Network Working Group
Internet-Draft
Updates: 4379 (if approved)
Intended status: Standards Track
Expires: June 8, 2015

J. Luo, Ed.
ZTE
L. Jin, Ed.

T. Nadeau, Ed.
Lucidvision
G. Swallow, Ed.
Cisco

December 5, 2014

Relayed Echo Reply mechanism for LSP Ping
draft-ietf-mpls-lsp-ping-relay-reply-06

Abstract

In some inter autonomous system (AS) and inter-area deployment scenarios for RFC 4379 "Label Switched Path (LSP) Ping and Traceroute", a replying LSR may not have the available route to the initiator, and the Echo Reply message sent to the initiator would be discarded resulting in false negatives or complete failure of operation of LSP Ping and Traceroute. This document describes extensions to LSP Ping mechanism to enable the replying Label Switching Router (LSR) to have the capability to relay the Echo Response by a set of routable intermediate nodes to the initiator. This document updates RFC 4379.

Table of Contents

1. Introduction

This document describes the extensions to the Label Switched Path (LSP) Ping as specified in [RFC4379], by adding a relayed echo reply mechanism which could be used to detect data plane failures for the inter autonomous system (AS) and inter-area LSPs. The extensions are to update the [RFC4379]. Without these extensions, the ping functionality provided by [RFC4379] would fail in many deployed inter-AS scenarios, since the replying LSR in one AS may not have the available route to the initiator in the other AS. The mechanism in this document defines a new message type referred as "Relayed Echo Reply message", and a new TLV referred as "Relay Node Address Stack TLV".

This document is also to update [RFC4379], include updating of Echo Request sending procedure in section 4.3 of [RFC4379], Echo Request receiving procedure in section 4.4 of [RFC4379], Echo Reply sending procedure in Section 4.5 of [RFC4379], Echo Reply receiving procedure in section 4.6 of [RFC4379].

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 [RFC2119].

2. Motivation

LSP Ping [RFC4379] defines a mechanism to detect the data plane failures and localize faults. The mechanism specifies that the Echo Reply should be sent back to the initiator using an UDP packet with the IPv4/ IPv6 address of the originating LSR. This works in administrative domains where IP addresses reachability are allowed among LSRs, and every LSR is able to route back to the originating LSR. However, in practice, this is often not the case due to intra-provider routing policy, route hiding, and network address translation at autonomous system border routers (ASBR). In fact, it is almost uniformly the case that in inter-AS scenarios, it is not

allowed the distribution or direct routing to the IP addresses of any of the nodes other than the ASBR in another AS.

Figure 1 demonstrates a case where one LSP is set up between PE1 and PE2. If PE1's IP address is not distributed to AS2, a traceroute from PE1 directed to PE2 could fail if the fault exists somewhere between ASBR2 and PE2. Because P2 cannot forward packets back to PE1 given that it is an routable IP address in AS1 but not routable in AS2. In this case, PE1 would detect a path break, as the Echo Reply messages would not be received. Then localization of the actual fault would not be possible.

Note that throughout the document, routable address means that it is possible to route an IP packet to this address using the normal information exchanged by the IGP operating in the AS

```
+-------+   +-------+   +------+   +------+   +------+   +------+
|       |   |       |   |      |   |      |   |      |   |      |
|  PE1  +---+   P1  +---+ ASBR1+---+ ASBR2+---+  P2  +---+  PE2 |
|       |   |       |   |      |   |      |   |      |   |      |
+-------+   +-------+   +------+   +------+   +------+   +------+
<---------------AS1------------><---------------AS2----------->
<--------------------------- LSP ------------------------------>
```

Figure 1: Simple Inter-AS LSP Configuration

A second example that illustrates how [RFC4379] would be insufficient would be the inter-area situation in a seamless MPLS architecture [I-D.ietf-mpls-seamless-mpls] as shown below in Figure 2. In this example LSRs in the core network would not have IP reachable route to any of the ANs. When tracing an LSP from one AN to the remote AN, the LSR1/LSR2 node could not make a response to the Echo Request either, like the P2 node in the inter-AS scenario in Figure 1.

```
           +-------+   +-------+   +------+   +------+
           |       |   |       |   |      |   |      |
        +--+ AGN11 +---+ AGN21 +---+ ABR1 +---+ LSR1 +--> to AGN
       /   |       |  /|       |   |      |   |      |
+----+/    +-------+\/ +-------+   +------+  /+------+
| AN |              /\                     \/
+----+\    +-------+  \+-------+   +------+/\ +------+
       \   |       |   |       |   |      |  \|      |
        +--+ AGN12 +---+ AGN22 +---+ ABR2 +---+ LSR2 +--> to AGN
           |       |   |       |   |      |   |      |
           +-------+   +-------+   +------+   +------+
static route     ISIS L1 LDP             ISIS L2 LDP
<-Access-><--Aggregation Domain--><---------Core--------->
```

Figure 2: Seamless MPLS Architecture

This document describes extensions to the LSP Ping mechanism to facilitate a response from the replying LSR, by defining a mechanism that uses a relay node (e.g, ASBR) to relay the message back to the initiator. Every designated or learned relay node must be reachable to the next relay node or to the initiator. Using a recursive approach, relay node could relay the message to the next relay node until the initiator is reached.

The LSP Ping relay mechanism in this document is defined for unicast case. How to apply the LSP Ping relay mechanism in multicast case is out of the scope.

3. Extensions

[RFC4379] describes the basic MPLS LSP Ping mechanism, which defines two message types, Echo Request and Echo Reply message. This document defines a new message, Relayed Echo Reply message. This new message is used to replace Echo Reply message which is sent from the replying LSR to a relay node or from a relay node to another relay node.

A new TLV named Relay Node Address Stack TLV is defined in this document, to carry the IP addresses of the possible relay nodes for the replying LSR.

In addition, a new Return Code is defined to notify the initiator that the packet length is exceeded unexpected by the Relay Node Address Stack TLV.

It should be noted that this document focuses only on detecting the LSP which is set up using a uniform IP address family type. That is,

all hops between the source and destination node use the same address family type for their LSP ping control planes. This does not preclude nodes that support both IPv6 and IPv4 addresses simultaneously, but the entire path must be addressable using only one address family type. Supporting for mixed IPv4-only and IPv6-only is beyond the scope of this document.

3.1. Relayed Echo Reply message

The Relayed Echo Reply message is a UDP packet, and the UDP payload has the same format with Echo Request/Reply message. A new message type is requested from IANA.

```
New Message Type:
    Value    Meaning
    -----    -------
    TBD      MPLS Relayed Echo Reply
```

The use of TCP and UDP port number 3503 is described in [RFC4379] and has been allocated by IANA for LSP Ping messages. The Relayed Echo Reply message will use the same port number.

3.2. Relay Node Address Stack

The Relay Node Address Stack TLV is an optional TLV. It MUST be carried in the Echo Request, Echo Reply and Relayed Echo Reply messages if the echo reply relayed mechanism described in this document is required. Figure 3 illustrates the TLV format.

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                Type           |              Length           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Initiator Source Port       |   Number of Relayed Addresses |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
~                Stack of Relayed Addresses                     ~
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 3: Relay Node Address Stack TLV

- Type: to be assigned by IANA. A value should be assigned from 32768-49161 as suggested by [RFC4379] Section 3.

- Length: the length of the value field in octets.

- Initiator Source Port: the source UDP port that the initiator uses in the Echo Request message, and also the port that is expected to receive the Echo Reply message.

- Number of Relayed Addresses: an integer indicating the number of relayed addresses in the stack.

- Stack of Relayed Addresses: a list of relay node addresses.

The format of each relay node address is 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|        Address  Type         | Address Length|  Reserved   |K|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~           Relayed Address (0, 4, or 16 octects)               ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Type#	Address Type	Address Length
0	Unspecified	0
1	IPv4	4
2	IPv6	16

Reserved: This field is reserved and MUST be set to zero.

K bit: if the K bit is set to 1, then this sub-TLV MUST be kept in Relay Node Address Stack during TLV compress process described in section 4.2. The entry with Unspecified Address Type SHOULD NOT set K bit.

Having the K bit set in the relay node address entry causes that entry to be preserved in the Relay Node Address Stack TLV for the entire traceroute operation. A responder node MAY set the K bit to ensure its relay node address entry remains as one of the relay nodes in the Relay Node Address Stack TLV. The address with K bit set will always be a relay node address for the Relayed Echo Reply, see section 4.3. Some nodes could be configured to always set the K bit, or the module handling MPLS echo requests could discover its K bit use through topology awareness. One application scenario of K bit is given out in section 5.

Relayed Address: this field specifies the node address, either IPv4 or IPv6.

3.3. New Return Code

A new Return Code is used by the replying LSR to notify the initiator that the packet length is exceeded unexpected by the Relay Node Address Stack TLV.

New Return Code:

Value	Meaning
TBD	Response Packet length was exceeded by the Relay Node Address Stack TLV unexpected

4. Procedures

4.1. Sending an Echo Request

In addition to the procedures described in section 4.3 of [RFC4379], a Relay Node Address Stack TLV MUST be carried in the Echo Request message to facilitate the relay functionality.

When the Echo Request is first sent by the initiator, a Relay Node Address Stack TLV with the initiator address in the stack and its source UDP port MUST be included. That will ensure that the first relay node address in the stack will always be the initiator address.

For the subsequent Echo Request messages, the initiator would copy the Relay Node Address Stack TLV from the received Echo Reply message.

4.2. Receiving an Echo Request

In addition to the processes in section 4.4 of [RFC4379], the procedures of the Relay Node Address Stack TLV are defined here.

Upon receiving a Relay Node Address Stack TLV of the Echo Request message, the receiver MUST check the addresses of the stack in sequence from top to bottom (the first address in the stack will be the first one to be checked), to find out the first routable IP address. Those address entries behind of the first routable IP address in the address list with K bit set to 0 MUST be deleted, and the address entry of the replying LSR MUST be added at the bottom of the stack. The address entry added by the replying LSR MUST be same as the source IP address of Relay Echo Reply (section 4.3) or Echo Reply message (section 4.5) being sent. A second or more address entries could also be added if necessary, which depends on implementation. Those address entries with K bit set to 1 MUST be

kept in the stack. The updated Relay Node Address Stack TLV MUST be carried in the response message.

If the replying LSR is configured to hide its routable address information, the address entry added in the stack SHOULD be a blank entry with Address Type set to unspecified. The blank address entry in the receiving Echo Request SHOULD be treated as an unroutable address entry.

If the packet length was exceeded unexpectedly by the Relay Node Address Stack TLV, the TLV SHOULD be returned back unchanged in the Echo Reply message. And the new return code in section 3.3 SHOULD be used to notify the initiator of the situation.

An LSR not recognize the Relay Node Address Stack TLV, SHOULD ignore it according to section 3 of [RFC4379].

4.3. Originating an Relayed Echo Reply

To find out the next relay node address, the node SHOULD check the address items in Relay Node Address Stack TLV in sequence from top to down, and find the first IP routable address, e.g., A, and the last address with K bit set, e.g., B. If address A is before address B in Relay Node Address Stack TLV, then use address B as the next relay node address. Otherwise, use address A as the next relay node address. If there is no B existed, then use A as the next relay node address. If the resolved next relay node address is not routable, then sending of Relayed Echo Reply or Echo Reply will fail.

When the replying LSR receives an Echo Request, and the first IP address in the Relay Node Address Stack TLV is not the next relay node address, the replying LSR SHOULD send a Relayed Echo Reply message to the next relay node. The processing of Relayed Echo Reply is the same with the procedure of the Echo Reply described in Section 4.5 of [RFC4379], except the destination IP address and the destination UDP port. The destination IP address of the Relayed Echo Reply is set to the next relay node address from the Relay Node Address Stack TLV, and both the source and destination UDP port is set to 3503. The updated Relay Node Address Stack TLV described in section 4.2 MUST be carried in the Relayed Echo Reply message.

4.4. Relaying an Relayed Echo Reply

Upon receiving an Relayed Echo Reply message with its own address as the destination address in the IP header, the relay node SHOULD find out the next relay node address as described in section 4.3.

If the next relay node address is not the first one in the address list, e.g, another intermediate relay node, the relay node SHOULD send an Relayed Echo Reply message to this next relay node with the payload unchanged. The TTL of the Relayed Echo Reply SHOULD be copied from the received Relay Echo Reply and decremented by 1.

Note, the next relay node address MUST be located before the source IP address of the received Relayed Echo Reply which MUST be also in the stack, otherwise the Relayed Echo Reply SHOULD NOT be sent, so as to avoid potential loop.

4.5. Sending an Echo Reply

The Echo Reply is sent in two cases:

1. When the replying LSR receives an Echo Request, and the first IP address in the Relay Node Address Stack TLV is the next relay node address (section 4.3), the replying LSR would send an Echo Reply to the initiator. In addition to the procedure of the Echo Reply described in Section 4.5 of [RFC4379], the updated Relay Node Address Stack TLV described in section 4.2 MUST be carried in the Echo Reply.

2. When the intermediate relay node receives a Relayed Echo Reply, and the first IP address in the Relay Node Address Stack TLV is the next relay node address (section 4.3), the intermediate relay node would send the Echo Reply to the initiator with the UDP payload unchanged other than the Message Type field (change from type of Relayed Echo Reply to Echo Reply). The destination IP address of the Echo Reply is set to the first IP address in the stack, and the destination UDP port would be copied from the Initiator Source Port field of the Relay Node Address Stack TLV. The source UDP port should be 3503. The TTL of the Echo Reply SHOULD be copied from the received Relay Echo Reply and decremented by 1.

4.6. Receiving an Echo Reply

In addition to the processes in Section 4.6 of [RFC4379], the initiator would copy the Relay Node Address Stack TLV received in the Echo Reply to the next Echo Request.

4.7. Impact to Traceroute

Source IP address in Echo Reply and Relay Echo Reply is to be of the address of the node sending those packets, not the original responding node. Then the traceroute address output module will print the source IP address as below:

```
       if (Relay Node Address Stack TLV exists) {
           Print the last address in the stack;
       } else {
           Print the source IP address of Echo Reply message;
       }
```

5. LSP Ping Relayed Echo Reply Example

Considering the inter-AS scenario in Figure 4 below.

```
   +-------+   +-------+   +------+   +------+   +------+   +------+
   |       |   |       |   |      |   |      |   |      |   |      |
   |  PE1  +---+   P1  +---+ ASBR1+---+ ASBR2+---+  P2  +---+  PE2 |
   |       |   |       |   |      |   |      |   |      |   |      |
   +-------+   +-------+   +------+   +------+   +------+   +------+
   <----------------AS1------------><---------------AS2------------>
   <--------------------------- LSP ------------------------------->
```

Figure 4: Example Inter-AS LSP

In the example, an LSP has been created between PE1 to PE2. When performing LSP traceroute on the LSP, the first Echo Request sent by PE1 with outer-most label TTL=1, contains the Relay Node Address Stack TLV with PE1's address.

After processed by P1, P1's address will be added in the Relay Node Address Stack TLV address list following PE1's address in the Echo Reply.

PE1 copies the Relay Node Address Stack TLV into the next Echo Request when receiving the Echo Reply.

Upon receiving the Echo Request, ASBR1 checks the address list in the Relay Node Address Stack TLV in sequence, and finds out that PE1's address is routable. Then deletes P1's address, and adds its own address following PE1 address. As a result, there would be PE1's address followed by ASBR1's address in the Relay Node Address Stack TLV of the Echo Reply sent by ASBR1.

PE1 then sends an Echo Request with outer-most label TTL=3, containing the Relay Node Address Stack TLV copied from the received Echo Reply message. Upon receiving the Echo Request message, ASBR2 checks the address list in the Relay Node Address Stack TLV in sequence, and finds out that PE1's address is IP route unreachable, and ASBR1's address is the first routable one in the Relay Node

Address Stack TLV. So ASBR1 is the next relay node. ASBR2 adds its address as the last address item following ASBR1's address in Relay Node Address Stack TLV, sets ASBR1's address as the destination address of the Relayed Echo Reply, and sends the Relayed Echo Reply to ASBR1.

Upon receiving the Relayed Echo Reply from ASBR2, ASBR1 checks the address list in the Relay Node Address Stack TLV in sequence, and finds out that PE1's address is first routable one in the address list. So PE1 is the next relay node. Then ASBR1 sends an Echo Reply to PE1 with the payload of the received Relayed Echo Reply unchanged other than the Message Type field.

For the Echo Request with outer-most label TTL=4, P2 checks the address list in the Relay Node Address Stack TLV in sequence, and finds out that both PE1's and ASBR1's addresses are not IP routable, and ASBR2's address is the first routable address. Then P2 sends an Relayed Echo Reply to ASBR2 with the Relay Node Address Stack TLV containing four addresses, PE1's, ASBR1's, ASBR2's and P2's address in sequence.

Then according to the process described in section 4.4, ASBR2 sends the Relayed Echo Reply to ASBR1. Upon receiving the Relayed Echo Reply, ASBR1 sends an Echo Reply to PE1 which is IP routable. And as relayed by ASBR2 and ASBR1, the Echo Reply would finally be sent to the initiator PE1.

For the Echo Request with outer-most label TTL=5, the Echo Reply would relayed to PE1 by ASBR2 and ASBR1, similar to the case of TTL=4.

The Echo Reply from the replying node which has no IP reachable route to the initiator is finally transmitted to the initiator by multiple relay nodes.

In the case that the interface address of ASBR1 to P1 is IP1 which maybe an IPv4 private address and not IP routable for AS2, and the loopback address on ASRB1 is IP2 which is routable for AS2. Then when ASBR1 sends a Relayed Echo Reply, it will firstly add IP1 without K bit set in the Relay Node Address Stack TLV, and then add IP2 with K bit set in the stack TLV. Then ASBR2/P2 could relay the Relayed Echo Reply back first to IP2 which is routable for ASBR2/P2, then ASBR1 will send Echo Reply to PE1. Thanks for the K bit, the ASBR1 will not be skipped for message relay.

6. Security Considerations

The Relayed Echo Reply mechanism for LSP Ping creates an increased risk of DoS by putting the IP address of a target router in the Relay Node Address Stack. These messages then could be used to attack the control plane of an LSR by overwhelming it with these packets. A rate limiter SHOULD be applied to the well-known UDP port on the relay node as suggested in [RFC4379]. The node which acts as a relay node SHOULD validate the relay reply against a set of valid source addresses and discard packets from untrusted border router addresses. An implementation SHOULD provide such filtering capabilities.

If an operator wants to obscure their nodes, it is RECOMMENDED that they may replace the replying node address that originated the Echo Reply with blank address in Relay Node Address Stack TLV.

Other security considerations discussed in [RFC4379], are also applicable to this document.

7. Backward Compatibility

When one of the nodes along the LSP does not support the mechanism specified in this document, the node will ignore the Relay Node Address Stack TLV as described in section 4.2. Then the initiator may not receive the Relay Node Address Stack TLV in Echo Reply message from that node. In this case, an indication should be reported to the operator, and the Relay Node Address Stack TLV in the next Echo Request message should be copied from the previous Echo Request, and continue the ping process. If the node described above is located between the initiator and the first relay node, the ping process could continue without interruption.

8. IANA Considerations

IANA is requested to assign one new Message Type, one new TLV and one new Return Code.

8.1. New Message Type

This document requires allocation of one new message type from "Multi-Protocol Label Switching (MPLS) Label Switched Paths (LSPs) Ping Parameters" registry, the "Message Type" registry:

Value	Meaning
TBD	MPLS Relayed Echo Reply

The value should be assigned from the "Standards Action" range (0-191), and using the lowest free value within this range.

8.2. New TLV

This document requires allocation of one new TLV from "Multi-Protocol Label Switching (MPLS) Label Switched Paths (LSPs) Ping Parameters" registry, the "TLVs" registry:

Type	Meaning
TBD	Relay Node Address Stack TLV

A suggested value should be assigned from "Standards Action" range (32768-49161) as suggested by [RFC4379] Section 3, using the first free value within this range.

8.3. New Return Code

This document requires allocation of one new return code from "Multi-Protocol Label Switching (MPLS) Label Switched Paths (LSPs) Ping Parameters" registry, the "Return Codes" registry:

Value	Meaning
TBD	Response Packet length was exceeded unexpected by the Relay Node Address Stack TLV unexpected

The value should be assigned from the "Standards Action" range (0-191), and using the lowest free value within this range.

9. Acknowledgement

The authors would like to thank Carlos Pignataro, Xinwen Jiao, Manuel Paul, Loa Andersson, Wim Henderickx, Mach Chen, Thomas Morin, Gregory Mirsky, Nobo Akiya and Joel M. Halpern for their valuable comments and suggestions.

10. Contributors

Ryan Zheng
JSPTPD
371, Zhongshan South Road
Nanjing, 210006, China
Email: ryan.zhi.zheng@gmail.com

11. References

11.1. Normative References

[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.

11.2. Informative References

[I-D.ietf-mpls-seamless-mpls]
Leymann, N., Decraene, B., Filsfils, C., Konstantynowicz, M., and D. Steinberg, "Seamless MPLS Architecture", draft-ietf-mpls-seamless-mpls-07 (work in progress), June 2014.

Authors' Addresses

Jian Luo (editor)
ZTE
50, Ruanjian Avenue
Nanjing, 210012, China

Email: luo.jian@zte.com.cn

Lizhong Jin (editor)
Shanghai, China

Email: lizho.jin@gmail.com

Thomas Nadeau (editor)
Lucidvision

Email: tnadeau@lucidvision.com

George Swallow (editor)
Cisco
300 Beaver Brook Road
Boxborough , MASSACHUSETTS 01719, USA

Email: swallow@cisco.com