Inter-AS Option B between NVO3 and BGP/MPLS IP VPN network

draft-hao-bess-inter-nvo3-vpn-02.txt

Abstract

This draft describes the solution of inter-as option-B connection between NVO3 network and MPLS/IP VPN network. The ASBR located in NVO3 network is called ASBR-d, the control plane and data plane procedures at ASBR-d are specified in this document, there are some differences from traditional option-B ASBR defined in [RFC 4364].

Status of this Memo

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

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

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."
The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt.

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

Copyright Notice

Copyright (c) 2015 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.

Table of Contents

1. Introduction ................................................ 3
2. Conventions used in this document............................ 3
3. Reference model ............................................. 5
4. Option-A inter-as solution overview......................... 6
5. Vanilla Option-B inter-as solution overview................ 6
6. Vanilla Inter-As Option-B procedures....................... 7
   6.1. Using BGP MPLS/IP VPN protocol........................ 7
       6.1.1. DC to WAN direction............................ 8
       6.1.2. WAN to DC direction............................ 9
   6.2. Data plane procedures.................................. 10
       6.2.1. DC to WAN direction............................ 10
       6.2.2. WAN to DC direction............................ 11
       6.2.3. Data plane NVE Operations summary............ 11
   6.3. NVE-NVA architecture................................... 11
       6.3.1. DC to WAN direction............................ 12
       6.3.2. WAN to DC direction............................ 13
7. Partial Option-B solution.................................. 13
8. Inter-as option comparisons................................ 13
9. Security Considerations.................................... 14
10. IANA Considerations....................................... 14
11. References ................................................ 15
    11.1. Normative References............................... 15
    11.2. Informative References............................. 15
12. Acknowledgments .......................................... 15
1. Introduction

In cloud computing era, multi-tenancy has become a core requirement for data centers. Since NVO3 can satisfy multi-tenancy key requirements, this technology is being deployed in an increasing number of cloud data center network. NVO3 focuses on the construction of overlay networks that operate over an IP (L3) underlay transport network. It can provide layer 2 bridging and layer 3 IP service for each tenant. VXLAN and NVGRE are two typical NVO3 technologies. NVO3 overlay network can be controlled through centralized NVE-NVA architecture or through distributed BGP VPN protocol.

NVO3 has good scaling properties from relatively small networks to networks with several million tenant systems (TSs) and hundreds of thousands of virtual networks within a single administrative domain. In NVO3 network, 24-bit VNID is used to identify different virtual networks in a data center. In a data center network, each tenant may include one or more layer 2 virtual network and in normal cases each tenant corresponds to one routing domain (RD). Normally each layer 2 virtual network corresponds to one or more subnets.

To provide cloud service to external data center client, data center networks should be connected with WAN networks. BGP MPLS/IP VPN has already been widely deployed at WAN networks. Normally internal data center and external MPLS/IP VPN network belongs to different autonomous system (AS). This requires the setting up of inter-as connections at Autonomous System Border Routers (ASBRs) between NVO3 network and external MPLS/IP network.

Currently, a typical connection mechanism between a data center network and an MPLS/IP VPN network is similar to Inter-AS Option-A of RFC4364, but it has scalability issue if there is huge number of tenants in data center networks. To overcome the issue, inter-as Option-B between NVO3 network and BGP MPLS/IP VPN network is proposed in this draft.

2. Conventions used in this document

Network Virtualization Edge (NVE) - An NVE is the network entity that sits at the edge of an underlay network and implements network virtualization functions.

Tenant System - A physical or virtual system that can play the role of a host, or a forwarding element such as a router, switch,
firewall, etc. It belongs to a single tenant and connects to one or more VNs of that tenant.

VN - A VN is a logical abstraction of a physical network that provides L2 network services to a set of Tenant Systems.

RD - Route Distinguisher. RDs are used to maintain uniqueness among identical routes in different VRFs. The route distinguisher is an 8-octet field prefixed to the customer’s IP address. The resulting 12-octet field is a unique "VPN-IPv4" address.

RT - Route targets. It is used to control the import and export of routes between different VRFs.
3. Reference model

Figure 1 Reference model

Figure 1 shows an arbitrary Multi-AS VPN interconnectivity scenario between NVO3 network and BGP MPLS/IP VPN network. NVE1, NVE2, and ASBR-d forms NVO3 overlay network in internal DC. TS1 and TS2 connect to NVE1, TS3 and TS4 connect to NVE2. PE1 and ASBR-w forms MPLS IP/VPN network in external DC. CE1 and CE2 connect to PE1. The NVO3 network belongs to AS 1, the MPLS/IP VPN network belongs to AS 2.
There are two tenants in NVO3 network, TSs in tenant 1 can freely communicate with CEs in VPN-Red, TSs in tenant 2 can freely communicate with CEs in VPN-Green. TS1 and TS3 belong to tenant 1, TS2 and TS4 belong to tenant 2. CE1 belongs to VPN-Red, CE2 belongs to VPN-Green. VNID 10 and VNID 20 are used to identify tenant 1 and tenant 2 respectively. CE1 and CE2 have local IP prefix of 10.1.1.1/24 and 20.1.1.1/24 respectively.

4. Option-A inter-as solution overview

In Option-A inter-as solution, peering ASBRs are connected by multiple sub-interfaces, each ASBR acts as a PE, and thinks that the other ASBR is a CE. Virtual routing and forwarding (VRF) data bases (RIB/FIB) are configured at AS border routers (ASBR-d and ASBR-w) so that each ASBRs associate each such sub-interface with a VRF and use EBGP to distribute unlabelled IPv4 addresses to each other. In the data-plane, VLANs are used for tenant traffic separation. ASBR-d terminates NVO3 encapsulation for inter-subnet traffic from TS in internal DC to CE in external DC.

Option-A inter-as solution has following issues:

1. Up to 16 million (16M) gateway interfaces (virtual/physical) and 16M EBGP session need to exist between the ASBRs.

2. Up to 16M VRFs need to be supported on border routers.

3. Several million routing entries need to be supported on border routers.

Inter-as option-B between NVO3 network and MPLS IP/VPN network can be used to address these issues. As option-B proposed in this draft is for multi-as interconnection between heterogeneous networks, so there are some differences from traditional Inter-AS Option-B of RFC4364.

5. Vanilla Option-B inter-as solution overview

Similar to the solution described in section 10, part (b) of [RFC4364] (commonly referred to as Option-B) peering ASBRs are connected as private peers that are enabled to receive Labeled packets from trusted peers. An MP-BGP session is used to distribute the labeled VPN prefixes between the ASBRs. In data plane, the traffic that flows between the ASBRs is placed in MPLS tunnels. Traffic separation among different VPNs between the ASBRs relies on MPLS VPN Label. The advantage of this option is that it’s more
scalable, as there is no need to have separate interface and BGP session per VPN/Tenant.

As for the routing distribution process from DC to WAN side, MPLS VPN Label is allocated on ASBR-d per VN per NVE. As for the routing distribution process from WAN to DC side, VNID is allocated on ASBR-d per MPLS VPN Label receiving from per ASBR-w. As for the data plane process, NVO3 tunnel and MPLS VPN tunnel are stitched at ASBR-d. From DC to WAN side, NVO3 tunnel is terminated, VNID and MPLS VPN Label switching is performed by looking up outgoing forwarding table in section 6.1.2. From WAN to DC side, MPLS VPN tunnel is terminated, MPLS VPN Label and NVO3 tunnel switching is performed by looking up incoming forwarding table in section 6.1.1. ASBR-w has no difference with traditional RFC4364 based Option-B behavior, no VRF is created on the ASBR-d.

6. Vanilla Inter-As Option-B procedures

Each NVE operates as a layer 3 gateway for local connecting TS(s). Operators may configure single and unique VNID for each tenant network on all NVEs or configure NVEs to locally allocate VNID for each tenant on the NVEs, the VNID is called VNID-t.

Routing information for each tenant should be synchronized between NVO3 and MPLS VPN network. In internal DC NVO3 network, routing information synchronization between NVE and ASBR-d can be through either: a) BGP MPLS/IP VPN protocol running between the NVEs and the ASBR-d or b) NVE-NVA architecture.

In case a), it is a coupled solution, the NVE entity normally resides on hardware network device like TOR switch. VRFs can be created on each NVE to isolate IP routing information in control plane and IP forwarding process in data plane between different tenants, each VRF has its own IP routing table. The BGP routes are originated on NVE with either implied nexthop address of the BGP router or self-nexthop set.

In case b), it is a decoupled solution, the NVE entity normally resides on vSwitch. VRFs are created on NVA only for control plane information isolation between different tenants, while in data plane, unified flow tables are used for all tenants on each NVE.

6.1. Using BGP MPLS/IP VPN protocol

Each NVE is a BGP speaker. Operators configure VRF and RD/RT for each tenant network on each NVE. BGP MPLS/IP VPN protocol extension is running between NVEs and ASBR-d utilizing the [BGP Remote-Next-
Hop] attribute which specifies a set of remote tunnels (1 to N) that occur between two BGP speakers.

When an NVE advertises a prefix with RD/RT, tunnel encapsulation and VNID-t are carried in BGP update message [BGP Remote-Next-Hop]. The NVE BGP receiver imports the prefix according RD/RT and maintains the mapping of prefix and VNID plus tunnel encapsulation (for VXLAN and NVGRE, they are outer destination IP address and inner destination MAC) in VRF.

[Note: the [BGP Remote-Next-Hop] is a work-in-progress that is an individual draft. The IDR WG may modify this draft or adopt another that provides a similar mechanism to support remote next-hops. This draft will follow the IDR adoption of a remote next-hop solution.]

6.1.1. DC to WAN direction

1. NVE1 and NVE2 operate as a layer 3 gateway for local connecting TSs. They learn the local TS’s IP Address via ARP or other mode and then advertise local TS’s IP Address with local NVE’s NVO3 tunnel end points information to ASBR-d using [BGP Remote-Next-Hop]. The routing information from NVE1 and NVE2 are as follows.

```
+---------+------------+-------+---------+--------+
<table>
<thead>
<tr>
<th>Node</th>
<th>IP Prefix</th>
<th>RD</th>
<th>RT</th>
<th>VNID-t</th>
</tr>
</thead>
<tbody>
<tr>
<td>NVE1</td>
<td>TS1/32</td>
<td>RD-A</td>
<td>RT-A</td>
<td>10</td>
</tr>
<tr>
<td>NVE1</td>
<td>TS2/32</td>
<td>RD-B</td>
<td>RT-B</td>
<td>20</td>
</tr>
<tr>
<td>NVE2</td>
<td>TS3/32</td>
<td>RD-A</td>
<td>RT-A</td>
<td>10</td>
</tr>
<tr>
<td>NVE2</td>
<td>TS4/32</td>
<td>RD-B</td>
<td>RT-B</td>
<td>20</td>
</tr>
</tbody>
</table>
+---------+------------+-------+---------+--------+
```

Table 1 Routing information from NVE

2. When ASBR-d receives routing information from each NVE, it allocates MPLS VPN Label per tenant (VNID-t) per NVE and the RD and RT remain the same (see table 2 below for examples). Then the ASBR-d advertises the VPN route with new allocated MPLS VPN Label to ASBR-w. The allocated MPLS VPN label and its corresponding <NVE, VNID-t> pair forms incoming forwarding table which is used to forward MPLS traffic from WAN to DC side. As an example the incoming forwarding table on ASBR-d could be as follows:
6.1.2. WAN to DC direction

1. When ASBR-d receives routing information from ASBR-w, ASBR-d allocates VNID-d for each VPN Label, and then ASBR-w advertises the VPN route with new allocated VNID-d to each NVE (NVE1 and NVE2). The role of the VNID-d is similar to the role of Incoming VPN Label in traditional MPLS VPN Option-B based ASBR defined in [RFC 4364], it has local significance on ASBR-d, each VNID corresponds to a MPLS VPN Label received from peer ASBR-w. The allocated VNID-d and its corresponding out VPN Label forms an outgoing forwarding table which is used to forward NVO3 traffic from DC to WAN side. Assuming ASBR-d receives VPN Label 3000 and 4000 from ASBR-w allocated for VPN-Red and VPN-Green at PE1 respectively, the outgoing forwarding table on ASBR-d is as follows:
2. When each local NVE receives routing information from ASBR-d, it matches the Route Target Attribute in BGP MPLS/IP VPN protocol with local VRF’s import RT configuration and populates local VRF with these matched VPN routes (see table 3 above).

6.2. Data plane procedures

This section describes the step by step procedures of data forward between TS1 and CE1 for either: a) DC to WAN direction IP data flows, or b) WAN to DC direction IP data flows.

6.2.1. DC to WAN direction

1. TS1 sends traffic to NVE1, the destination IP is CE1’s IP address.

2. NVE1 looks up tenant 1’s IP forwarding table, then it gets NVO3 tunnel encapsulation information. The destination outer address is ASBR-d’s IP address, VNID is 10000 allocated by ASBR-d for VPN route of CE1 received from ASBR-w. NVE1 performs NVO3 encapsulation and sends the traffic to ASBR-d.

3. ASBR-d decapsulates NVO3 encapsulation and gets VNID 10000. Then it looks up outgoing forwarding table based on the VNID and gets MPLS VPN label 3000. Finally it pushes MPLS VPN label for the IP traffic and sends it to ASBR-w.

4. Then the traffic is forwarded to CE1 through regular MPLS VPN forwarding process.
6.2.2. WAN to DC direction

1. CE1 sends traffic to PE1, destination IP is TS1’s IP address. The traffic is forwarded to ASBR-d through regular MPLS VPN forwarding process. The incoming MPLS VPN label at ASBR-d is 1000 allocated by ASBR-d for tenant 1 at NVE1.

2. ASBR-d looks up incoming forwarding table and gets NVO3 encapsulation, then performs NVO3 encapsulation and sends the traffic to NVE1. The destination outer IP is NVE1’s IP, VNID is 10 corresponding to tenant 1.

3. NVE1 decapsulates NVO3 encapsulation, gets local IP forwarding table relying on VNID 10, and then sends the traffic to TS1.

6.2.3. Data plane NVE Operations summary

Each NVE maintains a lookup table per tenant, i.e. VNID-t and the received mappings from ASBR-d for each tenant. For the prefix that is from inside DC, the inner/outer mapping entry is the prefix <-> remote NVE IP address. For the prefix that is from outside DC, the inner/outer mapping entry is the prefix <-> VNID-d + ASBR1-d IP address.

When receiving a packet from a tenant system locally, NVE performs a lookup in the corresponding tenant table for the destination address on the packet. If the prefix results to single IP address, NVE will encapsulate the packet with VNID-t and IP address as outer IP address. If the prefix results to a VNID and IP address, NVE will encapsulate the packet with the VNID and IP address as outer IP address.

When receiving a packet from NVO3, NVE decapsulates the packet and find the attached tenant system based on the VNID and destination address on the packet, forward the decapsulated packet to the tenant system.

6.3. NVE-NVA architecture

In this architecture, the NVE control plane and forwarding functionality are decoupled. All NVEs in NVO3 network don’t need support distributed BGP VPN protocol (BGP Remote-Next-Hop), these NVEs have only data plane functionality and are controlled by
centralized NVA using openflow, ovsdb, i2rs, etc. The NVA runs IBGP VPN protocol for all the NVEs with ASBR-d utilizing the [BGP Remote-Next-Hop] attribute to pass along the tunnel endpoints and encapsulations associated with each NVE. The ASBR-d runs EBGP VPN protocol with peer ASBR-w. ASBR-d allocates MPLS VPN Label per tenant per NVE.

NVA maintains all tenant information, and originates BGP routes with the appropriate RD and AD. The NVA tenant information includes VNID-t to identify each tenant and the corresponding RD and RT. This information can be statically configured by operators or dynamically notified by cloud management systems. This information also includes all TS’s MAC/IP address and its attached NVE information.

```
-------  IBGP  -------  EBGP  -------
|NVA   | ------- |ASBR-d| ----------|ASBR-w |
-------  -------            -------
            |Southbound interface (Openflow,OVSDB, I2RS)|
            .
            .
            .
-------  -------
|NVE1|   |NVE2|
-------  -------
```

Figure 2 NVE-NVA Architecture

6.3.1. DC to WAN direction

1. NVA advertises all internal data center VPN routing information to ASBR-d, which includes RD, RT, VNID-t, IP prefix and the attached NVE IP address. The VNID-t and NVE IP address are used for traffic NVO3 encapsulation from ASBR-d to NVE.

2. ASBR-d allocates MPLS VPN Label per VNID per NVE and generates incoming forwarding table same as Table 2.
6.3.2. WAN to DC direction

1. ASBR-d receives VPN routing information from peer ASBR-w. ASBR-d allocates VNID-d, for each MPLS VPN Label receiving from ASBR-w and generates outgoing forwarding table same as Table 4. Then it advertises the VPN route to NVA, which includes RD, RT, VNID-l, IP prefix, and set itself as next hop. The VNID and ASBR-d IP address are used for traffic NVO3 encapsulation from NVE to ASBR-d.

2. NVA matches local Route Target configuration, imports VPN route to each tenant, and downloads flow table to corresponding NVE.

7. Partial Option-B solution

In vanilla option-B solution, each NVE need to maintain routing items corresponding to IP prefix located outside data center for north-south bound traffic forwarding. If there are some VPNs which have large number of IP prefix, it will cause much pressure on local NVEs. In this case, partial Option-B solution can be used.

In partial Option-B solution, default route is used for north-south bound traffic on each NVE. The traffic from each NVE will be forwarded to ASBR-d using NVO3 encapsulation, VNID is used to identify tenant VRF at ASBR-d. ASBR-d terminates the NVO3 encapsulation, looks up local VRF’s IP routing table, then performs MPLS encapsulation and sends to peer ASBR-w.

For the traffic from WAN to DC, ASBR-d needs to maintain all TS’s IP addresses and their attached NVE device in corresponding VRF. When the ASBR-d receives MPLS traffic from peer ASBR-w, MPLS encapsulation is terminated, looks up local VRF’s IP routing table, then performs NVO3 encapsulation and sends to local destination NVE.

From control plane perspective, EBGP VPN connection is terminated at ASBR-d, which means the ASBR doesn’t allocate new VNID-d for each MPLS VPN Label and advertise it to peer NVE in local AS, VRF is created on the ASBR-d, the VPN route from WAN side populates to local VRF.

8. Inter-as option comparisons

The document describes several inter-as implementation options between ASBR-d and ASBR-w. The following table illustrates the comparison among the implementation options.
<table>
<thead>
<tr>
<th></th>
<th>Option-A</th>
<th>Partial Option-B</th>
<th>Vanilla Option-B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub-interface</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>VRF</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Scalability</td>
<td>Worst</td>
<td>Middle</td>
<td>Best</td>
</tr>
<tr>
<td>Hardware</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Implementation at ASBR-d</td>
<td>No Upgrade</td>
<td>No Upgrade</td>
<td>Need Upgrade</td>
</tr>
</tbody>
</table>

Table 5 Inter-as option comparisons

Option-A design uses a regular VPN handoff between ASBR-d and ASBR-w. A sub-interface is required per a NVO instance in between. Both border routers perform the VRF lookup. Thus, the solution has a scalability concern. Existing hardware supports this solution.

Partial Option-B does not require sub-interfaces between ASBR-d and ASBR-w, only ASBR-d performs the VRF lookup, so it has better scalability than option A. Existing hardware can support this solution.

In the vanilla Option-B solution, there is no sub-interface between border routers and no VRF table on ASBR-d and ASBR-w. Tunnel stitching is performed on the ASBR-d. Thus this solution has the best scalability. From hardware perspective, the vanilla option-B needs ASBR-d hardware upgrade to support the tunnel stitching.

9. Security Considerations

Similar to the security considerations for inter-as Option-B in [RFC4364] the appropriate trust relationship must exist between NVO3 network and MPLS/IP VPN network. VPN-IPv4 routes in NVO3 network should neither be distributed to nor accepted from the public Internet, or from any BGP peers that are not trusted. For other general VPN Security Considerations, see [RFC4364].

10. IANA Considerations

This document requires no IANA actions. RFC Editor: Please remove this section before publication.
11. References

11.1. Normative References


11.2. Informative References


12. Acknowledgments

Authors like to thank Xiaohu Xu, Liang Xia, Shunwan Zhang, Yizhou Li, Lili Wang for his valuable inputs.

Authors’ Addresses

Weiguo Hao
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China
Email: haoweiguo@huawei.com