SPRING Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 18, 2018

Z. Ali
C. Filsfils
R. Gandhi
N. Kumar
Cisco Systems, Inc.
D. Steinberg
Steinberg Consulting
S. Salsano
Universita di Roma "Tor Vergata"
P. Ventre
CNIT
G. Naik
Drexel University
D. Voyer
D. Bernier
Bell Canada
February 14, 2018

Performance Measurement in Segment Routing Networks with IPv6 Data Plane (SRv6)
draft-ali-spring-srv6-pm-00.txt

Abstract

RFC 6374 specifies protocol mechanisms to enable efficient and accurate measurement of packet loss, one-way and two-way delay, as well as related metrics such as delay variation and channel throughput in MPLS networks. This document describes how these mechanisms can be used for performance measurement in Segment Routing with IPv6 data plane (SRv6) networks.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

Service provider's ability to satisfy Service Level Agreements (SLAs) depend on the ability to measure and monitor performance metrics for packet loss and one-way and two-way delay, as well as related metrics such as delay variation and channel throughput. The ability to monitor these performance metrics also provides operators with greater visibility into the performance characteristics of their networks, thereby facilitating planning, troubleshooting, and network performance evaluation.

[RFC6374] specifies protocol mechanisms to enable the efficient and accurate measurement of these performance metrics in MPLS networks. The One-Way Active Measurement Protocol (OWAMP) defined in [RFC4656] and Two-Way Active Measurement Protocol (TWAMP) defined in [RFC5357] provide capabilities for the measurement of various performance metrics in IP networks. However, mechanisms in [RFC6374] are more suitable for Segment Routing when using MPLS data plane [I-D.spring-sr-mpls-pm]. This document describes how these mechanisms can be used for Performance Measurement (PM) in Segment Routing with the IPv6 data plane (SRv6) networks.

2. Conventions Used in This Document

2.1. Key Word Definitions

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.

2.2. Abbreviations

DM: Delay Measurement.

ECMP: Equal Cost Multi-Path.

LM: Loss Measurement.

PM: Performance Measurement.

SID: Segment ID.

SL: Segment Left.

SRH: Segment Routing Header.

TC: Traffic Class.

UCMP: Unequal Cost Multi-Path.

2.3. Terminology and Reference Topology

In this document, the following simple topology is used for illustration.

```
                               --------
    +--------------------------| N100 |------------------------+
    |                          --------                        |
    |                                                          |
       ====== link1====== link3------ link5====== link9------
       ||N1||======||N2||======| N3 |======||N4||======| N5 |
       ||  ||------||  ||------|    |------||  ||------|    |
       ====== link2====== link4------ link6======link10------
                      |                       |
                      |        ------         |
                      +--------| N6 |---------+
                        link7  |    | link8
                               ------
```

Figure 1: Reference Topology

In the reference topology in Figure 1:

Nodes N1, N2, and N4 are SRv6 capable nodes.

Nodes N3, N5 and N6 are classic IPv6 nodes.

Node 100 is a controller.

Node Nk has a classic IPv6 loopback address Bk::/128

Node Nk has Ak::/48 for its local SID space from which Local SIDs are explicitly allocated.

The IPv6 address of the nth Link between node X and Y at the X side is represented as 99:X:Y::Xn. e.g., the IPv6 address of link6 (the 2nd link) between N3 and N4 at N3 in Figure 1 is 99:3:4:32. Similarly, the IPv6 address of link5 (the 1st link between N3 and N4) at node 3 is 99:3:4::31.

Ak::0 is explicitly allocated as the END function at Node k.

Ak::Cij is explicitly allocated as the END.X function at node k towards neighbor node i via jth Link between node i and node j. e.g.,

A2::C31 represents END.X at N2 towards N3 via link3 (the 1st link between N2 and N3). Similarly, A4::C52 represents the END.X at N4 towards N5 via link10.

<S1, S2, S3> represents a SID list where S1 is the first SID and S3 is the last SID. (S3, S2, S1; SL) represents the same SID list but encoded in the SRH format where the rightmost SID (S1) in the SRH is the first SID and the leftmost SID (S3) in the SRH is the last SID.

(SA, DA) (S3, S2, S1; SL) represents an IPv6 packet, SA is the IPv6 source address, DA the IPv6 destination address, (S3, S2, S1; SL) is the SRH header that includes the SID list <S1, S2, S3>.

SR policy is defined in Section 3 of [I-D.spring-segment-routing-policy].

3. Performance Delay Measurement

3.1. One-Way Delay Measurement

The one-way delay measurement for Packet IP network is defined in

[RFC7679]. It is further exemplified using the following Figure 2.

```
                                          ------
                                         |N100|
                                         |    |
                                          ------
                                            ^
                                            | Response Option-2
                 T1                 T2      |
            +-------+/     Query         \+-------+
            |       | - - - - - - - - - ->|       |
            |   N1  |=====================|   N4  |
            |       |<- - - - - - - - - - |       |
            +-------+\ Response Option-1 /+-------+
                 T4                 T3
```

Figure 2: Delay Measurement Reference Model

Nodes N1 and N4 may not be directly connected, as shown in the reference topology in Figure 1. When N1 and N4 are not directly connected, the one-way delay measurement reflects the delay observed by the packet over an arbitrary SRv6 segment-list (policy) [I-D.spring-segment-routing-policy]. In other words, the one-way delay is associated with the forward (N1 to N4) direction of the SRv6 segment-list.

The delay measurement can be performed using Active (using synthetic probe) mode and Passive (using data stream aka in-band) mode. In both modes, T1 refers to the time the packet is transmitted from N1. Timestamp is added as late as possible at the egress pipeline (in hardware) at node N1. T2 refers to the time the packet is received at N2. Timestamp at the receiver (N2) is added as soon as possible at the ingress pipeline (in hardware).

The one-way delay metric can be computed as follow [RFC7679], [RFC6374],

One-way delay = T2 - T1

Clock synchronization on the querier and responder nodes using the methods detailed in [RFC6374] is required.

Note that for one-way delay measurement, the receiver (node N4 in Figure 2) may send a response to the sender or to a controller (N100 in Figure 2). The controller may also request the querier (node N1 in Figure 2) to initiate delay measurement (this messaging is not shown in Figure 2 and is beyond the scope of this document).

3.2. Two-Way Delay Measurement

[RFC6374], Section 3.4 defines timestamp format that can be used for delay measurement. The IEEE 1588 Precision Time Protocol (PTP) timestamp format [IEEE1588] is used by default as described in Appendix A of [RFC6374], but it may require hardware support.

Note that for one-way delay measurement, Clock synchronization between the querier and responder nodes using methods detailed in [RFC6374] is required. Two-way delay measurement does not require clock to be synchronized between the querier and responder nodes.

3.3. Delay Measurement Message Format

[I-D.6man-segment-routing-header] defines Segment Routing Header (SRH) for SRv6. SRH can contain TLVs, as specified in [I-D.6man-segment-routing-header]. This document specifies Delay Measurement (DM) TLV that is carried in SRH for both one-way and two-way delay measurement. The DM TLV uses a modified DM message format specified in [RFC6374] and is defined as follows:

```
 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     |            RESERVED           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```



```
|Version| Flags |  Control Code |            RESERVED           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  QTF  |  RTF  | RPTF  |               Reserved                |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      Session Identifier           |    TC     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                          Timestamp 1                          |
|                                                               |
```

```
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
.                                                               .
.                                                               .
.                                                               .
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                           Timestamp 4                         |
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                           SUB-TLV Block                       ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 3: Delay Measurement TLV Format

The meanings of the fields are summarized in the following table.

Field	Meaning
Type	SRH DM TLV type (Value TBA)
Length	Total length of the TLV in bytes
Version	Protocol version
Flags	Message control flags
Control Code	Code identifying the query or response type
QTF	Querier timestamp format
RTF	Responder timestamp format
RPTF	Responder's preferred timestamp format
Reserved	Reserved for future specification
Session Identifier	Set arbitrarily by the querier
Traffic Class (TC) Field	Traffic Class being measured
Timestamp 1-4	64-bit timestamp values (see Section 3.4 in [RFC6374])
SUB-TLV Block	Optional block of Type-Length-Value fields

Reserved fields MUST be set to 0 and ignored upon receipt. The possible values for the remaining fields are as follows.

Version: Currently set to 1 (to identify definition of TC field in [RFC6374])

Flags: As specified in [RFC6374]. The T flag in a DM message is set

to 1.

Control Code: As specified in [RFC6374].

Message Length: Set to the total length of this message in bytes, including the Version, Flags, Control Code, and Message Length fields as well as the TLV Block, if any.

Querier Timestamp Format: The format of the timestamp values written by the querier, as specified in Section 3.4 of [RFC6374].

Responder Timestamp Format: The format of the timestamp values written by the responder, as specified in Section 3.4 of [RFC6374].

Responder's Preferred Timestamp Format: The timestamp format preferred by the responder, as specified in Section 3.4 of [RFC6374].

Session Identifier: Set arbitrarily in a query and copied in the response, if any. This field uniquely identifies a measurement operation (also called a session) that consists of a sequence of messages. All messages in the sequence have the same Session Identifier [RFC6374].

TC: Traffic Class being measured.

Timestamp 1-4 (T1-T4): The mapping of timestamps to the Timestamp 1-4 fields is designed to ensure that transmit timestamps are always written at the same fixed offset in the packet, and likewise for receive timestamps. This property is important for hardware processing.

SUB-TLV Block: Zero or more TLV fields. This document assumes the use of the DM message TLVs defined in [RFC6374].

3.3.1. Timestamping

[RFC6374], Section 3.4 defines timestamp format that can be used for delay measurement. The IEEE 1588 Precision Time Protocol (PTP) timestamp format [IEEE1588] is used by default as described in Appendix A of [RFC6374], but it may require hardware support. As an alternative, Network Time Protocol (NTP) timestamp format is also supported in [RFC6374].

Note that for one-way delay measurement, Clock synchronization between the querier and responder nodes using methods detailed in [RFC6374] is required. Two-way delay measurement does not require

clock to be synchronized between the querier and responder nodes.

3.4. One-Way Delay Measurement using Synthetic Probes

For delay measurement using synthetic probes, a DM TLV is inserted in the SRH to record the timestamps and END.OTP SID as described in the pseudo code in [I-D.spring-srv6-network-programming] are used to punt the probe packets.

3.4.1. Example Procedure

To measure one-way delay from node N1 over an SRv6 Policy [I-D.spring-segment-routing-policy] that goes through a segment-list <A2::C31, A4::C52> to node N4, the following procedure is followed:

- o Node N1 constructs a DM probe packet with (B1::0, A2::C31)(A4::C52, A2::C31, SL=1; NH=NONE, DM TLV). To punt the DM probe packet at node N4, node N1 inserts the END.OTP SID [I-D.spring-srv6-network-programming] just before the target SID A4::C52 in the SRH. Thus, the packet as it leaves node N1 looks like (B1::0, A2::C31)(A4::C52, A4::OTP, A2::C31; SL=2; NH=NONE, DM TLV (with T1 from N1)). The PM synthetic probe query message does not contain any payload data.

- o When node N4 receives the packet (B1::0, A4::OTP)(A4::C52, A4::OTP, A2::C31; SL=1; NH=NONE, DM TLV), it processes the END.OTP SID, as described in the pseudo code in [I-D.spring-srv6-network-programming]. In doing so, it punts the timestamped packet (with T2 from N4) to the Performance Measurement (PM) process in control plane for processing.

3.5. In-band One-Way Segment-by-Segment Delay Measurement

For delay measurement for in-band with data traffic, a DM TLV in the SRH to record timestamps and O-bit as described in [I-D.spring-srv6-network-programming] to punt a copy of the packet on every SRv6 nodes are used.

3.5.1. Example Procedure

Consider the case where the user wants to measure one-way delay from node N1 over an SRv6 Policy [I-D.spring-segment-routing-policy] that goes through a segment-list <A2::C31, A4::C52>. However, the user desired to measure delay in-band with data traffic on a segment-by-segment basis.

o To force a punt of the timestamped copy of the data packet at node

N2 and node N4, node N1 sets the O-bit in SRH at locally configured periodic measurement interval. The packet, as it leaves node 1, looks like (B1::0, A2::C31)(A4::C52, A2::C31; SL=1, Flags.O=1, DM TLV (with T1 from N1), NH=data payload type)(data payload). Here, the data payload refers to the actual data traffic going over the policy whose performance is being measured. Node N1 may optionally punt a timestamped copy of the packet with T1 to the local PM process in control plane.

o When node N2 receives the packet (B1::0, A2::C31)(A4::C52, A2::C31; SL=1, Flags.O=1, DM TLV, NH=data payload type)(data payload) packet, it processes the O-bit in SRH, as described in the pseudo code in [I-D.spring-srv6-network-programming]. A timestamped copy of the packet gets punted to the PM process in control plane for processing. Node N2 continues to apply the A2::C31 SID function on the original packet and forwards it, accordingly. As SRH.Flags.O=1, Node N2 also disables the PSP behavior, i.e., does not remove the SRH.

o The PM process in control plane at node N2 sends the copy of the timestamped packet (with DM TLV containing T1 from N1 and T2 from N2) to a locally configured controller or to the querier. Please note that, as mentioned in [I-D.spring-srv6-network-programming], if node N2 does not support the O-bit, it simply ignores it and processes the local SID, A2::C31. In this case, the controller will not get the performance data from the segments with the nodes that do not support the O-bit.

o When node N4 receives the packet (B1::0, A4::C52)(A4::C52, A2::C31; SL=0, Flags.O=1, DM TLV (containing T1 from N1); NH=data payload type)(data payload), it processes the O-bit in SRH, as described in the pseudo code in [I-D.spring-srv6-network-programming]. A timestamped copy of the packet gets punted to the PM process in control plane for processing.

o The PM process in control plane at node N2 sends the copy of the timestamped packet (with DM TLV containing T1 from N1 and T2 from N4) to a locally configured controller.

The controller processes the timestamped packet from each segment and computes the segment-by-segment one-way delay.

3.5.2. Node Capability

Support for O-bit is part of node capability advertisement. This enables node N1 and the controller N100 to know which segment nodes are capable of sending timestamped copy of packets.

4. Performance Loss Measurement

To be added.

5. Probe Reply Message

5.1. One-way Measurement Probe Reply

For one-way performance measurement [RFC7679], the PM querier node can receive "out-of-bands" probe replies by properly setting the UDP Return Object (URO) TLV in the probe message. The URO TLV (Type=131) is defined in [RFC7876] and includes the UDP-Destination-Port and IP Address. In particular, if the querier sets its own IP address in the URO TLV the probe response is sent back by the responder node to the querier node.

The PM process in the control plane on the responder node copies the content of the DM TLV into the payload of the PM reply message.

5.1.1. Probe Reply Message to Controller

As shown in Figure 1, if the querier node N1 requires the probe reply to be sent to the controller N100, it sets the IP address of N100 in the Address field of the URO TLV of the PM probe query message.

5.2. Two-way Measurement Probe Reply

To be added.

6. Security Considerations

TBA.

7. IANA Considerations

IANA is requested to allocate a value for the new SRH TLV Type for Delay Measurement.

8. References

8.1. Normative References

[IEEE1588] IEEE, "1588-2008 IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems", March 2008.

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

[RFC6374] Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS networks', RFC 6374, September 2011.

[RFC7876] Bryant, S., Sivabalan, S., and Soni, S., "UDP Return Path for Packet Loss and Delay Measurement for MPLS Networks", RFC 7876, July 2016.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", RFC 8174, May 2017.

[I-D.spring-srv6-network-programming] Filsfils, C. et al. "SRv6 Network Programming", draft-filsfils-spring-srv6-network-programming, work in progress.

[I-D.6man-segment-routing-header] Previdi, S., Filsfils, et al, "IPv6 Segment Routing Header (SRH)", draft-ietf-6man-segment-routing-header, work in progress.

8.2. Informative References

[RFC4656] Shalunov, S., Teitelbaum, B., Karp, A., Boote, J., and M.

Zekauskas, "A One-way Active Measurement Protoco (OWAMP)", RFC 4656, September 2006.

[RFC5357] Hedayat, K., Krzanowski, R., Morton, A., Yum, K., and J. Babiarz, "A Two-Way Active Measurement Protocol (TWAMP)", RFC 5357, October 2008.

[RFC7679] Almes, G., et al, "A One-Way Delay Metric for IP Performance Metrics (IPPM)', RFC 7679, January 2016.

[I-D.spring-segment-routing-policy] Filsfils, C., et al. "Segment Routing Policy for Traffic Engineering", draft-filsfils-spring-segment-routing-policy, work in progress.

[I-D.spring-sr-mpls-pm] Filsfils, C., et al. "Segment Routing Policy for Traffic Engineering", draft-gandhi-spring-sr-mpls-pm, work in progress.

Acknowledgments

To be added.

Contributors

Faisal Iqbal
Cisco Systems, Inc.
Email: faiqbal@cisco.com

Carlos Pignataro
Cisco Systems, Inc.
Email: cpignata@cisco.com

Authors' Addresses

Clarence Filsfils
Cisco Systems, Inc.
Email: cfilsfil@cisco.com

Zafar Ali
Cisco Systems, Inc.
Email: zali@cisco.com

Rakesh Gandhi
Cisco Systems, Inc.
Canada
Email: rgandhi@cisco.com

Nagendra Kumar
Cisco Systems, Inc.
Email: naikumar@cisco.com

Dirk Steinberg
Steinberg Consulting
Germany
Email: dws@dirksteinberg.de

Stefano Salsano
Universita di Roma "Tor Vergata"
Italy

Email: stefano.salsano@uniroma2.it

Pier Luigi Ventre
CNIT
Italy
Email: pierluigi.ventre@cnit.it

Gaurav Naik
Drexel University

USA
Email: gn@drexel.edu

Daniel Voyer
Bell Canada
Email: daniel.voyer@bell.ca

Daniel Bernier
Bell Canada
Email: daniel.bernier@bell.ca