Network Working Group
Internet-Draft
Intended status: Experimental
Expires: September 6, 2015

F. Xia
B. Sarikaya
Huawei Technologies Co., Ltd.
S. Fan
China Telecom
March 5, 2015

Quality of Service Marking and Framework in Overlay Networks
draft-xia-nvo3-vxlan-qosmarking-04.txt

Abstract

Overlay networks such as The Virtual eXtensible Local Area Network enable multiple tenants to operate in a data center. Each tenant needs to be assigned a priority group to prioritize their traffic using tenant based quality of service marking. Also, cloud carriers wish to use quality of service to differentiate different applications, i.e. legacy, traffic based marking. For these purposes, Quality of Service bits are assigned in the Virtual eXtensible Local Area Network outer Ethernet header. How these bits are assigned and are processed in the network are explained in detail. Also the document presents a quality of service framework for overlay networks such as Virtual eXtensible Local Area Network.

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 September 6, 2015.

Table of Contents

1. Introduction

Data center networks are being increasingly used by telecom operators as well as by enterprises. An important requirement in data center networks is multitenancy, i.e. multiple tenants each with their own isolated network domain. Virtual eXtensible Local Area Network (VXLAN) is a solution that is gaining popularity in industry [RFC7348]. VXLAN overlays a Layer 2 network over a Layer 3 network. Each overlay is identified by the VXLAN Network Identifier (VNI). VXLAN tunnel end point (VTEP) can be hosted at the the hypervisor on the server or higher above in the network. VXLAN encapsulation with a UDP header is only known to the VTEP, the Virtual Machines (VM) never sees it.

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).

Since VXLAN allows multiple tenants to operate, data center operators are facing the problem of treating their traffic. There is interest to provide different quality of service to the tenants based on their service level agreements, i.e. tenant-based QoS.

Cloud carriers have interest in different quality of service to different applications such as voice, video, network control applications, etc. In this case, quality of service marking can be done using deep packet inspection (DPI) in order to detect the type of application in each packet, i.e. application based QoS.

In this document, we develop Quality of Service marking solution for VXLAN as part of the Quality of Service framework for overlay multi-tenant networks as such it complements VXLAN architecture defined in [RFC7348]. The solution is compatible with IP level Differentiated Services model or diffserv described in [RFC2474] and [RFC2475]. Configuration guidelines are described in [RFC4594]. Diffserv interconnection classes and interconnection practice are described in [I-D.geib-tsvwg-diffserv-intercon].

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].

3. Problem Statement

In an overlay network such as VXLAN network multiple tenants are supported. There is interest in assigning different priority to each tenant’s traffic based on the premium that tenant pays, etc. In another words, cloud carriers would like to categorize tenants into different traffic classes such as diamond, gold, silver and bronze classes.

Cloud carriers wish to categorize the traffic based on the application such as voice, video, etc. Based on the type of the application different traffic classes may be identified and different priority levels can be assigned to each.

In order to do these, quality of service marking is needed in the overlay network.

The solution developed in this document is based on the Network Virtualization Edge (NVE) to do the marking at the encapsulation layer. The marking, especially the tenant based QoS marking has to be done when the frames are introduced by the virtual machines (VM) because it is the encapsulation layer that adds the tenant identification, e.g. VXLAN Network Identifier (VNI) to each frame coming from the VMs. Because of this tenant based Quality of Service marking SHOULD be done at the encapsulation layer.

4. QoS Marking Schemes in VXLAN

Three bits are reserved in VXLAN Outer Ethernet Header's C-Tag 802.1Q field/VLAN Tag Information field's priority code point (PCP) field shown as QoS-flag in Figure 1.

Three bits called QoS-flag are reserved to indicate the quality of service class that this packet belongs. These bits will be assigned according to the type of traffic carried in this flow, e.g. video, voice, critical application, etc. These assignments in relation to IP level Differentiated Services model, diffserv bits or DS field, are discussed in Section 7.

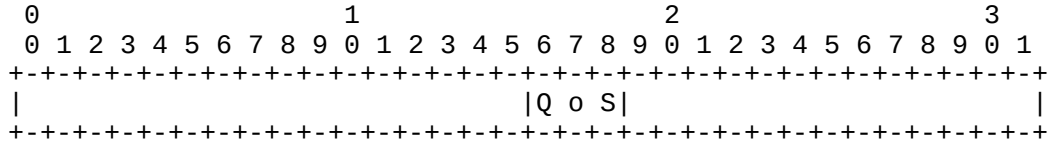

Figure 1: QoS Flag

Priority code point bits are assigned as follows in IEEE 802.1p [IEEE802.1D]:

001 - BK or background traffic

000 - BE or best effort traffic

010 - EE or Excellent Effort

011 - CA or Critical Applications

100 - VI or Video

101 - VO or Voice

110 - IC or Internetwork Control

111 - NC or Network Control

'111' has the highest priority while '001' has the lowest, for example, video traffic has higher priority than web surfing which is best effort traffic.

IEEE 802.1p [IEEE802.1D] based PCP marking is supported by most switches currently deployed that have the QoS capabilities.

We propose two different mappings to make use of the QoS field, tenant based marking or application based marking. Both of these markings are compatible with IEEE 802.1p PCP marking as well as the class selector codepoints defined by diffserv.

4.1. Tenant Based Marking

Tenant based marking is based on tenancy priorities. The cloud carrier categorizes its tenants into different groups such as diamond, gold, silver, bronze, standard and so on. All traffic for a diamond tenant has a high priority to be forwarded regardless of application types. The below is the mapping proposed in this document.

001 - Reserved

000 - Standard

010 - Bronze

011 - Silver

100 - Gold

101 - Diamond

110 - Emergency

111 - Reserved

The sender SHOULD assign bits 16-18 with bits assigned values as above if the quality of service treatment is needed on this packet. The sender SHOULD assign the same bit pattern to all the packets of the same tenant, i.e. the same VNI value.

4.2. Application Based Marking

This marking is based on application priorities. NVE uses some mechanism such as Deep Packet Inspection (DPI) to identify application types, and fills in the QoS field based on the identified application types. The below is a possible mapping.

001 - Reserved

000 - ftp/email

010 - Facebook, Twitter, Social Media

011 - Instant Message

100 - video

101 - voice

110 - High Performance computation

111 - Reserved

The sender SHOULD assign bits 16-18 with bits assigned values as above if the quality of service treatment is needed on this packet. The sender SHOULD assign the same bit pattern to all the packets of the same flow.

4.3. QoS Bits in Outer IP Header

Outer IPv4 header in VXLAN has type of service (TOS) field of 8 bits. Outer IPv6 header is VXLAN has traffic class field of 8 bits.

In case virtual machines are differentiated services capable, inner IP header contains per hop behavior (PHB) settings inline with diffserv RFCs. VXLAN NVE SHOULD copy inner IP header QoS bits into outer IP header bits.

In case virtual machines are not differentiated services capable, outer IP header QoS bits MUST be assigned by VXLAN NVE. VXLAN NVE SHOULD assign one of the class selector (CS) codepoints with value 'xxx000' corresponding to the marking explained above if the packet is a unicast packet. For more details, see Section 7.

5. Quality of Service Operation at VXLAN Decapsulation Point

There are two types of VXLAN packets receivers, that is, a VXLAN enabled NVE or a VXLAN gateway [I-D.sarikaya-nvo3-proxy-vxlan].

When the VXLAN enabled NVE receives the packet, it decapsulates the packet and delivers it downstream to a corresponding VM. If there are multiple packets to be processed, packets with high priority (that is higher QoS value) should be processed first.

The QoS operation is different for the VXLAN gateway processing. The gateway which provides VXLAN tunnel termination functions could be ToR/access switches or switches higher up in the data center network topology. 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. If inner VLAN is included in the VXLAN frame or a VLAN is supposed to be added based on configuration, the VXLAN gateway decapsulates the VXLAN packet and remarks the QoS field of the outgoing Ethernet frame based on VXLAN Outer Ethernet Header QoS bits. The switch SHOULD copy the Q-Flags of VXLAN Outer Ethernet Header into IEEE 802.1p Priory Code Point (PCP) field in VLAN tag.

6. Quality of Service Operation at VXLAN Encapsulation Point

There are two types of VXLAN packet senders, that is, a VXLAN enabled NVE or a VXLAN gateway.

For a VXLAN enabled NVE, the upstream procedure is:

Reception of Frames

The VXLAN enabled NVE receives an Ethernet packet from a hosting VM.

Lookup

Making use of the destination of the Ethernet packet, the VXLAN enabled NVE looks up MAC-NVE mapping table, and retrieves IP address of destination NVE.

Acquisition of QoS parameters

There are two different ways to acquire QoS parameters for VXLAN encapsulation. The first is a dynamic one which requires a VXLAN enabled NVE has Deep Packet Inspection (DPI) capability and can identify different application types. The second is a static one

which requires a VM manager to assign QoS parameters to different VNIs based on premium that different tenancies pay.

Encapsulation of frames

The NVE then encapsulates the packet using VXLAN format with acquired QoS parameters and VNI. The specific format is given in Section 4. After the frame is encapsulated it is sent out upstream to the network.

For a VXLAN gateway, packets are encapsulated using VXLAN format with QoS field in a similar way. Once the VXLAN gateway receives a packet from a non-VXLAN domain, it encapsulates the packet with QoS parameters which are acquired through DPI or priorities of tenancies.

7. QoS processing for VXLAN outer IP header

QoS is user experience of end-to-end network operation. A packet from VM A to VM B normally traverses such network entities sequentially as virtual switch A which is co-located with VM A, TOR switch A, aggregation switch A, a core switch, aggregation switch B, TOR switch B, virtual switch B. VXLAN processing only takes place in virtual switches, and all other network entities only execute IP forwarding. VXLAN QoS mapping to outer IP header at virtual switch A is needed to achieve end-to-end QoS.

Six bits of the Differentiated Services Field (DS field) are used as a codepoint (DSCP) to select the per hop behaviour (PHB) a packet experiences at each node in a Differentiated Services Domain [RFC2474]. DS field is 8 bits long, 6 bits of it are used as DSCP and two bits are unused. DS field is carried in both IPv4 and IPv6 packet headers.

Using 6 bits of DS field, it is possible to define 64 codepoints. A pool of 32 RECOMMENDED codepoints called Pool 1 is standardized by IANA. 8 of these 32 codepoints are called class selector (CS) codepoints, from CS0 corresponding to ’000000’ to CS7 corresponding to ’111000’ codepoints. There are also 12 assured forwarding (AF) codepoints [RFC2597], an expedited forwarding (EF) per hop behavior [RFC3246] and VOICE-ADMIT codepoint for capacity-admitted traffic [RFC5865]. In this document we use class selector codepoints. Other codepoints are out of scope.

When a packet forwarded from non-VXLAN domain to VXLAN domain through a VXLAN gateway, DSCP field of outer IP header for unicast packets should be marked based on VXLAN QoS using the class selector codepoints, i.e. ’xxx000’ that corresponds to VXLAN QoS marking, i.e.

with xxx assigned based on static or dynamic QoS markings defined above in Section 4.

8. Security Considerations

Special security considerations in [RFC7348] are applicable.

9. IANA considerations

None.

10. Acknowledgements

The authors are grateful to Brian Carpenter, David Black, Erik Nordmark for their constructive comments on our work.

11. References

11.1. 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.

[RFC0791] Postel, J., "Internet Protocol", STD 5, RFC 791, September 1981.

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

[RFC2474] Nichols, K., Blake, S., Baker, F., and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474, December 1998.

[RFC2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, December 1998.

[RFC2597] Heinanen, J., Baker, F., Weiss, W., and J. Wroclawski, "Assured Forwarding PHB Group", RFC 2597, June 1999.

[RFC3246] Davie, B., Charny, A., Bennet, J., Benson, K., Le Boudec, J., Courtney, W., Davari, S., Firoiu, V., and D. Stiliadis, "An Expedited Forwarding PHB (Per-Hop Behavior)", RFC 3246, March 2002.

[RFC4594] Babiarz, J., Chan, K., and F. Baker, "Configuration Guidelines for DiffServ Service Classes", RFC 4594, August 2006.

[RFC5865] Baker, F., Polk, J., and M. Dolly, "A Differentiated Services Code Point (DSCP) for Capacity-Admitted Traffic", RFC 5865, May 2010.

[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-02 (work in progress), October 2014.

[IEEE802.1D] IEEE, "Virtual Bridged Local Area Networks", IEEE Std 802.1D-2005, May 2006.

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.geib-tsvwg-diffserv-intercon] Geib, R. and D. Black, "DiffServ interconnection classes and practice", draft-geib-tsvwg-diffserv-intercon-08 (work in progress), November 2014.

[I-D.sarikaya-nvo3-proxy-vxlan] Sarikaya, B. and F. Xia, "Virtual eXtensible Local Area Network over IEEE 802.1Qbg", draft-sarikaya-nvo3-proxy-vxlan-00 (work in progress), October 2014.

Authors' Addresses

Frank Xia
Huawei Technologies Co., Ltd.
101 Software Avenue, Yuhua District
Nanjing, Jiangsu 210012, China

Phone: ++86-25-56625443
Email: xiayangsong@huawei.com

Behcet Sarikaya
Huawei Technologies Co., Ltd.
5340 Legacy Dr. Building 3
Plano, TX 75024

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

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

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