cisco Live!

ALL IN

#CiscoLive

The bridge to possible

Overlay Multicast in VXLAN EVPN

Understanding fundamental concepts and architecture

Tarique Shakil
Principal Technical Marketing Engineer Cloud Networking,
Cisco Systems

BRKDCN-3238

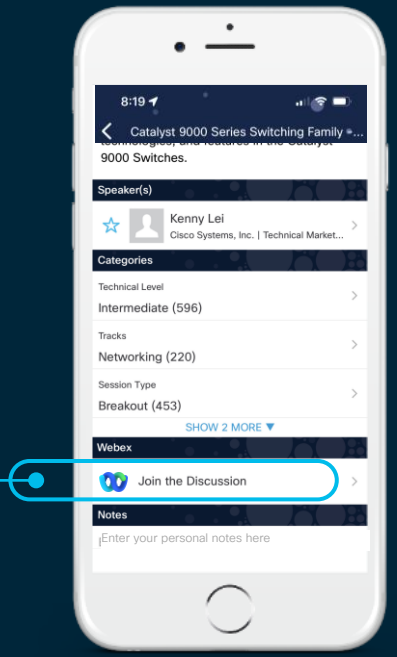
Cisco Webex App

Questions?

Use Cisco Webex App to chat
with the speaker after the session

How

1. Find this session in the Cisco Live Mobile App
2. Click “Join the Discussion”
3. Install the Webex App or go directly to the Webex space
4. Enter messages/questions in the Webex space



https://ciscolive.ciscoevents.com/ciscolivebot/#BRKDCN-3238

Webex spaces will be moderated by the speaker until June 17, 2022.

Agenda

- Multicast Routing Concepts
- VXLAN EVPN Multicast Forwarding
- MP-BGP NGMVPN Concepts
- VXLAN EVPN TRM Architecture
- VXLAN EVPN TRM Forwarding
- Configuring VXLAN EVPN TRM
- Summary

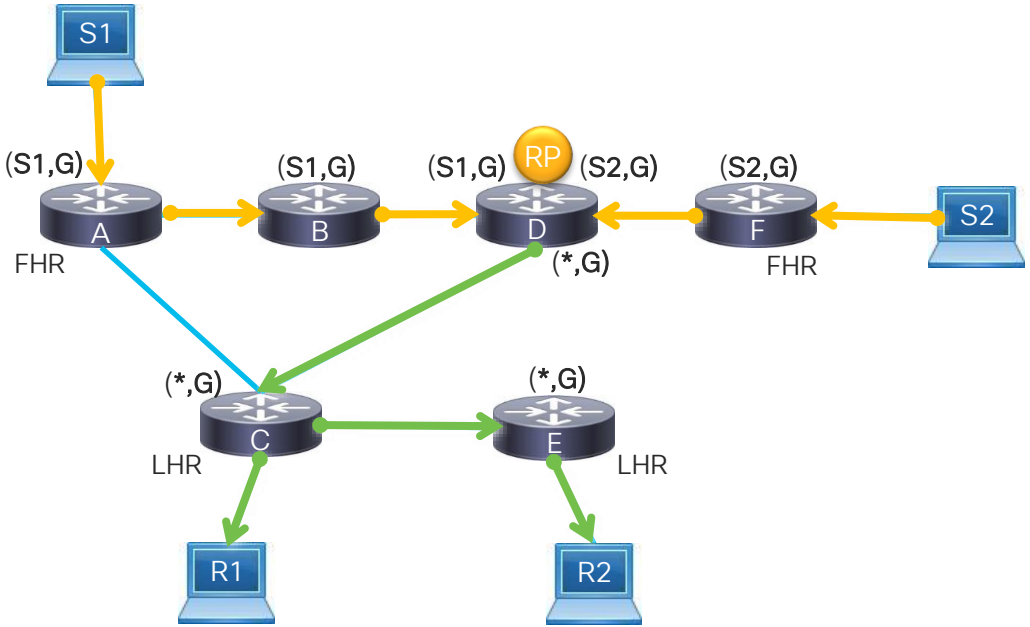
Multicast Routing Concepts

Multicast Terminology

- **First Hop Router (FHR):** Router closest to the source.
- **Last Hop Router (LHR):** Router closest to the receiver.
- **(*,G):** Multicast state in the router's MRIB from any source to group G. Represents a shared tree using a RP.
- **(S,G):** Multicast state in a router's MRIB for a source S to a group G.
- **Incoming Interface (IIF):** Interface towards a RP (*,G) or Source (S,G) based on URIB.
- **Outgoing Interface (OIF):** Interface list that communicate with receivers (received PIM join or IGMP membership).
- **RPF:** Reverse Path Forwarding. Loop avoidance check to the source or RP of Group.

Multicast Distribution Tree (MDT)

Shared Tree – PIM SM



(*,G) (AnySource, Group)

RP PIM Rendezvous Point

Source Tree

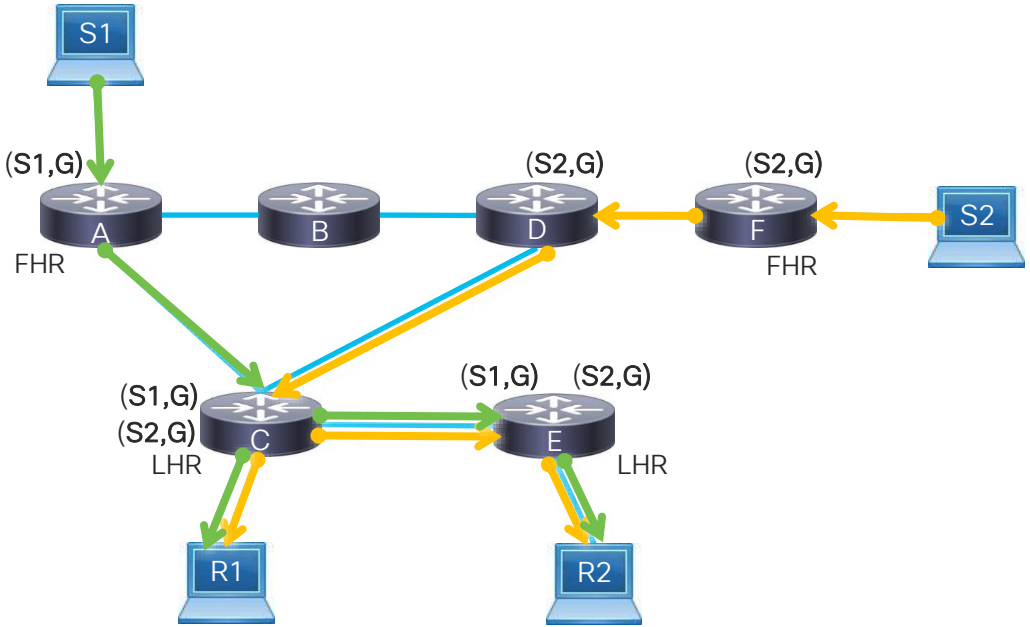
Shared Tree

- Every node should know who is the RP
- (*,G) consumes less memory, but may introduce sub-optimal path from source to all receivers*

*Usually optimized by switching to the source tree (default behavior)

Multicast Distribution Tree (MDT)

Shortest Path Tree – PIM SSM



(S,G) (Source, Group)

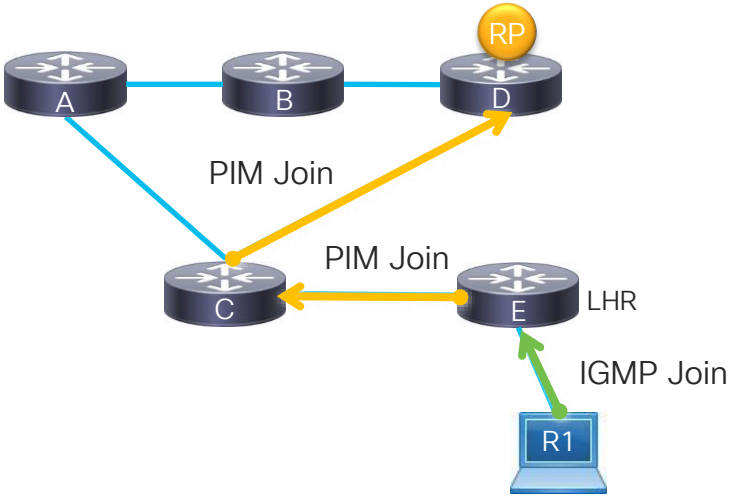
Source Tree (S2, G)

Source Tree (S1, G)

- No need for RP
- (S,G) consumes more memory, but is always optimal. Group address can be reused

PIM Join Triggered by IGMP Report

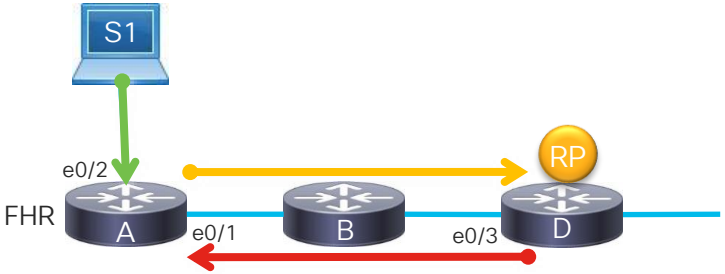
Receiver announcing interest in Multicast group.



- RPF Calculation
  - Based on IP address of tree root (Source or RP)
  - Determines where to send PIM Joins/Prunes
  - PIM Joins continue towards the root to build the multicast tree
  - Multicast data then flows down the tree

PIM Source Registering

FHR Source Registration



RP — PIM Rendezvous Point

Source Starts MCAST Feed

(* ,G)
(S,G) — FHR creates MRIB entry

FHR Generates PIM register to notify RP about new source (unicast tunnel)

RP Generates a PIM register-stop to notify FHR that registration was complete

Multicast Routing Table Entry

show ip mroute

MCAST Source IP Address

MCAST Group Address

Example of (S,G) type route

(10.2.11.4, 239.0.0.11), uptime: 6d09h, mrib ip pim nve

Incoming interface: Ethernet1/1, RPF nbr: 10.2.11.4

Outgoing interface list: (count: 2)

Ethernet1/2, uptime: 6d09h, pim

Ethernet1/3, uptime: 6d09h, pim

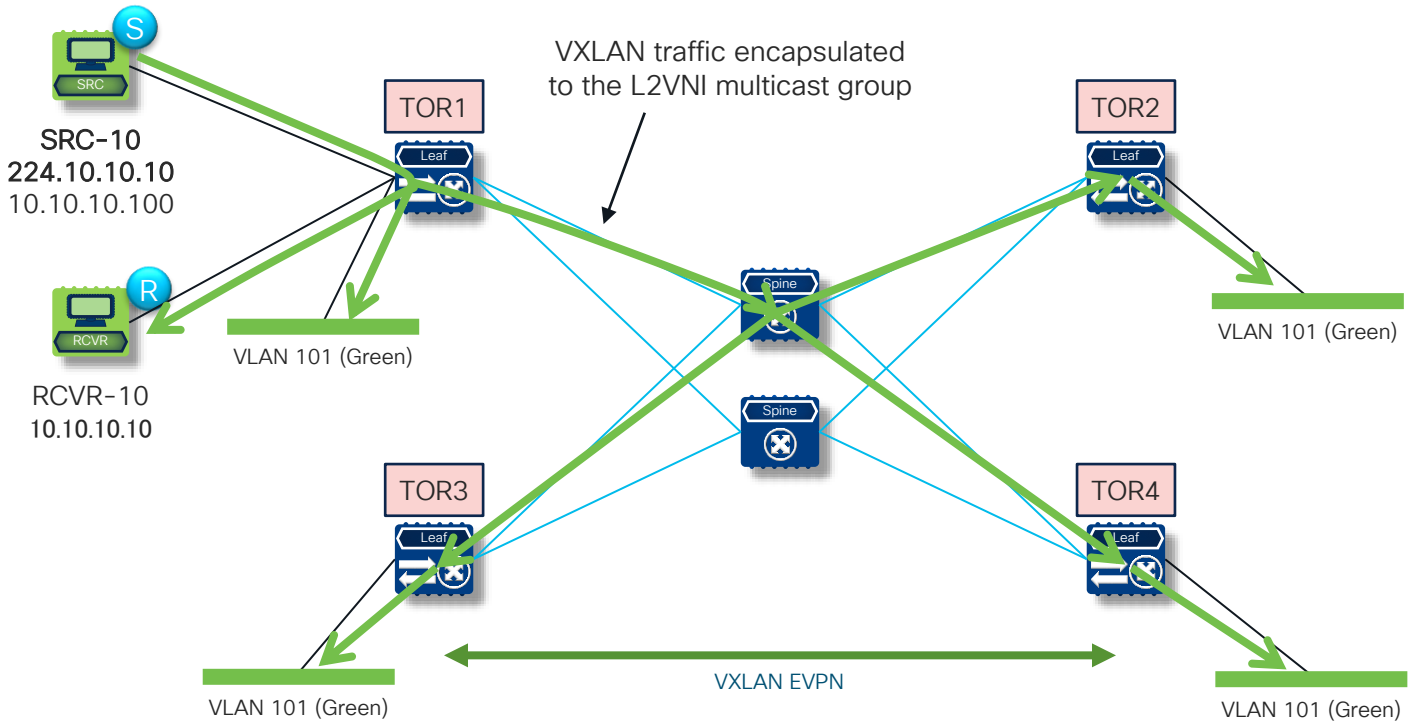
Towards the Source or RP

Towards the Receivers

VXLAN EVPN Multicast Forwarding

Same Subnet Forwarding no IGMP Snooping

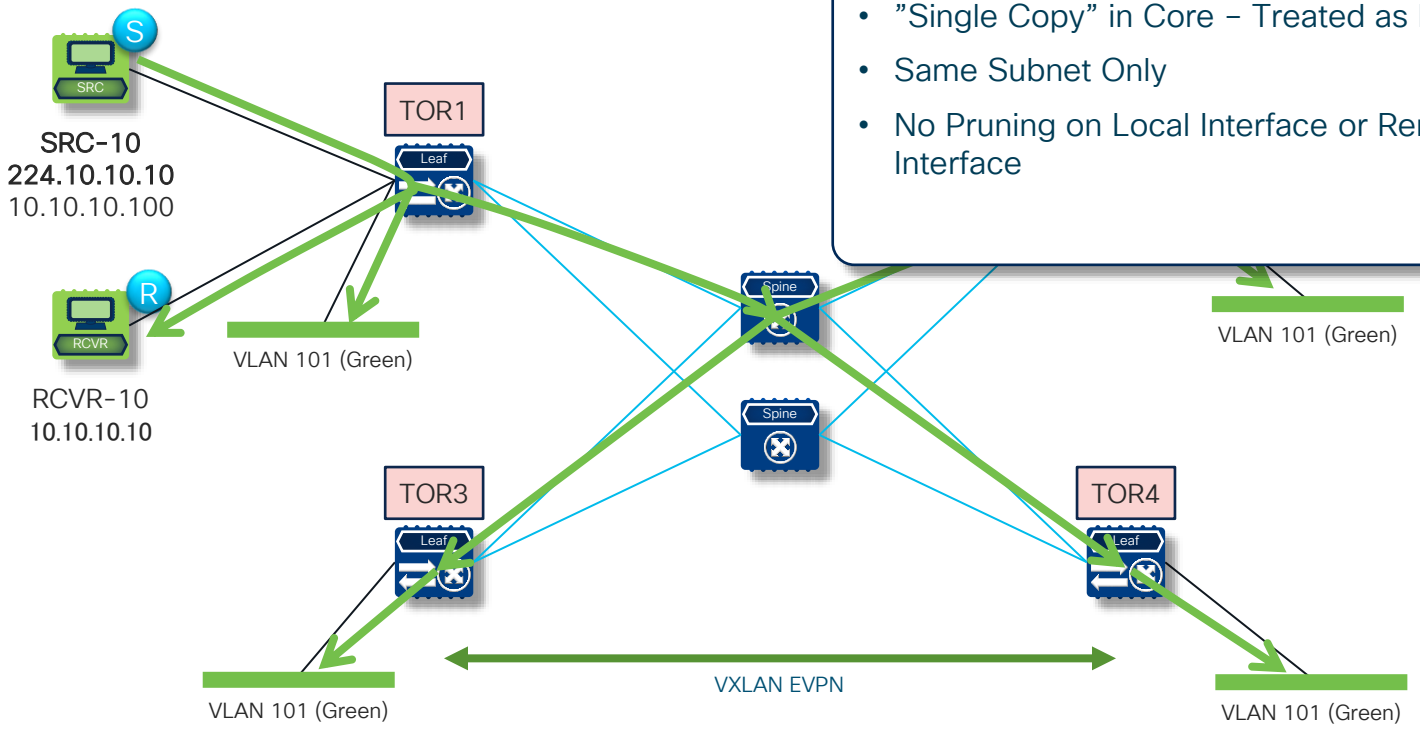
Default Multicast Forwarding in VXLAN Overlay



Same Subnet Forwarding no IGMP Snooping

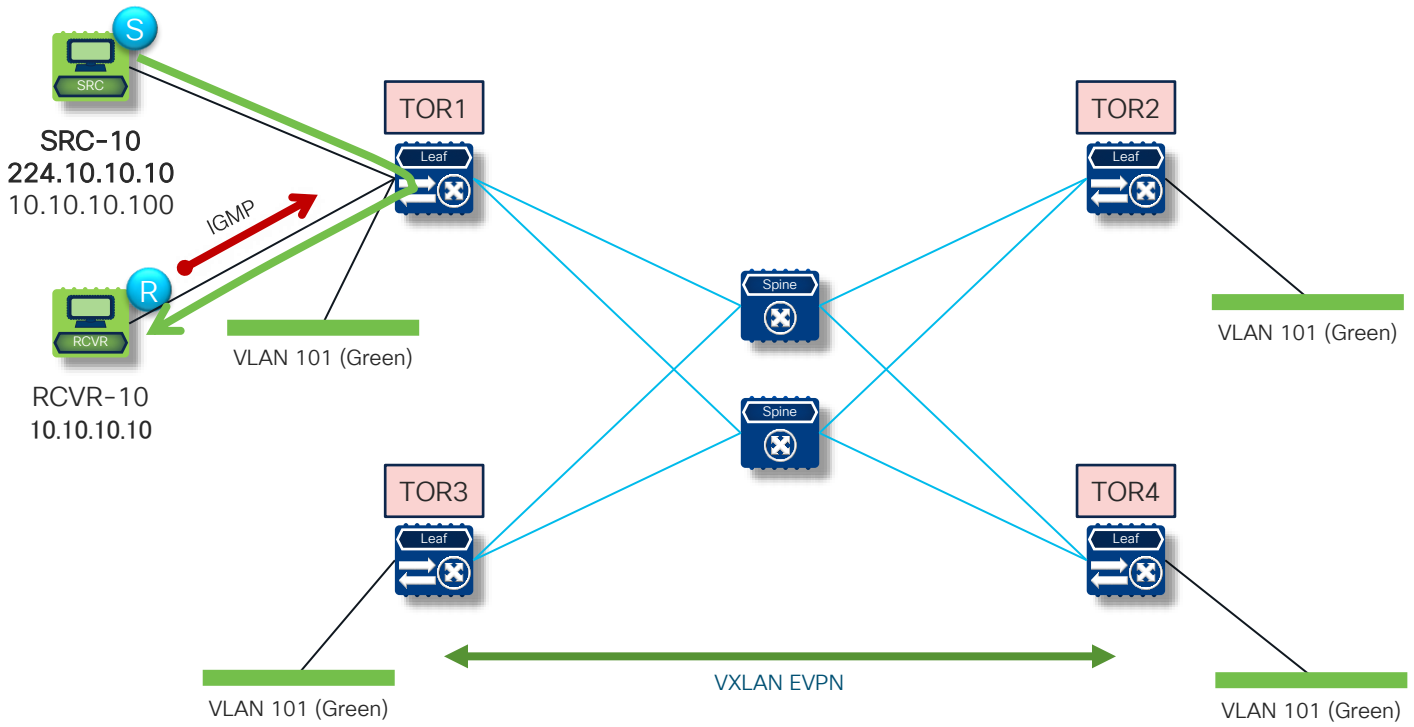
Default Multicast Forwarding in VXLAN Overlay

- ”Single Copy” in Core – Treated as BUM
- Same Subnet Only
- No Pruning on Local Interface or Remote VTEP Interface



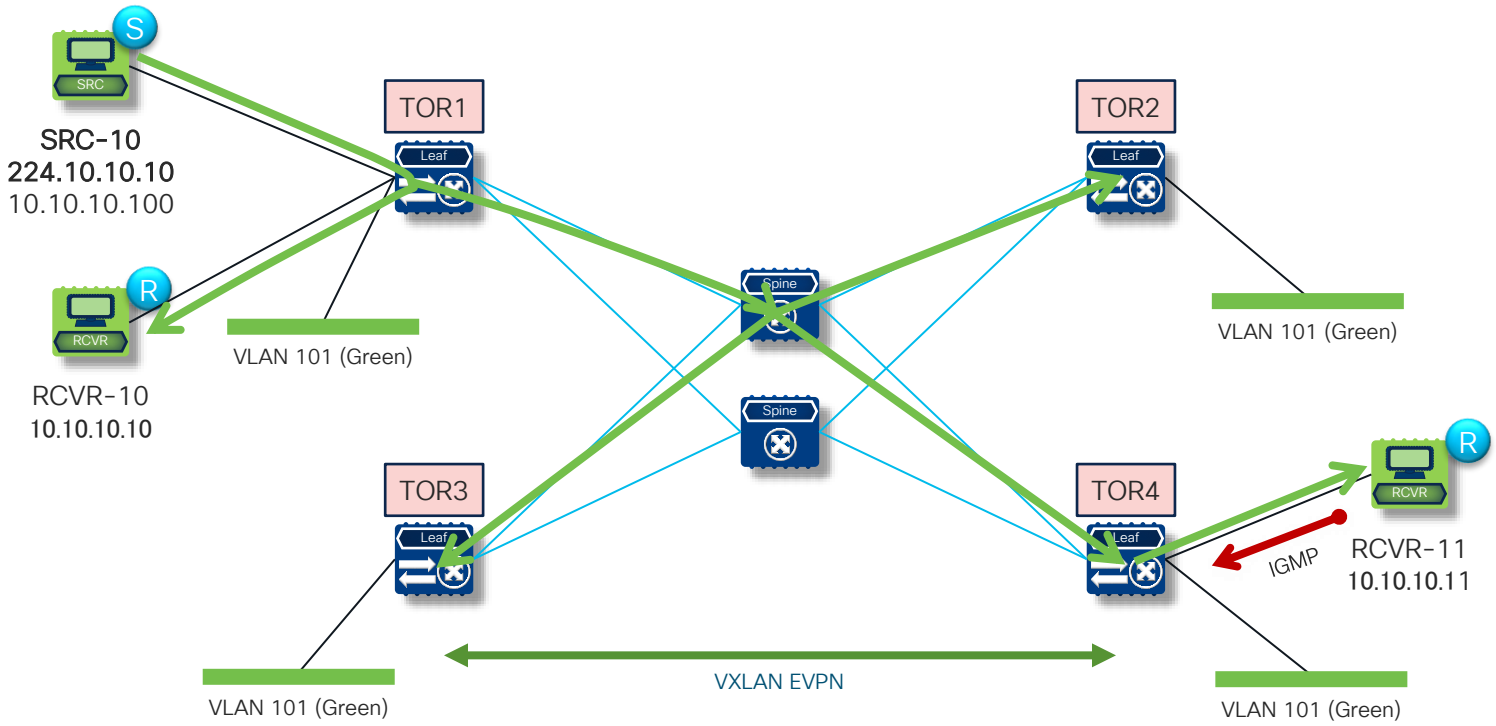
Same Subnet Forwarding with IGMP Snooping

Default Multicast Forwarding in VXLAN Overlay



Same Subnet Forwarding with IGMP Snooping

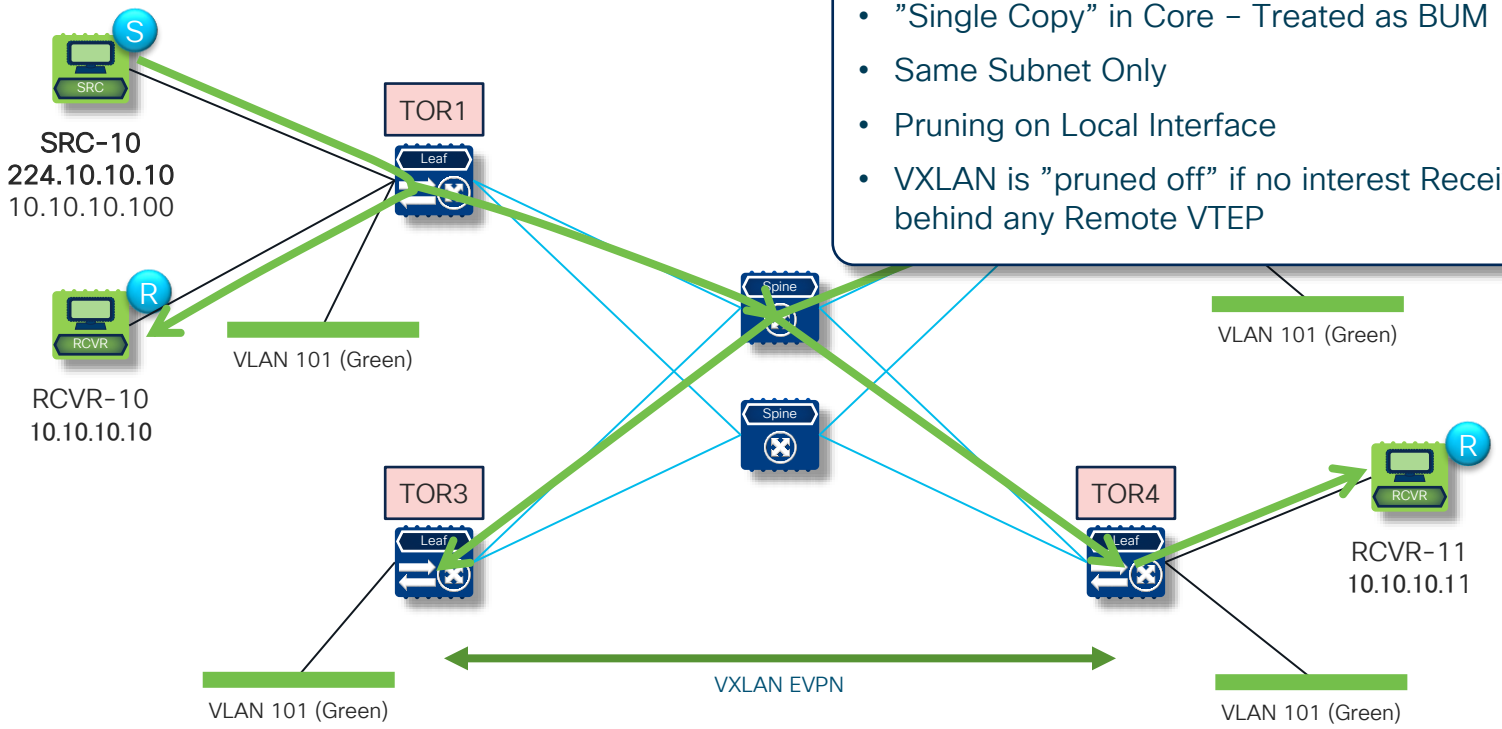
Default Multicast Forwarding in VXLAN Overlay



Same Subnet Forwarding with IGMP Snooping

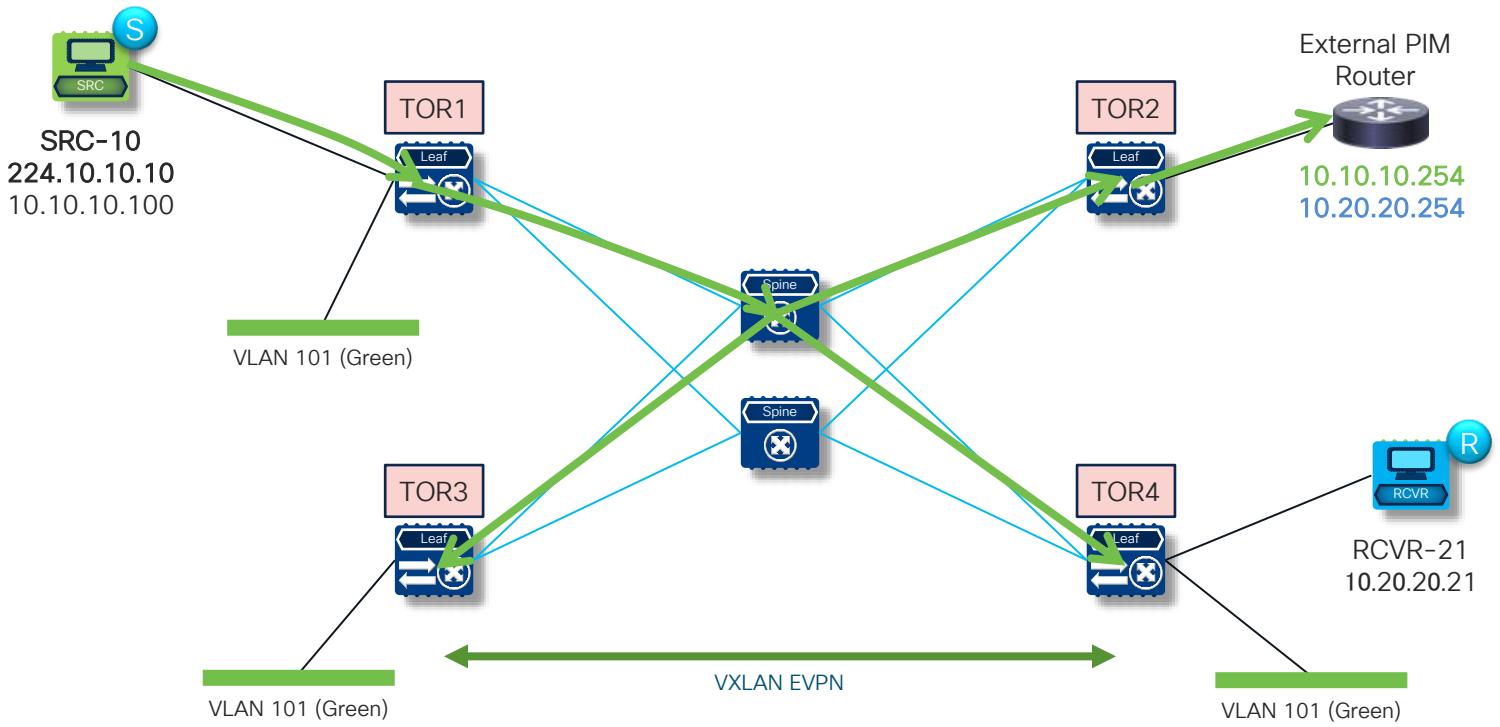
Default Multicast Forwarding in VXLAN Overlay

- ”Single Copy” in Core – Treated as BUM
- Same Subnet Only
- Pruning on Local Interface
- VXLAN is ”pruned off” if no interest Receiver exists behind any Remote VTEP



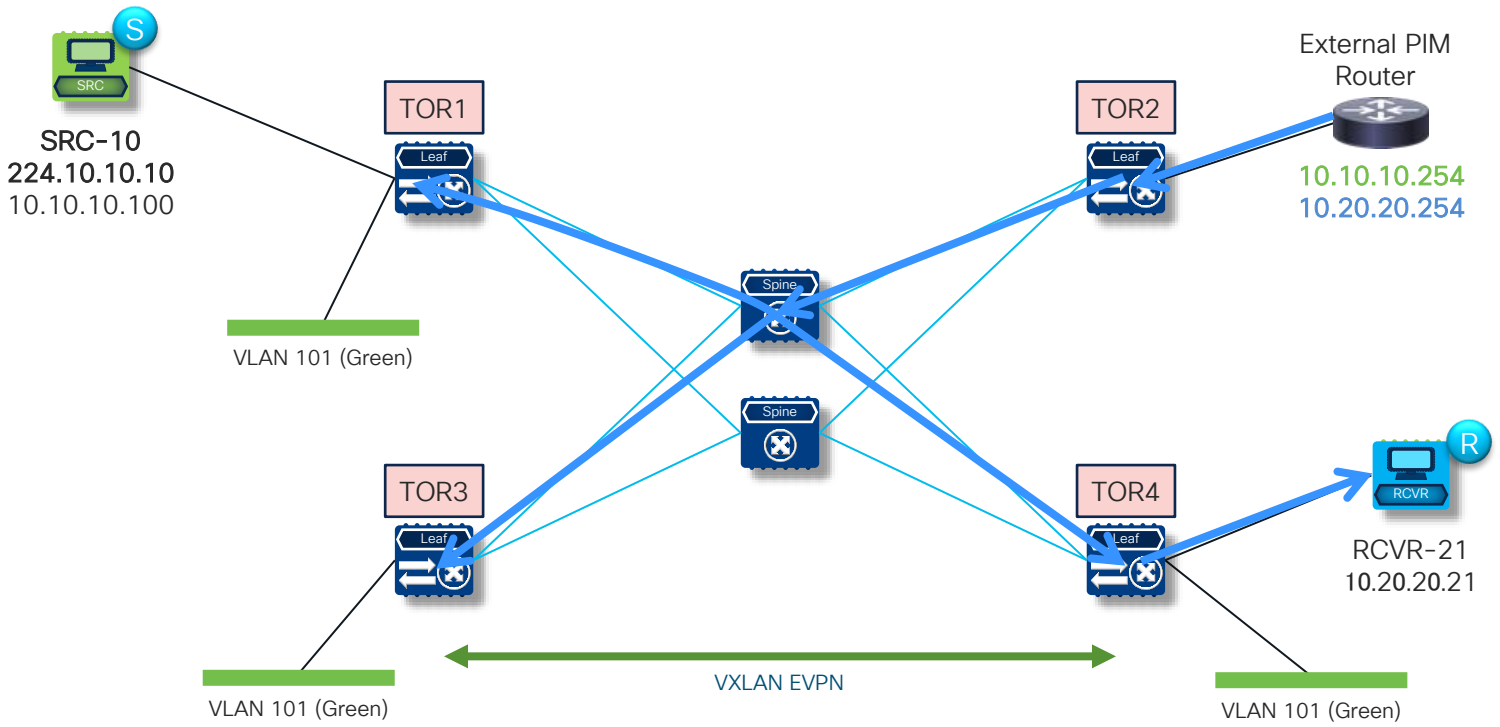
Different Subnet Forwarding – Router on-a-Stick

Default Multicast Forwarding in VXLAN Overlay



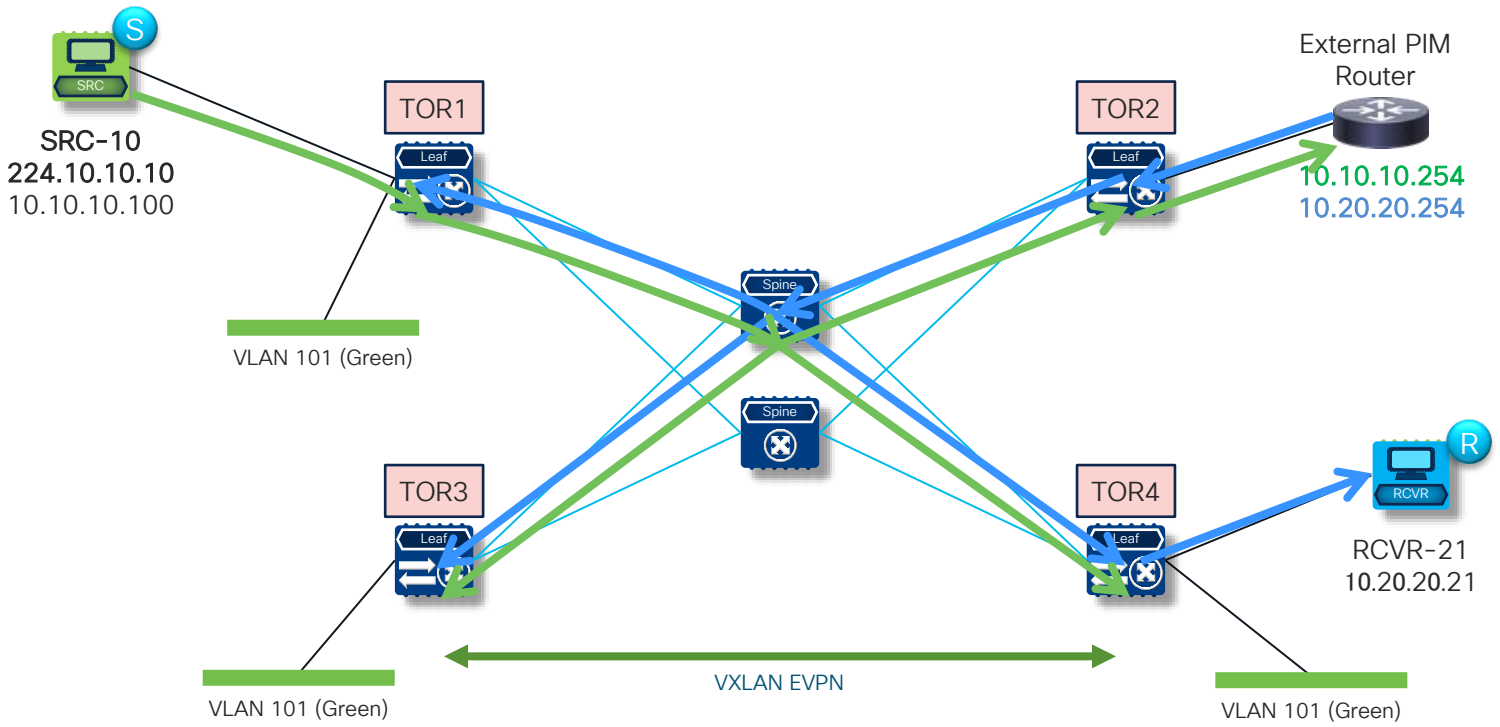
Different Subnet Forwarding – Router on-a-Stick

Default Multicast Forwarding in VXLAN Overlay



Different Subnet Forwarding – Router on-a-Stick

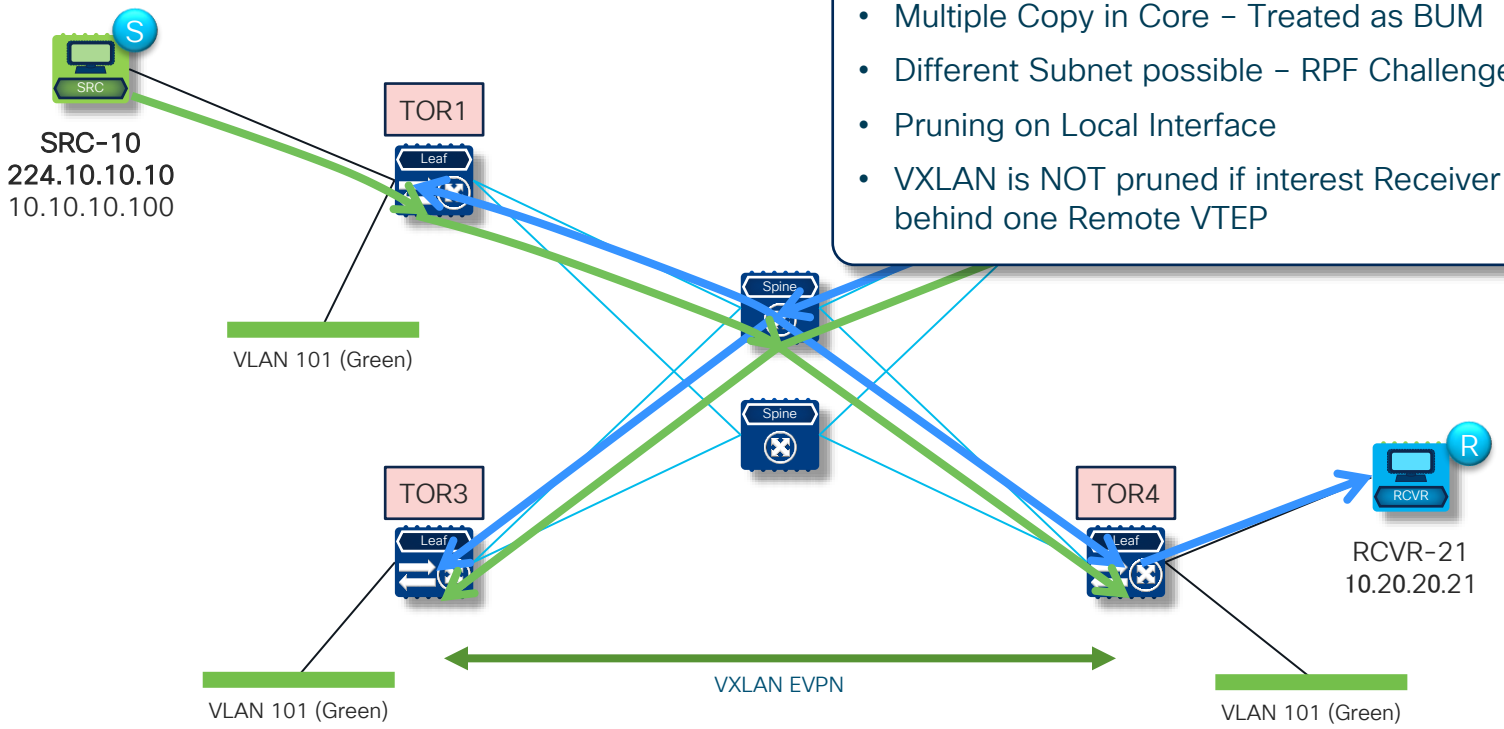
Default Multicast Forwarding in VXLAN Overlay



Different Subnet Forwarding – Router on-a-Stick

Default Multicast Forwarding in VXLAN Overlay

- Multiple Copy in Core – Treated as BUM
- Different Subnet possible – RPF Challenges
- Pruning on Local Interface
- VXLAN is NOT pruned if interest Receiver exists behind one Remote VTEP



MP-BGP NGMVPN Concepts

MP-BGP NGMVPN Terminology

- **Multicast Domain:** A group of VTEPs assigned to the same VRF that have multicast connectivity allowing multicast traffic to be sent between the sites.
- **Multicast VPN(MVPN):** A VRF that can route both unicast and multicast routing.
- **Multicast Distribution Trees (MDT):** A forwarding tree built to carry customer multicast traffic between edge routers connected to common VRFs in the same Multicast Domain.
- **Inclusive Trees:** A single MDT that carries traffic for all multicast traffic in a MVPN. Commonly known as Default-MDT.
- **Selective Trees:** An MDT that carries traffic for a specific or a set of multicast groups within an MVPN. Commonly known as Data-MDT.

MP-BGP NGMVPN Control Plane

- MP-BGP is used to exchange both unicast (AF EVPN) and multicast (AF MVPN) route information in a VXLAN BGP EVPN fabric.
- The RD is an 8-byte or 64-bit value with two parts. <autonomous system number>:<admin assigned value>.
- The RT is an 8-byte BGP extended community consisting of two parts. <AS # | IP Address>:<admin assigned value>.
- Like unicast VPNs, the RT ensures the c-multicast (tenant) routes are only imported to the correct VRFs.

MP-BGP NGMVPN Function

- The auto-discovery of remote PEs (VTEPs) participating in the same MVPN domain. This answers the questions “who are the members of my multicast domain?”.
- Tunnel type and ID information exchange between PEs (VTEPs) for the tunnel used for forwarding c-multicast (tenant) routes. The tunnel used to transport multicast traffic in MVPN is called the provider tunnel and distributed in BGP in an attribute called provider multicast service interface (PMSI). This answers the questions “Which tunnel do I send my multicast traffic on?”
- Exchanging of c-multicast (tenant) routing information. This answers the questions “Which multicast groups can receivers subscribe to and who are the sources for those groups?”

MP-BGP NGMVPN Packet Types

- MVPN Route Type 5 Source Active A-D.
  - Originated by the FHR / VTEP / PE with Active Source

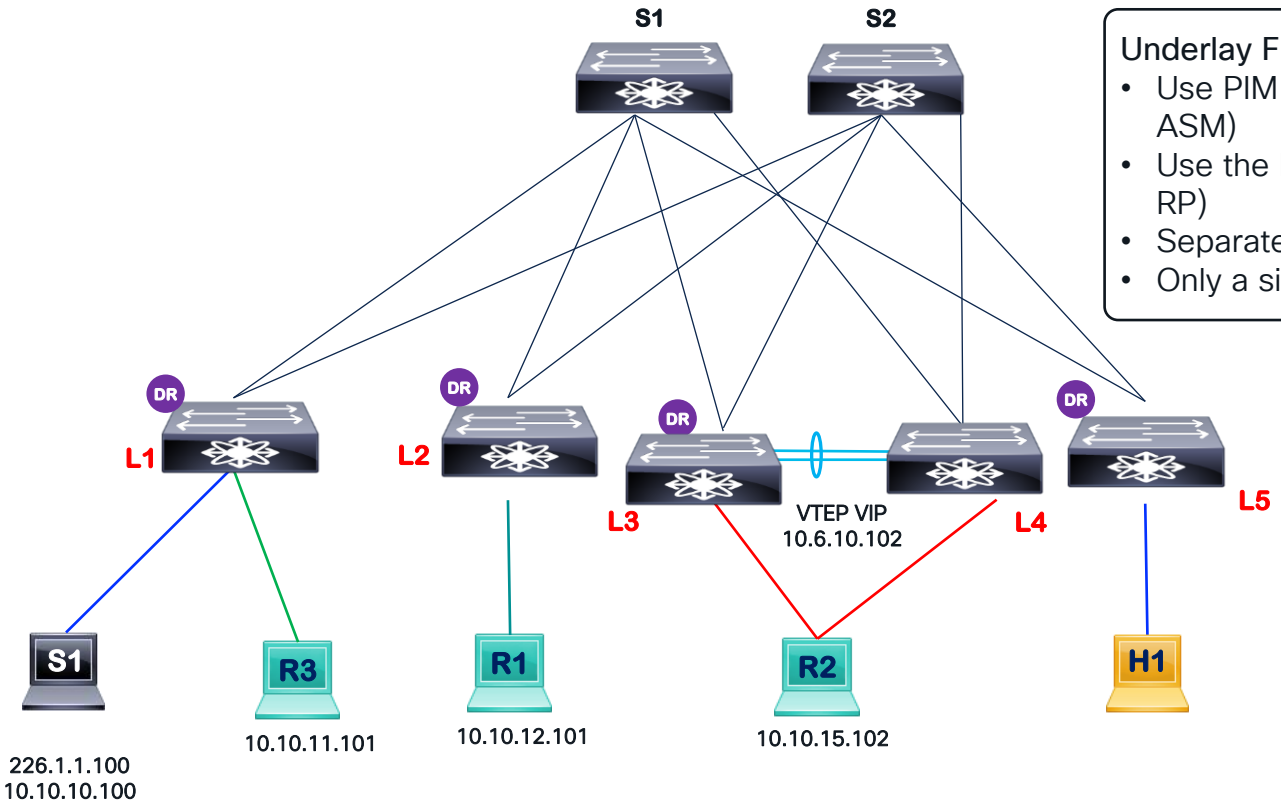
- MVPN Route Type 6 Shared tree Join
  - (*,G) Join Originated by LHR / VTEP/ PE with an interested receiver
  - Used with External RP Configuration

- MVPN Route Type 7 Source Tree Join
  - (S,G) Join by a LHR / VTEP/ PE after receiving a MVPN Type 5 Route

- Nexus 9000 NXOS implementation based on RFC 6513 and 6514

VXLAN EVPN TRM Architecture

VXLAN EVPN TRM Underlay Routing

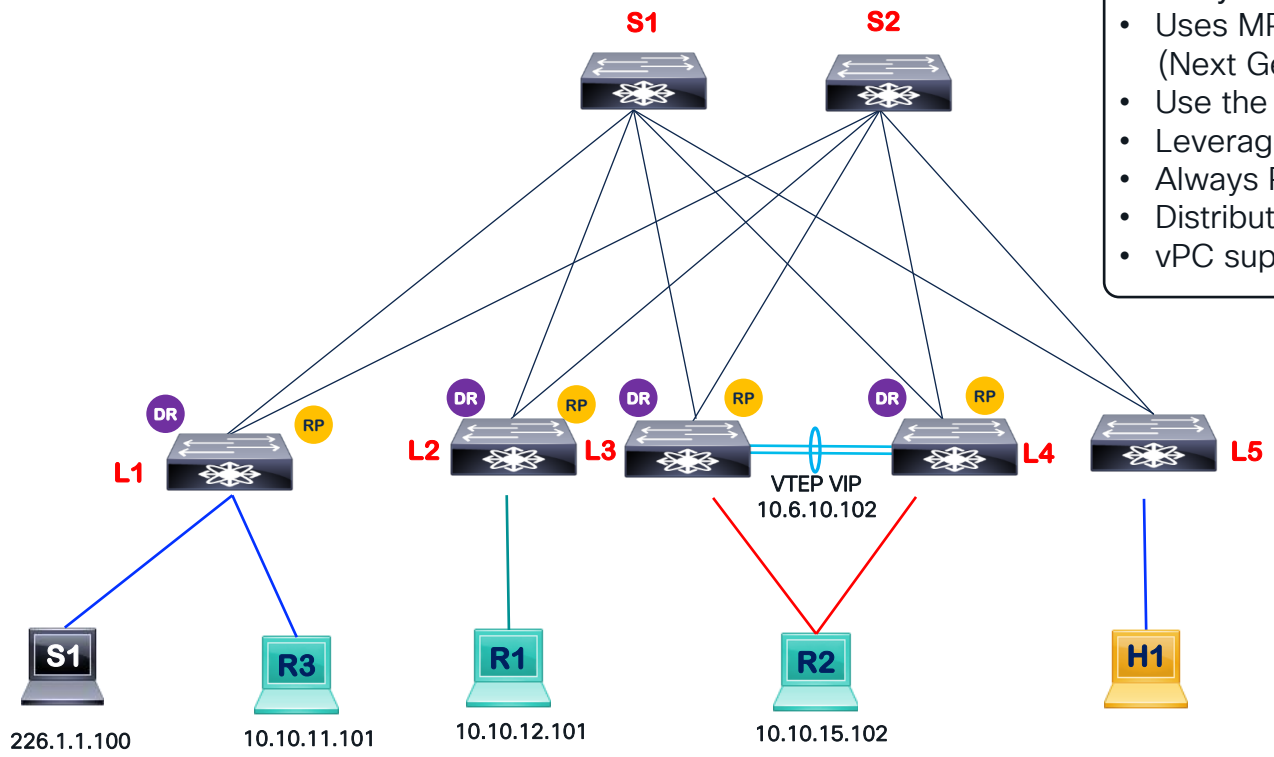


Underlay Functional Details:

- Use PIM based underlay for transport (PIM ASM)
- Use the RP defined in the underlay (PIM Anycast RP)
- Separate MCAST Group for each L3VNI(VRF)
- Only a single packet in the core

DR Designated Router

VXLAN EVPN TRM Overlay Routing



Overlay Functional Details:

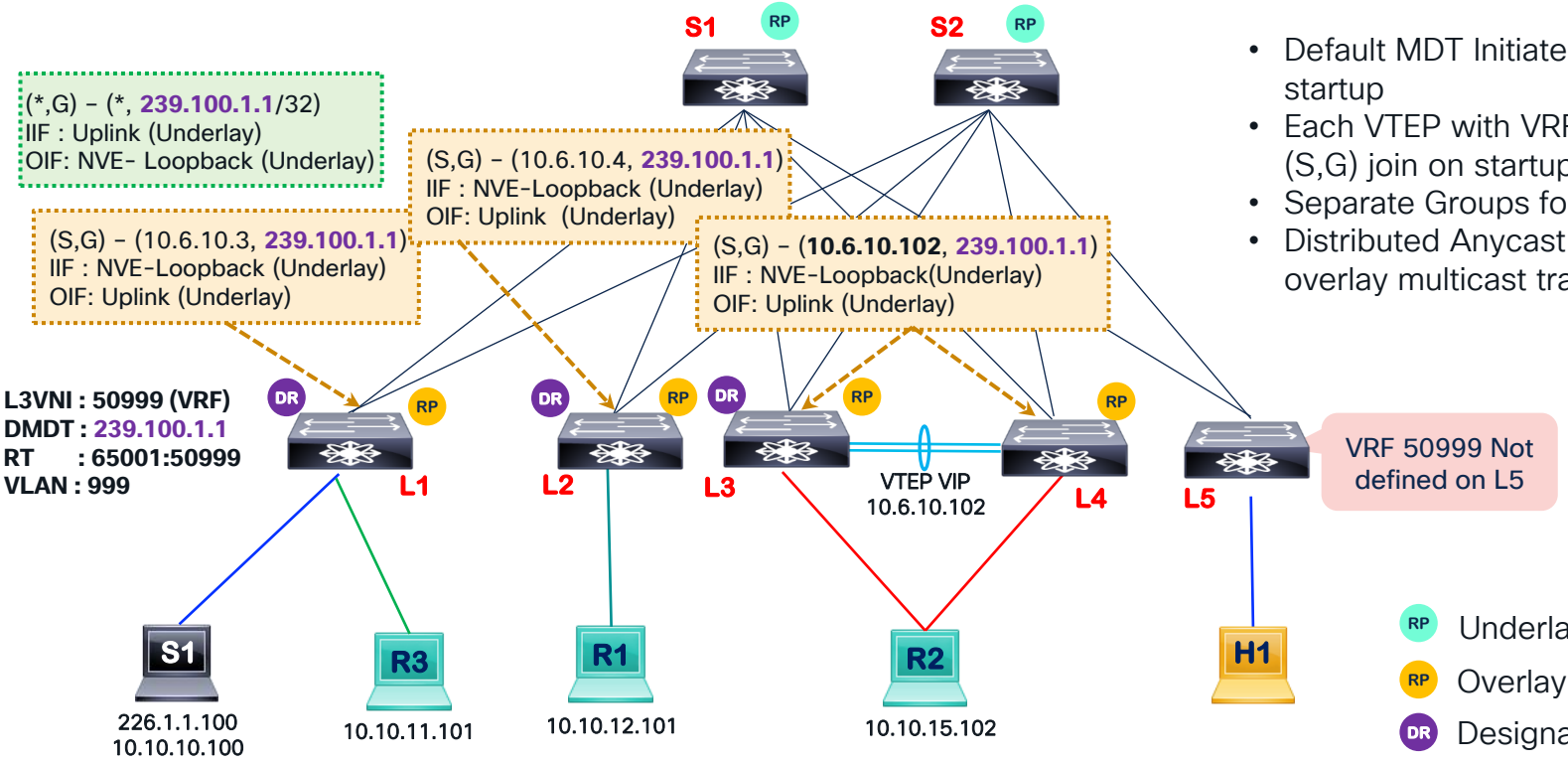
- Uses MP-BGP based ngMVPN control plane (Next Gen Multicast VPN)
- Use the Router Reflector in the Spine
- Leverages **RP-less** (in fabric RP) configuration
- Always Route approach (Per-VLAN config)
- Distributed Anycast Designated Router (DR)
- vPC support and non TRM VTEP Integration

RP Rendezvous Point

DR Designated Router

VXLAN EVPN TRM

Underlay Multicast Routing State



- Default MDT Initiates (*,G) join on VTEP startup
- Each VTEP with VRF configured initiates (S,G) join on startup
- Separate Groups for BUM and MDT
- Distributed Anycast RP on VTEPs for overlay multicast traffic.

RP Underlay Rendezvous Point

RP Overlay Rendezvous Point

DR Designated Router

Default MDT MRIB State

```
abcpod-1-dc1-leaf1# show ip mroute
IP Multicast Routing Table for VRF "default"

(*, 239.100.1.1/32), uptime: 1d20h, nve ip pim
  Incoming interface: Ethernet1/1, RPF nbr: 10.6.10.1
  Outgoing interface list: (count: 1)
    nve1, uptime: 1d20h, nve

(10.6.11.2/32, 239.100.1.1/32), uptime: 1d20h, nve mrib ip pim
  Incoming interface: loopback1, RPF nbr: 10.6.11.2
  Outgoing interface list: (count: 1)
    Ethernet1/1, uptime: 1d20h, pim
```

(*,G) State
G → DMDT

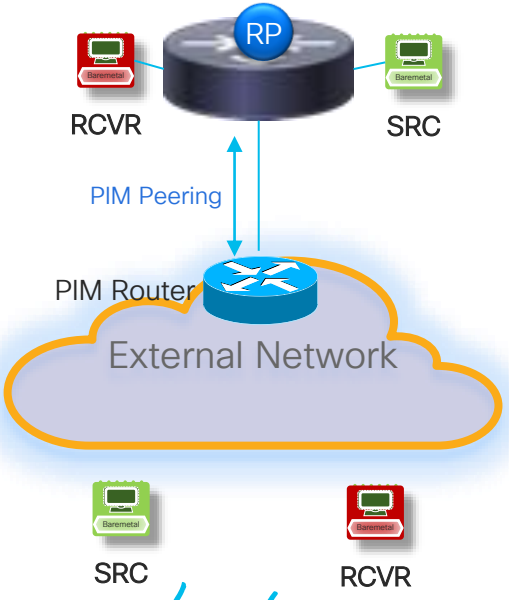
(S,G) State
G → DMDT

Tenant Routed Multicast

RP Deployment Models

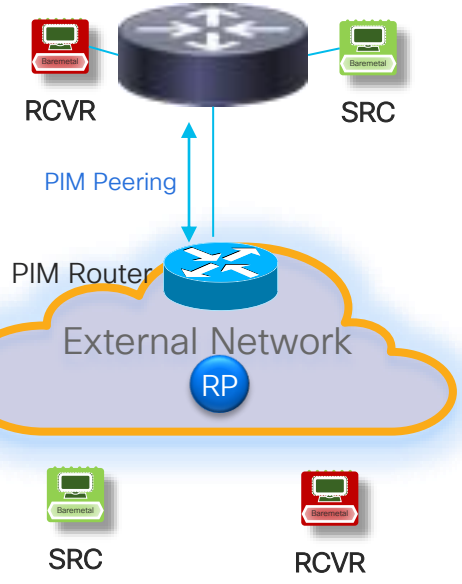
RP-Less

VXLAN EVPN Fabric as logical PIM Router (playing also the RP role)



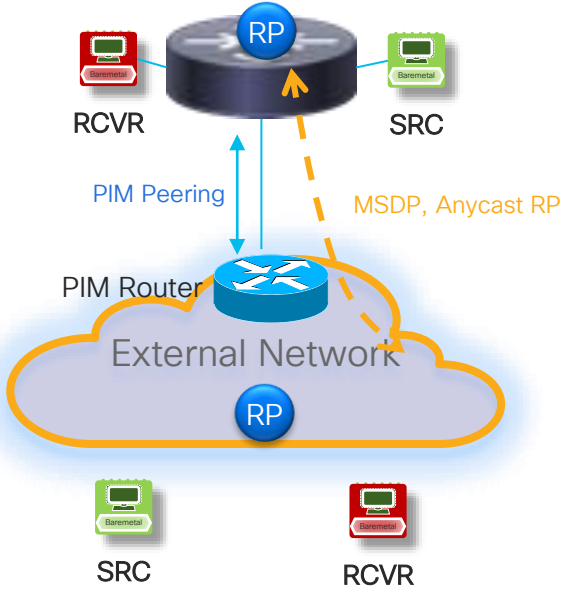
External RP

VXLAN EVPN Fabric as logical PIM Router



RP Anywhere

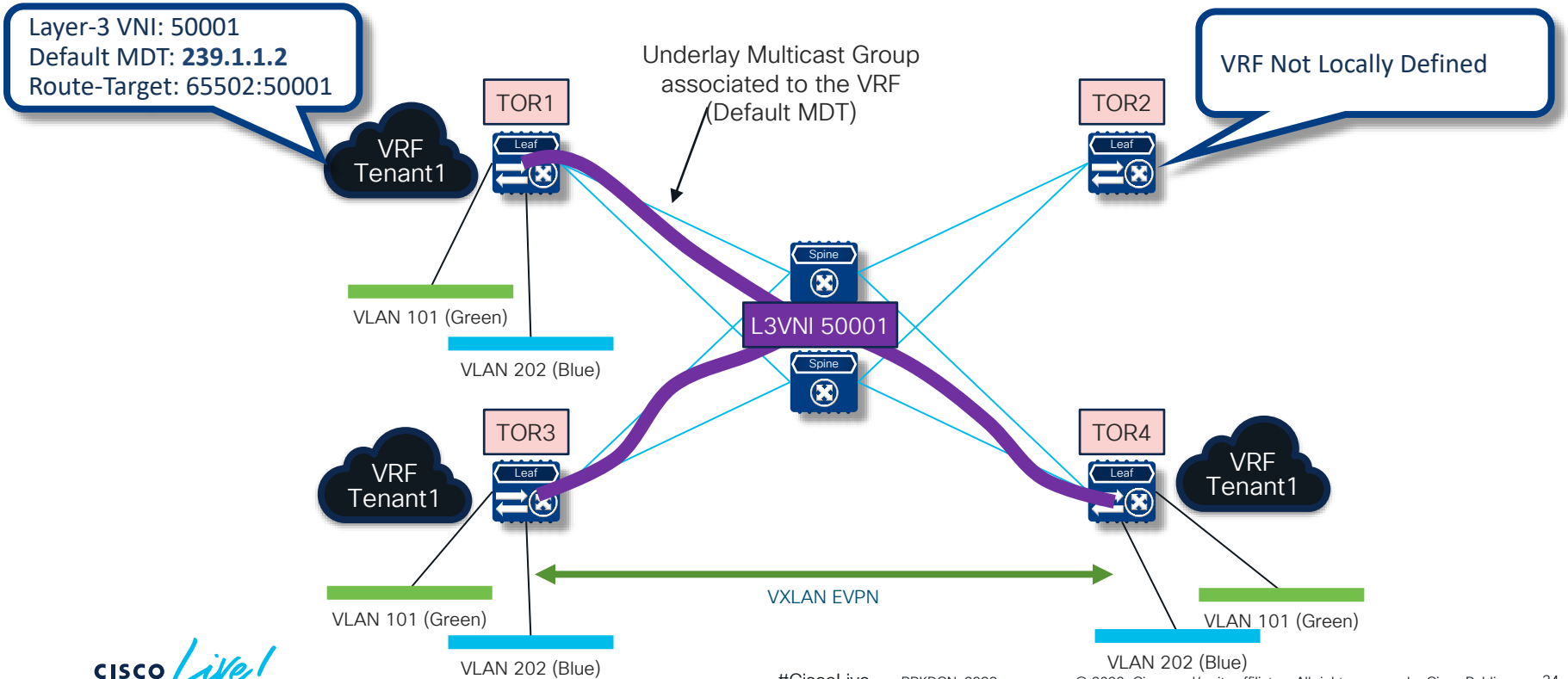
VXLAN EVPN Fabric as logical PIM Router (playing also the RP role)



VXLAN EVPN TRM Forwarding

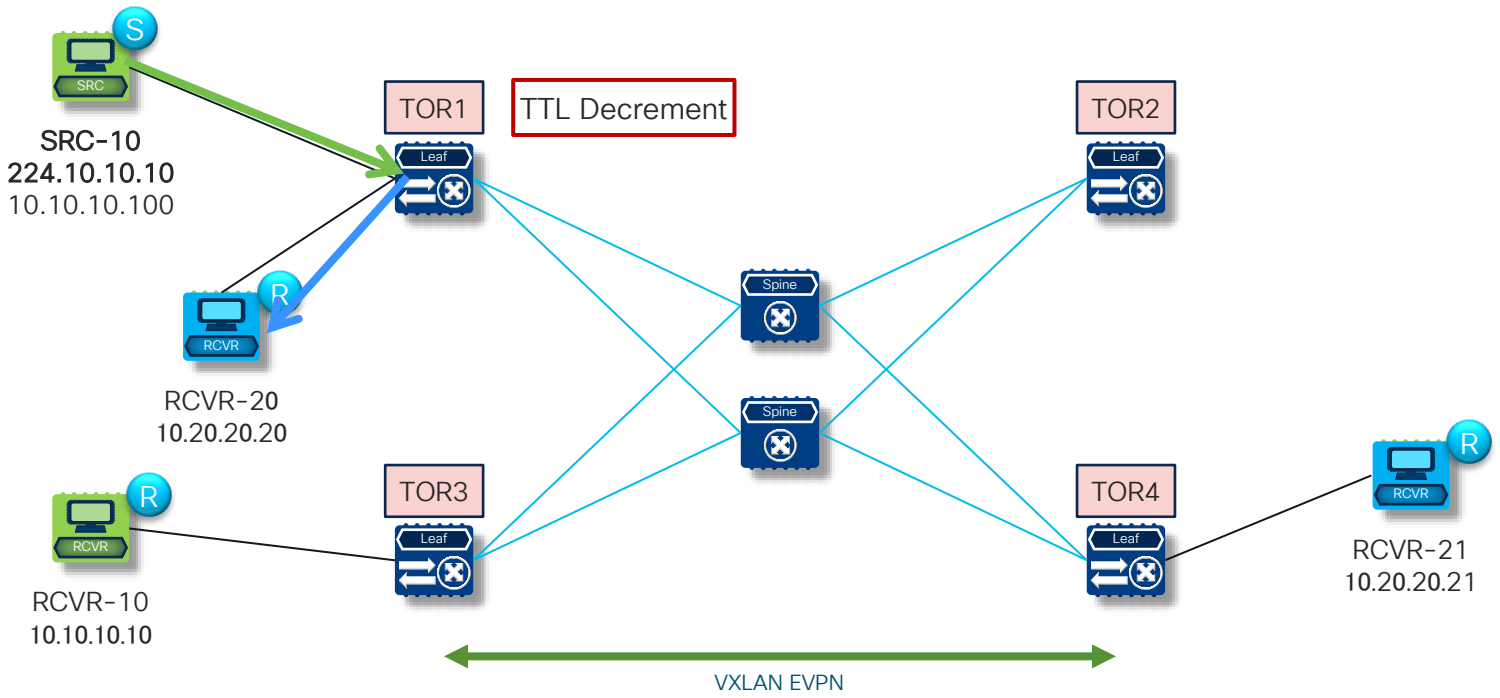
TRM Forwarding

Always Route Approach



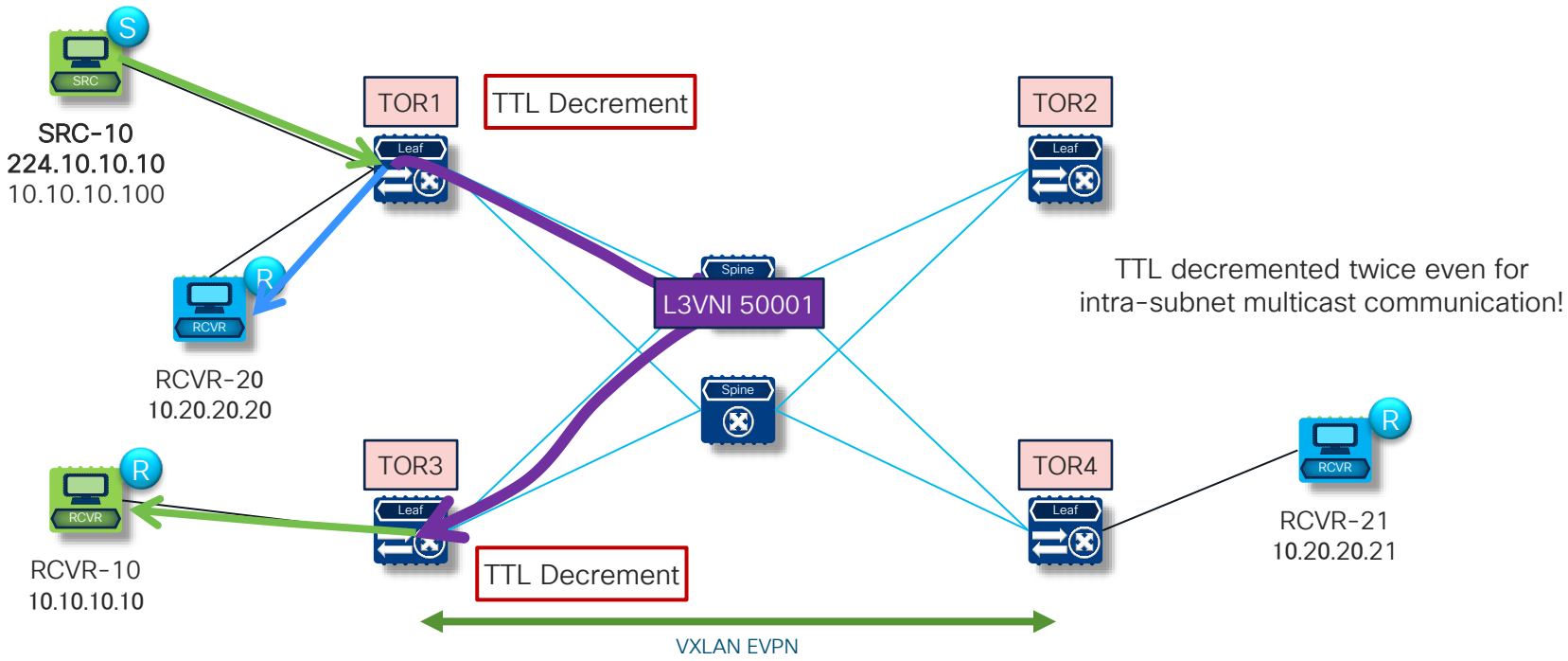
TRM Forwarding

Always Route Approach



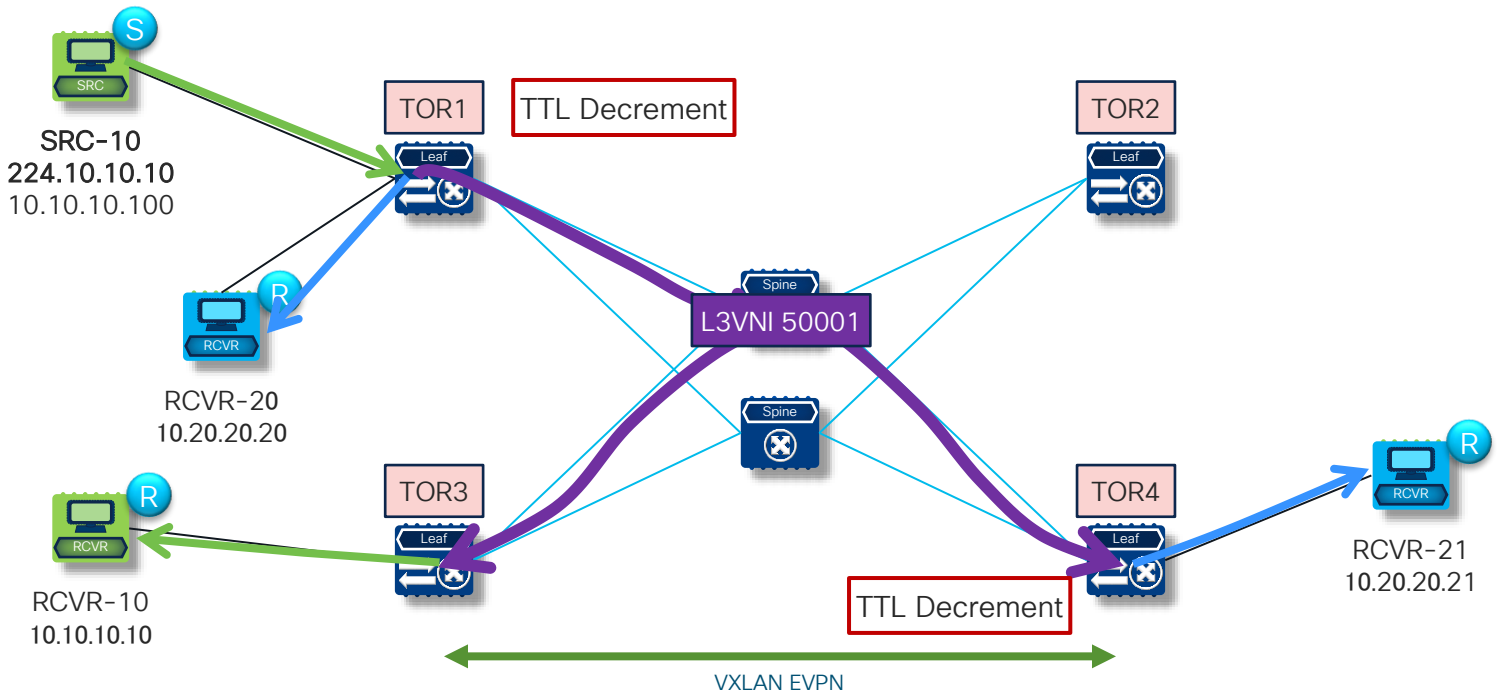
TRM Forwarding

Always Route Approach



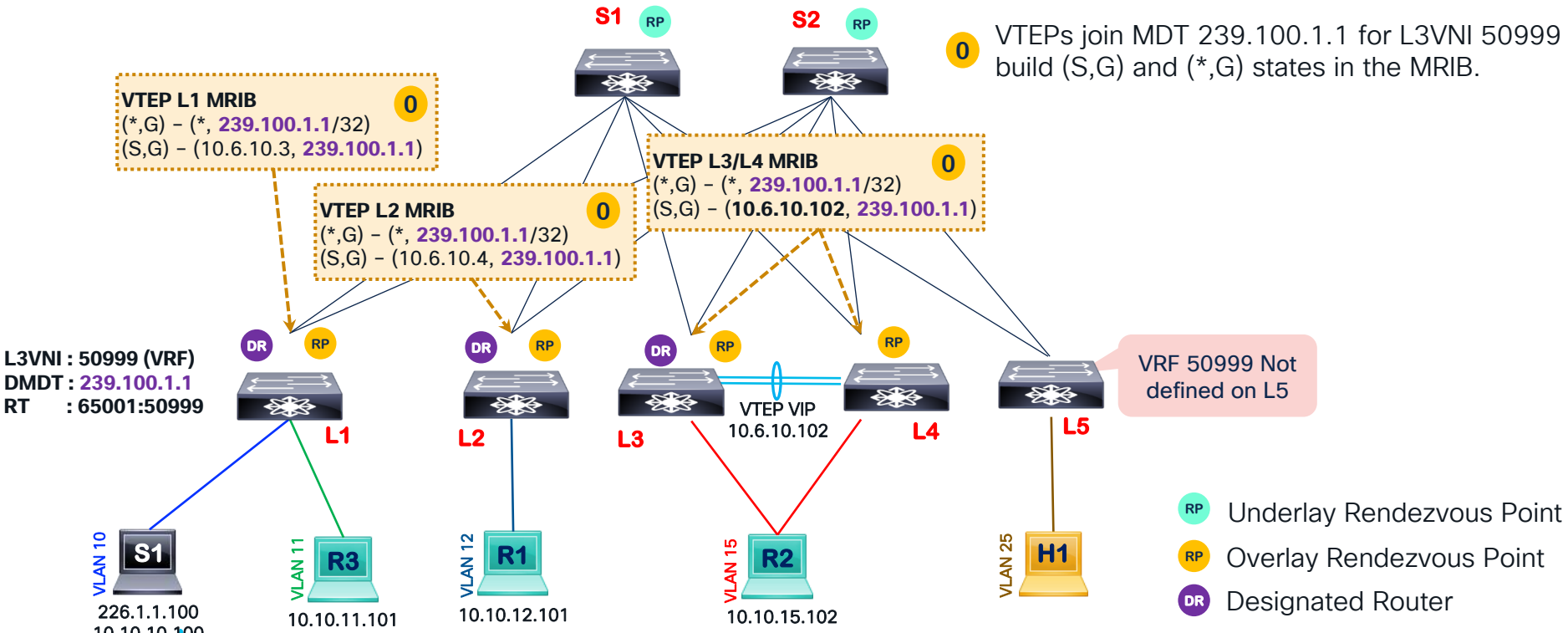
TRM Forwarding

Always Route Approach



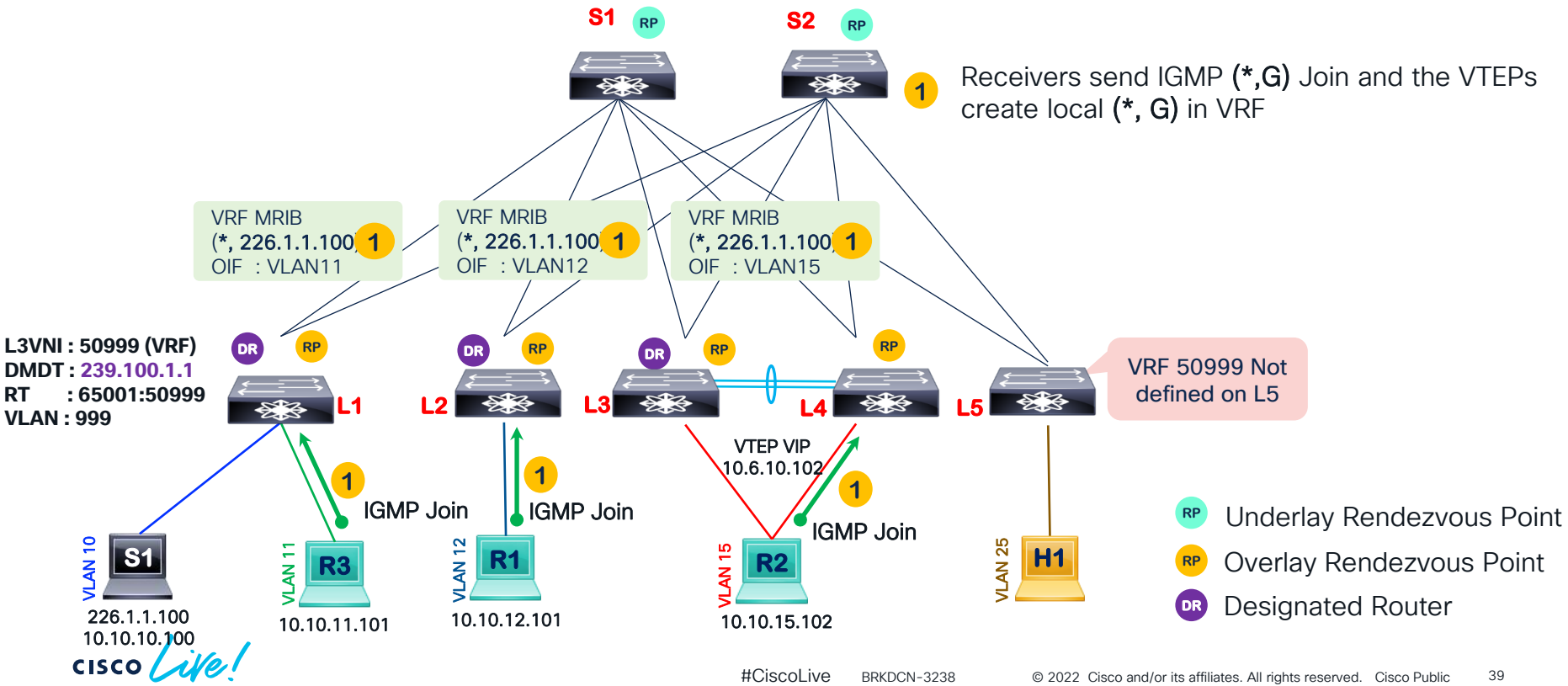
TRM Routing with Anycast RP

Underlay Multicast State



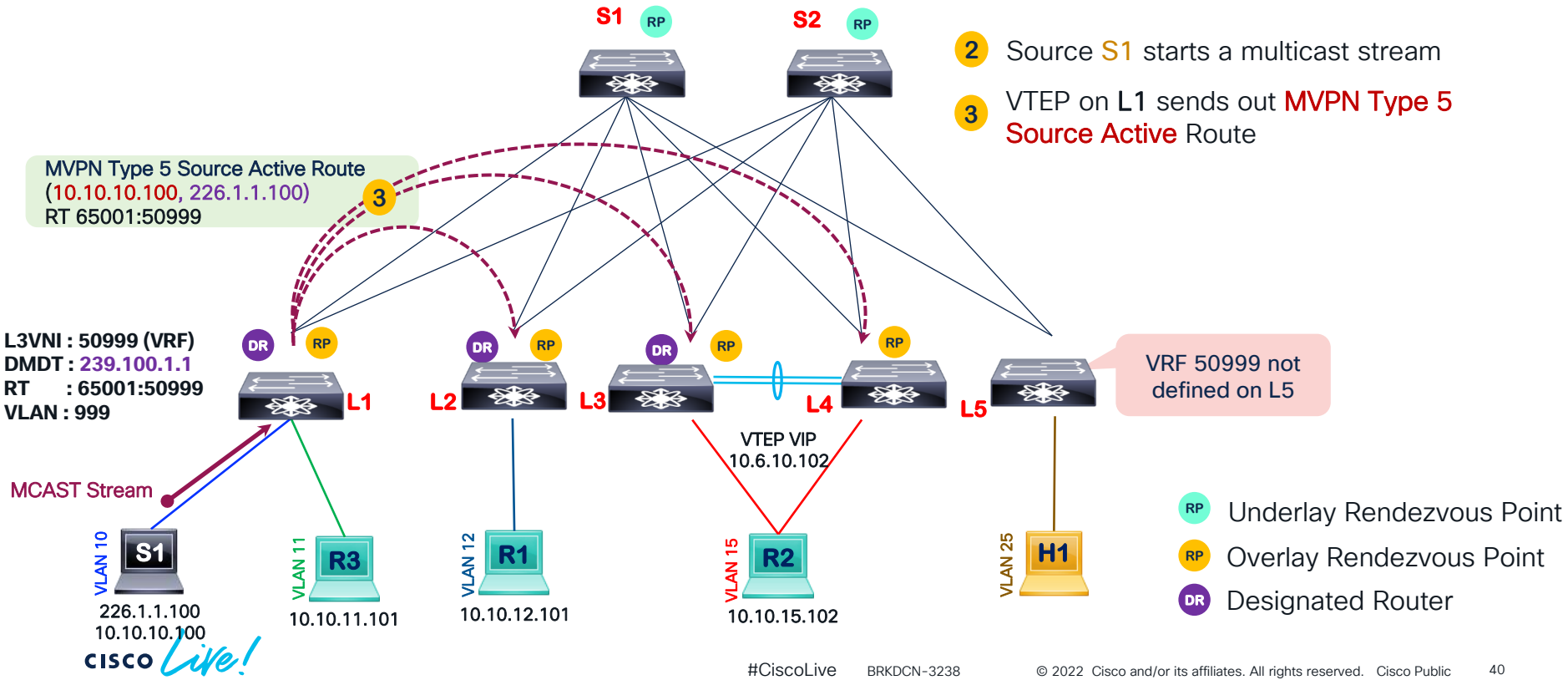
TRM Routing with Anycast RP

Packet Walk



TRM Routing with Anycast RP

Packet Walk



VRF Route Import

Extended Community

```
abcpod-1-dc1-bgw1# show bgp l2vpn evpn route-type 2 vrf tenant-1
Route Distinguisher: 10.6.10.4:3    (L3VNI 50999)
BGP routing table entry for [2]:[0]:[0]:[48]:[0050.0000.0c00]:[32]:[10.10.10.100]/272, version 109
Paths: (1 available, best #1)
Flags: (0x000202) (high32 00000000) on xmit-list, is not in l2rib/evpn, is not in HW

  Advertised path-id 1
  Path type: internal, path is valid, is best path, no labeled nexthop
             Imported from 10.6.10.2:32782:[2]:[0]:[0]:[48]:[0050.0000.0c00]:[32]:[10.10.10.100]/272
  AS-Path: NONE, path sourced internal to AS
    10.6.11.2 (metric 81) from 10.6.10.1 (10.6.10.1)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 30015 50999
      Extcommunity: RT:65001:30015 RT:65001:50999 Route-Import:10.6.11.2:999
```

MPVPN Route Type 5

MP-BGP RIB Leaf 1 (FHR)

```
abcpod-1-dc1-leaf1# show bgp ipv4 mvpn route-type 5 detail vrf Tenant-1
Route Distinguisher: 10.6.10.2:3    (L3VNI 50999)
BGP routing table entry for [5][10.10.10.100][226.1.1.100]/64, version 7
Paths: (1 available, best #1)
Flags: (0x000002) (high32 00000000) on xmit-list, is not in mvpn

  Advertised path-id 1
  Path type: local, path is valid, is best path, no labeled nexthop
  AS-Path: NONE, path locally originated
    0.0.0.0 (metric 0) from 0.0.0.0 (10.6.10.2)
      Origin IGP, MED not set, localpref 100, weight 32768
      Extcommunity: RT:65001:50999

  Path-id 1 advertised to peers:
    10.6.10.1
```

Overlay Multicast Group

Multicast Source IP

MVPN Type 5 Route

Route Target AS:L3VNI

MPVPN Route Type 5

MP-BGP RIB Leaf 2 (LHR)

```
abcpod-1-dc1-leaf2# show bgp ipv4 mvpn route-type 5 detail vrf tenant-1
Route Distinguisher: 10.6.10.3:3    (L3VNI 50999)
BGP routing table entry for [5][10.10.10.100][226.1.1.100]/64, version 10
Paths: (1 available, best #1)
Flags: (0x000002) (high32 00000000) on xmit-list, is not in mvpn, is not in HW

  Advertised path-id 1
  Path type: internal, path is valid, is best path, no labeled nexthop
             Imported from 10.6.10.2:3:[5][10.10.10.100][226.1.1.100]/64
  AS-Path: NONE, path sourced internal to AS
    10.6.11.2 (metric 81) from 10.6.10.1 (10.6.10.1)
      Origin IGP, MED not set, localpref 100, weight 0
      Extcommunity: RT:65001:50999
      Originator: 10.6.10.2 Cluster list: 10.6.10.1
```

Overlay Multicast Group

Multicast Source IP

MVPN Type 5 Route

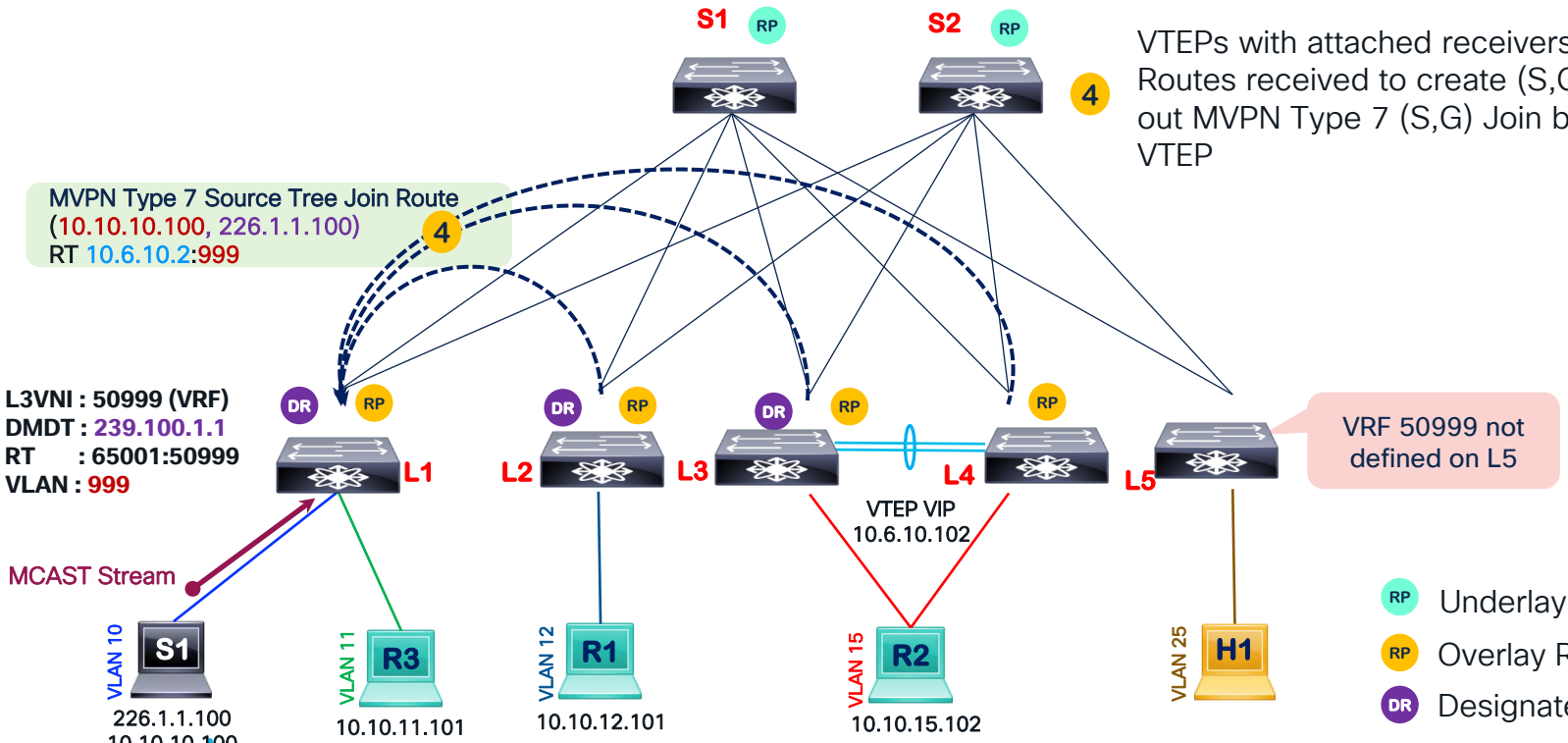
Import based on MVPN Type5 received

Route Target AS:L3VNI

Where the source is attached

TRM Routing with Anycast RP

Packet Walk



VTEPs with attached receivers use MVPN Type 5 Routes received to create (S,G) state and send out MVPN Type 7 (S,G) Join back to the source VTEP

RP Underlay Rendezvous Point

RP Overlay Rendezvous Point

DR Designated Router

MPVPN Route Type 7

MP-BGP RIB Leaf 1 (FHR)

```
abcpod-1-leaf1# show bgp ipv4 mvpn route-type 7
Route Distinguisher: 10.6.10.2:3    (L3VNI 50999)
BGP routing table entry for [7][10.10.10.100][226.1.1.100][65001]/96, version 824
Paths: (1 available, best #1)
Flags: (0x00001a) (high32 00000000) on xmit-list, is in mvpn, is not in HW

  Advertised path-id 1
  Path type: internal, path is valid, is best path, no labeled nexthop, in rib
          Imported from 10.6.10.3:32782:[7][10.10.10.100][226.1.1.100][65001]/96
  AS-Path: NONE, path sourced internal to AS
   10.6.10.3 (metric 3) from 10.6.10.1 (10.6.10.1)
     Origin IGP, MED not set, localpref 100, weight 0
     Extcommunity: RT:10.6.11.2:999
```

Multicast Source IP

Overlay Multicast Group

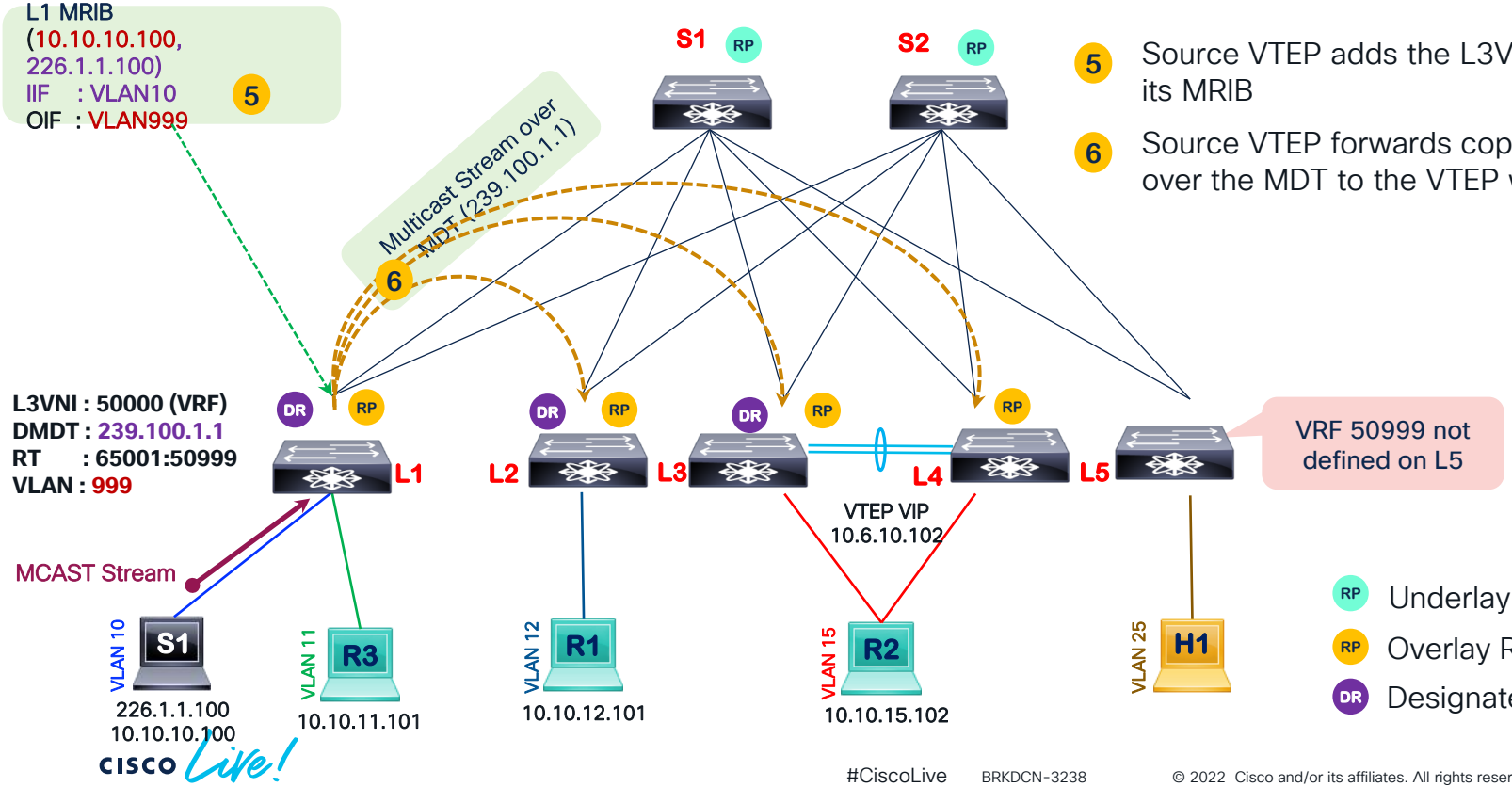
MVPN Type 7 Route

From where the import happened

VRI defines who will import

TRM Routing with Anycast RP

Packet Walk



5 Source VTEP adds the L3VNI SVI its OIF list in its MRIB

6 Source VTEP forwards copy of the stream over the MDT to the VTEP with receivers

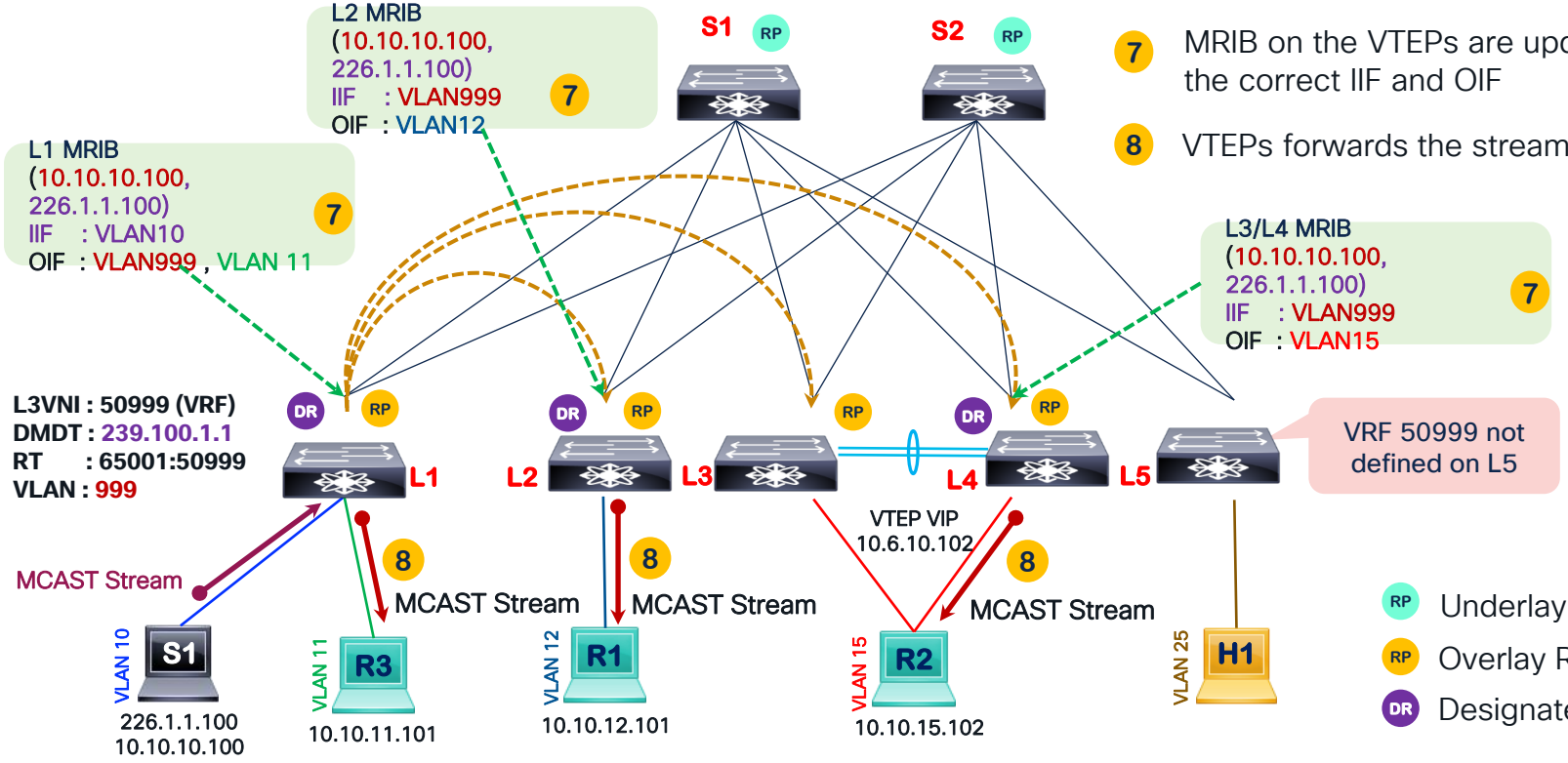
RP Underlay Rendezvous Point

RP Overlay Rendezvous Point

DR Designated Router

TRM Routing with Anycast RP

Packet Walk



7 MRIB on the VTEPs are updated to reflect the correct IIF and OIF

8 VTEPs forwards the stream towards the hosts

RP Underlay Rendezvous Point

RP Overlay Rendezvous Point

DR Designated Router

Configuring VXLAN EVPN TRM

VXLAN EVPN TRM Configuration Guidelines

- TRM uses an "Always Route" approach in the overlay.
- TRM requires an IPv4 underlay.
- TRM is only supported when PIM Any Source Multicast (ASM) is used in the underlay
- TRM is not supported with PIM BiDir in the underlay
  - PIM BiDir is supported for Unicast in the underlay
- TRM also supports IPv6 Multicast in the overlay as of NXOS 10.2.1
  - MLD snooping with VxLAN VLANs with TRM
- TRM only supports PIM ASM and PIM SSM in the overlay
- 224.0.0.0/24 subnet (local scope) is excluded from TRM and is always bridged
- TRM L3 Designated Router (DR) capability is supported on 2nd Gen hardware (Nexus 9200,9300 EX/FX/FX2/GX and 9700 EX and FX LC) switches

VXLAN EVPN TRM with Anycast RP

Feature and BGP Configuration

```
nv overlay evpn
feature bgp
feature pim
feature interface-vlan
feature vn-segment-vlan-based
feature nv overlay
feature ngmvpn

router bgp 65501
  neighbor 10.100.100.201
    remote-as 65501
    update-source loopback0
    address-family l2vpn evpn
      send-community both
    address-family ipv4 mvpn
      send-community both
```

”feature ngmvpn” will enable the Next-Generation Multicast VPN (ngMVPN) control-plane. New address-family commands become available in BGP.

VXLAN EVPN (”feature nv overlay” and ”nv overlay evpn”) has to be enabled first

”address-family ipv4 mvpn” enables ngMVPN Address-Family for Multicast signaling. ”**send community both**” ensures both standard and extended communities are exchanged for this address-family.

VXLAN EVPN TRM with Anycast RP

Tenant Distributed Anycast Gateway SVI Configurations

VRF Tenant1

```
interface vlan10
 vrf member Tenant1
 ip address 10.10.10.1/24 tag 12345
 ip pim sparse-mode
 ip pim neighbor-policy NONE*
 fabric forwarding mode anycast-gateway

interface vlan20
 vrf member Tenant1
 ip address 20.20.20.1/24 tag 12345
 ip pim sparse-mode
 ip pim neighbor-policy NONE*
 fabric forwarding mode anycast-gateway

interface vlan30
 vrf member Tenant1
 ip address 30.30.30.1/24 tag 12345
 ip pim sparse-mode
 ip pim neighbor-policy NONE*
 fabric forwarding mode anycast-gateway
```

”ip pim sparse-mode” enables IGMP and PIM on the SVI. This is required if Multicast Sources and/or Receivers exist in this VLAN

Create a ”ip pim neighbor-policy” to avoid forming PIM neighbor relationship with PIM Routers within the VLAN (Don’t use Distributed Anycast Gateway for PIM Peering).

VXLAN EVPN TRM with Anycast RP

Tenant VRF PIM Configurations

VRF Tenant1

```
vlan 2501
 vn-segment 50001

interface vlan2501
 vrf member Tenant1
 ip forward
 ip pim sparse-mode

interface loopback250
 vrf member Tenant1
 ip address 10.51.51.254/32 tag 12345
 ip pim sparse-mode

ip multicast overlay-spt-only

vrf context Tenant1
 ip pim rp-address 10.51.51.254
 vni 50001
 rd auto
 address-family ipv4 unicast
  route-target both auto
  route-target both auto evpn
```

"ip pim sparse-mode" enables Multicast Routing on the Tenant.

"ip address 10.51.51.254" defines the Overlay Multicast Rendezvous-Point (RP) IP address in the respective VRF. This IP address has to be advertised in the BGP EVPN control-plane of the VRF (i.e. redistribute).

"ip multicast overlay-spt-only" is needed for defining the distributed RP.

"ip pim rp-address" defines the Overlay Multicast Rendezvous-Point (RP) in the VRF

Note: The per-VRF Loopback for the RP configuration has to be configured on every Node that is running Tenant Routed Multicast (TRM). The Distributed RP is the recommended way for configuring the RP.

VXLAN EVPN TRM with Anycast RP

Tenant VRF VTEP Configurations

VRF Tenant1

```
vrf context Tenant1
 ip pim rp-address 10.51.51.254
 vni 50001
 rd auto
 address-family ipv4 unicast
  route-target both auto
  route-target both auto evpn
  route-target both auto mvpn

interface nve1
 source-interface loopback1
 host-reachability protocol bgp
 member vni 30010
  mcast-group 239.1.1.1
 member vni 30020
  mcast-group 239.1.1.1
 member vni 30030
  mcast-group 239.1.1.2
 member vni 50001 associate-vrf
  mcast-group 239.10.1.1
```

"route-target both auto mvpn" defines the BGP Route-Target that is added as an Extended Community attribute to the Customer Multicast (C-Multicast) routes (ngMVPN Route-Type 6 and 7)

Auto Route-Targets are constructed by the 2-byte Autonomous System Number and Layer-3 VNI (ASN:VNI).

"mcast-group" for the VRF VNI (Layer-3 VNI) builds the Default Multicast Distribution Tree (Default MDT).

The Multicast Group is used in the Underlay (Core) for all Multicast Routing within the associated L3VNI (VRF).

Note: Underlay Multicast Groups for L2VNI (Broadcast, Unknown Unicast), Default MDT and Data MDT should not be shared. Use separate, non overlapping Groups

Summary

Key Takeaways

- VXLAN EVPN TRM uses open standard VXLAN data plane with MP-BGP NGMVPN control plane for tenant multicast routing.
- A single MP-BGP control plane protocol is used for both unicast (AF EVPN) and multicasting (AF MVPN) routing in tenants in a VXLAN BGP EVPN Fabric.
- VXLAN EVPN TRM forwards using an ”Always Route” approach.
- VXLAN EVPN TRM supports various RP deployments models including Anycast RP, External RP and RP Anywhere allowing redundancy and ease of migration of RPs.
- IGMP maintains its current function as Host Reporting protocol.
- PIM operates in the tenant for tenant multicast domain and underlay for Data MDT for the tenant.

Technical Session Surveys

- Attendees who fill out a minimum of four session surveys and the overall event survey will get Cisco Live branded socks!
- Attendees will also earn 100 points in the Cisco Live Game for every survey completed.
- These points help you get on the leaderboard and increase your chances of winning daily and grand prizes.

Cisco Learning and Certifications

From technology training and team development to Cisco certifications and learning plans, let us help you empower your business and career. www.cisco.com/go/certs

Pay for Learning with Cisco Learning Credits

(CLCs) are prepaid training vouchers redeemed directly with Cisco.

Learn

Cisco U.
IT learning hub that guides teams and learners toward their goals

Cisco Digital Learning
Subscription-based product, technology, and certification training

Cisco Modeling Labs
Network simulation platform for design, testing, and troubleshooting

Cisco Learning Network
Resource community portal for certifications and learning

Train

Cisco Training Bootcamps
Intensive team & individual automation and technology training programs

Cisco Learning Partner Program
Authorized training partners supporting Cisco technology and career certifications

Cisco Instructor-led and Virtual Instructor-led training
Accelerated curriculum of product, technology, and certification courses

Certify

Cisco Certifications and Specialist Certifications
Award-winning certification program empowers students and IT Professionals to advance their technical careers

Cisco Guided Study Groups
180-day certification prep program with learning and support

Cisco Continuing Education Program
Recertification training options for Cisco certified individuals

Here at the event? Visit us at **The Learning and Certifications lounge at the World of Solutions**

Continue your education

- Visit the Cisco Showcase for related demos
- Book your one-on-one Meet the Engineer meeting
- Attend the interactive education with DevNet, Capture the Flag, and Walk-in Labs
- Visit the On-Demand Library for more sessions at www.CiscoLive.com/on-demand

Thank you

cisco Live!

ALL IN

#CiscoLive