NVO3 Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 26, 2019

G. Fioccola
Huawei Technologies
G. Mirsky
ZTE Corp.
T. Mizrahi
Huawei Network.IO Innovation Lab

October 23, 2018

Performance Measurement (PM) with Alternate Marking in Network Virtualization Overlays (NVO3)
draft-fmm-nvo3-pm-alt-mark-03

Abstract

This document describes how the alternate marking method can be used for performance measurement method in a Network Virtualization Overlays (NVO3) Domain. The description aims to be general for NVO3 encapsulations, but is focused to Geneve, recommended by the NVO3 design team [I-D.ietf-nvo3-encap].

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 https://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 April 26, 2019.

Table of Contents

1. Introduction

[RFC7365] provides a framework for Data Center (DC) Network Virtualization over Layer 3 (NVO3) tunnels. It is intended to aid in standardizing protocols and mechanisms to support large-scale network virtualization for data centers.

[RFC8321] describes a performance measurement method, which can be used to measure packet loss, latency, and jitter on live traffic. Since this method is based on marking consecutive batches of packets the method often referred to as the Alternate Marking Method (AMM).

This document defines how the alternate marking method can be used to measure packet loss and delay metrics of an NVO3 Domain.

2. Conventions used in this document

2.1. Terminology

AMM: Alternate Marking Method

OAM: Operations, Administration and Maintenance

NVO3: Network Virtualization Overlays

NVE: Network Virtualization Edge

VNI: Virtual Network Instance

DC: Data Center

NVA: Network Virtualization Authority

Geneve: Generic Network Virtualization Encapsulation

VXLAN: Virtual Extensible LAN

GUE: Generic UDP Encapsulation

2.2. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

3. OAM Performance Measurement in a NVO3 Domain

Figure 1 shows the generic reference model for a DC network virtualization over an L3 infrastructure while Figure 2 shows the generic reference model for the Network Virtualization Edge (NVE). Both Figures are taken from [RFC7365] and [RFC8014].

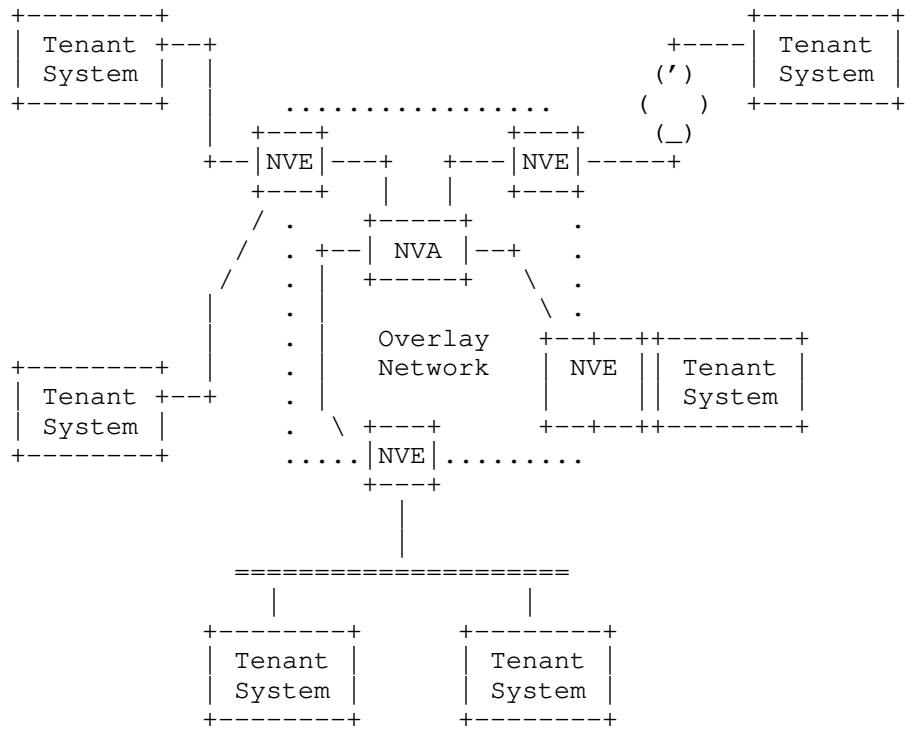

Figure 1: Generic Reference Model for DC Network Virtualization Overlays (RFC7365)

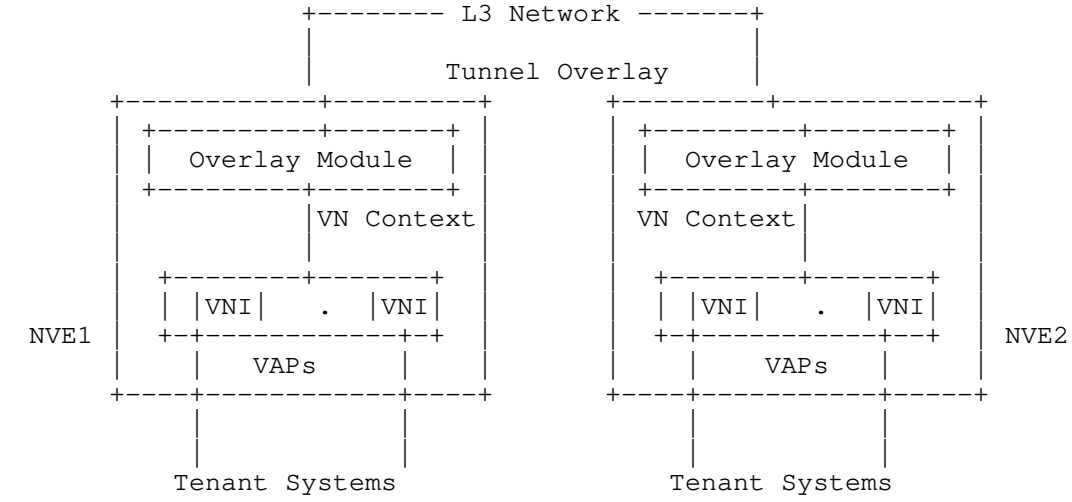

Figure 2: Generic NVE Reference Model (RFC7365)

L3 networks provide transport for an emulated Layer 2 created by NVE devices. The connectivity between the NVE devices is achieved with unicast and multicast tunneling methods. Then, the NVE devices present an emulated Layer 2 network to the Tenant End Systems at a Virtual Network Instance (VNI) through Virtual Access Points (VAPs). The NVE devices map Layer 2 unicast to Layer 3 unicast point-to-point tunnels and may either map Layer 2 multicast to Layer 3 multicast tunnels or may replicate packets onto multiple Layer 3 unicast tunnels.

The emulated Layer 2 network is provided by the NVE devices to which the Tenant End Systems are connected. This network of NVE can be operated by a single service provider or can span across multiple administrative domains. Likewise, the L3 Overlay Network can be operated by a single service provider or span across multiple administrative domains.

Each of the layers is responsible for its own OAM. Complex OAM relationships exist as a result of the hierarchical layering, but this is out of scope here.

When we refer to an OAM domain considered in this document we refer to a set of NVEs and the tunnels which interconnect them.

It is commonly agreed that NVO3 OAM Performance Management supports measurements (packet loss, latency, and jitter) per VNI between two NVE devices that support the same VNI within a given NVO3 domain.

4. The Mark Field in the NVO3 Header

This document defines a two-bit long field, referred to as Mark field (M), as part of the NVO3 Header and designated for the alternate marking performance measurement method [RFC8321]. The Mark field MUST NOT be used in defining forwarding and/or quality of service treatment of an NVO3 packet. The Mark field MUST be used only for the performance measurement of data traffic in the NVO3 layer. Since the field does not affect forwarding and/or quality of service treatment of packets, the alternate marking method in the NVO3 layer can be viewed as nearly-passive performance measurement method.

Figure 3 displays the format of the Mark field.

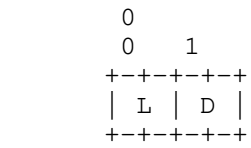

Figure 3: Mark field (M) format

where:

- L - Loss bit;
- D - Delay bit.

5. Theory of Operation

The marking method can be used in NVO3. For example, one can consider the NVO3 reference model presented in Figure 1. AMM can be applied at either ingress or egress NVE to detect performance degradation defect and localize it efficiently.

Using AMM, NVE1 creates distinct sub-flows. Each sub-flow consists of consecutive blocks that are unambiguously recognizable by a monitoring point at any component of the NVO3, e.g. NVE2 or NVE3, and can be measured to calculate packet loss and/or packet delay metrics.

Every NVO3 Header [I-D.ietf-nvo3-geneve], [I-D.ietf-nvo3-vxlan-gpe] and [I-D.ietf-nvo3-gue] can be considered for the application of AMM.

5.1. Single Mark Enabled Measurement

As explained in the [RFC8321], marking can be applied to delineate blocks of packets based either on the equal number of packets in a block or based on equal time interval. The latter method offers better control as it allows better account for capabilities of downstream nodes to report statistics related to batches of packets and, at the same time, time resolution that affects defect detection interval.

If the Single Mark measurement used, then the D flag MUST be set to zero on transmit and ignored by monitoring point.

The L flag is used to create alternate flows to measure the packet loss by switching the value of the L flag every N-th packet or at certain time intervals. Delay metrics MAY be calculated with the alternate flow using any of the following methods:

- First/Last Packet Delay calculation: whenever the marking, i.e. the value of L flag, changes a component of the NVO3 can store the timestamp of the first/last packet of the block. The timestamp can be compared with the timestamp of the packet that arrived in the same order through a monitoring point at a downstream component of the NVO3 to compute packet delay. Because timestamps collected based on order of arrival this method is sensitive to packet loss and re-ordering of packets

- Average Packet Delay calculation: an average delay is calculated by considering the average arrival time of the packets within a single block. A component of the NVO3 may collect timestamps for each packet received within a single block. Average of the timestamp is the sum of all the timestamps divided by the total number of packets received. Then the difference between averages calculated at two monitoring points is the average packet delay on that segment. This method is robust to out of order packets and also to packet loss (only a small error is introduced). This method only provides a single metric for the duration of the block and it doesn't give the minimum and maximum delay values. This limitation could be overcome by reducing the duration of the block by means of a highly optimized implementation of the method.

5.2. Double Mark Enabled Measurement

Double Mark method allows measurement of minimum and maximum delays for the monitored flow but it requires more nodal and network resources. If the Double Mark method used, then the L flag MUST be used to create the alternate flow, i.e. mark larger batches of packets. The D flag MUST be used to mark single packets to measure delay jitter.

The first marking (L flag alternation) is needed for packet loss and also for average delay measurement. The second marking (D flag is put to one) creates a new set of marked packets that are fully identified over NVO3, so that a component can store the timestamps of these packets; these timestamps can be compared with the timestamps of the same packets on another component of the NVO3 to compute packet delay values for each packet. The number of measurements can be easily increased by changing the frequency of the second marking. But the frequency of the second marking must be not too high in order to avoid out of order issues. This method is suitable to have not only the average delay but also the minimum and maximum delay values and, in wider terms, to know more about the statistic distribution of delay values.

5.3. Multiplexed Mark Enabled Measurement

There is also a scheme that provides the benefits of Double Mark method, but uses only one bit like Single Mark. This methodology is described in [I-D.mizrahi-ippm-compact-alternate-marking]. The concept is that in the middle of each block of packets with a certain value of the L flag, a single packet has the L flag inverted. So, by examining the stream, the packets with the inverted bit can be easily identified and employed for delay measurement. This Alternate Marking variation is advantageous because it requires only one bit from each packet, and such bits are always in short supply.

6. Multipoint Measurement Considerations

The Multipoint characteristics of the traffic within a given NVO3 Domain could be considered a valuable Use Case of [I-D.fioccola-ippm-multipoint-alt-mark].

7. The Mark Field in Geneve

[I-D.ietf-nvo3-geneve] defines the format of the Geneve Header.

The design team recommendations in [I-D.ietf-nvo3-encap] section 7 concluded that Geneve is most suitable as a starting point for the proposed standard for network virtualization.

In addition, the design team recommended addressing requirements for OAM considerations for alternate marking and for performance measurements that need 2 bits in the header. This document clarifies the need for the current OAM bit in the Geneve Header.

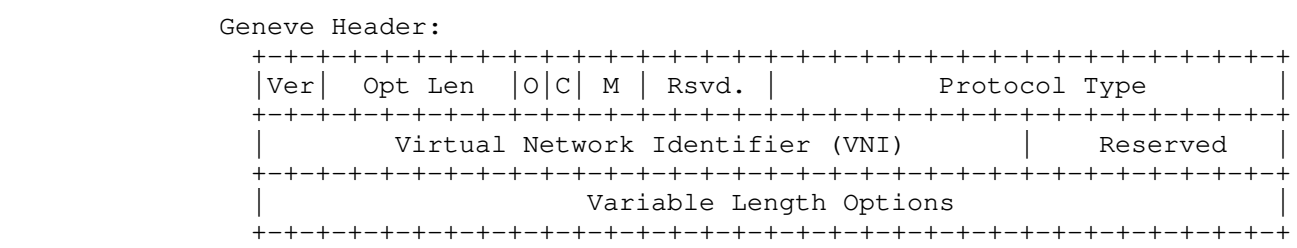

Figure 4: Geneve Header

This document defines a two-bit long field, referred to as the Mark field (M in Figure 4, as part of Geneve and designated for the alternate marking performance measurement method [RFC8321]. The Mark field MUST NOT be used in defining forwarding and/or quality of service treatment of a NVO3 packet. The Mark field MUST be used only for the performance measurement of data traffic in the NVO3 layer.

Since the field does not affect forwarding and/or quality of service treatment of packets, the alternate marking method in the NVO3 layer can be viewed as nearly-passive performance measurement method.

8. IANA Considerations

8.1. Mark Field in Geneve Header

This document requests IANA to allocate Mark field as two bits-long field from Geneve Header Reserved Bits [I-D.ietf-nvo3-geneve].

This document requests IANA to register values of the Mark field of Geneve as the following:

Bit Position	Marking	Description	Reference
0	L	Single Mark Measurement	This document
1	D	Double Mark Measurement	This document

Table 1: Mark field of Geneve

9. Security Considerations

This document lists the OAM requirement for the NVO3 domain and does not raise any security concerns or issues in addition to ones common to networking and NVO3.

10. Acknowledgement

The authors would like to thank Dale R. Worley for the contribution.

11. References

11.1. Normative References

[I-D.ietf-nvo3-encap]
Boutros, S., "NVO3 Encapsulation Considerations", draft-ietf-nvo3-encap-02 (work in progress), September 2018.

[I-D.ietf-nvo3-geneve]
Gross, J., Ganga, I., and T. Sridhar, "Geneve: Generic Network Virtualization Encapsulation", draft-ietf-nvo3-geneve-08 (work in progress), October 2018.

[I-D.ietf-nvo3-gue]
Herbert, T., Yong, L., and O. Zia, "Generic UDP Encapsulation", draft-ietf-nvo3-gue-05 (work in progress), October 2016.

[I-D.ietf-nvo3-vxlan-gpe]
Maino, F., Kreeger, L., and U. Elzur, "Generic Protocol Extension for VXLAN", draft-ietf-nvo3-vxlan-gpe-06 (work in progress), April 2018.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

11.2. Informative References

[I-D.fioccola-ippm-multipoint-alt-mark]
Fioccola, G., Cociglio, M., Sapio, A., and R. Sisto, "Multipoint Alternate Marking method for passive and hybrid performance monitoring", draft-fioccola-ippm-multipoint-alt-mark-04 (work in progress), June 2018.

[I-D.mizrahi-ippm-compact-alternate-marking]
Mizrahi, T., Arad, C., Fioccola, G., Cociglio, M., Chen, M., Zheng, L., and G. Mirsky, "Compact Alternate Marking Methods for Passive and Hybrid Performance Monitoring", draft-mizrahi-ippm-compact-alternate-marking-03 (work in progress), October 2018.

[RFC7365] Lasserre, M., Balus, F., Morin, T., Bitar, N., and Y. Rekhter, "Framework for Data Center (DC) Network Virtualization", RFC 7365, DOI 10.17487/RFC7365, October 2014, <https://www.rfc-editor.org/info/rfc7365>.

[RFC8014] Black, D., Hudson, J., Kreeger, L., Lasserre, M., and T. Narten, "An Architecture for Data-Center Network Virtualization over Layer 3 (NVO3)", RFC 8014, DOI 10.17487/RFC8014, December 2016, <https://www.rfc-editor.org/info/rfc8014>.

[RFC8321] Fioccola, G., Ed., Capello, A., Cociglio, M., Castaldelli, L., Chen, M., Zheng, L., Mirsky, G., and T. Mizrahi, "Alternate-Marking Method for Passive and Hybrid Performance Monitoring", RFC 8321, DOI 10.17487/RFC8321, January 2018, <https://www.rfc-editor.org/info/rfc8321>.

Authors' Addresses

Giuseppe Fioccola
Huawei Technologies
Riesstrasse, 25
Munich 80992
Germany

Email: giuseppe.fioccola@huawei.com

Greg Mirsky
ZTE Corp.

Email: gregimirsky@gmail.com

Tal Mizrahi
Huawei Network.IO Innovation Lab
Israel

Email: tal.mizrahi.phd@gmail.com