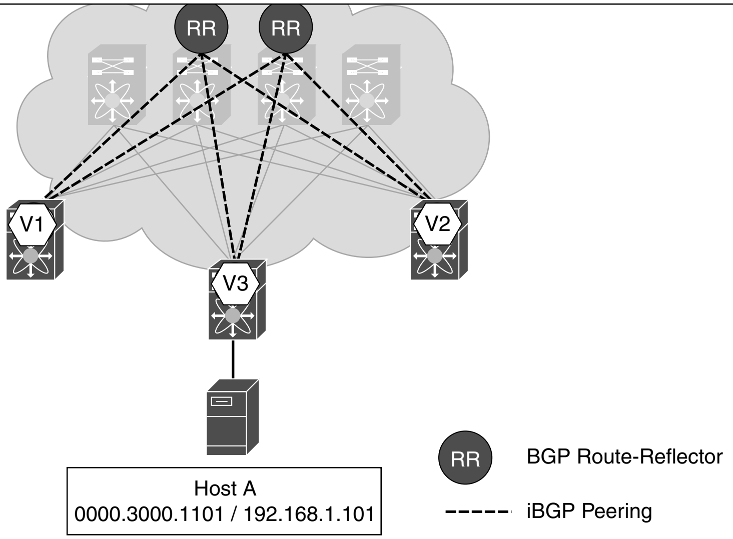


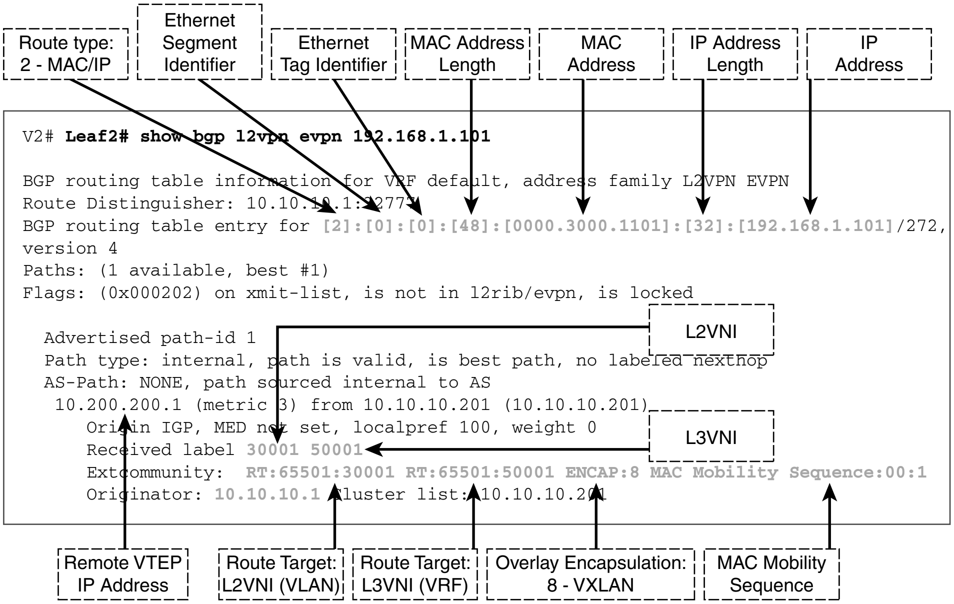
Figure 3-15 *Endpoint Mobility*

An endpoint may move multiple times over the course of its lifetime within a fabric. Every time the endpoint moves, the VTEP that detects its new location increments the sequence number by 1 and then advertises the host prefix for that endpoint into the BGP EVPN control plane. Because the BGP EVPN information is synced across all the VTEPs, every VTEP is aware of whether an endpoint is coming up for the first time or whether it is an endpoint move event based on the previous endpoint reachability information.

An endpoint may be powered off or become unreachable for a variety of reasons. If an endpoint has aged out of the ARP, MAC, and BGP tables, the extended community MAC mobility sequence is 0 as well. If the same endpoint reappears (from the fabric's and BGP EVPN's point of view), this reappearance is treated as a new learn event. In other words, the BGP EVPN control plane is aware of the current active endpoints and their respective locations, but the control plane does not store any history of the previous incarnations of the endpoints.

Whether an endpoint actually moves (a *hot* move) or an endpoint dies and another endpoint assumes its identity (a *cold* move), from the BGP EVPN control plane's point of view, the actions taken are similar. This is because the identity of an endpoint in the BGP EVPN control plane is derived from its MAC and/or IP addresses.

An endpoint move event results in either Reverse ARP (RARP) or Gratuitous ARP (GARP), signaling either by the endpoint or on behalf of the endpoint. In the case of RARP, the hypervisor or virtual switch typically sends out a RARP with the SMAC address set to the endpoint MAC address and the DMAC set to the broadcast MAC (FFFF.FFFF.FFFF). In the case of GARP, both the IP and MAC address of the endpoint are notified at the new location. This results in IP/MAC reachability of the endpoint being updated in the BGP EVPN control plane. On reception of this information at the old location (VTEP), an endpoint verification process is performed to determine whether the endpoint has indeed moved away from the old location. Once this validation occurs successfully, the old prefix with the old sequence number is withdrawn from the BGP EVPN control plane. In addition to the prefix being withdrawn from BGP EVPN, the ARP and/or MAC address tables are also cleared of the old location. Figure 3-16 illustrates a sample BGP EVPN output for a prefix associated with an end host that has moved within the fabric. In the figure, the fields of interest, including the MAC Mobility Sequence Number field, are highlighted.



```
V2# Leaf2# show bgp l2vpn evpn 192.168.1.101

BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 10.10.10.1:32777
BGP routing table entry for [2]:[0]:[0]:[48]:[0000.3000.1101]:[32]:[192.168.1.101]/272,
version 4
Paths: (1 available, best #1)
Flags: (0x000202) on xmit-list, is not in l2rib/evpn, is locked

  Advertised path-id 1
  Path type: internal, path is valid, is best path, no labeled nexthop
  AS-Path: NONE, path sourced internal to AS
   10.200.200.1 (metric 3) from 10.10.10.201 (10.10.10.201)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 30001 50001
      Extcommunity:  RT:65501:30001 RT:65501:50001 ENCAP:8 MAC Mobility Sequence:00:1
      Originator: 10.10.10.1 Cluster list: 10.10.10.201
```

Figure 3-16 *MAC Mobility Sequence*

During the verification procedure, if the endpoint responds at the old location as well, a potential duplicate endpoint is detected since the prefix is being seen at multiple locations. The default detection value for duplicate endpoint detection is “5 moves within

180 seconds." This means after the fifth move within the time frame of 180 seconds, the VTEP starts a "30-second hold" timer before restarting the verification and cleanup process. After the fifth time (5 moves within 180 seconds), the VTEP freezes the duplicate entry. With the Cisco NX-OS BGP EVPN implementation, these default detection values can be modified via user configuration.

By using the MAC mobility sequence numbers carried with the Route type 2 advertisement (MAC/IP advertisement), the BGP EVPN control plane can identify when a potential location change occurs for an endpoint. When a given prefix is determined to still be reachable, the control plane actively verifies whether a given endpoint has actually moved. The control plane provides a tremendous amount of value in not only verifying and cleaning up mobility events but also detecting duplicates within the VXLAN BGP EVPN network.

Virtual PortChannel (vPC) in VXLAN BGP EVPN

The bundling of multiple physical interfaces into a single logical interface between two chassis is referred to as a *port channel*, which is also known as a link aggregation group (LAG). Virtual PortChannel (vPC)[14] is a technology that provides Layer 2 redundancy across two or more physical chassis. Specifically, a single chassis is connected to two other chassis that are configured as a vPC pair. The industry term for this is Multi-Chassis Link Aggregation Group (MC-LAG). Recall that while the underlying transport network for VXLAN is built with a Layer 3 routing protocol leveraging ECMP, the endpoints still connect to the leafs or edge devices via classic Ethernet. It should be noted that the VXLAN BGP EVPN fabric does not mandate that endpoints must have redundant connections to the fabric. However, in most practical deployments, high availability is typically a requirement, and for that purpose, endpoints are connected to the edge devices via port channels.

Several protocols exist to form port channels, including static-unconditional configurations, protocols such as Port Aggregation Protocol (PAgP),[15] and industry standard protocols such as Link Aggregation Control Protocol (LACP).[16]

Port channels can be configured between a single endpoint and a single network switch. When ToR- or switch-level redundancy (as well as link-level redundancy) is a requirement, an MC-LAG is employed to provide the capability to connect an endpoint to multiple network switches.

vPCs allow interfaces of an endpoint to physically connect to two different network switches. From an endpoint perspective, they see a single network switch connected via a single port channel with multiple links. The endpoint connected to the vPC domain can be a switch, a server, or any other networking device that supports the IEEE 802.3 standard and port channels. vPC allows the creation of Layer 2 port channels that span two switches. vPCs consist of two vPC member switches connected by a peer link, with one being the primary and the other being the secondary. The system formed by the network switches is referred to as a *vPC domain* (see Figure 3-17).

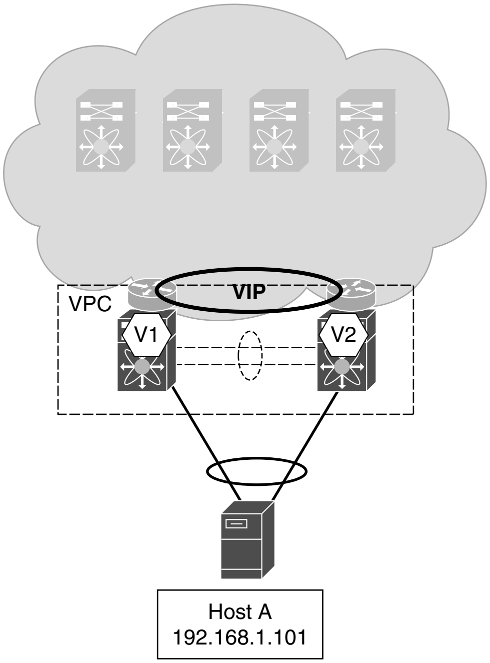
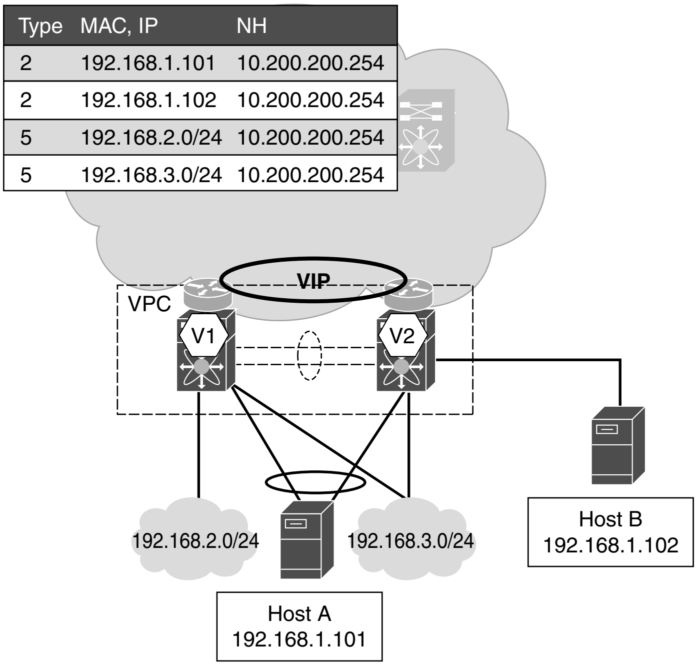


Figure 3-17 *Classic Ethernet vPC Domain*

The vPC primary and secondary members are initially configured as individual edge devices or VTEPs for VXLAN integration. For northbound traffic, the routed interfaces are part of the underlay network to provide reachability to each individual VTEP. Each VTEP is represented by an individual primary IP address (PIP) per VTEP.

With this in mind, the vPC feature first needs to be enabled, followed by the configuration of the vPC domain between the two vPC member network switches. Next, the vPC peer link (PL) and vPC peer keepalive (PKL) need to be connected between the two nodes. The underlay network also needs to be configured on the vPC peer link for both unicast and multicast (if multicast is used) underlay routing purposes.

A single VTEP is configured to represent the vPC domain in the VXLAN BGP EVPN fabric. To achieve the single VTEP, an anycast VTEP is configured with a common virtual IP address (VIP) that is shared across both switches that form the vPC domain. The anycast IP address is used by all the endpoints behind the vPC domain and is represented by a single anycast VTEP for the vPC domain.



Type	MAC, IP	NH
2	192.168.1.101	10.200.200.254
2	192.168.1.102	10.200.200.254
5	192.168.2.0/24	10.200.200.254
5	192.168.3.0/24	10.200.200.254

Figure 3-18 *vPC with VXLAN BGP EVPN*

A sample vPC domain is shown in Figure 3-18. VTEP V1 has IP address 10.200.200.1/32, and VTEP V2 has IP address 10.200.200.2/32. These are individual physical IP addresses (PIPs) on the NVE interface or the VTEP. A secondary IP address is added to the two VTEPs, which represents the VIP or anycast IP address (specifically 10.200.200.254/32). The secondary address depicts the location of any endpoint that is attached to the vPC pair. This address is the next hop advertised in the BGP EVPN control plane advertisement, representing the location of all local endpoints below a given vPC pair of switches.

The VIP is advertised from both vPC member switches so that both vPC members can receive traffic directly from any locally attached endpoints. Remote VTEPs can reach the VIP advertised by both the vPC member switches via ECMP over the underlay routed network. In this way, dual-attached endpoints can be reached as long as there is at least one path to any one of the vPC member switches or VTEPs.

In a vPC domain, both dual-attached endpoints and single-attached endpoints (typically referred to as orphans) are advertised as being reachable via the VIP associated with the anycast VTEP. Consequently, for orphan endpoints below a vPC member switch, a backup path is set up via the vPC peer link if the reachability to the spines is lost due to uplink failures. Hence, an underlay routing adjacency across the vPC peer link is recommended to address the failure scenarios.

If multicast is enabled on the underlay for carrying multidestination traffic, multicast routing (typically PIM) should also be enabled over the peer link. Of note, by default, the advertisement of an anycast VTEP VIP IP address as the next hop with BGP EVPN applies to the MAC/IP advertisements of Route type 2 as well as the IP prefix routes of Route type 5. In this way, all remote VTEPs can always see all BGP EVPN routes behind a vPC pair as being reachable via a single VTEP represented by the VIP. The VTEP associated with the VIP on the VPC peers of a given VPC domain is sometimes referred to as an anycast VTEP for that domain. If there are *N* leafs deployed in a VXLAN BGP EVPN fabric and all of them are configured as VPC peers, there are *N*/2 peers present in the network, represented by the respective anycast VTEPs, one per VPC domain.

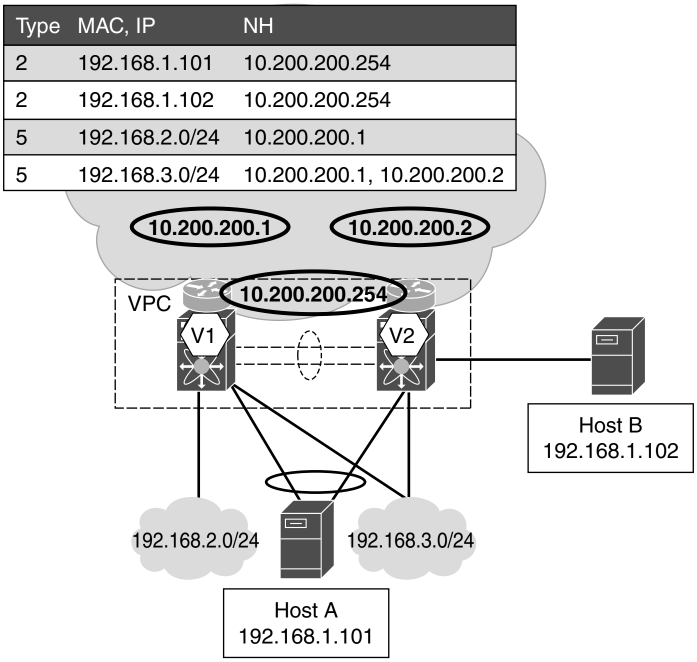
The term *anycast VTEP* should not be confused with the *anycast gateway*. Recall that the anycast gateway refers to the distributed IP anycast gateway for a given subnet that is shared by any and all leafs (including those configured as VPC peers) simultaneously with end host reachability information exchanged between the leafs via BGP EVPN. This is referred to as the *gateway* because end hosts within a subnet are configured with the default router set to the corresponding anycast gateway IP. Consequently, from an end host point of view, the default gateway's IP-to-MAC binding should remain the same, regardless of where the end host resides within the fabric. With Cisco NX-OS, all the anycast gateways also share the same globally configured anycast gateway MAC (AGM).

For a given vPC domain, since all ARPs, MACs, and ND entries are synced between the two vPC member switches, both switches are effectively advertising reachability for the same set of prefixes over BGP EVPN. A given vPC member switch thus ignores the BGP EVPN advertisements received from the adjacent vPC member switch since they are part of the same vPC domain because these advertisements are identified by an appropriate site-of-origin extended community attribute that they carry. In this way, only one VTEP or VIP needs to be advertised over BGP EVPN for all endpoints per vPC domain. For a VXLAN BGP EVPN network with *N* vPC pairs, each remote VTEP needs to know about only *N*–1 VTEP IPs or VIPs. Even with the presence of only one VIP acting as the anycast VTEP within a given vPC domain, the MP-BGP Route Distinguishers (RDs) are individual identifiers defined on a per-vPC member basis.

However, in some scenarios, IP prefixes may only be advertised by one of the two vPC member switches in a given vPC domain. For example, an individual loopback address may be configured under a VRF on each of the vPC member switches. In this case, because all reachability is advertised into BGP EVPN as being reachable via the anycast VTEP VIP, the traffic destined to a particular vPC member switch may reach its adjacent peer. After decapsulation, the traffic gets black-holed because it will suffer a routing table lookup miss within that VRF.

When IP address information is only advertised by one of the two vPC member switches, such problems may arise. This is because the return traffic needs to reach the correct vPC member switch. Advertising everything from the vPC domain as being reachable via the anycast VTEP IP address does not achieve this. This use case applies to scenarios having southbound IP subnets, orphan-connected IP subnets, individual loopback IP addresses, and/or DHCP relays. With all these use cases, Layer 3 routing adjacency on a per-VRF basis is required to ensure a routing exchange between the vPC domain members. This ensures that even if the packet reaches the incorrect vPC peer, after decapsulation the routing table lookup within the VRF does not suffer a lookup miss.

The configuration of this per-VRF routing adjacency between the vPC member switches may get a bit tedious. However, with BGP EVPN, an elegant solution to achieve similar behavior involves the use of a feature called *advertise-pip*. This feature is globally enabled on a per-switch basis. Recall that every vPC member switch has a unique PIP and a shared VIP address. With the advertise-pip feature enabled, all IP route prefix reachability is advertised from the individual vPC member switches using the PIP as the next hop in Route type 5 messages. Endpoint reachability corresponding to the IP/MAC addresses via Route type 2 messages continues to use the anycast VTEP or VIP as the next hop (see Figure 3-19). This allows per-switch routing advertisement to be individually advertised, allowing the return traffic to reach the respective VTEP in the vPC domain. In this way, a vPC member switch that is part of a given vPC domain also knows about individual IP prefix reachability of its adjacent vPC peer via BGP EVPN Route type 5 advertisements.



Type	MAC, IP	NH
2	192.168.1.101	10.200.200.254
2	192.168.1.102	10.200.200.254
5	192.168.2.0/24	10.200.200.1
5	192.168.3.0/24	10.200.200.1, 10.200.200.2

Figure 3-19 *Advertise PIP with vPC*

One additional discussion point pertains to the Router MAC carried in the extended community attributes of the BGP EVPN Route type 2 and Route type 5 messages. With advertise-pip, every vPC switch advertises two VTEPs' IP addresses (one PIP and one VIP). The PIP uses the switch Router MAC, and the VIP uses a locally derived MAC based on the VIP itself. Both of the vPC member switches derive the same MAC because they are each configured with the same shared VIP under the anycast VTEP. Because the Router MAC extended community is nontransitive, and the VIPs are unique within a given VXLAN BGP EVPN fabric, using locally significant router MACs for the VIPs is not a concern.

In summary, by using the advertise-pip feature with VXLAN BGP EVPN, the next-hop IP address of a VTEP is conditionally handled, depending on whether a Route type 2 (MAC/IP advertisement) or a Route type 5 (IP prefix route) is announced. Advertising PIP allows for efficient handling of vPC-attached endpoints as well as IP subnets specific to the use cases for Layer 2 and Layer 3 border connectivity or DHCP relays. Finally, it should be noted that some differences in the way vPC operates with the different Nexus platforms might exist. For an exhaustive list of the configuration required with VPC in VXLAN BGP EVPN environment for Nexus 9000 platform, please refer to Cisco's *Example of VXLAN BGP EVPN (EBGP)*.[17] And for Nexus 7000 and Nexus 5600 platforms, please refer to Cisco's *Forwarding configurations for Cisco Nexus 5600 and 7000 Series switches in the programmable fabric*.[18]

DHCP

Dynamic Host Configuration Protocol (DHCP)[19] is a commonly used protocol that provides IP address and other options for an endpoint. A DHCP server component is responsible for delivering IP address assignments and managing the binding between an endpoint MAC address and the provided IP address. In addition to distributing the IP address (DHCP scope) itself, additional options such as default gateway, DNS server, and various others (DHCP options) can be assigned to the endpoint that serves as a DHCP client. The requester in the case of DHCP is the DHCP client, which sends a discovery message to the DHCP server. The DHCP server then responds to this discover request with an offer message. Once this initial exchange is complete, a DHCP request (from client to server) is followed by a DHCP acknowledgment, and this completes the assignment of the IP address and subsequent information through the DHCP scope options. Figure 3-20 illustrates this DHCP process.

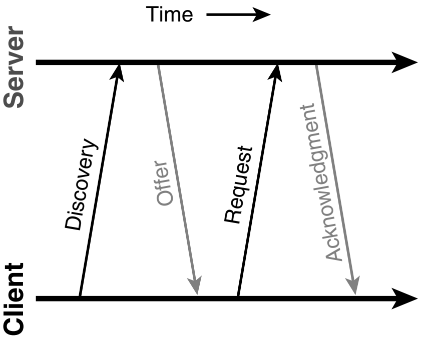


Figure 3-20 *DHCP Process*

Because the DHCP process relies on broadcasts to exchange information between DHCP client and DHCP server, the DHCP server needs to be in the same Layer 2 network in which the DHCP client resides. This is one deployment method for DHCP. In most cases, multiple IP subnets exist, and routers are in the path between a DHCP client and DHCP server. In this case, a DHCP relay agent is required. The DHCP relay agent is a DHCP protocol helper that snoops DHCP broadcasts and forwards these messages (specifically unicasts) to a specified DHCP server.

The DHCP relay is typically configured on the default gateway facing the DHCP client. The DHCP relay agent can support many different use cases for automated IP addressing, depending on whether the DHCP server resides in the same network, the same VRF, or a different VRF compared to the DHCP client. Having a per-tenant or per-VRF DHCP server is common in certain service provider data centers where strict segregation between tenants is a mandatory requirement. On the other hand, in enterprise environments, having a centralized DHCP server in a shared-services VRF is not uncommon. Traffic crosses tenant (VRF) boundaries with the DHCP server and DHCP client in a different VRF(s). In general, we can differentiate between three DHCP server deployment modes:

- **DHCP client and DHCP server within the same Layer 2 segment (VLAN/L2 VNI)**
  - No DHCP relay agent required
  - DHCP discovery from DHCP client is broadcast in the local Layer 2 segment (VLAN/L2 VNI)
  - As DHCP uses broadcasts, multidestination replication is required
- **DHCP client and DHCP server in different IP subnets but in the same tenant (VRF)**
  - DHCP relay agent required
  - DHCP discovery from DHCP client is snooped by the relay agent and directed to the DHCP server
- **DHCP client and DHCP server are in different IP subnets and different tenants (VRF)**
  - DHCP relay agent required with VRF selection
  - DHCP discovered from DHCP client is snooped by the relay agent and directed to the DHCP server into the VRF where the DHCP server resides

Two of the three DHCP server deployment modes relay data with a DHCP relay agent.[20] The DHCP server infrastructure is a shared resource in a multitenant network, and multitenancy support for DHCP services needs to be available. The same DHCP relay agent is responsible for handling the DHCP discovery message and DHCP request message from the DHCP client to the DHCP server and the DHCP offer message and DHCP acknowledgment message from the DHCP server to the requesting DHCP client.

Typically, the DHCP relay agent is configured at the same place where the default gateway IP address is configured for a given subnet. The DHCP relay agent uses the default gateway IP address in the GiAddr field in the DHCP payload for all DHCP messages that are relayed to the DHCP server. The GiAddr field in the relayed DHCP message is used for subnet scope selection so that a free IP address is picked from this subnet at the DHCP server and the respective DCHP options are assigned to the offer that will be returned from the server to the relay agent. The GiAddr field indicates the address to which the DHCP server sends out the response (that is, the DHCP offer or acknowledgment).

Additional considerations need to be taken into account when implementing the DHCP relay agent with the distributed IP anycast gateway in a VXLAN BGP EVPN fabric. This is because the same IP address is shared by all VTEPs that provide Layer 3 service in the form of a default gateway for a given network. Likewise, the same GiAddr field is consequently stamped in DHCP requests relayed by these VTEPs. As a result, the DHCP response from the DHCP server may not reach the same VTEP that had relayed the initial request due to the presence of an anycast IP. In other words, the response may go to any one of the VTEPs servicing the anycast IP, which is clearly undesirable (see Figure 3-21). If a unique IP address per VTEP exists, then that can be used in the GiAddr field, thereby ensuring that the server response is guaranteed to come back to the right VTEP. This is achieved via specification of an **ip dhcp relay source-interface** ***xxx*** command under the Layer 3 interface configured with the anycast gateway IP.

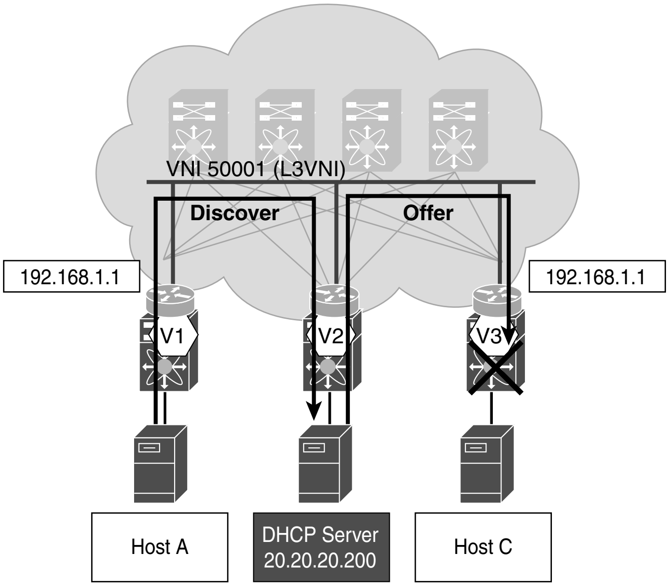


Figure 3-21 *GiAddr Problem with Distributed IP Anycast Gateway*

In this way, the GiAddr field is modified to carry the IP address of the specified source interface. Any given interface on the network switch can be used, as long as the IP address is unique within the network used for reaching the DHCP server. In addition, the IP address must be routable, so that the DHCP server must be able to respond to the individual and unique DHCP relay IP address identified by the GiAddr field. By accomplishing this, the DHCP messages from the client to the server are ensured to travel back and forth through the same network switch or relay agent.

It should be noted that typically the GiAddr field also plays a role in the IP subnet scope selection from which a free IP address is allocated and returned to the DHCP client. Because the GiAddr field has been changed based on the source interface specification, it no longer represents the DHCP scope selection and respective IP address assignment.

As a result, a different way to enable scope selection is required. DHCP option 82 (a vendor-specific option) was designed to put in circuit-specific information, and has been used in different deployments (see Figure 3-22). DHCP option 82 has two suboptions that derive the required IP subnet for the IP address assignment, which provides a different approach for DHCP subnet scope selection.

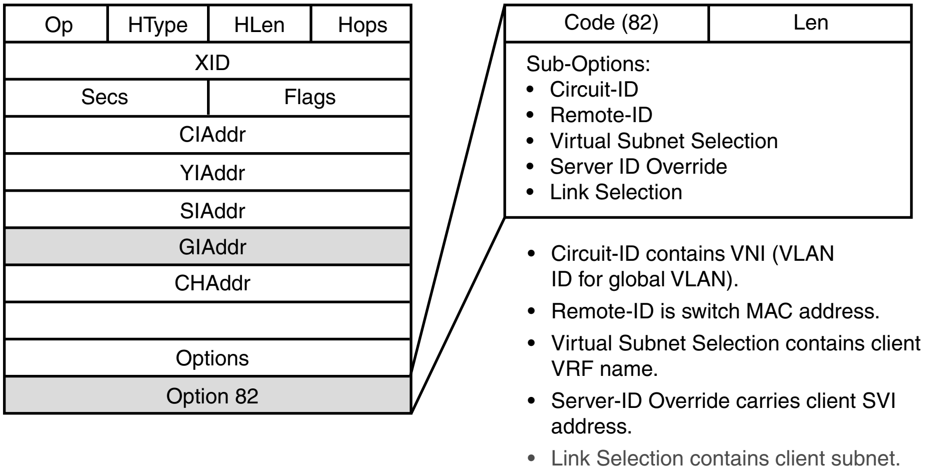

Op
HType
HLen
Hops
XID
Secs
Flags
CIAddr
YIAddr
SIAddr
GIAddr
CHAddr
Options
Option 82
Code (82)
Len
Sub-Options:
• Circuit-ID
• Remote-ID
• Virtual Subnet Selection
• Server ID Override
• Link Selection
• Circuit-ID contains VNI (VLAN ID for global VLAN).
• Remote-ID is switch MAC address.
• Virtual Subnet Selection contains client VRF name.
• Server-ID Override carries client SVI address.
• Link Selection contains client subnet.

Figure 3-22 *DHCP Option 82*

With DHCP option 82 with the Circuit-ID suboption, the VLAN or VNI information associated with the network in which the client resides is added to the DHCP messages, which provide this information to the DHCP server. The DHCP server needs to support the functionality to select the right DHCP scope, based on the Circuit-ID suboption. DHCP servers that support the DHCP option 82 with the Circuit-ID suboption include Microsoft's Windows 2012 DHCP server, ISC DHCP server (dhcpd), and Cisco Prime Network Registrar (CPNR).

With the second option, which is the preferred option, the DHCP scope selection is performed using the Link Selection suboption of DHCP option 82. Within the Link Selection suboption, the original client subnet is carried, thereby allowing the correct DHCP scope selection. DHCP servers that support the Link Selection suboption include dhcpd, Infoblox's DDI, and CPNR. For an exhaustive list of the DHCP relay configuration required in a VXLAN BGP EVPN network, refer to the vPC VTEP DHCP relay configuration example at Cisco.com.[21]

With the use of the distributed anycast gateway in a VXLAN BGP EVPN fabric and its multitenant capability, centralized DHCP services simplify the ability to provide IP address assignments to endpoints. With a unique source IP address for relaying the DHCP message, and with DHCP option 82, the IP address assignment infrastructure can be centrally provided even across the VRF or tenant boundary.

Summary

This chapter provides an in-depth discussion of the core forwarding capabilities of a VXLAN BGP EVPN fabric. For carrying broadcast, unknown unicast, and multicast (BUM) traffic, the chapter examines both multicast and ingress replication, coupled with the inclusive multicast Route type support with BGP EVPN. It also discusses enhanced forwarding features that allow for reducing ARP and unknown unicast traffic within the fabric. Perhaps the key benefit of a BGP EVPN fabric is the realization of a distributed anycast gateway at the ToR or leaf layer that allows for optimal forwarding of both Layer 2 and Layer 3 traffic within the fabric, using IRB semantics. Efficient handling of endpoint mobility with BGP EVPN is also described, as well as how to handle dual-attached endpoints with vPC-based deployments. Finally, the chapter concludes with a description of how DHCP relay functionality can be implemented in environments with distributed IP anycast gateway.

References

1. Internet Society. *Internet-drafts (I-Ds)*. www.ietf.org/id-info/.

2. Internet Society. *Request for comments (RFC)*. www.ietf.org/rfc.html.

3. Network Working Group. *DHCP relay agent information option*. 2001. tools.ietf.org/html/rfc3046.

4. Network Working Group. *Protocol Independent Multicast–Sparse Mode (PIM-SM) protocol specification (revised)*. 2006. tools.ietf.org/html/rfc4601.

5. Network Working Group. *Bidirectional Protocol Independent Multicast (BIDIR-PIM)*. 2007. tools.ietf.org/html/rfc5015.

6. Network Working Group. *Neighbor discovery for IP version 6 (IPv6)*. 2007. tools.ietf.org/html/rfc4861.

7. Rabadan, J., et al. *Operational aspects of proxy-ARP/ND in EVPN networks*. 2016. www.ietf.org/id/draft-ietf-bess-evpn-proxy-arp-nd-01.txt.

8. Cisco. *Proxy ARP*. 2008. www.cisco.com/c/en/us/support/docs/ip/dynamic-address-allocation-resolution/13718-5.html.

9. Network Working Group. *Cisco's Hot Standby Router Protocol (HSRP)*. 1998. www.ietf.org/rfc/rfc2281.txt.

10. Network Working Group. *Virtual Router Redundancy Protocol (VRRP)*. 2004. tools.ietf.org/html/rfc3768.

11. Cisco. *Cisco GLBP load balancing options*. 1998. www.cisco.com/c/en/us/products/collateral/ios-nx-os-software/ip-services/product_data_sheet0900aecd803a546c.html.

12. Cisco. *Anycast HSRP*. 2016. www.cisco.com/c/en/us/td/docs/switches/datacenter/sw/6_x/nx-os/fabricpath/configuration/guide/b-Cisco-Nexus-7000-Series-NX-OS-FP-Configuration-Guide-6x/b-Cisco-Nexus-7000-Series-NX-OS-FP-Configuration-Guide-6x_chapter_0100.html#concept_910E7F7E592D487F84C8EE81BC6FC14F.

13. Cisco. *VXLAN network with MP-BGP EVPN control plane design guide*. 2015. www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/guide-c07-734107.html#_Toc414541688.

14. Cisco. *Cisco NX-OS software Virtual PortChannel: Fundamental concepts 5.0*. 2014. www.cisco.com/c/en/us/products/collateral/switches/nexus-5000-series-switches/design_guide_c07-625857.html.

15. Finn, N. *Port Aggregation Protocol*. 1998. www.ieee802.org/3/trunk_study/april98/finn_042898.pdf.

16. Network Working Group 802.1. *802.1AX-2008—IEEE standard for local and metropolitan area networks—Link aggregation*. 2008. standards.ieee.org/findstds/standard/802.1AX-2008.html.

17. Cisco. *Example of VXLAN BGP EVPN (EBGP)*. 2016. www.cisco.com/c/en/us/td/docs/switches/datacenter/nexus9000/sw/7-x/vxlan/configuration/guide/b_Cisco_Nexus_9000_Series_NX-OS_VXLAN_Configuration_Guide_7x/b_Cisco_Nexus_9000_Series_NX-OS_VXLAN_Configuration_Guide_7x_chapter_0100.html#concept_53AC00F0DE6E40A79979F27990443953.

18. Cisco. *Forwarding configurations for Cisco Nexus 5600 and 7000 Series switches in the programmable fabric*. 2016. www.cisco.com/c/en/us/td/docs/switches/datacenter/pf/configuration/guide/b-pf-configuration/Forwarding-Configurations.html#concept_B54E8C5F82C14C7A9733639B1A560A01.

19. Network Working Group. *Dynamic Host Configuration Protocol.* 1997. www.ietf.org/rfc/rfc2131.txt.

20. Network Working Group. *DHCP relay agent information option*. 2001. tools.ietf.org/html/rfc3046.

21. Cisco. *vPC VTEP DHCP relay configuration example*. 2016. www.cisco.com/c/en/us/td/docs/switches/datacenter/nexus9000/sw/7-x/vxlan/configuration/guide/b_Cisco_Nexus_9000_Series_NX-OS_VXLAN_Configuration_Guide_7x/b_Cisco_Nexus_9000_Series_NX-OS_VXLAN_Configuration_Guide_7x_appendix_0110.html#id_16000.

Appendix A

VXLAN BGP EVPN Implementation Options

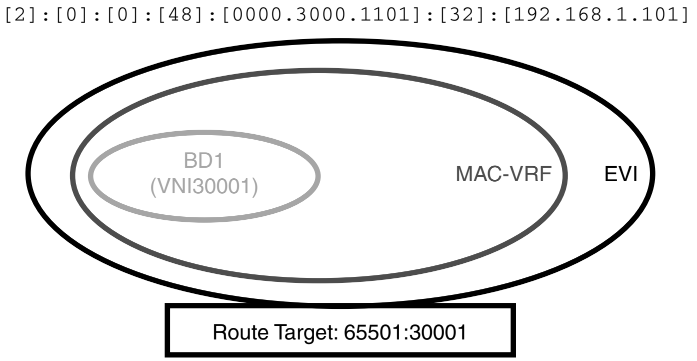
This book encompasses details on building data center fabrics with open standards, specifically VXLAN, BGP, and EVPN. Although the primary focus is specific to the Cisco NX-OS implementation, the book also explains the fundamentals associated with the base technologies of VXLAN, BGP, and EVPN. Nevertheless, with open standards, the associated documented definitions in IETF contain some prescribed implementation options and additional functionality that will be briefly discussed here.

This appendix goes over some specific EVPN implementation options that are mentioned as part of RFC 7432 BGP MPLS-Based Ethernet VPN[1], the EVPN Overlay draft-ietf-bess-evpn-overlay[2] draft, and the EVPN Prefix Advertisement draft-ietf-bess-evpn-prefix-advertisement[3] draft. The option for EVPN inter-subnet forwarding specified in draft-ietf-bess-evpn-inter-subnet-forwarding[4] that relates to symmetric and asymmetric Integrated Route and Bridge (IRB) options is excluded as it has been covered in detail in Chapter 3, “VXLAN/EVPN Forwarding Characteristics,” under the IRB discussion.

EVPN Layer 2 Services

EVPN defines different VLAN Layer 2 Services, as mentioned in RFC7432 Section 6. There are two options mentioned in Section 5.1.2 in draft-ietf-bess-evpn-overlay, which apply to the VXLAN-related implementation. The first option, implemented by Cisco, is called *VLAN-based* bundle service interface. The second option is called *VLAN-aware* bundle service interface.

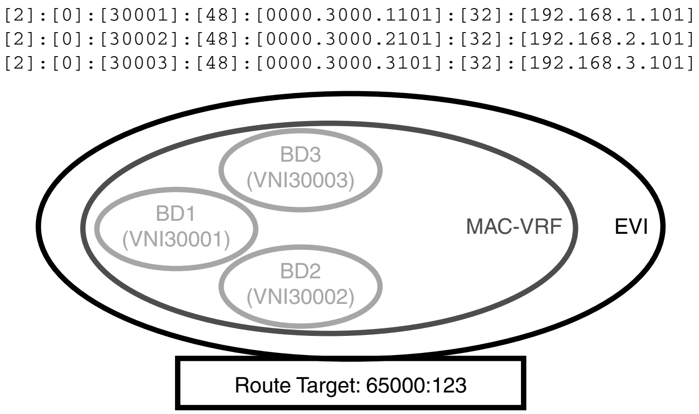
In the VLAN-based option, a *single* bridge domain is mapped to a *single* EVPN Virtual Instance (EVI). The EVI provides the Route Distinguisher (RD) and controls the import and export of the associated prefixes via Route Targets (RT) into the MAC-VRF and in turn into the bridge domain (see Figure A-1). When using the VLAN-based approach, the EVI corresponds to a single MAC-VRF in the control plane and a single VNI in the data plane, resulting in a 1:1 mapping of EVI, MAC-VRF, and bridge domain (VNI). The disadvantage of the VLAN-based implementation is the configuration requirement of one EVI per bridge domain. This becomes an advantage for import/export granularity on a per bridge domain basis. Example 5-3 in Chapter 5, “Multitenancy,” provides a configuration example for the Cisco VLAN-based implementation.



```
[2]:[0]:[0]:[48]:[0000.3000.1101]:[32]:[192.168.1.101]
```

Figure A-1 *VLAN-Based Bundle Services*

In the case of the VLAN-aware bundle service interface, multiple bridge domains can be mapped to a single EVI. The VLAN-aware implementation allows a single RD and RT set to be shared across multiple bridge domains that belong to the same EVI. This results in a 1:N mapping for the EVI to MAC-VRF to bridge domain (see Figure A-2); resulting in the VNI in the data plane being sufficient to identify the corresponding bridge domain. The VLAN-aware implementation reduces the configuration requirements for EVIs. The disadvantage of the VLAN-aware implementation is that it does not allow the granularity of prefix import/export on a per bridge domain basis.



```
[2]:[0]:[30001]:[48]:[0000.3000.1101]:[32]:[192.168.1.101]
[2]:[0]:[30002]:[48]:[0000.3000.2101]:[32]:[192.168.2.101]
[2]:[0]:[30003]:[48]:[0000.3000.3101]:[32]:[192.168.3.101]
```

Figure A-2 *VLAN-Aware Bundle Services*

The big difference between the VLAN-based and the VLAN-aware implementation is the use of the Ethernet Tag ID field in the control plane. The VLAN-based option requires this field to be zero (0). The VLAN-aware option specifies the Ethernet Tag ID must carry the identification of the respective bridge domain (VNI). The difference in Ethernet Tag ID usage is described in RFC7432 section 6.1: VLAN-Based Service Interface and

6.3: VLAN-Aware Service Interface. It is important to understand that both options are valid and conform to RFC 7432 but are not per se interoperable.

EVPN IP-VRF to IP-VRF Model

In some cases, IP prefix routes may be advertised for subnets and IPs behind an IRB. This use case is referred to as the "IP-VRF to IP-VRF" model. As part of the EVPN prefix-advertisement draft, draft-ietf-bess-evpn-prefix-advertisement, there are three implementation options for the IP-VRF to IP-VRF model. The draft provides two required components and one optional component. For draft conformity, it is necessary to follow one of the two required component models. The main focus here is on the two required models; the optional model will be touched upon briefly towards the end. Section 5.4 in draft-ietf-bess-evpn-prefix-advertisement describes the three models, which will be briefly introduced in this section.

The first model is called the *interface-less* model where the Route type 5 route contains the necessary information for an IP Prefix advertisement. The Cisco NX-OS implementation uses the interface-less model. Within the IP Prefix advertisement, all the IP routing information is included-such as the IP subnet, IP subnet length, and the next hop. In addition, the IP-VRF context is preserved in form of the Layer 3 VNI present in the Network Layer Reachability Information (NLRI). The interface-less model also includes the Router MAC of the next hop as part of the BGP extended community. With VXLAN being a MAC in IP/UDP encapsulation, the data plane encapsulation requires population of the inner source and DMAC addresses. Although the local router MAC is known and used as the inner-source MAC, the DMAC address needs to be populated based on the lookup of the destination prefix. The Router MAC extended community will provide this necessary information, as shown in Figure A-3. Note that with the interface-less model, the Gateway IP (GW IP) field is always populated to 0. Example 5-9 in Chapter 5 provides a configuration example for Cisco's interface-less implementation.

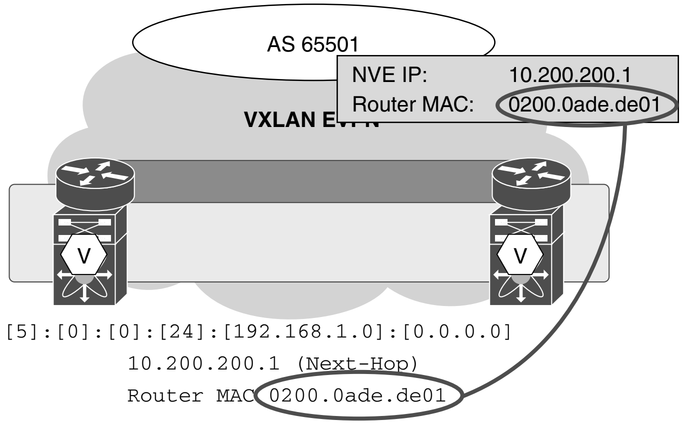
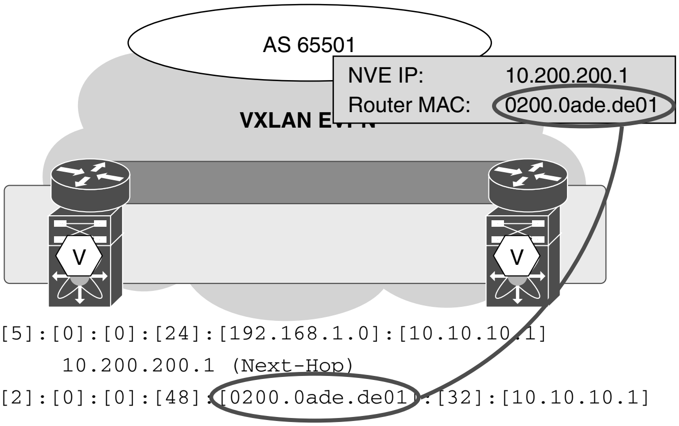


Figure A-3 *IP-VRF Interface-Less Model*

The second model is called *interface-full*, which has two sub-modes:

- Core-facing IRB
- Unnumbered Core-facing IRB

In both cases of the interface-full model, in addition to the Route type 5 prefix advertisement, a Route type 2 advertisement is also generated. With the interface-full core-facing IRB option, the Router MAC extended community is not part of the Route type 5 advertisement. Instead, the Gateway IP (GW IP) field is populated to be that of the core-facing IRB associated with the VTEP. In addition, the VNI field in the Route type 5 advertisement is set to 0. To complete the information required for the VXLAN encapsulation, a Route type 2 advertisement is generated, as shown in Figure A-4. The Route type 2 advertisement provides the IP, as well as MAC information for the next-hop core-facing IRB interface that in turn is used for the population of the VXLAN header. In addition, the VNI field corresponding to the tenant or IP-VRF is also carried in the Route type 2 advertisement.



```
[5]:[0]:[0]:[24]:[192.168.1.0]:[10.10.10.1]
     10.200.200.1 (Next-Hop)
[2]:[0]:[0]:[48]:[0200.0ade.de01]:[32]:[10.10.10.1]
```

Figure A-4 *IP-VRF Interface-Full Model with Core-Facing IRB*

In the optional interface-full model of unnumbered core-facing IRB (see Figure A-5), the Router MAC extended community is sent as part of the Route type 5 advertisement in a similar way as the interface-less model. However, the VNI field in the advertisement remains 0, as does the GW IP address field. The associated Route type 2 advertisement is keyed by the Router MAC address associated with the next-hop core-facing IRB interface and also carries the VNI associated with the tenant or IP-VRF. In this way, when traffic is destined to the Route type 5 advertised prefix, there is a recursive lookup performed to verify the next-hop router MAC information, and it is employed to populate the corresponding VXLAN header.

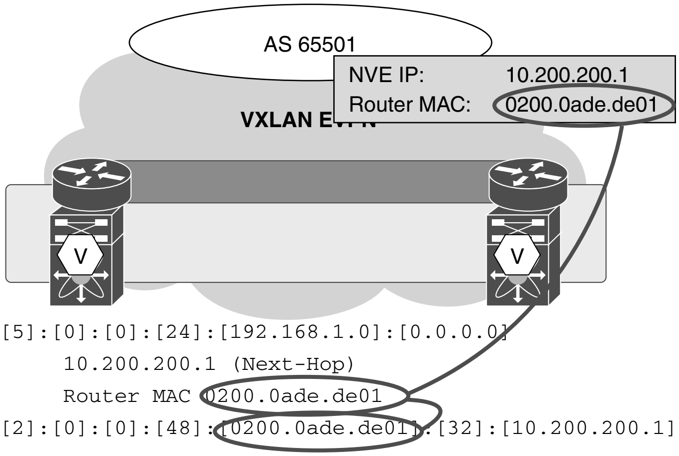

AS 65501
NVE IP: 10.200.200.1
Router MAC: 0200.0ade.de01
VXLAN EVPN
V
V
[5]:[0]:[0]:[24]:[192.168.1.0]:[0.0.0.0]
10.200.200.1 (Next-Hop)
Router MAC 0200.0ade.de01
[2]:[0]:[0]:[48]:[0200.0ade.de01]:[32]:[10.200.200.1]

Figure A-5 *IP-VRF Interface-Full Model with Unnumbered Core-Facing IRB*

The difference between the interface-less and the interface-full model is mainly the presence or absence of an additional Route type 2 advertisement. The interface-less option expects the Router MAC in the Route type 5 advertisement and will not leverage the Route type 2 information advertised by an interface-full core-facing IRB model. Similarly, an interface-full option would expect the additional Route type 2 advertisement. In both cases, the VXLAN data plane encapsulation would have the information necessary for populating the inner DMAC address.

In summary, the different IP-VRF to IP-VRF models are described in the EVPN prefix advertisement draft, draft-ietf-bess-evpn-prefix-advertisement in section 5.4. It is important to understand what is required for each model so that they conform to the IETF draft; the different models are not per-se interoperable.

References

1. *BGP MPLS-Based Ethernet VPN*. IETF Website, 2015. Retrieved from https://tools.ietf.org/html/rfc7432.

2. *A Network Virtualization Overlay Solution using EVPN*. IETF Website, 2016. Retrieved from https://tools.ietf.org/html/draft-ietf-bess-evpn-overlay.

3. *IP Prefix Advertisement in EVPN*. IETF Website, 2016. Retrieved from https://tools.ietf.org/html/draft-ietf-bess-evpn-prefix-advertisement.

4. *Integrated Routing and Bridging in EVPN*. IETF Website, 2015. Retrieved from https://tools.ietf.org/html/draft-ietf-bess-evpn-inter-subnet-forwarding.

Index