NV03 Working Group
Internet Draft
Intended Status: Informational

Tissa Senevirathne
Samer Salam
Deepak Kumar
Norman Finn

Cisco

Donald Eastlake
Sam Aldrin

Huawei

Expires September 2017

May 5, 2017

NVO3 Fault Management
draft-tissa-nvo3-oam-fm-04.txt

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), 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/ietf/1id-abstracts.txt.

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

This Internet-Draft will expire on November 6, 2017.

Copyright Notice

Copyright (c) 2017 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.

Abstract

This document specifies Fault Management solution for network virtualization overlay networks. Methods in this document follow the IEEE 802.1 CFM framework and reuse OAM tools where possible. Additional messages and TLVs are defined for IETF overlay OAM specific applications or where extensions beyond IEEE 802.1 CFM are required.

Table of Contents

1. Introduction

Conceptually, NVO3 architecture contains four separate layers, namely, Service (customer) Layer, Overlay Layer, Transport or Underlay Layer and Media Layer. Figure 1 below depicts the relationship between each of these layers.

Fault Monitoring, Fault Verification, Fault Isolation, Loss and delay measurements are integral part of NVO3 [nvo3oamReq]. These need to cover both unicast and multicast traffic streams.

For effective fault isolation, the overlay OAM solution should complement the OAM functions at adjacent layers, thereby leading to nested OAM model. Nested OAM allows operators to quickly and effectively troubleshoot and isolate data plane failures using a common OAM framework.

Common OAM message format and infrastructure makes it easier to accomplish nested OAM. IEEE Connectivity Fault Management (CFM) [8021Q] is widely used in the Ethernet world. The same technology has been extended by ITU-T [Y1731], Metro Ethernet Forum (MEF) and TRILL [TRILLFM].

A common OAM message channel can be shared between different technologies. This consistency between different OAM technologies promotes nested fault monitoring and isolation between technologies that share the same OAM framework.

This document uses the message format defined in IEEE 802.1ag Connectivity Fault Management (CFM) [8021Q] as the basis for the OAM messages.

This document presents the NVO3 OAM message structure, NVO3 frame identification and NV03 OAM tools. The NVO3 OAM Management Information Base (MIB) will be presented in a separate document.

The ITU-T Y.1731 [Y1731] standard utilizes the same messaging format as [8021Q] and for OAM messages where applicable. This document takes a similar stance and reuses [8021Q] and TRILL OAM [TRILLFM]. It is assumed that readers are familiar with [8021Q] and [Y1731].

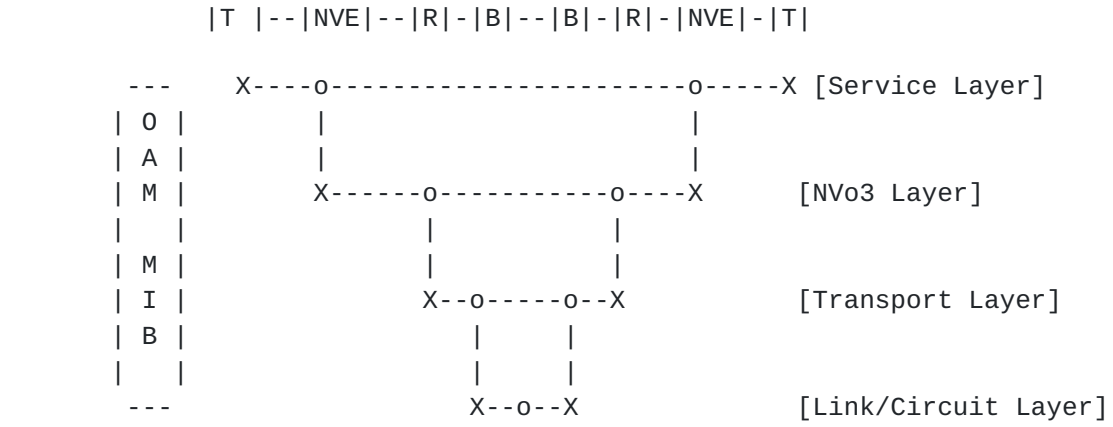


Figure 1 Layered OAM Architecture

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].

Acronyms used in the document include the following:

MP - Maintenance Point [8021Q]

MEP - Maintenance End Point [8021Q]

MIP - Maintenance Intermediate Point [8021Q]

MA - Maintenance Association [8021Q]

MD - Maintenance Domain [8021Q]

CCM - Continuity Check Message [8021Q]

LBM - Loop Back Message [8021Q]

PTM - Path Trace Message

MTV - Multi-destination Tree Verification Message

ECMP - Equal Cost Multipath

ISS - Internal Sub Layer Service [8021Q]

VNI - Virtual Network Instance [NVO3FRM]

NVE - Network Virtual Edge [NVO3FRM]

SAP - Service Access Point [8021Q]

3. NV03 OAM Layers

Figure 1 above depicts different layers within NVO3. Each of these layers has a unique scope within the common framework. In this section, we define functionality of each of these layers

Service Layer: Service Layer carries customer or user traffic. It is originated at Tenant systems and terminates at other Tenant system(s) within the same NVO3 context (VNI).

NVO3 Layer: NVO3 Layer carries service Layer traffic encapsulated in NVO3 format. NVO3 Layer originates at an NVE and terminates at another NVE.

Transport Layer: Transport Layer carries NVO3 Layer traffic encapsulated in its data format. The transport Layer can be IP, MPLS or any other protocol.

Link Layer: This is also known as circuit layer.
It carries Transport Layer traffic in a specific manner from one device to another.
Ethernet is an example of link layer.

4. General Format of NV03 OAM frames

For accurate monitoring and/or diagnostics, OAM Messages are required to follow the same data path as corresponding user packets. Additionally, NVEs are required to identify NVO3 frames and act on them, in addition to preventing the overlay OAM packets from leaking outside of the NVO3 domain [NVO3OAMREQ]. This document defines the format of the NVO3 overlay OAM messages.

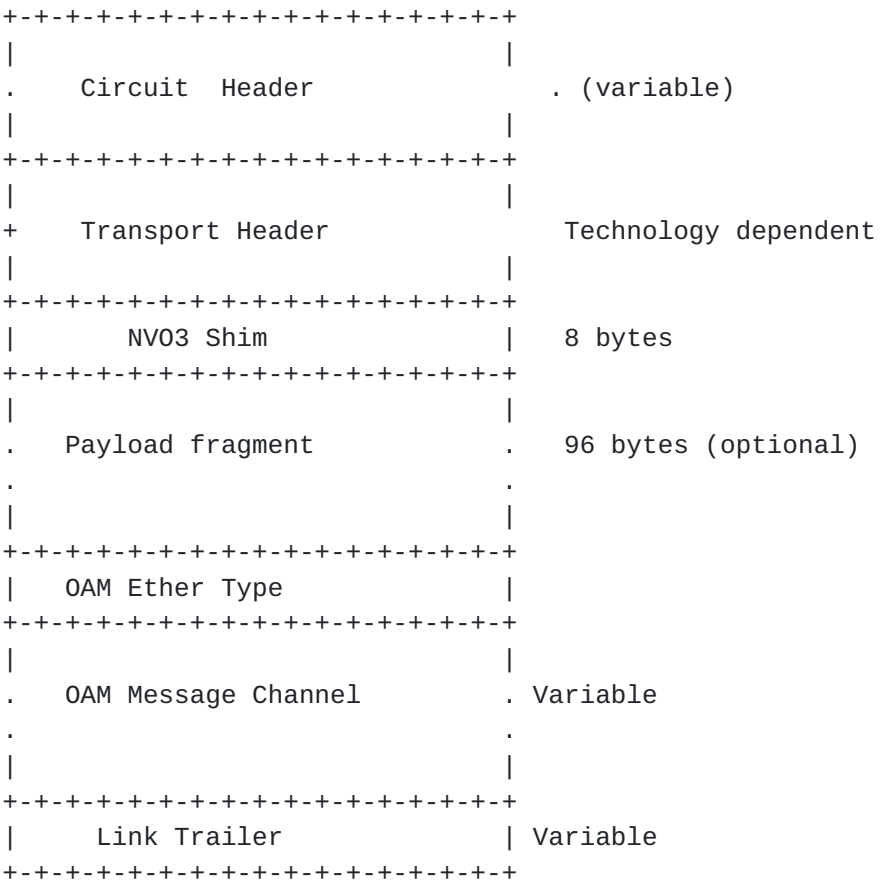


Figure 2 Format of NVO3 Overlay OAM Messages

Link Header: Media-dependent header. For Ethernet, this includes Destination MAC, Source MAC, VLAN (optional) and EtherType fields.

Transport Header: Header of the Transport Layer e.g. IP , MPLS, TRILL etc.

NVO3 Shim: This is a fixed sized field (size TBD 8 bytes [vxLAN]) that carries NVO3 specific information. Fields in this header will be utilized to identify OAM frames and other NVO3 specific operations. Please see section 4.2 below of NVO3 OAM specific operations.

Payload Fragment: This is an optional field. When included it has a fixed size of 96 bytes. The least significant bits of the field MUST be padded with zeros, up to 96 bytes, when the payload fragment is less than 96 bytes. The Payload Fragment field starts with the Inner.MacDA.

OAM Ether Type: OAM Ether Type is 16-bit EtherType that identifies the OAM Message channel that follows. This document specifies using the EtherType 0x8902 allocated for [8021Q] for this purpose. Identifying the OAM Message Channel with a dedicated EtherType allows the easy identification of the beginning of the OAM message channel across multiple standards.

OAM Message Channel: This is a variable size section that carries OAM related information. The message format defined in [8021Q] will be reused.

Link Trailer: Media-dependent trailer. For Ethernet, this is the FCS (Frame Check Sequence).

Note: In this draft we are proposing re-use of OAM Channel defined in RFC7455 and RFC7456.

4.1. NV03 Shim

NVO3 Shim is an 8 octet vector that carries series of flags and additional information [vxLAN]. Each of the flags identifies a specific operation.

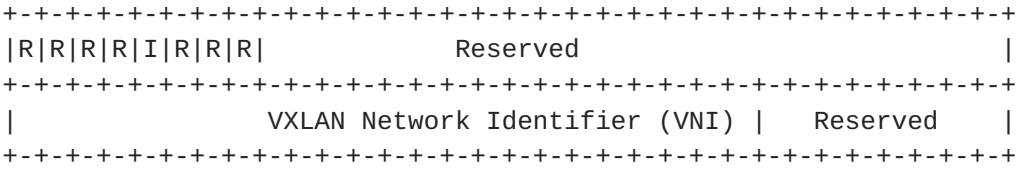


Figure 3 NVO3 Shim

4.2. Identification of OAM frames

Implementations that comply with this document MUST utilize "O" flag to identify NVO3 OAM frames. The "O" flag MUST NOT BE utilized for forwarding decisions such as the selection of which ECMP path to use.

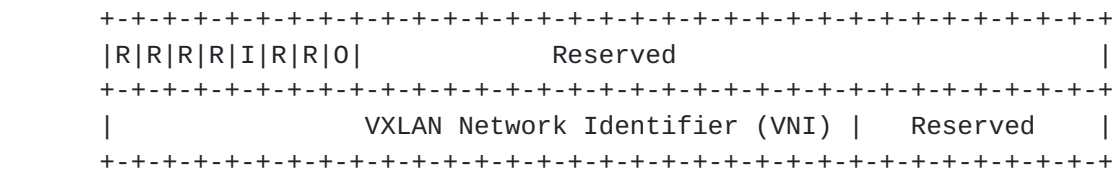


Figure 4 NVO3 shim with the "O" Flag

O (1 bit) - Indicates this is a possible OAM frame and is subject to specific handling as specified in this document. O bit is aligned with Vxlan GPE.

Payload Fragment is optional and if hardware is capable of matching OAM frame based on O bit then payload fragment handling doesn't require extra bit by checking etype. OAM Frame may have 96 octets of payload fragments immediately after the NVO3 shim or OAM Ethertype 0x8902 immediately follows the NVO3 shim after Inner DMAC and Inner SMAC. This can be determined by matching 0x8902 at right position after matching "O" bit, if it's not present, then look deeper after 96 bytes.

Payload Fragment is optional implementation but it's very important to track actual data path in scenario described below.

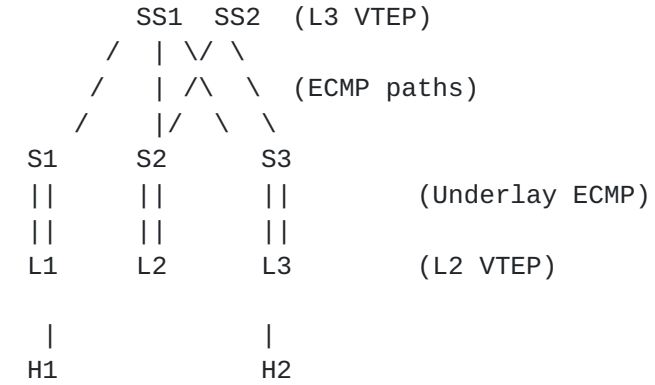


Figure 5 Payload Fragment use case

For H1 and H2 we can have bridge traffic and routed traffic on SSx. For Routed traffic inner header or payload Fragement is required to be looked to find the exact ECMP path.

All other fields carry the same meaning as defined in [vXLAN].

4.3 OAM Channel

OAM channel proposed to use is based of RFC7455 and RFC7456. Advantage of re-using this channel is header is flexible to support extensible functionality. It also provide backward compatibility mode for hardware which were developed before OAM became standard to do OAM functionality.

4.3.1 OAM Channel Functionality Summary

OAM Common Header is defined in section 8.1. MD-L allows multiple level of OAM to be possible, for eg:- IP layer Overlay can be at default level 3 and SFC layer OAM can be at Application Level 4.

Opcode provide hardware friendly and extensible extensions. Hardware friendly messages are defined without TLVs.

Summary of Opcode(s) to be considered.

1. CCM - Proactive Fault Monitoring (one-way)

2. LBM/LBR - on-demand Fault Verification (LBR can be sent out of band also and carry ICMP error code if required.)

3. PTM/PTR - on-demand Fault Isolation based on TTL expiry (works with non OAM capable underlay, ip unnumbered underlay, and provide Egress Interface to better isolate fault than traditional traceroute)

4. DMM/DMR - on-demand or pro-active Delay Measurement.

5. 1DM - On-demand or pro-active one way Delay Measurement.

6. LTM/LTR - on-demand Fault Isolation for scenario where all underlay switches are OAM capable to provide path via hardware forwarding without CPU intervention.

7. SLM/SLR - on-demand or pro-active Synthetic Delay Measurement.

8. 1SL - On-demand or pro-active one way Synthetic Delay Measurement.

Application Identifier TLV allow Return code and sub-code to carry ICMP Errors, It allows out of band or in-band communication flexibility and support OAM to be carried in multiple fragments if data request is very large.

4.3.2 Opcode Implementation Recommendation

CCM - Optional LBM/LBR - Mandatory PTM/PTR - Mandatory DMM/DMR -

Optional (As it's performance function) 1DM - Optional LTM/LTR - Optional SLM/SLR - Optional 1SL - Optional

5. Maintenance Associations (MA) in NVO3

[8021Q] defines a maintenance association as a logical relationship between a group of nodes. Each Maintenance Association (MA) is identified with a unique MAID of 48 bytes [8021Q]. CCM and other related OAM functions operate within the scope of an MA. The definition of MA is technology independent. MA is encoded in the technology independent part of the OAM message Hence the MAID, as defined in [8021Q], can be utilized for NVo3 OAM, without modifications. This also allows us to utilize CCM and LBM messages defined in [8021Q], as is.

In NVO3, an MA may contain two or more NVEs (MEPs). For unicast, it is likely that the MA contains exactly two MEPs that are the two end-points of the flow. For multicast, the MA may contain two or more MEPs.

For NVO3, in addition to all of the standard CFM MIB [8021Q] definitions, each MEP's MIB contains one or more flow definitions corresponding to the set of flows that the MEP monitors. Flow entropy specifies the VNI within the NVE.

MEPs can be created per VNI within the NVE.

We propose to augment the [8021Q] MIB to add NVO3 specific information. Figure 5, below depicts the augmentation of the CFM MIB to add the NVO3 specific Flow Entropy.

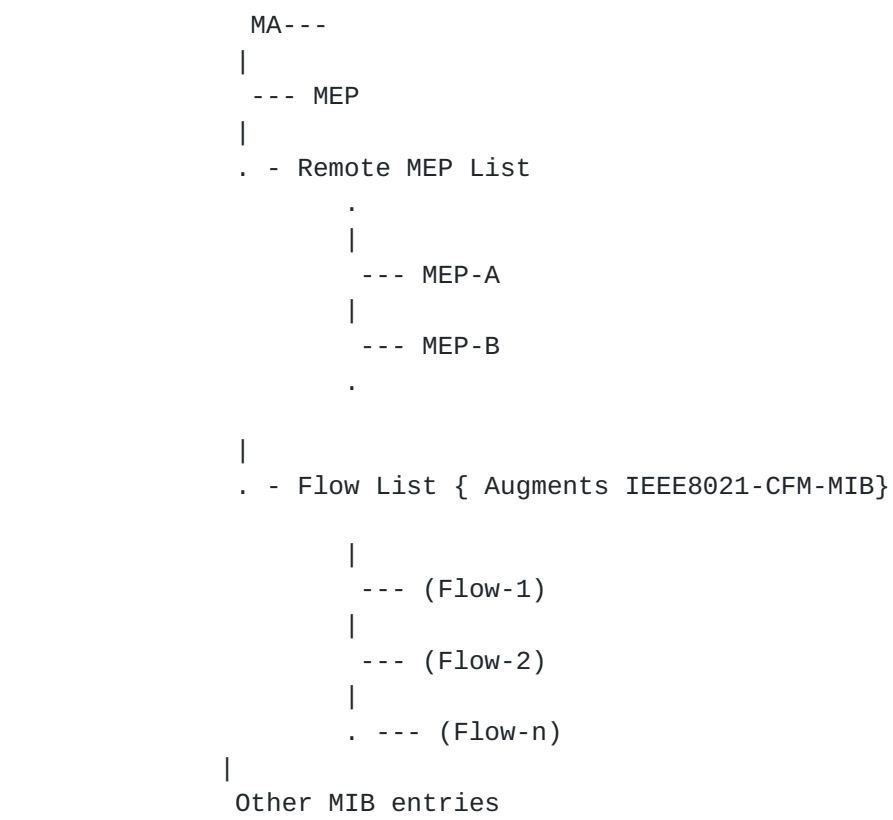


Figure 6 Correlation of NVO3 augmented MIB

The flow list contains multiple flows. Each flow contains a VNI and optional data payload. There can be more than one flow for a given VNI with different data payloads e.g. unicast vs. multicast or unicast with different data payloads.

6. MEP Addressing

In IEEE 802.1ag [8021Q], OAM messages address the target MEP by utilizing a unique MAC address. In NVO3, MEPs are created per VNI, MEPs are addressed by combination of Transport Layer address of the NVE and the VNI.

At the MEP, OAM packets go through a hierarchy of op-code de-multiplexers. The op-code de-multiplexers channel the incoming OAM packets to the appropriate message processor (e.g. LBM) The reader may refer to Figure 6 below for a visual depiction of these different de-multiplexers.

1. Identify the packets that need OAM processing at the Local Device as specified in Section 4.2.

   a. Identify the MEP that is associated with the VNI.

2. The MEP then validates the MD-LEVEL

   a. Redirect to MD-LEVEL De-multiplexer

3. MD-LEVEL de-multiplexer compares the MD-Level of the packet against the MD level of the local MEPs. (Note: there can be more than one MEP at the same MD-Level but belonging to different MAs)

   a. If the packet MD-LEVEL is equal to the configured MD-LEVEL of the MEP, then pass to the Opcode de-multiplexer

   b. If the packet MD-LEVEL is less than the configured MD-LEVEL of the MEP, discard the packet

   c. If the packet MD-LEVEL is greater than the configured MD-LEVEL of the MEP, then pass on to the next higher MD-LEVEL de-multiplexer, if available. Otherwise, if no such higher MD-LEVEL de-multiplexer exists, then forward the packet as normal data.

4. Opcode De-multiplexer compares the opcode in the packet with the supported opcodes

   a. If Op-code is CCM, LBM, LBR, PTM, PTR, MTVM, MTVR, then pass on to the correct Processor

   b. If Op-code is Unknown, then discard.

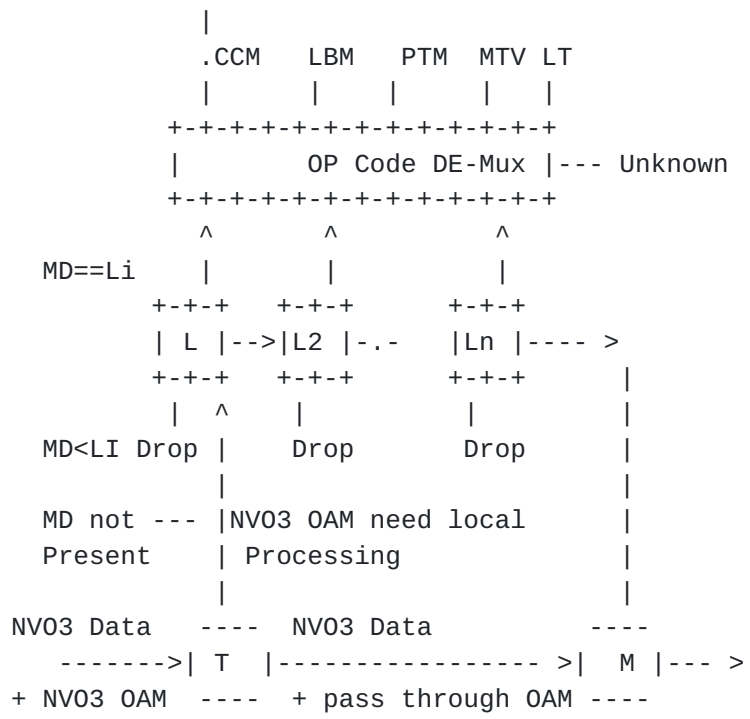


Figure 7 OAM De-Multiplexers at MEP for active SAP

Default MEPs are assumed to be created at NVE to handle OAM functionality as per Appendix A.

T : Denotes Tap, that identifies OAM frames that need local processing. These are the packets with OAM flag set AND OAM Ether type is present after the flow entropy of the packet

M : Is the post processing merge, merges data and OAM messages that are passed through. Additionally, the Merge component ensures, as explained earlier, that OAM packets are not forwarded out as native frames.

L : Denotes MD-Level processing. Packets with MD-Level less than the Level will be dropped. Packets with equal MD-Level are passed on to the opcode de-multiplexer. Others are passed on to the next level MD processors or eventually to the merge point (M).

NOTE: LBM, MTV and PT are not subject to MA de-multiplexers. These packets do not have an MA encoded in the packet. Adequate response can be generated to these packets, without loss of functionality, by any of the MEPs.

6.1. Use of MIP in NV03

Maintenance Intermediate Points (MIP) are mainly used for fault isolation. Link Trace Messages in [8021Q] utilize a well-known multicast MAC address and MIPs generate responses to Link Trace messages. Response to Link Trace messages or lack thereof can be used for fault isolation in NVO3.

As explained in section 10. below, a TTL expiry approach will be utilized for fault isolation and path tracing. The approach is very similar to the well-known IP trace-route approach. Hence, explicit addressing of MIPs is not required for the purpose of fault isolation.

Any given NVO3 device can have multiple MIPs located within the device. As such, a mechanism is required to identify which MIP should respond to an incoming OAM message.

A similar approach to that presented above for MEPs can be used for MIP processing. It is important to note that "M", the merge block of a MIP, does not prevent OAM packets leaking out as native frames. On edge interfaces, MEPs MUST be configured to prevent the leaking of overlay OAM packets out of the NVO3 domain.

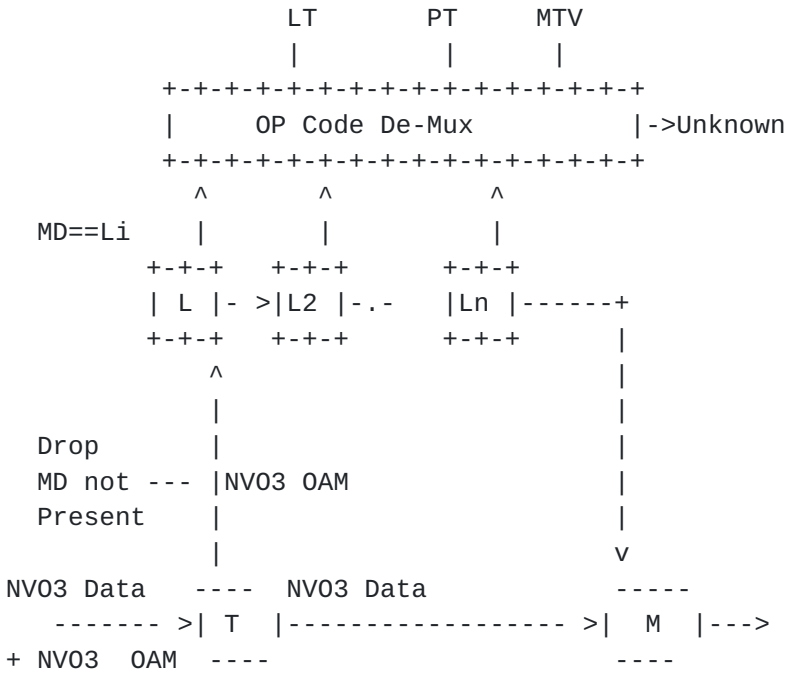


Figure 8 OAM De-Multiplexers at MIP for active SAP

T: TAP processing for MIP. All packets with OAM flag set are captured.

L : MD Level Processing, Packet with matching MD Level are "copied" to the Opcode de-multiplexer and original packet is passed on to the next MD level processor. Other packets are simply passed on to the next MD level processor, without copying to the OP code de-multiplexer.

M : Merge processor, merge OAM packets to be forwarded along with the data flow.

Packets that carry Path Trace Message (PTM) or Multi-destination Tree Verification (MTV) OpCodes are passed on to the respective processors.

Packets with unknown OpCodes are counted and discarded.

7. Continuity Check Message (CCM)

CCMs are used to monitor connectivity and configuration errors. [8021Q] monitors connectivity by having a MEP listening to periodic CCM messages received from its remote MEP partners in the MA. An [8021Q] MEP identifies cross-connect errors by comparing the MAID in the received CCM message with the MEP's local MAID. The MAID [8021Q] is a 48-byte field that is technology independent. Similarly, the MEPID is a 2-byte field that is independent of the technology. Given this generic definition of CCM fields, CCM as defined in [8021Q] can be utilized in NVO3 with no changes. NVO3 specific information may be carried in CCMs when encoded using IETF overlay specific TLVs or sub- TLVs. This is possible since CCMs are capable of carrying optional TLVs.

Unlike classical Ethernet environments with spanning tree, NVO3 supports multipath forwarding. The path taken by a packet depends on the Transport header and other parts of the packet. The Maintenance Association identifies the interested end-points (MEPs) of a given monitored path. For unicast there are only two MEPs per MA. For multicast there can be two or more MEPs in the MA. The Flow (i.e VNI and other parameters) values of the monitored flows are defined within the MA. CCM transmit logic will utilize these flow entropy values when constructing the CCM packets. Please see below for the theory of operation of CCM.

The CFM MIB of [8021Q] will be augmented with the definition of flow- entropy. Please see [TBDMIBD] for these and other NVO3 related OAM MIB definitions. The figure below depicts the correlation between MA, CCM and the flow.

CCM implementation is Optional.

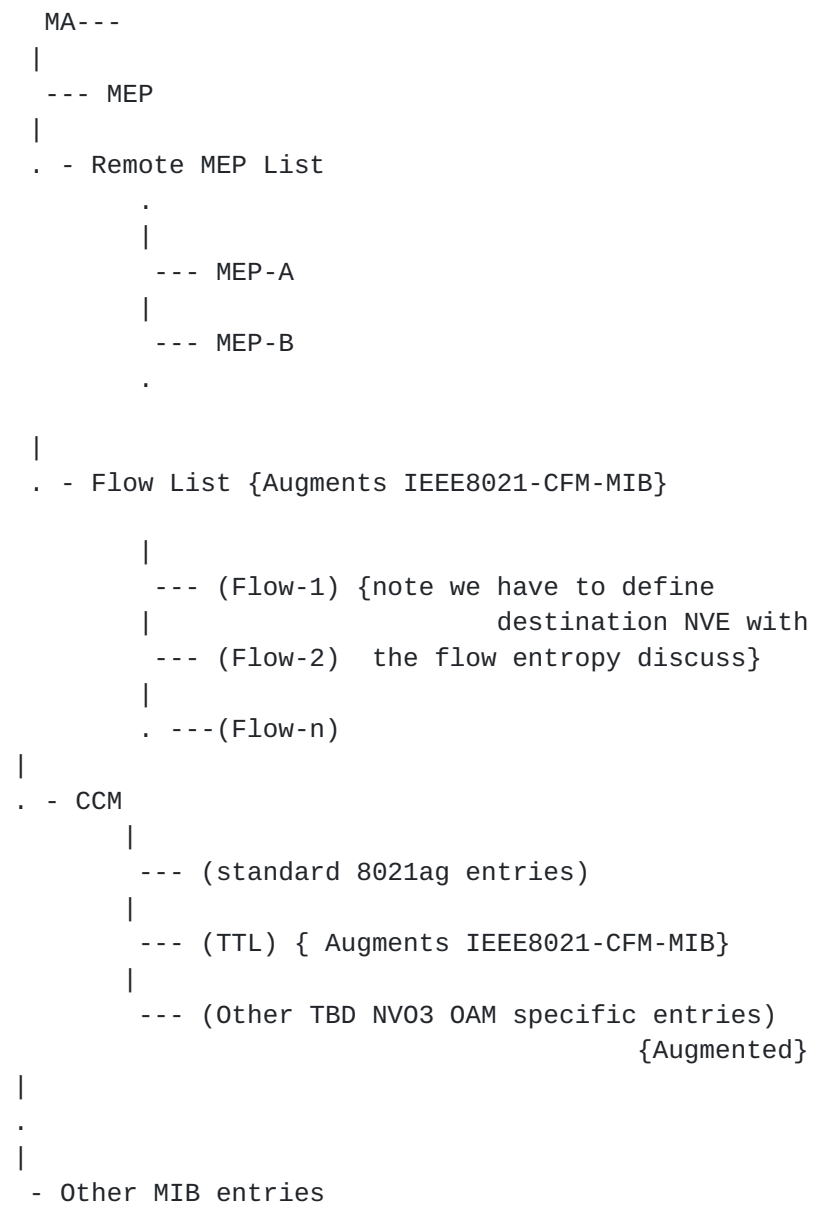


Figure 9 Augmentation of CCM MIB in NVO3

NOTE: Flow entropy field contain VNI and flow specific information

that affect the path selection and forwarding.

In a multi-pathing environment, a Flow - by definition - is unidirectional. A question may arise as to what flow entropy should be used in the response. CCMs are unidirectional and have no explicit reply; as such, the issue of the response flow entropy does not arise. In the transmitted CCM, each MEP reports local status using the Remote Defect Indication (RDI) flag. Additionally, a MEP may raise SNMP TRAPs [TBDMIB] as Alarms when a connectivity failure occurs.

8. NVO3 OAM Message Channel

The NVO3 OAM Message Channel can be divided into two parts: NVO3 OAM Message header and NVO3 OAM Message TLVs. Every OAM Message MUST contain a single OAM message header and a set of one or more specified OAM Message TLVs.

8.1. NVO3 OAM Message header

As discussed earlier, a common messaging framework between [8021Q], NVO3, and other similar standards such as Y.1731 can be accomplished by re-using the OAM message header defined in [8021Q].

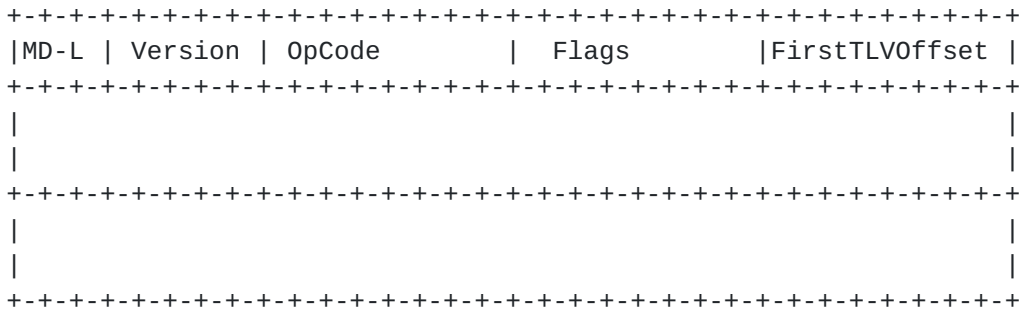


Figure 10 OAM Message Format

o MD-L: Maintenance Domain Level (3 bits). Identifies the maintenance domain level. For NVO3, in general, this field is set to a single value. If using Base Mode operations as defined in Appendix B, this field is set to 3. However, future extensions of NVO3, for example to support hierarchy, may create different MD-LEVELs and MD-L field must be appropriately set in those scenarios. (Please refer to [8021Q] for the definition of MD-Level) o Version: Indicates the version (5 bits), as specified in [8021Q]. This document does not require changing the Version defined in [8021Q].

o Flags: Includes operational flags (1 byte). The definition of flags is Opcode-specific and is covered in the applicable sections.

o FirstTLVOffset: Defines the location of the first TLV, in bytes, starting from the end of the FirstTLVOffset field (1 byte). (Refer to [8021Q] for the definition of the FirstTLVOffset.)

MD-L, Version, Opcode, Flags and FirstTLVOffset fields collectively are referred to as the OAM Message Header.

The Opcode specific information section of the OAM Message may contain Session Identification number, time-stamp, etc.

8.2. IETF Overlay OAM Opcodes

The following Opcodes are defined for IETF Overlay OAM. Each of the Opcodes indicates a separate type of OAM message. Details of the messages are presented in the related sections.

IETF OAM Message Opcodes:

64 : Path Trace Reply 65 : Path Trace Message 66 : Multicast Tree Verification Reply 67 : Multicast Tree Verification Message

8.3. Format of IETF Overlay OAM TLV

The same TLV format as defined in section 21.5.1 of [8021Q] is used for IETF Overlay OAM. The following figure depicts the general format of a IETF Overlay OAM TLV:

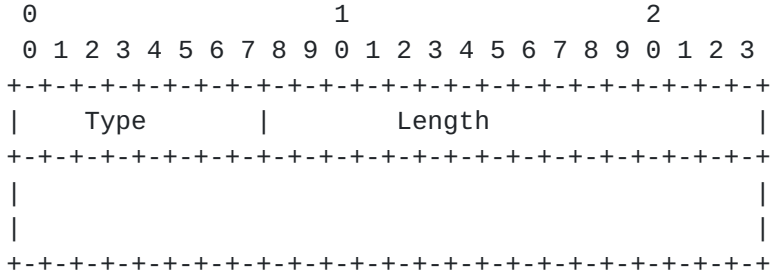


Figure 11 NVO3 OAM TLV

Type (1 octet): Specifies the Type of the TLV (see sections 8.4. for TLV types).

Length (2 octets): Specifies the length of the 'Value' field in

octets. Length of the 'Value' field can be either zero or more octets.

Value (variable): The length and the content of this field depend on the type of the TLV. Please refer to applicable TLV definitions for the details.

Semantics and usage of Type values allocated for TRILL OAM purpose are defined by this document and other future related documents.

8.4. IETF Overlay OAM TLVs

Overlay related TLVs are defined in this section. Previously defined [8021Q] TLVs are reused, where applicable. Block of 32 TLVs are allocated for the purpose of IETF defined standards

8.4.1. Common TLVs between 8201Q CFM and IETF Overlay OAM

The following TLVs are defined in [8021Q]. We propose to re-use them where applicable. The format and semantics of the TLVs are as defined in [8021Q].

Type	Name of TLV in [8021Q]
0	End TLV
1	Sender ID TLV
2	Port Status TLV
3	Data TLV
4	Interface Status TLV
5	Reply Ingress TLV
6	Reply Egress TLV
7	LTM Egress Identifier TLV
8	LTR Egress Identifier TLV
9-30	Reserved
31	Organization Specific TLV

8.4.2. IETF Overaly OAM Specific TLVs

As indicated above, a block of 32 TLVs will be requested to be reserved for IETF OAM purposes. Listed below is a summary of the IETF Overlay OAM TLVs and their corresponding codes. Format and semantics of OAM TLVs are defined in subsequent sections.

Type	TLV Name
64	OAM Application Identifier
65	Out of Band IP Address
66	Original Payload

67	Diagnostic VLAN
68	scope
69	Previous Device address
70	Next Hop Device List (ECMP)
71	Multicast Receiver Availability
72	Flow Identifier
73	Reflector Entropy
74 to 95	Reserved

8.4.3. OAM Application Identifier TLV

OAM Application Identifier TLV carries Overlay OAM application specific information. The Overlay OAM Application Identifier TLV MUST always be present and MUST be the first TLV in the OAM messages. Messages that do not include the OAM Application Identifier TLV as the first TLV MUST be discarded.

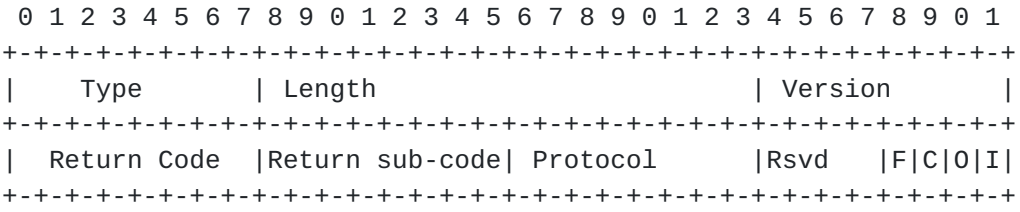


Figure 12 OAM Application Identifier TLV

Type (1 octet) = TBD-TLV-64 indicate that this is the IETF Overlay OAM Application Identifier TLV

Length (2 octets) = 6

IETF Overlay OAM Version (1 Octet), currently set to zero. Indicates the Overlay OAM version. IETF Overlay OAM version can be different than the [8021Q] version.

Return Code (1 Octet): Set to zero on requests. Set to an appropriate value in response messages.

Return sub-code (1 Octet): Return sub-code is set to zero on transmission of request message. Return sub-code identifies categories within a specific Return code. Return sub-code MUST be interpreted within a Return code.

Protocol: This indicates the overlay protocol on which the OAM is applied. In this document we cover NVO3

0 : TRILL

1 : NVO3

2 - 255 : reserved

F (1 bit): Final flag, when set, indicates this is the last response.

C (1 bit): Label error, if set indicates that the label (VLAN) in the flow entropy is different than the label included in the diagnostic TLV. This field is ignored in request messages and MUST only be interpreted in response messages.

O (1 bit): If set, indicates, OAM out-of-band response requested.

I (1 bit): If set, indicates, OAM in-band response requested.

NOTE: When both O and I bits are set to zero, indicates that no response is required (silent mode). User MAY specify both O and I or one of them or none. When both O and I bits are set response is sent both in-band and out-of-band.

8.4.4. Out Of Band Reply Address TLV

Out of Band Reply Address TLV specifies the address to which an out of band OAM reply message MUST be sent. When O bit in the Application Identifier TLV is not set, Out of Band Reply Address TLV is ignored.

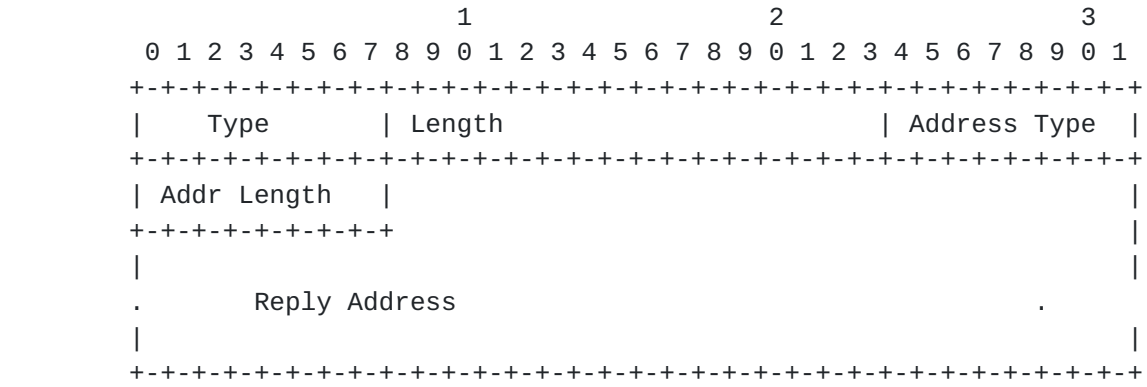


Figure 13 Out of Band IP Address TLV

Type (1 octet) = TBD-TLV-65

Length (2 octets) = Variable. Minimum length is 2.

Address Type (1 Octet): 0 - IPv4. 1 - IPv6. All other values

reserved.

Addr Length (1 Octet). 4 - IPv4. 16 - IPv6,

Reply Address (variable): Address where the reply needed to be sent. Length depends on the address specification.

8.4.5. Diagnostics Label TLV

Diagnostic label specifies the VNI which the OAM messages are generated. Receiving device MUST compare the VNI of the NVO3 shim to that of Diagnostic TLV. Label Error Flag in the response MUST be set when the two VLANs do not match.

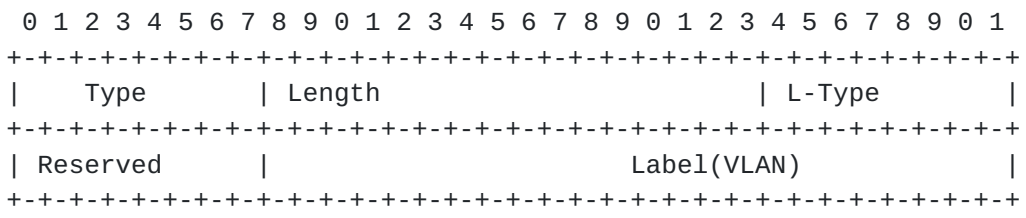


Figure 14 Diagnostic TLV

Type (1 octet) = TBD-TLV-66 indicates that this is the Diagnostic TLV

Length (2 octets) = 5

L-Type (Label type, 1 octet)

0- indicate 802.1Q 12 bit VLAN.

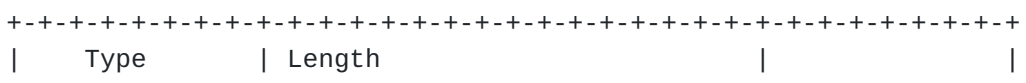
2 - 24 bit ISID

3 - VNI 24 bits

Label (24 bits): Either 12 bit VLAN or 24 bit ISID or 24 bit VNI.

Devices should perform Label error checking when Label TLV is not included in the OAM message.

8.4.6. Original Data Payload TLV



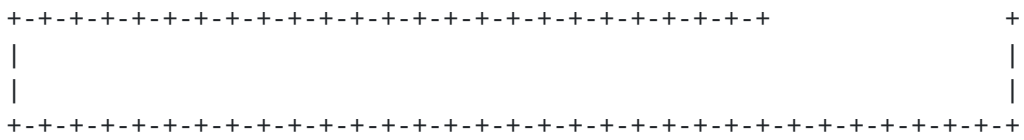


Figure 15 Original Data Payload TLV

Type (1 Octet) = TBD-TLV-67

Length (2 octets) = variable

8.4.7. Flow Identifier (flow-id) TLV

Flow Identifier (flow-id) uniquely identifies a specific flow. The flow-id value is unique per MEP and needs to be interpreted as such.

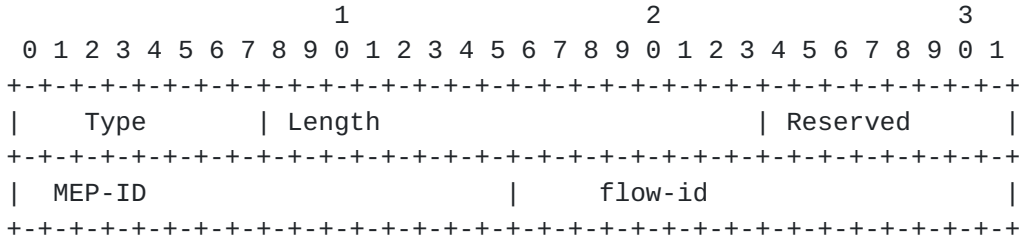


Figure 16 Flow Identifier TLV

Type (1 octet) = TBD-TLV-72 Length (2 octets) = 5.

Reserved (1 octet) set to 0 on transmission and ignored on reception.

MEP-ID (2 octets) = MEP-ID of the originator [8021Q].

Flow-id (2 octets) = uniquely identifies the flow per MEP. Different MEPs may allocate the same flow-id value. The {MEP-ID, flow-id} pair is globally unique.

Inclusion of the MEP-ID in the flow-id TLV allows inclusion of MEP-ID for messages that do not contain MEP-ID in OAM header. Applications may use MEP-ID information for different types of troubleshooting.

8.4.8. Reflector Entropy TLV

Reflector Entropy TLV is an optional TLV. This TLV, when present, tells the responder to utilize the Reflector Entropy specified

within the TLV as the flow-entropy of the response message.

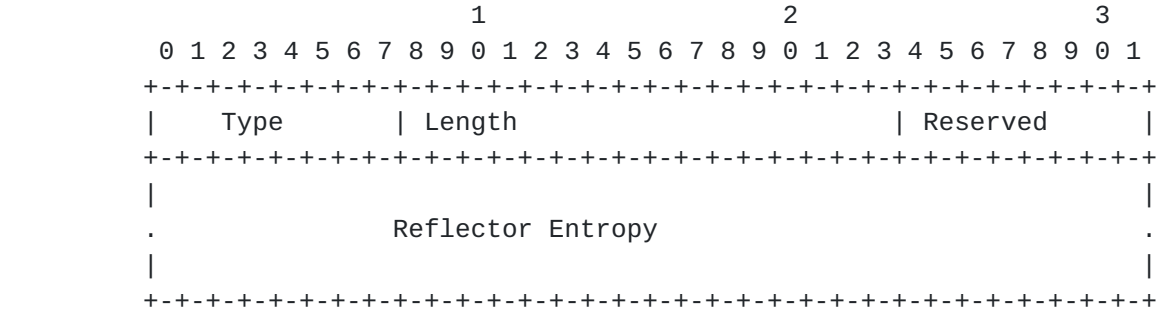


Figure 17 Reflector Entropy TLV

Type (1 octet) =TBD-TLV-73 Reflector Entropy TLV.

Length (1 octet) =97.

Reserved (1 octet) = set to zero on transmission and ignored by the recipient.

Reflector Entropy (96-octet) = Flow Entropy to be used by the responder. May be padded with zero if the desired flow entropy is less than 96 octets.

9. Loopback Message

9.1. Loopback OAM Message format

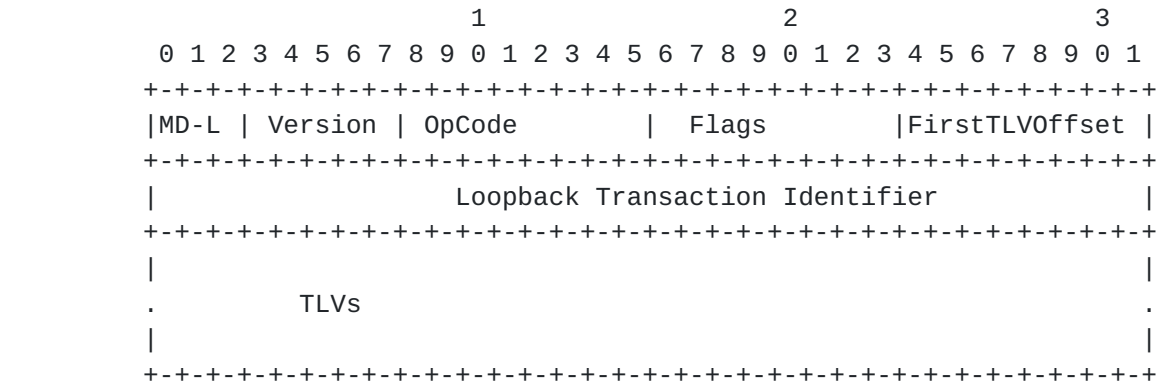


Figure 18 Loopback OAM Message Format

The above figure depicts the format of the Loopback Request and response messages as defined in [8021Q]. The Opcode for Loopback Message is set to 65 and the Opcode for the Reply Message is set to 64. The Session Identification Number is a 32-bit integer that allows the requesting Device to uniquely identify the corresponding session. Responding Device, without modification, MUST echo the received "Loopback Transaction Identifier" number..

9.2. Theory of Operation

9.2.1. Actions by Originator

Identifies the destination NVE address based on user specification or based on the specified destination, VNI and MAC

Constructs the flow entropy based on user specified parameters or implementation specific default parameters.

Constructs the OAM header: sets the opcode to Loopback message type (3). Assign applicable Loopback Transaction Identifier number for the request.

The Overlay OAM Version TLV MUST be included and with the flags set to applicable values.

Include following OAM TLVs, where applicable

- Out-of-band Reply address TLV
- Diagnostic Label TLV
- Sender ID TLV

Specify the Transport Layer parameters based on user inputs or default settings.

Dispatch the OAM frame for transmission.

Devices may continue to retransmit the request at periodic intervals, until a response is received or the re-transmission count expires. At each transmission Session Identification number MUST be incremented.

9.2.2. Intermediate Devices

Intermediate Devices forward the frame as a normal data frame and no special handling is required.

9.2.3. Destination Device

If the Loopback message is addressed to the local Device and satisfies the OAM identification criteria specified in section 4.2. then, the Device data plane forwards the message to the CPU for further processing.

The OAM application layer further validates the received OAM frame by checking for the presence of OAM-Ethertype and the MD Level. Frames that do not contain OAM-Ethertype MUST be discarded.

Construction of the OAM response:

OAM application encodes the received Transport header, NVO3 shim and payload fragment (when present) in the Original payload TLV and includes it in the OAM message.

Set the Return Code and Return sub code to applicable values. Update the OAM opcode to 2 (Loopback Message Reply)

Optionally, if the VNI identifier value of the received differs from the value specified in the diagnostic Label, set the Label Error Flag on OAM Application Identifier TLV.

Include the sender ID TLV (1)

If in-band response was requested, dispatch the frame to theNVO3 data plane with request-originator Transport Layer address as the destination address.

If out-of-band response was requested, dispatch the frame to the IP forwarding process.

10. Path Trace Message

The primary use of the Path Trace Message is for fault isolation. It may also be used for plotting the path taken from a given Device to another Device.

[8021Q] accomplishes the objectives of the NVO3 Path Trace Message using Link Trace Messages. Link Trace Messages utilize a well-known multicast MAC address. However, in NOV3 the transport Layer can be different technologies such as IP or MPLS etc. Hence, a definition of a new Path Trace message format is required for Overlay OAM. The Path Trace message is defined for the purpose.

The Path Trace Message has the same format as Loopback Message, but utilizes a different opcode set. Opcode for Path Trace Reply Message is 64 and Request Message is 65

Operation of the Path Trace message is identical to the Loopback message except that it is first transmitted with a TTL field value of 1. The sending device expects a Time Expiry Return-Code from the next hop or a successful response. If a Time Expiry Return-code is received as the response, the originator Device records the information received from intermediate node that generated the Time Expiry message and resends the message by incrementing the previous TTL value by 1. This process is continued until, a response is received from the destination Device or Path Trace process timeout occur or TTL reaches a configured maximum value.

10.1. Theory of Operation

10.1.1. Action by Originator Device Identifies the destination NVE address based on user specification or based on the specified destination, VNI and MAC

Constructs the NVO3 shim and the flow based on user specified parameters or implementation specific default parameters.

Construct the OAM header: Set the opcode to Path Trace Request message type (TBD-65). Assign an applicable Session Identification number for the request. Return-code and sub-code MUST be set to zero.

The OAM Application Identifier TLV MUST be included and set the flags to applicable values.

Include following OAM TLVs, where applicable

- Out-of-band IP address TLV
- Diagnostic Label TLV
- Include the Sender ID TLV

Specify the TTL of the transport header as 1 for the first request.

Dispatch the OAM frame to the data plane for transmission.

A Device (MEP) may continue to retransmit the request at periodic intervals, until a response is received or the re-transmission count expires. At each new re-transmission, the Session Identification

number MUST be incremented. Additionally, for responses received from intermediate devices (MIP), the device address and interface information MUST be recorded.

10.1.2. Intermediate Device

Path Trace Messages transit through Intermediate devices transparently, unless Hop-count has expired. OAM application layer further validates the received OAM frame by examining the presence of NVO3 OAM Flag and OAM-Ethertype and by examining the MD Level. Frames that do not contain OAM-Ethertype MUST be discarded.

Construction of the OAM response:

OAM application encodes the received Transport header and flow entropy in the Original payload TLV and include it in the OAM message.

Set the Return Code to (2) "Time Expired" and Return sub code to zero (0). Update the OAM opcode to 64 (Path Trace Message Reply).

If the VNI identifier value of the received OAM message differs from the value specified in the diagnostic Label, set the Label Error Flag on OAM Application Identifier TLV.

Include following TLVs

Reply Ingress TLV (5)

Reply Egress TLV (6)

Interface Status TLV (4)

Sender ID TLV (1)

If Label error detected, set C flag (Label error detected) in the version.

If in-band response was requested, dispatch the frame to the NVO3 data plane with request-originator Transport address as the destination address.

If out-of-band response was requested, dispatch the frame to the standard IP forwarding process.

10.1.3. Destination Device

Processing is identical to Loop back response processing in section 9.2.3. with the exception that OAM Opcode is set to Path Trace Reply (64).

11. Link Trace Message

Link Trace Message (LTM/LTR) procedure is defined in detail in [8021Q]. In this section I am covering the summary.

Link Trace Message are to be used when operator network all switches are OAM capable. In this scenario we will recommend 8021Q port based model for Link Trace Message and Reply Procedure as Bridge Brain is same as Path Trace message.

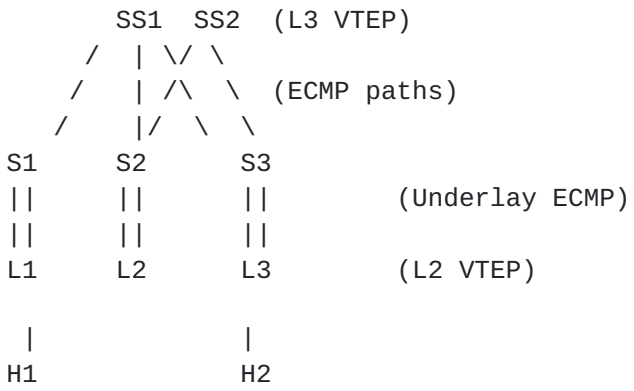


Figure 19 Payload Fragment use case

In Figure 19, if all switches in operator network are OAM capable then hardware friendly and accurate OAM solution for Fault isolation will be Link Trace.

11.1 MEP and MIP

By default if network is OAM capable all switch has virtual default MEP created on the NVE and MIP created on all fabric facing interface.

MEP can initiate and terminate the OAM message. MIP is stateless and passive and only Reply to OAM Message.

11.2 Initiator

Initiator Device generate Link Trace Message as per procedure

described in [8021Q] and encapsulated as per the draft.

As transit packet will hit the MIP on the egress port, a copy of packet is punted to cpu to generate Link Trace Reply (LTR) to the originating MEP and packet continue forwarding in hardware.

11.3 Intermediate Devices

Intermediate devices has MIP configured on all the fabric Interfaces.

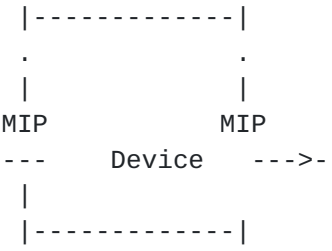


Figure 20 MIP configuration and traffic Flow

In above figure traffic is entering from left side of device hitting first MIP and copy of packet is punted to cpu to generate Link Trace Reply and LTM continue forwarding in hardware.

LTM hits right side MIP while exiting out of the device and copy of the packet is punted to cpu to generate Link Trace Reply and LTM continue forwarding.

11.4 Terminating Device

Terminating device has MIP and MEP configured, and on ingress interface packet will hit the MIP and copy of it's punted to generate LTR and packet continue forwarding in hardware.

Packet then hits the terminating MEP and LTR is generated.

11.5 Output

If for example path was like below

A Eth1 -- Eth2 B Eth3 --- Eth4 C

Link trace Message is generated from A towards C.

LTR will be received from
Eth1 A
Eth2 B

Eth3 B
Eth4 C
MEP C

In this way complete path is tracked in hardware forwarding.

12. Application of Continuity Check Message (CCM) in NVO3

Section 7. provides an overview of CCM Messages defined in [8021Q] and how they can be used within the NVO3 OAM. This section, presents the application and Theory of Operations of CCM within the NVO3 OAM framework. Readers are referred to [8021Q] for CCM message format and applicable TLV definitions and usages. Only the NVO3 specific aspects are explained below.

In NVO3, between any two given MEPs there can be multiple potential paths. Whereas in [8021Q], there is always a single path between any two MEPs at any given time. It is important that solutions to have the ability to monitor continuity over one or more paths.

CCM Messages are uni-directional, such that there is no explicit response to a received CCM message. Connectivity fault status is indicated by setting the applicable flags (e.g. RDI) of the CCM messages transmitted by an MEP.

It is important that the solution presented in this document accomplishes the requirements specified in [NVO3OAMREQ] within the framework of [8021Q] in a straightforward manner and with minimum changes. Section 8 above defines multiple flows within the CCM object, each corresponding to a flow that a given MEP wishes to monitor.

Receiving MEPs do not cross check whether a received CCM belongs to a specific flow from the originating MEP. Any attempt to track status of individual flows may explode the amount of state information that any given MEP has to maintain.

The obvious question arises: How does the originating device know which flow or flows are at fault?

This is accomplished with a combination of the RDI flag in the CCM header, flow-id TLV, and SNMP Notifications (Traps). Section 11.1 below discusses the procedure.

12.1. CCM Error Notification

Each MEP transmits 4 CCM messages per each flow. ([8021Q] detects

CCM fault when 3 consecutive CCM messages are lost). Each CCM Message has a unique sequence number and unique flow-identifier. The flow identifier is included in the OAM message via flow-id TLV.

When an MEP notices a CCM timeout from a remote MEP (MEP-A), it sets the RDI flag on the next CCM message it generates. Additionally, it logs and sends SNMP notification that contain the remote MEP Identification, flow-id and the Sequence Number of the last CCM message it received and if available, the flow-id and the Sequence Number of the first CCM message it received after the failure. Each MEP maintains a unique flow-id per each flow, hence the operator can easily identify flows that correspond to the specific flow-id.

The following example illustrates the above.

Assume there are two MEPs, MEP-A and MEP-B.

Assume there are 3 flows between MEP-A and MEP-B.

Let's assume MEP-A allocates sequence numbers as follows

Flow-1 Sequence={1,2,3,4,13,14,15,16,.. } flow-id=(1)

Flow-2 Sequence={5,6,7,8,17,18,19,20,.. } flow-id=(2)

Flow-3 Sequence={9,10,12,11,21,22,23,24,.. } flow-id=(3)

Let's Assume Flow-2 is at fault.

MEP-B, receives CCM from MEP-A with sequence numbers 1,2,3,4, but did not receive 5,6,7,8. CCM timeout is set to 3 CCM intervals in [8021Q]. Hence MEP-B detects the error at the 8'th CCM message. At this time the sequence number of the last good CCM message MEP-B has received from MEP-A is 4 and flow-id of the last good CCM Message is (1). Hence MEP-B will generate a CCM error SNMP notification with MEP-A and Last good flow-id (1) and sequence number 4.

When MEP-A switches to flow-3 after transmitting flow-2, MEP-B will start receiving CCM messages. In the foregoing example it will be CCM message with Sequence Numbers 9,10,11,12,21 and so on. When in receipt of a new CCM message from a specific MEP, after a CCM timeout, the NVO3 OAM will generate an SNMP Notification of CCM resume with remote MEP-ID and the first valid flow-id and the Sequence number after the CCM timeout. In the foregoing example, it is MEP-A, flow-id (3) and Sequence Number 9.

The remote MEP list under the CCM MIB Object is augmented to contain

"Last Sequence Number", flow-id and "CCM Timeout" variables. Last Sequence Number and flow-id are updated every time a CCM is received from a remote MEP. CCM Timeout variable is set when the CCM timeout occurs and is cleared when a CCM is received.

12.2. Theory of Operation

12.2.1. Actions by Originator Device

Derive the VNI and entropy based on flow entropy specified in the CCM Management object.

Construct the NVO3 CCM OAM header as specified in [8021Q].

OAM Version TLV MUST be included as the first TLV and set the flags to applicable values.

Include other TLVs specified in [8021Q]

Include the following optional OAM TLVs, where applicable

- Sender ID TLV

Specify the TTL of the NVO3 data frame per user specification or utilize an applicable Hop count value.

Dispatch the OAM frame to the NVO3 data plane for transmission.

An MEP transmits a total of 4 requests, each at CCM retransmission interval. At each transmission the Session Identification number MUST be incremented by one.

At the 5'th retransmission interval, flow entropy of the CCM packet is updated to the next flow entropy specified in the CCM Management Object. If current flow entropy is the last flow entropy specified, move to the first flow entropy specified and continue the process.

12.2.2. Intermediate Devices

Intermediate devices forward the frame as a normal data frame and no special handling is required.

12.2.3. Destination Device

If the CCM Message is addressed to the local MEP or multicast and satisfies OAM identification methods specified in sections 4.2. then the data plane forwards the message to the CPU for further processing.

The OAM application layer further validates the received OAM frame by examining the presence of OAM-Ethertype at the end of the flow entropy. Frames that do not contain OAM-Ethertype at the end of the flow entropy MUST be discarded.

Validate the MD-LEVEL and pass the packet to the Opcode de-multiplexer. The Opcode de-multiplexer delivers CCM packets to the CCM process.

The CCM Process performs processing specified in [8021Q].

Additionally the CCM process updates the CCM Management Object with the sequence number of the received CCM packet. Note: The last received CCM sequence number and CCM timeout are tracked per each remote MEP.

If the CCM timeout is true for the sending remote MEP, then clear the CCM timeout in the CCM Management object and generate the SNMP notification as specified above.

13. Security Considerations

Specific security considerations related methods presented in this document are currently under investigation.

14. IANA Considerations

Re-use the Type and TLV from RFC7455.

15. References

15.1. Normative References

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

[8021Q] IEEE, "Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks", IEEE Std 802.1Q-2011, August, 2011.

15.2. Informative References

[RFC4379] Kompella, K. et.al, "Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures", RFC 4379, February 2006.

[RFC6291] Andersson, L., et.al., "Guidelines for the use of the "OAM" Acronym in the IETF" RFC 6291, June 2011.

[Y1731] ITU, "OAM functions and mechanisms for Ethernet based networks", ITU-T G.8013/Y.1731, July, 2011.

[NVO3FRM] Lasserre, M., et.al., "Framework for DC Network Virtualization", Work in Progress, January, 2014.

[NVO3ARC] Black, D., et.al., "Architecture for Overlay Networks (NVO3)", Work in Progress, December, 2013.

[NVO3OAMREQ] Ashwood-Smith, P., "NVO3 Operations, Administrations and Maintenance Requirements", work in progress, January, 2014.

[NV03DPREQ] Bitar, N., "NVO3 Data Plane Requirements", Work in Progress, November, 2013.

16. Acknowledgments

This document was prepared using 2-Word-v2.0.template.dot.

Appendix A. Base Mode for NVO3 OAM

CFM, as defined in [8021Q], requires configuration of several parameters before the protocol can be used. These parameters include MAID, Maintenance Domain Level (MD-LEVEL) and MEPIDs. The Base Mode for NVO3 OAM defined here facilitates ease of use and provides out of the box plug-and-play capabilities.

All NVO3 compliant devices that support OAM specified in this document MUST support Base Mode operation.

All NVO3 compliant devices that support OAM MUST create a default MA with MAID as specified herein.

MAID [8021Q] has a flexible format and includes two parts: Maintenance Domain Name and Short MA name. In the Based Mode of operation, the value of the Maintenance Domain Name must be the character string "NVO3BaseMode" (excluding the quotes "). In Base Mode operation Short MA Name format is set to 2-octet integer format (value 3 in Short MA Format field) and Short MA name set to 65532

(0xFFFC).

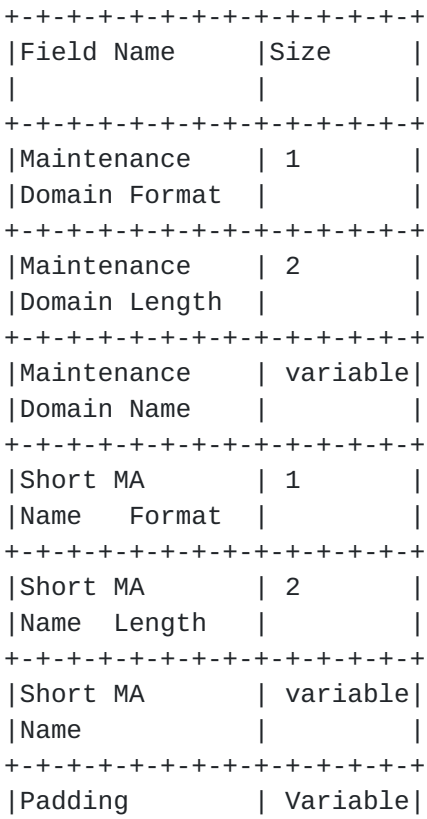
The Default MA belongs to MD-LEVEL 3.

In the Base Mode of operation, each NVO3 device creates a single UP MEP associated with a virtual OAM port located within the CPU with no physical layer (NULL PHY). The MEPID associated with this MEP is the 2-octet VNI Nickname.

By default, with the base mode OAM in place, all NVO3 devices operating in the Base Mode for NVO3 OAM are able to initiate LBM, PT and other OAM tools with no configuration.

Implementation MAY provide default flow to be included in OAM messages. Content of the default flow is outside the scope of this document.

Figure 18, below depicts encoding of MAID within the CCM messages.



+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

Figure 18 MAID structure as defined in [8021Q]

Maintenance Domain Name Format is set to Value: 4

Maintenance Domain Name Length is set to value: 13

Maintenance Domain Name is set to: NVO3BaseMode

Short MA Name Format is set to value: 3

Short MA Name Length is set to value: 2

Shirt MA Name is set to : FFFC

Padding : set of zero up to 48 octets of total length of the MAID.

Please refer to [8021Q] for details.

Authors' Addresses

Tissa Senevirathne
CISCO Systems
375 East Tasman Drive.
San Jose, CA 95134
USA.

Phone: +1 408-853-2291
Email: tsenevir@gmail.com

Norman Finn
CISCO Systems
510 McCarthy Blvd
Milpitas, CA 95035
USA

Email: nfinn@cisco.com

Samer Salam
CISCO Systems
595 Burrard St. Suite 2123

Vancouver, BC V7X 1J1, Canada

Email: ssalam@cisco.com

Deepak Kumar
CISCO Systems
510 McCarthy Blvd,
Milpitas, CA 95035, USA

Phone : +1 408-853-9760
Email: dekumar@cisco.com

Donald Eastlake
Huawei Technologies
155 Beaver Street
Milford, MA 01757

Phone: +1-508-333-2270
Email: d3e3e3@gmail.com

Sam Aldrin
Huawei Technologies
2330 Central Express Way
Santa Clara, CA 95951
USA

Email: aldrin.ietf@gmail.com