Network Working Group
Internet-Draft
Updates: 4379 (if approved)
Intended status: Standards Track
Expires: August 5, 2013

R. Zheng, Ed.
L. Jin, Ed.
ZTE
T. Nadeau, Ed.
Juniper Networks
G. Swallow, Ed.
Cisco

February 1, 2013

# Relayed Echo Reply mechanism for LSP Ping
# draft-zjns-mpls-lsp-ping-relay-reply-01

Abstract

In some inter-AS and inter-area deployment scenarios for 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 LSR to have the capability to relay the echo response by a set of routable intermediate nodes to the initiator during the traceroute process in inter-AS and inter-area scenarios.

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 August 5, 2013.

Table of Contents

## 1. Introduction

This document describes extensions to the LSP Ping and Traceroute as specified in [RFC4379] that add as a Relayed Echo Reply mechanism that can be used to detect data plane failures in inter-AS and inter-area MPLS LSPs. Prior to this extension, inter-AS functionality of [RFC4379] would fail in most deployment scenarios. A new message referred to as "Relayed Echo Reply message" and a new TLV referred to as "Relay Node Address Stack TLV" are defined in this draft to overcome these deficiencies.

### 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 data plane failures and localize faults. In the traceroute mode of LSP Ping procedure, the Echo Request message is send along the data plane between the originating LSR and one of the LSRs along the LSP, but is directed to be punted to the the control plane of each transit LSR. The control plane of the receiving LSR then responds directly to the originator using an Echo Reply massage with proper information are required to send to the initiator at each transit LSR. Each hop along the LSP is progressively probed by increasing the TTL of the Echo Request Message until the terminus of the LSP is reached. Using this mechanism, the LSP data plane is tested, and any resulting faults can be localized. Furthermore, this mechanism allows a network operator to create an accurate view of deployed LSP topology.

The original mechanism specifies that The Echo Reply be sent back to the initiator usig a UDP packet containing directed back to the IPv4/ IPv6 address of the originating LSR. This works in adminitrative domains allowing IP address reachability and routing back to the originating LSR. However, in practice, this is often not the case due to intra-provider routing policy, route hiding, network address translation at boundary autonomous system border routers (i.e.: ASBR), etc... In fact, it is almost uniformly the case that in inter-AS scenarios to not allow the distribution or direct routing to the IP addresses of any of the nodes other than the ASBR.

Figure 1 demonstrates how initiating a traceroute procedure on an ingress LSR (i.e.: PE1) of an LSP from PE1 to PE2, can be constructed between P nodes within an AS, which are then connected to ASBRs

interconnect both ASs. In this case, if private addresses were in use within AS2, a traceroute from PE1 directed to PE2 could fail if the fault exists somewhere between AS2 and PE2 because P2 cannot forward packets back to PE1 given that it is a private address within AS1. In this case, PE1 would detect a path break, as the Echo Request messages would not be responded to; however, localization of the actual fault would not be possible.

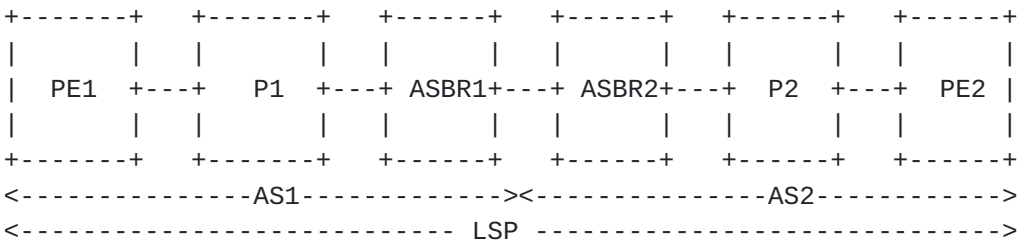



Figure 1: Simple Inter-AS LSP Configuration

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

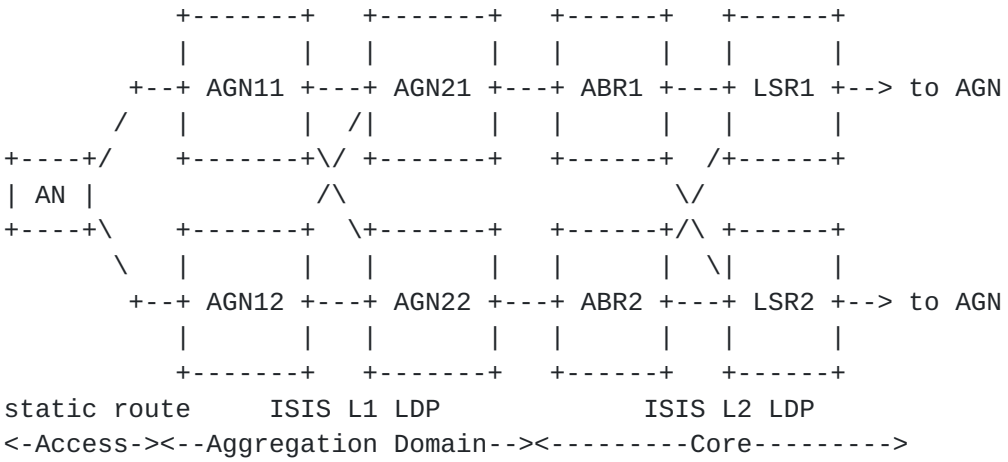



Figure 2: Seamless MPLS Architecture

This remainder of this document describes extensions to the LSP Ping mechanism to facillitate a response from the replying LSR using a simple mechanism that uses the ASBRs to relay the message back to the initiator. This approach will work because every subsequent AS can and must have a route back to a connected AS. Using a recursive approach, intermediate ASs can relay the message toward each other until the final AS is reached. At this point, the ASBR must have a route to the initiating LSR because it is directly attached to it. This is achieved by augmenting the replying LSR's LSP Ping algorithm to send a response to a relay node (as indicated by the Relay Node Address Stack TLV), and the response would be relayed to the next relay node (i.e.: ASBR), until it reaches the ultimate ASBR. At that point the ASBR should be able to resolve a local route to the initiator.

## 3. Extensions

RFC4379 describes the basic MPLS LSP Ping mechanism, which defines two message types. This draft 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 draft, 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 was exceeded by the Relay Node Address Stack TLV unexpected.

It should be noted that this document focuses only on detecting the LSP which is setup using a uniform type of IP address. That is, all hops between the originator and terminus use one address type of address) to address their control planes. This does not preclude nodes that support both IPv6 and IPv4 addresses simultaneously, but the entire path MUST be addressible using only one address family type. Support 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 TCP and UDP port number 3503 has been allocated in [RFC4379] 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 draft is required. Figure 3 illustrates the TLV format.

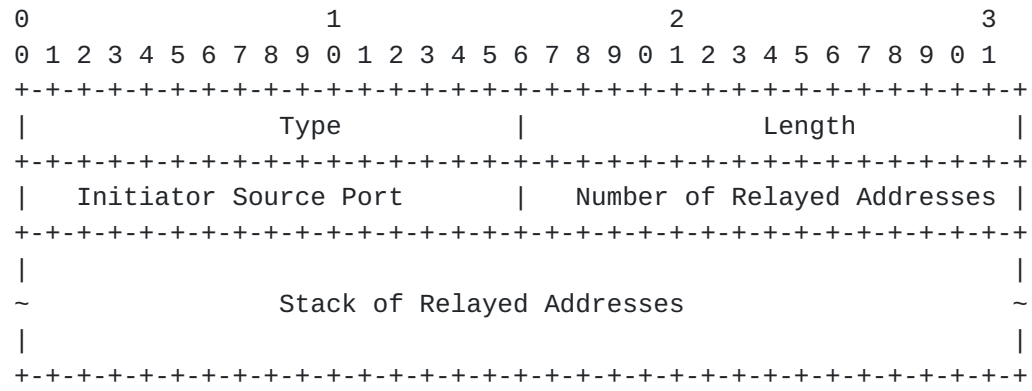



Figure 3: Relay Node Address Stack TLV

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

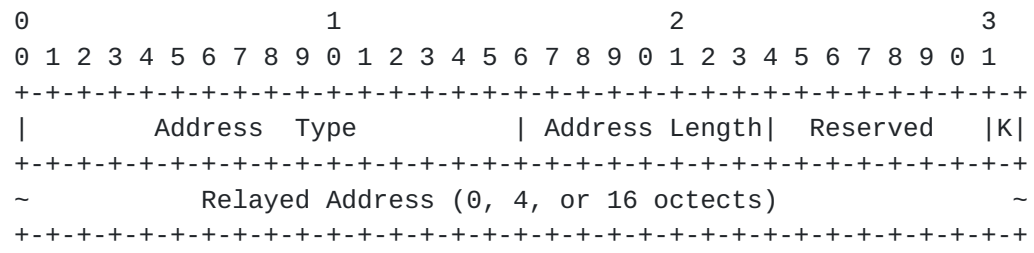
- Length: The Length of the Value field in octets.

- Initiator Source Port: The port that the initiator sends the Echo Request message, and also the port that 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:




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

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

K bit:

If the K bit is set to 1, then this sub-TLV SHOULD be kept in Relay Node Address Stack, SHOULD not be deleted in compress process of section 4.2. The K bit may be set by ASBRs which address would be kept in the stack if necessary.

If the K bit is set to 0, then this sub-TLV SHOULD be processed normally according to section 4.2.

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 was exceeded by the Relay Node Address Stack TLV unexpected.

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 for facillitate the relay functionality.

When the Echo Request is first sent by initiator supporting these extensions, a Relay Node Address Stack TLV with the initiator address in the stack and its source port MUST be included.

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 would check the addresses of the stack in sequence from top to bottom, i.e., the first address in the stack would be first one to be checked, to find out the first public routable IP address. Those address entries behind of the first routable IP address in the address list with K bit set to 0 would be deleted, and the address entry of the replying LSR would be added at the bottom of the stack. Those address entries with K bit set to 1 would be kept in the stack. The updated Relay Node Address Stack TLV would be carried in the response message.

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

If the packet length was exceeded by the Relay Node Address Stack TLV unexpectedly, the TLV SHOULD be returned back unchanged in the echo response message. And the new return code would help to notify the initiator of the situation.

If the first routable IP address is the first address in the stack, the replying LSR would respond an Echo Reply message to the initiator.

If the first routable IP address is of an intermediate node, other

than the first address in the stack, the replying LSR would send an Relayed Echo Reply instead of an Echo Reply in response.

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

### 4.3. Sending an Relayed Echo Reply

The Relayed Echo Reply is sent in two cases:

1. When the replying LSR received an Echo Request with the initiator IP address in the Relay Node Address Stack TLV is IP unroutable, the replying LSR would send an Relayed Echo Reply message to the first routable intermediate 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 of the message part. The destination IP address of the Relayed Echo Reply is set to the first routable IP address from the Relay Node Address Stack TLV, and the destination UDP port is set to 3503.

2. When the intermediate relay node received an Relayed Echo Reply with the initiator IP address in the Relay Node Address Stack TLV is IP unroutable, the intermediate relay node would send the Relayed Echo Reply to the next relay node with the content of the UDP packet unchanged. The destination IP address of the Relayed Echo Reply is set to the first routable IP address from the Relay Node Address Stack TLV. Both the source and destination UDP port should be 3503.

### 4.4. Receiving an Relayed Echo Reply

Upon receiving an Relayed Echo Reply message with its address as the destination address in the IP header, the relay node should check the address items in Relay Node Address Stack TLV in sequence and find the first routable node address.

If the first routable address is the top one of the address list, i.e., the initiator address, the relay node should send an Echo Reply message to the initiator containing the same payload with the Relayed Echo Reply message received.

If the first routable address is not the top one of the address list, i.e., another intermediate relay node, the relay node should send an Relayed Echo Reply message to this relay node with the payload unchanged.

### 4.5. Sending an Echo Reply

The Echo Reply is sent in two cases:

1. When the replying LSR received an Echo Request with the initiator IP address in the Relay Node Address Stack TLV is IP routable, 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 Relay Node Address Stack TLV would be carried in the Echo Reply.

2. When the intermediate relay node LSR received an Relayed Echo Reply with the initiator IP address in the Relay Node Address Stack TLV is IP routable, the intermediate relay node would send the Echo Reply to the initiator with the payload no changes other than the Message Type field. The destination IP address of the Echo Reply is set to the initiator IP address, 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.

### 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.

## 5. LSP Ping Relayed Echo Reply Example

Considering the inter-AS scenario in Figure 4 below.

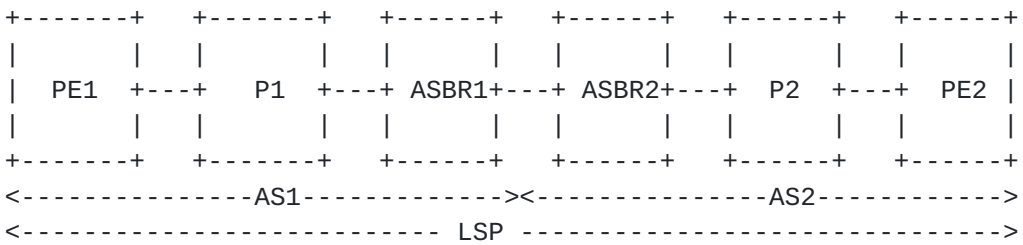



Figure 4: Example Inter-AS LSP

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 outter-most label TTL=1, contains the Relay Node Address

Stack TLV with the only address of PE1.

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 address is routable. Then deletes P1 address, and adds its own address following PE1 address. As a result, there would be PE1 address followed by ASBR1 address in the Relay Node Address Stack TLV of the Echo Reply sent by ASBR1.

PE1 then sends an Echo Request with outter-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 address is IP route unreachable, and ASBR1 address is the first routable one in the Relay Node Address Stack TLV. ASBR2 adds its address as the last address item following ASBR1 address in Relay Node Address Stack TLV, sets ASBR1 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 address is first routable one in the address list. Then ASBR1 send an Echo Reply to PE1 with the payload of received Relayed Echo Reply no changes other than the Message Type field.

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

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

For the Echo Request with outter-most label TTL=5, the echo response 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 reachable route to the initiator is finally transmitted to the initiator by multiple relay nodes.

## 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 they may replace the failed node that originated the Echo Reply with their own address.

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 draft, 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

New Message Type:

| Value | Meaning |
|---|---|
| TBD | MPLS Relayed Echo Reply |

### 8.2. New TLV

New TLV: Routable Relay Node Address TLV

| Type | Meaning |
|---|---|
| TBD | Relay Node Address Stack TLV |

A suggested value is assigned from 32768-49161 as suggested by RFC4379 Section 3.

### 8.3. New Return Code

New Return Code:

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

## 9. Acknowledgement

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

## 10. References

### 10.1. Normative References

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

[RFC4377] Nadeau, T., Morrow, M., Swallow, G., Allan, D., and S. Matsushima, "Operations and Management (OAM) Requirements for Multi-Protocol Label Switched (MPLS) Networks", RFC 4377, February 2006.

[RFC4378] Allan, D. and T. Nadeau, "A Framework for Multi-Protocol Label Switching (MPLS) Operations and Management (OAM)",

RFC 4378, February 2006.

[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.

### 10.2. Informative References

[ietf-mpls-seamless] Leymann, N., Decraene, B., Filsfils, C., Konstantynowicz, M. and D. Steinberg, "Seamless MPLS Architecture", draft-ietf-mpls-seamless-mpls-02 , October 2012.

## Authors' Addresses

Ryan Zheng (editor)
ZTE
50, Ruanjian Avenue
Nanjing, 210012, China

Email: zheng.zhi@zte.com.cn

Lizhong Jin (editor)
ZTE
889, Bibo Road
Shanghai, 201203, China

Email: lizho.jin@gmail.com

Thomas Nadeau (editor)
Juniper Networks
Westford, MA

Email: tnadeau@juniper.net

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

Email: swallow@cisco.com