BGP-EVPN for the Data Center and L3 DCI

Mayur Shetty (mashetty@cisco.com) , Pratima Kini (pkini@cisco.com)

November 30, 2015

Agenda

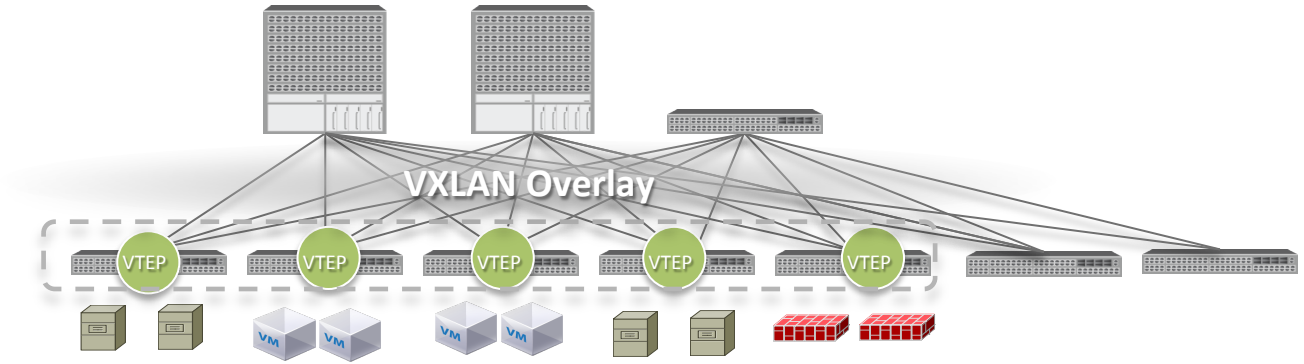
- What is EVPN ?
- What problems does VXLAN solve ?
- What problems does BGP EVPN solve ?
- Overview of BGP EVPN Route Types and associated use cases
  - Host IP/MAC distribution
  - IP Prefix route
- Layer-3 DCI
  - L3 handoff between VXLAN-EVPN and MPLS-L3VPN to extend the layer-3 network connectivity across Data Centers over a WAN

What is E-VPN?

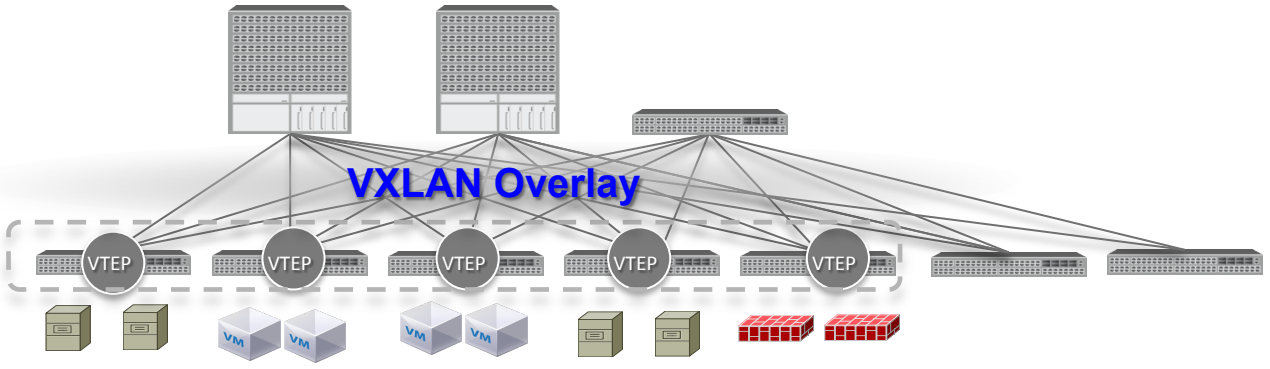
- Ethernet VPN (EVPN) – connect a group of customer sites using a virtual bridge.
  - Treat MAC addresses as routable addresses and distribute them in BGP
  - Uses Multi-protocol BGP
  - Initially started as next generation L2VPN solution for service provider networks
- Evolution of EVPN
  - Data center use cases
    - Multi-tenancy with virtualized hosts
    - Support of VXLAN and NVGRE encapsulations
    - Integrated routing and bridging
  - Support exchange of IP addresses and IP prefixes
- Status of EVPN
  - Standardization effort - IETF L2VPN work group
  - Multi-vendor support – core set of drafts co-authored by engineers from Cisco, Juniper, Alcatel-Lucent, Verizon, ATT, Bloomberg.
  - BGP MPLS based EVPN is RFC 7432 and extensions for DC is currently in draft stages.

What problems does VXLAN solve?

Customer Needs	VXLAN Enables
Any workload anywhere – VLANs limited by L3 boundaries	Any Workload anywhere- across Layer 3 boundaries
VM Mobility	Seamless VM Mobility
Scale above 4k Segments (VLAN limitation)	Scale up to 16M segments
Secure Multi-tenancy	Traffic & Address Isolation



Challenges with VXLAN Deployments



LIMITED SCALE

Flood and learn (BUM)- Inefficient Bandwidth Utilization

Resource Intensive – Large MAC Tables

CENTRALIZED

Centralized Gateways, Controller – Traffic Hair-pining

Sub-Optimal Traffic Flow

Barrier for Scaling out Large Data Centers and Cloud Deployments

What problems does BGP EVPN solve for VXLAN?

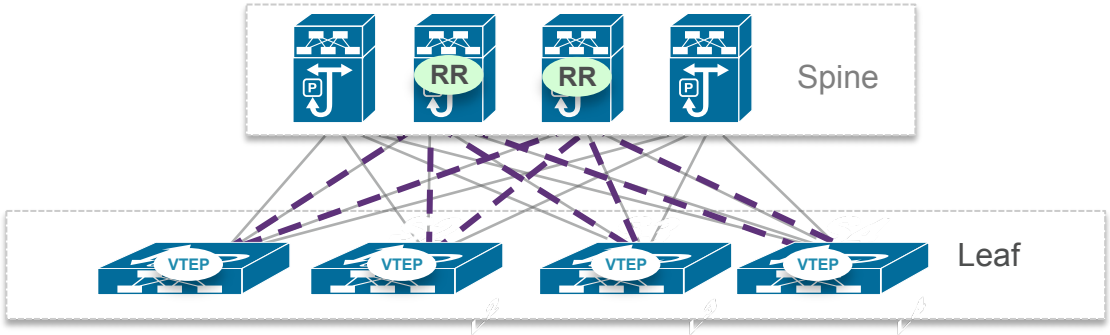
- Control plane for VXLAN overlays
- Optimize/eliminate flooding of
  - Unknown unicast traffic.
  - Protocol messages, e.g. ARP
- Virtual machine mobility with optimal forwarding
  - No hair pinning of traffic to previous location
- Active/active multi-homing with per flow load balancing
- Large scale multi-tenancy in control plane with characteristics of L3VPN
  - Route filtering and constrained route distribution
- Ingress replication of multi-destination traffic
  - Multicast free underlay

BGP-EVPN / VXLAN Terminology

1. Layer-2 VNI
   - VNI (VXLAN network identifier) carried in VXLAN packets bridged across VTEPs (VXLAN tunnel end point) . This VNI is configured per VLAN.
2. Layer-3 VNI
   - VNI carried in the VXLAN packets routed across VTEPs. This VNI is linked per Tenant VRF.
3. Anycast GW
   - All L3 VTEPs are configured with same mac and same subnet for host facing SVI.
4. VRF overlay VLAN
   - Every Tenant VRF will need a Vlan to be configured for VXLAN routing.
   - This VLAN is configured with L3-VNI.
5. VXLAN L2 Gateway
   - VTEP capable of switching VLAN->VXLAN, VXLAN->VLAN packets with in same VNI.
6. VXLAN L3 Gateway
   - VTEP capable of routing packets across different VNIs.

EVPN Control Plane – Reachability Distribution

EVPN Control Plane -- Host and Subnet Route Distribution



iBGP Adjacencies

RR

Route-Reflectors deployed for scaling purposes

BGP Update
- Host-MAC
- Host-IP
- Internal IP Subnet
- External Prefixes

- Use MP-BGP with EVPN Address Family on leaf nodes to distribute internal host MAC/IP addresses, subnet routes and external reachability information
- MP-BGP also used to distribute IP multicast groups information
- MP-BGP enhancements to carry up to 100s of thousands of routes with reduce convergence time

BGP EVPN Route Types

Route-Type	EVPN Routes	Purpose
1	Ethernet Auto-discovery Route	Mass withdrawal and Aliasing
2	**MAC/IP Route**	**Advertise host MAC and IP address**
3	**Inclusive Multicast Route**	**Tunnel end point discovery for setting up of replication list**
4	Ethernet Segment Route	Discovery of nodes in redundancy group and DF-election
5	**IP Prefix route**	**Advertise IP prefixes**

BGP EVPN MAC Route (Type – 2 Route)

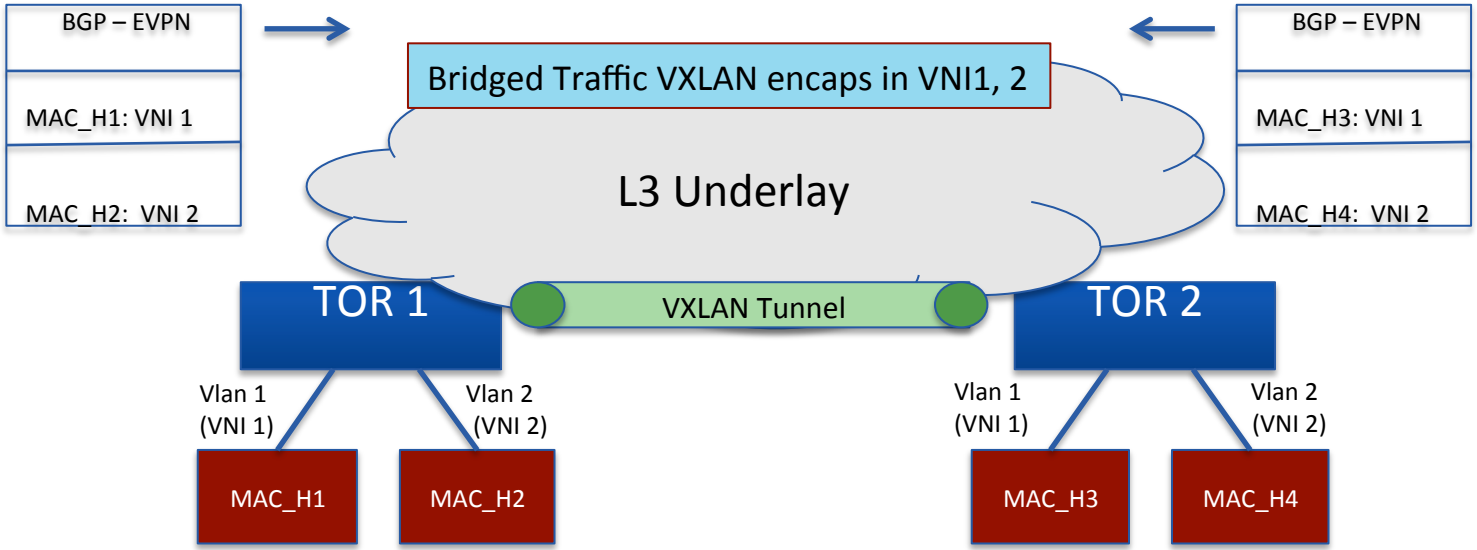
- RD: per VPN RD
- MAC Len: 48
- MAC Addr: Host-MAC
- IP Length: 32
- IP address: Host-IP
- Label1: VNI for BD
- Label2: VNI for VRF

Route Target
- RT for EVI
- RT for VRF

Tunnel Attribute
- Tunnel Type VXLAN
- MAC: Router MAC

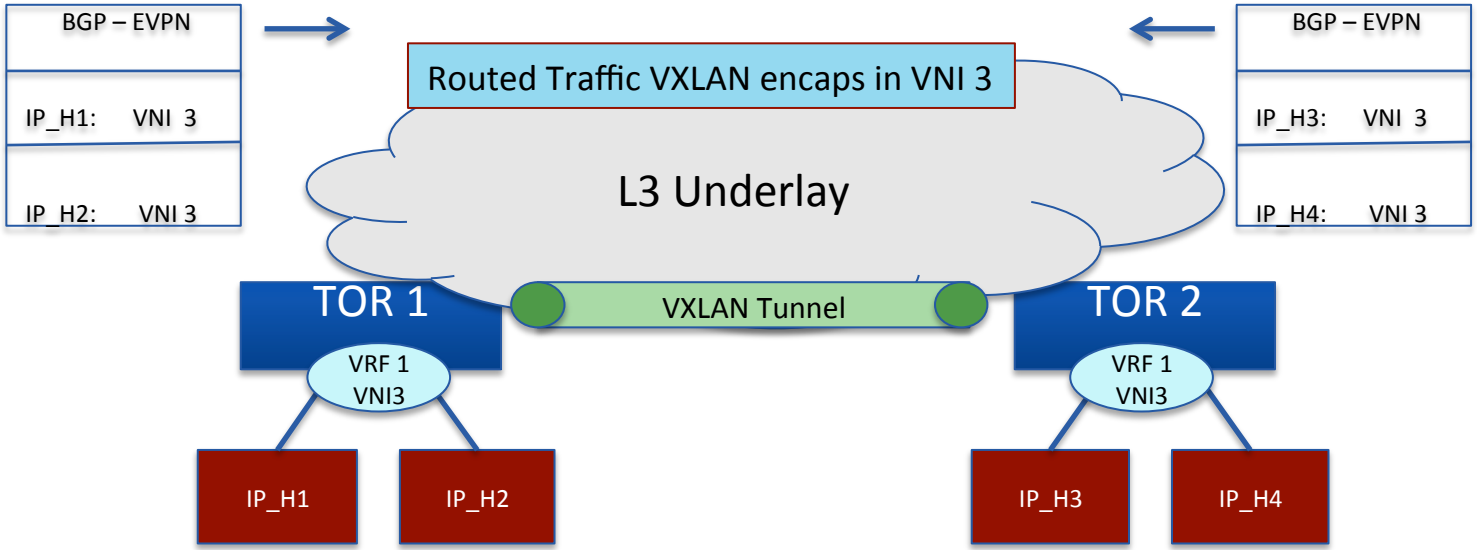
MAC Mobility Sequence Number

BGP EVPN Control Plane for VXLAN L2 Overlay (Type-2 Route)



Advertisement of MAC:VNI bindings from a TOR via BGP EVPN enables a remote TOR to send bridged traffic to that MAC using the VNI for that MAC to the TOR that advertised the MAC over VXLAN

BGP EVPN Control Plane for VXLAN L3 Overlay (Type-2 Route)



Advertisement of IP:VNI bindings from a TOR via BGP EVPN enables a remote TOR to send routed traffic to that IP using the VNI for that IP to the TOR that advertised the IP over VXLAN

BGP EVPN IP Prefix Route (Type – 5 Route)

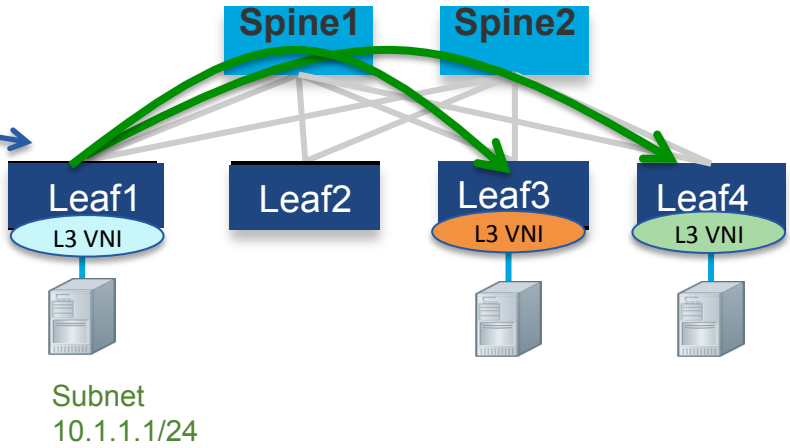
- RD: VRF RD
- Ethernet Segment: 0
- Ethernet Tag: 0
- IP Address Len: 0-32
- IP Address: IP prefix
- GW IP Address: 0
- Label : L3 VNI

Route Target
- RT for VRF

Tunnel Attribute
- Tunnel Type VXLAN
- MAC: Router MAC

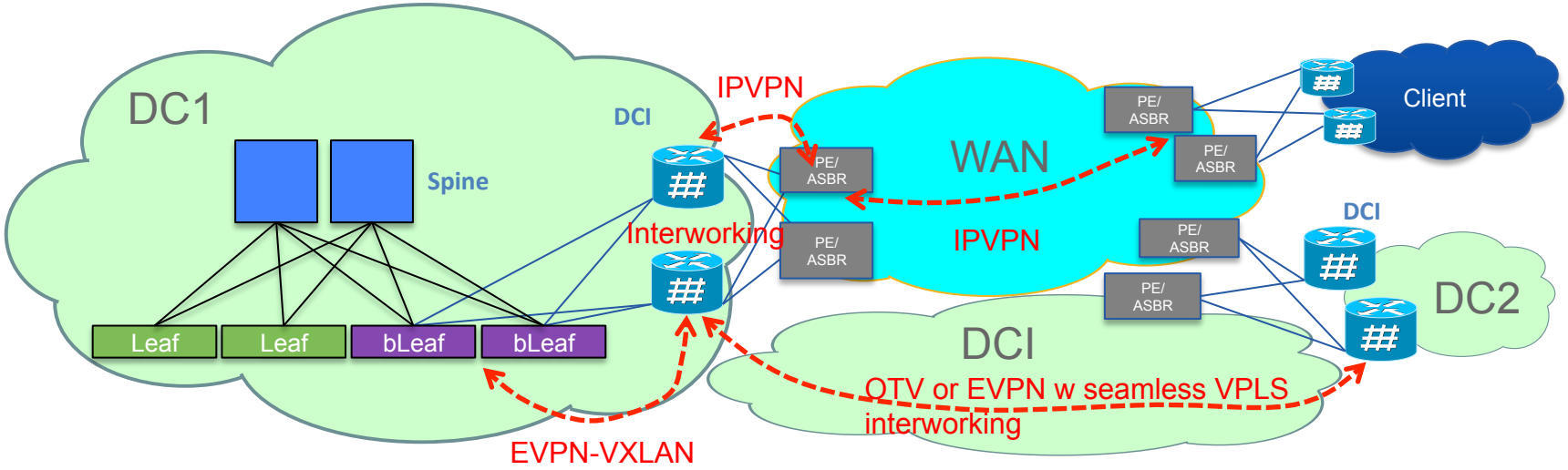
BGP EVPN Control Plane Type-5 Route

IP Prefix Route:
Prefix : 10.1.1.0/24
VNI : L3 VNI
Next Hop : IP-L1
Router MAC : System MAC of Leaf1

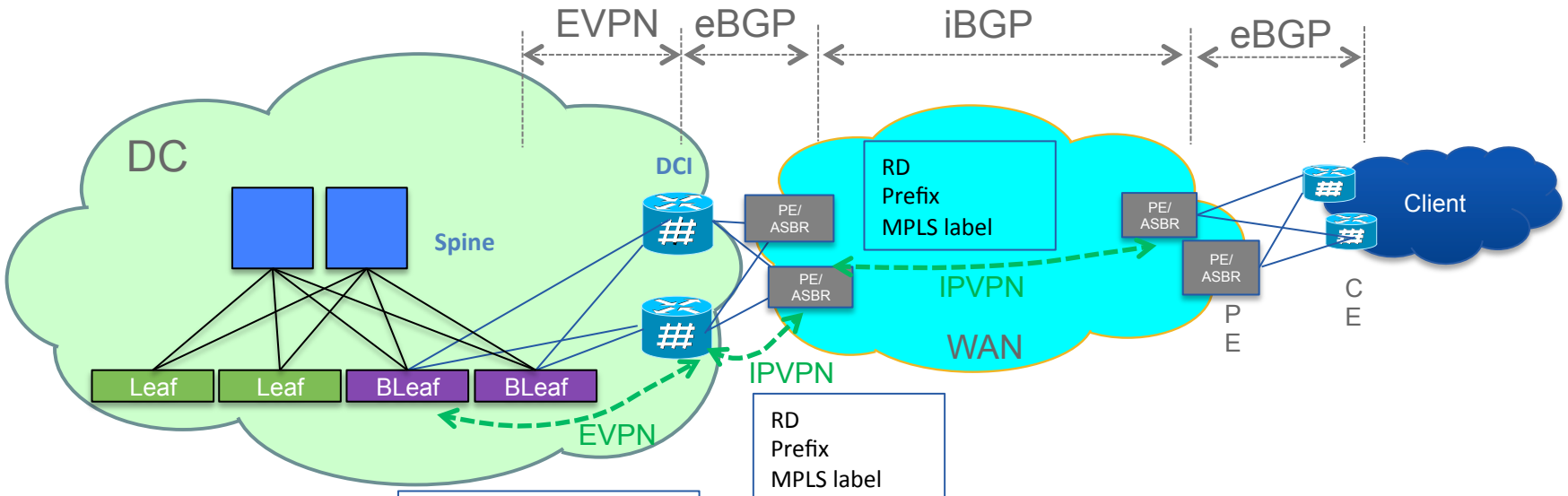


DCI Overview

- L3 DCI Service:
  - EVPN-VXLAN to IP VPN (unicast) Interworking on DCI
- L2 DCI Service:
  - EVPN-VXLAN to VPLS Interworking on DCI
  - EVPN-VXLAN to OTV Interworking on DCI
  - EVPN-VXLAN to EVPN-MPLS Interworking on DCI

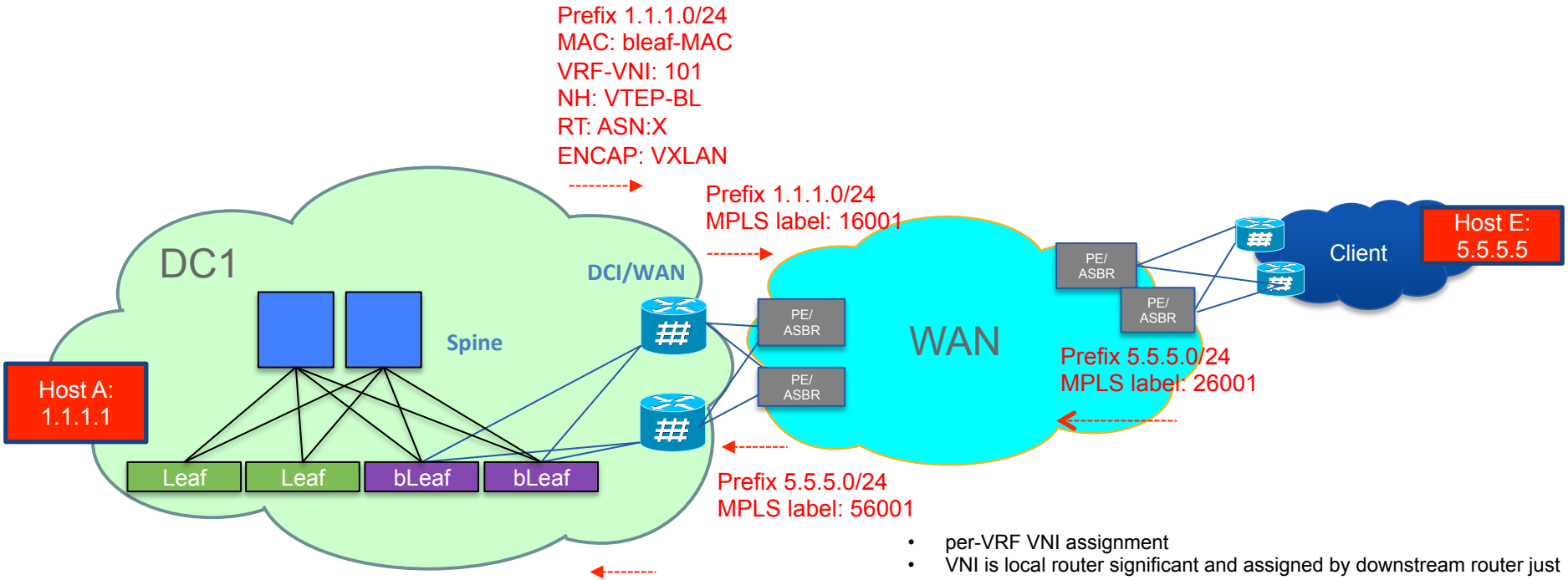


L3 DCI Service



- **VNID**: per-VRF encap, downstream assigned by BL and DCI
- **MAC**: next-hop router MAC (BL, DCI)
- **NH**: VTEP IP (BL, DCI)
- **RT**: coordinated between each DCI-BL pair
- **ENCAP:** VXLAN

L3 DCI Service – Control Plane



- Fabric can advertise aggregated prefix and specific host route to WAN
- WAN router will typically advertise default route into fabric

- per-VRF VNI assignment
- VNI is local router significant and assigned by downstream router just like MPLS VPN label
- Ingress and egress VNIs can be different for same VRF
- RT is unique per-VRF between fabric and WAN as the “glue”
- MAC is next-hop router MAC which could advertised global per-VTEP or per-VNI / VRF
- NH is set to VTEP IP
- ENCAP is set to VXLAN

L3 DCI Service – DCI Forwarding Plane

1. WAN -> Fabric

2. Fabric -> WAN

References

- VXLAN: A framework for overlaying Virtualized Layer 2 Networks over Layer 3 Networks -- RFC 7348
  - https://tools.ietf.org/html/rfc7348
- BGP MPLS based EVPN -- RFC 7432
  - https://tools.ietf.org/html/rfc7432
- Requirements for Ethernet VPN (EVPN)
  - https://tools.ietf.org/html/rfc7209
- A Network Virtualization Overlay Solution using EVPN
  - https://tools.ietf.org/html/draft-ietf-bess-evpn-overlay-02

TOMORROW starts here.