L2VPN Workgroup
INTERNET-DRAFT
Intended Status: Standards Track

A. Sajassi (Editor)
Cisco

J. Drake (Editor)
Juniper

Y. Rekhter
R. Shekhar
B. Schliesser
Juniper

Nabil Bitar
Verizon

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

Aldrin Isaac
Bloomberg

James Uttaro
AT&T

L. Yong
Huawei

W. Henderickx
Alcatel-Lucent

Expires: April 21, 2014

October 21, 2013

A Network Virtualization Overlay Solution using EVPN
draft-sd-l2vpn-evpn-overlay-02

Abstract

This document describes how EVPN can be used as an 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: MPLS over GRE, VXLAN, and NVGRE.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

Copyright and License Notice

Table of Contents

1 Introduction 4
1.1 Terminology 5
2 EVPN Features 5
3 Encapsulation Options for EVPN Overlays 6
3.1 VXLAN/NVGRE Encapsulation 6
3.1.1 Virtual Identifiers Scope 7
3.1.1.1 Data Center Interconnect with Gateway 7
3.1.1.2 Data Center Interconnect without Gateway 8
3.1.2 Virtual Identifiers to EVI Mapping 8
3.1.2.1 Auto Derivation of RT & RD 9
3.1.3 Constructing EVPN BGP Routes 10
3.1.3.1 Constructing E-VPN MAC Address Advertisement Route 11
3.2 MPLS over GRE 11
4 EVPN with Multiple Data Plane Encapsulations 12
5 NVE Residing in Hypervisor 12
5.1 Impact on EVPN BGP Routes & Attributes for VXLAN/NVGRE Encapsulation 12
5.2 Impact on EVPN Procedures for VXLAN/NVGRE Encapsulation . . 13
6 NVE Residing in ToR Switch 14
6.1 EVPN Multi-Homing Features 14
6.1.1 Multi-homed Ethernet Segment Auto-Discovery 14

6.1.2 Fast Convergence and Mass Withdraw 14
6.1.3 Split-Horizon 15
6.1.4 Aliasing and Backup-Path 15
6.1.5 DF Election 16
6.2 Impact on EVPN BGP Routes & Attributes 16
6.3 Impact on EVPN Procedures 16
6.3.1 Split Horizon 17
6.3.2 Aliasing and Backup-Path 18
7 Support for Multicast 18
8 Support for NVEs with data plane MAC learning 19
8.1 Advertising NVE capabilities 20
8.2 Advertising flood lists for ingress replication 20
9 Inter-AS 21
10 Acknowledgement 22
11 Security Considerations 22
12 IANA Considerations 22
13 References 22
11.1 Normative References 22
11.2 Informative References 23
Authors' Addresses 23

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 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 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.

1.1 Terminology

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 RFC 2119 [KEYWORDS].

NVE: Network Virtualization Endpoint

Virtual Identifier: refers to a VXLAN VNI or NVGRE VSID

2 EVPN Features

EVPN was originally designed to support the requirements detailed in [EVPN-REQ] 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 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 multi-homing 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) Mass withdraw is used. When a link between a CE and a PE fails, the PEs in all EVPNs associated with that failed link are notified via the withdrawal of a single EVPN route regardless of how many MAC addresses are located at the CE.

7) Route filtering and constrained route distribution are used to

ensure that the control plane traffic for a given EVPN is only distributed to the PEs in that EVPN.

8) The internal identifier of a broadcast domain, the Ethernet Tag, is a 32 bit number, which is mapped into whatever broadcast domain identifier, e.g., VLAN ID, is understood by the attaching CE device. This means that when 802.1q interfaces are used, there are up to 4096 distinct VLAN IDs for each attaching CE device in a given EVPN.

9) VM Mobility mechanisms ensure that all PEs in a given EVPN 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.

3 Encapsulation Options for EVPN Overlays

3.1 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 infrastructure between NVEs, VXLAN Tunnel End Point (VTEPs) in VXLAN and Network Virtualization Endpoint (NVEs) in NVGRE. Both of these technologies include the identifier of the specific NVO instance, Virtual Network Identifier (VNI) in VXLAN and Virtual Subnet Identifier (VSID), NVGRE, in each packet.

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

[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 [VXLAN]. In this scenario, the VTEP does not include an inner VLAN tag on frame encapsulation, and discards decapsulated frames with an inner VLAN tag. This mode of operation in [VXLAN] maps to VLAN Based Service in [EVPN], 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 maps to VLAN Bundle Service in [EVPN], where the VLANs of a given tenant get mapped to an EVI.

[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 [EVPN]. In other words, [NVGRE] prohibits the application of VLAN Bundle Service in [EVPN] and it only requires VLAN Based Service in [EVPN].

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. 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.

3.1.1 Virtual Identifiers Scope

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

3.1.1.1 Data Center Interconnect with Gateway

In the case where NVEs in different data centers need to be interconnected, and a Gateway is employed at the edge of the data center network, the NVEs should treat the VNI or VSID as a globally unique identifier within a data center. This is because the Gateway will provide the functionality of translating the VNI or VSID 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 / VSIDs between the values used in each of the data center networks and the values used in the WAN.

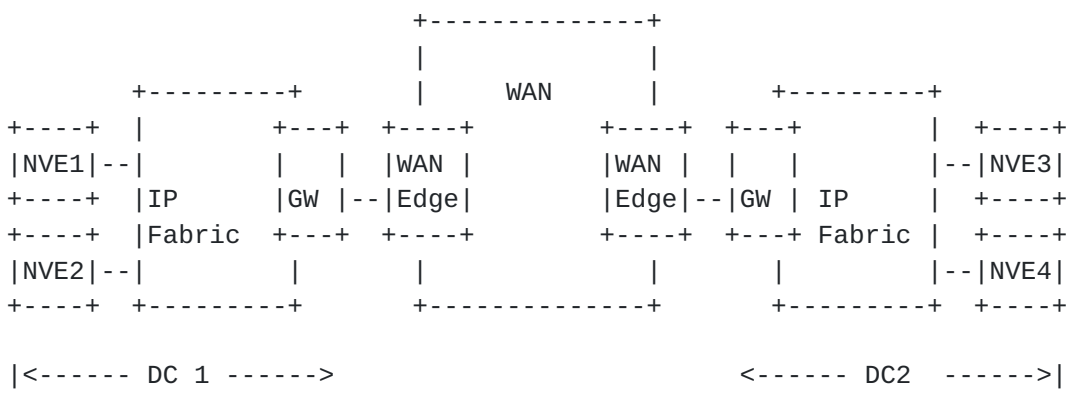


Figure 1: Data Center Interconnect with Gateway

3.1.1.2 Data Center Interconnect without Gateway

In the case where NVEs in different data centers need to be interconnected, and Gateways are not employed at the edge of the data center network, it is useful to treat the VNIs or VSIDs as locally significant identifiers (e.g., as an MPLS label). More specifically, the VNI or VSID value that is used by the transmitting NVE is allocated by the NVE that is receiving the traffic (in other words, this is a "downstream assigned" MPLS label). This allows the VNI or VSID space to be decoupled between different data center networks without the need for a dedicated Gateway at the edge of the data centers.

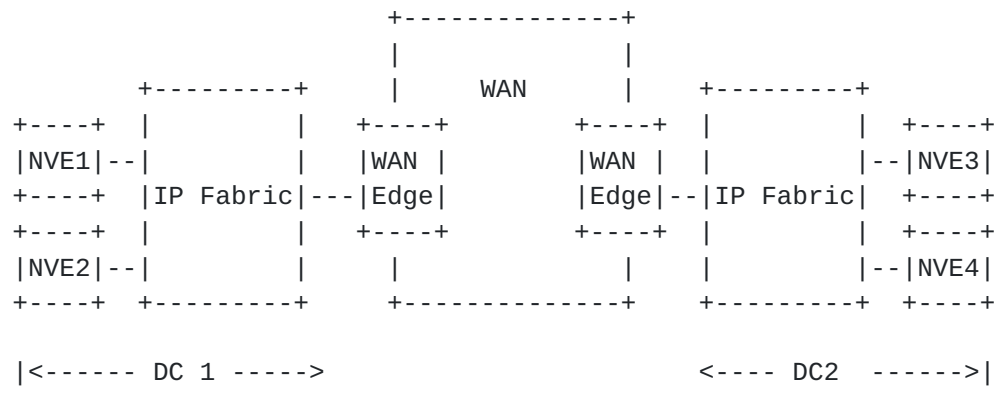


Figure 2: Data Center Interconnect without Gateway

3.1.2 Virtual Identifiers to EVI Mapping

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

1. Option 1: Single Virtual Identifier per EVI

In this option, every VNI or VSID is mapped to a unique EVI. As such, a BGP RD and RT is needed per VNI / VSID on every VTEP. 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 VTEPs that are interested in a given VNI (or VSID). The disadvantage of this model may be the provisioning overhead if RD and RT are not derived automatically from VNI (for VSID).

In this option, the MAC-VRF table is identified by the RT in the control plane (because Ethernet Tag field in the MAC route is set to zero) and by the VNI (or VSID) in the data-plane.

2. Option 2: Multiple Virtual Identifiers per EVI

In this option, multiple VNIs or VSIDs are mapped to a unique EVI. For example, if a tenant has multiple segments/subnets each represented by a VNI (or VSID), then all the VNIs (or VSIDs) for that tenant are mapped to a single EVI - e.g., the EVI in this case represents the tenant and not a subnet . The advantage of this model is that it doesn't require the provisioning of RD/RT per VNI or VSID which is a moot point if auto-derivation is used. The disadvantage of this model is that routes would be imported by VTEPs that may not be interested in a given VNI (or VSID).

In this option the MAC-VRF table is identified by the VNI (or VSID) in both the control plane and the data-plane.

3.1.2.1 Auto Derivation of RT & RD

When the option of a single VNI (or VSID) per EVI is used, it is important to auto-derive RD and RT for EVPN BGP routes in order to simplify configuration for data center operations. RD can be auto-derive as described in [EVPN] 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 (or VSIDs), 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 and VSIDs are administered independently which means VNI and VSID 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 or 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 remaining 7 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 : EVI
  1 : VXLAN
  2 : NVGRE
  3 : I-SID
  4 : VID

3.1.3 Constructing EVPN BGP Routes

In EVPN, an MPLS label distributed by the egress PE via the EVPN control plane and 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 to disposition that packet. This is very similar to the use of the VNI or VSID by the egress VTEP or NVE, respectively, with the difference being that an MPLS label has local significance and is distributed by the EVPN control plane, while a VNI or VSID typically has global significance.

As discussed in Section 3.1.1 above, there are scenarios in which it is desirable to use a locally significant value for VNI or VSID and in such such scenarios, MPLS label is advertised in EVPN BGP routes and it is used in VXLAN or NVGRE encapsulation as a 20-bit value for VNI or VSID.

This memo specifies that when EVPN is used with a VXLAN or NVGRE data plane and when a globally significant VNI or VSID is desirable, then for VNI-based mode (single VNI per EVI), the Ethernet Tag field of EVPN BGP routes (which is a 4-octet field) MUST be set to zero just like the VLAN-based mode in baseline EVPN. If VNI-aware bundle mode (multiple VNIs per EVI) is desired, then the Ethernet Tag field of EVPN BGP routes MUST be set to VNI (or VSID) accordingly just like VLAN-aware bundle mode in baseline EVPN. In both cases, the MPLS label field of the EVPN BGP routes MUST be set to zero.

This memo also specifies that when EVPN is used with a VXLAN or NVGRE data plane and when a locally significant VNI or VSID is desirable, then MPLS field of EVPN BGP routes (which is a 3-octet field) MUST be

used and Ethernet Tag field MUST be set to zero. In such scenarios, only VNI-based mode (single VNI per EVI) is supported.

In order to indicate that a VXLAN or NVGRE data plane encapsulation rather than MPLS label stack encapsulation is to be used, the BGP Encapsulation extended community defined in [RFC5512] is included with EVPN MAC route, Inclusive Multicast route, or per EVI Ethernet AD route advertised by an egress PE. Two new values, one for VXLAN and one for NVGRE, will be defined to extend the list of encapsulation types defined in [RFC5512]:

+ 3 - VXLAN Encapsulation
+ 4 - NVGRE Encapsulation

If BGP Encapsulation extended community is not present, then the default encapsulation MPLS encapsulation (or statically configured encapsulation) is used.

3.1.3.1 Constructing E-VPN MAC Address Advertisement Route

In EVPN, unicast MAC addresses are advertised via MAC Advertisement route. The Ethernet Tag field in this route is set zero for the VNI-based mode and set to VNI (or VSID) for VNI-aware bundle mode. The MPLS label field is set to zero. The encapsulation is set via the BGP Encapsulation extended community as described in section 3.1.3.

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 the route are set as per EVPN.

3.2 MPLS over GRE

The EVPN data-plane is modeled as an EVPN MPLS client layer sitting over an MPLS PSN tunnel. Some of the EVPN functions (split-horizon, aliasing and repair-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 key field MUST 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.

4 EVPN with Multiple Data Plane Encapsulations

The use of the BGP Encapsulation extended community allows each PE in a given EVPN to know whether the other PEs in that EVPN support MPLS label stack, VXLAN, and/or NVGRE data plane encapsulations. I.e., PEs in a given EVPN may support multiple data plane encapsulations.

If BGP Encapsulation extended community is not present, then the default MPLS encapsulation (or statically configured encapsulation) is used. However, if this attribute is present, then an ingress PE can send a frame to an egress PE only if the set of encapsulations advertised by the egress PE in the subject MAC Advertisement or Per EVI Ethernet AD route, forms a non-empty intersection with the set of encapsulations supported by the ingress PE, and it is at the discretion of the ingress PE which encapsulation to choose from this intersection.

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 EVPN to ensure that all of the PEs in that EVPN 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.

5 NVE Residing in Hypervisor

When a PE and its CEs are co-located in the same physical device, e.g., when the PE resides in a server and the CEs are its VMs, the links between them are virtual and they typically share fate; i.e., the subject CEs 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 need not be "visible" to any other PEs, and thus does not require any specific protocol mechanisms. The most common case of this is when the NVE resides in 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.

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

As discussed above, both [NVGRE] and [VXLAN] do not require the tenant VLAN tag to be sent in BGP routes. Therefore, the 4-octet

Ethernet Tag field in the EVPN BGP routes can be used to represent the globally significant value for VXLAN VNI or NVGRE VSID and MPLS field can be used to represent the locally significant value for VNI or VSID.

When the VXLAN VNI or NVGRE VSID is assumed to be a global value, one might question the need for the Route Distinguisher (RD) in the EVPN routes. In the scenario where all data centers are under a single administrative domain, and there is a single global VNI/VSID 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 [EVPN]. In other words, whenever there is more than one administrative domain for global VNI or VSID, then a non-zero RD MUST be used, or whenever the VNI or VSID 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 five out of the set of eight :

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

As mentioned in section 3.1.1, BGP Encapsulation extended community as defined in [RFC5512] SHOULD be used along with MAC Advertisement Route or Ethernet AD Route to indicate the supported encapsulations.

5.2 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 [EVPN].

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

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

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 [EVPN].

6 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 Active/Standby redundancy mode. If the servers are single-homed to the ToR switches, then the scenario becomes similar to that where the NVE resides in the hypervisor, as discussed in Section 5, as far as the required EVPN functionality.

[EVPN] 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 encapsulation (such as VXLAN or NVGRE) and what modifications are required.

6.1 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 [EVPN].

6.1.1 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.

6.1.2 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 adjacencies to point to the backup NVE(s).

6.1.3 Split-Horizon

Consider a station that is multi-homed to two or more NVEs on an Ethernet segment ES1, with all-active redundancy. If the station sends a multicast, broadcast or unknown unicast packet to a particular NVE, say NE1, then NE1 will forward that packet to all or subset of the other NVEs in the EVPN instance. In this case the NVEs, other than NE1, that the station is multi-homed to MUST drop the packet and not forward back to the station. This is referred to as "split horizon" filtering.

6.1.4 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 Active-Standby flag reset.

Backup-Path is a closely related function, albeit it applies to the case where the redundancy mode is Active/Standby. 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 Active-Standby flag set.

6.1.5 DF Election

Consider a station that is a host or a VM that is multi-homed directly to more than one NVE in an EVPN on a given Ethernet segment. One or more Ethernet Tags may be configured on the Ethernet segment. In this scenario only one of the PEs, referred to as the Designated Forwarder (DF), is responsible for certain actions:

- Sending multicast and broadcast traffic, on a given Ethernet Tag on a particular Ethernet segment, to the station.

- Flooding unknown unicast traffic (i.e. traffic for which an NVE does not know the destination MAC address), on a given Ethernet Tag on a particular Ethernet segment to the station, if the environment requires flooding of unknown unicast traffic.

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.

6.2 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 [EVPN] are used. As discussed in Section 3.1, the VSID or VNI is encoded in the Ethernet Tag field of the routes if globally significant or in the MPLS label field if locally significant.

As mentioned in section 3.1.1, BGP Encapsulation extended community as defined in [RFC5512] SHOULD be used along with MAC Advertisement Route or Ethernet AD Route to indicate the supported encapsulations.

6.3 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, let's 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 or VSID. In this case, the Split-Horizon and Aliasing functions are not required

but other functions such as multi-homed Ethernet segment auto-discovery, fast convergence and mass withdraw, repair path, and DF election are required. In this case, the impact of the use of the VXLAN/NVGRE encapsulation on the EVPN procedures is when the Backup-Path function is supported, as discussed next:

In EVPN, the NVEs connected to a multi-homed site using Active/Standby redundancy optionally advertise a VPN label, in the Ethernet A-D Route per EVI, used to send traffic to the backup NVE in the case where the primary NVE fails. In the case where VXLAN or NVGRE encapsulation is used, some alternative means that does not rely on MPLS labels is required to support Backup-Path. This is discussed in Section 4.3.2 below. If the Backup-Path function is not used, then the VXLAN/NVGRE encapsulation would have no impact on the EVPN procedures.

Second, let's consider the case of All-Active redundancy. In this case, out of the EVPN multi-homing features listed in section 4.1, 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.

6.3.1 Split Horizon

In EVPN, an MPLS label is used for split-horizon filtering to support active/active multi-homing where an ingress ToR switch adds a label corresponding to the site of origin (aka ESI MPLS Label) when encapsulating the packet. The egress ToR switch checks the ESI MPLS 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 MPLS 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 [EVPN] and the local bias procedures described in this memo, 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. 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.

6.3.2 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 AD routes are used for this purpose. The one used for Aliasing has a VPN scope and carries a VPN label but the one used for Backup-Path has Ethernet segment scope and doesn't carry any VPN specific info (e.g., Ethernet Tag and MPLS label are set to zero). The same two routes are used when VXLAN or NVGRE encapsulation is used with the difference that when Ethernet AD route is used for Aliasing with VPN scope, the Ethernet Tag field is set to VNI or VSID to indicate VPN scope (and MPLS field may be set to a VPN label if needed).

7 Support for Multicast

The E-VPN Inclusive Multicast BGP route is used to discover the multicast tunnels among the endpoints associated with a given VXLAN VNI or NVGRE VSID. The Ethernet Tag field of this route is used to encode the VNI for VLXAN or VSID for NVGRE. 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 3.1.1. 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 CEs 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 or VSID. Whereas, in the latter, a multicast tree is shared among multiple VNIs or VSIDs. This is done by having the NVEs advertise multiple Inclusive Multicast routes with different VNI or VSID encoded in the Ethernet Tag field, but with the same tunnel identifier encoded in the PMSI Tunnel attribute.

8 Support for NVEs with data plane MAC learning

In an overlay network it possible to have a mix of NVEs, such that only a subset of the NVEs are capable of participating in control plane MAC learning via EVPN. The other subset of NVEs would perform conversational MAC learning in data plane. It must be possible for NVEs with this mixed capability to still be part of the same overlay network.

If the administrative policy of an EVPN NVE requires for flooding of unknown unicast, then the following procedures are not needed; however, if the administrative policy of the EVPN NVE requires no flooding of unknown unicast, then for such a mixed overlay network to operate correctly, the following requirements MUST be met:

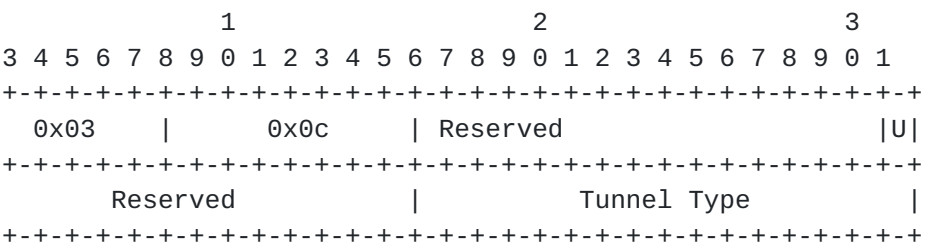
- When an NVE capable of doing control-plane MAC learning via EVPN wants to send an unknown unicast frame, it MUST send it to a subset of NVEs in the VNI that only have data plane MAC learning capability. This can be achieved by creating a flood list for each VNI to carry unknown unicast traffic, which only the subset of NVEs with data plane MAC learning are part of. Section 8.2 describes the procedure to accomplish this.

- Broadcast traffic MUST be sent to all NVEs in the VNI regardless of the MAC learning capability. A separate flood list for each VNI to carry broadcast traffic can be created for this, and all NVEs in the VNI would be part of this flood list. Section 8.2 describes the procedure to accomplish this.

- When an NVE capable of only data plane MAC learning wants to send an unknown unicast frame, it MUST send it to all NVEs in the VNI. This can be achieved by flooding the unknown unicast frame in the broadcast flood list (as described earlier).

8.1 Advertising NVE capabilities

BGP Encapsulation extended community is used to signal NVE capabilities. NVE capabilities are used to build different types of flood lists in the broadcast domain for optimal forwarding in the case of NVEs with mixed capabilities, as described in section 8. The reserved field of the BGP Encapsulation extended community is repurposed to indicate NVE capabilities as following:



U bit indicates that the NVE must be included in unknown unicast flood list

The Reserved fields must be set to zero and ignored on receipt.

8.2 Advertising flood lists for ingress replication

Flooding of unknown unicast, broadcast and multicast can either be achieved by using multicast trees in the underlay or using ingress replication. If IP multicast is used for flooding, separate flood lists, as described in section 8, can be created by using separate IP multicast groups for different flood lists. If ingress replication is used for flooding, then the EVPN capable NVEs must maintain separate flood lists depending on advertised NVE capability. Either way, there is a need to signal which NVEs are part of which flood lists. This section describes enhancements to BGP signaling required to achieve this.

The EVPN Inclusive Multicast Route along with NVE capabilities as described in section 8.1 can be used to build different flood lists. The Inclusive Multicast Route is encoded as follows: The Ethernet Tag field is set to the VNI for VXLAN and VSID for NVGRE. The Originator's IP address field is set to the NVE's IP address. The Next Hop field of the MP_REACH_NLRI attribute of the route is set to NVE's IP address. The Inclusive Multicast route is tagged with the PMSI tunnel attribute. The BGP Encapsulation extended community is included with U, B or K bit set as described in section 8.1 to enable an NVE to be part of a specific flood list depending on its

capabilities.

9 Inter-AS

For inter-AS operation, two scenarios must be considered:

- Scenario 1: The tunnel endpoint IP addresses are public
- Scenario 2: The tunnel endpoint IP addresses are private

In the first scenario, inter-AS operation is straight-forward and follows existing BGP inter-AS procedures. However, in the first scenario where the tunnel endpoint IP addresses are public, there may be security concern regarding the distribution of these addresses among different ASes. This security concern is one of the main reasons for having the so called inter-AS "option-B" in MPLS VPN solutions such as EVPN.

The second scenario is more challenging, because the absence of the MPLS client layer from the VXLAN encapsulation creates a situation where the ASBR has no fully qualified indication within the tunnel header as to where the tunnel endpoint resides. To elaborate on this, recall that with MPLS, the client layer labels (i.e. the VPN labels) are downstream assigned. As such, this label implicitly has a connotation of the tunnel endpoint, and it is sufficient for the ASBR to look up the client layer label in order to identify the label translation required as well as the tunnel endpoint to which a given packet is being destined. With the VXLAN encapsulation, the VNI is globally assigned and hence is shared among all endpoints. The destination IP address is the only field which identifies the tunnel endpoint in the tunnel header, and this address is privately managed by every data center network. Since the tunnel address is allocated out of a private address pool, then we either need to do a lookup based on VTEP IP address in context of a VRF (e.g., use IP-VPN) or terminate the VXLAN tunnel and do a lookup based on the tenant's MAC address to identify the egress tunnel on the ASBR. This effectively mandates that the ASBR to either run another overlay solution such as IP-VPN over MPLS/IP core network or to be aware of the MAC addresses of all VMs in its local AS, at the very least.

If VNIs/VSIDs have local significance, then the inter-AS operation can be simplified to that of MPLS and thus MPLS inter-AS option B and C can be leveraged in here. That's why the use of local significance VNIs/VSIDs (e.g., MPLS labels) are recommended for inter-AS operation of DC networks without gateways.

10 Acknowledgement

The authors would like to thank David Smith, John Mullooly, Thomas Nadeau for their valuable comments and feedback.

11 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]. 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.

12 IANA Considerations

13 References

11.1 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.

11.2 Informative References

[EVPN-REQ] Sajassi et al., "Requirements for Ethernet VPN (EVPN)", draft-ietf-l2vpn-evpn-req-01.txt, work in progress, October 21, 2012.

[NVGRE] Sridhavan, M., et al., "NVGRE: Network Virtualization using Generic Routing Encapsulation", draft-sridharan-virtualization-nvgre-01.txt, July 8, 2012.

[VXLAN] Dutt, D., et al, "VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", draft-mahalingam-dutt-dcops-vxlan-02.txt, August 22, 2012.

[EVPN] Sajassi et al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-02.txt, work in progress, February, 2012.

[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.

Authors' Addresses

Ali Sajassi
Cisco
Email: sajassi@cisco.com

John Drake
Juniper Networks
Email: jdrake@juniper.net

Nabil Bitar
Verizon Communications
Email : nabil.n.bitar@verizon.com

Aldrin Isaac
Bloomberg
Email: aisaac71@bloomberg.net

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

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

Ravi Shekhar
Juniper Networks
Email: rshekhar@juniper.net

Samer Salam
Cisco
Email: ssalam@cisco.com

Keyur Patel
Cisco
Email: Keyupate@cisco.com

Dhananjaya Rao
Cisco
Email: dhrao@cisco.com

Samir Thoria
Cisco
Email: sthoria@cisco.com