Network Working Group
Internet-Draft
Expires: November 16, 2015

B. Sarikaya
F. Xia
Huawei USA
S. Fan
China Telecom
May 15, 2015

DHCP Options for Configuring Tenant Identifier and Multicast Addresses in Overlay Networks
draft-sarikaya-nvo3-dhc-vxlan-multicast-03.txt

Abstract

This document defines DHCPv4 and DHCPv6 options for assigning VXLAN Network Identifier and multicast addresses for overlay networks such as the Virtual eXtensible Local Area Network (VXLAN). New DHCP options are defined which allow a Network Virtualization Edge to request any source multicast address and tenant identifier for the newly created virtual machine.

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). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

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."

This Internet-Draft will expire on November 16, 2015.

to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction

Data center networks are being increasingly used by telecom operators as well as by enterprises. Usually these networks are organized as one large Layer 2 network in a single building. In some cases such a network is extended geographically using Virtual Local Area Network (VLAN) technologies still as an even larger Layer 2 network connecting the virtual machines (VM), each with its own MAC address.

Another important requirement was growing demand for multitenancy, i.e. multiple tenants each with their own isolated network domain. In a data center hosting multiple tenants, each tenant may independently assign MAC addresses and VLAN IDs and this may lead to potential duplication. At the same time, it is usually the Layer 3 network between the VMs run on different servers, whether these VMs located inside the same DCs or different DCs. That is, the Layer3 switches inside DC and the Internet between DCs will isolate the Layer2 network. By the way, as the granularity is based on VM and the end-points may be deployed in different DCs, the existed VLAN ID will be a big barrier as the max number is 4096.

What we need is IP tunneling scheme based overlay network. In this document we assume the Virtual eXtensible Local Area Network (VXLAN) based overlay is used but the solution is valid for other overlay networks such as NVGRE and others.

VXLAN overlays a Layer 2 network over a Layer 3 network. Each overlay is identified by the VXLAN Network Identifier (VNI). This allows up to 16M VXLAN segments to coexist within the same administrative domain [RFC7348]. In VXLAN, each MAC frame is transmitted after encapsulation, i.e. an outer Ethernet header, an IPv4/IPv6 header, UDP header and VXLAN header are added. Outer Ethernet header indicates an IPv4 or IPv6 payload. VXLAN header contains 24-bit VNI.

It should be noted that in this document, VTEP plays the role of the Network Virtualization Edge (NVE) according to NVO3 architecture for overlay networks like VXLAN or NVGRE defined in [I-D.ietf-nvo3-arch]. NVE interfaces the tenant system underneath with the L3 network called the Virtual Network (VN).

In order to deliver tenant traffic, NVE needs to know the VXLAN Network Identifier (VNI) assigned to this tenant. In this document, we use Dynamic Host Configuration Protocol (DHCP) for this purpose. NVE needs VNI to encapsulate the packets that its virtual machines generate. Since multi-tenancy requires independent address space for each tenant. The address configuration mechanism, e.g. DHCP server needs to keep track of VNI values assigned when assigning addresses to the virtual machines (VM). It is possible for a VM to be configured the same IPv4/IPv6 address as another VM belonging to a different tenant.

As stated in NVO3 architecture document [I-D.ietf-nvo3-arch] for tenant multicast (or broadcast) traffic, an NVE MUST maintain a per-VN table of mappings and other information on how to deliver multicast (or broadcast) traffic. Since we assume that the underlying network supports IP multicast, the NVE could use IP multicast to deliver tenant traffic. In such a case, the NVE would need to know what IP underlay multicast address to use for a given VN. By the way, the underlying network maybe can NOT support IP multicast, in that case the VTEP should act as the node run multicast protocol itself, for instance, PIM. But it will be beyond the scope of this document.

In this document, we develop a protocol to assign multicast addresses to the VXLAN tunnel end points or NVEs using Dynamic Host Configuration Protocol (DHCP). Multicast communication in the overlay network is used for sending broadcast MAC frames, e.g. the Address Resolution Protocol (ARP) broadcast frame. Multicast

communication can also be used to transmit multicast frames and unknown MAC destination frames.

2. Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119]. The terminology in this document is based on the definitions in [RFC7348], [I-D.ietf-nvo3-arch] and [I-D.ietf-nvo3-nve-nva-cp-req].

3. Overview of the protocol

Multicast addresses to be assigned by the DHCP server are administratively scoped multicast addresses, in IPv4 [RFC2365] and in IPv6 [RFC4291]. The steps involved in the protocol are explained below for IPv4:

Creation of a VM

In this step, NVE receives a request from the Network Virtualization Authority (NVA)[I-D.ietf-nvo3-nve-nva-cp-req] to create a Virtual Machine with a tenant identifier, e.g. VXLAN Network Identifier or VNI.

DHCP Operation

NVE starts DHCP state machine by sending DHCPDISCOVER message to the default router, e.g. the Top of Rack (ToR) switch or the aggregation switch. ToR switch, aggregation switch could be DHCP server or most possibly DHCP relay with DHCP server located upstream. NVE MUST include Multicast Address options defined in this document. NVE sends the parameter request list option with the newly defined VXLAN Network Identifiers(VNI) DHCP Option to request a value for VXLAN Network Identifier. DHCP server replies with DHCPOFFER message. DHCP server sends VNI DHCP Option and administratively scoped IPv4 multicast address. NVE checks this message and if it sees the options it requested, DHCP server is confirmed to support the multicast address options. DHCPREQUEST message from NVE and DHCPACK message from DHCP server complete DHCP message exchange. Different VXLAN Network Identifiers (VNI) need different multicast groups (for load balancing). At the same time, different VNIs need different address spaces for VM, that is, two VMs belong to different VNIs probably have the same IP address.

NVE as Multicast Source

After receiving the required information, the NVE as multicast source communicates with the edge router in order to build the multicast tree.

NVE as Listener

After receiving the required information, the NVE as listener communicates with the edge router by sending IGMP Report to join the multicast group.

IPv6 operation is slightly different:

Creation of a VM

In this step, NVE receives a request from the NVA to create a Virtual Machine with a tenant identifier or VNI.

DHCP Operation

NVE starts DHCP state machine by sending DHCPv6 Solicit message to the default router, e.g. the Top of Rack (ToR) switch or aggregation switch. ToR switch/aggregation switch could be DHCP server or most possibly DHCP relay with DHCP server located upstream. NVE MUST include the Option Request Option with OPTION_VNI defined in this document to request a value for the tenant id or VXLAN Network Identifier. NVE MUST include DHCPv6 VXLAN Multicast Address Option or OPTION_MA to request the multicast address. DHCP server replies with DHCPv6 Advertise message. NVE checks this message and if it sees the options it requested, DHCP server is confirmed to support multicast address options. DHCPv6 Request message from NVE and DHCPv6 Reply message from DHCPv6 server complete DHCP message exchange.

NVE as Multicast Source

After receiving the required information, the NVE as multicast source communicates with the edge router in order to build the multicast tree.

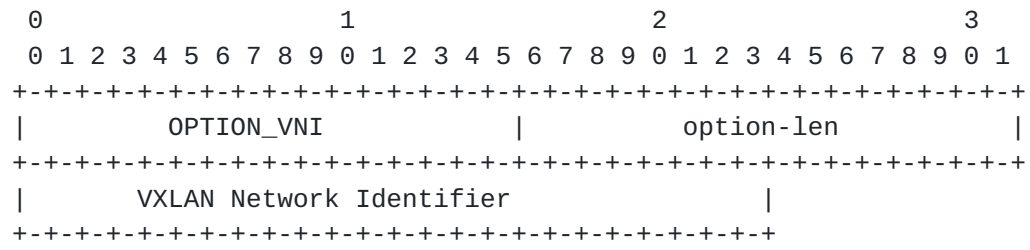
NVE as Listener

After receiving the required information, the NVE as listener communicates with the edge router by sending MLD Report to join the multicast group.

4. DHCPv6 Options

4.1. VXLAN Network Identifier Option

A DHCP VNI Option is defined as follows.



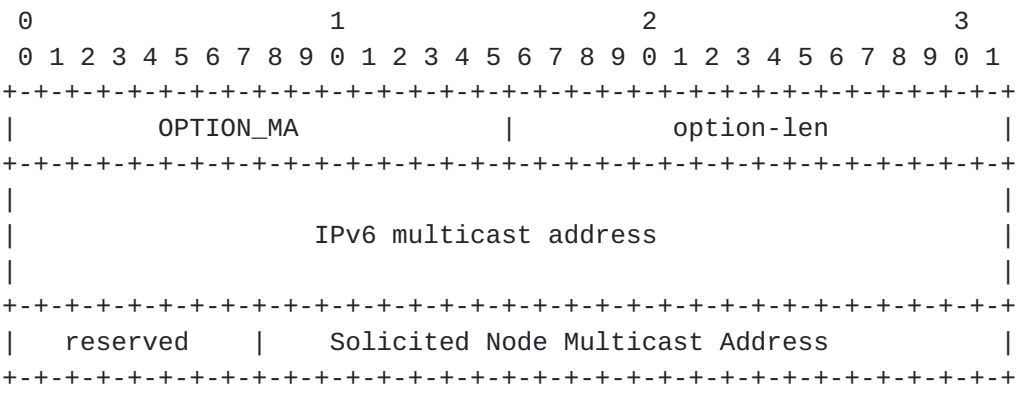
option-code OPTION_VNI (TBD).

option-len 7.

VXLAN Network Identifier 3.

4.2. DHCPv6 VXLAN Multicast Address Option

The option allows the NVE to send solicited-node multicast address to DHCP server and receive administratively scoped IPv6 multicast address.



option-code OPTION_MA (TBD).

option-len 24.

IPv6 multicast address An IPv6 address.

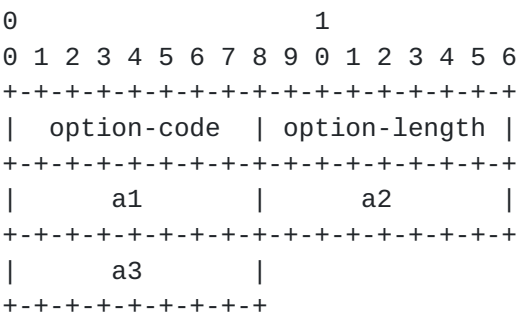
reserved must be set to zero

Solicited Node Multicast Address as in RFC 4861.

5. DHCPv4 Options

5.1. VXLAN Network Identifier Option

The option allows the NVE to send the VNI to DHCP server.



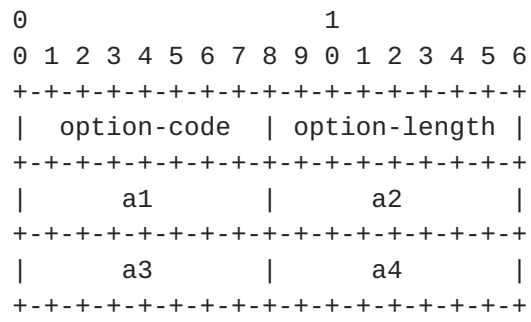
Option-code
VXLAN Network Identifier Option (TBD)

Option-len
3.

a1-a4
NVE as DHCP Client receives VNI value in a1-a3.

5.2. VXLAN Multicast Address Option

The option allows the NVE to receive administratively scoped IPv4 multicast address.



Option-code
VXLAN Multicast Address Option (TBD)

Option-len
4.

a1-a4
VTEP as DHCP Client sets a1-a4 to zero, DHCP server sets a1-a4 to the multicast address.

6. Client Operation

In DHCPv4, the client, VTEP MUST set 'htype' and 'chaddr' fields to specify the client link-layer address type and the link-layer address. The client must set the hardware type, 'htype' to 1 for Ethernet [RFC1700] and 'chaddr' is set to the MAC address of the virtual machine.

The client MUST request VXLAN Network Identifier Option in parameter request list to receive the VXLAN network identifier value assigned to the virtual machine by DHCP server.

In DHCPv6, the client MUST use OPTION_CLIENT_LINKLAYER_ADDR defined in [RFC6939] to send the MAC address. In this option, link-layer type MUST be set to 1 for Ethernet and link-layer address MUST be set to the MAC address of VM. Note that in [RFC6939], OPTION_CLIENT_LINKLAYER_ADDR is defined to be used in Relay-Forward DHCP message. In this document this option MUST be sent in DHCPv6 Solicit message.

The client MUST request IPv6 VXLAN Network Identifier Option OPTION_VNI in Option Request Option (ORO) to request the VXLAN network identifier value assigned to the virtual machine by DHCP server.

The Client MUST set IPv6 multicast address in DHCPv6 VXLAN Multicast Address Option, OPTION_MA to zero.

The client MUST set Solicited Node Multicast Address to zero if the neighbor discovery message is sent to all-nodes multicast address. The client MUST set Solicited Node Multicast Address to the low-order 24 bits of an address of the destination if the neighbor discovery message is sent to the solicited-node multicast address.

7. Server Operation

DHCPv4 server looks for VXLAN Network Identifier Option code in parameter request list option in DHCPDISCOVER message from the Client. DHCP server MUST add VXLAN Network Identifier Option in DHCP OFFER message with VNI value assigned in a1-a3.

If DHCPv4 server is configured to support VXLAN multicast address assignments, it SHOULD look for VXLAN Multicast Address Option in DHCPDISCOVER message. The server MUST return in VXLAN Multicast Address Option's a1-a4 an organization-local scope IPv4 multicast address (239.192.0.0/14) [RFC2365]. The server MUST use the VNI value for obtaining the organization-local scope IPv4 multicast address. VNI value is directly copied to 239.192.0.0/14 if the first

6 bits are zero, i.e. no overflow ranges need to be used. Otherwise, either of 239.0.0.0/10, 239.64.0.0/10 and 239.128.0.0/10 overflow ranges SHOULD be used. Note that these ranges can accomodate the VNI in its entirety.

DHCPv4 server MUST know and consider VXLAN Network Identifier (VNI) value before assigning VXLAN Multicast Address Option. The server gets VNI value by means out of scope in this document.

DHCPv6 server looks for VXLAN Network Identifier Option or OPTION_VNI value in Option Request Option of DHCPv6 Solicit message. DHCPv6 server MUST add VXLAN Network Identifier Option in DHCPv6 Advertise message with VNI value in VXLAN Identifier field.

If DHCPv6 server is configured to support VXLAN multicast address assignments it SHOULD look for DHCPv6 VXLAN Multicast Address Option in DHCPv6 Solicit message. The server MUST return in IPv6 multicast address field an Admin-Local scope IPv6 multicast address (FF04/16) by copying the VNI of the virtual machine to the least significant 24 bits of the group ID field and setting all other bits to zero if Solicited Node Multicast Address field received from the client was set to zero in DHCPv6 Advertise message. Otherwise the Solicited Node Multicast Address field is copied to bits 47-24 of the group ID field and all leading bits are set to zero.

DHCPv6 server MUST know and consider VXLAN Network Identifier (VNI) value before assigning VXLAN Multicast Address Option. The server gets VNI value by means out of scope in this document.

8. Security Considerations

The security considerations in [RFC2131], [RFC2132] and [RFC3315] apply. Special considerations in [RFC7348] are also applicable.

9. IANA considerations

IANA is requested to assign the OPTION_VNI and OPTION_MA and VXLAN Network Identifier and VXLAN Multicast Address Option Codes in the registry maintained for DHCPv4 and DHCPv6.

10. Acknowledgements

The authors are grateful to Bernie Volz and Ted Lemon for providing comments that helped us improve the document.

11. References

11.1. Normative References

[RFC1700] Reynolds, J. and J. Postel, "Assigned Numbers", RFC 1700, October 1994.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC2131] Droms, R., "Dynamic Host Configuration Protocol", RFC 2131, March 1997.

[RFC2132] Alexander, S. and R. Droms, "DHCP Options and BOOTP Vendor Extensions", RFC 2132, March 1997.

[RFC3315] Droms, R., Bound, J., Volz, B., Lemon, T., Perkins, C., and M. Carney, "Dynamic Host Configuration Protocol for IPv6 (DHCPv6)", RFC 3315, July 2003.

[RFC3956] Savola, P. and B. Haberman, "Embedding the Rendezvous Point (RP) Address in an IPv6 Multicast Address", RFC 3956, November 2004.

[RFC2365] Meyer, D., "Administratively Scoped IP Multicast", BCP 23, RFC 2365, July 1998.

[RFC4291] Hinden, R. and S. Deering, "IP Version 6 Addressing Architecture", RFC 4291, February 2006.

[RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, September 2007.

[RFC6939] Halwasia, G., Bhandari, S., and W. Dec, "Client Link-Layer Address Option in DHCPv6", RFC 6939, May 2013.

[I-D.ietf-nvo3-arch]
Black, D., Hudson, J., Kreeger, L., Lasserre, M., and T. Narten, "An Architecture for Overlay Networks (NVO3)", draft-ietf-nvo3-arch-03 (work in progress), March 2015.

[I-D.ietf-nvo3-nve-nva-cp-req]
Kreeger, L., Dutt, D., Narten, T., and D. Black, "Network Virtualization NVE to NVA Control Protocol Requirements", draft-ietf-nvo3-nve-nva-cp-req-03 (work in progress), October 2014.

11.2. Informative References

[RFC7348] Mahalingam, M., Dutt, D., Duda, K., Agarwal, P., Kreeger, L., Sridhar, T., Bursell, M., and C. Wright, "Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", RFC 7348, August 2014.

[I-D.sridharan-virtualization-nvgre] Garg, P. and Y. Wang, "NVGRE: Network Virtualization using Generic Routing Encapsulation", draft-sridharan-virtualization-nvgre-08 (work in progress), April 2015.

Authors' Addresses

Behcet Sarikaya
Huawei USA
1700 Alma Dr. Suite 500
Plano, TX 75075

Phone: +1 972-509-5599
Email: sarikaya@ieee.org

Frank Xia
Huawei USA
Nanjing, China

Phone: +1 972-509-5599
Email: xiayangsong@huawei.com

Shi Fan
China Telecom
Room 708, No.118, Xizhimennei Street
Beijing , P.R. China 100035

Phone: +86-10-58552140
Email: shifan@ctbri.com.cn