Network Working Group
Expires: May 2007

Thomas D. Nadeau
George Swallow
Cisco Systems, Inc.

October 2006

Detecting MPLS Data Plane Failures in Inter-AS and inter-provider Scenarios

draft-nadeau-mpls-interas-lspping-02.txt

Status of this Memo

Abstract

This document describes a simple and efficient mechanism that can be used to detect data plane failures in Multi-Protocol Label Switching Label Switched Paths that extend beyond a single Autonomous System and/or across multiple Service Provider network boundaries. This document describes extensions to the existing MPLS LSP Ping protocol to achieve these goals.

Table of Contents

1. Introduction

This document describes a simple and efficient mechanism that can be used to detect data plane failures in MPLS LSPs that span across multiple Autonomous System (AS) and service provider boundaries. At present, the existing MPLS LSP Ping protocol cannot handle all but one of these cases. This document first explains the scenarios where the existing protocol is inadequate, then describes information carried in extended MPLS "echo request" and "echo reply" messages; and finally describes enhanced mechanisms for transporting the echo reply, as well as processing it at intermediate points (both in an out of the originating AS).

An important consideration in this design is that MPLS echo requests follow the same data path that normal MPLS packets would traverse. MPLS echo requests are meant primarily to validate the data plane, and secondarily to verify the data plane against the control plane. Mechanisms to check the control plane are valuable, but are not covered in this document.

As is described in [RFC4379], to avoid potential Denial of Service attacks, it is recommended to regulate the LSP ping traffic going to the control plane. A rate limiter should be applied to the well-known UDP port defined below. Furthermore, due to the

fact that there are data exchanges between provider networks

which may wish to hide the details of their network, it is recommended that the inter-AS border routers provide operators with control over what information (i.e.: addresses) in these messages.

2. Terminology

2.1 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].

2.2 Terminology

Definitions of key terms for MPLS OAM are found in [RFC4378] and the reader is assumed to be familiar with those definitions which are not repeated here.

The following additional terms are useful to understand this document.

2.3 Acronyms

The following list of acronyms is a repeat of common acronyms defined in many other documents, and is provided here for convenience.

```
  CE: Customer Edge
  PE: Provider Edge
ASBR: Autonomous System Border Router
 DoS: Denial of service
ECMP: Equal Cost Multipath
 LDP: Label Distribution Protocol
 LSP: Label Switch Path
 LSR: Label Switch Router
 OAM: Operations and Management
OA&M: Operations, Administration and Maintenance.
RSVP: Resource reSerVation Protocol
  SP: Service Provider
```

3. Structure of This Document

The body of this memo contains four main parts: motivation, extensions to the MPLS echo request/reply packet format, inter-AS LSP ping operation, and a reliable return path. It is suggested that first-time readers skip the actual packet formats and read the Theory of Operation first; the document is structured the way it is to avoid forward references.

4. Motivation

The requirements specified in [RFC4377] stipulate that data plane OAM functions must be provided as solutions for service providers. These data plane test functions must not only function within an autonomous system (AS), but must also function across ASs. Furthermore, these tests must function correctly across ASs that span multiple Service Provider(SP) domains. At present, the data plane liveliness tools function in these capacities only in the narrow (and rarely used) case where the IP addresses of LSRs involved are known to each other. For example, when the IP addresses from one AS are exchanged through routing with other attached ASs. Another case includes the Layer-3 VPN inter-provider interconnection where the PE addresses are distributed between service providers. However, these cases are uncommon, and thus the existing LSP Ping [RFC4379] tool is unable to respond under most error condition configurations. For example consider the following configuration. Imagine that PE1 and PE2 are in two different provider domains. In this case, it is commonly desirable for providers to NOT distribute the IP addresses of any of the intermediate P routers between PE1 and PE2.

```
{--- AS1 ---}        {--- AS2 ---}
PE1--P-P--ASBR1----ASBR2--P-P--PE2
```

Now, imagine that the LSP that connects PE1 to PE2 contains a fault somewhere bewteen ASBR2 and PE2 as is indicated by 'X' between the two P routers:

```
{--- AS1 ---}        {--- AS2 ---}
PE1--P-P--ASBR1----ASBR2--P-X-P--PE2
```

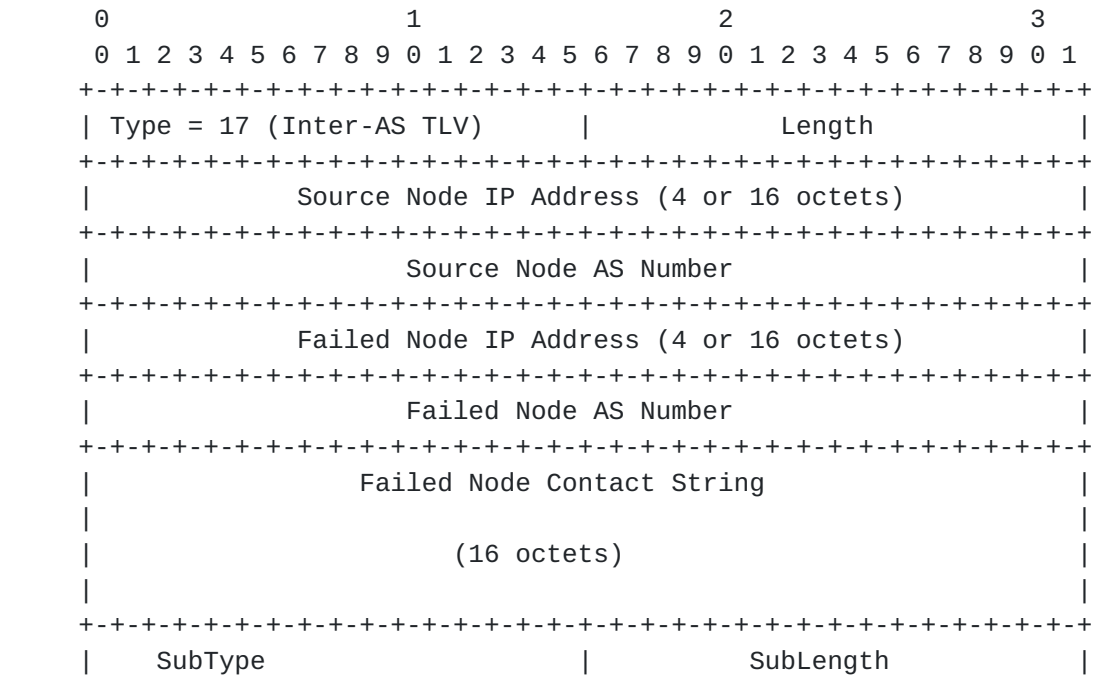
If an LSP Ping is initiated at PE1 with a destination of PE2 and a source of PE1, the packet is label switched correctly until it reaches the first P router within AS2. Here lets imagine that MPLS forwarding is disabled on the link between the two P routers. Upon discovering this while attempting to process the LSP Ping Request packet, the first P router will attempt to reply directly to PE1 with the appropriate error code 5. However, because the address of PE1 is actually private to AS1 by virtue of not being distributed by ASBR1 into AS2, the P router cannot correctly forward the reply to PE1. In this case,

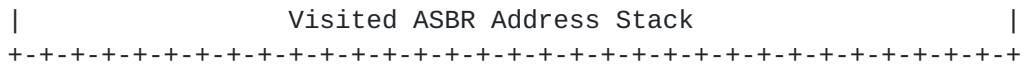
PE1 may surmise that some failure has occurred, but it cannot determine what the error is or where it exists. This clearly does not meet the requirements stipulted in [RFC4377]. This draft describes extensions to [RFC4379] that overcome the aforementioned limitations, and thus allow for the handling of inter-AS/provider cases.

5. Inter-AS TLVs

5.1. Inter-AS TLV

The Inter-AS TLV Reply Object is an optional TLV that is used to collect and report the ASBRs along the path of the LSP under test. Only one such object may appear in a Reply message. The purpose of this object is to allow the upstream router to relay a Reply message from ASBR to ASBR when a failure is detected. A router will use this TLV to look up the last ASBR as indicated as the top-most address on the address stack, that forwarded the Request message into its AS, and then forward the Reply to that router after popping the address from the stack. The Reply message will ultimately be relayed to the original soure of the request. This message has one format that contains the true source and destination addresses of the Request message, as well as a stack of ASBR addresses that were visited while forwarding this message. Type 17 is defined for this TLV (to be assigned by IANA).





SubType:

Sub-Type #	Length	Value Field
1	6	IPv4 Return Stack
2	6	IPv4 Trace Stack
3	6	Ipv6 Return Stack
4	6	IPv6 Trace Stack

Note that only combinations of 1+2 or 3+4 may be used.

Failed Node AS Number:

This field may contain the AS number in which the node where the failure was detected resides. If no AS number is indicated, this field MUST contain 0s.

Failed Node IP Address:

If the interface to the downstream LSR is numbered, then the Address Type MUST be set to IPv4 or IPv6, the Downstream IP Address MUST be set to either the downstream LSR's Router ID or the interface address of the downstream LSR, and the Downstream Interface Address MUST be set to the downstream LSR's interface address.

If the interface to the downstream LSR is unnumbered, the Address Type MUST be Unnumbered, the Downstream IP Address MUST be the downstream LSR's Router ID (4 octets), and the Downstream Interface Address MUST be set to the index assigned by the upstream LSR to the interface.

Failed Node AS Number:

This field may contain the AS number in which the node where the failure was detected resides. If no AS number is indicated, this field MUST contain 0s.

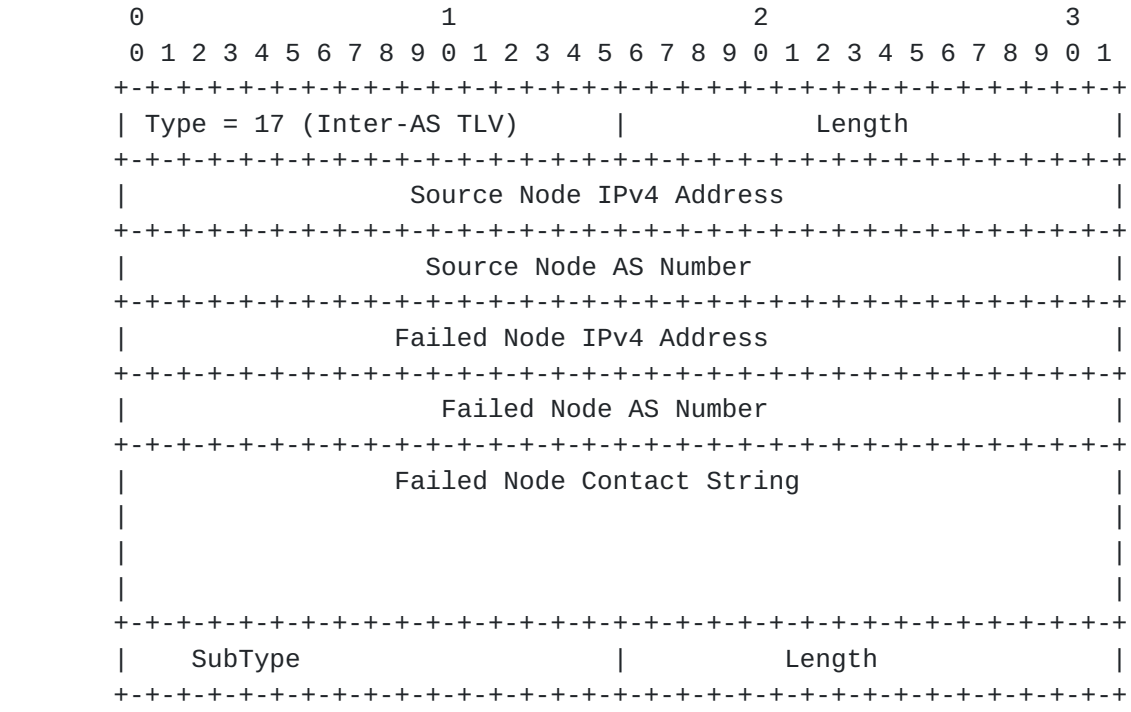
Failed Node Contact String:

This field may contains a string of ASCII characters inserted by the node where the failure was detected or by its closest ASBR. This field MUST indicate contact information such as a provider's international phone number and other relevant contact information in cases where local policy dictates that a provider will not

fill in the Failed Node AS number and/or the Failed Node Address. In all other cases, this field MUST contain 0s.

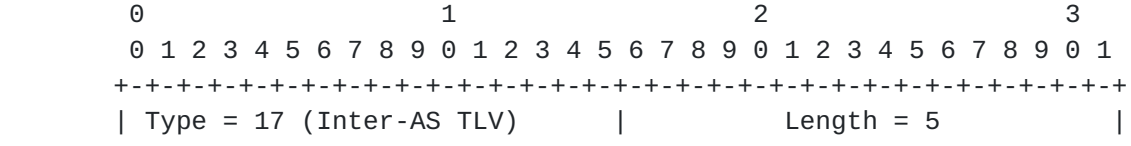
7.1.1 IPv4 Inter-AS TLV

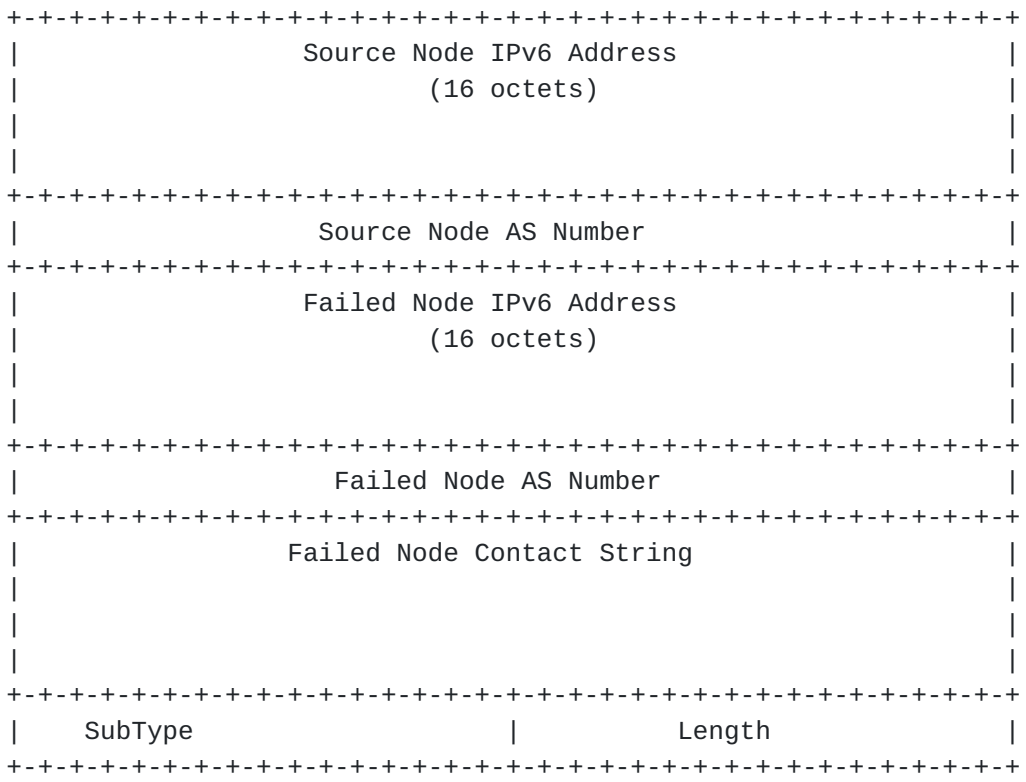
The value consists of four octets of an IPv4 prefix followed by one octet of prefix length in bits; the format is given below. The IPv4 prefix is in network byte order; if the prefix is shorter than 32 bits, trailing bits SHOULD be set to zero. .



7.1.1 IPv6 Inter-AS TLV

The value consists of 16 octets of an IPv6 prefix followed by one octet of prefix length in bits; the format is given below. The IPv6 prefix is in network byte order; if the prefix is shorter than 128 bits, trailing bits SHOULD be set to zero. This FEC is used if the protocol advertising the label is unknown, or may change during the course of the LSP. An example is ani nter-AS LSP that may be sig- naled by LDP in one AS, by RSVP-TE in another AS, and by BGP between the ASs, such as is common for inter-AS VPNs.





7.1.3 Visited ASBR Address Stack

The term Visited ASBR Address Stack applies to two stacks of IP addresses of the ASBRs along the path of an LSP called the Trace and Return Stacks. The two stacks have the same format; however they have slightly different semantics. Both stack objects are stacks of addresses that denote the list of visited ASBRs. They contain stack of a single field containing either an IPv4 address if the TLV SubType field is set to 1, or an IPv6 address as indicated by the TLV SubType field being set to 3.

The Return Stack is to be used in a destructive manner as a means of unwinding the path of ASBRs that were used to originally forward the Request. Each subsequent ASBR along the path that receives the reply should destructively remove itself from the stack.

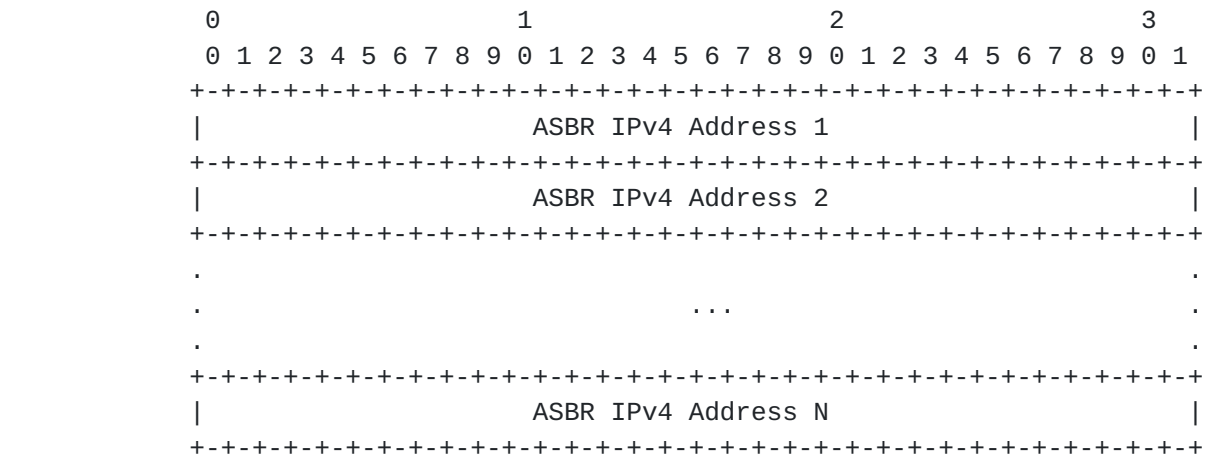
On the other hand, the Trace Stack MUST only be added to (i.e.: ASBR addresses pushed) and items never removed from this stack. This will allow the source to see the trace of the path of ASBRs once the Reply message is returned. In cases where policy dictates that ASBR addresses must be hidden, a value of all 0s MUST be

inserted into the stack, or the stack completely removed prior to

forwarding the Reply. It is prefered that a blank entry be left, as this will at least indicate that there was one hop without revealing its IP address.

IPv4 Trace and Visited Stack Objects

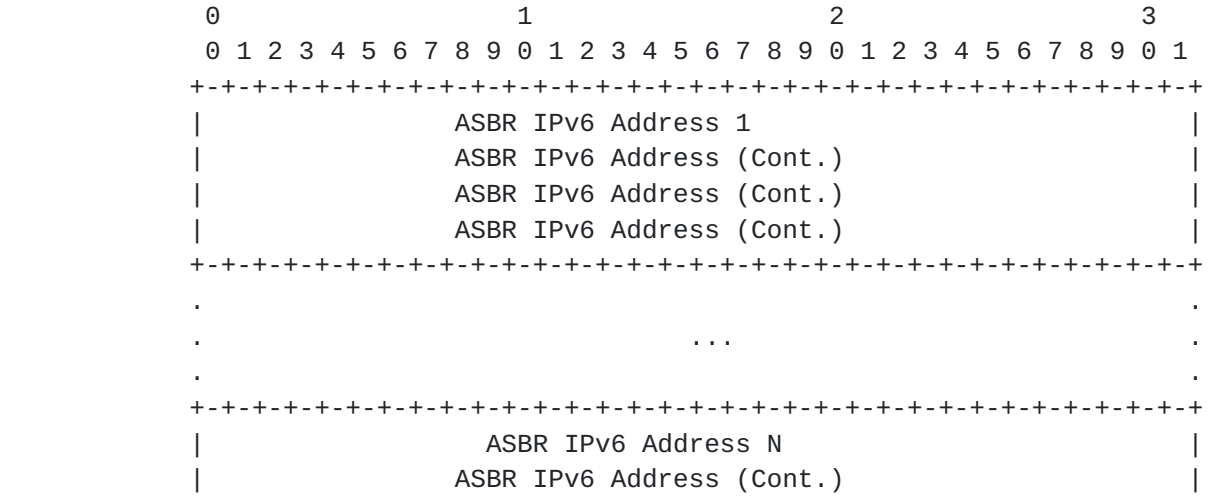
The Length is 4*N octets, N is the number of visited ASBRs. This object has the following format:



ASBR IPv4 Address 1, ASBR IPv4 Address 1, ... contain a valid IPv4 address.

IPv6 Trace and Visited Stack Objects

The Length is 16*N octets, N is the number of visited ASBRs. This object has the following format:



```
|               ASBR IPv6 Address (Cont.)                       |
|               ASBR IPv6 Address (Cont.)                       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

ASBR IPv6 Address 1, ASBR IPv6 Address 1, ... contain a valid IPv4 address.
IPv6

6. Error Code(s)

TBD

7. Theory of Operation

When tracing an LSP which spans multiple AS, an Inter-AS Reply Object is included in the Echo Request. Initially the object contains only the address of the source PE and a Trace stack with that same address. As the tracing progress each ASBR copies the trace stack as a reply stack, it then pushes its address to the trace stack. It includes both stacks in an Inter-AS Reply object and sends it in an Echo Reply message to the top address in the reply stack. The receiver of the Reply message then verifies that it is included in the reply stack. It then pops its address from the reply stack and re-addresses the Echo Reply message to the (new) top element of the reply stack. This is repeated until the source PE receives the Echo Reply.

7.1 Adjustments to Outgoing Labels

When an LSP request is sent from an originator, some adjustments may need to be made to outgoing labels:

Inter-AS cases:

A) VRF to VRF

The LSP terminates at the ASBR. These procedures do not apply.

B) EBGP redistribution of labeled VPN-IPv4 routes from AS to neighboring AS.

Tracing is performed by incrementing the VPN label begining at one. If TTL hiding is in effect, then tracing of PSN label is not necessary for these procedures.

C) Carrier's Carrier (CsC):

1) TTL Hiding
   a. Will work as is.
   b. Verification of the core must be done separately by core owners.
   c. Traceroute can trace both stubs of the 'carried' carrier.

2) No TTL hiding
   a. Set VPN TTL to 1.
   b. CsC CE or Ps would return to the CsC PE who would relay messages back to originator.
   c. For traceroute, set VPN TTL=1, and progressively increase the IGP TTL by 1 to probe.

7.2 Receiving Echo Replies

The existing packet processing algorithm as specified in [RFC4379] is enhanced as follows to support inter-AS/provider LSP ping/trace.

When an Echo Reply message is received:

1) If the packet is addressed to this router
   (i.e.: destination address == this router's router ID):

   a. If the original sender field TLV == this router's address, process normally. // today's functionality for a normal reply received by the src.

   b. Else this packet has been delivered to this router because it is an ASBR and needs to proxy for a P router in its AS to return the reply.

      If the inter-AS TLV is present,

      i. If the last visited AS is empty, set it to the ASBR's primary AS#.

      ii. If the stack is empty, this is an error case. The TLV SHOULD NOT be present if the stack is empty.

      iii. Else if the top-most address in the stack is this router's address.

         1. Pop it from the stack.
         2. Replace the packet's destination address with the next address in the stack.
         3. Replace the packet's src address with this ASBR's address.

4. Optionally, the ASBR may hide (i.e.: remove) information that its local policy has been configured for.
5. Look up the route/next-hop for this address and deliver the packet. The ASBR should be able to resolve the address because at this point unless there has been an error in the return path forwarding, then the packet should be at the border of the originating AS. If the look-up fails, drop the packet and notify the operator of this router that an error condition has occurred.

When an LSP ping request is received:

1) If this router is an ASBR
   a. Write the next entry in the Last Seen ASBR stack's address as the destination address of the packet and forward it to that address.
   b. Otherwise process normally as specified in the LSP ping draft.

8. Security Considerations

In addition to the Security Considerations from [RFC4379], here are at least two approaches to attacking LSRs using the mechanisms defined here.

One is a Denial of Service attack, by sending MPLS echo requests/replies to LSRs and thereby increasing their workload. The other is obfuscating the state of the MPLS data plane liveness by spoofing, hijacking, replaying or otherwise tampering with MPLS echo requests and replies.

Authentication will help reduce the number of seemingly valid MPLS echo requests, and thus cut down the Denial of Service attacks; beyond that, each LSR must protect itself.

Authentication sufficiently addresses spoofing, replay and most tampering attacks; one hopes to use some mechanism devised or suggested by the RPSec WG. It is not clear how to prevent hijacking (non-delivery) of echo requests or replies; however, if these messages are indeed hijacked, LSP ping will report that the data plane isn't working as it should.

It doesn't seem vital (at this point) to secure the data carried in MPLS echo requests and replies, although knowledge of the state of the MPLS data plane may be considered confidential by some.

9. IANA Considerations

[need to request some new Message Types, TLV Types, Return Codes]

9.1. Message Types, Reply Modes, Return Codes

It is requested that IANA maintain registries for Message Types, Reply Modes, Return Codes and Return Subcodes. Each of these can take values in the range 0-255. Assignments in the range 0-191 are via Standards Action; assignments in the range 192-251 are made via Expert Review; values in the range 252-255 are for Vendor Private Use, and MUST NOT be allocated.

If any of these fields fall in the Vendor Private range, a top-level Vendor Enterprise Code TLV MUST be present in the message.

9.2. TLVs

It is requested that IANA maintain registries for the Type field of top-level TLVs as well as for sub-TLVs. The valid range for each of these is 0-65535. Assignments in the range 0-16383 and 32768-49161 are made via Standards Action as defined in [RFC2434]; assignments in the range 16384-31743 and 49162-64511 are made via Expert Review (see below); values in the range 31744-32746 and 64512-65535 are for Vendor Private Use, and MUST NOT be allocated.

If a TLV or sub-TLV has a Type that falls in the range for Vendor Private Use, the Length MUST be at least 4, and the first four octets MUST be that vendor's SMI Enterprise Code, in network octet order. The rest of the Value field is private to the vendor.

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., Matshushima, S., "Operations and Management (OAM) Requirements for Multi-Protocol Label Switched (MPLS) Networks", RFC 4377, February 2006.

[RFC4378] Allan, D., Nadeau, T., "A Framework for Multi-Protocol Label Switching (MPLS) Operations and Management", RFC 4378, February 2006.

[RFC4379] Kompella, k., Swallow, G., "Detecting MPLS Data Plane

Liveness", RFC 4379, February 2006.

10.2 Informative References

[RFC2434] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 2434, October 1998.

11. Acknowledgment

The authors wish to acknowledge and thank the following individuals for their valuable comments to this document: Azhar Sayeed, Vanson Lim, and Mike Piecuch.

12. Authors' Addresses

Thomas D. Nadeau
Cisco Systems, Inc.
1414 Massachusetts Ave,
Boxboro, MA 01719
Phone: +1.978.936.1470
Email: tnadeau@cisco.com

George Swallow
Cisco Systems
1414 Massachusetts Ave,
Boxborough, MA 01719
Phone: +1 978 936 1398
Email: swallow@cisco.com

13. Intellectual Property Statement

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement this standard. Please address the information to the IETF at ietf-ipr@ietf.org.

14. Full Copyright Statement