Network Working Group Arthi Ayyangar (Juniper Networks)
Internet Draft Category: Standards Track
October 2003 Expires: April 2004

Inter-region MPLS Traffic Engineering
draft-ayyangar-inter-region-te-01.txt

Abstract

This draft proposes mechanisms for the establishment and maintenance of MPLS Traffic Engineering (TE) Label Switched Paths (LSP) that traverse multiple regions, where a region could be an IGP Area or an Autonomous System or a GMPLS overlay network. This document also outlines different mechanisms that an operator could employ within his region to facilitate the setup of these inter-region TE LSPs.

Conventions used in this document

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].

1. Terminology

In the context of this document we define a "TE LSP region", or just a "region" as either a single IGP area, or a single Autonomous System (AS). Note that a region may itself be formed of multiple regions. E.g. a region formed by an AS may be composed of multiple regions, each corresponding to an IGP area. Also, this proposal may be applicable to other inter-region TE LSP setup scenarios like TE LSP across networks with different addressing/routing realms or under different administrative domains, but these scenarios will not be discussed in detail in this document.

The notion of 'TE LSP nesting' refers to the ability to carry one or more inter-region TE LSPs within another intra-region TE LSP by using the MPLS label stacking property at the Head-end of the intra-region TE LSP. On the other hand, 'stitching a TE LSP' means to split an inter-region LSP and insert a different intra-region LSP, into the split. This implies a label swap operation at the Head-end of the intra-region LSP. Similar to [LSP_HIER], in the context of this document as well, the term FA-LSP always implies one or more LSPs nested within another LSP using the label stack construct. And we use the term 'LSP segment' when one LSP is split and another LSP is inserted into the split (LSP stitching).

Other terminologies used are:

LSP - An MPLS Label Switched Path

LSR - Label Switch Router

TE LSP - Traffic Engineering Label Switch Path (also referred as LSP in this draft)

TED - TE database

FA-LSP - Forwarding Adjacency LSP

GMPLS - Generalized MPLS

ABR - Area Border Router

AS - Autonomous System

ASBR - Autonomous System Border Router

ERO - Explicit Route Object

IGP - Interior Gateway Protocol

CSPF - Constraint-based Shortest Path First

RRO - Record Route Object

Protect TE LSP - a TE LSP requesting local protection in the context of [FAST-REROUTE]: the local protection desired bit of the SESSION-ATTRIBUTE object is set or a FAST-REROUTE object is inserted in the RSVP Path message

PLR - Point of Local Repair; i.e. the head end of a bypass tunnel or a detour LSP

MP - Merge Point; i.e. the LSR where the bypass tunnel or detour rejoin the protected LSP

SP - Service Provider

HE LSR - Head End Label Switching Router

2. Introduction

When TE LSPs span multiple regions, there can be multiple approaches taken by the solutions to setup such LSPs. The solution proposed in [INTER-AS] describes a way to signal a contiguous TE LSP that spans multiple ASes and exercise more control at the HE LSR of the TE LSP. On the other hand, this document focuses on signaling a TE LSP with LSP segments or FA-LSPs in different regions and exercise more localized (per region) control on the inter-region TE LSP. Also, the solution proposed here applies to TE LSP setup across regions other than ASes as well. There are also other differences between the two solutions that will not be detailed here.

This draft builds upon the constructs and mechanisms defined in [LSP-HIER] for the setup of TE LSPs spanning multiple regions. The idea is to separate the inter-region LSP into different segments at the region boundaries such that each region is completely in control of it's segment, with respect to its placement and maintenance. The signaling mechanisms described in [LSP-HIER] are re-used here to setup inter-region TE LSPs. This improves scaling of control plane by aggregating the end-to-end LSP requests. So, the control state corresponding to the inter-region TE LSP is only present on the region boundary LSRs. Also, this helps to achieve a common solution to solve the generic problem of establishing TE LSPs crossing different "regions", without any significant protocol changes. The operator of a region has complete control on the intra-region FA-LSPs (with nesting) or the LSP segments (with stitching) which transport the inter-region TE LSPs. Since these characteristics are also

usually desirable while setting up TE LSPs that traverse networks under different administrative control or networks with different addressing and routing realms, this solution is also applicable to the GMPLS overlay model ([GMPLS-OVERLAY]), where the client requesting the LSP setup would belong to a region different from the core network region.

[LSP-HIER] uses the Interface Switching Capabilities to construct regions and determine region boundaries. This document augments the definition of region in [LSP-HIER] by adding two new types of regions :- an IGP area (for OSPF)/ level (for ISIS) and an Autonomous System (AS). In the former case, the boundaries of the region are the Area Border Routers (ABRs) of the area that forms the region: the boundaries of the region could be determined by examining the IGP information. In the latter case, the boundaries of the region are the Autonomous System Border Routers (ASBRs) of the AS that forms the region: the boundaries of a region could be determined by examining the BGP information. Neither of these cases require the use of the Interface Switching Capabilities to construct regions and determine regions boundaries.

The draft abides by the requirements for Inter-AS Traffic Engineering listed in [INTER-AS-TE-REQTS] for inter-AS LSP setup where a region would correspond to an AS and the requirements for Inter-Area Traffic Engineering listed in [INTER-AREA-TE-REQTS] for inter-area TE LSP setup where a region would correspond to a single OSPF area or an ISIS level.

3. Assumptions and Requirements

- Each region in all the examples below is assumed to be capable of doing Traffic Engineering; i.e. running OSPF-TE or ISIS-TE and RSVP-TE and may itself be composed of several other regions (for instance when an AS is made of several IGP areas/levels).

- The inter-region LSPs are signaled using RSVP-TE ([RSVP-TE]).

- The path (ERO) for the inter-region LSP traversing multiple regions is either configured as a set of (loose and/or strict) hops or the boundary LSRs for each of the regions along the path are capable of dynamically finding the next-hop boundary LSR towards the LSP destination when the LSP setup request arrives. This process of being able to dynamically determine the next-hop boundary LSR for an LSP destination during TE LSP setup will be referred to as "auto-discovery" mechanism in the rest of this document. At the simplest, auto-discovery can be based on just reachability information. This would usually be accompanied with crankback mechanisms to try another

next-hop boundary LSR, when the path to one next-hop boundary LSR fails. While the ability of a region boundary LSR to auto-discover the next-hop boundary LSR needs to be a part of the complete inter-region TE solution, it will not be discussed here in any further detail. Also, this mechanism will be considered to be optional as far as inter-region TE LSP setup is concerned. In the absence of any auto-discovery mechanism, the configured ERO MUST at least include all the boundary LSRs of each region to be traversed along the path. Also, the addresses configured in the ERO MUST be reachable by the corresponding previous hop boundary LSR.

- The region boundary LSRs are assumed to be capable of expansion of a loose next-hop in the ERO by performing partial CSPF computation to that loose hop, instead of to the LSP destination. This way, TE LSP path computation is distributed among different regions and no topology information needs to be distributed between regions.

- The paths for the intra-region FA-LSPs or LSP segments, may be pre-configured or computed dynamically based on the arriving inter-region LSP setup request; depending on the requirements of the transit region. When the paths for the FA-LSPs/LSP segments are pre-configured, the constraints as well as other parameters like local protection scheme for the intra-region FA-LSP/LSP segment are also pre-configured. Some local algorithm can be used on the HE LSR of a FA-LSP to dynamically adjust the FA-LSP bandwidth based on the cumulative bandwidth requested by the inter-region TE LSPs. It is recommended to use a threshold triggering mechanism to avoid constant bandwidth readjustment as inter-region TE LSPs are set up and torn down.

- While certain constraints like bandwidth can be used across different regions, certain other TE constraints like resource affinity, color, metric; etc as listed in [RFC2702] could be translated differently in different regions. It is assumed that, at the region boundary LSRs, there will exist some sort of local mapping based on offline policy agreement, in order to translate such constraints across region boundaries. This would be something similar to 'Inter-AS TE Agreement Enforcement Polices' stated in [INTER-AS-TE-REQTS].

- The intra-region FA-LSPs or LSP segments in packet-switched networks are assumed to be unidirectional.

- When a region boundary LSR at the exit of a region receives a TE LSP setup request (Path message) for an inter-region TE LSP, then if this LSP had been nested or stitched at the entry region boundary LSR, then this exit boundary LSR can determine the corresponding FA-LSP or LSP segment from the received Path message. The signaling

mechanism used to signal an inter-region TE LSP being transported either over a FA-LSP or LSP segment is similar to that described in [LSP-HIER]. The way to identify an unnumbered FA is described in [RSVP-UNNUM]. The same mechanisms will be used here.

- The Record Route Object (RRO) is an optional object, and if the inter-region LSPs do not require features like protection which require the RRO in the RSVP messages, then the inter-region LSPs do not need to carry the RRO.

4. Basic Operation

4.1. Intra-region FA-LSP/LSP segment setup

FA-LSPs or LSP segments can be pre-configured on any region boundary LSR. But, when the FA-LSP/LSP segment setup is dynamic, then by default, only a region boundary LSR that receives an inter-region LSP setup request (Path) from a different region SHOULD trigger the setup of a FA-LSP or LSP segment in its region. So all boundary LSRs at the entry to a region, are candidates for dynamic intra-region FA-LSP/LSP segment setup. The HE LSR of an inter-region LSP can be treated as an exception to the above clause and MAY be considered as a candidate for dynamic FA-LSP/LSP segment setup. Also, the default behavior MAY be overridden by configuration, if required.

When a region boundary LSR receives a Path message with a loose next-hop in the ERO, then it carries out the following actions:

- It checks if the loose next-hop is accessible via the TED. If the loose next-hop is not present in the TED and an auto-discovery mechanism exists, then it will check if the next-hop at least has IP reachability (via IGP or BGP). If the next-hop is not reachable, then the LSR will be unable to propagate the Path message any further and will send back a PathErr upstream. If the next-hop is reachable, then it will find a region boundary LSR to get to the next-hop. In the absence of an auto-discovery mechanism, the region boundary LSR should be the loose next-hop in the ERO and hence should be accessible via the TED, otherwise path computation for the inter-region TE LSP will fail.

- If the next-hop boundary LSR is present in the TED, then if this region boundary LSR (receiving the LSP setup request) is a candidate LSR for intra-region FA-LSP/LSP segment setup, and if there is no FA-LSP/LSP segment from this LSR to the next-hop boundary LSR, (satisfying the constraints) it SHOULD signal a FA-LSP/LSP segment to the next-hop boundary LSR. If pre-configured FA-LSP(s) or LSP

segment(s) already exist, then it SHOULD try to select from among those intra-region LSPs. Depending on local policy, it MAY signal a new FA-LSP/LSP segment if this selection fails. If the FA-LSP/LSP segment is successfully signaled or selected, it propagates the inter-region Path message to the next-hop following the procedures described in [LSP-HIER]. If, for some reason the dynamic FA-LSP/LSP segment setup to the next-hop boundary LSR fails, a PathErr is sent upstream for the inter-region LSP. Similarly, if selection of a pre-configured FA-LSP/LSP segment fails and local policy prevents dynamic FA-LSP/LSP segment setup, then a PathErr is sent upstream for the inter-region TE LSP.

- If, however, this region boundary LSR is not a FA-LSP/LSP segment candidate, then it SHOULD simply compute a CSPF path upto the next-hop boundary LSR (carry out an ERO expansion to the next-hop boundary LSR) and propagate the Path message downstream. The outgoing ERO may be modified after an ERO expansion to the loose next-hop.

The above procedures do not apply when a region boundary LSR receives a Path message with strict next-hop. Also, detailed explanation of signaling a FA-LSP versus an LSP segment follows in Section 5.

Note that this mode of operation depends on TE LSP attributes requested in the LSP_ATTRIBUTES object (see [RSVP-ATTRIBUTES] and [INTER-AS]) of the inter-region TE LSP. As described in Section 1, the solution proposed in [INTER-AS] enables an operator to signal a contiguous inter-AS TE LSP. So, the following bit may be set in the Attributes Flag TLV of LSP_ATTRIBUTES object to control the desired behavior on the intermediate nodes.

0x02: Contiguous LSP required bit: when set this indicates that a boundary LSR MUST not perform any nesting or stitching on the TE LSP and the TE LSP MUST be routed as any other TE LSP (it must be contiguous end to end). When this bit is cleared, a boundary LSR can decide to perform either nesting or stitching.

The details of the LSP_ATTRIBUTES object can be found in [RSVP-ATTRIBUTES] and the above bit in the Attributes Flag TLV is defined in [INTER-AS].

4.2. Inter-area LSP setup

Several schemes have been proposed for inter-area TE (see [INTER-AREA]). This draft proposed an additional method. In this case, each region corresponds to a single IGP area and the ABRs will form the region boundaries.

A-B-C-D: inter-area TE LSP path

B, B', C, C': ABRs

```
     +--------------+ +-----------------+ +--------------+
     |Area1         | |   Area0         | |Area2         |
     |               B                   C               |
     |A             | |                 | |             D|
     |               B'                  C'              |
     |              | |                 | |              |
     +--------------+ +-----------------+ +--------------+
```

Let us consider a scenario where A initiates the setup of an inter-area TE LSP from A to D. When the Path message reaches B, B performs the following set of actions:

- It determines the egress LSR from its region along the LSP path (say C), either from the ERO or by using some constraint-aware auto-discovery mechanism or based simply on reachability information.

- B will check if it has a FA-LSP or LSP segment to C matching the constraints carried in the inter-region Path message. If not, B will setup a FA-LSP or LSP segment from B to C. Note that once the FA-LSP/LSP segment is setup, it may be advertised as a link within that region (see [LSP-HIER]) (area 0 in this example). The FA-LSP or LSP segment could have also been pre-configured.

- In the Path message for the FA-LSP/LSP segment, B also signals whether it will do a one-to-one LSP stitching or whether it will nest the inter-region LSP over the intra-region FA-LSP. The details of nesting versus stitching and how this behavior is signaled, are described in Section 5.

- Also, there could be multiple FA-LSPs/LSP segments between B and C. So, B needs to select one FA-LSP/LSP segment from these, for the inter-region LSP through Area 0. The mechanism and the criterion used to select the FA-LSP/LSP segment is local to B and will not be described here in detail. e.g. if we have multiple pre-configured FA-LSPs/LSP segments, a local policy may prefer to use FA-LSPs (nesting) for most inter-region LSP requests. And it may select the LSP segments (stitching) only for some specific inter-region LSPs.

- Once it has selected the FA-LSP/LSP segment for the inter-region LSP, using the signaling procedures described in [LSP-HIER] B sends the Path message for inter-region LSP to C. Note that irrespective of whether B does nesting or stitching, the Path message for the inter-region TE LSP is always forwarded to C. C then repeats the exact same procedures and the Path message for the inter-area TE LSP will reach

the destination D. If C cannot find a path obeying the set of constraints for the inter-region TE LSP, then C MUST send a PathErr message to B. Then B can in turn either select another FA-LSP/LSP segment to C if such an LSP exists or select another egress boundary LSR (C' in the example above). Note also, that B may be configured to forward the PathErr upto the inter-region HE LSR without trying to select another egress LSR.

- The Resv message for the inter-area TE LSP is sent back from D to A. When the Resv message arrives at C, depending on whether C is nesting or stitching, C will install the appropriate label actions for the packets arriving on the inter-region LSP. Similar procedures are carried out at B as well, while processing the Resv message.

As the Resv message for the inter-region LSP, traverses back from D to A, each LSR along the Path may record an address into the RRO object carried in the Resv. According to [RSVP-TE], the addresses in the RRO object may be a node or interface addresses. The link corresponding to an unnumbered FA-LSP/LSP segment will have the ingress and egress LSR Router-IDs as the link addresses ([RSVP-UNNUM]). So when C sends the Resv message to B, C will record it's Router ID in the RRO object. So, the inter-area TE LSP from A to D would have an RRO of A-B-C-D or A-<other hops>-B-C-D, depending on whether the source region is setting up a FA-LSP/LSP segment or not. If the FA-LSPs/LSP segments are numbered, then the addresses assigned to the FA-LSP/LSP segment will be recorded in the RRO object.

4.3. Inter-AS LSP setup

In this case, each region corresponds to a single AS and the ASBRs define the region boundaries. As per the ERO loose-hop expansion explained in Section 4.1, however, the FA-LSPs or LSP segments will only be originated by the ASBRs at the entry to the AS. A few examples of the inter-AS LSPs:

A to F (LSR to LSR), CE1 to CE2 (CE to CE), R1 to R2 or R1 to R3, CE1 to F (CE to ASBR)

AS1, AS2, AS3: Three ASes either belonging to same SP or different SPs

B,B',C,C',D,D',E,E': ASBRs

R1, R2, R3: Other LSRs within the respective ASes

```
         +---------------+  +-----------------+  +---------------+
         |AS1            |  |AS2              |  |AS3            |
```

```
         |             B |--|C              D |--|E              |
     CE1-|A              |  |                 |  |             F|-CE2
         |      R1     B'|--|C'     R2      D'|--|E'    R3       |
         |               |  |                 |  |               |
         +---------------+  +-----------------+  +---------------+
```

The procedures for establishing an inter-AS TE LSP are very similar to that of the inter-area TE LSP described above. The main difference here from the inter-area case, is the presence of the ASBR-ASBR link(s). If the ASBRs are connected by a single hop link, then they typically would not be running an IGP between them. Even when there are multiple hops between the ASBRs, although there would be some IGP running between the ASBRs; the TE information for the ASBR-ASBR links is not usually available. E.g. TE information corresponding to links B-C and B'-C' is not usually available. Also, two adjacent ASes could be running different IGPs. It is useful to have the TE information for the links originated by the ASBR to take into account the TE parameters on that link during CSPF computation on the ASBR. The proposal here is for the ASBR to add the TE information corresponding to the links that it originates in it's local TE database, although no real IGP adjacency exists on those links.

As a further optimization, if this TE information corresponding to the links originated by the ASBR is made available in the TED of other LSRs in the same region that the ASBR belongs to, then the CSPF computations for inter-AS TE LSP path in that region can also account for constraints on the ASBR-ASBR link. This will improve the chance of successful path computation upto the next AS in case of a bottleneck on some ASBR-ASBR links and it reduces one level of crankback. Note that no topology information is flooded and these links are not used in IGP SPF computations. Only the TE information for the links originated by the ASBR is advertised using standard IGP TE extensions ([OSPF-TE], [ISIS-TE]).

E.g. B and B' would add in their respective TEDs, TE information corresponding to links B-C and B'-C' respectively. Note that this may need some additional configuration on B and B' to associate their links to the nexthop ASBR nodes C and C' respectively in the TED. This is because the ASBRs usually only have EBGP nexthop information. Now, so that CSPF in AS1 can take into account the TE parameters for links B-C and B'-C' the ASBRs B and B' would need to advertise the TE information for B-C and B'-C' in AS1's IGP.

If the TE information for the ASBR links is not available on the ASBRs, then the ASBRs cannot perform any CSPF computations and any path that the TE LSP takes between the ASBRs (including bypass tunnels, detours) SHOULD be configured as a set of strict interface hops.

Similarly, if the TE information of the ASBR links is not advertised into the corresponding region, then we can compute a TE LSP path only upto the exit ASBR in that AS (instead of being able to compute a path upto the entry ASBR in the next AS). While the solution does not mandate the flooding of TE information for basic operation; it is preferable to do so. Also, this solution requires an ASBR to support RSVP-TE signaling in case it is participating in inter-AS TE LSP setup.

In the above topology, for an LSP setup from CE1 to CE2, the FA-LSPs/LSP segments may be setup between C-D and potentially E-F. The Path message in this case traverses along CE1-A-B-C-D-E-F-CE2. In the RRO sent in the Resv message, the ASBRs which are ingress into the AS (like C, E, C', E') can record the interface address corresponding to the ASBR-ASBR link in the RRO.

Between the ASBRs regular RSVP-TE signaling procedures are carried out. In case the ASBRs (say B and C) are more than one hop away, then instead of creating RSVP state for every inter-AS LSP traversing B and C; one MAY decided to aggregate these requests by setting up a FA-LSP between the ASBRs to nest the inter-AS LSP requests. As per the definition in 4.1, boundary LSR B, by default is not a candidate to initiate a FA-LSP or LSP segment setup. But this behavior MAY be overridden by configuration. In this case, the zone between the ASBRs is treated as another region.

5. Nesting versus Stitching at region boundaries

An LSR at the region boundary, may either 'nest' several inter-region LSPs into a single intra-region FA-LSP, or it may choose to split the inter-region LSP and insert an intra-region LSP segment into the split. We refer to the latter action as 'stitching' in this document. This is a decision local to the region. Stitching may be done either due to a local configuration or due to technology (e.g., wavelength LSPs). While a boundary LSR in one region may choose to stitch, a boundary LSR in another region could choose to nest. e.g. one may choose to stitch the inter-area LSPs and nest the inter-AS LSPs or vice versa.

LSP stitching as proposed in the document is considered to be a special case of LSP hierarchy and may be used when hierarchy is just not possible. E.g. if the inter-region TE LSP is a fiber-switched LSP, then no further hierarchy is possible and this LSP may be stitched to an intra-region FSC LSP segment.

The signaling mechanisms for both stitching and nesting are basically

similar to that proposed in [LSP-HIER], except for additional flag bits signaled to indicate stitching behavior, as explained below. Also, in both cases the intra-region TE LSP is a different LSP from the inter-region TE LSP. This gives the region more autonomy over the placement and maintenance of it's intra-region FA-LSP or LSP segments. As explained in later sections, this has benefits like containing re-optimizations within a region. The main difference is that nesting facilitates scaling of both control and forwarding planes by reducing both control and forwarding state on the core LSRs in the region.

Since it is more scalable and more suitable to most packet-switched networks, nesting will be the prefered default behavior within a region. If, however, the region boundary LSR desires to do one-to-one stitching, then it MUST indicate this in the Path message for the intra-region LSP segment. This signaling is needed so that the egress LSR for the LSP segment knows in advance, how the ingress for the LSP segment plans to map traffic on to the LSP segment. This will allow it to allocate the correct label(s) as explained below. Also, so that the HE LSR can ensure that correct stitching actions were carried out at the egress LSR, a new flag has been defined below in the RRO subobject to indicate that the LSP segment may be used for stitching.

In order to request LSP stitching, we define a new flag bit in the Attributes Flags TLV of the LSP_ATTRIBUTES object defined in [RSVP-ATTRIBUTES]:

0x04: LSP stitching desired

This flag will be set in the Attributes Flags TLV of the LSP_ATTRIBUTES object in the Path message for the intra-region LSP segment by the HE LSR of the LSP segment (region boundary LSR) that desires LSP stitching. This flag SHOULD not be modified by any other LSRs in that region.

An intra-region LSP segment can only be used for stitching if appropriate label actions were carried out at the agress LSR of the LSP segment. In order to indicate this to the HE LSR for the LSP segment, the following new flag bit is defined in the RRO sub-object:

0x20 : LSP segment stitching ready

If an egress LSR receiving a Path message, supports the LSP_ATTRIBUTES object and the Attributes Flags TLV as well, and also recognizes the 'LSP stitching desired' flag bit, but cannot support the requested stitching behavior, then it MUST send back a PathErr message with an error code of "Routing Problem" and an error sub-code=16 "Stitching unsupported" to the HE LSR of the intra-region LSP

segment.

If an egress LSR receiving a Path message with the 'LSP stitching desired' flag set, recognizes the object, the TLV and the flag and also supports the desired stitching behavior, then it MUST allocate a non-NULL label for that LSP segment in the corresponding Resv message. Now, so that the HE LSR can ensure that the correct label actions will be carried out by the egress LSR and that the LSP segment can be used for stitching, the egress LSR MUST set the 'LSP segment stitching ready' bit defined in the RRO sub-object. Also, when the egress LSR for the LSP segment receives a Path message for an inter-region LSP using this LSP segment, it SHOULD first check if it is also the egress for the inter-region TE LSP. If the egress LSR is the egress for both the intra-region LSP segment as well as the inter-region TE LSP, and it requires Penultimate Hop Popping (PHP), then the LSR MUST send back a Resv refresh for the intra-region LSP segment with a new label corresponding to the NULL label. The egress LSR SHOULD always allocate a NULL label in the Resv message for the inter-region TE LSP.

Finally, if the egress LSR for the intra-region LSP segment supports the LSP_ATTRIBUTES object but does not recognize the Attributes Flags TLV, or supports the TLV as well but does not recognize this particular flag bit, then it SHOULD simply ignore the above request.

An ingress LSR requesting stitching SHOULD examine the RRO sub-object flag corresponding to the egress LSR for the intra-region LSP segment, to make sure that stitching actions were carried out at the egress LSR. It MUST NOT use the LSP segment for stitching if the 'LSP segment stitching ready' flag is cleared.

An ingress LSR stitching an inter-region LSP to an LSP segment MUST ignore any Label received in the Resv for the inter-region LSP.

Example: In case of inter-AS LSP setup from CE1 to CE2 as described in 4.3, let us assume that the ASBR C is doing one-to-one LSP stitching as described above. So when C receives the inter-AS LSP Path message, it will first initiate the setup of an intra-region LSP segment to D, if not already present. In the Path message for this LSP segment, C will set the "LSP stitching desired" flag in the Attributes Flags TLV of the LSP_ATTRIBUTES object. When D receives this Path message, D will allocate a non-NULL label (real label for swap action) in the Resv message for this LSP segment. Also, D will set the "LSP segment stitching ready" flag in the RRO subobject in the Resv message. Once the LSP segment is signaled successfully, C will then forward the Path message for the inter-AS TE LSP to D, which propagates it further. Eventually as the Resv message for the

inter-AS LSP traverses back from E to D and reaches D, D will remember to swap the LSP segment label with the label received for the inter-AS LSP from E. Also, D will itself allocate a NULL label in the Resv message for the inter-AS TE LSP and sends the Resv message to C. C ignores the Label object in the Resv message received from D for the inter-AS TE LSP and remembers to swap the label that it allocates in the inter-AS Resv message sent to B with the label that it had received from say, LSR R2 for the intra-AS LSP segment. In this manner, the inter-AS TE LSP is stitched to an intra-region LSP segment in AS2. In this example, if the LSP destination for the inter-AS LSP had been D, then if this is a packet-switched LSP and if D requires PHP, then on receiving the Path message for the inter-AS LSP, D will re-send a Resv message for the intra-region LSP segment to say R2, by changing the Label to a NULL label.

6. Fast Recovery support using MPLS TE Fast Reroute

[FAST-REROUTE] describes two methods for local protection for a TE LSP in case of node or link failure. This section describes how these mechanisms work with the proposed solution for inter-region TE LSP setup. Most of the behavior described in the above Fast Reroute document is directly applicable to the inter-region TE LSP setup case. When a FA-LSP or LSP segment is setup a priori it's local protection scheme SHOULD be pre-decided; if it is signaled dynamically then the desired protection scheme MAY be derived from the inter-region TE LSP.

6.1. Failure within a region (link or node failure)

In case of any failure affecting a FA-LSP or LSP segment, the existing local protection procedures for recovery as defined in [FAST-REROUTE] are applicable directly and will take care of local protection of the FA-LSP/LSP segment without requiring any new extensions. Since an intra-region FA-LSP/LSP segment transports the inter-region TE LSP traffic, the inter-region traffic automatically gets protected by protecting the corresponding the intra-region FA-LSP/LSP segment. However, these local protection schemes using backups are usually temporary and the HE LSR for the protected LSP needs to be usually notified so that it can compute a new path around the failure.

The failure notification (RSVP Path Error/Notify message "Tunnel Locally Repaired") for the FA-LSP/LSP segment SHOULD be sent to the respective ingress LSR for that intra-region FA-LSP/LSP segment in that region. The ingress LSR for the FA-LSP/LSP segment will then try to re-route the FA-LSP/LSP segment around the failure, and the inter-region LSPs using the FA-LSP/LSP segment will start taking the new

path in that region. However, in case the HE LSR in that region is unable to find a path around the failure to re-route the intra-region FA-LSP/LSP segment, then a failure and repair notification stated above (PathErr) for all the affected inter-region TE LSPs MAY be propagated to the upstream region towards the HE LSR for the inter-region TE LSP. This could be a local policy decision. Other region boundary LSRs along the way could intercept the error message to do some kind of crankback. This two-phase approach tries to handle the failure first locally within a region as far as possible by intercepting the error notification at the region boundary LSR and re-routing the intra-region LSP. Only if that fails, do we propagate the error notification further upstream.

Alternatively, instead of intercepting the error notifications and following the above two-phase approach, one may choose to always send back error notifications back to the HE LSR for the inter-region TE LSP in the originating region. This could be a local policy decision. In any case, the TE LSP SHOULD be re-routed around the failure using the "make-before-break" approach.

Example: In case of the inter-AS TE LSP setup described in Section 4.3, let us assume that the FA-LSP/LSP segment traverses R2 in AS2, and is node-protected against the failure of R2. In that case, when R2 or the corresponding link to R2 fails, then the traffic will be locally protected by the corresponding detour or backup path (depending on the local protection scheme) associated with the protected FA-LSP/LSP segment. When the PathErr/Notify message "Tunnel Locally Repaired" reaches C, C may find a new path for the FA-LSP/LSP segment and signal it. During this time, the FA-LSP/LSP segment along the old path was locally repaired and so traffic will continue to take the backup path around the failure. Once the new path for the FA-LSP/LSP segment is successfully signaled the traffic is switched to the new path and the old path is torn down. Note that since the inter-region traffic is sent along the FA-LSP/LSP segment, that traffic has been protected as well.

6.2. Failure of link at region boundaries (ASBR-ASBR)

In the above example for inter-AS TE LSP setup, let us consider the failure of link B-C. The LSPs traversing these links are the inter-region LSPs. Depending on the type of local protection desired, B which is the PLR will have either a bypass tunnel or a detour along a path disjoint from link B-C. In order to protect the inter-region LSP(s) from a B-C link failure, a NHOP bypass tunnel (if the "facility backup" method is used) or a set of detours (one for every inter-region protect TE LSP) MUST be pre-configured at B. Since the zone between the ASBRs is not usually TE-enabled, it would usually not be possible for the ASBR to compute a complete path around the

protected link B-C. So backup paths with strict ERO configuration must be configured. If, however, the ASBRs do have the TE-information for the ASBR-ASBR links, then depending on the topology, a minimum path configuration specifying the loose hops may suffice.

So, in either case, the PLR would select the bypass tunnel or detour for a protect TE LSP, terminating at the NHOP. The NHOP is determined by looking at the RRO object received in the Resv for the protected inter-region TE LSP. e.g. B will select a bypass tunnel terminating on C by looking at the RRO at B, which would be C-D-E-F or <link address on C>-D-E-F.

Also, the node where the backup intersects the protected LSP (MP : Merge Point) may either be reachable directly from the PLR or it may reside in the other AS. e.g. the bypass could either take path B-B'-C'-C in which case the MP is C itself or the MP may also be D if the only available bypass tunnel path is B-B'-C'-D-C. Therefore, mechanisms like 'auto-discovery' of next-hop LSR and ERO loose-hop expansion with partial CSPF computation to the first reachable LSR may also be applicable to the backup path computation.

In this solution, the MP for the inter-region LSP will always be a region boundary LSR. This is because, the FA-LSP/LSP segment is a different LSP (different session) from the inter-region LSP, so the inter-region LSP backup can only intersect the protected LSP path at the region boundary LSRs.

6.3. Failure of node at region boundary

Let us again consider the example topology in 4.3 for inter-As LSP setup. This gives rise to the following scenarios for protection:

6.3.1. Protecting the last hop boundary LSR (ASBR) in a region

Example: protecting against failure of node D in AS2.

Considering the FA-LSP/LSP segment terminating at D, this is the last hop for the FA-LSP/LSP segment, so there can be no node-protection for D via the FA-LSP/LSP segment. However, as far as the inter-region LSP is concerned, its path is along A-B-C-D-E-F and the FA-LSP/LSP segment between C and D is a link. So for protecting against D's failure, C is the PLR and C will setup a bypass tunnel to the NNHOP for this LSP, which is E. Again the NNHOP is determined by examining the received RRO for the inter-region LSP. So one or more bypass tunnels following C-D'-E must be configured on C to protect against node D's failure. It is worth mentioning that this may add some additional constraints on the backup path since the bypass tunnel path needs to be diverse from the C-D-E path instead of just being

diverse from the X-D-E path where X is the upstream neighbor of D. The consequences are that the path is likely to be longer and if bandwidth protection is desired for instance ([FACILITY-BACKUP] more resources may be reserved in AS2 than necessary.

6.3.2. Protecting against the LSR at the entry to the region

Example: protecting against the failure of LSR C in AS2.

Again, in this case, the FA-LSP/LSP segment offers no protection; so one or more backups MUST be configured from the previous hop LSR in the inter-region LSP, i.e. B, to the NNHOP with respect to the inter-region LSP, i.e. D. A bypass tunnel B-B'-C'-D would protect against C's failure. Depending on whether auto-discovery mechanisms are available, and whether TE-information for ASBR-ASBR links is available, the configuration required on the PLR for the backup could be minimal or could require specifying the entire path. The same constraints as mentioned above apply in this case.

When the FA-LSP/LSP Segment is unnumbered, the Router ID of the boundary LSR will be recorded in the RRO object (see [RSVP-UNNUM]). However, if the FA-LSP/LSP segment is numbered, then bypass tunnel selection to protect an inter-region TE LSP with Fast Reroute "facility backup" ([FAST-REROUTE]) against the failure of an ASBR-ASBR link or an ASBR node would require the support of [NODE-ID].

7. Re-optimization of inter-region LSPs

In this solution, re-optimization is treated as a local matter to any region for both stitching as well as nesting case. The main reason being that the inter-region TE LSP is a different LSP (and therefore different RSVP session) from the intra-region FA-LSP or LSP segment in a region. Therefore, re-optimization in a region is done by re-optimizing the intra-region FA-LSPs/ LSP segments. Since the inter-region TE LSPs are transported using FA-LSPs/LSP segments across a region, optimality of the inter-region LSPs in a region is dependent on the optimality of the corresponding FA-LSPs/LSP segments. If, after an inter-region LSP is setup, a more optimal path is available within a region; the corresponding FA-LSP(s)/LSP segment(s) will be re-optimized using "make-before-break" techniques discussed in [RSVP-TE]. Re-optimization of the FA-LSP/LSP segment automatically re-optimizes the inter-region LSPs that the FA-LSP/LSP segment transports. Re-optimization parameters like frequency of re-optimization, criteria for re-optimization like metric or bandwidth availability; etc can vary from one region to another and can be configured as required, per FA-LSP/LSP segment if it is pre-configured or based on some global policy within the region.

So, in this scheme, since each region takes care of re-optimizing its own FA-LSPs/LSP segments, and therefore the corresponding inter-region TE LSPs, the make-before-break can happen locally and is not triggered by the HE LSR for the inter-region LSP. So, no additional RSVP signaling is required for LSP re-optimization and re-optimization is transparent to the HE LSR of the inter-region TE LSP.

If, however, an operator desires to manually trigger re-optimization at the HE LSR for the inter-region LSP, then this solution does not prevent that. A manual trigger for re-optimization at the HE LSR, SHOULD force a re-optimization thereby signaling a "new" path for the same LSP (along the optimal path) making use of the make-before-break procedure. In response to this new setup request, the boundary LSR may either initiate new LSP segment setup, in case the inter-region TE LSP is being stitched to the intra-region LSP segment or it may select an existing FA-LSP in case of nesting. When the LSP setup along the current optimal path is complete, the head end should switchover the traffic onto that path and the old path is eventually torn down. Note that the HE LSR does not know a priori whether a more optimal path exists. Such a manual trigger from the HE LSR of the inter-region TE LSP is, however, not considered to be a frequent occurrence.

8. Specific requirements for inter-AS TE LSP setup

In this section, we discuss the proposed solution in light of some of the requirements specific to inter-AS traffic engineering as listed in [INTER-AS-TE-REQTS]. This covers the Inter-AS TE requirements for a single SP administrative domain as well as across multiple SP administrative domains. Some of the requirements may have already been covered in previous sections and will not be discussed here again. Also, while this proposal applies to Inter-Area MPLS TE as well, the inter-AS requirements are seen as a super-set of the Inter-Area MPLS TE requirements described in [INTER-AREA-TE-REQTS], so this will not be discussed again in an inter-area TE context.

8.1. Support of diversely routed inter-region TE LSP

There may be a need to support diversely routed paths for an inter-region TE LSP, either for path protection or load balancing. RRO plays an important role in determining the path along which the TE LSP traverses. In case of an intra-area TE LSP, today this tells us the "links" traversed by a path. This information is used to compute other disjoint paths by excluding the above links in CSPF path computation. One must just mention that this method (also called "2 steps approach") does not always guaranty to find two diverse paths

even if such paths exist. Also this simple algorithm does not allow to find two paths such that the sum of their cost is minimal. In case of an inter-region path setup, it is important to note that CSPF computation may be distributed over different LSRs and also the path represented by the RRO, need not represent physical links, they could be other FA-LSPs/LSP segments. E.g. in 4.1 let us assume that the primary path for the LSP from A to D takes A-B-C-D. So a diverse path for this LSP may be signaled as A-B'-C'-D. The information for links to "exclude" could be signaled as specified in [RSVP-CONSTRAINTS] as a constraint or using the mechanisms described in [EXCLUDE-ROUTES].

8.2. Inter-AS MPLS TE Management

The failure detection and isolation mechanisms proposed by [LSPING] can be applied to inter-AS TE management as well. This solution does not present any issues for support of either [LSPING] or [MPLS-TTL].

8.3. Confidentiality

As mentioned in [INTER-AS-TE-REQTS], an inter-AS LSP can cross different AS boundaries. This solution does not require another region (AS) to expose any of its local addresses within the AS to another AS for any signaling needs. Even for local protection using Fast-Reroute no special requirements are imposed on the nodes within the AS to expose their addresses. This was discussed under Section 6.

8.4. Policy Control at the AS boundaries

Since this solution is based on per region FA-LSP/LSP segment based approach, it inherently provides more control to an SP in setting up FA-LSPs/LSP segments and deciding their attributes within his AS. Other than that, as stated in [INTER-AS-TE-REQTS], there may be other inter-AS TE agreements made with other ASes and implemented as local policy on the ASBR. Among other behaviors stated in the above document, mapping of constraints and other attributes from the inter-region LSP to the intra-region FA-LSP/LSP segment, nesting versus stitching behavior, overriding the default ASBR behavior with respect to signaling of FA-LSP/LSP segment; etc can also be exercised based on local policy configuration.

8.5. Scalability and Extensibility

This draft provides a common solution to be used for LSPs traversing regions of different types and hence can be used not only for inter-area and inter-AS TE, but also for setting up LSP paths across GMPLS overlay regions.

The nesting solution especially, adopts a scalable approach based on

LSP hierarchy, to aggregate several inter-region TE LSP requests onto a few intra-region FA-LSPs, thereby reducing both control state as well as forwarding state on the transit LSRs within a region.

9. Security Considerations

This document raises no new security concerns for RSVP. Existing security features can be used for signaling inter-region TE LSPs.

10. Intellectual Property Considerations

Juniper Networks may have intellectual property rights claimed in regard to some of the specification contained in this document.

11. Acknowledgements

Many thanks to Yakov Rekhter, Kireeti Kompella, Pedro Marques and Ina Minei for their initial discussions that contributed to the solution proposed in the draft. We would also like to thank Yakov Rekhter and Kireeti Kompella for their useful comments and suggestions towards the draft.

12. References

[LSP-HIER] Kompella K., Rekhter Y., "LSP Hierarchy with Generalized MPLS TE", (work in progress).

[FAST-REROUTE] Ping Pan, et al, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", (work in progress)

[INTER-AS] Vasseur, Zhang, "Inter-AS MPLS Traffic Engineering", (work in progress).

[GMPLS-OVERLAY] G. Swallow et al, "GMPLS RSVP Support for the Overlay Model", (work in progress).

[RSVP-ATTRIBUTES] Ferrel A. et al, "Encoding of Attributes for Multiprotocol Label Switching (MPLS) Label Switched Path (LSP) Establishment Using RSVP-TE", (work in progress).

[INTER-AREA] Kompella et al, "MPLS Inter-area Traffic Engineering", (work in progress).

[INTER-AS-TE-REQTS] Zhang et al, "MPLS Inter-AS Traffic Engineering

requirements", (work in progress).

[INTER-AREA-TE-REQTS] Boyle J., "Requirements for support of Inter-Area and Inter-AS MPLS Traffic Engineering", (work in progress).

[RSVP-TE] Awduche, et al, "Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.

[RFC2702] Awduche, et al, "Requirements for Traffic Engineering Over MPLS", RFC2702, September 1999.

[RSVP-UNNUM] Kompella K., Rekhter Y., "Signalling Unnumbered Links in RSVP-TE", RFC 3477, January 2003.

[OSPF-TE] Katz, D., Yeung, D., Kompella, K., "Traffic Engineering Extensions to OSPF", RFC 3630 (Updates RFC 2370), September 2003.

[ISIS-TE] Smit, H., Li, T., "IS-IS extensions for Traffic Engineering", (work in progress).

[FACILITY-BACKUP] Vasseur et al, "MPLS Traffic Engineering Fast reroute: bypass tunnel path computation for bandwidth protection", (work in progress).

[NODE-ID] Vasseur, Ali and Sivabalan, "Definition of an RRO node-id subobject", (work in progress).

[RSVP-CONSTRAINTS] Kompella, K., "Carrying Constraints in RSVP", (work in progress).

[EXCLUDE-ROUTES] Lee C.Y, Farrel A. et al, "Exclude Routes - Extension to RSVP-TE", (work in progress).

[LSPING] Kompella, K., Pan, P., Sheth, N., Cooper, D.,Swallow, G., Wadhwa, S., Bonica, R., " Detecting Data Plane Liveliness in MPLS", (work in progress).

[MPLS-TTL], Agarwal, et al, "Time to Live (TTL) Processing in MPLS Networks", RFC 3443 (Updates RFC 3032) ", January 2003.

13. Author Information

Arthi Ayyangar
Juniper Networks, Inc.
1194 N.Mathilda Ave
Sunnyvale, CA 94089
USA
e-mail: arthi@juniper.net

14. Full Copyright Notice

15. Acknowledgement

Funding for the RFC Editor function is currently provided by the Internet Society.