Central Directory Approach for Mapping VTEP IP Address to VM MAC/IP address in VXLAN
draft-sarikaya-nvo3-dhc-vxlan-centraldir-mapping-00.txt

Abstract

This document proposes a central database for the address resolution and neighbor discovery protocols in Virtual eXtensible Local Area Network or VXLAN environments. An entry is added to the database when a virtual machine is created and an IP address is assigned. When a hosted Virtual Machine makes an ARP/ND Request, the source Virtual VXLAN tunnel end point, after searching the central database, sends Virtual Machine’s address resolution and neighbor discovery replies in unicast to the hosted Virtual machine. The document also defines DHCPv4/v6 options for DHCPv4/v6 ARP/ND Directory Server IP Address and VXLAN Network Identifier.

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 August 18, 2014.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of
publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect 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 ................................................. 3
2. Terminology .................................................. 4
3. Overview of the protocol ..................................... 4
4. DHCP Options .................................................. 6
   4.1. VXLAN Network Identifier Option ......................... 6
   4.2. DHCPv6 ARP/ND Directory Server IP Address Option ..... 6
   4.3. DHCPv4 ARP/ND Directory Server IP Address Option ..... 7
5. Directory Lookup Operation .................................. 8
6. Creating and Maintaining Directory Operation ............ 8
7. Security Considerations ...................................... 9
8. IANA considerations .......................................... 9
9. Acknowledgements ............................................ 9
10. References .................................................. 9
   10.1. Normative References .................................. 9
   10.2. Informative References ................................. 10
Authors’ Addresses ............................................. 11
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 multi-tenancy, 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 is 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 similar to VNI. Although VXLAN terminology is used throughout, the protocol defined in this document applies to VXLAN as well as NVGRE.

In VXLAN, after hosts are configured, they start communication with external hosts and servers and Address Resolution Protocol (ARP) [RFC0826] in IPv4 and Neighbor Discovery (ND) [RFC4861] in IPv6 which are broadcast/multicast based are used to map the destination VXLAN tunnel end point (VTEP) IP address into the destination Virtual Machine (VM) MAC address.

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).

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. If 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. This issue is addressed in our document [sarikaya-nvo3-dhc-vxlan-multicast].

In VXLAN, the hosts are connected to a potentially large link and on such a network, broadcast/multicast communication may slow down the network operation. Also the underlying network may not support multicast. In those cases, there is merit in using a complimentary approach, i.e. having a central directory server that keeps all IP address/MAC address mappings and the hosts can send their ARP/ND Request messages in unicast to the directory server and get a reply from the server.

In this document, we develop a protocol to build a centralized directory for mapping VXLAN tunnel end point (VTEP) IP address to Virtual Machine (VM) MAC address. We consider two approaches: static versus dynamic. Static mapping is possible with the Virtual Machine (VM) Management Center that is responsible for the creation, configuration and Mobility of VMs. Such a VM management center can assign VM MAC and VTEP IP addresses, and it can also populate the ARP/ND directory with its configuration. However, dynamic approach is more desirable to creating a central database for the Address Resolution Protocol (ARP) or Neighbor Discovery protocols due to the dynamic creation/ deletion of VMs.

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] and [I-D.ietf-nvo3-arch].

3. Overview of the protocol

The steps involved in the protocol are explained below:
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 and a MAC address.

DHCPv4 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 Directory Server IP Address option 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 VM and server IP addresses to VTEP. VTEP checks this message and if it sees the options it requested, DHCP server is confirmed to support the directory address option. DHCPREQUEST message from VTEP and DHCPACK message from DHCP server complete DHCP message exchange.

DHCPv6 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 directory server address options. DHCPv6 Request message from VTEP and DHCPv6 Reply message from DHCPv6 server complete DHCP message exchange.

Updating the server

VTEP registers VM IP address, VXLAN Network Identifier and MAC address in the ARP/ND Directory Server. The server keeps a central directory of all VMs in VXLAN so that it can reply to the requests from VM’s VTEP.

ARP/ND with the server

After IP address configuration VM starts communication with other hosts. VM sends its ARP Request or Neighbor Solicitation messages to its VTEP. VTEP queries ARP/ND Directory Server. The result of a query includes destination VTEP, destination VM IP and destination VM MAC address. VTEP constructs an ARP/ND reply packet with the search results and sends the packet to the hosted...
VM. Hosted VM is now ready to communicate with the destination VM.

4. DHCP Options

4.1. VXLAN Network Identifier Option

Different VXLAN Network Identifiers (VNI) need different address spaces for VM, that is, two VMs belongs to different VNIIs probably have the same IP address.

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

```
+-----------------+-----------------+-----------------+-----------------+
| OPTION_VNI     | option-len      | VXLAN Identifier|
+-----------------+-----------------+-----------------+
```

option-code  OPTION_VNI (TBD).
option-len    7.

VXLAN Network Identifier 3.

4.2. DHCPv6 ARP/ND Directory Server IP Address Option

The option allows the VTEP to receive ARP/ND directory server IPv6 address. This option is used when VTEP makes a DHCP Request to receive Virtual Machine IPv6 address.
option-code  OPTION_DSA (TBD).

option-len  20.

IPv6 address  An IPv6 address.

4.3.  DHCPv4 ARP/ND Directory Server IP Address Option

The option allows the VTEP to receive ARP/ND directory server IPv4 address. This option is used when VTEP makes a DHCP Request to receive Virtual Machine IPv4 address.

0 1
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| option-code | option-length |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| a1 | a2 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| a3 | a4 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

Option-code  VXLAN ARP/ND Directory Server Address Option (TDB)

Option-len  4.

a1-a4

VTEP as DHCP Client sets a1-a4 to zero, DHCP server sets a1-a4 to the ARP/ND Directory Server address.
5. Directory Lookup Operation

The steps involved in the directory lookup are explained below.

Creation of a ARP/ND Packet

In this step, a newly created VM sends out an ARP/ND packet with multicast (broadcast) destination address.

Directory Receives Search Request

The VTEP captures the packet and extracts the IP address. VTEP sends an LDAP Search Request to the directory server using the directory server IPv4/v6 address for the IP address and VXLAN Network Identifier.

VTEP Sends Reply to ARP/ND Request

VTEP receives the results of the search, i.e. destination VM MAC address, destination VTEP IP address, destination VM IPv4/v6 address in a Search Result message from the directory server. Source VTEP constructs an ARP Reply/ Neighbor Advertisement message with the VM IPv4/v6 address and sends it to the hosted VM.

VM Receives ARP/ND Reply

VM receives ARP/ND Reply from the source VTEP and starts communication with the destination VM.

6. Creating and Maintaining Directory Operation

The directory is created and maintained dynamically. VTEPs use DHCP and Lightweight Directory Access Protocol (LDAP) in this process [RFC4511].

VMs are created dynamically as needed in the data center by the VTEP. VTEP which hosts the VM sends a DHCPv4 DHCPDISCOVER or DHCPv6 Solicit message to get an IPv4/v6 address for the VM.

VTEP MUST add VXLAN Network Identifier Option and DHCPv4/v6 ARP/ND Directory Server IP Address Option to DHCP message in order to receive the directory server IP address. After VTEP received the Directory Server IP Address after the creation of the first VM, it MAY not include DHCPv4/v6 ARP/ND Directory Server IP Address Option to DHCP message.
VTEP now must update the directory server to record IP address of VM along with MAC address of VM, VTEP IP v4/v6 address and VM IPv4/v6 address. VTEP uses LDAP for this purpose. VTEP as the directory client sends an AddRequest to the directory server. VTEP MUST receive an AddResponse with success from the server.

When VM is shut down, VTEP must delete the directory entry for this VM. VTEP deletes the corresponding entry using LDAP. VTEP as the directory client sends an DelRequest to the directory server. VTEP MUST receive an DelResponse with success from the server.

7. Security Considerations

The security considerations in [RFC2131], [RFC2132] and [RFC3315] apply. Special considerations in [I-D.mahalingam-dutt-dcops-vxlan] are also applicable. [RFC4513]

8. IANA considerations

IANA is requested to assign the OPTION_VNI and OPTION_DSA and VXLAN Network Identifier and ARP/ND Directory Server IP Address Option Codes in the registry maintained for DHCPv4 and DHCPv6.

9. Acknowledgements

10. References

10.1. Normative References


10.2. Informative References


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