

NV03 Workgroup  
Internet Draft  
Intended status: Informational

J. Rabadan, Ed.  
M. Bocci  
Nokia

S. Boutros  
WMware

Expires: May 3, 2018

October 30, 2017

Applicability of EVPN to NV03 Networks  
draft-rabadan-nvo3-evpn-applicability-00

Abstract

In NV03 networks, Network Virtualization Edge (NVE) devices sit at the edge of the underlay network and provide Layer-2 and Layer-3 connectivity among Tenant Systems (TSes) of the same tenant. The NVEs need to build and maintain mapping tables so that they can deliver encapsulated packets to their intended destination NVE(s). While there are different options to create and disseminate the mapping table entries, NVEs may exchange that information directly among themselves via a control-plane protocol, such as EVPN. EVPN provides an efficient, flexible and unified control-plane option that can be used for Layer-2 and Layer-3 Virtual Network (VN) service connectivity. This document describes the applicability of EVPN to NV03 networks and how EVPN solves the challenges in those networks.

Status of this Memo

This Internet-Draft is submitted 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

Internet-Draft

EVPN Applicability for NV03

October 30, 2017

<http://www.ietf.org/ietf/lid-abstracts.txt>

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

This Internet-Draft will expire on May 3, 2018.

## Copyright Notice

Copyright (c) 2017 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to [BCP 78](#) 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

<a href="#">1. Introduction</a>	<a href="#">3</a>
<a href="#">2. EVPN and NV03 Terminology</a>	<a href="#">3</a>
<a href="#">3. Why Is EVPN Needed In NV03 Networks?</a>	<a href="#">5</a>
<a href="#">4. Applicability of EVPN to NV03 Networks</a>	<a href="#">7</a>
<a href="#">4.1. EVPN Route Types used in NV03 Networks</a>	<a href="#">7</a>
<a href="#">4.2. EVPN Basic Applicability For Layer-2 Services</a>	<a href="#">8</a>
<a href="#">4.2.1. Auto-Provisioning of the NVE services</a>	<a href="#">9</a>
<a href="#">4.2.2. Remote NVE Auto-Discovery</a>	<a href="#">10</a>
<a href="#">4.2.3. Distribution Of Tenant MAC and IP Information</a>	<a href="#">10</a>
<a href="#">4.3. EVPN Basic Applicability for Layer-3 Services</a>	<a href="#">11</a>
4.4. EVPN as a Control Plane for NV03 Encapsulations and GENEVE	<a href="#">13</a>
<a href="#">4.5. EVPN OAM and application to NV03</a>	<a href="#">14</a>
<a href="#">4.6. EVPN as the control plane for NV03 security</a>	<a href="#">14</a>
<a href="#">4.7. Advanced EVPN Features For NV03 Networks</a>	<a href="#">14</a>
<a href="#">4.7.1. Virtual Machine (VM) Mobility</a>	<a href="#">14</a>
4.7.2. MAC Protection, Duplication Detection and Loop	

Protection . . . . .	<a href="#">15</a>
4.7.3. Reduction/Optimization of BUM Traffic In Layer-2 Services . . . . .	<a href="#">15</a>
4.7.4. Ingress Replication (IR) Optimization For BUM Traffic .	<a href="#">16</a>
<a href="#">4.7.5.</a> EVPN Multi-homing . . . . .	<a href="#">17</a>

4.7.6. EVPN Recursive Resolution for Inter-Subnet Unicast Forwarding . . . . .	<a href="#">18</a>
<a href="#">4.7.7.</a> EVPN Optimized Inter-Subnet Multicast Forwarding . . .	<a href="#">19</a>
<a href="#">4.7.8.</a> Data Center Interconnect (DCI) . . . . .	<a href="#">19</a>
<a href="#">5.</a> Conclusion . . . . .	<a href="#">20</a>
<a href="#">6.</a> Conventions used in this document . . . . .	<a href="#">20</a>
<a href="#">7.</a> Security Considerations . . . . .	<a href="#">20</a>
<a href="#">8.</a> IANA Considerations . . . . .	<a href="#">20</a>
<a href="#">9.</a> References . . . . .	<a href="#">20</a>
<a href="#">9.1</a> Normative References . . . . .	<a href="#">21</a>
<a href="#">9.2</a> Informative References . . . . .	<a href="#">21</a>
<a href="#">10.</a> Acknowledgments . . . . .	<a href="#">23</a>
<a href="#">11.</a> Contributors . . . . .	<a href="#">23</a>
<a href="#">12.</a> Authors' Addresses . . . . .	<a href="#">23</a>

## [1.](#) Introduction

In NV03 networks, Network Virtualization Edge (NVE) devices sit at the edge of the underlay network and provide Layer-2 and Layer-3 connectivity among Tenant Systems (TSes) of the same tenant. The NVEs need to build and maintain mapping tables so that they can deliver encapsulated packets to their intended destination NVE(s). While there are different options to create and disseminate the mapping table entries, NVEs may exchange that information directly among themselves via a control-plane protocol, such as EVPN. EVPN provides an efficient, flexible and unified control-plane option that can be used for Layer-2 and Layer-3 Virtual Network (VN) service connectivity.

In this document, we assume that the EVPN control-plane module resides in the NVEs. The NVEs can be virtual switches in hypervisors, TOR/Leaf switches or Data Center Gateways. Note that Network Virtualization Authorities (NVAs) may be used to provide the

forwarding information to the NVEs, and in that case, EVPN could be used to disseminate the information across multiple federated NVAs. The applicability of EVPN would then be similar to the one described in this document. However, for simplicity, the description assumes control-plane communication among NVE(s).

## [2.](#) EVPN and NV03 Terminology

- o NVE: Network Virtualization Edge is a network entity that sits at the edge of an underlay network and implements L2 and/or L3 network virtualization functions. The network-facing side of the NVE uses

the underlying L3 network to tunnel tenant frames to and from other NVEs. The tenant-facing side of the NVE sends and receives Ethernet frames to and from individual Tenant Systems. In this document, an NVE could be implemented as a virtual switch within a hypervisor, a switch or a router, and runs EVPN in the control-plane.

- o EVI: or EVPN Instance. It is a Layer-2 Virtual Network that uses an EVPN control-plane to exchange reachability information among the member NVEs. It corresponds to a set of MAC-VRFs of the same tenant. See MAC-VRF in this section.
- o BD: or Broadcast Domain, it corresponds to a tenant IP subnet. If no suppression techniques are used, a BUM frame that is injected in a BD will reach all the NVEs that are attached to that BD. An EVI may contain one or multiple BDs depending on the service model [[RFC7432](#)]. This document will use the term BD to refer to a tenant subnet.
- o EVPN VLAN-based service model: it refers to one of the three service models defined in [[RFC7432](#)]. It is characterized as a BD that uses a single VLAN per physical access port to attach tenant traffic to the BD. In this service model, there is only one BD per EVI.
- o EVPN VLAN-bundle service model: similar to VLAN-based but uses a bundle of VLANs per physical port to attach tenant traffic to the BD. As in VLAN-based, in this model there is a single BD per EVI.
- o EVPN VLAN-aware bundle service model: similar to the VLAN-bundle

model but each individual VLAN value is mapped to a different BD. In this model there are multiple BDs per EVI for a given tenant. Each BD is identified by an "Ethernet Tag", that is a control-plane value that identifies the routes for the BD within the EVI.

- o IP-VRF: an IP Virtual Routing and Forwarding table, as defined in [[RFC4364](#)]. It stores IP Prefixes that are part of the tenant's IP space, and are distributed among NVEs of the same tenant by EVPN. Route-Distinguisher (RD) and Route-Target(s) (RTs) are required properties of an IP-VRF. An IP-VRF is instantiated in an NVE for a given tenant, if the NVE is attached to multiple subnets of the tenant and local inter-subnet-forwarding is required across those subnets.
- o MAC-VRF: a MAC Virtual Routing and Forwarding table, as defined in [[RFC7432](#)]. The instantiation of an EVI (EVPN Instance) in an NVE. Route-distinguisher (RD) and Route-Target(s) (RTs) are required properties of a MAC-VRF and they are normally different than the ones defined in the associated IP-VRF (if the MAC-VRF has an IRB

interface).

- o BT: a Bridge Table, as defined in [[RFC7432](#)]. A BT is the instantiation of a BD in an NVE. When there is a single BD on a given EVI, the MAC-VRF is equivalent to the BT on that NVE.
- o AC: Attachment Circuit or logical interface associated to a given BT. To determine the AC on which a packet arrived, the NVE will examine the physical/logical port and/or VLAN tags (where the VLAN tags can be individual c-tags, s-tags or ranges of both).
- o IRB: Integrated Routing and Bridging interface. It refers to the logical interface that connects a BD instance (or a BT) to an IP-VRF and allows to forward packets with destination in a different subnet.
- o ES: Ethernet Segment. When a Tenant System (TS) is connected to one or more NVEs via a set of Ethernet links, then that set of links is referred to as an 'Ethernet segment'. Each ES is represented by a unique Ethernet Segment Identifier (ESI) in the NV03 network and the ESI is used in EVPN routes that are specific to that ES.

- o NV03 or Overlay tunnels: Network Virtualization Over Layer-3 tunnels. In this document, NV03 tunnels or simply Overlay tunnels will be used interchangeably. Both terms refer to a way to encapsulate tenant frames or packets into IP packets whose IP Source Addresses (SA) or Destination Addresses (DA) belong to the underlay IP address space, and identify NVEs connected to the same underlay network. Examples of NV03 tunnel encapsulations are VXLAN [[RFC7348](#)], [[GENEVE](#)] or MPLSoUDP [[RFC7510](#)].
- o VNI: Virtual Network Identifier. Irrespective of the NV03 encapsulation, the tunnel header always includes a VNI that is added at the ingress NVE (based on the mapping table lookup) and identifies the BT at the egress NVE. This VNI is called VNI in VXLAN or GENEVE, VSID in nvGRE or Label in MPLSoGRE or MPLSoUDP. This document will refer to VNI as a generic Virtual Network Identifier for any NV03 encapsulation.
- o BUM: Broadcast, Unknown unicast and Multicast frames.
- o SA and DA: they refer to Source Address and Destination Address. They are used along with MAC or IP, e.g. IP SA or MAC DA.

### 3. Why Is EVPN Needed In NV03 Networks?

Data Centers have adopted NV03 architectures mostly due to the issues

discussed in [[RFC7364](#)]. The architecture of a Data Center is nowadays based on a CLOS design, where every Leaf is connected to a layer of Spines, and there is a number of ECMP paths between any two leaf nodes. All the links between Leaf and Spine nodes are routed links, forming what we also know as an underlay IP Fabric. The underlay IP Fabric does not have issues with loops or flooding (like old Spanning Tree Data Center designs did), convergence is fast and ECMP provides a fairly optimal bandwidth utilization on all the links.

On this architecture and as discussed by [[RFC7364](#)] multi-tenant intra-subnet and inter-subnet connectivity services are provided by NV03 tunnels, being VXLAN [[RFC7348](#)] or [[GENEVE](#)] two examples of such tunnels.

Why is a control-plane protocol along with NV03 tunnels required?

There are three main reasons:

- a) Auto-discovery of the remote NVEs that are attached to the same BD as the ingress NVE is.
- b) Dissemination of the MAC/IP host information so that mapping tables can be populated on the remote NVEs.
- c) Advanced features such as MAC Mobility, MAC Protection, BUM traffic reduction/suppression, Multi-homing, etc.

A possible approach to achieve points (a) and (b) above is "Flood and Learn". "Flood and Learn" refers to not using a specific control-plane on the NVEs, but rather "Flood" BUM traffic from the ingress NVE to all the egress NVEs attached to the same BD. The egress NVEs may then use data path MAC SA "Learning" on the frames received over the NV03 tunnels. When the destination host replies back and the frames arrive at the NVE that initially flooded BUM frames, the NVE will also "Learn" the MAC SA of the frame encapsulated on the NV03 tunnel. This approach has the following drawbacks:

- o In order to Flood a given BUM frame, the ingress NVE must know the IP addresses of the remote NVEs attached to the same BD. This may be done as follows:
  - The remote tunnel IP addresses can be statically provisioned on the ingress NVE. If the ingress NVE receives an access BUM frame for the BD, it will do ingress replication and will send the frame to all the configured egress NVE IP DAs in the BD.
  - All the NVEs attached to the same BD can subscribe to an underlay IP Multicast Group that is dedicated to that BD. When an ingress NVE receives a BUM frame on an ingress AC, it will send a single

copy of the frame encapsulated into an NV03 tunnel, using the multicast address as IP DA of the tunnel. This solution requires PIM in the underlay network and the association of individual BDs to underlay IP multicast groups.

- o "Flood and Learn" solves the issues of auto-discovery and learning of the MAC to VNI/tunnel IP mapping on the NVEs for a given BD. However, it does not provide a solution for advanced features and

it does not scale well.

EVPN provides a unified control-plane that solves the NVE auto-discovery, tenant MAP/IP dissemination and advanced features in a scalable way and keeping the independence of the underlay IP Fabric, i.e. there is no need to enable PIM in the underlay network and maintain multicast states for tenant BDs.

[Section 4](#) describes how to apply EVPN to meet the control-plane requirements in an NV03 network.

#### [4.](#) Applicability of EVPN to NV03 Networks

This section discusses the applicability of EVPN to NV03 networks. The intend is not to provide a comprehensive explanation of the protocol itself but give an introduction and point at the corresponding reference document, so that the reader can easily find more details if needed.

##### [4.1.](#) EVPN Route Types used in NV03 Networks

EVPN supports multiple Route Types and each type has a different function. For convenience, Table 1 shows a summary of all the existing EVPN route types and its usage. We will refer to these route types as RT-x throughout the rest of the document, where x is the type number included in the first column of Table 1.



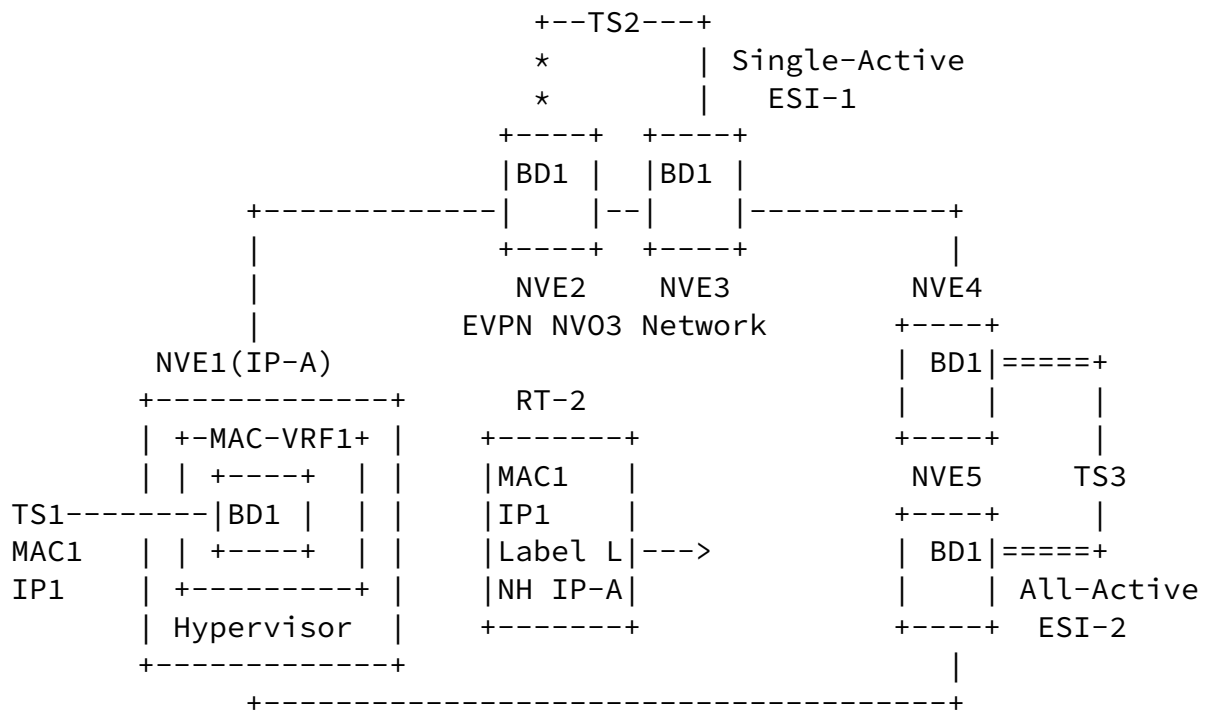
Type	Description	Usage
1	Ethernet Auto-Discovery	Multi-homing: Per-ES: Mass withdrawal Per-EVI: aliasing/backup
2	MAC/IP Advertisement	Host MAC/IP dissemination Supports MAC mobility and protection
3	Inclusive Multicast Ethernet Tag	NVE discovery and BUM flooding tree setup
4	Ethernet Segment	Multi-homing: DF Election
5	IP Prefix	IP Prefix dissemination
6	Selective Multicast Ethernet Tag	Indicate interest for a multicast S,G or *,G
7	IGMP Join Synch	Multi-homing: S,G or *,G state synch
8	IGMP Leave Synch	Multi-homing: S,G or *,G leave synch
9	Per-Region I-PMSI A-D	BUM tree creation across regions
10	S-PMSI A-D	Multicast tree for S,G or *,G states
11	Leaf A-D	Used for responses to explicit tracking

Table 1 EVPN route types

#### 4.2. EVPN Basic Applicability For Layer-2 Services

Although the applicability of EVPN to NV03 networks spans multiple documents, EVPN's baseline specification is [RFC7432]. [RFC7432] allows multipoint layer-2 VPNs to be operated as [RFC4364] IP-VPNs, where MACs and the information to setup flooding trees are distributed by MP-BGP. Based on [RFC7432], [EVPN-OVERLAY] describes how to use EVPN to deliver Layer-2 services specifically in NV03 Networks.

Figure 1 represents a Layer-2 service deployed with an EVPN BD in an NV03 network.



In a simple NV03 network, such as the example of Figure 1, these are the basic constructs that EVPN uses for Layer-2 services (or Layer-2 Virtual Networks):

- o BD1 is an EVPN Broadcast Domain for a given tenant and TS1, TS2 and TS3 are connected to it. The five represented NVEs are attached to BD1 and are connected to the same underlay IP network. That is, each NVE learns the remote NVEs' loopback addresses via underlay routing protocol.
- o NVE1 is deployed as a virtual switch in a Hypervisor with IP-A as underlay loopback IP address. The rest of the NVEs in Figure 1 are physical switches and TS2/TS3 are multi-homed to them. TS1 is a virtual machine, identified by MAC1 and IP1.

#### 4.2.1. Auto-Provisioning of the NVE services

When deploying a Layer-2 Service for a tenant in an NV03 network, all the NVEs attached to the same subnet must be configured with a MAC-VRF and the BD for the subnet, as well as certain parameters for them. Note that, if the EVPN service model is VLAN-based or VLAN-bundle, implementations do not normally have a specific provisioning for the BD (since it is in that case the same construct as the MAC-

EVPN allows auto-deriving as many MAC-VRF parameters as possible. As an example:

- o The MAC-VRF's RT and RD for the EVPN routes may be auto-derived. Section 5.1.2.1 in [\[EVPN-OVERLAY\]](#) specifies how to auto-derive a MAC-VRF's RT as long as VLAN-based service model is implemented. [\[RFC7432\]](#) specifies how to auto-derive the RD.
- o The EVPN Multi-homing parameters can also be auto-derived. If the TS is connected to the multi-homed NVEs via LAG and LACP, the ESI can be auto-derived from the LACP information as described in [\[RFC7432\], section 5](#). The ESI-import Route-Target can also be auto-derived based on the procedures of [\[RFC7432\], section 7.6](#).

#### [4.2.2](#). Remote NVE Auto-Discovery

Auto-discovery is one of the basic capabilities of EVPN. Auto-discovery via MP-BGP is used to discover the remote NVEs attached to a given BD, NVEs participating in a given redundancy group, the tunnel encapsulation types supported by an NVE, etc.

In particular, when a new MAC-VRF and BD are enabled, the NVE will advertise a new RT-3. Besides other fields, the RT-3 will encode the IP address of the advertising NVE, the Ethernet Tag (which is zero in case of VLAN-based and VLAN-bundle models) and also a PMSI Tunnel Attribute (PTA) that indicates the information about the intended way to deliver BUM traffic for the BD.

In the example of Figure 1, when MAC-VRF1/BD1 are enabled, NVE1 will send an RT-3 including its own IP address, Ethernet-Tag for BD1 and the PTA. Assuming Ingress Replication (IR), the RT-3 will include an identification for IR in the PTA and the VNI the NVEs must use to send BUM traffic to the advertising NVE. The other NVEs in the BD, will import the RT-3 and will add NVE1's IP address to the flooding list for BD1. Note that the RT-3 is also sent with a BGP encapsulation attribute [\[TUNNEL-ENCAP\]](#) that indicates what NV03 encapsulation the remote NVEs should use when sending BUM traffic to NVE1.

Refer to [\[RFC7432\]](#) for more information about the RT-3 and forwarding of BUM traffic, and to [\[EVPN-OVERLAY\]](#) for its considerations on NV03 networks.

#### [4.2.3.](#) Distribution Of Tenant MAC and IP Information

Tenant MAC/IP information is advertised to remote NVEs using RT-2s.

Following the example of Figure 1:

- o In a given EVPN BD, TSes' MAC addresses are first learned at the NVE they are attached to, via data path or management plane learning. In Figure 1 we assume NVE1 learns MAC1/IP1 in the management plane (for instance, via Cloud Management System) since the NVE is a virtual switch. NVE2, NVE3, NVE4 and NVE4 are TOR/Leaf switches and they normally learn MAC addresses via data path.
- o Once NVE1's BD1 learns MAC1/IP1, NVE1 advertises that information along with a VNI and Next Hop IP-A in an RT-2. The EVPN routes are advertised using the RD/RTs of the MAC-VRF where the BD belongs. All the NVEs in BD1 learn local MAC/IP addresses and advertise them in RT-2 routes in a similar way.
- o The remote NVEs can then add MAC1 to their mapping table for BD1 (BT). For instance, when TS3 sends frames to NVE4 with MAC DA = MAC1, NVE4 does a MAC lookup on the BT that yields IP-A and Label L. NVE4 can then encapsulate the frame into an NV03 tunnel with IP-A as the tunnel IP DA and L as the Virtual Network Identifier. Note that the RT-2 may also contain the host's IP address (as in the example of Figure 1). While the MAC of the received RT-2 is installed in the BT, the IP address may be installed in the Proxy-ARP/ND table (if enabled) or in the ARP/IP-VRF tables if the BD has an IRB. See [section 4.7.3.](#) to see more information about Proxy-ARP/ND and [section 4.3.](#) for more details about IRB and Layer-3 services.

Refer to [\[RFC7432\]](#) and [\[EVPN-OVERLAY\]](#) for more information about the RT-2 and forwarding of known unicast traffic.

### 4.3. EVPN Basic Applicability for Layer-3 Services

[IP-PREFIX] and [INTER-SUBNET] are the reference documents that describe how EVPN can be used for Layer-3 services. Inter Subnet Forwarding in EVPN networks is implemented via IRB interfaces between BDs and IP-VRFs. As discussed, an EVPN BD corresponds to an IP subnet. When IP packets generated in a BD are destined to a different subnet (different BD) of the same tenant, the packets are sent to the IRB attached to local BD in the source NVE. As discussed in [INTER-SUBNET], depending on how the IP packets are forwarded between the ingress NVE and the egress NVE, there are two forwarding models: Asymmetric and Symmetric.

The Asymmetric model is illustrated in the example of Figure 2 and it requires the configuration of all the BDs of the tenant in all the NVEs attached to the same tenant. In that way, there is no need to

advertise IP Prefixes between NVEs since all the NVEs are attached to all the subnets. It is called Asymmetric because the ingress and egress NVEs do not perform the same number of lookups in the data plane. In Figure 2, if TS1 and TS2 are in different subnets, and TS1 sends IP packets to TS2, the following lookups are required in the data path: a MAC lookup (on BD1's table), an IP lookup (on the IP-VRF) and a MAC lookup (on BD2's table) at the ingress NVE1 and then only a MAC lookup at the egress NVE. The two IP-VRFs in Figure 2 are not connected by tunnels and all the connectivity between the NVEs is done based on tunnels between the BDs.

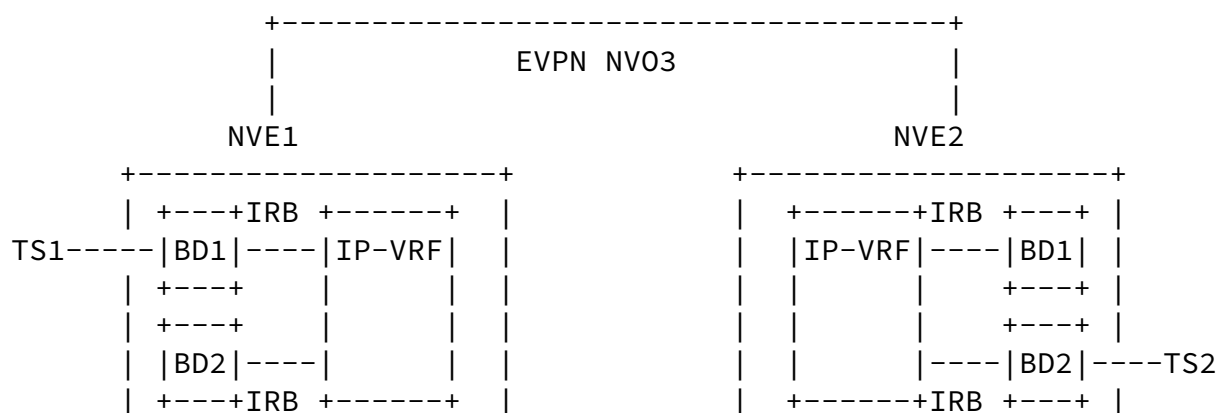




Figure 2 EVPN for L3 in an NV03 Network - Asymmetric model

In the Symmetric model, depicted in Figure 3, there are the same data path lookups at the ingress and egress NVEs. For example, if TS1 sends IP packets to TS3, the following data path lookups are required: a MAC lookup at NVE1's BD1 table, an IP lookup at NVE1's IP-VRF and then IP lookup and MAC lookup at NVE2's IP-VRF and BD3 respectively. In the Symmetric model, the Inter Subnet connectivity between NVEs is done based on tunnels between the IP-VRFs.

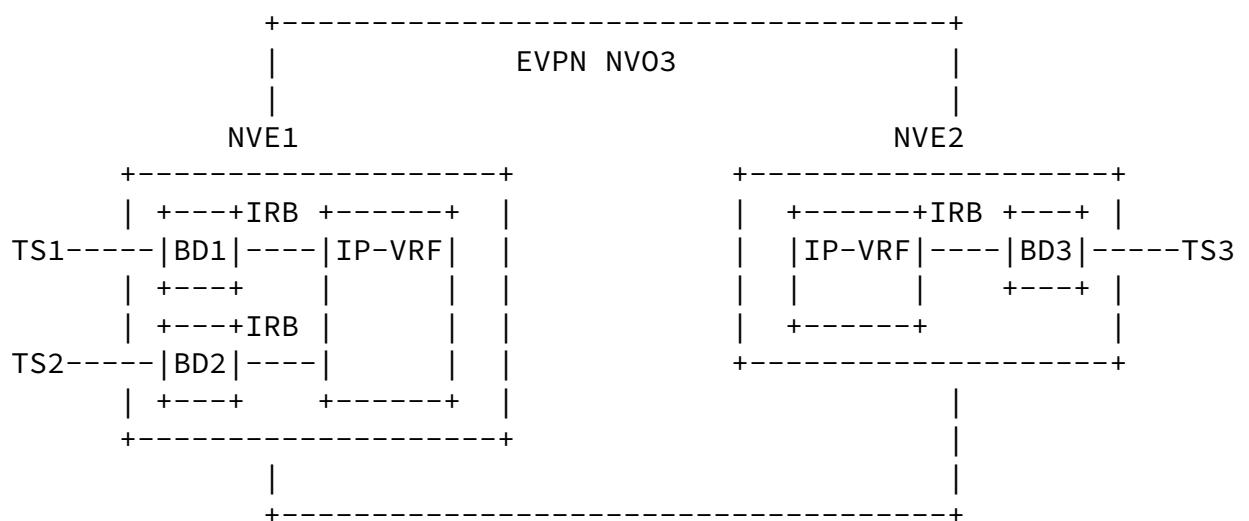


Figure 3 EVPN for L3 in an NV03 Network - Symmetric model

The Symmetric model scales better than the Asymmetric model because

it does not require the NVEs to be attached to all the tenant's subnets. However, it requires the use of NV03 tunnels on the IP-VRFs and the exchange of IP Prefixes between the NVEs in the control plane. EVPN uses RT-2 and RT-5 routes for the exchange of host IP routes (in the case of RT-2 and RT-5) and IP Prefixes (RT-5s) of any length. As an example, in Figure 3, NVE2 needs to advertise TS3's host route and/or TS3's subnet, so that the IP lookup on NVE1's IP-VRF succeeds.

[INTER-SUBNET] specifies the use of RT-2s for the advertisement of host routes. Section 4.4.1 in [[IP-PREFIX](#)] specifies the use of RT-5s for the advertisement of IP Prefixes in an "Interface-less IP-VRF-to-IP-VRF Model".

#### [4.4](#). EVPN as a Control Plane for NV03 Encapsulations and GENEVE

[EVPN-OVERLAY] describes how to use EVPN for NV03 encapsulations, such as VXLAN, nvGRE or MPLSoGRE. The procedures can be easily applicable to any other NV03 encapsulation, in particular GENEVE.

The NV03 working group has been working on different data plane encapsulations. The Generic Network Virtualization Encapsulation [[GENEVE](#)] has been recommended to be the proposed standard for NV03 Encapsulation. The EVPN control plane can signal the GENEVE encapsulation type in the BGP Tunnel Encapsulation Extended Community (see [[TUNNEL-ENCAP](#)]).

The NV03 encapsulation design team has made a recommendation in

[NV03-ENCAP] for a control plane to:

- 1- Negotiate a subset of GENEVE option TLVs that can be carried on a GENEVE tunnel
- 2- Enforce an order for GENEVE option TLVs and
- 3- Limit the total number of options that could be carried on a GENEVE tunnel.

The EVPN control plane can easily extend the BGP Tunnel Encapsulation

Attribute sub-TLV [[TUNNEL-ENCAP](#)] to specify the GENEVE tunnel options that can be received or transmitted over a GENEVE tunnels by a given NVE.

#### [4.5.](#) EVPN OAM and application to NV03

EVPN OAM (as in [[EVPN-LSP-PING](#)]) defines mechanisms to detect data plane failures in an EVPN deployment over an MPLS network. These mechanisms detect failures related to P2P and P2MP connectivity, for multi-tenant unicast and multicast L2 traffic, between multi-tenant access nodes connected to EVPN PE(s), and in a single-homed, single-active or all-active redundancy model.

In general, EVPN OAM mechanisms defined for EVPN deployed in MPLS networks are equally applicable for EVPN in NV03 networks.

#### [4.6.](#) EVPN as the control plane for NV03 security

EVPN can be used to signal the security protection capabilities of a sender NVE, as well as what portion of an NV03 packet (taking a GENEVE packet as an example) can be protected by the sender NVE, to ensure the privacy and integrity of tenant traffic carried over the NV03 tunnels.

#### [4.7.](#) Advanced EVPN Features For NV03 Networks

This section describes how EVPN can be used to deliver advanced capabilities in NV03 networks.

##### [4.7.1.](#) Virtual Machine (VM) Mobility

[RFC7432] replaces the traditional Ethernet Flood-and-Learn behavior among NVEs with BGP-based MAC learning, which in return provides more control over the location of MAC addresses in the BD and consequently advanced features, such as MAC Mobility. If we assume that VM

Mobility means the VM's MAC and IP addresses move with the VM, EVPN's MAC Mobility is the required procedure that facilitates VM Mobility. According to [[RFC7432](#)] [section 15](#), when a MAC is advertised for the first time in a BD, all the NVEs attached to the BD will store



Sequence Number zero for that MAC. When the MAC "moves" within the same BD but to a remote NVE, the NVE that just learned locally the MAC, increases the Sequence Number in the RT-2's MAC Mobility extended community to indicate that it owns the MAC now. That makes all the NVE in the BD change their tables immediately with no need to wait for any aging timer. EVPN guarantees a fast MAC Mobility without flooding or black-holes in the BD.

#### [4.7.2.](#) MAC Protection, Duplication Detection and Loop Protection

The advertisement of MACs in the control plane, allows advanced features such as MAC protection, Duplication Detection and Loop Protection.

[RFC7432] MAC Protection refers to EVPN's ability to indicate - in an RT-2 - that a MAC must be protected by the NVE receiving the route. The Protection is indicated in the "Sticky bit" of the MAC Mobility extended community sent along the RT-2 for a MAC. NVEs' ACs that are connected to subject-to-be-protected servers or VMs may set the Sticky bit on the RT-2s sent for the MACs associated to the ACs. Also statically configured MAC addresses should be advertised as Protected MAC addresses, since they are not subject to MAC Mobility procedures.

[RFC7432] MAC Duplication Detection refers to EVPN's ability to detect duplicate MAC addresses. A "MAC move" is a relearn event that happens at an access AC or through an RT-2 with a Sequence Number that is higher than the stored one for the MAC. When a MAC moves a number of times N within an M-second window between two NVEs, the MAC is declared as Duplicate and the detecting NVE does not re-advertise the MAC anymore.

While [[RFC7432](#)] provides MAC Duplication Detection, it does not protect the BD against loops created by backdoor links between NVEs. However, the same principle (based on the Sequence Number) may be extended to protect the BD against loops. When a MAC is detected as duplicate, the NVE may install it as a black-hole MAC and drop received frames with MAC SA and MAC DA matching that duplicate MAC. Loop Protection is described in [[LOOP](#)].

#### [4.7.3.](#) Reduction/Optimization of BUM Traffic In Layer-2 Services

In BDs with a significant amount of flooding due to Unknown unicast

and Broadcast frames, EVPN may help reduce and sometimes even suppress the flooding.

In BDs where most of the Broadcast traffic is caused by ARP (Address Resolution Protocol) and ND (Neighbor Discovery) protocols on the TSes, EVPN's Proxy-ARP and Proxy-ND capabilities may reduce the flooding drastically. The use of Proxy-ARP/ND is specified in [PROXY-ARP-ND].

Proxy-ARP/ND procedures along with the assumption that TSes always issue a GARP (Gratuitous ARP) or an unsolicited Neighbor Advertisement message when they come up in the BD, may drastically reduce the unknown unicast flooding in the BD.

The flooding caused by TSes' IGMP/MLD or PIM messages in the BD may also be suppressed by the use of IGMP/MLD and PIM Proxy functions, as specified in [IGMP-MLD-PROXY] and [PIM-PROXY]. These two documents also specify how to forward IP multicast traffic efficiently within the same BD, translate soft state IGMP/MLD/PIM messages into hard state BGP routes and provide fast-convergence redundancy for IP Multicast on multi-homed Ethernet Segments (ESes).

#### 4.7.4. Ingress Replication (IR) Optimization For BUM Traffic

When an NVE attached to a given BD needs to send BUM traffic for the BD to the remote NVEs attached to the same BD, IR is a very common option in NV03 networks, since it is completely independent of the multicast capabilities of the underlay network. Also, if the optimization procedures to reduce/suppress the flooding in the BD are enabled ([section 4.7.3](#)), in spite of creating multiple copies of the same frame at the ingress NVE, IR may be good enough. However, in BDs where Multicast (or Broadcast) traffic is significant, IR may be very inefficient and cause performance issues on virtual-switch-based NVEs.

[OPT-IR] specifies the use of AR (Assisted Replication) NV03 tunnels in EVPN BDs. AR retains the independence of the underlay network while providing a way to forward Broadcast and Multicast traffic efficiently. AR uses AR-REPLICATORS that can replicate the Broadcast/Multicast traffic on behalf of the AR-LEAF NVEs. The AR-LEAF NVEs are typically virtual-switches or NVEs with limited replication capabilities. AR can work in a single-stage replication mode (Non-Selective Mode) or in a dual-stage replication mode (Selective Mode). Both modes are detailed in [OPT-IR].

In addition, [OPT-IR] also describes a procedure to avoid sending

need that type of traffic. This is done by enabling PFL (Pruned Flood Lists) on a given BD. For instance, an virtual-switch NVE that learns all its local MAC addresses for a BD via Cloud Management System, does not need to receive the BD's Unknown unicast traffic. PFLs help optimize the BUM flooding in the BD.

#### [4.7.5](#). EVPN Multi-homing

Another fundamental concept in EVPN is multi-homing. A given TS can be multi-homed to two or more NVEs for a given BD, and the set of links connected to the same TS is defined as Ethernet Segment (ES). EVPN supports single-active and all-active multi-homing. In single-active multi-homing only one link in the ES is active. In all-active multi-homing all the links in the ES are active for unicast traffic. Both modes support load-balancing:

- o Single-active multi-homing means per-service load-balancing to/from the TS, for example, in Figure 1, for BD1 only one of the NVEs can forward traffic from/to TS2. For a different BD, the other NVE may forward traffic.
- o All-active multi-homing means per-flow load-balancing for unicast frames to/from the TS. That is, in Figure 1 and for BD1, both NVE4 and NVE5 can forward known unicast traffic to/from TS3. For BUM traffic only one of the two NVEs can forward traffic to TS3, and both can forward traffic from TS3.

There are two key aspects of EVPN multi-homing:

- o DF (Designated Forwarder) election: the DF is the NVE that forwards the traffic to the ES in single-active mode. In case of all-active, the DF is the NVE that forwards the BUM traffic to the ES.
- o Split-horizon function: prevents the TS from receiving echoed BUM frames that the TS itself sent to the ES. This is especially relevant in all-active ESes, where the TS may forward BUM frames to a non-DF NVE that can flood the BUM frames back to the DF NVE

and then the TS. As an example, in Figure 1, assuming NVE4 is the DF for ES-2 in BD1, BUM frames sent from TS3 to NVE5 will be received at NVE4 and, since NVE4 is the DF for DB1, it will forward them back to TS3. Split-horizon allows NVE4 (and any multi-homed NVE for that matter) to identify if an EVPN BUM frame is coming from the same ES or different, and if the frame belongs to the same ES2, NVE4 will not forward the BUM frame to TS3, in spite of being the DF.

While [[RFC7432](#)] describes the default algorithm for the DF Election, [[HRW-DF](#)], [[PREF-DF](#)] and [[AC-DF](#)] specify other algorithms and procedures that optimize the DF Election.

The Split-horizon function is specified in [[RFC7432](#)] and it is carried out by using a special ESI-label that it identifies in the data path, all the BUM frames being originated from a given NVE and ES. Since the ESI-label is an MPLS label, it cannot be used in all the non-MPLS NV03 encapsulations, therefore [[EVPN-OVERLAY](#)] defines a modified Split-horizon procedure that is based on the IP SA of the NV03 tunnel, known as "Local-Bias". It is worth noting that Local-Bias only works for all-active multi-homing, and not for single-active multi-homing.

#### [4.7.6](#). EVPN Recursive Resolution for Inter-Subnet Unicast Forwarding

[Section 4.3](#). describes how EVPN can be used for Inter Subnet Forwarding among subnets of the same tenant. RT-2s and RT-5s allow the advertisement of host routes and IP Prefixes (RT-5) of any length. The procedures outlined by [section 4.3](#). are similar to the ones in [[RFC4364](#)], only for NV03 tunnels. However, [[EVPN-PREFIX](#)] also defines advanced Inter Subnet Forwarding procedures that allow the resolution of RT-5s to not only BGP next-hops but also "overlay indexes" that can be a MAC, a GW IP or an ESI, all of them in the tenant space.

Figure 4 illustrates an example that uses Recursive Resolution to a GWIP as per [[IP-PREFIX](#)] [section 4.4.2](#). In this example, IP-VRFs in NVE1 and NVE2 are connected by a SBD (Supplementary BD). An SBD is a BD that connects all the IP-VRFs of the same tenant, via IRB, and has no ACs. NVE1 advertises the host route TS2-IP/L (IP address and Prefix Length of TS2) in an RT-5 with overlay index GWIP=IP1. Also,



the efficient forwarding of IP multicast traffic from any BD to any other BD (or even to the same BD where the Source resides). The [OISM] procedures are supported along with EVPN multi-homing, and for any tree allowed on NV03 networks, including IR or AR.

#### [4.7.8. Data Center Interconnect \(DCI\)](#)

Tenant Layer-2 and Layer-3 services deployed on NV03 networks must be extended to remote NV03 networks that are connected via non-NOV3 WAN networks (mostly MPLS based WAN networks). [EVPN-DCI] defines some architectural models that can be used to interconnect NV03 networks via MPLS WAN networks.

When NV03 networks are connected by MPLS WAN networks, [EVPN-DCI] specifies how EVPN can be used end-to-end, in spite of using a different encapsulation in the WAN.

Even if EVPN can also be used in the WAN for Layer-2 and Layer-3 services, there may be a need to provide a Gateway function between EVPN for NV03 encapsulations and IPVPN for MPLS tunnels. [EVPN-IPVPN] specifies the interworking function between EVPN and IPVPN for

unicast Inter Subnet Forwarding. If Inter Subnet Multicast Forwarding is also needed across an IPVPN WAN, [OISM] describes the required interworking between EVPN and MVPN.

## [5. Conclusion](#)

EVPN provides a unified control-plane that solves the NVE auto-discovery, tenant MAP/IP dissemination and advanced features required by NV03 networks, in a scalable way and keeping the independence of the underlay IP Fabric, i.e. there is no need to enable PIM in the underlay network and maintain multicast states for tenant BDs.

This document justifies the use of EVPN for NV03 networks, discusses its applicability to basic Layer-2 and Layer-3 connectivity requirements, as well as advanced features such as MAC-mobility, MAC Protection and Loop Protection, multi-homing, DCI and much more.

## 6. Conventions used in this document

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](#) [[RFC2119](#)].

In this document, these words will appear with that interpretation only when in ALL CAPS. Lower case uses of these words are not to be interpreted as carrying [RFC-2119](#) significance.

In this document, the characters ">>" preceding an indented line(s) indicates a compliance requirement statement using the key words listed above. This convention aids reviewers in quickly identifying or finding the explicit compliance requirements of this RFC.

## 7. Security Considerations

This section will be added in future versions.

## 8. IANA Considerations

None.

## 9. References

Rabadan et al. Expires May 3, 2018 [Page 20]

---

Internet-Draft EVPN Applicability for NV03 October 30, 2017

### 9.1 Normative References

[RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", [RFC 7432](#), DOI 10.17487/RFC7432, February 2015, <<http://www.rfc-editor.org/info/rfc7432>>.

[RFC7365] Lasserre, M., Balus, F., Morin, T., Bitar, N., and Y. Rekhter, "Framework for Data Center (DC) Network Virtualization", [RFC 7365](#), DOI 10.17487/RFC7365, October 2014, <<http://www.rfc-editor.org/info/rfc7365>>.

[RFC7364] Narten, T., Ed., Gray, E., Ed., Black, D., Fang, L., Kreeger, L., and M. Napierala, "Problem Statement: Overlays for Network Virtualization", [RFC 7364](https://www.rfc-editor.org/info/rfc7364), DOI 10.17487/RFC7364, October 2014, <<http://www.rfc-editor.org/info/rfc7364>>.

## 9.2 Informative References

[IP-PREFIX] Rabadan et al., "IP Prefix Advertisement in EVPN", [draft-ietf-bess-evpn-prefix-advertisement-08](#), work in progress, October, 2017.

[INTER-SUBNET] Sajassi et al., "IP Inter-Subnet Forwarding in EVPN", [draft-ietf-bess-evpn-inter-subnet-forwarding-03](#), work in progress, February, 2017

[EVPN-USAGE] Rabadan et al., "Usage and applicability of BGP MPLS based Ethernet VPN", work in progress, [draft-ietf-bess-evpn-usage-06](#), August 2017

[EVPN-OVERLAY] Sajassi-Drake et al., "A Network Virtualization Overlay Solution using EVPN", work in progress, [draft-ietf-bess-evpn-overlay-08](#), March 2017

[GENEVE] Gross et al., "Geneve: Generic Network Virtualization Encapsulation", [draft-ietf-nvo3-geneve-05](#), work in progress, September 2017

[NV03-ENCAP] Boutros et al., "NV03 Encapsulation Considerations", [draft-ietf-nvo3-encap-01](#), work in progress, October 2017

[TUNNEL-ENCAP] Rosen et al., "The BGP Tunnel Encapsulation Attribute", [draft-ietf-idr-tunnel-encaps-03](#), work in progress, May 31, 2016.

Rabadan et al.

Expires May 3, 2018

[Page 21]

---

Internet-Draft

EVPN Applicability for NV03

October 30, 2017

[EVPN-LSP-PING] Jain et al., "LSP-Ping Mechanisms for EVPN and PBB-EVPN", [draft-jain-bess-evpn-lsp-ping-05](#), work in progress, July 2017

[LOOP] Rabadan et al., "Loop Protection in EVPN networks", [draft-snr-bess-evpn-loop-protect-00](#), work in progress, July 2017



[PROXY-ARP-ND] Rabadan et al., "Operational Aspects of Proxy-ARP/ND in EVPN Networks", [draft-ietf-bess-evpn-proxy-arp-nd-03](#), work in progress, October 2017

[IGMP-MLD-PROXY] Sajassi et al., "IGMP and MLD Proxy for EVPN", [draft-ietf-bess-evpn-igmp-mld-proxy-00](#), work in progress, March 2017

[PIM-PROXY] Rabadan et al., "PIM Proxy in EVPN Networks", [draft-skr-bess-evpn-pim-proxy-01](#), work in progress, October 2017

[OPT-IR] Rabadan et al., "Optimized Ingress Replication solution for EVPN", [draft-ietf-bess-evpn-optimized-ir-02](#), work in progress, August 2017

[HRW-DF] Mohanty et al., "A new Designated Forwarder Election for the EVPN", [draft-ietf-bess-evpn-df-election-03](#), work in progress, October 2017

[PREF-DF] Rabadan et al., "Preference-based EVPN DF Election", [draft-ietf-bess-evpn-pref-df-00](#), work in progress, June 2017

[AC-DF] Rabadan et al., "AC-Influenced Designated Forwarder Election for EVPN", [draft-ietf-bess-evpn-ac-df-02](#), work in progress, October 2017

[OISM] Lin et al., "EVPN Optimized Inter-Subnet Multicast (OISM) Forwarding", [draft-lin-bess-evpn-irb-mcast-04](#), work in progress, October 2017

[EVPN-DCI] Rabadan et al., "Interconnect Solution for EVPN Overlay networks", [draft-ietf-bess-dci-evpn-overlay-05](#), work in progress, July 2017

[BUM-UPDATE] Zhang et al., "Updates on EVPN BUM Procedures", [draft-ietf-bess-evpn-bum-procedure-updates-02](#), work in progress, September 2017

[EVPN-IPVPN] Rabadan-Sajassi et al., "EVPN Interworking with IPVPN", [draft-rabadan-sajassi-bess-evpn-ipvpn-interworking-00](#), work in progress, October 2017

[RFC7348] Mahalingam, M., Dutt, D., Duda, K., Agarwal, P., Kreeger,

L., Sridhar, T., Bursell, M., and C. Wright, "Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", [RFC 7348](#), DOI 10.17487/RFC7348, August 2014, <<http://www.rfc-editor.org/info/rfc7348>>.

[RFC7510] Xu, X., Sheth, N., Yong, L., Callon, R., and D. Black, "Encapsulating MPLS in UDP", [RFC 7510](#), DOI 10.17487/RFC7510, April 2015, <<http://www.rfc-editor.org/info/rfc7510>>.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", [RFC 4364](#), DOI 10.17487/RFC4364, February 2006, <<http://www.rfc-editor.org/info/rfc4364>>.

## [10](#). Acknowledgments

## [11](#). Contributors

## [12](#). Authors' Addresses

Jorge Rabadan  
Nokia  
777 E. Middlefield Road  
Mountain View, CA 94043 USA  
Email: [jorge.rabadan@nokia.com](mailto:jorge.rabadan@nokia.com)

Sami Boutros  
VMware  
Email: [sboutros@vmware.com](mailto:sboutros@vmware.com)

Matthew Bocci  
Nokia  
Email: [matthew.bocci@nokia.com](mailto:matthew.bocci@nokia.com)

