L2VPN Workgroup
INTERNET-DRAFT
Intended Status: Standards Track

A. Sajassi (Editor)
Cisco
J. Drake (Editor)
Juniper
N. Bitar
Nokia
R. Shekhar
Juniper
J. Uttaro
AT&T
W. Henderickx
Nokia

Expires: December 10, 2016

June 10, 2016

A Network Virtualization Overlay Solution using EVPN
draft-ietf-bess-evpn-overlay-04

Abstract

This document describes how Ethernet VPN (EVPN) [RFC7432] can be used as an Network Virtualization Overlay (NVO) solution and explores the various tunnel encapsulation options over IP and their impact on the EVPN control-plane and procedures. In particular, the following encapsulation options are analyzed: VXLAN, NVGRE, and MPLS over GRE.

Table of Contents

1 Introduction

In the context of this document, a Network Virtualization Overlay (NVO) is a solution to address the requirements of a multi-tenant data center, especially one with virtualized hosts, e.g., Virtual Machines (VMs). The key requirements of such a solution, as described in [Problem-Statement], are:

- Isolation of network traffic per tenant

- Support for a large number of tenants (tens or hundreds of thousands)

- Extending L2 connectivity among different VMs belonging to a given tenant segment (subnet) across different PODs within a data center or between different data centers

- Allowing a given VM to move between different physical points of attachment within a given L2 segment

The underlay network for NVO solutions is assumed to provide IP connectivity between NVO endpoints (NVEs).

This document describes how Ethernet VPN (EVPN) can be used as an NVO solution and explores applicability of EVPN functions and procedures. In particular, it describes the various tunnel encapsulation options for EVPN over IP, and their impact on the EVPN control-plane and procedures for two main scenarios:

a) when the NVE resides in the hypervisor, and
b) when the NVE resides in a Top of Rack (ToR) device

Note that the use of EVPN as an NVO solution does not necessarily mandate that the BGP control-plane be running on the NVE. For such scenarios, it is still possible to leverage the EVPN solution by using XMPP, or alternative mechanisms, to extend the control-plane to the NVE as discussed in [L3VPN-ENDSYSTEMS].

The possible encapsulation options for EVPN overlays that are analyzed in this document are:

- VXLAN and NVGRE
- MPLS over GRE

Before getting into the description of the different encapsulation options for EVPN over IP, it is important to highlight the EVPN solution's main features, how those features are currently supported,

and any impact that the encapsulation has on those features.

2 Specification of Requirements

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

3 Terminology

NVO: Network Virtualization Overlay

NVE: Network Virtualization Endpoint

VNI: Virtual Network Identifier (for VXLAN)

VSID: Virtual Subnet Identifier (for NVGRE)

EVPN: Ethernet VPN

EVI: An EVPN instance spanning the Provider Edge (PE) devices participating in that EVPN.

MAC-VRF: A Virtual Routing and Forwarding table for Media Access Control (MAC) addresses on a PE.

Ethernet Segment (ES): When a customer site (device or network) is connected to one or more PEs via a set of Ethernet links, then that set of links is referred to as an 'Ethernet segment'.

Ethernet Segment Identifier (ESI): A unique non-zero identifier that identifies an Ethernet segment is called an 'Ethernet Segment Identifier'.

Ethernet Tag: An Ethernet tag identifies a particular broadcast domain, e.g., a VLAN. An EVPN instance consists of one or more broadcast domains.

PE: Provider Edge device.

Single-Active Redundancy Mode: When only a single PE, among all the PEs attached to an Ethernet segment, is allowed to forward traffic to/from that Ethernet segment for a given VLAN, then the Ethernet segment is defined to be operating in Single-Active redundancy mode.

All-Active Redundancy Mode: When all PEs attached to an Ethernet

segment are allowed to forward known unicast traffic to/from that Ethernet segment for a given VLAN, then the Ethernet segment is defined to be operating in All-Active redundancy mode.

4 EVPN Features

EVPN was originally designed to support the requirements detailed in [RFC7209] and therefore has the following attributes which directly address control plane scaling and ease of deployment issues.

1) Control plane traffic is distributed with BGP and Broadcast and Multicast traffic is sent using a shared multicast tree or with ingress replication.

2) Control plane learning is used for MAC (and IP) addresses instead of data plane learning. The latter requires the flooding of unknown unicast and ARP frames; whereas, the former does not require any flooding.

3) Route Reflector is used to reduce a full mesh of BGP sessions among PE devices to a single BGP session between a PE and the RR. Furthermore, RR hierarchy can be leveraged to scale the number of BGP routes on the RR.

4) Auto-discovery via BGP is used to discover PE devices participating in a given VPN, PE devices participating in a given redundancy group, tunnel encapsulation types, multicast tunnel type, multicast members, etc.

5) All-Active multihoming is used. This allows a given customer device (CE) to have multiple links to multiple PEs, and traffic to/from that CE fully utilizes all of these links. This set of links is termed an Ethernet Segment (ES).

6) When a link between a CE and a PE fails, the PEs for that EVI are notified of the failure via the withdrawal of a single EVPN route. This allows those PEs to remove the withdrawing PE as a next hop for every MAC address associated with the failed link. This is termed 'mass withdrawal'.

7) BGP route filtering and constrained route distribution are leveraged to ensure that the control plane traffic for a given EVI is only distributed to the PEs in that EVI.

8) When a 802.1Q interface is used between a CE and a PE, each of the VLAN ID (VID) on that interface can be mapped onto a bridge table (for upto 4094 such bridge tables). All these bridge tables may be

mapped onto a single MAC-VRF (in case of VLAN-aware bundle service).

9) VM Mobility mechanisms ensure that all PEs in a given EVI know the ES with which a given VM, as identified by its MAC and IP addresses, is currently associated.

10) Route Targets are used to allow the operator (or customer) to define a spectrum of logical network topologies including mesh, hub & spoke, and extranets (e.g., a VPN whose sites are owned by different enterprises), without the need for proprietary software or the aid of other virtual or physical devices.

11) Because the design goal for NVO is millions of instances per common physical infrastructure, the scaling properties of the control plane for NVO are extremely important. EVPN and the extensions described herein, are designed with this level of scalability in mind.

5 Encapsulation Options for EVPN Overlays

5.1 VXLAN/NVGRE Encapsulation

Both VXLAN and NVGRE are examples of technologies that provide a data plane encapsulation which is used to transport a packet over the common physical IP infrastructure between Network Virtualization Edges (NVEs) - e.g., VXLAN Tunnel End Points (VTEPs) in VXLAN network. Both of these technologies include the identifier of the specific NVO instance, Virtual Network Identifier (VNI) in VXLAN and Virtual Subnet Identifier (VSID) in NVGRE, in each packet. In the remainder of this document we use VNI as the representation for NVO instance with the understanding that VSID can equally be used if the encapsulation is NVGRE unless it is stated otherwise.

Note that a Provider Edge (PE) is equivalent to a NVE/VTEP.

VXLAN encapsulation is based on UDP, with an 8-byte header following the UDP header. VXLAN provides a 24-bit VNI, which typically provides a one-to-one mapping to the tenant VLAN ID, as described in [RFC7348]. In this scenario, the ingress VTEP does not include an inner VLAN tag on the encapsulated frame, and the egress VTEP discards the frames with an inner VLAN tag. This mode of operation in [RFC7348] maps to VLAN Based Service in [RFC7432], where a tenant VLAN ID gets mapped to an EVPN instance (EVI).

VXLAN also provides an option of including an inner VLAN tag in the encapsulated frame, if explicitly configured at the VTEP. This mode of operation can map to VLAN Bundle Service in [RFC7432] because all

the tenant's tagged frames map to a single bridge table / MAC-VRF, and the inner VLAN tag is not used for lookup by the disposition PE when performing VXLAN decapsulation as described in section 6 of [RFC7348].

[NVGRE] encapsulation is based on [GRE] and it mandates the inclusion of the optional GRE Key field which carries the VSID. There is a one-to-one mapping between the VSID and the tenant VLAN ID, as described in [NVGRE] and the inclusion of an inner VLAN tag is prohibited. This mode of operation in [NVGRE] maps to VLAN Based Service in [RFC7432].

As described in the next section there is no change to the encoding of EVPN routes to support VXLAN or NVGRE encapsulation except for the use of BGP Encapsulation extended community to indicate the encapsulation type (e.g., VxLAN or NVGRE). However, there is potential impact to the EVPN procedures depending on where the NVE is located (i.e., in hypervisor or TOR) and whether multi-homing capabilities are required.

5.1.1 Virtual Identifiers Scope

Although VNIs are defined as 24-bit globally unique values, there are scenarios in which it is desirable to use a locally significant value for VNI, especially in the context of data center interconnect:

5.1.1.1 Data Center Interconnect with Gateway

In the case where NVEs in different data centers need to be interconnected, and the NVEs need to use VNIs as a globally unique identifiers within a data center, then a Gateway needs to be employed at the edge of the data center network. This is because the Gateway will provide the functionality of translating the VNI when crossing network boundaries, which may align with operator span of control boundaries. As an example, consider the network of Figure 1 below. Assume there are three network operators: one for each of the DC1, DC2 and WAN networks. The Gateways at the edge of the data centers are responsible for translating the VNIs between the values used in each of the data center networks and the values used in the WAN.

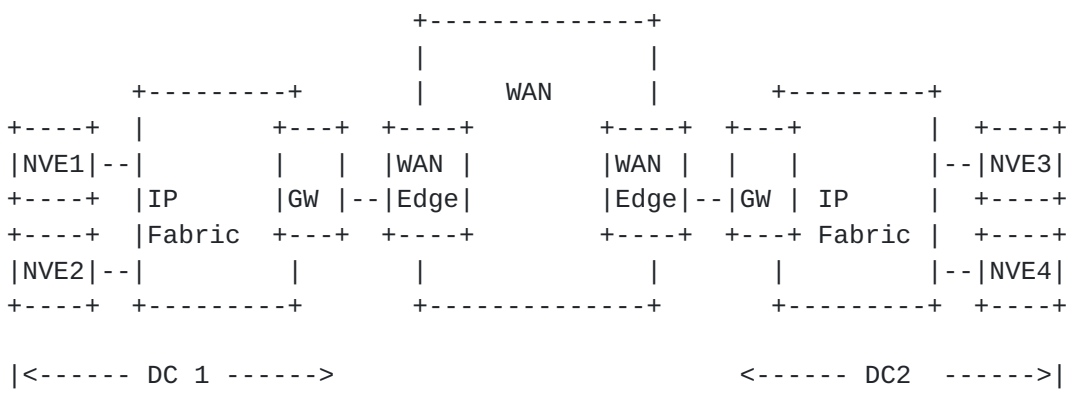


Figure 1: Data Center Interconnect with Gateway

5.1.1.2 Data Center Interconnect without Gateway

In the case where NVEs in different data centers need to be interconnected, and the NVEs need to use locally assigned VNIs (e.g., similar to MPLS labels), then there may be no need to employ Gateways at the edge of the data center network. More specifically, the VNI value that is used by the transmitting NVE is allocated by the NVE that is receiving the traffic (in other words, this is similar to "downstream assigned" MPLS label). This allows the VNI space to be decoupled between different data center networks without the need for a dedicated Gateway at the edge of the data centers. This topics is covered in section 10.2.

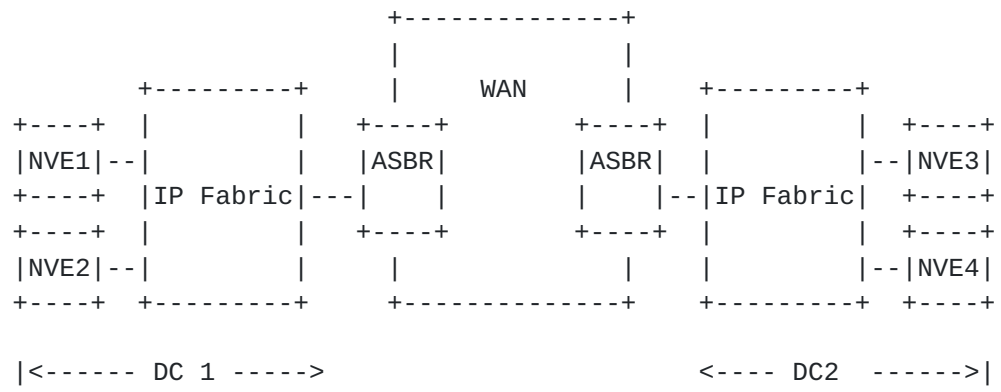


Figure 2: Data Center Interconnect with ASBR

5.1.2 Virtual Identifiers to EVI Mapping

When the EVPN control plane is used in conjunction with VXLAN (or NVGRE encapsulation), two options for mapping the VXLAN VNI (or NVGRE VSID) to an EVI are possible:

1. Option 1: Single Subnet per EVI

In this option, a single subnet represented by a VNI is mapped to a unique EVI. This corresponds to the VLAN Based service in [RFC7432], where a tenant VLAN ID gets mapped to an EVPN instance (EVI). As such, a BGP RD and RT is needed per VNI on every NVE. The advantage of this model is that it allows the BGP RT constraint mechanisms to be used in order to limit the propagation and import of routes to only the NVEs that are interested in a given VNI. The disadvantage of this model may be the provisioning overhead if RD and RT are not derived automatically from VNI.

In this option, the MAC-VRF table is identified by the RT in the control plane and by the VNI in the data-plane. In this option, the specific the MAC-VRF table corresponds to only a single bridge table.

2. Option 2: Multiple Subnets per EVI

In this option, multiple subnets each represented by a unique VNI are mapped to a single EVI. For example, if a tenant has multiple segments/subnets each represented by a VNI, then all the VNIs for that tenant are mapped to a single EVI - e.g., the EVI in this case represents the tenant and not a subnet . This corresponds to the VLAN-aware bundle service in [RFC7432]. The advantage of this model is that it doesn't require the provisioning of RD/RT per VNI. However, this is a moot point if option 1 with auto-derivation is used. The disadvantage of this model is that routes would be imported by NVEs that may not be interested in a given VNI.

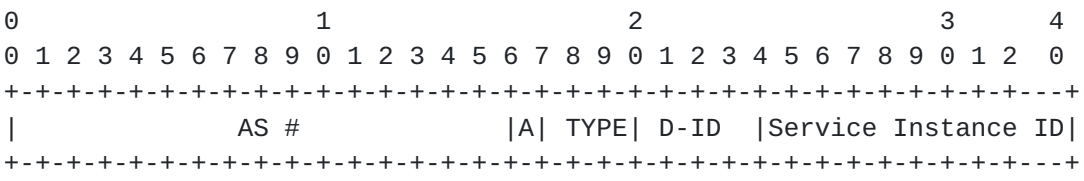
In this option the MAC-VRF table is identified by the RT in the control plane and a specific bridge table for that MAC-VRF is identified by the <RT, Ethernet Tag ID> in the control plane. In this option, the VNI in the data-plane is sufficient to identify a specific bridge table - e.g., no need to do a lookup based on VNI and Ethernet Tag ID fields to identify a bridge table.

5.1.2.1 Auto Derivation of RT

When the option of a single VNI per EVI is used, it is important to auto-derive RT for EVPN BGP routes in order to simplify configuration for data center operations. RD can be auto generated as described in [RFC7432] and RT can be auto-derived as described next.

Since a gateway PE as depicted in figure-1 participates in both the DCN and WAN BGP sessions, it is important that when RT values are auto-derived for VNIs, there is no conflict in RT spaces between DCN

and WAN networks assuming that both are operating within the same AS. Also, there can be scenarios where both VXLAN and NVGRE encapsulations may be needed within the same DCN and their corresponding VNIs are administered independently which means VNI spaces can overlap. In order to ensure that no such conflict in RT spaces arises, RT values for DCNs are auto-derived as follow:



- 2 bytes of global admin field of the RT is set to the AS number.

- Three least significant bytes of the local admin field of the RT is set to the VNI, VSID, I-SID, or VID.

- The most significant bit of the local admin field of the RT is set as follow:
  0: auto-derived
  1: manually-derived

- The next 3 bits of the most significant byte of the local admin field of the RT identifies the space in which the other 3 bytes are defined. The following spaces are defined:
  0 : VID
  1 : VXLAN
  2 : NVGRE
  3 : I-SID
  4 : EVI
  5 : dual-VID

- The remaining 4 bits of the most significant byte of the local admin field of the RT identifies the domain-id. The default value of domain-id is zero indicating that only a single numbering space exist for a given technology. However, if there are more than one number space exist for a given technology (e.g., overlapping VXLAN spaces), then each of the number spaces need to be identify by their corresponding domain-id starting from 1.

5.1.3 Constructing EVPN BGP Routes

In EVPN, an MPLS label is distributed by the egress PE via the EVPN control plane and is placed in the MPLS header of a given packet by

the ingress PE. This label is used upon receipt of that packet by the egress PE for disposition of that packet. This is very similar to the use of the VNI by the egress NVE, with the difference being that an MPLS label has local significance while a VNI typically has global significance. Accordingly, and specifically to support the option of locally assigned VNIs, the MPLS label field in the MAC Advertisement, Ethernet AD per EVI, and Inclusive Multicast Ethernet Tag routes is used to carry the VNI. For the balance of this memo, the MPLS label field will be referred to as the VNI field. The VNI field is used for both local and global VNIs, and for either case the entire 24-bit field is used to encode the VNI value.

For the VLAN-based service (a single VNI per MAC-VRF), the Ethernet Tag field in the MAC/IP Advertisement, Ethernet AD per EVI, and Inclusive Multicast route MUST be set to zero just as in the VLAN Based service in [RFC7432].

For the VLAN-aware bundle service (multiple VNIs per MAC-VRF with each VNI associated with its own bridge table), the Ethernet Tag field in the MAC Advertisement, Ethernet AD per EVI, and Inclusive Multicast route MUST identify a bridge table within a MAC-VRF and the set of Ethernet Tags for that EVI needs to be configured consistently on all PEs within that EVI. For local VNIs, the value advertised in the Ethernet Tag field MUST be set to a VID just as in the VLAN-aware bundle service in [RFC7432]. Such setting must be done consistently on all PE devices participating in that EVI within a given domain. For global VNIs, the value advertised in the Ethernet Tag field SHOULD be set to a VNI as long as it matches the existing semantics of the Ethernet Tag, i.e., it identifies a bridge table within a MAC-VRF and the set of VNIs are configured consistently on each PE in that EVI.

In order to indicate that which type of data plane encapsulation (i.e., VXLAN, NVGRE, MPLS, or MPLS in GRE) is to be used, the BGP Encapsulation extended community defined in [TUNNEL-ENCAP]and [RFC5512] is included with all EVPN routes (i.e. MAC Advertisement, Ethernet AD per EVI, Ethernet AD per ESI, Inclusive Multicast Ethernet Tag, and Ethernet Segment) advertised by an egress PE. Five new values have been assigned by IANA to extend the list of encapsulation types defined in [TUNNEL-ENCAP] and they are listed in section 13.

The MPLS encapsulation tunnel type, listed in section 13, is needed in order to distinguish between an advertising node that only supports non-MPLS encapsulations and one that supports MPLS and non-MPLS encapsulations. An advertising node that only supports MPLS encapsulation does not need to advertise any encapsulation tunnel

types; i.e., if the BGP Encapsulation extended community is not present, then either MPLS encapsulation or a statically configured encapsulation is assumed.

The Ethernet Segment and Ethernet AD per ESI routes MAY be advertised with multiple encapsulation types as long as they use the same EVPN multi-homing procedures - e.g., the mix of VXLAN and NVGRE encapsulation types is a valid one but not the mix of VXLAN and MPLS encapsulation types.

The Next Hop field of the MP_REACH_NLRI attribute of the route MUST be set to the IPv4 or IPv6 address of the NVE. The remaining fields in each route are set as per [RFC7432].

5.2 MPLS over GRE

The EVPN data-plane is modeled as an EVPN MPLS client layer sitting over an MPLS PSN-tunnel server layer. Some of the EVPN functions (split-horizon, aliasing, and backup-path) are tied to the MPLS client layer. If MPLS over GRE encapsulation is used, then the EVPN MPLS client layer can be carried over an IP PSN tunnel transparently. Therefore, there is no impact to the EVPN procedures and associated data-plane operation.

The existing standards for MPLS over GRE encapsulation as defined by [RFC4023] can be used for this purpose; however, when it is used in conjunction with EVPN the GRE key field SHOULD be present, and SHOULD be used to provide a 32-bit entropy field. The Checksum and Sequence Number fields are not needed and their corresponding C and S bits MUST be set to zero. A PE capable of supporting this encapsulation, should advertise its EVPN routes along with the Tunnel Encapsulation extended community indicating MPLS over GRE encapsulation, as described in previous section.

6 EVPN with Multiple Data Plane Encapsulations

The use of the BGP Encapsulation extended community per [TUNNEL-ENCAP] and [RFC5512] allows each NVE in a given EVI to know each of the encapsulations supported by each of the other NVEs in that EVI. i.e., each of the NVEs in a given EVI may support multiple data plane encapsulations. An ingress NVE can send a frame to an egress NVE only if the set of encapsulations advertised by the egress NVE in the subject MAC/IP Advertisement or per EVI Ethernet AD route, forms a non-empty intersection with the set of encapsulations supported by the ingress NVE, and it is at the discretion of the ingress NVE which encapsulation to choose from this intersection. (As noted in

section 5.1.3, if the BGP Encapsulation extended community is not present, then the default MPLS encapsulation or a statically configured encapsulation is assumed.)

An ingress node that uses shared multicast trees for sending broadcast or multicast frames MUST maintain distinct trees for each different encapsulation type.

It is the responsibility of the operator of a given EVI to ensure that all of the NVEs in that EVI support at least one common encapsulation. If this condition is violated, it could result in service disruption or failure. The use of the BGP Encapsulation extended community provides a method to detect when this condition is violated but the actions to be taken are at the discretion of the operator and are outside the scope of this document.

7 NVE Residing in Hypervisor

When a NVE and its hosts/VMs are co-located in the same physical device, e.g., when they reside in a server, the links between them are virtual and they typically share fate; i.e., the subject hosts/VMs are typically not multi-homed or if they are multi-homed, the multi-homing is a purely local matter to the server hosting the VM and the NVEs, and need not be "visible" to any other NVEs residing on other servers, and thus does not require any specific protocol mechanisms. The most common case of this is when the NVE resides on the hypervisor.

In the sub-sections that follow, we will discuss the impact on EVPN procedures for the case when the NVE resides on the hypervisor and the VXLAN (or NVGRE) encapsulation is used.

7.1 Impact on EVPN BGP Routes & Attributes for VXLAN/NVGRE Encapsulation

In the scenario where all data centers are under a single administrative domain, and there is a single global VNI space, the RD MAY be set to zero in the EVPN routes. However, in the scenario where different groups of data centers are under different administrative domains, and these data centers are connected via one or more backbone core providers as described in [NOV3-Framework], the RD must be a unique value per EVI or per NVE as described in [RFC7432]. In other words, whenever there is more than one administrative domain for global VNI, then a non-zero RD MUST be used, or whenever the VNI value have local significance, then a non-zero RD MUST be used. It is recommend to use a non-zero RD at all time.

When the NVEs reside on the hypervisor, the EVPN BGP routes and attributes associated with multi-homing are no longer required. This

reduces the required routes and attributes to the following subset of four out of eight:

- MAC/IP Advertisement Route
- Inclusive Multicast Ethernet Tag Route
- MAC Mobility Extended Community
- Default Gateway Extended Community

However, as noted in section 8.6 of [RFC7432] in order to enable a single-homing ingress NVE to take advantage of fast convergence, aliasing, and backup-path when interacting with multi-homed egress NVEs attached to a given Ethernet segment, the single-homing ingress NVE SHOULD be able to receive and process Ethernet AD per ES and Ethernet AD per EVI routes.

7.2 Impact on EVPN Procedures for VXLAN/NVGRE Encapsulation

When the NVEs reside on the hypervisors, the EVPN procedures associated with multi-homing are no longer required. This limits the procedures on the NVE to the following subset of the EVPN procedures:

1. Local learning of MAC addresses received from the VMs per section 10.1 of [RFC7432].

2. Advertising locally learned MAC addresses in BGP using the MAC/IP Advertisement routes.

3. Performing remote learning using BGP per Section 10.2 of [RFC7432].

4. Discovering other NVEs and constructing the multicast tunnels using the Inclusive Multicast Ethernet Tag routes.

5. Handling MAC address mobility events per the procedures of Section 16 in [RFC7432].

However, as noted in section 8.6 of [RFC7432] in order to enable a single-homing ingress NVE to take advantage of fast convergence, aliasing, and back-up path when interacting with multi-homed egress NVEs attached to a given Ethernet segment, a single-homing ingress NVE SHOULD implement the ingress node processing of Ethernet AD per ES and Ethernet AD per EVI routes as defined in sections 8.2 Fast Convergence and 8.4 Aliasing and Backup-Path of [RFC7432].

8 NVE Residing in ToR Switch

In this section, we discuss the scenario where the NVEs reside in the

Top of Rack (ToR) switches AND the servers (where VMs are residing) are multi-homed to these ToR switches. The multi-homing may operate in All-Active or Single-Active redundancy mode. If the servers are single-homed to the ToR switches, then the scenario becomes similar to that where the NVE resides on the hypervisor, as discussed in Section 7, as far as the required EVPN functionality are concerned.

[RFC7432] defines a set of BGP routes, attributes and procedures to support multi-homing. We first describe these functions and procedures, then discuss which of these are impacted by the VxLAN (or NVGRE) encapsulation and what modifications are required.

8.1 EVPN Multi-Homing Features

In this section, we will recap the multi-homing features of EVPN to highlight the encapsulation dependencies. The section only describes the features and functions at a high-level. For more details, the reader is to refer to [RFC7432].

8.1.1 Multi-homed Ethernet Segment Auto-Discovery

EVPN NVEs (or PEs) connected to the same Ethernet Segment (e.g. the same server via LAG) can automatically discover each other with minimal to no configuration through the exchange of BGP routes.

8.1.2 Fast Convergence and Mass Withdraw

EVPN defines a mechanism to efficiently and quickly signal, to remote NVEs, the need to update their forwarding tables upon the occurrence of a failure in connectivity to an Ethernet segment (e.g., a link or a port failure). This is done by having each NVE advertise an Ethernet A-D Route per Ethernet segment for each locally attached segment. Upon a failure in connectivity to the attached segment, the NVE withdraws the corresponding Ethernet A-D route. This triggers all NVEs that receive the withdrawal to update their next-hop adjacencies for all MAC addresses associated with the Ethernet segment in question. If no other NVE had advertised an Ethernet A-D route for the same segment, then the NVE that received the withdrawal simply invalidates the MAC entries for that segment. Otherwise, the NVE updates the next-hop adjacency list accordingly.

8.1.3 Split-Horizon

If a server is multi-homed to two or more NVEs (represented by an Ethernet segment ES1) and operating in an all-active redundancy mode, sends a BUM packet (ie, Broadcast, Unknown unicast, or Multicast) packet to one of these NVEs, then it is important to ensure the packet is not looped back to the server via another NVE connected to

this server. The filtering mechanism on the NVE to prevent such loop and packet duplication is called "split horizon filtering'.

8.1.4 Aliasing and Backup-Path

In the case where a station is multi-homed to multiple NVEs, it is possible that only a single NVE learns a set of the MAC addresses associated with traffic transmitted by the station. This leads to a situation where remote NVEs receive MAC advertisement routes, for these addresses, from a single NVE even though multiple NVEs are connected to the multi-homed station. As a result, the remote NVEs are not able to effectively load-balance traffic among the NVEs connected to the multi-homed Ethernet segment. This could be the case, for e.g. when the NVEs perform data-path learning on the access, and the load-balancing function on the station hashes traffic from a given source MAC address to a single NVE. Another scenario where this occurs is when the NVEs rely on control plane learning on the access (e.g. using ARP), since ARP traffic will be hashed to a single link in the LAG.

To alleviate this issue, EVPN introduces the concept of Aliasing. This refers to the ability of an NVE to signal that it has reachability to a given locally attached Ethernet segment, even when it has learnt no MAC addresses from that segment. The Ethernet A-D route per EVI is used to that end. Remote NVEs which receive MAC advertisement routes with non-zero ESI SHOULD consider the MAC address as reachable via all NVEs that advertise reachability to the relevant Segment using Ethernet A-D routes with the same ESI and with the Single-Active flag reset.

Backup-Path is a closely related function, albeit it applies to the case where the redundancy mode is Single-Active. In this case, the NVE signals that it has reachability to a given locally attached Ethernet Segment using the Ethernet A-D route as well. Remote NVEs which receive the MAC advertisement routes, with non-zero ESI, SHOULD consider the MAC address as reachable via the advertising NVE. Furthermore, the remote NVEs SHOULD install a Backup-Path, for said MAC, to the NVE which had advertised reachability to the relevant Segment using an Ethernet A-D route with the same ESI and with the Single-Active flag set.

8.1.5 DF Election

If a host is multi-homed to two or more NVEs on an Ethernet segment operating in all-active redundancy mode, then for a given EVI only one of these NVEs, termed the Designated Forwarder (DF) is

responsible for sending it broadcast, multicast, and, if configured for that EVI, unknown unicast frames.

This is required in order to prevent duplicate delivery of multi-destination frames to a multi-homed host or VM, in case of all-active redundancy.

In NVEs where .1Q tagged frames are received from hosts, the DF election is performed on host VLAN IDs (VIDs). It is assumed that for a given Ethernet Segment, VIDs are unique and consistent (e.g., no duplicate VIDs exist).

In GWs where VxLAN encapsulated frames are received, the DF election is performed on VNIs. Again, it is assumed that for a given Ethernet Segment, VNIs are unique and consistent (e.g., no duplicate VNIs exist).

8.2 Impact on EVPN BGP Routes & Attributes

Since multi-homing is supported in this scenario, then the entire set of BGP routes and attributes defined in [RFC7432] are used. The setting of the Ethernet Tag field in the MAC Advertisement, Ethernet AD per EVI, and Inclusive Multicast routes follows that of section 5.1.3. Furthermore, the setting of the VNI field in the MAC Advertisement and Ethernet AD per EVI routes follows that of section 5.1.3.

8.3 Impact on EVPN Procedures

Two cases need to be examined here, depending on whether the NVEs are operating in Active/Standby or in All-Active redundancy.

First, lets consider the case of Active/Standby redundancy, where the hosts are multi-homed to a set of NVEs, however, only a single NVE is active at a given point of time for a given VNI. In this case, the aliasing is not required and the split-horizon may not be required, but other functions such as multi-homed Ethernet segment auto-discovery, fast convergence and mass withdraw, backup path, and DF election are required.

Second, let's consider the case of All-Active redundancy. In this case, out of all the EVPN multi-homing features listed in section 8.1, the use of the VXLAN or NVGRE encapsulation impacts the split-horizon and aliasing features, since those two rely on the MPLS client layer. Given that this MPLS client layer is absent with these types of encapsulations, alternative procedures and mechanisms are

needed to provide the required functions. Those are discussed in detail next.

8.3.1 Split Horizon

In EVPN, an MPLS label is used for split-horizon filtering to support All-Active multi-homing where an ingress NVE adds a label corresponding to the site of origin (aka ESI Label) when encapsulating the packet. The egress NVE checks the ESI label when attempting to forward a multi-destination frame out an interface, and if the label corresponds to the same site identifier (ESI) associated with that interface, the packet gets dropped. This prevents the occurrence of forwarding loops.

Since the VXLAN or NVGRE encapsulation does not include this ESI label, other means of performing the split-horizon filtering function MUST be devised. The following approach is recommended for split-horizon filtering when VXLAN (or NVGRE) encapsulation is used.

Every NVE track the IP address(es) associated with the other NVE(s) with which it has shared multi-homed Ethernet Segments. When the NVE receives a multi-destination frame from the overlay network, it examines the source IP address in the tunnel header (which corresponds to the ingress NVE) and filters out the frame on all local interfaces connected to Ethernet Segments that are shared with the ingress NVE. With this approach, it is required that the ingress NVE performs replication locally to all directly attached Ethernet Segments (regardless of the DF Election state) for all flooded traffic ingress from the access interfaces (i.e. from the hosts). This approach is referred to as "Local Bias", and has the advantage that only a single IP address needs to be used per NVE for split-horizon filtering, as opposed to requiring an IP address per Ethernet Segment per NVE.

In order to prevent unhealthy interactions between the split horizon procedures defined in [RFC7432] and the local bias procedures described in this document, a mix of MPLS over GRE encapsulations on the one hand and VXLAN/NVGRE encapsulations on the other on a given Ethernet Segment is prohibited.

8.3.2 Aliasing and Backup-Path

The Aliasing and the Backup-Path procedures for VXLAN/NVGRE encapsulation is very similar to the ones for MPLS. In case of MPLS, two different Ethernet A-D routes are used for this purpose. The one used for Aliasing has a VPN scope (per EVI) and carries a VPN label but the one used for Backup-Path has Ethernet segment scope (per ES) and doesn't carry any VPN specific info (e.g., Ethernet Tag and MPLS

label are set to zero). In case of VxLAN/NVGRE, the same two routes are used for the Aliasing and the Backup-Path. In case of Aliasing, the Ethernet Tag and VNI fields in Ethernet A-D per EVI route is set as described in section 5.1.3.

9 Support for Multicast

The E-VPN Inclusive Multicast BGP route is used to discover the multicast tunnels among the endpoints associated with a given EVI (e.g., given VNI) for VLAN-based service and a given <EVI,VLAN> for VLAN-aware bundle service. The Ethernet Tag field of this route is set as described in section 5.1.3. The Originating router's IP address field is set to the NVE's IP address. This route is tagged with the PMSI Tunnel attribute, which is used to encode the type of multicast tunnel to be used as well as the multicast tunnel identifier. The tunnel encapsulation is encoded by adding the BGP Encapsulation extended community as per section 5.1.1. The following tunnel types as defined in [RFC6514] can be used in the PMSI tunnel attribute for VXLAN/NVGRE:

+ 3 - PIM-SSM Tree
+ 4 - PIM-SM Tree
+ 5 - BIDIR-PIM Tree
+ 6 - Ingress Replication

Except for Ingress Replication, this multicast tunnel is used by the PE originating the route for sending multicast traffic to other PEs, and is used by PEs that receive this route for receiving the traffic originated by hosts connected to the PE that originated the route.

In the scenario where the multicast tunnel is a tree, both the Inclusive as well as the Aggregate Inclusive variants may be used. In the former case, a multicast tree is dedicated to a VNI. Whereas, in the latter, a multicast tree is shared among multiple VNIs. This is done by having the NVEs advertise multiple Inclusive Multicast routes with different VNI encoded in the Ethernet Tag field, but with the same tunnel identifier encoded in the PMSI Tunnel attribute.

10 Data Center Interconnections - DCI

For DCI, the following two main scenarios are considered when connecting data centers running evpn-overlay (as described here) over MPLS/IP core network:

- Scenario 1: DCI using GWs
- Scenario 2: DCI using ASBRs

The following two subsections describe the operations for each of these scenarios.

10.1 DCI using GWs

This is the typical scenario for interconnecting data centers over WAN. In this scenario, EVPN routes are terminated and processed in each GW and MAC/IP routes are always re-advertised from DC to WAN but from WAN to DC, they are not re-advertised if unknown MAC address (and default IP address) are utilized in NVEs. In this scenario, each GW maintains a MAC-VRF (and/or IP-VRF) for each EVI. The main advantage of this approach is that NVEs do not need to maintain MAC and IP addresses from any remote data centers when default IP route and unknown MAC routes are used - i.e., they only need to maintain routes that are local to their own DC. When default IP route and unknown MAC route are used, any unknown IP and MAC packets from NVEs are forwarded to the GWs where all the VPN MAC and IP routes are maintained. This approach reduces the size of MAC-VRF and IP-VRF significantly at NVEs. Furthermore, it results in a faster convergence time upon a link or NVE failure in a multi-homed network or device redundancy scenario, because the failure related BGP routes (such as mass withdraw message) do not need to get propagated all the way to the remote NVEs in the remote DCs. This approach is described in details in section 3.4 of [DCI-EVPN-OVERLAY].

10.2 DCI using ASBRs

This approach can be considered as the opposite of the first approach and it favors simplification at DCI devices over NVEs such that larger MAC-VRF (and IP-VRF) tables are need to be maintained on NVEs; whereas, DCI devices don't need to maintain any MAC (and IP) forwarding tables. Furthermore, DCI devices do not need to terminate and processed routes related to multi-homing but rather to relay these messages for the establishment of an end-to-end LSP path. In other words, DCI devices in this approach operate similar to ASBRs for inter-AS options B. This requires locally assigned VNIs to be used just like downstream assigned MPLS VPN label where for all practical purposes the VNIs function like 24-bit VPN labels. This approach is equally applicable to data centers (or Carrier Ethernet networks) with MPLS encapsulation.

In inter-AS option B, when ASBR receives an EVPN route from its DC over iBGP and re-advertises it to other ASBRs, it re-advertises the EVPN route by re-writing the BGP next-hops to itself, thus losing the identity of the PE that originated the advertisement. This re-write of BGP next-hop impacts the EVPN Mass Withdraw route (Ethernet A-D per ES) and its procedure adversely. However, it does not impact EVPN Aliasing mechanism/procedure because when the Aliasing routes (Ether

A-D per EVI) are advertised, the receiving PE first resolves a MAC address for a given EVI into its corresponding <ES,EVI> and subsequently, it resolves the <ES,EVI> into multiple paths (and their associated next hops) via which the <ES,EVI> is reachable. Since Aliasing and MAC routes are both advertised per EVI basis and they use the same RD and RT (per EVI), the receiving PE can associate them together on a per BGP path basis (e.g., per originating PE) and thus perform recursive route resolution - e.g., a MAC is reachable via an <ES,EVI> which in turn, is reachable via a set of BGP paths, thus the MAC is reachable via the set of BGP paths. Since on a per EVI basis, the association of MAC routes and the corresponding Aliasing route is fixed and determined by the same RD and RT, there is no ambiguity when the BGP next hop for these routes is re-written as these routes pass through ASBRs - i.e., the receiving PE may receive multiple Aliasing routes for the same EVI from a single next hop (a single ASBR), and it can still create multiple paths toward that <ES, EVI>.

However, when the BGP next hop address corresponding to the originating PE is re-written, the association between the Mass Withdraw route (Ether A-D per ES) and its corresponding MAC routes cannot be made based on their RDs and RTs because the RD for Mass Withdraw route is different than the one for the MAC routes. Therefore, the functionality needed at the ASBRs and the receiving PEs depends on whether the Mass Withdraw route is originated and whether there is a need to handle route resolution ambiguity for this route. The following two subsections describe the functionality needed by the ASBRs and the receiving PEs depending on whether the NVEs reside in a Hypervisors or in TORs.

10.2.1 ASBR Functionality with NVEs in Hypervisors

When NVEs reside in hypervisors as described in section 7.1, there is no multi-homing and thus there is no need for the originating NVE to send Ethernet A-D per ES or Ethernet A-D per EVI routes. However, as noted in section 7, in order to enable a single-homing ingress NVE to take advantage of fast convergence, aliasing, and backup-path when interacting with multi-homing egress NVEs attached to a given Ethernet segment, the single-homing NVE SHOULD be able to receive and process Ethernet AD per ES and Ethernet AD per EVI routes. The handling of these routes are described in the next section.

10.2.2 ASBR Functionality with NVEs in TORs

When NVEs reside in TORs and operate in multi-homing redundancy mode, then as described in section 8, there is a need for the originating NVE to send Ethernet A-D per ES route(s) (used for mass withdraw) and Ethernet A-D per EVI routes (used for aliasing). As described above, the re-write of BGP next-hop by ASBRs creates ambiguities when

Ethernet A-D per ES routes are received by the remote NVE in a different ASBR because the receiving NVE cannot associated that route with the MAC/IP routes of that Ethernet Segment advertised by the same originating NVE. This ambiguity inhibits the function of mass-withdraw per ES by the receiving NVE in a different AS.

As an example consider a scenario where CE is multi-homed to PE1 and PE2 where these PEs are connected via ASBR1 and then ASBR2 to the remote PE3. Furthermore, consider that PE1 receives M1 from CE1 but not PE2. Therefore, PE1 advertises Eth A-D per ES1, Eth A-D per EVI1, and M1; whereas, PE2 only advertises Eth A-D per ES1 and Eth A-D per EVI1. ASBR1 receives all these five advertisements and passes them to ASBR2 (with itself as the BGP next hop). ASBR2, in turn, passes them to the remote PE3 with itself as the BGP next hop. PE3 receives these five routes where all of them have the same BGP next-hop (i.e., ASBR2). Furthermore, the two Ether A-D per ES routes received by PE3 have the same info - i.e., same ESI and the same BGP next hop. Although both of these routes are maintained by the BGP process in PE3 (because they have different RDs and thus treated as different BGP routes), information from only one of them is used in the L2 routing table (L2 RIB).

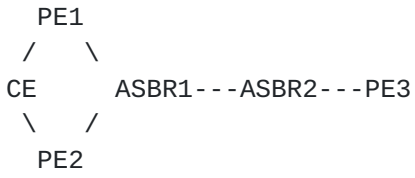


Figure 1: Inter-AS Option B

Now, when the AC between the PE2 and the CE fails and PE2 sends NLRI withdrawal for Ether A-D per ES route and this withdrawal gets propagated and received by the PE3, the BGP process in PE3 removes the corresponding BGP route; however, it doesn't remove the associated info (namely ESI and BGP next hop) from the L2 routing table (L2 RIB) because it still has the other Ether A-D per ES route (originated from PE1) with the same info. That is why the mass-withdraw mechanism does not work when doing DCI with inter-AS option B. However, as described previoulsy, the aliasing function works and so does "mass-withdraw per EVI" (which is associated with withdrawing the EVPN route associated with Aliasing - i.e., Ether A-D per EVI route).

In the above example, the PE3 receives two Aliasing routes with the same BGP next hop (ASBR2) but different RDs. One of the Alias route

has the same RD as the advertised MAC route (M1). PE3 follows the route resolution procedure specified in [RFC7432] upon receiving the two Aliasing route - ie, it resolves M1 to <ES, EVI1> and subsequently it resolves <ES,EVI1> to a BGP path list with two paths along with the corresponding VNIs/MPLS labels (one associated with PE1 and the other associated with PE2). It should be noted that even though both paths are advertised by the same BGP next hop (ASRB2), the receiving PE3 can handle them properly. Therefore, M1 is reachable via two paths. This creates two end-to-end LSPs from PE3 to PE1 for M1 such that when PE3 wants to forward traffic destined to M1, it can load balanced between the two paths. Although route resolution for Aliasing routes with the same BGP next hop is not explicitly mentioned in [RFC7432], the is the expected operation and thus it is elaborated here.

When the AC between the PE2 and the CE fails and PE2 sends NLRI withdrawal for Ether A-D per EVI routes and these withdrawals get propagated and received by the PE3, the PE3 removes the Aliasing route and updates the path list - ie, it removes the path corresponding to the PE2. Therefore, all the corresponding MAC routes for that <ES,EVI> that point to that path list will now have the updated path list with a single path associated with PE1. This action can be considered as the mass-withdraw at the per-EVI level. The mass-withdraw at per-EVI level has longer convergence time than the mass-withdraw at per-ES level; however, it is much faster than the convergence time when the withdraw is done on a per-MAC basis.

In summary, it can be seen that aliasing (and backup path) functionality should work as is for inter-AS option B without requiring any addition functionality in ASBRs or PEs. However, the mass-withdraw functionality falls back from per-ES mode to per-EVI mode for inter-AS option B - i.e., PEs receiving mass-withdraw route from the same AS use Ether A-D per ES route; whereas, PEs receiving mass-withdraw route from different AS use Ether A-D per EVI route.

11 Acknowledgement

The authors would like to thank Aldrin Isaac, David Smith, John Mullooly, Thomas Nadeau for their valuable comments and feedback. The authors would also like to thank Jakob Heitz for his contribution on section 10.2.

12 Security Considerations

This document uses IP-based tunnel technologies to support data

plane transport. Consequently, the security considerations of those tunnel technologies apply. This document defines support for VXLAN and NVGRE encapsulations. The security considerations from those documents as well as [RFC4301] apply to the data plane aspects of this document.

As with [RFC5512], any modification of the information that is used to form encapsulation headers, to choose a tunnel type, or to choose a particular tunnel for a particular payload type may lead to user data packets getting misrouted, misdelivered, and/or dropped.

More broadly, the security considerations for the transport of IP reachability information using BGP are discussed in [RFC4271] and [RFC4272], and are equally applicable for the extensions described in this document.

If the integrity of the BGP session is not itself protected, then an imposter could mount a denial-of-service attack by establishing numerous BGP sessions and forcing an IPsec SA to be created for each one. However, as such an imposter could wreak havoc on the entire routing system, this particular sort of attack is probably not of any special importance.

It should be noted that a BGP session may itself be transported over an IPsec tunnel. Such IPsec tunnels can provide additional security to a BGP session. The management of such IPsec tunnels is outside the scope of this document.

13 IANA Considerations

IANA has allocated the following BGP Tunnel Encapsulation Attribute Tunnel Types:

8	VXLAN Encapsulation
9	NVGRE Encapsulation
10	MPLS Encapsulation
11	MPLS in GRE Encapsulation
12	VXLAN GPE Encapsulation

14 References

14.1 Normative References

[KEYWORDS] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC4271] Y. Rekhter, Ed., T. Li, Ed., S. Hares, Ed., "A Border

Gateway Protocol 4 (BGP-4)", January 2006.

[RFC4272] S. Murphy, "BGP Security Vulnerabilities Analysis.", January 2006.

[RFC4301] S. Kent, K. Seo., "Security Architecture for the Internet Protocol.", December 2005.

[RFC5512] Mohapatra, P. and E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", RFC 5512, April 2009.

[RFC7432] Sajassi et al., "BGP MPLS Based Ethernet VPN", RFC 7432, February 2014

14.2 Informative References

[RFC7209] Sajassi et al., "Requirements for Ethernet VPN (EVPN)", RFC 7209, May 2014

[RFC7348] Mahalingam, M., et al, "VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", RFC 7348, August 2014

[NVGRE] Garg, P., et al., "NVGRE: Network Virtualization using Generic Routing Encapsulation", draft-sridharan-virtualization-nvgre-07.txt, November 11, 2014

[Problem-Statement] Narten et al., "Problem Statement: Overlays for Network Virtualization", draft-ietf-nvo3-overlay-problem-statement-01, September 2012.

[L3VPN-ENDSYSTEMS] Marques et al., "BGP-signaled End-system IP/VPNs", draft-ietf-l3vpn-end-system, work in progress, October 2012.

[NOV3-FRWK] Lasserre et al., "Framework for DC Network Virtualization", draft-ietf-nvo3-framework-01.txt, work in progress, October 2012.

[DCI-EVPN-OVERLAY] Rabadan et al., "Interconnect Solution for EVPN Overlay networks", draft-ietf-bess-dci-evpn-overlay-02, work in progress, February 29, 2016.

[TUNNEL-ENCAP] Rosen et al., "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-02, work in progress, May 31, 2016.

Contributors

S. Salam K. Patel D. Rao S. Thoria D. Cai Cisco

Y. Rekhter R. Shekhar Wen Lin Nischal Sheth Juniper

L. Yong Huawei

Authors' Addresses

Ali Sajassi
Cisco
Email: sajassi@cisco.com

John Drake
Juniper Networks
Email: jdrake@juniper.net

Nabil Bitar
Nokia
Email : nabil.bitar@nokia.com

R. Shekhar
Juniper
Email: rshekhar@juniper.net

James Uttaro
AT&T
Email: uttaro@att.com

Wim Henderickx
Alcatel-Lucent
e-mail: wim.henderickx@nokia.com