INTERNET-DRAFT
Intended Status: Standard Track

Sami Boutros
Jerome Catrouillet
VMware

Expires: January 8, 2020

July 7, 2019

# MAC move over Geneve encapsulation
## draft-boutros-nvo3-mac-move-over-geneve-01

## Abstract

This document specifies a mechanism to signal Media Access Control (MAC) addresses move over a Network Virtualization Overlays over Layer 3 (NVO3) virtual tunnel. Such notification is useful in redundancy scenarios when a Layer 2 service that was active on a Network Virtualization Edge (NVE) fails over to a standby NVE. This notification can be used only when data plane mac learning is enabled over the NVO3 tunnels.

## Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/1id-abstracts.html

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html

## Copyright and License Notice

Copyright (c) 2019 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

In multi-homing scenarios a Layer 2 service can be multi-homed to more than one Network virtualization Edge (NVE). Only one NVE can be active for a given Layer 2 service, and a standby NVE can be chosen to take over the Layer 2 service when the active NVE goes down. The mechanisms to elect which NVE will be active or standby to provide single active redundancy for a given Layer 2 service is outside the scope of this document.

When a standby NVE gets activated, Standby NVE needs to send a MAC Move message to all remote NVE(s) that spans this L2 service over the Geneve tunnels to Move all MAC learned in data plane via the old active NVE.

The MAC Move message will contain the NVE Identifier(s) of the old Active NVE and the new active NVE.

MAC Move can be used to optimize network convergence and reduce blackholes, when an active NVE hosting a logical L2 service fails over to a standby NVE.

The protocol defined in this document addresses possible loss of the MAC Move messages due to network congestion, but does not guarantee delivery.

In the event that MAC Move messages does not reach the intended target, the fallback to MAC re-learning or as a last resort aging out of MAC addresses in the absence of frames from the sources, will resume the traffic via new active NVE.

## 1.1 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 RFC 2119 [RFC2119].

## 1.2 Abbreviations

NVO3 Network Virtualization Overlays over Layer 3

OAM Operations, Administration, and Maintenance

TLV Type, Length, and Value

VNI Virtual Network Identifier

NVE Network Virtualization Edge





NVA Network Virtualization Authority

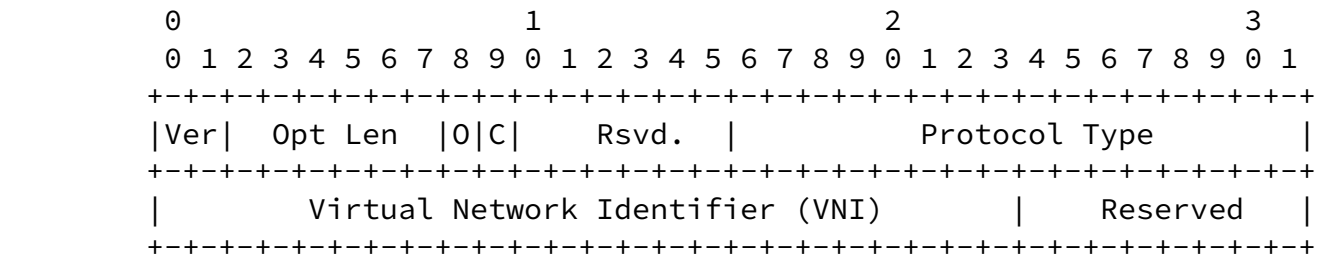
NIC Network interface card

VTEP Virtual Tunnel End Point

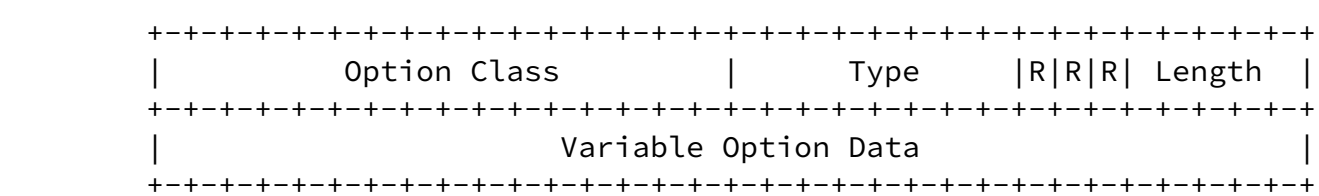
Transit device Underlay network devices between NVE(s).

## 2. MAC Move Frame Format

Geneve Header:




Geneve Option Header:



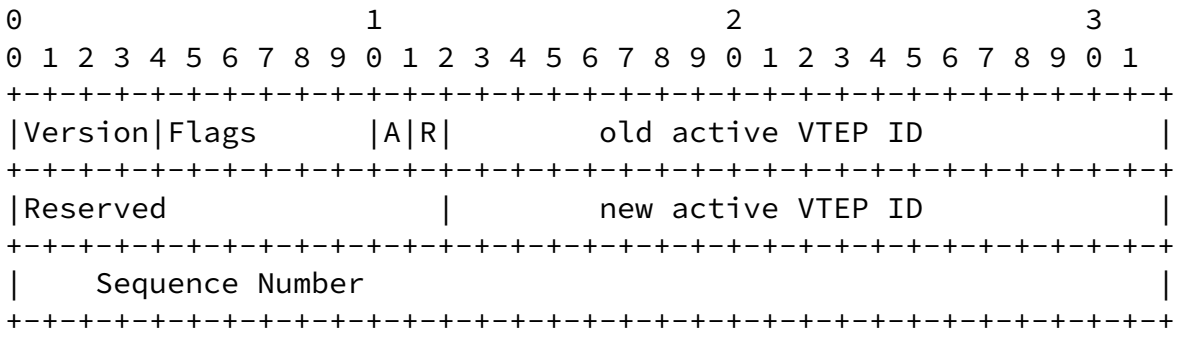

Option Class = To be assigned by IANA (TBA).

Type = TBA.

'C' bit, Endpoints must drop if they do not recognize this option)

Length = 2 (8 bytes)

Variable option data:




Version (4 bits): Initially the Version will be 0.

A (1 bit): Set by a receiver to acknowledge receipt and processing of a MAC Move message.





R (1 bit): Set to indicate if the sender is requesting reset of the sequence numbers. The sender sets this bit when it has no local record of previous send and expected receive sequence numbers.

Flags(6): Reserved and should be set to 0.

VTEP ID (20 bits): Identifies old and new active NVE(s).

Sequence Number (32) bits: For overflow detection a sequence number that exceeds (0x7FFFFFFF) is considered an overflow and reset to 1.

## 3. Operation

This section describes how the initial MAC Move Messages are sent and retransmitted, as well as how the messages are processed and retransmitted messages are identified. The mechanisms described are very similar to the one defined in [RFC 7769].

## 3.1 Operation of Sender

At the NVE , each L2 logical switch identified by a VNI is associated with a counter to keep track of the sequence number of the transmitted MAC Move messages. Whenever a node sends a MAC Move message, it increments the transmitted sequence-number counter and includes the new sequence number in the message.

The transmit sequence number is initialized to 1 at the onset, after the wrap and after the sequence number reset request receipt. Hence the transmit sequence number is set to 2 in the first MAC Move message sent after the sequence number is initialized to 1.

The sender expects an ACK from the receiver within a retransmit time interval, which can be either a default (1 second) or a configured value. If the ACK is not received within the Retransmit time, the sender retransmits the message with the same sequence number as the original message. The retransmission MUST cease when an ACK is received. In order to avoid continuous re-transmissions in the absence of acknowledgements, the sender MUST cease retransmission after a small number of transmissions, two retries is RECOMMENDED.

Alternatively, an increasing backoff delay with a larger number of retries MAY be implemented to improve scaling issues.

During the period of retransmission, if a need to send a new MAC Move message with updated sequence number arises, then retransmission of the older unacknowledged Move message MUST be suspended and





retransmit time for the new sequence number MUST be initiated. In essence, a sender engages in retransmission logic only for the most recently sent Move message for a given L2 Logical Switch identified by a VNI.

In the event that the L2 logical switch is deleted and re-added or the VTEP node is restarted with new configuration, the NVE may lose information about the previously sent sequence number. This becomes

problematic for the remote peer as it will continue to ignore the received MAC Move messages with lower sequence numbers. In such cases, it is desirable to reset the sequence numbers, the reset R-bit is set in the first MAC Move to notify the remote peer to reset the send and receive sequence numbers. The R-bit must be cleared in subsequent MAC Move messages after the acknowledgement is received.

### 3.2 Operation of Receiver

Each L2 logical switch identified by a VNI is associated with a receive sequence number per remote NVE to keep track of the expected sequence number of the MAC Move message.

Whenever a MAC Move message is received, and if the sequence number on the message is greater than the value in the receive sequence number of this remote NVE, the MAC addresses learned from the NVE associated with the NVE identifier in the message are moved to be associated with the new active NVE identifier, and the receive sequence number of the remote NVE is updated with the received sequence number. The receiver sends an ACK with the same sequence number in the received message.

If the sequence number in the received message is smaller than or equal to the value in the receive sequence number per remote NVE, the MAC Move is not processed. However, an ACK with the received sequence number MUST be sent as a response to stop the sender retransmission.

A MAC Move message with the R-bit set MUST be processed by resetting the receive sequence number of the remote NVE, and Moving the MACs as described above. The acknowledgement is sent with the R-bit cleared.

## 4. Security Considerations

This document does not introduce any additional security constraints.

## 5. IANA Considerations

IANA is requested to assign a new option class from the "Geneve Option Class" registry for the Geneve MAC Move option.

| Option Class | Description |
| --- | --- |

# 6 References

## 6.1 Normative References

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

## 6.2 Informative References

[Geneve] "Generic Network Virtualization Encapsulation", [I-D.ietf-nvo3-geneve]
[RFC 7769] "MAC Address Withdrawal over Static PW", [RFC 7769]

## Authors' Addresses

Sami Boutros
VMware
Email: boutross@vmware.com

Jerome Catrouillet
VMware
Email: jcatrouillet@vmware.com

Ankur Sharma
VMware
Email: ankursharma@vmware.com