

nvo3
Internet-Draft
Intended status: Standards Track
Expires: July 19, 2014

N. Kumar
C. Pignataro
D. Rao
Cisco Systems, Inc.
S. Aldrin
Huawei Technologies, Inc.
January 15, 2014

Detecting NV03 Overlay Data Plane failures
draft-kumar-nvo3-overlay-ping-01

Abstract

This document describes a simple and efficient mechanism to perform L2 or L3 VN Context validation and to detect any data plane failures in IPv4 or IPv6 based overlay network providing L2 or L3 virtualized network.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of [BCP 78](#) and [BCP 79](#).

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on July 19, 2014.

Copyright Notice

Copyright (c) 2014 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

Internet-Draft Detecting NV03 Overlay Data Plane Failures January 2014

the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1.	Introduction	2
2.	Requirements notation	3
3.	Terminology	3
4.	Packet Format	3
4.1.	Return Code and Return Subcode	4
4.2.	TLV Format	5
4.2.1.	Target Object TLV	5
4.2.2.	Downstream Detailed Mapping Extension	5
4.2.3.	Multipath Information Encoding	6
5.	NV03 Echo Packet Indicator	7
5.1.	When the core is IPv6 network	7
5.2.	When the core is IPv4 network	8
6.	Theory of Operation	8
6.1.	Sending NV03 Echo Request	8
6.2.	Receiving NV03 Echo Request	8
6.2.1.	Transit Node procedure	9
6.2.2.	Edge Node procedure	10
6.3.	Sending NV03 Echo Reply	10
6.4.	Receiving NV03 Echo Reply	11
6.5.	Dealing with Equal-Cost-Multi-Path (ECMP)	11
7.	Connectivity verification between Tenant system	12
8.	IANA Considerations	12
8.1.	Message Types, Reply Modes, Return Codes	12
8.2.	TLVs	12
9.	Security Considerations	12
10.	Acknowledgement	13
11.	References	13
11.1.	Normative References	13
11.2.	Informative References	13
	Authors' Addresses	13

[1.](#) Introduction

I.D-ietf-nvo3-framework [[I-D.ietf-nvo3-framework](#)] specifies a framework that defines mechanism to support large scale network virtualization by connecting L2 or L3 virtualized network over L3 tunnels. Various tunneling options like IPv4, IPv6 or MPLS can be used in underlying network.

[Section 3.8](#) of [I.D-ietf-nvo3-dataplane-requirement] specifies the requirement of OAM tool that performs connectivity verification and fault isolation along with revealing ECMP paths between NVE nodes. While the mechanism described in [RFC4379](#) [[RFC4379](#)] helps with

Internet-Draft Detecting NV03 Overlay Data Plane Failures January 2014

satisfying this OAM requirement when MPLS tunnel is used, there is no native way to achieve the same when IPv4 or IPv6 is used as tunneling option.

This document describes a simple and efficient mechanism to perform L2 or L3 VN Context validation and to detect any data plane failures in IPv4 or IPv6 overlay network by re-purposing and extending MPLS Ping mechanism defined in [RFC4379](#) [[RFC4379](#)]. This document also describes the mechanism to reveal all available paths (multi path) between any ingress and egress NVE nodes.

[2.](#) Requirements notation

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [[RFC2119](#)].

[3.](#) Terminology

L2 VN: Layer 2 Virtual Network

L3 VN: Layer 3 Virtual Network

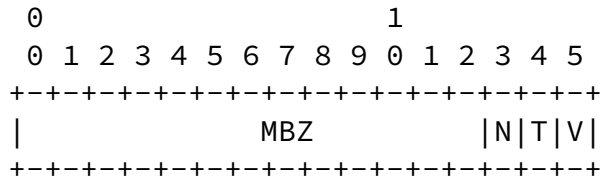
NVE: Network Virtualization Edge

ECMP: Equal Cost multiple path

[4.](#) Packet Format

NV03 PATH Ping packet is a IPv4 or IPv6 UDP packet and the basic structure of the packet remains the same as mentioned in [Section 3 of RFC4379](#) [[RFC4379](#)].

This document introduces a new flag in Global Flags field defined in [RFC4379](#) [[RFC4379](#)]. The new format of the Global Flags field is:



The V flag is described in [RFC4379](#) [RFC4379] and T flag is described in [RFC6425](#) [RFC6425].

The N flag (NV03 PATH Ping) MUST be set in echo request and reply packet only when it is used to validate NV03 Path.

The Message Type is one of the following:

Value	Meaning
-----	-----
11	NV03 Echo Request
12	NV03 Echo Reply

The Reply Mode can be one of the following:

Value	Meaning
-----	-----
11	Do not Reply
12	Reply via IPv4/IPv6 UDP packet

Return codes and Subcodes are described in [section 4.1](#).

The Sender's Handle, Sequence Number, TimeStamp sent and TimeStamp Received field are as mentioned in [Section 3 of RFC4379](#) [RFC4379].

The TLV format is same as mentioned in [\[RFC4379\]](#) and this document introduce a new TLV described later.

[4.1](#). Return Code and Return Subcode

Responder uses Return code field to reply with validity check or any error message to Initiator. It doesnt carry any meaning in Echo Request and should be set as zero.

The Return Code can be one of the following:

Value	Meaning
-----	-----
100	No Return Code
101	Malformed Echo Request Received
102	One or more TLVs not understood
103	Egress for the Target
104	No control plane mapping for the Target Object <RSC>
105	Downstream Detailed Mapping mismatch
108	Packet-Forward-OK

The Return Subcode contains the pointer in Target Object TLV for which the Replying node doesn't have a control plane mapping. For example, when NVE receives Target Object TLV with multiple Sub-TLVs and if NVE doesn't have an entry for second Sub-TLV should include 2 as RSC value.

[4.2.](#) TLV Format

The document introduces the below list of TLVs used in NV03 Echo packets:

Type	Value Field
-----	-----
101	Target Object

[4.2.1.](#) Target Object TLV

Target Object TLV is a list of Sub-TLVs that carries the element against which the path or control plane validation is done.

This document defines the below Target Object Sub-TLVs:

Sub-Type	Length	Value Field
-----	-----	-----
1	5	IPv4 prefix
2	17	IPv6 Prefix
3	variable	L2 VN ID
4	variable	L3 VN ID

NV03 ECHO Request MUST have a Target Object TLV with atleast one Sub-TLV which describe the egress node about the element to be validated.

For example, if NVE X wanted to verify that MAC M1 is associated with VN ID VN1, it carries relevant information like VN ID and the MAC address in Sub-TLV type 3 and send to egress NVE. Egress NVE on receiving NV03 Echo Request will validate the Target Object and will reply back with respective Return Code.

New Sub-TLVs can be proposed as and when required in future.

4.2.2. Downstream Detailed Mapping Extension

This document extends the Downstream Detailed Mapping TLV defined in [Section 3.3 of RFC6424](#) [RFC6424] to be used in NV03 scenarios with IPv4 or IPv6 based core network. This document introduces a new DS flag and the format is as below:

```
0 1 2 3 4 5 6 7
+--+--+--+--+--+
|Rsvd(MBZ)|P|I|N|
+--+--+--+--+--+
```

Flag	Name and Meaning
----	-----
P	Set when used in NV03 Ping

The P flag (NV03 Path ping) MUST be set in Downstream Detailed Mapping TLV only when it is used in NV03 scenarios. When P flag is set, I flag MUST NOT be set.

For simplicity, The DDMAP with N flag set in DS flag will be referred as NV03 DDMAP in this document.

This document also defines the below new multipath types to be used in NV03 Path ping.

Type	Meaning	Multipath information
-----	-----	-----
11	UDP Port Mask	UDP Port and bit mask
12	Flow Label Mask	IPv6 Flow label and bit mask

Multipath Information

UDP port or IPv6 Flow label range encoded according to the Multipath type. The next section explains the encoding details.

4.2.3. Multipath Information Encoding

Based on the Multipath type, the Multipath Information encodes Flow label range or UDP port range that will exercise each path. The Multipath encoding follows the same procedure specifies in [Section 3.3.1 of RFC4379](#) [RFC4379]. For completeness, it is explained in this document with UDP port range and IPv6 FLOW label range.

Multipath type 1 encodes UDP port range. The UDP prefix is formatted as a base UDP port value with non-prefix low-order bits set to zero. Since the UDP port is 16 bits, the leading 16 bits are set as zeros. The maximum prefix (including leading zeros) length can be 27. Following the prefix is a mask of length $2^{(32-\text{prefix-length})}$ bits. Each bit set to one represents a valid UDP port. UDP port values of all the odd numbers between 32704 and 3267 would be encoded as below:

0	1	2	3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1			
+-+-+-----	+-+-+-----	+-+-+-----	+-+-+-----
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 0 0 0			
+-+-+-----	+-+-+-----	+-+-+-----	+-+-+-----
0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1			
+-+-+-----	+-+-+-----	+-+-+-----	+-+-+-----
0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1			

Extension Header with NV03 OAM flag set. Any node on replying back with NV03 Echo Reply MUST include IPv6 OAM Extension Header with NV03 OAM flag set. Any transit node on receiving IPv6 destined packet with TTL=1 SHOULD interpret the payload as NV03 ECHO packet if IPv6 OAM Extension header is present.

[5.2.](#) When the core is IPv4 network

Any transit node on receiving IPv4 destined packet with TTL=1 SHOULD interpret the payload as NV03 ECHO packet if UDP port is 3503.

[6.](#) Theory of Operation

NV03 Echo Request and Reply packet are UDP packet with destination port as 3503. This section describes the procedure in Initiating and responding nodes.

[6.1.](#) Sending NV03 Echo Request

Initiator MUST include the respective Sub-TLV for the target(s) to be validated (L2 or L3 VNI or NVE address) in Target Object TLV. It also MUST set the N flag in Global Flags ([Section 4](#)) and set the message type as 11. The Reply mode is set to the desired mode (type 11 or 12); Return code and Return Subcode are set to zero. The Sender's Handle and Sequence number are set by Initiator.

The source UDP port can be chosen by the Initiator and the destination UDP port is set to 3503. The IP header is set as follows: the source IP address is a routable address of the sender; the destination IP address is the target egress NVE address. When the core is IPv6 network, Initiator MUST include IPv6 OAM Extension header.

In ping mode (Connectivity check), the IP TTL is set to 255. In traceroute mode (Fault isolation), the IP TTL is set successively from 1 and MUST stop sending the Request if it receives a reply with Return code 103 or 104.

[6.2.](#) Receiving NV03 Echo Request

Sending an NV03 Echo Request to control plane for payload processing is done by IP TTL expiration in case of IPv4 and a combination of IP TTL expiration and IPv6 OAM Extension header incase of IPv6. The control plane further identifies it as NV03 Echo packet by a combination of UDP destion port 3503 and N flag in Global flag field ([Section 4](#)).

Any node on receiving NV03 Echo Request MUST send Echo Reply with Return code "Malformed Echo Request received" to Initiator if the packet fails sanity check. If the sanity check for NV03 Echo Request is fine, Any node should store the below temporary variables:

- o Interface-I: Interface over which Echo Request is received.
- o Address-A: Local address on Interface-I.
- o Index-I: Interface Index for upstream node interface connected to Interface-I.
- o Destination-D: Destination address of the NV03 Echo Request.

[6.2.1.](#) Transit Node procedure

Any transit node on receiving NV03 Echo Request should perform the below:

1. Transit node MUST only check the first (or top) Sub-TLV and MUST NOT iterate to other Sub-TLVs beneath. Ideally the first Sub-TLV will be IPv4 prefix or IPv6 prefix Sub-TLV on which the transit node is required to act upon. It MUST set the Return code as "One or more TLVs not understood" if the first (or top) Sub-TLV in Target Object TLV is not understood.
2. If the TLV is understood, Transit node MUST perform longest match lookup in its local forwarding table and set the Return code as "Replying node has no control plane mapping for the Target Object" and Return Subcode as 1, if there is no matching entry for the prefix in Sub-TLV in its local forwarding table.
3. If the received NV03 Echo Request has NV03 Downstream Detailed Mapping TLV MUST set the Return code as "Downstream Detailed Mapping mismatch" if any of the below fails:
 - * When Address Type is 1, Address-A SHOULD match Downstream Interface Index and Router ID or Address-A SHOULD match Downstream Address.
 - * When Address Type is 2 or 4, Index-I SHOULD match Downstream Interface Index and Router ID SHOULD match Downstream Address.
 - * When Address Type is 3, Address-A SHOULD match Downstream Interface Index and Router ID SHOULD match Downstream Address.

4. If the IP prefix in Sub-TLV matches any entry in local forwarding table and if step 4 satisfies the received NV03 DDMAP or if there

is no NV03 DDMAP received in Echo Request, Transit node MUST set the Return code as "Packet-Forward-OK" and set DDMAP field for each available multipath egress interface in forwarding table.

5. If the received NV03 Echo Request has Multipath information sub-TLV in NV03 DDMAP, Transit node MUST reply as mentioned [section 5.5](#).

[6.2.2](#). Edge Node procedure

Any Edge node (NVE) on receiving NV03 Echo Request should perform the below:

1. If the sanity check is fine, NVE MUST check all the Sub-TLVs and MUST set the Return code as "One or more TLVs not understood" if any of the TLV is not understood.
2. If the TLVs are understood, NVE node MUST send Echo Reply with Return code "Replying node has no control plane mapping for the Target Object" and Return Subcode as 1, if there is no matching entry in local forwarding table to take a forwarding decision.
3. NVE node MUST set the Return code as "Replying node is the egress for the Target" if the IP address in first (or top) Sub-TLV in Target Object TLV is locally configured to this node and there is no further sub-TLV in Target Object TLV
4. If the Target Object TLV have more than one Sub-TLV, NVE MUST validate all the Sub-TLVs and set the Return code as "Replying node has no control plane mapping for the Target Object" and Return Subcode as pointer of the Sub-TLV which fails the validation.
5. If the received Echo Request has NV03 Downstream Detailed Mapping TLV MUST check as mentioned in step 4 of [section 6.2.1](#).
6. If the validation for all Sub-TLVs in Target Object TLV is fine, NVE MUST set Return code as "Replying node is the egress for the Target" and SHOULD NOT include NV03 DDMAP.

[6.3.](#) Sending NV03 Echo Reply

NV03 Echo Reply is a UDP packet and MUST be sent only in response to received NV03 Echo Request. The format of NV03 Echo Reply is same as Echo request.

Responder MUST fill the DDMAP field, Return Code and Return Subcode from previous section. It MUST also set the N flag in Global Flags

and set the message type as 12. The Sender's Handle, Sequence Number field MUST be copied from the received Echo Request.

The source UDP port is set to 3503 and the source UDP port of received Echo Request is copied to destination UDP port of Echo Reply. The IP header is set as follows: the source IP address is a routable address of the responder; the destination IP address is copied from source address of Echo Request and IP TTL is set to 255. When the core is IPv6 network, Responder MUST include IPv6 OAM Extension Header.

[6.4.](#) Receiving NV03 Echo Reply

Any node should receive NV03 Echo Reply only in response to an NV03 Echo Request that it sent. Initiator MUST drop the packet if it fails sanity check. If the sanity check is fine, the Echo Reply should be mapped with the respective Echo Request using the destination port and Sender's Handle. If there is no match, the Echo Reply MUST be ignored. Else, it checks the Sequence Number to match the iteration.

In traceroute mode, If the Echo Reply contains NV03 DDMAP, it SHOULD copy the same to subsequent Echo Request(s).

[6.5.](#) Dealing with Equal-Cost-Multi-Path (ECMP)

For redundancy and load balancing purpose, It is common to see multiple equal cost paths between ingress and egress NVE and it is a local matter to transit node to decide the egress interface based on local hashing algorithm. It is common to see deployment with routers that support load-balancing based on UDP ports or based on IPv6 Flow label. So it is useful to have the OAM tool to exercise all possible

paths between ingress and egress NVEs.

This can be achieved using Multipath Information Sub-TLV in NV03 DDMAP. This can be used as follows:

- o When the core is IPv6 network, the Initiator will send Echo Request in traceroute mode (start with TTL as 1 and increment for subsequent message) with Multipath type set to 2. Any transit node will include multipath encoding for each downstream interface in a way that the local hashing decision based on IPv6 flow label will use the respective downstream path. Initiator will then send NV03 Echo Request with respective Flow label to exercise these paths.
- o When the core is Ipv4 network, the Initiator will send Echo Request in traceroute mode with Multipath type set to 1. Any

transit node will include multipath encoding for each downstream interface in a way that local hashing decision based on UDP port will use the respective downstream path. initiator will then send NV03 Echo Request with respective UDP port as source port to exercise these paths.

[7.](#) Connectivity verification between Tenant system

As like other overlay ping mechanism, the approach discussed in this document will help with connectivity verification between NVE nodes and control plane validation on NVE nodes.

Any dataplane programming corruption for VN context details in NVE nodes will be in different layer and may need end-to-end connectivity verification procedures.

[8.](#) IANA Considerations

This document reuse UDP port 3503 for NV03 Echo packets.

[8.1.](#) Message Types, Reply Modes, Return Codes

This document request to assign the Message Types and Reply mode mentioned in [Section 4](#) and Return code mentioned in [Section 4.1](#)

[8.2.](#) TLVs

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

Type	Sub-Type	Value Field
-----	-----	-----
101		Target Object
	1	IPv4 Prefix
	2	IPv6 Prefix
	3	L2 VN ID
	4	L3 VN ID

[9.](#) Security Considerations

The security consideration for NV03 Ping is similar to ICMP or LSP Ping. AS like ICMP or LSP ping, NV03 may be exposed to Denial-of-service attacks and it is RECOMMENDED to regulate the NV03 Ping packet flow to control plane. A rate limiter SHOULD be applied to avoid any attack

As like ICMP or LSP Ping, a traceroute can be used to obtain network information. It is RECOMMENDED that the implementation check the

source address of the Echo messages against any local secured list like access list before processing the message further

[10.](#) Acknowledgement

The authors would like to thank Lizhong Jin for his review and comments.

[11.](#) References

[11.1.](#) Normative References

[I-D.ietf-nvo3-framework]

Lasserre, M., Balus, F., Morin, T., Bitar, N., and Y. Rekhter, "Framework for DC Network Virtualization", [draft-ietf-nvo3-framework-04](#) (work in progress), November 2013.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate

Requirement Levels", [BCP 14](#), [RFC 2119](#), March 1997.

- [RFC4379] Kompella, K. and G. Swallow, "Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures", [RFC 4379](#), February 2006.
- [RFC6424] Bahadur, N., Kompella, K., and G. Swallow, "Mechanism for Performing Label Switched Path Ping (LSP Ping) over MPLS Tunnels", [RFC 6424](#), November 2011.
- [RFC6425] Saxena, S., Swallow, G., Ali, Z., Farrel, A., Yasukawa, S., and T. Nadeau, "Detecting Data-Plane Failures in Point-to-Multipoint MPLS - Extensions to LSP Ping", [RFC 6425](#), November 2011.

[11.2](#). Informative References

- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", [RFC 3031](#), January 2001.

Authors' Addresses

Nagendra Kumar
Cisco Systems, Inc.
7200 Kit Creek Road
Research Triangle Park, NC 27709
US

Email: naikumar@cisco.com

Kumar, et al.

Expires July 19, 2014

[Page 13]

Internet-Draft Detecting NV03 Overlay Data Plane Failures January 2014

Carlos Pignataro
Cisco Systems, Inc.
7200 Kit Creek Road
Research Triangle Park, NC 27709-4987
US

Email: cpignata@cisco.com

Dhananjaya Rao
Cisco Systems, Inc.

170 W Tasman Drive
San Jose, CA 95138
US

Email: dhrao@cisco.com

Sam Aldrin
Huawei Technologies, Inc.
1188 Central Express Way
Santa Clara, CA 95051
US

Email: aldrin.ietf@gmail.com