Workgroup: SPRING
Internet-Draft:
draft-filsfils-spring-path-tracing-srmpls-03
Published: 18 November 2023
Intended Status: Standards Track
Expires: 21 May 2024
Authors: C. Filsfils
Cisco Systems, Inc.
A. Abdelsalam, Ed.
Cisco Systems, Inc.
P. Camarillo, Ed.
Cisco Systems, Inc.
I. Meilik
Broadcom
M. Valentine
Goldman Sachs
R. Geib
Deutsche Telekom
J. Desmarais
Colt Technology Services

# Path Tracing in SR-MPLS networks

## Abstract

Path Tracing provides a record of the packet path as a sequence of interface ids. In addition, it provides a record of end-to-end delay, per-hop delay, and load on each interface that forwards the packet.

Path Tracing has the lowest MTU overhead compared to alternative proposals such as [INT], [RFC9197], [I-D.song-opsawg-ifit-framework], and [I-D.kumar-ippm-ifa].

Path Tracing supports fine grained timestamp. It has been designed for linerate hardware implementation in the base pipeline.

This document defines the Path Tracing specification for the SR-MPLS dataplane. The Path Tracing specification for the SRv6 dataplane is defined in [I-D.filsfils-spring-path-tracing].

## 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 https://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 21 May 2024.

**Copyright Notice**

Copyright (c) 2023 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 (https://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 Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

**Table of Contents**

1. Introduction
2. Terminology
  2.1. Requirements Language
3. PT Source Node Dataplane Behavior
4. PT Midpoint Node Dataplane Behavior
5. PT Sink Node Dataplane Behavior
6. PT Headers
  6.1. MPLS Hop-by-Hop Path Tracing Option
7. Benefits
8. Security Considerations
9. IANA Considerations
10. References
  10.1. Normative References
  10.2. Informative References
Contributors
Authors' Addresses

## 1. Introduction

Path Tracing provides a record of the packet path as a sequence of interface ids. In addition, it provides a record of end-to-end delay, per-hop delay, and load on each interface that forwards the packet.

Path Tracing has the lowest MTU overhead compared to alternative proposals such as [INT], [RFC9197], [I-D.song-opsawg-ifit-framework], and [I-D.kumar-ippm-ifa].

Path Tracing supports fine grained timestamp. It has been designed for linerate hardware implementation in the base pipeline.

Path Tracing is applicable to both SR-MPLS [RFC8660], as well as SRv6 [RFC8986]. This document defines the Path Tracing specification

for the SR-MPLS dataplane. The SRv6 dataplane is detailed in [I-D.filsfils-spring-path-tracing].

## 2. Terminology

The following terms used within this document are defined in [RFC6790], [RFC8402], [RFC8754], [RFC8986], [I-D.decraene-mpls-slid-encoded-entropy-label-id] and [I-D.filsfils-spring-path-tracing]: Segment Routing (SR), SR Domain, Segment Identifier (SID), SR-MPLS SID, SR Policy, Segment Routing Header (SRH), SR source node, transit node, SR Endpoint, SA, DA, EL, ELI, ELC, PT, PT Probing Instance, PT Source, PT Midpoint, PT Sink, RC, MCD, SRH PT-TLV, TEF.

The following terms are used in this document as defined below:

MPLS HbH-PT: MPLS Hop-by-Hop Path Tracing Option used for Path Tracing. It contains a stack of MCDs. It is defined in Section 6.1 of this document.

SEL: Structured Entropy Label as defined in [I-D.decraene-mpls-slid-encoded-entropy-label-id].

TEF Label: MPLS Label bound to Timestamp, Encapsulation and Forward (TEF) behavior. The allocation of the TEF Label is out of scope of this document.

PTI: PT Indicator is a flag bit used to indicate the presence of the MPLS HbH-PT after the BoS Label and triggers PT behavior at a PT Midpoint.

### 2.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

## 3. PT Source Node Dataplane Behavior

For each configured PT Probing Instance, according to the probe-rate, the PT Source generates a PT probe packet as follows:

```
S01. Generate a new packet
S02. Push an SRH PT-TLV
S03. Set the session ID field of the SRH PT-TLV as per
     PT Probing Instance configuration
S04. Set the Sequence Number field of SRH PT-TLV and
     increase local counter
S05. Push an MPLS HbH-PT header
S06. Set all bits of MCD Stack of the MPLS HbH-PT header to zero
S07. Set the VER field of the MPLS HbH-PT to 0x2
S08. Set the value of Opt Data Len field as per
     the PT Probing Instance configuration
S09. Push an MPLS Structured Entropy Label (SEL)
S10. Set the PTI flag in the ELC field of the SEL
S11. Set the value of the SEL entropy field as per
     the PT Probing Instance configuration
S12. Set Bottom of Stack bit (S) of the SEL to 1
S13. Push an MPLS Entropy Indicator Label (ELI)
S14. Push an MPLS TEF Label as per the PT Probing
     Instance configuration
S15. Set the TC and TTL value of the TEF Label as per
     PT Probing Instance configuration
S16. Push an SR-MPLS transport Label stack as per the
     PT Probing Instance configuration
S17. Set the TC and TTL value of the SR-MPLS transport Labels
     as per PT Probing Instance configuration
S18. Add padding bytes after SRH PT-TLV to reach the desired
     packet size as per the MTU sweeping range configuration in
     the PT Probing Instance configuration
S19. Perform MPLS lookup using the topmost label to
     determine the Outgoing Interface (IFACE-OUT)
S20. Record Transmit 64-bit timestamp (SRC.T64) in the
     T64 field of the SRH PT-TLV
S21. Record IFACE-OUT ID (SRC.OIF) in the IF_ID field
     of the SRH PT-TLV
S22. Record IFACE-OUT Load (SRC.OIL) in the IF_LD field
     of the SRH PT-TLV
S23. Forward the packet via IFACE-OUT
```

Notes:

*The pseudocode describes local processing at a node. An implementation of the pseudocode is compliant as long as the externally observable wire protocol is as described in the pseudocode.

## 4. PT Midpoint Node Dataplane Behavior

When an MPLS LSR router receives an MPLS packet with SEL, the MPLS LSR router processes the SEL as follows:

```
S01. When processing SEL {
S02.    Use Entropy field to compute ECMP hash and decide IFACE-OUT
S03.    IF (SEL[ELC].PTI == 1 and SEL[BOS] == 1) {
S04.       Compute the Midpoint MCD for IFACE-OUT
S05.       Locate the MPLS HbH-PT immediately after SEL
S06.       MPLS_HbH-PT.MCD_Stack[3:Opt_Data_Len -1] =
           MPLS_HbH-PT.MCD_Stack[0:Opt_Data_Len -4]
              //Shift MCD Stack 3Bytes to the right
S07.       MPLS_HbH-PT.MCD_Stack[0:2] = MCD[0:2]
             //i.e., Push the MCD at the beginning of the Stack
S08.    }
S09. }
```

Notes:

*The PT Midpoint behavior MUST be implemented in the normal pipeline to experience the regular datapath (i.e., linerate). Offloading the processing of this option to either the slow-path or a co-processors is not acceptable and yields invalid results.

## 5. PT Sink Node Dataplane Behavior

We define a new MPLS Label bound to an SRv6 Policy with Timestamp, Encapsulation and Forward ("TEF Label" for short). When Node N receives an MPLS packet with topmost Label is TEF Label, N performs the TEF behavior to the MPLS packet.

```
S01. Record Rx 64-bit timestamp (SNK.T64)
S02. Record incoming interface ID (Sink.IIF)
S03. Record incoming interface Load (Sink.IIL)
S04. Push a new IPv6 header
S05. Set the IPv6 SA to the Sink node loopback
S06. Set the IPv6 DA to the first SID in the SRv6 SID List
S07. Set the IPv6 Next Header field to 43 (SRH)
S08. Append an SRH
S09. Set the SRH Next Header field to 137 (MPLS)
S10. Write the SID list in the SRH
S11. Append an SRH PT-TLV
S12. Set the session ID field of the SRH PT-TLV to zero
S13. Set the Sequence Number field of the SRH PT-TLV to zero
S14. Write Sink.T64 in the T64 field of the SRH PT-TLV
S15. Write Sink.IIF in the IF_ID field of the SRH PT-TLV
S16. Write Sink.IIL in the IF_LD field of the SRH PT-TLV
S17. Perform an IPv6 lookup and forward the packet
```

Notes:

*The pseudocode describes local processing at a node. An implementation of the pseudocode is compliant as long as the

externally observable wire protocol is as described in the pseudocode.

## 6. PT Headers

### 6.1. MPLS Hop-by-Hop Path Tracing Option

We define a new header called MPLS Hop-by-Hop Path Tracing option ("MPLS HbH-PT" for short). The header is used to collect the MCD of each PT Midpoint on the packet path. The MPLS HbH-PT has the following format:

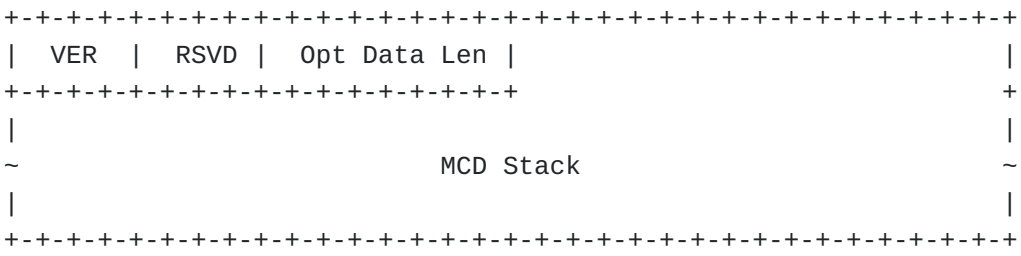


Figure 1: IPv6 Hop-by-Hop Path Tracing Option Format

Where:

*VER: In MPLS, the first nibble after the Label stack indicates the packet IP protocol version. VER is set to 0x2.

*RSVD: Reserved 4-bits. Currently not used.

*Opt Data Len: carries the length of MCD stack (in bytes). Used by PT Midpoint to determine the MCD stack shift value.

*MCD Stack: used to collect the MCDs from PT Midpoints

Note: The MPLS Hop-by-Hop Path Tracing option has a variable length. The operator, upon configuring the Source node behavior, MUST select an option length that is supported by all the routers in the network.

## 7. Benefits

*Insignificant MTU overhead:

-PT has the lowest MTU overhead compared to alternative solutions such as [INT], [RFC9197], [I-D.song-opsawg-ifit-framework], and [I-D.kumar-ippm-ifa].

*Linerate and HW friendliness:

- -Designed for linerate hardware implementation, using the regular forwarding pipeline. No offloading to co-processors whose databases might defer from forwarding pipeline.

- -Leverages mature hardware capabilities (basic shift operation); no packet resizing at every node along the path

*Scalable Fine-grained Timestamp:

- -64-bits timestamp at PT SRC and PT SNK

- -8-bits truncated timestamp at PT Midpoint leveraging flexible per-outgoing-link template allowing diverse link types in the same measurement (e.g., DC, metro, WAN)

*Scalable Load measurement

## 8. Security Considerations

TBD

## 9. IANA Considerations

TBD

## 10. References

### 10.1. Normative References

**[I-D.decraene-mpls-slid-encoded-entropy-label-id]**
Decraene, B., Filsfils, C., Henderickx, W., Saad, T., Beeram, V. P., and L. Jalil, "Using Entropy Label for Network Slice Identification in MPLS networks.", Work in Progress, Internet-Draft, draft-decraene-mpls-slid-encoded-entropy-label-id-05, 12 December 2022, <https://datatracker.ietf.org/doc/html/draft-decraene-mpls-slid-encoded-entropy-label-id-05>.

**[I-D.filsfils-spring-path-tracing]**
Filsfils, C., Abdelsalam, A., Camarillo, P., Yufit, M., Graf, T., Su, Y., Matsushima, S., Valentine, M., and Dhamija, "Path Tracing in SRv6 networks", Work in Progress, Internet-Draft, draft-filsfils-spring-path-tracing-05, 23 October 2023, <https://

datatracker.ietf.org/doc/html/draft-filsfils-spring-path-tracing-05>.

**[RFC2119]** Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

**[RFC6790]** Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, DOI 10.17487/RFC6790, November 2012, <https://www.rfc-editor.org/info/rfc6790>.

**[RFC8174]** Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

**[RFC8402]** Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <https://www.rfc-editor.org/info/rfc8402>.

**[RFC8660]** Bashandy, A., Ed., Filsfils, C., Ed., Previdi, S., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing with the MPLS Data Plane", RFC 8660, DOI 10.17487/RFC8660, December 2019, <https://www.rfc-editor.org/info/rfc8660>.

**[RFC8754]** Filsfils, C., Ed., Dukes, D., Ed., Previdi, S., Leddy, J., Matsushima, S., and D. Voyer, "IPv6 Segment Routing Header (SRH)", RFC 8754, DOI 10.17487/RFC8754, March 2020, <https://www.rfc-editor.org/info/rfc8754>.

**[RFC8986]** Filsfils, C., Ed., Camarillo, P., Ed., Leddy, J., Voyer, D., Matsushima, S., and Z. Li, "Segment Routing over IPv6 (SRv6) Network Programming", RFC 8986, DOI 10.17487/RFC8986, February 2021, <https://www.rfc-editor.org/info/rfc8986>.

### 10.2. Informative References

**[I-D.kumar-ippm-ifa]**
Kumar, J., Anubolu, S., Lemon, J., Manur, R., Holbrook, H., Ghanwani, A., Cai, D., Ou, H., Li, Y., and X. Wang, "Inband Flow Analyzer", Work in Progress, Internet-Draft, draft-kumar-ippm-ifa-07, 7 September 2023, <https://datatracker.ietf.org/doc/html/draft-kumar-ippm-ifa-07>.

**[I-D.song-opsawg-ifit-framework]** Song, H., Qin, F., Chen, H., Jin, J., and J. Shin, "Framework for In-situ Flow Information

Telemetry", Work in Progress, Internet-Draft, draft-song-opsawg-ifit-framework-21, 23 October 2023, <https://datatracker.ietf.org/doc/html/draft-song-opsawg-ifit-framework-21>.

**[INT]** "In-band Network Telemetry (INT) Dataplane Specification", 2020, <https://github.com/p4lang/p4-applications/blob/master/docs/INT_v2_1.pdf>.

**[RFC9197]** Brockners, F., Ed., Bhandari, S., Ed., and T. Mizrahi, Ed., "Data Fields for In Situ Operations, Administration, and Maintenance (IOAM)", RFC 9197, DOI 10.17487/RFC9197, May 2022, <https://www.rfc-editor.org/info/rfc9197>.

## Contributors

Jisu Bhattacharya
Cisco Systems, Inc.
United States of America

Email: jisu@cisco.com

Rakesh Gandhi
Cisco Systems, Inc.
Canada

Email: rgandhi@cisco.com

Shay Zadok
Broadcom
Israel

Email: shay.zadok@broadcom.com

Mark Yufit
Broadcom
Israel

Email: mark.yufit@broadcom.com

Bart Janssens
Colt Technology Services
Belgium

Email: Bart.Janssens@colt.net

## Authors' Addresses

Clarence Filsfils
Cisco Systems, Inc.

Belgium

Email: cf@cisco.com

Ahmed Abdelsalam (editor)
Cisco Systems, Inc.
Italy

Email: ahabdels@cisco.com

Pablo Camarillo Garvia (editor)
Cisco Systems, Inc.
Spain

Email: pcamaril@cisco.com

Israel Meilik
Broadcom
Israel

Email: israel.meilik@broadcom.com

Mike Valentine
Goldman Sachs
United States of America

Email: michael.j.valentine@gs.com

Ruediger Geib
Deutsche Telekom
Germany

Email: Ruediger.Geib@telekom.de

Jonathan Desmarais
Colt Technology Services
United Kingdom

Email: Jonathan.Desmarais@colt.net