Weiguo Hao
Lucy Yong
Huawei
S. Hares
Hickory Hill Consulting

Internet Draft
Intended status: Informational
Expires: January 2015
July 1, 2014

Inter-AS Option B between NVO3 and BGP/MPLS IP VPN network
draft-hao-l3vpn-inter-nvo3-vpn-00.txt

Status of this Memo

Copyright Notice

Abstract

This draft describes option-B inter-as connection between NVO3 network and MPLS/IP VPN network. Comparing to traditional Option-B inter-as connection defined in [RFC 4364], this draft provides enhancement for heterogeneous network multi-as connection, the control plane and data plane procedures in NVO3 network are newly designed.

Table of Contents

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 VN ID is used to identify different virtual networks, theoretically 16M virtual networks can be supported 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

```
|---------------------------------------------------|
|  ------           AS1                             |
|  | TS1| -                                         |
|  ------  -                                        |
|            - ------    ------                     |
|            - |NVE1| -- |TOR1|----------------     |
|  ------  -   ------    ------               |     |
|  | TS2|-                                    |     |
|  ------                                     |     |
|                                         ---------  |
|                           |------------ | ASBR1|  |
|  ------                   |             ---------  |
|  | TS3| -                 |                 |     |
|  ------  -                |                 |     |
|            - ------    ------               |     |
|            - |NVE2| -- |TOR2|               |     |
|  ------  -   ------    ------               |     |
|  | TS4|-                                    |     |
|  ------                                     |     |
|----------------------------------------------------
|                                             |     |
|  ------                                     |     |
|  | CE1| -                                   |     |
|  ------  -                                  |     |
|            - ------                     ---------  |
|            - | PE1| --------------------| ASBR2|  |
|   ------  -  ------                     ---------  |
|   | CE2|-                                         |
|   ------          AS2                             |
|---------------------------------------------------|
```

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 ASBR1 forms NVO3 overlay network in internal DC. TS1 and TS2 connect to NVE1, TS3 and TS4 connect to NVE2. PE1 and ASBR2 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.

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 (ASBR1 and ASBR2) so that each ASBRs associate each such sub-interface with a VRF and use EBGP to distribute unlabeled IPv4 addresses to each other. In the data-plane, VLANs are used for tenant traffic separation. In normal case ASBR1 also acts as NVO3 layer 3 gateway, it can terminate NVO3 encapsulation for inter-subnet traffic between TS in internal DC and CE in external DC.

For the traffic from internal DC to external DC, the forwarding process of ASBR1 at internal DC is as follows:

1. Terminates NVO3 encapsulation and gets VN ID.

2. Finds corresponding VRF relying on the VN ID.

3. Looks up IP forwarding table in the VRF, and then forwards the traffic to a sub-interface connecting to peer ASBR.

The forwarding process of ASBR2 at MPLS/VPN network is as follows:

1. Finds corresponding VRF based on sub-interface.

2. Looks up IP forwarding table in the VRF, encapsulates the traffic with MPLS VPN label, and then sends the traffic to MPLS VPN network.

For the traffic from external DC to internal DC, the traffic forwarding process is similar to the above process.

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. Due to it is for multi-as interconnection between heterogeneous networks, so there are some differences from traditional Inter-AS Option-B of RFC4364.

5. 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 by one or more sub-interfaces that are enabled to receive MPLS traffic. 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 upon MPLS tunnels, traffic separation among different VPNs between the ASBRs relies on MPLS VPN Label.

In this solution, the procedures in MPLS/IP VPN network are same as defined in [RFC4364], but the procedures in NVO3 network need to be newly designed to support inter-as Option-B.

The advantage of this option is that it’s more scalable, as there is no need to have one sub-interface and BGP session per VPN/Tenant.

6. Inter-As Option-B procedures

The TS and CE information in above figure 1 are as follows:

TS	Tenant	IP Address	VN ID
TS1	1	10.1.1.2	10
TS2	2	20.1.1.2	20
TS3	1	10.1.1.3	10
TS4	2	20.1.1.3	20

Table 1 TS information in NVO3 network

CE	Route Distinguisher	Route Target	IP Address	Mask
CE1	VPN-Red1	1:1	30.1.1.1	24
CE2	VPN-Green1	2:2	40.1.1.1	24

Table 2 CE information in MPLS/IP VPN network

Section 6.1 below describes the route distribution process for this option, and section 6.2 describes the data forwarding process.

NVO3 network can pass routing data for the NVEs (IP Address, VN ID) through either: a) RFC 4364 running between the NVEs and the ASBR1, or b) NVE-NVA architecture. Therefore, the routing distribution process is different for these two options. Section 6.1.1 describes the routing distribution procedures using RFC 4364 on NVO3 network, and section 6.1.2 describes the procedures using NVE-NVA architecture.

The Data plane process is same in these two cases.

6.1. Routing distribution procedures

6.1.1. Using RFC 4364

The route distribution in NVO3 network makes use of the BGP multiple tunnel identifiers [BGP Remote-Next-Hop] to create an RFC4364 Option-B solution. This section provides a step by step explanation of the process.

In internal DC network, VRF1 and VRF2 are created on NVE1 and NVE2 to isolate IP forwarding process between tenant 1 and tenant 2. Route distinguishers (RD) and RT are specified for each VRF on these

NVEs. BGP MPLS/IP VPN protocol extension is running between NVEs and ASBR1 utilizing the [BGP Remote-Next-Hop] which describes the BGP MPLS/IP VPN protocol extension detail to specify a set of remote tunnels (1 to N) that occur between two BGP speakers. The VRF configuration information on each NVE are as follows:

NVE	Tenant	Route Distinguisher	Route Target
NVE1	1	VPN-Red2	1:1
NVE1	2	VPN-Green2	2:2
NVE2	1	VPN-Red3	1:1
NVE2	2	VPN-Green3	2:2

6.1.1.1. Internal DC to external DC direction

1. NVE1 and NVE2 operate as a layer 3 gateway for local connecting TS. NVE1 and NVE2 learn the local TS's IP Address via ARP, and advertise this information to the ASBR1. The routing information from NVE1 and NVE2 are as follows:

NVE	RD:IP Prefix	Route Target	VN ID
NVE1	VPN-Red2:10.1.1.2/32	1:1	10
NVE1	VPN-Green2:20.1.1.2/32	2:2	20
NVE2	VPN-Red3:10.1.1.3/32	1:1	10
NVE2	VPN-Green3:20.1.1.3/32	2:2	20

Routing information sent form NVE1 and NVE2

2. ASBR1 allocates MPLS VPN Label per tenant (VN ID) per NVE and the RD and RT remain the same. Then the ASBR1 advertises the VPN route with new allocated MPLS VPN Label to ASBR2. The allocated MPLS VPN label and its corresponding NVE+VN ID forms incoming forwarding table which is used to forward MPLS traffic from external DC to internal DC. The incoming forwarding table on ASBR1 is as follows:

MPLS VPN Label	NVE + VN ID
1000	NVE1 + 10
2000	NVE1 + 20
1001	NVE2 + 10
2001	NVE2 + 20

Incoming forwarding table

3. ASBR2 allocates new local VPN label for each receiving VPN label from ASBR1 firstly, then ASBR2 advertises the VPN route with new allocated MPLS VPN Label to PE1. As for data plane forwarding process, the new local VPN label are in VPN label, the receiving VPN label from ASBR1 are out VPN label. The VPN label switch table on ASBR2 is as follows:

In VPN Labels	Out VPN Labels
1000	1000
2000	2000
1001	1001
2001	2001

VPN label switch table

4. PE1 matches the Route Target Attribute in BGP MPLS/IP VPN protocol with local VRF’s import RT configuration. Then it populates local VRF with these matched VPN routes. The routing tables of VPN-Red and VPN-Green are as follows:

IP Prefix	VPN Label	BGP Peer
10.1.1.2/32	1000	ASBR2
10.1.1.3/32	1001	ASBR2

Routing table in VPN-Red

IP Prefix	VPN Label	BGP Peer
20.1.1.2/32	2000	ASBR2
20.1.1.3/32	2001	ASBR2

Routing table in VPN-Green

6.1.1.2. External DC to internal DC direction

1. PE1 learns IP prefix from CE1 and CE2 and populates these IP prefix to local VRF. Then the PE allocates VPN Label for these IP prefix and announces VPN routing information with allocated VPN Label to ASBR1. The VPN routing information are as follows:

Route Target	IP Prefix	VPN Label
1:1	VPN-Red1:30.1.1.1/24	3000
2:2	VPN-Green1:40.1.1.1/24	4000

2. ASBR2 allocates new local VPN label for each receiving VPN label from PE1, then ASBR2 advertises the VPN route with new allocated MPLS VPN Label to ASBR1. As for data plane forwarding process, the new local VPN label are in VPN label, the receiving VPN label from ASBR1 are out VPN label. The VPN label switch table on ASBR2 is as follows:

In VPN Label	Out VPN Label
3000	3000
4000	4000

3. ASBR1 allocates VN ID for each VPN Label receiving from ASBR2, and then ASBR2 advertises the VPN route with new allocated VN ID to each NVE (NVE1 and NVE2). The role of the VN ID is similar to the role of In VPN Label in ASBR1, it has local significance on ASBR1, each VN ID corresponds to per MPLS VPN Label on peer ASBR2; The VN ID space should be assigned in beforehand and should be orthogonal to the VN ID space for tenant identification(for example, assuming ASBR1 has local connecting TSs of tenant 1 to tenant 100, VN ID 1 to 100 are allocated for these tenants, other VN ID other than 1 to 100 can be allocated for outgoing forwarding table purpose). The allocated VN ID and its corresponding out VPN Label forms an outgoing forwarding table which is used to forward NVO3 traffic from internal DC to external DC. The outgoing forwarding table on ASBR1 is as follows:

VN ID	Out VPN Label
10000	3000
10001	4000

Outgoing forwarding table

4. NVE1 matches the Route Target Attribute in BGP MPLS/IP VPN protocol with local VRF’s import RT configuration. Then it populates local VRF with these matched VPN routes. The routing tables of tenant 1 and tenant 2 are as follows:

IP Prefix	VN ID	Dest Outer IP
30.1.1.1/24	10000	ASBR1

Tenant 1 routing table on NVE1 and NVE2

IP Prefix	VN ID	Dest Outer IP
40.1.1.1/24	10001	ASBR1

Tenant 2 routing table on NVE1 and NVE2

6.1.2. NVE-NVA architecture

No distributed BGP VPN protocol (RFC4364) is running on all NVEs and ASBR1 in NVO3 network, NVEs and ASBR1 are controlled by centralized NVA. The NVA runs EBGP VPN protocol with peer ASBR2 and exchanges VPN routing information between NVO3 network and MPLS/IP VPN network.

NVA maintains tenant information collected from all tenants. This information includes VN ID 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.

NVA also maintains all TS’s MAC/IP address and its attached NVE information for each tenant.

```
                ------     EBGP     --------
                |NVA | ------------- |ASBR2 |
                ------               --------
                  .
                  . Southbound interface(Openflow,OVSDB,etc)
   .........................
   .          .           .
   .          .           .
   .          .           .
------     ------      -------
|NVE1|     |NVE2|      |ASBR1|
------     ------      -------
```

Figure 2 NVE-NVA Archetecture

6.1.2.1. Internal DC to external DC direction

1. NVA allocates MPLS VPN Label per tenant per NVE.

2. NVA advertises all internal data center VPN routing information to peer ASBR2, which includes RD, IP prefix, RT, and MPLS VPN Label.

3. NVA downloads incoming forwarding table to ASBR1.

6.1.2.2. External DC to internal DC direction

1. NVA receives VPN routing information from peer ASBR2.

2. NVA allocates VN ID for each MPLS VPN Label receiving from ASBR2.

3. NVA downloads outgoing forwarding table to ASBR1.

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

6.2. Data plane procedures

This section describes the step by step procedures of data forward for either: a) internal DC to external DC IP data flows, or b) the external DC to internal DC IP data flows.

6.2.1. Internal DC to external DC direction

1. TS1 sends traffic to NVE1, the destination IP is CE1’s IP address of 30.1.1.1.

2. NVE1 looks up VRF1’s IP forwarding table, then it gets NVO3 tunnel encapsulation information. The destination outer address is ASBR1’s IP address, VN ID is 10000. NVE1 performs NVO3 encapsulation and sends the traffic to ASBR1.

3. ASBR1 decapsulates NVO3 encapsulation and gets VN ID 10000. Then it looks up outgoing forwarding table based on the VN ID and gets MPLS VPN label 3000. Finally it pushes MPLS VPN label for the IP traffic and sends it to ASBR2.

4. ASBR2 swaps VPN Label, then sends the traffic to PE1 through IGP tunnel.

5. PE1 terminates IGP tunnel, pops MPLS VPN label 3000, looks up local IP forwarding table in VRF1, and then forwards the traffic to CE1.

6.2.2. External DC to internal DC direction

1. CE1 sends traffic to PE1, destination IP is TS1's IP address of 10.1.1.2.

2. PE1 looks up local IP forwarding table in VPN-RED, pushes MPLS VPN label 1000, then searches IGP tunnel, then the PE sends the traffic to ASBR2 through IGP tunnel.

3. ASBR2 terminates IGP tunnel, swaps MPLS VPN label, then sends the traffic to ASBR1.

4. ASBR1 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, VN ID is 10.

5. NVE1 decapsulates NVO3 encapsulation, gets local VRF1 relying on VN ID 10, looks up local IP forwarding table in VRF1, then sends the traffic to TS1.

7. 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].

8. IANA Considerations

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

9. References

9.1. Normative References

[1] [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate

Requirement Levels", BCP 14, RFC 2119, March 1997.

9.2. Informative References

[1] [RFC4364] E. Rosen, Y. Rekhter, " BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

[2] [NVA] D.Black, etc, "An Architecture for Overlay Networks (NVO3)", draft-ietf-nvo3-arch-01, February 14, 2014

[3] [BGP Remote-Next-Hop] G. Van de Velde,etc, "BGP Remote-Next-Hop", draft-vandevelde-idr-remote-next-hop-05, January, 2014

[4] [RFC7047] B. Pfaff, B. Davie,"The Open vSwitch Database Management Protocol", RFC 7047, December 2013

[5] [OpenFlow1.3]OpenFlow Switch Specification Version 1.3.0 (Wire Protocol 0x04). June 25, 2012. (https://www.opennetworking.org/images/stories/downloads/sdn-resources/onf-specifications/openflow/openflow-spec-v1.3.0.pdf)

10. Acknowledgments

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

Authors' Addresses

Weiguo Hao
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China
Phone: +86-25-56623144
Email: haoweiguo@huawei.com

Lucy Yong
Huawei Technologies
Phone: +1-918-808-1918
Email: lucy.yong@huawei.com

Susan Hares
Huawei Technologies
Phone: +1-734-604-0323
Email: shares@ndzh.com.