NVO3 working group
Internet Draft
Category: Standard Track
Expires: May 24,

L. Xia
Weiguo Hao
Huawei
Greg Mirsky
Ericsson

November 24, 2014

Detecting NVO3 Overlay Point-to-Multipoint Data Plane failures
draft-xia-nvo3-overlay-p2mp-ping-01

Abstract

This draft refers to P2MP LSP ping, and describes NVO3 overlay P2MP ping mechanisms by extending the existing NVO3 overlay P2P ping mechanisms for applying to NVO3 overlay P2MP path in order to simplify implementation and network operation.

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/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 May 24, 2015.

Table of Contents

1. Introduction

For NVO3 network, comprehensive OAM tool set is very important. NVO3 point-to-point (P2P) ping is a simple and efficient mechanism for detecting data plane failures of layer 2 (L2) or layer 3 (L3) virtual network overlaid on IPv4 or IPv6 underlay networks. [NVO3OVERLAYOAM] and [NVO3OVERLAYPING] have described P2P ping mechanism for various overlay technologies (i.e., VXLAN [RFC7348], NVGRE [NVGRE], etc) of NVO3 network.

NVO3 P2P ping follows the basic idea of LSP ping described in [RFC4379]. Which is modeled after the ping/traceroute paradigm: ping (ICMP Echo Request [RFC792]) is used for connectivity verification, and traceroute is used for hop-by-hop fault localization as well as path tracing. In the ping mode, the NVO3 Echo Request message is sent by an NVE within a tested overlay path, along the same data path to the destination NVE as other packets. Upon reception, NVO3 Echo Request is passed to the control plane of the destination NVE to verify if it is indeed an egress NVE for the tested overlay path. In the traceroute mode, the NVO3 Echo Request message is sent in band and reach the control plane of each transit node. The node performs various checks to verify it is indeed a transit node for this path. If error or ping reply timeout is found in one transit node, fault localization is achieved. Also it can trace the underlay path that is exercised by any given overlay path. NVO3 Echo Reply message is used by egress NVEs or transit nodes to report the results to the ingress NVE. It can be sent in band or out-of-band. Only NVO3 Echo Request message must have the same NVO3 overlay encapsulation as regular data plane packets to ensure they share the same data path.

NVO3 network needs to support Broadcast, Unknown unicast, Multicast (BUM) traffics. [NVO3MCAST] discusses four methods to handle BUM traffic in NVO3 network:

1. No multicast support.

2. Replication at the source NVE.

3. Replication at a centralized multicast service node.

4. IP multicast in the underlay.

NVO3 point-to-multipoint (P2MP) ping should be supported for each tenant's multicast service's connectivity verification, fault

localization and path tracing. The NVO3 P2MP ping must have the ability to:

- Support the connectivity verification from an arbitrary NVE to either a specific set of NVEs or all NVEs overlaid on the underlay Multicast Distribution Tree (MDT);

- Support the fault localization in the underlay MDT;

- Support the path tracing of the underlay MDT exercised by any given overlay path.

This draft draws on P2MP LSP ping [RFC6425] solution, and describes NVO3 overlay P2MP ping mechanisms by extending the existing NVO3 overlay P2P ping mechanisms for applying to NVO3 overlay P2MP path.

1.1. 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].

1.2. Terminology

This document uses the terms defined in NVO3 framework [RFC7365], [RFC6425] and [NVO3OVERLAYPING].

BUM - Broadcast, Unknown unicast, Multicast

IGMP - Internet Group Management Protocol

MDT - Multicast Distribution Tree

MLD - Multicast Listener Discover

NVE - Network Virtualization Edge

PIM-DM - Protocol Independent Multicast-Dense Mode

PIM-SM - Protocol Independent Multicast-Sparse Mode

PIM-SSM - Protocol Independent Multicast-Source-Specific Multicast

2. Overview

2.1. NVO3 Multicast Methods

This section gives a simple summary of current existed NVO3 multicast methods.

When the NVO3 underlay network supports multicast, the underlay MDT is established according to the topology of tenant network and BUM packets type:

- For broadcast or unknown unicast packets, all NVEs in a L2 overlay network are involved. Each NVE within a L2 overlay network of a tenant network can be the ingress NVE for the MDT, the rest NVEs connected to other parts of the same L2 overlay network are all egress NVEs for that MDT;

- For L3 multicast packets, possibly only part of NVEs in a tenant network are involved. The NVE connected to the multicast source node is the ingress NVE; other NVEs connected to the nodes that need to receive multicast traffic are all egress NVEs.

In this method, the ingress NVE directly encapsulates the BUM packets with the appropriate IP multicast address in the tunnel encapsulation header for delivery to the desired set of egress NVEs. All egress NVEs that belong to the same multicast group MUST send IGMP/MLD packets to underlay network edge devices to trigger underlay network MDT establishment.

When the NVO3 underlay network does not support multicast, the mechanism of head-end replication at the ingress NVE or centralized replication at the multicast service node can be used. These two mechanisms use unicast NVO3 encapsulation to send BUM traffic to all destination NVEs and don't rely on multicast forwarding capability of underlay network.

2.2. NVO3 P2MP Ping Mechanisms

Comparing to NVO3 P2P overlay services, NVO3 P2MP overlay services are more complex. To meet the requirements mentioned in Section 1, the NVO3 P2MP ping mechanisms need some extensions to the current NVO3 ping mechanisms.

This draft refers to [RFC6425] and describes the following extensions:

- The packet format extension to the NVO3 overlay P2MP Echo Request/Reply messages;

- The extended mechanisms of NVO3 overlay P2MP ping operations;

- Different operation mechanism for two scenarios - with or without underlay multicast support;

- Special considerations for traceroute function;

- Support for the hierarchical NVE scenario.

NVO3 overlay P2MP ping packet is an IPv4 or IPv6 UDP packet identified by well known UDP Port 3503, and the basic structure of the packet remains the same as defined in Section 3 of [RFC4379].

3. Packet Format

This document also references Section 4 of [NVO3OVERLAYPING] as the extended specification to the payload of inner UDP of NVO3 Echo Request/Reply packet format, which includes:

- a new "N" flag (NVO3 PATH Ping) in Global Flags field for identifying a NVO3 ping Echo message;

- The extended specification of Message Type, Reply Mode, Return Code and Return Subcode in the contents of the Echo UDP message part;

- Target Object TLV: A list of newly defined sub-TLVs (i.e., IPv4/IPv6 prefix sub-TLVs for egress NVE address, L2/L3 VN ID sub-TLVs for L2/L3 VNI) for validating the target objects of NVO3 P2P path;

- Downstream Detailed Mapping Extension and related Multipath information encoding algorithm.

Besides the above extensions, NVO3 P2MP ping mechanisms need to define new TLVs and sub-TLVs to support the functionalities specific to its P2MP application.

3.1. Identifying the P2MP NVO3 Overlay Path

3.1.1. No Multicast Support by NVO3 Underlay Network

If NVO3 underlay network does not support multicast, NVO3 overlay network replicates and sends BUM traffic at ingress NVE or multicast

service node by unicast. NVO3 P2MP ping mechanisms require ingress NVE or multicast service node to use NVO3 P2P ping sessions to all or desired subset of egress NVEs. The ingress NVE (or multicast service node) MUST use the Target Object TLVs for P2P ping sessions with target egress NVEs. It's noted that, for the multicast service node case, the E2E P2MP ping session also includes a P2P ping sub-session between the ingress NVE and the multicast service node.

3.1.2. Multicast Support by NVO3 Underlay Network

3.1.2.1. Broadcast and Unknown Unicast Packets Case

If NVO3 underlay network supports multicast, a new Target Object TLV--L2 VN ID sub-TLV is defined to validate the VN context of L2 network for broadcast or unknown unicast packets. The example TLV format for VXLAN is as followed:

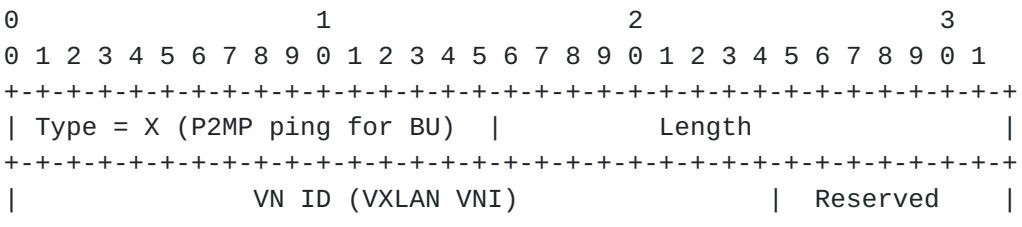


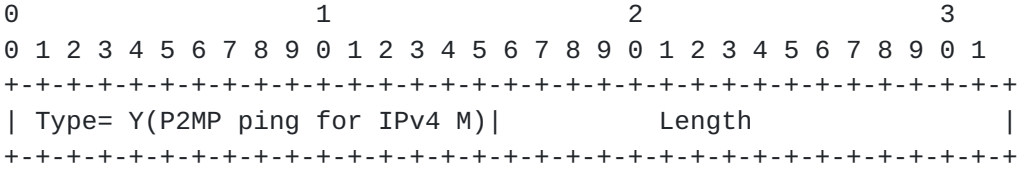
Figure 1: L2 VN ID sub-TLV for VXLAN

Note: BU - Broadcast, Unknown unicast.

The above TLV format can be a general format used in P2MP ping Echo message for all NVO3 overlay technologies (i.e. VXLAN, NVGRE, etc) to validate the VN context of L2 network for the case of broadcast or unknown unicast packets.

3.1.2.2. Multicast Packets Case

For L3 multicast case, two new Target Object TLVs: VN IPv4/IPv6 multicast sub-TLVs are defined for IPv4 and IPv6 overlay network as followed:



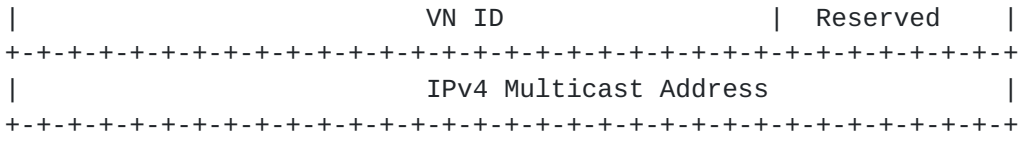


Figure 2: VN multicast sub-TLV for IPv4 overlay network

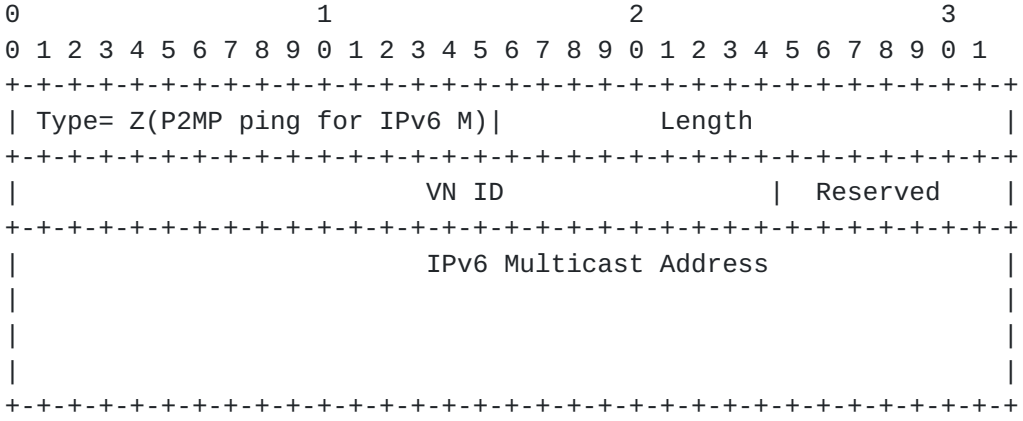


Figure 3: VN multicast sub-TLV for IPv6 overlay network

Note: M - Multicast.

In this case, in addition to the VN ID used for the validation of L2/L3 VN context, IPv4/IPv6 multicast address is needed to check if the target NVEs are on the path to the receiver of the corresponding MDT. NVEs MUST be able to snoop IGMP/MLD messages in order to remember it has multicast receiver attached.

3.2. Limiting the Scope of Responses

To limit the scope of responses, NVO3 P2MP ping mechanisms MUST include specific TLV in Echo Request message to make the appointed target NVEs to respond Echo Request message, and other NVEs do not respond it.

Section 3.2 of [RFC6425] defines the P2MP Responder Identifier TLV and four sub-TLVs (IPv4/IPv6 Egress Address P2MP Responder sub-TLVs, IPv4/IPv6 Node Address P2MP Responder sub-TLVs). The responding node upon receiving any of these sub-TLVs would respond or not respond to the Echo Request message thus achieving goal of limiting the scope of responses.

To use IPv4/IPv6 Egress Address P2MP Responder sub-TLVs properly, the nodes that locate upstream in the MDT must know all the egress NVE IP addresses. Current multicast protocols (i.e., PIM-SM, PIM-DM, IGMP, MLD, etc) cannot meet this requirement. So, these two sub-TLVs cannot be used for the NVO3 P2MP ping mechanisms.

Without any extra requirement, IPv4/IPv6 Node Address P2MP Responder sub-TLVs definition and related operations remain the same for NVO3 P2MP ping mechanisms.

3.3. Other TLVs and Flags

The Echo Jitter TLV and the intended behavior of the responding node upon receiving it, defined in Section 3.3 of [RFC6425], remain the same to help to limit congestion of Echo replies.

The Respond Only If TTL Expired Flag in the Global Flags field defined in Section 3.4 of [RFC6425] is inherited here with one change: TTL to be checked is in the outer IP header.

4. Theory of Operation

This section refers to Section 4 of [RFC6425] to describe how the NVO3 P2MP Echo messages are processed at various nodes, e.g. ingress NVE, transit nodes, etc. This section also discusses the different mechanisms for ping mode and traceroute mode.

Note that the main goal of this section is to describe the changes that existing P2MP MPLS ping operations need to make in order to make it applicable for NVO3 network. The unchanged operations are listed here briefly. The details can refer to Section 4 of [RFC6425].

4.1. No Multicast Support by NVO3 Underlay Network

For ping mode, NVO3 P2MP Echo Request messages are replicated at ingress NVE or multicast service node and sent each copy of the message to destination NVEs through unicast NVO3 encapsulation. This mode is more often used by Connectivity Verification function.

For traceroute mode, NVO3 P2MP Echo Request messages are replicated at ingress NVE or multicast service node, and then each copy of the message will be sent to desired transit nodes through unicast encapsulation with specific TTL value in outer IP header. This mode can be used for Fault Localization or Path Tracing functions.

Ingress NVE or multicast service node can send Echo Request messages to one or a set of members of egress NVEs, or their transit nodes by

their choice. Because all the Echo Request messages are sent by unicast, the operations onto them remain the same as with NVO3 P2P ping mechanisms defined in [NVO3OVERLAYPING].

In addition, ingress NVE is still possible to send multiple Echo Request messages to multiple target nodes and gets many Echo Reply messages in a short period, which will cause congestion on the ingress NVE. So, the Echo Jitter TLV and according random delay reply mechanism defined in [RFC6425] should be supported for this case.

4.2. Multicast Support by NVO3 Underlay Network

Because NVO3 underlay network supports multicast protocol, NVO3 overlay network can use the underlay MDT to transfer BUM traffics and the NVO3 P2MP Echo Request messages. The elements in the MDT include ingress node, transit node, branch node, bud node, egress node. They have different operations on the P2MP Echo Request messages which are described in the following sections.

4.2.1. Ingress NVE Operations

Ingress NVE should follow the procedure defined in [RFC4379] to construct the P2MP Echo Request message. The packet format MUST be the extended format described in Section 3 of this document and it MUST be IP encapsulated.

Echo Request message will contain a Target Object TLV to identify NVE membership. L2 VN ID sub-TLV is for broadcast or unknown unicast case, VN IPv4/IPv6 multicast sub-TLVs is for L3 multicast case.

Echo Request message can contain IPv4/IPv6 Node Address P2MP Responder sub-TLVs to limit responses from only those targeted nodes under the ping mode.

The Echo Jitter TLV can also be contained to limit the congestion in the ingress NVE.

4.2.2. Responding Node Operations

Except for the ingress NVE, other nodes potentially are to act as a responding node.

Usually the Echo Request message will be addressed to the egress and/or bud nodes. In case of TTL Expiry, i.e. in the traceroute mode, the Echo Request message may stop at branch or transit nodes. In

both scenarios, the Echo Request message will be passed on to the control plane to generate the Echo Reply message.

After the rate-limit control and sanity check, the responding node MUST determine how to reply based on the Reply Mode field. Then, the responding node MUST determine if it is on the underlay MDT in question which the overlay P2MP path is tunneled on by checking the destination multicast address against the control plane:

- If the responding node is not on the underlay MDT in question, it MUST send Echo Reply with Return Code "Replying node has no control plane mapping for the Target Object" and Return Subcode as 1 which implies the MDT address check failure, as defined in [NVO3OVERLAYPING];

- If the node is on the underlay MDT in question, the node MUST check whether or not the Echo Request is directed to it:

    - If a P2MP Responder Identifier TLV is present, then the node MUST follow the procedures defined in Section 3.2 to determine whether or not it should respond to the request;

    - If the P2MP Responder Identifier TLV is not present (or, in the error case, is present, but does not contain any sub-TLVs), then the node MUST respond according to [NVO3OVERLAYPING] processing rules.

The following sections describe the expected values of Return Codes for various nodes on the underlay MDT which the overlay P2MP path is tunneled on. Note that the Return Code might change based on the presence of a Responder Identifier TLV or Downstream Detailed Mapping TLV.

4.2.2.1. Responses from Transit and Branch Nodes

The presence of a Responder Identifier TLV does not influence the choice of the Return Code. To report success, the Return Code MAY be set to "Packet-Forward-OK". For error conditions, use appropriate values defined in [NVO3OVERLAYPING].

The presence of a Downstream Detailed Mapping TLV will influence the choice of Return Code. As per [RFC6424], the Return Code may be set to "See DDM TLV for Return Code and Return Subcode". The Return Code for each Downstream Detailed Mapping TLV will depend on the downstream path as described in [RFC6424].

There will be a Downstream Detailed Mapping TLV for each downstream path being reported in the Echo Reply. Hence, for transit nodes, there will be only one such TLV, and for branch nodes, there will be more than one.

4.2.2.2. Responses from Egress Nodes

The presence of a Responder Identifier TLV does not influence the choice of the Return Code. To report success, the Return Code MAY be set to "Egress for the Target", as defined in [NVO3OVERLAYPING]. For error conditions, use appropriate values defined in [NVO3OVERLAYPING].

To check whether a NVE is an egress NVE for an overlay P2MP path, two cases need to be differentiated according to previous analysis:

- o Broadcast or unknown unicast overlay path case: L2 VN ID sub-TLV should be contained in the Echo Request message, which includes the L2 VN ID need to be validated. An NVE determines if it is the egress NVE for this VN by checking if there is an entry of this VN ID existed in the NVE. If it is the egress NVE, MUST return the success response;

- o L3 multicast overlay path case: VN IPv4/IPv6 multicast sub-TLV should be contained in the Echo Request message. Except for the same VN ID validation procedure as above, NVE also needs to check if it is on the path to a receiver of this IPv4/IPv6 multicast address, which is included in VN IPv4/IPv6 multicast sub-TLV. Only if all the validations are OK, the NVE thinks itself an egress NVE and MUST return the success response.

To save the MDT number in the NVO3 underlay network, multiple overlay P2MP paths may share the same underlay MDT. The egress NVEs of one overlay P2MP path of them may be a subset of the egress nodes of MDT. When the NVO3 P2MP Echo Request messages for this overlay P2MP path are sent to all the egress nodes of MDT and get response from them, some egress NVEs that do not belong to this overlay P2MP path SHOULD reply with Return Code "Replying node has no control plane mapping for the Target Object" and Return Subcode as 1 to the ingress NVE. Note that this does not mean the error condition for the egress NVE.

The presence of the Downstream Detailed Mapping TLV does not influence the choice of Return Code. Egress NVEs do not put in any Downstream Detailed Mapping TLV in the Echo Reply as per [RFC6424].

4.2.2.3. Responses from Bud Nodes

The processing at bud nodes is more complex than other types of MDT nodes. The bud node behaves as either an egress node or a transit node, or a combination of an egress and branch node. This behavior is determined by the presence of any Responder Identifier TLV. Similarly, the Downstream Detailed Mapping TLV can influence the Return Code values.

To determine the behavior of the bud node, use the following rules:

- If the Responder Identifier TLV is not present, then the node MUST behave as a combination of egress and branch node;

- If the Responder Identifier TLV containing a Node Address sub-TLV is present, and:

  - If the address specified in the sub-TLV matches to an address in the node, then the node MUST behave like a combination of egress and branch node;

  - If the address specified in the sub-TLV does not match any address in the node, then no reply SHOULD be sent.

Once the node behavior has been determined, the possible values for Return Codes are as follows:

- If the node is behaving as a combination of egress and branch node, and:

  - If a Downstream Detailed Mapping TLV is not present, then for a success response, the Return Code SHOULD be set to "Egress for the Target". For error conditions, use appropriate values defined in [NVO3OVERLAYPING];

  - If a Downstream Detailed Mapping TLV is present, then for a success response, the Return Code SHOULD be set to "Egress for the Target". For error conditions, use appropriate values defined in [NVO3OVERLAYPING]. The Return Code for the each Downstream Detailed Mapping TLV will depend on the downstream path as described in [RFC6424]. There will be a Downstream Detailed Mapping TLV for each downstream path from the node.

4.3. Special Considerations for Traceroute

Comparing to P2P ping, P2MP ping has a new problem of how to figure out the end of the traceroute processing. The relevant background can refer to Section 4.3.1 of [RFC6425]. For the two cases of NVO3 P2MP ping:

- o Replication at ingress NVE or multicast service node: Because the initiating node has a priori knowledge about the number of egress nodes and their addresses, it is possible to continue processing until a valid reply has been received from each end point, provided that the replies can be matched correctly to the egress NVEs;

- o Underlay network multicast supporting: the ingress NVE might not always know about all of the egress NVEs. Hence, there might not be a definitive way to estimate the end of processing for traceroute. Therefore, a configurable upper limit on TTL values is a good solution for this case. By this way, the user can choose the depth to which the tree will be probed.

The other problems of traceroute mode for MPLS P2MP ping also exist for NVO3 P2MP ping. The specific mechanisms defined in Sections 4.3.2, 4.3.3, 4.3.4, 4.3.5 of [RFC6425] are all applicable to NVO3 P2MP ping.

5. Hierarchical NVE Scenario

TBD

6. Security Considerations

TBD.

7. Acknowledgements

8. IANA Considerations

8.1. TLVs

The TLVs and Sub-TLVs requested by this draft for IANA consideration are the following:

Type	Sub-Type	Value Field
XXX		Target Object
	X	L2 VN ID
	X	VN IPv4/IPv6 multicast

9. References

9.1. Normative References

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

[RFC4379] Allan, D. and T. Nadeau, "A Framework for Multi-Protocol Label Switching (MPLS) Operations and Management (OAM)", RFC 4378, February 2006

[RFC792] Postel, J., "Internet Control Message Protocol", STD 5, RFC 792, September 1981

[RFC6425] Saxena, S., et al, "Detecting Data Plane Failures in Point-to-Multipoint Multiprotocol Label Switching (MPLS) - Extensions to LSP", RFC 6425, November 2011

[RFC6424] Bahadur, N., Kompella, K., and G. Swallow, "Mechanism for Performing LSP-Ping over MPLS Tunnels", RFC 6424, November 2011

[RFC7348] Mahalingam, M., Dutt, D., et al, "Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", RFC7348, August, 2014

[RFC7365] LASSERRE, M., Motin, T., et al, "Framework for DC Network Virtualization", RFC7365, October, 2014

9.2. Informative References

[NVO3OVERLAYOAM] Jain, P., et al, "Generic Overlay OAM and Datapath Failure Detection", draft-jain-nvo3-overlay-oam-02, work in progress, October, 2014

[NVO3OVERLAYPING] Kumar, N., "Detecting NVO3 Overlay Data Plane failures", draft-kumar-nvo3-overlay-ping-01, work in progress, January, 2014

[NVGRE] Sridharan, M., et al, "NVGRE: Network Virtualization using Generic Routing Encapsulation", draft-sridharan-virtualization-nvgre-06, work in progress, October, 2014

[NVO3MCAST] Ghanwani, A., et al, "Framework of Supporting Applications Specific Multicast in NVO3", draft-ghanwani-nvo3-app-mcast-framework-00, work in progress, June, 2014

Authors' Addresses

Liang Xia
Huawei Technologies

Email: frank.xialiang@huawei.com

Weiguo Hao
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China
Phone: +86-25-56623144
Email: haoweiguo@huawei.com

Greg Mirsky
Ericsson

EMail: gregory.mirsky@ericsson.com