Internet Engineering Task Force (IETF)
Request for Comments: 6374
Category: Standards Track
ISSN: 2070-1721
September 2011

D. Frost
S. Bryant
Cisco Systems

# Packet Loss and Delay Measurement for MPLS Networks

## Abstract

Many service provider 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.  This measurement capability also provides operators with greater visibility into the performance characteristics of their networks, thereby facilitating planning, troubleshooting, and network performance evaluation.  This document specifies protocol mechanisms to enable the efficient and accurate measurement of these performance metrics in MPLS networks.

## Status of This Memo

## Copyright Notice

Table of Contents

## 1.  Introduction

Many service provider 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.  This measurement capability also provides operators with greater visibility into the performance characteristics of their networks, thereby facilitating planning, troubleshooting, and network performance evaluation.  This document specifies protocol mechanisms to enable the efficient and accurate measurement of these performance metrics in MPLS networks.

This document specifies two closely related protocols, one for packet loss measurement (LM) and one for packet delay measurement (DM). These protocols have the following characteristics and capabilities:

o  The LM and DM protocols are intended to be simple and to support efficient hardware processing.

- The LM and DM protocols operate over the MPLS Generic Associated Channel (G-ACh) [RFC5586] and support measurement of loss, delay, and related metrics over Label Switched Paths (LSPs), pseudowires, and MPLS sections (links).

- The LM and DM protocols are applicable to the LSPs, pseudowires, and sections of networks based on the MPLS Transport Profile (MPLS-TP), because the MPLS-TP is based on a standard MPLS data plane. The MPLS-TP is defined and described in [RFC5921], and MPLS-TP LSPs, pseudowires, and sections are discussed in detail in [RFC5960]. A profile describing the minimal functional subset of the LM and DM protocols in the MPLS-TP context is provided in [RFC6375].

- The LM and DM protocols can be used both for continuous/proactive and selective/on-demand measurement.

- The LM and DM protocols use a simple query/response model for bidirectional measurement that allows a single node -- the querier -- to measure the loss or delay in both directions.

- The LM and DM protocols use query messages for unidirectional loss and delay measurement. The measurement can be carried out either at the downstream node(s) or at the querier if an out-of-band return path is available.

- The LM and DM protocols do not require that the transmit and receive interfaces be the same when performing bidirectional measurement.

- The DM protocol is stateless.

- The LM protocol is "almost" stateless: loss is computed as a delta between successive messages, and thus the data associated with the last message received must be retained.

- The LM protocol can perform two distinct kinds of loss measurement: it can measure the loss of specially generated test messages in order to infer the approximate data-plane loss level (inferred measurement) or it can directly measure data-plane packet loss (direct measurement). Direct measurement provides perfect loss accounting, but may require specialized hardware support and is only applicable to some LSP types. Inferred measurement provides only approximate loss accounting but is generally applicable.

The direct LM method is also known as "frame-based" in the context of Ethernet transport networks [Y.1731]. Inferred LM is a generalization of the "synthetic" measurement approach currently in development for Ethernet networks, in the sense that it allows test messages to be decoupled from measurement messages.

- The LM protocol supports measurement in terms of both packet counts and octet counts.

- The LM protocol supports both 32-bit and 64-bit counters.

- The LM protocol can be used to measure channel throughput as well as packet loss.

- The DM protocol supports multiple timestamp formats, and provides a simple means for the two endpoints of a bidirectional connection to agree on a preferred format. This procedure reduces to a triviality for implementations supporting only a single timestamp format.

- The DM protocol supports varying the measurement message size in order to measure delays associated with different packet sizes.

The One-Way Active Measurement Protocol (OWAMP) [RFC4656] and Two-Way Active Measurement Protocol (TWAMP) [RFC5357] provide capabilities for the measurement of various performance metrics in IP networks. These protocols are not streamlined for hardware processing and rely on IP and TCP, as well as elements of the Network Time Protocol (NTP), which may not be available or optimized in some network environments; they also lack support for IEEE 1588 timestamps and direct-mode LM, which may be required in some environments. The protocols defined in this document thus are similar in some respects to, but also differ from, these IP-based protocols.

## 1.1. Applicability and Scope

This document specifies measurement procedures and protocol messages that are intended to be applicable in a wide variety of circumstances and amenable to implementation by a wide range of hardware- and software-based measurement systems. As such, it does not attempt to mandate measurement quality levels or analyze specific end-user applications.

## 1.2. Terminology

| Term | Definition |
|---|---|
| ACH | Associated Channel Header |
| DM | Delay Measurement |
| ECMP | Equal Cost Multipath |
| G-ACh | Generic Associated Channel |
| LM | Loss Measurement |
| LSE | Label Stack Entry |
| LSP | Label Switched Path |
| NTP | Network Time Protocol |
| OAM | Operations, Administration, and Maintenance |
| PTP | Precision Time Protocol |
| TC | Traffic Class |

## 1.3. 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].

## 2. Overview

This section begins with a summary of the basic methods used for the bidirectional measurement of packet loss and delay. These measurement methods are then described in detail. Finally, a list of practical considerations is discussed that may come into play to inform or modify these simple procedures. This section is limited to theoretical discussion; for protocol specifics, the reader is referred to Sections 3 and 4.

### 2.1. Basic Bidirectional Measurement

The following figure shows the reference scenario.

```
                          T1              T2
                +-------+/     Query       \+-------+
                |       | - - - - - - - - ->|       |
                |   A   |===================|   B   |
                |       |<- - - - - - - - - |       |
                +-------+\     Response    /+-------+
                          T4              T3
```

This figure shows a bidirectional channel between two nodes, 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 channel.

In this situation, A can arrange to measure the packet loss over the channel 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 channel 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 channel from A (B_RxP) and the count of packets transmitted prior to time T3 over the channel 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 channel 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 channel 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 channel. 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.

## 2.2. Packet Loss Measurement

Suppose a bidirectional channel exists between the nodes A and B. The objective is to measure at A the following two quantities associated with the channel:

A_TxLoss (transmit loss): the number of packets transmitted by A over the channel but not received at B;

A_RxLoss (receive loss): the number of packets transmitted by B over the channel 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], ...) over the channel 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 channel 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 channel at the time this message is received (excluding the message itself).

At this point, B transmits the message back to A, recording within it the following value:

B_TxP[n]: the total count of packets transmitted by B over the channel 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 channel 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.

(Strictly speaking, it is not necessary that the fourth count, A_RxP[n], actually be written in the message, but this is convenient for some implementations and useful if the message is to be forwarded on to an external measurement system.)

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 they represent the total transmit and receive loss over the channel 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. For example, consider the values of the A_TxP counter at sequence numbers 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 channel, 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. This bound will be referred to as the MaxLMInterval of the channel. It is clear that the MaxLMInterval will be a more restrictive constraint in the case of direct LM and for smaller counter sizes.

The loss measurement approach described in this section has the characteristic of being stateless at B and "almost" stateless at A. Specifically, A must retain the data associated with the last LM response received, in order to use it to compute loss when the next response arrives. This data MAY be discarded, and MUST NOT be used as a basis for measurement, if MaxLMInterval elapses before the next response arrives, because in this case an unambiguous measurement cannot be made.

The foregoing discussion has assumed the counted objects are packets, but this need not be the case. In particular, octets may be counted instead. This will, of course, reduce the MaxLMInterval accordingly.

In addition to absolute aggregate loss counts, the individual loss counts yield other metrics, such as the average loss rate over any multiple of the measurement interval. An accurate loss rate can be determined over time even in the presence of anomalies affecting individual measurements, such as those due to packet misordering (Section 4.2.10).

Note that an approach for conducting packet loss measurement in IP networks is documented in [RFC2680]. This approach differs from the one described here, for example by requiring clock synchronization between the measurement points and lacking support for direct-mode LM.

## 2.3. Throughput Measurement

If LM query messages contain a timestamp recording their time of transmission, this data can be combined with the packet or octet counts to yield measurements of the throughput offered and delivered over the channel during the interval in terms of the counted units.

For a bidirectional channel, for example, given any two LM response messages (separated in time by not more than the MaxLMInterval), the difference between the counter values tells the querier the number of units successfully transmitted and received in the interval between the timestamps. Absolute offered throughput is the number of data units transmitted and absolute delivered throughput is the number of data units received. Throughput rate is the number of data units sent or received per unit time.

Just as for loss measurement, the interval counts can be accumulated to arrive at the absolute throughput of the channel since the start of the measurement operation or be used to derive related metrics such as the throughput rate. This procedure also enables out-of-service throughput testing when combined with a simple packet generator.

## 2.4. Delay Measurement

Suppose a bidirectional channel exists between the nodes A and B. The objective is to measure at A one or more of the following quantities associated with the channel:

o The one-way delay associated with the forward (A to B) direction of the channel;

o The one-way delay associated with the reverse (B to A) direction of the channel;

o The two-way delay (A to B to A) associated with the channel.

The one-way delay metric for packet networks is described in [RFC2679]. In the case of two-way delay, there are actually two possible metrics of interest. The "two-way channel delay" is the sum of the one-way delays in each direction and reflects the delay of the channel itself, irrespective of processing delays within the remote endpoint B. The "round-trip delay" is described in [RFC2681] and includes in addition any delay associated with remote endpoint processing.

Measurement of the one-way delay quantities requires that the clocks of A and B be synchronized, whereas the two-way delay metrics can be measured directly even when this is not the case (provided A and B have stable clocks).

A measurement is accomplished by sending a Delay Measurement (DM) query message over the channel to B that 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 transmits the message 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.

(Strictly speaking, it is not necessary that the fourth timestamp, T4, actually be written in the message, but this is convenient for some implementations and useful if the message is to be forwarded on to an external measurement system.)

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

two-way channel delay = (T4 - T1) - (T3 - T2)

and the round-trip delay as

round-trip delay = T4 - T1.

If the clocks of A and B are known at A to be synchronized, then both one-way delay values, as well as the two-way channel delay, can be computed at A as

forward one-way delay = T2 - T1

reverse one-way delay = T4 - T3

two-way channel delay = forward delay + reverse delay.

Note that this formula for the two-way channel delay reduces to the one previously given, and clock synchronization is not required to compute this metric.

## 2.5. Delay Variation Measurement

Inter-Packet Delay Variation (IPDV) and Packet Delay Variation (PDV) [RFC5481] are performance metrics derived from one-way delay measurement and are important in some applications. IPDV represents the difference between the one-way delays of successive packets in a stream. PDV, given a measurement test interval, represents the difference between the one-way delay of a packet in the interval and that of the packet in the interval with the minimum delay.

IPDV and PDV measurements can therefore be derived from delay measurements obtained through the procedures in Section 2.4. An important point regarding delay variation 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 with one another.

## 2.6. Unidirectional Measurement

In the case that the channel from A to (B1, ..., Bk) (where B2, ..., Bk refers to the point-to-multipoint case) 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 used for bidirectional channels, 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 channel in essentially the same way as described previously, that is:

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 channel from a terminal node B back to A is available, the LM and DM message responses can be communicated to A via this channel so that the measurements can be carried out at A.

### 2.7. Dyadic Measurement

The basic procedures for bidirectional measurement assume that the measurement process is conducted by and for the querier node A. Instead, it is possible, with only minor variation of these procedures, to conduct a dyadic or "dual-ended" measurement process in which both nodes A and B perform loss or delay measurement based on the same message flow. This is achieved by stipulating that A copy the third and fourth counter or timestamp values from a response message into the third and fourth slots of the next query, which are otherwise unused, thereby providing B with equivalent information to that learned by A.

The dyadic procedure has the advantage of halving the number of messages required for both A and B to perform a given kind of measurement, but comes at the expense of each node’s ability to control its own measurement process independently, and introduces additional operational complexity into the measurement protocols. The quantity of measurement traffic is also expected to be low relative to that of user traffic, particularly when 64-bit counters are used for LM. Consequently, this document does not specify a dyadic operational mode. However, it is still possible, and may be useful, for A to perform the extra copy, thereby providing additional information to B even when its participation in the measurement process is passive.

### 2.8. Loopback Measurement

Some bidirectional channels may be placed into a loopback state such that messages are looped back to the sender without modification. In this situation, LM and DM procedures can be used to carry out measurements associated with the circular path. This is done by generating "queries" with the Response flag set to 1.

For LM, the loss computation in this case is:

```
A_Loss[n-1,n] = (A_TxP[n] - A_TxP[n-1]) - (A_RxP[n] - A_RxP[n-1])
```

For DM, the round-trip delay is computed. In this case, however, the remote endpoint processing time component reflects only the time required to loop the message from channel input to channel output.

### 2.9. Measurement Considerations

A number of additional considerations apply in practice to the measurement methods summarized above.

#### 2.9.1. Types of Channels

There are several types of channels in MPLS networks over which loss and delay measurement may be conducted. The channel type may restrict the kinds of measurement that can be performed. In all cases, LM and DM messages flow over the MPLS Generic Associated Channel (G-ACh), which is described in detail in [RFC5586].

Broadly, a channel in an MPLS network may be either a link, a Label Switched Path (LSP) [RFC3031], or a pseudowire [RFC3985]. Links are bidirectional and are also referred to as MPLS sections; see [RFC5586] and [RFC5960]. Pseudowires are bidirectional. Label Switched Paths may be either unidirectional or bidirectional.

The LM and DM protocols discussed in this document are initiated from a single node: the querier. A query message may be received either by a single node or by multiple nodes, depending on the nature of the channel. In the latter case, these protocols provide point-to-multipoint measurement capabilities.

#### 2.9.2. Quality of Service

Quality of Service (QoS) capabilities, in the form of the Differentiated Services architecture, apply to MPLS as specified in [RFC3270] and [RFC5462]. Different classes of traffic are distinguished by the three-bit Traffic Class (TC) field of an MPLS Label Stack Entry (LSE). Delay measurement applies on a per-traffic-class basis, and the TC values of LSEs above the G-ACh Label (GAL) that precedes a DM message are significant. Packet loss can be measured with respect either to the channel as a whole or to a specific traffic class.

#### 2.9.3. Measurement Point Location

The location of the measurement points for loss and delay within the sending and receiving nodes is implementation dependent but directly affects the nature of the measurements. For example, a sending implementation may or may not consider a packet to be "lost", for LM purposes, that was discarded prior to transmission for queuing-related reasons; conversely, a receiving implementation may or may not consider a packet to be "lost", for LM purposes, if it was physically received but discarded during receive-path processing. The location of delay measurement points similarly determines what, precisely, is being measured. The principal consideration here is that the behavior of an implementation in these respects MUST be made clear to the user.

### 2.9.4. Equal Cost Multipath

Equal Cost Multipath (ECMP) is the behavior of distributing packets across multiple alternate paths toward a destination. The use of ECMP in MPLS networks is described in BCP 128 [RFC4928]. The typical result of ECMP being performed on an LSP that is subject to delay measurement will be that only the delay of one of the available paths is, and can be, measured.

The effects of ECMP on loss measurement will depend on the LM mode. In the case of direct LM, the measurement will account for any packets lost between the sender and the receiver, regardless of how many paths exist between them. However, the presence of ECMP increases the likelihood of misordering both of LM messages relative to data packets and of the LM messages themselves. Such misorderings tend to create unmeasurable intervals and thus degrade the accuracy of loss measurement. The effects of ECMP are similar for inferred LM, with the additional caveat that, unless the test packets are specially constructed so as to probe all available paths, the loss characteristics of one or more of the alternate paths cannot be accounted for.

### 2.9.5. Intermediate Nodes

In the case of an LSP, it may be desirable to measure the loss or delay to or from an intermediate node as well as between LSP endpoints. This can be done in principle by setting the Time to Live (TTL) field in the outer LSE appropriately when targeting a measurement message to an intermediate node. This procedure may fail, however, if hardware-assisted measurement is in use, because the processing of the packet by the intermediate node occurs only as the result of TTL expiry, and the handling of TTL expiry may occur at a later processing stage in the implementation than the hardware-assisted measurement function. The motivation for conducting measurements to intermediate nodes is often an attempt to localize a problem that has been detected on the LSP. In this case, if intermediate nodes are not capable of performing hardware-assisted measurement, a less accurate -- but usually sufficient -- software-based measurement can be conducted instead.

### 2.9.6. Different Transmit and Receive Interfaces

The overview of the bidirectional measurement process presented in Section 2 is also applicable when the transmit and receive interfaces at A or B differ from one another. Some additional considerations, however, do apply in this case:

- If different clocks are associated with transmit and receive processing, these clocks must be synchronized in order to compute the two-way delay.

- The DM protocol specified in this document requires that the timestamp formats used by the interfaces that receive a DM query and transmit a DM response agree.

- The LM protocol specified in this document supports both 32-bit and 64-bit counter sizes, but the use of 32-bit counters at any of the up to four interfaces involved in an LM operation will result in 32-bit LM calculations for both directions of the channel.

### 2.9.7. External Post-Processing

In some circumstances, it may be desirable to carry out the final measurement computation at an external post-processing device dedicated to the purpose. This can be achieved in supporting implementations by, for example, configuring the querier, in the case of a bidirectional measurement session, to forward each response it receives to the post-processor via any convenient protocol. The unidirectional case can be handled similarly through configuration of the receiver or by including an instruction in query messages for the receiver to respond out-of-band to the appropriate return address.

Post-processing devices may have the ability to store measurement data for an extended period and to generate a variety of useful statistics from them. External post-processing also allows the measurement process to be completely stateless at the querier and responder.

### 2.9.8. Loss Measurement Modes

The summary of loss measurement at the beginning of Section 2 made reference to the "count of packets" transmitted and received over a channel. If the counted packets are the packets flowing over the channel in the data plane, the loss measurement is said to operate in "direct mode". If, on the other hand, the counted packets are selected control packets from which the approximate loss characteristics of the channel are being inferred, the loss measurement is said to operate in "inferred mode".

Direct LM has the advantage of being able to provide perfect loss accounting when it is available. There are, however, several constraints associated with direct LM.

For accurate direct LM 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. If these "synchronization conditions" do not hold, the LM message counters will not reflect the true state of the data plane, with the result that, for example, the receive count of B may be greater than the transmit count of A, and attempts to compute loss by taking the difference will yield an invalid result. This requirement for synchronization between LM message counters and the data plane may require special support from hardware-based forwarding implementations.

A limitation of direct LM is that it may be difficult or impossible to apply in cases where the channel is an LSP and the LSP label at the receiver is either nonexistent or fails to identify a unique sending node. The first case happens when Penultimate Hop Popping (PHP) is used on the LSP, and the second case generally holds for LSPs based on the Label Distribution Protocol (LDP) [RFC5036] as opposed to, for example, those based on Traffic Engineering extensions to the Resource Reservation Protocol (RSVP-TE) [RFC3209]. These conditions may make it infeasible for the receiver to identify the data-plane packets associated with a particular source and LSP in order to count them, or to infer the source and LSP context associated with an LM message. Direct LM is also vulnerable to disruption in the event that the ingress or egress interface associated with an LSP changes during the LSP’s lifetime.

Inferred LM works in the same manner as direct LM except that the counted packets are special control packets, called test messages, generated by the sender. Test messages may be either packets explicitly constructed and used for LM or packets with a different primary purpose, such as those associated with a Bidirectional Forwarding Detection (BFD) [RFC5884] session.

The synchronization conditions discussed above for direct LM also apply to inferred LM, the only difference being that the required synchronization is now between the LM counters and the test message generation process. Protocol and application designers MUST take these synchronization requirements into account when developing tools for inferred LM, and make their behavior in this regard clear to the user.

Inferred LM provides only an approximate view of the loss level associated with a channel, but is typically applicable even in cases where direct LM is not.

### 2.9.9. Loss Measurement Scope

In the case of direct LM, where data-plane packets are counted, there are different possibilities for which kinds of packets are included in the count and which are excluded. The set of packets counted for LM is called the "loss measurement scope". As noted above, one factor affecting the LM scope is whether all data packets are counted or only those belonging to a particular traffic class. Another is whether various "auxiliary" flows associated with a data channel are counted, such as packets flowing over the G-ACh. Implementations MUST make their supported LM scopes clear to the user, and care must be taken to ensure that the scopes of the channel endpoints agree.

### 2.9.10. Delay Measurement Accuracy

The delay measurement procedures described in this document are designed to facilitate hardware-assisted measurement and to function in the same way whether or not such hardware assistance is used. The measurement accuracy will be determined by how closely the transmit and receive timestamps correspond to actual packet departure and arrival times.

As noted in Section 2.4, measurement of one-way delay requires clock synchronization between the devices involved, while two-way delay measurement does not involve direct comparison between non-local timestamps and thus has no synchronization requirement. The measurement accuracy will be limited by the quality of the local clock and, in the case of one-way delay measurement, by the quality of the synchronization.

### 2.9.11. Delay Measurement Timestamp Format

There are two significant timestamp formats in common use: the timestamp format of the Network Time Protocol (NTP), described in [RFC5905], and the timestamp format used in the IEEE 1588 Precision Time Protocol (PTP) [IEEE1588].

The NTP format has the advantages of wide use and long deployment in the Internet, and it was specifically designed to make the computation of timestamp differences as simple and efficient as possible. On the other hand, there is now also 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 that identify the timestamp formats used by the two devices involved in a DM operation. This implies that a node attempting to carry out a DM operation may be faced with the problem of computing with and possibly reconciling different timestamp formats. To ensure interoperability, it is necessary that support of at least one timestamp format is mandatory. This specification requires the support of the IEEE 1588 PTP format. Timestamp format support requirements are discussed in detail in Section 3.4.

## 3. Message Formats

Loss Measurement and Delay Measurement messages flow over the MPLS Generic Associated Channel (G-ACh) [RFC5586]. Thus, a packet containing an LM or DM message contains an MPLS label stack, with the G-ACh Label (GAL) at the bottom of the stack. The GAL is followed by an Associated Channel Header (ACH), which identifies the message type, and the message body follows the ACH.

This document defines the following ACH Channel Types:

MPLS Direct Loss Measurement (DLM)
MPLS Inferred Loss Measurement (ILM)
MPLS Delay Measurement (DM)
MPLS Direct Loss and Delay Measurement (DLM+DM)
MPLS Inferred Loss and Delay Measurement (ILM+DM)

The message formats for direct and inferred LM are identical. The formats of the DLM+DM and ILM+DM messages are also identical.

For these channel types, the ACH SHALL NOT be followed by the ACH TLV Header defined in [RFC5586].

The fixed-format portion of a message MAY be followed by a block of Type-Length-Value (TLV) fields. The TLV block provides an extensible way of attaching subsidiary information to LM and DM messages. Several such TLV fields are defined below.

All integer values for fields defined in this document SHALL be encoded in network byte order.

### 3.1. Loss Measurement Message Format

The format of a Loss Measurement message, which follows the Associated Channel Header (ACH), is 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
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |Version| Flags |  Control Code |        Message Length         |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       | DFlags|  OTF  |                   Reserved                    |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |                       Session Identifier          |    DS     |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |                        Origin Timestamp                       |
       |                                                               |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |                           Counter 1                           |
       |                                                               |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       .                                                               .
       .                                                               .
       .                                                               .
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |                           Counter 4                           |
       |                                                               |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       ~                           TLV Block                           ~
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Loss Measurement Message Format

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

| Field | Meaning |
|---|---|
| Version | Protocol version |
| Flags | Message control flags |
| Control Code | Code identifying the query or response type |
| Message Length | Total length of this message in bytes |
| Data Format Flags (DFlags) | Flags specifying the format of message data |
| Origin Timestamp Format (OTF) | Format of the Origin Timestamp field |
| Reserved | Reserved for future specification |
| Session Identifier | Set arbitrarily by the querier |
| Differentiated Services (DS) Field | Differentiated Services Code Point (DSCP) being measured |
| Origin Timestamp | 64-bit field for query message transmission timestamp |
| Counter 1-4 | 64-bit fields for LM counter values |
| TLV Block | Optional block of Type-Length-Value fields |

The possible values for these fields are as follows.

Version: Currently set to 0.

Flags: The format of the Flags field is shown below.

```
+-+-+-+-+
|R|T|0|0|
+-+-+-+-+
```

Loss Measurement Message Flags

The meanings of the flag bits are:

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

T: Traffic-class-specific measurement indicator. Set to 1 when the measurement operation is scoped to packets of a particular traffic class (DSCP value), and 0 otherwise. When set to 1, the DS field of the message indicates the measured traffic class.

0: Set to 0.

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

For a Query:

0x0: In-band Response Requested. Indicates that this query has been sent over a bidirectional channel and the response is expected over the same channel.

0x1: Out-of-band Response Requested. Indicates that the response should be sent via an out-of-band channel.

0x2: No Response Requested. Indicates that no response to the query should be sent. This mode can be used, for example, if all nodes involved are being controlled by a Network Management System.

For a Response:

Codes 0x0-0xF are reserved for non-error responses. Error response codes imply that the response does not contain valid measurement data.

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

0x2: 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.

0x3: Notification - Initialization in Progress. Indicates that the query was processed but this response does not contain valid measurement data because the responder’s initialization process has not completed.

0x4: Notification - Data Reset Occurred. Indicates that the query was processed, but a reset has recently occurred that may render the data in this response inconsistent relative to earlier responses.

0x5: Notification - Resource Temporarily Unavailable. Indicates that the query was processed, but resources were unavailable to complete the requested measurement and that, consequently, this response does not contain valid measurement data.

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 channel.

0x13: Error - Unsupported Data Format. Indicates that the operation failed because the data format specified in the query is not supported.

0x14: Error - Authentication Failure. Indicates that the operation failed because the authentication data supplied in the query was missing or incorrect.

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

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

0x17: Error - Unsupported Mandatory TLV Object. Indicates that the operation failed because a TLV Object received in the query and marked as mandatory is not supported.

0x18: Error - Unsupported Query Interval. Indicates that the operation failed because the query message rate exceeded the configured threshold.

0x19: Error - Administrative Block. Indicates that the operation failed because it has been administratively disallowed.

0x1A: Error - Resource Unavailable. Indicates that the operation failed because node resources were not available.

0x1B: Error - Resource Released. Indicates that the operation failed because node resources for this measurement session were administratively released.

0x1C: Error - Invalid Message. Indicates that the operation failed because the received query message was malformed.

0x1D: Error - Protocol Error. Indicates that the operation failed because a protocol error was found in the received query message.

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.

DFlags: The format of the DFlags field is shown below.

```
+-+-+-+-+
|X|B|0|0|
+-+-+-+-+
```

Data Format Flags

The meanings of the DFlags bits are:

X: Extended counter format indicator. Indicates the use of extended (64-bit) counter values. Initialized to 1 upon creation (and prior to transmission) of an LM Query and copied from an LM Query to an LM response. Set to 0 when the LM message is transmitted or received over an interface that writes 32-bit counter values.

B: Octet (byte) count. When set to 1, indicates that the Counter 1-4 fields represent octet counts. The octet count applies to all packets within the LM scope (Section 2.9.9), and the octet count of a packet sent or received over a channel includes the total length of that packet (but excludes headers, labels, or framing of the channel itself). When set to 0, indicates that the Counter 1-4 fields represent packet counts.

0: Set to 0.

Origin Timestamp Format: The format of the Origin Timestamp field, as specified in Section 3.4.

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.

DS: When the T flag is set to 1, this field is set to the DSCP value [RFC3260] that corresponds to the traffic class being measured. For MPLS, where the traffic class of a channel is identified by the three-bit Traffic Class in the channel’s LSE [RFC5462], this field

SHOULD be set to the Class Selector Codepoint [RFC2474] that corresponds to that Traffic Class. When the T flag is set to 0, the value of this field is arbitrary, and the field can be considered part of the Session Identifier.

Origin Timestamp: Timestamp recording the transmit time of the query message.

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.

The mapping of counter types such as A_TxP to the Counter 1-4 fields is designed to ensure that transmit counter values are always written at the same fixed offset in the packet, and likewise for receive counters. This property may be important for hardware processing.

When a 32-bit counter value is written to one of the counter fields, that value SHALL be written to the low-order 32 bits of the field; the high-order 32 bits of the field MUST, in this case, be set to 0.

TLV Block: Zero or more TLV fields.

## 3.2. Delay Measurement Message Format

The format of a Delay Measurement message, which follows the Associated Channel Header (ACH), is 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
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
      |Version| Flags |  Control Code |        Message Length         |
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
      |  QTF  |  RTF  | RPTF  |              Reserved                 |
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
      |                 Session Identifier                |    DS     |
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
      |                           Timestamp 1                         |
      |                                                               |
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
      .                                                               .
      .                                                               .
      .                                                               .
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
      |                           Timestamp 4                         |
      |                                                               |
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
      ~                           TLV Block                           ~
      +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Delay Measurement Message Format

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

| Field | Meaning |
|---|---|
| Version | Protocol version |
| Flags | Message control flags |
| Control Code | Code identifying the query or response type |
| Message Length | Total length of this message in bytes |
| 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 |
| Differentiated Services (DS) Field | Differentiated Services Code Point (DSCP) being measured |
| Timestamp 1-4 | 64-bit timestamp values |
| 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 0.

Flags: As specified in Section 3.1. The T flag in a DM message is set to 1.

Control Code: As specified in Section 3.1.

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.

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

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

Session Identifier: As specified in Section 3.1.

DS: As specified in Section 3.1.

Timestamp 1-4: Referring to Section 2.4, 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.

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.

TLV Block: Zero or more TLV fields.

### 3.3. Combined Loss/Delay Measurement Message Format

The format of a combined Loss and Delay Measurement message, which follows the Associated Channel Header (ACH), is 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
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |Version| Flags |  Control Code |        Message Length         |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       | DFlags|  QTF  |  RTF  | RPTF  |           Reserved            |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |                      Session Identifier           |    DS     |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |                           Timestamp 1                         |
       |                                                               |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       .                                                               .
       .                                                               .
       .                                                               .
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |                           Timestamp 4                         |
       |                                                               |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |                           Counter 1                           |
       |                                                               |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       .                                                               .
       .                                                               .
       .                                                               .
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |                           Counter 4                           |
       |                                                               |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       ~                           TLV Block                           ~
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Loss/Delay Measurement Message Format

The fields of this message have the same meanings as the corresponding fields in the LM and DM message formats, except that the roles of the OTF and Origin Timestamp fields for LM are here played by the QTF and Timestamp 1 fields, respectively.

### 3.4. Timestamp Field Formats

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

0: Null timestamp format.  This value is a placeholder indicating that the timestamp field does not contain a meaningful timestamp.

1: Sequence number. This value indicates that the timestamp field is to be viewed as a simple 64-bit sequence number. This provides a simple solution for applications that do not require a real absolute timestamp, but only an indication of message ordering; an example is LM exception detection.

2: Network Time Protocol version 4 64-bit timestamp format [RFC5905]. 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.

3: Low-order 64 bits of the IEEE 1588-2008 (1588v2) Precision Time Protocol timestamp format [IEEE1588]. This truncated format consists of a 32-bit seconds field followed by a 32-bit nanoseconds field, and is the same as the IEEE 1588v1 timestamp format.

Timestamp formats of n < 64 bits in size SHALL be encoded in the 64-bit timestamp fields specified in this document using the n high-order bits of the field. The remaining 64 - n low-order bits in the field SHOULD be set to 0 and MUST be ignored when reading the field.

To ensure that it is possible to find an interoperable mode between implementations, it is necessary to select one timestamp format as the default. The timestamp format chosen as the default is the truncated IEEE 1588 PTP format (format code 3 in the list above); this format MUST be supported. The rationale for this choice is discussed in Appendix A. Implementations SHOULD also be capable of reading timestamps written in NTPv4 64-bit format and reconciling them internally with PTP timestamps for measurement purposes. Support for other timestamp formats is OPTIONAL.

The implementation MUST make clear which timestamp formats it supports and the extent of its support for computation with and reconciliation of different formats for measurement purposes.

### 3.5. TLV Objects

The TLV Block in LM and DM messages consists of zero or more objects with the following format:

```
 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     |         Value                 ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

TLV Format

The Type and Length fields are each 8 bits long, and the Length field indicates the size in bytes of the Value field, which can therefore be up to 255 bytes long.

The Type space is divided into Mandatory and Optional subspaces:

| Type Range | Semantics |
|---|---|
| 0-127 | Mandatory |
| 128-255 | Optional |

Upon receipt of a query message including an unrecognized mandatory TLV object, the recipient MUST respond with an Unsupported Mandatory TLV Object error code.

The types defined are as follows:

| Type | Definition |
|---|---|
| Mandatory | |
| 0 | Padding - copy in response |
| 1 | Return Address |
| 2 | Session Query Interval |
| 3 | Loopback Request |
| 4-126 | Unallocated |
| 127 | Experimental use |
| | |
| Optional | |
| 128 | Padding - do not copy in response |
| 129 | Destination Address |
| 130 | Source Address |
| 131-254 | Unallocated |
| 255 | Experimental use |

### 3.5.1. Padding

The two padding objects permit the augmentation of packet size; this is mainly useful for delay measurement. The type of padding indicates whether the padding supplied by the querier is to be copied to, or omitted from, the response. Asymmetrical padding may be useful when responses are delivered out-of-band or when different maximum transmission unit sizes apply to the two components of a bidirectional channel.

More than one padding object MAY be present, in which case they MUST be contiguous. The Value field of a padding object is arbitrary.

### 3.5.2. Addressing

The addressing objects have the following format:

```
     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     |        Address Family         |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    ~                            Address                            ~
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Addressing Object Format

The Address Family field indicates the type of the address, and it SHALL be set to one of the assigned values in the "IANA Address Family Numbers" registry.

The Source and Destination Address objects indicate the addresses of the sender and the intended recipient of the message, respectively. The Source Address of a query message SHOULD be used as the destination for an out-of-band response unless some other out-of-band response mechanism has been configured, and unless a Return Address object is present, in which case the Return Address specifies the target of the response.  The Return Address object MUST NOT appear in a response.

### 3.5.3. Loopback Request

The Loopback Request object, when included in a query, indicates a request that the query message be returned to the sender unmodified. This object has a Length of 0.

Upon receiving the reflected query message back from the responder, the querier MUST NOT retransmit the message.  Information that uniquely identifies the original query source, such as a Source Address object, can be included to enable the querier to differentiate one of its own loopback queries from a loopback query initiated by the far end.

This object may be useful, for example, when the querier is interested only in the round-trip delay metric.  In this case, no support for delay measurement is required at the responder at all, other than the ability to recognize a DM query that includes this object and return it unmodified.

### 3.5.4. Session Query Interval

The Value field of the Session Query Interval object is a 32-bit unsigned integer that specifies a time interval in milliseconds.

```
     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     |            Session Query      >
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    <        Interval (ms)          |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Session Query Interval Object Format

This time interval indicates the interval between successive query messages in a specific measurement session. The purpose of the Session Query Interval (SQI) object is to enable the querier and responder of a measurement session to agree on a query rate. The procedures for handling this object SHALL be as follows:

1. The querier notifies the responder that it wishes to be informed of the responder's minimum query interval for this session by including the SQI object in its query messages, with a Value of 0.

2. When the responder receives a query that includes an SQI object with a Value of 0, the responder includes an SQI object in the response with the Value set to the minimum query interval it supports for this session.

3. When the querier receives a response that includes an SQI object, it selects a query interval for the session that is greater than or equal to the Value specified in the SQI object and adjusts its query transmission rate accordingly, including in each subsequent query an SQI object with a Value equal to the selected query interval. Once a response to one of these subsequent queries has been received, the querier infers that the responder has been apprised of the selected query interval and MAY then stop including the SQI object in queries associated with this session.

Similar procedures allow the query rate to be changed during the course of the session by either the querier or the responder. For example, to inform the querier of a change in the minimum supported query interval, the responder begins including a corresponding SQI object in its responses, and the querier adjusts its query rate if necessary and includes a corresponding SQI object in its queries until a response is received.

Shorter query intervals (i.e., higher query rates) provide finer measurement granularity at the expense of additional load on measurement endpoints and the network; see Section 6 for further discussion.

# 4. Operation

## 4.1. Operational Overview

A loss or delay measurement operation, also called a session, is controlled by the querier and consists of a sequence of query messages associated with a particular channel and a common set of measurement parameters. If the session parameters include a response request, then the receiving node or nodes will (under normal conditions) generate a response message for each query message received, and these responses are also considered part of the session. All query and response messages in a session carry a common session identifier.

Measurement sessions are initiated at the discretion of the network operator and are terminated either at the operator’s request or as the result of an error condition. A session may be as brief as a single message exchange, for example when a DM query is used by the operator to "ping" a remote node, or it may extend throughout the lifetime of the channel.

When a session is initiated for which responses are requested, the querier SHOULD initialize a timer, called the SessionResponseTimeout, that indicates how long the querier will wait for a response before abandoning the session and notifying the user that a timeout has occurred. This timer persists for the lifetime of the session and is reset each time a response message for the session is received.

When a query message is received that requests a response, a variety of exceptional conditions may arise that prevent the responder from generating a response that contains valid measurement data. Such conditions fall broadly into two classes: transient exceptions from which recovery is possible and fatal exceptions that require termination of the session. When an exception arises, the responder SHOULD generate a response with an appropriate Notification or Error control code according to whether the exception is, respectively, transient or fatal. When the querier receives an Error response, the session MUST be terminated and the user informed.

A common example of a transient exception occurs when a new session is initiated and the responder requires a period of time to become ready before it can begin providing useful responses. The response control code corresponding to this situation is Notification -

Initialization in Progress. Typical examples of fatal exceptions are cases where the querier has requested a type of measurement that the responder does not support or where a query message is malformed.

When initiating a session, the querier SHOULD employ the Session Query Interval mechanism (Section 3.5.4) to establish a mutually agreeable query rate with the responder. Responders SHOULD employ rate-limiting mechanisms to guard against the possibility of receiving an excessive quantity of query messages.

## 4.2. Loss Measurement Procedures

### 4.2.1. Initiating a Loss Measurement Operation

An LM operation for a particular channel consists of sending a sequence (LM[1], LM[2], ...) of LM query messages over the channel at a specific rate and processing the responses received, if any. As described in Section 2.2, the packet loss associated with the channel 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 channel or be used to derive related metrics such as the average loss rate.

The query message transmission rate MUST be sufficiently high, given the LM message counter size (which can be either 32 or 64 bits) and the speed and minimum packet size of the underlying channel, that the ambiguity condition noted in Section 2.2 cannot arise. In evaluating this rate, the implementation SHOULD assume that the counter size is 32 bits unless explicitly configured otherwise or unless (in the case of a bidirectional channel) all local and remote interfaces involved in the LM operation are known to be 64-bit-capable, which can be inferred from the value of the X flag in an LM response.

### 4.2.2. Transmitting a Loss Measurement Query

When transmitting an LM Query, the Version field MUST be set to 0. The R flag MUST be set to 0. The T flag SHALL be set to 1 if, and only if, the measurement is specific to a particular traffic class, in which case the DS field SHALL identify that traffic class.

The X flag MUST be set to 1 if the transmitting interface writes 64-bit LM counters and otherwise MUST be set to 0 to indicate that 32-bit counters are written. The B flag SHALL be set to 1 to indicate that the counter fields contain octet counts or to 0 to indicate packet counts.

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

The Session Identifier field can be set arbitrarily.

The Origin Timestamp field SHALL be set to the time at which this message is transmitted, and the Origin Timestamp Format field MUST be set to indicate its format, according to Section 3.4.

The Counter 1 field SHOULD be set to the total count of units (packets or octets, according to the B flag) transmitted over the channel prior to this LM Query, or to 0 if this is the beginning of a measurement session for which counter data is not yet available. The Counter 2 field MUST be set to 0. If a response was previously received in this measurement session, the Counter 1 and Counter 2 fields of the most recent such response MAY be copied to the Counter 3 and Counter 4 fields, respectively, of this query; otherwise, the Counter 3 and Counter 4 fields MUST be set to 0.

### 4.2.3. Receiving a Loss Measurement Query

Upon receipt of an LM Query message, the Counter 2 field SHOULD be set to the total count of units (packets or octets, according to the B flag) received over the channel prior to this LM Query. If the receiving interface writes 32-bit LM counters, the X flag MUST be set to 0.

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.

In the case of a fatal exception that prevents the requested measurement from being made, the error SHOULD be reported, via either a response, if one was requested, or else as a notification to the user.

### 4.2.4. Transmitting a Loss Measurement Response

When constructing a Response to an LM Query, the Version field MUST be set to 0. The R flag MUST be set to 1. The value of the T flag MUST be copied from the LM Query.

The X flag MUST be set to 0 if the transmitting interface writes 32-bit LM counters; otherwise, its value MUST be copied from the LM Query. The B flag MUST be copied from the LM Query.

The Session Identifier, Origin Timestamp, and Origin Timestamp Format fields 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 units transmitted over the channel 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.

### 4.2.5. Receiving a Loss Measurement Response

Upon in-band receipt of an LM Response message, the Counter 2 field is set to the total count of units received over the channel prior to this LM Response. If the receiving interface writes 32-bit LM counters, the X flag is set to 0. (Since the life of the LM message in the network has ended at this point, it is up to the receiver whether these final modifications are made to the packet. If the message is to be forwarded on for external post-processing (Section 2.9.7), then these modifications MUST be made.)

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. If the Control Code indicates an error condition, or if the response message is invalid, the LM operation MUST be terminated and an appropriate notification to the user generated.

### 4.2.6. Loss Calculation

Calculation of packet loss is carried out according to the procedures in Section 2.2. The X flag in an LM message informs the device performing the calculation whether to perform 32-bit or 64-bit arithmetic. If the flag value is equal to 1, all interfaces involved in the LM operation have written 64-bit counter values, and 64-bit arithmetic can be used. If the flag value is equal to 0, at least one interface involved in the operation has written a 32-bit counter value, and 32-bit arithmetic is carried out using the low-order 32 bits of each counter value.

Note that the semantics of the X flag allow all devices to interoperate regardless of their counter size support. Thus, an implementation MUST NOT generate an error response based on the value of this flag.

#### 4.2.7. Quality of Service

The TC field of the LSE corresponding to the channel (e.g., LSP) being measured SHOULD be set to a traffic class equal to or better than the best TC within the measurement scope to minimize the chance of out-of-order conditions.

#### 4.2.8. G-ACh Packets

By default, direct LM MUST exclude packets transmitted and received over the Generic Associated Channel (G-ACh). An implementation MAY provide the means to alter the direct LM scope to include some or all G-ACh messages. Care must be taken when altering the LM scope to ensure that both endpoints are in agreement.

#### 4.2.9. Test Messages

In the case of inferred LM, the packets counted for LM consist of test messages generated for this purpose, or of some other class of packets deemed to provide a good proxy for data packets flowing over the channel. The specification of test protocols and proxy packets is outside the scope of this document, but some guidelines are discussed below.

An identifier common to both the test or proxy messages and the LM messages may be required to make correlation possible. The combined value of the Session Identifier and DS fields SHOULD be used for this purpose when possible. That is, test messages in this case will include a 32-bit field that can carry the value of the combined Session Identifier + DS field present in LM messages. When TC-specific LM is conducted, the DS field of the LSE in the label stack of a test message corresponding to the channel (e.g., LSP) over which the message is sent MUST correspond to the DS value in the associated LM messages.

A separate test message protocol SHOULD include a timeout value in its messages that informs the responder when to discard any state associated with a specific test.

### 4.2.10. 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. Note that in the absence of ECMP, packet misordering within a traffic class is a relatively rare event.

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 channel 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.

An implementation SHOULD identify a threshold value that indicates the upper bound of lost packets measured in a single computation beyond which the interval is considered unmeasurable. This is called the "MaxLMIntervalLoss threshold". It is clear that this threshold should be no higher than the maximum number of packets (or bytes) the channel is capable of transmitting over the interval, but it may be lower. Upon encountering an unmeasurable interval, the LM state (i.e., data values from the last LM message received) SHOULD be discarded.

With regard to lost LM messages, the MaxLMInterval (see Section 2.2) indicates the maximum amount of time that can elapse before the LM state is discarded. If some messages are lost, but a message is subsequently received within MaxLMInterval, its timestamp or sequence number will quantify the loss, and it MAY still be used for measurement, although the measurement interval will in this case be longer than usual.

If an LM message is received that has a timestamp less than or equal to the timestamp of the last LM message received, this indicates that an exception has occurred, and the current interval SHOULD be considered unmeasurable unless the implementation has some other way of handling this condition.

## 4.3. Delay Measurement Procedures

### 4.3.1. Transmitting a Delay Measurement Query

When transmitting a DM Query, the Version and Reserved fields MUST be set to 0. The R flag MUST be set to 0, the T flag MUST be set to 1, 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 channel is unidirectional, this field MUST NOT be set to 0x0 (Query: In-band Response Requested).

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 3.4. The Responder Timestamp Format and Responder’s Preferred Timestamp Format fields MUST be set to 0.

The Session Identifier field can be set arbitrarily. The DS field MUST be set to the traffic class being measured.

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 Timestamp 2 field MUST be set to 0. If a response was previously received in this measurement session, the Timestamp 1 and Timestamp 2 fields of the most recent such response MAY be copied to the Timestamp 3 and Timestamp 4 fields, respectively, of this query; otherwise, the Timestamp 3 and Timestamp 4 fields MUST be set to 0.

### 4.3.2. 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 was 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.

In the case of a fatal exception that prevents the requested measurement from being made, the error SHOULD be reported, via either a response, if one was requested, or else as a notification to the user.

### 4.3.3. Transmitting a Delay Measurement Response

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

The Session Identifier 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 3.4. Furthermore, the RTF field MUST be set equal to either the QTF or the RPTF field. See Section 4.3.5 for guidelines on the 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 3.4 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.

### 4.3.4. Receiving a Delay Measurement Response

Upon in-band receipt of a DM Response message, the Timestamp 2 field is set to the time at which this DM Response was received. (Since the life of the DM message in the network has ended at this point, it is up to the receiver whether this final modification is made to the packet. If the message is to be forwarded on for external post-processing (Section 2.9.7), then these modifications MUST be made.)

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. If the Control Code indicates an error condition, or if the response message is invalid, the DM operation MUST be terminated and an appropriate notification to the user generated.

### 4.3.5. 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 channel. The procedures for making this determination SHALL be as follows.

Upon sending an initial DM Query over a channel, 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.

#### 4.3.5.1. Single-Format Procedures

When an implementation supports only one timestamp format, the procedures above reduce to the following simple behavior:

- All DM Queries are transmitted with the same QTF;
- All DM Responses are transmitted with the same RTF, and the RPTF is always set equal to the RTF;
- All DM Responses received with RTF not equal to QTF are discarded;
- On a unidirectional channel, all DM Queries received with QTF not equal to the supported format are discarded.

### 4.3.6. Quality of Service

The TC field of the LSE corresponding to the channel (e.g., LSP) being measured MUST be set to the value that corresponds to the DS field in the DM message.

## 4.4. Combined Loss/Delay Measurement Procedures

The combined LM/DM message defined in Section 3.3 allows loss and delay measurement to be carried out simultaneously. This message SHOULD be treated as an LM message that happens to carry additional timestamp data, with the timestamp fields processed as per delay measurement procedures.

# 5. Implementation Disclosure Requirements

This section summarizes the requirements placed on implementations for capabilities disclosure. The purpose of these requirements is to ensure that end users have a clear understanding of implementation capabilities and characteristics that have a direct impact on how loss and delay measurement mechanisms function in specific situations. Implementations are REQUIRED to state:

- METRICS: Which of the following metrics are supported: packet loss, packet throughput, octet loss, octet throughput, average loss rate, one-way delay, round-trip delay, two-way channel delay, packet delay variation.

- MP-LOCATION: The location of loss and delay measurement points with respect to other stages of packet processing, such as queuing.

- CHANNEL-TYPES: The types of channels for which LM and DM are supported, including LSP types, pseudowires, and sections (links).

- QUERY-RATE: The minimum supported query intervals for LM and DM sessions, both in the querier and responder roles.

- LOOP: Whether loopback measurement (Section 2.8) is supported.

- LM-TYPES: Whether direct or inferred LM is supported, and for the latter, which test protocols or proxy message types are supported.

- LM-COUNTERS: Whether 64-bit counters are supported.

- LM-ACCURACY: The expected measurement accuracy levels for the supported forms of LM, and the expected impact of exception conditions such as lost and misordered messages.

- LM-SYNC: The implementation’s behavior in regard to the synchronization conditions discussed in Section 2.9.8.

- LM-SCOPE: The supported LM scopes (Sections 2.9.9 and 4.2.8).

- DM-ACCURACY: The expected measurement accuracy levels for the supported forms of DM.

- DM-TS-FORMATS: The supported timestamp formats and the extent of support for computation with and reconciliation of different formats.

## 6. Congestion Considerations

An MPLS 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 channel naturally results in a degree of additional load placed on both the network and the terminal nodes of the channel. 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 channel 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 proactive 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.

## 7. Manageability Considerations

The measurement protocols described in this document are intended to serve as infrastructure to support a wide range of higher-level monitoring and diagnostic applications, from simple command-line diagnostic tools to comprehensive network performance monitoring and analysis packages. The specific mechanisms and considerations for protocol configuration, initialization, and reporting thus depend on the nature of the application.

In the case of on-demand diagnostics, the diagnostic application may provide parameters such as the measurement type, the channel, the query rate, and the test duration when initiating the diagnostic; results and exception conditions are then reported directly to the application. The system may discard the statistics accumulated during the test after the results have been reported or retain them to provide a historical measurement record.

Alternatively, measurement configuration may be supplied as part of the channel configuration itself in order to support continuous monitoring of the channel’s performance characteristics. In this case, the configuration will typically include quality thresholds depending on the service level agreement, the crossing of which will trigger warnings or alarms, and result reporting and exception notification will be integrated into the system-wide network management and reporting framework.

## 8. Security Considerations

This document describes procedures for the measurement of performance metrics over a pre-existing MPLS path (a pseudowire, LSP, or section). As such, it assumes that a node involved in a measurement operation has previously verified the integrity of the path and the identity of the far end using existing MPLS mechanisms such as Bidirectional Forwarding Detection (BFD) [RFC5884]; tools, techniques, and considerations for securing MPLS paths are discussed in detail in [RFC5920].

When such mechanisms are not available, and where security of the measurement operation is a concern, reception of Generic Associated Channel messages with the Channel Types specified in this document SHOULD be disabled. Implementations MUST provide the ability to disable these protocols on a per-Channel-Type basis.

Even when the identity of the far end has been verified, the measurement protocols remain vulnerable to injection and man-in-the-middle attacks. The impact of such an attack would be to compromise the quality of performance measurements on the affected path. An attacker positioned to disrupt these measurements is, however, capable of causing much greater damage by disrupting far more critical elements of the network such as the network control plane or user traffic flows. At worst, a disruption of the measurement protocols would interfere with the monitoring of the performance aspects of the service level agreement associated with the path; the existence of such a disruption would imply that a serious breach of basic path integrity had already occurred.

If desired, such attacks can be mitigated by performing basic validation and sanity checks, at the querier, of the counter or timestamp fields in received measurement response messages. The minimal state associated with these protocols also limits the extent of measurement disruption that can be caused by a corrupt or invalid message to a single query/response cycle.

Cryptographic mechanisms capable of signing or encrypting the contents of the measurement packets without degrading the measurement performance are not currently available. In light of the preceding discussion, the absence of such cryptographic mechanisms does not raise significant security issues.

Users concerned with the security of out-of-band responses over IP networks SHOULD employ suitable security mechanisms such as IPsec [RFC4301] to protect the integrity of the return path.

## 9. IANA Considerations

Per this document, IANA has completed the following actions:

- Allocation of Channel Types in the "PW Associated Channel Type" registry

- Creation of a "Measurement Timestamp Type" registry

- Creation of an "MPLS Loss/Delay Measurement Control Code" registry

- Creation of an "MPLS Loss/Delay Measurement Type-Length-Value (TLV) Object" registry

### 9.1. Allocation of PW Associated Channel Types

As per the IANA considerations in [RFC5586], IANA has allocated the following Channel Types in the "PW Associated Channel Type" registry:

| Value | Description | TLV Follows | Reference |
|---|---|---|---|
| 0x000A | MPLS Direct Loss Measurement (DLM) | No | RFC 6374 |
| 0x000B | MPLS Inferred Loss Measurement (ILM) | No | RFC 6374 |
| 0x000C | MPLS Delay Measurement (DM) | No | RFC 6374 |
| 0x000D | MPLS Direct Loss and Delay Measurement (DLM+DM) | No | RFC 6374 |
| 0x000E | MPLS Inferred Loss and Delay Measurement (ILM+DM) | No | RFC 6374 |

### 9.2. Creation of Measurement Timestamp Type Registry

IANA has created a new "Measurement Timestamp Type" registry, with format and initial allocations as follows:

| Type | Description | Size in Bits | Reference |
|---|---|---|---|
| 0 | Null Timestamp | 64 | RFC 6374 |
| 1 | Sequence Number | 64 | RFC 6374 |
| 2 | Network Time Protocol version 4 64-bit Timestamp | 64 | RFC 6374 |
| 3 | Truncated IEEE 1588v2 PTP Timestamp | 64 | RFC 6374 |

The range of the Type field is 0-15.

The allocation policy for this registry is IETF Review.

### 9.3. Creation of MPLS Loss/Delay Measurement Control Code Registry

IANA has created a new "MPLS Loss/Delay Measurement Control Code" registry. This registry is divided into two separate parts, one for Query Codes and the other for Response Codes, with formats and initial allocations as follows:

Query Codes

| Code | Description | Reference |
|---|---|---|
| 0x0 | In-band Response Requested | RFC 6374 |
| 0x1 | Out-of-band Response Requested | RFC 6374 |
| 0x2 | No Response Requested | RFC 6374 |

Response Codes

| Code | Description | Reference |
|---|---|---|
| 0x0 | Reserved | RFC 6374 |
| 0x1 | Success | RFC 6374 |
| 0x2 | Data Format Invalid | RFC 6374 |
| 0x3 | Initialization in Progress | RFC 6374 |
| 0x4 | Data Reset Occurred | RFC 6374 |
| 0x5 | Resource Temporarily Unavailable | RFC 6374 |
| 0x10 | Unspecified Error | RFC 6374 |
| 0x11 | Unsupported Version | RFC 6374 |
| 0x12 | Unsupported Control Code | RFC 6374 |
| 0x13 | Unsupported Data Format | RFC 6374 |
| 0x14 | Authentication Failure | RFC 6374 |
| 0x15 | Invalid Destination Node Identifier | RFC 6374 |
| 0x16 | Connection Mismatch | RFC 6374 |
| 0x17 | Unsupported Mandatory TLV Object | RFC 6374 |
| 0x18 | Unsupported Query Interval | RFC 6374 |
| 0x19 | Administrative Block | RFC 6374 |
| 0x1A | Resource Unavailable | RFC 6374 |
| 0x1B | Resource Released | RFC 6374 |
| 0x1C | Invalid Message | RFC 6374 |
| 0x1D | Protocol Error | RFC 6374 |

IANA has indicated that the values 0x0 - 0xF in the Response Code section are reserved for non-error response codes.

The range of the Code field is 0 - 255.

The allocation policy for this registry is IETF Review.

## 9.4. Creation of MPLS Loss/Delay Measurement TLV Object Registry

IANA has created a new "MPLS Loss/Delay Measurement TLV Object" registry, with format and initial allocations as follows:

| Type | Description | Reference |
|---|---|---|
| 0 | Padding - copy in response | RFC 6374 |
| 1 | Return Address | RFC 6374 |
| 2 | Session Query Interval | RFC 6374 |
| 3 | Loopback Request | RFC 6374 |
| 127 | Experimental use | RFC 6374 |
| 128 | Padding - do not copy in response | RFC 6374 |
| 129 | Destination Address | RFC 6374 |
| 130 | Source Address | RFC 6374 |
| 255 | Experimental use | RFC 6374 |

IANA has indicated that Types 0-127 are classified as Mandatory, and that Types 128-255 are classified as Optional.

The range of the Type field is 0 - 255.

The allocation policy for this registry is IETF Review.

## 10. Acknowledgments

The authors wish to thank the many participants of the MPLS working group who provided detailed review and feedback on this document. The authors offer special thanks to Alexander Vainshtein, Loa Andersson, and Hiroyuki Takagi for many helpful thoughts and discussions, to Linda Dunbar for the idea of using LM messages for throughput measurement, and to Ben Niven-Jenkins, Marc Lasserre, and Ben Mack-Crane for their valuable comments.

## 11. References

### 11.1. 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", BCP 14, RFC 2119, March 1997.

[RFC2474] Nichols, K., Blake, S., Baker, F., and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474, December 1998.

[RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.

[RFC3270] Le Faucheur, F., Wu, L., Davie, B., Davari, S., Vaananen, P., Krishnan, R., Cheval, P., and J. Heinanen, "Multi-Protocol Label Switching (MPLS) Support of Differentiated Services", RFC 3270, May 2002.

[RFC5462] Andersson, L. and R. Asati, "Multiprotocol Label Switching (MPLS) Label Stack Entry: "EXP" Field Renamed to "Traffic Class" Field", RFC 5462, February 2009.

[RFC5586] Bocci, M., Vigoureux, M., and S. Bryant, "MPLS Generic Associated Channel", RFC 5586, June 2009.

[RFC5905] Mills, D., Martin, J., Burbank, J., and W. Kasch, "Network Time Protocol Version 4: Protocol and Algorithms Specification", RFC 5905, June 2010.

### 11.2. Informative References

[RFC2679] Almes, G., Kalidindi, S., and M. Zekauskas, "A One-way Delay Metric for IPPM", RFC 2679, September 1999.

[RFC2680] Almes, G., Kalidindi, S., and M. Zekauskas, "A One-way Packet Loss Metric for IPPM", RFC 2680, September 1999.

[RFC2681] Almes, G., Kalidindi, S., and M. Zekauskas, "A Round-trip Delay Metric for IPPM", RFC 2681, September 1999.

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.

[RFC3260] Grossman, D., "New Terminology and Clarifications for Diffserv", RFC 3260, April 2002.

[RFC3985] Bryant, S. and P. Pate, "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, March 2005.

[RFC4301] Kent, S. and K. Seo, "Security Architecture for the Internet Protocol", RFC 4301, December 2005.

[RFC4656] Shalunov, S., Teitelbaum, B., Karp, A., Boote, J., and M. Zekauskas, "A One-way Active Measurement Protocol (OWAMP)", RFC 4656, September 2006.

[RFC4928] Swallow, G., Bryant, S., and L. Andersson, "Avoiding Equal Cost Multipath Treatment in MPLS Networks", BCP 128, RFC 4928, June 2007.

[RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.

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

[RFC5481] Morton, A. and B. Claise, "Packet Delay Variation Applicability Statement", RFC 5481, March 2009.

[RFC5884] Aggarwal, R., Kompella, K., Nadeau, T., and G. Swallow, "Bidirectional Forwarding Detection (BFD) for MPLS Label Switched Paths (LSPs)", RFC 5884, June 2010.

[RFC5920] Fang, L., "Security Framework for MPLS and GMPLS Networks", RFC 5920, July 2010.

[RFC5921] Bocci, M., Bryant, S., Frost, D., Levrau, L., and L. Berger, "A Framework for MPLS in Transport Networks", RFC 5921, July 2010.

[RFC5960] Frost, D., Bryant, S., and M. Bocci, "MPLS Transport Profile Data Plane Architecture", RFC 5960, August 2010.

[RFC6375] Frost, D., Ed. and S. Bryant, Ed., "A Packet Loss and Delay Measurement Profile for MPLS-Based Transport Networks", RFC 6375, September 2011.

[Y.1731] ITU-T Recommendation Y.1731, "OAM Functions and Mechanisms for Ethernet based Networks", February 2008.

## Appendix A. Default Timestamp Format Rationale

This document initially proposed the Network Time Protocol (NTP) timestamp format as the mandatory default, as this is the normal default timestamp in IETF protocols and thus would seem the "natural" choice. However, a number of considerations have led instead to the specification of the truncated IEEE 1588 Precision Time Protocol (PTP) timestamp as the default. NTP has not gained traction in industry as the protocol of choice for high-quality timing infrastructure, whilst IEEE 1588 PTP has become the de facto time transfer protocol in networks that are specially engineered to provide high-accuracy time distribution service. The PTP timestamp format is also the ITU-T format of choice for packet transport networks, which may rely on MPLS protocols. Applications such as one-way delay measurement need the best time service available, and converting between the NTP and PTP timestamp formats is not a trivial transformation, particularly when it is required that this be done in real time without loss of accuracy.

The truncated IEEE 1588 PTP format specified in this document is considered to provide a more than adequate wrap time and greater time resolution than it is expected will be needed for the operational lifetime of this protocol. By truncating the timestamp at both the high and low order bits, the protocol achieves a worthwhile reduction in system resources.

## Authors' Addresses

Dan Frost
Cisco Systems

EMail: danfrost@cisco.com

Stewart Bryant
Cisco Systems

EMail: stbryant@cisco.com