Network Working Group B. Sarikaya
Internet-Draft F. Xia
Expires: May 10, 2014 Huawei USA
November 06, 2013

Virtual eXtensible Local Area Network over IEEE 802.1Qbg
draft-sarikaya-proxy-vxlan-00.txt

Abstract

In data centers there is interest in offloading network functions to the switches in order to keep the server focused on computation not networking. IEEE 802.1Qbg or Virtual Ethernet Port Aggregator (VEPA) at the hypervisor simply forces each VM frame sent out to the external switch regardless of destination. In this case, the eXtensible Local Area Network operation or proxying at a higher level switch is needed. Communication functions of the eXtensible Local Area Network are moved above to the Top of Rack switches. Top of Rack switch encapsulates the packets and directs them to their destination. Packets from the eXtensible Local Area Network servers are decapsulated before sending it to the destination proxy servers.

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 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.

One deployment strategy for VXLAN is to upgrade data center server hypervisors for VXLAN compatibility. Data center servers that can not be upgraded can also be given VXLAN capability using proxying. For proxying to work, IEEE 801.1Qbg [IEEE802.1Qbg] or Virtual Ethernet Port Aggregator (VEPA) functionality is needed in legacy server hypervisors.

In a virtual server environment the most common way to provide Virtual Machine (VM) switching connectivity is a Virtual Ethernet Bridge (VEB) or a vSwitch. VEP acts similar to a Layer 2 hardware switch providing inbound/outbound and inter-VM communication. VEP aggregates multiple VMs traffic across a set of links as well as provides frame delivery between VMs based on MAC address. .

However VEB lacks network management, monitoring and security, IEEE 801.1Qbg or Virtual Ethernet Port Aggregator (VEPA) provides a simple solution. VEPA simply sends each VM frame out to the external switch regardless of destination to be handled by an external switch, i.e. Proxy VXLAN switch.

VXLAN is a server-based network virtualization solution, and hypervisors are responsible for all networking work. At the same time, IEEE 801.1Qbg [IEEE802.1Qbg] follows a total different philosophy that servers should do as little as possible networking job, and it defines a way for virtual switches to send all traffic and forwarding decisions to the adjacent physical switch. This removes the burden of VM forwarding decisions and network operations from the host CPU. It also leverages the advanced management capabilities in the access or aggregation layer switches.

In this document, we develop Proxy VXLAN switch behavior.

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. Problem Statement

In a virtual server environment the most common way to provide Virtual Machine (VM) switching connectivity is a Virtual Ethernet Bridge (VEB) or a vSwitch. VEB acts similar to a Layer 2 hardware switch providing inbound/outbound and inter-VM communication. VEP aggregates multiple VMs traffic across a set of links as well as provides frame delivery between VMs based on MAC address. There are a number of disadvantages for VEB solution. First of all, vSwitch consumes valuable CPU and memory bandwidth. The higher the traffic load, the greater the number of CPU and memory cycles required to move traffic through the vSwitch, reducing the ability to support larger numbers of VMs in a physical server. Secondly, the solution lacks network-based visibility.vSwitches have a limited feature set. They don't provide local traffic visibility or have capabilities for enterprise data monitoring, security, or network management. Finaly, it lacks network policy enforcement. Modern external switches have many advanced features such as port security, quality of service (QoS), and access control lists (ACL). But vSwitches often do not have, or have limited support for such features.

To solve the management challenges with VEBs, Edge Virtual Bridging (EVB) in the IEEE 802.1Qbg standard was proposed. The primary goals of EVB are to combine the best of software and hardware vSwitches with the best of external L2 network switches. EVB is based on VEPA (Virtual Ethernet Port Aggregator) technology. It is a way for virtual switches to send all traffic and forwarding decisions to the adjacent physical switch. This removes the burden of VM forwarding decisions and network operations from the host CPU. It also leverages the advanced management capabilities in the access or aggregation layer switches.

VXLAN idea is mainly developed on vSwitch not IEEE 802.1Qbg. To accommodate IEEE 802.1Qbg, Proxy VXLAN is proposed in this document.

4. Proxy VXLAN Architecture

Proxy VXLAN is composed of servers that can host virtual machines. The servers are not involved in any communications required by VXLAN. This function is moved to the switches above the server. Top of Rack switches are examples of switches that can host proxy functions, i.e. VXLAN Tunnel End Point, VTEPs. Servers support IEEE 802.1Qbg.

VTEPs receive raw packets from the servers and send packets upstream after VXLAN encapsulation. Proxy VXLAN is assumed to be connected to VXLAN enabled servers. VTEP in a Proxy VLAN architecture always tags the outgoing frames to let VXLAN enabled servers know that these frames are proxied.

A given ToR switch hosting VTEP can serve one of more legacy servers. Virtual machine creation/deletion is done by the management center.

5. Overview of the protocol

The steps involved in the protocol are explained below:

Encapsulation/Decapsulation of Frames

In a hybrid Proxy VXLAN, when a frame is received on the VXLAN connected interface, the proxy switch decapsulates the frame and forwards the packet to the non VXLAN server. When an incoming frame from the non-VXLAN interface is received, the proxy switch encapsulates it and forwards it to the VXLAN server.

Virtual Machine Creation

Virtual machine creation is initiated by the management center. The management center notifies a given non-VXLAN server to create a VM. The center assigned a MAC address and VXLAN Network Identifier to the VM.

VTEP Notification Management Center notifies the ToR switch that is responsible for the server of this newly created virtual machine. The center sends MAC address, VNI of the virtual machine to the ToR switch which will act as the VTEP for this VM.

Virtual Machine Operation Virtual machine execution usually starts with ARP/ND Request to get the IP address of the destination virtual machine. After ARP/ND, virtual machine enters into IP communication with the destination virtual machine.

```
Core switches          +----+  +----+  +----+
                       |    |  |    |  |    |
```

```
                          |    |  |    |  |    |
                          +----+  +--+-+  +---\+
   Management             --           |         \\
   Center             ---            |            \
                   --               |             \
                ---                 |              \\
              --                   |                 \
         +----+                  +-+--+            +--\-+
         |    | ToR switch       |    |ToR switch  |    |ToR switch
         |    |--VTEP1           |    |VTEP2       |    |VTEP3
         +----+  -+          +---+----+\           +-/--+
           +--+   |     +--+-|     +--+ \\     +--+//
           |  +---|     |  |       |  +---\    |  /
           +--+         +--+       +--+        +--+
            |            |          |           |
        +---------+ +---------+ +---------+ +---------+
        |+--+ +--+| |+--+ +--+| |+--+ +--+| |+--+ +--+|
        ||vm| |vm|| ||vm| |vm|| ||vm| |vm|| ||vm| |vm||
        |+--+ +--+| |+--+ +--+| |+--+ +--+| |+--+ +--+|
        +---------+ +---------+ +---------+ +---------+
```

Figure 1: Proxy VXLAN Architecture

6. Encapsulation/Decapsulation Operation

In a hybrid Proxy VXLAN, when a frame is received on the VXLAN connected interface, the proxy switch removes the VXLAN header. It checks the destination MAC address of the inner Ethernet frame and forwards the packet to a physical port based on this MAC address. When an incoming frame from the non-VXLAN interface is received, the proxy switch first adds a VXLAN header. VXLAN Network ID (VNI) is set to the value which is provided by the management center. I Flag is set to 1. A new flag, P flag is set to 1 to indicate that this frame is coming from Proxy VXLAN.

Source port is assigned by the proxy switch. Destination port is set to 4789. UDP checksum is set to zero.

Source IPv4/v6 address is the proxy switch IPv4/v6 address. Destination IPv4/v6 address is obtained based on the inner destination MAC address. It is a multicast address if the incoming frame belongs to ARP/ND or multicast communication. Otherwise the proxy switch looks up its ARP/ND cache to find the IP address corresponding to the inner destination MAC address and places the result in the destination IPv4/IPv6 address.

When an incoming frame from the non-VXLAN interface is received, the proxy switch checks if the destination is within the host (another VM

in the same VLAN). In that case the frame is forwarded back down the port it was received on.

7. Virtual Machine Creation

Virtual machines are created by the management center. The management center creates a virtual machine, assigns a server to it. Usually each server may host more than one virtual machine.

When a virtual machine is created, the management center assigns a MAC address and its VXLAN Network Identifier. The center sends MAC address and VNI to the server.

8. VXLAN Tunnel Endpoint Notification

At the time when the virtual machine is created, the management center also notifies the ToR switch hosting a VTEP that is responsible for the server in which the VM was created. ToR switch receives MAC address and VNI value for this virtual machine.

ToR switch is responsible for keeping all MAC address/VNI values for each virtual machine that it serves. These values are used in encapsulating the packets coming from the virtual machines and in virtual machine operation.

9. Virtual Machine Mobility and Operation

In Proxy VXLAN, virtual machine mobility can be achieved using the following steps:

Step 1. Source VTEP is notified with destination VTEP of the moving VM,

Step 2 Source VTEP tunnels all packets for the VM to destination VTEP

Step 3 When the VM is ready, it would send gratuitous ARP to all VMs

Step 4 When the source VTEP receives the gratuitous ARP, it removes the VM MAC from its original forwarding table and stops tunneling for this virtual machine .

When a VM is created or after VM is moved, VM starts its operation, e.g. by sending ARP/ND packets. Non-VXLAN server sends the packet to the upstream switch which finally reaches the ToR switch hosting the VTEP[architecture]. VTEP normally converts this packet, i.e. broadcast packet into a multicast packet, encapsulates it and sends it out to VXLAN enabled servers. How ARP/ND packets are processed is out of scope.

10. Security Considerations

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

11. 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.

12. Acknowledgements

13. References

13.1. Normative References

[RFC0826] Plummer, D., "Ethernet Address Resolution Protocol: Or converting network protocol addresses to 48.bit Ethernet address for transmission on Ethernet hardware", STD 37, RFC 826, November 1982.

[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.

[RFC4511] Sermersheim, J., "Lightweight Directory Access Protocol (LDAP): The Protocol", RFC 4511, June 2006.

[RFC4513] Harrison, R., "Lightweight Directory Access Protocol (LDAP): Authentication Methods and Security Mechanisms", RFC 4513, June 2006.

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

[IEEE802.1Qbg]

IEEE, "Edge Virtual Bridging", IEEE Std 802.1Qbg-2012, May 2012.

13.2. 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-05 (work in progress), October 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
5340 Legacy Dr. Building 3
Plano, TX 75024

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

Frank Xia
Huawei USA
Nanjing, China

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