NVO3 Workgroup
Internet Draft
Intended status: Informational

J. Rabadan, Ed.
M. Bocci
Nokia

S. Boutros
WMware

A. Sajassi
Cisco

Expires: August 13, 2018

February 9, 2018

Applicability of EVPN to NVO3 Networks
draft-rabadan-nvo3-evpn-applicability-01

Abstract

In NVO3 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 NVO3 networks and how EVPN solves the challenges in those networks.

This Internet-Draft will expire on August 13, 2018.

Table of Contents

1. Introduction

In NVO3 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 NVO3 Terminology

o EVPN: Ethernet Virtual Private Networks, as described in [RFC7432].

- o PE: Provider Edge router.

- o NVO3 or Overlay tunnels: Network Virtualization Over Layer-3 tunnels. In this document, NVO3 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 NVO3 tunnel encapsulations are VXLAN [RFC7348], [GENEVE] or MPLSoUDP [RFC7510].

- o VXLAN: Virtual eXtensible Local Area Network, an NVO3 encapsulation defined in [RFC7348].

- o GENEVE: Generic Network Virtualization Encapsulation, an NVO3 encapsulation defined in [GENEVE].

- o CLOS: a multistage network topology described in [CLOS1953], where all the edge switches (or Leafs) are connected to all the core switches (or Spines). Typically used in Data Centers nowadays.

- o ECMP: Equal Cost Multi-Path.

- 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-Distinghisher (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-distinghisher (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 NVO3 network and the ESI is used in EVPN routes that are specific to that ES.

- o DF and NDF: they refer to Designated Forwarder and Non-Designated Forwarder, which are the roles that a given PE can have in a given ES.

- o VNI: Virtual Network Identifier. Irrespective of the NVO3 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 NVO3 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.

- o RT and RD: they refer to Route Target and Route Distinguisher.

- o PTA: Provider Multicast Service Interface Tunnel Attribute.

- o RT-1, RT-2, RT-3, etc.: they refer to Route Type followed by the type number as defined in the IANA registry for EVPN route types.

- o TS: Tenant System.

- o ARP and ND: they refer to Address Resolution Protocol and Neighbor Discovery protocol.

3. Why Is EVPN Needed In NVO3 Networks?

Data Centers have adopted NVO3 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 NVO3 tunnels, being VXLAN [RFC7348] or [GENEVE] two examples of such tunnels.

Why is a control-plane protocol along with NVO3 tunnels required?

There are three main reasons:

a) Auto-discovery of the remote NVEs that are attached to the same VPN instance (Layer-2 and/or Layer-3) 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 and ARP/ND traffic reduction/suppression, Multi-homing, Prefix Independent Convergence (PIC) like functionality, Fast Convergence, etc.

A possible approach to achieve points (a) and (b) above for multipoint Ethernet services, 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 NVO3 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 NVO3 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 a BUM frame for the BD on an ingress AC, 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 NVO3 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 NVO3 network.

4. Applicability of EVPN to NVO3 Networks

This section discusses the applicability of EVPN to NVO3 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 NVO3 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: ES auto-discovery and 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 NVO3 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 NVO3 Networks.

Figure 1 represents a Layer-2 service deployed with an EVPN BD in an NVO3 network.

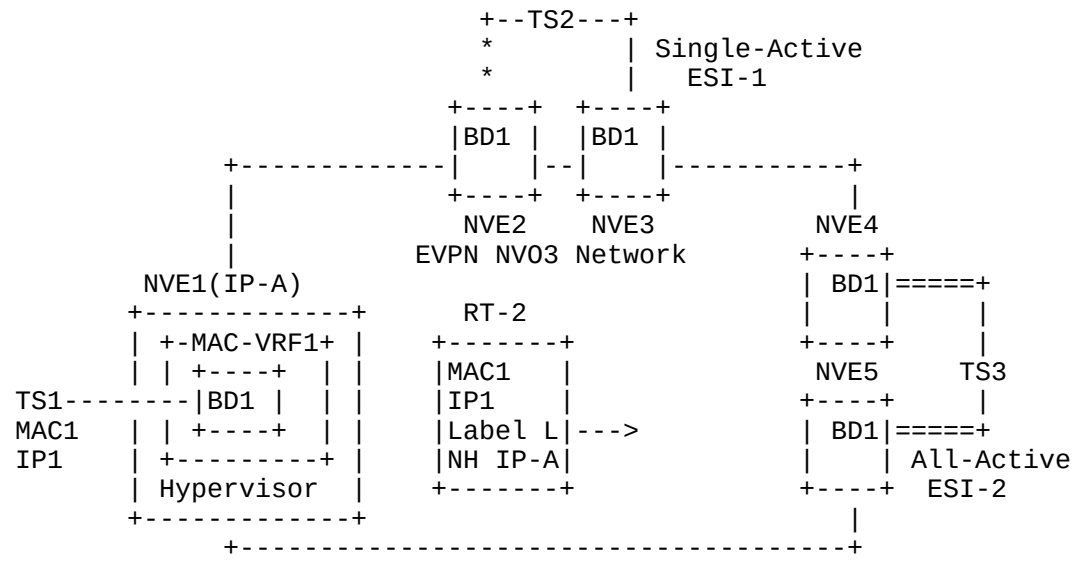

Figure 1 EVPN for L2 in an NVO3 Network - example

In a simple NVO3 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-Discovery and Auto-Provisioning of ES, Multi-Homing PEs and NVE services

Auto-discovery is one of the basic capabilities of EVPN. The provisioning of EVPN components in NVEs is significantly automated, simplifying the deployment of services and minimizing manual operations that are prone to human error.

These are some of the Auto-Discovery and Auto-Provisioning capabilities available in EVPN:

- o Automation on Ethernet Segments (ES): an ES is defined as a group of NVEs that are attached to the same TS or network. An ES is identified by an Ethernet Segment Identifier (ESI) in the control plane, but neither the ESI nor the NVEs that share the same ES are required to be manually provisioned in the local NVE:

  - If the multi-homed TS or network are running protocols such as LACP (Link Aggregation Control Protocol), MSTP (Multiple-instance Spanning Tree Protocol), G.8032, etc. and all the NVEs in the ES can listen to the protocol PDUs to uniquely identify the multi-homed TS/network, then the ESI can be "auto-sensed" or "auto-provisioned" following the guidelines in [RFC7432] section 5.

  - As described in [RFC7432], EVPN can also auto-derive the BGP parameters required to advertise the presence of a local ES in the control plane (RT and RD). Local ESes are advertised using RT-4s and the ESI-import Route-Target used by RT-4s can be auto-derived based on the procedures of [RFC7432], section 7.6.

  - By listening to other RT-4s that match the local ESI and import RT, an NVE can also auto-discover the other NVEs participating in the multi-homing for the ES.

  - Once the NVE has auto-discovered all the NVEs attached to the same ES, the NVE can automatically perform the DF Election algorithm (which determines the NVE that will forward traffic to the multi-homed TS/network). EVPN guarantees that all the NVEs in the ES have a consistent DF Election.

- o Auto-provisioning of services: when deploying a Layer-2 Service for a tenant in an NVO3 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-VRF). EVPN allows auto-deriving as many MAC-VRF parameters as possible. As an example, 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.

4.2.2. Remote NVE Auto-Discovery

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 NVO3 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 NVO3 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:

- 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.

- 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.

- 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 NVO3 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.

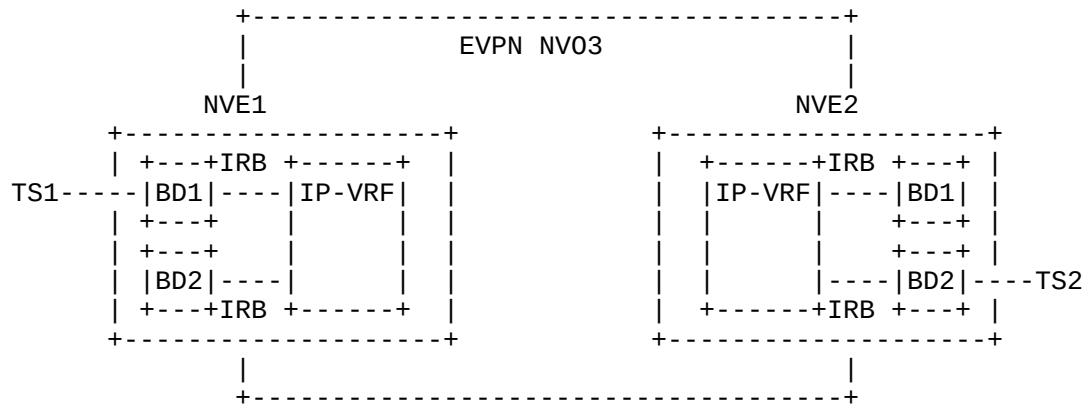

Figure 2 EVPN for L3 in an NVO3 Network - Asymmetric model

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.

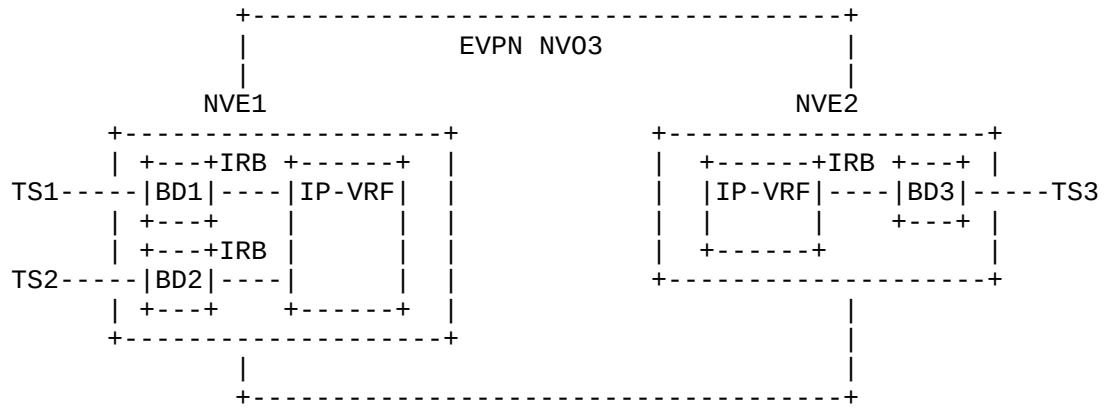

Figure 3 EVPN for L3 in an NVO3 Network - Symmetric model

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 NVO3 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 NVO3 Encapsulations and GENEVE

[EVPN-OVERLAY] describes how to use EVPN for NVO3 encapsulations, such us VXLAN, nvGRE or MPLSoGRE. The procedures can be easily applicable to any other NVO3 encapsulation, in particular GENEVE.

The NVO3 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 NVO3 Encapsulation. The EVPN control plane can signal the GENEVE encapsulation type in the BGP Tunnel Encapsulation Extended Community (see [TUNNEL-ENCAP]).

The NVO3 encapsulation design team has made a recommendation in [NVO3-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. [EVPN-GENEVE] describes the EVPN control plane extensions to support GENEVE.

4.5. EVPN OAM and application to NVO3

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 NVO3 networks.

4.6. EVPN as the control plane for NVO3 security

EVPN can be used to signal the security protection capabilities of a sender NVE, as well as what portion of an NVO3 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 NVO3 tunnels.

4.7. Advanced EVPN Features For NVO3 Networks

This section describes how EVPN can be used to deliver advanced capabilities in NVO3 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 NVO3 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) NVO3 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 Broadcast, Multicast or Unknown unicast to certain NVEs that don’t 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:

- 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.

- All-active multi-homing means per-flow load-balanding 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:

- 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.

- 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 NVO3 encapsulations, therefore [EVPN-OVERLAY] defines a modified Split-horizon procedure that is based on the IP SA of the NVO3 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 NVO3 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, IP1 is advertised in an RT-2 associated to M1, VNI-S and BGP next-hop NVE1. Upon importing the two routes, NVE2 installs TS2-IP/L in the IP-VRF with a next-hop that is the GWIP IP1. NVE2 also installs M1 in the SBD, with VNI-S and NVE1 as next-hop. If TS3 sends a packet with IP DA=TS2, NVE2 will perform a Recursive Resolution of the RT-5 prefix information to the forwarding information of the correlated RT-2. The RT-5’s Recursive Resolution has several advantages such as better convergence in scaled networks (since multiple RT-5s can be invalidated with a single withdrawal of the overlay index route) or the ability to advertise multiple RT-5s from an overlay index that can move or change dynamically. [EVPN-PREFIX] describes a few use-cases.

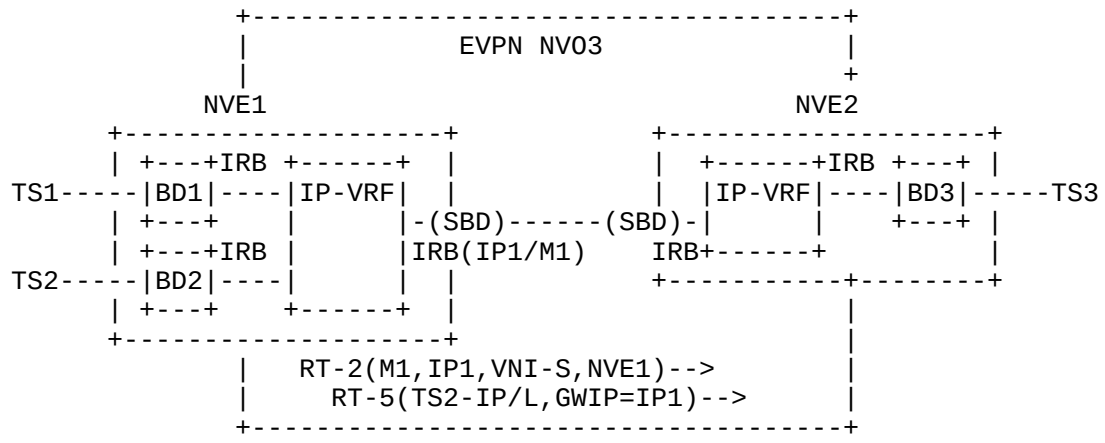

Figure 4 EVPN for L3 - Recursive Resolution example

4.7.7. EVPN Optimized Inter-Subnet Multicast Forwarding

The concept of the SBD described in section 4.7.6 is also used in [OISM] for the procedures related to Inter Subnet Multicast Forwarding across BDs of the same tenant. For instance, [OISM] allows 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 NVO3 networks, including IR or AR. [OISM] also describes the interoperability between EVPN and other multicast technologies such as MVPN (Multicast VPN) and PIM for inter-subnet multicast.

[EVPN-MVPN] describes another potential solution to support EVPN to MVPN interoperability.

4.7.8. Data Center Interconnect (DCI)

Tenant Layer-2 and Layer-3 services deployed on NVO3 networks must be extended to remote NVO3 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 NVO3 networks via MPLS WAN networks.

When NVO3 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 NVO3 encapsulations and IPVPN for MPLS tunnels. [EVPN-IPVPN] specifics 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 NVO3 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 NVO3 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", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

7. Security Considerations

This section will be added in future versions.

8. IANA Considerations

None.

9. References

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, DOI 10.17487/RFC7364, October 2014, <http://www.rfc-editor.org/info/rfc7364>.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

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

[NVO3-ENCAP] Boutros et al., "NVO3 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.

[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 at 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>.

[CLOS1953] Clos, C., "A Study of Non-Blocking Switching Networks", The Bell System Technical Journal, Vol. 32(2), DOI 10.1002/j.1538-7305.1953.tb01433.x, March 1953.

[EVPN-GENEVE] Boutros et al., "EVPN control plane for Geneve", draft-boutros-bess-evpn-geneve-01, work in progress, February 2018.

[EVPN-MVPN] Sajassi et al., "Seamless Multicast Interoperability between EVPN and MVPN PEs", draft-sajassi-bess-evpn-mvpn-seamless-interop-00, work in progress, July 2017.

10. Acknowledgments

11. Contributors

12. Authors' Addresses

Jorge Rabadan (Editor)
Nokia
777 E. Middlefield Road
Mountain View, CA 94043 USA

Email: jorge.rabadan@nokia.com

Sami Boutros
VMware
Email: sboutros@vmware.com

Matthew Bocci
Nokia
Email: matthew.bocci@nokia.com

Ali Sajassi
Cisco
Email: sajassi@cisco.com