MPLS Working Group
Internet-Draft
Intended status: Standards Track
Expires: June 12, 2018

H. Sitaraman
V. Beeram
Juniper Networks
T. Parikh
Verizon
T. Saad
Cisco Systems

December 9, 2017

Signaling RSVP-TE tunnels on a shared MPLS forwarding plane
draft-sitaraman-mpls-rsvp-shared-labels-03.txt

Abstract

As the scale of MPLS RSVP-TE networks has grown, so the number of Label Switched Paths (LSPs) supported by individual network elements has increased. Various implementation recommendations have been proposed to manage the resulting increase in control plane state.

However, those changes have had no effect on the number of labels that a transit Label Switching Router (LSR) has to support in the forwarding plane. That number is governed by the number of LSPs transiting or terminated at the LSR and is directly related to the total LSP state in the control plane.

This document defines a mechanism to prevent the maximum size of the label space limit on an LSR from being a constraint to control plane scaling on that node. That is, it allows many more LSPs to be supported than there are forwarding plane labels available.

This work introduces the notion of pre-installed 'per Traffic Engineering (TE) link labels' that can be shared by MPLS RSVP-TE LSPs that traverse these TE links. This approach significantly reduces the forwarding plane state required to support a large number of LSPs. This couples the feature benefits of the RSVP-TE control plane with the simplicity of the Segment Routing MPLS forwarding plane.

This document also introduces the ability to mix different types of label operations along the path of an LSP, thereby allowing the ingress router or an external controller to influence how to optimally place a LSP in the network.

Requirements Language

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

14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

Status of This Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on June 12, 2018.

Copyright Notice

Copyright (c) 2017 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction

The scaling of RSVP-TE [RFC3209] control plane implementations can be improved by adopting the guidelines and mechanisms described in [RFC2961] and [I-D.ietf-teas-rsvp-te-scaling-rec]. These documents do not make any difference to the forwarding plane state required to handle the control plane state. The forwarding plane state remains unchanged and is directly proportional to the total number of Label Switching Paths (LSPs) supported by the control plane.

This document describes a mechanism that prevents the size of the platform specific label space on a Label Switching Router (LSR) from being a constraint to pushing the limits of control plane scaling on that node.

This work introduces the notion of pre-installed 'per Traffic Engineering (TE) link labels' that are allocated by an LSR. Each such label is installed in the MPLS forwarding plane with a 'pop' operation and the instruction to forward the received packet over the TE link. An LSR advertises this label in the Label object of a Resv

message as LSPs are set up and they are recorded hop by hop in the Record Route object (RRO) of the Resv message as it traverses the network. To make use of this feature, the ingress Label Edge Router (LER) pushes a stack of labels [RFC3031] as received in the RRO. These 'TE link labels' can be shared by MPLS RSVP-TE LSPs that traverse the same TE link.

This forwarding plane behavior fits in the MPLS architecture [RFC3031] and is same as that exhibited by Segment Routing (SR) [I-D.ietf-spring-segment-routing] when using an MPLS forwarding plane and a series of adjacency segments. This work couples the feature benefits of the RSVP-TE control plane with the simplicity of the Segment Routing MPLS forwarding plane. The RSVP-TE tunnels that use this shared forwarding plane can co-exist with MPLS-SR LSPs [I-D.ietf-spring-segment-routing-mpls] as described in [I-D.ietf-teas-sr-rsvp-coexistence-rec].

RSVP-TE using a shared MPLS forwarding plane offers the following benefits:

1. Shared Labels: The transit label on a TE link is shared among RSVP-TE tunnels traversing the link and is used independent of the ingress and egress of the LSPs.

2. Faster LSP setup time: No forwarding plane state needs to be programmed during LSP setup and teardown resulting in faster time for provisioning and deprovisioning LSPs.

3. Hitless re-routing: New transit labels are not required during make-before-break (MBB) in scenarios where the new LSP instance traverses the exact same path as the old LSP instance. This saves the ingress LER and the services that use the tunnel from needing to update the forwarding plane with new tunnel labels and so makes MBB events faster. Periodic MBB events are relatively common in networks that deploy the 'auto-bandwidth' feature on RSVP-TE LSPs to monitor bandwidth utilization and periodically adjust LSP bandwidth.

4. Mix and match labels: Both 'TE link labels' and regular labels can be used on transit hops for a single RSVP-TE tunnel (see Section 6). This allows backward compatibility with transit LSRs that provide regular labels in Resv messages.

No additional extensions are required to routing protocols (IGP-TE) in order to support this shared MPLS forwarding plane. Functionalities such as bandwidth admission control, LSP priorities, preemption, auto-bandwidth and Fast Reroute continue to work with this forwarding plane.

The signaling procedures and extensions discussed in this document do not apply to Point to Multipoint (P2MP) RSVP-TE Tunnels.

2. Terminology

The following terms are used in this document:

TE link label: An incoming label at an LSR that will be popped by the LSR with the packet being forwarded over a specific outgoing TE link to a neighbor.

Shared MPLS forwarding plane: An MPLS forwarding plane where every participating LSR uses TE link labels on every LSP.

Segment Routed RSVP-TE tunnel: An MPLS RSVP-TE tunnel that requests the use of a shared MPLS forwarding plane at every hop of the LSP.

3. Allocation of TE Link Labels

An LSR that participates in a shared MPLS forwarding plane MUST allocate a unique TE link label for each TE link. When an LSR encounters a TE link label at the top of the label stack it MUST pop the label and forward the packet over the TE link to the downstream neighbor on the RSVP-TE tunnel.

Multiple TE link labels MAY be allocated for the TE link to accommodate tunnels requesting no protection, link-protection and node-protection over the specific TE link.

Implementations that maintain per label bandwidth accounting at each hop must aggregate the reservations made for all the LSPs using the shared TE link label.

4. Segment Routed RSVP-TE Tunnel Setup

This section provides an example of how the RSVP-TE signaling procedure works to set up a tunnel utilizing a shared MPLS forwarding plane. The sample topology below is used to explain the example. Labels shown at each node are TE link labels that, when present at the top of the label stack, indicate that they should be popped and that the packet should be forwarded on the TE link to the neighbor.

```
 +---+100  +---+150  +---+200  +---+250  +---+
 | A |-----| B |-----| C |-----| D |-----| E |
 +---+     +---+     +---+     +---+     +---+
   |110      |450      |550      |650      |850
   |         |         |         |         |
   |         |400      |500      |600      |800
   |       +---+     +---+     +---+     +---+
   +-------| F |-----|G  |-----|H  |-----|I  |
           +---+300  +---+350  +---+700  +---+
```

Figure 1: Sample Topology - TE Link Labels

Consider two tunnels:

RSVP-TE tunnel T1: From A to E on path A-B-C-D-E

RSVP-TE tunnel T2: From F to E on path F-B-C-D-E

Both tunnels share the TE links B-C, C-D, and D-E.

RSVP-TE is used to signal the setup of tunnel T1 (using the TE link label attributes flag defined in Section 10.2). When LSR D receives the Resv message from the egress LER E, it checks the next-hop TE link (D-E) and provides the TE link label (250) in the Resv message for the tunnel placing the label value in the Label object and also in the Label subobject carried in the RRO and setting the TE link label flag as defined in Section 10.3.

Similarly, LSR C provides the TE link label (200) for the TE link C-D, and LSR B provides the TE link label (150) for the TE link B-C.

For tunnel T2, the transit LSRs provide the same TE link labels as described for tunnel T1 as the links B-C, C-D, and D-E are common between the two LSPs.

The ingress LERs (A and F) will push the same stack of labels (from top of stack to bottom of stack) {150, 200, 250} for tunnels T1 and T2 respectively.

It should be noted that a transit LSR does not swap the top TE link label on an incoming packet (the label that it advertised in the Resv message it sent). All it has to do is pop the top label and forward the packet.

The values in the Label subobjects in the RRO are of interest to the ingress LERs in order to construct the stack of labels to impose on the packets.

If, in this example, there was another RSVP-TE tunnel T3 from F to I on path F-B-C-D-E-I, then this would also share the TE links B-C, C-D, and D-E and additionally traverse link E-I. The label stack used by F would be {150, 200, 250, 850}. Hence, regardless of the ingress and egress LERs from where the LSPs start and end, they will share LSR labels at shared hops in the shared MPLS forwarding plane.

There MAY be local operator policy at the ingress LER that influences the maximum depth of the label stack that can be pushed for a Segment Routed RSVP-TE tunnel. Prior to signaling the LSP, the ingress LER may decide that it would be unable to push a label stack containing one label for each hop along the path. In this case the LER can choose either to not signal a Segment Routed RSVP-TE tunnel (using normal LSP signaling instead), or can adopt the techniques described in Section 5 or Section 6.

5. Delegating Label Stack Imposition

One or more transit LSRs can assist the ingress LER by imposing part of the label stack required for the path. Consider the example in Figure 2 with an RSVP-TE tunnel from A to L on path A-B-C-D-E-F-G-H-I-J-K-L. In this case, the LSP is too long for LER A to impose the full label stack, so it uses the assistance of delegation hops LSR D and LSR I to impose parts of the label stack.

Each delegation hop allocates a delegation label to represent a set of labels that will be pushed at this hop. When a packet arrives at a delegation hop LSR with a delegation label, the LSR pops the label and pushes a set of labels before forwarding the packet.

```
                               1250d
 +---+100p  +---+150p  +---+200p  +---+250p  +---+300p  +---+
 | A |------| B |------| C |------| D |------| E |------| F |
 +---+      +---+      +---+      +---+      +---+      +---+
                                                          |350p
                                                          |
                               1500d                      |
 +---+  600p+---+  550p+---+  500p+---+  450p+---+  400p+---+
 | L |------| K |------| J |------| I |------| H |------+ G +
 +---+      +---+      +---+      +---+      +---+      +---+

        Notation : <Label>p - TE link label
                   <Label>d - delegation label
```

Figure 2: Delegating Label Stack Imposition

5.1. Stacking at the Ingress

When delegation labels come into play, there are two stacking approaches that the ingress can choose from. Section 7 explains how the label stack can be constructed.

5.1.1. Stack to Reach Delegation Hop

In this approach, the stack pushed by the ingress carries a set of labels that will take the packet to the first delegation hop. When this approach is employed, the set of labels represented by a delegation label at a given delegation hop will include the corresponding delegation label from the next delegation hop. As a result, this delegation label can only be shared among LSPs that are destined to the same egress and traverse the same downstream path.

This approach is shown in Figure 3. The delegation label 1250 represents the stack {300, 350, 400, 450, 1500} and the delegation label 1500 represents the label stack {550, 600}.

```
 +---+                 +---+                 +---+
 | A |-----.......-----| D |-----.......-----| I |----.....
 +---+                 +---+                 +---+

                  Pop 1250 &            Pop 1500 &
  Push                  Push                  Push
 ......                ......                ......
 : 150:          1250->: 300:          1500->: 550:
 : 200:                : 350:                : 600:
 :1250:                : 400:                ......
 ......                : 450:
                       :1500:
                       ......
```

Figure 3: Stack to Reach Delegation Hop

With this approach, the ingress LER A will push {150, 200, 1250} for the tunnel in Figure 2. At LSR D, the delegation label 1250 will get popped and {300, 350, 400, 450, 1500} will get pushed. And at LSR I, the delegation label 1500 will get popped and the remaining set of labels {550, 600} will get pushed.

5.1.2. Stack to Reach Egress

In this approach, the stack pushed by the ingress carries a set of labels that will take the packet all the way to the egress so that all the delegation labels are part of the stack. When this approach is employed, the set of labels represented by a delegation label at a given delegation hop will not include the corresponding delegation label from the next delegation hop. As a result, this delegation label can be shared among all LSPs traversing the segment between the two delegation hops.

The downside of this approach is that the number of hops that the LSP can traverse is dictated by the label stack push limit of the ingress.

This approach is shown in Figure 4. The delegation label 1250 represents the stack {300, 350, 400, 450} and the delegation label 1500 represents the label stack {550, 600}.

```
 +---+                +---+                +---+
 | A |-----.....-----| D |-----.....-----| I |-----.....
 +---+                +---+                +---+

             Pop 1250 &           Pop 1500 &
  Push                  Push                 Push
 ......               ......               ......
 : 150:         1250->: 300:         1500->: 550:
 : 200:               : 350:               : 600:
 :1250:               : 400:               ......
 :1500:               : 450:
 ......               ......
                      |1500|
                      ......
```

Figure 4: Stack to reach egress

With this approach, the ingress LER A will push {150, 200, 1250, 1500} for the tunnel in Figure 2. At LSR D, the delegation label 1250 will get popped and {300, 350, 400, 450} will get pushed. And at LSR I, the delegation label 1500 will get popped and the remaining set of labels {550, 600} will get pushed. The signaling extension required for the ingress to indicate the chosen stacking approach is defined in Section 10.6.

5.2. Explicit Delegation

In this delegation option, the ingress LER can explicitly delegate one or more specific transit LSRs to handle pushing labels for a certain number of their downstream hops. In order to accurately pick the delegation hops, the ingress needs to be aware of the label stack depth push limit of each of the transit LSRs prior to initiating the signaling sequence. The mechanism by which the ingress or controller (hosting the path computation element) learns this information is outside the scope of this document.

The signaling extension required for the ingress LER to explicitly delegate one or more specific transit hops is defined in Section 10.4. The extension required for the delegation hop to indicate that the recorded label is a delegation label is defined in Section 10.5.

5.3. Automatic Delegation

In this approach, the ingress LER lets the downstream LSRs automatically pick suitable delegation hops during the initial signaling sequence. The ingress does not need to be aware up front of the label stack depth push limit of each of the transit LSRs. The delegation hops are picked based on a per-hop signaled attribute called the Effective Transport Label-Stack Depth (ETLD) as described in the next section.

5.3.1. Effective Transport Label-Stack Depth (ETLD)

The ETLD is signaled as a per-hop attribute in the Path message [RFC7570]. When automatic delegation is requested, the ingress MUST populate the ETLD with the maximum number of transport labels that it can potentially send to its downstream hop. This value is then decremented at each successive hop. If a node is reached where the ETLD set from the previous hop is 1, then that node MUST select itself as the delegation hop. If a node is reached and it is determined that this hop cannot receive more than one transport label, then that node MUST select itself as the delegation hop. If there is a node or a sequence of nodes along the path of the LSP that do not support ETLD, then the immediate hop that supports ETLD MUST select itself as the delegation hop. The ETLD MUST be decremented at each non-delegation transit hop by either 1 or some appropriate number based on the limitations at that hop. At each delegation hop, the ETLD MUST be reset to the maximum number of transport labels that the hop can send and the ETLD decrements start again at each successive hop until either a new delegation hop is selected or the egress is reached. The net result is that by the time the Path message reaches the egress, all delegation hops are selected. During

the Resv processing, at each delegation hop, a suitable delegation label is selected (either an existing label is reused or a new label is allocated) and recorded in the Resv message.

Consider the example shown in Figure 5. Let's assume ingress LER A can push up to 3 transport labels while the remaining nodes can push up to 5 transport labels. The ingress LER A signals the initial Path message with ETLD set to 3. The ETLD value is adjusted at each successive hop and signaled downstream as shown. By the time the Path message reaches the egress LER L, LSRs D and I are automatically selected as delegation hops.

```
        ETLD:3    ETLD:2    ETLD:1    ETLD:5    ETLD:4
        ----->    ----->    ----->    ----->    ----->
                                  1250d
    +---+100p +---+150p +---+200p +---+250p +---+300p +---+
    | A |-----| B |-----| C |-----| D |-----| E |-----| F |  ETLD:3
    +---+     +---+     +---+     +---+     +---+     +---+    |
                                                        |350p  |
                                                        |      |
                                  1500d                 |      |
    +---+ 600p+---+ 550p+---+ 500p+---+ 450p+---+ 400p+---+    v
    | L |-----| K |-----| J |-----| I |-----| H |-----+ G +
    +---+     +---+     +---+     +---+     +---+     +---+

        ETLD:3    ETLD:4    ETLD:5    ETLD:1    ETLD:2
        <-----    <-----    <-----    <-----    <-----
```

Figure 5: ETLD

Signaling extension for the ingress LER to request automatic delegation is defined in Section 10.4. The extension for signaling the ETLD is defined in Section 10.7. The extension required for the delegation hop to indicate that the recorded label is a delegation label is defined in Section 10.5.

6. Mixing TE Link Labels and Regular Labels in an RSVP-TE Tunnel

Labels can be mixed across transit hops in a single MPLS RSVP-TE LSP. Certain LSRs can use TE link labels and others can use regular labels. The ingress can construct a label stack appropriately based on what type of label is recorded from every transit LSR.

```
                          (#)       (#)
    +---+100  +---+150  +---+200  +---+250  +---+
    | A |-----| B |-----| C |-----| D |-----| E |
    +---+     +---+     +---+     +---+     +---+
      |110      |450      |550      |650      |850
      |         |         |         |         |
      |         |400      |500      |600      |800
      |       +---+     +---+     +---+     +---+
      +-------| F |-----|G  |-----|H  |-----|I  |
              +---+300  +---+350  +---+700  +---+

            Notation : (#) denotes regular labels
                       Other labels are TE link labels
```

Figure 6: Sample Topology - TE Link Labels and Regular Labels

If the transit LSR allocates a regular label to be sent upstream in the Resv, then the label operation at the LSR is a swap to the label received from the downstream LSR. If the transit LSR is using a TE link label to be sent upstream in the Resv, then the label operation at the LSR is a pop and forward regardless of any label received from the downstream LSR. There is no change in the behavior of a penultimate hop popping (PHP) LSR [RFC3031].

Section 7 explains how the label stack can be constructed. For example, the LSP from A to I using path A-B-C-D-E-I will use a label stack of {150, 200}.

7. Construction of Label Stacks

The ingress LER or delegation hop MUST check the type of label received from each transit hop as recorded in the RRO in the Resv message and generate the appropriate label stack to reach the next delegation hop or the egress.

The following logic could be used by the node constructing the label stack:

> Each RRO label sub-object SHOULD be processed starting with the label sub-object from the first downstream hop. Any label provided by the first downstream hop MUST always be pushed on the label stack regardless of the label type. If the label type is a TE link label, then any label from the next downstream hop MUST also be pushed on the constructed label stack. If the label type is a regular label, then any label from the next downstream hop MUST NOT be pushed on the constructed label stack. If the label type is a delegation label, then the stacking procedure stops at

that delegation hop. Approaches in Section 5.1 SHOULD be used to determine how the delegation labels are pushed in the label stack.

8. Facility Backup Protection

The following section describe how link and node protection works with facility backup protection [RFC4090] for the Segment Routed RSVP-TE tunnels.

8.1. Link Protection

To provide link protection at a Point of Local Repair (PLR) with a shared MPLS forwarding plane, the LSR SHOULD allocate a separate TE link label for the TE link that will be used for RSVP-TE tunnels that request link-protection from the ingress. No signaling extensions are required to support link protection for RSVP-TE tunnels over the shared MPLS forwarding plane.

At each LSR, link protected TE link labels can be allocated for each TE link and a link protecting facility backup LSP can be created to protect the TE link. The link protected TE link label can be sent by the LSR for LSPs requesting link-protection over the specific TE link. Since the facility backup terminates at the next-hop (merge point), the incoming label on the packet will be what the merge point expects.

Consider the network shown in Figure 7. LSR B can install a facility backup LSP for the link protected TE link label 151. When the TE link B-C is up, LSR B will pop 151 and send the packet to C. If the TE link B-C is down, the LSR can pop 151 and send the packet via the facility backup to C.

```
      101(*)     151(*)     201(*)     251(*)
 +---+100   +---+150   +---+200   +---+250   +---+
 | A |------| B |------| C |------| D |------| E |
 +---+      +---+      +---+      +---+      +---+
   |110       |450       |550       |650       |850
   |          |          |          |          |
   |          |400       |500       |600       |800
   |        +---+      +---+      +---+      +---+
   +--------| F |------|G  |------|H  |------|I  |
            +---+300   +---+350   +---+700   +---+

  Notation : (*) denotes link protection TE link labels
```

Figure 7: Link Protection Topology

8.2. Node Protection

The solutions for the PLR to provide node-protection for the Segment Routed RSVP-TE tunnel will be explained in a future version of this document.

9. Quantifying TE Link Labels

This section quantifies the number of labels required in the forwarding plane to provide sharing of labels across Segment Routed RSVP-TE tunnels. An MPLS RSVP-TE tunnel offers either no protection, link protection, or node protection and only one of these labels is required per tunnel during signaling. The scale of the number of TE link labels required per LSR can be deduced as follows:

- For an LSR having X neighbors reachable across Y interfaces, the number of unprotected TE link labels is X.

- For a PLR having X neighbors reachable across Y interfaces, the number of link protected TE link labels is X.

- For a PLR having X neighbors, each having Nx neighbors (i.e. next-nexthops for the PLR), number of node protected TE link labels is SUM_OF_ALL(Nx).

The total number of TE link labels is given by:
Unprotected TE link labels +
link protected TE link labels +
node protected TE link labels = 2X + SUM_OF_ALL(Nx)

10. Protocol Extensions

10.1. Requirements

The functionality discussed in this document imposes the following requirements on the signaling protocol.

- The Ingress of the LSP SHOULD have the ability to mandate/request the use and recording of TE link labels at all hops along the path of the LSP.

- When the use of TE link labels is mandated/requested for the path:

  * the node recording the TE link label SHOULD have the ability to indicate if the recorded label is a TE link label.

* the ingress SHOULD have the ability to delegate label stack imposition by:

  + explicitly mandating specific hops to be delegation hops (or)

  + requesting automatic delegation.

* When explicit delegation is mandated or automatic delegation is requested:

  + the ingress SHOULD have the ability to indicate the chosen stacking approach (and)

  + the delegation hop SHOULD have the ability to indicate that the recorded label is a delegation label.

10.2. Attribute Flags TLV: TE Link Label

Bit Number (TBD1): TE Link Label

The presence of this in the LSP_ATTRIBUTES/LSP_REQUIRED_ATTRIBUTES object of a Path message indicates that the ingress has requested/ mandated the use and recording of TE link labels at all hops along the path of this LSP. When a node that does not cater to the mandate receives a Path message carrying the LSP_REQUIRED_ATTRIBUTES object with this flag set, it MUST send a PathErr message with an error code of 'routing problem' and an error value of 'TE link label usage failure'.

10.3. RRO Label Subobject Flag: TE Link Label

Bit Number (TBD2): TE Link Label

The presence of this flag indicates that the recorded label is a TE link label. This flag MUST be used by a node only if the use and recording of TE link labels is requested/mandated for the LSP.

10.4. Attribute Flags TLV: LSI-D

Bit Number (TBD3): Label Stack Imposition - Delegation (LSI-D)

Automatic Delegation: The presence of this flag in the LSP_ATTRIBUTES object of a Path message indicates that the ingress has requested automatic delegation of label stack imposition. This flag MUST be set in the LSP_ATTRIBUTES object of a Path message only if the use and recording of TE link labels is requested/mandated for this LSP.

Explicit Delegation: The presence of this flag in the HOP_ATTRIBUTES subobject [RFC7570] of an ERO object in the Path message indicates that the hop identified by the preceding IPv4 or IPv6 or Unnumbered Interface ID subobject has been picked as an explicit delegation hop. The HOP_ATTRIBUTES subobject carrying this flag MUST have the R (Required) bit set. This flag MUST be set in the HOP_ATTRIBUTES subobject of an ERO object in the Path message only if the use and recording of TE link labels is requested/mandated for this LSP. If the hop is not able to comply with this mandate, it MUST send a PathErr message with an error code of 'routing problem' and an error value of 'label stack imposition failure'.

10.5. RRO Label Subobject Flag: Delegation Label

Bit Number (TBD4): Delegation Label

The presence of this flag indicates that the recorded label is a delegation label. This flag MUST be used by a node only if the use and recording of TE link labels and delegation are requested/mandated for the LSP.

10.6. Attributes Flags TLV: LSI-D-S2E

Bit Number (TBD5): Label Stack Imposition - Delegation - Stack to reach egress (LSI-D-S2E)

The presence of this flag in the LSP_ATTRIBUTES object of a Path message indicates that the ingress has chosen to use the "Stack to reach egress" approach for stacking. The absence of this flag in the LSP_ATTRIBUTES object of a Path message indicates that the ingress has chosen to use the "Stack to reach delegation hop" approach for stacking. This flag MUST be set in the LSP_ATTRIBUTES object of a Path message only if the use and recording of TE link labels and delegation are requested/mandated for this LSP.

10.7. Attributes TLV: ETLD

The format of the ETLD Attributes TLV is shown in Figure 8. The Attribute TLV Type is TBD6.

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|         Reserved                              |     ETLD      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 8: The ETLD Attributes TLV

The presence of this TLV in the HOP_ATTRIBUTES subobject of an RRO object in the Path message indicates that the hop identified by the preceding IPv4 or IPv6 or Unnumbered Interface ID subobject supports automatic delegation. This attribute MUST be used only if the use and recording of TE link labels is requested/mandated and automatic delegation is requested for the LSP. The ETLD field specifies the maximum number of transport labels that this hop can potentially send to its downstream hop.

11. OAM Considerations

MPLS LSP ping and traceroute [RFC8029] are applicable for Segment Routed RSVP-TE tunnels. The existing procedures allow for the label stack imposed at a delegation hop to be reported back in the Label Stack Sub-TLV in the MPLS echo reply for traceroute.

12. Acknowledgements

The authors would like to thank Adrian Farrel, Kireeti Kompella, Markus Jork and Ross Callon for their input from discussions.

Adrian Farrel provided a review and text suggestion for clarity and readability.

13. Contributors

The following individuals contributed to this document:

Raveendra Torvi
Juniper Networks
Email: rtorvi@juniper.net

Chandra Ramachandran
Juniper Networks
Email: csekar@juniper.net

George Swallow
Email: swallow.ietf@gmail.com

14. IANA Considerations

14.1. Attribute Flags: TE Link Label, LSI-D, LSI-D-S2E

IANA manages the 'Attribute Flags' registry as part of the 'Resource Reservation Protocol-Traffic Engineering (RSVP-TE) Parameters' registry located at http://www.iana.org/assignments/rsvp-te-parameters. This document introduces three new Attribute Flags.

Bit No.	Name	Attribute FlagsPath	Attribute FlagsResv	RRO	ERO	Reference
TBD1	TE Link Label	Yes	No	No	No	This document (Section 11.2)
TBD3	LSI-D	Yes	No	No	Yes	This document (Section 11.4)
TBD5	LSI-D-S2E	Yes	No	No	No	This document (Section 11.6)

14.2. Attribute TLV: ETLD

IANA manages the "Attribute TLV Space" registry as part of the 'Resource Reservation Protocol-Traffic Engineering (RSVP-TE) Parameters' registry located at http://www.iana.org/assignments/rsvp-te-parameters. This document introduces a new Attribute TLV.

Type	Name	Allowed on LSP ATTRIBUTES	Allowed on LSP REQUIRED ATTRIBUTES	Allowed on LSP Hop Attributes	Reference
TBD6	ETLD	No	No	Yes	This document (Section 11.7)

14.3. Record Route Label Sub-object Flags: TE Link Label, Delegation Label

IANA manages the 'Record Route Object Sub-object Flags' registry as part of the 'Resource Reservation Protocol-Traffic Engineering (RSVP-TE) Parameters' registry located at http://www.iana.org/assignments/rsvp-te-parameters. This registry currently does not include Label Sub-object Flags. This document requests the addition of a new sub-registry for Label Sub-object Flags as shown below.

Flag	Name	Reference
0x1	Global Label	RFC 3209
TBD2	TE Link Label	This document (Section 11.3)
TBD4	Delegation Label	This document (Section 11.5)

15. Security Considerations

This document does not introduce new security issues. The security considerations pertaining to the original RSVP protocol [RFC2205] and RSVP-TE [RFC3209] and those that are described in [RFC5920] remain relevant.

16. References

16.1. Normative References

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

[RFC2205] Braden, R., Ed., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, DOI 10.17487/RFC2205, September 1997, <https://www.rfc-editor.org/info/rfc2205>.

[RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, DOI 10.17487/RFC3031, January 2001, <https://www.rfc-editor.org/info/rfc3031>.

[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, <https://www.rfc-editor.org/info/rfc3209>.

[RFC4090] Pan, P., Ed., Swallow, G., Ed., and A. Atlas, Ed., "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, DOI 10.17487/RFC4090, May 2005, <https://www.rfc-editor.org/info/rfc4090>.

[RFC7570] Margaria, C., Ed., Martinelli, G., Balls, S., and B. Wright, "Label Switched Path (LSP) Attribute in the Explicit Route Object (ERO)", RFC 7570, DOI 10.17487/RFC7570, July 2015, <https://www.rfc-editor.org/info/rfc7570>.

[RFC8029] Kompella, K., Swallow, G., Pignataro, C., Ed., Kumar, N., Aldrin, S., and M. Chen, "Detecting Multiprotocol Label Switched (MPLS) Data-Plane Failures", RFC 8029, DOI 10.17487/RFC8029, March 2017, <https://www.rfc-editor.org/info/rfc8029>.

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

16.2. Informative References

[I-D.ietf-spring-segment-routing]
Filsfils, C., Previdi, S., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", draft-ietf-spring-segment-routing-13 (work in progress), October 2017.

[I-D.ietf-spring-segment-routing-mpls]
Filsfils, C., Previdi, S., Bashandy, A., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing with MPLS data plane", draft-ietf-spring-segment-routing-mpls-11 (work in progress), October 2017.

[I-D.ietf-teas-rsvp-te-scaling-rec]
Beeram, V., Minei, I., Shakir, R., Pacella, D., and T. Saad, "Techniques to Improve the Scalability of RSVP Traffic Engineering Deployments", draft-ietf-teas-rsvp-te-scaling-rec-08 (work in progress), October 2017.

[I-D.ietf-teas-sr-rsvp-coexistence-rec]
Sitaraman, H., Beeram, V., Minei, I., and S. Sivabalan, "Recommendations for RSVP-TE and Segment Routing LSP co-existence", draft-ietf-teas-sr-rsvp-coexistence-rec-01 (work in progress), June 2017.

[RFC2961] Berger, L., Gan, D., Swallow, G., Pan, P., Tommasi, F., and S. Molendini, "RSVP Refresh Overhead Reduction Extensions", RFC 2961, DOI 10.17487/RFC2961, April 2001, <https://www.rfc-editor.org/info/rfc2961>.

[RFC5920] Fang, L., Ed., "Security Framework for MPLS and GMPLS Networks", RFC 5920, DOI 10.17487/RFC5920, July 2010, <https://www.rfc-editor.org/info/rfc5920>.

Authors' Addresses

Harish Sitaraman
Juniper Networks
1133 Innovation Way
Sunnyvale, CA 94089
US

Email: hsitaraman@juniper.net

Vishnu Pavan Beeram
Juniper Networks
10 Technology Park Drive
Westford, MA 01886
US

Email: vbeeram@juniper.net

Tejal Parikh
Verizon
400 International Parkway
Richardson, TX 75081
US

Email: tejal.parikh@verizon.com

Tarek Saad
Cisco Systems
2000 Innovation Drive
Kanata, Ontario K2K 3E8
Canada

Email: tsaad@cisco.com