

NV03 Workgroup
INTERNET-DRAFT
Intended Status: Standards Track

Ali Sajassi
Samer Salam
Keyur Patel
Cisco

Nabil Bitar
Verizon

Wim Henderickx
Alcatel-Lucent

Expires: April 22, 2013

October 22, 2012

A Network Virtualization Overlay Solution using E-VPN
draft-sajassi-nvo3-evpn-overlay-01

Abstract

This document describes how E-VPN can be used as an NVO solution and explores the various tunnel encapsulation options and their impact on the E-VPN control-plane and procedures. In particular, the following three 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."

The list of current Internet-Drafts can be accessed at <http://www.ietf.org/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at <http://www.ietf.org/shadow.html>

INTERNET DRAFT

E-VPN Overlay

October 22, 2012

Copyright and License Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to [BCP 78](http://trustee.ietf.org/license-info) and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1	Introduction	4
1.1	Terminology	4
2	E-VPN Main Features	5
2.1	Multi-homed Ethernet Segment Auto-Discovery	5
2.2	Fast Convergence and Mass Withdraw	5
2.3	Split-Horizon	5
2.4	Aliasing	6
2.5	DF Election	6
3	Encapsulation Options for E-VPN Overlays	7
3.1	MPLS over GRE	7
3.1.1	Benefits of MPLS over GRE	7
3.2	VXLAN/NVGRE Encapsulation	8
3.2.1	Impact on E-VPN Routes for VXLAN/NVGRE Encapsulation . .	8
3.2.2	Impact on E-VPN Procedures for VXLAN/NVGRE Encapsulation	9
3.2.2.1	NVE with No Redundancy	9
3.2.2.2	NVE with Active/Standby Redundancy	10
3.2.2.3	NVE with All-Active Redundancy	10
3.2.3	Support for Multicast	13
3.2.4	Inter-AS Challenges	13
4	Comparison between MPLSoGRE and VXLAN/NVGRE Encapsulation . . .	14
5	Acknowledgement	15
6	Security Considerations	15
7	IANA Considerations	15

8	References	15
8.1	Normative References	15
8.2	Informative References	15
	Authors' Addresses	16

INTERNET DRAFT

E-VPN Overlay

October 22, 2012

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 (i.e. Virtual Machines or VMs). The key requirements of such a solution as described in [[Problem-Statement](#)] are:

- Isolation of network traffic per tenant
- Support of 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

The underlay network for NVO solutions is assumed to provide IP connectivity.

This document describes how E-VPN can be used as an NVO solution and explores the various tunnel encapsulation options for E-VPN over IP, and their impact on the E-VPN control-plane and procedures. Note that the use of E-VPN as an NVO solution does not necessarily mandate that the BGP control-plane be running on the NVE. This may not be desirable, for e.g., when the NVE resides on the hypervisor. For such scenarios, it is still possible to leverage the E-VPN 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 E-VPN overlays that are analyzed in this document are:

- MPLS over GRE
- VXLAN and NVGRE

Before getting into the description of the different encapsulation options for E-VPN over IP, it is important to highlight the E-VPN solution main features, how those features are currently supported, and any impact that the encapsulation may have 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](#)].

[2](#) E-VPN Main Features

In this section, we will recap the main features of E-VPN, to highlight the encapsulation dependencies. The section only describes the features and functions at high-level. For more details, the reader is to refer to [[E-VPN](#)].

[2.1](#) Multi-homed Ethernet Segment Auto-Discovery

E-VPN NV Edge devices (NVEs) connected to the same Ethernet segment (e.g. server) can automatically discover each other with minimal to no configuration through the exchange of BGP routes.

[2.2](#) Fast Convergence and Mass Withdraw

E-VPN 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. 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).

[2.3](#) Split-Horizon

Consider a station that is multi-homed to two or more NVEs on an Ethernet segment ESI, 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 E-VPN 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. In order to achieve this split horizon function, every multicast, broadcast or unknown unicast packet is encapsulated with an MPLS label that identifies the Ethernet segment of origin (i.e. the segment from which the frame entered the E-VPN network). This label is referred to as the ESI MPLS label, and is distributed using the "Ethernet A-D route per Ethernet Segment". This route is imported by the PEs connected to the Ethernet Segment and also by the PEs that have at least one E-VPN instance in common with the Ethernet Segment in the route. The disposition PEs rely on the value of the ESI MPLS label to determine whether or not a flooded

frame is allowed to egress a specific Ethernet segment.

[2.4](#) Aliasing

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 PEs are connected to the multi-homed segment. As a result, the remote PEs 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 PEs 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 PE. Another scenario where this occurs is

when the PEs 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, E-VPN introduces the concept of 'Aliasing'. Aliasing 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 PEs which receive MAC advertisement routes with non-zero ESI SHOULD consider the advertised MAC address as reachable via all PEs which have advertised reachability to the relevant Segment using Ethernet A-D routes with the same ESI (and Ethernet Tag if applicable) and with the Active-Standby flag reset.

[2.5](#) DF Election

Consider a station that is a host or a VM that is multi-homed directly to more than one NVE in an E-VPN 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

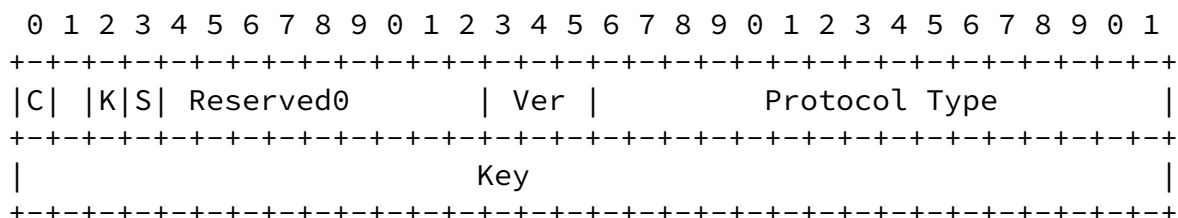
redundancy.

[3](#) Encapsulation Options for E-VPN Overlays

[3.1](#) MPLS over GRE

The E-VPN data-plane is modeled as an E-VPN MPLS client layer sitting

over an MPLS PSN tunnel. The Split-Horizon and Aliasing functions of E-VPN are tied to the MPLS client layer. In order to keep the E-VPN procedures intact and data-plane operation as is, an ideal encapsulation would allow the E-VPN MPLS client layer to be carried over an IP PSN tunnel transparently - i.e., without any changes. The existing standards-based GRE encapsulation as defined by [RFC2890] and [RFC2784] provides such a solution:



The Key field can be used to provide 32-bit entropy field.

The C (Checksum Present) and S (Sequence Number Present) bits in the GRE header are set to zero. The K bit is set to 1.

[MPLSoUDP] discusses using a UDP header instead of the GRE header to transport MPLS client layer over an IP PSN tunnel. The main advantage for doing so is for better load-balancing capabilities over existing IP networks, where some core routers can perform ECMP based on the UDP header but not based on the GRE Key field. However, the routers that are capable of supporting [NVGRE] encapsulation, can also perform load-balancing based on the GRE key which accommodates a 32-bit entropy value; whereas, UDP encapsulation accommodates a 16-bit entropy value.

3.1.1 Benefits of MPLS over GRE

The benefits of using the MPLS over GRE encapsulation are as follows:

- Uses existing standard for transporting MPLS over IP.
- Uses E-VPN control plane (BGP routes and attributes), as well as E-VPN procedures and functions exactly as is.
- Consistent with L3VPN over IP ([RFC 4797](#))
- The MPLS label can be a global value (instead of downstream

assigned) just like VXLAN or NVGRE service-instance ID.

- Provides seamless interoperability with E-VPN PEs. There is no need for a gateway device.

[3.2](#) VXLAN/NVGRE Encapsulation

If either the VXLAN or NVGRE encapsulation were to be used with the E-VPN control plane, there will be an impact on the E-VPN client layer and the associated procedures and BGP routes. In order to assess this impact, the first step is to identify which subset of the service interfaces defined in [\[E-VPN\]](#) is needed for the NVO solutions defined in [\[VXLAN\]](#) and [\[NVGRE\]](#). Then we need to examine how the E-VPN BGP routes and procedures should be modified to support these service interfaces with the new encapsulation.

[E-VPN] defines the following four service interface types:

- VLAN Based Service Interface
- VLAN Bundle Service Interface
- Port-based Service Interface
- VLAN Aware Bundle Service Interface

For a detailed description of these service interface types, refer to [EVPN-REQ] and [\[E-VPN\]](#). As described in [\[E-VPN\]](#), the first three service interface types don't require encoding the VLAN Tag in the BGP routes, because there is a one-to-one mapping between an EVI and a broadcast domain represented by a virtual network or a virtual segment.

[NVGRE] requires only VLAN-based service interface and it clearly describes that the tenant VLAN Tag (inner VLAN Tag) is not part of the encapsulated frames because there is a one-to-one mapping between Virtual Subnet Identifier (VSID) and the inner VLAN ID.

The [\[VXLAN\]](#) default mode of operation only requires VLAN-based service interface, as it specifies that the VTEP does not include an inner VLAN tag upon encapsulation; moreover, the decapsulated frames with an inner VLAN tag should get discarded. However, [\[VXLAN\]](#) provides an option of including an inner VLAN tag in the encapsulated packet if it is configured explicitly at the VTEP. If an inner VLAN tag is included, then VXLAN requires a VLAN-bundle service interface. However, as discussed above, this service interface type does not require that the tenant VLAN tag be sent in the BGP routes.

[3.2.1](#) Impact on E-VPN Routes 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 32-bit Ethernet tag field in the E-VPN BGP routes can be used to represent NVGRE VSID or VXLAN VNI. This is not accidental, but rather by design: The Ethernet Tag field in E-VPN was designed not just for C-tagged or S-tagged interfaces [802.1Q] but also for I-tagged interfaces [802.1ah] where an I-SID is a 24-bit entity representing a virtual segment just like VSID or VNI. Therefore, there is no need to re-purpose the MPLS label field in the E-VPN BGP routes and this field can be omitted in the E-VPN BGP routes. The length field of the NLRI in E-VPN routes will be three octets shorter for VXLAN and NVGRE encapsulations.

Since VXLAN VNI or NVGRE VSID is assumed to be a global value, one might question the need for the Route Distinguisher (RD) in the E-VPN 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 can be set to zero in the E-VPN routes. However, in the scenarios where different group 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 [[E-VPN](#)]. In other words, whenever, there is more than one administrative domain for VNI or VSID, then a non-zero RD MUST be used.

[3.2.2](#) Impact on E-VPN Procedures for VXLAN/NVGRE Encapsulation

In order to analyze the impact of the VXLAN/NVGRE encapsulation on E-VPN procedures, we must distinguish three NVE redundancy models:

- No redundancy
- Active/Standby redundancy
- All-active redundancy

The impact of the encapsulation varies depending on the employed model.

[3.2.2.1](#) NVE with No Redundancy

This is the scenario where, for e.g., the NVE is implemented on the hypervisor. In this case, neither the Split-Horizon nor the Aliasing functions are required or applicable. Therefore, the choice of VXLAN/NVGRE encapsulation has no impact on E-VPN procedures.

For all practical purposes, in this scenario, the only difference

INTERNET DRAFT

E-VPN Overlay

October 22, 2012

between the choice of GRE or VXLAN/NVGRE encapsulation is in the size of the entropy field (32-bits vs. 16 bits).

[3.2.2.2](#) NVE with Active/Standby Redundancy

This is the scenario 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 as well, the Split-Horizon function is not required. However, in order to support fast convergence in case where the primary NVE fails, the Aliasing function of E-VPN is needed. Note that Aliasing in this scenario is used to quickly identify the backup NVE rather than being used for traffic load-balancing. In this case, the impact of the use of the VXLAN/NVGRE encapsulation on the E-VPN procedures is as discussed in [Section 3.2.2.3.2](#), with the difference being that a remote NVE uses the received Ethernet A-D routes to build primary and backup paths to the advertising NVEs, instead of a load-balancing path-list.

If fast convergence is not required or not used, then the VXLAN/NVGRE encapsulation would have no impact on the E-VPN procedures.

[3.2.2.3](#) NVE with All-Active Redundancy

Out of the E-VPN features listed in [section 2](#), 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.

[3.2.2.3.1](#) Split Horizon

In E-VPN, an MPLS label is used for split-horizon filtering to support active/active multi-homing where an ingress NV Edge device (NVE) adds a label corresponding to the site of origin (aka ESI MPLS Label) when encapsulating the packet. The egress NVE 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. One way of supporting this function is to assign an IP address for each site of origin (e.g., for each ESI in the E-VPN terminology) and advertise this IP address in the BGP Remote-Next-Hop attribute associated with the E-VPN Ethernet A-D route (refer to [section 3.2.3](#) for details). The "Active-Standby" bit in the flags of

the ESI MPLS Label Extended Community MUST be set to 0 to indicate active/active multi-homing and the MPLS label field MUST be set to zero to indicate that IP address in the BGP Remote-Next-Hop attribute will be used for split-horizon filtering. The ingress NVE uses the IP address associated with a given site as the source IP address for all traffic originating from said site. The egress NVE will program its egress ACL with this IP address for the interfaces corresponding to that same site.

Although the impact in control plane is minimal and the existing E-VPN BGP routes can be used with minimum modifications to its corresponding procedures, the same cannot be said in terms of network operations, management, and data plane. The use of IP addresses to represent the site of origin requires many IP addresses to be allocated and configured on a single NVE. For example a TOR with N interfaces may require one IP address per interface in worst case which may impact management and operational aspects of the Data Center Network. Also, the data-plane operation for Split-Horizon filtering will be different from that of MPLS client layer and it cannot be assumed that platforms/ASICs that support Split-Horizon filtering based on MPLS label can also support such function based on IP addresses. However, there are alternative options for performing such Split-Horizon filtering function when doing VXLAN/NVGRE encapsulation, while retaining a single IP address per NVE, and those will be described in a future revision of this document.

It should be noted that such filtering function is not required when doing active/standby multi-homing where load-balancing from a tenant can still be performed on a per VLAN basis - e.g., different VLANs are active on different NVEs connected to a multi-homed site. Furthermore, active/active multi-homing is primarily applicable when NVEs are on physical devices as opposed to on the hypervisor. For

example, [\[VXLAN\]](#) describes the use of physical devices as VXLAN gateways to connect a legacy network with a VXLAN overlay network. In such scenarios, one would expect: a) that the number of such gateways is not very large and/or b) that not all of them require active/active multi-homing.

[3.2.2.3.2](#) Aliasing

In E-VPN, the NVEs connected to a multi-homed site optionally advertise a VPN label used to load-balance traffic between NVEs, even when a given MAC address is learnt by only a single NVE connected to the site. In the case where VXLAN or NVGRE encapsulation is used, some alternative means that does not rely on MPLS labels is required to support aliasing. One solution would be to rely on the IP address per site assignment depicted in the previous section for aliasing as well: Effectively every NVE advertises an Ethernet A-D route for a

given site with the BGP Remote-Next-Hop attribute set to an IP address that has a 1:1 mapping to the site. The remote NVEs resolve an ESI (site ID) to a list of IP addresses corresponding to that site. Furthermore, a given MAC address that is associated with an ESI, in turn, gets resolved to this list of IP addresses. When a remote NVE wants to forward a packet for a given MAC address, it selects one of IP addresses from the list (using a hash value for load balancing) and encapsulates the packet using that IP address as the destination IP address in the VXLAN or NVGRE encapsulation. The source IP address will be that of the source multi-homed site. In case where the source site is single homed, the source IP address will be the loopback address of the NVE.

[3.2.2.3.3](#) Tunnel Endpoint Identification

To accommodate the Split Horizon as well as Aliasing functions of E-VPN, multiple IP tunnel endpoints (one per site) must be associated with the same NVE. As such, the mechanisms of [\[RFC5512\]](#) cannot be used to specify the tunnel endpoint and encapsulation, since those mechanisms only allow a single tunnel endpoint IP address to be associated with the BGP speaker. To alleviate this, the BGP Remote-Next-Hop attribute defined in [\[REMOTE-NH\]](#) can be used. Two new Tunnel Types would be required for VXLAN and NVGRE.

This attribute will be carried with the E-VPN Ethernet A-D route. The IP address field of this attribute serves two functions:

- It indicates the tunnel endpoint destination IP address that must be used when load-balancing traffic associated with a given site (i.e. ESI).
- It is used to build the egress ACL for filtering multi-destination traffic on multi-homed Ethernet Segments. In this context, the IP address is the tunnel endpoint source address.

It is worth noting that for multi-homed Ethernet segments, the NVE will always advertise an Ethernet A-D route with the Remote-Next-Hop attribute, in addition to the MAC Advertisement routes. In this case, the NVEs which receive the routes derive the tunnel endpoint IP address for a given MAC address as follows:

- 1- The NVE identifies the Ethernet Segment Identifier (ESI) associated with the MAC address, as encoded in the MAC Advertisement route.
- 2- The NVE then sets the tunnel endpoint IP address for that MAC to the value encoded in the Remote-Next-Hop attribute of the Ethernet AD

route advertised for the ESI identified in step 1.

On the other hand, for single-homed Ethernet segments, the NVE will only advertise the MAC Advertisement routes. In this latter case, the tunnel endpoint IP address is derived from the BGP Next-Hop attribute associated with the MAC Advertisement route.

[3.2.3](#) Support for Multicast

The E-VPN Inclusive Multicast BGP route can be used to discover the multicast endpoints associated with a given VXLAN VNI or NVGRE VSID. The Ethernet Tag field of this route is used to encode the VNI or VSID. 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 following tunnel types can be used for VXLAN/NVGRE:

- PIM-SSM Tree

- PIM-SM Tree
- BIDIR-PIM Tree
- Ingress Replication

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.

[3.2.4](#) Inter-AS Challenges

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.

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.

Even 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 E-VPN.

Using MPLS over GRE encapsulation addresses both of these concerns.

[4](#) Comparison between MPLSoGRE and VXLAN/NVGRE Encapsulation

The comparison between MPLSoGRE and VXLAN/NVGRE encapsulation depends on the required functionality on NVEs. If the hosts are single-homed to NVEs without any need to support redundancy group on NVEs, or if the hosts are multi-homed to two or more NVEs with active/standby redundancy but without the need for fast convergence upon a failure, then both MPLSoGRE and VXLAN/NVGRE do equally well with E-VPN control plane.

If we need to support active/standby multi-homing with fast convergence upon a failure or if we need to support active/active multi-homing, then MPLSoGRE encap can provide these additional functionality without any impact to E-VPN routes and procedures. Furthermore, it can provide complete support for inter-AS operation and complete set of E-VPN functions without impacting IP address assignment and management of the underlying network. However, VXLAN/NVGRE impacts E-VPN routes and procedures as well as the underlying data plane behavior as noted above. Furthermore, there are implications to IP address assignments, security, and inter-AS operations. It should be noted that the additional requirements on the data plane behavior as well as the above implications are the consequence of the functionality that need to be supported and

independent of the control-plane choice.

As noted previously, there are existing core switches that do not support ECMP by hashing the GRE key; however, vast majority of

existing core switches support ECMP by hashing UDP header; therefore, VXLAN encapsulation can provide better ECMP functions for these existing switches. Thus, the choice for overlay encapsulation depends on needed functionality, inter-AS scenarios, security requirements, and the ECMP capabilities of the core switches.

[5](#) Acknowledgement

The authors would like to thank John Mullooly and Dave Smith for providing value comments and feedbacks.

[6](#) Security Considerations

[7](#) IANA Considerations

[8](#) References

[8.1](#) Normative References

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

[REMOTE-NH] Van de Velde et al., "BGP Remote-Next-Hop", [draft-vandevelde-idr-remote-next-hop-01.txt](#), work in progress, July 2012.

[8.2](#) Informative References

[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.

[E-VPN] Sajassi et al., "BGP MPLS Based Ethernet VPN", [draft-ietf-l2vpn-evpn-01.txt](#), work in progress, February, 2012.

[Problem-Statement] Narten et al., "Problem Statement: Overlays for Network Virtualization", [draft-ietf-nvo3-overlay-problem-statement-00](#), September 2012.

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

Authors' Addresses

Ali Sajassi
Cisco
Email: sajassi@cisco.com

Samer Salam
Cisco
595 Burrard Street
Vancouver, BC V7X 1J1, Canada
Email: ssalam@cisco.com

Keyur Patel
Cisco
170 West Tasman Drive
San Jose, CA 95134, US
Email: Keyupate@cisco.com

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

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

