10+ Years of EVPN

L3 VPN – A short history…

- MPLS Layer 3 IP VPN
  - RFC 2547 published March 1999, updated to RFC4364 published Feb 2006
  - BGP based control plane using stack label to identify VPN (VRF).
  - New concepts of Route-targets (RT) and Route-Discriminator (RD)
  - Cross vendor support Cisco and Juniper

- The simplicity didn't last … with competing Multicast VPN draft
  - Rosen draft - PIM in the Core to build mGRE tunnels with IP forwarding (2009)
  - NG-MVPN (RFC 6513)
    - BGP control-plane building on RFC 2547 with MPLS forwarding plane (2012)
    - But multiple different deployment models +20 profiles for transport and PMSI

Cross vendor consensus for IP VPN but interop challenges for MVPN due to complexity and vendor disparity

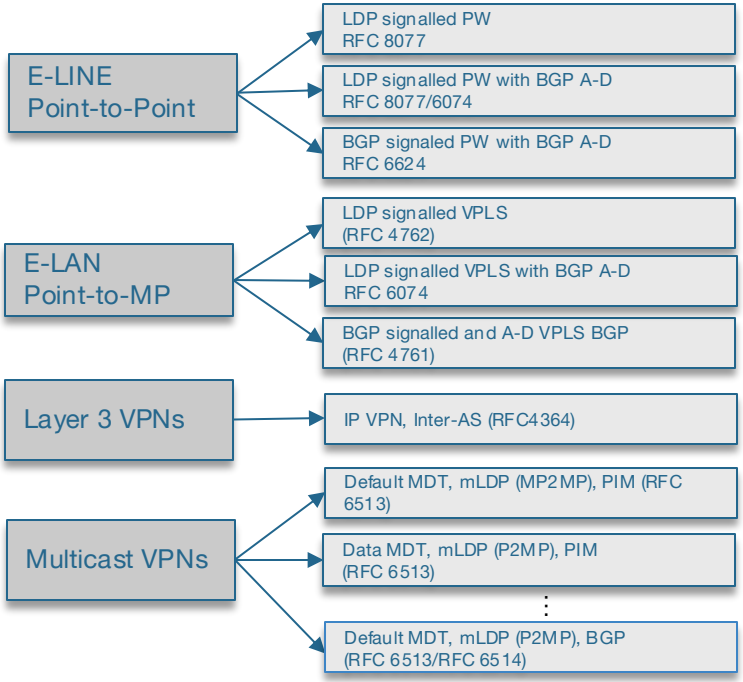
L2 VPN – A short history…

- Simple task of creating a P-to-P L2 VPNs between two logical interfaces
  - No MAC learning, no routing, just a logical pipe to replicate a dark fiber circuit
  - Now commonly referred to as a Pseudowire

- Two competing drafts – Kompella brothers from ALU and Juniper
  - RFC 4762 LDP based control-plane to signal MPLS label to the identify the L2 VPN.
    - Manual peer discovery later replaced with BGP for auto-discovery (RFC 6624)
  - RFC 4761, BGP based control-plane for signaling the MPLS label of the L2 VPN similar to L3VPN.
    - Since it's BGP, auto-discovery solved easily.

More complexity with the vendor's preferring one standard over another meaning interop severely restricted

Why have one answer when you can have multiple…



Complexity Madness

- Unique solutions for IP and Ethernet Services
- Vendor specific approaches and separate IETF drafts
- Multiple control planes, signaling models and BGP AFs
- Proprietary solutions to fill the gaps – multi-homing?

Waste of engineering resources, customer complexity with major interop challenges

Let Start Again…

- Learning our lessons and now that we understand the requirements, what if we could start again…

The Answer was – Ethernet VPN (EVPN)

- 2012 Requirements document for deploying Ethernet VPN services
  - Multi-vendor supported draft published March 2012 - draft-ietf-l2vpn-evpn-req-00
  - Defines requirements for Ethernet P2P and P2MP services, optimal forwarding, fast convergence, load-balancing, multi-homing and flood suppression
  - After 7 subsequent revisions, published as informational Standard **RFC 7209 in May 2014**

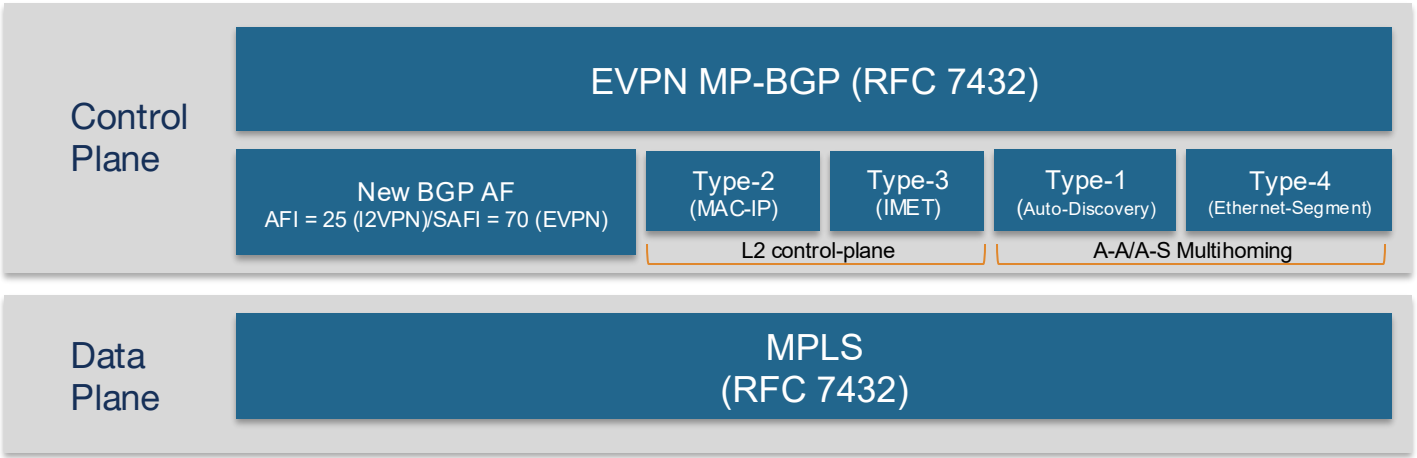

Internet Working Group
Internet Draft
Category: Standards Track
Ali Sajassi(Editor)
Samer Salam
Clarence Filsfils
Cisco
N. Aggarwal(Editor)
Juniper Networks
Nabil Bitar
Verizon
Jim Uttaro
AT&T
Aldrin Isaac
Bloomberg
Wim Henderickx
Alcatel-Lucent
Expires: September 27, 2012
March 27, 2012
Requirements for Ethernet VPN (E-VPN)
draft-ietf-l2vpn-evpn-req-00.txt

- 2012 Specifications for implementing Ethernet VPN services
  - Specification based on the requirements drafts/RFC 7209
  - Multi-vendor support, 1st draft published 2012 - draft-ietf-l2vpn-evpn-00
  - Defines a new BGP control-plane, AF and multihoming, extendable to **ALL** Ethernet Services
  - After 11 subsequent revisions, published as informational Standard **RFC 7432 in Feb 2015**

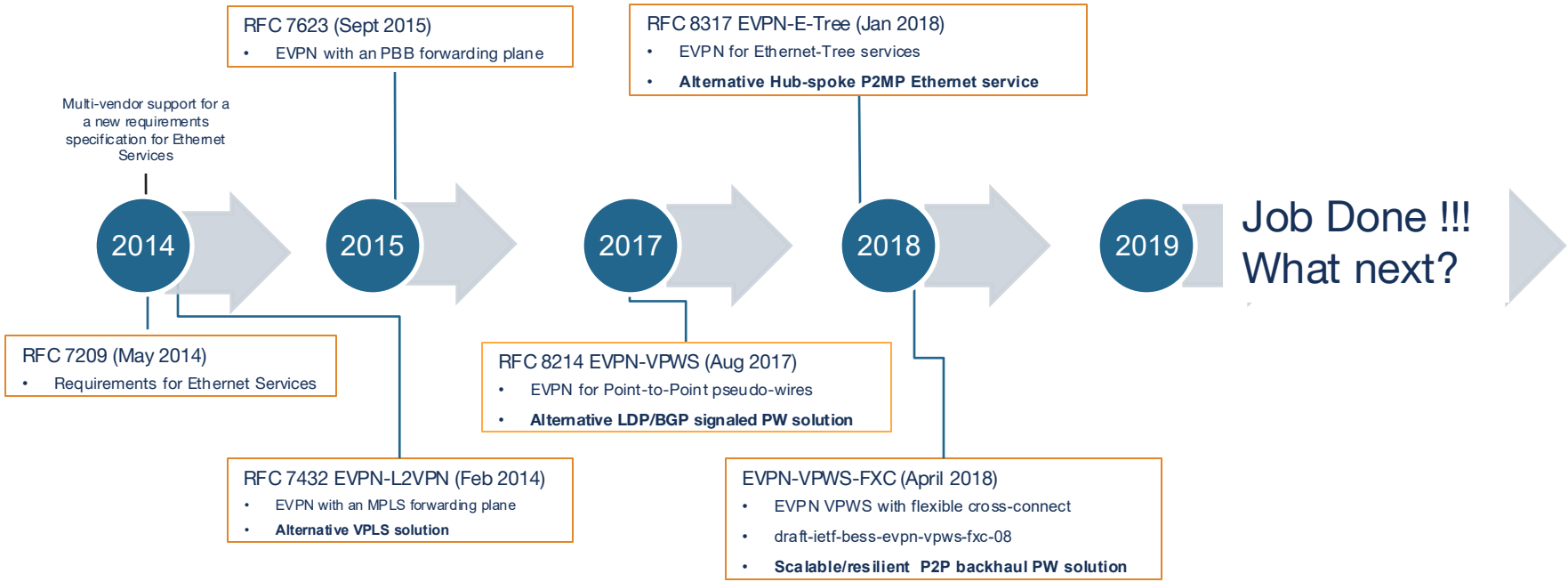

Network Working Group
INTERNET-DRAFT
Category: Standards Track
Expires: August 25, 2012
R. Aggarwal
Ashtan
A. Sajassi
Cisco
J. Uttaro
AT&T
W. Henderickx
Alcatel-Lucent
A. Isaac
Bloomberg
N. Bitar
Verizon
F. Balus
Alcatel-Lucent
K. Shekhar
J. Drake
Juniper Networks
S. Boutros
K. Patel
Cisco
February 24, 2012
BGP MPLS Based Ethernet VPN
draft-ietf-l2vpn-evpn-00

What did we get…

- New BGP control-plane for Ethernet Services – RFC 7432
  - New BGP address family, with support for multi-tenancy through RT/RD constructs
  - BGP control-plane for advertising MAC addresses and MAC-IP bindings
  - ARP suppression techniques to reduce flooding
  - Auto-discovery of peer PE nodes in the same MAC-VRF
  - Standardized multi-homing model for All-Active and Active-Standby forwarding models



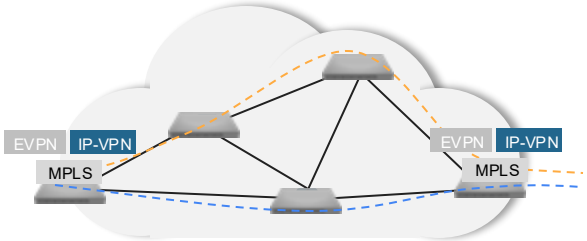
And we are off…



Market dynamics…

Traditional

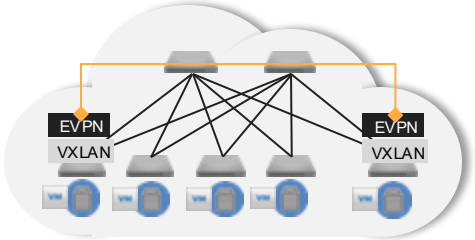
Telco/SP Market
MPLS transport for FRR and TE



- Telco Backbone requirement for VPN services
  - Layer 2 Ethernet Services
  - Layer 3, IP and multicast services
- MPLS transport required for VPN services
  - Scarcity of BW and available links
  - Driving TE and FRR requirements

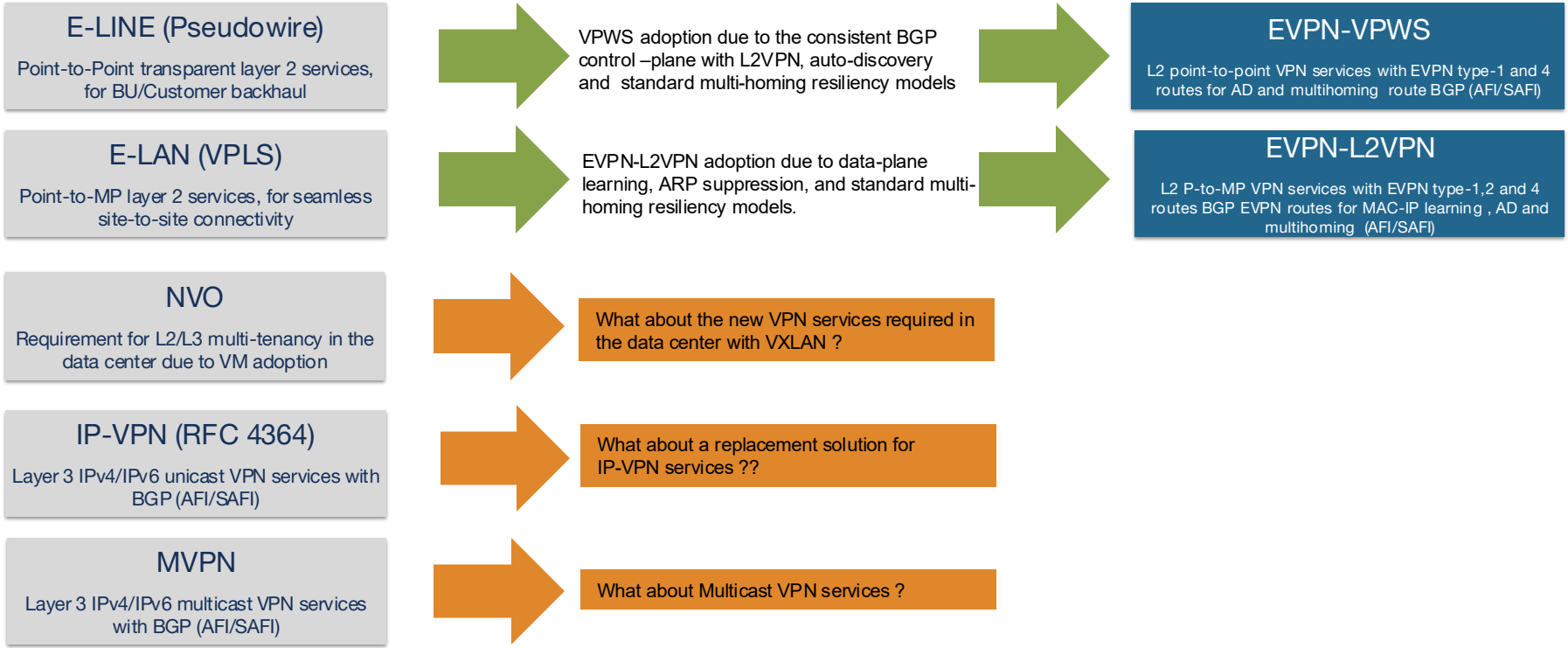
New-2014+

Telco Cloud/Enterprise DC
Virtualized Data Center

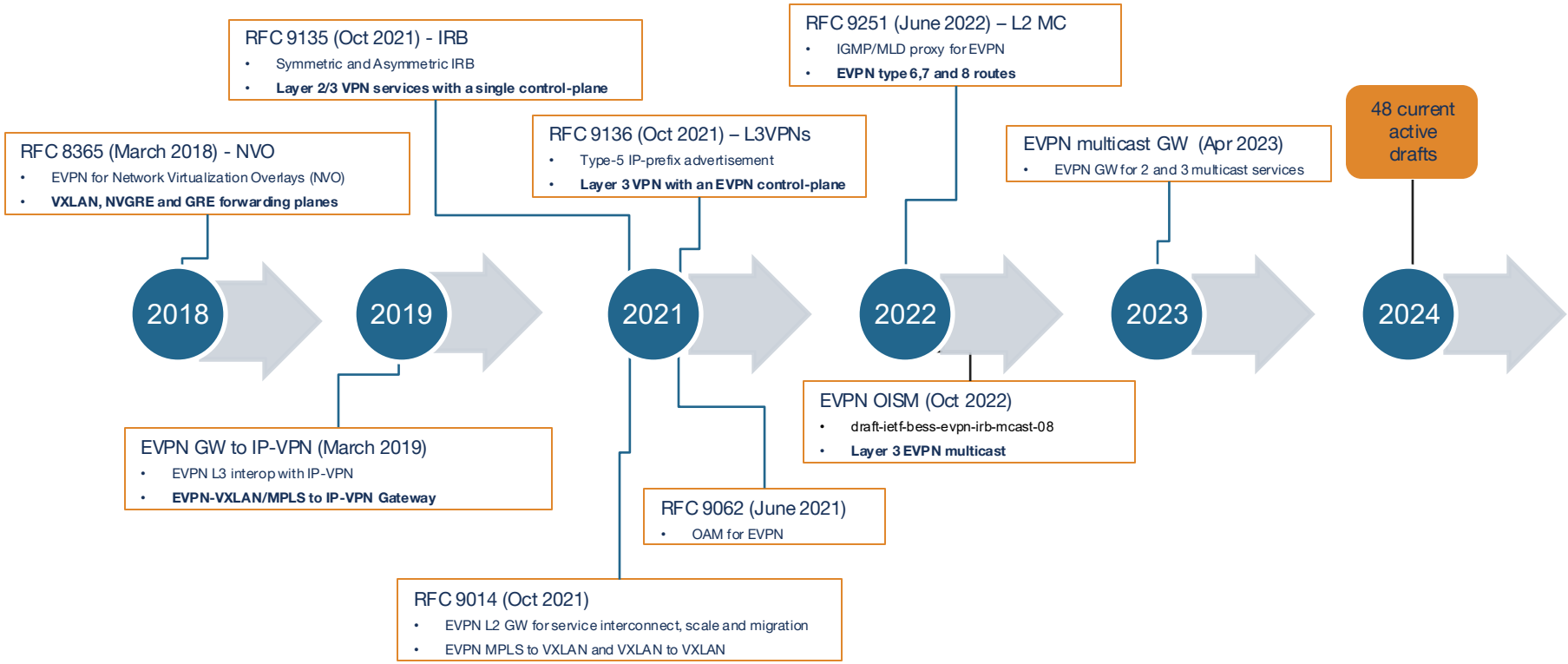


- New Virtualized DC requirement for VPN services
  - Driving an “any server and anywhere” model
  - Combined Layer 2 and 3 VPN Services
- Low-cost transport for the DC – VXLAN
  - BW freely available with ECMP for fast-failover
  - Complex costly TE and FRR benefits of MPLS NOT required

What next?

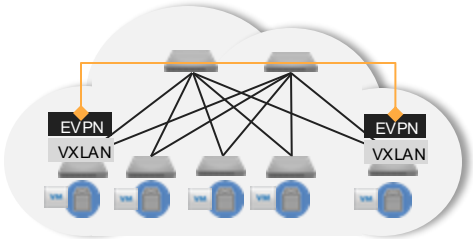

E-LINE (Pseudowire)
Point-to-Point transparent layer 2 services, for BU/Customer backhaul
VPWS adoption due to the consistent BGP control –plane with L2VPN, auto-discovery and standard multi-homing resiliency models
EVPN-VPWS
L2 point-to-point VPN services with EVPN type-1 and 4 routes for AD and multihoming route BGP (AFI/SAFI)
E-LAN (VPLS)
Point-to-MP layer 2 services, for seamless site-to-site connectivity
EVPN-L2VPN adoption due to data-plane learning, ARP suppression, and standard multi-homing resiliency models.
EVPN-L2VPN
L2 P-to-MP VPN services with EVPN type-1,2 and 4 routes BGP EVPN routes for MAC-IP learning , AD and multihoming (AFI/SAFI)
NVO
Requirement for L2/L3 multi-tenancy in the data center due to VM adoption
What about the new VPN services required in the data center with VXLAN ?
IP-VPN (RFC 4364)
Layer 3 IPv4/IPv6 unicast VPN services with BGP (AFI/SAFI)
What about a replacement solution for IP-VPN services ??
MVPN
Layer 3 IPv4/IPv6 multicast VPN services with BGP (AFI/SAFI)
What about Multicast VPN services ?

And we are off again…



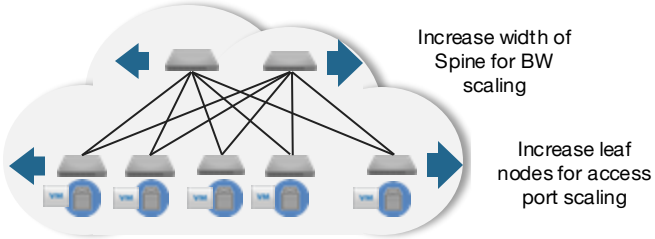
Problems to be addressed in the Data Center

- Consistent East-to-West traffic performance
- Multi-tenancy across a shared infrastructure
- Any server anywhere to optimize DC resources - L2 and 3 VPNs
- Resiliency and redundant server and application connectivity
- Multicast VPNs - media and financial customers
- Scale HW density radix deployments
- Seamless L2 and L3 VPNs across any transport



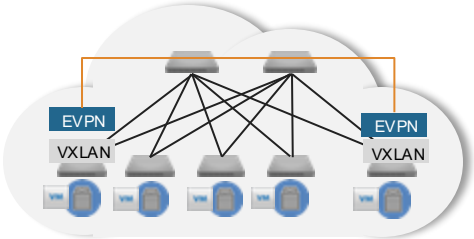
EVPN in the Data Center

- IP Leaf & Spine Fabric
  - Leaf nodes residing at the top of each rack
  - Spine IP transport for interconnecting all leaf nodes
  - All leaf nodes are equidistant
  - Consistent east-to-west performance & hop count
  - Scale-out
    - Increase access port density by adding more leaf nodes
    - Increase leaf node bandwidth and/or max. leaf nodes by adding more spines



Consistent throughput for inter-rack communication

- EVPN with VXLAN (RFC 8365)
  - EVPN for multi-tenancy Layer 2 and 3 VPNs
  - Layer 2 and 3 VPNs across all leaf nodes
  - BGP control-plane learning, MAC, MAC-IP and IP-prefix
  - Spine nodes act as simple IP transport
  - Support for both IPv4 and IPv6 underlay



EVPN-VXLAN control plane for L2 and L3 VPNs across the shared infrastructure

Why VXLAN and not MPLS?

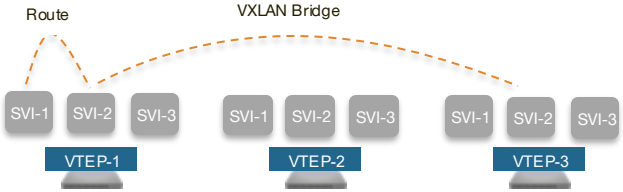
- **V**irtualized e**X**tensible **L**ocal **A**rea **N**etwork – RFC 7348
  - MAC in IP encapsulation for bridging traffic over IP
  - VXLAN header contains a 24-bit VNI label for transporting L2 and L3 VPN
  - UDP Source port a hash of the inner Payload for Entropy

- Simple encapsulation model supported on low-cost high BW Merchant Silicon
  - 24-bit VNI label suitable delivering L2 and L3 VPN across a IPv4/IPv6 transport
  - Payload entropy for ECMP load-balancing across the uplinks of a leaf-spine fabric
  - No DC requirements for TE due to the available BW – MPLS unnecessary costly
  - No DC requirement for FRR due to ECMP architecture of a leaf-spine topology

EVPN – Integrated Routing and Bridging (IRB)

- Optimal first-hop routing with Integrated Routing and Bridging (IRB) at the leaf
  - RFC 9135 published Oct 2021 - Integrated Routing and Bridging (IRB) with EVPN
  - Involves type-2 (MAC-IP) and type-5 (IP-prefixes) routes
  - Two forwarding model Asymmetric and Symmetrical IRB

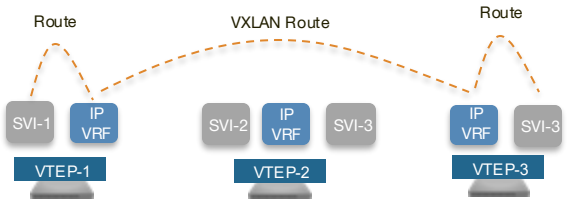
Asymmetric IRB



- SVI/VNIs configured on all VTEPs with an anycast IP.
- Routing at first-hop VTEP, bridging to remote host
- Return traffic, routed at the remote first-hop VTEP
- Asymmetric traffic path, scale concern SVIs on all VTEPs

Symmetric IRB

Industry approach



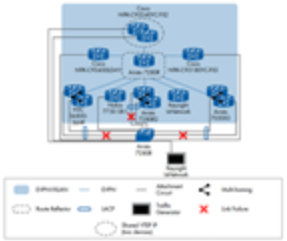
- SVI/VNIs only on the relevant VTEPs with an anycast IP.
- Routing at first-hop VTEP, into intermediate shared IP-VRF
- Routed at the remote VTEP into the destination subnet
- Symmetrical forwarding path, better scale, SVI & ARP only where required

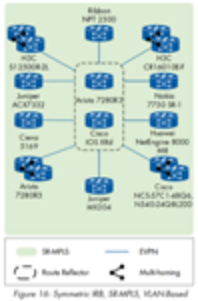
EVPN – Integrated Routing and Bridging (IRB)

- Vendor's initially supported one or both models with restricted interop
  - **Symmetric IRB** now the common model adopted across vendors
  - For improved scale; reduction in MAC routes and next-Hops
    - Only need to bridge when necessary – Asymmetric IRB always requires a bridging action
    - Host routes in same the VRF sharing the same next-hop

EANTC 2024

Symmetric IRB the **only** tested model across MPLS and VXLAN testbeds

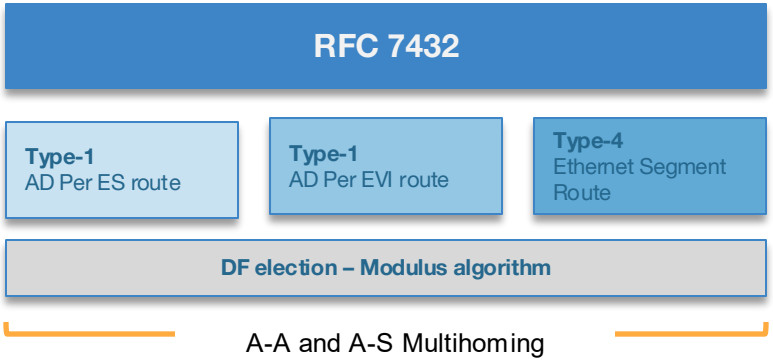


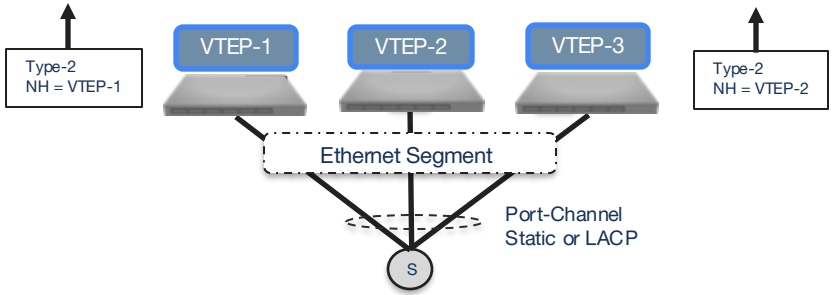


New IETF draft proposal to allow Symmetric and Asymmetric to interop

EVPN Interoperability Modes draft-ietf-bess-evpn-modes-interop-02

EVPN – New multi-homing proposals



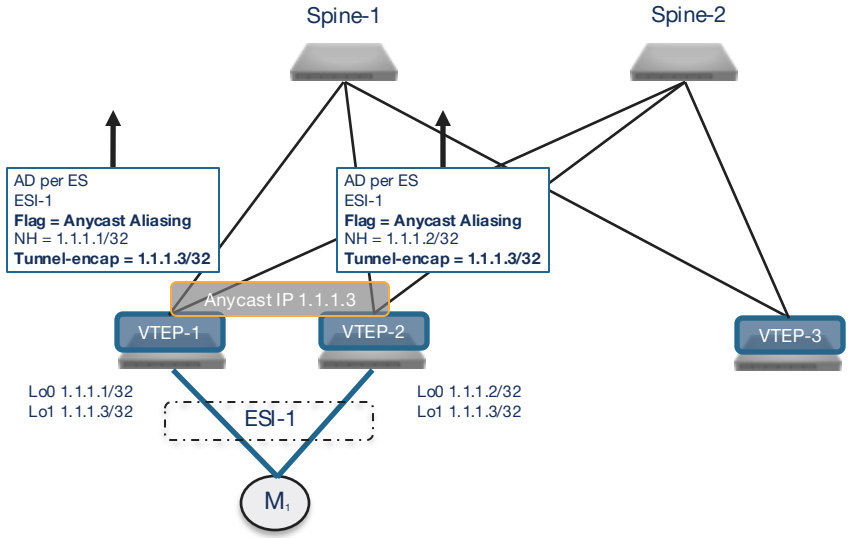


Draft	Overview
Framework for Ethernet VPN Designated Forwarder Election Extensibility **RFC 8584**	New DF algorithm community, HRW algorithm and AC-Influenced DF Election Capability
Preference based DF Election **Preference-based EVPN DF Election**	Preference based DF election algorithm using the DF algorithm community
Anycast Aliasing for multihoming **Draft-rabnag-bess-evpn-anycast-aliasing-00**	EVPN multi-homing with an anycast next-hop for ECMP in the underlay
IP aliasing **draft-sajassi-bess-evpn-ip-aliasing-09**	Layer 3 Fast Convergence and Aliasing and backup path for type-2 MAC-IP and type-5 (Ip-prefix) route
Fast Recovery for EVPN Designated Forwarder Election **draft-ietf-bess-evpn-fast-df-recovery**	Synchronised DF election to avoid temporary loops and packet duplication using NTP or PTP
EVPN Port-Active Redundancy Mode **draft-ietf-bess-evpn-mh-pa-10**	Port active multi-homing, physical interface of the ES placed in an active or disable state based on the DF election, can be used with LACP LAG

EVPN – DC multi-homing optimization

- **New proposed Anycast Aliasing draft**
  - draft-rabnag-bess-evpn-anycast-aliasing
  - Node's of the ES configured with a unique and shared VTEP IP address
  - Shared VTEP IP advertised as tunnel-encap attribute in the type-1 (AD per ES) route
  - Ingress VTEPs use the shared IP as next-hop for MACs learnt on the ES .

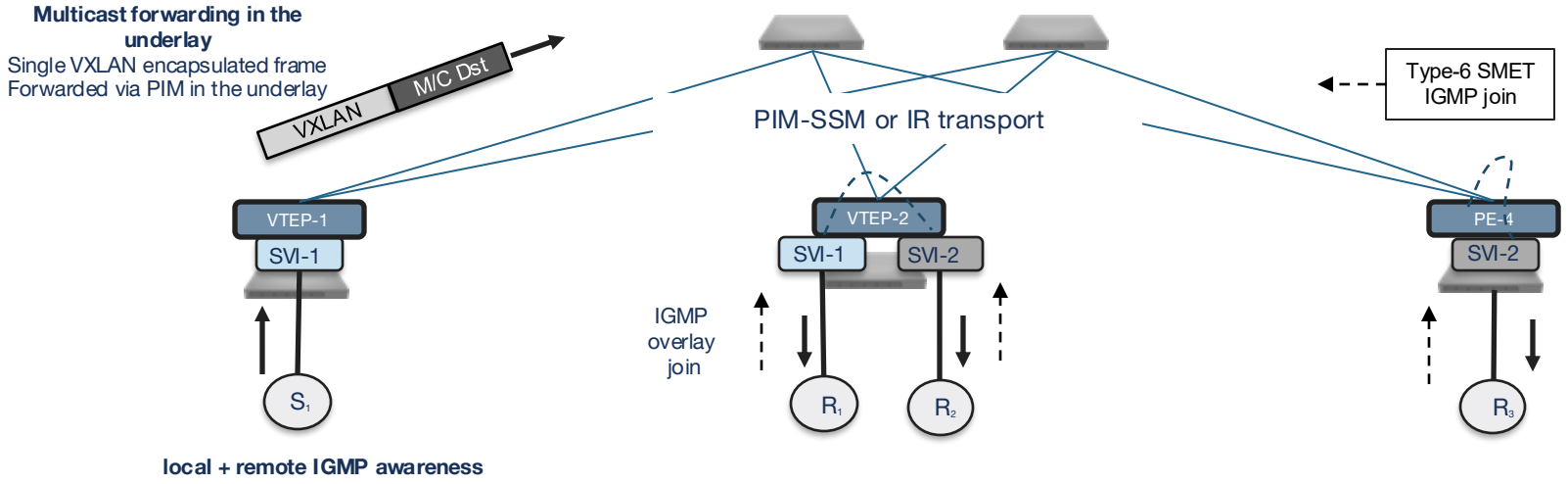
- **Benefits of the approach**
  - ECMP is now performed in the underlay via shared IP
  - Removes the need for type-1 per EVI route advertisement
  - Provides FRR in the event of local link failure on the ES



Still an early draft. Ongoing discussions within the IETF (Arista co-author)

EVPN multicast for optimal forwarding

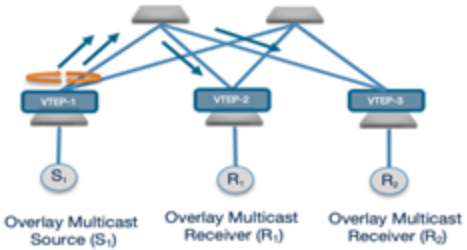
Standard based solution	New EVPN route types	Optimized multicast routing	PIM-SSM or IR transport
RFC 9251 RFC 9572 RCF 9625	**Type-6** for advertising local IGMP joins. Type-7 and 8 for supporting A-A multi-homing sources and receivers	OISM model, single copy to interested VTEP, egress VTEP responsible for routing to local receivers	VXLAN encapsulation with underlay PIM multicast or Ingress Replication for transporting the overlay multicast



EVPN Multicast – Transport models

Ingress Replication

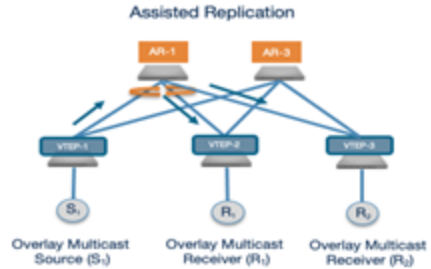
Generate a unique VXLAN unicast copy to each interested remote VTEP for the overlay multicast group



- High BW utilization on the uplink between leaf and spine
- Asynchronous delivery, due to replication model
- Suitable for low bandwidth apps

Assisted Replication

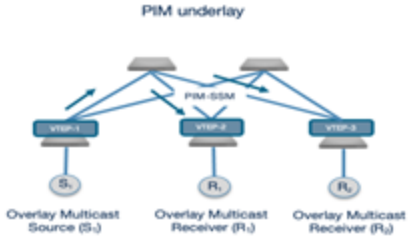
Off-load the replication process to the Spine node, which is now generates the unique unicast VXLAN packets to each interested remote VTEP



- BW on the uplink between spine and leaf reduced
- Still Asynchronous delivery, due to replication model
- Limited benefit over IR model

PIM (P2MP) transport

Enable PIM in the underlay, interested VTEPs join an associated multicast group in the underlay. Ingress VTEP VXLAN encaps with a multicast underlay destination, single VXLAN packet

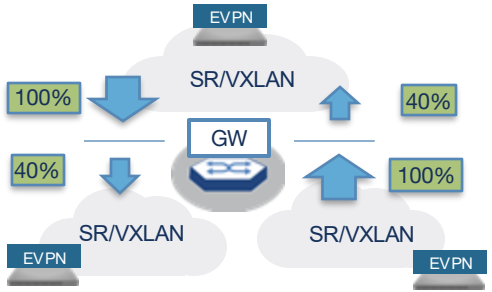


- BW on the uplink between spine and leaf reduced
- Fair synchronous delivery, no packet rewrite
- Suitable for high bandwidth, latency sensitive apps

EVPN – why do we need GWs?

EVPN scaling

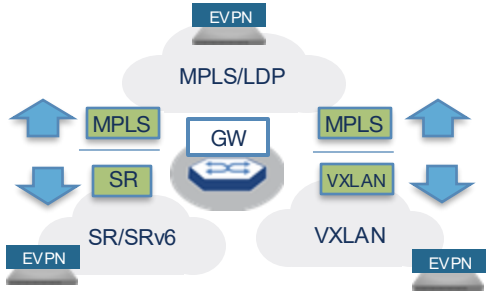
EVPN GW for hierarchical scaling when interconnecting multiple EVPN domain



- Provide L2 and L3 VPN connect across domains
- Hierarchy to reduce EVPN state across domains
- Improve HW table utilization (MAC, ARP etc)
- Improve SW memory utilization (RIB routes)

Transport Transition

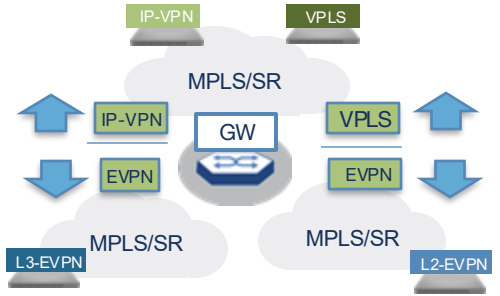
EVPN GW to provide seamless L2 and L3 VPN across any Data Center, WAN or SD-WAN transport



- Transport agnostic EVPN services
- Seamless L2 and L3 VPNs across the WAN
- DC Border GW connected to the MPLS/SR WAN
- ASBR GW for WAN upgrade from MPLS to SR

Service Migration and interop

EVPN GW to provide service interop with traditional IP, multicast and Ethernet services



- Migration from traditional Ethernet & IP services
- Seamless L2 and L3 VPNs between services
- Migration path from IP-VPN to L3-EVPN
- Migration path from VPLS to L2-EVPN

EVPN GW – Background

- EVPN GW functionality is standard based solution defined across multiple RFCs/Drafts
  - Covering support for both Layer 2 (type-2 & 3), L3 (type-5) and multicast (OISM) in the future
  - With support for different BGP AFs (migration) and data-plane encapsulations (VXLAN, MPLS, PBB, SRv6)
  - Widely adopted Industry approach (see EANTC testing) (with some exception)

Draft	Overview	
draft-ietf-bess-evpn-ipvpn-interworking-11	EVPN GW for L3 interconnect interop between EVPN-VXLAN/MPLS and IP-VPN	Industry adopted L3 GW
RFC 9014	EVPN GW for L2 interconnect between EVPN domains (VXLAN and MPLS encap)	Industry adopted L2 GW
draft-rabnic-bess-evpn-mcast-eeg-04	EVPN GW for L2 and L3 multicast interconnect EVPN domains (VXLAN and MPLS encap)	Future
draft-sr-bess-evpn-dpath-04	D-path for loop-free route advertisement across domain with an EVPN GW (VXLAN and MPLS encap)	
draft-sharma-bess-multi-site-evpn-05 (expired)	EVPN DCI solution focused only on EVPN-VXLAN no MPLS or AF support	Other

EVPN GW – EANTC interop testing

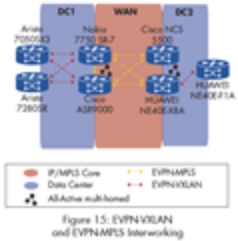
- **Extensive RFC 9014 multi-vendor interop testing**
  - Early testing focused on WAN/MPLS GW solution
  - Now multi-vendor support for multiple encapsulations (MPLS and VLANs)
  - Verified EVPN GW for both MPLS and VXLAN encapsulations
  - Industry direction for EVPN scaling and DCI, given the choice of AFs and encapsulation options

https://www.youtube.com/watch?v=RNi3QRXK3XU

2019 Interop testing
EVPN-VXLAN to EVPN-MPLS GW Interconnect

2020 Interop testing
EVPN-VXLAN to EVPN-MPLS GW Interconnect

2021 Interop testing
EVPN-VXLAN to EVPN-MPLS GW Interconnect



2022 Interop testing
EVPN-VXLAN to EVPN-VXLAN GW Interconnect

2023 Interop testing
EVPN-VXLAN to EVPN-VXLAN GW Interconnect

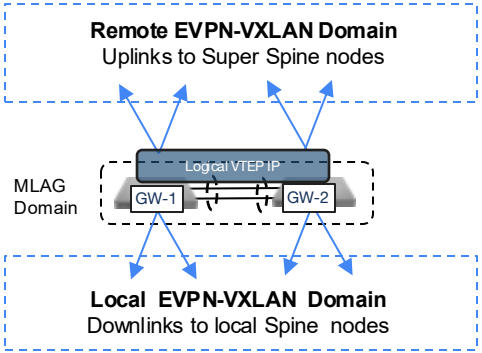
2024 Interop testing
EVPN-VXLAN to EVPN-VXLAN GW Interconnect

Multi-vendor support, industry direction is RFC 9014

EVPN GW – Redundancy models

EVPN-GW with MLAG

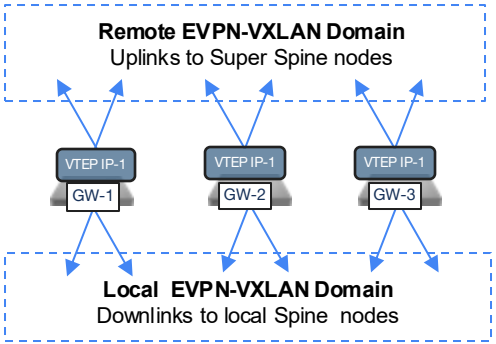
(VXLAN to VXLAN)



- Logical VTEP IP next-hop of EVPN routes
- ECMP load-balancing in the underlay
- Peer-link required between the nodes
- Support for directly attached hosts
- Limit single MLAG domain per-site

EVPN-GW with Anycast IP

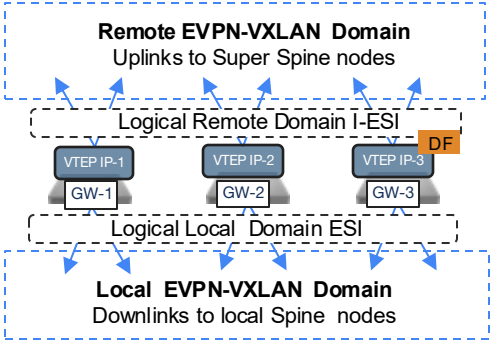
(VXLAN to VXLAN)



- Anycast IP next-hop of EVPN routes
- ECMP load-balancing in the underlay
- No peer-link required between the nodes
- No support for directly attached hosts
- Support for 16 GW nodes per-site

EVPN-GW with All-Active

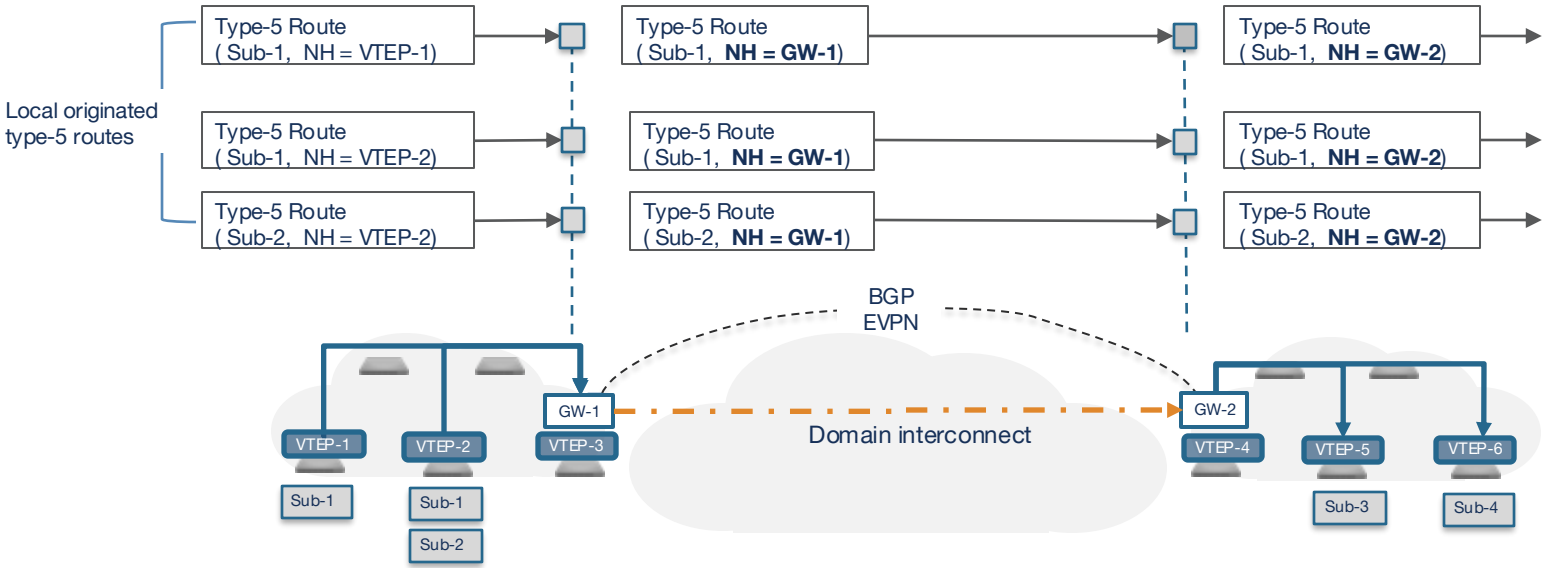
(VXLAN to VXLAN)



- GW nodes configured in an I-ESI
- ECMP load-balancing in the overlay
- No peer-link required between the nodes
- Support for directly attached hosts
- Support for 16 GW nodes per-site

EVPN L3 GW – next-hop self model

- EVPN GW Layer 3 forwarding model defined in *draft-ietf-bess-dci-evpn-overlay* draft
- The GW re-advertises local type-5 routes, with next-hop changed to the GW
- The loopback IPs (NH) of the local domain VTEPs are hidden by the GW
- The GW does NOT perform a best-path export and RD rewrite (NOT re-originated)
- If the GW has multiple ECMP routes to the same ip-prefix within the DC (different RDs for each leaf), each individual type-5 route will be re-advertised by the GW - not a single route



EVPN GW Multicast

- **Support for OISM on the GW**
  - Type-6 routes can be advertised across domains, seamless Layer 2 & 3 multicast connectivity
  - Removes the need for the PEG m-route workaround.
  - All-Active redundancy support in the initial release for L2 multicast
  - Arista co-author of the new IETF draft