BESS Working Group
W. Hao
L. Yong
S. Hares
Internet Draft
Huawei
Osama Zia
Microsoft
Muhammad Durrani
Cisco
Intended status: Standard Track
September 10, 2015
Expires: March 10, 2016

Inter-AS Option C between NVO3 and BGP/MPLS IP VPN network
draft-hao-bess-inter-nvo3-vpn-optionc-03.txt

Abstract

This draft describes inter-as option-C solution between NVO3 network and MPLS/IP VPN network. Transport layer stitching solution should be provided. Also to ensure VPNv4 route exchange correctly between local NVE and remote PE, VNID space should be partitioned, only the VNIDs of lower 1 Million can be used for interconnection with outer MPLS VPN network using option-C solution, the rest 15 Million VNIDs can only be used for intra DC.

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.

Table of Contents

1. Introduction 2
2. Conventions used in this document..... 4
3. Reference model 5
4. Traditional Option-C [RFC4364] Recap..... 6
5. Inter-As Option-C Solution..... 6
5.1. EBGP process for transport layer stitching..... 7
5.1.1. UDP based overlay network..... 7
5.1.2. GRE based overlay network..... 8
5.2. VPN routes exchange..... 9
5.3. Data forwarding process..... 9
5.3.1. Data flow from TS1 to CE1..... 9
5.3.2. Data flow from CE1 to TS1..... 10
6. NVE-NVA architecture..... 10
6.1. EBGP process for transport layer stitching..... 11
6.2. VPN route exchange..... 11
7. Security Considerations..... 12
8. IANA Considerations 12
9. References 12
9.1. Normative References..... 12
9.2. Informative References..... 13
10. Acknowledgments 13

1. Introduction

In cloud computing era, multi-tenancy has become a core requirement for data centers. Since Network Virtualization Overlays (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 [RFC7348] and NVGRE [NVGRE] are two typical NVO3 technologies. In NVO3 network, 24-bit VNID (or VSID) is used to identify different virtual networks, theoretically 16M virtual networks can be supported in a data center. MPLS Over GRE and MPLS In UDP [RFC7510] are another two technologies to construct the overlay network, 20-bit MPLS Label is used as virtual networks identification. 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 a data center network, each tenant may include one or more layer 2 virtual network. In normal case, each tenant corresponds to one routing domain (RD), each layer 2 virtual network corresponds to one or more subnets.

To provide cloud service to external data center customers, data center networks should be connected to the WAN network. BGP MPLS/IP VPN are widely deployed technologies on WAN networks. Normally internal data center and external MPLS/IP VPN network are different Autonomous System (AS).

In multiple NVO3 data center inter-connecting scenario, the traffic across data center normally are carried over BGP MPLS/IP VPN network. This also requires an applicable inter-as solution between NVO3 network and external MPLS/IP network which can meet scale demands on existing and future NVO3 data center.

Similar to the Inter-as connection method defined in RFC4364, there are three different ways of handling this case, they are option-A, option-B and option-C respectively in order of increasing scalability.

Option-A is a back-to-back VRFs solution. Using option-A, EBGP session per VPN is created on peering ASBRs. In the data-plane, VLANs are used for tenant traffic separation. It has the lowest scalability among the three solutions. Compared to option-A solution, option-B solution has more scalability. But using option-B, ASBRs need to maintain and distribute all VPN prefixes. In the data plane, ASBRs need to perform MPLS VPN Label switching. Because MPLS VPN Label switching table space on ASBRs is limited, it still has scalability limitation for large VPN network. Option-C solution is a most scalable option through separating VPNv4 and PE prefixes exchange, the ASBRs don't need to maintain and distribute the

customers VPN prefixes. The ASBR is only used to exchange the service provider(SP) internal IP.

This draft is to propose inter-as option-C solution between NVO3 network and external BGP MPLS/IP VPN network. Compared to the traditional option-C solution defined in [RFC4364], it is for heterogeneous network interconnection, the control plane and data plane procedures in NVO3 network should be newly specified.

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.

VNID - Virtual Network Identifier (for VxLAN)

VSID - Virtual Subnet Identifier (for NVGRE)

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

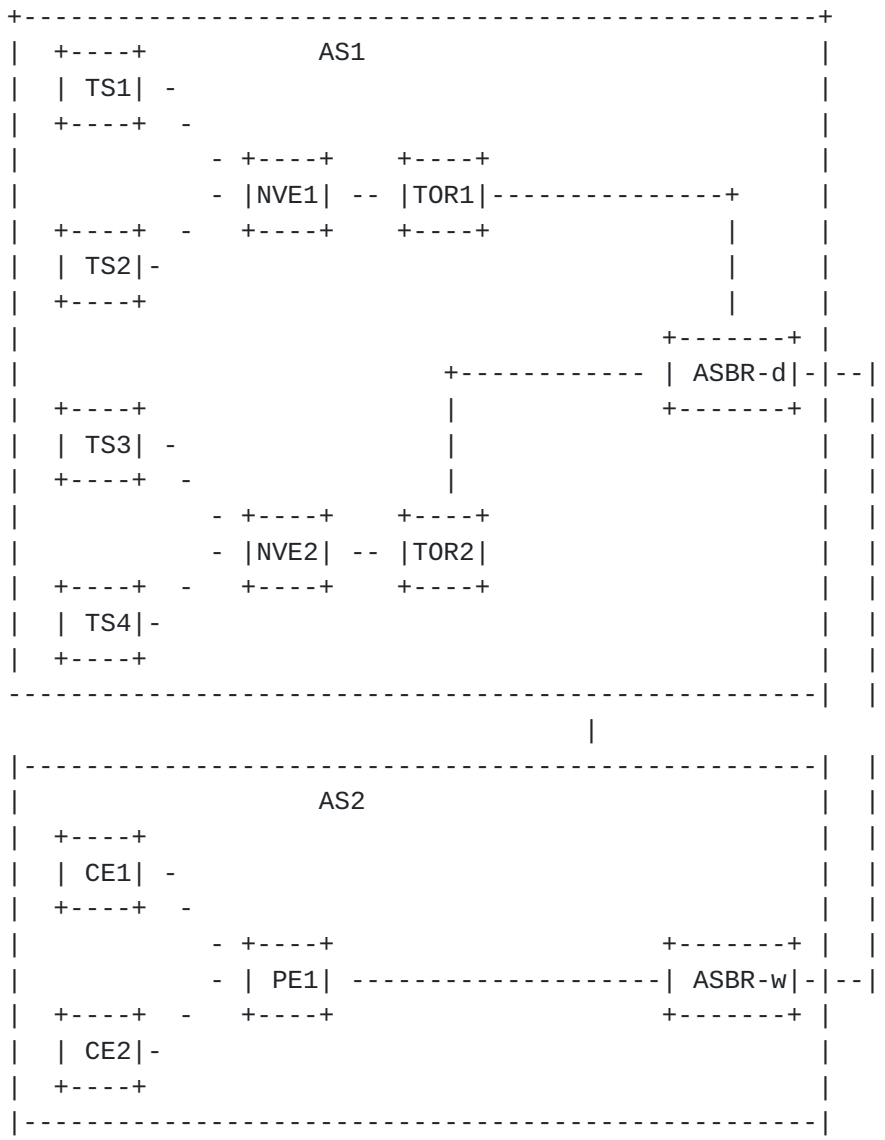

Figure 1 Reference model

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 is in AS 1, the MPLS/IP VPN network is in 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 tenant1 and tenant2 respectively. PE1 assigned MPLS VPN Label 1000 and 2000 for the routes from CE1 and CE2 respectively.

4. Traditional Option-C [RFC4364] Recap

In traditional Option-C defined in [RFC4364], an MP-EBGP session between the end PEs in source and destination ASs is used for the redistribution of VPN-IPv4 routes. Labeled IPv4 routes are redistributed by EBGP between neighboring autonomous systems , inter-AS Option-C uses BGP as the label distribution protocol. Through this solution, VPN connectivity is maintained while keeping VPN-IPv4 routes out of the ASBRs, an ASBR only need maintain labeled IPv4/32 routes to the PE routers within its AS. If the /32 routes for the PE routers are NOT made known to the P routers(other than the ASBRs), then a packet's ingress PE need to put a three-label stack on it. The bottom label is assigned by the egress PE, corresponding to the packet's destination address in a particular VRF. The middle label is assigned by the ASBR, corresponding to the /32 route to the egress PE. The top label is assigned by the ingress PE's IGP Next Hop, corresponding to the /32 route to the ASBR.

5. Inter-As Option-C Solution

Each NVE operates as default layer 3 gateway to connect locally attached TS(s). VRFs are created on each NVE to isolate IP data plane forwarding table between various attached tenants. The VNID(or VSID and MPLS Label) is used as Tenant identification .

Similar to traditional Option-C defined in [RFC4364], an end to end tunnel path from NVE to PE as transport layer should be established through EBGPs (ASBR-ASBR) and two IBGPs in local DCs (between ASBR-PE and ASBR-NVE), and MP-BGP will be established over the tunnel so that VPN-IPv4 routes can be exchanged between the NVE and PE without AS awareness. Unlike traditional Option-C BGP label switched path, the tunnel path has two segments, one segment is the NVO3 tunnel from NVE to ASBR-d in NVO3 network, another segment is traditional BGP LSP from ASBR-d to PE in WAN network, the two segments should be stitched together at ASBR-d as per recommended implementation. The behavior on ASBR-w and PEs in MPLS VPN network has no implementation differences compared to the behavior of ASBR and PEs in traditional RFC4364 based MPLS VPN Option-C network.

5.1. EBGP process for transport layer stitching

This section will describe the EBGP procedures for the transport layer forwarding path stitching.

In WAN to DC direction, when ASBR-d receives labeled IPv4/32 routes from ASBR-w, one of the several allocation methods can be used for tunnel stitching among them few are IP, UDP port and GRE key allocation method . The method chosen by operator depends on the data center network type and the network scale. For the UDP based network of VXLAN [RFC7348] and MPLS In UDP [RFC7510], either IP allocation method or UDP port allocation method can be used. UDP port allocation should be within UDP ephemeral port range and one UDP port maps to a label. For NVGRE network, only IP allocation method can be used. For MPLS Over GRE network, either IP allocation method or GRE key allocation method can be used.

In DC to WAN direction, the transport layer stitching solution is same for all kinds of NVO3 network. In this solution, ASBR-d announces labeled IPv4/32 routes to ASBR-w for each NVE where unique MPLS Label is allocated. . The allocated MPLS Label and NVE IP address mapping forms incoming forwarding table which is used to stitch BGP LSP and NVO3 tunnel for inbound traffic forwarding, i.e., from external DC to internal DC.

5.1.1. UDP based overlay network

Both VXLAN and MPLS In UDP are UDP based encapsulations. For the outbound traffic from NVE to ASBR-d, there are two options at ASBR-d, i.e., the ASBR-d only accepts the traffic with standard destination UDP port(4789 for VXLAN [RFC7348], 6635 for MPLS In UDP [RFC7510]) or non-standardized destination UDP port in outer UDP header encapsulation.

In WAN to DC direction, if standard destination UDP port solution is used, when ASBR-d receives labeled IPv4/32 routes from ASBR-w, IP address allocation method should be used. The ASBR allocates an IP address per MPLS Label specified for a particular route defined in [RFC3107] to identify each remote PE, and then advertises the IPv4/32 route(indicating remote PE reachability) to all local NVEs with the VXLAN or MPLS In UDP tunnel attribute. [TUNNELENCAP] defines the relevant TLVs and sub-TLVs for the Tunnel Encapsulation Attribute. The local NVEs will encapsulate transport layer header using the Tunnel Encapsulation Attribute for the outbound traffic from internal DC to external DC, the ASBR-d generated IP is the

destination IP in NVO3 tunnel outer header, the UDP port is the standard well-known port for VXLAN and MPLS In UDP. The IP pool should be configured beforehand on ASBR-d. The new allocated IP and MPLS Label mapping forms outgoing forwarding table on ASBR-d which is used to stitch NVO3 tunnel and BGP LSP for outbound traffic forwarding . If non-standard destination UDP port is used, ASBR-d can allocate the combination of IP and UDP port(or only UDP port) per MPLS Label to identify each remote PE, and then advertises the IPv4/32 route received from ASBR-w to all local NVEs with the Tunnel Encapsulation Attribute. For each NVE, the destination IP and the destination port in NVO3 tunnel outer header is the new allocated IP and the new allocated port respectively. The new allocated IP and UDP port combination (or only UDP port) and MPLS Label mapping forms outgoing forwarding table on ASBR-d. This method is called UDP allocation method, the allocated UDP port range should be configured beforehand on ASBR-d.

In summary, IP allocation method has more IP address consumption than the UDP allocation method. If there is large number of remote PEs in WAN network, the UDP allocation method is suggested to be used to enhance network scalability.

5.1.2. GRE based overlay network

Both NVGRE and MPLS Over GRE are GRE based encapsulations. The GRE key field can be used to convey application-specific key value. In NVGRE, the key field has been used to convey 24-bit Virtual Subnet Identifier (VSID) as tenant identification, so for NVGRE, the GRE key field can't be used for the stitching purpose and only IP allocation method can be used. In MPLS Over GRE, the GRE key field has not been used explicitly by an application and can be used for the transport layer stitching at ASBR-d, i.e., GRE key allocation method can be used to conserve IP address space.

In WAN to DC direction, for MPLS Over GRE, when ASBR-d receives labeled IPv4/32 routes from ASBR-w, the ASBR can allocate a GRE key per MPLS Label to identify each remote PE, and then advertises the IPv4/32 route to all local NVEs with the Tunnel Encapsulation Attribute. The new allocated GRE key and MPLS Label correspondence forms outgoing forwarding table on ASBR-d. This method is called GRE key allocation method.

If ASBR-d needs to change IP address, UDP port or GRE key for a particular /32 route, it should advertising a new route with the

same NLRI and a new Tunnel Encapsulation Attribute to refresh all NVEs's local information.

5.2. VPN routes exchange

Each NVE and remote PE should establish MP-EBGP session for the announcement of VPN-IPv4 routes through RFC4364. Route distinguishers (RD) and RT are specified for each VRF on each NVE and PE.

Each NVE advertises all local VPN route to remote PEs using tenant identification VNID (or VSID and MPLS Label) as MPLS VPN Label. These remote PEs deal with the NVE as regular PE, they match RT and populates these VPN route to local VRF. For the traffic from remote CE to local TS, ingress PE uses the VNID as bottom label in MPLS encapsulation. Because VNID field is 24 bits, to ensure these NVEs and PEs interworking, VNID length should not be beyond 20 bits, i.e., VNID value must not be larger than 1 Million. In proposed implementation NVO3 network the VNID space should be partitioned, only the VNIDs of lower 1 Million can be used for interconnection with outer MPLS VPN network, the rest 15 Million VNIDs can only be used for intra DC.

Each MPLS VPN PE also advertises all local VPN route to peer NVEs, these NVEs match RT and populates these VPN route to local VRF. For the traffic from local TS to remote CE, because ingress NVE doesn't support MPLS encapsulation, it encoded the MPLS VPN Label advertised from remote PE as VNID in NVO3 encapsulation.

5.3. Data forwarding process

When VXLAN network and UDP port allocation method are used in data center, the procedures of data forwarding between TS1 and CE1 in figure 1 will be described step by step as follows.

5.3.1. Data flow from TS1 to CE1

1. TS1 sends a packet to NVE1, destination IP is CE1's IP.

2. NVE1 acquires local VRF relying on packet input interface, then looks up the VRF's routing table corresponding to tenant 1, performs NVO3 encapsulation, and sends the encapsulated packet out to ASBR-d. The MPLS VPN Label associated with the packet's destination address is encoded in VNID field. VXLAN tunnel destination IP and destination UDP port are the IP address and UDP port allocated on ASBR-d associated with the /32 routes for the remote PE routers that the remote CE attached to.

3. ASBR-d decapsulates the VXLAN received packet and performs MPLS encapsulation. Two Labels should be pushed in the MPLS encapsulation, BGP LSP Label as top Label and MPLS VPN Label as bottom Label. BGP LSP Label is acquired by looking up outgoing stitching table, MPLS VPN Label is copied from VNID.

4. ASBR-w swaps BGP MPLS Label, and push IGP Label and sends the packet out to PE1. MPLS VPN Label remains unchanged.

5. PE1 pops all MPLS Label, finds local VRF relying on bottom MPLS VPN Label, performs looks up in local VRF IP forwarding table , and then sends the packet out to CE1.

5.3.2. Data flow from CE1 to TS1

1. CE1 sends a packet to PE1, destination IP is TS1's IP.

2. PE1 acquires local VRF interface relying on packet input interface where CE1 egress out the packet, then launches a lookup in VRF's routing table. It pushes three-label stack on the outgoing packet. The bottom label is the tenant VNID corresponds to TS1, the VNID is 10. The middle label is assigned by the ASBR-w, associating with the /32 route for the egress NVE1. The top label is assigned by the ingress PE's IGP Next Hop, corresponding to the /32 route to ASBR-w.

3. ASBR-w pops top IGP Label, swaps middle BGP Label, and then sends the packet out to ASBR-d.

4. ASBR-d decapsulates MPL packet, performs VXLAN encapsulation and then sends the packet to egress NVE1. The egress NVE's IP address is acquired by performing a looking up in the stitching table, VNID is copied from the bottom MPLS VPN Label.

5. NVE1 decapsulates incoming NVO3 encapsulated packet, looks up local VRF interface based on VNID, then performs a look up in the routing table and forwards the packet out to destination TS1.

6. NVE-NVA architecture

In this architecture, the NVE control plane and forwarding functionality are decoupled. All NVEs in NVO3 network don't need to support BGP protocol; these NVEs have only data plane functionality and are controlled by centralized NVA using openflow, ovsdb, i2rs, etc. The NVA runs BGP with ASBR-d to exchange plain IP route to

IPv4/32 of each WAN PE associated with the BGP tunnel encapsulation attribute. The NVA also runs MP-BGP protocol [RFC4364] with peer PE for all the NVEs to exchange VPNv4 route, VNID is used as MPLS VPN Label. ASBR-d can choose IP allocation, UDP allocation or GRE key allocation method for the transport stitching.

NVA maintains all tenant information, and originates BGP routes with the appropriate RD and RT. The NVA tenant information includes VNID(or VSID and MPLS Label) to identify each tenant and the corresponding RD and RT. This information can be statically configured by operators or dynamically allocated. This information also includes all TS's MAC/IP address and its attached NVE information.

6.1. EBGP process for transport layer stitching

DC to WAN direction:

1. ASBR-d allocates BGP MPLS Label per NVE.

2. ASBR-d advertises BGP Label routing information to peer ASBR-w. ASBR-d generates incoming stitching table <new allocated BGP MPLS Label, NVE IP>.

WAN to DC direction:

1. ASBR-d receives BGP Label routing information from peer ASBR-w.

2. ASBR-d allocates NVO3 Tunnel IP, UDP port or GRE key for each MPLS Label received from ASBR-w, the ASBR-d announces the IPv4/32 Route to NVA with the Tunnel Encapsulation Attribute.

3. ASBR-d generates outgoing stitching table<new allocated Tunnel IP(or UDP port and GRE key), received MPLS Label>.

6.2. VPN route exchange

NVA advertises all internal data center tenant routing information to remote PEs using RFC 4364, which includes RD, RT, IP prefix, and MPLS VPN Label, the tenant identification of VNID is used as MPLS VPN Label.

Each remote MPLS VPN PE also advertises local VPN routes to NVA. NVA acquires NVO3 Tunnel IP(or UDP port and GRE key) allocated by

ASBR-d corresponding to the PE, matches RT attribute and populates the VPN routes to local VRF.

Then the NVA downloads corresponding VPN forwarding table including <destination IP prefix/Mask, NVO3 Tunnel IP(or UDP port and GRE key), VNID>to each NVE.

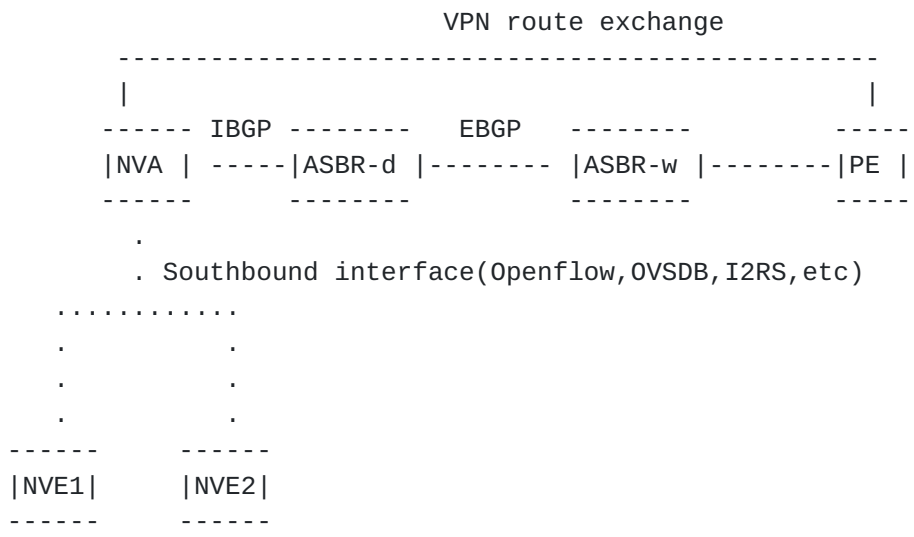

Figure 2 NVE-NVA Architecture

7. Security Considerations

Internal IP (Loopback IP for PE/NVE) addresses a network is advertised and visible in another network, which is a security risk. Most operators wants to prevent any external visibility and access into their internal devices IP. option C is suggested to be deployed within a single SP or enterprise with both MPLS and NVO3 networks.

8. IANA Considerations

NA.

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.

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

[3] [RFC3107] Y. Rekhter,E. Rosen, ''Carrying Label Information in BGP-4'', RFC 3107, May 2001

9.2. Informative References

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

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

[3] [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)

[4] [RFC7348] M. Mahalingam, etc, "Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", RFC7348, August 2014.

[5] [NVGRE] P. Garg, etc, "NVGRE: Network Virtualization using Generic Routing Encapsulation", draft-sridharan-virtualization-nvgre-08, April 13, 2015.

[6] [TUNNELENCAP] E. Rosen, etc, "Using the BGP Tunnel Encapsulation Attribute without the BGP Encapsulation SAFI", draft-rosen-idr-tunnel-encaps-00, June, 2015.

10. Acknowledgments

Authors like to thank Thomas Morin, Shunwan Zhuang, Haibo Wang, Jie Dong for their 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.

Osama Zia
Microsoft
Email: osamaz@microsoft.com

Muhammad Durrani
Cisco
Phone: +1-408-527-6921
Email: mdurrani@cisco.com