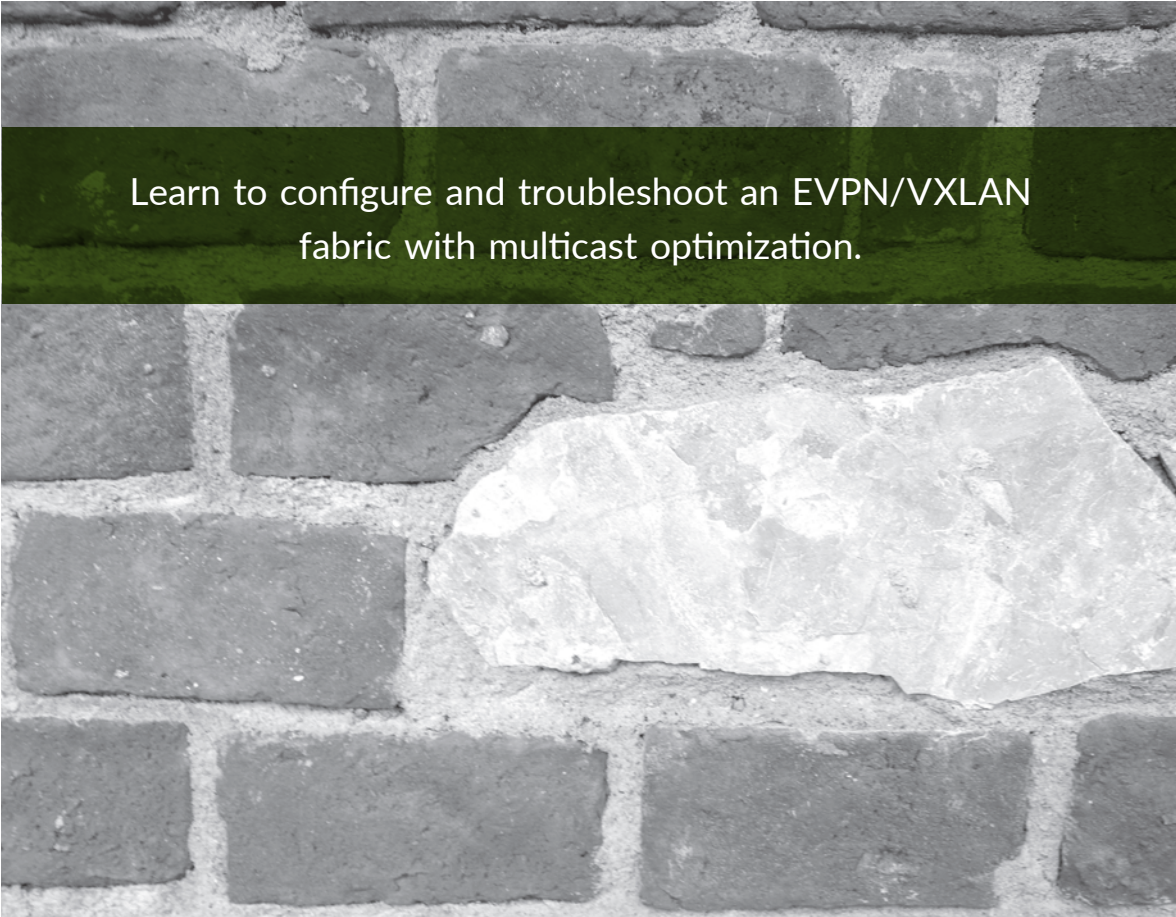


DAY ONE: DEPLOYING OPTIMIZED MULTICAST IN EVPN/VXLAN



Learn to configure and troubleshoot an EVPN/VXLAN
fabric with multicast optimization.

By Vikram Nagarajan, Princy Elizabeth, and Himanshu Agarwal

DAY ONE: DEPLOYING OPTIMIZED MULTICAST IN EVPN/VXLAN

This *Day One* book provides a quick background on multicast and EVPN technologies and then delves into EVPN BUM forwarding rules within a VLAN, and the problems of 'flooding everywhere'. Various multicast optimization procedures such as Assisted Replication and Selective (SMET) Forwarding are discussed in a staged manner. The authors document the procedures of solving problems in multihoming, the cornerstone of EVPN, and the benefits are illustrated quantitatively with examples to describe the advantages of various optimizations. Filled with network diagrams and configuration examples, *Day One: Deploying Optimized Multicast in EVPN/VXLAN* is a must read for optimizing multicast in modern network fabrics.

"The authors of this book have taken two famously complex protocols, made them work together, and explained the results with a clarity and a precision rarely seen in this industry. You'll put this book down with a renewed energy to take your EVPN and multicast skills to the next level. A true gift to the network engineering community."

Chris Parker, JNCIE-SP #2981, Senior Network Engineer at Nominet, Juniper Ambassador

IT'S DAY ONE AND YOU HAVE A JOB TO DO, SO LEARN:

- How multicast works in an EVPN-VXLAN fabric.
- How multicast traffic is routed from one VLAN to another.
- How Assisted Replication (AR) helps improve resource utilization and link-utilization.
- How to appreciate the optimization on access-side interfaces with IGMP-Snooping.
- How to appreciate the optimization in EVPN core using Selective (SMET) Forwarding.
- How to appreciate the optimizations used in conjunction with a large-scale EVPN multicast deployment with Inter-VLAN and External Multicast.
- Configure and troubleshoot optimized Intra-VLAN & Inter-VLAN Multicast.
- Configure and troubleshoot the devices to work with an External Multicast Source/Receiver sitting outside the EVPN Fabric.
- Configure and Troubleshoot an EVPN-VXLAN Fabric with multicast optimization enabled with Assisted-Replication.



Juniper Networks Books are focused on network reliability and efficiency. Peruse the complete library at www.juniper.net/dayone.

JUNIPER
NETWORKS

Day One: Deploying Optimized Multicast in EVPN/VXLAN

by Vikram Nagarajan, Princy Elizabeth, and Himanshu Agarwal

Part 1: Background

<i>Chapter 1: Multicast Primer</i>	7
<i>Chapter 2: EVPN Primer</i>	16
<i>Chapter 3: EVPN Base Configuration in DC Fabric Topology</i>	26

Part 2: EVPN Intra-VLAN Multicast

<i>Chapter 4: EVPN Intra-VLAN Multicast Without Optimization</i>	32
<i>Chapter 5: Assisted Replication</i>	52
<i>Chapter 6: EVPN Intra-VLAN Multicast with Optimization</i>	65
<i>Chapter 7: EVPN Intra-Multicast Optimization with Multihoming</i>	86
<i>Chapter 8: Assisted Replication with SMET</i>	108

Part 3: EVPN Inter-VLAN Multicast

<i>Chapter 9: EVPN Inter-VLAN Multicast Routing without Optimization</i>	124
<i>Chapter 10: EVPN Inter-VLAN Multicast Routing with Optimization</i>	134
<i>Chapter 11: External Multicast with PIM IRB</i>	151

<i>Appendix: Base Configurations</i>	176
--	-----

© 2020 by Juniper Networks, Inc. All rights reserved.

Juniper Networks and Junos are registered trademarks of Juniper Networks, Inc. in the United States and other countries. The Juniper Networks Logo and the Junos logo, are trademarks of Juniper Networks, Inc. All other trademarks, service marks, registered trademarks, or registered service marks are the property of their respective owners. Juniper Networks assumes no responsibility for any inaccuracies in this document. Juniper Networks reserves the right to change, modify, transfer, or otherwise revise this publication without notice.

Published by Juniper Networks Books

Authors: Vikram Nagarajan, Princy Elizabeth, Himanshu Agarwal

Technical Reviewers: Victor Ganjian, Sudarsanan C, Wen Lin, Stefano Anchini, Bob Hayes, Christoph Plum, Michal Styszynski, Aldrin Isaac, Cantemir Olaru, Anoop Kumar Sahu, Ramesh Kandula, Beth Montoya, Barbara Matsu-mura

Editor in Chief: Patrick Ames

Copyeditor: Nancy Koerbel

Printed in the USA by Vervante Corporation.

Version History: v1, Feb. 2020

2 3 4 5 6 7 8 9 10

About the Authors

Vikram Nagarajan is a Senior Staff Engineer with the Routing Protocols group in Juniper Networks, Bengaluru, India. Having worked for close to 15 years in various VPN and multicast technologies, he recently hopped onto the EVPN bandwagon to partake in solving complex problems in the data center. He was involved in pioneering EVPN multicast solutions from their early days. He likes to call himself “the Multicast Philosopher.”

Princy Elizabeth is a Senior Staff Engineer with Juniper Networks, Bengaluru, India. She has over 15 years of experience designing, implementing, and troubleshooting L2/L3 Multicast features. She has worked from “Day One” in the design and implementation of EVPN Multicast on Junos, and this Day One book is an attempt to share that EVPN multicast expertise with readers.

Himanshu Agarwal is currently a Staff Engineer in Juniper’s Routing Business Unit in Bengaluru, India. He has 13 years of experience in testing and troubleshooting a wide range of routing and Ethernet switching technologies. Himanshu holds a Bachelors of Technology in Electronics and Communications and a post-graduate diploma in Embedded System Design.

Authors’ Acknowledgments

Vikram Nagarajan: I earnestly thank Patrick Ames for bringing out the authors in us by his encouragement, humor, and also the subtle nudge to get us back to the project, just at the right time when we were drifting away. Thanks also to our copyeditor, Nancy Koerbel. I thank my manager Ramesh Kandula for bringing out the best in me in various projects. My gratitude to my co-authors Himanshu Agarwal and Princy Elizabeth in making the overall journey in EVPN multicast pleasant and fulfilling. Thanks to the technical reviewers who were diligent in going over the details in the book and providing questions, thus giving us confidence that the book is indeed purposeful. I am grateful to my colleague and friend Chandra for his profound wisdom and guidance. I would like to thank my wife Vibha, my mother, and my children, Vasudev and Vittal, for motivating me by eagerly asking as to when they can lay hands on the book. Finally, I dedicate this effort to my late father for being my inspiration and a very close friend.

Princy Elizabeth: At the outset, I would like to thank my co-authors, Himanshu Agarwal for conceiving this book and Vikram Nagarajan for his infectious enthusiasm and persistence in driving it to completion. I am also grateful to my manager Ramesh Kandula for supporting and encouraging our idea to write this book. A big thank you to all the technical reviewers for taking the time to review and provide their valuable feedback. Many thanks to our Editor in Chief, Patrick Ames, for his guidance, patience and his well-timed nudges that pushed us to give our best, and to our copyeditor, Nancy Koerbel, for her assistance with the development of the book. Finally, I would like to thank my family: my parents - for everything I am today, my husband and mother-in-law - for supporting and encouraging me in everything I do, including writing this book, and my sons, Evan and Johan, for teaching me something new every day.

Himanshu Agarwal: I would also like to thank Editor in Chief Patrick Ames, and our copyeditor, Nancy Koerbel, for their assistance and guidance with the book. I would like to sincerely thank my co-authors Vikram and Princy. I would also like to thank my manager, Stanzin Takpa, for encouraging me to work on this book. I would like to thank my wife Archika and son Vihaan for their patience during this journey. I also want to thank my mother and late father for always keeping confidence in me.

Welcome to Day One

This book is part of the *Day One* library, produced and published by Juniper Networks Books. *Day One* books cover the Junos OS and Juniper Networks network-administration with straightforward explanations, step-by-step instructions, and practical examples that are easy to follow.

- Download a free PDF edition at <http://www.juniper.net/dayone>
- PDF books are available on the Juniper app: **Junos Genius**
- Purchase the paper edition at Vervante Corporation (www.vervante.com).

Key Multicast Resources

The Juniper TechLibrary supports multicast technologies with its excellent documentation. It is thorough and kept up-to-date with the latest technologies and changes. This book is no substitute for that body of information. The authors assume that you have familiarity with the *Multicast Protocols Feature Guide*: https://www.juniper.net/documentation/en_US/junos/information-products/pathway-pages/config-guide-multicast/config-guide-multicast.html.

An excellent source of information for how to deploy EVPN-VXLAN in a data center environment is *Day One: Using Ethernet VPNS For Data Center Interconnect*, available at www.juniper.net/dayone.

Another excellent source of information on EVPN technology is *This Week: Data Center Deployment With EVPN/VXLAN*, available at www.juniper.net/dayone.

What You Need to Know Before Reading This Book

- You need a basic understanding of Junos and the Junos CLI, including configuration changes using edit mode. See the *Day One* books at www.juniper.net/books for a variety of books at all skill levels.
- You need a basic understanding of multicast and multicast protocols such as PIM, IGMP, etc.
- You should have some experience with EVPN and VXLAN technologies.
- This book assumes that you have a basic understanding of networking and will be able to configure and troubleshoot IP Addressing, unicast, and multicast routing and switching.
- The book has configuration sections and details about the verification of multicast in EVPN-VXLAN environment. It is highly recommended to find a lab with Junos routers to practice on.

After Reading This Book You Will Learn

After reading this book you will:

- Understand how multicast works in an EVPN-VXLAN Fabric.
- Understand how multicast traffic is routed from one VLAN to the other.
- Understand Assisted Replication (AR) and how it helps improve resource utilization and link-utilization
- How to appreciate the optimization on access-side interfaces with IGMP-Snooping
- How to appreciate the optimization in EVPN core using Selective (SMET) Forwarding
- How to appreciate the optimizations used in conjunction with a large-scale EVPN multicast deployment with Inter-VLAN and External Multicast
- Configure and troubleshoot optimized Intra-VLAN & Inter-VLAN Multicast.
- Configure and troubleshoot the devices to work with External Multicast Source/Receiver sitting outside the EVPN Fabric.
- Configure and Troubleshoot a EVPN-VXLAN Fabric with multicast optimization enabled with Assisted-Replication.

How This Book Is Set Up

This book gives a quick background on multicast, PIM, and EVPN technologies in Part I. It is intended as more of a refresher and a very high level primer.

Part II, delves into EVPN BUM and multicast forwarding rules within a VLAN and the problems of flooding everywhere. Various optimization procedures are discussed in a staged manner with Assisted Replication and Selective (SMET) Forwarding. Procedures are described in solving problems in multihoming, the cornerstone of EVPN. The benefits are illustrated quantitatively with examples to describe what the various optimizations bring to the table.

In Part III, the reader is taken through Inter-VLAN Multicast routing procedures riding on PIM as the protocol. It also explains how the optimizations introduced in Part II play a significant role in reducing bandwidth and resource consumption. Lastly, various nuances and procedures for multicast are described in relation to connecting the DC Fabric to the external world.

Chapter 1

Multicast Primer

This chapter describes the basics of multicast and explores multicast addressing and the procedures for multicast forwarding using the Physical Interface Module (PIM) and Internet Group Management Protocol (IGMP) protocols. Though the book delves mostly into Layer 2 multicast and the relevant optimizations in EVPN environments, for the sake of completeness this chapter provides background on the multicast technology.

Types of Traffic

There are three types of traffic delivery: unicast, broadcast, and multicast. These different types of traffic forwarding are used for various purposes in computer networks.

Unicast

Unicast traffic is the most predominant. It includes both Layer 2 and Layer 3 unicast. So traffic destined for an IP address, say, 150.1.1.1, will be delivered using hop-by-hop mechanisms. The paths are determined by unicast routing protocols in a Layer 3 network. In a Layer 2 network, the unicast frames are sent to a specific known MAC address.

Broadcast

Broadcast traffic is mainly used in Layer 2 networks. These are packets that are sent to *all* hosts in the network. Therefore, when a packet is sent to the link-local broadcast address, it is sent to all the hosts in the network. A classic example of

broadcast traffic is an ARP packet that is used to determine the IP-to-MAC address mapping of the destination host. Broadcasting packets at Layer 3 is generally avoided because it can potentially flood the Internet.

Multicast

Multicast traffic is forwarded to a *group* of hosts in a Layer 2 or Layer 3 network and is heavily used in both Layer 2 and Layer 3 networks. In Layer 2 networks, the traffic is sent to those Hosts that are interested in receiving traffic to a particular multicast group. In Layer 3 networks, multicast traffic is steered over different Layer 3 routers to reach the interested listener hosts.

Internet Group Management Protocol (IGMP)

IGMP is a protocol used within a LAN or Layer 3-subnet to express listener interest. Hosts send out IGMP reports to dedicated listeners interested in a particular group. These reports are received by a Layer 3 router. These Layer 3 routers maintain state for this group and forward multicast traffic destined to the group. The Layer 3 routers send out periodic IGMP queries to solicit reports from the hosts. The essence of this book is a paradigm of IGMP snooping enabled at Layer 2 switches for optimized multicast forwarding, and this is covered in detail in later chapters.

Multicast Addressing

Let's briefly visit the addresses earmarked for multicast applications.

Layer 3 Multicast Addressing

Layer 3 multicast addresses are classified as Class-D addresses and range from 224.0.0.0 to 239.255.255.255. Few multicast addresses among these ranges are reserved for specific purposes, although there are special addresses, 224.0.0/24 for example, that are meant for a specific purpose. This is called the *link-local multicast addresses range*, and the traffic for these addresses is to be constrained within the subnet. For example, 224.0.0.5 and 224.0.0.13 are used as the destination multicast address for OSPF and PIM to exchange hellos.

The multicast host applications look to *join* a particular multicast group address, say 235.1.1.1, and expect multicast traffic to be delivered on this address, which they consume. For example, an IPTV host may subscribe for a channel by expressing listener interest for a particular specified group. Also, different channels can be deployed by using different multicast groups.

Layer 2 Multicast Addressing

In the unicast world, a particular L3-IP address maps to a MAC address in Layer 2. In multicast, the particular L3-multicast address is mapped to a L2-multicast address. The L3-multicast address is used in multicast routing, while the L2-multicast address is used in L2-multicast forwarding of the L2-frame.

A host that is interested in a particular group, say 228.11.1.2, expresses its listener interest by sending out an IGMP packet for that group. Also, its hardware programs the L2 MAC address in its hardware corresponding to 228.11.1.2 with, say, MAC address E40B 0102, such that when the multicast traffic arrives with this destination MAC, the host is able to accept the packet. There are certain nuances in the L3-IP to L2-MAC multicast address mapping related to collision, but let's skip that for now.

Overall, the multicast addresses in Layer 3 and Layer 2 are used by hosts and routers alike to get the multicast traffic signaling and forwarding working.

Protocol Independent Multicast (PIM - ASM)

Let's review how multicast in Layer 3 networks works, again touching upon only the basic procedures. This background information will be relevant in the chapters in Part 2 that look at the big picture of how multicast is used in data center fabrics.

Any-source multicast (ASM) is a scenario where the listeners or hosts do not know the multicast source information. All they know is the group information for the desired traffic. There is a central device called a *rendezvous point* (RP) that acts as a liaison between the multicast listeners and the sources for the group.

Typical Multicast Topology

Consider the topology in Figure 1.1, where there is a multicast source behind R1. R1 is called the first-hop router because it is the first hop to the source. There is also R6, which is called the last-hop router, because it is the last hop from the source and the router nearest to the listener. Typically, the RP is placed centrally in the network. Unicast routing is enabled in the network, and PIM rides on existing unicast routing topology to steer multicast. The unicast reachability can be provided by any of the routing-protocols such as OSPF, ISIS, BGP, or static routing.

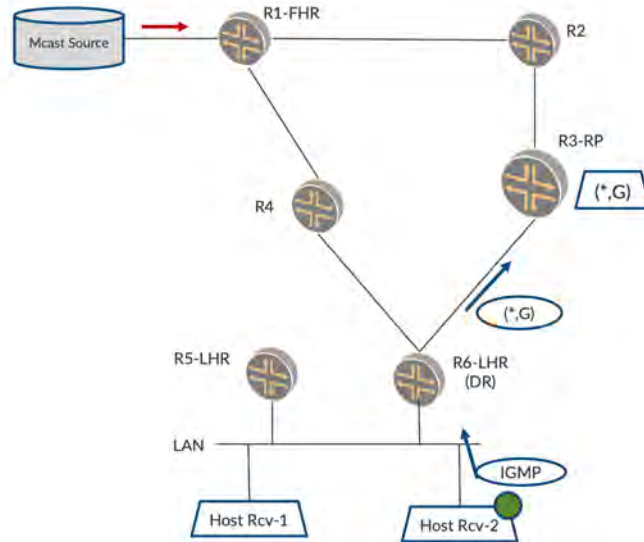


Figure 1.1

Typical Multicast Topology

Listener Sends IGMP Report

Figure 1.1 shows a listener Host-Rcv-2 on the LAN behind R6-LHR. When this host is interested in a particular multicast group, say 235.1.1.1, it sends an IGMP report for this group G.

PIM Hellos and PIM Neighbor Discovery

Before we describe what R6 does with the report, let's quickly touch upon PIM procedures for neighbor discovery and designated router (DR) election. PIM devices send out PIM hellos to discover each other. In Figure 1.1, R6 as well as R5 are on the LAN interface. Both R5 and R6 discover each other on the LAN.

If both devices pulled traffic and forwarded it on the LAN, it would result in duplicates. So R5 and R6 enter into a PIM-DR election and the PIM-DR alone creates and propagates state for that subnet. The PIM-NDR may create state but does not propagate it to pull traffic. This DR election is based on a priority field carried in the PIM hello messages. If this is not explicitly carried, the device with the lowest IP address on the interface is elected the PIM-DR.

LHR Creates PIM (*,G) State and Sends PIM Join Towards RP

Let's resume the path of the Join. In Figure 1.1, both R5 and R6 receive the IGMP Report and create a PIM (*,G) state. However, R5, being PIM-NDR, does not propagate the state. R6, being the PIM-DR for the LAN, propagates the state by sending a PIM (*,G) Join towards the RP. Since the host and R6 do not know the location of the source(s), R6 looks to send the PIM Join to RP.

Source Sends Multicast Traffic – FHR Sends PIM Register to RP

Let's go to the next step. Say there is a multicast source that starts sending the multicast traffic. R1, the FHR, sees the multicast traffic from the source. R1 does not have the information of the interested listeners, so it conveys the information about the source to the RP. R1 sends this multicast packet into a unicast tunnel and sends it to the RP. This is called a *PIM Register Message* and you can see it in Figure 1.2.

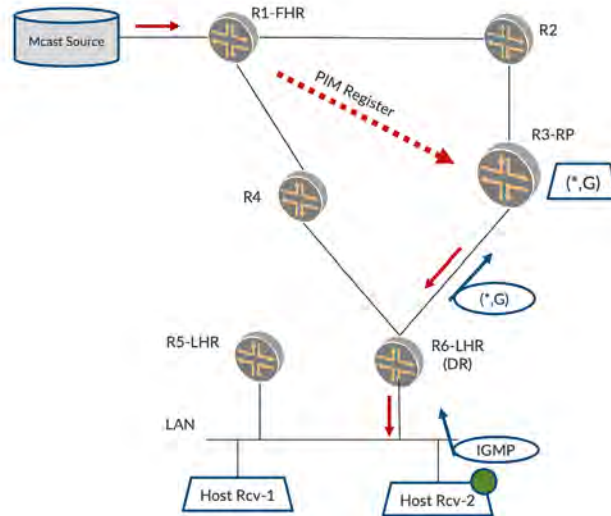


Figure 1.2

Traffic on Rendezvous-Point Tree

Traffic Flows Over Rendezvous-Point Tree (RPT)

In Figure 1.2, the RP receives the multicast packet over a unicast tunnel as PIM register packets. The RP decapsulates the multicast traffic and checks if there are any listeners for the group. Since it has a $(*,G)$ Join state created by virtue of the downstream Join from R6, the RP forwards the multicast traffic to R6 over the shared tree. R6 now forwards the traffic over the LAN and the Host-Rcv-2 receives it.

So far so good. But it's not a good thing that RP receives the multicast traffic over a unicast tunnel. Also, now that R6 has come to know of the multicast source, it can possibly join the multicast source itself and pull the traffic over the shortest path. The two procedures below allow us to move towards this.

RP Joins Source Tree

Now, in Figure 1.3, R3-RP, having known the location of the source, looks to join the source natively. To achieve this it sends a PIM (S,G) Join towards the source to R2. R2 propagates the (S,G) Join towards R1. Now R1 forwards traffic towards R2, and in turn, R2 towards R3-RP. Then R3-RP informs R1-FHR by sending a *PIM Register Stop* to stop forwarding traffic on the unicast tunnel. R3-RP forwards the traffic it receives natively to R6-LHR.

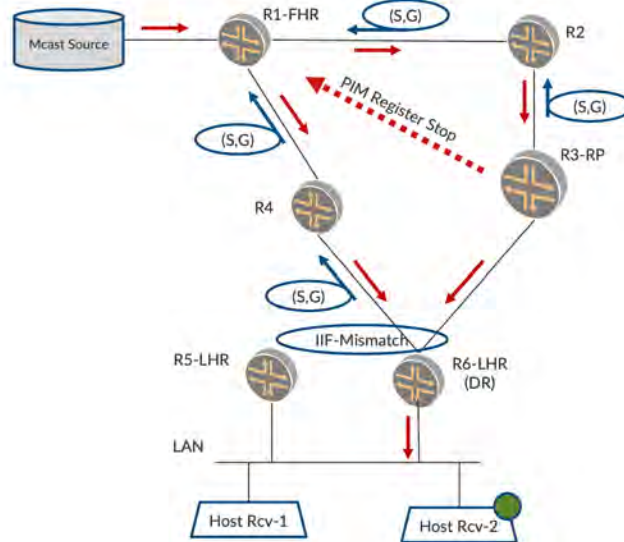


Figure 1.3

Traffic on Source Tree

LHR Joins Shortest-Path Tree

At around the same time, R6 realizes the location of the Source and it sends a (S,G) Join towards the Source to R4 to pull traffic over the shortest-path tree (SPT). R4 sends the (S,G) Join to R1. Now, R1 forwards the traffic to R4, and R4 in turn, forwards the traffic to R6. Thus R6 has built a SPT where traffic is forwarded with the least latency.

LHR Switches Traffic From RPT to SPT

Now R6, per Figure 1.3, receives traffic over both SPT as well as RPT. This is detected by an *Incoming Interface Mismatch* event that occurs on R6. It should not forward traffic from both trees onto the LAN. It picks the traffic on SPT and forwards it on the LAN.

Also, since the SPT has been formed, it wants to *prune* the shared tree. Towards this, R6 sends a (*,G) + (S,G, RPT_Prune) message towards RP as shown in Figure 1.4. This indicates to the RP that the listener is interested in traffic for this group G. However, since it has joined SPT for this particular source S, RP skips forwarding traffic from this source S. If any other source comes alive, say S2, RP forwards the traffic on the shared tree.



Figure 1.4

IGMP Leave and PIM Prune

PIM-ASM Summary

Phew! Too much detail in too short a time. If this seemed overwhelming, take heart that we will quickly describe PIM-SSM and that should keep things simple. This section on PIM-ASM was included for the sake of providing some relevant background that may be helpful in understanding later chapters. If you don't fully understand the procedures, you should still be in good shape because this book mainly deals with multicast in Layer 2.

PIM-SSM

We just explored PIM-ASM and its procedures, so now let's explore PIM-SSM, which is simpler in terms of procedures. In Figure 1.5 the listener sends an IGMPv3 report. This report has the information on both Group and Source, so R6 can simply send the (S,G) Join towards the source to R4. R4 sends (S,G) towards R1. R1 forwards the traffic towards R4, and R4 in turn forwards traffic towards R6. There is only the source tree. R6 forwards the traffic to listeners on the LAN.

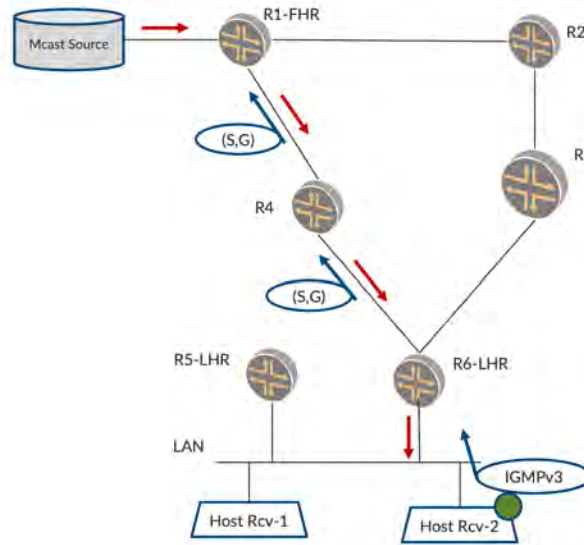


Figure 1.5

PIM-SSM

Later, when traffic starts, R1 forwards the traffic towards R4, and R4 in turn forwards traffic towards R6. There is only the source tree. R6 forwards the traffic to listeners on the LAN.

Though PIM-SSM simplifies the procedures in the network, it places the responsibility of the source awareness on the listeners. Since it is very difficult to determine the source information a priori, PIM-ASM is widely deployed because the complexity is handled by the network but the hosts can just express interest in the particular *channel* or group alone.

This book deals mainly with IGMPv2 hosts. Therefore those hosts send (*,G) Reports rather than IGMPv3 hosts which send (S,G) Reports.

Other Multicast Technologies

What we just described is IGMP and PIM. There are other technologies that transport multicast, over a L3-VPN, for example, and BGP-MVPNs or Rosen MVPNs are used to transport multicast in conjunction with PIM. Also, in the MPLS world, multicast is transported by MLDP or RSVP-TE P2MP tunnels that perform the replication over the MPLS core. The procedures for these are beyond the scope of the book.

Summary

Multicast is increasingly being deployed for large scale applications. The dominant force driving multicast seems to be IPTV applications where the challenges of sending-to-many and increasing bandwidth are being addressed by multicast technologies such as:

- Multimedia content delivery systems
- Stock-ticker applications in stock exchanges
- IPTV for live television distribution and televised company meetings
- Bulk File Distribution to deliver operating system images/patches.

This chapter started to cover some ground on multicast procedures. It was by no means comprehensive – multicast as a technology has evolved over two decades and has many nuances. So this chapter was intended to provide the basics and get the reader conversant with terms that appear often later in the book. In Chapter 2 we will conduct a similar exercise on EVPN, and provide a quick primer to understand the terms and key topics at a very high level.

Chapter 2

EVPN Primer

This chapter discusses the basics of EVPN in a very cursory manner, illustrating the key concepts in order to serve as a foundation for future chapters. Relevant links with detailed explanations of topics you may wish to pursue further are included.

Traditionally, L2-switched environments have been confined to a single site and technologies like VPLS have helped stretch the VLANs in such environments. However, L2-switched environments have major challenges with respect to data plane learning in the MPLS core and are unable to achieve active/active multi-homing. VPLS supports only active/standby multi-homing. Also, VPLS is a MPLS-only technology.

Ethernet VPN (EVPN) is a standards-based technology that provides virtual multi-point bridged connectivity between different Layer 2 domains over an IP or IP/MPLS backbone network. Like other VPN technologies, such as IP VPN and virtual private LAN service (VPLS), EVPN instances are configured on provider edge (PE) routers to maintain logical service separation between customers. The PE routers connect to customer edge (CE) devices, which can be routers, switches, or hosts.

The PE routers then exchange reachability information using Multiprotocol Border Gateway Protocol (MP-BGP) and encapsulated traffic is forwarded between PE routers. Because elements of the architecture are common with other VPN technologies, you can seamlessly introduce and integrate EVPN into existing service environments.

EVPN became the widely adopted paradigm to stretch the VLANs by virtue of several characteristics including:

- Control plane MAC learning using BGP signaling mechanisms
- Active/active multi-homing with ESIs
- MAC-Mass withdrawal
- Aliasing of unicast traffic from Ingress
- BGP-based policy to fine-tune to different customer requirements and scale
- Multitenancy
- MAC address mobility

EVPNoMPLS was intended to introduce control plane MAC learning using BGP signaling to exchange the MAC routes. Also, active/active multi-homing was addressed with EVPN with the help of Ethernet Segment Identifiers (ESI). While EVPNoMPLS was targeted for service provider networks, VxLAN was rapidly evolving as a data center technology. To be able to stretch the VLANs in a VxLAN data center cloud, EVPN was a good choice with several other advantages that EVPN brought to the table. EVPNoVXLAN soon became the choice for stretching Layer 2 in enterprise data center fabrics.

Let's begin by briefly introducing the basic building blocks of a data center fabric.

Building Blocks of EVPNoVXLAN

VLAN/Bridge Domain

A virtual LAN (VLAN) is a broadcast domain that is partitioned and isolated in a network at the data link layer (Layer 2). VLANs work by applying tags to network frames and handling these tags – creating the appearance and functionality of network traffic that is physically on a single network but acts as if it is split between separate networks. VLANs keep network applications separate despite being connected to the same network, and without requiring multiple sets of cabling and devices.

VLANs allow you to group hosts together even if they are not connected to the same switch. Because VLAN membership can be configured through software, this simplifies network design and deployment. Many deployments use VLANs to separate their customers' private zones from each other. That allows each customer's servers to be grouped together in a single network segment while being located anywhere in their data center.

VXLAN

Virtual Extensible LAN (VXLAN) is a virtualization technology that addresses the scalability problems associated with large cloud computing deployments. It uses a VLAN-like encapsulation technique to encapsulate Layer 2 Ethernet frames within Layer 4 UDP datagrams, using 4789 as the default IANA-assigned destination UDP port number. VXLAN endpoints that terminate VXLAN tunnels can be virtual or physical switch ports. They are known as VXLAN tunnel endpoints (VTEPs).

VXLAN standardizes as an overlay encapsulation protocol. It increases scalability up to 16 million logical networks and allows for Layer 2 adjacency across IP networks. Multicast or unicast with head-end replication is used to flood broadcast, unknown unicast, and multicast (BUM) traffic.

VXLAN has evolved very well within the industry and has been used as an encapsulation mechanism with EVPN, leading to EVPNoVXLAN procedures.

EVPN VRF (MAC-VRF or EVI)

EVPN is built on a classic VPN model using MP-BGP extensions. The concept of a VRF instance (virtual routing and forwarding) is inherited from the L3VPN/L2VPN world into EVPN; for example, different EVPN instances (EVIs) can be created for different customers and separate routing and forwarding tables will be maintained for each. This is achieved by employing classic route-distinguisher/route-target mechanisms.

EVI Route-Distinguisher and Route-Target

A route distinguisher is an address qualifier unique on an EVPN PE device. It is used to distinguish distinct virtual private network (VPN) routes of separate customers who connect to the provider. Traditionally, the route distinguisher is an 8-octet field prefixed to the customer's IPv4 address. The resulting 12-octet field is a unique VPN-IPv4 address. The usage of the RD field in EVPN ensures that the MAC/ESI/(S,G) prefixes are unique across the different VRFs in a EVPN device.

The route target is an 8-byte field which is a BGP-extended Communities Attribute. This is utilized to configure the prefixes that are to be suitably imported and exported on the VRFs of the EVPN devices. It helps in keeping the VRF routes constrained within the VRF, such that different customers' prefixes and traffic do not intersperse with each other.

EVPNoVXLAN

EVPNoVXLAN is the best of both worlds in traditional L3/L2VPNs and the VXLAN paradigm. By combining the two, you can extend VLAN network using VXLAN and also achieve a L2VPN-like segregation using MP-BGP mechanisms for the EVPN family.

Even though VXLAN is an independent technology used to mitigate the scale limitations of VLANs, VXLAN also provides a Layer 3 tunneling paradigm to carry the Layer 2 frames within. This is leveraged with EVPNoVXLAN. That is to say, EVPNoVXLAN is used to stretch the VLANs using VXLAN as the underlay.

MPLS Label Vis-à-vis VXLAN Identifier (VNI)

In classic Layer 3 VPNs using MP-BGP and MPLS, MPLS is the transport. With MP-BGP extensions you carve out services with VRFs. For each VRF, there is a label assigned, called *service label*, which is used to forward and receive packets within a VRF across the sites.

In EVPNoVXLAN, there is no MPLS. For the transport there are VXLAN and VTEP interface tunnels. To identify a service, (in this case, perhaps a customer-VLAN), you need an identifier to encapsulate and decapsulate the traffic. Towards this, the VNI (VXLAN identifier) segregates the VLANs over the EVPN core.

Sample EVPNoVXLAN Topology

Consider the topology shown in Figure 2.1 where an EVPNoVXLAN fabric is used to stretch the VLANs spread over several sites. In typical data centers, the VLANs are present in the same geographical site but present in other floors of the same building or in various grouped buildings. The placement of the VLANs over the sites provides a mechanism for managing the network while offering good resilience at scale.

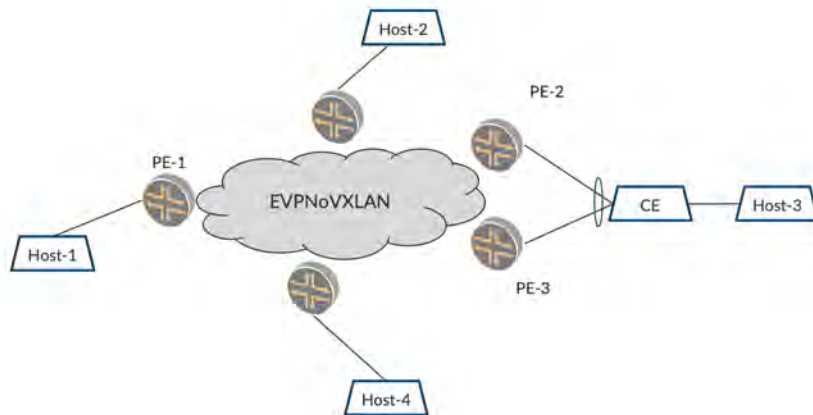


Figure 2.1

EVPNoVXLAN Topology

For example, there are several VLANs of different customers behind an EVPN PE, say PE1 to begin with. Over the course of time, there is another site (building/floor, etc.,) behind PE2, where the *same* VLANs are relevant and there is a need for the VLANs to be stretched. Also, to provide resiliency, PE3 is connected in a multi-homed manner such that PE2 and PE3 can coordinate and provide redundancy and resiliency.

The VLANs behind PEs are learned and their information is suitably exchanged with each other. Also, VXLAN tunnels are built using VTEP interfaces such that traffic received on the VLAN by PE1 can be forwarded on the VTEP tunnel to PE2, and PE2 forwards onto the respective VLAN. This tunneling mechanism with VTEP interfaces is provided by the VXLAN underlay, while EVPN service for the VLANs forms the overlay.

MORE? Here's a good overview of EVPN/VXLAN. https://www.juniper.net/documentation/en_US/junos/topics/concept/evpn-vxlan-data-plane-encapsulation.html.

Different VLAN Services with EVPN

EVPN supports three ways to stretch the VLANs over the EVPN core.

- *EVPN VLAN-based service*: This supports the mapping of one routing instance of type EVPN to one VLAN. There is only one bridge table that corresponds to the one VLAN. This type is typically used in EVPN/MPLS service provider networks.
- *EVPN VLAN-bundle service*: Here you have one EVPN instance EVI mapped to a single bridge-domain. This bridge-domain can carry several VLANs 'bundled' within it. This is achieved by configuring a VLAN-ID-list or VLAN-range that is to be carried in the bridge-domain. In this scheme, the VLANs are passed through over the EVPN core. This is sometimes called *port-based pseudowire* (PW) and is used mostly in EVPN/MPLS service provider networks for PW services.
- *EVPN VLAN-Aware bundle service*: With VLAN-Aware bundle service, there is one EVPN instance EVI configured with several bridge-domains. Each of the bridge-domains is mapped to a single VLAN. This way you have each of the VLANs carried with separate VNIs, providing a nice segregation across the VLANs. This scheme is typically used in data center deployments and is used in this book for illustration purposes.

Layer 2 Traffic Types

Layer 2 traffic is broadly divided into unicast, broadcast, unknown unicast, and multicast:

- Broadcast traffic is destined to all hosts in the VLAN. This traffic has a specific destination address (0xFFFF) and is flooded onto all the ports by the L2 switch.
- Unicast traffic occurs when the destination MAC is unicast MAC and the switch knows the outgoing port for the destination MAC.
- Unknown unicast traffic is when the destination MAC is unicast MAC and the switch does not know the outgoing port for the destination MAC. When the switch receives such traffic from a port, it sends out an ARP packet seeking the destination to reply back. This ARP packet has a specific destination MAC address. This packet is flooded onto all ports (excluding the one on which the packet arrived) by making copies for each of the ports.
- Multicast traffic has destination MAC addresses in the specified range of multicast addresses. The equivalent IP range for multicast address is Class-D 224.0.0.0 to 239.0.0.0. When the switch receives such a packet, it floods the packet onto all the ports by making one copy for each port. This book deals mainly with the handling of this L2 multicast traffic, the problems with flooding, and the optimizations to mitigate them.

Data-Plane vis-à-vis Control-Plane MAC Learning

Traditional L2-stretched networks with VPLS only support data plane learning as shown in Figure 2.2. The VPLS PE devices learn of the MAC addresses on the L2-access interfaces. They also learn of the remote MAC addresses that are behind other VPLS PE devices over the pseudowire interfaces. Both of these happen in data plane.

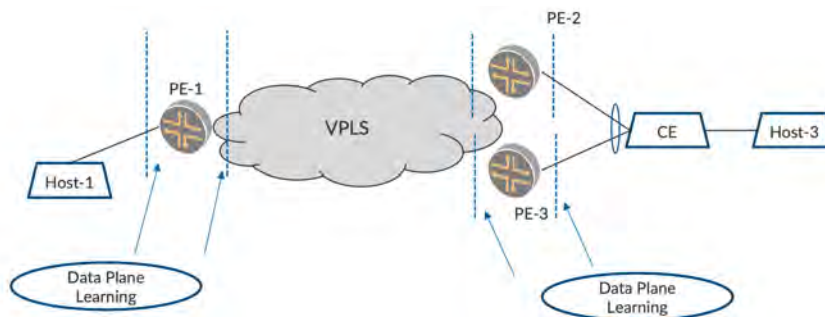


Figure 2.2

Data Plane Learning

PE-1 learns of the incoming MAC addresses from Host-1 over the L2-access interface while it learns of the incoming MAC addresses from Host-3 over the VPLS pseudowire interface. The data-plane component keeps itself busy with the learning. Typically, several remote PEs and MAC addresses have to be learned from the pseudowires coming from each of them.

EVPN brings control plane learning to the table, and the MACs that are learned on the access interface are advertised to the other PEs using MP-BGP (EVPN Type 2) routes. In Figure 2.3, PE-3 advertises EVPN Type 2 routes for the MACs that it learned over the interface towards Host-3. PE-1, upon receiving these Type 2 routes, installs the MAC routes in its forwarding information base, forwarding table (FIB), thus obviating the need for MAC learning on the EVPN core interface. Likewise, PE-1 advertises EVPN Type 2 routes for the MACs that it learned over the interface to Host-3 which PE-3 programs in its FIB.

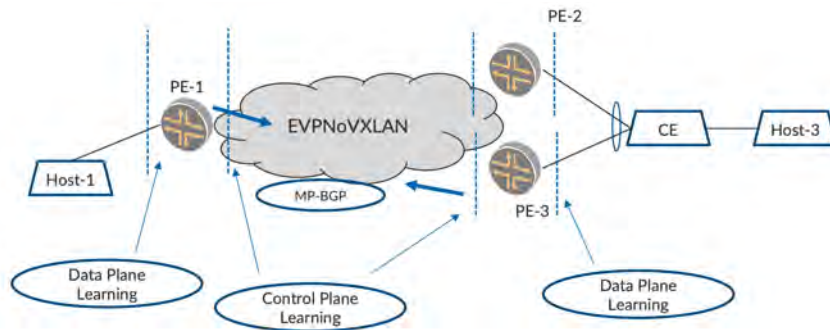


Figure 2.3

Control Plane Learning

EVPN Multihoming with Ethernet Segment Identifier

One of the main features that EVPN brings to networking is active/active multihoming. Historically, with VPLS, there was only single/active multihoming to ensure the classic L2-SPT tree rules are met to avoid loops. With EVPN and its control plane learning feature, this problem is circumvented, and with active/active multihoming there are these gains:

- **Unicast Traffic Load Balancing (MAC Aliasing):** This ensures that traffic from ingress can be load balanced nicely over two different paths to the multihomed nodes. The traffic load is shared between both the nodes. Also, both the access links towards the CE are utilized.
- **BUM Traffic Load Sharing:** Since both the multihomed nodes and the access links can be utilized, BUM traffic can be load shared, too. However, for a single VLAN, care must be taken that only one node forwards the traffic. This is

achieved by designated forwarder (DF) election. When you have several VLANs, the BUM traffic load is nicely shared by the multihomed nodes. This is achieved by one of the multihomed nodes being the DF for some VLANs and the other multihomed node being the DF for the other VLANs.

- **Resiliency:** When one of the multihomed nodes goes down or one of the multihomed links goes down, the unicast/BUM traffic can converge to the other alive node with minimal traffic loss.
- All the multihomed-related information is exchanged via BGP routes in the control plane so that the complexity of the redundancy and convergence are moved out of the data plane.

In Figure 2.4, if PE-2 and PE-3 are to be multihomed, there has to be a construct to suggest the same to the other PEs. This is achieved by binding the two PEs and the multihomed link in a redundancy link pair. This is referred in EVPN as the Ethernet Segment Identifier (ESI).

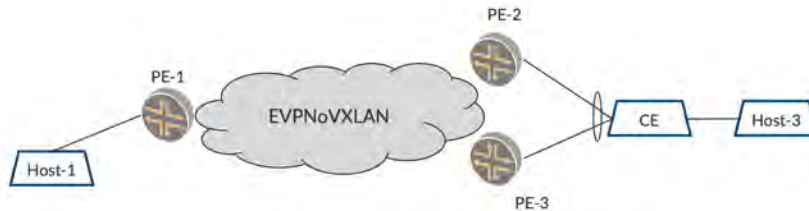


Figure 2.4 EVPN Multihoming

EVPN Type 1 Route Per ESI

The two PEs, PE-2 and PE-3, are configured with the same ESI value on the multihomed link. When there are several multihomed links, each of the multihomed links are configured with a separate ESI. Thus, we achieve a correlation between the multihomed nodes and the link that they are multihomed.

This ESI information is exchanged separately using EVPN Type 1 routes. The other PEs realize upon receiving the Type 1 route that the two PEs are multihomed on that particular ESI. Also, when the multihomed nodes advertise a MAC route using a Type 2 route, they add the ESI value in the Type 2 route such that the other PEs know that the particular MAC is behind a set of multihomed nodes.

MAC Aliasing

Let's look at the ingress PE, from PE-1's point of view. Say there are several MACs behind the CE. These MACs are learned by PE-2 and PE-3 and they are advertised in control plane using EVPN Type 2 MAC Route. Since PE-2 and PE-3 have learned these MACs over the ESI link, these PEs add the ESI value in the Type 2 route.

The ingress PE, when it receives the Type 2 MAC routes realize that these MACs are reachable over both PE-2 and PE-3. This is because PE-2 and PE-3 would have advertised a Type 1 route for the ESI. Once PE1 forms a mapping for the MAC route, it can install the forwarding such that the incoming packets from Host-1 can be load balanced between PE-2 and PE-3. This load balancing usually happens based on the incoming packet's tuple (source-IP, destination-IP, etc.). Thus, when there are several flows going from Host-1 towards Host-3, these are nicely load balanced between the two paths to the two nodes. This way, both paths, PE-1 to PE-2, and PE-1 to PE-3, are utilized as the traffic processing and forwarding load is shared between PE-2 and PE-3.

MAC Mass Withdrawal with ESI

In addition to helping with MAC aliasing, the paradigm of ESI has one another important benefit. When ESI is configured on a link, be it single-homed or multi-homed, it helps in faster convergence. Consider Figure 2.4. In this case, say the single-homed ingress PE, the PE-1, is configured with an ESI towards Host-1. PE-1 advertises a Type 1 route for that ESI. Also, PE-1 learns and advertises MACs on ESI with Type 2 with the ESI information.

Later, when the link goes down, PE-1 would first withdraw the Type 1 route. Other PEs realize that the ESI behind PE-1 has gone down, and hence would cleanup all the MAC entries learned from the peer. This results in better convergence because once the MAC entries are deleted, ARP can begin and the new MACs can be discovered behind a new PE.

This is a typical use case where a site of MACs is moved from one floor of the building to another. If all these MACs have to be withdrawn individually, it results in service disruption. By virtue of the Type 1 AD route withdrawal, the MACs get cleaned up, BUM flooding begins, and the new PE that the MACs have moved behind is learned quickly.

EVPN Type-4 Route for DF/NDF Calculation

When it comes to BUM forwarding (typically, these are ARP request packets and multicast packets), care should be taken that only one of the multihomed nodes sends the packets to the CE. If both of the nodes forward the BUM packets to the CE, duplicates would occur and this can cause problems for the hosts.

To have a single forwarder for BUM packets alone, there is a DF election procedure. This is achieved by an EVPN Type 4 route, which carries the ESI information. This Type 4 route for an ESI is used only for electing the DF while the Type 1 route per ESI is used to carry additional information, like Split Horizon Label, etc., for MPLS.

Once the DF election is performed, one of the multihomed nodes becomes the DF and the other multihomed nodes become NDF. The NDF nodes do not forward the BUM traffic that arrives from the core. We will visit further rules related to this in subsequent chapters.

NOTE One thing to remember in active/active mode, is that unicast traffic is forwarded by both nodes while BUM traffic is forwarded only by the DF node.

EVPN Type-3 Route for BUM Forwarding

The EVPN peers exchange EVPN Type 3 routes to exchange the VLAN information. This EVPN route is used to build inclusive multicast tunnels between the Ingress and the other EVPN PEs that host the same VLAN. The BUM traffic is forwarded by Ingress using Ingress replication. For example, the Ingress replicates the incoming packet and sends one copy each to the remote PE that hosts the VLAN. Based on the EVPN Type 3 route that is exchanged, the Ingress replication tunnels are built.

Chapter Summary

This chapter provided information on the basic building blocks of EVPN, the Layer 2 types of traffic, and a cursory walk through the different MP-BGP NLRI routes that help in achieving different goals. The benefits of control plane learning were discussed, as well as how multihoming features are achieved using NLRI routes. We touched upon BUM forwarding, of which multicast in particular will be detailed throughout the rest of the book.

This chapter was intended to provide a background of EVPN and is by no means comprehensive. For details on the characteristics and behavior of EVPN, and the differentiators that it brings to the table, it will be best to visit the links provided in the Appendix.

Chapter 3 provides the basic configuration to bring up EVPN in a DC fabric. This includes configuring the underlay and the EVPN overlay, and the VLANs and the ESI information.

Chapter 3

EVPN Base Configuration in DC Fabric Topology

The complete topology as shown in Figure 3.1 consists of:

- The data center fabric:
 - Five EVPN LEAF PEs: LEAF-1, LEAF-2, LEAF-3, LEAF-4, and LEAF-5
 - Two EVPN Border LEAF PEs: BL-1 and BL-2
 - Two Lean Spine devices: SPINE-1 and SPINE-2
- External multicast world:
 - One PIM Gateway: PIM-GW
 - One PIM RP: PIM-RP
- Three CE devices for multihoming some of the end Hosts: CE-1, CE-2, and CE-3.
- Eight router tester (RT) ports to simulate multicast traffic source(s) and receivers.

We will configure these devices as required.

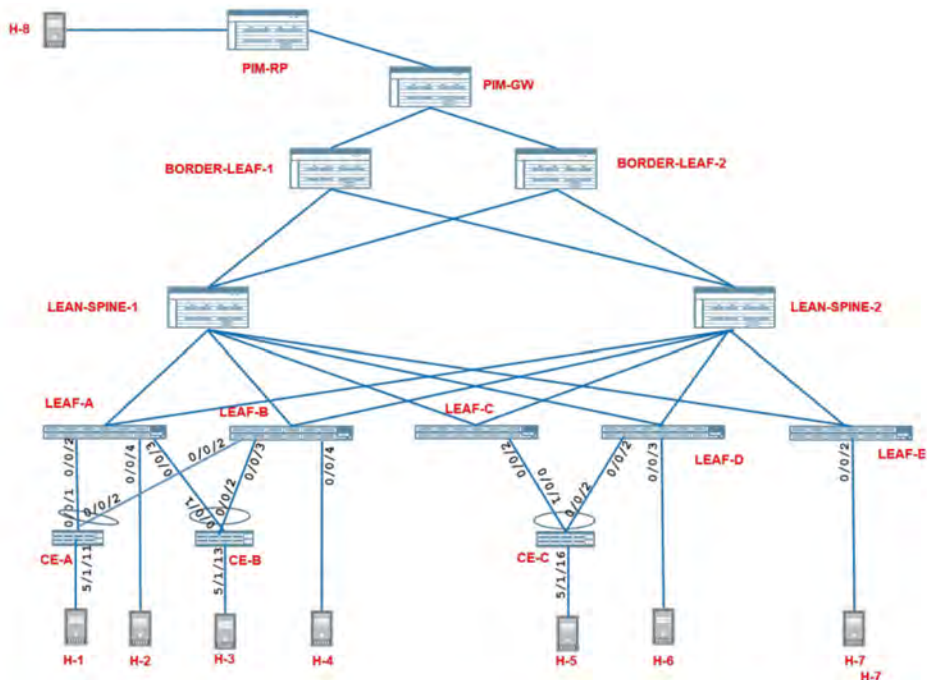


Figure 3.1 Chapter 3 Topology

Base EVPN Configuration

Configuring the Underlay

LEAF-1

Configure the underlay interfaces:

```
set interfaces xe-0/0/0 description "TO SPINE-1"
set interfaces xe-0/0/0 unit 0 family inet address 8.8.8.2/24
set interfaces xe-0/0/1 description "TO SPINE-2"
set interfaces xe-0/0/1 unit 0 family inet address 13.13.13.2/24
set interfaces lo0 unit 0 family inet address 105.105.105.105/32
commit
```

Configure BGP for routing in the underlay:

```
set routing-options router-id 105.105.105.105
set routing-options autonomous-system 65005
set policy-options policy-statement EXPORT-L0 term 1 from protocol direct
set policy-options policy-statement EXPORT-L0 term 1 then accept
set protocols bgp group UNDERLAY family inet any
```

```

set protocols bgp group UNDERLAY export EXPORT-L0
set protocols bgp group UNDERLAY local-as 65005
set protocols bgp group UNDERLAY neighbor 8.8.8.1 description SPINE-1
set protocols bgp group UNDERLAY neighbor 8.8.8.1 peer-as 65003
set protocols bgp group UNDERLAY neighbor 13.13.13.1 description SPINE-2
set protocols bgp group UNDERLAY neighbor 13.13.13.1 peer-as 65004
commit

```

The underlay configuration on other LEAF PEs and border LEAF PEs will be similar and is provided in the Appendix.

SPINE-1

Configure the underlay interfaces:

```

set interfaces xe-0/0/0 description "T0 BL-1"
set interfaces xe-0/0/0 unit 0 family inet address 5.5.5.2/24
set interfaces xe-0/0/1 description "T0 BL-2"
set interfaces xe-0/0/1 unit 0 family inet address 7.7.7.2/24
set interfaces xe-0/0/2 description "T0 LEAF-1"
set interfaces xe-0/0/2 unit 0 family inet address 13.13.13.1/24
set interfaces xe-0/0/3 description "T0 LEAF-2"
set interfaces xe-0/0/3 unit 0 family inet address 14.14.14.1/24
set interfaces xe-0/0/4 description "T0 LEAF-3"
set interfaces xe-0/0/4 unit 0 family inet address 15.15.15.1/24
set interfaces xe-0/0/5 description "T0 LEAF-4"
set interfaces xe-0/0/5 unit 0 family inet address 16.16.16.1/24
set interfaces xe-0/0/6 description "T0 LEAF-5"
set interfaces xe-0/0/6 unit 0 family inet address 17.17.17.1/24
set interfaces lo0 unit 0 family inet address 104.104.104.104/32
commit

```

Configure BGP for routing in the underlay:

```

set routing-options router-id 104.104.104.104
set routing-options autonomous-system 65004
set policy-options policy-statement EXPORT-L0 term 1 from protocol direct
set policy-options policy-statement EXPORT-L0 term 1 then accept
set protocols bgp group UNDERLAY family inet any
set protocols bgp group UNDERLAY export EXPORT-L0
set protocols bgp group UNDERLAY local-as 65004
set protocols bgp group UNDERLAY neighbor 5.5.5.1 description BL-1
set protocols bgp group UNDERLAY neighbor 5.5.5.1 peer-as 65001
set protocols bgp group UNDERLAY neighbor 7.7.7.1 description BL-2
set protocols bgp group UNDERLAY neighbor 7.7.7.1 peer-as 65002
set protocols bgp group UNDERLAY neighbor 13.13.13.2 description LEAF-1
set protocols bgp group UNDERLAY neighbor 13.13.13.2 peer-as 65005
set protocols bgp group UNDERLAY neighbor 14.14.14.2 description LEAF-2
set protocols bgp group UNDERLAY neighbor 14.14.14.2 peer-as 65006
set protocols bgp group UNDERLAY neighbor 15.15.15.2 description LEAF-3
set protocols bgp group UNDERLAY neighbor 15.15.15.2 peer-as 65007
set protocols bgp group UNDERLAY neighbor 16.16.16.2 description LEAF-4
set protocols bgp group UNDERLAY neighbor 16.16.16.2 peer-as 65008
set protocols bgp group UNDERLAY neighbor 17.17.17.2 description LEAF-5
set protocols bgp group UNDERLAY neighbor 17.17.17.2 peer-as 65009
commit

```

We will go over the configuration for SPINE-1. The configuration on SPINE-2 will be similar and is provided in the Appendix.

Configuring the Overlay

We will now configure the overlay on the EVPN PEs.

LEAF-1

Configure I-BGP for routing in the overlay:

```
set protocols bgp group OVERLAY type internal
set protocols bgp group OVERLAY local-address 105.105.105.105
set protocols bgp group OVERLAY family evpn signaling
set protocols bgp group OVERLAY local-as 65000
set protocols bgp group OVERLAY neighbor 101.101.101.101 description BL-1
set protocols bgp group OVERLAY neighbor 102.102.102.102 description BL-2
set protocols bgp group OVERLAY neighbor 106.106.106.106 description LEAF-2
set protocols bgp group OVERLAY neighbor 107.107.107.107 description LEAF-3
set protocols bgp group OVERLAY neighbor 108.108.108.108 description LEAF-4
set protocols bgp group OVERLAY neighbor 109.109.109.109 description LEAF-5
commit
```

Configure the customer interfaces and VLANs:

```
set chassis aggregated-devices ethernet device-count 2
set interfaces xe-0/0/2 gigether-options 802.3ad ae0
set interfaces xe-0/0/3 gigether-options 802.3ad ae1
set interfaces xe-0/0/4 description "T0 Host-2"
set interfaces xe-0/0/4 unit 0 family ethernet-switching interface-mode trunk
set interfaces xe-0/0/4 unit 0 family ethernet-switching VLAN members VLAN-101
set interfaces ae0 description "T0 CE-1"
set interfaces ae0 aggregated-ether-options lACP active
set interfaces ae0 aggregated-ether-options lACP periodic fast
set interfaces ae0 aggregated-ether-options lACP system-id 00:11:11:11:11:11
set interfaces ae0 unit 0 family ethernet-switching interface-mode trunk
set interfaces ae0 unit 0 family ethernet-switching VLAN members VLAN-101
set interfaces ae1 description "T0 CE-2"
set interfaces ae1 aggregated-ether-options lACP active
set interfaces ae1 aggregated-ether-options lACP periodic fast
set interfaces ae1 aggregated-ether-options lACP system-id 00:22:22:22:22:22
set interfaces ae1 unit 0 family ethernet-switching interface-mode trunk
set interfaces ae1 unit 0 family ethernet-switching VLAN members VLAN-101
set interfaces ae1 unit 0 family ethernet-switching VLAN members VLAN-102
set vlans VLAN-101 VLAN-id 101
set vlans VLAN-101 vxlan vni 101
set vlans VLAN-102 VLAN-id 102
set vlans VLAN-102 vxlan vni 102
commit
```

Configure EVPN to extend the customer VLANs:

```
set protocols evpn encapsulation vxlan
set protocols evpn extended-vni-list 101
set protocols evpn extended-vni-list 102
set switch-options vtep-source-interface lo0.0
set switch-options route-distinguisher 105.105.105.105:1
set switch-options vrf-target target:1:1
commit
```

Configure ESIs for multihomed interfaces:

```
set interfaces ae0 esi 00:11:11:11:11:11:11:11:11
set interfaces ae0 esi all-active
set interfaces ae1 esi 00:22:22:22:22:22:22:22:22
set interfaces ae1 esi all-active
commit
```

The overlay configuration on other EVPN LEAF PEs will be similar and is provided in the Appendix.

BL-1

Configure I-BGP for routing in the overlay:

```
set protocols bgp group OVERLAY type internal
set protocols bgp group OVERLAY local-address 102.102.102.102
set protocols bgp group OVERLAY family evpn signaling
set protocols bgp group OVERLAY local-as 65000
set protocols bgp group OVERLAY neighbor 101.101.101.101 description BL-1
set protocols bgp group OVERLAY neighbor 105.105.105.105 description LEAF-1
set protocols bgp group OVERLAY neighbor 106.106.106.106 description LEAF-2
set protocols bgp group OVERLAY neighbor 107.107.107.107 description LEAF-3
set protocols bgp group OVERLAY neighbor 108.108.108.108 description LEAF-4
set protocols bgp group OVERLAY neighbor 109.109.109.109 description LEAF-5
commit
```

Configure the customer VLANs:

```
set vlans VLAN-101 VLAN-id 101
set vlans VLAN-101 vxlan vni 101
set vlans VLAN-102 VLAN-id 102
set vlans VLAN-102 vxlan vni 102
commit
```

Configure EVPN to extend the customer VLANs:

```
set protocols evpn encapsulation vxlan
set protocols evpn extended-vni-list 101
set protocols evpn extended-vni-list 102
set switch-options vtep-source-interface lo0.0
set switch-options route-distinguisher 101.101.101.101:1
set switch-options vrf-target target:1:1
commit
```

Configure the L3 VRF and its interfaces:

```
set interfaces irb unit 101 virtual-gateway-accept-data
set interfaces irb unit 101 virtual-gateway-esi 00:66:66:66:66:66:66:66:66
set interfaces irb unit 101 virtual-gateway-esi all-active
set interfaces irb unit 101 family inet address 18.18.18.1/24 virtual-gateway-address 18.18.18.100
set interfaces irb unit 102 virtual-gateway-accept-data
set interfaces irb unit 102 virtual-gateway-esi 00:77:77:77:77:77:77:77:77
set interfaces irb unit 102 virtual-gateway-esi all-active
set interfaces irb unit 102 family inet address 19.19.19.1/24 virtual-gateway-address 19.19.19.100
set interfaces lo0 unit 1 family inet address 101.101.101.102/32
set routing-instances VRF-1 instance-type virtual-router
set routing-instances VRF-1 interface irb.101
```

```

set routing-instances VRF-1 interface irb.102
set routing-instances VRF-1 interface lo0.1
set vlans VLAN-101 l3-interface irb.101
set vlans VLAN-102 l3-interface irb.102
commit

```

Configure the L3 multicast protocols on the L3 VRF:

```

set routing-instances VRF-1 protocols pim interface lo0.1 mode sparse
set routing-instances VRF-1 protocols pim rp local address 101.101.101.102
set routing-instances VRF-1 protocols pim interface all mode sparse
set routing-instances VRF-1 protocols pim interface irb.102 priority 2
commit

```

The overlay configuration on BL-2 will be similar and is provided in the Appendix.

Configuring the CE devices

CE-1

Configure the interfaces and bridge-domains:

```

set chassis aggregated-devices ethernet device-count 1
set interfaces ge-0/0/0 description "T0 Host-1"
set interfaces ge-0/0/0 unit 0 family bridge interface-mode trunk
set interfaces ge-0/0/0 unit 0 family bridge VLAN-id-list 101
set interfaces ge-0/0/0 unit 0 family bridge VLAN-id-list 102
set interfaces ge-0/0/1 gigether-options 802.3ad ae0
set interfaces ge-0/0/2 gigether-options 802.3ad ae0
set interfaces ae0 aggregated-ether-options lacp active
set interfaces ae0 aggregated-ether-options lacp periodic fast
set interfaces ae0 description "T0 LEAF-1_LEAF-2"
set interfaces ae0 unit 0 family bridge interface-mode trunk
set interfaces ae0 unit 0 family bridge VLAN-id-list 101
set interfaces ae0 unit 0 family bridge VLAN-id-list 102
set bridge-domains BD-101 domain-type bridge
set bridge-domains BD-101 VLAN-id 101
set bridge-domains BD-102 domain-type bridge
set bridge-domains BD-102 VLAN-id 102
commit

```

The overlay configuration on other CE devices will be similar and is provided in the Appendix, and refer to the Appendix for complete base configuration on all devices in the topology.

Chapter 4

EVPN Intra-VLAN Multicast Without Optimization

This chapter explores how non-optimized multicast works in an EVPN data center fabric. It first describes multicast in a topology with single-homed EVPN devices to allow us to become conversant with the procedures and terminologies. Later the chapter explores procedures for multicast forwarding in a topology with multi-homed EVPN devices (like DF/NDF-based forwarding and local-bias based forwarding).

By the end of this chapter, you should have a fair understanding of:

- Intra-subnet multicast in an EVPN DC fabric
- L2-switched multicast procedures in EVPN multihomed topologies
- Overall multicast forwarding rules for L2 multicast in EVPN

Physical Topology For an EVPN Fabric

Figure 4.1 illustrates a typical IP-CLOS physical topology. Typically, there are two Border LEAF devices (BL) and several LEAF devices (in the order of hundreds). Also, there are two (or four) Leaf Spine (LS) devices that participate in the EVPN underlay for physical connectivity between devices in the EVPN fabric.

Typically, within the fabric, the multicast sources and hosts reside behind LEAF devices.

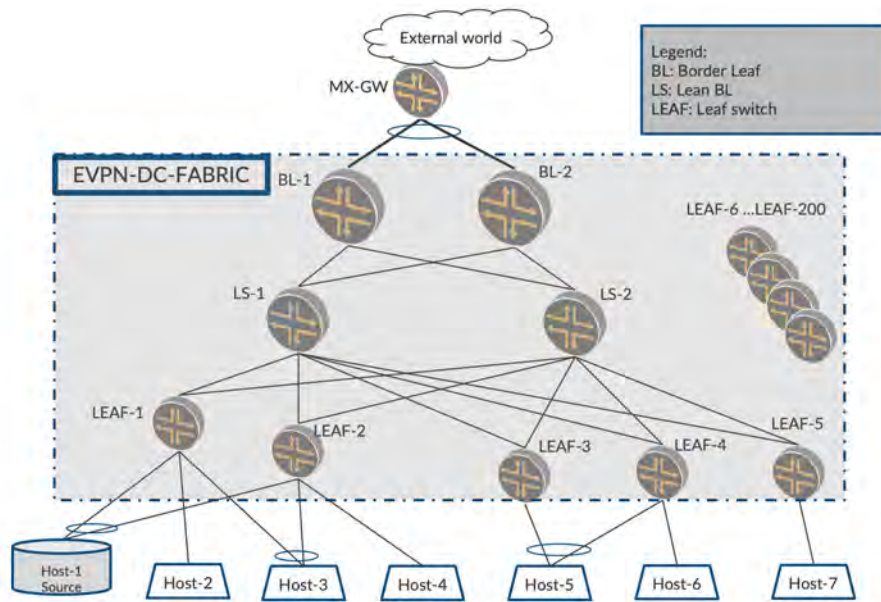


Figure 4.1 Typical IP-CLOS Physical Topology

Logical Topology

For illustrating the procedures for non-optimized multicast in this chapter, Figure 4.2 shows a logical topology where an EVPN network is configured only on the BL and LEAF devices. Typically, the Lean Spine layer helps with underlay alone and is not configured with EVPN.

Initially, we will describe single-homed hosts. The LEAF devices from LEAF-6 to LEAF-200 have access links and have hosts behind them. To explain the principles and procedures, we will use LEAF-1 to LEAF-5. The same procedures would apply to the rest of the LEAF devices, too.

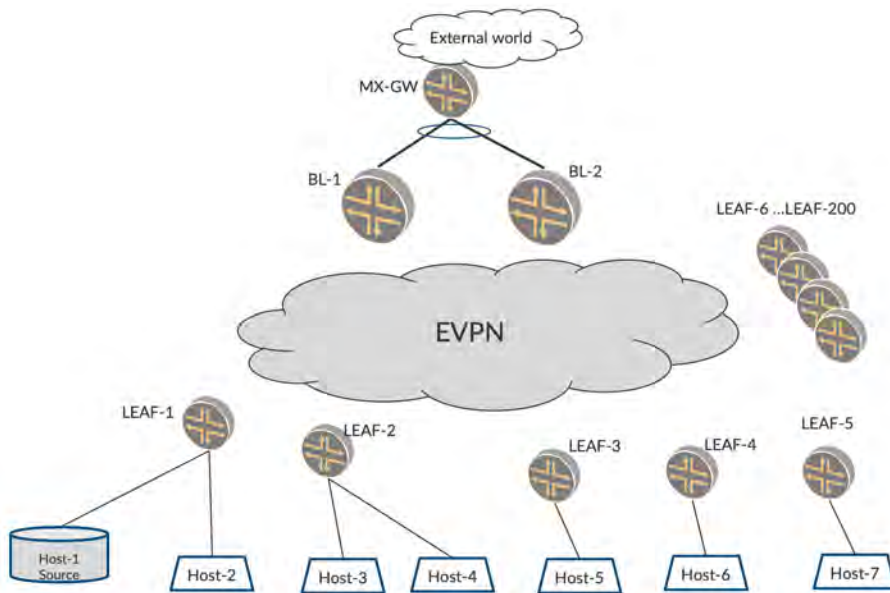


Figure 4.2 Logical Singlehomed Topology

Intra-subnet Multicast without Optimization

In this section, let's explore the procedures for Intra-subnet multicast forwarding within the DC fabric on a single VLAN (L2-switched multicast). In this topology, there is only one VLAN, VLAN-101. Typically, four sets of behaviors have to be considered with respect to multicast forwarding. (The configuration for the EVPN devices is included later in this chapter.)

- No sources and listeners exist
- Sources do not exist but listeners exist
- Sources exist but no listeners exist
- Sources and listeners exist

No Sources and Listeners Exist

This is straightforward in that it is a fabric with no multicast listeners or sources started yet.

Sources Do Not Exist but Listeners Exist

In this case, the listeners (hosts represented by Host-3 and Host-5 in Figure 4.3) are interested in traffic for a particular group G1, say 235.1.1.1. Since there is no multicast traffic started for that group yet, the listeners don't receive any traffic.

Sources Exist but No Listeners Exist

Consider a case where the multicast source (represented by Host-1 Source in Figure 3) has started to send the traffic for group G1, say 235.1.1.1, but there are no listeners yet for that group. In this case of intra-subnet multicast forwarding, the LEAF-1 that receives the multicast traffic will forward (L2-switch) the traffic to all the Layer 2 access interfaces on that VLAN. ie., towards Host-2. (*SH-FWD*). Also, LEAF-1 will not forward the traffic back to Host-1 based on Split Horizon. (*ACCESS-SPLIT-HRZN*).

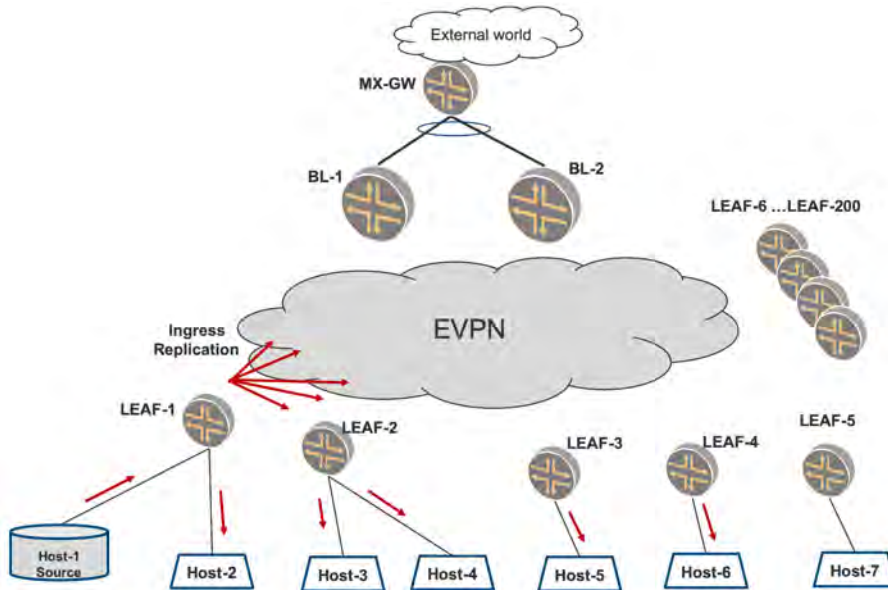


Figure 4.3 Layer 2 Multicast Traffic Flooding

In Figure 4.3, LEAF-1 will forward the traffic towards the EVPN core to all the EVPN PEs using Ingress Replication. The PEs to which traffic is sent are determined by the remote PEs' participation in the VLAN, VLAN-101. (*CORE-IMET-FWD*)

The participation of a LEAF in a VLAN is determined by virtue of EVPN Type-3s received from them. You can see all LEAFs except the LEAF-5 host, VLAN-101. Therefore, LEAF-5 would not send a Type-3 for VLAN-101. The multicast traffic will be Ingress Replicated to both BLs and all LEAFs except LEAF-5. (*CORE-IMET-SKIP*)

All the LEAF and BL devices that received the traffic from the EVPN core will forward this traffic onto all their access interfaces, irrespective of the presence of listeners as shown in Figure 4.3. LEAF-2 will forward towards Host-3 and Host-4.

LEAF-3 will forward towards Host-5, and so on. Even when there are no listeners in the VLAN, you can see that the multicast traffic is *'flooded everywhere'*.

This may not be desirable because such a high volume of traffic unnecessarily flooding everywhere may affect bandwidth utilization in the core and the access interfaces of different EVPN devices in the fabric. Flooding is one of the characteristics of EVPN multicast without optimization. Later chapters describe how optimization procedures help address this flooding problem.

Sources Exist and Listeners Exist

Consider the case where the source is sending traffic for group G1, say 235.1.1.1, and there are listeners in the fabric as shown in Figure 4.4. Therefore, Host-3 and Host-5 are interested in traffic for group G1 as shown. Listener interest is conveyed by the hosts by sending an IGMP report. Listener interest is represented in Figure 4.4 by a green circle on the host.

These IGMP hosts are expressing interest in traffic for group G1 from any source by sending IGMP report (*,G1). As described in the last section, by virtue of flooding multicast traffic everywhere, it reaches the interested listeners in the fabric and is consumed by them.

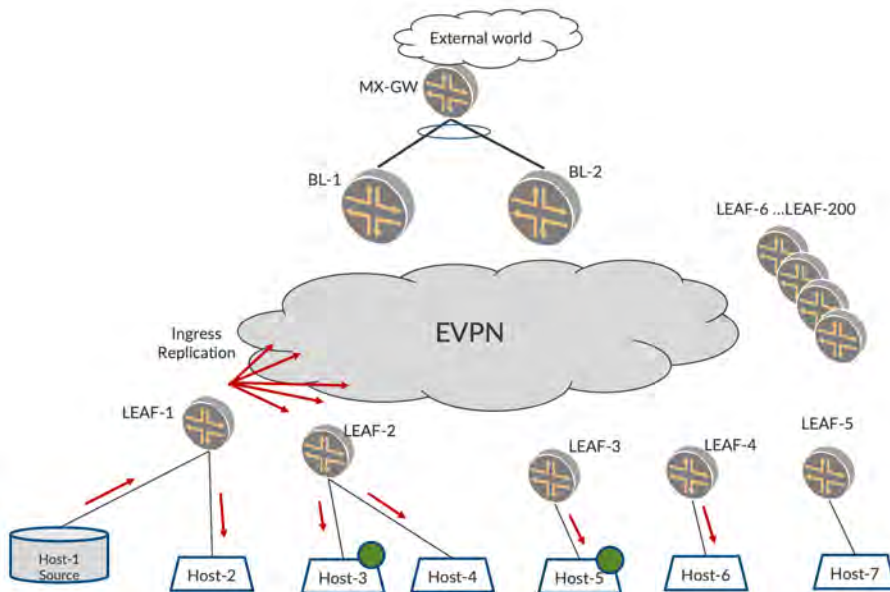


Figure 4.4

L2 Multicast Traffic Flooding with Listeners

Intra-subnet Multicast in EVPN Multihomed Topologies

One of the advantages of EVPN deployment is the multihoming feature with Ethernet Segment Identifier (ESI). An ESI helps in grouping a set of links on different EVPN devices such that the devices that host a particular ESI will consider themselves multihomed on that ESI. This is achieved by the devices exchanging BGP EVPN Type 1 and Type-4 routes and deducing the multihomedness for the ESI.

Typically, the set of links that are grouped as an ESI are terminated on the CE in an aggregated Ethernet lag (AE) interface. Overall, the EVPN ESI paradigm will ensure redundancy and load balancing features from the multihomed PEs towards the CE. The AE interface bundle on the CE will ensure redundancy and load balancing features towards the multihomed PEs.

Overall, for the case of multihomed listeners, the objective is to ensure that the listeners do not receive duplicate copies of the same traffic. For the case of multihomed sources, the objective is to ensure that the traffic is not looped back towards the source. Consider the logical topology for EVPN Multihoming in Figure 4.5.

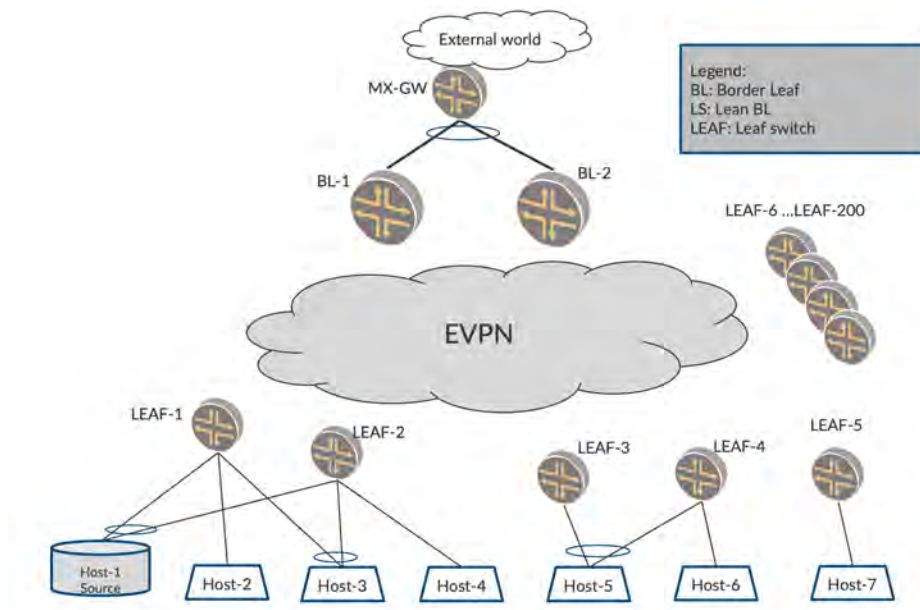


Figure 4.5 Logical Multihomed Topology

EVPN multihoming with ESI is a L2 feature. Therefore, two PE are considered multihomed on an ESI (per VLAN per EVPN instance). This section describes how LEAF devices are multihomed. BL devices running PIM L3-routing being multihomed to tenant routers (titled as External Multicast) are described in detail in later chapters.

We shall describe this in order to illustrate the BUM forwarding rules:

- Listener is behind multihomed LEAFs and source is single-homed.
- Source is behind multihomed LEAF.
- Source is behind a LEAF that has multihomed listener on a different ESI.

Listeners Behind Multihomed LEAF Devices and Source are Singlehomed

Consider the topology in Figure 4.6 where two LEAF devices, LEAF-3 and LEAF-4, are multihomed to Host-5. The rationale for multihoming Host-5 to LEAF-3 and LEAF-4 is that, in case of failure of one of the LEAFs, the traffic resumes over the other LEAF as soon as possible (resiliency). Also, if there are multiple unicast flows going from LEAF-1 to Host-5, some flows will be sent over LEAF-3 and others over LEAF-4 (load balancing).

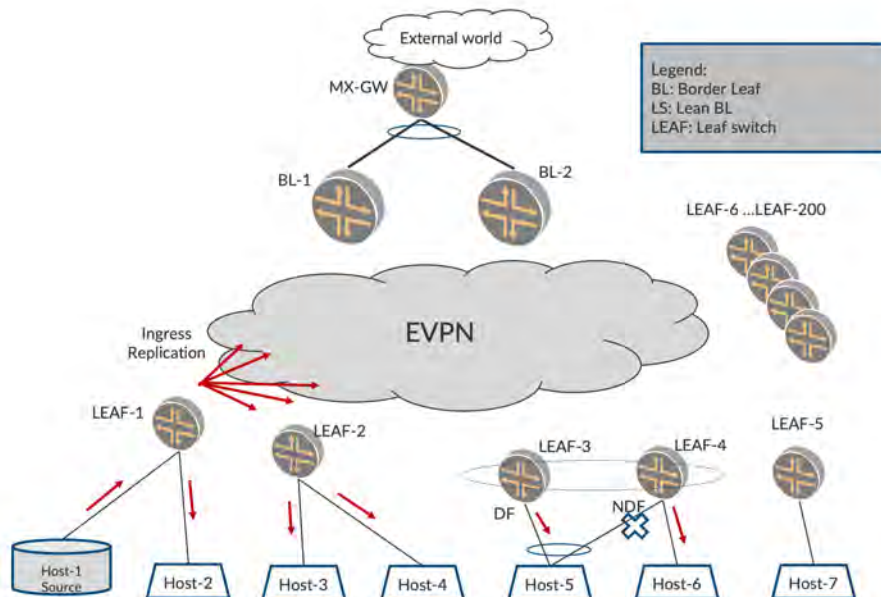


Figure 4.6 Listener Behind Multihomed PEs

For multicast traffic, care should be taken that the traffic from EVPN core is not sent by both LEAF-3 and LEAF-4, lest it result in duplicates for Host-5. (*Duplicates are a worse problem for multicast applications than traffic loss!*)

The Ingress LEAF, LEAF-1, will flood multicast traffic to both LEAF-3 and LEAF-4. Both LEAFs receive the traffic. However, only one LEAF should forward the traffic onto the ESI interface where they have realized the multihomed relationship.

To achieve this, the multihomed LEAFs determine the DF for a ESI amongst the PEs that are multihomed on that ESI. This determination is done by running a DF election algorithm (MOD based, local-preference, etc.,) based on Type-4 routes. Once this determination is completed, one amongst the multihomed PEs for the ESI is elected as the DF and the other multihomed PEs for that ESI are marked as NDF (non-DF).

When multicast traffic arrives from the EVPN core, rules for forwarding are as below:

- On a single-homed access interface, flood the traffic.
- On a multihomed access interface, if elected as DF, flood the traffic.
- On a multihomed access interface, if marked as NDF, don't flood the traffic.

In Figure 4.6 LEAF-3 and LEAF-4, who are being multihomed towards Host-5 over an ESI, run an election for DF. Say LEAF-3 is elected as the DF and LEAF-4 is NDF for the ESI. When LEAF-3 receives traffic from the core (Ingress Replicated from LEAF-1), it floods to Host-5, since it is the DF on the ESI (*CLASSICAL-DF-NDF*).

LEAF-4, on receiving traffic from the EVPN core, does not flood the traffic towards Host-5 since it has marked the ESI as NDF (*CLASSICAL-DF-NDF*). Thus, Host-5 does not receive duplicates. LEAF-4, however, floods the traffic to Host-6 since it is a singly-homed access interface.

Multicast Source Behind Multihomed LEAF Devices

Consider Figure 4. 7 where the multicast source is multihomed to two LEAF devices over an ESI where LEAF-1 is the NDF and LEAF-2 is the DF. When the source starts sending traffic, by virtue of an AE interface hashing, the traffic can be sent on either of the members of the LAG bundle. Therefore, Host-1 can send the traffic to either LEAF-1 or to LEAF-2.

Let's say the multicast traffic is sent to LEAF-1. If no special handling for this scenario is undertaken, the following would occur based on (*CLASSICAL-DF-NDF*) procedures described earlier. LEAF-1 will send traffic towards the core. LEAF-2 on receiving this traffic from the core will flood on interfaces where it is the elected DF.

In this case, since LEAF-2 is the elected DF towards Host-1, it will end up forwarding back to the source. This is not correct behavior. Hence, special handling is required such that LEAF-2 does not send back the traffic on the ESI that it is multihomed to LEAF-1.

MPLS has the split-horizon label to handle such scenarios. How can this problem be addressed in VXLAN?

If it was possible for LEAF-2 to deduce that the packet was sent by LEAF-1, LEAF-2 can program its forwarding such that if packets come in from LEAF-1, it will skip forwarding on those interfaces that are multihomed with LEAF-1. Let's examine this in detail.

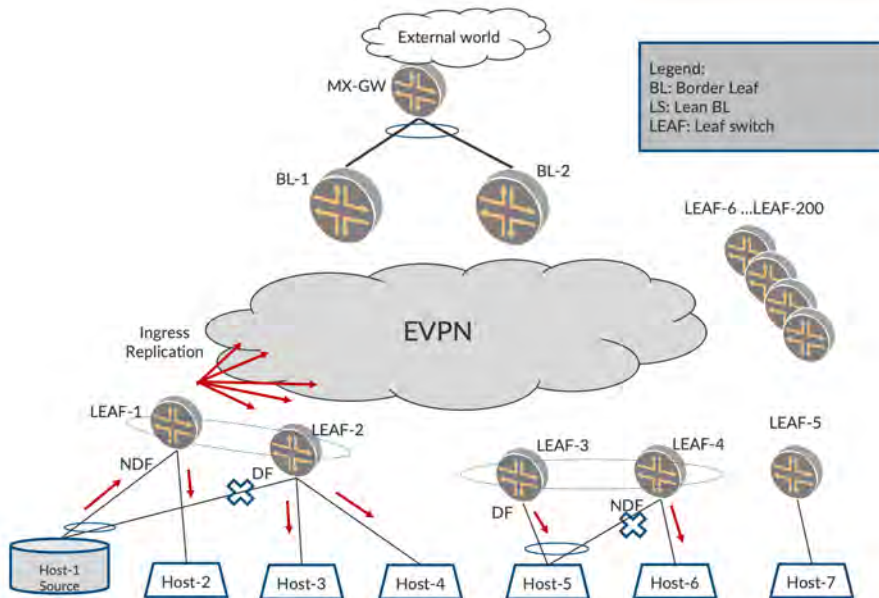


Figure 4.7

Source Behind Multihomed PEs

LEAF-1 sends the multicast packets with the source-VTEP IP of LEAF-1, say S-VTEP-IP-1. LEAF-2 knows that LEAF-1 has S-VTEP-IP-1 as its source-VTEP from BGP Type-3 routes.

LEAF-2 walks its access interfaces/ESIs and builds a *multihomed interface list with LEAF-1* where it is multihomed to LEAF-1. LEAF-2 then builds a rule in forwarding such that, when a packet arrives with S-VTEP-IP-1, it will skip forwarding on this multihomed interface list with LEAF-1. Such forwarding rules are to be built for each LEAF that LEAF-2 is multihomed with. Let's refer to this as (*DST-LOCAL-BIAS*).

With the forwarding being programmed, when LEAF-2 receives traffic from S-VTEP-IP-LEAF-1, the traffic is not sent back to the multihomed interface but flooded towards Host-3 and Host-4 (since these interfaces are not multihomed with LEAF-1).

Source is Behind a LEAF That Has a Multihomed Listener

Before we revisit the multihomed forwarding rules and rewrite the overall multicast L2 forwarding rules, we need to consider one more topology where an EVPN LEAF device has a local source and has a multihomed interface that are on different interfaces/ESIs.

Consider the topology shown in Figure 4.8 where LEAF-1 has a multicast source on a single-homed interface and a listener Host-3 on a multihomed interface shared with LEAF-2. Based on an election, let's say that LEAF-1 is the NDF and LEAF-2 is the DF for the ESI.

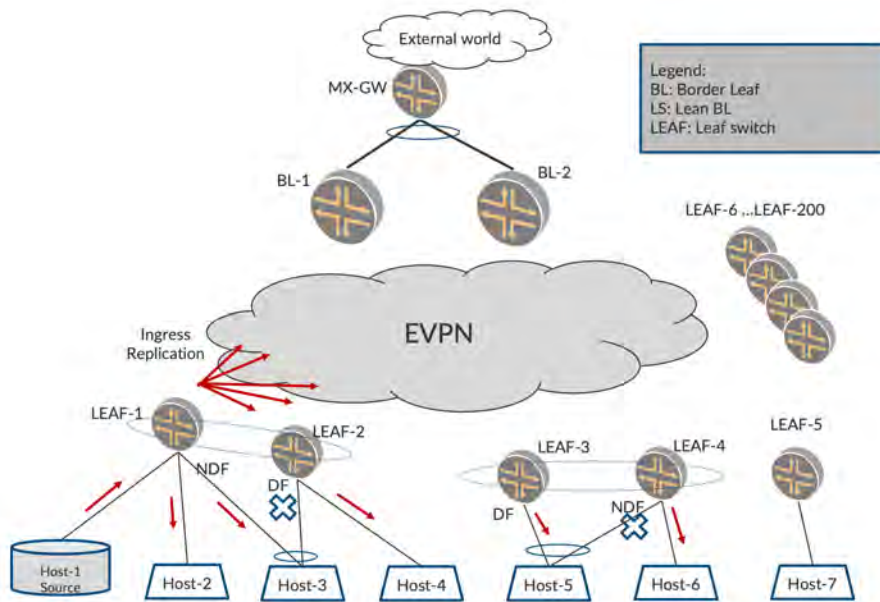


Figure 4.8 Source Behind Multihomed Listener

Per rules described earlier, LEAF-2, on receiving the traffic from the core, will not forward the traffic onto the multihomed interface list with LEAF-1. Therefore, LEAF-2 will not forward the traffic to Host-3.

LEAF-1 being the NDF towards Host-3 is an uncanny situation here. LEAF-2 will not forward the traffic to Host-3 due to (*DST-LOCAL-BIAS*). This is to ensure that traffic received on multihomed interfaces does not get looped back to the

source. However, LEAF-1 being NDF towards Host-3 cannot forward to Host-3 based on (*CLASSICAL-DF-NDF*) rules. How is Host-3 to get traffic?

To address this scenario, let's deviate a little from the (*CLASSICAL-DF-NDF*) rules. When LEAF-1 receives *traffic from a local access interface, it will flood the traffic onto __all__ the other local access interfaces* irrespective of whether it is DF on the target interface or not. This can be referred to as (*SRC-LOCAL-BIAS*).

In Figure 4. 8, LEAF-1 on receiving traffic from the source from the access interface will flood on all its access interfaces irrespective of DF/NDF. Therefore LEAF-1 will forward the traffic to Host-3 despite being NDF on the interface since the traffic arrived from an access interface. Thus, Host-3 will receive the traffic from LEAF-1.

LEAF-2 on receiving the traffic from the core from LEAF-1, will determine the s-VTEP to S-VTEP-IP-LEAF-1 and will skip forwarding to 'MH interface list to LEAF-1' (*DST-LOCAL-BIAS*). Thus, LEAF-2 will __not__ forward to Host-3. LEAF-2 will forward to Host-4 since it is not a multihomed interface with LEAF-1.

Putting It All Together for Intra-VLAN Multicast

Based on the forwarding rules described so far in this chapter, let's correlate the behavior in our sample topology *vis-à-vis* the forwarding rules. Please refer to the Traffic Verification section for statistics.

In Figure 4.9 LEAF-1 is the NDF on the MH-interfaces that go to Host-1 and Host-3. LEAF-3 is the DF on the MH interface that goes to Host-5. Host-1 is the source of multicast traffic. Host-1, based on the hash of its AE bundle, can send the traffic to either LEAF-1 or LEAF-2. Say it sends to LEAF-1.

Here, LEAF-1 performs the actions below:

- Does not send the traffic back to Host-1 (*ACCESS-SPLIT-HRZN*)
- Ingress Replicates traffic to all remote PEs over VTEP (*CORE-IMET-FWD*)
- Does not send to LEAF-5 as no Type-3 for VLAN-101. (*CORE-IMET-SKIP*)
- Sends traffic to Host-2 since it is single-homed interface (*SH-FWD*)
- Sends to Host-3, though it is NDF and it is access traffic (*SRC-LOCAL-BIAS*)

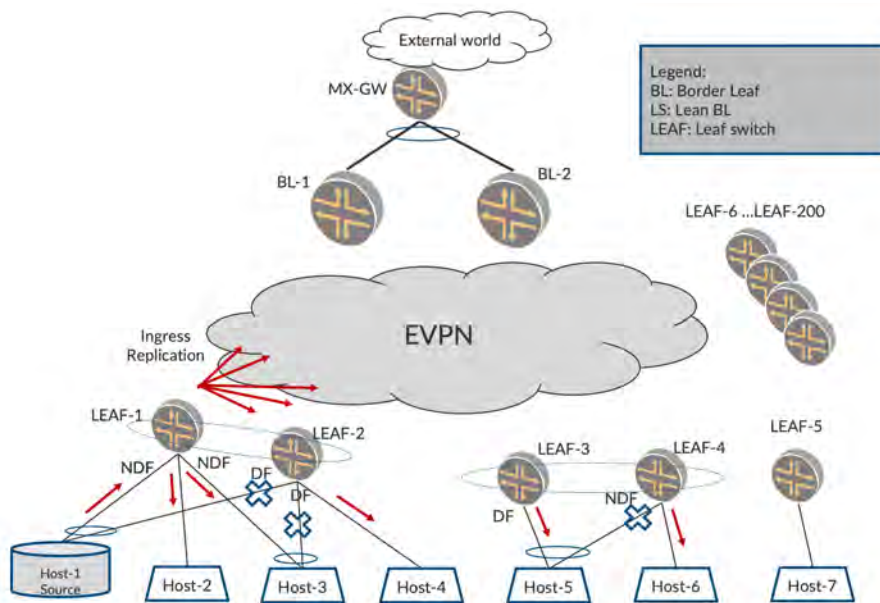


Figure 4.9 Forwarding Rules

LEAF-2, on receiving the traffic from core, performs the actions listed below:

- Does not send traffic back to core (CORE-SPLIT-HRZN)
- Sends traffic towards Host-4 since it is a single-homed interface. (SH-FWD)
- Does __not__ send traffic to Host-1 (DST-LOCAL-BIAS)
- Does __not__ send the traffic to Host-3 (DST-LOCAL-BIAS)

LEAF-3, on receiving traffic from core, performs the actions below:

- Sends to Host-5 since it is DF on that interface (CLASSICAL-DF-NDF)

LEAF-4, on receiving traffic from core, performs the actions below:

- Does __not__ send to Host-5 since it is NDF (CLASSICAL-DF-NDF)
- Forwards the traffic to Host-6 since it is a single-homed interface (SH-FWD)
- LEAF-5 does not receive traffic from core since it does not host VLAN-1.

Overall Forwarding Rules For Multicast Traffic in EVPN

Let's enhance the rules stated earlier in this chapter and take multihomed into account.

(*CLASSICAL-DF-NDF*)

- When traffic arrives from EVPN core from a PE who I am __not__ multihomed with
- On a single-homed access interface, flood the traffic.
- On a multihomed access interface, if elected as DF, flood the traffic
- On a multihomed access interface, if marked as NDF, don't flood the traffic

(*DST-LOCAL-BIAS*)

- When traffic arrives from EVPN core from a PE, say PE-X, who I am multihomed with
- On a single-homed access interface, flood the traffic.
- On a multihomed access interface where multihomed with PE-X, don't flood traffic, irrespective of DF/NDF.
- On a multihomed access interface where not multihomed with PE-X, flood if I am DF on that interface.
- On a multihomed access interface where not multihomed with PE-X, do not flood if I am NDF on that interface.

(*SRC-LOCAL-BIAS*)

- When traffic arrives from access interface:
- Flood on all the other access interfaces irrespective of DF/NDF.
- Ingress Replicate to core to all PEs that host the VLAN.

The above procedures, (2) and (3), for forwarding BUM traffic are generally referred to as *local-bias*. As the name suggests, when traffic arrives on a local access interface, the LEAF device is given the bias to forward it onto other access interfaces irrespective of DF/NDF. The remote multihomed devices do not forward onto the multihomed access interface irrespective of DF/NDF. Therefore, the local PE is given preference to forwarding over the remote PE (hence the term *local-bias*).

We have to keep in mind that these rules of local bias are applicable only in EVPN-VXLAN. With EVPN-MPLS, the usage of the split-horizon label per ESI addresses the multihomed scenarios.

NOTE The procedures for split-horizon label are beyond the scope of this book.

Chapter Summary

This chapter has explored different BUM forwarding rules in single-homed and multihomed topologies. We illustrated how multicast traffic is flooded everywhere throughout the EVPN fabric towards the core and all access interfaces. Multihoming forwarding rules help to ensure that listeners do not receive duplicates or are looped back. These rules take into account the DF/NDF status of an ESI, whether the traffic arrived from a PE with which it is multihomed, and whether traffic arrived on an access interface.

Chapter 5 explores the challenges with Ingress Replication and how it can be mitigated. Configurations and verifications now follow.

Configuration

Figure 4.10 is the reference topology.

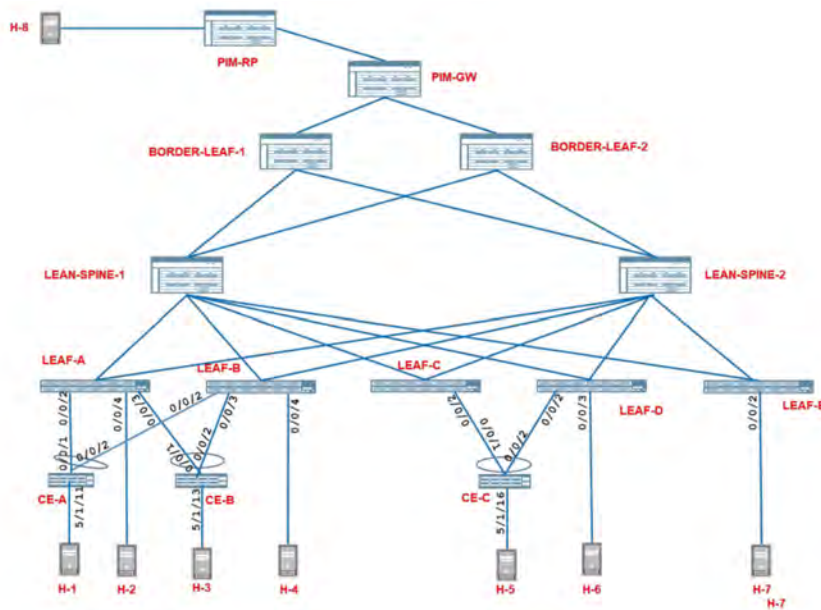


Figure 4.10

Reference Topology

Now let's focus on the configuration and verification of the intra-VLAN Multicast functionality described in this chapter, including multihoming in EVPNvVXLAN and the commands used. In this section, we will see the EVPN intra-VLAN multicast traffic forwarding behavior, particularly the *flood everywhere* aspect in the absence of any optimizations.

The basic underlay and overlay configurations listed in Chapter 3 are sufficient for this section. For all of our intra VLAN discussions, we will focus on VLAN-101 (VLAN id 101, VNI 101). Please note that LEAF-5 does not Host VLAN-101/VNI-101.

Traffic Verification

From Host-1, start sending multicast traffic at 10 pps (packets per second) for group 225.1.1.1 in VLAN-101. Note that, as of now, no receivers have actually expressed interest in receiving this traffic.

From the RT statistics in Figure 4.11, you can see that Host-1 sends traffic at 10 pps, which is received by all the Hosts within the DC that are part of VLAN-101 (Host-2 to Host-6), though none of them are interested in the traffic. Host-7 alone, which is not part of VLAN-101, is spared from the traffic. Host-8 is outside of the fabric and it can be ignored for now.

	Stat Name	Port Name	Link State	Frames Tx. Rate	Valid Frames Rx. Rate
1	10.216.45.202/Card20/Port01	HOST-1	Link Up	10	0
2	10.216.45.202/Card03/Port01	HOST-2	Link Up	0	10
3	10.216.45.202/Card20/Port02	HOST-3	Link Up	0	10
4	10.216.45.202/Card03/Port02	HOST-4	Link Up	0	10
5	10.216.45.202/Card20/Port03	HOST-5	Link Up	0	10
6	10.216.45.202/Card03/Port03	HOST-6	Link Up	0	10
7	10.216.45.202/Card03/Port04	HOST-7	Link Up	0	0
8	10.216.45.202/Card20/Port04	HOST-8	Link Up	0	0

Figure 4.11

RT Stats

Multicast Traffic Outputs - LEAF-1

Host-1 is multihomed to LEAF-1 and LEAF-2. So the traffic from Host-1 towards LEAF-1/LEAF-2 may be load balanced and can arrive on either LEAF-1 or LEAF-2. In our case, the multicast traffic arrives on access interface, ae0 on LEAF-1.

The traffic is not flooded back on the incoming interface, ae0.0:
(ACCESS-SPLIT-HRZN).

On LEAF-1, the traffic is forwarded on the other single-homed access interface, xe-0/0/4 towards Host-2: (SH-FWD).

The traffic is forwarded on all multihomed access interfaces, ae1.0 on LEAF-1, irrespective of the DF/NDF status (SRC-LOCAL-BIAS):

```
lab@LEAF-1> monitor interface traffic detail
Interface  Link  Input packets      (pps)  Output packets      (pps)  Description
...
xe-0/0/4   Up    0                  (0)    2474                (10)   T0 Host-2
...
ae0        Up    2467              (10)    0                  (0)    T0 CE-1
ae1        Up    0                 (0)    2470                (10)   T0 CE-2
...
```

The multicast traffic is also forwarded on the VTEPs towards BL-1 (101.101.101.101) and BL-2 (102.102.102.102).

Also forwarded on the VTEPs towards LEAF-2 (106.106.106.106), LEAF-3 (107.107.107.107), and LEAF-4 (108.108.108.108): (CORE-IMET-FWD).

The traffic is not forwarded on the VTEP towards LEAF-5 (109.109.109.109) (CORE-IMET-SKIP):

```
lab@LEAF-1> show interfaces vtep extensive | grep "VXLAN Endpoint Type: Remote|Output packets.*pps"
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 106.106.106.106...
Output packets: 2488 9 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 101.101.101.101...
Output packets: 2488 9 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 108.108.108.108...
Output packets: 2488 9 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 107.107.107.107...
Output packets: 2489 10 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 109.109.109.109...
Output packets: 0 0 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 102.102.102.102...
Output packets: 2489 10 pps
```

Multicast Traffic Outputs - LEAF-2

The multicast traffic arriving on LEAF-2 is not forwarded on the multihomed access interfaces, ae0 and ae1, although it is the DF on both of these interfaces since the source PE, LEAF-1, is a multihomed peer on the ESI of these interfaces (DST-LOCAL-BIAS). This ensures that there is no looping of traffic towards the multihomed source, Host-1, and no traffic duplication towards the multihomed Host, Host-3.

The traffic is forwarded on xe-0/0/4.0 (SH-FWD) towards the single-homed Host, Host-4:

```
lab@LEAF-2> monitor interface traffic detail
Interface  Link  Input packets      (pps)  Output packets      (pps)  Description
...
xe-0/0/4   Up    0                  (0)    2478                (10)   T0 Host-4
ae0        Up    553                (0)    213                 (0)    T0 CE-1
ae1        Up    0                  (0)    208                 (0)    T0 CE-2
...
```

The traffic is not sent back on any of the VTEPs. Though VTEPs are part of the flood next hop, the split horizon rules for BUM traffic arriving from a core ensure that the traffic is not sent back to the core. (CORE-SPLIT-HRZN).

Multicast Traffic Outputs - LEAF-3

The traffic arriving on LEAF-3 is not forwarded on the multihomed access interface, ae0.0, since it is the NDF (CLASSICAL-DF-NDF). This ensures that the multihomed Host, Host-5, does not receive duplicate traffic:

```
lab@LEAF-3> monitor interface traffic detail
Interface  Link  Input packets      (pps)      Output packets      (pps) Description
...
ae0        Up    0                  (0)         208                 (0) T0 CE-3
...
```

The traffic is also not sent back on any of the VTEPs (CORE-SPLIT-HRZN).

Multicast Traffic Outputs - LEAF-4

The multicast traffic arriving on LEAF-4 is forwarded on the multihomed access interface, ae0.0, since it is the DF and LEAF-1 is not a multihomed peer (see section 3.4.1): (CLASSICAL-DF-NDF).

The traffic is also forwarded on xe-0/0/3.0 (SH-FWD) towards the single-homed Host, Host-6:

```
lab@LEAF-4> monitor interface traffic detail
Interface  Link  Input packets      (pps)      Output packets      (pps) Description
...
xe-0/0/3    Up    0                  (0)         2480                (10) T0 Host-6
ae0         Up    505                (0)         2481                (10) T0 CE-3
...
```

The traffic is not sent back on any of the VTEPs (CORE-SPLIT-HRZN).

Multicast Traffic Outputs - LEAF-5

LEAF-5, since it does not Host the VLAN-101, does not receive the traffic from LEAF-1 at all.

Multicast Traffic Outputs - BL-1 and BL-2

The behavior of the two border LEAF devices will become significant only once we reach the chapter about inter-VLAN traffic forwarding. So until then, we will ignore the traffic forwarding behavior on these devices.

Detailed Control Plane Verification

Verifying the Flood Routes

For each VLAN, a PE builds a flood next hop consisting of all its access interfaces for that VLAN, and the VTEPs corresponding to the EVPN peers from which it has received Type-3 routes for that VLAN. This next hop is used to flood the multicast traffic in the VLAN.

For instance, on LEAF-1, the flood next hop is comprised of VTEPs corresponding to:

- BL-1 (vtep.32770)
- BL-2 (vtep.32774)
- LEAF-2 (vtep.32769)
- LEAF-3 (vtep.32772)
- LEAF-4 (vtep.32771)

NOTE The VTEP corresponding to LEAF-5 (vtep.32773) is not present in the flood next hop for VLAN-101.

```
lab@LEAF-1> show ethernet-switching flood route re-flood bridge-domain VLAN-101
Flood route prefix: 0x30000/51
Flood route type: RE_FLOOD
Flood route owner: __re_flood__
Flood group name: __re_flood__
Flood group index: 65534
Nexthop type: comp
Nexthop index: 1735
Flooding to:
  Name      Type      NhType      Index
  __all_ces__  Group      comp        1734
  Composition: split-horizon
  Flooding to:
    Name      Type      NhType      Index
    xe-0/0/4.0  CE        ucst        1732
    ae1.0       CE        ucst        1715
    ae0.0       CE        ucst        1714
Flooding to:
  Name      Type      NhType      Index
  __ves__    Group      comp        1724
  Composition: flood-to-all
  Flooding to:
    Name      Type      NhType      Index
    vtep.32769  CORE_FACING  venh        1751
    vtep.32770  CORE_FACING  venh        1756
    vtep.32771  CORE_FACING  venh        1761
    vtep.32772  CORE_FACING  venh        1762
    vtep.32774  CORE_FACING  venh        1765
```

The flood next hop information on the other PEs will be similar to the above.

Verification of Multihoming State

Each PE Hosting the ES, performs a DF election for the ES.

For the active/active Ethernet segments multihomed to LEAF-1 and LEAF-2, therefore, 00:11:11:11:11:11:11:11:11 and 00:22:22:22:22:22:22:22:22, LEAF-2 (106.106.106.106) is elected DF.

For the active/active Ethernet segment multihomed to LEAF-3 and LEAF-4, i.e. 00:33:33:33:33:33:33:33:33:33:33, LEAF-4 (108.108.108.108) is elected DF.

In most cases we will verify the states for a single ESI on LEAF-1 by itself. The states for other ESIs and on other PEs are similar and are left as an exercise for the reader.

DF Verification

Let's verify the DF/NDF state on LEAF-1:

```
lab@LEAF-1> show evpn instance designated-forwarder esi 00:11:11:11:11:11:11:11:11:11:11
...
  ESI: 00:11:11:11:11:11:11:11:11:11:11
    Designated forwarder: 106.106.106.106

lab@LEAF-1> show evpn instance designated-forwarder esi 00:22:22:22:22:22:22:22:22:22:22
...
  ESI: 00:22:22:22:22:22:22:22:22:22:22
    Designated forwarder: 106.106.106.106
```

Verification For ESI Status in Detail

The following command may be used to see more details on the DF election for an ES:

```
lab@LEAF-1> show evpn instance esi 00:11:11:11:11:11:11:11:11:11:11 extensive
...
  Number of local interfaces: 4 (4 up)
    Interface name  ESI                               Mode           Status    AC-Role
    .local..4      00:00:00:00:00:00:00:00:00:00:00  single-homed   Up        Root
    ae0.0          00:11:11:11:11:11:11:11:11:11:11  all-active     Up        Root
    ae1.0          00:22:22:22:22:22:22:22:22:22:22  all-active     Up        Root
    xe-0/0/4.0     00:00:00:00:00:00:00:00:00:00:00  single-homed   Up        Root
...
  ESI: 00:11:11:11:11:11:11:11:11:11:11
    Status: Resolved by IFL ae0.0
    Local interface: ae0.0, Status: Up/Forwarding
    Number of remote PEs connected: 1
      Remote PE      MAC label  Aliasing label  Mode
      106.106.106.106  0          0               all-active
    DF Election Algorithm: MOD based
    Designated forwarder: 106.106.106.106
    Backup forwarder: 105.105.105.105
...
  Router-ID: 105.105.105.105
...
```

Verifying ESI State: Forwarding/Blocked

Now let's verify the ESI state.

LEAF-1, since it is not the DF on multihomed access interfaces ae0.0 and ae01, marks them in "Blocking" mode for BUM traffic, while LEAF-2, being the DF, marks these interfaces in "Forwarding" mode:

```
lab@LEAF-1> show interfaces ae0.0 extensive
...
... EVPN multi-homed status: Blocking BUM Traffic to ESI, ...
...
    Local Bias Logical Interface Name: vtep.32769, Index: 558, VXLAN Endpoint Address: 106.106.106.106

lab@LEAF-2> show interfaces ae0.0 extensive
...
... EVPN multi-homed status: Forwarding, ...
...
    Local Bias Logical Interface Name: vtep.32769, Index: 556, VXLAN Endpoint Address: 105.105.105.105
```

The PE(s) to which this interface is multihomed can also be seen in the output. This information is used on this PE when applying the DST-LOCAL-BIAS for traffic arriving from the core.

Chapter 5

Assisted Replication

In Chapter 4 we explored the different procedures for multicast data forwarding like ingress replication, DF/NDF forwarding, and local bias rules.

Typically, data centers have several top-of-rack (TOR/LEAF) switches (in the order of hundreds) that have less port density, have less forwarding/processing capability, and are less expensive. The QFX5110, for instance, is such a device, and is the right fit for a LEAF role.

Also, a data center physical topology looks similar to Figure 5.1 and has the following components:

- **LEAF Layer:** There are several LEAFs, (QFX 5110s) that connect to the VM Hosts, bare-metal servers (BMS), etc.
- **Border LEAF:** There are couple of LEAF devices called Border-LEAF (BL) devices (QFX10K) that are used to connect the DC fabric to the outside world over a gateway (MX-GW).
- **Lean Spine:** There are couple of lean spine devices (LS) (QFX10K) that are used to physically interconnect the participating EVPN devices. These lean spine devices serve as physical interconnects only, do not host any access interfaces, and generally do not participate in an EVPN overlay. They are part of the BGP underlay.

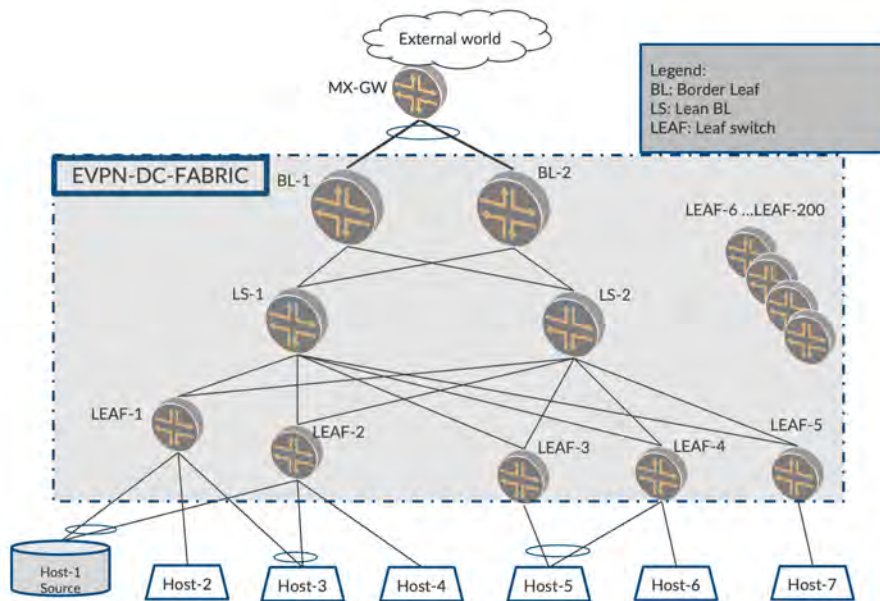


Figure 5.1 *Typical Data Center Physical Topology*

Ingress Replication Characteristics and Challenges

When multicast source sends traffic for different flows, say, for IPTV applications, the traffic rate is typically very high (in order of 4Mbps per flow, for example, HD streams). When such a high rate of traffic arrives on LEAF-1, LEAF-1 shall Ingress replicate the traffic to all other PEs in the fabric. To achieve this, for each packet LEAF-1 receives on the access interface, it replicates one copy per remote PE and sends it over the EVPN core as shown in Figure 5.2.

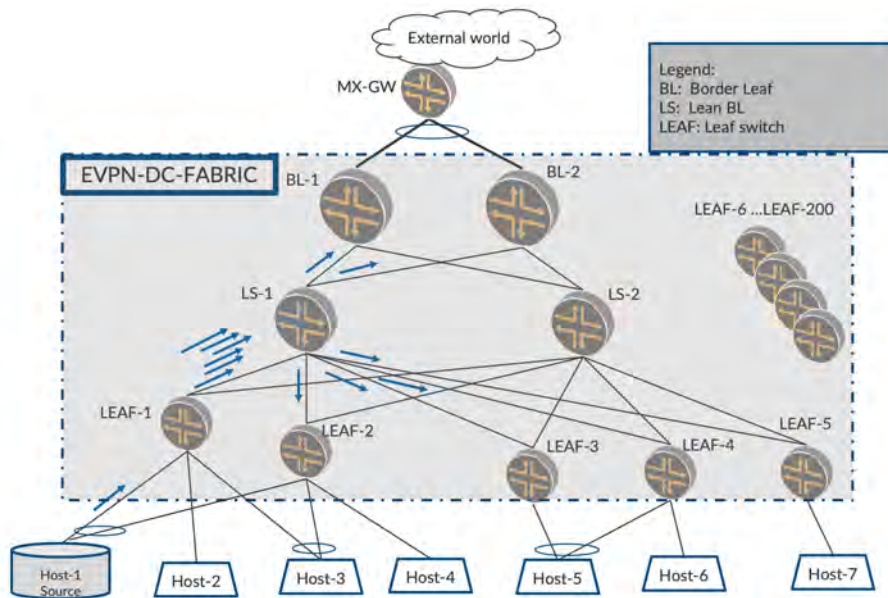


Figure 5.2 Ingress Replication Traffic Flow In the Underlay

With, say, 1000 flows at 4 Mbps per flow, and about 200 PEs in the Fabric, LEAF-1 sends out ($1000 * 200 * 4 \text{ Mbps} = 800\text{Gbps}$). Two situations arise here.

Since LEAF-1 is a device with less capacity, it is burdened to make 200 copies (one per PE) of the incoming heavy traffic flow. This can potentially melt the Ingress.

Though the overlay traffic is 800 Gbps distributed to 200 PEs, (4 Gbps per PE in the logical topology), LEAF-1 sends all the 800 Gbps traffic on the physical link between LEAF-1 and LS-1. Thereafter, LS-1, forwards these packets to each of the other PEs at 4 Gbps each. The link between LS-1 to LEAF-3, let's say, carries 4Gbps, which is expected and is not a problem. However, the link between LEAF-1 and LS-1 is severely overutilized leaving no space for other applications and potential traffic drops.

We have just illustrated this situation with one multicast source. If there are several sources behind other LEAFs sending traffic at high rates, many LEAFs will be effected as the links that go from the LEAF to the LS will also choke.

Assisted Replication

Using the above scenario, let's explore if there is anything that can be done to mitigate the problems of (i) the inability of LEAF-1 to replicate to several PEs, with the traffic coming in at high rate, and (ii) the overutilization of the link between LEAF-1 and LS-1, since that link has to carry the packets destined to all PEs in the underlay.

Wouldn't it be nice if LS-1, which typically is a high-end device, could '*assist*' in the responsibility of replication to other PEs? Yes. If LS-1 can assist LEAF-1 with the load of replication to other PEs, both the above problems are addressed. How?

Let's say LEAF-1 sends only one copy of the packet to LS-1, and LS-1 takes over the role of replication and replicates to other EVPN devices in the fabric. The LEAF-1 needs to send only one packet to the LS-1, so the replication load on LEAF-1 and the number of packets that flow over the link between LEAF-1 and LS-1 is reduced by a factor of 200.

Of course, LS-1 should have the capability and configuration to '*assist*' such replication. Also, LS-1 should exchange this capability with other EVPN devices and the other devices to offload the responsibility of replication to LS-1. Which tunnels the traffic should be forwarded onto will also be coordinated.

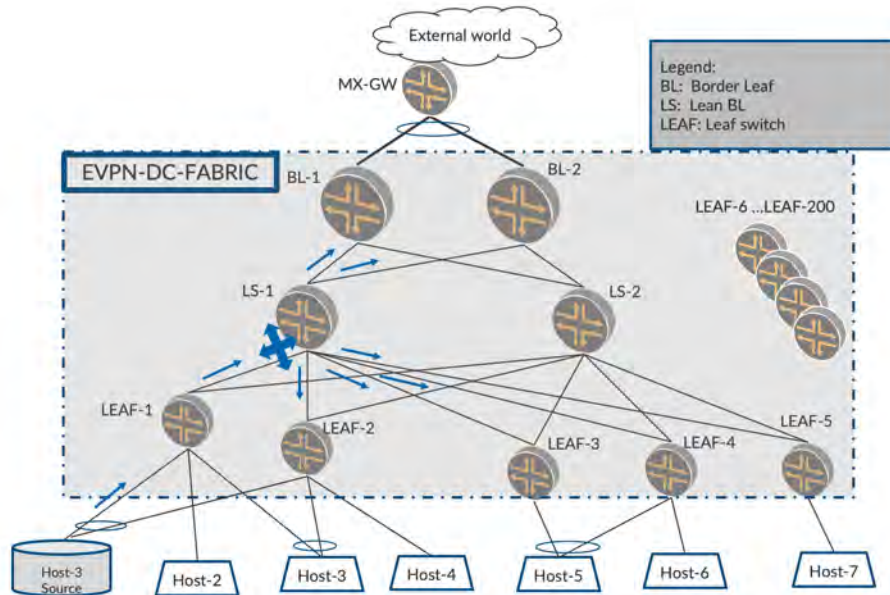


Figure 5.3

Assisted Replication Traffic Flow In the Underlay

Enter assisted replication. In this chapter, we describe *assisted replication* (AR), its characteristics, and the benefits it brings to the table in reducing load and complexity. In later chapters, once we have introduced optimized multicast, we shall also explore how AR, in conjunction with optimized multicast, effectively makes the DC Fabric operate in a slick manner for multicast traffic.

An Important Thing to Note

AR functionality is optional. That is to say, it is not mandatory to deploy AR to be able to optimize multicast traffic (a discussion of which is coming in later chapters). Some customers ensure multicast traffic is first optimized in the fabric, and then proceed with addressing the problems of load and link bandwidth utilization. However, for academic reasons, AR fits in well in this chapter in terms of a staged understanding of the different mechanisms of EVPN multicast.

Building Blocks of Assisted Replication

There are three components of the AR solution in the fabric as shown in Figure 5.4:

- the role of AR-LEAF
- the role of AR-Replicator
- the role of RNVE (Regular Network Virtualization Equipment): a device that does not support the AR feature

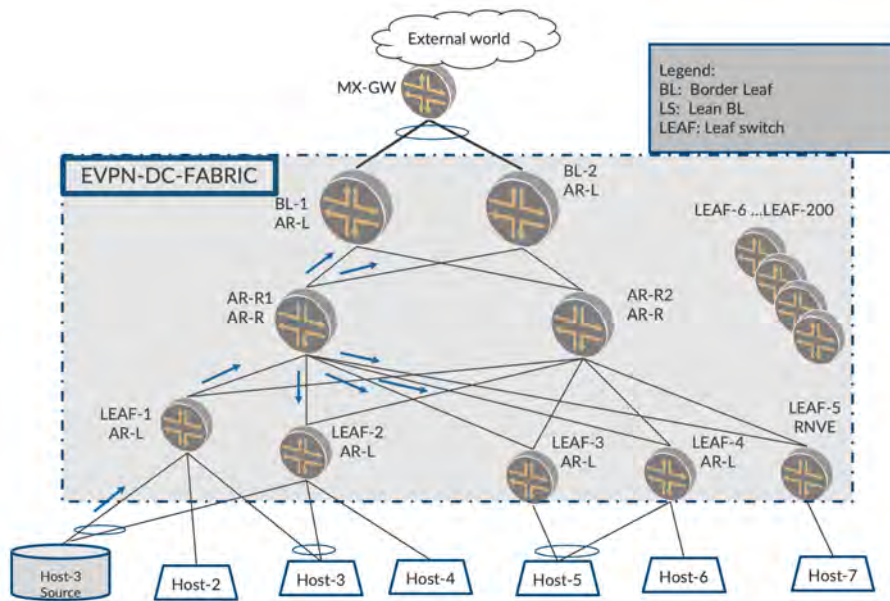


Figure 5.4

Assisted Replication Roles

The AR-LEAF in our topology will be LEAFs, and the AR Replicator role will be LS-1 and LS-2. LS-1 and LS-2, prior to AR, participate in the underlay build-up of the fabric. With AR, these devices will perform the additional roles of AR replicator in the overlay, too. *This way, there is no additional introduction of hardware and the existing device is utilized to solve the two hard problems of Ingress Replication described earlier.*

LEAF-1, LEAF-2, LEAF-3, and LEAF-4 are configured as AR-LEAF. (AR-L)

BL-1 and BL-2 are configured as AR-LEAF, too.

AR-R1 and AR-R2 are configured as the AR-Replicators. (AR-R)

There is no AR configuration on LEAF-5. So this makes LEAF-5, the RNVE.

Requirements of Assisted Replication In Fabric

The EVPN PEs should be designated with a specific AR role using a configuration (AR-LEAF/Replicator) or lack thereof. (RNVE)

The EVPN devices should be able to detect the presence of other AR devices and their roles.

The AR-capable devices (therefore, AR-LEAFs and AR-Replicators) should accommodate RNVE devices (therefore, AR-incapable devices).

The AR-LEAF PEs should be able to build AR-tunnels to AR-replicators to send multicast traffic to AR-Replicators.

The AR-Replicator PEs should be able to use existing IR-tunnels to send replicated multicast traffic to other AR-LEAF devices and RNVEs.

The AR-LEAF and AR-Replicators should be able to use the existing IR-tunnels to receive traffic from RNVEs.

Ensure EVPN multihoming behavior and forwarding are at parity with IR.

There should be provisions to support multiple AR-Replicator roles for load balancing of the replication. The AR-LEAF devices should be able to load-balance amongst the multiple AR-Replicator devices. In typical customer deployments, there would be two or four AR-Replicators in the Fabric for load balancing and resiliency.

Tunnel Building For Assisted Replication In Fabric

AR-R1 and AR-R2, by virtue of being replicators, advertise an AR Type-3 route in addition to the usual IR Type-3 of today. The AR Type-3 route parameters are used in building the AR-tunnels from the AR-LEAF to the AR-Replicator.

LEAFs 1-4, the AR-LEAFs, on receiving the AR-Type-3 route from AR-R1 and AR-R2, deduce the presence of replicators and build the AR-tunnel to the replicators.

The AR-LEAF detects multiple replicators, by virtue of how they build the load balancing capability in their AR-tunnels. Therefore, some flows will be sent to AR-R1 for replication, others to AR-R2.

AR-LEAF and AR-Replicators and the RNVE build the IR-tunnels with each other based on existing Type-3 routes.

The AR-LEAF and AR-Replicators keep their IR-tunnels to each other for two reasons: (1) for a LEAF to receive replicated traffic from AR-R, and, (2) to fall back to IR when the AR-tunnel fails for any reason.

Assisted Replication Traffic Forwarding Rules

Assisted Replication Traffic Forwarding with Traffic Arriving From Behind AR-LEAF

When multicast traffic for a flow arrives on the access interface on LEAF-1, LEAF-1 sends the traffic to one of the replicators, such as AR-R1. LEAF-1 sends only one copy of this traffic to AR-R1 on the AR-tunnel.

AR-R1, upon receiving the traffic on the AR-tunnel, replicates this packet to the other AR-LEAFs, other ARs and RNVEs, (LEAF-2, LEAF-3, LEAF-4, BL-1, BL-2), (AR-2), and (LEAF-5). The replicated traffic is sent to the AR-LEAFs and the RNVEs on the IR tunnel.

The receiving PEs get this replicated traffic in their existing IR tunnel. Hence, the behavior is same as with existing IR tunnels. Therefore, they receive the multicast traffic on the IR tunnel and forward it to access interfaces.

Assisted Replication Traffic Forwarding with Traffic Arriving From Behind RNVE

When multicast traffic arrives on the access interface of RNVE (LEAF-5), the RNVE itself replicates and sends one copy each of the traffic to all the other LEAFs, ARs, and BLs (classic Ingress Replication) on the IR tunnel. *Remember, there is no AR-configuration or AR-tunnel on RNVE.*

The AR replicators, upon receiving this traffic from RNVE, should not replicate further to ensure duplicates do not occur. How is the AR replicator able to distinguish which traffic to replicate (from behind the AR-LEAF) and which traffic not to replicate (from behind RNVE)? AR-1 figures this out by checking on the tunnel type on which the traffic arrived. Therefore, if traffic arrived on the AR-tunnel, it replicates to other PEs, and if it arrives on the IR-tunnel, it does not replicate.

Cardinal AR/IR Tunnel Forwarding Rules Summary

AR-Replicator

- On AR-R, if traffic arrived on AR-tunnel, replicate:
 - send replicated traffic to LEAF, RNVE, and AR on IR-tunnel
- On AR-R, if traffic arrived on IR-tunnel, do not replicate.

AR-LEAF

- On LEAF-1, for access traffic, send one copy to AR-R on AR-tunnel
- On LEAF-1, traffic arriving on IR, forward on access (existing behavior).
- RNVE: Existing IR Forwarding and Receiving Rules apply

Assisted Replication In a Multihoming Environment

EVPN Multihoming Local Bias Refresher

Please revisit Chapter 4's coverage of EVPN multicast forwarding in multihomed topologies. Here's a quick version: let's say that in Figure 5.5, LEAF-1 and LEAF-2 are multihomed on ESI-1, and that multicast traffic is arriving on ESI-1 on LEAF-1. When LEAF-1 Ingress replicates the packet to its multihomed peer LEAF-2, LEAF-2 should not forward on ESI-1 as duplicates/loops would occur. How does LEAF-2 avoid forwarding on ESI-1?

LEAF-2 looks at the source-IP of the packet and figures out that it has originated from LEAF-1. Based on this, LEAF-2 does not forward on the ESIs that are multihomed with LEAF-1. In this case, ESI-1. LEAF-2 would forward on other single-homed interfaces and on other ESIs where it is not multihomed with LEAF-1.

Assisted Replication In a Multihomed Environment – AR-R Should Retain the Src-IP of the Replicated Packet

Given the above, when we introduce AR in the fabric, we need to be careful with the packet that the AR replicates and sends. That is to say, it is paramount that the source-IP of the replicated packet is retained as LEAF-1 for local bias to work correctly.

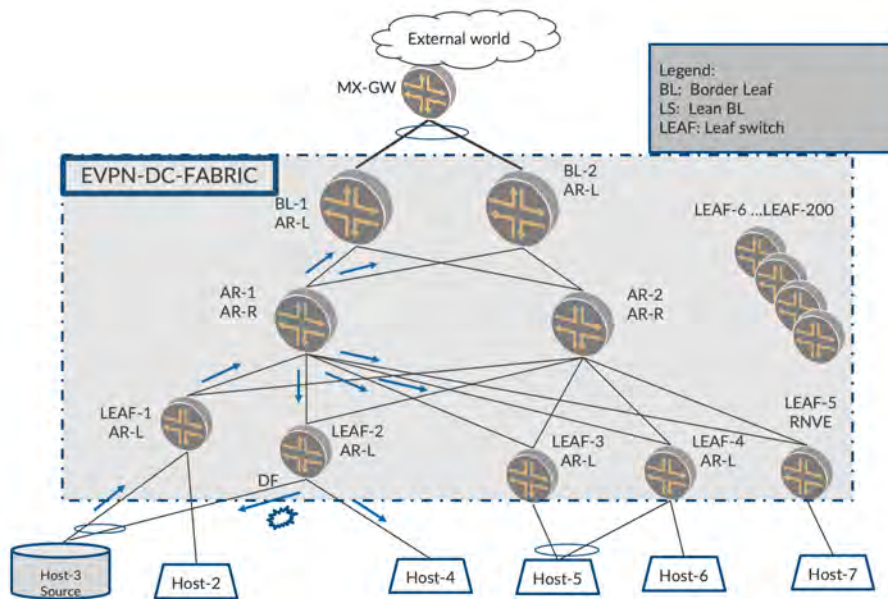


Figure 5.5 Assisted Replication In Multihomed Environments

If the AR-R sets its own IP (AR-R-IP) as the Src-IP of the replicated packet, LEAF-2 would end up sending on ESI-1 (since LEAF-2 thinks it is a packet from AR-R (core) and not from LEAF-1).

If this occurs, the LEAF-2 is in a spot because it cannot really tell if this replicated packet originated from LEAF-1 (multihomed) or from LEAF-3 (not multihomed). LEAF-2 will treat it as regular BUM packet coming in from the core and forward it to the ESI (since it is DF on the ESI) causing duplicates/loops on the MH-ESI.

It may not always be possible for the AR-Replicator to retain the Src-IP of the incoming packet onto the replicated packet (on some platforms).

Assisted replication functionality on AR-R is about receiving a packet on one VTEP (AR-tunnel) and then replicating it on other VTEPs in (IR-tunnel). When the AR device replicates and sends out the packet on VTEP following classic forwarding procedures, on some platforms the AR device can place only its own IP, (AR-R-IP) as the Src-IP.

Typically, when a PE Ingress replicates traffic arriving on access, it builds multiple copies and sends to the remote PEs with the Src-IP of the outer VXLAN header as its own IP. Prior to AR, there was never a need for transforming a packet arriving on core and sending back to the core with the retained Src-IP.

Overall, Ingress replication was always about taking an incoming packet from access interface and replicating it towards the core stamping its own Src-IP in the outer header. It may have been hard to add additional functionality to be able to take an incoming packet from the core interface and replicate it towards the core interface itself while also stamping the Src-IP as the incoming packet's Src-IP.

NOTE Since special handling is involved and this is unique forwarding behavior, some platforms do not have the capacity to retain the Src-IP of LEAF-1.

Enhanced Assisted Replication Procedure to the Rescue

AR provides a huge benefit in that the limitation of some platforms described in the above section as not able to retain the Src-IP should not preclude adoption of AR. If it was somehow possible for the LEAF/Replicator devices to cooperate and handle this situation, it would be worthwhile. Fortunately, this can be solved, and it is referred to as *Enhanced Assisted Replication*.

Capability Exchange

In its Type-3 AR route the AR-Replicator advertises its ability to retain the Src-IP by way of an extended community value. A value of *Base-AR* mode in the community would mean, 'I can retain the Src-IP of LEAF-1' and a value of *Enhanced-AR* mode in the community would mean, 'I am not capable of retaining the Src-IP of LEAF-1.'

AR-R Capable of Retaining the Source-IP of LEAF-1

If AR-R is capable of retaining Src-IP, then all is well. The AR-R appropriately sends the community value suggesting that it is in Base AR mode. Based on the exchange, the AR-LEAF devices and AR-Replicator devices follow the procedures described earlier in the plain vanilla AR procedures. We are good in multihomed scenarios, too.

Since the Source-IP of LEAF-1 is retained in the replicated packet, LEAF-2, upon receiving packet from the ARR, figures out the Src-IP of the packet to be that of LEAF-1. With this, it performs local-bias and skips forwarding on the ESI.

AR-R Not Capable of Retaining the Source-IP of LEAF-1

If AR-R is not capable of retaining Src-IP, then the AR-R sets the community value suggesting that it is in Enhanced-AR-mode. The AR-LEAF devices on hearing this community get into Enhanced-AR-mode. In this case of enhanced-AR mode, extra procedures are required of both AR-R and AR-L as shown in Figure 5.6.

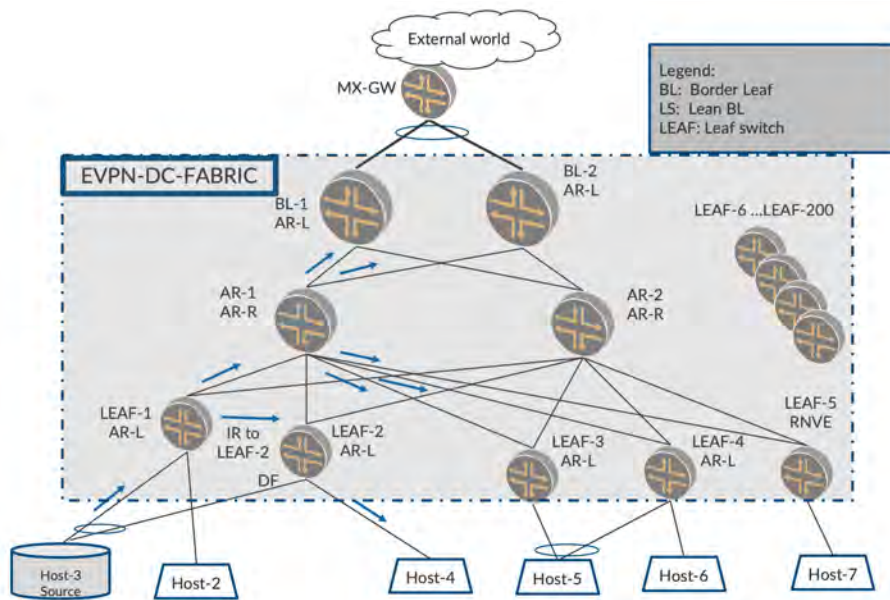


Figure 5.6 Enhanced AR Mode

In Figure 5.6, AR-R deduces which AR-LEAF PEs are multihomed on a VLAN. This information is available and can be deduced from the Type 1 route per EVI that is advertised for ESIs. MAC-aliasing uses this information already.

Once AR-R deduces LEAF-1 and LEAF-2 are multihomed, when it receives traffic from LEAF-1 and replicates to other LEAFs, AR-R *skips* send the replicated packet to LEAF-2 alone, therefore, the multihomed peers that LEAF-1 is multihomed to.

Since AR-R has skipped sending to LEAF-2, there should be some other means of sending that traffic to LEAF-2. So LEAF-1, the AR-LEAF, sends one copy of the access traffic to the replicator and one another copy to LEAF-2 alone over IR-tunnel. (to its MH peers on that VLAN). This needs to be sent to LEAF-2 because LEAF-2 may have other single-homed interfaces where this traffic needs to be sent.

When LEAF-2 receives traffic from LEAF-1 it will look at the Src-IP of the packet and will do what's needed (DST-LOCAL-BIAS). Since LEAF-2 has already directly received from LEAF-1, it can perform the local-bias and not send on the ESI link but can forward on the single-homed interfaces.

This way we have arrived at the middle ground of deriving the benefits of AR as it is also handled in multihoming scenarios.

Overall Assisted Replication Procedures

AR-Replicator:

- *On AR-R, if traffic arrived on AR-tunnel, replicate*
 - Send replicated traffic to LEAF, RNVE, and AR on IR-tunnel
- On AR-R, if traffic arrived on IR-tunnel, do not replicate

AR-LEAF:

- *On LEAF-1, for access traffic, send one copy to AR-1 on AR-tunnel*
 - LEAF-1, traffic arriving on IR, forward on access (existing behavior)
- RNVE: Existing IR Forwarding and Receiving Rules apply

Enhanced-AR- Additional Forwarding Rules, in Addition to Base-AR mode:

- AR-R skips sending the replicated packet to LEAF-2.
 - LEAF-1 makes an extra copy to LEAF-2 alone, and sends over IR-tunnel.

Chapter Summary

This chapter explored assisted replication, its benefits, basic procedures, platform limitations with EVPN-MH scenarios, and procedures to detect and accommodate it.

With AR we transferred the load of replication from low-end LEAF devices to capable high-end replicator devices effectively reducing replication load on LEAF and avoiding the overuse of the link bandwidth between the LEAFs and the Lean Spine. The existing underlay Lean Spine performs the additional role of Replicator.

In multihoming environments AR has two modes, base and enhanced. We have capability exchange occurring with an extended community in Type-3 AR route.

In Base-AR mode, where the Src-IP of the LEAF is retained, classical local bias, multihoming procedures work automatically to ensure duplicates do not occur as the Src-IP of the originating LEAF is retained by the AR-R.

In Enhanced-AR mode where the Src-IP of the LEAF is not retained, enhanced AR procedures are followed to ensure duplicates do not occur. This is achieved by the AR-R sending the replicated packet to the multihomed peer and the Ingress PE sending one copy to AR-R and one another copy to the multihomed peer alone.

In ensuing chapters, we'll explore the first steps towards optimization of multicast. We will begin by observing how traffic towards the access side is optimized followed by how traffic towards the core is optimized.

In Chapters 3 and 4 we explored flooding of multicast traffic everywhere. Therefore, the traffic is sent to all Egress PEs and access interfaces, and the Egress PEs in turn send to their access interfaces.

As was mentioned earlier, understanding the concepts of AR or its procedures is not a prerequisite to proceed with getting to different multicast optimizations in the next chapters. As we introduce each facet of optimization, we describe how it would fit in if AR is also deployed along with optimization.

Chapter 6

EVPN Intra-VLAN Multicast with Optimization

Chapters 4 and 5 explored EVPN Multicast procedures where L2-switched traffic was *'flooded everywhere'*. This flooding to remote PEs that don't have any listeners behind them, and the PEs flooding traffic onto all the access interfaces, even if there are no listeners, is clearly undesirable.

One downside to this unrestrained flooding is that all the LEAF devices are burdened with processing traffic. That is not useful. Some LEAF devices may be low-end devices or may even be virtual LEAF devices. If these devices are subjected to heavy multicast traffic, they will choke other traffic applications and degrade performance.

Also, redundant traffic that gets sent on access interfaces may affect the link utilization, degrading other genuine flows for which there are listeners. The flooding of traffic onto EVPN core with Ingress Replication to PEs that do not have listeners behind them results in core bandwidth consumption and an excess load on the Ingress LEAF towards creating multiple redundant copies.

In this chapter we describe mechanisms that ensure traffic is forwarded only on to those access interfaces where there is listener interest. We also describe procedures where the traffic is Ingress Replicated over the EVPN core to only those remote EVPN PEs that have expressed listener interest. Let's explore the procedures that will help mitigate the unbridled flooding of multicast traffic everywhere. Towards this end, we introduce IGMP-snooping, Selective Multicast Forwarding, BGP Type-6 SMET Route, and more.

These procedures help optimize multicast forwarding in the DC Fabric. Care is taken to ensure that multicast traffic is forwarded only towards those PEs or access interfaces that have listener interest for a particular flow. As a result, there is a substantial gain in terms of link and core bandwidth utilization, reduction in Ingress Replication Load, and reduction in Egress processing load.

When these optimization techniques are used along with AR (described in an earlier chapter), we get the benefits of optimization and also the transfer of replication load to the AR-Replicator.

Due to legacy devices, in some cases it may not be possible to have this optimization capability on all LEAFs. In this chapter we describe the procedures that help to accommodate such devices and ensure multicast forwarding occurs in the right manner.

EVPN with IGMP-Snooping

Multicast optimization on a L2-switch is described as such because EVPN as a technology enables stretching the L2-domain over remote devices. An EVPN LEAF device plays the role of an L2-switch on its access interfaces. With IGMP snooping enabled on an EVPN LEAF device (see Figure 6.1) the multicast traffic is sent only on those interfaces that have listeners behind them.

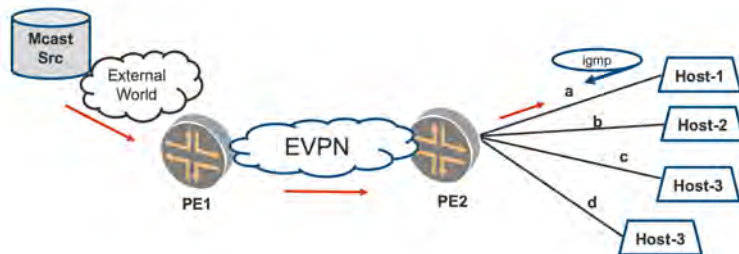


Figure 6.1

EVPN with IGMP-Snooping

Let's say that in Figure 6.1, PE2 is configured with EVPN and IGMP-snooping. When multicast traffic sent from the Mcast Source reaches PE2 over EVPN, PE2 should forward multicast traffic only to interface 'a', since it has received an IGMP Report (listener interest) only from interface 'a'.

Traffic is not flooded onto other access interfaces b, c, and d.

PE2 is an EVPN device that is enabled with IGMP-snooping on the bridge-domain. PE2 snoops membership information on the access interface and tracks the association. When multicast traffic arrives from core or from access interface, the traffic is flooded only on those interfaces where the IGMP state is learned.

Primer on IGMP-Snooping

In this section, let's explore the L2 multicast optimization techniques that have been widely deployed in the switching world for more than a decade. This multicast optimization on the access interfaces is achieved with IGMP Snooping. This has been used extensively in the L2-switching world where there is a need to forward traffic onto only those interfaces with listeners behind them.

In a typical L2 switching environment, multicast traffic is always forwarded on all interfaces. In Figure 6.2's topology, the switch has five access interfaces: a, b, c, d, and x. Say there is no L2-switched multicast optimization configured on the switch.

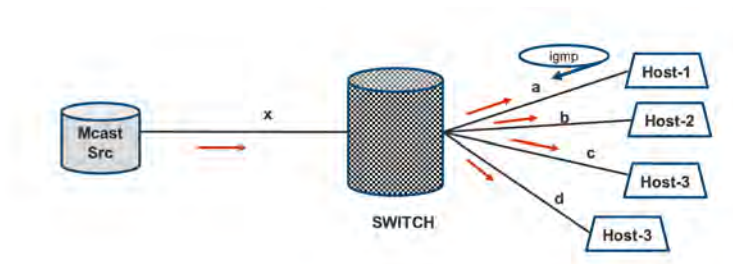


Figure 6.2 *Need for IGMP-Snooping*

Let's assume Host1 alone is interested in traffic for group G and sends an IGMP report for for (*,G). IGMP is a protocol used by hosts and adjacent routers/switches to establish multicast group memberships. With an IGMP report for group G, the host expresses listener interest for group G to the nearest switch/router on the LAN.

On a switch without multicast optimization, when the source starts sending traffic on x for group G, the switch will flood the traffic onto all four interfaces – a, b, c, and d – even though Hosts 2, 3, and 4 are not interested in that group.

In a different case, when there are no listeners for the group and the source sends traffic, the switch will still flood the traffic onto all four interfaces: a, b, c, and d. This is clearly undesirable.

Now let's say the switch is configured with IGMP snooping. With IGMP snooping, when an IGMP Report/Join is received (snooped) on interface a, the switch will snoop the (*,G) report and create an associated IGMP state for G with interface a.

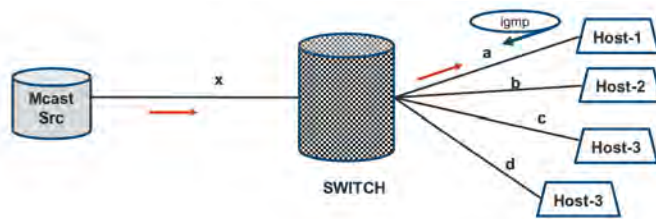


Figure 6.3 Optimized L2-multicast Forwarding with IGMP-Snooping

When multicast traffic arrives on interface ‘x’ for group G, the switch will forward the traffic only to interface ‘a’ based on the (*,G) state that it maintained for interface ‘a’. (*ACCESS-SNOOP*). Thus, the unnecessary flooding of traffic onto interfaces b, c, and d is avoided.

Also, if there are no IGMP reports from any of the interfaces (a,b,c,d) and if traffic arrives on interface ‘x’, with IGMP snooping enabled, the switch will not forward traffic to any of the interfaces (a,b,c,d).

Typically the switch will have several ports. With IGMP-snooping enabled on the switch, the bandwidth usage on the interfaces is considerably reduced. Also, the hosts that are not interested in the group are not burdened with unnecessary traffic. Needless to say, the replication load of the switch to create redundant copies is also reduced by virtue of IGMP-snooping.

The switch periodically sends out ‘IGMP General Queries’ to solicit listener interest by way of IGMP reports from hosts. When a host is no longer interested in the particular flow G, the host sends out an IGMP Leave message. Upon receiving the IGMP Leave message, the switch sends out a ‘Group Specific Query’ for the particular group G, to solicit if there are still other hosts interested in group G.

Towards this end, the switch starts a timer called ‘Last Membership Query’. Once the timer expires, the switch is now sure of the absence of listener interest on the interface. Based on this, the switch removes the IGMP state for the group for the interface.

Selective Multicast Forwarding Towards the Core

So far we have discussed multicast traffic optimization on the access side. It is imperative that the traffic that is flooded onto the core using Ingress Replication is optimized, too, such that only those PEs that have listeners behind them actually receive traffic. This is referred to as *EVPN Selective Multicast Forwarding*.

Consider Figure 6.4 where all the LEAF devices have VLAN VLAN-101/VNI-101 enabled. LEAF-3 does not have a listener for group G, behind it. But what if there were hundreds of LEAFs like LEAF-3 that do not have listeners behind them for the group G?

If the Ingress PE, LEAF-1 in Figure 6.4, excludes these LEAFs in its Ingress Replication Flood List, then we would have achieved Selective Multicast Forwarding for group G. This results in conserving core bandwidth and also in exempting the LEAFs that don't have listeners behind them from traffic processing overhead.

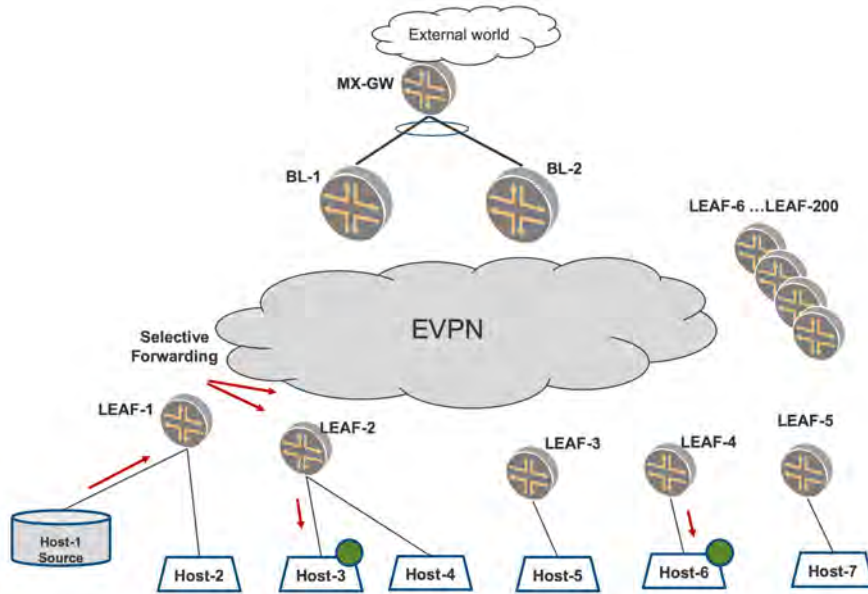


Figure 6.4

Selective Multicast Forwarding

With selective forwarding, in Figure 6.4 LEAF-1 will L2-switch the traffic from the source only to LEAF-2 and LEAF-4. Further, LEAF-2 will flood the multicast traffic only on those access interfaces where there are listeners (by virtue of IGMP-snooping on access (*ACCESS-SNOOP*)). Therefore, LEAF-2 sends the traffic only to Host-3 but not towards Host-4. LEAF-4 floods the traffic only to Host-6. No other access interfaces carry the traffic. Also, no other LEAF devices receive the traffic over the core.

Three kinds of benefits accrue:

- LEAF-1's Ingress Replication load is considerably reduced, for example, replicating traffic to only two LEAFs as against sending to 200 LEAFs.
- Core bandwidth utilization is considerably reduced. Each superfluous replicated packet would have had to go to non-interested LEAF devices only to be dropped.
- LEAF-3 and other such LEAFs that do not have listeners are spared from processing this unnecessary traffic, thereby freeing them up for other tasks.

When a host behind a LEAF joins a group G, the Ingress sends the traffic for group G to the LEAF. Similarly, when the host sends out IGMP Leave for group G, the Ingress stops sending the traffic for group G to the LEAF. In other words, the Ingress forwards traffic for a group G to a LEAF only as long as there is a remote listener interest for the group behind that LEAF.

Basic Procedures for Selective Multicast Forwarding

Now let's describe the mechanism with which Selective Multicast Forwarding is accomplished. As described in the IGMP-snooping Primer section at the beginning of this chapter, the EVPN LEAF devices snoop the reports and keep the IGMP state associated with the interface.

BGP Type-6 SMET (Selective Multicast Ethernet Tag) Route

The EVPN LEAF device, when it receives a report for a group G1 from any of the access interfaces on a VLAN, will originate a BGP EVPN Type-6 NLRI route for G1. This Type-6 route contains Route-Distinguisher, Ethernet-Tag, Multicast Group address, and originating IP address as shown in Figure 6.5.

6	1.1.1.1:22	601	235.1.1.1	0.0.0.0	10.1.1.1
Route Type	Route Distinguisher (RD)	E-TAG	Group Address	Source Address	Originating IP address

Figure 6.5

Type-6 Route Fields

This route also carries the RT-target community that identifies the EVPN instance such that only those EVPN PEs that host the relevant EVPN instance will import this route.

The E-Tag field carries the VLAN for which listener interest for the Group address is relevant. For a group G in a VLAN V-101, there may be few or several LEAF devices that have listener interest. Each of these LEAF devices will originate this Type-6 route. Since their Route Distinguishers and Originating IP address will be different, the Ingress will be able to clearly identify the specific interested LEAF devices and build the Selective Floodlist accordingly.

The term outgoing interface list (OIL) is traditionally used to refer to the list interfaces on which multicast traffic for a specific group G1 has to be sent out. In our discussions, the OIL for a group G1 on an Ingress will be the VTEP interfaces that point to the specific interested LEAF devices for that group.

This BGP Type-6 route can be considered as the EVPN equivalent of an IGMP report received on the access, therefore this Type-6 route is the equivalent of IGMP (*,G1) report in the EVPN core.

In Figure 6.6, LEAF-2, when it receives the IGMP (*,G1) report from Host-3 on VLAN-101 from access, will originate an EVPN Type-6 route for VLAN VLAN-101 and group G1. Similarly, LEAF-4 will also originate a Type-6 for group G1 on VLAN-101.

NOTE It is worth noting that LEAF-2 will not forward the IGMP-report into the core as it has already signaled the listener interest with Type-6. This Type-6 route also helps in avoiding unnecessary IGMP reports being refreshed in the core.

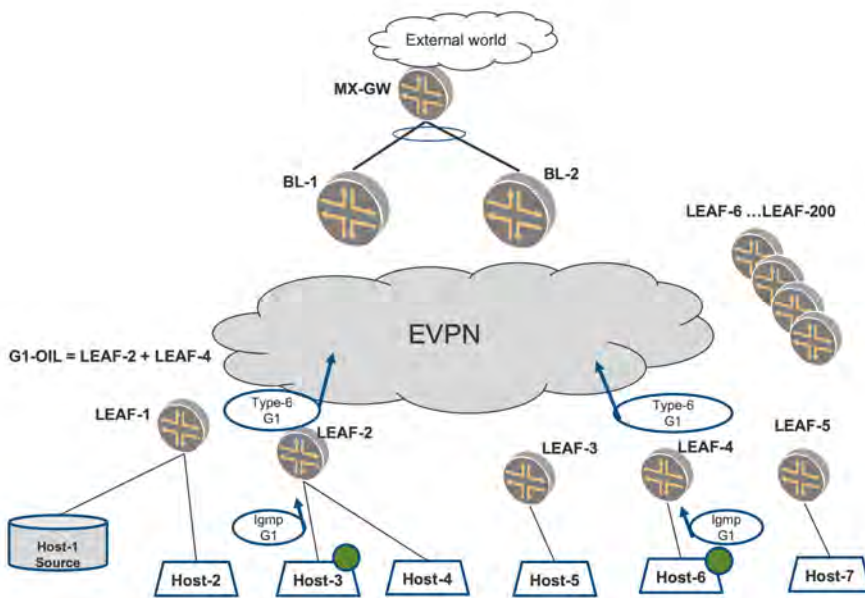


Figure 6.6

BGP EVPN Type-6 Route

LEAF-1 will import the Type-6 routes for group G1 from LEAF-2 and LEAF-4. Further, LEAF-1 will install a multicast snooping forwarding entry specifically for G1 that includes only LEAF-2 and LEAF-4 in its OIL. When traffic arrives for group G1 from the source, the traffic will hit this specific multicast forwarding entry for G1 and will be Ingress Replicated selectively to only LEAF-2 and LEAF-4 alone.

An Example with Two Different Groups G1 and G2

Figure 6.7 depicts hosts that are interested in two different groups, say G1 and G2. The hosts Host-3 and Host-5 are interested in group G1, while Host-2 and Host-6 are interested in group G2. The respective Type-6 routes for group G1 are originated by LEAF-2 and LEAF-3, while for group G2, by LEAF-1 and LEAF-4.

From Ingress PE LEAF-1's point-of-view, the OIL for G2 will include LEAF-4 and also access interface 'x' towards Host-2. The Ingress PE will build an OIL that has a VTEP interface towards LEAF-4 and also an access interface 'x' towards Host-2. Traffic for group G2 will not be sent to any of the 200 LEAF devices except LEAF-4 and onto 'x' towards Host-2. This is a significant optimization *vis-à-vis* the Chapter 4 sections.

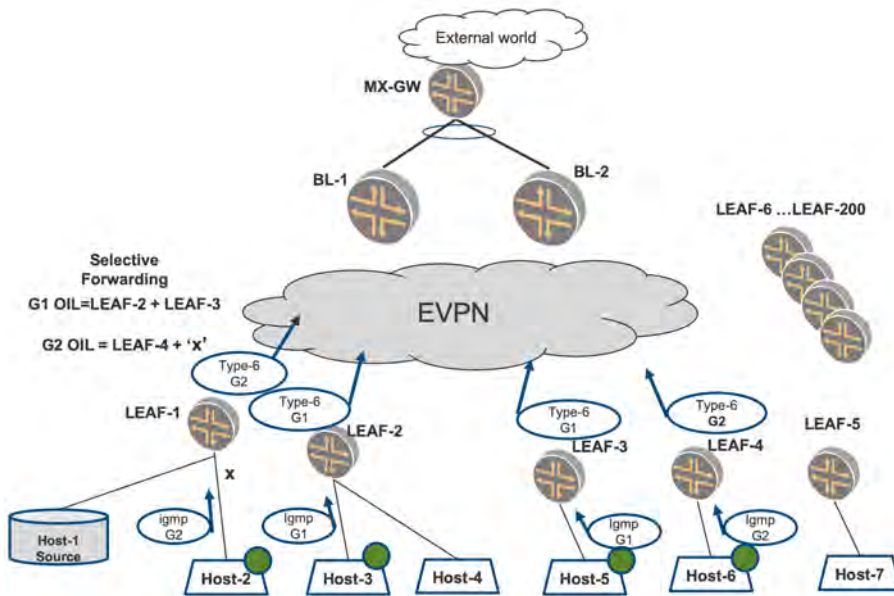


Figure 6.7

Two Different Groups G1 and G2

Suppose there is a group G3 (not shown in Figure 6.7) for which there are no listeners anywhere in the fabric. The Ingress, LEAF-1, does not receive any Type-6 routes for G3 and hence does not install a specific forwarding entry for G3. When a source sends traffic for G3 towards LEAF-1, then, LEAF-1 will not forward the traffic for G3 towards the core at all.

Often there is high volume multicast traffic that is sent for several groups but there are no active listeners for such groups. With the mechanism of Selective Multicast Forwarding, the high-volume traffic is not sent towards the core to any of the 200 LEAF devices for such groups, thus saving considerable core bandwidth.

BGP Type-6 NLRI SMET Route and Selective Multicast Forwarding

The EVPN Type-3 route is also referred to as an *inclusive multicast Ethernet tag* (IMET) route because based on this route traffic is forwarded in an inclusive manner, for example, ‘flooded’ to the LEAF devices irrespective of listener interest.

The EVPN Type-6 route is referred to as a *Selective Multicast Ethernet Tag* (SMET) route.

The subtle usage of terms here does make a difference:

- **SMET Route:** This refers to the BGP Type-6 NLRI route that the EVPN LEAF devices originate when there is a listener interest for a particular group G in a particular VLAN.
- **Selective Multicast (SMET) Forwarding** (sometimes called SMET Forwarding): This is the capability of Ingress EVPN PE that builds a specific multicast snooping forwarding entry for a group G1 and the OIL with specific interested the LEAF devices’ VTEPs, based on the remote EVPN Type-6 routes that it has heard from the LEAF devices. (*CORE-SMET-FWD*)
- **IGMP-snooping and Selective (SMET) Forwarding** are techniques for optimizing multicast traffic. For optimized multicast forwarding in the core to work well, it is imperative that both the above ‘a’ and ‘b’ are supported on the appropriate devices. For example, a EVPN device that is connected to listeners should support at least SMET Route capability. An EVPN device that is Ingress should support both SMET Route capability and Selective Multicast (SMET) Forwarding paradigm.

Since the Type-6 routes get sent from the LEAF to Spine, periodic refreshes of the IGMP reports over the core are avoided. On the access interface, when the listeners send IGMP Leave for a group G, the LEAF will send out a Group Specific Query to ensure no other listeners exist for that group. Once the LEAF has affirmed that listener interest has gone away, it withdraws the advertised Type-6 route. This results in the Ingress updating its OIL with the remote LEAF removed from the list.

An IGMP-snooping state for a group on access side, tracks each L2-interface on which a listener report for G is received. It is worthwhile to mention here that Type-6 for a group G is representative of the listener interest on that VLAN on the LEAF.

As long as there is at least one access L2 interface where there is listener interest, a Type-6 route for group G is originated. Also, when there is no access L2 interface in the VLAN on which there is listener interest for group G, the Type-6 for group G on that VLAN is withdrawn.

So far, we have described only single-homed scenarios where optimized forwarding is illustrated. With multihoming there are challenges that are explored in Chapter 7.

EVPN Multicast Flags Community: SMET Feature Capability Exchange

Many times, in practical scenarios, there is a mix of devices that support selective (SMET) forwarding and those that do not support it. Also, some LEAF devices may not be enabled with IGMP-snooping. Thus there is a need to exchange the capability so that the participating EVPN devices can interoperate well.

Towards this end, EVPN Multicast Flags extended community (MF-COM) was introduced to help the exchanging capabilities of the EVPN devices in the fabric so that the Ingress PE can accommodate those devices that don't support the functionality and that don't interoperate well. You can see the header for EVPN MF-COM in Figure 6.8.

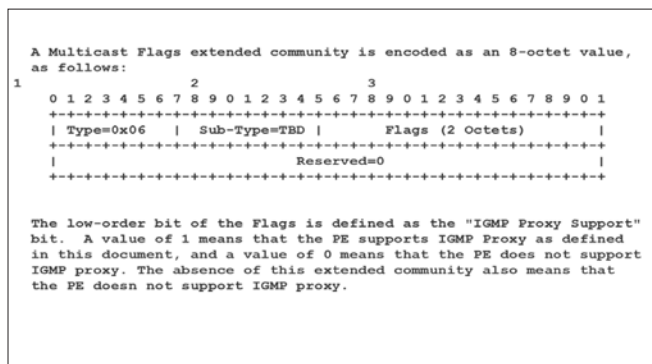


Figure 6.8 EVPN MF-COM Header

An EVPN device that is enabled with IGMP-snooping and supports BGP SMET Route Type-6 will add a BGP Extended community called EVPN Multicast Flags (EVPN-MF) to its EVPN Type-3 route.

In addition, this device will set the least significant bit (LSB) of the octet in the community to 1 to suggest support for SMET capability. A device that supports IGMP-snooping and supports BGP SMET Route Type-6 capability must add the EVPN-MF community and set the LSB in the octet to 1.

Also, if a device is not IGMP-snooping enabled or does not support BGP SMET Route Type-6 or if it wants traffic flooded to it irrespective of listeners, it will either skip adding the EVPN-MF community or add the EVPN-MF community but reset the LSB of the octet in the community to 0 to suggest traffic be flooded to this PE.

Once this community is exchanged in EVPN Type-3 route, the participating PEs have information on which LEAF devices have to be included for all traffic, and to which LEAFs have to be selectively forwarded.

Additional Procedures for Selective Multicast Forwarding

Forward L2-switched Traffic to L3-PIM Spines

When a LEAF receives traffic from its access interface for group G1, the LEAF should forward traffic selectively only to those EVPN PE devices that have originated a Type-6 for that group G1.

In addition to forwarding traffic to the LEAF devices that have originated Type-6 for group G1, the Ingress should forward the traffic to the L3-PIM devices, too. If there are no remote listeners for the group, the traffic must still be sent to the L3-PIM devices alone.

The rationale for this is as follows. A L3-PIM device performs Inter-subnet routing from the source-VLAN onto other receiver VLANs (Inter-VLAN Multicast will be illustrated in detail in Chapter 7). So, the L3-PIM device should attract and receive the traffic on the source-VLAN all the time. Also, the L3-PIM device that is PIM DR should register the multicast source to the PIM-RP. So the traffic should reach the L3-PIM device all the time (see Figure 6.9).

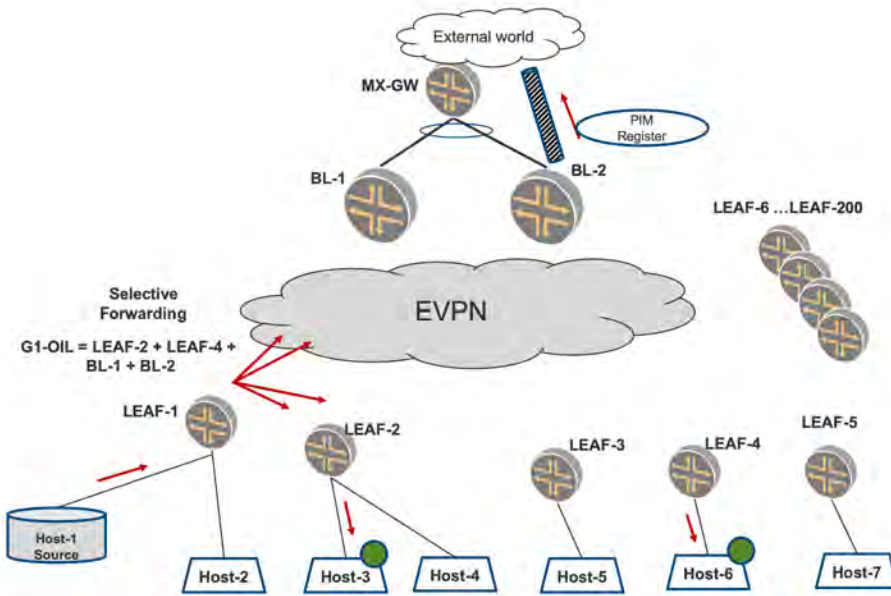


Figure 6.9

Include L3-PIM Devices

How Is a L3-PIM Device Detected?

In the last section, the Ingress forwards the traffic for group G1 to L3-PIM devices. This L3-PIM device is detected by virtue of the L3-PIM device resetting the LSB in EVPN-MF community in a BGP Type-3 route.

It is a bit of a paradox that the L3-PIM device is enabled with IGMP-snooping, capable of originating SMET Type-6 routes as well as capable of Selective (SMET) Forwarding. However, the L3-PIM device has to attract traffic for the purposes of Inter-VLAN routing and registering the multicast source with the RP.

Towards this end, the L3-PIM either skips adding the EVPN-MF community or adds the EVPN-MF community but resets the LSB in the octet to 0. Any LEAF Ingress PE that receives such a Type-3 will flood all the traffic to this L3-PIM device.

Accommodating Devices That Do Not Support Snooping/SMET Route

In certain deployments, there may be few LEAF devices that are not enabled with IGMP-snooping. Also, there may be some EVPN LEAF devices that are enabled with IGMP-snooping, but the devices do not have the support for originating BGP Type-6 route.

Such legacy LEAF devices do not convey remote listener interest multicast interest to other PEs. Let's suppose that LEAF-5 in Figure 6.10 is a legacy device that does not support IGMP-snooping.

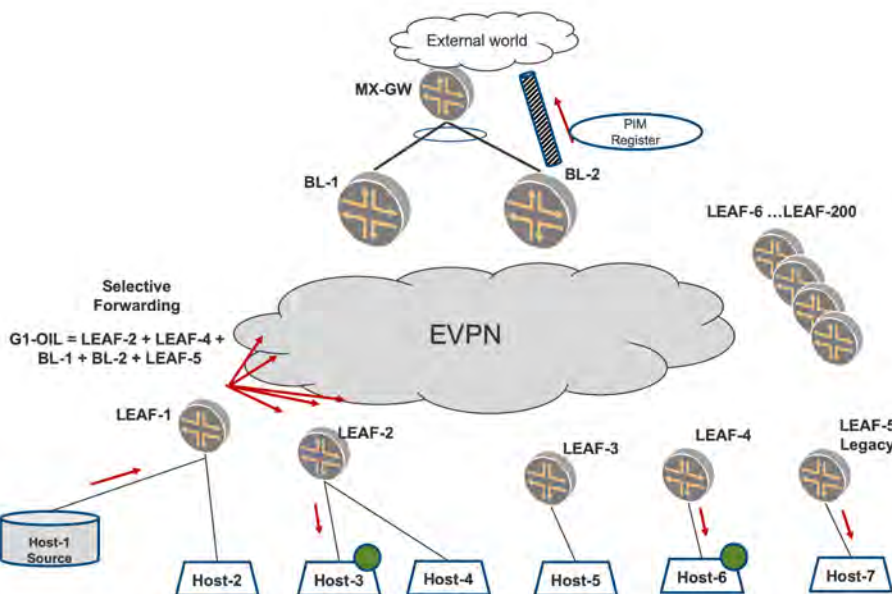


Figure 6.10

Include Legacy Devices

However, such legacy LEAF devices may have listeners interested in traffic for the group G1. Since they do not originate a Type-6 route, the Ingress does not have a way to determine if there are listeners behind such legacy PEs or not. If the Ingress does not forward the traffic to the legacy PEs, the listeners behind the legacy PEs will not get the traffic. So the Ingress has to flood the traffic to the legacy PEs and in this way ensure compatibility with devices that do not support optimized multicast procedures.

How is a Legacy Device Detected?

In an earlier section, we saw that the Ingress forwards the traffic for group G1 to legacy LEAF devices. This legacy LEAF is detected by virtue of an absence of an EVPN-MF community in a BGP Type-3 route.

The Ingress device, when it learns of the LEAF devices from Type-3, will check for this community. If Ingress detects that the EVPN-MF is not present, the Ingress will deduce that it is a legacy device. Based on this, the Ingress will always flood the traffic to this legacy device.

In the Figure 6.10, for G1, LEAF-1 will have in its OIL, LEAF-2, LEAF-4, the BLs, and LEAF-5. Other LEAFs that don't have listeners behind them will not be part of the OIL.

Forwarding Prefix for L3-PIM and Legacy Devices to Flood Traffic

To achieve flooding of traffic in the special cases described earlier, where the egress PEs want to attract traffic for *__all__* groups, a default multicast prefix route of 224/4 is installed with its OIL containing L3-PIM and *legacy* devices. This will cause traffic for flows that don't have a specific route entry installed (for example, 235.1.1.1) to hit this prefix 224/4 and reach the L3-PIM devices. If there were listeners existing for group 235.1.1.1, the OIL for 235.1.1.1 will include the L3-PIM and *legacy* devices.

Flood 224/24 Groups (for example, OSPF/LDP Hellos) Everywhere

Consider special well-known multicast groups like 224.0.0.5, 224.0.0.13. These are well-known multicast groups for different purposes, for example, 224.0.0.5 is a multicast group used for OSPF Hellos, 224.0.0.13 for PIM Hellos. Such well known groups are specified to be in the range of 224.0.0.0/24 (224.0.0.1 to 224.0.0.255).

The traffic that is destined for such groups should be flooded everywhere because LEAF devices do not originate Type-6 for such groups. For example, there may be L3-OSPF devices behind LEAF-1 and LEAF-2. The OSPF devices have to discover each other using hellos. These hello messages have to be forwarded to each other by the LEAFs. So, the Ingress PE installs a 224.0.0.0/24 specific forwarding entry that has in its next hop *__all__* the PEs in the fabric. This happens without any need for configuration.

Flood User-configured Groups Everywhere

There may be scenarios where the operator may want some multicast groups to be flooded everywhere irrespective of the presence of listeners. These groups may not fall in the 224/24 range. Suppose there is a group 230.1.1.1 for which the operator desires the traffic to be flooded everywhere. This can be accomplished by having a configuration on the Ingress. The configuration will contain a list of groups for which the traffic should be flooded. When traffic does arrive for 230.1.1.1, irrespective of listeners on access or over the core, the traffic is flooded everywhere.

Rules for Selective Multicast Forwarding

So the rules of Selective Multicast Forwarding for traffic destined for group G1 are enhanced in the following manner:

- L2-forwards the multicast traffic for group G1 towards those EVPN PEs that have originated a BGP Type-6 SMET route for that group G1.
- L2-forwards the multicast traffic for group G1 towards L3-spine devices for the purpose of Inter-subnet multicast and registering the multicast source with PIM-RP.
- L2-forwards the multicast traffic for group G1 towards the legacy LEAF devices that are not enabled with IGMP snooping or those that do not support BGP Type-6 SMET Route origination.
- L2-forwards the multicast traffic for group in range 224.0.0/24, to __all__ EVPN PEs in the fabric.
- L2-forwards the multicast traffic for user-configured groups to __all__ EVPN PEs in the fabric.

Chapter Summary

This chapter explored optimized multicast with IGMP-snooping and Selective (SMET) Forwarding. In large scale multicast deployments with a high volume of traffic, these optimizations play a key role in conserving bandwidth in both core and access, and also in the replication and packet processing load on the participating EVPN devices.

It also explored different signaling procedures with SMET Type-6 routes and how interoperability is achieved with legacy devices. We explored the different outliers that an EVPN device takes into consideration when building the list of PEs, therefore, L3-PIM devices, legacy devices, etc., and also the special forwarding rules needed for link local multicast addresses 224/4.

SMET Forwarding is enabled when IGMP-snooping is enabled. It may be the case that some groups need to be flooded everywhere. This can be achieved by a configuration where the operator can cherry-pick the group addresses where traffic needs to be flooded.

In Chapter 7 we will explore optimized multicast forwarding in multihomed scenarios..

Configuration

The reference topology is shown in Figure 6.11.

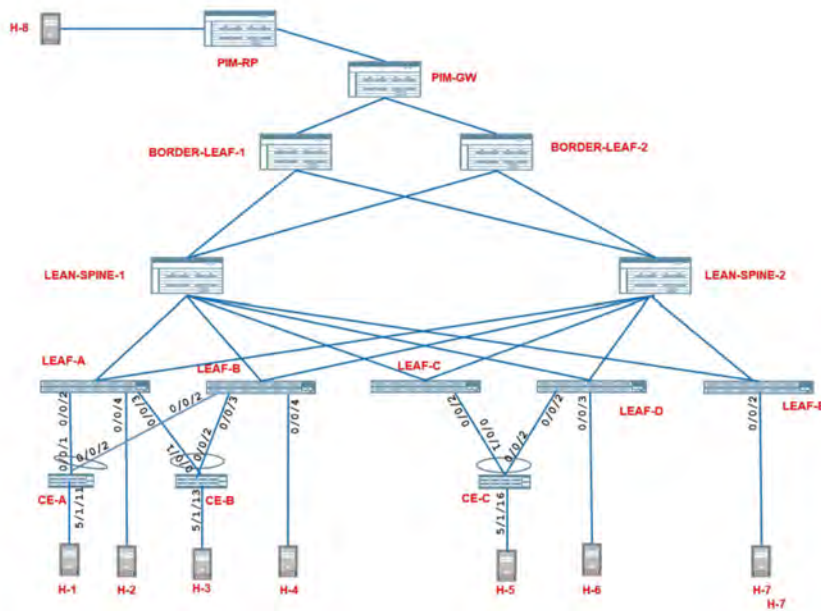


Figure 6.11

Chapter 6 Reference Topology

Having reviewed intra-VLAN multicast traffic optimization for EVPN, let's now see it in action by enabling IGMP-snooping and Selective (SMET) Forwarding and observing how traffic forwarding is optimized.

Configure IGMP-Snooping

Turn on IGMP-snooping on the EVPN PEs. Upon enabling IGMP-snooping on a BD/VLAN level, SMET Type-6 route origination Selective (SMET) Forwarding is enabled, too.

Configure VLAN-101 on LEAF-5. However, since we want LEAF-5 to simulate a legacy device that does not support IGMP snooping, we do not turn on IGMP-snooping on any of the VLANs on LEAF-5.

Before we begin, stop the traffic, and the receivers on RT, and load the following configuration snippets on the following devices:

LEAF-1, LEAF-2, LEAF-3, LEAF-4:

BL-1, BL-2:

```
set protocols igmp-snooping VLAN VLAN-101
set protocols igmp-snooping VLAN VLAN-102
```

Configure VLAN-101 on LEAF-5

Load this configuration snippets on LEAF-5:

```
set interfaces xe-0/0/2 unit 0 family ethernet-switching VLAN members VLAN-101
set vlans VLAN-101 VLAN-id 101
set vlans VLAN-101 vxlan vni 101
set protocols evpn extended-vni-list 101
```

Traffic Verification

From Host-1, start sending multicast traffic at 10 pps for group 225.1.1.1 in VLAN-101.

On Host-6 that is single-homed to LEAF-4, start a receiver for the multicast group, 225.1.1.1 on VLAN-101.

Traffic Statistics on Router Tester

From the RT statistics in Figure 6.12, you can see that the traffic sent by Host-1 at 10 pps is now being received only by the interested receiver, Host-6, and the legacy device, Host-7, in VLAN-101.

	Stat Name	Port Name	Link State	Frames Tx, Rate	Valid Frames Rx, Rate
1	10.216.45.202/Card20/Port01	HOST-1	Link Up	10	0
2	10.216.45.202/Card03/Port01	HOST-2	Link Up	0	0
3	10.216.45.202/Card20/Port02	HOST-3	Link Up	0	0
4	10.216.45.202/Card03/Port02	HOST-4	Link Up	0	0
5	10.216.45.202/Card20/Port03	HOST-5	Link Up	0	0
6	10.216.45.202/Card03/Port03	HOST-6	Link Up	0	10
7	10.216.45.202/Card03/Port04	HOST-7	Link Up	0	10
8	10.216.45.202/Card20/Port04	HOST-8	Link Up	0	0

Figure 6.12

RT Stats

Multicast Traffic Outputs - LEAF-1

As it did previously, the load balanced multicast traffic arrives on the access interface ae0 on LEAF-1. However, the multicast traffic arriving on LEAF-1 is no longer forwarded on any of the access interfaces on LEAF-1, since there are no receivers, thus avoiding waste of access side bandwidth:

```
lab@LEAF-1> monitor interface traffic detail
Interface  Link  Input packets      (pps)      Output packets      (pps) Description
...
xe-0/0/4    Up      0      (0)      2564      (0) T0 Host-2
...
ae0         Up    9467    (10)      0      (0) T0 CE-1
ae1         Up      0      (0)    2560      (0) T0 CE-2
...
```

The multicast traffic is forwarded on the VTEPs towards Border-LEAF PEs (101.101.101.101 and 102.102.102.102) and LEAF-5 (109.109.109.109).

The traffic is also sent on the VTEP towards LEAF-4 (108.108.108.108) that has an interested receiver.

LEAF-2 (106.106.106.106) and LEAF-3 (107.107.107.107) that do not have any interested receivers are spared of the traffic:

```
lab@LEAF-1> show interfaces vtep extensive | grep "VXLAN Endpoint Type: Remote|Output packets.*pps"
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 106.106.106.106...
Output packets: 4713 0 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 101.101.101.101...
Output packets: 9375 10 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 108.108.108.108...
Output packets: 5653 10 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 107.107.107.107...
Output packets: 4713 0 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 109.109.109.109...
Output packets: 4664 9 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 102.102.102.102...
Output packets: 9375 9 pps
```

Multicast Traffic Outputs - LEAF-4

The access side IGMP-snooping functionality ensures that the multicast traffic arriving on LEAF-4 is forwarded on the single-homed interface, xe-0/0/3.0, that now has a receiver, but it is not forwarded on the multihomed interface ae0.0 that does not have a receiver:

```
lab@LEAF-4> monitor interface traffic detail
Interface  Link  Input packets      (pps)      Output packets      (pps) Description
...
xe-0/0/3    Up      0      (0)    5660      (10) T0 Host-6
ae0         Up    505     (0)    2481      (0) T0 CE-3
...
```

Multicast Traffic Outputs - LEAF-5

LEAF-5 being a legacy device and being configured with VLAN-101, now receives the traffic and floods it on its access interface, xe-0/0/2.0, though it does not have a receiver:

```
lab@LEAF-5> monitor interface traffic detail
Interface  Link  Input packets      (pps)      Output packets      (pps) Description
...
xe-0/0/2    Up      0      (0)      4666      (10) T0 Host-7
...
```

Multicast Traffic Outputs – BL-1 and BL-2

Again, we will ignore the traffic forwarding behavior on these devices until we get to Chapter 4.

Detailed Control Plane Verification

Verification of Base EVPN IGMP Snooping State

Verify that IGMP-snooping has been enabled on VLAN-101 on all the EVPN PEs, except LEAF-5:

```
lab@LEAF-1> show evpn instance extensive
Instance: default-switch
...
Number of bridge domains: 2
VLAN  Domain ID  ... IPv4 SG sync  IPv4 IM core nexthop ...
101   101        ... Enabled           131082 ...
102   102        ... Enabled           131084 ...
```

Similar outputs can be seen on LEAF-2, LEAF-3, LEAF-4, BL-1, and BL-2.

Now let's verify that the Type-3 route being advertised by the IGMP-snooping enabled regular LEAF EVPN PEs (LEAF-1 through LEAF-4) now carries the Multicast Flags extended community with the "IGMP Proxy Support" bit set:

```
lab@LEAF-1> show route extensive table default-switch.evpn.0 evpn-ethernet-tag-id 101 match-
prefix 3:* protocol bgp| grep "entry|Communities"
...
3:106.106.106.106:1::101::106.106.106.106/248 IM (1 entry, 1 announced)
Communities: target:1:1 encapsulation:vxlan(0x8) evpn-mcast-flags:0x1:snooping-
enabled
3:107.107.107.107:1::101::107.107.107.107/248 IM (1 entry, 1 announced)
Communities: target:1:1 encapsulation:vxlan(0x8) evpn-mcast-flags:0x1:snooping-
enabled
3:108.108.108.108:1::101::108.108.108.108/248 IM (1 entry, 1 announced)
Communities: target:1:1 encapsulation:vxlan(0x8) evpn-mcast-flags:0x1:snooping-
enabled
```

Verify that the non-IGMP-snooping enabled LEAF-5 continues to advertise

Type-3 routes without the Multicast Flags Extended community or with the “IGMP Proxy Support” bit reset in the Multicast Flags extended community:

```
lab@LEAF-1> show route extensive table default-switch.evpn.0 evpn-ethernet-tag-id 101 match-
prefix 3:109* protocol bgp| grep "entry|Communities"
3:109.109.109.109:1::102::109.109.109.109/248 IM (1 entry, 1 announced)
Communities: target:1:1 encapsulation:vxlan(0x8)
```

Verify that border LEAF EVPN PEs (BL-1 and BL-2) that are enabled with IGMP-snooping and L3-multicast (PIM), advertise Type-3 routes without the Multicast Flags Extended community or with the “IGMP Proxy Support” bit reset in the Multicast Flags extended community:

```
lab@LEAF-1> show route extensive table default-switch.evpn.0 evpn-ethernet-tag-id 101 match-
prefix 3:* protocol bgp| grep "entry|Communities"
3:101.101.101.101:1::101::101.101.101.101/248 IM (1 entry, 1 announced)
Communities: target:1:1 encapsulation:vxlan(0x8)
3:102.102.102.102:1::101::102.102.102.102/248 IM (1 entry, 1 announced)
Communities: target:1:1 encapsulation:vxlan(0x8)
```

Verify that the snooping-enabled LEAF PEs have programmed their multicast forwarding table such that all unknown multicast ingress traffic is forwarded towards the border LEAF PEs and towards the non-IGMP-snooping enabled PEs:

```
lab@LEAF-1> show multicast snooping route extensive VLAN VLAN-101
```

```
...
Group: 224.0.0.0/4
Source: *
VLAN: VLAN-101
Mesh-group: __all_ces__
Downstream interface list:
evpn-core-nh -(131082)
...
```

```
lab@LEAF-1> show evpn multicast-snooping next-hops 131082 detail
```

```
...
ID          Refcount KRefCount Downstream interface Addr
131082      3          1 vtep.32770-(1756)
              vtep.32773-(1764)
              vtep.32774-(1765)
```

Verification of EVPN IGMP Proxy State with Remote Receivers

Verify that on LEAF-4, the IGMP group membership has been learned on the VLAN-101 interface xe-0/0/3.0 by snooping the IGMP reports:

```
lab@LEAF-4> show igmp snooping membership VLAN VLAN-101 225.1.1.1
Instance: default-switch
VLAN: VLAN-101
...
Interface: xe-0/0/3.0, Groups: 1
Group: 225.1.1.1
Group mode: Exclude
Source: 0.0.0.0
```

```
Last reported by: 18.18.18.70
Group timeout:    209 Type: Dynamic
```

Verify that LEAF-4, having learned local IGMP membership for the group 225.1.1.1, builds local EVPN IGMP-proxy state and originates Type 6 IGMP proxy routes to notify remote PEs of its interest in receiving multicast traffic for the group:

```
lab@LEAF-4> show igmp snooping evpn proxy VLAN VLAN-101 225.1.1.1
Instance: default-switch
  Bridge-Domain: VLAN-101, VN Identifier: 101
    Group      Source      Local      Remote
    225.1.1.1  0.0.0.0      1          0

lab@LEAF-4> show evpn igmp-snooping proxy l2-domain-id 101 group 225.1.1.1 extensive
Instance: default-switch
  VN Identifier: 101
    Group      Source      Local      Remote      Corenh      Flood
    225.1.1.1  0.0.0.0      1          0            0            0

lab@LEAF-4> show route table default-switch.evpn.0 evpn-ethernet-tag-id 101 match-
prefix 6:* extensive protocol evpn
default-switch.evpn.0: 83 destinations, 83 routes (83 active, 0 holddown, 0 hidden)
6:108.108.108.108:1::101::225.1.1.1::108.108.108.108/520 (1 entry, 1 announced)
  *EVPN      Preference: 170
...
      Protocol next hop: 108.108.108.108
...
      Communities: encapsulation:vxlan(0x8)
      IGMP flags: 0xa
```

Verify that all remote PEs process this EVPN Type-6 and learn that LEAF-4 is an interested remote EVPN receiver for the group:

```
lab@LEAF-1> show route table default-switch.evpn.0 evpn-ethernet-tag-id 101 match-
prefix 6:* extensive
default-switch.evpn.0: 82 destinations, 82 routes (82 active, 0 holddown, 0 hidden)
6:108.108.108.108:1::101::225.1.1.1::108.108.108.108/520 (1 entry, 1 announced)
  *BGP      Preference: 170/-101
      Route Distinguisher: 108.108.108.108:1
...
      Source: 108.108.108.108
      Protocol next hop: 108.108.108.108
...
      Communities: target:1:1 encapsulation:vxlan(0x8)
      Import Accepted
      IGMP flags: 0xa
...

lab@LEAF-1> show evpn igmp-snooping proxy l2-domain-id 101 group 225.1.1.1 extensive
Instance: default-switch
  VN Identifier: 101
    Group      Source      Local      Remote      Corenh      Flood
    225.1.1.1  0.0.0.0      0          1          131087      0

lab@LEAF-1> show igmp snooping evpn proxy VLAN VLAN-101 225.1.1.1
Instance: default-switch
```

```

Bridge-Domain: VLAN-101, VN Identifier: 101
  Group      Source      Local      Remote
  225.1.1.1  0.0.0.0      0          1

```

Verification of Multicast Forwarding State with Remote Receiver

Verify that the multicast forwarding state created for group 225.1.1.1 in LEAF-4 includes the interested single-homed interface xe-0/0/3.0:

```
lab@LEAF-4> show multicast snooping route extensive VLAN VLAN-101 group 225.1.1.1
```

```

...
Group: 225.1.1.1/32
  Source: *
  VLAN: VLAN-101
  Mesh-group: __all_ces__
  Downstream interface list:
    evpn-core-nh -(131082) xe-0/0/3.0 -(1730)
...

```

Verify that on all other PEs, LEAF-4 (vtep.32771) has been added to the EVPN core next hop for group 225.1.1.1. Note that in addition BL-1 (vtep.32770), BL-2 (vtep.32774), and LEAF-5 (vtep.32773) will also be present in the EVPN core next hop for the group:

```
lab@LEAF-1> show evpn igmp-snooping proxy l2-domain-id 101 group 225.1.1.1 extensive
```

```

Instance: default-switch
  VN Identifier: 101
    Group      Source      Local      Remote      Corenh      Flood
    225.1.1.1  0.0.0.0      0          1          131087      0

```

```
lab@LEAF-1> show evpn multicast-snooping next-hops 131087 detail
```

```

...
ID          Refcount KRefCount Downstream interface Addr
131087      3          1 vtep.32770-(1756)
              vtep.32771-(1761)
              vtep.32773-(1764)
              vtep.32774-(1765)

```

Verify that the multicast forwarding state created for group 225.1.1.1 includes the EVPN core next hop shown above:

```
lab@LEAF-1> show multicast snooping route extensive VLAN VLAN-101 group 225.1.1.1
```

```

...
Group: 225.1.1.1/32
  Source: *
  VLAN: VLAN-101
  Mesh-group: __all_ces__
  Downstream interface list:
    evpn-core-nh -(131087)
...

```

Chapter 7

EVPN Intra-Multicast Optimization with Multihoming

Multicast optimization in EVPN provides several benefits in terms of bandwidth conservation and reduction in the replication/processing load on LEAF devices. By virtue of tracking IGMP Join state in the access interfaces (built from IGMP reports), and in the EVPN core (built from remote BGP Type-6 SMET routes), the EVPN device selectively forwards the multicast traffic only to the interfaces/PEs that have listeners behind it.

Such tracking of state and forwarding based on reports and SMET routes brings forth a scenario with EVPN multihoming where additional procedures are required to synchronize the state between the multihomed EVPN devices. This chapter describes the challenges that come about with enabling IGMP-snooping with EVPN multihoming and describes the procedures that address those challenges.

EVPN Multihoming Problem with IGMP-Snooping

First let's consider Figure 7.1, where LEAF-3, LEAF-4, and LEAF-5 are multihomed on an ESI.

Let's suppose that LEAF-4 is the DF and LEAF-3 and LEAF-5 are Non-DFs. LEAF-2 and LEAF-6 are other single-homed PEs.

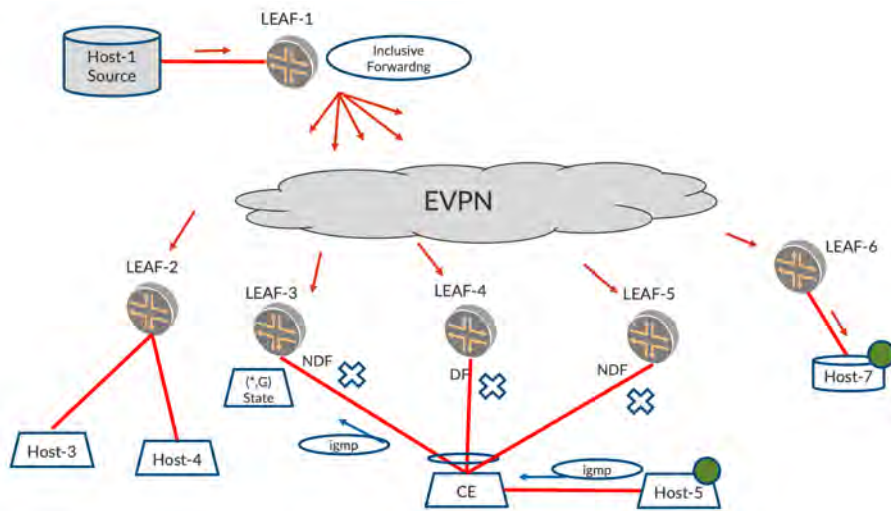


Figure 7.1 IGMP-Snooping with Multihoming Scenario

To describe the problem here, let's consider that LEAF-1 is the Ingress and it performs Inclusive (IMET) forwarding, so therefore LEAF-1 floods the traffic to the core instead of selectively forwarding based on Type-6 SMET routes.

Suppose Host-5 is an IGMP Host interested in traffic for group G and hence sends IGMP (*,G) reports. The CE that is multihomed to the three LEAFs over an aggregated interface (AE) bundle may send the report to any one of the multihomed LEAFs. This is by virtue of the hashing that the CE performs on the packet based on the tuple consisting of Source IP, Destination IP, Source MAC, etc. Okay, let's suppose that CE sends the IGMP (*,G) report to LEAF-3.

Upon receiving the (*,G) report, LEAF-3 creates an IGMP state for (*,G) such that if traffic for G arrives from the core or other access interfaces, it will forward the traffic to the CE and hence to Host-5. However, LEAF-3 is Non-DF per EVPN multihoming procedures in our example. Since the report did not reach LEAF-4 or LEAF-5, they do not create IGMP states for (*,G).

When LEAF-1 starts sending traffic for group G using Inclusive Forwarding, all the multihomed LEAFs will receive the traffic from core. By (CLASSICAL-DF-NDF) rules, only the DF forwards the traffic to access interface. Here, LEAF-4 is the DF.

However, since LEAF-4 does not have IGMP state for (*,G), LEAF-4 does not forward the traffic. LEAF-3, though it has the IGMP state for (*,G), does not forward the traffic because it is the non-DF on the multihomed link. LEAF-5, does not have the IGMP state as also it is an NDF.

Hence, we end up in a situation where neither of the multihomed PEs forward the traffic to Host-5, which is clearly undesirable.

If for some groups, by virtue of hash-tuple on CE, some IGMP reports reached LEAF-4, the DF, LEAF-4 will create a (*,G) state, and when traffic arrives from the core, LEAF-4, the DF, will forward traffic to CE and Rcv-1. This would hold in steady state.

There is still a problem, however. If LEAF-4 goes down or the multihomed interface on LEAF-4 goes down, LEAF-3 may be elected as the DF. Since LEAF-3 does not have IGMP (*,G) state, it will not forward the traffic. The refresh of the report may be sent to LEAF-5, the non-DF, and we may end up in the same state as before.

Multihoming Problem with Selective Forwarding

The problem with Inclusive Forwarding is that when the traffic from the EVPN core reaches all the multihomed peers, it does not get forwarded to the receiver because the LEAF-3 that has the IGMP state is the non-DF, and the LEAF-4 that is the DF does not have the IGMP state. Let's review Figure 7.2.

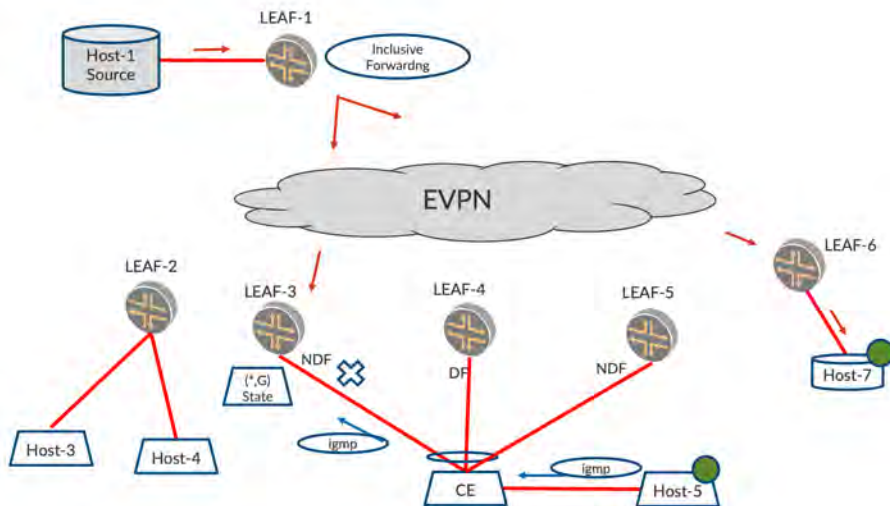


Figure 7.2 IGMP-Snooping with Multihoming Scenario with Selective Forwarding

This is a problem with Selective Forwarding, too. If LEAF-3 and LEAF-6 originated the BGP Type-6 SMET route to reflect the state that it learned on the access, LEAF-1 will forward the multicast traffic only to LEAF-3 and LEAF-6. In this case, LEAF-3, though it has the IGMP state for group, does not forward the traffic to the receiver because it is the non-DF.

With Inclusive Forwarding, and no IGMP-snooping (in other words, BUM flooding procedures), this is not a problem because traffic will be flooded by LEAF-1 and will reach all Egress LEAF devices. The LEAF that is the DF will forward the traffic to the Host.

Since IGMP-Snooping and Selective (SMET) Forwarding mitigates the ‘flood everywhere’ problem, it is worthwhile if we address this particular situation that arises due to IGMP-snooping and EVPN multihoming in the following sections on EVPN route types that can help solve this problem.

BGP Type-7 Join-Sync Route to Synchronize IGMP State in Multihoming PEs

To address the problem described here, we need the IGMP state for the group G to be synchronized among the multihomed PEs. If the IGMP state is synchronized among the multihomed PEs, the DF will have the (*,G) state and hence will forward the traffic towards the receivers.

The IGMP state for a particular group, learned over an ESI interface, is synchronized with other MH peers that host the same ESI by virtue of BGP Type-7 Join-Sync routes. Based on this exchange, the multihomed EVPN PEs can synchronize IGMP-state and install forwarding state for the group. By doing this, any of the relevant multihomed PEs will be in a position to forward the traffic on being elected the DF for that ESI.

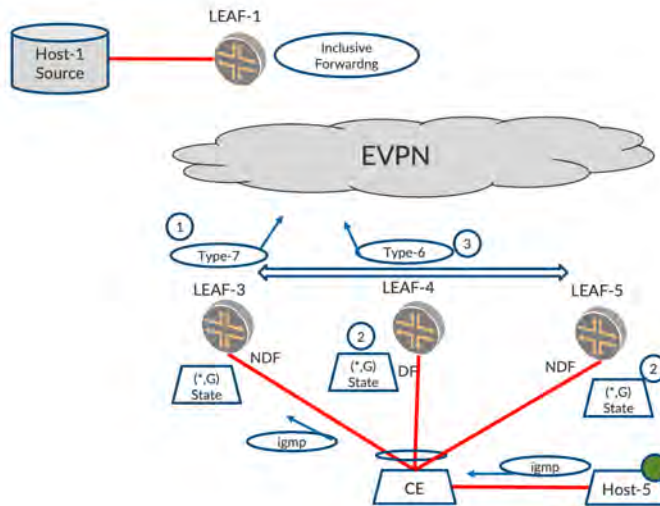


Figure 7.3

BGP Type-7 IGMP Join-Sync Route Control Plane Procedures

In Figure 7.3, the IGMP-report has reached LEAF-3. As part of event-[1], LEAF-3 originates a BGP Type-7 with the VLAN, group, and ESI information. This Type-7 route is imported only by LEAF-4 and LEAF-5 since these LEAFs host the ESI. Other LEAFs do not import this Type-7 route.

When LEAF-4 and LEAF-5 receive this Type-7 route, they build IGMP state for that particular group with the outgoing interface as the multihomed ESI interface (event-[2]). This way, the IGMP state for the group G is synchronized among the multihomed PEs.

By virtue of the IGMP state learned for the VLAN, the procedures described in Chapter 6 will result in the multihomed PEs originating Type-6 routes for the group/VLAN. (event-[3]).

EVPN Type-7 NLRI Route

The Type-7 route consists of a route distinguisher, VLAN, Multicast Group and Source, ESI, and an originating IP. The route also carries the ESI value in an extended community. This helps to ensure that only those PEs that host the ESI import this Type-7 route.

The Type-7 route also carries a community called EVI-COM. This community identifies the EVPN instance on which the IGMP state has to be synchronized. The typical route target that is added in Type 2 and Type-6 routes is not added because we want the Type-7 routes to be imported only on those PEs that host the particular ESI in question.

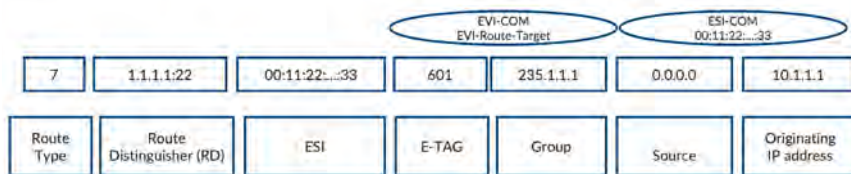


Figure 7.4

Type-7 Route Header

Tracking of Type-7 Routes from Peers

An EVPN PE creates local IGMP state based on the incoming IGMP report on a particular ESI. It may be the case that there are multiple hosts behind the same ESI sending reports for a group. It may be that the reports landed on different multihomed peers. In this case, the PEs that received IGMP report from access-side ESI interface originate Type-7.

When an IGMP state timeout occurs on one LEAF, it should not delete the state immediately. Instead, it should check if there are any other remote Type-7 routes advertised for the same VLAN/Group/ESI/EVI. If so, the original Type-7 is withdrawn but the IGMP state is retained.

Type-7 Withdraw Semantics

A Type-7 route is withdrawn when the IGMP state has timed out (hosts haven't refreshed the state on the ESI). A LEAF device, on receiving the Type-7 withdraw, should be careful in clearing the state. This Type-7 withdraw may be due to the ESI link being down or the originator node itself going down. The receiving PE has to ascertain that it is indeed a state timeout and then proceed to clear the state, or traffic drops will occur.

Non-DF to Originate Type-6 SMET Route for Better Convergence

Typically, upon receiving this Type-7 route only the DF needs to originate Type-6, since it is supposed to pull traffic from the core and forward the traffic to access. However, if the DF node fails or the ESI link on the DF fails, the new DF needs to be elected and then it has to originate the Type-6 SMET route. Also, the Ingress has to include the new DF in its outgoing PE list. These can result in considerable latency because BGP message exchange is involved.

This can be mitigated if the non-DF PEs also originate the Type-6 SMET route such that they pull the traffic from Ingress in steady state, but do not forward for access by virtue of being non-DF. Later, when the DF fails, the new DF, once elected, will have the traffic already arriving from the core. So, all the new DF needs to do is to forward the traffic to access. This results in considerable gain in convergence.

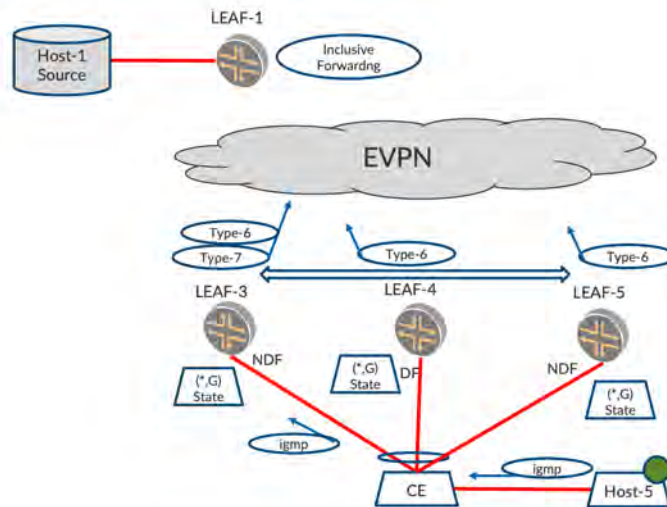


Figure 7.5

BGP Type-7 IGMP Join-Sync Route Control Plane Procedures for Better Convergence

Problem with IGMPv2 Leave

In earlier sections, we described how the IGMP (*,G) Report state is synchronized among the multihomed PEs. This section describes a problem wherein handling of an IGMP Leave message in a multihomed scenario needs special consideration.

IGMP Leave Primer

Just as an IGMP Report message for (*,G) is used to convey listener interest for a particular group G, the IGMP Leave message is used to convey withdrawal of listener interest for the group by the host.

When an IGMP Leave message is received by a switch, it needs to clear the IGMP (*,G) state that is used for forwarding such that traffic is no longer sent to the receivers. Before it clears the state, the switch needs to ensure that there are no other hosts interested in the group.

Towards this end, the switch sends out a last member query (LMQ) for group G to solicit any reports from other interested hosts. If the switch does not receive any report for the group G until a particular time period, LMQ Interval, the switch clears the state for (*,G). If any other host sends a report before the interval, the state is retained.

What Is Leave Latency?

Multicast applications are very sensitive to unwanted traffic. For instance, an IPTV Host may desire channel-1 by sending an IGMP report for group G1 and it may be receiving traffic. Say the bandwidth of the access interface has provision for traffic from one channel. When the user switches to channel-2, the host will withdraw interest for channel-1 by sending a Leave for group G1 and sending an IGMP report for group G2.

The switch has to process the Leave message for G1, check if other listeners exist, and if there are no other listeners, clear the state for G1 and stop forwarding to G1. Soon after, when the switch receives the IGMP report for group G2, it should create state for (*,G2) and start forwarding traffic for G2.

If for some reason the switch takes a longer time to clear state for G1, but creates for G2 and forwards traffic for both groups, the provisioned bandwidth may not be sufficient to carry both channels and may cause distortion on the IPTV host. Also, it may be that the host may not be capable of handling traffic for two groups.

Overall, it is imperative that the switch reacts soon enough on receiving a Leave message, does due diligence on LMQ, and clears state if no listeners exist. This delay in clearing state is referred to as *Leave Latency* and is an important factor for consideration in IPTV and stock-ticker applications.

If a Leave message is ‘lost on wire’ or not processed, the traffic will be forwarded to the listener. The switch, if it does not receive the IGMP reports periodically, say every 60 seconds, will clear the state after the IGMP-timeout interval, for example after 210 seconds.

So if a Leave message is lost or not handled, traffic will keep flowing until IGMP-timeout occurs. Leave Latency is given more careful consideration than the IGMP learning rate due to sensitivity of the IPTV hosts.

A Subtle Characteristic of the IGMPv2 Leave Message

IGMP Report messages have a source IP as the source of the host and the destination IP field as the Multicast Group address G that the host is interested in.

However, an IGMP Leave message has a destination IP field as 224.0.0.2, and the group that is desired to be withdrawn is present inside the payload of the Leave message. Historically, the reason has been the case with IGMPv2. Since IGMPv2 hosts are widely prevalent and have been working very well, it has remained this way.

Why is This Relevant to Optimized Multicast with Multihoming?

As discussed earlier, the CE will hash the incoming packets based on Source-IP, Destination IP, Source-MAC, etc. Since the destination IP field addresses of the Report are not the same as that of the Leave message, it so happens that the Report is sent to one multihomed EVPN PE, but the Leave message is sent to another multihomed EVPN PE.

IGMPv3

In the next section, we describe the challenge that arises due to the IGMPv2 Leave message having the destination address of 224.0.0.2. IGMPv3 is immune to this challenge because the IGMPv3 report and IGMPv3 Leave messages are carried in the payload and the destination address of both the Report and the Leave message is 224.0.0.22. Soon you will understand why this difference is relevant.

How Does This Cause a Problem in a Multihoming Scenario?

In Figure 7.7, the IGMP Report may be sent to LEAF-3. Later when Host-5 is not interested in traffic for group G, it sends out an IGMPv2 Leave for (*,G). This IGMP Leave for the same group may be sent to LEAF-5 because the destination IP address field for Leave is 224.0.0.2, while that of the IGMP report is the group itself. This is because the CE includes the destination IP address of the packet in calculating the hash tuple.

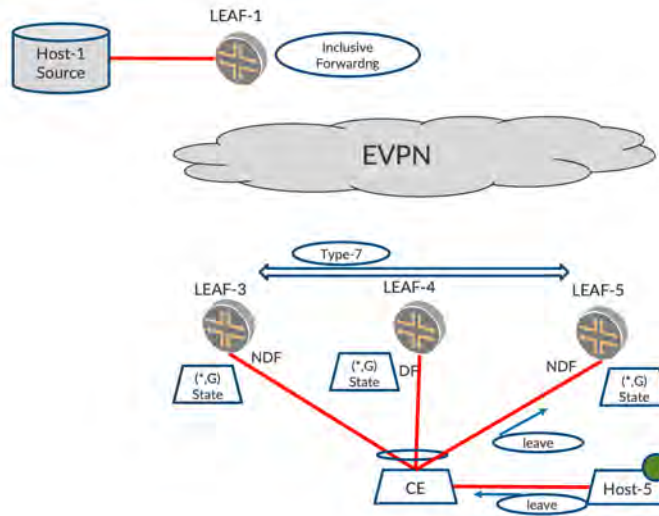


Figure 7.7 Problem with IGMPv2 Leave in a Multihoming Scenario

Let's look at this in more detail. In Figure 7.7, LEAF-3 received the IGMP Report on the access interface, created IGMP (*,G) state (locally learned), and originated a Type-7 route. LEAF-4 and LEAF-5 imported the Type-7 and create an IGMP (*,G) state (remotely learned). So far so good.

Consider the scenario when the leave message reaches LEAF-5. What is the expected behavior here? An LMQ has to be sent towards CE to solicit for reports and if no reports arrive before a particular interval, the state has to be cleared from all the multihomed PEs, particularly the DF, so that traffic forwarding is stopped for group G.

What are our options? If we let existing IGMP Leave processing rules alone kick in, LEAF-5, on hearing Leave, will send an LMQ. What if there are no hosts interested anymore? LEAF-3, where the report initially landed, has not heard of the Leave. So LEAF-3 will not clear state for a good 210 seconds and the Type-7 will remain advertised for that long.

LEAF-5 will be in a flux where it has received a Leave message on the access, but it also has a remotely learned state. LEAF-5 cannot withdraw a Type-7 since it was not the originator of the Type-7 route in the first place, so the problem gets more involved. LEAF-5 retains the state since there is a remote Type-7 from LEAF-3. LEAF-4, and the DF, will retain the translated IGMP (*,G) state and keep forwarding for 210 seconds.

Overall, even though the host sent a Leave and it indeed reached the multihomed PE set, the Leave latency is still 210 seconds. This is clearly undesirable, multihoming and optimization features notwithstanding.

It gets more complicated when another host sends a report before the LMQ and this report lands on a different PE, say LEAF-4. Now LEAF-3 and LEAF-4 have locally learned states, while LEAF-5 has a remotely learned state and a Leave to handle. In this scenario, the states have to reconcile such that the state of IGMP (*,G) is retained on all PEs.

So it is imperative to address this Leave handling scenario as much as an IGMP-Join sync. This problem is hard to solve relying on existing IGMP Leave handling procedures alone because there is a need to synchronize the Leave. Also, while IGMP Join (*,G) is a state that can be translated to BGP routes and exchanged, the Leave message is indeed an event and not a state.

With Leave messages, the earlier learned state is cleared. Hence, this appears as an event for the LEAF that got a Leave. Traditionally BGP notifies ADD/DELETE/UPDATE of a state. This Leave message is supposed to result in withdrawal of BGP route and state, but the fact that the Leave arrives on the non-originator of the state makes it difficult to convey the withdrawal of the state.

This event also has to be conveyed among the multihomed PEs such that they are in sync with the latest activity on the access ESI interface for the group to ensure Leave Latency is within acceptable limits. Different combinations also have to be able to work without much Leave Latency, the most common being (i) a IGMP Join followed by a Leave, or (ii), a Join followed by a Leave and followed by another Join, etc.

This problem is addressed by introducing another BGP route called Type-8 Leave-Sync route such that the multihomed PEs are in sync with each other about the state for the group (*,G).

BGP Type-8 Leave-Sync Route to Reconcile IGMP State

To address the problem described in the preceding section, let's explore the approach. The objective of the solution is the below:

- When a Join is received from access, the IGMP join state should be synced between all multihomed PEs.
- When a Leave is received from access on a multihomed setup, the LMQ timer should be started on all multihomed peers and the LMQ should be sent on the access by the DF.
- If no Join is received until the LMQ timer expires, then, the IGMP state should be deleted on all multihomed peers.
- If an override IGMP join is received before the LMQ timer expires, then, the IGMP state should be retained on all multihomed peers.

In Figure 7.8, a Join was received on the access from Host-5 on LEAF-3 and state was synchronized between the multihomed peers using a Type-7 Join-Sync route. Now, let's say, Host-5 sends an IGMP leave and this IGMP leave arrives on LEAF-5 (event-[1]).

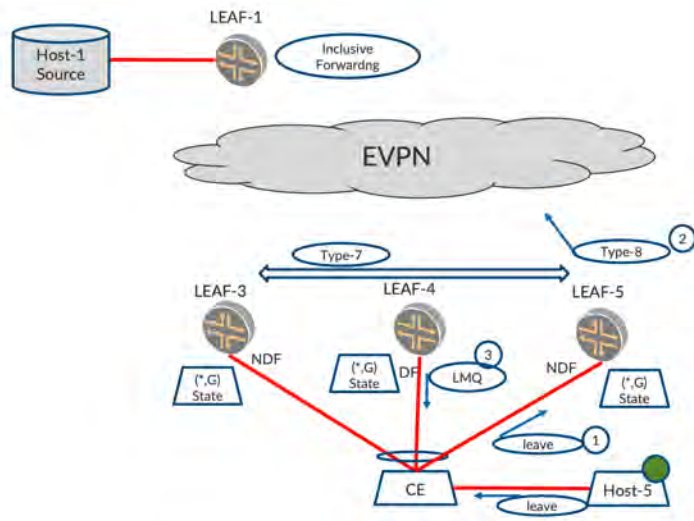


Figure 7.8 BGP Type-8 Leave-Sync Route

Upon receiving a Leave, LEAF-5 originates a Type-8 route (event-[2]) which has the NLRI fields the same as that of a Type-7 (VLAN, ESI, Group, source, origin IP, etc.). The trigger for the origination of Type-8 is a Leave message but there is no trigger for withdrawal of this route. Thus LEAF-5 starts a 'Leave-Sent-Timer' to withdraw the route once its purpose of synchronizing Leave is served.

Other multihomed peers, LEAF-3 and LEAF-4, receive the Type-8. Based upon this, all the multihomed PEs look to send out the send LMQ and start the LMQ timer. Since this LMQ is a multicast message, only the DF will send out the LMQ on the ESI. (event-[3]).

Two events can occur with this LMQ timer.

- Event-A: No Joins received on the LAN for this group before LMQ expires.
- Event-B: Join received on the LAN for this group before LMQ expires.

Event-A: No Joins Received on LAN in Response to LMQ (Join + Leave)

In Figure 7.9, after sending LMQ and starting the LMQ timer, the multihomed Peers wait to see if any IGMP report is refreshed for that group. If no reports were received and the LMQ timer expires (event-[1]).

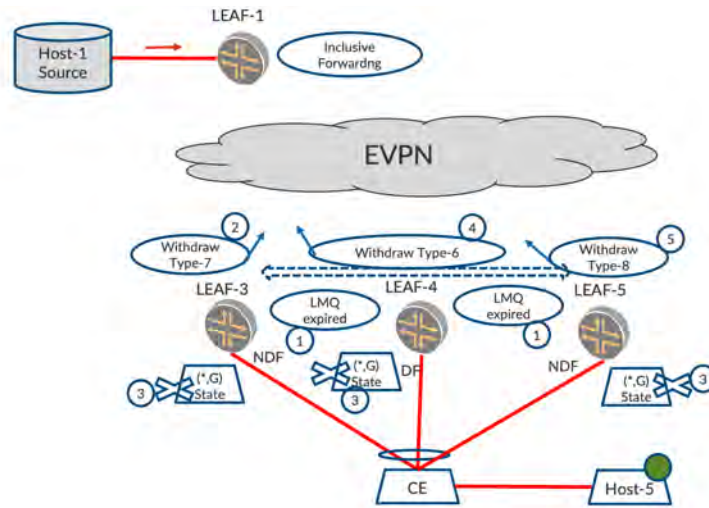


Figure 7.9

Event A: No Joins Received on LAN in Response to LMQ

In Figure 7.9, the originator of the Type-7 withdraws its Type-7 route (event-[2]). On the Type-7 remote route withdrawal, the multihomed peers delete the IGMP state, thus stopping forwarding traffic to the access. (event-[3]).

Based on the local state going away, if there is no other interested access interface in the VLAN for that group, the multihomed devices withdraw their earlier originated Type-6 route. (event-[4]).

With the Type-8 route, the states are now cleared and LEAF-5 withdraws its Type-8 route (event-[5]) to avoid any stale lingering of Type-8, thus clearing all the states for that flow.

Event-B: Join Received Before LMQ Timer Expires (Join + Leave + Join)

In Figure 7.10, after sending LMQ and starting the LMQ timer, the multihomed Peers wait for some time to see if any IGMP report is refreshed for that group. Suppose that before the LMQ timer expires another host sends an IGMP report on the LAN. This report may arrive on the earlier originator of Type-7, LEAF-3, or on a new originator, say, LEAF-4.

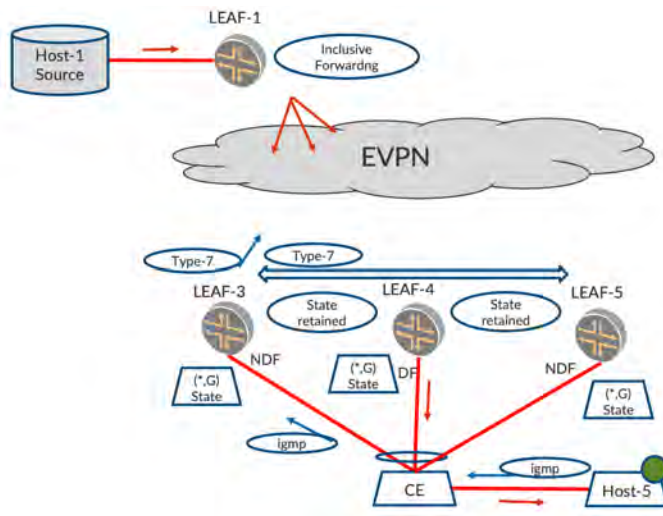


Figure 7.10

Event B: Join Received on LAN Before LMQ Expires

If Refresh Report Arrived on LEAF-3, the Earlier Type-7 Originator

LEAF-3, on receiving the refreshed report on access, stops its LMQ timer. It does not withdraw its Type-7 route. Later, on other multihomed peers, when the LMQ timer expires they check if there is a remote Type-7 for the group. Since there is a remote Type-7 for the group, the multihomed peers will retain the IGMP state and continue forwarding the traffic.

If Refresh Report Arrived on LEAF-4, Which Is Not the Earlier Type-7 Originator

On receiving the refreshed report on access, LEAF-4 stops its LMQ timer. It also originates its Type-7 route, which gets synced to other multihomed peers. Once they've received the remote Type-7 route, other multihomed peers import this into their table.

When LMQ expires (since its report was not refreshed), the originated Type-7 on LEAF-3 is withdrawn. However, since there is at least one other remote Type-7 (from LEAF-4), LEAF-3 retains the IGMP state.

And on LEAF-5, when LMQ expires, there is at least one remote Type-7 (from LEAF-4). So LEAF-5 also retains the IGMP state.

Putting It All Together for Optimized Multicast

Figure 7.11 repeats our original topology to bring back the big picture. Host-3 and Host-5, the multihomed hosts, are interested in traffic. The multihomed LEAF devices synchronize the IGMP state and ensure correct forwarding occurs. The Traffic Verification section of this chapter illustrates this scenario in detail.

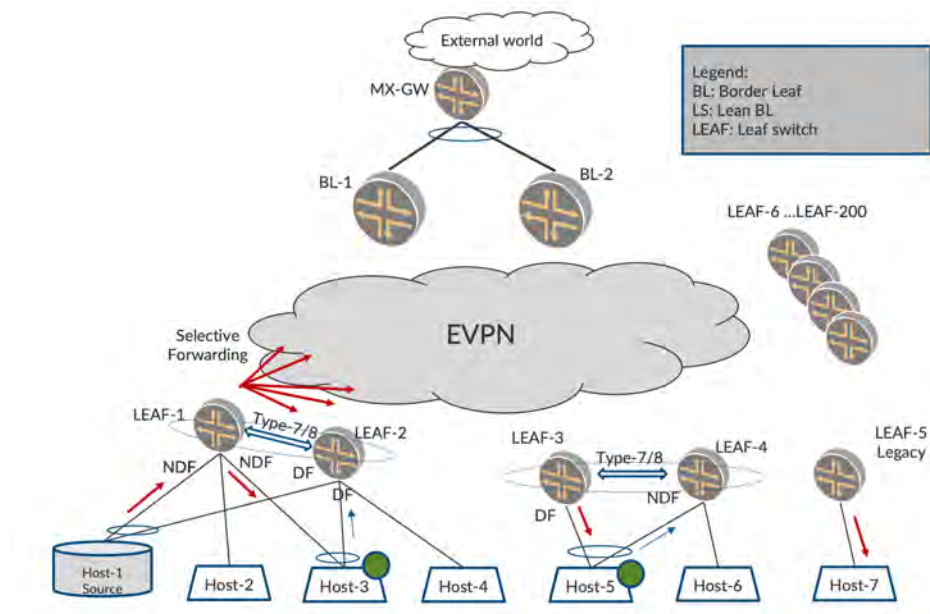


Figure 7.11 Optimized Multicast Forwarding

Chapter Summary

This chapter explored the nuances associated with optimizing multicast in multihomed topologies. To accomplish this, it detailed how IGMP reports received on an ESI link are synchronized amongst the multihomed peers that host the ESI. It also examined the origination of the Type-6 route by the multihomed peers for the sake of convergence.

Also of interest was the challenge that arises by virtue of the IGMP Leave message reaching a PE different from the one where IGMP Report was received, and how this is addressed by Type-8 route.

With Types-6/7/8, we have ensured multicast optimization in the core and access. We have also addressed the multihoming challenges. In Chapter 8 we will explore how the combination of these optimizations with AR brings in the best of both worlds.

Configuration

Figure 7.12 depicts the reference topology.

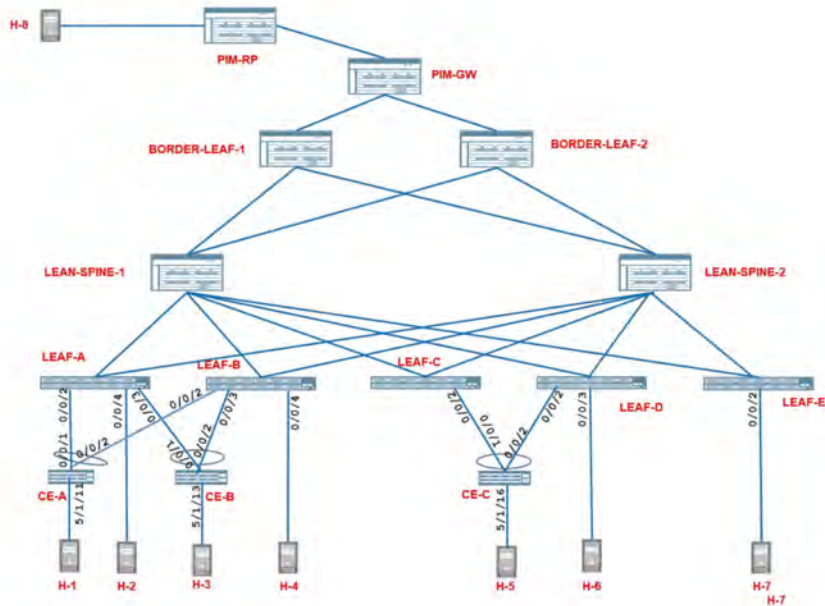


Figure 7.12 Reference Topology

Configuration

The configurations done in Chapter 6 are sufficient for this section.

Traffic Verification

Keeping the existing traffic that we started in Chapter 6 from Host-1, and the receiver on Host-6; on Host-3, that is multihomed to LEAF-1 and LEAF-2. Let's now start a receiver for the multicast group, 225.1.1.1 on VLAN-101.

From the RT statistics in Figure 7.13, you can see that the traffic sent by Host-1 at 10 pps is now being received by the interested single-homed receivers: Host-6, the interested multihomed receiver Host-3, and the legacy device, Host-7 in VLAN-101.

	Stat Name	Port Name	Link State	Frames Tx. Rate	Valid Frames Rx. Rate
1	10.216.45.202/Card20/Port01	HOST-1	Link Up	10	0
2	10.216.45.202/Card03/Port01	HOST-2	Link Up	0	0
3	10.216.45.202/Card20/Port02	HOST-3	Link Up	0	10
4	10.216.45.202/Card03/Port02	HOST-4	Link Up	0	0
5	10.216.45.202/Card20/Port03	HOST-5	Link Up	0	0
6	10.216.45.202/Card03/Port03	HOST-6	Link Up	0	10
7	10.216.45.202/Card03/Port04	HOST-7	Link Up	0	10
8	10.216.45.202/Card20/Port04	HOST-8	Link Up	0	0

Figure 7.13

RT Stats

Multicast Traffic Outputs - LEAF-1

As before, the load balanced multicast traffic arrives on access interface, ae0 on LEAF-1. LEAF-1 now forwards this traffic on its access interface ae1.0, on which it has learned an interested IGMP receiver. Recall that the SRC-LOCAL-BIAS rules allow LEAF-1 to forward the traffic on ae1.0 irrespective of the DF/NDF status:

```
lab@LEAF-1> monitor interface traffic detail
Interface  Link  Input packets  (pps)  Output packets  (pps)  Description
...
xe-0/0/4   Up    0              (0)    2564            (0)    T0 Host-2
...
ae0        Up    14467          (10)   0               (0)    T0 CE-1
ae1        Up    0              (0)    7260            (10)   T0 CE-2
...
```

The multicast traffic is forwarded on the VTEPs towards Border-LEAF PEs (101.101.101.101 and 102.102.102.102) and LEAF-5 (109.109.109.109). The traffic is also sent on the VTEPs towards LEAF-2 (106.106.106.106) and LEAF-4 (108.108.108.108) since they have interested receivers.

LEAF-3 (107.107.107.107) that does not have any interested receivers is still spared of the traffic:

```
lab@LEAF-1> show interfaces vtep extensive | grep "VXLAN Endpoint Type: Remote|Output packets.*pps"
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 106.106.106.106...
Output packets: 7238 9 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 101.101.101.101...
Output packets: 14854 9 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 108.108.108.108...
Output packets: 10638 10 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 107.107.107.107...
Output packets: 4719 0 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 109.109.109.109...
Output packets: 10144 10 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 102.102.102.102...
Output packets: 14855 9 pps
```

Multicast Traffic Outputs - LEAF-2

The access side IGMP-snooping functionality ensures that the multicast traffic arriving on LEAF-2 is not forwarded on the single-homed interface, xe-0/0/4.0, or on ae0.0, which does not have a receiver.

Though the multihomed access interface ae1.0 has a receiver, recall that the DST-LOCAL-BIAS rules ensure that the multicast traffic is not forwarded on this interface. This ensures that there is no traffic duplication towards the multihomed host, Host-2:

```
lab@LEAF-2> monitor interface traffic detail
Interface  Link  Input packets      (pps)  Output packets      (pps)  Description
...
xe-0/0/4   Up    0                  (0)    247                  (0)    T0 Host-4
ae0        Up    553                (0)    213                  (0)    T0 CE-1
ae1        Up    0                  (0)    208                  (0)    T0 CE-2
...
```

Multicast Traffic Outputs - LEAF-4, LEAF-5, and BL Devices

There is no change in the traffic forwarding behavior on these devices. The outputs are therefore omitted for the sake of brevity.

Detailed Control Plane Verification

Verification of EVPN Join-Sync with Multihomed Receivers

Since Host-3 is multihomed, the IGMP report may reach either LEAF-1 or LEAF-2. In our case, it reaches LEAF-1.

Let's verify that on LEAF-1, the IGMP group membership has been learned on VLAN-101 interface ae1.0 by snooping the IGMP reports:

```
lab@LEAF-1> show igmp snooping membership VLAN VLAN-101 225.1.1.1
Instance: default-switch
VLAN: VLAN-101
```

```

...
Interface: ae1.0, Groups: 1
  Group: 225.1.1.1
    Group mode: Exclude
    Source: 0.0.0.0
    Last reported by: 18.18.18.40
    Group timeout:    226 Type: Dynamic

lab@LEAF-1> show igmp snooping evpn database VLAN VLAN-101 225.1.1.1 interface ae1.0
Instance: default-switch
  Bridge-Domain: VLAN-101, VN Identifier: 101
    Group IP: 225.1.1.1, Source IP: 0.0.0.0
      Core NH: 131073
      Access OIF Count: 1
        Interface  Local  Remote
        ae1.0      1      0

lab@LEAF-1> show evpn igmp-snooping database l2-domain-id 101 group 225.1.1.1 extensive
Instance: default-switch
  VN Identifier: 101
    Group IP: 225.1.1.1, Source IP: 0.0.0.0
      Access OIF Count: 1
        Interface  ESI  Local  Remote
        ae1.0      00:22:22:22:22:22:22:22:22:22  1      0

```

Verify that LEAF-1 has originated an EVPN Type 7 Join-Sync route corresponding to this locally learned IGMP group membership on the multihomed interface ae1.0:

```

lab@LEAF-1> show route table __default_evpn__.evpn.0 evpn-ethernet-tag-id 101 match-
prefix 7:* extensive
__default_evpn__.evpn.0: 7 destinations, 7 routes (7 active, 0 holddown, 0 hidden)
7:105.105.105.105:1::2222222222222222::101::225.1.1.1::105.105.105.105/600 (1 entry, 1 announced)
  *EVPN Preference: 170

...
  Protocol next hop: 105.105.105.105

...
  Communities: encapsulation:vlan(0x8) es-import-target:22-22-22-22-22-22 evi-rt:1:1

...
  IGMP flags: 0xa

```

Verify that LEAF-2 has processed this EVPN Type 7 Join-Sync route from LEAF-1 and learned the remote membership on ae1.0:

```

lab@LEAF-2> show route table __default_evpn__.evpn.0 evpn-ethernet-tag-id 101 match-
prefix 7:* extensive
__default_evpn__.evpn.0: 7 destinations, 7 routes (7 active, 0 holddown, 0 hidden)
7:105.105.105.105:1::2222222222222222::101::225.1.1.1::105.105.105.105/600 (1 entry, 1 announced)
  *BGP Preference: 170/-101
    Route Distinguisher: 105.105.105.105:1

...
  Source: 105.105.105.105
  Protocol next hop: 105.105.105.105

...
  Communities: encapsulation:vlan(0x8) es-import-target:22-22-22-22-22-22 evi-rt:1:1
  Import Accepted
  IGMP flags: 0xa

...

```

```
lab@LEAF-2> show evpn igmp-snooping database l2-domain-id 101 group 225.1.1.1 extensive
Instance: default-switch
VN Identifier: 101
```

```
Group IP: 225.1.1.1, Source IP: 0.0.0.0
```

```
Access OIF Count: 1
```

Interface	ESI	Local	Remote
ae1.0	00:22:22:22:22:22:22:22:22	0	1

```
lab@LEAF-2> show igmp snooping evpn database VLAN VLAN-101 225.1.1.1 interface ae1.0
Instance: default-switch
```

```
Bridge-Domain: VLAN-101, VN Identifier: 101
```

```
Group IP: 225.1.1.1, Source IP: 0.0.0.0
```

```
Core NH: 131082
```

```
Access OIF Count: 1
```

Interface	Local	Remote
ae1.0	0	1

Verification of EVPN IGMP Proxy State with Multihomed Receivers

Verify that LEAF-2, having learned local IGMP membership for the group 225.1.1.1, via EVPN Type 7 Join-Sync routes, builds local EVPN IGMP-proxy state and originates a Type 6 IGMP Proxy route to notify remote PEs of its interest in receiving multicast traffic for the group.

Note that LEAF-2 would also learn remote proxy states based on the Type 6 originated by LEAF-4 (in Chapter 6) and LEAF-1 (now):

```
lab@LEAF-2> show igmp snooping evpn proxy VLAN VLAN-101 225.1.1.1
Instance: default-switch
```

```
Bridge-Domain: VLAN-101, VN Identifier: 101
```

Group	Source	Local	Remote
225.1.1.1	0.0.0.0	1	1

```
lab@LEAF-2> show evpn igmp-snooping proxy l2-domain-id 101 group 225.1.1.1 extensive
Instance: default-switch
```

```
VN Identifier: 101
```

Group	Source	Local	Remote	Corenh	Flood
225.1.1.1	0.0.0.0	1	2	131089	0

```
lab@LEAF-2> show route table default-switch.evpn.0 evpn-ethernet-tag-id 101 match-
prefix 6:* extensive protocol evpn
default-switch.evpn.0: 84 destinations, 84 routes (84 active, 0 holddown, 0 hidden)
6:106.106.106.106:1::101::225.1.1.1::106.106.106.106/520 (1 entry, 1 announced)
*EVPN Preference: 170
```

```
...
```

```
Protocol next hop: 106.106.106.106
```

```
...
```

```
Communities: encapsulation:vxlan(0x8)
IGMP flags: 0xa
```

Similarly, due to its local IGMP join, LEAF-1 would also originate a Type 6 route for the group 225.1.1.1. LEAF-1 would also learn remote proxy states based on the Type 6 originated by LEAF-4 (in Chapter 6) and LEAF-2 (now):

```

lab@LEAF-1> show igmp snooping evpn proxy VLAN VLAN-101 225.1.1.1
Instance: default-switch
  Bridge-Domain: VLAN-101, VN Identifier: 101
    Group      Source      Local      Remote
    225.1.1.1  0.0.0.0      1          1

lab@LEAF-1> show evpn igmp-snooping proxy l2-domain-id 101 group 225.1.1.1 extensive
Instance: default-switch
  VN Identifier: 101
    Group      Source      Local      Remote      Corenh      Flood
    225.1.1.1  0.0.0.0      1          2          131088      0

lab@LEAF-1> show route table default-switch.evpn.0 evpn-ethernet-tag-id 101 match-
prefix 6:* extensive protocol evpn
default-switch.evpn.0: 84 destinations, 84 routes (84 active, 0 holddown, 0 hidden)
6:105.105.105.105:1::101::225.1.1.1::105.105.105.105/520 (1 entry, 1 announced)
  *EVPN      Preference: 170

...
      Protocol next hop: 105.105.105.105

...
      Communities: encapsulation:vlan(0x8)
      IGMP flags: 0xa

```

Verify that all remote PEs process the EVPN Type 6 routes and learn LEAF-1 and LEAF-2 as interested remote EVPN receivers for the group. We did this in Chapter 6 for the Type-6 from LEAF-4. So this is left as an exercise for the reader.

Verification of Multicast Forwarding State

Verify that the multicast forwarding state created for the group 225.1.1.1 in LEAF-1 and LEAF-2 now includes the interested multihomed interface ae1.0:

```

lab@LEAF-1> show multicast snooping route extensive VLAN VLAN-101 group 225.1.1.1
...
Group: 225.1.1.1/32
  Source: *
  VLAN: VLAN-101
  Mesh-group: __all_ces__
    Downstream interface list:
      evpn-core-nh -(131088) ae1.0 -(1715)

...
lab@LEAF-2> show multicast snooping route extensive VLAN VLAN-101 group 225.1.1.1
...
Group: 225.1.1.1/32
  Source: *
  VLAN: VLAN-101
  Mesh-group: __all_ces__
    Downstream interface list:
      evpn-core-nh -(131089) ae1.0 -(1715)

...

```

Verify that on LEAF-1, in addition to LEAF-4 (vtep.32771), now LEAF-2 (vtep.32769) has also been added to the EVPN core next hop for the group 225.1.1.1. Note that in addition, BL-1 (vtep.32770), BL-2 (vtep.32774), and LEAF-5 (vtep.32773), will also be present in the EVPN core next hop for the group:

```
lab@LEAF-1> show evpn igmp-snooping proxy l2-domain-id 101 group 225.1.1.1 extensive
Instance: default-switch
```

```
VN Identifier: 101
```

Group	Source	Local	Remote	Corenh	Flood
225.1.1.1	0.0.0.0	1	2	131088	0

```
lab@LEAF-1> show evpn multicast-snooping next-hops 131088 detail
```

```
...
ID          Refcount KRefCount Downstream interface Addr
131088      3          1 vtep.32769-(1751)
               vtep.32770-(1756)
               vtep.32771-(1761)
               vtep.32773-(1764)
               vtep.32774-(1765)
```

Verify that the multicast forwarding state created for the group 225.1.1.1 has been updated to now include the EVPN core next hop seen above:

```
lab@LEAF-1> show multicast snooping route extensive VLAN VLAN-101 group 225.1.1.1
```

```
...
Group: 225.1.1.1/32
Source: *
VLAN: VLAN-101
Mesh-group: __all_ces__
Downstream interface list:
evpn-core-nh -(131088) ae1.0 -(1715)
...
```

Chapter 8

Assisted Replication with SMET

Multicast So Far

Let's quickly revisit the different features described earlier in this book to provide context for this chapter on AR+SMET:

- *Non-optimized multicast*: This is the simplest case, where multicast traffic is flooded everywhere (both to access and to all PEs in the core).
- *Assisted Replication*: This is the case where the replication load is transferred from the LEAF devices to the replicator devices. In this case, too, the traffic is flooded to all the access interfaces and to all the EVPN PEs in the core; only the replication load is transferred from the LEAF to the replicator, while the replicator floods the traffic to all PEs.
- *IGMP-Snooping and Selective (SMET) Forwarding*: In this case we enabled IGMP-snooping on a VLAN, by virtue of which traffic is forwarded only on those access interfaces where there is listener interest for the group. Also, traffic is forwarded only to those EVPN PEs in the core that have listener interest behind them (signaled by Type-6) – in this scheme, the replication load towards EVPN core, though significantly reduced due to selective forwarding, is still on the LEAF.

AR + SMET

In a data center fabric with several LEAF devices and a high volume of multicast traffic, it may be effective to have a combination of the features of IGMP-Snooping, Selective (SMET) Forwarding, and AR to conserve core access bandwidth and also transfer replication load from LEAF devices. That is to say, it is a good practice to configure both AR and IGMP-snooping in the fabric to get the best of both worlds. This is commonly referred to as AR+SMET.

With this scheme, both AR and IGMP-snooping will be enabled on LEAF, LS, and BL devices. The LEAF devices and BL devices will have the role of AR-LEAF, while the LS-1 device will be configured with the role of AR-Replicator.

For the AR-Replicators to help optimized forwarding in the core, they will be enabled with EVPN. Therefore these devices originate and process EVPN NLRI routes.

AR+SMET Roles and Configuration

The LEAF and BL devices in this scheme are configured as follows:

- IGMP-snooping on access interface: Traffic is forwarded only on those access interfaces where there is listener interest for the group. (IGMP Report)
- Selective (SMET) Forwarding: Traffic is forwarded only towards those EVPN PEs that express listener interest for the group with Type-6 and to PIM devices.
- AR-LEAF Role: The LEAF based on AR-LEAF Role, sends the traffic only to the Replicator. AR-LEAF is configured on the LEAFs.

The LS devices in this scheme are configured as follows:

- EVPN Family: In prior chapters, the Lean Spine devices were participating in the underlay. In Chapter 5, the Lean-Spine device participated in EVPN in order to originate Type-3 AR and IR NLRI Routes and build a provider tunnel. In this case, too, an EVPN family would need to be enabled. Additionally, with IGMP-snooping, the Type-6 routes received from the LEAFs will be processed to build a next hop (OIL), which has only the interested listeners.
- AR-Replicator: To help transfer the load from the LEAF devices to the AR-R. The Replicator configuration is enabled on the LS devices.
- Selective (SMET) Forwarding: The Replicator devices forward the traffic only to those EVPN PEs that have expressed listener interest. This is achieved by configuring IGMP-snooping on the VLANs.
- Enhanced AR Mode: The AR devices, when they are not able to retain the source-IP of the incoming packet, have to ensure that the packet is not sent to the multihomed peers of the PE that originated the traffic. Please refer to the section, Assisted Replication in Multihoming Environment, in Chapter 5.

This is relevant with SMET as well. For example, when we have AR plus SMET, the LEAF devices should forward the packet to its multihomed peers and the AR-R should skip forwarding to the multihomed peers of the PE that originated the traffic. Additionally, by virtue of SMET forwarding, the LEAF should send the traffic to the multihomed PEs only if there is listener interest for the group.

Summary Configuration of Features:

- BL and LEAF devices: IGMP-Snooping + AR-LEAF Role
- LS devices: IGMP-Snooping + AR-Replicator Role

AR+SMET Procedure

Consider Figure 8.1, where a listener comes up behind LEAF-4 and LEAF-5 for group G1. LEAF-4 and LEAF-5 send a Type-6 route for the group G1. This Type-6 is received by all EVPN PEs. LEAF-1 and AR-Replicators build SMET forwarding state for this group G1 with the relevant LEAFs alone in the OIL.

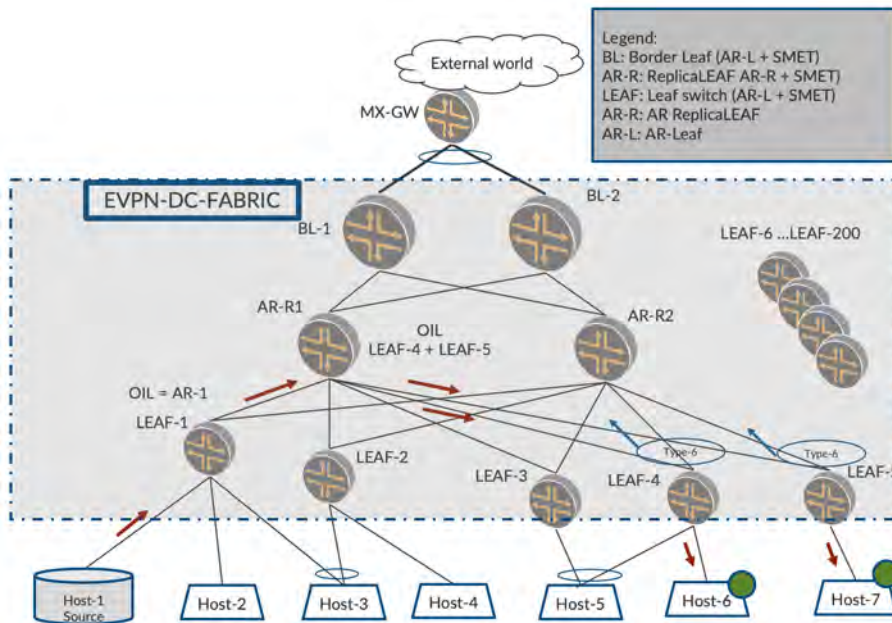


Figure 8.1 AR Plus SMET AR+SMET

The overall objective is that the traffic from behind LEAF-1 has to be forwarded only to LEAF-4 and LEAF-5. Also, the objective is that LEAF-1 sends only one copy of the traffic to AR-R1 and AR-R1 replicates to LEAF-4 and LEAF-5.

LEAF-1, based on SMET forwarding principles, adds the OIL for group G since there is remote listener interest for group G1. Since LEAF-1 is AR-LEAF, it adds only AR-R1 in its OIL. AR-R1 based on SMET forwarding principles adds the OIL for group G with LEAF-4 and LEAF-5.

LEAF-1 sends one copy to AR-1 and this is replicated by AR-1 to LEAF-4 and LEAF-5 alone. Traffic is not sent to the LEAF devices (LEAF-2, LEAF-4, LEAF-6, LEAF-200) thereby conserving core bandwidth, replication, and processing loads.

AR Plus SMET Enhanced Mode Procedures

The procedures for AR plus SMET in Enhanced Mode are similar to what was described earlier in this chapter, since the AR-R is not capable of retaining the Src-IP of the incoming packet from LEAF-1, AR-R skips replicating to all the LEAFs that are multihomed to LEAF-1. To be able to keep the local-bias behavior intact, LEAF-1 ingress replicates the packet to LEAF-2.

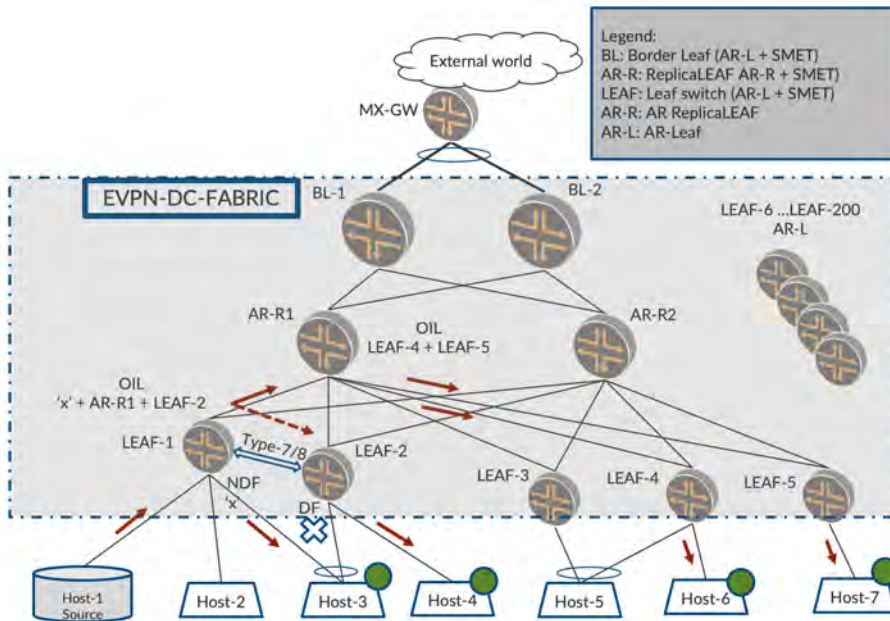


Figure 8.2

AR Plus SMET Enhanced Mode

There is listener interest from Host-3, Host-4, Host-6, and Host-7. We can see that Host-3 and Host-4 are behind LEAF-2 which is multihomed to LEAF-1. Also, the report from Host-3 would be synchronized using the Type-7/8 principles described in Chapter 7.

Now LEAF-1 has received Type-6 from LEAF-2, LEAF-4, and LEAF-5. Since LEAF-1 is configured as an AR-Leaf and also has deduced that it is multihomed to LEAF-2, it sends one copy to AR-R1 over AR-Tunnel and one copy to LEAF-2 over IR-Tunnel. Also, since traffic is coming from access interface, LEAF-1 forwards the traffic to Host-3 due to (*SRC-LOCAL-BIAS*).

On receiving the packet from LEAF-1, LEAF-2 does not forward to Host-3 despite being DF (*DST-LOCAL-BIAS*) since the packet arrived from multihomed-peer (S-VTEP-IP). However, LEAF-2 forwards the traffic to Host-4 since it is a single-homed interface (*SH-FWD*).

AR+SMET Benefit Illustration

Let's consider a case where there are 200 LEAFs in a DC-Fabric. There is a high volume of multicast traffic for 20 groups G1, and each group has a traffic rate of 1 Mbps. There are 10 LEAFs in the fabric interested in each group. Let's characterize the behavior with each mechanism.

Number of LEAFs in Fabric: $N = 200$

Number of groups: $G = 20$

Traffic Rate: $R = 1$ Mbps

Number of LEAFs interested in traffic: $T = 10$

Non-optimized Multicast

Core bandwidth consumption:

$$(N * G * R) = (200 * 20 * 1) = 4000 \text{ Mbps}$$

Replication Load on LEAF hosting the source:

$$(N * G) = 200 * 20 = 4000 \text{ times}$$

Link bandwidth consumption between LEAF and Leaf-Spine:

$$(N * G * R) = (200 * 20 * 1) = 4000 \text{ Mbps}$$

Assisted Replication (AR):

Core bandwidth consumption:

$$(N * G * R) = (200 * 20 * 1) = 4000 \text{ Mbps}$$

Replication Load on LEAF hosting the source:

$$(1 * G) = 1 * 20 = 20 \text{ times}$$

Link bandwidth consumption between LEAF and Leaf-Spine:

$$(1 * G * R) = (1 * 20 * 1) = 20 \text{ Mbps}$$

Optimized Multicast (SMET Forwarding) without AR

Core bandwidth consumption:

$$(T * G * R) = (10 * 20 * 1) = 200 \text{ Mbps}$$

Replication Load on LEAF hosting the source:

$$(T * G) = 10 * 20 = 200 \text{ times}$$

Link bandwidth consumption between LEAF and Lean-Spine:

$$(T * G * R) = (10 * 20 * 1) = 200 \text{ Mbps}$$

AR + SMET

Core bandwidth consumption:

$$(T * G * R) = (10 * 20 * 1) = 200 \text{ Mbps}$$

Replication Load on LEAF for each packet received from access:

$$(1 * G) = 1 * 20 = 20 \text{ times}$$

Link bandwidth consumption between LEAF and Lean-Spine:

$$(1 * G * R) = (1 * 20 * 1) = 20 \text{ Mbps}$$

You can see that with AR+SMET, the overall core bandwidth consumption is significantly reduced. Also, the link utilization between LEAF and the Lean Spine device is considerably reduced, and the replication load on LEAF is reduced.

Number of TORs in the fabric: N = 200	Number of Groups: G = 20
Number of TORs interested in the fabric: T = 10	Traffic Rate: R = 1 Mbps

Intra-VLAN Multicast	Non-Optimized Multicast	AR	SMET	AR + SMET	Gain Factor: AR+SMET vis-a-vis Non-optimized
Core Bandwidth consumption (in Mbps)	4000	4000	200	200	20
Replication Load on TOR hosting the source	4000	20	200	20	200
Link Bandwidth consumption between TOR and Lean Spine (in Mbps)	4000	20	200	20	200

By way of this scenario, we have achieved the objectives of optimized multicast and also replication load transfer from LEAF to AR-R, as graphed in Figure 8.3

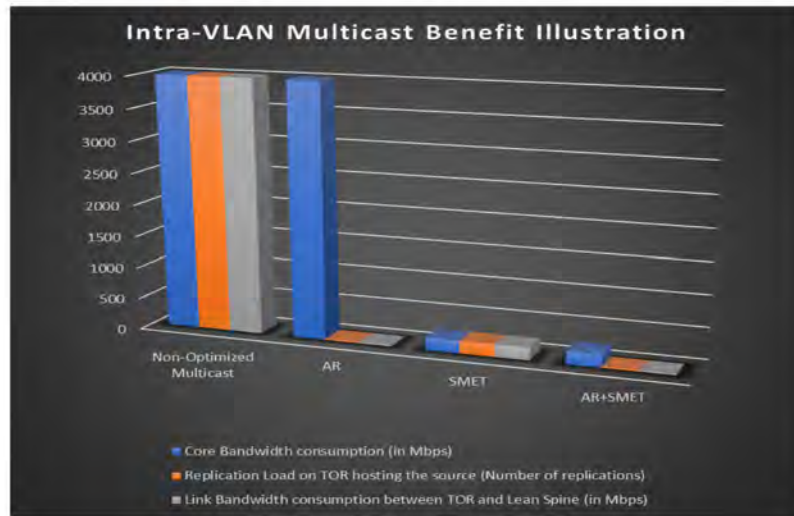


Figure 8.3 Intra-VLAN Multicast Benefit Illustration

The optimizations described thus far are applicable for a single-VLAN. When we visit Inter-VLAN routing/and forwarding, all of these optimizations will play a key role in significantly reducing the overall traffic.

Chapter Summary

This chapter explored AR+SMET and the procedures for the same. It addressed how an AR in conjunction with SMET and IGMP-snooping significantly reduces the core-bandwidth utilization, link bandwidth utilization between LEAF and Leaf-Spine, the Replication Load on the LEAF, and the processing load of LEAFs from unnecessary traffic. We considered a typical use case and calculated the traffic utilization and replication load on different devices in order to appreciate the benefits of these optimization techniques when deployed in unison.

Configuration

Figure 8.4 illustrates our reference topology.

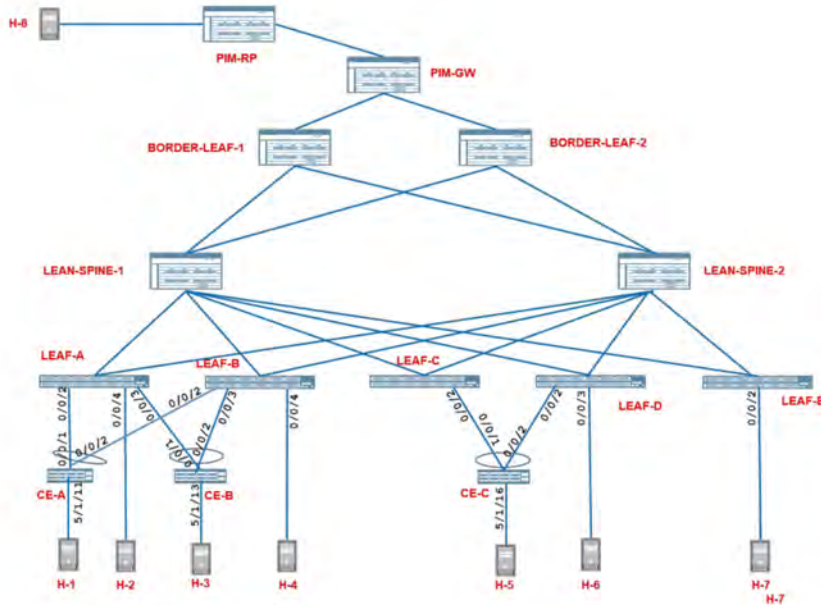


Figure 8.4 Reference Topology

Now let's see how the use of AR further optimizes our IGMP-snooping enabled network.

Configuration Details

Let's turn on AR on the EVPN PEs.

We will configure all our existing IGMP-snooping enabled EVPN PEs (LEAF-1 through LEAF-4, BL-1, and BL-2) as AR LEAF (AR-LEAF) devices and configure the erstwhile "Lean" spine PEs as EVPN PEs and use them as AR Replicator (AR-Replicator) devices for the network.

Configuring the EVPN overlay on SPINE-1

Copy and paste this configuration on SPINE-1.

Configure I-BGP for routing in the overlay:

```

set protocols bgp group OVERLAY type internal
set protocols bgp group OVERLAY local-address 103.103.103.103
set protocols bgp group OVERLAY family evpn signaling
set protocols bgp group OVERLAY local-as 65000
set protocols bgp group OVERLAY neighbor 101.101.101.101 description BL-1
set protocols bgp group OVERLAY neighbor 102.102.102.102 description BL-2
set protocols bgp group OVERLAY neighbor 104.104.104.104 description SPINE-1
set protocols bgp group OVERLAY neighbor 105.105.105.105 description LEAF-1
set protocols bgp group OVERLAY neighbor 106.106.106.106 description LEAF-2
set protocols bgp group OVERLAY neighbor 107.107.107.107 description LEAF-3
set protocols bgp group OVERLAY neighbor 108.108.108.108 description LEAF-4
set protocols bgp group OVERLAY neighbor 109.109.109.109 description LEAF-5
commit

```

Configure the customer VLANs:

```

set vlans VLAN-101 VLAN-id 101
set vlans VLAN-101 vxlan vni 101
set vlans VLAN-102 VLAN-id 102
set vlans VLAN-102 vxlan vni 102
commit

```

Configure EVPN to extend the customer VLANs:

```

set protocols evpn encapsulation vxlan
set protocols evpn extended-vni-list 101
set protocols evpn extended-vni-list 102
set switch-options vtep-source-interface lo0.0
set switch-options route-distinguisher 103.103.103.103:1
set switch-options vrf-target target:1:1
commit

```

Configure IGMP-snooping on the customer VLANs:

```

set protocols igmp-snooping VLAN VLAN-101
set protocols igmp-snooping VLAN VLAN-102
commit

```

Configuring the EVPN Overlay on SPINE-2

Copy and paste the below configuration on SPINE-2.

Configure I-BGP for routing in the overlay:

```

set protocols bgp group OVERLAY type internal
set protocols bgp group OVERLAY local-address 104.104.104.104
set protocols bgp group OVERLAY family evpn signaling
set protocols bgp group OVERLAY local-as 65000
set protocols bgp group OVERLAY neighbor 101.101.101.101 description BL-1
set protocols bgp group OVERLAY neighbor 102.102.102.102 description BL-2
set protocols bgp group OVERLAY neighbor 103.103.103.103 description SPINE-1
set protocols bgp group OVERLAY neighbor 105.105.105.105 description LEAF-1
set protocols bgp group OVERLAY neighbor 106.106.106.106 description LEAF-2
set protocols bgp group OVERLAY neighbor 107.107.107.107 description LEAF-3
set protocols bgp group OVERLAY neighbor 108.108.108.108 description LEAF-4
set protocols bgp group OVERLAY neighbor 109.109.109.109 description LEAF-5
commit

```

Configure the customer VLANs:

```

set vlans VLAN-101 VLAN-id 101
set vlans VLAN-101 vxlan vni 101

```

```
set vlans VLAN-102 VLAN-id 102
set vlans VLAN-102 vxlan vni 102
commit
```

Configure EVPN to extend the customer VLANs:

```
set protocols evpn encapsulation vxlan
set protocols evpn extended-vni-list 101
set protocols evpn extended-vni-list 102
set switch-options vtep-source-interface lo0.0
set switch-options route-distinguisher 104.104.104.104:1
set switch-options vrf-target target:1:1
commit
```

Configure IGMP-snooping on the customer VLANs:

```
set protocols igmp-snooping VLAN VLAN-101
set protocols igmp-snooping VLAN VLAN-102
commit
```

Configuring Spine PEs as Overlay BGP Neighbors on the LEAF and Border-LEAF PEs

Copy and paste the below configuration on the following devices:

LEAF-1, LEAF-2, LEAF-3, LEAF-4, BL-1, BL-2

```
set protocols bgp group OVERLAY neighbor 103.103.103.103 description SPINE-1
set protocols bgp group OVERLAY neighbor 104.104.104.104 description SPINE-2
commit
```

Configuring LEAF and Border-LEAF PEs as AR-LEAF

Copy and paste the below configuration on the following devices:

LEAF-1, LEAF-2, LEAF-3, LEAF-4, BL-1, BL-2

```
set protocols evpn assisted-replication LEAF
commit
```

Configuring SPINE-1 as AR-Replicator

Copy and paste the below configuration on SPINE-1.

Configure a secondary loopback address on SPINE-1:

```
set interfaces lo0 unit 0 family inet address 103.103.103.103/32 primary
set interfaces lo0 unit 0 family inet address 103.103.103.113/32
commit
```

Configure AR-Replicator using the secondary loopback IP as the AR-Replicator IP:

```
set protocols evpn assisted-replication replicator inet 103.103.103.113
set protocols evpn assisted-replication replicator vxlan-encapsulation-source-ip ingress-replication-ip
commit
```

Configuring SPINE-2 as AR-Replicator

Copy and paste the below configuration on SPINE-2.

Configure a secondary loopback address on SPINE-2:

```
set interfaces lo0 unit 0 family inet address 104.104.104.104/32 primary
set interfaces lo0 unit 0 family inet address 104.104.104.114/32
commit
```

Configure AR-Replicator using the secondary loopback IP as the AR-Replicator IP:

```
set protocols evpn assisted-replication replicator inet 104.104.104.114
set protocols evpn assisted-replication replicator vxlan-encapsulation-source-ip ingress-
replication-ip
commit
```

Traffic Verification

In Chapter 7 we started traffic from Host-1 and started receivers on Host-6 and Host-3. Having now turned on AR, let's see how the traffic forwarding has changed in the network.

From the RT statistics in Figure 8.5, you can see that the traffic sent by Host-1 at 10 pps continues to be received by the interested single-homed receiver, Host-6, the interested multihomed receiver Host-3, and the legacy device, Host-7 in VLAN-101.

	Stat Name	Port Name	Link State	Frames Tx, Rate	Valid Frames Rx, Rate
1	10.216.45.202/Card20/Port01	HOST-1	Link Up	10	0
2	10.216.45.202/Card03/Port01	HOST-2	Link Up	0	0
3	10.216.45.202/Card20/Port02	HOST-3	Link Up	0	10
4	10.216.45.202/Card03/Port02	HOST-4	Link Up	0	0
5	10.216.45.202/Card20/Port03	HOST-5	Link Up	0	0
6	10.216.45.202/Card03/Port03	HOST-6	Link Up	0	10
7	10.216.45.202/Card03/Port04	HOST-7	Link Up	0	10
8	10.216.45.202/Card20/Port04	HOST-8	Link Up	0	0

Figure 8.5

RT Stats

So what has changed? Let's look at LEAF-1 again.

Multicast Traffic Outputs - LEAF-1

Just as before, the load-balanced multicast traffic arrives on access interface, ae0 on LEAF-1 and is forwarded on its access interface ae1.0 on which it has learned an IGMP report.

But things have changed in the core. Unlike before, now LEAF-1 sends out only two copies of the packet towards the core, thereby reducing the replication load on LEAF-1 and also conserving the physical link bandwidth between LEAF-1 and the Spine.

The multicast traffic is sent towards one of the AR-Replicator Spines, in our case SPINE-2. Note that this traffic is sent to the AR-IP of SPINE-2 (104.104.104.114).

In addition, the traffic is also sent on the VTEP towards the multi-homing peer PE, LEAF-2 (106.106.106.106):

```
lab@LEAF-1> show interfaces vtep extensive | grep "VXLAN Endpoint Type:.*Remote|Output packets.*pps"
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 106.106.106.106...
Output packets:          844          9 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 107.107.107.107...
Output packets:          0          0 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 108.108.108.108...
Output packets:        2847          0 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 101.101.101.101...
Output packets:        3536          0 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 102.102.102.102...
Output packets:        3536          0 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 109.109.109.109...
Output packets:        3536          0 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 103.103.103.113...
Output packets:         222          0 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 103.103.103.103...
Output packets:          0          0 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 104.104.104.104...
Output packets:          0          0 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 104.104.104.114...
Output packets:        1521         10 pps
```

Multicast Traffic Outputs - SPINE-2

The multicast traffic load balanced between the AR-Replicators by LEAF-1 is received by SPINE-2 on its AR-tunnel, vtep.32782:

```
lab@SPINE-2> show interfaces vtep extensive | grep "VXLAN Endpoint Type:AR Remote|Input packets.*pps"
...
VXLAN Endpoint Type: AR Remote, VXLAN Endpoint Address: 105.105.105.105, L2 Routing Instance:
default-switch, L3 Routing Instance: default
Input packets:          1414          9 pps
...
```

SPINE-2 selectively replicates this traffic only to the interested AR-LEAF PEs. The traffic is forwarded on the VTEPs towards Border-LEAF PEs (101.101.101.101 and 102.102.102.102) and on the VTEP towards LEAF-5 (109.109.109.109).

The traffic is also sent on the VTEP towards LEAF-4 (108.108.108.108) that has interested receivers *AND* is not multihomed to the source PE, LEAF-1:

```
lab@SPINE-2> show interfaces vtep extensive | grep "VXLAN Endpoint Type:.*Remote|Output packets.*pps"
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 108.108.108.108...
Output packets:        1418         10 pps
```

```

VXLAN Endpoint Type: AR Remote, VXLAN Endpoint Address: 108.108.108.108...
Output packets: 0 0 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 101.101.101.101...
Output packets: 1412 9 pps
VXLAN Endpoint Type: AR Remote, VXLAN Endpoint Address: 101.101.101.101...
Output packets: 0 0 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 102.102.102.102...
Output packets: 1419 10 pps
VXLAN Endpoint Type: AR Remote, VXLAN Endpoint Address: 102.102.102.102...
Output packets: 0 0 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 109.109.109.109...
Output packets: 1419 9 pps
VXLAN Endpoint Type: AR Remote, VXLAN Endpoint Address: 109.109.109.109...
Output packets: 0 0 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 103.103.103.113...
Output packets: 0 0 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 103.103.103.103...
Output packets: 6 0 pps
VXLAN Endpoint Type: AR Remote, VXLAN Endpoint Address: 103.103.103.103...
Output packets: 0 0 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 105.105.105.105...
Output packets: 6 0 pps
VXLAN Endpoint Type: AR Remote, VXLAN Endpoint Address: 105.105.105.105...
Output packets: 0 0 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 106.106.106.106...
Output packets: 6 0 pps
VXLAN Endpoint Type: AR Remote, VXLAN Endpoint Address: 106.106.106.106...
Output packets: 0 0 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 107.107.107.107...
Output packets: 6 0 pps
VXLAN Endpoint Type: AR Remote, VXLAN Endpoint Address: 107.107.107.107...
Output packets: 0 0 pps

```

Multicast Traffic Outputs – LEAF-2, LEAF-4, LEAF-5, BL-1 and BL-2

There is no change in the traffic forwarding behavior on these devices, but for the fact that, except LEAF-2 that continues to receive traffic on the VTEP from LEAF-1, all other devices receive the traffic on the VTEP from SPINE-2 (104.104.104.104).

The outputs are therefore omitted for the sake of brevity.

Detailed Control Plane Verification

For the sake of brevity, we will focus on LEAF-1 for our AR-LEAF verifications and on SPINE-2 for our AR-Replicator verifications below. The state of the other AR-LEAFs and AR-Replicator PEs will be similar.

Verification of base EVPN Assisted-Replication State

Verify that the AR-R SPINE-2 has set up AR-tunnels corresponding to each remote PE, to receive traffic sent by them using the AR-IP.

For each AR-tunnel, the regular remote VTEPs corresponding to the PE(s) with which the source-PE is multihomed and the regular remote VTEP corresponding to source-PE itself, can also be seen in the output (Local Bias Logical Interface...). This information is used by the AR-R to avoid replicating the traffic arriving from this source PE on the AR-tunnel to multihomed peers of the source PE or to the source PE itself:

```
lab@SPINE-2> show interfaces vtep extensive | grep "VXLAN Endpoint Type: AR Remote|vtep"
...
Logical interface vtep.32771 (Index 559) (SNMP ifIndex 558) (HW Token 4294967295) (Generation 177)
VXLAN Endpoint Type: AR Remote, VXLAN Endpoint Address: 108.108.108.108, L2 Routing
Local Bias Logical Interface Name: vtep.32770, Index: 558, VXLAN Endpoint Address: 108.108.108.108
Local Bias Logical Interface Name: vtep.32785, Index: 585, VXLAN Endpoint Address: 107.107.107.107
...
Logical interface vtep.32773 (Index 561) (SNMP ifIndex 560) (HW Token 4294967295) (Generation 179)
VXLAN Endpoint Type: AR Remote, VXLAN Endpoint Address: 101.101.101.101, L2 Routing
Local Bias Logical Interface Name: vtep.32772, Index: 560, VXLAN Endpoint Address: 101.101.101.101
...
Logical interface vtep.32775 (Index 575) (SNMP ifIndex 562) (HW Token 4294967295) (Generation 181)
VXLAN Endpoint Type: AR Remote, VXLAN Endpoint Address: 102.102.102.102, L2 Routing
Local Bias Logical Interface Name: vtep.32774, Index: 574, VXLAN Endpoint Address: 102.102.102.102
...
Logical interface vtep.32777 (Index 577) (SNMP ifIndex 564) (HW Token 4294967295) (Generation 183)
VXLAN Endpoint Type: AR Remote, VXLAN Endpoint Address: 109.109.109.109, L2 Routing Instance:
default-switch, L3 Routing Instance: default
Local Bias Logical Interface Name: vtep.32776, Index: 576, VXLAN Endpoint Address: 109.109.109.109
...
Logical interface vtep.32780 (Index 580) (SNMP ifIndex 567) (HW Token 4294967295) (Generation 186)
VXLAN Endpoint Type: AR Remote, VXLAN Endpoint Address: 103.103.103.103, L2 Routing
Local Bias Logical Interface Name: vtep.32779, Index: 579, VXLAN Endpoint Address: 103.103.103.103
...
Logical interface vtep.32782 (Index 582) (SNMP ifIndex 569) (HW Token 4294967295) (Generation 188)
VXLAN Endpoint Type: AR Remote, VXLAN Endpoint Address: 105.105.105.105, L2 Routing Instance:
default-switch, L3 Routing Instance: default
Local Bias Logical Interface Name: vtep.32781, Index: 581, VXLAN Endpoint Address: 105.105.105.105
Local Bias Logical Interface Name: vtep.32783, Index: 583, VXLAN Endpoint Address: 106.106.106.106
...
Logical interface vtep.32784 (Index 584) (SNMP ifIndex 571) (HW Token 4294967295) (Generation 190)
VXLAN Endpoint Type: AR Remote, VXLAN Endpoint Address: 106.106.106.106, L2 Routing
Local Bias Logical Interface Name: vtep.32783, Index: 583, VXLAN Endpoint Address: 106.106.106.106
Local Bias Logical Interface Name: vtep.32781, Index: 581, VXLAN Endpoint Address: 105.105.105.105
...
Logical interface vtep.32786 (Index 586) (SNMP ifIndex 573) (HW Token 4294967295) (Generation 192)
VXLAN Endpoint Type: AR Remote, VXLAN Endpoint Address: 107.107.107.107, L2 Routing Instance:
default-switch, L3 Routing Instance: default
Local Bias Logical Interface Name: vtep.32785, Index: 585, VXLAN Endpoint Address: 107.107.107.107
Local Bias Logical Interface Name: vtep.32770, Index: 558, VXLAN Endpoint Address: 108.108.108.108
```

Verify that the AR-R EVPN PEs (SPINE-1 and SPINE-2) advertise a Replicator-1R Type-3 route with the Originating Router's IP address set to their respective AR-IPs.

The PMSI Tunnel Attribute (PTA) in this route should have the Tunnel-Type set as AR-Replication and the 2-bit type field in the flags field of the PTA set to represent the AR-Role as AR-Replicator (01).

The route also carries the multicast flags extended community. In addition to the “IGMP Proxy Support” bit, this route has the Extended-MH-AR bit also set:

```
lab@LEAF-1> show route extensive table default-switch.evpn.0 evpn-ethernet-tag-id 101 match-
prefix 3:* protocol bgp | grep "entry|PMSI|Communities"
...
3:103.103.103.103:1::101::103.103.103.113/248 IM (1 entry, 1 announced)
PMSI: Flags 0x8: Label 6: Type ASSISTED-REPLICATION 103.103.103.113 Role AR-
REPLICATOR
Communities: target:1:1 encapsulation:vlan(0x8) evpn-mcast-flags:0x4:extended-MH-AR
...
3:104.104.104.104:1::101::104.104.104.114/248 IM (1 entry, 1 announced)
PMSI: Flags 0x8: Label 6: Type ASSISTED-REPLICATION 104.104.104.114 Role AR-
REPLICATOR
Communities: target:1:1 encapsulation:vlan(0x8) evpn-mcast-flags:0x4:extended-MH-AR
...
```

Similar outputs can be seen on all the EVPN PEs.

Verify that the AR-Leaf EVPN PEs (LEAF-1, LEAF-2, LEAF-3, LEAF-4, LEAF-5, BORDER-LEAF-1, and BORDER-LEAF-2) advertise their regular Type-3 route. However, the 2-bit type field in the flags field of the PMSI Tunnel Attribute (PTA) in this route will have to be set to represent the AR-Role as AR-Leaf (01):

```
lab@LEAF-1> show route extensive table default-switch.evpn.0 evpn-ethernet-tag-id 101 match-
prefix 3:* protocol evpn | grep "entry|PMSI"
...
3:105.105.105.105:1::101::105.105.105.105/248 IM (1 entry, 1 announced)
PMSI: Flags 0x10: Label 101: Type INGRESS-REPLICATION 105.105.105.105 AR-LEAF
...
```

Similar outputs can be seen on all the AR-LEAF EVPN PEs.

Verify that the AR-LEAF LEAF-1 has set up VTEPs to send traffic to the AR-R PEs, SPINE-1, and SPINE-2 using their advertised AR-IP:

```
lab@LEAF-1> show interfaces vtep extensive | grep "VXLAN Endpoint Type: Remote|vtep"
...
Logical interface vtep.32775 (Index 577) (SNMP ifIndex 563) (HW Token 4294967295) (Generation 210)
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 103.103.103.113, L2 Routing Instance: default-
switch, L3 Routing Instance: default
...
Logical interface vtep.32778 (Index 580) (SNMP ifIndex 566) (HW Token 4294967295) (Generation 213)
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 104.104.104.114, L2 Routing Instance: default-
switch, L3 Routing Instance: default
```

Verification of EVPN Assisted-Replication State with Snooping

Verify that LEAF-1 has learned SPINE-1 (AR-IP 103.103.103.113) and SPINE-2 (AR-IP 104.104.104.114) as AR-Replicators:

```
lab@LEAF-1> show evpn multicast-snooping assisted-replication replicators l2-domain-id 101
Instance: default-switch
AR Role: AR LEAF
VN Identifier: 101
```

```

Operational Mode: Extended AR
Replicator IP  Nexthop Index  Interface  Mode
103.103.103.113 1789      vtep.32775 Extended AR
104.104.104.114 1783      vtep.32778 Extended AR

```

Verify that LEAF-1 has added LEAF-2 (106.106.106.106) to its list of multihomed peer PEs to whom traffic should continue to be ingress replicated from LEAF-1:

```

lab@LEAF-1> show evpn multicast-snooping assisted-replication multihomed-peers extensive
Instance: default-switch
Neighbor address: 106.106.106.106
  Nexthop Index: 1734
Interface: vtep.32769
  Local Multihomed ESIs
    00:11:11:11:11:11:11:11:11:11
    00:22:22:22:22:22:22:22:22:22

```

Verification of Multicast Forwarding State

Verify that on LEAF-1, the EVPN core next hop for the group 225.1.1.1 now includes only the load balancing next hop to the AR-R PEs and the VTEP to the multihomed peer PE, LEAF-2 (vtep.32769):

```

lab@LEAF-1> show evpn igmp-snooping proxy l2-domain-id 101 group 225.1.1.1 extensive
Instance: default-switch
  VN Identifier: 101
    Group      Source      Local  Remote  Corenh      Flood
    225.1.1.1  0.0.0.0      1      2      131084      0

```

```

lab@LEAF-1> show evpn multicast-snooping next-hops 131084 detail
...
ID          Refcount KRefCount Downstream interface Addr
131084      3          1 131081
                        vtep.32769-(1734)

```

```

lab@LEAF-1> show evpn multicast-snooping assisted-replication next-hops l2-domain-id 101
Instance: default-switch
AR Role: AR LEAF
  VN Identifier: 101
    Load Balance Nexthop Index: 131081
      Load balance to:
        Nexthop Index  Interface  AR IP
        1789           vtep.32775  103.103.103.113
        1783           vtep.32778  104.104.104.114

```

Verify that the multicast forwarding state created for the group 225.1.1.1 has now been updated to include the new EVPN core-next hop seen above:

```

lab@LEAF-1> show multicast snooping route extensive VLAN VLAN-101 group 225.1.1.1
...
Group: 225.1.1.1/32
Source: *
VLAN: VLAN-101
Mesh-group: __all_ces__
  Downstream interface list:
    evpn-core-nh -(131084) ae1.0 -(1715)

```

Chapter 9

EVPN Inter-VLAN Multicast Routing without Optimization

Part 2 of this book explored how Intra-VLAN multicast works in an EVPN data center fabric. It also explored different optimization procedures like IGMP-snooping, Selective (SMET) Forwarding, and Assisted Replication (AR). These procedures are about traffic forwarding rules within a VLAN.

Traditionally, customers have several VLANs in their data center fabric, for example, an Admin department in VLAN-A, Payroll in VLAN-B, Engineering in VLAN-C, etc. In IPTV applications, a set of subscribers may be hosted in one VLAN and another set in another VLAN and so on. The number of VLANs in an EVPN fabric can range from 20 to around 1000.

There may be a need to forward multicast traffic across VLANs. The sources of the multicast traffic may be present in some VLANs, while listeners may be present in some or all VLANs. There may be a need for routing the multicast traffic from one VLAN to the interested listeners in other VLANs.

PIM has been the de facto protocol for Inter-VLAN multicast routing and operates at the Level 3 layer. For inter-subnet multicast, we will first describe how PIM on an external M-Router can be used to achieve multicast routing across the VLANs in the fabric. Then we will describe how L3 routing can be achieved by running L3 multicast routing with PIM on IRB on the EVPN device itself.

The model that we will describe is called *Centrally Routed Bridged* (CRB) model for multicast, where routing is performed in a central device as opposed to performing routing in the *Edge Routed Bridge* (ERB) model. It may be worth noting that unicast can be configured to work in an ERB model, while multicast routing is configured to work in a CRB model.

By the end of this chapter, you should have a fair understanding of:

- Inter-subnet multicast in an EVPN data center fabric using External M-Router.
- Inter-subnet Multicast in an EVPN data center with L3-PIM routing on devices with IRB.
- How the ‘*flood everywhere*’ problem is compounded with Inter-VLAN multicast.

We also describe the procedures of Inter-Subnet Multicast Routing without any Intra-VLAN optimization. This is first to illustrate the procedures for Inter-VLAN multicast alone, and how the problem of ‘flood everywhere’ is compounded manifold in Inter-VLAN scenarios.

In Chapter 10 we throw in the optimization techniques of IGMP-snooping, Selective Forwarding, and Assisted Replication, and illustrate the bandwidth conservation that can be achieved through using those techniques.

Inter-subnet Multicast

There are two ways L3 inter-subnet multicast can be deployed:

- a.) Perform L3-routing on a peripheral device (External M-Router/Tenant Router)
- b.) Perform L3-routing on a device that is part of the fabric using IRB

Though (b) is the preferred and widely deployed method for Inter-subnet multicast, (a) is sometimes used to achieve the same result when devices in the fabric do not support IRB routing, or, if for security reasons, operators prefer L3-routing of multicast on a peripheral device. Once we have finished exploring (a) in the next section, later chapters and sections will focus on (b).

Inter-subnet Multicast with External Multicast Router

Consider the topology depicted in Figure 9.1. A data center fabric has two VLANs, v-red and v-blue. The Border Leaf (BL) devices do not perform L3-multicast routing but may perform unicast routing (achieved by configuring IRB with routing protocols for unicast but not enabled with PIM on the IRB interface).

There is a ‘PIM M-Router’ device on the periphery of the fabric that is assigned the responsibility for L3-inter-subnet multicast routing for the fabric. One of the EVPN devices, BL-2, is connected to the PIM M-Router. BL-2 and the PIM M-Router host all VLANs in the fabric and BL-2 L2 switches the traffic and listener interest from fabric to the PIM M-Router.

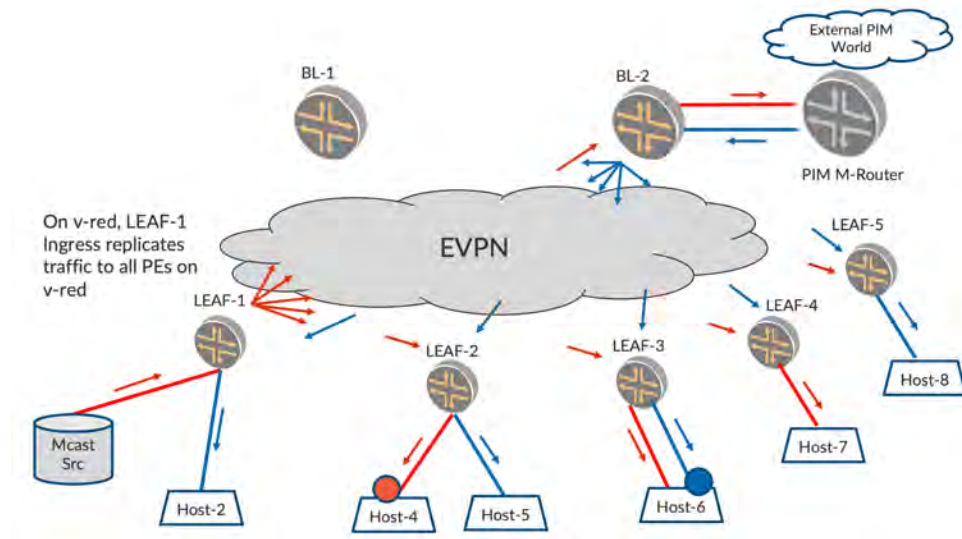


Figure 9.1 Inter-subnet Multicast with External Multicast Router

Bridge Domains (BDs) Not Everywhere

It is worth noting that the LEAF devices need not host all the VLANs in the fabric. For example, LEAF-1 can host VLANs 1 to 50, while LEAF-2 can Host VLANs 20 to 70, and so on. Still, multicast routing will work between the source-VLAN to the listener VLANs as long as the BL-2 and the PIM M-Router host all the VLANs in the fabric.

In Figure 9.1, the multicast source is on v-red behind LEAF-1. There are hosts Host-2, Host-4, Host-5, and Host-6 that have v-blue. Also, hosts Host-4, Host-6, and Host-7 have v-red. It can be observed that LEAF-1, LEAF-2, and LEAF-3 host both VLANs, while LEAF-4 and LEAF-5 do not host all the VLANs..

No Source or Listeners Started

If there are not listeners and sources in any of the VLANs, there is nothing to do.

Source Alone Starting Up On v-red

There are no listeners anywhere in the fabric. This is similar to what we have seen when the traffic on v-red is flooded to all the PEs and the access interfaces on v-red. Also, this traffic is flooded by BL-2 towards M-Router.

Listeners alone coming up

Say there is no traffic started and there are listeners coming up for group G1 on v-red from Host-4 and on v-blue from Host-6 (shown in Figure 9.1 by circles). These reports are sent across the EVPN core and reach the BL-2, are sent to the PIM M-Router, and create PIM states on this M-Router device for (*,G1) on both VLANs, blue and red.

Source started on v-red and listeners on v-red and v-blue

By virtue of listener interest for group G1 on v-red from Host-4 and on v-blue from Host-6, PIM (*,G1) states are created on the PIM M-Router. When the source starts sending traffic on v-red, this traffic reaches the PIM M-Router device on v-red. This PIM M-Router device routes the multicast traffic from the v-red to v-blue.

Post routing, the M-Router device forwards the packet on v-blue. This traffic reaches BL-2 and gets flooded towards the EVPN core on v-blue. The LEAFs that host v-blue receive the traffic and flood it to the listeners; LEAF-1, LEAF-2, LEAF-3, and LEAF-5 flood the traffic received on v-blue to Host-2, Host-5, Host-6, and Host-8, respectively. It is worth noting that the LEAF devices flood the traffic to hosts irrespective of active listener interest.

If we have several VLANs (say, v-orange, v-green, etc.), and there is at least one listener in each of these VLANs, the traffic will get routed by the M-Router from v-red onto all the listener VLANs. Also, post routing, the routed multicast traffic will be forwarded towards the EVPN core onto all the listener VLANs.

Flooding and Multicast Traffic Hair-pinning

Let's take a closer look at LEAF-1.

It can be observed that LEAF-1 floods the traffic on v-red. This traffic goes to the M-Router and gets routed onto v-blue and returns to LEAF-1. Then LEAF-1 floods the traffic to v-blue access interfaces. The multicast traffic has to hop to the periphery of the data center fabric to get routed and has to return via the fabric to the interested listeners.

This is a characteristic of CRB-Multicast. The main reason for deployment of CRB-Multicast routing is that since the central device (BL-2) performs the Multicast routing, the LEAFs are *not* required to host all the VLANs.

Also, since Snooping/Selective forwarding is turned off, LEAF-2 and LEAF-5 forward traffic to Host-5 and Host-8 on v-blue, even though there is no listener interest for group G1. Also, BL-2 forwards traffic to all LEAFs irrespective of the presence of listener interest behind them.

It is worth noting that with optimizations in multicast, the ‘*flood everywhere*’ aspect of the problem can be addressed. Overall, with optimizations, the following are possible with AR, SMET forwarding, and IGMP-snooping:

- the core-bandwidth utilization can be conserved on each VLAN
- the link bandwidth utilization can be conserved on each VLAN
- the access side link utilization can be conserved on each VLAN
- the replication load on LEAF and BLs is reduced for each VLAN

However, the issue of inter-subnet multicast traffic hopping all up until M-Router and returning back to the LEAFs (‘*hair-pinning*’) cannot be alleviated with the optimizations.

It may make sense here to explain that the alternative to the ‘hair-pinning’ problem is ERB Multicast. However, it has several nuances that have to be taken into consideration when it comes to scenarios of BDs-not-everywhere, External Multicast, and the need for PIM and IRB on all participating devices.

Since with CRB-Multicast, (a) BDs-not-everywhere is allowed, (b) PIM is required on only one device, and (c) External Multicast works seamlessly and uses traditional PIM procedures, CRB-Multicast is preferred for deployments today. Also, since the ‘*flood everywhere*’ problem is mitigated significantly with the optimizations in Part I of this book, applications today can be effectively run with CRB-Multicast.

Inter-subnet Multicast with PIM and IRB On Spines

In an earlier section, we explored how inter-subnet multicast is achieved with the use of an external M-Router placed in the periphery of the EVPN data center fabric. In this section, let’s explore how we can have L3-routing PIM on IRB functionality on BL devices within the fabric (see Figure 9.2). This approach is preferred over using M-Router, because the need for an external M-Router is obviated and existing participating EVPN BLs take up the role of L3-PIM multicast routing. Also, the number of hops the packets are subjected to is reduced. Additional enhancements for optimization become possible with this model.

In Figure 9.2 the PIM L3-routing functionality can be performed on BL-2 itself, instead of a PIM M-Router. For this to work, BL-2 has to host all the VLANs in the fabric and has to be enabled with PIM on IRBs for all the participating VLANs.

The routing of the multicast traffic from v-red to v-blue will be performed by BL-2 as BL-2 is the PIM DR for v-blue. Post routing, BL-2 will flood the traffic on v-blue to all LEAFs, and in turn to the access interfaces LEAFs that are on v-blue. The traffic forwarding on LEAFs is similar that discussed earlier in this chapter.

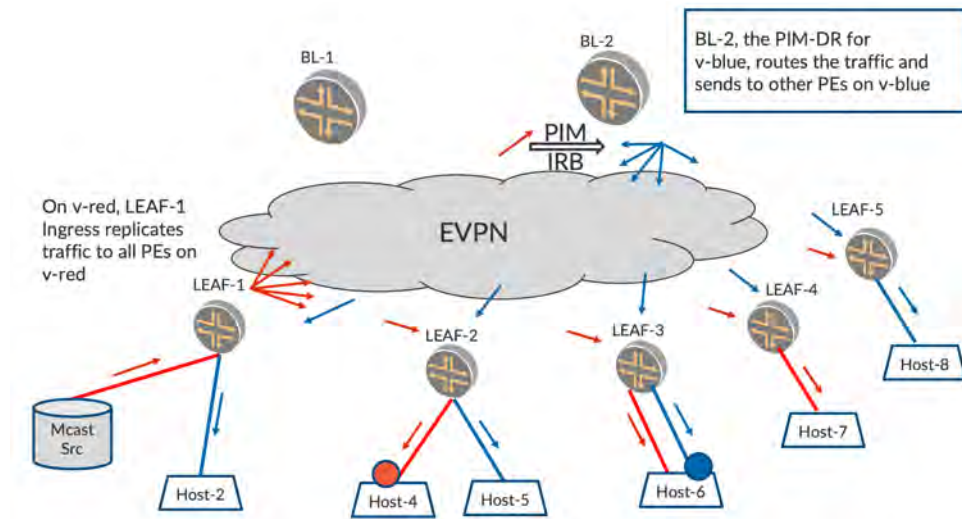


Figure 9.2 Inter-subnet Multicast with BL Running PIM on IRB

The earlier section related to Flooding and Traffic Hair-pinning is applicable in this scheme as well. For example, the traffic on v-red has to reach BL-2, get routed onto other listener VLANs, and return to the LEAFs.

An IRB Multicast Primer

When a multicast L2-frame reaches an EVPN device on a VLAN that has IRB enabled on that VLAN, the multicast L2-frame is L2-switched to other access interfaces on the EVPN device on that source-VLAN.

In addition, the L2-headers are removed and the L3-packet inside is punted to the L3-component of the IRB interface of that VLAN. This will make it appear as if the L3-packet has reached the IRB interface on the EVPN device. From this point on, the L3 procedures will apply on the packet that arrived on IRB interface.

Consider Figure 9.2 where there are listeners on v-blue at Host-6. When these IGMP reports (multicast L2-frames) reach SPINE-2, the L3-IGMP report will be punted to IRB interface on v-blue on BL-2. This will lead to IGMP and PIM (*,G1) state creation on IRB.blue on BL-2.

When Mcast Src starts sending traffic on v-red, this traffic reaches BL-2 and is punted to IRB.red. Now PIM L3-multicast routing occurs on BL-2, therefore traffic from IRB.red is routed to IRB.blue using PIM. Post routing into IRB.blue, the traffic is Ingress replicated on v-blue to all EVPN PEs that participate in v-blue.

Thus the multicast traffic reaches the LEAF devices on v-blue and the LEAF devices in turn flood the routed packets on v-blue to all the access interfaces.

To achieve this, PIM needs to be enabled on IRB interfaces on the BL devices. IRB interfaces are the representative interfaces of a L2-bridge domain. Typically, there is a VLAN/bridge-domain and it has multiple L2-interfaces associated with the BD. When traffic arrives on the L2-interface for that BD, traditional L2-switching occurs on that VLAN. In addition, L3-layer functionality of unicast route lookup or multicast routing with PIM occurs at the IRB interface.

All PIM-related procedures and protocols are applicable on IRB interfaces. So BL-1 and BL-2 see each other as PIM neighbors on each of the IRB interfaces (VLANs). By virtue of EVPN emulating a LAN, the IRB interface acts as the L3-interface towards the LAN for the spine devices. PIM Joins/Hellos are sent and received over the IRBs and PIM DRs are also elected.

Inter-subnet Multicast with Multiple VLANs

In earlier sections we described how Inter-subnet multicast works between two VLANs. Typically, in multicast applications, there will be several VLANs where there are interested listeners for the group G1. When the traffic for the group arrives, this source-traffic is routed onto all VLANs using multicast IRB procedures and forwarded onto the listener IRBs and Ingress replicated on all the interested VLANs.

If a single BL takes the load for routing traffic onto all VLANs, then that BL will be overloaded. It is desirable that the L3-routing load is shared across multiple BL devices.

Towards this end, two BLs are typically deployed for PIM L3-routing purposes. Classical PIM procedures are utilized to elect PIM DR for different VLANs. Therefore, for some set of VLANs, BL-1 will be the PIM DR while for other VLANs, BL-2 will be the PIM DR. The BL that is the non-DR (NDR) *on the listener IRB* will not route the traffic onto that VLAN.

Something to Remember About PIM DR-ship

PIM DR-ship is always relevant when building the outgoing interface list. The responsibilities of a PIM device that is a PIM DR on an IRB.600 are:

- When traffic arrives on that IRB.600, register the source to the PIM-RP.
- When an IGMP report is received on IRB.600 for group G1, create a PIM (*,G1) Join state on IRB.600.
- When traffic arrives on any other IRB, say IRB.800, if there are listeners on IRB.600, then route the traffic from IRB.800 to IRB.600 and Ingress replicate the traffic on VLAN 600 to other PEs that Host VLAN-600.

PIM DR election, by default, is based on the IP address of the BL devices. The BL with the highest IP address on a particular IRB interface is elected as the DR. It may be that the IP address of one BL device is always higher on all IRBs. In this case, to effectively load-balance across VLANs, it is a good idea to configure the DR priority value under the IRB.

Configuring the DR priority value on an IRB interface will cause the DR-priority value to be carried in the PIM hello message. This DR-priority value will override the DR election based on IP address. This way, different L3-PIM BL devices can be configured as the DR for different IRBs.

An Example with Multiple Spines Sharing Load Using PIM DRship

Let's consider a topology like Figure 9.3, where there are three VLANs. VLAN v-red has a multicast source behind LEAF-1. There are two other VLANs, v-blue and v-green, where there are listeners behind some LEAFs. BL-1 is the PIM DR for v-red and v-blue, and BL-2 is the PIM DR for v-green.

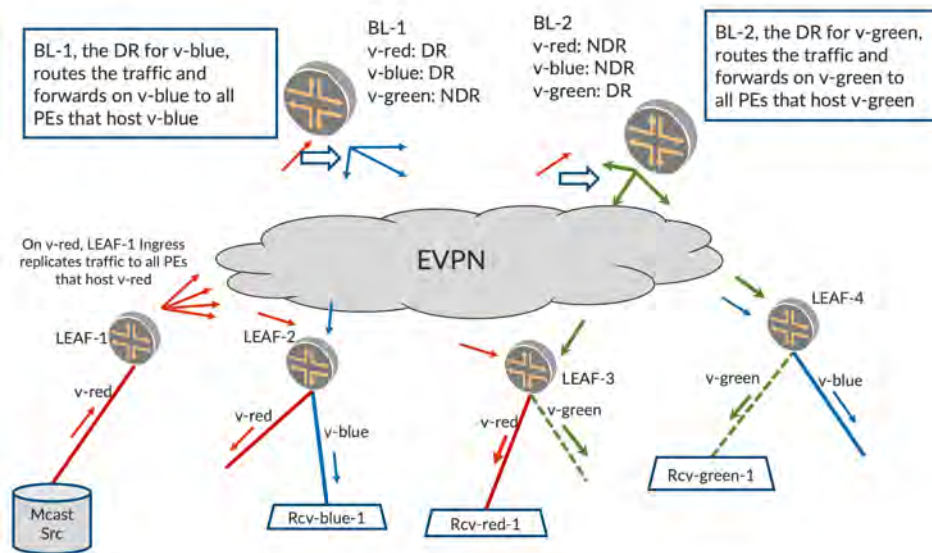


Figure 9.3 *Inter-subnet Multicast with Multiple VLANs*

When IGMP reports are received on v-blue, the BL-1 (PIM DR for v-blue) creates PIM (*,G) state for IRB.blue. Similarly, when IGMP reports are received on v-green, the BL-2 (PIM DR for v-green) creates PIM (*,G) state for IRB.green.

When Mcast Src sends traffic on IRB.red, the L2-switching based forwarding will ensure that traffic is forwarded everywhere on v-red. When traffic arrives on BL-1

on v-red (IRB.red), BL-1 will L3-route the traffic onto v-blue (since it is the DR for v-blue). Post routing, BL-1 will Ingress replicate the packet onto v-blue to all PEs that host v-blue. BL-1, being the non-DR for v-green will not route the packet onto v-green.

Similarly, when BL-2 receives the traffic on v-red (IRB.red), it will L3-route the traffic onto v-green (since it is the DR for v-green). Post routing, BL-2 will Ingress replicate the packet onto v-green to all PEs that Host v-green. BL-2, being the non-DR for v-blue, will not route the packet onto v-blue.

Thus multiple BLs are used to load share the role of PIM L3-routing in the fabric. In a fabric with, say, 100 VLANs, BL-1 may be PIM DR for VLANs 1-50 and BL-2 may be DR for VLANs 51-100. BL-1 routes the traffic onto 50 VLANs while BL-2 does so for the rest.

Flood-everywhere Problem Compounded with Inter-subnet Multicast

Kindly note that since there is no optimization in use so far, the L2-switched traffic is forwarded everywhere on the source-VLAN. On the receiving VLANs, as long as there is a *single* receiver existing, a PIM (*,G) join state gets created on the BL and this results in the BL routing the traffic onto that receiver VLAN.

This results in all the PEs that host the receiver-VLAN getting the traffic, and these PEs in turn forwarding the traffic onto the access interfaces irrespective of whether there are existing listeners or not. That is to say, the non-optimized multicast paradigm to flood everywhere in Intra-VLAN multicast is compounded in Inter-subnet multicast scenarios when the routed traffic is flooded everywhere, too. In later chapters we discuss how optimized multicast mitigates this flooding of traffic.

For now, let's briefly take a typical use case and describe the bandwidth calculations. Consider the following use case where optimizations are turned off:

- Number of LEAFs in Fabric: 'N' = 200
- Number of groups: 'G' = 20
- Traffic Rate: 'R' = 1 pps
- Number of VLANs 'M' = 500

The core-bandwidth utilization on the source-VLAN will be:

$$(N * G * R) = (200 * 20 * 1) = 4 \text{ Kbps}$$

The routed multicast traffic is sent to all the VLANs. The core-bandwidth utilization of the Fabric will be:

$$(N * G * R * M) = (200 * 20 * 1 * 500) = 2 \text{ Mbps}$$

This is increasingly referred to as the $(M * N)$ replication problem where 'M' is the number of VLANs and 'N' is the number of LEAFs that the traffic has to be replicated to. Without optimization, on each of the 'M' VLANs the traffic is flooded to each of the 'N' LEAFs. Also, on each LEAF, traffic is flooded onto the access interfaces on each of the 'M' VLANs.

Chapter Summary

This chapter explored the mechanism of Inter-VLAN multicast in an EVPN data center fabric. We turned OFF the optimizations to illustrate how the problems with non-optimized multicast get compounded with multiple VLANs in the fabric. In the next chapter, we illustrate how the optimizations play a significant role in mitigating the *'flood everywhere'* problem.

Chapter 10

EVPN Inter-VLAN Multicast Routing with Optimization

Chapter 9 explored the principles of Inter-VLAN multicast routing in an EVPN DC Fabric with PIM, using external M-Router and with PIM enabled on IRB on the BL devices. It also explored the characteristics of Inter-VLAN multicast in the absence of the optimizations like IGMP-Snooping, Selective (SMET) Forwarding, and Assisted Replication. In this chapter, we explore what happens if we turn ON these optimizations.

With optimizations, the basic L3-routing procedures for Inter-subnet Multicast do not change. Post routing the routed multicast traffic is selectively forwarded to only those PEs that expressed listener interest for that group. The optimized Intra-VLAN multicast routing procedures are followed for each of the receiver VLANs.

Inter-VLAN Multicast with Selective (SMET) Forwarding

In Figure 10.1, LEAF-2 originates Type-6 on v-red and LEAF-3 originates Type-6 on v-blue based on received IGMP reports. BL-2, which is the PIM-DR for v-blue, routes the traffic from v-red onto v-blue. Post routing, BL-2 L2-forwards the traffic on v-blue to only LEAF-3 and BL-1 per Optimized Intra-subnet Multicast procedures described in Chapter 6. The LEAFs that do not have listeners for v-blue, LEAF-1, LEAF-2, and LEAF-5 are spared from receiving routed multicast traffic for v-blue.

Inter-VLAN Multicast with Multiple Listener VLANs

In an EVPN fabric, there will be several receiver VLANs where the traffic has to be routed. Selective Multicast Forwarding (SMET) can be appreciated in a scenario where there are several VLANs with Inter-VLAN multicast forwarding. Consider

Figure 10.2 where there are two receiver VLANs, namely v-blue and v-green. The number of VLANs is in the order of hundreds. And let's say BL-1 is the PIM-DR for blue and BL-2 is the PIM-DR for green.

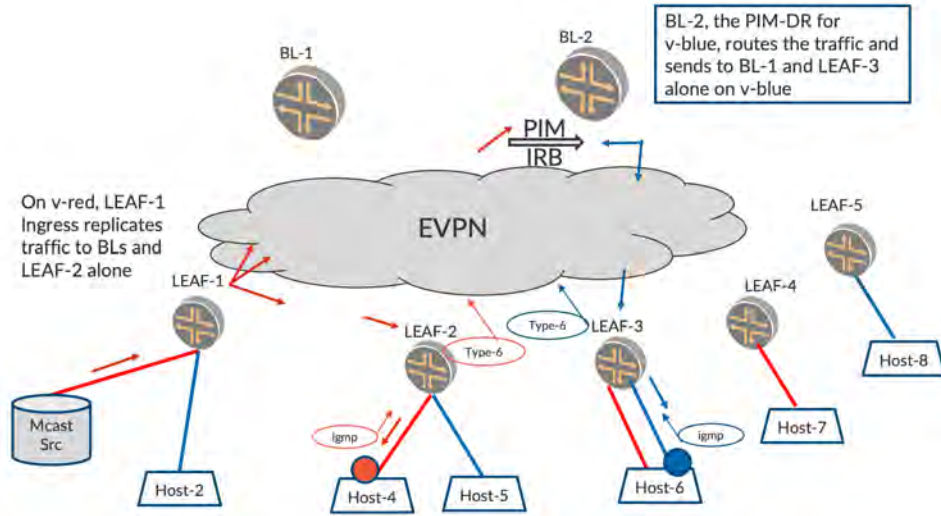


Figure 10.1 Inter-subnet Multicast with BL Running PIN on IRB with SMET

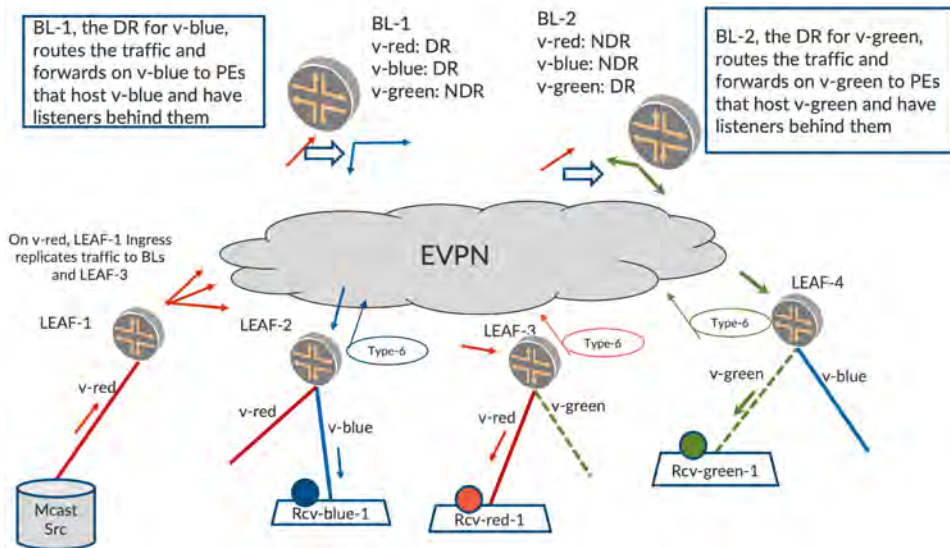


Figure 10.2 Inter-subnet Multicast with Multiple VLANs (with SMET Forwarding)

With optimized multicast enabled, LEAF-2 will originate Type-6 (v-blue) and LEAF-4 will originate Type-6 (v-green). Also, LEAF-3 will originate Type-6 (v-red). When the source starts sending traffic on v-red, LEAF-1 will L2-forward the traffic to the two BLs and to LEAF-3 on v-red. LEAF-3 will forward to its access Host Rcv-red-1.

When BL-1 receives the traffic on v-red it will route the traffic onto v-blue since it is PIM-DR on v-blue. Post routing on v-blue will L2-forward the traffic to the other BL and LEAF-2 per Intra-subnet selective multicast procedures. LEAF-2 will forward the traffic on v-blue to its access interface to Rcv-blue-1.

Similarly, when BL-2 receives the traffic on v-red it will route the traffic onto v-green since it is the PIM-DR on v-green. Post routing on v-green will L2-forward the traffic to the other BL and to LEAF-4 per selective multicast procedures. LEAF-4 will forward the traffic on its access interface to Rcv-green-1.

BL-1 will receive the same traffic from BL-2 on irb.v-green. This traffic will be dropped due to IIF-Mismatch since BL-1's RPF to source is over irb.v-red.

Thus, traffic is sent only to the required PEs and only onto those access interfaces on the receiver VLANs, thus conserving bandwidth and resources on each receiver VLAN.

Inter-VLAN Multicast with AR and SMET

Now let's use a typical EVPN data center fabric illustrated in Figure 10.3, where the BL devices are enabled with PIM on IRBs. Though BL L3-PIM devices are high-end, when several VLANs and multiple Ingress Replications have to be performed, to send to several LEAFs in the fabric there is a possibility of the BLs getting overwhelmed. Even if the BL devices held up well, in terms of replication, the link between the BLs and the Lean-Spine may not be able to carry the multiple replicated copies.

In addition to SMET, enabling AR in the fabric brings in additional gains in transferring the replication load from BL devices to the AR-Replicator and also reducing the bandwidth on the link between the BL and the AR-R.

In Figure 10.3, LEAF-2 and LEAF-4 have listener interest on v-blue while LEAF-2 and LEAF-5 have listener interest on v-green. Say BL-1 is the PIM DR for v-blue and BL-2 is PIM DR for v-green, and we have AR plus SMET optimization enabled in the data center fabric.

When the source sends traffic on v-red, LEAF-1, being the AR-LEAF, sends only a copy of the packet to the AR-Replicator, for example, AR-1. Now AR-1 replicates this packet to the BLs since BLs are running PIM, and resets the bit in Multicast Flags Community (MF-COM). Please see Chapter 6.

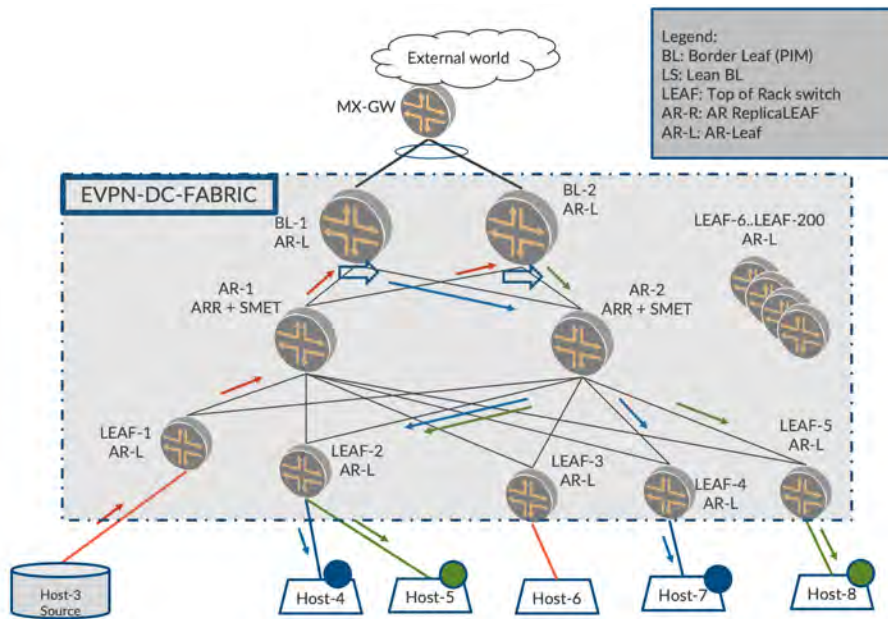


Figure 10.3

Inter-VLAN Multicast with AR Plus SMET

BL-1, being a PIM enabled device and the PIM DR on v-blue, routes the traffic from v-red to v-blue. Typically, without AR, BL would have replicated two copies to be sent to LEAF-2 and LEAF-4. Also, the link between BL-1 and AR-2 would be carrying two copies of the same packet. With AR, BL-1, being the AR-LEAF, sends only one packet to one of the AR-Replicators, say AR-2. Now AR-2 replicates the packet on v-blue to LEAF-2 and LEAF-4.

Similarly, BL-2, the PIM DR on v-green, routes the packet from v-red to v-green. Since BL-2 is an AR-LEAF, instead of making multiple copies it sends one copy to the replicator AR-2. AR-2 duly replicates and sends the packet on v-green to LEAF-2 and LEAF-5.

LEAF-3 and the rest of the LEAFs (LEAF-6 to LEAF-200) are spared from receiving any traffic. Needless to say, the access interfaces on the LEAFs that don't have listeners are also spared of traffic.

Thus, AR and SMET, when used in an Inter-VLAN multicast environment bring in the following benefits on *each of the listener VLANs*:

- Core bandwidth conservation
- Reduced Replication Load on LEAF and BLs
- Selective Forwarding by the LEAFs, BLs, and the AR-Replicators

- Reduced Link-utilization between LEAF/BL and the Lean Spines
- Access side bandwidth conservation
- Reduced load of processing unnecessary multicast traffic received from the core

In a scaled setup, such schemes of optimization help in reducing the load on packet replication, processing, and also the core-bandwidth utilization. Thus, AR plus SMET brings to EVPN data center fabric the equivalent of traditional selective P2MP multicast.

There's also load sharing at various levels:

- Leaf-layer: The LEAF devices pick one of the replicators for replication for different flows. By virtue of this, the multiple underlay links between the LEAF and the replicators are shared equitably.
- BL-Layer: Different BL devices are PIM DRs for different VLANs. Thus the load of routing for different VLANs is shared amongst the BL devices.

When the LEAF devices are multihomed only the DF forwards, thus conserving the access bandwidth. The DF election is thus performed so that different LEAF devices among a multihomed set become DFs for different VLANs.

Optimized *vis-à-vis* Non-optimized Inter-VLAN Mcast

Let's make some comparisons similar to how we calculated in Chapter 8. Consider a case where there are 200 LEAFs in a data center fabric. Say there is high volume of multicast traffic for 20 groups, each group has traffic rate of 1 Mbps, and there are 10 LEAFs in the fabric interested in each group in each VLAN. Further, assume there are 500 VLANs in the data center fabric. Let's characterize the behavior with each mechanism.

- Number of LEAFs in Fabric: $N = 200$
- Number of groups: $G = 20$
- Traffic Rate: $R = 1$ Mbps
- Number of LEAFs interested in traffic per VLAN per group: $T = 10$
- Number of VLANs in Fabric $M = 500$

Non-optimized Multicast

Core bandwidth consumption:

$$(N * G * R * M) = (200 * 20 * 1 * 500) = 2000 \text{ Gbps}$$

Replication Load on BL:

$$(N * G * M) = 200 * 20 * 500 = 2M \text{ times}$$

Link bandwidth consumption between BL and Lean-Spine:
 $(N * G * R * M) = (200 * 20 * 1 * 500) = 2000 \text{ Gbps}$

Assisted Replication

Core bandwidth consumption:
 $(N * G * R * M) = (200 * 20 * 1 * 500) = 2000 \text{ Gbps}$

Replication Load on BL:
 $(1 * G * M) = 1 * 20 * 500 = 10\text{K times}$

Link bandwidth consumption between BL and Lean-Spine:
 $(1 * G * R * M) = (1 * 20 * 1 * 500) = 10 \text{ Gbps}$

Optimized Multicast (SMET Forwarding) without AR

Core bandwidth consumption:
 $(T * G * R * M) = (10 * 20 * 1 * 500) = 100 \text{ Gbps}$

Replication Load on BL for each packet received from core:
 $(T * G * M) = (10 * 20 * 500) = 100\text{K times}$

Link bandwidth consumption between LEAF and Lean-Spine:
 $(T * G * R * M) = (10 * 20 * 1 * 500) = 100,000 = 100 \text{ Gbps}$

AR + SMET

Core bandwidth consumption:
 $(T * G * R * M) = (10 * 20 * 1 * 500) = 100 \text{ Gbps}$

Replication Load on BL for each packet received from core:
 $(1 * G * M) = (1 * 20 * 500) = 10\text{K times}$

Link bandwidth consumption between BL and Lean-Spine:
 $(1 * G * R * M) = (1 * 20 * 1 * 500) = 10 \text{ Gbps}$

With AR and SMET you can see that the overall core bandwidth consumption is significantly reduced. Also, the utilization between BL and the Lean Spine device is considerably reduced. See Table 10.1 and Figures 10.4 and 10.5. Also, the replication load on BLs is reduced.

Table 10.1 *Bandwidth Consumption*

Number of VLANs in the fabric: 500	
Number of TORs in the fabric: N = 200	Number of Groups: G = 20
Number of TORs interested in the fabric: T = 10	Traffic Rate: R = 1 Mbps

Inter-VLAN Multicast	Non-Optimized Multicast	AR	SMET	AR + SMET	Gain Factor: AR+SMET vis-a-vis Non-optimized
Core Bandwidth consumption (in Gbps)	2000	2000	100	100	20
Replication Load on TOR hosting the source	2000K	10K	100K	10K	200
Link Bandwidth consumption between TOR and Lean Spine (in Gbps)	2000	10	100	10	200

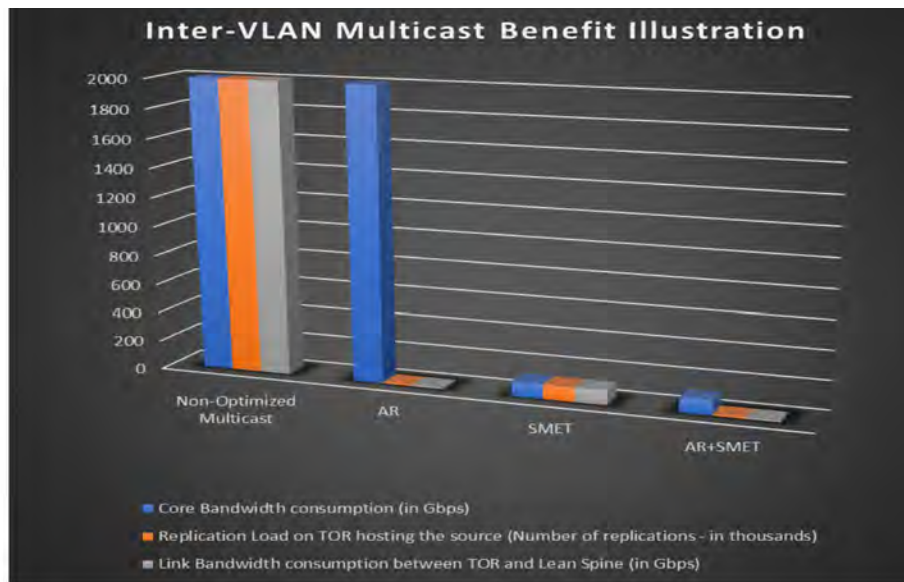


Figure 10.4

Inter-VLAN Multicast

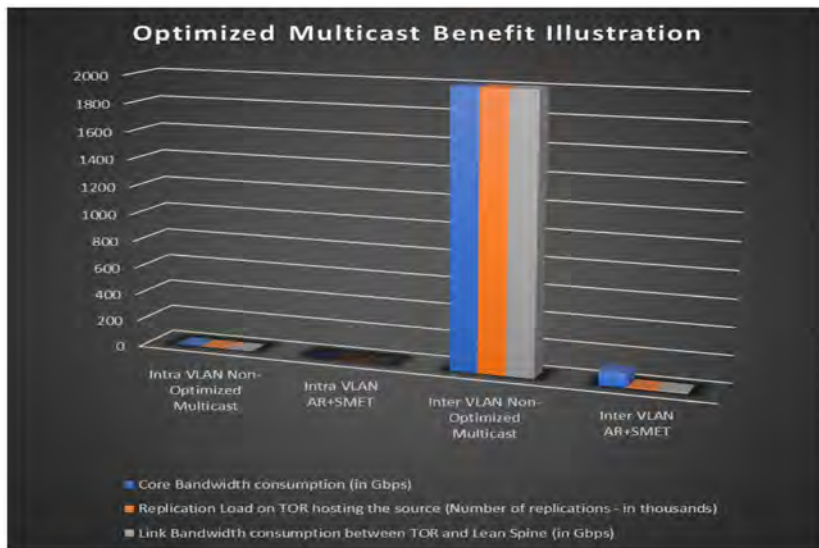


Figure 10.5 Optimized Multicast Benefits

Chapter Summary

This chapter has explored Inter-VLAN Multicast behavior when deployed with optimization. Inter-VLAN multicast, per se, has no procedural change. However, with the optimization schemes, the benefits accrue since these are applicable on each of the listener VLANs where traffic is routed. We also explored the different benefits quantitatively in a typical use case and illustrated how all of the optimization techniques from Part I of the book are playing a crucial role in Inter-VLAN multicast.

So having covered both Intra-VLAN and Inter-VLAN multicast within a data center fabric, it's time to explore the different ways this fabric can be connected to the outside world and multicast traffic can be sent/received from/to the fabric.

Configuration and Verification

So far our sources and receivers have all been on the same VLAN, VLAN-101. Having explored inter-VLAN multicast, we are now ready to look at the multicast behavior when we have the source(s) and receiver(s) within the DC but on different VLANs.

Before beginning, let's stop all sources and receivers that were previously started.

Configuration

Though we have, so far, focused solely on VLAN-101, our base configuration has always included VLAN-102. So the configurations done in Chapter 9 are sufficient for this section and we can move on directly to traffic verification. Remember also that our configurations already include SMET and AR optimizations.

Traffic Verification

As before, start sending multicast traffic from Host-1 at 10 pps (packets per second) for group 225.1.1.1 in VLAN-101.

On Host-6 and Host-3, start receivers for the multicast group, 225.1.1.1, but this time on both VLAN-101 and VLAN-102.

From the RT statistics, you can see that the traffic sent by Host-1 at 10 pps is now received by the interested receivers, Host-6 and Host-3, and the legacy device, Host-7, in both VLAN-101 and VLAN-102, thereby resulting in 20 pps of incoming traffic on each of them.

	Stat Name	Port Name	Link State	Frames Tx. Rate	Valid Frames Rx. Rate
# 1	10.216.45.202/Card20/Port01	HOST-1	Link Up	10	0
2	10.216.45.202/Card03/Port01	HOST-2	Link Up	0	0
3	10.216.45.202/Card20/Port02	HOST-3	Link Up	0	20
4	10.216.45.202/Card03/Port02	HOST-4	Link Up	0	0
5	10.216.45.202/Card20/Port03	HOST-5	Link Up	0	0
6	10.216.45.202/Card03/Port03	HOST-6	Link Up	0	20
7	10.216.45.202/Card03/Port04	HOST-7	Link Up	0	20
8	10.216.45.202/Card20/Port04	HOST-8	Link Up	0	0

Figure 10.6

RT Stats

Multicast Traffic Outputs - LEAF-1, LEAF-2, LEAF-4, LEAF-5, SPINE-2 (VLAN-101)

The traffic forwarding behavior for the multicast traffic arriving in VLAN-101 will be same as before on these devices. So this flow contributes to 10 pps of the traffic seen on Host-2, Host-6, and Host-8.

Note that the traffic is also replicated by SPINE-2 to BL-1 and BL-2. We have so far postponed looking at the traffic forwarding behavior on these devices. Now that inter VLAN procedures are understood, let's see what happens to the traffic on these devices.

BL-1, BL-2

VLAN-101 does not have any access interfaces on the border-LEAF devices, hence there is no further switching of the multicast traffic on either one. However, a multicast (PIM+IGMP) enabled IRB is associated with the VLAN on both these devices.

Since BL-2 has a higher IP address, it is elected as the PIM-DR on irb.101 (VLAN-101).

On irb.102, since we have configured an explicit DR-priority on BL-1, in spite of having lower IP, it is elected as PIM DR on irb.102 (VLAN-102):

```
lab@BL-1> show pim interfaces instance VRF-1
Stat = Status, V = Version, NbrCnt = Neighbor Count,
S = Sparse, D = Dense, B = Bidirectional,
DR = Designated Router, DDR = Dual DR, DistDR = Distributed DR,
P2P = Point-to-point link, P2MP = Point-to-Multipoint,
Active = Bidirectional is active, NotCap = Not Bidirectional Capable
Name                Stat Mode IP V State          NbrCnt JoinCnt(sg/*g) DR address
irb.101              Up   S    4 2 NotDR,NotCap    1 1/0      18.18.18.2
irb.102             Up   S    4 2 DR,NotCap     1 0/0      19.19.19.1
lo0.1                Up   S    4 2 DR,NotCap        0 0/0      101.101.101.102
pime.32770           Up   S    4 2 P2P,NotCap      0 0/0
```

```
lab@BL-2> show pim interfaces instance VRF-1
Stat = Status, V = Version, NbrCnt = Neighbor Count,
S = Sparse, D = Dense, B = Bidirectional,
DR = Designated Router, DDR = Dual DR, DistDR = Distributed DR,
P2P = Point-to-point link, P2MP = Point-to-Multipoint,
Active = Bidirectional is active, NotCap = Not Bidirectional Capable
Name                Stat Mode IP V State          NbrCnt JoinCnt(sg/*g) DR address
irb.101              Up   S    4 2 DR,NotCap        1 1/0      18.18.18.2
irb.102             Up   S    4 2 NotDR,NotCap 1 0/0      19.19.19.1
lo0.1                Up   S    4 2 DR,NotCap        0 0/0      102.102.102.103
pime.32769           Up   S    4 2 P2P,NotCap      0 0/0
```

Multicast Traffic Outputs – BL-1, BL-2

Since there is receiver interest on irb.102, BL-1 being the PIM DR on irb.102, routes the traffic arriving on irb.101 into irb.102 (i.e. VLAN-102):

```
lab@BL-1> show multicast route extensive instance VRF-1
Instance: VRF-1 Family: INET
Group: 225.1.1.1
  Source: 18.18.18.30/32
  Upstream interface: irb.101
  Downstream interface list:
    irb.102
  Number of outgoing interfaces: 1
  Session description: Unknown
  Statistics: 1 kbps, 10 pps, 917043 packets
  Next-hop ID: 131102
  Upstream protocol: PIM
  Route state: Active
```

```
Forwarding state: Forwarding
Cache lifetime/timeout: 360 seconds
Wrong incoming interface notifications: 33
Uptime: 00:03:15
```

BL-1 then forwards the traffic routed into VLAN-102 towards one of the AR-Replicator Spines: in this case SPINE-1 (AR-IP = 103.103.103.113).

In addition, the traffic is also sent on the VTEP towards the BL-2 (102.102.102.102) since it is multihomed to BL-1 (refer to Chapter 5: Enhanced-AR Forwarding Rules):

```
lab@BL-1> show interfaces vtep extensive | grep "Output packets.*pps|VXLAN Endpoint Address"
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 102.102.102.102...
Output packets: 4194 10 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 105.105.105.105...
Output packets: 57 0 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 106.106.106.106...
Output packets: 57 0 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 107.107.107.107...
Output packets: 0 0 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 108.108.108.108...
Output packets: 55 0 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 109.109.109.109...
Output packets: 57 0 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 103.103.103.113...
Output packets: 4212 9 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 103.103.103.103...
Output packets: 1 0 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 104.104.104.104...
Output packets: 0 0 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 104.104.104.114...
Output packets: 3996 0 pps
```

Multicast Traffic Outputs - SPINE-1 (VLAN-102)

SPINE-1 selectively replicates the AR-tunnel traffic received from BL-1 in VLAN-102 to LEAF-1(105.105.105.105), LEAF-2(106.106.106.106), and LEAF-4(108.108.108.108) which have interested receivers, and legacy device LEAF-5(109.109.109.109):

```
lab@SPINE-1> show interfaces vtep extensive | grep "Output.*packets.*pps|VXLAN Endpoint Type"
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 108.108.108.108...
Output packets: 18136 9 pps
VXLAN Endpoint Type: AR Remote, VXLAN Endpoint Address: 108.108.108.108...
Output packets: 0 0 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 101.101.101.101...
Output packets: 5601 0 pps
VXLAN Endpoint Type: AR Remote, VXLAN Endpoint Address: 101.101.101.101...
Output packets: 0 0 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 102.102.102.102...
Output packets: 3525 0 pps
VXLAN Endpoint Type: AR Remote, VXLAN Endpoint Address: 102.102.102.102...
Output packets: 0 0 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 109.109.109.109...
```

```
Output packets:          18137          10 pps
VXLAN Endpoint Type: AR Remote, VXLAN Endpoint Address: 109.109.109.109...
Output packets:          0          0 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 105.105.105.105...
Output packets:          10984          10 pps
VXLAN Endpoint Type: AR Remote, VXLAN Endpoint Address: 105.105.105.105...
Output packets:          0          0 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 106.106.106.106...
Output packets:          10980          10 pps
VXLAN Endpoint Type: AR Remote, VXLAN Endpoint Address: 106.106.106.106...
Output packets:          0          0 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 107.107.107.107...
Output packets:          1614          0 pps
VXLAN Endpoint Type: AR Remote, VXLAN Endpoint Address: 107.107.107.107...
Output packets:          0          0 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 104.104.104.104...
Output packets:          10536          0 pps
VXLAN Endpoint Type: AR Remote, VXLAN Endpoint Address: 104.104.104.104...
Output packets:          0          0 pps
VXLAN Endpoint Type: Remote, VXLAN Endpoint Address: 104.104.104.114...
Output packets:          0          0 pps
```

Multicast Traffic Outputs - LEAF-1 (VLAN-102)

LEAF-1 receives 10 pps of multicast traffic from SPINE-1 in VLAN-102.

Though LEAF-1 has learned an interested IGMP receiver in VLAN-102 on its access interface ae1.0, CLASSICAL-DF-NDF rules prevent LEAF-1 from forwarding the multicast traffic on this interface.

The other access interfaces, ae0.0 and xe-0/0/4.0, do not have any receiver interest and IGMP-snooping procedures ensure that the traffic is not forwarded on these interfaces:

```
lab@LEAF-1> monitor interface traffic detail
Interface  Link  Input packets      (pps)  Output packets      (pps)  Description
...
xe-0/0/4   Up    0                  (0)    2564                 (0)    T0 Host-2
...
ae0        Up    16487              (10)   0                    (0)    T0 CE-1
ae1        Up    0                  (0)    9260                 (10)   T0 CE-2
...
```

So LEAF-1 effectively forwards only 10 pps in VLAN-101 to Host-3.

Multicast Traffic Outputs – LEAF-2 (VLAN-102)

LEAF-2 receives 10 pps of multicast traffic from SPINE-1 in VLAN-102.

Following CLASSICAL-DF-NDF rules and IGMP-snooping procedures on VLAN-102, LEAF-2 being the DF, forwards the multicast traffic to the interested IGMP receiver on its access interface ae1.0.

The other access interfaces, ae0.0 and xe-0/0/4.0, do not have any receiver interest and IGMP-snooping procedures ensure that the traffic is not forwarded on these interfaces:

```
lab@LEAF-2> monitor interface traffic detail
Interface  Link  Input packets      (pps)  Output packets      (pps)  Description
...
xe-0/0/4    Up      0      (0)      0      (0)  T0 Host-4
...
ae0         Up      7      (0)      0      (0)  T0 CE-1
ae1         Up     16      (0)    9561    (10) T0 CE-2
...
```

LEAF-2 forwards 10 pps in VLAN-102 to Host-3. Thus LEAF-1 and LEAF-2 together account for the total 20 pps traffic seen on Host-3.

Multicast Traffic Outputs – LEAF-4 (VLAN-102)

LEAF-4 receives 10 pps of multicast traffic from SPINE-1 in VLAN-102.

The access side IGMP-snooping functionality ensures that the multicast traffic arriving on LEAF-4 on VLAN-102 is forwarded on the single-homed interface xe-0/0/3.0 that has a receiver, but not on the multihomed interface ae0.0 that does not have a receiver.

Recall that traffic for VLAN-101 was also being forwarded on the interested single-homed interface xe-0/0/3.0. Thus, these together account for the 20 pps traffic seen egressing xe-0/0/3.0 and received on Host-6.

```
lab@LEAF-2> monitor interface traffic detail
Interface  Link  Input packets      (pps)  Output packets      (pps)  Description
...
xe-0/0/3    Up      0      (0)    8283    (20) T0 Host-6
ae0         Up      7      (0)      0      (0)  T0 CE-3
...
```

Multicast Traffic Outputs - LEAF-5

LEAF-5 being a legacy device receives the traffic in VLAN-102 and floods it on its access interface, xe-0/0/2.0, even though it does not have a receiver.

Recall that traffic for VLAN-101 was also being forwarded on interface xe-0/0/2.0. Thus these together account for the 20 pps traffic seen egressing xe-0/0/2.0 and received on Host-7:

```
lab@LEAF-5> monitor interface traffic detail
Interface  Link  Input packets      (pps)  Output packets      (pps)  Description
...
xe-0/0/2    Up      0      (0)    8666    (20) T0 Host-7
...
```

Detailed Control Plane Verification

So far we have focused on the L2 multicast states built on the PEs. For traffic to be routed, in addition to the L2 multicast states, the L3 multicast states also need to be set correctly up on the PEs that are configured with IRBs and are acting as multicast routers in our case these are the Border-Leaf devices, BL-1 and BL-2. Once the multicast traffic is routed at the Border-Leaf PEs and forwarded towards the LEAF PEs, the L2 switching of traffic on the LEAF PEs occurs using the intra-VLAN states similar to what we have already looked at in detail in earlier chapters.

Verification of Layer 3 IGMP State

Verify that on BL-1, in addition to the IGMP snooping proxy state learned on VLAN-101 and VLAN-102, the IGMP group membership from EVPN has been learned on the corresponding L3 interfaces, irb.101 and irb.102, as well:

```
lab@BL-1> show evpn igmp-snooping proxy
Instance: default-switch
  VN Identifier: 101
    Group IP: 225.1.1.1, Source IP: 0.0.0.0
  VN Identifier: 102
    Group IP: 225.1.1.1, Source IP: 0.0.0.0
```

```
lab@BL-1> show igmp group detail
Interface: irb.101, Groups: 1
  Group: 225.1.1.1
    Group mode: Exclude
    Source: 0.0.0.0
    Source timeout: 0
    Last reported by: Local
    Group timeout: 0 Type: EVPN
```

```
...
Interface: irb.102, Groups: 1
  Group: 225.1.1.1
    Group mode: Exclude
    Source: 0.0.0.0
    Source timeout: 0
    Last reported by: Local
    Group timeout: 0 Type: EVPN
```

...

Verify that the same states have also been learned on BL-2:

```
lab@BL-2> show evpn igmp-snooping proxy
Instance: default-switch
  VN Identifier: 101
    Group IP: 225.1.1.1, Source IP: 0.0.0.0
  VN Identifier: 102
    Group IP: 225.1.1.1, Source IP: 0.0.0.0
```

```
lab@BL-2> show igmp group detail
Interface: irb.102, Groups: 1
  Group: 225.1.1.1
    Group mode: Exclude
```

```

Source: 0.0.0.0
Source timeout: 0
Last reported by: Local
Group timeout: 0 Type: EVPN

```

...

```

Interface: irb.101, Groups: 1
  Group: 225.1.1.1
    Group mode: Exclude
    Source: 0.0.0.0
    Source timeout: 0
    Last reported by: Local
    Group timeout: 0 Type: EVPN

```

...

Verification of Layer 3 PIM State

Verify that on BL-1 and BL-2 PIM state has been created for the group:

```

lab@BL-1> show pim join extensive instance VRF-1
Instance: PIM.VRF-1 Family: INET
R = Rendezvous Point Tree, S = Sparse, W = Wildcard
Group: 225.1.1.1
  Source: *
  RP: 101.101.101.102
  Flags: sparse,rptree,wildcard
  Upstream interface: Local
  Upstream neighbor: Local
  Upstream state: Local RP
  Uptime: 00:05:44
  Downstream neighbors:
    Interface: irb.101
      18.18.18.1 State: Join Flags: SRW Timeout: Infinity
      Uptime: 00:05:44 Time since last Join: 00:05:44
    Interface: irb.102
      19.19.19.1 State: Join Flags: SRW Timeout: Infinity
      Uptime: 00:05:44 Time since last Join: 00:05:44
  Number of downstream interfaces: 2
  Number of downstream neighbors: 2
Group: 225.1.1.1
  Source: 18.18.18.30
  Flags: sparse,spt
  Upstream interface: irb.101
  Upstream neighbor: Direct
  Upstream state: Local Source, Local RP, No Prune to RP
  Keepalive timeout: 335
  Uptime: 00:20:58
  Downstream neighbors:
    Interface: irb.102
      19.19.19.2 State: Join Flags: S Timeout: Infinity
      Uptime: 00:05:53 Time since last Join: 00:05:53
  Number of downstream interfaces: 1
  Number of downstream neighbors: 1

```

```

lab@BL-2> show pim join extensive instance VRF-1
Instance: PIM.VRF-1 Family: INET
R = Rendezvous Point Tree, S = Sparse, W = Wildcard
Group: 225.1.1.1
  Source: *

```

```

RP: 102.102.102.103
Flags: sparse,rptree,wildcard
Upstream interface: Local
Upstream neighbor: Local
Upstream state: Local RP
Uptime: 00:05:53
Downstream neighbors:
  Interface: irb.102
    19.19.19.2 State: Join Flags: SRW Timeout: Infinity
    Uptime: 00:05:53 Time since last Join: 00:05:53
  Interface: irb.101
    18.18.18.2 State: Join Flags: SRW Timeout: Infinity
    Uptime: 00:05:53 Time since last Join: 00:05:53
Number of downstream interfaces: 2
Number of downstream neighbors: 2
Group: 225.1.1.1
Source: 18.18.18.30
Flags: sparse,spt
Upstream interface: irb.101
Upstream neighbor: Direct
Upstream state: Local Source, Local RP, No Prune to RP
Keepalive timeout: 328
Uptime: 00:21:07
Downstream neighbors:
  Interface: irb.101
    18.18.18.2 State: Join Flags: S Timeout: Infinity
    Uptime: 00:05:53 Time since last Join: 00:05:53
Number of downstream interfaces: 1
Number of downstream neighbors: 1

```

Verification of Layer 3 Multicast Forwarding State

Verify that BL-1 receives the traffic in the source VLAN-101 on irb.101, and being the PIM DR on the receiver VLAN, VLAN-102 (irb.102) routes the traffic into VLAN-102 via irb.102:

```

lab@BL-1> show multicast route extensive instance VRF-1
Instance: VRF-1 Family: INET
Group: 225.1.1.1
  Source: 18.18.18.30/32
  Upstream interface: irb.101
  Downstream interface list:
    irb.102
  Number of outgoing interfaces: 1
  Session description: Unknown
  Statistics: 1 kbps, 10 pps, 917043 packets
  Next-hop ID: 131102
  Upstream protocol: PIM
  Route state: Active
  Forwarding state: Forwarding
  Cache lifetime/timeout: 360 seconds
  Wrong incoming interface notifications: 33
  Uptime: 00:03:15

```

Verify that BL-2 receives the traffic in the source VLAN-101 on irb.101, but since it is not the PIM DR on the receiver VLAN, VLAN-102 (irb.102), it does not route the traffic into VLAN-102 via irb.102:

```

lab@BL-2> show multicast route extensive instance VRF-1
Instance: VRF-1 Family: INET
Group: 225.1.1.1
  Source: 18.18.18.30/32
  Upstream interface: irb.101
  Number of outgoing interfaces: 0
  Session description: Unknown
  Statistics: 1 kbps, 10 pps, 10575 packets
  Next-hop ID: 0
  Upstream protocol: PIM
  Route state: Active
  Forwarding state: Forwarding
  Cache lifetime/timeout: 360 seconds
  Wrong incoming interface notifications: 0
  Uptime: 00:21:07

```

Verification of Layer 2 Multicast Forwarding State in the Routed VLAN

On BL-1, the traffic thus routed into VLAN-102 via irb.102 will then be forwarded in VLAN-102 based on the L2 multicast forwarding state created for group 225.1.1.1, and source 18.18.18.30 in VLAN-102:

Since we are using AR, verify that the traffic is forwarded via the load balancing next hop to one of the AR-Rs:

```

lab@BL-1> show multicast snooping route extensive VLAN VLAN-102 group 225.1.1.1 source-
prefix 18.18.18.30
Group: 225.1.1.1/32
  Source: 18.18.18.30/32
  Vlan: VLAN-102
  Mesh-group: __all_ces__
    Downstream interface list:
      evpn-core-nh -(131092)
...
lab@BL-1> show evpn multicast-snooping next-hops 131092 detail
...
ID          Refcount KRefCount Downstream interface Addr
131092      3          1 131084
                               Flags 0x2100 type 0x18 members 0/0/1/0/0
                               Address 0xe531928
lab@BL-1> show evpn multicast-snooping assisted-replication next-hops index 131084
Instance: default-switch
AR Role: AR Leaf
  VN Identifier: 102
  Load Balance Nexthop Index: 131084
  Load balance to:
    Nexthop Index  Interface      AR IP
    1777           vtep.32777    103.103.103.113
    1794           vtep.32780    104.104.104.114

```

The AR-R receiving this traffic in VLAN-102, will selectively forward it in VLAN-102 to the interested PEs. This is plain intra-VLAN forwarding that we have discussed in previous chapters; the verification is left as an exercise for the reader.

Chapter 11

External Multicast with PIM IRB

Earlier chapters in this book explored how non-optimized and optimized inter-subnet multicast works in an EVPN data center fabric. This chapter looks into how external multicast (multicast from and to outside of EVPN data center fabric) works in an optimized way following PIM and EVPN procedures. Since optimization is a paradigm within the fabric, the procedures described for External Multicast are the same for fabric topologies whether multicast is optimized or not.

The EVPN data center fabric typically has two L3-PIM devices (BL-1 and BL-2) for inter-subnet multicast as described in Chapter 9. The two BL devices are deployed for redundancy purposes. It is commonplace that such an EVPN data center fabric is to be connected to the outside world to send and receive multicast traffic to and from the data center.

By the end of this chapter you should have a fair understanding of the different methods in which the data center fabric can be connected to the outside world and the procedures that are involved.

The following two approaches are considered best practice for external multicast deployments.

- Using classic L3-links (therefore, IP addresses configured on each link)
- Using L2-link and family bridge (using a MVLAN with EVPN multihomed procedures)

In this chapter, we build on the building blocks of PIM and EVPN multihoming procedures to describe traffic flows in external multicast scenarios. There are two scenarios.

- Source in the outside world: Listeners inside the fabric.
- Source inside the fabric: Listeners outside the fabric.

Once we understand the procedures for the above, we can describe the flows for sources existing inside or/and outside the fabric with listeners existing inside or/and outside the fabric.:

- Source in the outside world: Listeners inside the fabric as well as in the outside world
- Source inside the fabric: Listeners inside the fabric as well as in the outside world

Thus we can deploy multicast sources and listeners inside and outside of the fabric, and be able to explain the procedures involved in signaling and multicast traffic forwarding from sources and listeners.

External Multicast with Layer 3 Connectivity

To enable this effort, external multicast using PIM with classic L3 connectivity is described. And for this discussion, let's call the device that connects the fabric to the outside world *Gateway* (PIM-GW shown in Figure 11.1).

There are two BL devices in the fabric that run PIM on IRB. The PIM-GW can be connected to both BL-1 and BL-2 over classic L3 links. The PIM-GW sees BL-1 as a PIM neighbor on one L3-interface, say interface x, on subnet 10.1.1.0/24. The PIM-GW sees BL-2 as another PIM neighbor on another L3-interface, say interface y, on subnet 50.1.1.0/24. Unicast reachability is configured between the BLs, PIM-GW, PIM-RP, etc.

It's possible that the PIM-GW can be connected to only one Spine (either by network design or due to a link failure). In addition, the BLs can be connected to multiple GWs over multiple L3 links. And when link 'y' between PIM-GW and BL-2 goes down, BL-2 should have unicast reachability to PIM-GW and PIM-RP. This is typically ensured by having a dedicated L3-link between BL-1 and BL-2. The reachability can be ensured by configuring a unicast routing protocol on one of the IRBs on the BLs.

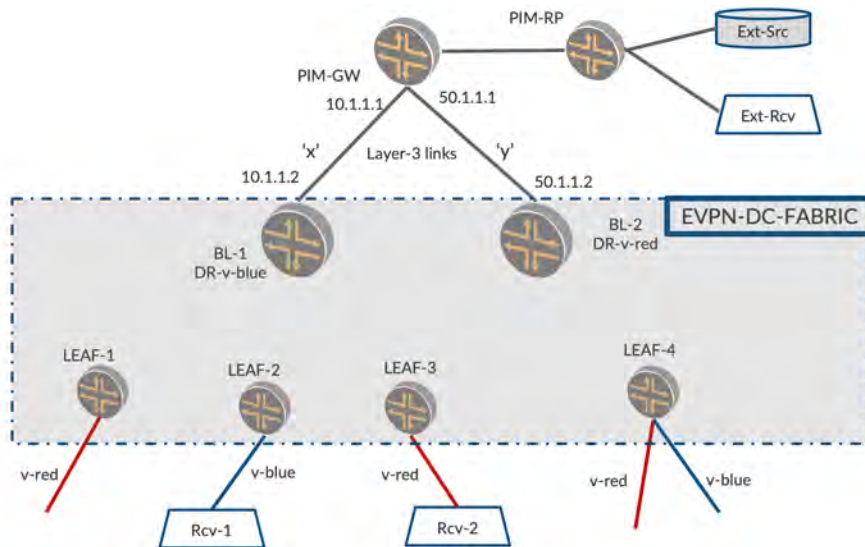


Figure 11.1 Topology: External Multicast with L3 Connectivity

The PIM-GW is the device that connects the EVPN-DC-Fabric to the outside world. The PIM-GW may itself be the PIM-RP or may be connected over multiple routers to a PIM-RP. The PIM-RP may be connected to external multicast sources and listeners directly or over multiple hops of routers. Procedures and packet flows for the following will be discussed:

- Listeners inside the fabric. Source outside the fabric
- Source inside the fabric: Listeners outside the fabric

Layer 3 Connectivity: Listener Inside the Fabric. Source Outside the Fabric

As shown by the red and blue VLANs in Figure 11.2 below, there is listener interest for group G, let's say 235.1.1.1. For the purpose of this example, Rcv-1 sends an IGMP report on v-blue and Rcv-2 sends an IGMP report on v-red.

On receiving the IGMP reports, LEAF-2 and LEAF-3 send equivalent Type-6 routes on v-red and v-blue. On receiving the Type-6 routes, BL-1 and BL-2 create PIM (*,G) states on IRB.red and IRB.blue, respectively, due to PIM DRship.

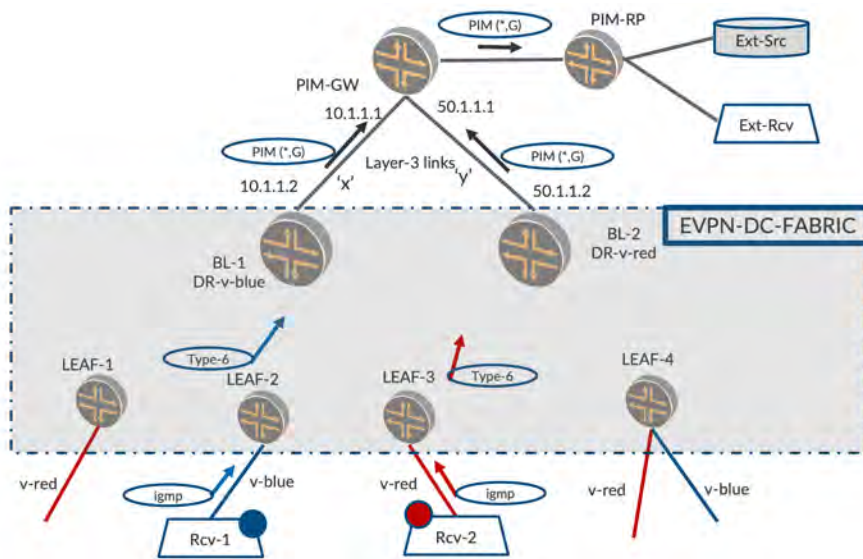


Figure 11.2 L3 Connectivity: Listener Inside and Source Outside the Fabric

When the BLs look to propagate the Join towards RP, they perform unicast route lookup towards RP address and send a PIM (*,G) Join towards RP. BL-1 sends the PIM Join on interface ‘x’ and BL-2 sends the join on interface ‘y’ towards the PIM-GW based on the best unicast path to RP.

PIM-GW receives (*,G) Join on both of its interfaces 'x' and 'y' and creates PIM (*,G) state with OIL as a list having both 'x' and 'y'. PIM-GW performs lookup towards RP and sends the Join to the device PIM-RP.

Overall, PIM states are created on BLs, PIM-GW, and PIM-RP with the appropriate PIM Join states having corresponding OILs.

Multicast Traffic Forwarding

Once the PIM Join lands in PIM-RP it resides there and gets periodically refreshed from downstream. When a multicast source becomes alive, PIM-RP gets to know of the source by virtue of PIM Registration (not shown Figure 11.3, as this is plain vanilla PIM behavior). After this, PIM-RP sends a (S,G) Join towards the source.

When the multicast source outside the fabric (Ext-Src) starts sending traffic, (shown as dotted lines in Figure 11.3), the traffic reaches RP. Since RP has a PIM Join state created for the group, as described in an earlier section, RP sends the traffic towards PIM-GW. Now, PIM-GW, on receiving traffic from RP, routes onto its OIL, the two L3 interfaces, ‘x’ and ‘y’.

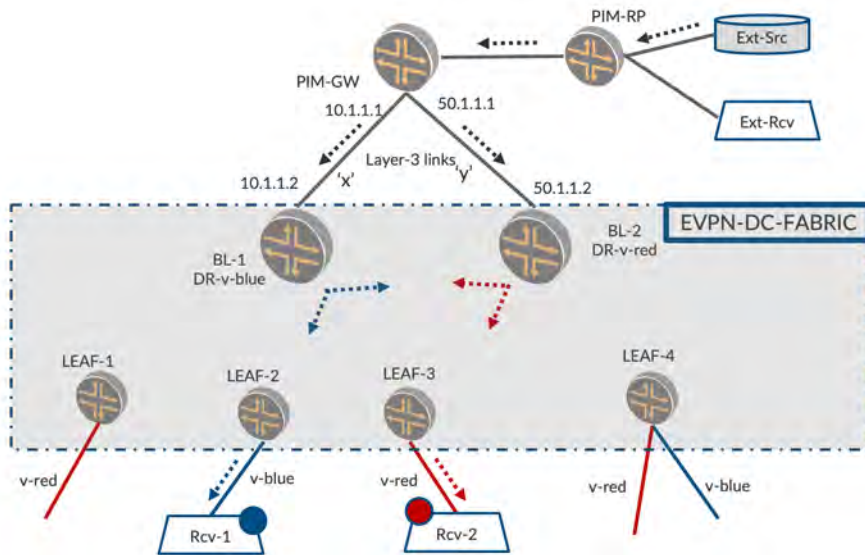


Figure 11.3 L3 Connectivity: Listener Inside and Source Outside the Fabric - TF

When the BLs receive the traffic on the L3 links, they route the traffic from the L3-link onto the IRBs and forward the traffic to the EVPN core. So BL-1 routes traffic coming on interface 'x' onto IRB.blue, and sends the traffic towards LEAF-2 and BL-2 on v-blue per procedures in Chapter 6. Similarly, BL-2 routes traffic coming on interface 'y' onto IRB.red, and sends the traffic towards LEAF-3 and BL-1 on v-red.

This routing on BL-1 from interface 'x' onto IRB.blue is similar to Inter-VLAN multicast routing. PIM routes at Layer 3 from 'x' onto IRB.blue. The routed multicast traffic on IRB.blue is sent to interested listener PEs using Selective Forwarding. If there are any local interfaces on BL-1 on v-blue, the routed traffic will be flooded to it too.

Layer 3 Connectivity: Listener Outside the Fabric. Source Inside the Fabric

In Figure 11.4, we have Ext-Rcv, a listener outside the fabric and Mcast-Src-1, inside the fabric. When listener Ext-Rcv sends an IGMP report, a PIM (*,G) Join state is created on RP. (The PIM Join from a router connected to Ext-Rcv may arrive at RP hop-by-hop.)

When Mcast-Src-1 starts sending traffic on v-red, LEAF-1 floods the traffic towards the EVPN core with Selective Forwarding procedures. This traffic reaches IRB.red on BL-1 and BL-2. The PIM DR for IRB.red, BL-2, performs PIM registration for the (S,G) information.

The PIM-RP, on learning of the source, looks to join the multicast source natively and hence initiates PIM (S,G) Join towards source. This PIM (S,G) Join reaches PIM-GW.

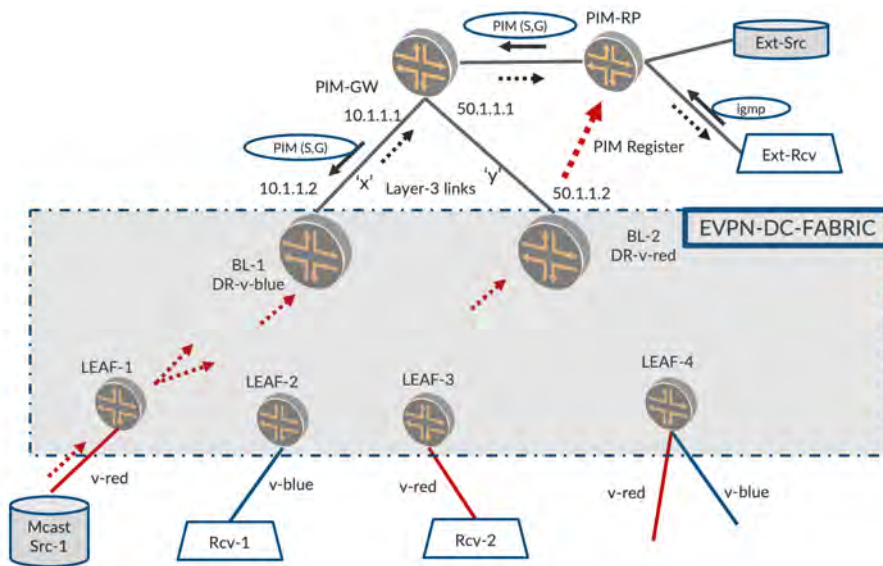


Figure 11.4

L3 Connectivity: Source Inside DC: Listener Outside DC

PIM-GW now conducts unicast route lookup to the source to pull traffic. PIM-GW has two paths to reach source over 'x' and 'y.' PIM-GW picks one of the interfaces to send the PIM (S,G) Join, say 'x,' and sends the Join. PIM-GW does not send the Join on 'y.'

When BL-1 receives the PIM (S,G) Join on 'x', it routes the traffic from IRB.red to 'x'. Once traffic reaches PIM-GW on interface 'x', PIM-GW sends the traffic to RP and in turn, RP to Ext-Rcv, thus reaching a listener outside the fabric.

PIM-GW, may have a scheme to load balance the PIM Joins sent over the two interfaces. Some (S,G)s may be sent to BL-1 and others may be sent to BL-2. This load balancing can be performed based on Join counts on the interfaces or based on prefix-hashing of the (S,G) tuple.

External Multicast with L2 Connectivity - MVLAN

Earlier we explored the nuances of connecting the EVPN fabric to the external world over L3-connectivity. Now let's explore the connectivity of EVPN fabric to the external world over L2-connectivity.

What does 'L2-connectivity' mean in this context? The physical links between GW and BLs 'x' and 'y' will remain the same as earlier. Instead of configuring separate subnets between PIM-GW and Spines, in this scheme we will configure the PIM-GW and BLs to be on the same *bridge*. Yes, we will mimic a LAN behavior between PIM-GW and BLs so that they see each other on the same subnet.

Though there are only two physical links (PIM-GW to BL-1 and PIM-GW to BL-2), the 'bridge' or 'emulated LAN' behavior will ensure that even when, say, the link between PIM-GW and SPINE-2 goes down, the devices can see each other on the subnet.

Sounds nice. How is this achieved? EVPN multihoming, of course. How? Begin by reviewing Figure 11.5.

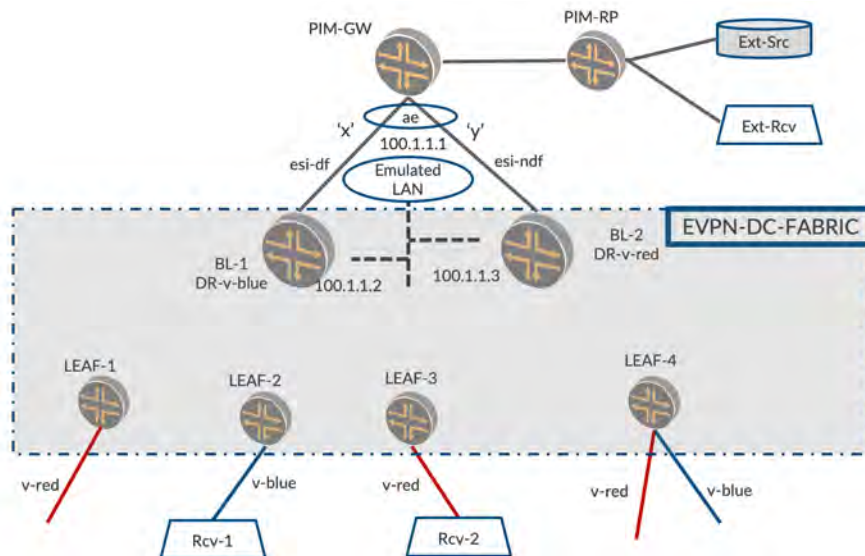


Figure 11.5 Topology: External Multicast with L2 MVLAN Connectivity

On BLs, we configured an EVPN VLAN with appropriate VLAN-ID and VNID for this purpose. This is a typical EVPN VLAN— nothing special. Since this VLAN is going to build the emulated LAN towards PIM-GW for us, however, let's call it a MVLAN (multicast-VLAN). *Please keep in mind that MVLAN is just a normal VLAN in terms of configuration and procedures.* For easier understanding, we'll call it *MVLAN* since it is the VLAN used to connect the GW and BLs in a subnet (LAN).

Layer 2 Connectivity: MVLAN Configuration and Procedures

To configure a MVLAN with VLAN and VNIDs, add interface 'x' on BL-1 as part of the MVLAN by configuring family 'bridge.' Also, add interface 'y' on BL-2 as part of the MVLAN by configuring family 'bridge.'

EVPN Multi-homing: Since the BLs are multihomed on the L2 interfaces 'x' and 'y', EVPN ESI is configured. Say, BL-1's interface 'x' is elected EVPN-DF. As with classic EVPN multi-homing, the other end of the multihomed PEs should be an AE bundle. So, configure the AE bundle on the PIM-GW.

IRB.MVLAN and Routing protocols: Having come thus far, let's make it appear as three PIM routers on a LAN. Towards this end enable IRB on the MVLANs on the BLs. Configure L3-IP addresses on the IRBs and on the AE interface on the PIM-GW. Only the three IP addresses should be part of the same subnet. (for example, 100.1.1.1/24, 100.1.1.2/24, 100.1.1.3/24).

Properties of MVLAN: This MVLAN has all the properties of a regular LAN, therefore any BUM packet sent on the MVLAN will be received by all the devices on the MVLAN. Routing protocols can be enabled on this MVLAN. When configuring OSPF and PIM on these IRB interfaces on BLs and on AE interface on PIM-GW, you can see that they see each other as OSPF and PIM neighbors. Now that we have built an emulated LAN, let's explore some first principles of how it came to be.

Layer 2 Connectivity: PIM-GW to BLs Multicast Forwarding

When the device PIM-GW sends a multicast packet, say, an OSPF hello packet (destination address is a multicast address 224.0.0.5) on an AE bundle interface, the packet will be sent on any one of the member links of the AE bundle.

If the packet is sent on interface 'y,' when it reaches BL-2, BL-2 sees it as a multicast L2-frame arriving on the MVLAN since family 'bridge' is configured on this interface.

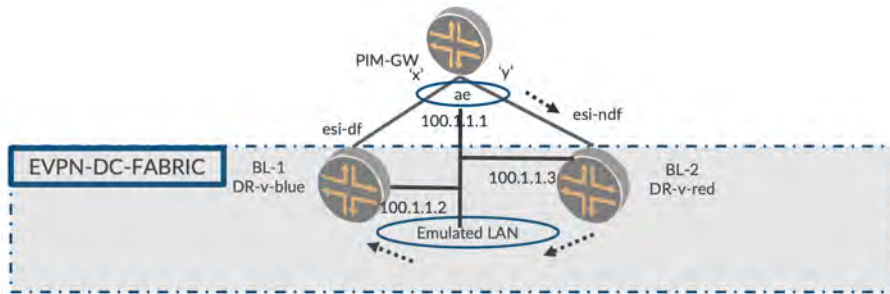


Figure 11.6 Traffic from PIM-GW Reaches Both BLs

BL-2 has two responsibilities:

- a.) L2-switch one copy of the frame towards EVPN core using Ingress Replication.
- b.) Remove L2-headers and punt the L3-packet to the IRB.MVLAN interface.

Due to 'a', this L2-frame reaches BL-1 over the EVPN core. Upon receiving this L2-frame, BL-1 has two responsibilities:

- a.) L2-switch one copy of the frame to access (accounting for DF and local bias).
- b.) Remove L2-headers and punt the L3-packet to the IRB.MVLAN interface.

You can see from all of this that the OSPF hello packet from PIM-GW is received on both of the BLs' IRB interface. The above procedures apply for any multicast packets that come from PIM-GW towards BLs.

The same procedures apply for multicast packets that transit PIM-GW and reach BLs. Therefore, when PIM-RP forwards multicast traffic to PIM-GW and PIM-GW forwards it over the AE interface towards BLs, both BLs will receive the traffic on IRB.MVLAN.

Suppose interface 'y' goes down. PIM-GW, when forwarding traffic on the AE bundle towards the BLs will be seamless because the PIM-GW will now send on the other member link of the bundle, namely 'x'. This traffic will reach BL-1 over the access interface and will reach BL-2 over the EVPN emulated LAN. Thus, both Spines will receive traffic when a link goes down.

Layer2 Connectivity: BL to the Other BL and PIMGW Multicast Forwarding

Let us quickly explore how the multicast packets from one BL reach the other BL and the PIM-GW (see Figure 11.7). Suppose BL-2 sends a PIM hello packet on IRB.MVLAN. This means that the L3-packet will be added with L2-headers and will be sent on all the access interfaces and towards the EVPN core.

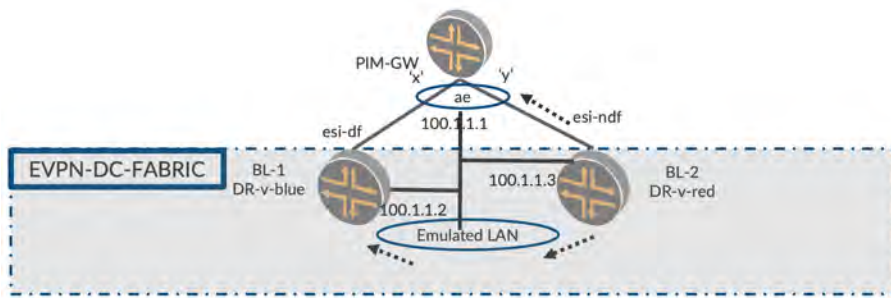


Figure 11.7 Traffic from BL-1 to the Other BL-2 and PIM-GW

A quick rule whose rationale will be made clear soon: *L3-Multicast Packets originated by the EVPN device or L3-routed multicast packets are to be sent on an access interface irrespective of whether the device is DF/NDF, or not on an ESI.*

So when BL-2 sends the multicast PIM hello packet, it is sent out on interface ‘y’ towards PIM-GW, though it is an NDF. Also, this packet is sent over the EVPN core and reaches BL-1. Based on the procedures for IRB, the hello packet reaches IRB.MVLAN on BL-1. SPINE-1 does not forward the L2-frame to PIM-GW due to local bias.

This explains the rationale for our quick rule. If BL-2 had not forwarded to PIM-GW due to being an NDF, and with BL-1 also not forwarding to PIM-GW due to local-bias, the PIM-GW will not have received the traffic/hello at all. The same rationale applies for L3-routed multicast traffic.

The same procedures can be used to describe how routed L3-multicast traffic reaches both the PIM-GW and the other BL. Therefore, when L3-multicast routing is done from IRB.v-red to IRB.MVLAN, by, say, BL-2, the traffic will be sent out on interface ‘y’ towards PIM-GW and also be sent towards the EVPN core to BL-1. BL-1 would not forward on ‘x’ due to local bias rules.

Summary

EVPN BUM Forwarding towards the core using Ingress Replication serves as a conduit for packets that are sent or received on the MVLAN. Thus, a packet sent on the MVLAN reaches all the devices on the LAN.

External Multicast with MVLAN: Traffic Flows

Earlier we introduced how an emulated LAN is achieved between the BLs and PIM-GW using EVPN-MH procedures. With these principles, we can walk through the scenarios of the listener and the source being inside or outside the fabric.

Layer 2 Connectivity: Listener Inside the Fabric. Source Outside the Fabric

In Figure 11.8 we have listener interest in the fabric on v-red and v-blue. LEAF-2 and LEAF-3 send Type-6 routes for blue and red, respectively.

Based on incoming Type-6 routes from LEAF devices, BL-1 and BL-2 create PIM (*,G) Join states on IRB.blue and IRB.red, respectively. Now when the BLs do a route lookup to source, the unicast route points to IRB.MVLAN. So the BLs send the Joins on IRB.MVLAN. When it reaches PIM-GW this PIM Join creates a state for PIM (*,G) with its OIL as 'ae' interface. PIM-GW propagates the Join to PIM-RP leading to PIM (*,G) state creation on PIM-RP.

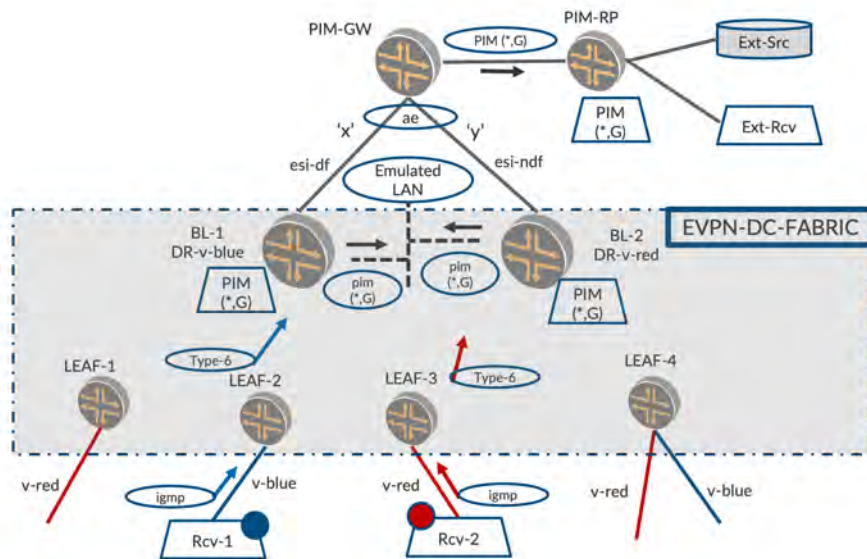


Figure 11.8 L2-Connectivity: Listener inside Fabric: Source Outside Fabric - Signaling

A quick discussion of nuance on PIM Join... the PIM Joins are multicast packets sent with destination 224.0.0.13, so when BL-2 deduces that the unicast route is PIM-GW and sends a PIM Join, it will be received by both BL-1 and PIM-GW. It is imperative that only PIM-GW processes that Join and BL-1 does not. Towards this end, BL-2 adds a 'upstream-neighbor-address' field in PIM Join TLV and populates it with PIM-GW. Based on this field, PIM-GW alone processes the Join, while BL-1 ignores the PIM Join sent by BL-2.

Multicast Traffic Forwarding

When multicast traffic is started from Ext-Src, the traffic reaches PIM-RP and PIM-GW. PIM-GW forwards the traffic over the AE interface. As we can see in Figure 11.9, the traffic forwarding is equivalent to sending on the LAN. This traffic reaches both BLs on IRB.MVLAN. BL-1 and BL-2 route the multicast traffic from IRB.MVLAN onto IRB.v-blue and IRB.v-red, respectively, thus reaching the listeners Rcv-1 and Rcv-2.

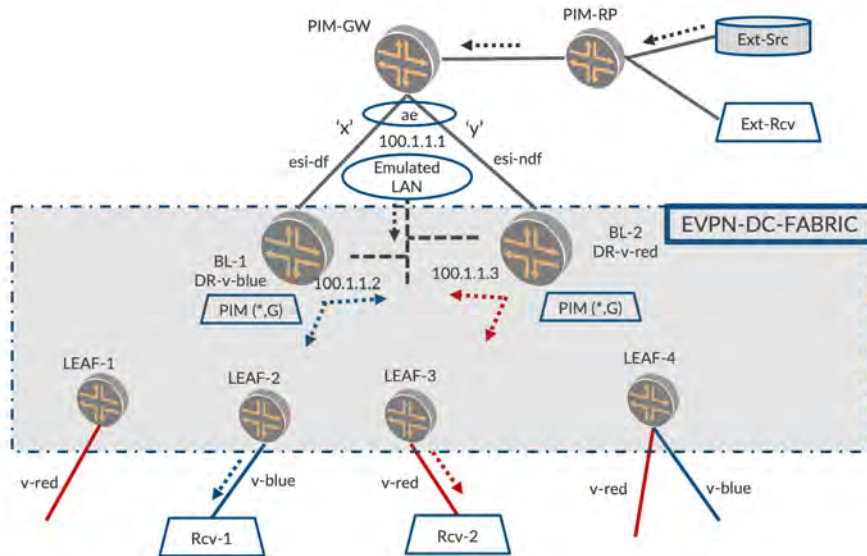


Figure 11.9

L2-Connectivity: Listener inside Fabric: Source Outside Fabric – Traffic Forwarding

The PIM state on say, BL-1, will have upstream interfaces as IRB.MVLAN and OIL as IRB.blue. The multicast traffic that reached IRB.MVLAN will be routed onto IRB.blue and sent to the interested LEAF devices (LEAF-2) using SMET forwarding.

NOTE In L3-connectivity, traffic was routed from interface 'x' onto IRB.blue. In this case, the traffic is routed from IRB.MVLAN onto IRB.v-blue. The PIM procedures are the same with only the incoming interface being different.

NOTE There's an IIF-Mismatch: BL-1's post routing onto IRB.blue will send it over the EVPN core. This traffic will reach BL-2 on v-blue. BL-2 will not forward to PIM-GW due to local-bias rules. Also, BL-2 will receive this traffic on IRB.blue. Since BL-2 has installed a state with incoming interface (IIF) as IRB.MVLAN, it will result in interface mismatch event (IIF-MISMATCH) and BL-2 will drop the packet.

Layer 2 Connectivity: Source Inside the Fabric. Listener Outside the Fabric

Let's explore the scenario and procedures when the source is inside the fabric and the listener is outside: see Figure 11.10. When Ext-Rcv sends an IGMP report, it reaches PIM-RP and the state is created on PIM-RP. When the source is inside the fabric, Mcast-Src-1 is started, and LEAF-1 ingress replicates the traffic on v-red to the BL devices.

BL-2 which is the PIM DR for v-red, sends a PIM Register to the PIM-RP. Thus, PIM-RP comes to know of the source. PIM-RP looks to join towards the source and sends a (S,G) Join towards PIM-GW. PIM-GW, upon receiving the PIM Join, looks to join the source.

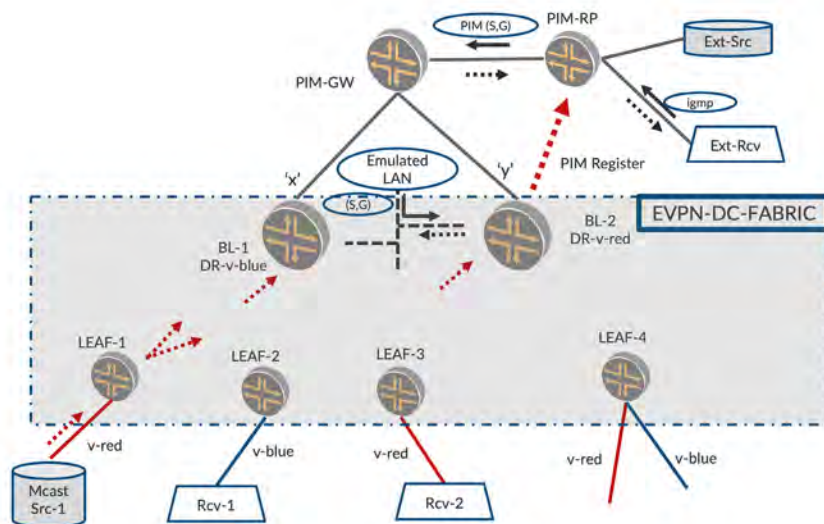


Figure 11.10 L2 Connectivity: Source Inside, Listener Outside Fabric

When PIM-GW does unicast route lookup to the source, the upstream interface will be AE. Since it sees two PIM neighbors, and both are equidistant to the source (ECMP), PIM-GW will select one BL-1 to target the Join. If the PIM-GW picked BL-2, per PIM procedures described previously, PIM-GW will populate BL-2's IP address in the 'upstream-neighbor-addr' field in the Join message. BL-1 will drop this Join since it deduces that the join is 'not-for-me'. BL-2 will accept this join ('for-me'), process it, and create state with OIL=IRB.MVLAN.

BL-2, which is receiving traffic on v-red will route the traffic onto IRB.MVLAN. This traffic reaches PIM-GW on AE and PIM-GW forwards to RP. The PIM-RP forwards it to the external listener.

Comparison Between L3-links and L2-MVLAN

Which connectivity to choose can be a matter of a personal choice for the operator. Some deployments may already have one connectivity configured for unicast and multicast may be deployed on top. Multicast traffic forwarding will work in both scenarios, though the procedures and states will be slightly different.

If you have a choice, it may be preferable to configure L2-connectivity because of better convergence. With L3-connectivity, reacting to link-down events requires L3 processing and updating, while with L2-connectivity, the link-down events can be handled seamlessly with AE interface handling and EVPN DF/NDF procedures.

With L3-connectivity, when a link between BL-2 and PIM-GW goes down, BL-2 needs unicast reachability. If this path goes over one of the IRBs in the fabric, it can cause confusion (though it will work correctly). To avoid this, it may be preferable for there to be a dedicated L3 link between BL-1 and BL-2. When a new BL is added, this new BL needs physical connectivity to PIM-GW and other BL to provide unicast reachability.

With L2-connectivity, once the M-VLAN is configured (as a normal EVPN VNI configuration), we can benefit from all the procedures from an emulated LAN. As long as there is one physical link being up between the BL and PIM-GW, multicast will work seamlessly. In the future, when the number of BLs is increased from two to four, these need not have physical connectivity as long as the M-VLAN is configured on the BLs. There is one downside with L2-connectivity that the DR may be BL-2 but the traffic from PIM-GW may reach BL-1 and then BL-2, thus taking an extra physical hop though on the same subnet.

Chapter Summary

This chapter was loaded with discussions of the different procedures and packet flows involved in connecting the EVPN data center fabric to the outside world.

It explored multicast traffic forwarding to and from the data center fabric using PIM and EVPN MH procedures. From this you should be able to understand how to deploy real-time multicast in an EVPN data center fabric in a centrally routed model.

Configuration and Verification

Let's look at the multicast behavior when the receiver or source are outside the data center. If you are in your lab following along, stop all sources and receivers that were previously started.

Configuration

In Chapter 10 we looked at inter VLAN routing of multicast traffic, though since our sources and receivers were both within the data center, we did not focus too much on the PIM protocol configurations and machinery involved with multicast routing – in particular, the PIM-FHR (first-hop router) and RP configuration. In fact, we were able to get away with both border-leaf PEs configured as local RPs.

However, now that we are looking at external multicast, let's make things more realistic. Towards this end we will configure a PIM-GW device which marks the beginning of our “external” world, configure a PIM RP outside the data center, and modify the configuration on BL-1 and BL-2 such that one of them acts as PIM FHR for multicast traffic originating within the data center. The configurations on all other devices remain the same.

Configuring the PIM-GW

Copy and paste the below configuration on MX-PIM-GW.

Configure a VLAN to connect to the DC:

```
set bridge-domains BD-1000 vlan-id 1000
commit
```

Configure the interfaces:

```
set interfaces lo0 unit 0 family inet address 110.110.110.110/32
set interfaces ge-0/0/1 description "T0 PIM-RP"
set interfaces ge-0/0/1 unit 0 family inet address 20.20.20.1/24
set interfaces ge-0/0/0 description "T0 BORDER_LEAF-1"
set interfaces ge-0/0/0 gigether-options 802.3ad ae0
set interfaces ge-0/0/2 description "T0 BORDER_LEAF-2"
set interfaces ge-0/0/2 gigether-options 802.3ad ae0
set chassis aggregated-devices ethernet device-count 1
set interfaces ae0 aggregated-ether-options lacp active
set interfaces ae0 aggregated-ether-options lacp periodic fast
set interfaces ae0 unit 0 family bridge interface-mode trunk
set interfaces ae0 unit 0 family bridge vlan-id-list 1000
set interfaces irb unit 1000 family inet address 30.30.30.10/24
set bridge-domains BD-1000 routing-interface irb.1000
commit
```

Configure OSPF for unicast routing:

```
set protocols ospf area 0.0.0.0 interface irb.1000
set protocols ospf area 0.0.0.0 interface lo0.0 passive
set protocols ospf area 0.0.0.0 interface ge-0/0/1.0
```

Configure PIM for L3 multicast routing:

```
set protocols pim rp static address 111.111.111.111
set protocols pim interface lo0.0
set protocols pim interface irb.1000
set protocols pim interface ge-0/0/1.0
commit
```

Configuring the PIM-RP

Configure the interfaces:

```
set chassis fpc 0 pic 0 tunnel-services
set interfaces ge-0/0/0 description "T0 MX-PIM-GW"
set interfaces ge-0/0/0 unit 0 family inet address 20.20.20.2/24
set interfaces ge-0/0/1 description "T0 IXIA-A"
set interfaces ge-0/0/1 unit 0 family inet address 21.21.21.1/24
set interfaces lo0 unit 0 family inet address 111.111.111.111/32
commit
```

Configure OSPF for unicast routing:

```
set protocols ospf area 0.0.0.0 interface ge-0/0/0.0
set protocols ospf area 0.0.0.0 interface lo0.0 passive
set protocols ospf area 0.0.0.0 interface ge-0/0/1.0
commit
```

Configure PIM for L3 multicast routing:

```
set protocols pim rp local address 111.111.111.111
set protocols pim interface ge-0/0/0.0 mode sparse
set protocols pim interface ge-0/0/1.0 mode sparse
set protocols pim interface lo0.0 mode sparse
commit
```

Configure tunnel services so that this PIM RP device can decapsulate PIM registers:

```
set chassis fpc 0 pic 0 tunnel-services
commit
```

Configuring the Border-Leaf PEs

Configuration on BL-1.

Configure tunnel services so that this device can encapsulate PIM registers:

```
set chassis fpc 0 pic 0 tunnel-services
commit
```

Configure the interface towards the PIM-GW:

```
set interfaces xe-0/0/0 gigether-options 802.3ad ae0
set interfaces ae0 description "T0 MX-PIM-GW"
set interfaces ae0 esi 00:44:44:44:44:44:44:44
set interfaces ae0 esi all-active
set interfaces ae0 aggregated-ether-options lACP active
```

```
set interfaces ae0 aggregated-ether-options lacp periodic fast
set interfaces ae0 aggregated-ether-options lacp system-id 00:44:44:44:44
set interfaces ae0 unit 0 family ethernet-switching interface-mode trunk
set interfaces ae0 unit 0 family ethernet-switching vlan members VLAN-1000
commit
```

Configure a new VLAN (let's call this the M-VLAN) to connect to the PIM-GW:

```
set vlans VLAN-1000 vlan-id 1000
set vlans VLAN-1000 vxlan vni 1000
commit
```

Configure EVPN to extend the M-VLAN:

```
set protocols evpn extended-vni-list 1000
commit
```

Configure IGMP-snooping in the M-VLAN and make the interface towards the PIM-GW an M-router interface:

```
set protocols igmp-snooping vlan VLAN-1000 interface xe-0/0/0.0 multicast-router-interface
commit
```

Configure an M-VLAN IRB:

```
set interfaces irb unit 1000 virtual-gateway-accept-data
set interfaces irb unit 1000 virtual-gateway-esi 00:55:55:55:55:55:55:55:55:55
set interfaces irb unit 1000 virtual-gateway-esi all-active
set interfaces irb unit 1000 family inet address 30.30.30.1/24 virtual-gateway-address 30.30.30.100
set routing-instances VRF-1 interface irb.1000
set vlans VLAN-1000 l3-interface irb.1000
commit
```

Configure non-passive OSPF on the M-VLAN IRB:

```
set routing-instances VRF-1 protocols ospf area 0.0.0.0 interface irb.1000
commit
```

Modify the PIM RP configuration:

```
delete routing-instances VRF-1 protocols pim rp
set routing-instances VRF-1 protocols pim rp static address 111.111.111.111
commit
```

Copy and paste the below configuration on BL-2:

BL-2:

```
Configure tunnel services so that this device can encapsulate PIM registers:
set chassis fpc 0 pic 0 tunnel-services
commit
```

Configure the interface towards the PIM-GW:

```
set interfaces xe-0/0/0 gigether-options 802.3ad ae0
set interfaces ae0 description "TO MX-PIM-GW"
set interfaces ae0 esi 00:44:44:44:44:44:44:44:44:44
set interfaces ae0 esi all-active
set interfaces ae0 aggregated-ether-options lacp active
```

```
set interfaces ae0 aggregated-ether-options lACP periodic fast
set interfaces ae0 aggregated-ether-options lACP system-id 00:44:44:44:44
set interfaces ae0 unit 0 family ethernet-switching interface-mode trunk
set interfaces ae0 unit 0 family ethernet-switching vlan members VLAN-1000
commit
```

Configure a new VLAN (let's call this the M-VLAN) to connect to the PIM-GW:

```
set vlans VLAN-1000 vlan-id 1000
set vlans VLAN-1000 vxlan vni 1000
commit
```

Configure EVPN to extend the M-VLAN:

```
set protocols evpn extended-vni-list 1000
commit
```

Configure IGMP-snooping in the M-VLAN and make the interface towards the PIM-GW an M-router interface:

```
set protocols igmp-snooping vlan VLAN-1000 interface xe-0/0/0.0 multicast-router-interface
commit
```

Configure an M-VLAN IRB:

```
set interfaces irb unit 1000 virtual-gateway-accept-data
set interfaces irb unit 1000 virtual-gateway-esi 00:55:55:55:55:55:55:55:55:55
set interfaces irb unit 1000 virtual-gateway-esi all-active
set interfaces irb unit 1000 family inet address 30.30.30.2/24 virtual-gateway-address 30.30.30.100
set routing-instances VRF-1 interface irb.1000
set vlans VLAN-1000 l3-interface irb.1000
commit
```

Configure non-passive OSPF on the M-VLAN IRB:

```
set routing-instances VRF-1 protocols ospf area 0.0.0.0 interface irb.1000
commit
```

Modify the PIM RP configuration:

```
delete routing-instances VRF-1 protocols pim rp
set routing-instances VRF-1 protocols pim rp static address 111.111.111.111
commit
```

Verification

Receiver Outside the DC and Source Within the DC

As you did before, start sending multicast traffic from Host-1 at 10 pps (packets per second) for group 225.1.1.1 in VLAN-101. On Host-6 and Host-3, start receivers for the multicast group, 225.1.1.1 on both VLAN-101 and VLAN-102.

You can see from the RT statistics in Figure 11.11 that, just as before, the traffic sent by Host-1 at 10 pps is now received by the interested receivers, Host-6 and

Host-3, and the legacy device, Host-7, in both VLAN-101 and VLAN-102, thereby resulting in 20 pps of incoming traffic on each of them.

In addition, you can see that the traffic is also received by the interested receiver outside the data center, Host-8, resulting in 10 pps of incoming traffic on Host-8.

	Stat Name	Port Name	Link State	Frames Tx, Rate
1	10.216.45.202/Card20/Port01	HOST-1	Link Up	1
2	10.216.45.202/Card03/Port01	HOST-2	Link Up	
3	10.216.45.202/Card20/Port02	HOST-3	Link Up	
4	10.216.45.202/Card03/Port02	HOST-4	Link Up	
5	10.216.45.202/Card20/Port03	HOST-5	Link Up	
6	10.216.45.202/Card03/Port03	HOST-6	Link Up	
7	10.216.45.202/Card03/Port04	HOST-7	Link Up	
8	10.216.45.202/Card20/Port04	HOST-8	Link Up	

Figure 11.11 RT Stats

Multicast Traffic Outputs - LEAF-1, LEAF-2, LEAF-4, LEAF-5, SPINE-2 (VLAN-101): the traffic forwarding behavior on LEAF-1, LEAF-2, LEAF-3, LEAF-4, SPINE-1, and SPINE-2 remains the same and was skipped for the sake of brevity.

Multicast Traffic Outputs – BL-1, BL-2: As before, BL-1 being PIM DR on irb.102, routes the traffic arriving on irb.101 into irb.102 to serve the VLAN-102 receivers within the DC:

```
lab@BL-1> show multicast route extensive instance VRF-1
Instance: VRF-1 Family: INET
Group: 225.1.1.1
  Source: 18.18.18.30/32
  Upstream interface: irb.101
  Downstream interface list:
    irb.102
  Number of outgoing interfaces: 1
  Session description: Unknown
  Statistics: 1 kbps, 10 pps, 917043 packets
  Next-hop ID: 131102
  Upstream protocol: PIM
  Route state: Active
  Forwarding state: Forwarding
  Cache lifetime/timeout: 360 seconds
  Wrong incoming interface notifications: 33
  Uptime: 00:03:15
```

Since the external receiver interest is notified to BL-2 via the PIM Join in the M-VLAN, VLAN-1000, BL-2 routes the traffic arriving on irb.101 into irb.1000 (i.e. VLAN-1000):

```
lab@BL-2> show multicast route extensive instance VRF-1
Instance: VRF-1 Family: INET
Group: 225.1.1.1
```

```

Source: 18.18.18.30/32
Upstream interface: irb.101
Downstream interface list:
  irb.1000
Number of outgoing interfaces: 1
Session description: Unknown
Statistics: 1 kbps, 10 pps, 1970 packets
Next-hop ID: 131108
Upstream protocol: PIM
Route state: Active
Forwarding state: Forwarding
Cache lifetime/timeout: 310 seconds
Wrong incoming interface notifications: 2
Uptime: 00:04:32

```

Chapter 10 showed what happens to the traffic routed in VLAN-102 by BL-1. So now, let's look at what happens to the traffic routed to irb.1000 by BL-2.

The traffic routed into VLAN-1000 is sent out on the access interface, ae0 towards the PIM-GW:

```

lab@BL-2> show interfaces ae0 extensive
...
Logical interface ae0.0 (Index 573) (SNMP ifIndex 568) (Generation 215)
Flags: Up SNMP-Traps 0x24024000 Encapsulation: Ethernet-Bridge
Statistics      Packets      pps      Bytes      bps
Bundle:
  Input :          6376          0      450534      296
  Output:          4926         11      423788      6328
...

```

Multicast Traffic Outputs – PIM-GW

PIM-GW routes the traffic arriving on irb.1000 onto its interface, ge-0/0/1.0, towards the PIM-RP:

```

lab@PIM-GW> show multicast route extensive
Instance: master Family: INET
Group: 225.1.1.1
  Source: 18.18.18.30/32
  Upstream interface: irb.1000
  Downstream interface list:
    ge-0/0/1.0
  Number of outgoing interfaces: 1
  Session description: Unknown
  Statistics: 1 kbps, 10 pps, 415 packets
  Next-hop ID: 1048577
  Upstream protocol: PIM
  Route state: Active
  Forwarding state: Forwarding
  Cache lifetime/timeout: 360 seconds
  Wrong incoming interface notifications: 0
  Uptime: 00:00:41

```

Multicast Traffic Outputs – PIM-RP

PIM-RP routes the traffic received from PIM-GW onto its interface, ge-0/0/1.0, towards the external receiver, HOST-8:

```
lab@PIM-RP> show multicast route extensive
Instance: master Family: INET
Group: 225.1.1.1
  Source: 18.18.18.30/32
  Upstream interface: ge-0/0/0.0
  Downstream interface list:
    ge-0/0/1.0
  Number of outgoing interfaces: 1
  Session description: Unknown
  Statistics: 1 kBps, 10 pps, 323 packets
  Next-hop ID: 1048578
  Upstream protocol: PIM
  Route state: Active
  Forwarding state: Forwarding
  Cache lifetime/timeout: 360 seconds
  Wrong incoming interface notifications: 1
  Uptime: 00:00:33
```

Receiver Within the DC and Source Outside the DC

Now let's look at the case where the source is outside the DC.

We will stop all sources and receivers that have been previously started, and you should, too, in your own lab.

Now start sending multicast traffic from Host-8 at 10 pps for group 225.1.1.1. As before, on Host-6 and Host-3, start receivers for the multicast group, 225.1.1.1 on both VLAN-101 and VLAN-102.

Traffic Statistics on RT

From the RT statistics, you can see that the traffic sent by Host-8 at 10 pps is received by the interested receivers, Host-6 and Host-3, and the legacy device, Host-7, in both VLAN-101 and VLAN-102 thereby resulting in 20 pps of incoming traffic on each of them.

Multicast Traffic Outputs – PIM-RP

PIM-RP routes the traffic received on interface ge-0/0/1.0 from HOST-8 onto its interface, ge-0/0/0.0, towards PIM-GW.

Multicast Traffic Outputs – PIM-GW

PIM-GW routes the traffic received on interface ge-0/0/1.0 from PIM-RP onto its interface, irb.1000 (VLAN-1000), towards the DC.

The traffic routed into VLAN-1000 will be forwarded on the interface ae0 and will be load balanced towards BL-1 or BL-2. In our case, we see that it is load balanced on interface ge-0/0/2 towards BL-2:

```
lab@PIM-GW> show interfaces ae0 extensive
...
Logical interface ae0.0 (Index 337) (SNMP ifIndex 563) (Generation 146)
Flags: Up SNMP-Traps 0x24024000 Encapsulation: Ethernet-Bridge
Statistics          Packets      pps        Bytes      bps
Bundle:
  Input :           4481          0       278030       336
  Output:           7348         10       364552      3960
...
Link:
  ge-0/0/0.0
    Input :           0           0           0           0
    Output:           0           0           0           0
  ge-0/0/2.0
    Input :           4481          0       278030       336
    Output:           7348         10       364552      3960
...
```

Multicast Traffic Outputs – BL-1, BL-2

BL-2 switches the external multicast traffic arriving on the access interface, ae0 in the MVLAN, VLAN-1000 towards BL-1. In addition, being PIM DR on irb.101, BL-2 also routes the traffic arriving on irb.1000 into irb.101 to serve the VLAN-101 receivers within the DC:

```
lab@BL-2> show multicast route extensive instance VRF-1
Instance: VRF-1 Family: INET
Group: 225.1.1.1
  Source: 21.21.21.2/32
  Upstream interface: irb.1000
  Downstream interface list:
    irb.101
  Number of outgoing interfaces: 1
  Session description: Unknown
  Statistics: 1 kBps, 10 pps, 6808 packets
  Next-hop ID: 131070
  Upstream protocol: PIM
  Route state: Active
  Forwarding state: Forwarding
  Cache lifetime/timeout: 360 seconds
  Wrong incoming interface notifications: 0
  Uptime: 00:15:23
```

BL-1, receives the external multicast traffic forwarded by BL-2 on the VTEP in the M-VLAN (VLAN-1000, irb.1000). Since the traffic arrives from the core, it is not sent back to the core (CORE-IMET-SKIP). Local bias procedures (DST-LOCAL-BIAS) prevent the traffic from being sent back on the access interface, ae0 towards the PIM-GW.

Being PIM DR on irb.102, BL-1 routes the traffic arriving on irb.1000 into irb.102 to serve the VLAN-102 receivers within the DC:

```

lab@BL-1> show multicast route extensive instance VRF-1
Instance: VRF-1 Family: INET
Group: 225.1.1.1
  Source: 21.21.21.2/32
  Upstream interface: irb.1000
  Downstream interface list:
    irb.102
  Number of outgoing interfaces: 1
  Session description: Unknown
  Statistics: 1 kbps, 10 pps, 6408 packets
  Next-hop ID: 1310890
  Upstream protocol: PIM
  Route state: Active
  Forwarding state: Forwarding
  Cache lifetime/timeout: 360 seconds
  Wrong incoming interface notifications: 0
  Uptime: 00:15:30

```

The external multicast traffic thus routed into VLAN-101 (by BL-2) and VLAN-102 (by BL-1) is optimally switched in the data center with the help of AR and SMET procedures. We did detailed verifications for this in Chapter 10. The verification of this flow is therefore left as an exercise for the reader.

NOTE Here we also see the advantage of the DR-Priority configuration we completed earlier in sharing the load of routing the multicast traffic into different customer VLANs, between the two BORDER-LEAF devices.

Detailed Control Plane Verification

Receiver Outside the DC and Source Within the DC

Verify that the PIM-GW sends the (S,G) PIM Join towards one of the border leaf PEs in order to pull traffic from the data center for the external receiver. In our case, the Join is being sent towards BL-2 via irb.1000:

```

lab@PIM-RP> show pim join extensive 225.1.1.1 source 18.18.18.30
Group: 225.1.1.1
  Source: 18.18.18.30
  Flags: sparse,spt
  Upstream interface: irb.1000
  Upstream neighbor: 30.30.30.2 (assert winner)
  Upstream state: Join to Source, Prune to RP
  Keepalive timeout:
  Uptime: 00:00:43
  Downstream neighbors:
    Interface: ge-0/0/1.0 (assert winner)
      20.20.20.2 State: Join Flags: S Timeout: 167
      Uptime: 00:00:43 Time since last Join: 00:00:43
      Assert Winner: 20.20.20.1 Metric: 2 Pref: 10 Timeout: 42
...

```

Verify that, BL-2 alone builds PIM (S,G) state with PIM-RP in its downstream in-

terface list:

```
lab@BL-2> show pim join extensive 225.1.1.1 source 18.18.18.30
Group: 225.1.1.1
  Source: 18.18.18.30
  Flags: sparse,spt
  Upstream interface: irb.101
  Upstream neighbor: Direct
  Upstream state: Local Source, Prune to RP
  Keepalive timeout: 335
  Uptime: 00:01:07
  Downstream neighbors:
    Interface: irb.101
      18.18.18.2 State: Join Flags: S   Timeout: Infinity
      Uptime: 00:01:07 Time since last Join: 00:01:07
    Interface: irb.1000 (assert winner)
      30.30.30.10 State: Join Flags: S Timeout: 203
      Uptime: 00:01:07 Time since last Join: 00:00:07
      Assert Winner: 30.30.30.2 Metric: 0 Pref: 0 Timeout: 17
  Number of downstream interfaces: 2
  Number of downstream neighbors: 2
```

```
lab@BL-1> show pim join extensive 225.1.1.1 source 18.18.18.30
Group: 225.1.1.1
  Source: 18.18.18.30
  Flags: sparse,spt
  Upstream interface: irb.101
  Upstream neighbor: Direct
  Upstream state: Local Source, Local RP
  Keepalive timeout: 345
  Uptime: 00:00:55
  Downstream neighbors:
    Interface: irb.102
      19.19.19.2 State: Join Flags: S   Timeout: Infinity
      Uptime: 00:01:07 Time since last Join: 00:00:55
  Number of downstream interfaces: 1
  Number of downstream neighbors: 1
```

Receiver Within the DC and Source Outside the DC

Verify that BL-1 and BL-2 send (*, G) and (S,G) PIM Joins towards PIM-GW on behalf the receivers within the DC:

```
lab@BL-2> show pim join extensive 225.1.1.1 star-g instance VRF-1
Instance: PIM.VRF-1 Family: INET
R = Rendezvous Point Tree, S = Sparse, W = Wildcard
Instance: PIM.VRF-1 Family: INET
R = Rendezvous Point Tree, S = Sparse, W = Wildcard
Group: 225.1.1.1
  Source: *
  RP: 111.111.111.111
  Flags: sparse,rptree,wildcard
  Upstream interface: irb.1000
  Upstream neighbor: 30.30.30.10
  Upstream state: Join to RP
  Uptime: 00:15:24
  Downstream neighbors:
```

```
Interface: irb.101
  18.18.18.2 State: Join Flags: SRW Timeout: Infinity
  Uptime: 00:14:07 Time since last Join: 00:14:07
Number of downstream interfaces: 1
Number of downstream neighbors: 1
lab@BL-2> show pim join extensive 225.1.1.1 source 21.21.21.2 instance VRF-1
Instance: PIM.VRF-1 Family: INET
R = Rendezvous Point Tree, S = Sparse, W = Wildcard
Group: 225.1.1.1
Source: 21.21.21.2
Flags: sparse,spt
Upstream interface: irb.1000
Upstream neighbor: 30.30.30.10
Upstream state: Join to Source, No Prune to RP
Keepalive timeout: 310
Uptime: 00:15:23
Downstream neighbors:
  Interface: irb.101
    18.18.18.2 State: Join Flags: S Timeout: Infinity
    Uptime: 00:14:07 Time since last Join: 00:14:07
Number of downstream interfaces: 1
Number of downstream neighbors: 1
```

Appendix

Base Configurations

This Appendix lists the base configurations on all devices (refer to Chapter 3). To quickly configure the boxes and get started, copy and paste the configurations on the devices as indicated.

Base Configuration on SPINE-1

```
set interfaces xe-0/0/0 description "T0 BL-1"
set interfaces xe-0/0/0 unit 0 family inet address 4.4.4.2/24
set interfaces xe-0/0/1 description "T0 BL-2"
set interfaces xe-0/0/1 unit 0 family inet address 6.6.6.2/24
set interfaces xe-0/0/2 description "T0 LEAF-1"
set interfaces xe-0/0/2 unit 0 family inet address 8.8.8.1/24
set interfaces xe-0/0/3 description "T0 LEAF-2"
set interfaces xe-0/0/3 unit 0 family inet address 9.9.9.1/24
set interfaces xe-0/0/4 description "T0 LEAF-3"
set interfaces xe-0/0/4 unit 0 family inet address 10.10.10.1/24
set interfaces xe-0/0/5 description "T0 LEAF-4"
set interfaces xe-0/0/5 unit 0 family inet address 11.11.11.1/24
set interfaces xe-0/0/6 description "T0 LEAF-5"
set interfaces xe-0/0/6 unit 0 family inet address 12.12.12.1/24
set interfaces lo0 unit 0 family inet address 103.103.103.103/32
set policy-options policy-statement EXPORT-L0 term 1 from protocol direct
set policy-options policy-statement EXPORT-L0 term 1 then accept
set routing-options router-id 103.103.103.103
set routing-options autonomous-system 65003
set protocols bgp group UNDERLAY family inet any
set protocols bgp group UNDERLAY export EXPORT-L0
set protocols bgp group UNDERLAY local-as 65003
set protocols bgp group UNDERLAY neighbor 4.4.4.1 description BL-1
set protocols bgp group UNDERLAY neighbor 4.4.4.1 peer-as 65001
set protocols bgp group UNDERLAY neighbor 6.6.6.1 description BL-2
```

```

set protocols bgp group UNDERLAY neighbor 6.6.6.1 peer-as 65002
set protocols bgp group UNDERLAY neighbor 8.8.8.2 description LEAF-1
set protocols bgp group UNDERLAY neighbor 8.8.8.2 peer-as 65005
set protocols bgp group UNDERLAY neighbor 9.9.9.2 description LEAF-2
set protocols bgp group UNDERLAY neighbor 9.9.9.2 peer-as 65006
set protocols bgp group UNDERLAY neighbor 10.10.10.2 description LEAF-3
set protocols bgp group UNDERLAY neighbor 10.10.10.2 peer-as 65007
set protocols bgp group UNDERLAY neighbor 11.11.11.2 description LEAF-4
set protocols bgp group UNDERLAY neighbor 11.11.11.2 peer-as 65008
set protocols bgp group UNDERLAY neighbor 12.12.12.2 description LEAF-5
set protocols bgp group UNDERLAY neighbor 12.12.12.2 peer-as 65009

```

Base Configuration on SPINE-2

```

set interfaces xe-0/0/0 description "TO BL-1"
set interfaces xe-0/0/0 unit 0 family inet address 5.5.5.2/24
set interfaces xe-0/0/1 description "TO BL-2"
set interfaces xe-0/0/1 unit 0 family inet address 7.7.7.2/24
set interfaces xe-0/0/2 description "TO LEAF-1"
set interfaces xe-0/0/2 unit 0 family inet address 13.13.13.1/24
set interfaces xe-0/0/3 description "TO LEAF-2"
set interfaces xe-0/0/3 unit 0 family inet address 14.14.14.1/24
set interfaces xe-0/0/4 description "TO LEAF-3"
set interfaces xe-0/0/4 unit 0 family inet address 15.15.15.1/24
set interfaces xe-0/0/5 description "TO LEAF-4"
set interfaces xe-0/0/5 unit 0 family inet address 16.16.16.1/24
set interfaces xe-0/0/6 description "TO LEAF-5"
set interfaces xe-0/0/6 unit 0 family inet address 17.17.17.1/24
set interfaces lo0 unit 0 family inet address 104.104.104.104/32
set policy-options policy-statement EXPORT-L0 term 1 from protocol direct
set policy-options policy-statement EXPORT-L0 term 1 then accept
set routing-options router-id 104.104.104.104
set routing-options autonomous-system 65004
set protocols bgp group UNDERLAY family inet any
set protocols bgp group UNDERLAY export EXPORT-L0
set protocols bgp group UNDERLAY local-as 65004
set protocols bgp group UNDERLAY neighbor 5.5.5.1 description BL-1
set protocols bgp group UNDERLAY neighbor 5.5.5.1 peer-as 65001
set protocols bgp group UNDERLAY neighbor 7.7.7.1 description BL-2
set protocols bgp group UNDERLAY neighbor 7.7.7.1 peer-as 65002
set protocols bgp group UNDERLAY neighbor 13.13.13.2 description LEAF-1
set protocols bgp group UNDERLAY neighbor 13.13.13.2 peer-as 65005
set protocols bgp group UNDERLAY neighbor 14.14.14.2 description LEAF-2
set protocols bgp group UNDERLAY neighbor 14.14.14.2 peer-as 65006
set protocols bgp group UNDERLAY neighbor 15.15.15.2 description LEAF-3
set protocols bgp group UNDERLAY neighbor 15.15.15.2 peer-as 65007
set protocols bgp group UNDERLAY neighbor 16.16.16.2 description LEAF-4
set protocols bgp group UNDERLAY neighbor 16.16.16.2 peer-as 65008
set protocols bgp group UNDERLAY neighbor 17.17.17.2 description LEAF-5
set protocols bgp group UNDERLAY neighbor 17.17.17.2 peer-as 65009

```

Base Configuration on LEAF-1

```

set chassis aggregated-devices ethernet deviCE-3ount 2
set interfaces xe-0/0/0 description "T0 SPINE-1"
set interfaces xe-0/0/0 unit 0 family inet address 8.8.8.2/24
set interfaces xe-0/0/1 description "T0 SPINE-2"
set interfaces xe-0/0/1 unit 0 family inet address 13.13.13.2/24
set interfaces xe-0/0/2 gigether-options 802.3ad ae0
set interfaces xe-0/0/3 gigether-options 802.3ad ae1
set interfaces xe-0/0/4 description "T0 Host-2"
set interfaces xe-0/0/4 unit 0 family ethernet-switching interface-mode trunk
set interfaces xe-0/0/4 unit 0 family ethernet-switching VLAN members VLAN-101
set interfaces xe-0/0/4 unit 0 family ethernet-switching VLAN members VLAN-102
set interfaces ae0 description "T0 CE-1"
set interfaces ae0 esi 00:11:11:11:11:11:11:11:11
set interfaces ae0 esi all-active
set interfaces ae0 aggregated-ether-options lACP active
set interfaces ae0 aggregated-ether-options lACP periodic fast
set interfaces ae0 aggregated-ether-options lACP system-id 00:11:11:11:11:11
set interfaces ae0 unit 0 family ethernet-switching interface-mode trunk
set interfaces ae0 unit 0 family ethernet-switching VLAN members VLAN-101
set interfaces ae0 unit 0 family ethernet-switching VLAN members VLAN-102
set interfaces ae1 description "T0 CE-2"
set interfaces ae1 esi 00:22:22:22:22:22:22:22:22
set interfaces ae1 esi all-active
set interfaces ae1 aggregated-ether-options lACP active
set interfaces ae1 aggregated-ether-options lACP periodic fast
set interfaces ae1 aggregated-ether-options lACP system-id 00:22:22:22:22:22
set interfaces ae1 unit 0 family ethernet-switching interface-mode trunk
set interfaces ae1 unit 0 family ethernet-switching VLAN members VLAN-101
set interfaces ae1 unit 0 family ethernet-switching VLAN members VLAN-102
set interfaces lo0 unit 0 family inet address 105.105.105.105/32
set policy-options policy-statement EXPORT-L0 term 1 from protocol direct
set policy-options policy-statement EXPORT-L0 term 1 then accept
set routing-options router-id 105.105.105.105
set routing-options autonomous-system 65005
set protocols bgp group UNDERLAY family inet any
set protocols bgp group UNDERLAY export EXPORT-L0
set protocols bgp group UNDERLAY local-as 65005
set protocols bgp group UNDERLAY neighbor 8.8.8.1 description SPINE-1
set protocols bgp group UNDERLAY neighbor 8.8.8.1 peer-as 65003
set protocols bgp group UNDERLAY neighbor 13.13.13.1 description SPINE-2
set protocols bgp group UNDERLAY neighbor 13.13.13.1 peer-as 65004
set protocols bgp group OVERLAY type internal
set protocols bgp group OVERLAY local-address 105.105.105.105
set protocols bgp group OVERLAY family evpn signaling
set protocols bgp group OVERLAY local-as 65000
set protocols bgp group OVERLAY neighbor 101.101.101.101 description BL-1
set protocols bgp group OVERLAY neighbor 102.102.102.102 description BL-2
set protocols bgp group OVERLAY neighbor 106.106.106.106 description LEAF-2
set protocols bgp group OVERLAY neighbor 107.107.107.107 description LEAF-3
set protocols bgp group OVERLAY neighbor 108.108.108.108 description LEAF-4
set protocols bgp group OVERLAY neighbor 109.109.109.109 description LEAF-5
set protocols evpn encapsulation vxlan
set protocols evpn extended-vni-list 101
set protocols evpn extended-vni-list 102
set switch-options vteP-source-interface lo0.0

```

```

set switch-options route-distinguisher 105.105.105.105:1
set switch-options vrf-target target:1:1
set vlans VLAN-101 VLAN-id 101
set vlans VLAN-101 vxlan vni 101
set vlans VLAN-102 VLAN-id 102
set vlans VLAN-102 vxlan vni 102

```

Base Configuration on LEAF-2

```

set chassis aggregated-devices ethernet device-count 2
set interfaces xe-0/0/0 description "TO SPINE-1"
set interfaces xe-0/0/0 unit 0 family inet address 9.9.9.2/24
set interfaces xe-0/0/1 description "TO SPINE-2"
set interfaces xe-0/0/1 unit 0 family inet address 14.14.14.2/24
set interfaces xe-0/0/2 gigether-options 802.3ad ae0
set interfaces xe-0/0/3 gigether-options 802.3ad ae1
set interfaces xe-0/0/4 description "TO Host-4"
set interfaces xe-0/0/4 unit 0 family ethernet-switching interface-mode trunk
set interfaces xe-0/0/4 unit 0 family ethernet-switching VLAN members VLAN-101
set interfaces xe-0/0/4 unit 0 family ethernet-switching VLAN members VLAN-102
set interfaces ae0 description "TO CE-1"
set interfaces ae0 esi 00:11:11:11:11:11:11:11:11:11
set interfaces ae0 esi all-active
set interfaces ae0 aggregated-ether-options lacp active
set interfaces ae0 aggregated-ether-options lacp periodic fast
set interfaces ae0 aggregated-ether-options lacp system-id 00:11:11:11:11:11
set interfaces ae0 unit 0 family ethernet-switching interface-mode trunk
set interfaces ae0 unit 0 family ethernet-switching VLAN members VLAN-101
set interfaces ae0 unit 0 family ethernet-switching VLAN members VLAN-102
set interfaces ae1 description "TO CE-2"
set interfaces ae1 esi 00:22:22:22:22:22:22:22:22:22
set interfaces ae1 esi all-active
set interfaces ae1 aggregated-ether-options lacp active
set interfaces ae1 aggregated-ether-options lacp periodic fast
set interfaces ae1 aggregated-ether-options lacp system-id 00:22:22:22:22:22
set interfaces ae1 unit 0 family ethernet-switching interface-mode trunk
set interfaces ae1 unit 0 family ethernet-switching VLAN members VLAN-101
set interfaces ae1 unit 0 family ethernet-switching VLAN members VLAN-102
set interfaces lo0 unit 0 family inet address 106.106.106.106/32
set policy-options policy-statement EXPORT-L0 term 1 from protocol direct
set policy-options policy-statement EXPORT-L0 term 1 then accept
set routing-options router-id 106.106.106.106
set routing-options autonomous-system 65006
set protocols bgp group UNDERLAY family inet any
set protocols bgp group UNDERLAY export EXPORT-L0
set protocols bgp group UNDERLAY local-as 65006
set protocols bgp group UNDERLAY neighbor 9.9.9.1 description SPINE-1
set protocols bgp group UNDERLAY neighbor 9.9.9.1 peer-as 65003
set protocols bgp group UNDERLAY neighbor 14.14.14.1 description SPINE-2
set protocols bgp group UNDERLAY neighbor 14.14.14.1 peer-as 65004
set protocols bgp group OVERLAY type internal
set protocols bgp group OVERLAY local-address 106.106.106.106
set protocols bgp group OVERLAY family evpn signaling
set protocols bgp group OVERLAY local-as 65000
set protocols bgp group OVERLAY neighbor 101.101.101.101 description BL-1

```

```

set protocols bgp group OVERLAY neighbor 102.102.102.102 description BL-2
set protocols bgp group OVERLAY neighbor 105.105.105.105 description LEAF-1
set protocols bgp group OVERLAY neighbor 107.107.107.107 description LEAF-3
set protocols bgp group OVERLAY neighbor 108.108.108.108 description LEAF-4
set protocols bgp group OVERLAY neighbor 109.109.109.109 description LEAF-5
set protocols evpn encapsulation vxlan
set protocols evpn extended-vni-list 101
set protocols evpn extended-vni-list 102
set switch-options vtep-source-interface lo0.0
set switch-options route-distinguisher 106.106.106.106:1
set switch-options vrf-target target:1:1
set vlans VLAN-101 VLAN-id 101
set vlans VLAN-101 vxlan vni 101
set vlans VLAN-102 VLAN-id 102
set vlans VLAN-102 vxlan vni 102

```

Base Configuration on LEAF-3

```

set chassis aggregated-devices ethernet device-count 2
set interfaces xe-0/0/0 description "TO SPINE-1"
set interfaces xe-0/0/0 unit 0 family inet address 10.10.10.2/24
set interfaces xe-0/0/1 description "TO SPINE-2"
set interfaces xe-0/0/1 unit 0 family inet address 15.15.15.2/24
set interfaces xe-0/0/2 gigether-options 802.3ad ae0
set interfaces ae0 description "TO CE-3"
set interfaces ae0 esi 00:33:33:33:33:33:33:33:33:33
set interfaces ae0 esi all-active
set interfaces ae0 aggregated-ether-options lacp active
set interfaces ae0 aggregated-ether-options lacp periodic fast
set interfaces ae0 aggregated-ether-options lacp system-id 00:33:33:33:33:33
set interfaces ae0 unit 0 family ethernet-switching interface-mode trunk
set interfaces ae0 unit 0 family ethernet-switching VLAN members VLAN-101
set interfaces ae0 unit 0 family ethernet-switching VLAN members VLAN-102
set interfaces lo0 unit 0 family inet address 107.107.107.107/32
set policy-options policy-statement EXPORT-L0 term 1 from protocol direct
set policy-options policy-statement EXPORT-L0 term 1 then accept
set routing-options router-id 107.107.107.107
set routing-options autonomous-system 65007
set protocols bgp group UNDERLAY family inet any
set protocols bgp group UNDERLAY export EXPORT-L0
set protocols bgp group UNDERLAY local-as 65007
set protocols bgp group UNDERLAY neighbor 10.10.10.1 description SPINE-1
set protocols bgp group UNDERLAY neighbor 10.10.10.1 peer-as 65003
set protocols bgp group UNDERLAY neighbor 15.15.15.1 description SPINE-2
set protocols bgp group UNDERLAY neighbor 15.15.15.1 peer-as 65004
set protocols bgp group OVERLAY type internal
set protocols bgp group OVERLAY local-address 107.107.107.107
set protocols bgp group OVERLAY family evpn signaling
set protocols bgp group OVERLAY local-as 65000
set protocols bgp group OVERLAY neighbor 101.101.101.101 description BL-1
set protocols bgp group OVERLAY neighbor 102.102.102.102 description BL-2
set protocols bgp group OVERLAY neighbor 105.105.105.105 description LEAF-1
set protocols bgp group OVERLAY neighbor 106.106.106.106 description LEAF-2
set protocols bgp group OVERLAY neighbor 108.108.108.108 description LEAF-4
set protocols bgp group OVERLAY neighbor 109.109.109.109 description LEAF-5

```

```

set protocols evpn encapsulation vxlan
set protocols evpn extended-vni-list 101
set protocols evpn extended-vni-list 102
set switch-options vtep-source-interface lo0.0
set switch-options route-distinguisher 107.107.107.107:1
set switch-options vrf-target target:1:1
set vlans VLAN-101 VLAN-id 101
set vlans VLAN-101 vxlan vni 101
set vlans VLAN-102 VLAN-id 102
set vlans VLAN-102 vxlan vni 102

```

Base Configuration on LEAF-4

```

set chassis aggregated-devices ethernet device-count 2
set interfaces xe-0/0/0 description "TO SPINE-1"
set interfaces xe-0/0/0 unit 0 family inet address 11.11.11.2/24
set interfaces xe-0/0/1 description "TO SPINE-2"
set interfaces xe-0/0/1 unit 0 family inet address 16.16.16.2/24
set interfaces xe-0/0/2 gigether-options 802.3ad ae0
set interfaces xe-0/0/3 description "TO Host-6"
set interfaces xe-0/0/3 unit 0 family ethernet-switching interface-mode trunk
set interfaces xe-0/0/3 unit 0 family ethernet-switching VLAN members VLAN-101
set interfaces xe-0/0/3 unit 0 family ethernet-switching VLAN members VLAN-102
set interfaces ae0 description "TO CE-3"
set interfaces ae0 esi 00:33:33:33:33:33:33:33
set interfaces ae0 esi all-active
set interfaces ae0 aggregated-ether-options lACP active
set interfaces ae0 aggregated-ether-options lACP periodic fast
set interfaces ae0 aggregated-ether-options lACP system-id 00:33:33:33:33:33
set interfaces ae0 unit 0 family ethernet-switching interface-mode trunk
set interfaces ae0 unit 0 family ethernet-switching VLAN members VLAN-101
set interfaces ae0 unit 0 family ethernet-switching VLAN members VLAN-102
set interfaces lo0 unit 0 family inet address 108.108.108.108/32
set policy-options policy-statement EXPORT-LO term 1 from protocol direct
set policy-options policy-statement EXPORT-LO term 1 then accept
set routing-options router-id 108.108.108.108
set routing-options autonomous-system 65008
set protocols bgp group UNDERLAY family inet any
set protocols bgp group UNDERLAY export EXPORT-LO
set protocols bgp group UNDERLAY local-as 65008
set protocols bgp group UNDERLAY neighbor 11.11.11.1 description SPINE-1
set protocols bgp group UNDERLAY neighbor 11.11.11.1 peer-as 65003
set protocols bgp group UNDERLAY neighbor 16.16.16.1 description SPINE-2
set protocols bgp group UNDERLAY neighbor 16.16.16.1 peer-as 65004
set protocols bgp group OVERLAY type internal
set protocols bgp group OVERLAY local-address 108.108.108.108
set protocols bgp group OVERLAY family evpn signaling
set protocols bgp group OVERLAY local-as 65000
set protocols bgp group OVERLAY neighbor 101.101.101.101 description BL-1
set protocols bgp group OVERLAY neighbor 102.102.102.102 description BL-2
set protocols bgp group OVERLAY neighbor 105.105.105.105 description LEAF-1
set protocols bgp group OVERLAY neighbor 106.106.106.106 description LEAF-2
set protocols bgp group OVERLAY neighbor 107.107.107.107 description LEAF-3
set protocols bgp group OVERLAY neighbor 109.109.109.109 description LEAF-5
set protocols evpn encapsulation vxlan

```

```

set protocols evpn extended-vni-list 101
set protocols evpn extended-vni-list 102
set switch-options vtep-source-interface lo0.0
set switch-options route-distinguisher 108.108.108.108:1
set switch-options vrf-target target:1:1
set vlans VLAN-101 VLAN-id 101
set vlans VLAN-101 vxlan vni 101
set vlans VLAN-102 VLAN-id 102
set vlans VLAN-102 vxlan vni 102

```

Base Configuration on LEAF-5

```

set interfaces xe-0/0/0 description "TO SPINE-1"
set interfaces xe-0/0/0 unit 0 family inet address 12.12.12.2/24
set interfaces xe-0/0/1 description "TO SPINE-2"
set interfaces xe-0/0/1 unit 0 family inet address 17.17.17.2/24
set interfaces xe-0/0/2 description "TO Host-7"
set interfaces xe-0/0/2 unit 0 family ethernet-switching interface-mode trunk
set interfaces xe-0/0/2 unit 0 family ethernet-switching VLAN members VLAN-102
set interfaces xe-0/0/2 unit 0 family ethernet-switching filter input SH-INGRESS
set interfaces xe-0/0/2 unit 0 family ethernet-switching filter output SH-EGRESS
set interfaces lo0 unit 0 family inet address 109.109.109.109/32
set policy-options prefix-list MCAST-GROUP 225.1.1.1/32
set policy-options policy-statement EXPORT-L0 term 1 from protocol direct
set policy-options policy-statement EXPORT-L0 term 1 then accept
set firewall family ethernet-switching filter SH-INGRESS term 1 from destination-prefix-list MCAST-
GROUP
set firewall family ethernet-switching filter SH-INGRESS term 1 then accept
set firewall family ethernet-switching filter SH-INGRESS term 1 then count SH-INGRESS-COUNTER
set firewall family ethernet-switching filter SH-INGRESS term 2 then accept
set firewall family ethernet-switching filter SH-EGRESS term 1 from destination-prefix-list MCAST-
GROUP
set firewall family ethernet-switching filter SH-EGRESS term 1 then accept
set firewall family ethernet-switching filter SH-EGRESS term 1 then count SH-EGRESS-COUNTER
set firewall family ethernet-switching filter SH-EGRESS term 2 then accept
set routing-options router-id 109.109.109.109
set routing-options autonomous-system 65009
set protocols bgp group UNDERLAY family inet any
set protocols bgp group UNDERLAY export EXPORT-L0
set protocols bgp group UNDERLAY local-as 65009
set protocols bgp group UNDERLAY neighbor 12.12.12.1 description SPINE-1
set protocols bgp group UNDERLAY neighbor 12.12.12.1 peer-as 65003
set protocols bgp group UNDERLAY neighbor 17.17.17.1 description SPINE-2
set protocols bgp group UNDERLAY neighbor 17.17.17.1 peer-as 65004
set protocols bgp group OVERLAY type internal
set protocols bgp group OVERLAY local-address 109.109.109.109
set protocols bgp group OVERLAY family evpn signaling
set protocols bgp group OVERLAY local-as 65000
set protocols bgp group OVERLAY neighbor 101.101.101.101 description BL-1
set protocols bgp group OVERLAY neighbor 102.102.102.102 description BL-2
set protocols bgp group OVERLAY neighbor 105.105.105.105 description LEAF-1
set protocols bgp group OVERLAY neighbor 106.106.106.106 description LEAF-2
set protocols bgp group OVERLAY neighbor 107.107.107.107 description LEAF-3
set protocols bgp group OVERLAY neighbor 108.108.108.108 description LEAF-4
set protocols bgp group OVERLAY neighbor 103.103.103.103 description SPINE-1

```

```

set protocols bgp group OVERLAY neighbor 104.104.104.104 description SPINE-2
set protocols evpn encapsulation vxlan
set protocols evpn extended-vni-list 102
set switch-options vtep-source-interface lo0.0
set switch-options route-distinguisher 109.109.109.109:1
set switch-options vrf-target target:1:1
set vlans VLAN-102 VLAN-id 102
set vlans VLAN-102 vxlan vni 102

```

Base Configuration on BL-1

```

set interfaces xe-0/0/1 description "TO SPINE-1"
set interfaces xe-0/0/1 unit 0 family inet address 4.4.4.1/24
set interfaces xe-0/0/2 description "TO SPINE-2"
set interfaces xe-0/0/2 unit 0 family inet address 5.5.5.1/24
set interfaces irb unit 101 virtual-gateway-accept-data
set interfaces irb unit 101 virtual-gateway-esi 00:66:66:66:66:66:66:66
set interfaces irb unit 101 virtual-gateway-esi all-active
set interfaces irb unit 101 family inet address 18.18.18.1/24 virtual-gateway-address 18.18.18.100
set interfaces irb unit 102 virtual-gateway-accept-data
set interfaces irb unit 102 virtual-gateway-esi 00:77:77:77:77:77:77:77
set interfaces irb unit 102 virtual-gateway-esi all-active
set interfaces irb unit 102 family inet address 19.19.19.1/24 virtual-gateway-address 19.19.19.100
set interfaces lo0 unit 0 family inet address 101.101.101.101/32
set interfaces lo0 unit 1 family inet address 101.101.101.102/32
set policy-options policy-statement EXPORT-L0 term 1 from protocol direct
set policy-options policy-statement EXPORT-L0 term 1 then accept
set routing-instances VRF-1 instance-type virtual-router
set routing-instances VRF-1 interface irb.101
set routing-instances VRF-1 interface irb.102
set routing-instances VRF-1 interface lo0.1
set routing-instances VRF-1 protocols pim rp local address 101.101.101.102
set routing-instances VRF-1 protocols pim interface all mode sparse
set routing-options router-id 101.101.101.101
set routing-options autonomous-system 65001
set protocols bgp group UNDERLAY family inet any
set protocols bgp group UNDERLAY export EXPORT-L0
set protocols bgp group UNDERLAY local-as 65001
set protocols bgp group UNDERLAY neighbor 4.4.4.2 description SPINE-1
set protocols bgp group UNDERLAY neighbor 4.4.4.2 peer-as 65003
set protocols bgp group UNDERLAY neighbor 5.5.5.2 description SPINE-2
set protocols bgp group UNDERLAY neighbor 5.5.5.2 peer-as 65004
set protocols bgp group OVERLAY type internal
set protocols bgp group OVERLAY local-address 101.101.101.101
set protocols bgp group OVERLAY family evpn signaling
set protocols bgp group OVERLAY local-as 65000
set protocols bgp group OVERLAY neighbor 102.102.102.102 description BL-2
set protocols bgp group OVERLAY neighbor 105.105.105.105 description LEAF-1
set protocols bgp group OVERLAY neighbor 106.106.106.106 description LEAF-2
set protocols bgp group OVERLAY neighbor 107.107.107.107 description LEAF-3
set protocols bgp group OVERLAY neighbor 108.108.108.108 description LEAF-4
set protocols bgp group OVERLAY neighbor 109.109.109.109 description LEAF-5
set protocols evpn encapsulation vxlan
set protocols evpn default-gateway no-gateway-community
set protocols evpn extended-vni-list 101

```

```

set protocols evpn extended-vni-list 102
set switch-options vtep-source-interface lo0.0
set switch-options route-distinguisher 101.101.101.101:1
set switch-options vrf-target target:1:1
set vlans VLAN-101 VLAN-id 101
set vlans VLAN-101 l3-interface irb.101
set vlans VLAN-101 vxlan vni 101
set vlans VLAN-102 VLAN-id 102
set vlans VLAN-102 l3-interface irb.102
set vlans VLAN-102 vxlan vni 102

```

Base Configuration on BL-2

```

set interfaces xe-0/0/1 description "T0 SPINE-1"
set interfaces xe-0/0/1 unit 0 family inet address 6.6.6.1/24
set interfaces xe-0/0/2 description "T0 SPINE-2"
set interfaces xe-0/0/2 unit 0 family inet address 7.7.7.1/24
set interfaces irb unit 101 virtual-gateway-accept-data
set interfaces irb unit 101 virtual-gateway-esi 00:66:66:66:66:66:66:66:66:66
set interfaces irb unit 101 virtual-gateway-esi all-active
set interfaces irb unit 101 family inet address 18.18.18.2/24 virtual-gateway-address 18.18.18.100
set interfaces irb unit 102 virtual-gateway-accept-data
set interfaces irb unit 102 virtual-gateway-esi 00:77:77:77:77:77:77:77:77:77
set interfaces irb unit 102 virtual-gateway-esi all-active
set interfaces irb unit 102 family inet address 19.19.19.2/24 virtual-gateway-address 19.19.19.100
set interfaces lo0 unit 0 family inet address 102.102.102.102/32
set interfaces lo0 unit 1 family inet address 102.102.102.103/32
set policy-options policy-statement EXPORT-L0 term 1 from protocol direct
set policy-options policy-statement EXPORT-L0 term 1 then accept
set routing-instances VRF-1 instance-type virtual-router
set routing-instances VRF-1 interface irb.101
set routing-instances VRF-1 interface irb.102
set routing-instances VRF-1 interface lo0.1
set routing-instances VRF-1 protocols pim rp local address 102.102.102.103
set routing-instances VRF-1 protocols pim interface all mode sparse
set routing-options router-id 102.102.102.102
set routing-options autonomous-system 65002
set protocols bgp group UNDERLAY family inet any
set protocols bgp group UNDERLAY export EXPORT-L0
set protocols bgp group UNDERLAY local-as 65002
set protocols bgp group UNDERLAY neighbor 6.6.6.2 description SPINE-1
set protocols bgp group UNDERLAY neighbor 6.6.6.2 peer-as 65003
set protocols bgp group UNDERLAY neighbor 7.7.7.2 description SPINE-2
set protocols bgp group UNDERLAY neighbor 7.7.7.2 peer-as 65004
set protocols bgp group OVERLAY type internal
set protocols bgp group OVERLAY local-address 102.102.102.102
set protocols bgp group OVERLAY family evpn signaling
set protocols bgp group OVERLAY local-as 65000
set protocols bgp group OVERLAY neighbor 101.101.101.101 description BL-1
set protocols bgp group OVERLAY neighbor 105.105.105.105 description LEAF-1
set protocols bgp group OVERLAY neighbor 106.106.106.106 description LEAF-2
set protocols bgp group OVERLAY neighbor 107.107.107.107 description LEAF-3
set protocols bgp group OVERLAY neighbor 108.108.108.108 description LEAF-4
set protocols bgp group OVERLAY neighbor 109.109.109.109 description LEAF-5
set protocols evpn encapsulation vxlan

```

```

set protocols evpn default-gateway no-gateway-community
set protocols evpn extended-vni-list 101
set protocols evpn extended-vni-list 102
set switch-options vtep-source-interface lo0.0
set switch-options route-distinguisher 102.102.102.102:1
set switch-options vrf-target target:1:1
set vlans VLAN-101 VLAN-id 101
set vlans VLAN-101 l3-interface irb.101
set vlans VLAN-101 vxlan vni 101
set vlans VLAN-102 VLAN-id 102
set vlans VLAN-102 l3-interface irb.102
set vlans VLAN-102 vxlan vni 102

```

Base Configuration on CE-1

```

set chassis aggregated-devices ethernet device-count 1
set interfaces ge-0/0/0 description "T0 Host-1"
set interfaces ge-0/0/0 unit 0 family bridge interface-mode trunk
set interfaces ge-0/0/0 unit 0 family bridge VLAN-id-list 101
set interfaces ge-0/0/0 unit 0 family bridge VLAN-id-list 102
set interfaces ge-0/0/1 gigether-options 802.3ad ae0
set interfaces ge-0/0/2 gigether-options 802.3ad ae0
set interfaces ae0 description "T0 LEAF-1_LEAF-2"
set interfaces ae0 aggregated-ether-options lacp active
set interfaces ae0 aggregated-ether-options lacp periodic fast
set interfaces ae0 unit 0 family bridge interface-mode trunk
set interfaces ae0 unit 0 family bridge VLAN-id-list 101
set interfaces ae0 unit 0 family bridge VLAN-id-list 102
set bridge-domains BD-101 domain-type bridge
set bridge-domains BD-101 VLAN-id 101
set bridge-domains BD-102 domain-type bridge
set bridge-domains BD-102 VLAN-id 102

```

Base Configuration on CE-2

```

set chassis aggregated-devices ethernet device-count 1
set interfaces ge-0/0/0 description "T0 Host-3"
set interfaces ge-0/0/0 unit 0 family bridge interface-mode trunk
set interfaces ge-0/0/0 unit 0 family bridge VLAN-id-list 101
set interfaces ge-0/0/0 unit 0 family bridge VLAN-id-list 102
set interfaces ge-0/0/1 gigether-options 802.3ad ae0
set interfaces ge-0/0/2 gigether-options 802.3ad ae0
set interfaces ae0 description "T0 LEAF-1_LEAF-2"
set interfaces ae0 aggregated-ether-options lacp active
set interfaces ae0 aggregated-ether-options lacp periodic fast
set interfaces ae0 unit 0 family bridge interface-mode trunk
set interfaces ae0 unit 0 family bridge VLAN-id-list 101
set interfaces ae0 unit 0 family bridge VLAN-id-list 102
set bridge-domains BD-101 domain-type bridge
set bridge-domains BD-101 VLAN-id 101
set bridge-domains BD-102 domain-type bridge
set bridge-domains BD-102 VLAN-id 102

```

Base Configuration on CE-3

```
set chassis aggregated-devices ethernet device-count 1
set interfaces ge-0/0/0 description "T0 Host-5"
set interfaces ge-0/0/0 unit 0 family bridge interface-mode trunk
set interfaces ge-0/0/0 unit 0 family bridge VLAN-id-list 101
set interfaces ge-0/0/0 unit 0 family bridge VLAN-id-list 102
set interfaces ge-0/0/1 gigether-options 802.3ad ae0
set interfaces ge-0/0/2 gigether-options 802.3ad ae0
set interfaces ae0 description "T0 LEAF-3_LEAF-4"
set interfaces ae0 aggregated-ether-options lacp active
set interfaces ae0 aggregated-ether-options lacp periodic fast
set interfaces ae0 unit 0 family bridge interface-mode trunk
set interfaces ae0 unit 0 family bridge VLAN-id-list 101
set interfaces ae0 unit 0 family bridge VLAN-id-list 102
set bridge-domains BD-101 domain-type bridge
set bridge-domains BD-101 VLAN-id 101
set bridge-domains BD-102 domain-type bridge
set bridge-domains BD-102 VLAN-id 102
```