Network Working Group
Internet-Draft
Expires: March 30, 2014

B. Sarikaya
F. Xia
Huawei USA

September 26, 2013

DHCP Options for Configuring Multicast Addresses in VXLAN
draft-sarikaya-dhc-vxlan-multicast-02.txt

Abstract

This document defines DHCPv4 and DHCPv6 options for assigning multicast addresses for the Tunnel End Point in the Virtual eXtensible Local Area Network (VXLAN) environments. New DHCP options are defined which allow a VXLAN Tunnel End Point to request any source multicast address for the newly created virtual machine, the address of the Rendezvous Point (RP) and possibly address(es) for the virtual machine.

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). 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 March 30, 2014.

Copyright Notice

Table of Contents

1. Introduction

Data center networks are being increasingly used by telecom operators as well as by enterprises. Currently 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.

What we need is IP based tunneling scheme based overlay network called Virtual eXtensible Local Area Network (VXLAN). VXLAN overlays a Layer 2 network over a Layer 3 network. Each overlay, identified by the VXLAN Network Identifier (VNI). This allows up to 16M VXLAN segments to coexist within the same administrative domain [I-D.mahalingam-dutt-dcops-vxlan]. 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.

VXLAN tunnel end point (VTEP) is the hypervisor on the server which houses the VM. VXLAN encapsulation is only known to the VTEP, the VM never sees it. Also the tunneling is stateless, each MAC frame is encapsulated independent on any other MAC frame.

Instead of using UDP header, Generic Routing Encapsulation (GRE) encapsulation can be used. A 24-bit Virtual Subnet Identifier (VSID) is placed in the GRE key field. The resulting encapsulation is called Network Virtualization using Generic Routing Encapsulation (NVGRE) [I-D.sridharan-virtualization-nvgre]. Note that VSID is simlar to VNI. Although VXLAN terminology is used throughout, the protocol defined in this document applies to VXLAN as well as NVGRE.

In this document, we develop a protocol to assign multicast addresses to the VXLAN tunnel end points using Dynamic Host Configuration Protocol (DHCP). Multicast communication in VXLAN 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 [I-D.mahalingam-dutt-dcops-vxlan]

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, VTEP receives a request from the Management Node to create a Virtual Machine with a VXLAN Network Identifier.

DHCP Operation

VTEP starts DHCP state machine by sending DHCPDISCOVER message to the default router, e.g. the Top of Rack (ToR) switch. ToR switch could be DHCP server or most possibly DHCP relay with DHCP server located upstream. VTEP MUST include the VXLAN Multicast Address and VXLAN Rendezvous Point Address options defined in this document. VTEP sends the VXLAN Network Identifier in the newly defined VNI DHCP Option. DHCP server replies with DHCPOFFER message. DHCP server sends administratively scope IPv4 multicast address and RP address to VTEP. VTEP checks this message and if it sees the options it requested, DHCP server is confirmed to support the multicast address options. DHCPREQUEST message from VTEP and DHCPACK message from DHCP server complete DHCP message exchange.

VTEP as Multicast Source

After receiving the required information, the VTEP as multicast source communicates with the Rendezvous Point in order to build the multicast tree.

VTEP as Listener

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

IPv6 operation is slightly different:

Creation of a VM

In this step, VTEP receives a request from the Management Node to create a Virtual Machine with a VXLAN Network Identifier.

DHCP Operation

VTEP starts DHCP state machine by sending DHCPv6 Solicit message to the default router, e.g. the Top of Rack (ToR) switch. ToR switch could be DHCP server or most possibly DHCP relay with DHCP server located upstream. VTEP MUST include the options defined in this document.DHCP server replies with DHCPv6 Advertise message. VTEP checks this message and if it sees the options it requested, DHCP server is confirmed to support multicast address options. DHCPv6 Request message from VTEP and DHCPv6 Reply message from DHCPv6 server complete DHCP message exchange.

VTEP as Multicast Source

After receiving the required information, the VTEP as multicast source communicates with the Rendezvous Point in order to build the multicast tree.

VTEP as Listener

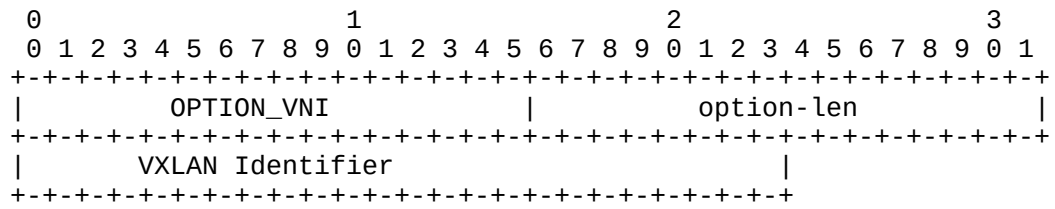
After receiving the required information, the VTEP as listener communicates with the edge router by sending MLD Report to join the multicast group.

4. DHCPv6 Options

4.1. VXLAN Network Identifier Option

Different VXLAN Network Identifiers (VNI) need different multicast groups, and even rendezvous point addresses (for load balancing). At the same time, different VNIs need different address spaces for VM, that is, two VMs belongs to different VNIs probably have the same IP address.

Because of the reasons stated above, a DHCP VNI Option is defined as follows.



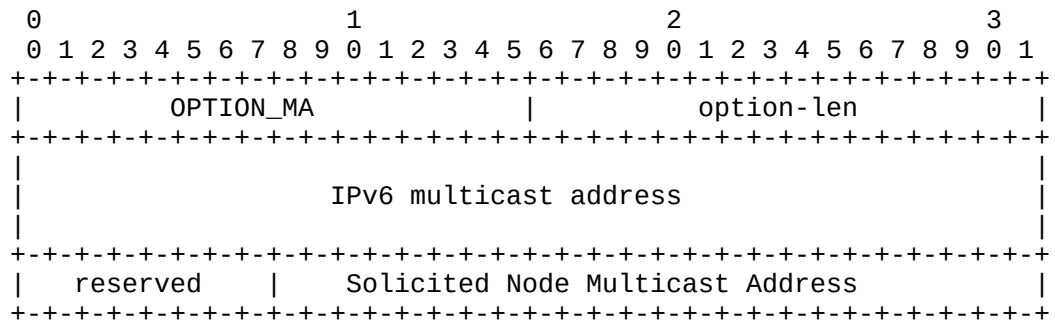
option-code OPTION_VNI (TBD).

option-len 7.

VXLAN Network Identifier 3.

4.2. IPv6 multicast address for the VNI Option

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



option-code OPTION_MA (TBD).

option-len 24.

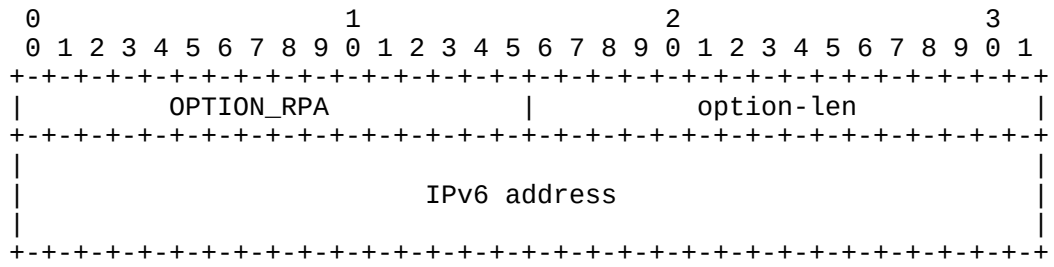
IPv6 multicast address An IPv6 address.

reserved must be set to zero

Solicited Node Multicast Address as in RFC 4861.

4.3. IPv6 Rendezvous Point Address Option

The option allows the VTEP to receive RP address for Any Source Multicast group from DHCP server.



option-code OPTION_RPA (TBD).

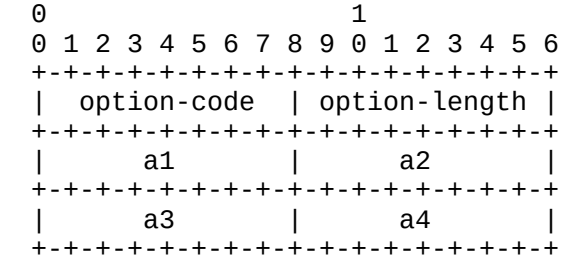
option-len 20.

IPv6 multicast address An IPv6 address

5. DHCPv4 Options

5.1. VXLAN Network Identifier Option

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



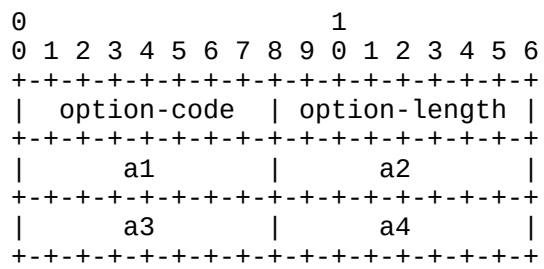
Option-code
VXLAN Network Identifier Option (TDB)

Option-len
4.

a1-a4
VTEP as DHCP Client sets a1-a3 to VNI and a4 to zero.

5.2. VXLAN Multicast Address Option

The option allows the VTEP to send the VNI DHCP server and receive administratively scoped IPv4 multicast address.



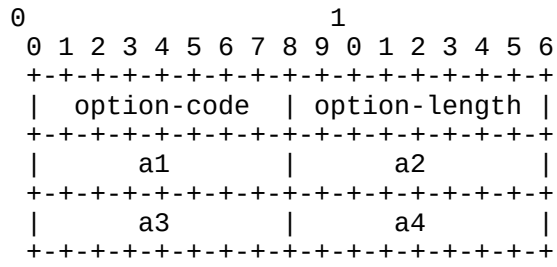
Option-code
VXLAN Multicast Address Option (TDB)

Option-len
4.

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

5.3. VXLAN Rendezvous Point Address Option

This option is used to receive VXLAN Rendezvous Point address from DHCP server.



Option-code
VXLAN Rendezvous Point Address Option (TBD)

Option-len
4.

a1-a4
VXLAN Rendezvous Point Address an IPv4 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 set VXLAN Multicast Address Option to zero. The client MUST set VXLAN Rendezvous Point Address Option to zero. The client MUST set VXLAN Network Identifier Option to the VXLAN network identifier assigned to the virtual machine.

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 set IPv6 VNI Option OPTION_VNI to the VXLAN network identifier assigned to the virtual machine.

The Client MUST set IPv6 multicast address for the VNI Option's multicast address field to zero.

The client MUST set IPv6 Rendezvous Point Address Option's IPv6 multicast address field 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

If DHCPv4 server is configured to support VXLAN multicast address assignments, it SHOULD look for VXLAN Multicast Address Option and VXLAN Rendezvous Point 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.

The server MUST set VXLAN Rendezvous Point Address Option's VXLAN Rendezvous Point Address field to IPv4 unicast address of the Rendezvous Point for the any source multicast Rendezvous Point router. How this assignment is done is out of scope.

If DHCPv6 server is configured to support VXLAN multicast address assignments it SHOULD look for IPv6 multicast address for the VNI Option and IPv6 Rendezvous Point 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. 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.

The server MUST assign IPv6 Rendezvous Point Address Option's IPv6 address field to the Rendezvous Point router's address in charge of this multicast group. The unicast address MUST BE assigned according to the rules defined in [RFC3956].

8. Security Considerations

The security considerations in [RFC2131], [RFC2132] and [RFC3315] apply. Special considerations in [I-D.mahalingam-dutt-dcops-vxlan] are also applicable.

9. IANA considerations

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

10. Acknowledgements

The authors are grateful to Bernie Volz 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.

11.2. Informative References

[I-D.mahalingam-dutt-dcops-vxlan]
Mahalingam, M., Dutt, D., Duda, K., Agarwal, P., Kreeger, L., Sridhar, T., Bursell, M., and C. Wright, "VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", draft-mahalingam-dutt-dcops-vxlan-04 (work in progress), May 2013.

[I-D.sridharan-virtualization-nvgre]
Sridharan, M., Greenberg, A., Wang, Y., Garg, P., Venkataramiah, N., Duda, K., Ganga, I., Lin, G., Pearson, M., Thaler, P., and C. Tumuluri, "NVGRE: Network Virtualization using Generic Routing Encapsulation", draft-sridharan-virtualization-nvgre-03 (work in progress), August 2013.

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