MPLS Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 8, 2017

G. Mirsky
ZTE Corp.
S. Ruffini
E. Gray
Ericsson
J. Drake
Juniper Networks
S. Bryant
Huawei
A. Vainshtein
ECI Telecom
March 7, 2017

Residence Time Measurement in MPLS network
draft-ietf-mpls-residence-time-15

Abstract

This document specifies a new Generic Associated Channel for Residence Time Measurement and describes how it can be used by time synchronization protocols within a MPLS domain.

Residence time is the variable part of the propagation delay of timing and synchronization messages; knowing what this delay is for each message allows for a more accurate determination of the delay to be taken into account in applying the value included in a Precision Time Protocol event message.

Status of This Memo

This Internet-Draft will expire on September 8, 2017.

Copyright Notice

Table of Contents

1. Introduction

Time synchronization protocols, e.g., Network Time Protocol version 4 (NTPv4) [RFC5905] and Precision Time Protocol (PTP) Version 2 [IEEE.1588.2008], define timing messages that can be used to synchronize clocks across a network domain. Measurement of the cumulative time that one of these timing messages spends transiting the nodes on the path from ingress node to egress node is termed Residence Time and it is used to improve the accuracy of clock synchronization. Residence Time is the sum of the difference between the time of receipt at an ingress interface and the time of transmission from an egress interface for each node along the network path from an ingress node to an egress node. This document defines a new Generic Associated Channel (G-ACh) value and an associated residence time measurement (RTM) message that can be used in a Multi-Protocol Label Switching (MPLS) network to measure residence time over a Label Switched Path (LSP).

This document describes RTM over an LSP signaled using RSVP-TE [RFC3209]. Using RSVP-TE, the LSP's path can be either explicitly specified or determined during signaling. Although it is possible to use RTM over an LSP instantiated using Label Distribution Protocol [RFC5036], that is outside the scope of this document.

Comparison with alternative proposed solutions such as [I-D.ietf-tictoc-1588overmpls] is outside the scope of this document.

1.1. Conventions used in this document

1.1.1. Terminology

MPLS: Multi-Protocol Label Switching

ACH: Associated Channel

TTL: Time-to-Live

G-ACh: Generic Associated Channel

GAL: Generic Associated Channel Label

NTP: Network Time Protocol

ppm: parts per million

PTP: Precision Time Protocol

BC: Boundary Clock

LSP: Label Switched Path

OAM: Operations, Administration, and Maintenance

RRO: Record Route Object

RTM: Residence Time Measurement

IGP: Internal Gateway Protocol

BGP-LS: Border Gateway Protocol - Link State

1.1.2. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

2. Residence Time Measurement

Packet Loss and Delay Measurement for MPLS Networks [RFC6374] can be used to measure one-way or two-way end-to-end propagation delay over LSP or PW. But these measurements are insufficient for use in some applications, for example, time synchronization across a network as defined in the PTP. In PTPv2 [IEEE.1588.2008], the residence time is accumulated in the correctionField of the PTP event message, as defined in [IEEE.1588.2008] and referred to as using a one-step clock, or in the associated follow-up message (or Delay_Resp message associated with the Delay_Req message), referred to as using a two-step clock (see the detailed discussion in Section 2.1).

IEEE 1588 uses this residence time to correct for the transit times of nodes on an LSP, effectively making the transit nodes transparent.

This document proposes a mechanism that can be used as one type of on-path support for a clock synchronization protocol or to perform one-way measurement of residence time. The proposed mechanism accumulates residence time from all nodes that support this extension along the path of a particular LSP in the Scratch Pad field of an RTM

message (Figure 1). This value can then be used by the egress node to update, for example, the correctionField of the PTP event packet carried within the RTM message prior to performing its PTP processing.

2.1. One-step Clock and Two-step Clock Modes

One-step mode refers to the mode of operation where an egress interface updates the correctionField value of an original event message. Two-step mode refers to the mode of operation where this update is made in a subsequent follow-up message.

Processing of the follow-up message, if present, requires the downstream end-point to wait for the arrival of the follow-up message in order to combine correctionField values from both the original (event) message and the subsequent (follow-up) message. In a similar fashion, each two-step node needs to wait for the related follow-up message, if there is one, in order to update that follow-up message (as opposed to creating a new one). Hence the first node that uses two-step mode MUST do two things:

1. Mark the original event message to indicate that a follow-up message will be forthcoming. This is necessary in order to

    Let any subsequent two-step node know that there is already a follow-up message, and

    Let the end-point know to wait for a follow-up message;

2. Create a follow-up message in which to put the RTM determined as an initial correctionField value.

IEEE 1588v2 [IEEE.1588.2008] defines this behavior for PTP messages.

Thus, for example, with reference to the PTP protocol, the PTPType field identifies whether the message is a Sync message, Follow_up message, Delay_Req message, or Delay_Resp message. The 10 octet long Port ID field contains the identity of the source port [IEEE.1588.2008], that is, the specific PTP port of the boundary clock connected to the MPLS network. The Sequence ID is the sequence ID of the PTP message carried in the Value field of the message.

PTP messages also include a bit that indicates whether or not a follow-up message will be coming. This bit MAY be set by a two-step mode PTP device. The value MUST NOT be unset until the original and follow-up messages are combined by an end-point (such as a Boundary Clock).

For compatibility with PTP, RTM (when used for PTP packets) must behave in a similar fashion. It should be noted that the handling of Sync event messages and of Delay_Req/Delay_Resp event messages that cross a two-step RTM node is different. The following outlines the handling of PTP Sync event message by the two-step RTM node. The details of handling Delay_Resp/Delay_Req PTP event messages by the two-step RTM node are discussed in Section 2.1.1. As a summary, a two-step RTM capable egress interface will need to examine the S-bit in the Flags field of the PTP sub-TLV (for RTM messages that indicate they are for PTP) and - if it is clear (set to zero), it MUST set the S bit and create a follow-up PTP Type RTM message. If the S bit is already set, then the RTM capable node MUST wait for the RTM message with the PTP type of follow-up and matching originator and sequence number to make the corresponding residence time update to the Scratch Pad field. The wait period MUST be reasonably bounded.

Thus, an RTM packet, containing residence time information relating to an earlier packet, also contains information identifying that earlier packet.

In practice, an RTM node operating in two-step mode behaves like a two-steps transparent clock.

A one-step capable RTM node MAY elect to operate in either one-step mode (by making an update to the Scratch Pad field of the RTM message containing the PTP event message), or in two-step mode (by making an update to the Scratch Pad of a follow-up message when presence of a follow-up is indicated), but MUST NOT do both.

Two main subcases identified for an RTM node operating as a two-step clock are described in the following sub-sections.

2.1.1. RTM with Two-step Upstream PTP Clock

If any of the previous RTM capable nodes or the previous PTP clock (e.g., the Boundary Clock (BC) connected to the first node), is a two-step clock, the residence time is added to the RTM packet that has been created to include the second PTP packet (i.e., follow-up message in the downstream direction), if the local RTM-capable node is also operating as a two-step clock. This RTM packet carries the related accumulated residence time and the appropriate values of the Sequence ID and Port ID (the same identifiers carried in the original packet) and the Two-step Flag set to 1.

Note that the fact that an upstream RTM-capable node operating in the two-step mode has created a follow-up message does not require any subsequent RTM capable node to also operate in the two-step mode, as

long as that RTM-capable node forwards the follow-up message on the same LSP on which it forwards the corresponding previous message.

A one-step capable RTM node MAY elect to update the RTM follow-up message as if it were operating in two-step mode, however, it MUST NOT update both messages.

A PTP Sync packet is carried in the RTM packet in order to indicate to the RTM node that residence time measurement must be performed on that specific packet.

To handle the residence time of the Delay_Req message on the upstream direction, an RTM packet must be created to carry the residence time on the associated downstream Delay_Resp message.

The last RTM node of the MPLS network, in addition to updating the correctionField of the associated PTP packet, must also react properly to the two-step flag of the PTP packets.

2.1.2. Two-step RTM with One-step Upstream PTP Clock

When the PTP network connected to the MPLS operates in one-step clock mode and an RTM node operates in two-step mode, the follow-up RTM packet must be created by the RTM node itself. The RTM packet carrying the PTP event packet needs now to indicate that a follow-up message will be coming.

The egress RTM-capable node of the LSP will be removing RTM encapsulation and, in case of two-step clock mode being indicated, will generate PTP messages to include the follow-up correction as appropriate (according to the [IEEE.1588.2008]). In this case, the common header of the PTP packet carrying the synchronization message would have to be modified by setting the twoStepFlag field indicating that there is now a follow up message associated to the current message.

3. G-ACh for Residence Time Measurement

RFC 5586 [RFC5586] and RFC 6423 [RFC6423] define the G-ACh to extend the applicability of the Pseudowire Associated Channel (ACH) [RFC5085] to LSPs. G-ACh provides a mechanism to transport OAM and other control messages over an LSP. Processing of these messages by selected transit nodes is controlled by the use of the Time-to-Live (TTL) value in the MPLS header of these messages.

The message format for Residence Time Measurement (RTM) is presented in Figure 1

```
     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
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |0 0 0 1|Version|   Reserved    |           RTM G-ACh           |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                                                               |
    |                        Scratch Pad                            |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |            Type               |             Length            |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                       Sub-TLV (optional)                      |
    ~                                                               ~
    |                                                               |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 1: RTM G-ACh message format for Residence Time Measurement

- First four octets are defined as G-ACh Header in [RFC5586]

- The Version field is set to 0, as defined in RFC 4385 [RFC4385].

- The Reserved field MUST be set to 0 on transmit and ignored on receipt.

- The RTM G-ACh field, value (TBA1) to be allocated by IANA, identifies the packet as such.

- The Scratch Pad field is 8 octets in length. It is used to accumulate the residence time spent in each RTM capable node transited by the packet on its path from ingress node to egress node. The first RTM-capable node MUST initialize the Scratch Pad field with its residence time measurement. Its format is IEEE double precision and its units are nanoseconds. Note that depending on whether the timing procedure is one-step or two-step operation (Section 2.1), the residence time is either for the timing packet carried in the Value field of this RTM message or for an associated timing packet carried in the Value field of another RTM message.

- The Type field identifies the type and encapsulation of a timing packet carried in the Value field, e.g., NTP [RFC5905] or PTP [IEEE.1588.2008]. This document asks IANA to create a sub-registry in Generic Associated Channel (G-ACh) Parameters Registry called "MPLS RTM TLV Registry" Section 7.2.

- The Length field contains the length, in octets, of the of the timing packet carried in the Value field.

- o The optional Value field MAY carry a packet of the time synchronization protocol identified by Type field. It is important to note that the packet may be authenticated or encrypted and carried over LSP edge to edge unchanged while the residence time is accumulated in the Scratch Pad field.

- o The TLV MUST be included in the RTM message, even if the length of the Value field is zero.

3.1. PTP Packet Sub-TLV

Figure 2 presents the format of a PTP sub-TLV that MUST be included in the Value field of an RTM message preceding the carried timing packet when the timing packet is PTP.

```
     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            |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                         Flags                         |PTPType|
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                            Port ID                            |
    |                                                               |
    |                               +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                               |           Sequence ID         |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2: PTP Sub-TLV format

where Flags field has format

```
     0                   1                   2
     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
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |S|                      Reserved                       |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 3: Flags field format of PTP Packet Sub-TLV

- o The Type field identifies PTP packet sub-TLV and is set to 1 according to Section 7.3.

- o The Length field of the PTP sub-TLV contains the number of octets of the Value field and MUST be 20.

- o The Flags field currently defines one bit, the S-bit, that defines whether the current message has been processed by a two-step node,

where the flag is cleared if the message has been handled exclusively by one-step nodes and there is no follow-up message, and is set if there has been at least one two-step node and a follow-up message is forthcoming.

- The PTPType field indicates the type of PTP packet carried in the TLV. PTPType is the messageType field of the PTPv2 packet whose values are defined in Table 19 of [IEEE.1588.2008].

- The 10 octet long Port ID field contains the identity of the source port.

- The Sequence ID is the sequence ID of the PTP message carried in the Value field of the message.

Tuple of PTPType, Port ID, and Sequence ID uniquely identifies PTP control packet encapsulated in RTM message and are used in two-step RTM mode Section 2.1.1.

4. Control Plane Theory of Operation

The operation of RTM depends upon TTL expiry to deliver an RTM packet from one RTM capable interface to the next along the path from ingress node to egress node. This means that a node with RTM capable interfaces MUST be able to compute a TTL which will cause the expiry of an RTM packet at the next node with RTM capable interfaces.

4.1. RTM Capability

Note that the RTM capability of a node is with respect to the pair of interfaces that will be used to forward an RTM packet. In general, the ingress interface of this pair must be able to capture the arrival time of the packet and encode it in some way such that this information will be available to the egress interface of a node.

The supported mode (one-step or two-step) of any pair of interfaces is determined by the capability of the egress interface. For both modes, the egress interface implementation MUST be able to determine the precise departure time of the same packet and determine from this, and the arrival time information from the corresponding ingress interface, the difference representing the residence time for the packet.

An interface with the ability to do this and update the associated Scratch Pad in real-time (i.e., while the packet is being forwarded) is said to be one-step capable.

Hence while both ingress and egress interfaces are required to support RTM for the pair to be RTM-capable, it is the egress interface that determines whether or not the node is one-step or two-step capable with respect to the interface-pair.

The RTM capability used in the sub-TLV shown in Figure 4 and Figure 5 is thus a non-routing related capability associated with the interface being advertised based on its egress capability. The ability of any pair of interfaces on a node that includes this egress interface to support any mode of RTM depends on the ability of the ingress interface of a node to record packet arrival time and convey it to the egress interface on the node.

When a node uses an IGP to support the RTM capability advertisement, the IGP sub-TLV MUST reflect the RTM capability (one-step or two-step) associated with the advertised interface. Changes of RTM capability are unlikely to be frequent and would result, for example, from operator's decision to include or exclude a particular port from RTM processing or switch between RTM modes.

4.2. RTM Capability Sub-TLV

[RFC4202] explains that the Interface Switching Capability Descriptor describes the switching capability of an interface. For bi-directional links, the switching capabilities of an interface are defined to be the same in either direction. I.e., for data entering the node through that interface and for data leaving the node through that interface. That principle SHOULD be applied when a node advertises RTM Capability.

A node that supports RTM MUST be able to act in two-step mode and MAY also support one-step RTM mode. Detailed discussion of one-step and two-step RTM modes is contained in Section 2.1.

4.3. RTM Capability Advertisement in Routing Protocols

4.3.1. RTM Capability Advertisement in OSPFv2

The format for the RTM Capability sub-TLV in OSPF is presented in Figure 4

```
     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            |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    | RTM |  Value       ...
    +-+-+-+-+-+-+-+-+-+-+- ...
```

Figure 4: RTM Capability sub-TLV in OSPFv2

- Type value (TBA2) will be assigned by IANA from appropriate registry for OSPFv2 Section 7.4.

- Length value equals number of octets of the Value field.

- Value contains variable number of bit-map fields so that overall number of bits in the fields equals Length * 8.

- Bits are defined/sent starting with Bit 0. Additional bit-map field definitions that may be defined in the future SHOULD be assigned in ascending bit order so as to minimize the number of bits that will need to be transmitted.

- Undefined bits MUST be transmitted as 0 and MUST be ignored on receipt.

- Bits that are NOT transmitted MUST be treated as if they are set to 0 on receipt.

- RTM (capability) - is a three-bit long bit-map field with values defined as follows:

  - 0b001 - one-step RTM supported;

  - 0b010 - two-step RTM supported;

  - 0b100 - reserved.

The capability to support RTM on a particular link (interface) is advertised in the OSPFv2 Extended Link Opaque LSA described in Section 3 [RFC7684] via the RTM Capability sub-TLV.

Its Type value will be assigned by IANA from the OSPF Extended Link TLV Sub-TLVs registry Section 7.4, that will be created per [RFC7684] request.

4.3.2. RTM Capability Advertisement in OSPFv3

The capability to support RTM on a particular link (interface) can be advertised in OSPFv3 using LSA extensions as described in [I-D.ietf-ospf-ospfv3-lsa-extend]. The sub-TLV SHOULD use the same format as in Section 4.3.1. The type allocation and full details of exact use of OSPFv3 LSA extensions is for further study.

4.3.3. RTM Capability Advertisement in IS-IS

The capability to support RTM on a particular link (interface) is advertised in a new sub-TLV which may be included in TLVs advertising Intermediate System (IS) Reachability on a specific link (TLVs 22, 23, 222, and 223).

The format for the RTM Capabilities sub-TLV is presented in Figure 5

```
     0                   1                   2
     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 ...
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+...
    |      Type     |     Length    | RTM |   Value      ...
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+...
```

Figure 5: RTM Capability sub-TLV

- o Type value (TBA3) will be assigned by IANA from the Sub-TLVs for TLVs 22, 23, 141, 222, and 223 registry for IS-IS Section 7.5.
- o Definitions, rules of handling, and values for fields Length and Value are as defined in Section 4.3.1
- o RTM (capability) - is a three-bit long bit-map field with values defined in Section 4.3.1.

4.3.4. RTM Capability Advertisement in BGP-LS

The format for the RTM Capabilities TLV is as presented in Figure 4.

Type value TBA9 will be assigned by IANA from the BGP-LS Node Descriptor, Link Descriptor, Prefix Descriptor, and Attribute TLVs sub-registry Section 7.6.

Definitions, rules of handling, and values for fields Length, Value, and RTM are as defined in Section 4.3.1.

The RTM Capability will be advertised in BGP-LS as a Link Attribute TLV associated with the Link NLRI as described in section 3.3.2 of [RFC7752].

4.4. RSVP-TE Control Plane Operation to Support RTM

Throughout this document we refer to a node as RTM capable node when at least one of its interfaces is RTM capable. Figure 6 provides an example of roles a node may have with respect to RTM capability:

```
 -----     -----     -----     -----     -----     -----     -----
 | A |-----| B |-----| C |-----| D |-----| E |-----| F |-----| G |
 -----     -----     -----     -----     -----     -----     -----
```

Figure 6: RTM capable roles

- A is a BC with its egress port in Master state. Node A transmits IP encapsulated timing packets whose destination IP address is G.

- B is the ingress LER for the MPLS LSP and is the first RTM capable node. It creates RTM packets and in each it places a timing packet, possibly encrypted, in the Value field and initializes the Scratch Pad field with its residence time measurement

- C is a transit node that is not RTM capable. It forwards RTM packets without modification.

- D is RTM capable transit node. It updates the Scratch Pad field of the RTM packet without updating the timing packet.

- E is a transit node that is not RTM capable. It forwards RTM packets without modification.

- F is the egress LER and the last RTM capable node. It removes the RTM ACH encapsulation and processes the timing packet carried in the Value field using the value in the Scratch Pad field. In particular, the value in the Scratch Pad field of the RTM ACH is used in updating the Correction field of the PTP message(s). The LER should also include its own residence time before creating the outgoing PTP packets. The details of this process depend on whether or not the node F is itself operating as one-step or two-step clock.

- G is a Boundary Clock with its ingress port in Slave state. Node G receives PTP messages.

An ingress node that is configured to perform RTM along a path through an MPLS network to an egress node MUST verify that the selected egress node has an interface that supports RTM via the egress node's advertisement of the RTM Capability sub-TLV, as covered in Section 4.3. In the Path message that the ingress node uses to instantiate the LSP to that egress node, it places an LSP_ATTRIBUTES

Object [RFC5420] with RTM_SET Attribute Flag set, as described in Section 7.8, which indicates to the egress node that RTM is requested for this LSP. The RTM_SET Attribute Flag SHOULD NOT be set in the LSP_REQUIRED_ATTRIBUTES object [RFC5420], unless it is known that all nodes recognize the RTM attribute (but need not necessarily implement it), because a node that does not recognize the RTM_SET Attribute Flag would reject the Path message.

If an egress node receives a Path message with the RTM_SET Attribute Flag in LSP_ATTRIBUTES object, the egress node MUST include an initialized RRO [RFC3209] and LSP_ATTRIBUTES object where the RTM_SET Attribute Flag is set and the RTM_SET TLV Section 4.4.1 is initialized. When the Resv message is received by the ingress node, the RTM_SET TLV will contain an ordered list, from egress node to ingress node, of the RTM capable nodes along the LSP's path.

After the ingress node receives the Resv, it MAY begin sending RTM packets on the LSP's path. Each RTM packet has its Scratch Pad field initialized and its TTL set to expire on the closest downstream RTM capable node.

It should be noted that RTM can also be used for LSPs instantiated using [RFC3209] in an environment in which all interfaces in an IGP support RTM. In this case the RTM_SET TLV and LSP_ATTRIBUTES Object MAY be omitted.

4.4.1. RTM_SET TLV

RTM capable interfaces can be recorded via RTM_SET TLV. The RTM_SET sub-object format is of generic Type, Length, Value (TLV), presented in Figure 7 .

```
  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    |I|        Reserved             |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 ~                             Value                             ~
 |                                                               |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 7: RTM_SET TLV format

The type value (TBA4) will be assigned by IANA from its RSVP-TE Attributes TLV Space sub-registry Section 7.7.

The Length contains the total length of the sub-object in bytes, including the Type and Length fields.

The I bit flag indicates whether the downstream RTM capable node along the LSP is present in the RRO.

The Reserved field must be zeroed on initiation and ignored on receipt.

The content of an RTM_SET TLV is a series of variable-length sub-TLVs. Only a single RTM_SET can be present in a given LSP_ATTRIBUTES object. The sub-TLVs are defined in Section 4.4.1.1 below.

The following processing procedures apply to every RTM capable node along the LSP. In this paragraph, an RTM capable node is referred to as a node for sake of brevity. Each node MUST examine the Resv message for whether the RTM_SET Attribute Flag in the LSP_ATTRIBUTES object is set. If the RTM_SET flag is set, the node MUST inspect the LSP_ATTRIBUTES object for presence of an RTM_SET TLV. If more than one is found, then the LSP setup MUST fail with generation of the ResvErr message with Error Code Duplicate TLV (Section 7.9) and Error Value that contains Type value in its 8 least significant bits. If no RTM_SET TLV is found, then the LSP setup MUST fail with generation of the ResvErr message with Error Code RTM_SET TLV Absent Section 7.9. If one RTM_SET TLV has been found, the node will use the ID of the first node in the RTM_SET in conjunction with the RRO to compute the hop count to its downstream node with reachable RTM capable interface. If the node cannot find a matching ID in the RRO, then it MUST try to use the ID of the next node in the RTM_SET until it finds the match or reaches the end of the RTM_SET TLV. If a match has been found, the calculated value is used by the node as the TTL value in the outgoing label to reach the next RTM capable node on the LSP. Otherwise, the TTL value MUST be set to 255. The node MUST add an RTM_SET sub-TLV with the same address it used in the RRO sub-object at the beginning of the RTM_SET TLV in the associated outgoing Resv message before forwarding it upstream. If the calculated TTL value has been set to 255, as described above, then the I flag in the node's RTM_SET TLV MUST be set to 1 before the Resv message is forwarded upstream. Otherwise, the I flag MUST be cleared (0).

The ingress node MAY inspect the I bit flag received in each RTM_SET TLV contained in the LSP_ATTRIBUTES object of a received Resv message. The presence of the RTM_SET TLV with the I bit field set to 1 indicates that some RTM nodes along the LSP could not be included in the calculation of the residence time. An ingress node MAY choose to resignal the LSP to include all RTM nodes or simply notify the user via a management interface.

There are scenarios when some information is removed from an RRO due to policy processing (e.g., as may happen between providers) or the RRO is limited due to size constraints. Such changes affect the core

assumption of this method and the processing of RTM packets. RTM SHOULD NOT be used if it is not guaranteed that the RRO contains complete information.

4.4.1.1. RTM_SET Sub-TLVs

The RTM Set sub-object contains an ordered list, from egress node to ingress node, of the RTM capable nodes along the LSP's path.

The contents of a RTM_SET sub-object are a series of variable-length sub-TLVs. Each sub-TLV has its own Length field. The Length contains the total length of the sub-TLV in bytes, including the Type and Length fields. The Length MUST always be a multiple of 4, and at least 8 (smallest IPv4 sub-object).

Sub-TLVs are organized as a last-in-first-out stack. The first-out sub-TLV relative to the beginning of RTM_SET TLV is considered the top. The last-out sub-TLV is considered the bottom. When a new sub-TLV is added, it is always added to the top.

The RTM_SET TLV is intended to include the subset of the RRO sub-TLVs that represents those egress interfaces on the LSP that are RTM-capable. After a node chooses an egress interface to use in the RRO sub-TLV, that same egress interface, if RTM-capable, SHOULD be placed into the RTM_SET TLV using one of the IPv4 sub-TLV, IPv6 sub-TLV, or Unnumbered Interface sub-TLV. The address family chosen SHOULD match that of the RESV message and that used in the RRO; the unnumbered interface sub-TLV is used when the egress interface has no assigned IP address. A node MUST NOT place more sub-TLVs in the RTM_SET TLV than the number of RTM-capable egress interfaces the LSP traverses that are under that node's control. Only a single RTM_SET sub-TLV with the given Value field MUST be present in the RTM_SET TLV. If more than one sub-TLV with the same value (e.g., a duplicated address) is found the LSP setup MUST fail with the generation of a ResvErr message with the Error Code "Duplicate sub-TLV" Section 7.9 and Error Value contains 16-bit value composed of (Type of TLV, Type of sub-TLV).

Three kinds of sub-TLVs for RTM_SET are currently defined.

4.4.1.1.1. IPv4 Sub-TLV

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Type       |     Length    |             Reserved          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       IPv4 address                            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 8: IPv4 sub-TLV format

Type

0x01 IPv4 address

Length

The Length contains the total length of the sub-TLV in bytes, including the Type and Length fields. The Length is always 8.

IPv4 address

A 32-bit unicast host address.

Reserved

Zeroed on initiation and ignored on receipt.

4.4.1.1.2. IPv6 Sub-TLV

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Type       |     Length    |             Reserved          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
|                         IPv6 address                          |
|                                                               |
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 9: IPv6 sub-TLV format

Type

0x02 IPv6 address

Length

The Length contains the total length of the sub-TLV in bytes, including the Type and Length fields. The Length is always 20.

IPv6 address

A 128-bit unicast host address.

Reserved

Zeroed on initiation and ignored on receipt.

4.4.1.1.3. **Unnumbered Interface Sub-TLV**

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Type       |     Length    |            Reserved           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                           Node ID                             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        Interface ID                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 10: IPv4 sub-TLV format

Type

0x03 Unnumbered interface

Length

The Length contains the total length of the sub-TLV in bytes, including the Type and Length fields. The Length is always 12.

Node ID

The Node ID interpreted as Router ID as discussed in Section 2 [RFC3477].

Interface ID

The identifier assigned to the link by the node specified by the Node ID.

Reserved

Zeroed on initiation and ignored on receipt.

5. Data Plane Theory of Operation

After instantiating an LSP for a path using RSVP-TE [RFC3209] as described in Section 4.4, the ingress node MAY begin sending RTM packets to the first downstream RTM capable node on that path. Each RTM packet has its Scratch Pad field initialized and its TTL set to expire on the next downstream RTM-capable node. Each RTM-capable node on the explicit path receives an RTM packet and records the time at which it receives that packet at its ingress interface as well as the time at which it transmits that packet from its egress interface. These actions should be done as close to the physical layer as possible at the same point of packet processing striving to avoid introducing the appearance of jitter in propagation delay whereas it should be accounted as residence time. The RTM-capable node determines the difference between those two times; for one-step operation, this difference is determined just prior to or while sending the packet, and the RTM-capable egress interface adds it to the value in the Scratch Pad field of the message in progress. Note, for the purpose of calculating a residence time, a common free running clock synchronizing all the involved interfaces may be sufficient, as, for example, 4.6 ppm accuracy leads to 4.6 nanosecond error for residence time on the order of 1 millisecond. This may be acceptable for applications where the target accuracy is in the order of hundreds of ns. As an example, several applications being considered in the area of wireless applications are satisfied with an accuracy of 1.5 microseconds [ITU-T.G.8271].

For two-step operation, the difference between packet arrival time (at an ingress interface) and subsequent departure time (from an egress interface) is determined at some later time prior to sending a subsequent follow-up message, so that this value can be used to update the correctionField in the follow-up message.

See Section 2.1 for further details on the difference between one-step and two-step operation.

The last RTM-capable node on the LSP MAY then use the value in the Scratch Pad field to perform time correction, if there is no follow-up message. For example, the egress node may be a PTP Boundary Clock synchronized to a Master Clock and will use the value in the Scratch Pad field to update PTP's correctionField.

6. Applicable PTP Scenarios

This approach can be directly integrated in a PTP network based on the IEEE 1588 delay request-response mechanism. The RTM capable nodes act as end-to-end transparent clocks, and typically boundary clocks, at the edges of the MPLS network, use the value in the

Scratch Pad field to update the correctionField of the corresponding PTP event packet prior to performing the usual PTP processing.

7. IANA Considerations

7.1. New RTM G-ACh

IANA is requested to reserve a new G-ACh as follows:

Value	Description	Reference
TBA1	Residence Time Measurement	This document

Table 1: New Residence Time Measurement

7.2. New RTM TLV Registry

IANA is requested to create a sub-registry in the Generic Associated Channel (G-ACh) Parameters Registry called "MPLS RTM TLV Registry". All code points in the range 0 through 127 in this registry shall be allocated according to the "IETF Review" procedure as specified in [RFC5226]. Code points in the range 128 through 191 in this registry shall be allocated according to the "First Come First Served" procedure as specified in [RFC5226]. This document defines the following new values RTM TLV types:

Value	Description	Reference
0	Reserved	This document
1	No payload	This document
2	PTPv2, Ethernet encapsulation	This document
3	PTPv2, IPv4 Encapsulation	This document
4	PTPv2, IPv6 Encapsulation	This document
5	NTP	This document
6-127	Unassigned	
128 - 191	Unassigned	
192 - 254	Private Use	This document
255	Reserved	This document

Table 2: RTM TLV Type

7.3. New RTM Sub-TLV Registry

IANA is requested to create a sub-registry in the MPLS RTM TLV Registry, requested in Section 7.2, called "MPLS RTM Sub-TLV Registry". All code points in the range 0 through 127 in this registry shall be allocated according to the "IETF Review" procedure as specified in [RFC5226]. Code points in the range 128 through 191 in this registry shall be allocated according to the "First Come First Served" procedure as specified in [RFC5226]. This document defines the following new values RTM sub-TLV types:

Value	Description	Reference
0	Reserved	This document
1	PTP	This document
2-127	Unassigned	
128 - 191	Unassigned	
192 - 254	Private Use	This document
255	Reserved	This document

Table 3: RTM Sub-TLV Type

7.4. RTM Capability sub-TLV in OSPFv2

IANA is requested to assign a new type for RTM Capability sub-TLV from the OSPFv2 Extended Link TLV Sub-TLVs registry as follows:

Value	Description	Reference
TBA2	RTM Capability	This document

Table 4: RTM Capability sub-TLV

7.5. IS-IS RTM Capability sub-TLV

IANA is requested to assign a new Type for the RTM Capability sub-TLV from the Sub-TLVs for TLVs 22, 23, 141, 222, and 223 registry as follows:

Type	Description	22	23	141	222	223	Reference
TBA3	RTM Capability	y	y	n	y	y	This document

Table 5: IS-IS RTM Capability sub-TLV Registry Description

7.6. RTM Capability TLV in BGP-LS

IANA is requested to assign a new code point for the RTM Capability TLV from the BGP-LS Node Descriptor, Link Descriptor, Prefix Descriptor, and Attribute TLVs sub-registry in its Border Gateway Protocol - Link State (BGP-LS) Parameters registry as follows:

TLV Code Point	Description	IS-IS TLV/Sub-TLV	Reference
TBA9	RTM Capability	22/TBA3	This document

Table 6: RTM Capability TLV in BGP-LS

7.7. RTM_SET Sub-object RSVP Type and sub-TLVs

IANA is requested to assign a new Type for the RTM_SET sub-object from the RSVP-TE Attributes TLV Space sub-registry as follows:

Type	Name	Allowed on LSP_ATTRIBUTES	Allowed on LSP_REQUIRED_ATTRIBUTES	Allowed on LSP Hop Attributes	Reference
TBA4	RTM_SET sub-object	Yes	No	No	This document

Table 7: RTM_SET Sub-object Type

IANA requested to create a new sub-registry for sub-TLV types of the RTM_SET sub-object. All code points in the range 0 through 127 in this registry shall be allocated according to the "IETF Review" procedure as specified in [RFC5226]. Code points in the range 128 through 191 in this registry shall be allocated according to the "First Come First Served" procedure as specified in [RFC5226]. This

document defines the following new values of RTM_SET object sub-object types:

Value	Description	Reference
0	Reserved	This document
1	IPv4 address	This document
2	IPv6 address	This document
3	Unnumbered interface	This document
4-127	Unassigned	
128 - 191	Unassigned	
192 - 254	Private Use	This document
255	Reserved	This document

Table 8: RTM_SET object sub-object types

7.8. RTM_SET Attribute Flag

IANA is requested to assign new flag from the RSVP-TE Attribute Flags registry

Bit No	Name	Attribute Flags Path	Attribute Flags Resv	RRO	ERO	Reference
TBA5	RTM_SET	Yes	Yes	No	No	This document

Table 9: RTM_SET Attribute Flag

7.9. New Error Codes

IANA is requested to assign new Error Codes from RSVP Error Codes and Globally-Defined Error Value Sub-Codes registry

Error Code	Meaning	Reference
TBA6	Duplicate TLV	This document
TBA7	Duplicate sub-TLV	This document
TBA8	RTM_SET TLV Absent	This document

Table 10: New Error Codes

8. Security Considerations

Routers that support Residence Time Measurement are subject to the same security considerations as defined in [RFC4385] and [RFC5085] .

In addition - particularly as applied to use related to PTP - there is a presumed trust model that depends on the existence of a trusted relationship of at least all PTP-aware nodes on the path traversed by PTP messages. This is necessary as these nodes are expected to correctly modify specific content of the data in PTP messages and proper operation of the protocol depends on this ability. In practice, this means that those portions of messages cannot be covered by either confidentiality or integrity protection. Though there are methods that make it possible in theory to provide either or both such protections and still allow for intermediate nodes to make detectable but authenticated modifications, such methods do not seem practical at present, particularly for timing protocols that are sensitive to latency and/or jitter.

The ability to potentially authenticate and/or encrypt RTM and PTP data for scenarios both with and without participation of intermediate RTM/PTP-capable nodes is left for further study.

While it is possible for a supposed compromised node to intercept and modify the G-ACh content, this is an issue that exists for nodes in general - for any and all data that may be carried over an LSP - and is therefore the basis for an additional presumed trust model associated with existing LSPs and nodes.

Security requirements of time protocols are provided in RFC 7384 [RFC7384].

9. Acknowledgments

Authors want to thank Loa Andersson, Lou Berger, Acee Lindem, Les Ginsberg, and Uma Chunduri for their thorough reviews, thoughtful comments and, most of all, patience.

10. References

10.1. Normative References

[IEEE.1588.2008]
"Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems", IEEE Standard 1588, July 2008.

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

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, DOI 10.17487/RFC3209, December 2001, <http://www.rfc-editor.org/info/rfc3209>.

[RFC3477] Kompella, K. and Y. Rekhter, "Signalling Unnumbered Links in Resource ReSerVation Protocol - Traffic Engineering (RSVP-TE)", RFC 3477, DOI 10.17487/RFC3477, January 2003, <http://www.rfc-editor.org/info/rfc3477>.

[RFC4385] Bryant, S., Swallow, G., Martini, L., and D. McPherson, "Pseudowire Emulation Edge-to-Edge (PWE3) Control Word for Use over an MPLS PSN", RFC 4385, DOI 10.17487/RFC4385, February 2006, <http://www.rfc-editor.org/info/rfc4385>.

[RFC5085] Nadeau, T., Ed. and C. Pignataro, Ed., "Pseudowire Virtual Circuit Connectivity Verification (VCCV): A Control Channel for Pseudowires", RFC 5085, DOI 10.17487/RFC5085, December 2007, <http://www.rfc-editor.org/info/rfc5085>.

[RFC5420] Farrel, A., Ed., Papadimitriou, D., Vasseur, JP., and A. Ayyangarps, "Encoding of Attributes for MPLS LSP Establishment Using Resource Reservation Protocol Traffic Engineering (RSVP-TE)", RFC 5420, DOI 10.17487/RFC5420, February 2009, <http://www.rfc-editor.org/info/rfc5420>.

[RFC5586] Bocci, M., Ed., Vigoureux, M., Ed., and S. Bryant, Ed., "MPLS Generic Associated Channel", RFC 5586, DOI 10.17487/RFC5586, June 2009, <http://www.rfc-editor.org/info/rfc5586>.

[RFC5905] Mills, D., Martin, J., Ed., Burbank, J., and W. Kasch, "Network Time Protocol Version 4: Protocol and Algorithms Specification", RFC 5905, DOI 10.17487/RFC5905, June 2010, <http://www.rfc-editor.org/info/rfc5905>.

[RFC6423] Li, H., Martini, L., He, J., and F. Huang, "Using the Generic Associated Channel Label for Pseudowire in the MPLS Transport Profile (MPLS-TP)", RFC 6423, DOI 10.17487/RFC6423, November 2011, <http://www.rfc-editor.org/info/rfc6423>.

[RFC7684] Psenak, P., Gredler, H., Shakir, R., Henderickx, W., Tantsura, J., and A. Lindem, "OSPFv2 Prefix/Link Attribute Advertisement", RFC 7684, DOI 10.17487/RFC7684, November 2015, <http://www.rfc-editor.org/info/rfc7684>.

[RFC7752] Gredler, H., Ed., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP", RFC 7752, DOI 10.17487/RFC7752, March 2016, <http://www.rfc-editor.org/info/rfc7752>.

10.2. Informative References

[I-D.ietf-ospf-ospfv3-lsa-extend] Lindem, A., Mirtorabi, S., Roy, A., and F. Baker, "OSPFv3 LSA Extendibility", draft-ietf-ospf-ospfv3-lsa-extend-13 (work in progress), October 2016.

[I-D.ietf-tictoc-1588overmpls] Davari, S., Oren, A., Bhatia, M., Roberts, P., and L. Montini, "Transporting Timing messages over MPLS Networks", draft-ietf-tictoc-1588overmpls-07 (work in progress), October 2015.

[ITU-T.G.8271] "Packet over Transport aspects - Synchronization, quality and availability targets", ITU-T Recomendation G.8271/Y.1366, July 2016.

[RFC4202] Kompella, K., Ed. and Y. Rekhter, Ed., "Routing Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4202, DOI 10.17487/RFC4202, October 2005, <http://www.rfc-editor.org/info/rfc4202>.

[RFC5036] Andersson, L., Ed., Minei, I., Ed., and B. Thomas, Ed., "LDP Specification", RFC 5036, DOI 10.17487/RFC5036, October 2007, <http://www.rfc-editor.org/info/rfc5036>.

[RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, DOI 10.17487/RFC5226, May 2008, <http://www.rfc-editor.org/info/rfc5226>.

[RFC6374] Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks", RFC 6374, DOI 10.17487/RFC6374, September 2011, <http://www.rfc-editor.org/info/rfc6374>.

[RFC7384] Mizrahi, T., "Security Requirements of Time Protocols in Packet Switched Networks", RFC 7384, DOI 10.17487/RFC7384, October 2014, <http://www.rfc-editor.org/info/rfc7384>.

Authors' Addresses

Greg Mirsky
ZTE Corp.

Email: gregimirsky@gmail.com

Stefano Ruffini
Ericsson

Email: stefano.ruffini@ericsson.com

Eric Gray
Ericsson

Email: eric.gray@ericsson.com

John Drake
Juniper Networks

Email: jdrake@juniper.net

Stewart Bryant
Huawei

Email: stewart.bryant@gmail.com

Alexander Vainshtein
ECI Telecom

Email: Alexander.Vainshtein@ecitele.com; Vainshtein.alex@gmail.com