Packet Loss and Delay Measurement for the MPLS Transport Profile
draft-frost-mpls-tp-loss-delay-00

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt.

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

This Internet-Draft will expire on April 21, 2010.

Copyright Notice

Copyright (c) 2009 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 in effect on the date of publication of this document (http://trustee.ietf.org/license-info). Please review these documents carefully, as they describe your rights and restrictions with respect to this document.

Abstract

An essential Operations, Administration and Maintenance requirement of the MPLS Transport Profile (MPLS-TP) is the ability to monitor performance metrics for packet loss and one-way and two-way delay for
MPLS-TP pseudowires, Label Switched Paths, and Sections. This document specifies protocol mechanisms to facilitate the efficient and accurate measurement of these performance metrics.

Requirements Language

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].
Table of Contents

1. Introduction .................................................. 4
   1.1. Review of Requirements ................................. 4
       1.1.1. Requirements for Packet Loss Measurement ............ 4
       1.1.2. Requirements for Delay Measurement ................. 4
   1.2. Terminology ............................................ 5
2. Overview ..................................................... 5
   2.1. Implementation Considerations .......................... 7
   2.2. Packet Loss Measurement ................................ 7
   2.3. Delay Measurement ........................................ 9
       2.3.1. Timestamp Format .................................... 10
   2.4. Delay Variation Measurement ............................. 11
   2.5. Unidirectional Connections ............................... 11
3. Packet Format .................................................. 11
   3.1. Loss Measurement Message Format ........................ 12
   3.2. Delay Measurement Message Format ....................... 14
   3.3. Timestamp Field Formats ................................ 16
4. Operation ..................................................... 17
   4.1. Loss Measurement Procedures ............................. 17
       4.1.1. Initiating a Loss Measurement Operation ............ 17
       4.1.2. Transmitting a Loss Measurement Query ............. 17
       4.1.3. Receiving a Loss Measurement Query ................ 18
       4.1.4. Transmitting a Loss Measurement Response .......... 18
       4.1.5. Receiving a Loss Measurement Response ............. 18
       4.1.6. Scope of Packet Loss Counters ...................... 19
       4.1.7. Message Loss and Packet Misorder Conditions ........ 19
   4.2. Delay Measurement Procedures ............................ 20
       4.2.1. Transmitting a Delay Measurement Query ............. 20
       4.2.2. Receiving a Delay Measurement Query ............... 20
       4.2.3. Transmitting a Delay Measurement Response .......... 21
       4.2.4. Receiving a Delay Measurement Response ............. 22
       4.2.5. Timestamp Format Negotiation ....................... 22
5. A Uni-format Implementation ................................... 23
6. Congestion Considerations ..................................... 23
7. Security Considerations ....................................... 24
8. IANA Considerations ........................................... 24
9. References .................................................... 24
   9.1. Normative References .................................... 24
   9.2. Informative References ................................... 25
Authors’ Addresses .............................................. 25
1.  Introduction

The MPLS Transport Profile (MPLS-TP) [I-D.ietf-mpls-tp-framework] comprises the set of protocol functions that meet the requirements [RFC5654] for the application of MPLS to transport networks.

The document [I-D.ietf-mpls-tp-oam-requirements] specifies Operations, Administration and Maintenance (OAM) definitions and requirements for the measurement of packet loss and one-way and two-way delay for MPLS-TP pseudowires (PWs), Label Switched Paths (LSPs), and Sections. For convenience these definitions and requirements are summarized in the following subsections.

1.1.  Review of Requirements

1.1.1.  Requirements for Packet Loss Measurement

The MPLS-TP OAM tool-set MUST provide a function to enable the quantification of packet loss ratio over a PW, LSP or Section.

Packet loss ratio is the ratio of the user packets not delivered to the total number of user packets transmitted during a defined time interval. The number of user packets not delivered is the difference between the number of user packets transmitted by an End Point and the number of user packets received at an End Point.

This function MAY either be performed pro-actively or on-demand. It SHOULD be performed between End Points of PWs, LSPs and Sections. It SHOULD be possible to rely on user traffic to perform that functionality.

The protocol solution(s) developed to perform this function MUST apply to point-to-point bidirectional (associated and co-routed) LSPs, point-to-point unidirectional LSPs and point-to-multipoint LSPs.

1.1.2.  Requirements for Delay Measurement

The MPLS-TP OAM tool-set MUST provide a function to enable the quantification of the one-way, and if appropriate, the two-way, delay of a PW, LSP or Section.

- One-way delay is the time elapsed from the start of transmission of the first bit of a packet by an End Point until the reception of the last bit of that packet by the other End Point.

- Two-way delay is the time elapsed from the start of transmission of the first bit of a packet by a End Point until the reception of
the last bit of that packet by the same End Point, when loop-back is performed at the other End Point.

This function SHOULD be performed on-demand and MAY be performed proactively. It SHOULD be performed between End Points of PWs, LSPs and Sections.

In addition to co-routed bidirectional LSPs, the protocol solution(s) developed to perform this function MUST also apply to point-to-point associated bidirectional LSPs, point-to-point unidirectional LSPs and point-to-multipoint LSPs but only to enable the quantification of the one-way delay.

1.2. Terminology

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACH</td>
<td>Associated Channel Header</td>
</tr>
<tr>
<td>DM</td>
<td>Delay Measurement</td>
</tr>
<tr>
<td>G-ACh</td>
<td>Generic Associated Channel</td>
</tr>
<tr>
<td>LM</td>
<td>Loss Measurement</td>
</tr>
<tr>
<td>LSP</td>
<td>Label Switched Path</td>
</tr>
<tr>
<td>LSR</td>
<td>Label Switching Router</td>
</tr>
<tr>
<td>MPLS-TP</td>
<td>MPLS Transport Profile</td>
</tr>
<tr>
<td>OAM</td>
<td>Operations, Administration and Maintenance</td>
</tr>
<tr>
<td>PW</td>
<td>Pseudowire</td>
</tr>
</tbody>
</table>

2. Overview

The basic procedures for measuring loss and delay over a bidirectional connection are conceptually simple. The following figure shows the reference scenario.

```
  +-------+/      Query       \+-------+
  |       |  - - - - - - - - - - ->    |
  |   A   |  ==============            |   B   |
  |       |  <- - - - - - - - - - -   |
  +-------+
```

T1                  T2

Figure 1

The figure shows a bidirectional connection between two LSRs, A and B, and illustrates the temporal reference points T1-T4 associated with a measurement operation that takes place at A. The operation consists of A sending a query message to B, and B sending back a
response. Each reference point indicates the point in time at which either the query or the response message is transmitted or received over the connection.

In this situation, A can arrange to measure the packet loss over the connection in the forward and reverse directions by sending Loss Measurement (LM) query messages to B each of which contains the count of packets transmitted prior to time T1 over the connection to B (A_TxP). When the message reaches B, it appends two values and reflects the message back to A: the count of packets received prior to time T2 over the connection from A (B_RxP), and the count of packets transmitted prior to time T3 over the connection to A (B_TxP). When the response reaches A, it appends a fourth value, the count of packets received prior to time T4 over the connection from B (A_RxP).

These four counter values enable A to compute the desired loss statistics. Because the transmit count at A and the receive count at B (and vice versa) may not be synchronized at the time of the first message, and to limit the effects of counter wrap, the loss is computed in the form of a delta between messages.

To measure at A the delay over the connection to B, a Delay Measurement (DM) query message is sent from A to B containing a timestamp recording the instant at which it is transmitted, i.e. T1. When the message reaches B, a timestamp is added recording the instant at which it is received (T2). The message can now be reflected from B to A, with B adding its transmit timestamp (T3) and A adding its receive timestamp (T4). These four timestamps enable A to compute the one-way delay in each direction, as well as the two-way delay for the connection. The one-way delay computations require that the clocks of A and B be synchronized; mechanisms for clock synchronization are outside the scope of this document.

In the case of a unidirectional connection (i.e. a unidirectional point-to-point or point-to-multipoint MPLS-TP LSP) rooted at A, the first half of each of the above procedures can be carried out to measure the forward one-way loss and delay associated with the LSP. At this point the measurement can either take place at the terminal node(s) of the connection rather than at A, or an out-of-band connection can be used, if available, to communicate the data back to A.

LM and DM messages flow over the Generic Associated Channel (G-ACh) [RFC5586] of an MPLS-TP connection (pseudowire, LSP or Section).

[[N1: The term "connection" is used in this document to mean an MPLS-TP PW, LSP, or Section. Either this or another term will be]}
The challenge in carrying out the above procedures lies with the implementation. For accurate loss measurement to occur, packets must not be sent between the time the transmit count for an outbound LM message is determined and the time the message is actually transmitted. Similarly, packets must not be received and processed between the time an LM message is received and the time the receive count for the message is determined. For accurate delay measurement, timestamps must be recorded in DM messages at a point in time as close as possible to when the message is actually transmitted or received over the connection.

These accuracy requirements imply that a hardware-based forwarding implementation may require hardware support for the processing of LM and DM messages. An important consideration of the LM/DM protocol and message format is therefore support for efficient hardware processing.

In situations where such accuracy is not required, or the necessary level of support is not available, an implementation MAY still generate and respond to LM and DM messages but SHOULD make its accuracy limitations clear to the user. In general the DM procedures described in this document remain viable under these conditions, but the procedures for LM may be inadequate. An alternate approach to LM in such situations is to assemble an approximate view of connection quality through sustained invasive generation of test messages alongside client traffic. Such alternative procedures are outside the scope of this document.

2.2. Packet Loss Measurement

Suppose a bidirectional connection such as an MPLS-TP pseudowire, bidirectional LSP, or Section exists between the LSRs A and B. The objective is to measure at A the following two quantities associated with the connection:

- $A_{\text{TxLoss}}$ (transmit loss): the number of packets transmitted by A over the connection but not received at B;
- $A_{\text{RxLoss}}$ (receive loss): the number of packets transmitted by B over the connection but not received at A.

This is accomplished by initiating a Loss Measurement (LM) operation at A, which consists of transmission of a sequence of LM query messages ($LM[1], LM[2], \ldots$) over the connection at a specified rate,
such as one every 100 milliseconds. Each message LM[n] contains the following value:

A_TxP[n]: the total count of packets transmitted by A over the connection prior to the time this message is transmitted.

When such a message is received at B, the following value is recorded in the message:

B_RxP[n]: the total count of packets received by B over the connection at the time this message is received (excluding the message itself).

At this point, B inserts an appropriate response code into the message and transmits it back to A, recording within it the following value:

B_TxP[n]: the total count of packets transmitted by B over the connection prior to the time this response is transmitted.

When the message response is received back at A, the following value is recorded in the message:

A_RxP[n]: the total count of packets received by A over the connection at the time this response is received (excluding the message itself).

The transmit loss A_TxLoss[n-1,n] and receive loss A_RxLoss[n-1,n] within the measurement interval marked by the messages LM[n-1] and LM[n] are computed by A as follows:

A_TxLoss[n-1,n] = (A_TxP[n] - A_TxP[n-1]) - (B_RxP[n] - B_RxP[n-1])
A_RxLoss[n-1,n] = (B_TxP[n] - B_TxP[n-1]) - (A_RxP[n] - A_RxP[n-1])

where the arithmetic is modulo the counter size.

The derived values

A_TxLoss = A_TxLoss[1,2] + A_TxLoss[2,3] + ...
A_RxLoss = A_RxLoss[1,2] + A_RxLoss[2,3] + ...

are updated each time a response to an LM message is received and processed, and represent the total transmit and receive loss over the connection since the LM operation was initiated.

When computing the values A_TxLoss[n-1,n] and A_RxLoss[n-1,n] the possibility of counter wrap must be taken into account. Consider for
example the values of the A_TxP counter at times n-1 and n. Clearly if A_TxP[n] is allowed to wrap to 0 and then beyond to a value equal to or greater than A_TxP[n-1], the computation of an unambiguous A_TxLoss[n-1,n] value will be impossible. Therefore the LM message rate MUST be sufficiently high, given the counter size and the speed and minimum packet size of the underlying connection, that this condition cannot arise. For example, a 32-bit counter for a 100 Gbps link with a minimum packet size of 64 bytes can wrap in 2^32 / (10^11/(64*8)) = ~22 seconds, which is therefore an upper bound on the LM message interval under such conditions.

2.3. Delay Measurement

Suppose a bidirectional connection such as an MPLS-TP pseudowire, bidirectional LSP, or Section exists between the LSRs A and B. The objective is to measure at A one or more of the following quantities associated with the connection:

- The one-way delay associated with the forward (A to B) direction of the connection;
- The one-way delay associated with the reverse (B to A) direction of the connection;
- The two-way delay (A to B to A) associated with the connection.

Of course, if the first two quantities are known then the third is immediate, being just their sum. Measurement of the one-way delay quantities, however, requires that the clocks of A and B be synchronized, whereas the two-way delay can be measured directly even when this is not the case (provided A and B have stable clocks).

The measurement is accomplished by sending a Delay Measurement (DM) query message over the connection to B which contains the following timestamp:

T1: the time the DM query message is transmitted from A.

When the message arrives at B, the following timestamp is recorded in the message:

T2: the time the DM query message is received at B.

At this point B inserts an appropriate response code into the message and transmits it back to A, recording within it the following timestamp:
T3: the time the DM response message is transmitted from B.

When the message arrives back at A, the following timestamp is recorded in the message:

T4: the time the DM response message is received back at A.

At this point, A can compute the two-way delay associated with the connection as

\[
\text{two-way delay} = (T4 - T1) - (T3 - T2).
\]

If the clocks of A and B are known at A to be synchronized, then all three delay values can be computed at A as

forward one-way delay = T2 - T1

reverse one-way delay = T4 - T3

two-way delay = forward delay + reverse delay.

2.3.1. Timestamp Format

There are at least two significant timestamp formats in common use: the timestamp format of the Internet standard Network Time Protocol (NTP), described in [RFC1305] and [RFC2030], and the timestamp format used in the IEEE 1588 Precision Time Protocol (PTP) [IEEE1588].

[[N2: There are actually two PTP timestamp formats: the 1588v1 format consists of a 32-bit seconds field and a 32-bit nanoseconds field; in 1588v2 the seconds field was extended to 48 bits. --DF]]

The NTP format has the advantages of wide use and long deployment in the Internet, and was specifically designed to make the computation of timestamp differences as simple and efficient as possible. On the other hand, there is also now a significant deployment of equipment designed to support the PTP format.

The approach taken in this document is therefore to include in DM messages fields which identify the timestamp formats used by the two devices involved in a DM operation. This implies that an LSR attempting to carry out a DM operation may be faced with the problem of computing with and possibly reconciling different timestamp formats. Support for multiple timestamp formats is OPTIONAL. An implementation SHOULD, however, make clear which timestamp formats it supports and the extent of its support for computation with and reconciliation of different formats for purposes of delay measurement.
In accordance with Internet standards for network time, the NTP timestamp format is the default format used in DM messages. This format MUST be supported.

2.4. Delay Variation Measurement

Packet Delay Variation [RFC3393] is another performance metric important in some applications. The PDV of a pair of packets within a stream of packets is defined for a selected pair of packets in the stream going from measurement point MP1 to measurement point MP2. The PDV is the difference between the one-way delay of the selected packets.

A PDV measurement can therefore be derived from successive delay measurements obtained through the procedures in Section 2.3. An important point regarding PDV measurement, however, is that it can be carried out based on one-way delay measurements even when the clocks of the two systems involved in those measurements are not synchronized.

2.5. Unidirectional Connections

In the case that the connection from A to (B1, ..., Bk) is unidirectional, i.e. is a unidirectional LSP, LM and DM measurements can be carried out at B1, ..., Bk instead of at A.

For LM this is accomplished by initiating an LM operation at A and carrying out the same procedures as for bidirectional connections, except that no responses from B1, ..., Bk to A are generated. Instead, each terminal node B uses the A_TxP and B_RxP values in the LM messages it receives to compute the receive loss associated with the connection in essentially the same way as described previously, i.e.

\[ B_{RxLoss}[n-1,n] = (A_{TxP}[n] - A_{TxP}[n-1]) - (B_{RxP}[n] - B_{RxP}[n-1]) \]

For DM, of course, only the forward one-way delay can be measured and the clock synchronization requirement applies.

Alternatively, if an out-of-band connection from a terminal node B back to A is available, the LM and DM message responses can be communicated to A via this connection so that the measurements can be carried out at A.

3. Packet Format

Loss Measurement and Delay Measurement messages flow over the Generic
Associated Channel (G-ACh) [RFC5586] of an MPLS-TP connection (pseudowire, LSP or Section).

[[N3: The question of ACH TLV usage and the manner of supporting metadata such as authentication keys and node identifiers is deliberately omitted. These issues will be addressed in a future version of the document. --DF]]

3.1. Loss Measurement Message Format

The format of a Loss Measurement message, beginning with the Associated Channel Header (ACH), is as follows:

```
+-----------------+-----------------+-----------------+-----------------+
|       0         |       1         |       2         |       3         |
+-----------------+-----------------+-----------------+-----------------+
| 0 0 0 1|Version|   Reserved    |  0xHH (MPLS-TP Loss)          |
+-----------------+-----------------+-----------------+-----------------+
|Version| Flags |  Control Code |          Reserved             |
+-----------------+-----------------+-----------------+-----------------+
|                         Querier Context                       |
+-----------------+-----------------+-----------------+-----------------+
|                           Counter 1                           |
+-----------------+-----------------+-----------------+-----------------+
|                           Counter 4                           |
+-----------------+-----------------+-----------------+-----------------+
```

Figure 2: Loss Measurement Message Format

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

<table>
<thead>
<tr>
<th>Field</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Version</td>
<td>Protocol version</td>
</tr>
<tr>
<td>Flags</td>
<td>Message control flags</td>
</tr>
<tr>
<td>Control Code</td>
<td>Code identifying the query or response type</td>
</tr>
<tr>
<td>Reserved</td>
<td>Reserved for future specification</td>
</tr>
<tr>
<td>Querier Context</td>
<td>Set arbitrarily by the querier</td>
</tr>
<tr>
<td>Counter 1-4</td>
<td>64-bit packet counter values in network byte order</td>
</tr>
</tbody>
</table>
The possible values for these fields are as follows.

Version: Currently set to 0.

Flags: Each bit represents a message control flag. The flags, listed in left-to-right (most- to least-significant-bit) order, are:

Q/R: Set to 0 for a Query and 1 for a Response.

Remaining bits: Reserved for future specification and set to 0.

Control Code: Set as follows according to whether the message is a Query or a Response as identified by the Q/R flag.

For a Query:

0x0: Query (in-band response requested). Indicates that this query has been sent over a bidirectional connection and the response is expected over the same connection.

0x1: Query (out-of-band response requested). Indicates that the response should be sent via an out-of-band channel.

0x2: Query (no response requested). Indicates that no response to the query should be sent.

For a Response:

0x1: Success. Indicates that the operation was successful.

0x8: Notification - Data Format Invalid. Indicates that the query was processed but the format of the data fields in this response may be inconsistent. Consequently these data fields MUST NOT be used for measurement.

0x10: Error - Unspecified Error. Indicates that the operation failed for an unspecified reason.

0x11: Error - Unsupported Version. Indicates that the operation failed because the protocol version supplied in the query message is not supported.

0x12: Error - Unsupported Control Code. Indicates that the operation failed because the Control Code requested an operation that is not available for this connection.

0x13: Error - Authentication Failure. Indicates that the operation failed because the authentication data supplied in
the query was missing or incorrect.

0x14: Error - Invalid Source Node Identifier. Indicates that the operation failed because the Source Node Identifier supplied in the query is not expected.

0x15: Error - Invalid Destination Node Identifier. Indicates that the operation failed because the Destination Node Identifier supplied in the query is not the identifier of this node.

0x16: Error - Connection Mismatch. Indicates that the operation failed because the connection identifier supplied in the query did not match the connection over which the query was received.

0x17: Error - Query Rate Exceeded. Indicates that the operation failed because the rate of query messages exceeded the configured threshold.

0x18: Error - Administrative Block. Indicates that the operation failed because it has been administratively disallowed.

0x19: Error - Temporary Resource Exhaustion. Indicates that the operation failed because node resources were not available.

Reserved: Currently set to 0.

Querier Context: Set arbitrarily in a query and copied in the response.

Counter 1-4: Referring to Section 2.2, when a query is sent from A, Counter 1 is set to A_TxP and the other counter fields are set to 0. When the query is received at B, Counter 2 is set to B_RxP. At this point, B copies Counter 1 to Counter 3 and Counter 2 to Counter 4, and re-initializes Counter 1 and Counter 2 to 0. When B transmits the response, Counter 1 is set to B_TxP. When the response is received at A, Counter 2 is set to A_RxP. All counter values MUST be in network byte order.

3.2. Delay Measurement Message Format

The format of a Delay Measurement message, beginning with the Associated Channel Header (ACH), is as follows:
The meanings of the fields following the ACH are summarized in the following table.

<table>
<thead>
<tr>
<th>Field</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Version</td>
<td>Protocol version</td>
</tr>
<tr>
<td>Flags</td>
<td>Message control flags</td>
</tr>
<tr>
<td>Control Code</td>
<td>Code identifying the query or response type</td>
</tr>
<tr>
<td>QTF</td>
<td>Querier timestamp format</td>
</tr>
<tr>
<td>RTF</td>
<td>Responder timestamp format</td>
</tr>
<tr>
<td>RPTF</td>
<td>Responder’s preferred timestamp format</td>
</tr>
<tr>
<td>Resv (Reserved)</td>
<td>Reserved for future specification</td>
</tr>
<tr>
<td>Querier Context</td>
<td>Set arbitrarily by the querier</td>
</tr>
<tr>
<td>Timestamp 1-4</td>
<td>128-bit timestamp values</td>
</tr>
<tr>
<td>Padding</td>
<td>Optional padding</td>
</tr>
</tbody>
</table>

The possible values for these fields are as follows.

Version: Currently set to 0.
Flags: As specified in Section 3.1.

Control Code: As specified in Section 3.1.

Querier Timestamp Format: The format of the timestamp values written by the querier, as specified in Section 3.3.

Responder Timestamp Format: The format of the timestamp values written by the responder, as specified in Section 3.3.

Responder’s Preferred Timestamp Format: The timestamp format preferred by the responder, as specified in Section 3.3.

Resv (Reserved): Currently set to 0.

Querier Context: Set arbitrarily in a query and copied in the response.

Timestamp 1-4: Referring to Section 2.3, when a query is sent from A, Timestamp 1 is set to T1 and the other timestamp fields are set to 0. When the query is received at B, Timestamp 2 is set to T2. At this point, B copies Timestamp 1 to Timestamp 3 and Timestamp 2 to Timestamp 4, and re-initializes Timestamp 1 and Timestamp 2 to 0. When B transmits the response, Timestamp 1 is set to T3. When the response is received at A, Timestamp 2 is set to T4. The actual formats of the timestamp fields written by A and B are indicated by the Querier Timestamp Format and Responder Timestamp Format fields respectively.

Padding: One or more octets of padding may optionally follow the Timestamp 4 field in a query, in order to allow for delay measurement based on packets of a particular size. The values of the pad octets, if present, are arbitrary, and if any are present they will be copied in the response.

The next version of this document will describe a mechanism to allow the querier to specify whether the responder should include padding in the response.

3.3. Timestamp Field Formats

The following timestamp format field values are specified in this document:

0x0: Network Time Protocol version 4 timestamp format [RFC2030]. This format consists of a 32-bit seconds field followed by a 32-bit fractional seconds field, so that it can be regarded as a fixed-point 64-bit quantity.
0x2: IEEE 1588-2008 Precision Time Protocol timestamp format [IEEE1588]. This format consists of a 48-bit seconds field followed by a 32-bit nanoseconds field.

In accordance with Internet standards for network time, the NTP timestamp format is the default format used in Delay Measurement messages. This format MUST be supported. Support for other timestamp formats is OPTIONAL.

Timestamp formats of n < 128 bits in size SHALL be encoded in the 128-bit timestamp fields specified in this document using the n high-order bits of the field. The remaining 128 - n low-order bits in the field SHOULD be set to 0 and MUST be ignored when reading the field.

4. Operation

4.1. Loss Measurement Procedures

4.1.1. Initiating a Loss Measurement Operation

An LM operation for a particular MPLS-TP connection consists of sending a sequence (LM[1], LM[2], ...) of LM query messages over the connection at a specific rate and processing the responses received, if any. As described in Section 2.2, the packet loss associated with the connection during the operation is computed as a delta between successive messages; these deltas can be accumulated to obtain a running total of the packet loss for the connection. The query message transmission rate MUST be sufficiently high, given the 64-bit LM message counter size and the speed and minimum packet size of the underlying connection, that the ambiguity condition noted in Section 2.2 cannot arise.

4.1.2. Transmitting a Loss Measurement Query

When transmitting an LM Query over an MPLS-TP connection, the Version and Reserved fields MUST be set to 0. The Q/R flag MUST be set to 0 and the remaining flag bits MUST be set to 0.

The Control Code field MUST be set to one of the values for Query messages listed in Section 3.1; if the connection is unidirectional, this field MUST NOT be set to 0x0 (Query: in-band response requested).

The Querier Context field can be set arbitrarily.

The Counter 1 field SHOULD be set to the total count of packets transmitted over the connection prior to this LM Query. The
remaining Counter fields MUST be set to 0.

4.1.3. Receiving a Loss Measurement Query

Upon receipt of an LM Query message, the Counter 2 field SHOULD be set to the total count of packets received over the connection prior to this LM Query.

At this point the LM Query message must be inspected. If the Control Code field is set to 0x2 (no response requested), an LM Response message MUST NOT be transmitted. If the Control Code field is set to 0x0 (in-band response requested) or 0x1 (out-of-band response requested), then an in-band or out-of-band response, respectively, SHOULD be transmitted unless this has been prevented by an administrative, security or congestion control mechanism.

4.1.4. Transmitting a Loss Measurement Response

When constructing a Response to an LM Query, the Version and Reserved fields MUST be set to 0. The Q/R flag MUST be set to 1 and the remaining flag bits MUST be set to 0.

The Querier Context field MUST be copied from the LM Query. The Counter 1 and Counter 2 fields from the LM Query MUST be copied to the Counter 3 and Counter 4 fields, respectively, of the LM Response.

The Control Code field MUST be set to one of the values for Response messages listed in Section 3.1. The value 0x10 (Unspecified Error) SHOULD NOT be used if one of the other more specific error codes is applicable.

If the response is transmitted in-band, the Counter 1 field SHOULD be set to the total count of packets transmitted over the connection prior to this LM Response. If the response is transmitted out-of-band, the Counter 1 field MUST be set to 0. In either case, the Counter 2 field MUST be set to 0.

4.1.5. Receiving a Loss Measurement Response

Upon in-band receipt of an LM Response message, the Counter 2 field SHOULD be set to the total count of packets received over the connection prior to this LM Response.

Upon out-of-band receipt of an LM Response message, the Counter 1 and Counter 2 fields MUST NOT be used for purposes of loss measurement.

If the Control Code in an LM Response is anything other than 0x1 (Success), the counter values in the response MUST NOT be used for
purposes of loss measurement. When the Control Code indicates an error condition, the LM operation SHOULD be suspended and an appropriate notification to the user generated. If a temporary error condition is indicated, the LM operation MAY be restarted automatically.

4.1.6. Scope of Packet Loss Counters

By default the packet counts appearing in LM messages on a connection MUST include packets transmitted and received over the Generic Associated Channel (G-ACh) associated with the connection. An implementation MAY provide the means to change the scope of the LM counters to exclude some or all G-ACh messages. Care must be taken in this case to ensure that the scopes of the counters at both ends of a connection agree.

4.1.7. Message Loss and Packet Misorder Conditions

Because an LM operation consists of a message sequence with state maintained from one message to the next, LM is subject to the effects of lost messages and misordered packets in a way that DM is not. Because this state exists only on the querier, the handling of these conditions is, strictly speaking, a local matter. This section, however, presents RECOMMENDED procedures for handling such conditions.

The first kind of anomaly that may occur is that one or more LM messages may be lost in transit. The effect of such loss is that when an LM Response is next received at the querier, an unambiguous interpretation of the counter values it contains may be impossible, for the reasons described at the end of Section 2.2. Whether this is so depends on the number of messages lost and the other variables mentioned in that section, such as the LM message rate and the connection parameters.

Another possibility is that LM messages are misordered in transit, so that for instance the response to LM[n] is received prior to the response to LM[n-1]. A typical implementation will discard the late response to LM[n-1], so that the effect is the same as the case of a lost message.

Finally, LM is subject to the possibility that data packets are misordered relative to LM messages. This condition can result, for example, in a transmit count of 100 and a corresponding receive count of 101. The effect here is that the A_TxLoss[n-1,n] value (for example) for a given measurement interval will appear to be extremely (if not impossibly) large. The other case, where an LM message arrives earlier than some of the packets, simply results in those
packets being counted as lost, which is usually what is desired.

Perhaps the simplest way to detect and handle the case of lost or out-of-order LM messages is to incorporate a sequence number in each message. Such a sequence number can be inserted within the bounds of the Querier Context field provided for implementation-specific use. An implementation adopting this approach can now take the following actions:

[[N4: Text to be added here about handling the above conditions with sequence numbers and thresholds. --DF]]

4.2. Delay Measurement Procedures

4.2.1. Transmitting a Delay Measurement Query

When transmitting a DM Query over an MPLS-TP connection, the Version and Reserved fields MUST be set to 0. The Q/R flag MUST be set to 0 and the remaining flag bits MUST be set to 0.

The Control Code field MUST be set to one of the values for Query messages listed in Section 3.1; if the connection is unidirectional, this field MUST NOT be set to 0x0 (Query: in-band response requested).

The Querier Context field can be set arbitrarily.

The Querier Timestamp Format field MUST be set to the timestamp format used by the querier when writing timestamp fields in this message; the possible values for this field are listed in Section 3.3. The Responder Timestamp Format and Responder’s Preferred Timestamp Format fields MUST be set to 0.

The Timestamp 1 field SHOULD be set to the time at which this DM Query is transmitted, in the format indicated by the Querier Timestamp Format field. The other timestamp fields MUST be set to 0.

One or more pad octets with arbitrary values MAY follow the Timestamp 4 field.

4.2.2. Receiving a Delay Measurement Query

Upon receipt of a DM Query message, the Timestamp 2 field SHOULD be set to the time at which this DM Query is received.

At this point the DM Query message must be inspected. If the Control Code field is set to 0x2 (no response requested), a DM Response message MUST NOT be transmitted. If the Control Code field is set to
0x0 (in-band response requested) or 0x1 (out-of-band response requested), then an in-band or out-of-band response, respectively, SHOULD be transmitted unless this has been prevented by an administrative, security or congestion control mechanism.

4.2.3. Transmitting a Delay Measurement Response

When constructing a Response to a DM Query, the Version and Reserved fields MUST be set to 0. The Q/R flag MUST be set to 1 and the remaining flag bits MUST be set to 0.

The Querier Context and Querier Timestamp Format (QTF) fields MUST be copied from the DM Query. The Timestamp 1 and Timestamp 2 fields from the DM Query MUST be copied to the Timestamp 3 and Timestamp 4 fields, respectively, of the DM Response.

The Responder Timestamp Format (RTF) field MUST be set to the timestamp format used by the responder when writing timestamp fields in this message, i.e. Timestamp 4 and (if applicable) Timestamp 1; the possible values for this field are listed in Section 3.3. Furthermore, the RTF field MUST be set equal either to the QTF or the RPTF field. See Section 4.2.5 for guidelines on selection of the value for this field.

The Responder’s Preferred Timestamp Format (RPTF) field MUST be set to one of the values listed in Section 3.3 and SHOULD be set to indicate the timestamp format with which the responder can provide the best accuracy for purposes of delay measurement.

The Control Code field MUST be set to one of the values for Response messages listed in Section 3.1. The value 0x10 (Unspecified Error) SHOULD NOT be used if one of the other more specific error codes is applicable.

If the response is transmitted in-band, the Timestamp 1 field SHOULD be set to the time at which this DM Response is transmitted. If the response is transmitted out-of-band, the Timestamp 1 field MUST be set to 0. In either case, the Timestamp 2 field MUST be set to 0.

If the response is transmitted in-band and the Control Code in the message is 0x1 (Success), then the Timestamp 1 and Timestamp 4 fields MUST have the same format, which will be the format indicated in the Responder Timestamp Format field.

Padding SHALL be included in the response if, and only if, padding was present in the DM Query, in which case the response padding MUST be identical to the query padding.
4.2.4. Receiving a Delay Measurement Response

Upon in-band receipt of a DM Response message, the Timestamp 2 field SHOULD be set to the time at which this DM Response is received.

Upon out-of-band receipt of a DM Response message, the Timestamp 1 and Timestamp 2 fields MUST NOT be used for purposes of delay measurement.

If the Control Code in a DM Response is anything other than 0x1 (Success), the timestamp values in the response MUST NOT be used for purposes of delay measurement. When the Control Code indicates an error condition, an appropriate notification to the user SHOULD be generated.

4.2.5. Timestamp Format Negotiation

In case either the querier or the responder in a DM transaction is capable of supporting multiple timestamp formats, it is desirable to determine the optimal format for purposes of delay measurement on a particular connection. The procedures for making this determination SHALL be as follows.

Upon sending an initial DM Query over a connection, the querier sets the Querier Timestamp Format (QTF) field to its preferred timestamp format.

Upon receiving any DM Query message, the responder determines whether it is capable of writing timestamps in the format specified by the QTF field. If so, the Responder Timestamp Format (RTF) field is set equal to the QTF field. If not, the RTF field is set equal to the Responder’s Preferred Timestamp Format (RPTF) field.

The process of changing from one timestamp format to another at the responder may result in the Timestamp 1 and Timestamp 4 fields in an in-band DM Response having different formats. If this is the case, the Control Code in the response MUST NOT be set to 0x1 (Success). Unless an error condition has occurred, the Control Code MUST be set to 0x2 (Notification - Data Format Invalid).

Upon receiving a DM Response, the querier knows from the RTF field in the message whether the responder is capable of supporting its preferred timestamp format: if it is, the RTF will be equal to the QTF. The querier also knows the responder’s preferred timestamp format from the RPTF field. The querier can then decide whether to retain its current QTF or to change it and repeat the negotiation procedures.
5. A Uni-format Implementation

Editor’s note. This text on the execution of the protocol on simple hardware need further thought and will be updated in the next version of this document.

A simple implementation of this protocol that only understands one time format MAY discard all Query messages with a QTF type that it does not support. Similarly a simple implementation may discard all Response messages with an RTF type that it does not support. Such an implementation would only successfully execute a delay measurement if both the query and response systems were configured to use identical formats.

6. Congestion Considerations

An MPLS-TP network may be traffic-engineered in such a way that the bandwidth required both for client traffic and for control, management and OAM traffic is always available. The following congestion considerations therefore apply only when this is not the case.

The proactive generation of Loss Measurement and Delay Measurement messages for purposes of monitoring the performance of an MPLS-TP connection naturally results in a degree of additional load placed on both the network and the terminal nodes of the connection. When configuring such monitoring, operators should be mindful of the overhead involved and should choose transmit rates that do not stress network resources unduly; such choices must be informed by the deployment context. In case of slower links or lower-speed devices, for example, lower Loss Measurement message rates can be chosen, up to the limits noted at the end of Section 2.2.

In general, lower measurement message rates place less load on the network at the expense of reduced granularity. For delay measurement this reduced granularity translates to a greater possibility that the delay associated with a connection temporarily exceeds the expected threshold without detection. For loss measurement, it translates to a larger gap in loss information in case of exceptional circumstances such as lost LM messages or misordered packets.

When carrying out a sustained measurement operation such as an LM operation or continuous pro-active DM operation, the querier SHOULD take note of the number of lost measurement messages (queries for which a response is never received) and set a corresponding Measurement Message Loss Threshold. If this threshold is exceeded, the measurement operation SHOULD be suspended so as not to exacerbate
the possible congestion condition. This suspension SHOULD be accompanied by an appropriate notification to the user so that the condition can be investigated and corrected.

From the receiver perspective, the main consideration is the possibility of receiving an excessive quantity of measurement messages. An implementation SHOULD employ a mechanism such as rate-limiting to guard against the effects of this case. Authentication procedures can also be used to ensure that only queries from authorized devices are processed.

7. Security Considerations

There are two main types of security considerations associated with the exchange of performance monitoring messages such as those described in this document: the possibility of a malicious or misconfigured device generating an excessive quantity of messages, causing service impairment; and the possibility of an unauthorized device learning the data contained in or implied by such messages.

The first consideration is discussed in Section 6. If reception of performance-related data by unauthorized devices is an operational concern, message authentication procedures such as those described in [xref] should be used to ensure that only queries from authorized devices are processed.

8. IANA Considerations

A future version of this document will detail IANA considerations for:

- ACH Channel Types for LM and DM messages
- Timestamp format registry
- LM and DM Control Codes

9. References

9.1. Normative References


[RFC5654] Niven-Jenkins, B., Brungard, D., Betts, M., Sprecher, N.,
9.2. Informative References

[I-D.ietf-mpls-tp-framework]
Bocci, M., Bryant, S., Frost, D., and L. Levrau, "A Framework for MPLS in Transport Networks",


RFC 2030, October 1996.

RFC 3393, November 2002.

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

Authors’ Addresses

Dan Frost (editor)
Cisco Systems

Email: danfrost@cisco.com