BESS

W. Hao
Y. Li
L. Wang

Internet Draft

Huawei Technologies Ltd.

Intended status: Standards Track
Expires: December 2015

June 5, 2015

Multicast Group Address Auto-Provisioning For NVO3 Network
draft-hao-bess-mcast-auto-nvo3-00.txt

Abstract

This document provides dynamic underlay multicast group address provisioning mechanism for each VNI or combination of VNI and overlay multicast group address. The underlay multicast group address allocation function is provided on NVA(centralized point), NVE communicates with NVA using BGP protocol extension.

Status of this Memo

This Internet-Draft is submitted 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/1id-abstracts.html

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

Copyright Notice

Table of Contents

1. Introduction

Network virtualization over Layer 3 (NVO3) is a technology that is used to address issues that arise in building large, multitenant data centers that make extensive use of server virtualization [RFC7364]. VXLAN and NVGRE are two typical NVO3 technologies. Both of these technologies include 24 bits Virtual Network Identifier (VNI) as tenant identification. NVO3 overlay network can be controlled through centralized NVE-NVA architecture or through distributed BGP VPN protocol.

For multicast traffic handling, using the multicast capabilities of the underlying Network is one possible way. In this method, the underlay supports IP multicast and the ingress NVE encapsulates the packet with the appropriate underlay IP multicast address in the tunnel encapsulation header for delivery to the desired set of NVEs. The underlay multicast distribution tree is called P-multicast tree, the underlay multicast group is called P-multicast group. The

protocol in the underlay to construct P-multicast tree could be any variant of Protocol Independent Multicast (PIM), or protocol dependent multicast, such as [ISIS-Multicast]. The method is mentioned in the NVO3 architecture [NVO3-ARCH] and multicast framework [MCASTFM] documents. In this method, for layer 2 directly connecting TSs, each NVE acts as a multicast router and supports

proper mapping of IGMP/MLD's messages to the messages needed by the underlay IP multicast protocols.

For broadcast, unknown unicast and bidirectional multicast application traffic in each VN, the associated VTEPs should act as both the source and destination of the traffic, bidirectional P-Multicast tree offer better scalability then PIM-SM/SSM with the number of flows required being g. Also Bidir tree is share tree, in regular data center spine-leaf network architecture, share tree has same optimal forwarding path as source tree established by PIM-SM/SSM. So in most cases, the underlay P-multicast tree in NVO3 network should be Bidir tree than source tree. Both BIDIR-PIM and ISIS-Multicast protocol are suitable to construct the bidirectional P-multicast tree. Each bidirectional P-multicast tree corresponds to one P-Multicast group address.

To transport overlay BUM(broadcast, unknown unicast and multicast) traffic, we need to have a mapping between the VNI/VNI plus C-Multicast group and the P-Multicast group that it will use. The overlay multicast group is called C-Multicast group. There are multiple mapping methods as follows:

1. Dedicated inclusive tree. In this case, a multicast tree is dedicated to a VNI, a separate underlay multicast group address is allocated for each VNI.

2. Aggregate inclusive tree. In this case, a multicast tree is shared by multiple VNIs, a shared multicast group address is allocated for multiple VNIs.

3. Dedicated selective tree. In this case, a multicast tree is dedicated to a combination of VNI plus overlay multicast group, a separate underlay multicast group address is allocated for a combination.

4. Aggregate selective tree. In this case, a multicast tree is shared

by multiple combinations of VNI plus overlay multicast group, a shared underlay multicast group address is allocated for multiple combinations.

When inclusive tree solution is used, if a VN has multiple TSs and these TSs spread over multiple NVEs, then these NVEs should ensure same P-multicast group address is provisioned for the VN, i.e., P-Multicast group provisioning consistency should be ensured among multiple NVEs connecting to same underlay multicast tree. For VN and P-Multicast group mapping, it can be done at the management layer which provided to the individual VTEPs through a management channel

or through control plane protocol which is introduced in this document.

In selective tree solution, because NVE can't get the list of participants for each C-multicast group ahead of time, the mapping between C-multicast group and P-multicast group on each NVE can't be configured statically through a management channel, P-multicast group and the mapping between P-multicast group and C-multicast group should be provided dynamically using control plane protocol.

This draft proposes a control plane method to dynamically allocate P-multicast group for each NVE on a centralized point like NVA, the communication protocol between each NVE and NVA uses BGP EVPN protocol extension. In the future, other protocols also can be considered.

2. Terminology

EVI: An EVPN instance spanning the Provider Edge (PE) devices participating in that EVPN.

Ethernet Tag: An Ethernet tag identifies a particular broadcast domain, e.g., a VLAN. An EVPN instance consists of one or more broadcast domains.

NVA: Network Virtualization Authority

NVE: Network Virtualization Edge

NVGRE: Network Virtualization using GRE

VXLAN: Virtual eXtensible LAN

3. Solution overview

P-multicast group address allocation function is provided on NVA(centralized point), NVA and each NVE establish BGP EVPN session in beforehand for communication. NVA configures P-multicast group address pool and allocation policy ahead of time. A non-exhaustive list of allocation policies on NVA are described as follows:

1. Per VN/VN plus C-Multicast group per P-Multicast group.

2. Per NVE sets per P-Multicast group. If multiple VNs attach to same NVE devices, then a same P-Multicast group is allocated for these VNs.

3. All VN share same P-Multicast group, but per P-Multicast group allocated per VN plus C-Multicast group.

EVPN can be used for NVO3 network for both unicast and multicast traffic forwarding [EVPN-OVERLAY]. VNI to EVPN EVI mapping supports 1:1 model and N:1 model.

The E-VPN Multicast BGP route combined with a new BGP Extended Community attribute(P-Multicast Allocation Extended Community) is used for P-multicast group address auto-provisioning process. The new BGP Extended Community attribute is defined to identify the group address request and reply message, the attribute may be advertised along with the E-VPN Inclusive Multicast BGP route and the E-VPN Selective Multicast BGP route [EVPN-SELMCST]. The E-VPN Inclusive Multicast BGP route is used to discover the multicast tunnels among the NVEs associated with a VNI. The E-VPN Selective Multicast BGP route is used to discover the multicast tunnels among the NVEs associated with a VNI plus C-multicast group [EVPN-SELMCST].

The P-Multicast group allocation interaction process between NVE and NVA is as follows:

1. When a NVE detects a local VN creation or first C-multicast joining event, the NVE sends P-multicast group address request

message to NVA(centralized point). PMSI Tunnel attribute isn't needed to be associated with the route. Request flag in the new BGP Extended Community attribute is set.

2. NVA allocates P-multicast group address relying on local policy, and then sends P-multicast group address reply message to original NVE. The allocated P-multicast group address is carried in the tunnel identifier of PMSI Tunnel attribute. Reply flag in the new BGP Extended Community attribute is set. NVA records the mapping between P-Multicast group and NVE's information. The NVE's information includes the NVE's VTEP IP address, VNID or combination of VNID plus C-multicast group.

3. The NVE sends E-VPN Multicast BGP route to other NVEs through regular EVPN process for multicast tunnel discovery [RFC7432]. The Multicast route is tagged with the PMSI Tunnel attribute, which is used to encode the type of multicast tunnel to be used as well as the multicast tunnel identifier which fills the allocated P-multicast group address.

The P-Multicast group release interaction process between NVE and NVA is as follows:

1. When a NVE detects a local VN deletion or last C-multicast leaving event, the NVE sends P-multicast group address withdraw message to NVA(centralized point). PMSI Tunnel attribute isn't needed to be associated with the route. Request flag in the new BGP Extended Community attribute is set.

2. NVA releases local record of the NVE's information. If this is the last NVE using the P-Multicast group, the NVA will release the P-Multicast group to pool for future re-allocation.

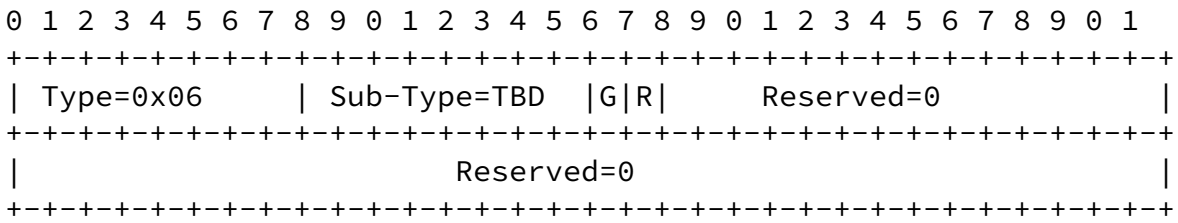
For the network wide global unique VNID, RD field can be set to zero in the allocation request and reply message, NVA relies on the VNID to allocate P-Multicast group address. For NVE local VNID, RD should be used to identify each virtual network, NVA relies on the global unique RD to allocate P-Multicast group address.

4. BGP Protocol extension

4.1. P-Multicast Allocation Extended Community

This Extended Community is a new transitive Extended Community having a Type field value of 0x06 and the Sub-Type TBD. It may be advertised along with Inclusive Multicast Ethernet Tag Routes or Selective Multicast Ethernet Tag Route.

Each P-Multicast Allocation extended community is encoded as an 8-octet value, as follows:



The "G" bit of 1 indicates "Request" message. The "R" bit of 1 indicates "Reply" message.

5. Security Considerations

NVA processes all NVE's P-Multicast group address allocation request message, it is vulnerable for attacking by inappropriate NVE in data center. NVE's identity should be strictly inspected on NVA, possible security solution need to be further researched.

6. IANA Considerations

IANA is requested to allocate the following EVPN Extended Community sub-type besides [RFC7432].

0x00	MAC Mobility	[RFC7432]

0x01	ESI Label	[RFC7432]
0x02	ES-Import Route Target	[RFC7432]
TBD	P-Multicast Allocation	[This document]

6.1. Normative References

[1] [NVO3MFM] A. Ghanwani, et al, "A Framework for Multicast in NVO3",draft-ietf-nvo3-mcast-framework-00, work in progress. May 10, 2015.

[2] [EVPN-OVERLAY] A. Sajassi, et al, "A Network Virtualization Overlay Solution using EVPN"draft-sd-l2vpn-evpn-overlay-03, work in progress. June 18, 2014.

[3] [RFC7432] Sajassi et al., "BGP MPLS Based Ethernet VPN", RFC7432, February 2015.

[4] [NVO3-ARCH] Narten, T. et al.," An Architecture for Overlay Networks (NVO3)", work in progress, February 2014.

[5] [NVO3-ARCH] J. Zhang, Z. Li, ''Selective Multicast in EVPN'', draft-zhang-l2vpn-evpn-selective-mcast-01, work in progress, July 2014.

6.2. Informative References

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

[2] [NVGRE] Sridharan, M. et al., "NVGRE: Network virtualization using Generic Routing Encapsulation", work in progress.

[3] [ISIS-Multicast] L. Yong, et al, "ISIS Protocol Extension For Building Distribution Trees", work in progress. Oct 2013.

7. Acknowledgments

The authors wish to acknowledge the important contributions of Yisong Liu, Shunwan Zhuang and Qiandeng Liang.

Authors' Addresses

Weiguo Hao
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China

Email: haoweiguo@huawei.com

Yizhou Li
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China
Email: liyizhou@huawei.com

Lili Wang
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China
Email: lily.wong@huawei.com