The comfortable complexity of Overlays

2025/05/15

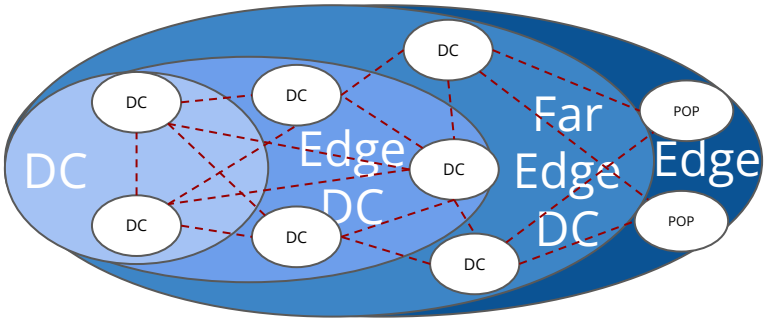
Sergey Kolobov

About

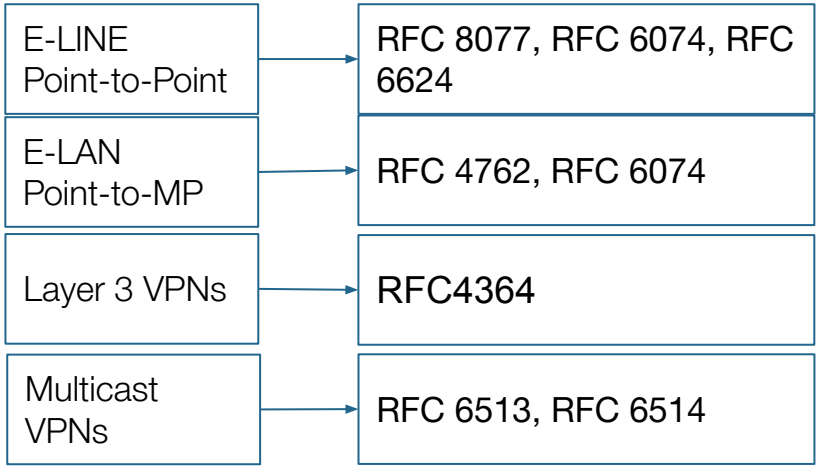
- Sergey Kolobov
- Tech Lead Systems Engineer @arista.com
- Supports Cloud providers, Hyperscale infrastructure for AI, DC, Enterprise
- Appreciates good documentation and clear communication
- Past: https://www.linkedin.com/in/sergey-kolobov-78992476/

Possible future directions?

- Transport aware applications - limited L2 requirements
- IPv6 native L3 LS fabric
- Basic & simplified: BGP, QoS, BFD
- SRv6 Underlay-aware overlay RFC9252 Color Community

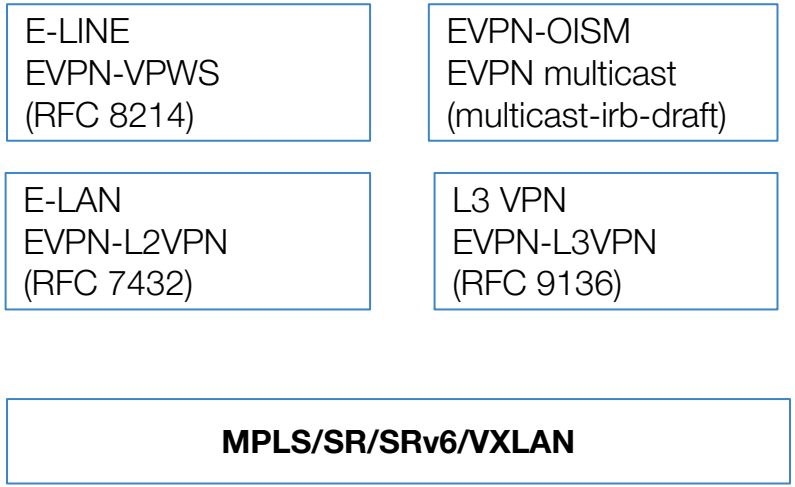


Simplification of control plane



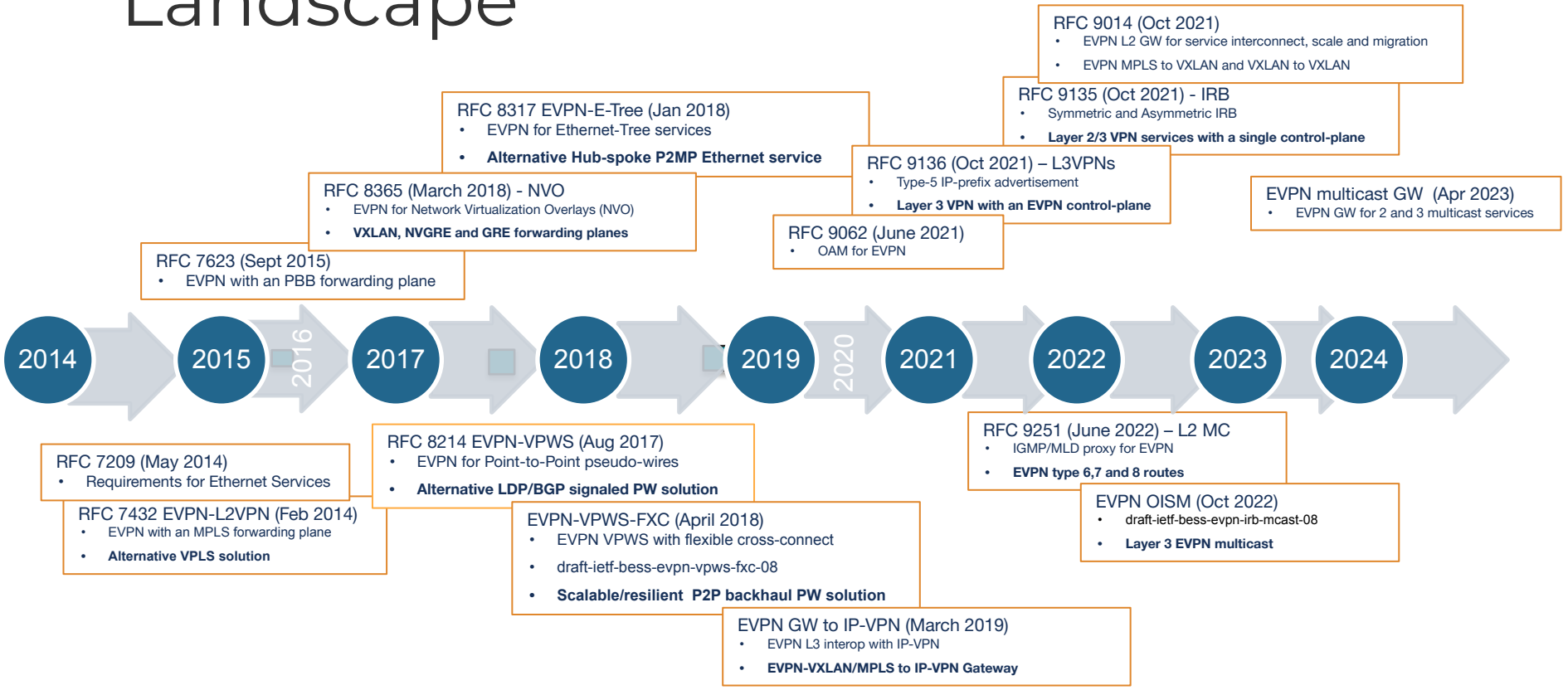
Overly complex, multiple signaling protocols, BGP AFs each with vendor specific implementations

EVPN BGP control plane
(single AFI/SAFI)



Simplify, single BGP AF and EVPN control plane for all layer 2 and 3 services types

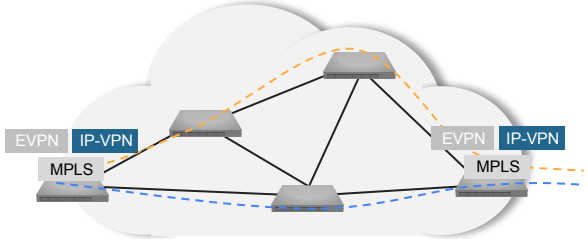
Landscape



48 current active drafts

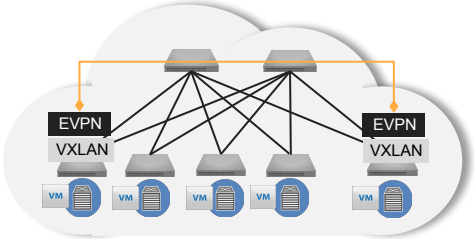
Market dynamics

Telco/SP Market
MPLS transport for FRR and TE



- MPLS
- Scarcity of BW and links
- Driving TE and FRR requirements

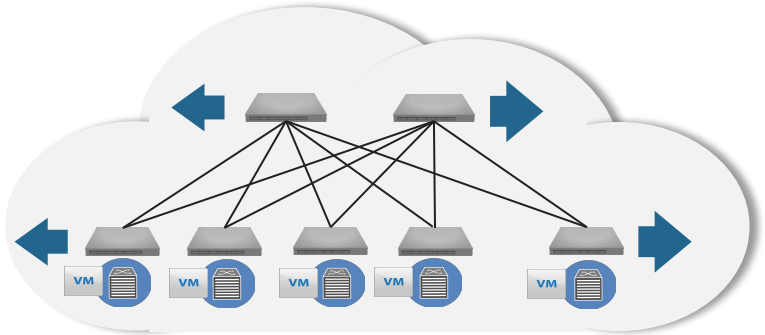
Telco Cloud/Enterprise DC
Virtualized Data Center



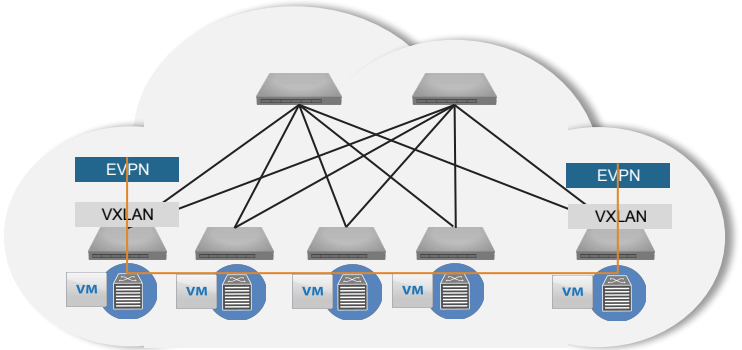
- “Any server and anywhere”
- Lower-cost VXLAN
- Costly TE and FRR benefits of MPLS NOT required

EVPN in the Data Center

- IP Leaf & Spine Fabric
- All leaves are equidistant
- Consistent East-to-West performance & hop count

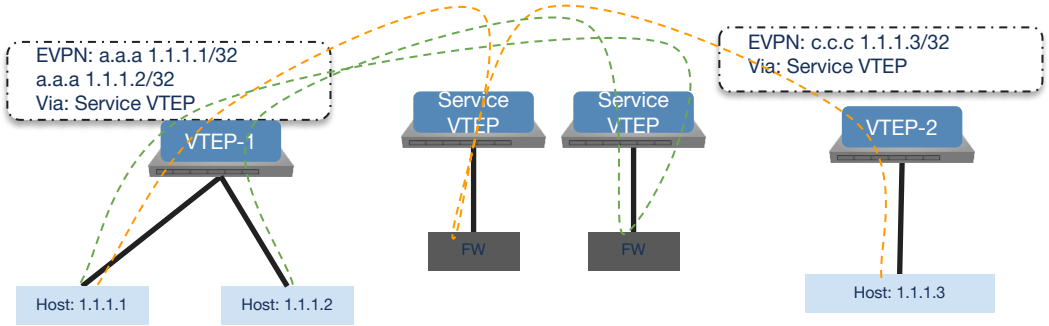


- EVPN with VXLAN (RFC 8365)
- Multi-tenancy
- L2 & L3 VPNs across all leaf nodes
- BGP control-plane learning



Enterprise use case

- Isolation of tenants, domains and hosts
- Host based segmentation
- Switches and FWs – Policy Enforcement Points
- EVPN VXLAN E-TREE RFC8317bis
- draft-smith-vxlan-group-policy

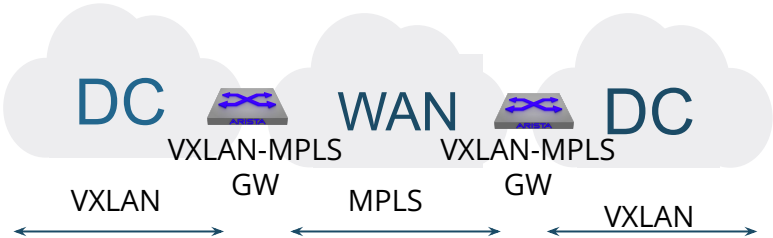


L2 in HPC/AI Use case

- Single tenant/ administration domain
- Legacy applications or services
- Purpose built pure L2 protocols: TTP over Ethernet for exascale lossy AI network
  - https://github.com/teslamotors/ttpoe
  - Big fabrics TCP/IP is too slow – bound by CPU SW kernel
  - PFC affects the global network

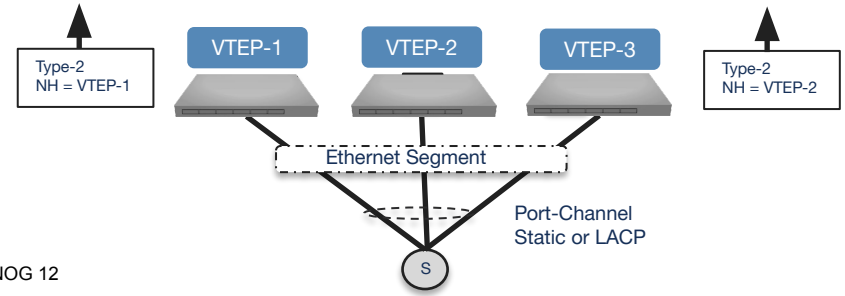
Telco/Cloud use case

- Multitenancy , DCI with EVPN MPLS
- L3 Leaf-Spine fabric with EVPN VXLAN
  - ECMP requirement within a fabric
  - EVPN RT-5 GW Index
- EVPN to the tenant - scalability concerns
- Leveraging EVPN MPLS to VXLAN Gateway

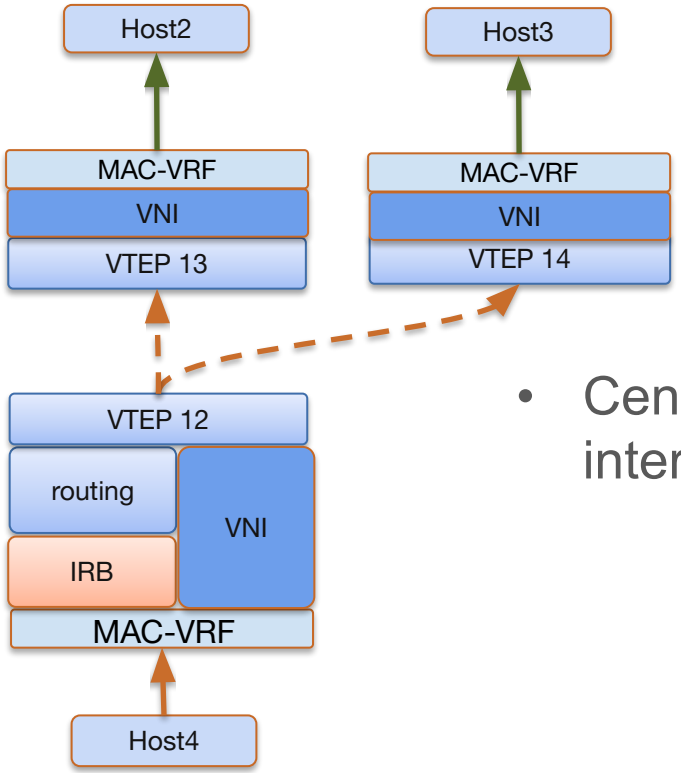


EVPN – Multi-homing options

- RFC 7432
- RFC 8584
- draft-rabnag-bess-evpn-anycast-aliasing
- draft-sajassi-bess-evpn-ip-aliasing
- draft-ietf-bess-evpn-fast-df-recovery
- draft-ietf-bess-evpn-mh-pa-10



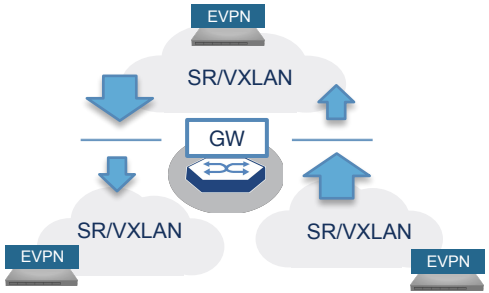
L2 VTEPs and Centralized Gateway



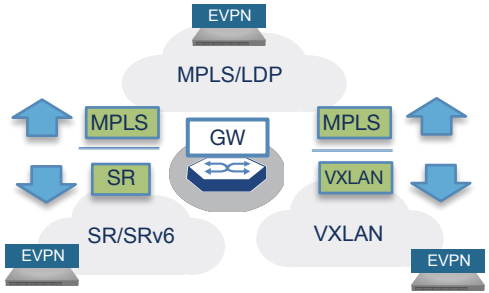
- VTEP 13 and VTEP 14 are L2 only VTEPs
- Centralized GW VTEP 12 is configured with an IRB interface

EVPN – why do we need GWs?

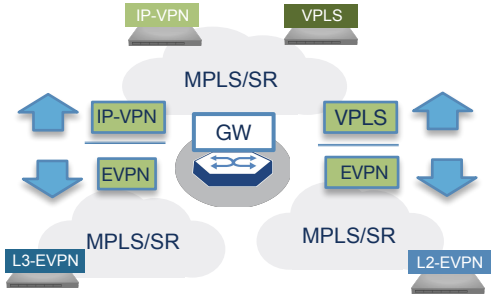
EVPN GW for hierarchical scaling



Transport Transition



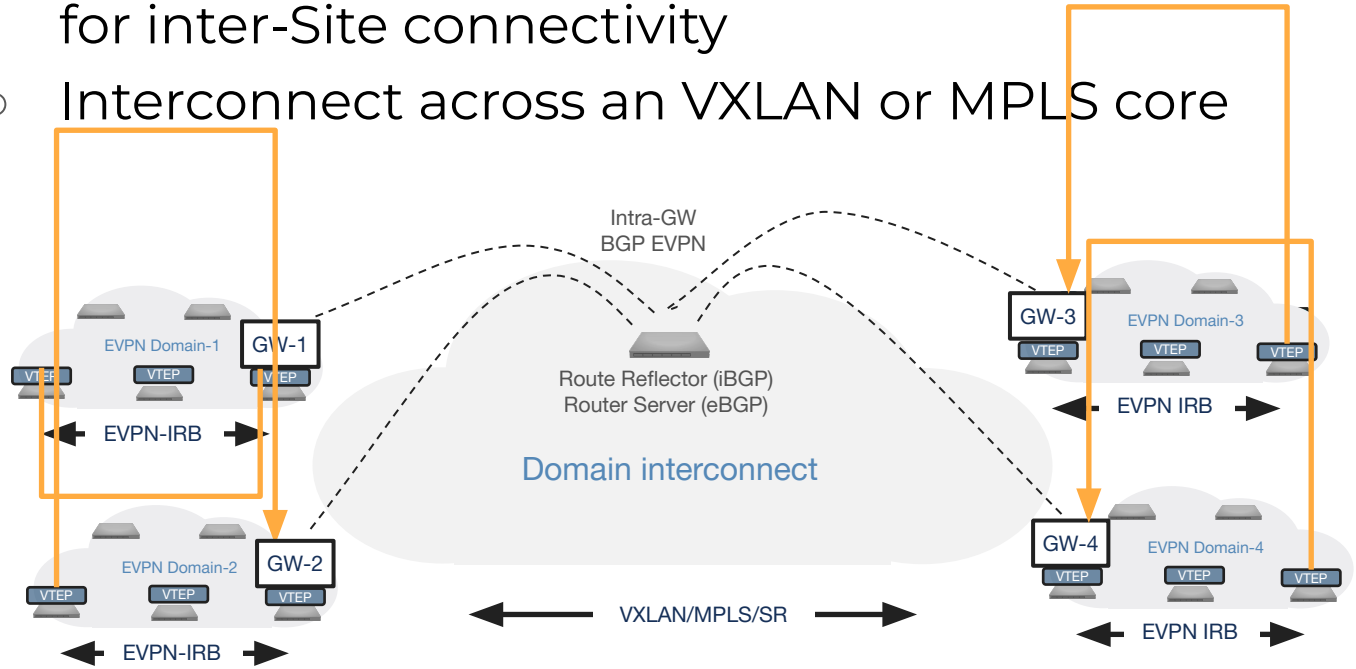
Service Migration and interop



- L2 L3 VPN
- Hierarchy reduce EVPN state
- Improve HW table utilization (MAC, ARP etc)
- Improve SW memory utilization (RIB routes)

EVPN GW for Hierarchical scaling

- EVPN GW Solution
  - Multiple domains, improved DC scale with inter-POD or for inter-Site connectivity
  - Interconnect across an VXLAN or MPLS core



EVPN GW – Background

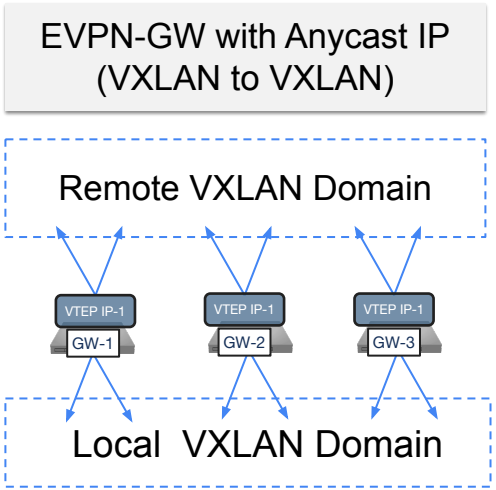
- Standard based solution defined across multiple RFCs/Drafts
  - Support for different BGP AFs and encapsulations (VXLAN, MPLS, SRv6)
  - Widely adopted Industry approach (see EANTC testing)*

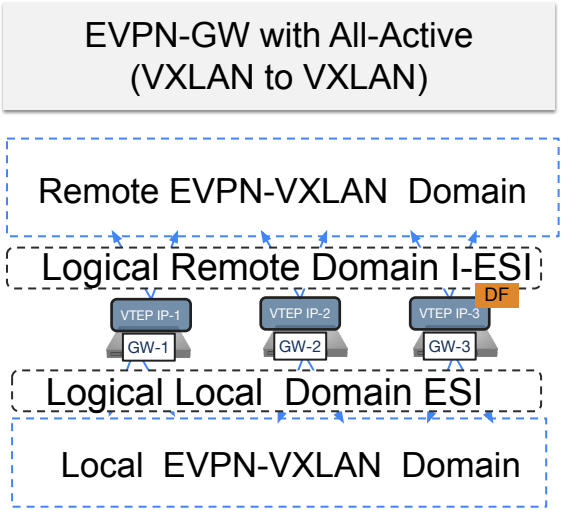
EANTC
EUROPEAN ADVANCED NETWORKING TEST CENTER

- draft-ietf-bess-evpn-ipvpn-interworking-11
- RFC 9014
- draft-rabnic-bess-evpn-mcast-eeg
- draft-sr-bess-evpn-dpath
- draft-sharma-bess-multi-site-evpn

chinog12
CHICAGO NETWORK OPERATORS GROUP

EVPN GW – Redundancy models



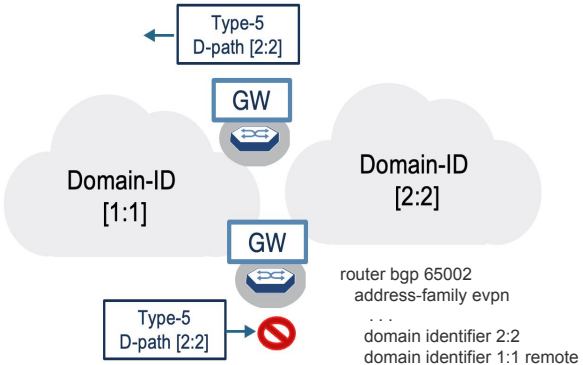


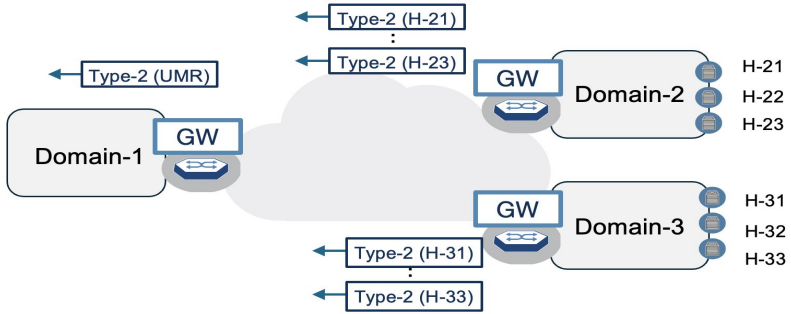
- Anycast IP next-hop of EVPN routes
- ECMP load-balancing in the underlay
- No directly attached hosts

- GW nodes configured in an I-ESI
- ECMP load-balancing in the overlay
- Support for directly attached hosts

EVPN GW – Extensions

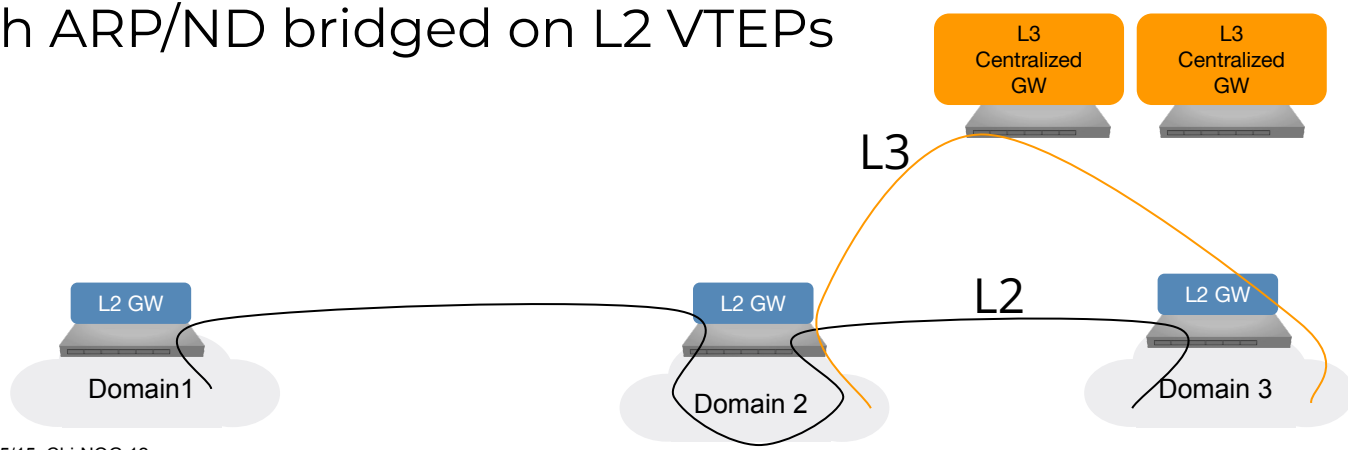
- Unknown MAC Route (UMR)
- RT/RD import and export model for type-5 routes
- D-Path support for type-5 routes





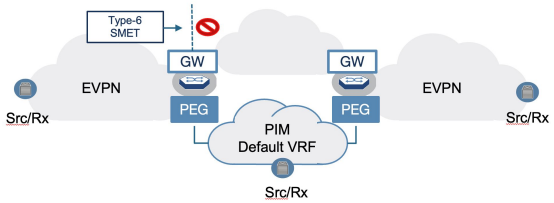
EVPN GW to scale out in a DC

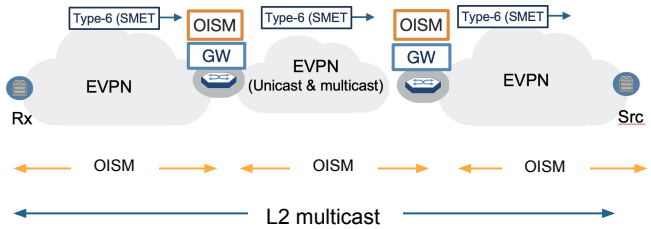
- L2 network is divided into domains by all-active multi-domain GWs
- L2 VTEPs EVPN All-Active Multihoming GW for Multi-Domains
- L3 network is flat single domain with centralised routing
  - Border GW - L3 VTEP EVPN all-active GW
- BUM control with ARP/ND bridged on L2 VTEPs



EVPN GW Multicast – Extensions

- EVPN GW support for multicast
  - Multicast VRF route leaking with OISM
  - EVPN OISM V6 underlay (SSM only) with V4/V6 overlay
  - EVPN Multicast with A-A Multi-homing
- OISM on the GW - Optimized Inter-Subnet Multicast
  - Type-6 routes across domains, L2 & L3 multicast
  - draft-rabnic-bess-evpn-mcast-eeg





Overlays simplified

- Multitenancy and L2 over L3 IP fabrics are provided with various overlay technologies
- EVPN is a simplification for control planes, supporting multiple service types
    a. Gateway for scaling, transport transition, and service migration
    b. Multi-homing options and gateway redundancy models
    c. Gateway extensions, such as UMR and multicast support

Thank you!