Network Working Group
Internet Draft
Intended Status: Experimental
Expires: February 2012

M.Mahalingam
VMware
D.Dutt
Cisco
K.Duda
Arista
P.Agarwal
Broadcom
L. Kreeger
Cisco
T. Sridhar
VMware
M.Bursell
Citrix
C.Wright
Red Hat

August 26, 2011

VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks

draft-mahalingam-dutt-dcops-vxlan-00.txt

Status of this Memo

Abstract

This document describes Virtual eXtensible Local Area Network (VXLAN), which is used to address the need for overlay networks within virtualized data centers accommodating multiple tenants. The scheme and the related protocols can be used in cloud service provider and enterprise data center networks.

Table of Contents

1. Introduction

Server virtualization has placed increased demands on the physical network infrastructure. At a minimum, there is a need for more MAC address table entries throughout the switched Ethernet network due to potential attachment of hundreds of thousands of Virtual Machines (VMs), each with its own MAC address.

Second, the VMs may be grouped according to their Virtual LAN (VLAN). In a data center one might need thousands of VLANs to partition the traffic according to the specific group that the VM may belong to. The current VLAN limit of 4094 is inadequate in such situations. A related requirement for virtualized environments is having the Layer 2 network scale across the entire data center or even between data centers for efficient allocation of compute, network and storage resources. Using traditional approaches like Spanning Tree Protocol (STP) for a loop free topology can result in a large number of disabled links in such environments.

Another type of demand that is being placed on data centers is the need to host multiple tenants, each with their own isolated network domain. This is not economical to realize with dedicated infrastructure, so network administrators opt to implement this over a shared network. A concomitant problem is that each tenant may independently assign MAC addresses and VLAN IDs leading to potential duplication of these on the physical network.

The last scenario is the case where the network operator prefers to use IP for interconnection of the physical infrastructure (e.g. to achieve multipath scalability through Equal Cost Multipath [ECMP]) while still preserving the Layer 2 model for inter-VM communication.

The scenarios described above lead to a requirement for an overlay network. This overlay would be used to carry the MAC traffic from the individual VMs in an encapsulated format over a logical "tunnel".

This document details a framework termed Virtual eXtensible Local Area Network (VXLAN) which provides such an encapsulation scheme to address the various requirements specified above.

1.1. Acronyms & Definitions

ACL - Access Control List

ECMP - Equal Cost Multipath

IGMP - Internet Group Management Protocol

PIM - Protocol Independent Multicast

STP - Spanning Tree Protocol

ToR - Top of Rack

TRILL - Transparent Interconnection of Lots of Links

VXLAN - Virtual eXtensible Local Area Network

VXLAN Segment - VXLAN Layer 2 overlay network over which VMs communicate

VXLAN Overlay Network - another term for VXLAN Segment

VXLAN Gateway - an entity which forwards traffic between VXLAN and non-VXLAN environments

VTEP - VXLAN Tunnel End Point - an entity which originates or terminates VXLAN tunnels

VLAN - Virtual Local Area Network

VM - Virtual Machine

VNI - VXLAN Network Identifier (or VXLAN Segment ID)

2. Conventions used in this document

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 RFC-2119 [RFC2119].

In this document, these words will appear with that interpretation only when in ALL CAPS. Lower case uses of these words are not to be interpreted as carrying RFC-2119 significance.

3. VXLAN Problem Statement

This section details the problems that VXLAN is intended to address. The focus is on the networking infrastructure within the data center and the issues related to them.

3.1. Limitations imposed by Spanning Tree & VLAN Spaces

Current Layer 2 networks use the Spanning Tree Protocol (STP) to avoid loops in the network due to duplicate paths. STP will turn off links to avoid the replication and looping of frames. Some data center operators see this as a problem with Layer 2 networks in general since with STP they are effectively paying for more ports and links than they can use. In addition, resiliency due to multipathing is not available with the STP model. Newer initiatives like TRILL are being proposed to help with multipathing and thus surmount some of the problems with STP. STP limitations may be avoided by configuring servers within a rack to be on the same Layer 3 network with switching happening at Layer 3 both within the rack and between racks. However, this is incompatible with a Layer 2 model for inter-VM communication.

Another characteristic of Layer 2 data center networks is their use of Virtual LANs (VLANs) to provide broadcast isolation. A 12 bit VLAN ID is used in the Ethernet data frames to divide the larger Layer 2 network into multiple broadcast domains. This has served well for several data centers which are limited to less than 4094 VLANs. With the growing adoption of virtualization, this upper limit is seeing pressure. Moreover, due to STP, several data centers limit the number of VLANs that could be used. In addition, requirements for multitenant environments accelerate the need for larger VLAN limits, as discussed in Section 3.3.

3.2. Multitenant Environments

Cloud computing involves on demand elastic provisioning of resources for multitenant environments. The most common example of cloud computing is the public cloud, where a cloud service provider offers these elastic services to multiple customers over the same infrastructure.

Isolation of network traffic by tenant could be done via Layer 2 or Layer 3 networks. For Layer 2 networks, VLANs are often used to segregate traffic - so a tenant could be identified by its own VLAN, for example. Due to the large number of tenants that a cloud provider might service, the 4094 VLAN limit is often inadequate. In

addition, there is often a need for multiple VLANs per tenant, which exacerbates the issue.

Another use case is cross pod expansion. A pod typically consists of one or more racks of servers with its associated network and storage connectivity. Tenants may start off on a pod and, due to expansion, require servers/VMs on other pods, especially the case when tenants on the other pods are not fully utilizing all their resources. This use case requires a "stretched" Layer 2 environment connecting the individual servers/VMs.

Layer 3 networks are not a complete solution for multi tenancy either. Two tenants might use the same set of Layer 3 addresses within their networks which requires the cloud provider to provide isolation in some other form. Further, requiring all tenants to use IP excludes customers relying on direct Layer 2 or non-IP Layer 3 protocols for inter VM communication.

3.3. Inadequate Table Sizes at ToR Switch

Today's virtualized environments place additional demands on the MAC address tables of Top of Rack (ToR) switches which connect to the servers. Instead of just one MAC address per server link, the ToR now has to learn the MAC addresses of the individual VMs (which could range in the 100s per server). This is a requirement since traffic from/to the VMs to the rest of the physical network will traverse the link to the switch. A typical ToR switch could connect to 24 or 48 servers depending upon the number of its server facing ports. A data center might consist of several racks, so each ToR switch would need to maintain an address table for the communicating VMs across the various physical servers. This places a much larger demand on the table capacity compared to non-virtualized environments.

If the table overflows, the switch may stop learning new addresses until idle entries age out, leading to significant flooding of unknown destination frames.

4. Virtual eXtensible Local Area Network (VXLAN)

VXLAN (Virtual eXtensible Local Area Network) addresses the requirements of Layer 2 and Layer 3 data center network infrastructure in the presence of VMs in a multitenant environment. It runs over the existing networking infrastructure and provides a means to "stretch" a Layer 2 network. In short, VXLAN is a Layer 2 overlay scheme over a Layer 3 network. Each overlay is termed a

VXLAN segment. Only VMs within the same VXLAN segment can communicate with each other. Each VXLAN segment is scoped through a 24 bit segment ID, hereafter termed the VXLAN Network Identifier (VNI). This allows up to 16M VXLAN segments to coexist within the same administrative domain.

The VNI scopes the inner MAC frame originated by the individual VM. Thus, you could have overlapping MAC addresses across segments but never have traffic "cross over" since the traffic is isolated using the VNI qualifier. This qualifier is in an outer header envelope over the inner MAC frame originated by the VM. In the following sections, the term "VXLAN segment" is used interchangeably with the term "VXLAN overlay network".

Due to this encapsulation, VXLAN could also be termed a tunneling scheme to overlay Layer 2 networks on top of Layer 3 networks. The tunnels are stateless, so each frame is encapsulated according to a set of rules. The end point of the tunnel (VTEP) discussed in the following sections is located within the hypervisor on the server which houses the VM. Thus, the VNI and VXLAN related tunnel/outer header encapsulation are known only to the VTEP - the VM never sees it (see Figure 1). Note that it is possible that VTEPs could also be on a physical switch or physical server and could be implemented in software or hardware. One use case where the VTEP is a physical switch is discussed in Section 6 VXLAN on VXLAN deployment scenarios.

The following sections discuss typical traffic flow scenarios in a VXLAN environment using one type of control scheme - data plane learning. Here, the association of VM's MAC to VTEP's IP is discovered via source learning. Multicast is used for carrying unknown destination, broadcast and multicast frames. VXLAN

4.1. Unicast VM to VM communication

Consider a VM within a VXLAN overlay network. This VM is unaware of VXLAN. To communicate with a VM on a different host, it sends a MAC frame destined to the target as before. The VTEP on the physical host looks up the VNI to which this VM is associated. It then determines if the destination MAC is on the same segment. If so, an outer header comprising an outer MAC, outer IP address and VXLAN header (see Figure 1 in Section 5 for frame format) are inserted in front of the original MAC frame. The final packet is transmitted out to the destination, which is the IP address of the remote VTEP connecting the destination VM addressed by the inner MAC destination address.

Upon reception, the remote VTEP verifies that the VNI is a valid one and is used by the destination VM. If so, the packet is stripped of its outer header and passed on to the destination VM. The destination VM never knows about the VNI or that the frame was transported with a VXLAN encapsulation.

In addition to forwarding the packet to the destination VM, the remote VTEP learns the Inner Source MAC to outer Source IP address mapping. It stores this mapping in a table so that when the destination VM sends a response packet, there is no need for an "unknown destination" flooding of the response packet.

Determining the MAC address of the destination VM prior to the transmission by the VM is performed as with non-VXLAN environments except as described below. Broadcast frames are used but are encapsulated within a multicast packet, as detailed in the next section.

4.2. Broadcast Communication and Mapping to Multicast

Consider the VM on the source host attempting to communicate with the destination VM using IP. Assuming that they are both on the same subnet, the VM sends out an ARP broadcast frame. In the non-VXLAN environment, this frame would be sent out using MAC broadcast which all switches carrying that VLAN.

With VXLAN, a header including the VXLAN VNI is inserted at the beginning of the packet along with the IP header and UDP header. However, this broadcast packet is sent out to the IP multicast group on which that VXLAN overlay network is realized. VXLAN

To realize this, we need to have a mapping between the VXLAN VNI and the IP multicast group that it will use. This mapping is done at the management layer and provided to the individual VTEPs through a management channel. Using this mapping, the VTEP can provide IGMP membership reports to the upstream switch/router to join/leave the VXLAN related IP multicast groups as needed. This will enable pruning of the leaf nodes for specific multicast traffic addresses based on whether a member is available on this host using that specific multicast address. In addition, use of multicast routing protocols like PIM will provide efficient multicast trees within the Layer 3 network.

The VTEP will use (*,G) joins. This is needed as the set of VXLAN tunnel sources is unknown and may change often, as the VMs come up/go down across different hosts. A side note here is that since

each VTEP can act as both the source and destination for multicast packets, a protocol like PIM-bidir would be more efficient.

The destination VM sends a standard ARP response using IP unicast. This frame will be encapsulated back to the VTEP connecting the originating VM using IP unicast VXLAN encapsulation. This is possible since the mapping of the ARP response's destination MAC to the VXLAN tunnel end point IP was learned earlier through the ARP request.

Another point to note is that multicast frames and "unknown MAC destination" frames are also sent using the multicast tree, similar to the broadcast frames.

4.3. Physical Infrastructure Requirements

When IP multicast is used within the network infrastructure, a multicast routing protocol like Protocol Independent Multicast Sparse Mode (PIM-SM) is used by the individual Layer 3 IP routers/switches within the network. This is used to build efficient multicast forwarding trees so that multicast trees are only sent to those hosts which have requested to receive them.

Similarly, there is no requirement that the actual network connecting the source VM and destination VM should be a Layer 3 network - VXLAN can also work over Layer 2 networks. In either case, efficient multicast replication within the Layer 2 network can be achieved using IGMP snooping.

5. VXLAN Frame Format

The VXLAN frame format is shown below. Parsing this from the bottom, there is an inner MAC frame with its own Ethernet header with source, destination MAC addresses along with the Ethernet type plus an optional VLAN tag (see Section 6 for further details of inner VLAN tag handling).

The inner MAC frame is encapsulated with the following three headers starting from the innermost header.

O VXLAN Header: This is an 8 byte field which has:

o Flags (8 bits) where the I flag MUST be set to 1 for a valid VXLAN Network ID (VNI). The remaining 7 bits (designated "R") are reserved fields and MUST be set to zero.

o VXLAN Segment ID/VXLAN Network Identifier (VNI) - this is a 24 bit value used to designate the individual VXLAN overlay network on which the communicating VMs are situated. VMs in different VXLAN overlay networks cannot communicate.

o Reserved fields (24 bits and 8 bits) - MUST be set to zero.

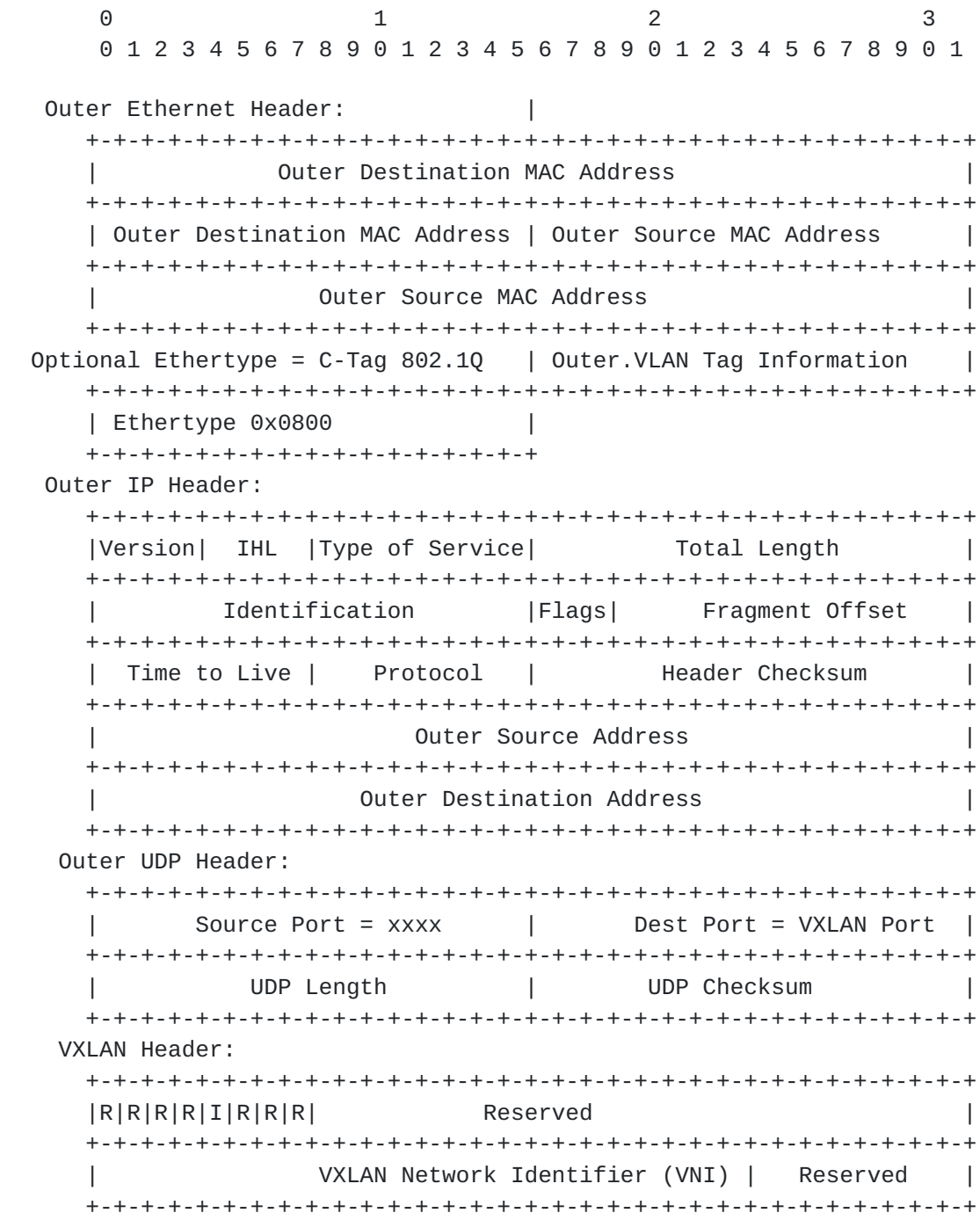
O Outer UDP Header: This is the outer UDP header with a source port provided by the VTEP and the destination port being a well known UDP port to be obtained by IANA assignment. It is recommended that the source port be a hash of the inner Ethernet frame's headers to obtain a level of entropy for ECMP/load balancing of the VM to VM traffic across the VXLAN overlay.

The UDP checksum field SHOULD be transmitted as zero. When a packet is received with a UDP checksum of zero, it MUST be accepted for decapsulation. Optionally, if the encapsulating endpoint includes a non-zero UDP checksum, it MUST be correctly calculated across the entire packet including the IP header, UDP header, VXLAN header and encapsulated MAC frame. When a decapsulating endpoint receives a packet with a non-zero checksum it MAY choose to verify the checksum value. If it chooses to perform such verification, and the verification fails, the packet MUST be dropped. If the decapsulating destination chooses not to perform the verification, or performs it successfully, the packet MUST be accepted for decapsulation.

O Outer IP Header: This is the outer IP header with the source IP address indicating the IP address of the VTEP over which the communicating VM (as depicted by the inner source MAC address) is running. The destination IP address is the IP address of the VTEP connecting the communicating VM as depicted by the inner destination MAC address.

O Outer Ethernet Header (example): Figure 1 shows an example of a common encapsulation of the entire IP packet with the VXLAN encapsulation inside an outer Ethernet header. The destination MAC

address in this frame may be the address of the target VTEP or an intermediate Layer 3 router. The outer VLAN tag is optional. If present, it may be used for delineating VXLAN traffic on the LAN.



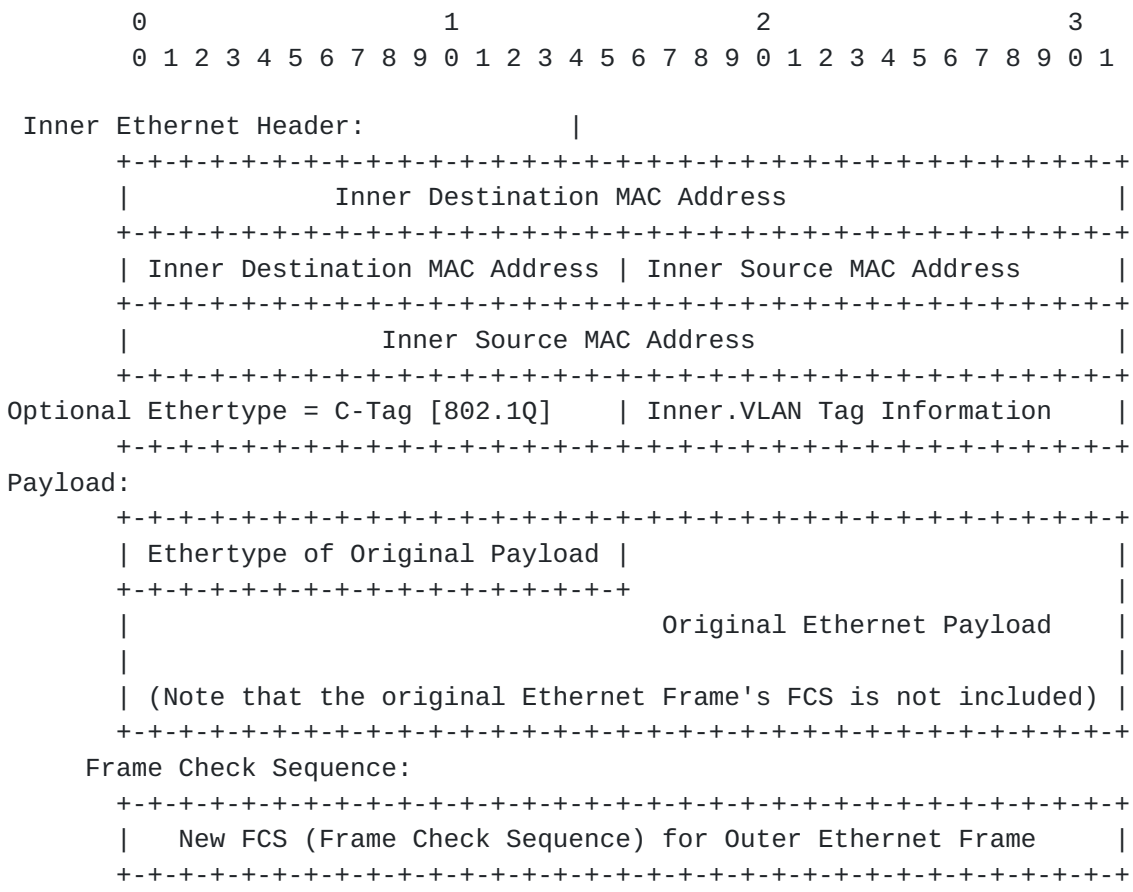


Figure 1 VXLAN Frame Format

The frame format above shows tunneling of Ethernet frames using IPv4 for transport. Use of VXLAN with IPv6 transport will be addressed in a future version of this draft.

6. VXLAN Deployment Scenarios

VXLAN is typically deployed in data centers on virtualized hosts, which may be spread across multiple racks. The individual racks may be parts of a different Layer 3 network or they could be in one Layer 2 network. The VXLAN segments/overlay networks are overlaid on top of these Layer 2 or Layer 3 networks.

Consider Figure 2 below depicting two virtualized servers attached to a Layer 3 infrastructure. The servers could be on the same rack, or on different racks or potentially across data centers within the same administrative domain. There are 4 VXLAN overlay networks identified by the VNIs 22, 34, 74 and 98. Consider the case of VM1-1 in Server 1 and VM2-4 on Server 2 which are on the same VXLAN

overlay network identified by VNI 22. The VMs do not know about the overlay networks and transport method since the encapsulation and decapsulation happen transparently at the VTEPs on Servers 1 and 2. The other overlay networks and the corresponding VMs are: VM1-2 on Server 1 and VM2-1 on Server 2 both on VNI 34, VM1-3 on Server 1 and VM2-2 on Server 2 on VNI 74, and finally VM1-4 on Server 1 and VM2-3 on Server 2 on VNI 98.

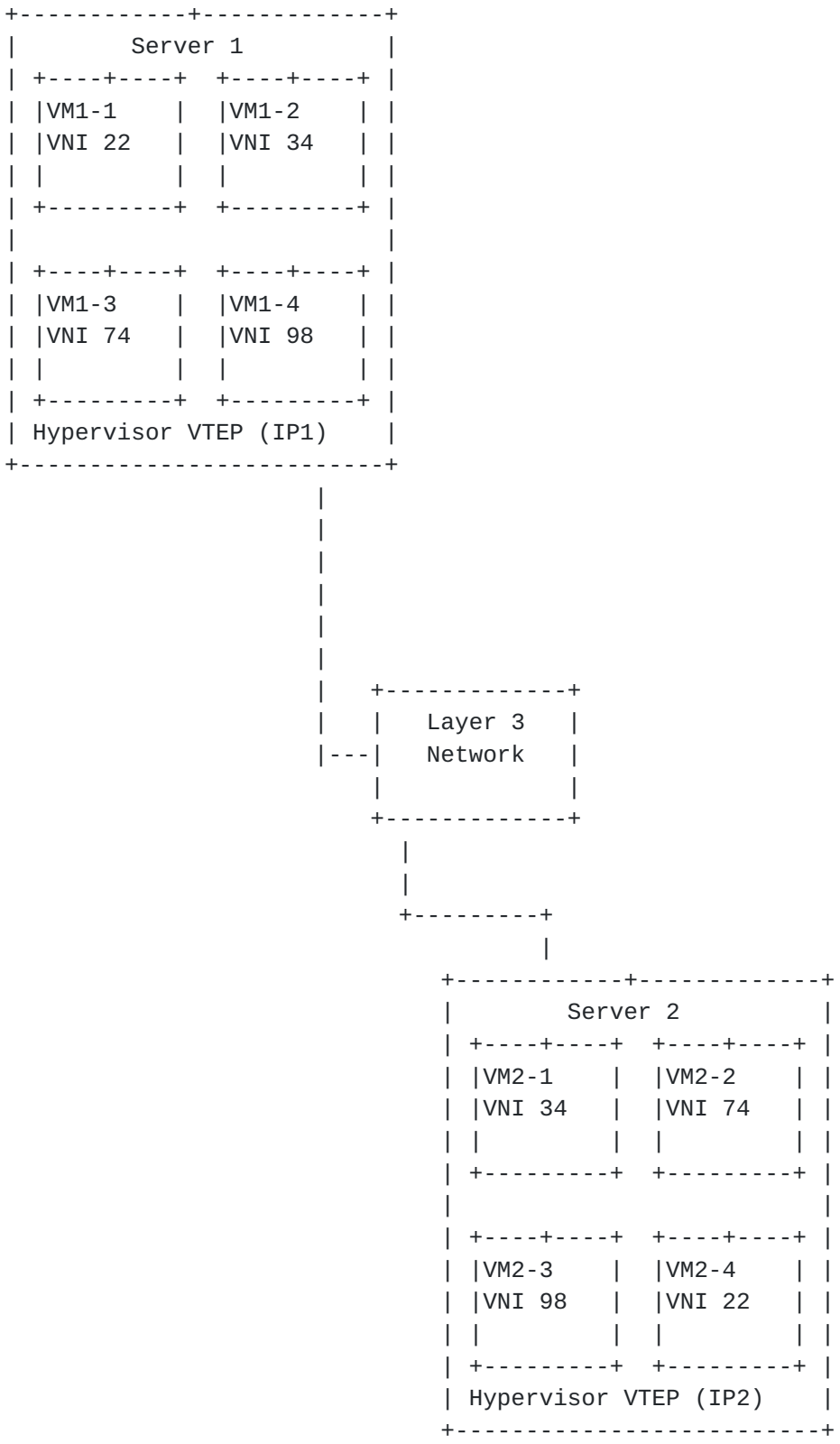

Server 1
VM1-1
VNI 22
VM1-2
VNI 34
VM1-3
VNI 74
VM1-4
VNI 98
Hypervisor VTEP (IP1)
Layer 3
Network
Server 2
VM2-1
VNI 34
VM2-2
VNI 74
VM2-3
VNI 98
VM2-4
VNI 22
Hypervisor VTEP (IP2)

Figure 2 VXLAN Deployment - VTEPs across a Layer 3 Network

One deployment scenario is where the tunnel termination point is a physical server which understands VXLAN. Another scenario is where nodes on a VXLAN overlay network need to communicate with nodes on legacy networks which could be VLAN based. These nodes may be physical nodes or virtual machines. To enable this communication, a network can include VXLAN gateways (see Figure 3 below with a switch acting as a VXLAN gateway) which forward traffic between VXLAN and non-VXLAN environments.

Consider Figure 3 for the following discussion. For incoming frames on the VXLAN connected interface, the gateway strips out the VXLAN header and forwards to a physical port based on the destination MAC address of the inner Ethernet frame. Decapsulated frames with the inner VLAN ID SHOULD be discarded unless configured explicitly to be passed on to the non-VXLAN interface. In the reverse direction, incoming frames for the non-VXLAN interfaces are mapped to a specific VXLAN overlay network based on the VLAN ID in the frame. Unless configured explicitly to be passed on in the encapsulated VXLAN frame, this VLAN ID is removed before the frame is encapsulated for VXLAN.

These gateways which provide VXLAN tunnel termination functions could be ToR/access switches or switches higher up in the data center network topology - e.g. core or even WAN edge devices. The last case (WAN edge) could involve a Provider Edge (PE) router which terminates VXLAN tunnels in a hybrid cloud environment. Note that in all these instances, the gateway functionality could be implemented in software or hardware.

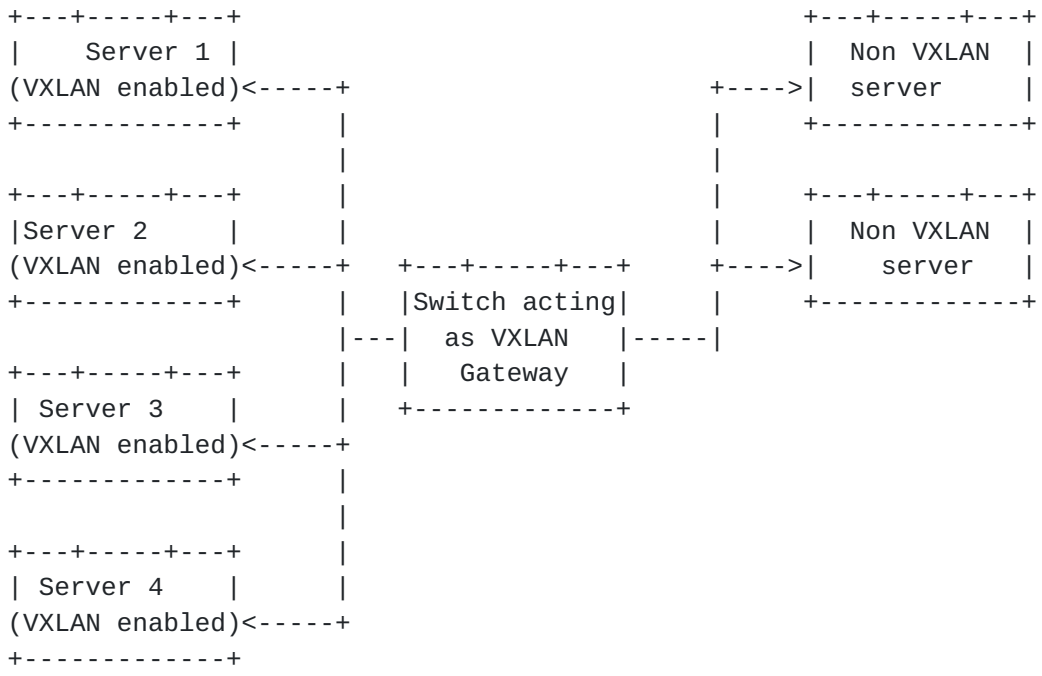


Figure 3 VXLAN Deployment - VXLAN Gateway

6.1. Inner VLAN Tag Handling

Inner VLAN Tag Handling in VTEP and VXLAN Gateway should conform to the following:

Decapsulated VXLAN frames with the inner VLAN tag SHOULD be discarded unless configured otherwise. On the encapsulation side, a VTEP SHOULD NOT include an inner VLAN tag on tunnel packets unless configured otherwise. When a VLAN-tagged packet is a candidate for VXLAN tunneling, the encapsulating VTEP SHOULD strip the VLAN tag unless configured otherwise.

7. Security Considerations

Traditionally, layer 2 networks can only be attacked from 'within' by rogue endpoints - either by having inappropriate access to a LAN and snooping on traffic or by injecting spoofed packets to 'take over' another MAC address or by flooding and causing denial of service. A MAC-over-IP mechanism for delivering Layer 2 traffic significantly extends this attack surface. This can happen by rogues

injecting themselves into the network by subscribing to one or more multicast groups that carry broadcast traffic for VXLAN segments and also by sourcing MAC-over-UDP frames into the transport network to inject spurious traffic, possibly to hijack MAC addresses.

This proposal does not, at this time, incorporate specific measures against such attacks, relying instead on other traditional mechanisms layered on top of IP. This section, instead, sketches out some possible approaches to security in the VXLAN environment.

Traditional Layer 2 attacks by rogue end points can be mitigated by limiting the management and administrative scope of who deploys and manages VMs/gateways in a VXLAN environment. In addition, such administrative measures may be augmented by schemes like 802.1X for admission control of individual end points. Also, the use of the UDP based encapsulation of VXLAN enables exploiting the 5 tuple based ACLs (Access Control Lists) functionality in physical switches.

Tunneled traffic over the IP network can be secured with traditional IPSEC mechanisms that authenticate and optionally encrypt VXLAN traffic. This will, of course, need to be coupled with an authentication infrastructure for authorized endpoints to obtain and distribute credentials.

VXLAN overlay networks are designated and operated over the existing LAN infrastructure. To ensure that VXLAN end points and their VTEPs are authorized on the LAN, it is recommended that a VLAN be designated for VXLAN traffic and the servers/VTEPs send VXLAN traffic over this VLAN to provide a measure of security.

In addition, VXLAN requires proper mapping of VNIs and VM membership in these overlay networks. It is expected that this mapping be done and communicated to the management entity on the VTEP and the gateways using existing secure methods.

8. IANA Considerations

An IANA port will be requested for the VXLAN destination UDP port.

9. Conclusions

This document has introduced VXLAN, an overlay framework for transporting MAC frames generated by VMs in isolated Layer 2 networks over an IP network. Through this scheme, it is possible to stretch Layer 2 networks across Layer 3 networks. This finds use in virtualized data center environments where Layer 2 networks may need to span across the entire data center, or even between data centers.

10. References

10.1. Normative References

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

10.2. Informative References

[RFC4601] Fenner, B., Handley, M., Holbrook, H., and Kouvelas, I., "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification", RFC 4601, August 2006.

[RFC5015] Handley, M., Kouvelas, I., Speakman, T., and Vicisano, L., "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, October 2007.

[RFC4541] Christensen, M., Kimball, K., and Solensky, F., "Considerations for Internet Group Management Protocol (IGMP) and Multicast Listener Discovery (MLD) Snooping Switches", RFC 4541, May 2006.

11. Acknowledgments

The authors wish to thank Ajit Sanzgiri for contributions to the Security Considerations section and editorial inputs, Joseph Cheng, Margaret Petrus and Milin Desai for their editorial reviews, inputs and comments.

Authors' Addresses

Mallik Mahalingam
VMware Inc.
3401 Hillview
Palo Alto, CA 94304

Email: mallik@vmware.com

Dinesh G. Dutt
Cisco Systems, Inc.
170 W. Tasman Avenue
Palo Alto, CA 94304

Email: ddutt@cisco.com

Kenneth Duda
Arista Networks
5470 Great America Parkway
Santa Clara, CA 95054

Email: kduda@aristanetworks.com

Puneet Agarwal
Broadcom Corporation
3151 Zanker Road
San Jose, CA 95134

Email: pagarwal@broadcom.com

Lawrence Kreeger
Cisco Systems, Inc.
170 W. Tasman Avenue
Palo Alto, CA 94304

Email: kreeger@cisco.com

T. Sridhar
VMware Inc.
3401 Hillview
Palo Alto, CA 94304

Email: tsridhar@vmware.com

Mike Bursell
Citrix Systems Research & Development Ltd.
Building 101
Cambridge Science Park
Milton Road
Cambridge CB4 0FY
United Kingdom

Email: mike.bursell@citrix.com

Chris Wright
Red Hat Inc.
1801 Varsity Drive
Raleigh, NC 27606

Email: chrisw@redhat.com