L3VPN Weiguo Hao
Lucy Yong
S. Hares
Internet Draft Huawei
R. Raszuk
L. Fang
Microsoft
Shahram Davari
Broadcom
Intended status: Informational September 17, 2014
Expires: March 2015

Inter-AS Option B between NVO3 and BGP/MPLS IP VPN network
draft-hao-l3vpn-inter-nvo3-vpn-01.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, they are different from traditional option-B ASBR defined in [RFC 4364].

This Internet-Draft will expire on March 17, 2015.

Copyright Notice

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|-                                     |     |
|  +----+                                      |     |
|                                         +--------+ |
|                           +------------ | ASBR-d| |
|  +----+                   |             +--------+ |
|  | TS3| -                 |                  |     |
|  +----+  -                |                  |     |
|            - +----+    +----+                |     |
|            - |NVE2| -- |TOR2|                |     |
|  +----+  -   +----+    +----+                |     |
|  | TS4|-                                     |     |
|  +----+                                      |     |
|-----------------------------------------------------
|                                              |     |
|  +----+                                      |     |
|  | CE1| -                                    |     |
|  +----+  -                                   |     |
|            - +----+                     +--------+ |
|            - | PE1| --------------------| ASBR-w| |
|   +----+  -  +----+                     +--------+ |
|   | 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 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. VN ID 10 and VN ID 20 are used to identify tenant1 and tenant2 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 unlabeled 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. Because 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. 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.

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, VN ID is allocated per MPLS VPN Label receiving from ASBR-w on ASBR-d. From data plane perspective, VN ID and MPLS VPN Label switching is performed on ASBR-d, ASBR-w has no difference with traditional RFC4364 based Option-B behavior, no VRF is created on the ASBR-d.

6. Inter-As Option-B procedures

Each NVE operates as default layer 3 gateway for local connecting TS(s). VRFs are created on each NVE to isolate IP forwarding process between different tenants. At least a L3 VN ID is used to identify each tenant.

Routing data for each tenant should be synchronized between NVO3 and MPLS VPN network. In internal DC NVO3 network, routing data synchronization between NVE and ASBR-d can be through either: a) RFC 4364 running between the NVEs and the ASBR1, or b) NVE-NVA architecture.

The Data plane process is same in these two cases.

6.1. Using RFC 4364

Route distinguishers (RD) and RT are specified for each VRF on each NVE. BGP MPLS/IP VPN protocol extension is running between NVEs and ASBR-d utilizing the [BGP Remote-Next-Hop] which describes the BGP MPLS/IP VPN protocol extension details to specify a set of remote tunnels (1 to N) that occur between two BGP speakers.

6.1.1. DC to WAN direction

1. NVE1 and NVE2 advertise local TS's IP Address to ASBR-d. NVE1 and NVE2 learn the local TS's IP Address via ARP or other mode.

2. When ASBR-d receives route data from each NVE, it allocates MPLS VPN Label per tenant (VN ID) per NVE and the RD and RT remain the same. 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, VN ID> pair forms incoming forwarding table which is used to forward MPLS traffic from external DC to internal DC. The incoming forwarding table on ASBR-d is as follows:

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

Incoming forwarding table

6.1.2. WAN to DC direction

1. When ASBR-d receives route data from ASBR-w, ASBR-d allocates VN ID for each VPN Label, and then ASBR-w 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 Incoming VPN Label in vanilla ASBR, it has local significance on ASBR-d, each VN ID corresponds to a MPLS VPN Label on peer ASBR-w; The VN ID space should be assigned in beforehand and should be orthogonal to the VN ID space for tenant identification(for example, assuming ASBR-d 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. Assuming ASBR-d receives VPN Label 3000 and 4000 from ASBR-w, the outgoing forwarding table on ASBR-d is as follows:

VN ID	Out VPN Label
10000	3000
10001	4000

Outgoing forwarding table

2. When each local NVE receives route data 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.

6.2. NVE-NVA architecture

No distributed BGP VPN protocol (RFC4364) is running on all NVEs and ASBR-d in NVO3 network, NVEs and ASBR-d are controlled by centralized NVA. The NVA runs EBGP VPN protocol with peer ASBR-w 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 Architecture

6.2.1. DC to WAN direction

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

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

3. NVA downloads incoming forwarding table to ASBR-d.

6.2.2. WAN to DC direction

1. NVA receives VPN routing information from peer ASBR-w.

2. NVA allocates VN ID for each MPLS VPN Label receiving from ASBR-w.

3. NVA downloads outgoing forwarding table to ASBR-d.

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

7. Enhanced Option-B solution

At WAN network side, if there is a VPN with multiple IP prefixs, VPN route synchronization to local NVE located in data center network will cause a lot pressure on it. In this case, the procedures above at ASBR-d can be enhanced as follows.

EBGP VPN connection for this VPN is terminated at ASBR-d, which means the ASBR doesn't allocate new VN ID 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. For the traffic from DC to WAN side, IP forwarding process is performed, VRF is selected based on VN ID, and then the traffic will be MPLS encapsulated and send to peer ASBR-w.

8. 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].

9. IANA Considerations

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

10. References

10.1. Normative References

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

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

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

10.2. Informative References

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

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

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

[4] [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)

11. Acknowledgments

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

Authors' Addresses

Weiguo Hao
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China
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.

Robert Raszuk
Email: robert@raszuk.net

Luyuan Fang
Microsoft
Email: lufang@microsoft.com

Shahram Davari
Broadcom
Davari@Broadcom.com