Network Work group
Internet-Draft
Intended status: Standards Track
Expires: July 6, 2014

N. Kumar
G. Swallow
C. Pignataro
N. Akiya
Cisco Systems, Inc.
S. Kini
Ericsson
H. Gredler
Juniper Networks
M. Chen
Huawei

January 02, 2014

Label Switched Path (LSP) Ping/Trace for Segment Routing Networks Using MPLS Dataplane

draft-kumarkini-mpls-spring-lsp-ping-00

Abstract

Segment Routing architecture leverages the source routing and tunneling paradigm and can be directly applied to MPLS dataplane. A node steers a packet through a controlled set of instructions called segments, by prepending the packet with Segment Routing header.

The segment assignment and forwarding semantic nature of Segment Routing raises additional consideration for connectivity verification and fault isolation in LSP with Segment Routing architecture. This document illustrates the problem and describe a mechanism to perform LSP Ping and Traceroute on Segment Routing network over MPLS dataplane.

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 6, 2014.

Table of Contents

1. Introduction

[I-D.filsfils-rtgwg-segment-routing] introduces and explains Segment Routing architecture that leverages the source routing and tunneling paradigm. A node steers a packet through a controlled set of

instructions called segments, by prepending the packet with SR header. A detailed definition about Segment Routing architecture is available in draft-filsfils-rtgwg-segment-routing and different use-cases are discussed in draft-filsfils-rtgwg-segment-routing-use-cases.

The Segment Routing architecture can be directly applied to MPLS dataplane in a way that, the segment will be of 20-bits size and SR header is the label stack.

Multi Protocol Label Switching (MPLS) has defined in [RFC4379] a simple and efficient mechanism to detect data plane failures in Label Switched Paths (LSP) by specifying information to be carried in an MPLS "echo request" and "echo reply" for the purposes of fault detection and isolation, and mechanisms for reliably sending the echo reply. The functionality is modeled after the ping/traceroute paradigm (ICMP echo request [RFC0792]) and is typically referred to as MPLS-ping and MPLS-traceroute.

Unlike LDP or RSVP which are the other well-known MPLS control plane protocols, segment assignment in Segment Routing architecture is not hop-by-hop basis.

This nature of Segment Routing raises additional consideration for connectivity verification and fault isolation in Segment Routing network. This document illustrates the problem and describe a mechanism to perform LSP Ping and Traceroute on Segment Routing network over MPLS dataplane.

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. Challenges with Existing mechanism

This document defines Target FEC Stack sub-TLVs and explains how they can be used to tackle below challenges.

3.1. Path validation in Segment Routing networks

[RFC4379] defines the OAM machinery that helps with connectivity check and fault isolation in MPLS dataplane path with the use of various Target FEC Stack Sub-TLV that are carried in MPLS Ping packets and used by the responder for FEC validation. While it is obvious that new Sub-TLVs need to be assigned, the unique nature of

Segment Routing architecture raises a need for additional machinery for path validation. This section discuss the challenges as below:

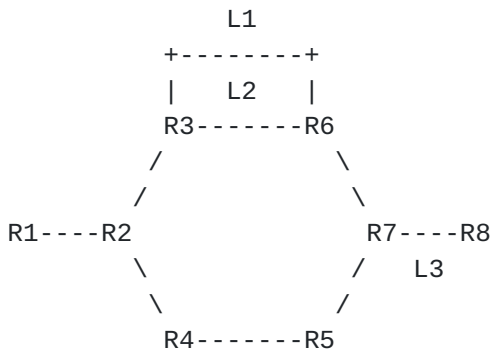


Figure 1: Segment Routing network

500x --> Node Segment ID for Router X
(Ex: 5006 is node segment ID for R6)
9axy --> Adj Segment ID from Router X to Y over link a
(Ex: 9136 is Adj segment ID from R3 to R6 via link 1)

The forwarding semantic of Adjacency segment is to pop the segment and send the packet to a specific neighbor over a specific link. A malfunctioning node may forward packets using Adjacency segment to incorrect neighbor or over incorrect link. Exposed segment (after incorrectly forwarded Adjacency segment) might still allow such packet to traverse to intended destination, yet intended strict traversal has been broken.

Assume in above topology, R1 sends traffic with segment stack as {9124, 5007, 9378} so that the path taken will be R1-R2-R4-R5-R7-R8. If the adjacency segment 9124 is misprogrammed in R2 to send the packet to R1 or R3, it will still be delivered to R8 but is not via the expected path.

MPLS traceroute may help with detecting such deviation in above mentioned scenario. However, in a different example, it may not be helpful if R3, due to misprogramming, forwards packet with adjacency segment 9236 via link L1 while it is expected to be forwarded over Link L2.

3.2. Service Label

One of the proposals for source routed LSPs is to include service labels in the MPLS label stack. These service labels are used to apply a service (as indicated by the service label) to the packet at

the intermediate LSRs along the explicit-route. Since these labels are part of the MPLS label stack these have implications on MPLS OAM. This document describes how the procedures of [RFC4379] can be applied to in the absence of service-labels in Section xx. Additional considerations for service labels are included in Section yy and requires further discussion.

4. Segment Routing Sub-TLV Format

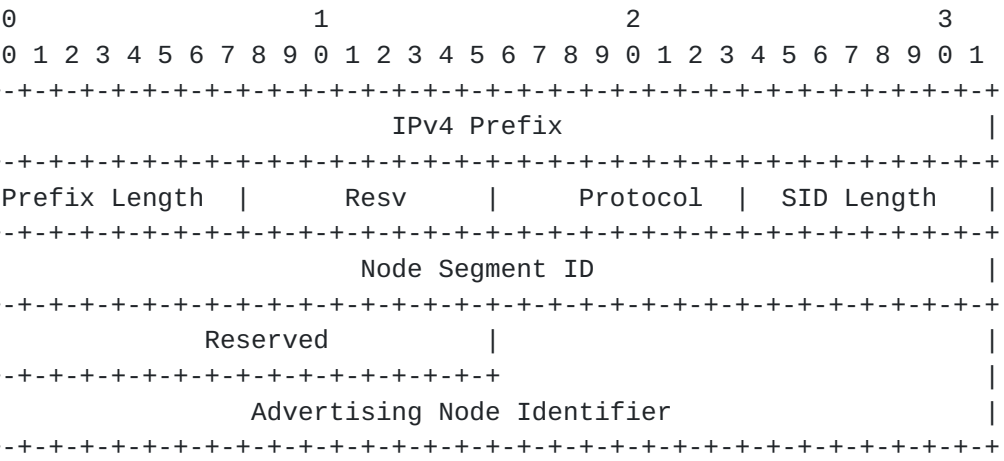
The format of the following FEC Sub-TLVs follows the philosophy of Target FEC Stack TLV carrying FECs corresponding to each label in the label stack. When operated with the procedures defined in [RFC4379], this allows LSP ping/traceroute operations to function when Target FEC Stack TLV contains more FECs than received label stack at responder nodes.

Type	Sub-Type	Value Field
1	TBD1	IPv4 Prefix Node Segment ID
	TBD2	IPv6 Prefix Node Segment ID
	TBD3	Adjacency Segment ID

Service Segments and FRR will be considered in future version.

4.1. IPv4 Prefix Node Segment ID

The format is as below:



IPv4 Prefix

This field carries the IPv4 prefix to which the Node Segment ID is assigned.

Prefix Length

One octet of prefix length in bits.

Protocol

Set to 1 if the IGP protocol is OSPF and 2 if IGP protocol is ISIS.

SID Length

Set to 3 or 4 depending on the Segment ID.

Node Segment ID

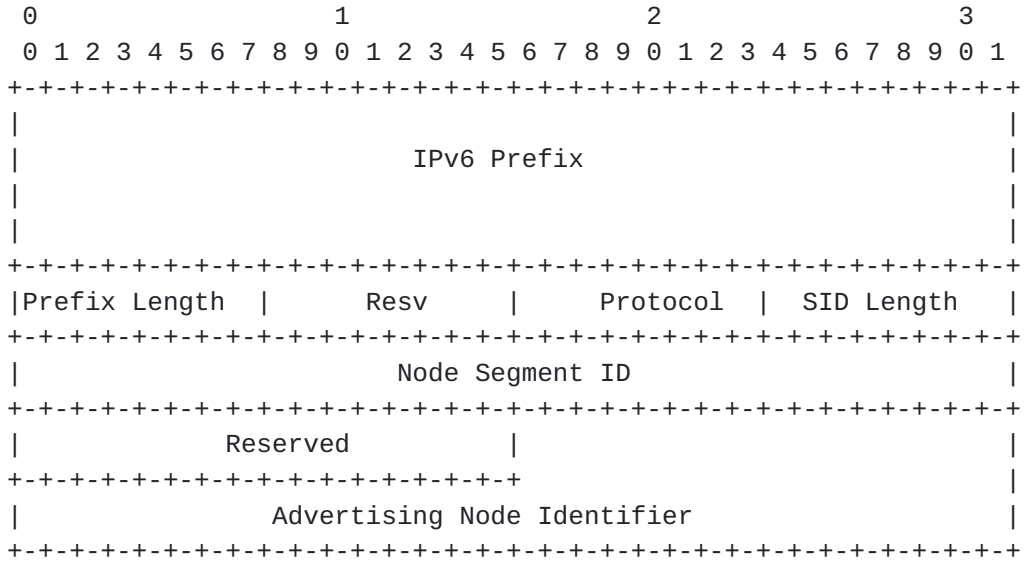
This field carries the Node segment ID. If the SID Length is 3, then the 20 rightmost bits represent the segment. If length is 4, then the value represent a 32 bits Segment ID.

Advertising Node Identifier

Specifies the advertising node identifier. When Protocol is set to 1, then the 32 rightmost bits represent OSPF Router ID and if protocol is set to 2, this field carries 48 bit ISIS System ID.

4.2. IPv6 Prefix Node Segment ID

The format is as below:



IPv6 Prefix

This field carries the IPv6 prefix to which the Node Segment ID is assigned.

Prefix Length

One octet of prefix length in bits.

Protocol

Set to 1 if the IGP protocol is OSPF and 2 if IGP protocol is ISIS.

SID Length

Set to 3 or 4 depending on the Segment ID.

Node Segment ID

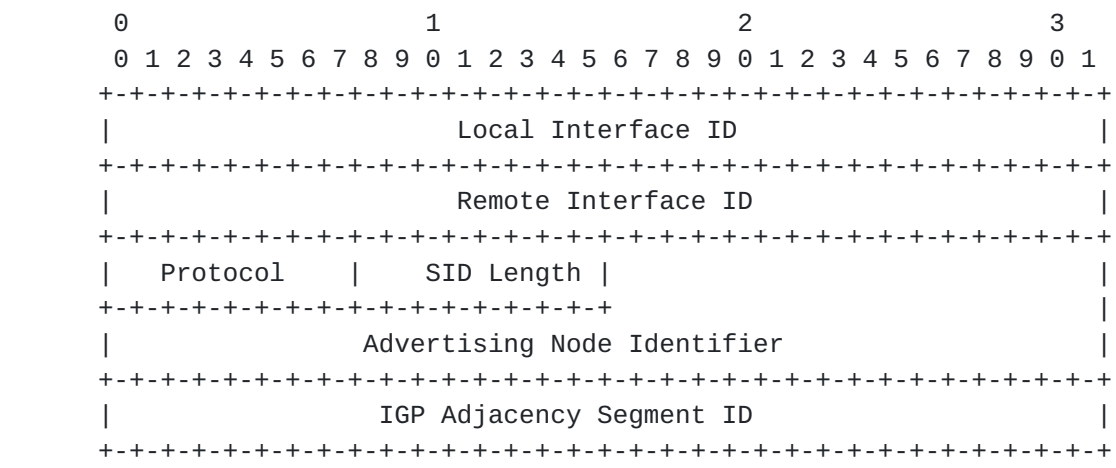
This field carries the Node segment ID. If the SID Length is 3, then the 20 rightmost bits represent the segment. If length is 4, then the value represent a 32 bits Segment ID.

Advertising Node Identifier

Specifies the advertising node identifier. When Protocol is set to 1, then the 32 rightmost bits represent OSPF Router ID and if protocol is set to 2, this field carries 48 bit ISIS System ID.

4.3. IGP Adjacency Segment ID

The format is as below:



Local Interface ID

An identifier that is assigned by local LSR for a link on which Adjacency SID is bound. If the Adj-SID represents parallel adjacencies (Section 3.3.2.1 of [I-D.filsfils-rtgwg-segment-routing]) this field MUST be set to zero.

Remote Interface ID

An identifier that is assigned by remote LSR for a link on which Adjacency SID is bound. If the Adj-SID represents parallel adjacencies (Section 3.3.2.1 of [I-D.filsfils-rtgwg-segment-routing]) this field MUST be set to zero.

Protocol

Set to 1 if the IGP protocol is OSPF and 2 if IGP protocol is ISIS

SID Length

Set to 3 or 4 depending on the Segment ID.

Advertising Node Identifier

Specifies the advertising node identifier. When Protocol is set to 1, then the 32 rightmost bits represent OSPF Router ID and if protocol is set to 2, this field carries 48 bit ISIS System ID.

IGP Adjacency Segment ID

This field carries the adjacency segment ID.

5. Extension to Downstream Mapping TLV

In an echo reply, the Downstream Mapping TLV [RFC4379] is used to report for each interface over which a FEC could be forwarded. For an FEC, there are multiple protocols that may be used to distribute label mapping. The "Protocol" field of the Downstream Mapping TLV is used to return the protocol that is used to distribute a specific a label. The following protocols are defined in section 3.2 of [RFC4379]:

Protocol #	Signaling Protocol
0	Unknown
1	Static
2	BGP
3	LDP
4	RSVP-TE

With segment routing, OSPF or ISIS can be used for label distribution, this document adds two new protocols as follows:

Protocol #	Signaling Protocol
5	OSPF
6	ISIS

6. Procedures

This section describes aspects of LSP ping/traceroute operations that require further considerations beyond [RFC4379].

6.1. FECs in Target FEC Stack TLV

When LSP echo request packets are generated by an initiator, FECs carried in Target FEC Stack TLV may need to or desire to have deviating contents. This document outlines expected Target FEC Stack TLV construction mechanics by initiator for known scenarios.

Ping

Initiator MUST include FEC(s) corresponding to the destination segment.

Initiator MAY include FECs corresponding to some or all of segments imposed in the label stack by the initiator to communicate the segments traversed.

Traceroute

Initiator MUST initially include FECs corresponding to all of segments imposed in the label stack.

When a received echo reply contains FEC Stack Change TLV with one or more of original segment(s) being popped, initiator MAY remove corresponding FEC(s) from Target FEC Stack TLV in the next (TTL+1) traceroute request.

When a received echo reply does not contain FEC Stack Change TLV, initiator MUST NOT attempt to remove FEC(s) from Target FEC Stack TLV in the next (TTL+1) traceroute request. Note that Downstream Label field of DSMAP/DDMAP contains hints on how initiator may be able to update the contents of next Target FEC Stack TLV. However, such hints are ambiguous without full understanding of PHP capabilities.

6.2. FEC Stack Change TLV

The network node which advertised the node segment ID is responsible for generating FEC Stack Change TLV of &pop& operation for node segment ID, regardless of if PHP is enabled or not.

The network node that is immediate downstream of the node which advertised the adjacency segment ID is responsible for generating FEC Stack Change TLV of &pop& operation for adjacency segment ID.

6.3. PHP, Adjacency SID Pop, Implicit NULL

Forwarding behavior of node segment ID PHP is equivalent to usage of implicit Null in MPLS protocols that embraces downstream label allocation scheme. Adjacency segment ID is also similar in a sense that it can be thought as nexthop destined locally allocated segment that has PHP enabled. Procedures described in Section 4.4 of [RFC4379] relies on Stack-D and Stack-R explicitly having Implicit Null value. It may simplify implementations to reuse Implicit Null for node segment ID PHP and adjacency segment ID cases. However, it is technically incorrect for Implicit Null value to externally appear. Therefore, implicit Null MUST NOT be placed in Stack-D and Interface and Label Stack TLV for node segment ID PHP and adjacency segment ID cases.

6.4. Segment Protocol Check

If the Target FEC Sub-TLV at FEC-stack-depth is TBD1 (IPv4 Prefix Node Segment ID), set Best return code to (error code TBD) if any below conditions fail:

* Validate that Advertising Node Identifier of Protocol is a local node.

* Validate that Node Segment ID is advertised for IPv4 Prefix by the Protocol with Advertising Node Identifier.

* Validate that Node Segment ID is advertisement of PHP bit.

If the Target FEC Sub-TLV at FEC-stack-depth is TBD2 (IPv6 Prefix Node Segment ID), set Best return code to (error code TBD) if any below conditions fail:

* Validate that Advertising Node Identifier of Protocol is a local node.

* Validate that Node Segment ID is advertised for IPv6 Prefix by the Protocol with Advertising Node Identifier.

* Validate that Node Segment ID is advertised of PHP bit.

If the Target FEC sub-TLV at FEC-stack-depth is TBD3 (Adjacency Segment ID), set Best return code to (error code TBD) if any below conditions fail:

* Validate that Remote Interface ID matches the local identifier of the interface on which the packet was received.

* Validate that IGP Adjacency SID is advertised by Advertising Node Identifier of Protocol in local IGP database.

6.5. TTL Consideration for traceroute

LSP Traceroute operation can properly traverse every hop of Segment Routing network in Uniform Model described in [RFC3443]. If one or more LSRs employ Short Pipe Model described in [RFC3443], then LSP Traceroute may not be able to properly traverse every hop of Segment Routing network due to absence of TTL copy operation when outer label is popped. In such scenario, following TTL manipulation technique MAY be used.

When tracing a LSP according to the procedures in [RFC4379] the TTL is incremented by one in order to trace the path sequentially along the LSP. However when a source routed LSP has to be traced there are as many TTLs as there are labels in the stack. The LSR that initiates the traceroute SHOULD start by setting the TTL to 1 for the tunnel in the LSP's label stack it wants to start the tracing from, the TTL of all outer labels in the stack to the max value, and the TTL of all the inner labels in the stack to zero. Thus a typical

start to the traceroute would have a TTL of 1 for the outermost label and all the inner labels would have TTL 0. If the FEC Stack TLV is included it should contain only those for the inner stacked tunnels. The lack of an echo response or the Return Code/Subcode should be used to diagnose the tunnel as described in [RFC4379]. When the tracing of a tunnel in the stack is complete, then the next tunnel in the stack should be traced. The end of a tunnel can be detected from the "Return Code" when it indicates that the responding LSR is an egress for the stack at depth 1. Thus the traceroute procedures in [RFC4379] can be recursively applied to traceroute a source routed LSP.

7. Issues with non-forwarding labels

Source stacking can be optionally used to apply services on the packet at a LSR along the path, where a label in the stack is used to trigger service application. A data plane failure detection and isolation mechanism should provide its functionality without applying these services. This is mandatory for services that are stateful, though for stateless services [RFC4379] could be used as-is. It MAY also provide a mechanism to detect and isolate faults within the service function itself.

To prevent services from being applied to an "echo request" packet, the TTL of service labels MUST be 0. However TTL processing rules of a service label must be the same as any MPLS label. Due to this a TTL of 0 in the service label would prevent the packet from being forwarded beyond the LSR that provides the service. To avoid this problem, the originator of the "echo request" must remove those service labels from the stack up to the tunnel that is being currently traced. In other words the ingress must remove all service-labels above the label of the tunnel being currently traced, but retain service labels below it when sending the echo request. Note that load balancing may affect the path when the service labels are removed, resulting in a newer path being traversed. However this new path is potentially different only up to the LSR that provides the service. Since this portion of the path was traced when the tunnels above this tunnel in the stack were traced and followed the exact path as the source routed LSP, this should not be a major concern. Sometimes the newer path may have a problem that was not in the original path resulting in a false positive. In such a case the original path can be traversed by changing the label stack to reach the intermediate LSR with labels that route along each hop explicitly.

8. IANA Considerations

To be Updated.

9. Security Considerations

To be Updated.

10. Acknowledgement

The authors would like to thank Stefano Previdi for his review and comments.

The authors wold like to thank Loa Andersson for his comments and recommendation to merge drafts.

11. Contributing Authors

Tarek Saad
Cisco Systems
Email: tsaad@cisco.com

Siva Sivabalan
Cisco Systems
Email: msiva@cisco.com

12. References

12.1. Normative References

[I-D.filsfils-rtgwg-segment-routing-use-cases]
Filsfils, C., Francois, P., Previdi, S., Decraene, B., Litkowski, S., Horneffer, M., Milojevic, I., Shakir, R., Ytti, S., Henderickx, W., Tantsura, J., Kini, S., and E. Crabbe, "Segment Routing Use Cases", draft-filsfils-rtgwg-segment-routing-use-cases-02 (work in progress), October 2013.

[I-D.filsfils-rtgwg-segment-routing]
Filsfils, C., Previdi, S., Bashandy, A., Decraene, B., Litkowski, S., Horneffer, M., Milojevic, I., Shakir, R., Ytti, S., Henderickx, W., Tantsura, J., and E. Crabbe, "Segment Routing Architecture", draft-filsfils-rtgwg-segment-routing-01 (work in progress), October 2013.

[I-D.gredler-spring-mpls]
Gredler, H., Rekhter, Y., Jalil, L., and S. Kini, "Supporting Source/Explicitly Routed Tunnels via Stacked LSPs", draft-gredler-spring-mpls-02 (work in progress), October 2013.

[RFC0792] Postel, J., "Internet Control Message Protocol", STD 5, RFC 792, September 1981.

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

[RFC3443] Agarwal, P. and B. Akyol, "Time To Live (TTL) Processing in Multi-Protocol Label Switching (MPLS) Networks", RFC 3443, January 2003.

[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.

12.2. Informative References

[RFC6291] Andersson, L., van Helvoort, H., Bonica, R., Romascanu, D., and S. Mansfield, "Guidelines for the Use of the "OAM" Acronym in the IETF", BCP 161, RFC 6291, June 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.

Authors' Addresses

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

Email: naikumar@cisco.com

George Swallow
Cisco Systems, Inc.
1414 Massachusetts Ave
Boxborough, MA 01719
US

Email: swallow@cisco.com

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

Email: cpignata@cisco.com

Nobo Akiya
Cisco Systems, Inc.
2000 Innovation Drive
Kanata, ON K2K 3E8
Canada

Email: nobo@cisco.com

Sriganesh Kini
Ericsson

Email: sriganesh.kini@ericsson.com

Hannes Gredler
Juniper Networks

Email: hannes@juniper.net

Mach(Guoyi) Chen
Huawei

Email: mach.chen@huawei.com