L2VPN Workgroup
INTERNET-DRAFT
Intended Status: Standards Track

Ali Sajassi
Samer Salam
Keyur Patel
Cisco

Wim Henderickx
Alcatel-Lucent

Nabil Bitar
Verizon

John Drake
Yakov Rakhter
Juniper

Aldrin Isaac
Bloomberg

Jim Uttaro
AT&T

Expires: April 22, 2012 October 22, 2012

E-VPN Seamless Interoperability with IP-VPN
draft-sajassi-l2vpn-evpn-ipvpn-interop-01

Abstract

E-VPN can be an integral part of an Integrated Routing and Bridging (IRB) solution which is capable of performing optimum unicast and multicast forwarding not just for L2 traffic but also for L3 traffic. This document describes how an IRB solution based on E-VPN can interoperate seamlessly with the IP-VPN solution over MPLS and IP networks.

Status of this Memo

Table of Contents

1 Introduction

E-VPN can be an integral part of an Integrated Routing and Bridging (IRB) solution which is capable of performing optimum unicast and multicast forwarding not just for L2 traffic (intra-subnet forwarding), as described in the baseline draft [E-VPN], but also is capable of performing optimum unicast and multicast forwarding for L3 traffic (inter-subnet forwarding) as described in [DC-MOBILITY].

Such IRB capability is of high relevance in data center applications where performing either L2 or L3 forwarding alone may not be sufficient.

1.1 Shortcomings of L2-Only Solution

Figure-1 depicts a Data Center Network (DCN) using IP overlay where the PE functionality (and IP tunnel encapsulation) are either residing on physical Top of Rack (ToR) switches or on virtual hypervisor-based switches. In this document, we refer to these PE devices (either physical or virtual) that provide IP overlay tunneling as Network Virtualized Endpoints (NVEs). The DCN is connected to the Internet and/or enterprise/SP MPLS/IP core network via gateway (GW) nodes.

```
                                   ,---------.
                                 ,'           `.
                                (    IP/MPLS    )
                                 `.           ,'
                                   `-+------+'
                                +--+--+   +-+---+
                                | GW  |+-+| GW  |
                                +-+---+   +-----+
                                   /
                            +----+---+   +---+-----+
                            | Core   |   |  Core   |
                            | IP SW  |   | IP SW   |
                            +-+----`.+   +-+---+---+
                               /         .'
                         +---+--+   +-`.+--+  +--+----+
                         | ToR  |   | ToR  |  |  ToR  |
                         +-+--`.+   +-+-`.-+  +-+--+--+
                          /        /        /
                       __/_       /        /_     ____
                      :VSw :    :VSw :     :VSw :   :VSw :
                      '----'    '----'     '----'   '----'
                         |         |          |        |
                        VM1       VM2        VM3      VM4
```

Figure 1: A typical DC network

If Network Virtualization Endpoints (NVEs) were only to provide L2 service (and forwarding), then for two VMs on two different subnets, which need to communicate with each other, their packets need to be forwarded to a router (either physical or virtual). In the above diagram, the packets from the VMs need to be forwarded all the way to one of the GW devices to perform L3 forwarding. This is generally sub-optimal because the two VMs may be connected to the same virtual switch or the same TOR where L3 switching could have been performed locally. Even if the two VMs are located in different PODs within the same DC, and the traffic between the two VMs requires transitioning a core switch, adding a GW for L3 switching adds additional hops to the data path. However, if an NVE has IRB capability, then it can perform optimum L2 forwarding for intra-subnet traffic and optimum L3 forwarding for inter-subnet traffic, delivering optimum forwarding of unicast and multicast packets at all time.

1.2 Shortcomings of L3-Only Solution

Consider the scenario where a server is multi-homed to several ToR devices using an Ethernet Link Aggregation Group with LACP [802.1AX]

and the VMs are connected to a virtual bridge on the server - i.e., there is an Ethernet bridge on the data path between the VMs and the TORs. The ToRs are acting as NVEs. In this scenario, the LAG spans across multiple PE devices (NVEs) and IGMP joins for the same multicast group can arrives at both PEs. As such, DF election and split-horizon filtering functions are required on the ToRs belonging to the same LAG in order to avoid loops and packet duplication. However, the existing IP-VPN solution does not provide such capabilities that are available in the E-VPN solution. Therefore, these ToR devices cannot be simple L3VPN PEs.

Assuming that the above shortcoming is addressed by adding DF election and split-horizon filtering to IP-VPN, several other issues will continue to exist with L3-only solution, particularly when attempting to rely on L3 forwarding for intra-subnet traffic:

1. With L3 forwarding, in the absence of a default route, unknown IP destination addresses are dropped. Furthermore, an IP default route directs a particular traffic flow to a single next-hop or outbound interface. This means that L3 forwarding cannot support the forwarding semantics of a subnet broadcast.

2. With L3 forwarding, the MAC header is link-local and MAC addresses are swapped on a hop-by-hop basis. This means that if an NVE resorts to L3 forwarding of intra-subnet traffic, then all hosts within the same subnet will receive traffic with the source MAC addresses set to the NVE's address(es) instead of the originating hosts' MAC addresses. As a result, any higher layer application which relies on the source MAC address for identifying the communicating endpoint will break, as it will no longer be able to tell apart the hosts within the subnet based on their MAC addresses. This essentially creates an address aliasing problem. A related issue, that results from the MAC address being rewritten by the NVE, is that the hosts can no longer perform duplicate MAC address detection.

3. With L3 forwarding, the IP TTL is decremented with every routed hop. Some applications rely on this fundamental behavior to confine traffic to the originating subnet, by setting the TTL to 1 on transmission. Such applications will no longer work when intra-subnet traffic is L3 forwarded.

4. IPv6 link-local addressing and duplicate address detection [RFC4862] assumes and relies upon L2 connectivity within the subnet. These mechanisms will break if the NVE performs L3 intra-subnet forwarding.

5. Finally last but not least, there are non-IP applications that require L2 forwarding or there are applications that rely on end host

MAC addresses.

1.3 Combined L2 & L3 Solution: IRB

An IRB solution based on E-VPN can address the shortcomings of L2-only as well as L3-only solutions, and provide optimum forwarding for both inter and intra subnet switching, not only within a DCN but across different DCNs. This E-VPN based solution fits well for DCN overlay and DCI applications, but typical deployments will include IP-VPN PEs that E-VPN PEs need to inter-operate with, such as:

1) IP-VPN client sites accessing cloud services
2) Communication with IP-VPN ToRs/VSw
3) Communication with IP-VPN GWs

Therefore, interoperability with IP-VPN PEs is of paramount importance.

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 [RFC2119].

2 Seamless Interoperability with IP-VPN PEs

2.1 Interoperability Use-Cases

There are three use-cases that require interoperability between E-VPN and IP-VPN. Those are discussed next.

```
                             +---+    Enterprise Site 1
                             |PE1|----- H1
                             +---+
                               /
                         ,---------.             Enterprise Site 2
                       ,'           `.    +---+
                      (    MPLS/IP    )---|PE2|-----  H2
                       `.   Core    ,'    +---+
                         `-+------+'
                        / /      \ \
                      +---+       \ \
                 ,----|GW |.       \ \
               ,'     +---+ `.      ,---------.
              (      DCN 1    )   ,'           `.
               `.           ,'   (      DCN2     )
                 `-+------+'      `.           ,'
                   __/__            `-+------+'
                  :NVE1 :           __/__   __\__
                  '-----'          :NVE2 :  :NVE3 :
                   |  |            '-----'  '-----'
                  VM1 VM             |  |      |
                                    VM3 VM4   VM5
```

Figure 2: Interoperability Use-Cases

2.1.1 IP-VPN Clients Access to Cloud Services

An SP offering IP-VPN services to an enterprise may wish to expand its service offering to include Cloud services, while leveraging its existing MPLS/IP infrastructure. The SP may deploy E-VPN on the NVE in order to support L2 connectivity between VMs. If the number of VPNs and routes per VPN is not high, the SP may extend the L3 edge to the NVE and implement that function on the ToR switch. An alternative would be to have the NVE be on the hypervisor, in higher scale scenarios. Either way, distributing the L3 edge to the NVE renders it possible to avoid having an IP-VPN GW for the DCN. For this scenario, interoperability between the E-VPN NVE and IP-VPN PE is required in order to enable the new service offering.

For e.g., consider Figure 1 where an IP-VPN service is being offered between Enterprise sites 1 and 2. PE1 and PE2 act as IP-VPN PEs. Furthermore, assume that DCN2 employ E-VPN (i.e. NVE2 and NVE3 are E-VPN PEs). For the SP to offer Cloud service, interoperability between the IP-VPN PEs and E-VPN NVEs is required.

2.1.2 Communication with IP-VPN NVEs

In certain deployments, where only L3 connectivity is required by

certain hosts (e.g. VMs), the NVEs associated with those hosts may employ IP-VPN functionality only. An example of this would be running the IP-VPN PE functionality on the hypervisor using the mechanisms of [L3VPN-ENDSYSTEM]. Other VMs may require both L2 as well as L3 connectivity. The NVEs associated with those latter VMs would employ E-VPN. In order to allow for inter subnet communication between both categories of VMs (i.e. those which require L3 connectivity only and those requiring both L2 as well as L3 connectivity), interoperability is required between the IP-VPN and the E-VPN NVEs.

To illustrate this with an example, consider the network of Figure 1. VM5 requires L3 connectivity only, and subsequently NVE3 employs IP-VPN PE functionality solely. VM3 requires both L2 and L3 connectivity, hence, NVE2 is employing E-VPN PE functionality. For VM3 to be able to optimally communicate with VM5, seamless interoperability between IP-VPN and E-VPN is required.

2.1.3 Communication with IP-VPN GWs

The DCN may include an IP-VPN GW in order to confine the routing tables of the NVEs to L3 routes that are local to the DCN. The NVEs, in this case, would have default routes pointing to the GW. When the NVEs need to provide L2 as well as L3 connectivity to the associated VMs, they must run E-VPN PE functionality. In order for the IP-VPN GW to learn reachability to the VMs local to the DCN, interoperability is required between E-VPN NVEs and the IP-VPN GW.

As an example, consider the network of Figure 1 where the GW of DCN1 is an IP-VPN gateway. If NVE1 employs E-VPN PE functionality, then interoperability between E-VPN and IP-VPN is required for connectivity between NVE1 and the GW.

2.2 Characteristics of Seamless Interoperability

Seamless interoperability between E-VPN and IP-VPN must meet the following characteristics:

- Be completely transparent to the operation of the IP-VPN PE. In other words, the IP-VPN PE would not even be aware that it is communicating with an E-VPN endpoint. As such, no upgrade to the IP-VPN nodes is required, not even a software upgrade.

- Be optimal from data-plane forwarding perspective. This means that a gateway function is not required in order to normalize the encapsulation to Ethernet in order to support the interoperability. To elaborate on this: it is always possible to have an E-VPN PE interoperate with an IP-VPN PE using a normalized Ethernet L2 hand-off between the two. This however, requires that the MPLS

encapsulation be terminated on each PE, with the added overhead of unnecessarily performing MPLS imposition and disposition on both PEs. A side-effect of this gateway approach is that the host MAC addresses will be visible to the E-VPN, and this may create scalability bottlenecks, especially in virtualized data center environments because of sheer number of host MAC addresses.

3 An IRB Solution Based on E-VPN

An IRB solution based on E-VPN can meet data center network requirements in terms of:

- Providing optimal forwarding for intra-subnet (L2) traffic.

- Providing optimal forwarding for inter-subnet (L3) traffic, by avoiding the need for a centralized L3 GW. This is because the E-VPN MAC Advertisement route can carry an IP address in addition to the MAC address.

- Support for light-weight multicast using ingress replication, in cases where multicast applications are not required or dominant.

- Support for optimal multicast delivery through P2MP tunnels, when required, to optimize DCN resources.

- Support for multi-homing with active/active redundancy and per-flow load-balancing using multi-chassis LAG.

- Support for network-based as well as host-based overlay models.

- Support for consistent policy-based forwarding for both L2 and L3 forwarded traffic.

3.1 E-VPN PE Model for Seamless Interoperability

This section describes the PE data-plane model required to achieve seamless interoperability.

The E-VPN PE establishes a many-to-one mapping between EVIs and a VRF. For a given EVI, it is possible to have multiple associated bridge-domains using the VLAN-aware bundling service interface, as defined in [EVPN-REQ]. Each bridge-domain maps to a unique IP subnet within a VRF context. The following figure depicts the model where there are N VRFs corresponding to N tenants, with each tenant having 2 EVIs and up to M subnets (bridge domains) per EVI.

Note that this PE model provides flexibility for a wide gamut of deployment options. For example, one end of the spectrum would be

with a single EVI per tenant being mapped to a single VRF. Each EVI hosts multiple bridge-domains (one bridge-domain per subnet). This model allows for L2 traffic segregation between different subnets in addition to L3 connectivity among those subnets, as long as global Service VLANs are assigned per tenant (this uses VLAN-aware bundling service in E-VPN). The other end of the spectrum is with multiple EVIs per tenant all mapped to a single VRF. Each EVI hosts a single bridge-domain in this latter case. This model allows for L2 traffic segregation between subnets in addition to L3 connectivity among those subnets without the need for globally assigned Service VLANs (this uses VLAN-based service in E-VPN).

```
       +---------------------------------------------+
       |                                             |
       |      +-----------+         +-----------+    |
       |      |   EVIn    |---------|  VRF n    |    |
       |   +------------+ |     +------------+  |    |
       |   |  +-----+   | |     |            |  |    |
       |   |  | BD1 |   | |     |            |  |    |
       |   |  +-----+   | |     |    VRF 1   |  |    |
       |   |     ..     +-------+            |  |    |
       |   |  +-----+   | |     |            |  |    |
       |   |  | BDm |   | |     |            |  |    |
       |   |  +-----+   | |     |            |  |    |
       |   |   EVI 1    |-+     |            |  |    |
       |   +------------+       |            |  |    |
       |                        |            |  |    |
       |     +------------+     |            |  |    |
       |     |   EVIn*2   |     |            |  |    |
       |   +------------+ |     |            |  |    |
       |   |  +-----+   | |-----|            |  |    |
       |   |  |BDm+1|   | |     |            |  |    |
       |   |  +-----+   | |     |            |  |    |
       |   |     ..     +-------+            |  |    |
       |   |  +-----+   | |     |            |  |    |
       |   |  |BDm*2|   | |     |            |  |    |
       |   |  +-----+   | |     |            |  |    |
       |   |   EVI 2    |-+     |            |--+    |
       |   +------------+       +------------+       |
       |                                             |
       |                 E-VPN PE                    |
       +---------------------------------------------+
```

Figure 3: E-VPN PE Model for Seamless Interoperability with IP-VPN

One way to visualize this model is to consider a bridged virtual interface (BVI) to be associated with every bridge-domain in a given EVI. The BVI is an L3 routed interface (hence terminates L2). All the

BVIs associated with a given EVI are placed in the same VRF.

The IP forwarding table in a given VRF is shared in between E-VPN and IP-VPN. When an E-VPN MAC advertisement route is received by the PE, the MAC address associated with the route is used to populate the bridge-domain MAC table, whereas the IP address associated with the route is used to populate the corresponding VRF. For intra-subnet forwarding, the PE consults the bridge-domain MAC table whereas for inter-subnet forwarding the PE performs the lookup in the associated VRF.

When an E-VPN packet is received by a PE, it decapsulates the MPLS header and then performs a lookup on the destination MAC address. If the MAC address corresponds to one of its BVI interfaces, the PE deduces that the packet must be inter-subnet routed. Hence, the PE performs an IP lookup in the associated VRF table. However, if the destination MAC address does not correspond to a BVI, then the PE concludes that this packet needs to be intra-subnet switched, and no further IP lookup is needed.

3.2 IP-VPN BGP support on E-VPN PEs

The E-VPN PE learns host (e.g. VM) MAC addresses via normal bridge learning, and host IP addresses either via snooping of control traffic (e.g. ARP, DHCP..) or gleaning of data traffic. Once the PE learns a new MAC/IP address tuple, it advertises two routes for that tuple:

- An E-VPN MAC Advertisement route using the E-VPN AFI/SAFI and associated NLRI, which is used to advertise reachability to other remote E-VPN nodes. The MAC route advertises both the IP and MAC addresses of the host.

- An IP-VPN route using IP-VPN AFI/SAFI and associated NLRIs, which is used to advertise reachability to remote IP-VPN speakers. The IP-VPN route advertises only the IP address of the end-station.

Given that on the E-VPN PEs there is a one-to-one mapping between an E-VPN Instance (EVI) and a VRF, the same BGP RT and RD are used for both E-VPN and IP-VPN routes. Received E-VPN routes carry both IP and MAC addresses. The MAC addresses are injected into BD tables whereas the IP addresses are injected into VRFs. When an E-VPN speaker receives an IP-VPN route from a remote IP-VPN speaker, it installs the associated IP address in the appropriate VRF. It should be noted that when a MAC address is installed in the EVI, it is only installed in a single BD associated with the subnet corresponding to the Ethernet Tag encoded in the E-VPN MAC route.

If, for a given tenant, the IP-VPN PEs only need to share IP-VPN routes for a subset of the subnets with their E-VPN PEs counterparts, then one RT is used as a common RT between IP-VPN and E-VPN PEs for the common subnets and a different one or more RTs are used by the E-VPN PEs for the other tenant subnets that don't need to share routes with the IP-VPN PEs. If further topology constraint is needed among E-VPN and IP-VPN PEs, then instead of a common RT, one can use additional RTs to satisfy the topology constraint.

3.3 Handling Multi-Destination Traffic:

A key issue is how to handle multi-destination traffic, since E-VPN uses an MPLS label for split-horizon, and the equivalent does not exist in IP-VPN. This can be solved in two different ways, depending on whether the network uses LSM or Ingress Replication:

For LSM, two different sets of P2MP multicast trees can be used by the E-VPN PEs. One tree set encompasses only the IP-VPN endpoints whereas the second set includes only the E-VPN speakers. When an E-VPN PE receives a multi-destination frame, it sends a copy on each of the two trees associated with a given EVI/VRF. When the PE sends traffic on the IP-VPN tree, it does not include the split-horizon label since the IP-VPN endpoints do not understand this label. Note that this does not create any adverse side-effects because an E-VPN PE and an IP-VPN will never be combined in the same Redundancy Group (i.e. will never be multi-homed to the same Ethernet Segment), and as such the split-horizon filtering is never required on the IP-VPN PEs.

For ingress replication, the E-VPN PE sends the right label stack depending on the capability of the receiving (i.e. egress) PE. When replicating to IP-VPN endpoints, the ingress PE simply does not include any split-horizon labels.

3.2.1 Further optimization on RR

It is possible to optimize the number of routes that are advertised by a given E-VPN speaker for a specific host address, by leveraging extra intelligence on the BGP route reflector. A future version of this document will describe the detailed procedures to achieve this.

5 Acknowledgement

6 Security Considerations

7 IANA Considerations

8 References

8.1 Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

8.2 Informative References

[EVPN] Sajassi et al., "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-00.txt, work in progress, February, 2012.

[DC-MOBILITY] Aggarwal et al., "Data Center Mobility based on BGP/MPLS, IP Routing and NHRP", draft-raggarwa-data-center-mobility-03.txt, work in progress, June, 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

Aldrin Isaac
Bloomberg
aldrin.isaac@gmail.com

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

John E. Drake
Juniper Networks
Email: jnadeau@juniper.net

Yakov Rekhter
Juniper Networks
Email: yakov@juniper.net