JP Vasseur (Editor)
Cisco Systems, Inc.
Arthi Ayyangar (Editor)
Juniper Networks

IETF Internet Draft
Expires: August, 2004

February, 2004

draft-vasseur-ccamp-inter-area-as-te-00.txt

Inter-area and Inter-AS MPLS Traffic Engineering

Status of this Memo

Abstract

This document proposes a set of signaling and routing mechanisms to establish and maintain generalized (packet and non-packet) MPLS Traffic Engineering Label Switched Path (MPLS TE LSPs) that span multiple areas or Autonomous Systems.. Each mechanism is described along with its applicability to the set of requirements.

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].

Content

1. Terminology

LSR: Label Switch Router

LSP: MPLS Label Switched Path

PCE: Path Computation Element. An LSR in charge of computing TE LSP path for which it is not the Head-end. For instance, an ABR (inter-area) or an ASBR (Inter-AS) can play the role of PCE.

PCC: Path Computation Client (any head-end LSR) requesting a path computation from the Path Computation Element.

Local Repair: local protection techniques used to repair TE LSPs quickly when a node or link along the LSPs path fails.

Protected LSP: an LSP is said to be protected at a given hop if it has one or multiple associated backup tunnels originating at that hop.

Bypass Tunnel: an LSP that is used to protect a set of LSPs passing over a common facility.

PLR: Point of Local Repair. The head-end of a bypass tunnel.

MP: Merge Point. The LSR where bypass tunnels meet the protected LSP.

NHOP Bypass Tunnel: Next-Hop Bypass Tunnel. A backup tunnel which bypasses a single link of the protected LSP.

NNHOP Bypass Tunnel: Next-Next-Hop Bypass Tunnel. A backup tunnel which bypasses a single node of the protected LSP.

Fast Reroutable LSP: any LSP for which the "Local protection desired" bit is set in the Flag field of the SESSION_ATTRIBUTE object of its

Path messages or signaled with a FAST-REROUTE object.

CSPF: Constraint-based Shortest Path First.

Inter-AS MPLS TE LSP: A TE LSP whose head-end LSR and tail-end LSR do not reside within the same Autonomous System (AS), or whose head-end LSR and tail-end LSR are both in the same AS but the TE LSPÆs path may be across different ASes. Note that this definition also applies to TE LSP whose Head-end and Tail-end LSRs reside in different sub-ASes (BGP confederations).

Inter-area MPLS TE LSP: A TE LSP where the head-end LSR and tail-end LSR do not reside in the same area or both the head-end and tail end LSR reside in the same area but the TE LSP transits one or more different areas along the path.

ABR Routers: routers used to connect two IGP areas (areas in OSPF or L1/L2 in IS-IS)

Interconnect routers or ASBR routers: routers used to connect together ASes of a different or the same Service Provider via one or more Inter-AS links.

Boundary LSR: a boundary LSR is either an ABR in the context of inter-area MPLS TE or an ASBR in the context of inter-AS MPLS TE.

TED: MPLS Traffic Engineering Database

In this document, the term inter-area/AS TE LSP refers to an inter-area or an inter-AS MPLS Traffic Engineering Label Switched Path.

The notion of ææTE LSP nestingÆÆ refers to the ability to carry one or more inter-area/AS TE LSPs within another intra-area/AS TE LSP by using the MPLS label stacking property at the intra-area/AS outer TE LSPÆs head-end LSR. On the other hand, ææstitching a TE LSPÆÆ means to split an inter-area/AS TE LSP and insert a different intra-area/AS LSP, into the split. This implies a label swap operation at the stitching point (head-end of the intra-area/AS TE 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. We use the term ææLSP segmentÆÆ in the context of LSP stitching (when one LSP is split and another LSP is inserted into the split).

2. Introduction

Considering the set of requirements for inter-area and inter-AS Traffic Engineering respectively listed in [INTER-AREA-TE-REQS] and [INTER-AS-TE-REQS], this document proposes a set of mechanisms to establish and maintain MPLS Traffic Engineering Label Switched Paths that span multiple areas in the context of inter-area MPLS TE or multiple ASes or sub-ASes (with BGP confederations) in the context of inter-AS MPLS TE. The mechanisms proposed in this document could also be applicable to MPLS TE domains other than areas and ASes as well.

According to the wide set of requirements defined in [INTER-AS-TE-REQS] and [INTER-AREA-TE-REQS], coming up with a single solution covering all the requirements is certainly possible but may not be desired: indeed, as described in [INTER-AS-TE-REQS] the spectrum of deployment scenarios is quite large and designing a solution addressing the super-set of all the requirements would lead to provide a rich set of mechanisms not required in several cases. Depending on the deployment scenarios of a SP, certain requirements stated above may be strict while certain other requirements may be relaxed.

There are two aspects to a TE LSP setup: the TE LSP path computation and the signaling. There are different ways in which path computation

for an inter-area/AS TE LSP could be done. For example, if the requirement is for an end-to-end constraint-based shortest path for the inter-area/AS TE LSP, then a mechanism using one or more distributed PCEs could be used to obtain an optimal path across different areas/ASes. Alternatively, one could also use some static or discovery mechanisms to determine the next boundary LSR per area/AS as the inter-area/AS TE LSP is being signaled. Other offline mechanisms for path computation are not precluded either. Depending on the requirements of the SP, one may adopt either of these techniques for inter-area/AS path computation. Hence, once the TE LSP path is obtained, this document provides three different types of inter-area/inter-AS TE LSP which are signaled by different means: contiguous, nested or stitched. Depending on the needs of the SP networks, one may choose either of these mechanisms to signal the TE LSP. In case of inter-AS (inter-provider) TE LSP setup, since different SPs may have different needs and may choose different TE policies in their network, this document provides a way to communicate some requirements that the head-end LSR originating the TE LSP may have for the ASes that the TE LSP transit. Also, with TE LSPs crossing AS boundaries or administrative domains, it is assumed that there will be some form of Policy control at the administrative

boundaries.

In section 11, the applicability and evaluation criteria of each solution proposed in this document with respect to the set of requirements defined in [INTER-AS-TE-REQS] and [INTER-AREA-TE-REQS] are analyzed.

3. General assumptions

In the rest of this document, we make the following set of assumptions:

1) Assumptions common to inter-area and inter-AS TE:

- Each area or AS 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). An AS may itself be composed of several other sub-AS(es) (BGP confederations) or areas/levels.

- The inter-area/AS LSPs are signaled using RSVP-TE ([RSVP-TE]).

- The path (ERO) for the inter-area/AS TE LSP traversing multiple areas/ASes may be signaled as a set of (loose and/or strict) hops. The hops may identify:
  - The complete strict path end to end across different areas/ASes
  - The complete strict path in the source area/AS followed by boundary LSRs (and domain identifiers, e.g. AS numbers)
  - The complete list of boundary LSRs along the path
  - The current boundary LSR and the LSP destination

In this case, the set of (loose or strict) hops can either be statically configured on the Head-end LSR or dynamically computed. In the former case, the resulting path is statically configured on the Head-end LSR. In the latter case (dynamic computation), two methods described in this document can be used:
  - A distributed path computation involving some PCEs (e.g ABR/ASBR) resulting in globally optimal path consisting of strict and/or loose hops,
  - Some Auto-discovery mechanism based on BGP and/or IGP information yielding the next-hop boundary LSR (ABR/ASBR) along the path as the LSP is being signaled, along with crankback mechanisms.

- Furthermore, the boundary LSRs are assumed to be capable of performing local path computation for expansion of a loose next-hop in the signaled ERO if the path is not signaled by the head-end LSR as a set of strict hops or if the strict is for example an AS number. This can be done by performing a CSPF computation to that loose hop, instead of to the LSP destination or by making use of some PCEs. In any case, no topology or resource information needs to be distributed between areas/ASes, which is critical to preserve IGP/BGP scalability.

- The paths for the intra-area/AS FA-LSPs or LSP segments or for a contiguous TE LSP within the area/AS, may be pre-configured or computed dynamically based on the arriving inter-area/AS LSP setup request; depending on the requirements of the transit area/AS. Note that this capability is explicitly specified as a requirement in [INTER-AS-TE-REQS]. 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-area/AS FA-LSP/LSP segment are also pre-configured. Some local algorithm can be used on the head-end LSR of a FA-LSP to dynamically adjust the FA-LSP bandwidth based on the cumulative bandwidth requested by the inter-area/AS TE LSPs. It is RECOMMENDED to use a threshold triggering mechanism to avoid constant bandwidth readjustment as inter-area/AS TE LSPs are set up and torn down.

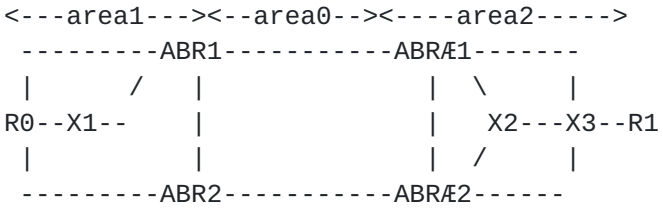
- While certain constraints like bandwidth can be used across different areas/ASes, certain other TE constraints like resource affinity, color, metric, etc. as listed in [RFC2702] could be translated at areas/ASes boundaries. If required, it is assumed that, at the area/AS boundary LSRs, there will exist some sort of local mapping based on offline policy agreement, in order to translate such constraints across area/AS boundaries. It is expected that such an assumption particularly applies to inter-AS TE: for example, the local mapping would be similar to the Inter-AS TE Agreement Enforcement Polices stated in [INTER-AS-TE-REQTS].

- When an area/AS boundary LSR at the exit of an area/AS receives a TE LSP setup request (Path message) for an inter-area/AS TE LSP, then if this LSP had been nested or stitched at the entry area/AS 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-area/AS 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 are used here.

2) Example of topology for the inter-area TE case:



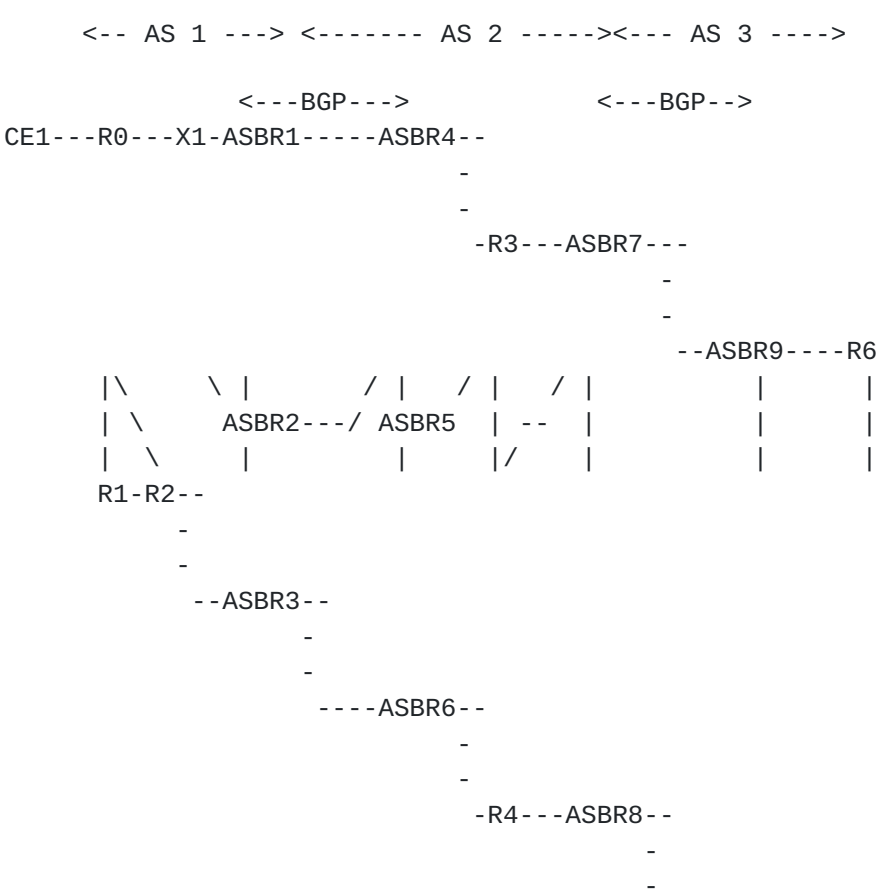
- ABR1, ABR2, ABRÆ1 and ABRÆ2 are ABRs
- X1: an LSR in area 1
- X2, X3: LSRs in area 2

Note:

- The terminology used in the example corresponds to OSPF but the set of mechanisms proposed in this document equally applies to IS-IS.
- Just a few routers in each area are depicted in the diagram above for the sake of simplicity.

3) Example of topology for the inter-AS TE case:

We will consider the following general case, built on a superset of the various scenarios defined in [INTER-AS-TE-REQS]:



<======= Inter-AS TE LSP(LSR to LSR)===========>
or

<======== Inter-AS TE LSP (CE to ASBR =>

or

<================= Inter-AS TE LSP (CE to CE)===============>

The diagram above covers all the inter-AS TE deployment cases described in [INTER-AS-TE-REQS].

Assumptions:

- Three interconnected ASes, respectively AS1, AS2, and AS3. Note that AS3 might be AS1 in some scenarios described in [INTER-AS-TE-REQS],

- The various ASBRs are BGP peers, without any IGP running on the single hop link interconnecting the ASBRs,

- Each AS runs an IGP (IS-IS or OSPF) with the required IGP TE extensions (see [OSPF-TE] and [IS-IS-TE]). In other words, the ASes are TE enabled,

- Each AS can be made of several areas. In this case, the TE LSP will rely on the inter-area TE techniques to compute and set up a TE LSP traversing multiple IGP areas. For the sake of simplicity, each routing domain will be considered as single area in this document, but the solutions described in this document does not prevent the use of multi-area techniques.

- An inter-AS TE LSP T1 originated at R0 in AS1 and terminating at R6 in AS3,

- An inter-AS TE LSP T2 originated at CE1 (connected to R0) and terminating at CE2 (connected to R7),

- A set of backup tunnels:

o B1 from X1 to ASBR4 following the path X1-ASBR2-ASBR4 and protecting against a failure of the ASBR1 node,

o B2 from ASBR1 to ASBR4 following the path ASBR1-ASBR2-ASBR4 and protecting against a failure of the ASBR1-ASBR4 link,

o B3 from ASBR1 to R3 following the path ASBR1-ASBR2-ASBR3-ASBR6-ASBR5-R3 and protecting against a failure of the ASBR4 node.

o B4 from ASBR1 to ASBR7 following the path ASBR1-ASBR2-ASBR3-ASBR6-R4-ASBR7 and protecting against a failure of the ASBR4 node.

- In the example above, ASBR1, ASBR8 and ASBR9 could have the function of PCE for respectively the ASes 1, 2 and 3 (the notion of PCE applies to the scenario 2 of this document).

4. Notion of contiguous, nested and stitched TE LSPs

A contiguous TE LSP is defined as a TE LSP spanning multiple IGP areas/levels or ASes, which must be considered as a unique end-to-end TE LSP. By contrast, a stitched or nested TE LSP is made of up multiple

LSP pieces within each area/AS which are either stitched or nested together at area/AS boundaries to form an inter-area/AS TE LSP.

In case of a contiguous TE LSP, it is expected to provide more control at the head-end LSR that originates the inter-area/AS TE LSP. On the other hand, in case of the stitched or nested TE LSP, the control of the TE LSP is performed on a per-area or per-AS basis. This difference is possible because in the latter case (stitching and nesting) the intra-area/AS TE LSP is a different TE LSP from the inter-area/AS TE LSP. The term æLSP segmentÆ is used when one TE LSP is split and another LSP is inserted into the split. And the term FA-LSP is used when one or more TE LSPs are carried within another LSP. Both stitched and nested TE LSPs are signaled using mechanisms defined in [LSP-HIER].

While signaling a contiguous TE LSP is different from signaling a stitched TE LSP, in the forwarding plane at the boundary LSR, both involve a label swap operation. However, nesting multiple inter-area/AS LSPs into another intra-area/AS LSP, is done using the MPLS label stacking construct.

It is desirable in mixed environments making use of different

techniques (contiguous, stitched or nested TE LSPs) to provide the ability for the head-end LSR of the inter-area/AS TE LSP to signal its requirement regarding the nature of the inter-area/AS TE LSP (contiguous, stitched, nested) on a per-LSP basis. For the sake of illustration, a Head-end LSR, may desire to prevent stitching or nesting for a traffic sensitive inter-area/AS TE LSPs that require a path control on the head-end LSR. On the other hand, the head-end LSR may decide to avoid any tight control.[LSP-ATTRIBUTES] defines the format of the attribute flags TLV included in the LSP-ATTRIBUTE object carried in an RSVP Path message which is used for the purpose of signaling the inter-area/AS TE LSP characteristics.

The following bits of the attribute flags TLV is defined for this purpose:

0x01: Contiguous LSP required bit: this flag is set by the head-end LSR that originates the inter-AS/area TE LSP if it desires a contiguous end-to-end TE LSP. When set, this indicates that a boundary LSR MUST not perform any stitching or nesting 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 stitching or nesting. A mid-point LSR not supporting contiguous TE LSP MUST send a Path Error message upstream with error sub-code=17 ææContiguous LSP type not supportedÆÆ. This bit MUST not be modified by any downstream node.

Additionally, in case of a non-contiguous inter-area/AS LSP, if the inter-area/AS TE LSP is being stitched into another intra-area/AS TE LSP, it is sometimes required to explicitly signal the stitching behavior in the ææintra-area/ASÆÆ LSP segment within that area/AS. The following bit of the attributes flags TLV is defined for this purpose:

0x02: LSP stitching required bit: this flag is set by the boundary LSR for the intra-area/AS LSP segment which is local to that area/AS. This flag SHOULD not be modified by any other LSR in that area/AS. If the egress LSR for the intra-area/AS LSP segment does not understand this flag then it will simply forward the object unmodified and will send a Path Error message upstream with error sub-code=16.

Further signaling details for TE LSP stitching are described in section 5.3.

Note: in some cases, it may be desirable for the head-end LSR to exert some control on the ability for the boundaries LSRs to make use of crankback. See [CRANKBACK] for the definition of those bits. When

crankback is allowed, the boundary LSR can either decide to forward the Path Error message upstream to the head-end LSR or try to select another egress boundary LSR (which is also referred to as crankback). When crankback is not allowed, a boundary LSR, when receiving a Path Error message from a downstream boundary LSR MUST propagate the Path Error message up to the inter-area/AS head-end LSR.

5. Scenario 1: Next-hop resolution during inter-area/AS TE LSP set up (per-area/AS path computation)

Regardless of whether the inter-area/AS TE LSP is a contiguous or stitched or nested TE LSP, a similar set of mechanisms for local TE LSP path computation (next hop resolution) and setup can be used.

When an ABR/ASBR receives a Path message with a loose next-hop in the ERO, then it carries out the following actions:

1) It checks if the loose next-hop is accessible via the TED. If the loose next-hop is not present in the TED, 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 an ABR/ASBR to get to the next-hop. In the absence of an auto-discovery mechanism, the ABR/ASBR should be the loose next-hop in the ERO and hence should be accessible via the TED, otherwise path computation for the inter-area/AS TE LSP will fail.

2) If the next-hop boundary LSR is present in the TED.

a) Case of a contiguous TE LSP (ææContiguous LSP required bitÆÆ of the attribute flags TLV included in the LSP-ATTRIBUTE object is set). In that case, the ABR/ASBR just performs an ERO expansion after having computed the path to the next loose hop (ABR/ASBR) that obeys the set of required constraints. If no path satisfying the set of constraints can be found then a Path Error MUST be sent for the inter-area/AS TE LSP.

b) Case of stitched or nested LSP (ææContiguous LSP required bitÆÆ of the attribute flags TLV included in the LSP-ATTRIBUTE object is cleared).

i) if this ABR/ASBR (receiving the LSP setup request) is a candidate LSR for intra-area 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-area/AS 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-area/AS 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-area/AS LSP. Similarly, if selection of a preconfigured FA-LSP/LSP segment fails and local policy prevents dynamic FA-LSP/LSP segment setup, then a PathErr is sent upstream for the inter-area/AS TE LSP.

ii) If, however, this boundary LSR is not a FA-LSP/LSP segment candidate, then it SHOULD simply compute a CSPF path up to 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 boundary LSR receives a Path message with strict next-hop.

5.1. Example with an inter-area TE LSP (based on the assumption described in section 3).

In this example, R0 sets up an inter-area TE LSP T1 to R1.

5.1.1. Case 1: T1 is a contiguous TE LSP

When the path message reaches ABR1, it first determines the egress LSR from its area 0 along the LSP path (say ABRÆ1), either directly from the ERO (if for example the next hop ABR is specified as a loose hop in the ERO) or by using some constraint-aware auto-discovery mechanism.

In the former case, every inter-AS TE LSP path is defined as a set of loose and strict hops but at least the ASBRs traversed by the inter-AS TE LSP MUST be specified as loose hops on the Head-End LSR.

- Example 1 (set of strict hops end to end): R0-X1-ABR1-ABRÆ1-X2-X3-R1

- Example 2 (set of loose hops): R0-ABR1(loose)-ABRÆ1(loose)-R1(loose)
- Example 3 (mix of strict and loose hops): R0-X1-ASBR1-ABRÆ1(loose)-X2-X3-R1

At least, the set of ABRs from the TE LSP head-end to the Tail-End MUST be present in the ERO as a set of loose hops. Optionally, a set of paths can be configured on the head-end LSR, ordered by priority. Each priority path can be associated with a different set of constraints. Typically, it might be desirable to systematically have a last resort option with no constraint to ensure that the inter-area TE LSP could always be set up if at least a path exist between the inter-area TE LSP source and destination. Note that in case of set up failure or when an RSVP Path Error is received indicating the TE LSP has suffered a failure, an implementation might support the possibility to retry a particular path option a specific amount of time (optionally with dynamic intervals between each trial) before trying a lower priority path option. Any path can be defined as a set of loose and strict hops. In other words,
in some cases, it might be desirable to rely on the dynamic path computation in some area, and exert a strict control on the path in other areas (defining strict hops).

Example of configuration of T1 on R0 in dynamic mode: T1 Path: R0-R6(loose)

Once it has computed the path up to the next ABR, ABR1 sends the Path message for the inter-area TE LSP to ABRÆ1. ABRÆ1 then repeats the exact same procedures and the Path message for the inter-area TE LSP will reach the destination R1. If ABRÆ1 cannot find a path obeying the set of constraints for the inter-area TE LSP, then ABRÆ1 MUST send a PathErr message to ABR1. Then ABR1 can in turn select another egress boundary LSR (ABRÆ2 in the example above) if crankback is allowed for this inter-area TE LSP (see [CRANBACK]). If crankback is not allowed for that inter-area TE LSP or if ABR1 has been configured not to perform crankback, then ABR1 MUST forward a PathErr up to the inter-area head-end LSR (R0) without trying to select another egress LSR.

5.1.2. Case 2: T2 is a stitched or nested TE LSP

When the path message reaches ABR1, it first determines the egress LSR from its area 0 along the LSP path (say ABRÆ1), either directly from the ERO or by using some constraint-aware auto-discovery mechanism.

ABR1 will check if it has a FA-LSP or LSP segment to ABRÆ1 matching the constraints carried in the inter-area Path message. If not, ABR1 will setup a FA-LSP or LSP segment from ABR1 to ABRÆ1. Note that once the FA-LSP/LSP segment is setup, it may be advertised as a link within that area (see [LSP-HIER]) (area 0 in this example). The FA-LSP or LSP segment could have also been pre-configured.

If the inter-area LSP is a packet LSP and ABR1 desires to do one-to-one stitching, then it will signal this explicitly in the Path message for the intra-area LSP segment as described in section 5.3.

Also, there could be multiple FA-LSPs/LSP segments between ABR1 and ABRÆ1. So, ABR1 needs to select one FA-LSP/LSP segment from these, for the inter-area LSP through area 0. The mechanism and the criterion used to select the FA-LSP/LSP segment is local to ABR1 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-area/AS LSP requests. And it may select the LSP segments (stitching) only for some specific inter-area LSPs.

Once it has selected the FA-LSP/LSP segment for the inter-area LSP, using the signaling procedures described in [LSP-HIER], ABR1 sends the Path message for inter-area TE LSP to ABRÆ1. Note that irrespective of whether ABR1 does nesting or stitching, the Path message for the inter-area TE LSP is always forwarded to ABRÆ1. ABRÆ1 then repeats the exact same procedures and the Path message for the inter-area TE LSP will reach the destination R1. If ABRÆ1 cannot find a path obeying the set of constraints for the inter-area TE LSP, then ABRÆ1 MUST send a PathErr message to ABR1. Then ABR1 can in turn either select another FA-LSP/LSP segment to ABRÆ1 if such an LSP exists or select another egress boundary LSR (ABRÆ2 in the example above) if crankback is allowed for this inter-area TE LSP (see [CRANBACK]). If crankback is not allowed for that inter-area TE LSP or if ABR1 has been configured not to perform crankback, then ABR1 MUST forward a PathErr up to the inter-area head-end LSR (R0) without trying to select another egress LSR.

5.1.3. Processing of the Resv message (common procedure for contiguous and stitched/nested LSPs)

The Resv message for the inter-area TE LSP is sent back from R1 to R0. When the Resv message arrives at ABRÆ1, depending on whether ABRÆ1 is nesting or stitching, ABRÆ1 will install the appropriate label actions for the packets arriving on the inter-area LSP. Similar procedures are carried out at ABR1 as well, while processing the Resv message.

As the Resv message for the inter-area LSP traverses back from R1 to R0, 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 ABRÆ1 sends the Resv message to ABR1, ABRÆ1 will record its Router ID in the RRO object. So, the inter-area TE LSP from R0 to R1 would have an RRO of R0-ABR1-ABRÆ1-R1 or R0-<other hops>-ABR1-ABRÆ1-R1, depending on whether the source area is setting up a FA-LSP/LSP segment or signaling a contiguous TE LSP. 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.

5.2. Example with an inter-AS TE LSP (based on the assumption described in section 3).

The procedures for establishing an inter-AS TE LSP are very similar to those of the inter-area TE LSP described above. The main difference here from the inter-area case, is the presence of ASBR-ASBR link(s).

The links interconnecting ASBRs are usually not TE enabled and no IGP is running at the AS boundaries.

An implementation supporting inter-AS MPLS TE MUST obviously allow the set up of inter-AS TE LSP over the region interconnecting multiple ASBRs. In other words, an ASBR compliant with this document MUST support the set up of TE LSP over ASBR to ASBR links, performing all the usual operations related to MPLS Traffic Engineering (call admission control, à) as defined in [RSVP-TE]. So the limitation (1) MUST be removed. Regarding the second limitation (2), in the very vast majority of the cases, two SPs do not run an IGP between ASBRs. Although this imposes for the two ASBRs to be interconnected via single hop link, this does not constitute a severe limitation.

An interesting optimization consists in allowing the ASBRs to flood the TE information related to the ASBR-ASBR link(s) although no IGP TE is enabled over those links (and so there is no IGP adjacency over the ASBR-ASBR links). This allows a head-end LSR to make a more appropriate route selection up to the first ASBR in the next hop AS in the case of scenario 1 and will significantly reduce the number of signaling steps in route computation. This also allows the entry ASBR in an AS to make a more appropriate route selection up to the entry ASBR in the next hop AS taking into account constraints associated with the ASBR-ASBR links. Moreover, this reduces the risk of call set up failure due to inter-ASBR links not satisfying the inter-AS TE set of constraints. Note that the TE information is only related to the ASBR-ASBR links. In other words, the TE LSA/LSP flooded by the ASBR includes not only the links

contained in the AS but also the ASBR-ASBR links.

Note that no summarized TE information is leaked between ASes in any of the proposed scenarios in this document.

Example:

For instance, in the diagram depicted above, when ASBR1 floods its IGP TE LSA (opaque LSA for OSPF)/LSP (TLV 22 for IS-IS) in its routing domain, it reflects the reservation states and TE properties of the following links: X1-ASBR1, ASBR1-ASBR4, ASBR1-ASBR2.

Thanks to such an optimization, the inter-ASBRs TE link information corresponding to the links originated by the ASBR is made available in the TED of other LSRs in the same area/AS that the ASBR belongs to, hence the CSPF computations for an inter-AS TE LSP path can also take into account the ASBR-ASBR link. This will improve the chance of successful path computation up to the next AS (ASBR4, 5 or 6 in this example) 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.

5.2.1. Case 1: T1 is a contiguous TE LSP

The inter-AS TE path may be configured on the head-end LSR as a set of strict hops, loose hops or a combination of both.

- Example 1 (set of strict hops end to end): R0-X1-ASBR1-ASBR4-ASBR5-R3-ASBR7-ASBR9-R6
- Example 2 (set of loose hops): R0-ASBR4(loose)-ASBR9(loose)-R6(loose)
- Example 3 (mix of strict and loose hops): R0-R2-ASBR3-ASBR2-ASBR1-ASBR4(loose)-ASBR10(loose)-ASBR9-R6

When a next hop is a loose hop, a dynamic path calculation (also called ERO expansion) is required taking into account the topology and TE information of its own AS and the set of TE LSP constraints. In the example 1 above, the inter-AS TE LSP path is statically configured as a set of strict hops, so in this case, no dynamic computation is required. In the example 2, a per-AS path computation is performed, respectively on R0 for AS1, ASBR4 for AS2 and ASBR9 for AS3.

Note that when an LSR has to perform an ERO expansion, the next hop must either belong to the same AS, or must be the ASBR directly connected to the next hops AS. In this later case, the ASBR reachability must be announced in the IGP TE LSA/LSP originated by its neighboring ASBR. In the example 2 above, the TE LSP path is defined as: R0-ASBR4(loose)-ASBR9(loose)-R6(loose). This implies that the ERO expansion performed by R0 must compute the path from R0 to ASBR4. As stated in section 6.2, the TE reservation state related to the ASBR1-ASBR4 link is flooded in AS1 by ASBR1. In addition, ASBR1 MUST also announce the IP address of ASBR4 specified in the T1 path configuration.

If an auto-discovery mechanism is available, every LSR receiving an RSVP Path message, will have to determine automatically the next hop ASBR, based on the IGP/BGP reachability of the TE LSP destination. With such a scheme, the head-end LSR and every downstream ASBR loose hop

(except the last loose hop that computes the path to the final destination) automatically computes the path up to the next ASBR, the next loose hop based on the IGP/BGP reachability of the TE LSP destination. If a particular destination is reachable via multiple loose hops (ASBRs), local heuristics may be implemented by the head-end

LSR/ASBRs to select the next hop an ASBR among a list of possible choices (closest exit point, metric advertised for the IP destination (ex: OSPF LSA External - Type 2), local policy,...). Once the next ASBR has been determined, an ERO expansion is performed as in the previous case.

Once it has computed the path up to the next ASBR, ASBR1 sends the Path message for the inter-area TE LSP to ASBR4 (supposing that ASBR4 is the selected next hop ASBR). ASBR4 then repeats the exact same procedures and the Path message for the inter-AS TE LSP will reach the destination R1. If ASBR4 cannot find a path obeying the set of constraints for the inter-AS TE LSP, then ASBR4 MUST send a PathErr message to ASBR1. Then ASBR1 can in turn either select another ASBR (ASBR5 in the example above) if crankback is allowed for this inter-AS TE LSP (see [CRANBACK]). If crankback is not allowed for that inter-AS TE LSP or if ASBR1 has been configured not to perform crankback, then ABR1 MUST forward a PathErr up to the head-end LSR (R0) without trying to select another egress LSR. In this case, the head-end LSR can in turn select another sequence of loose hops, if configured. Alternatively, the head-end LSR may decide to retry the same path; this can be useful in case of set up failure due an outdated IGP TE database in some downstream AS. An alternative could also be for the head-end LSR to retry to same sequence of loose hops after having relaxed some constraint(s).

5.2.2. Case 2: T1 is a stitched or nested TE LSP

The signaling procedures are very similar to the inter-area LSP setup case described in section 5.1.2. In this case, the FA-LSPs or LSP segments will only be originated by the ASBRs at the entry to the AS.

In the example provided in section 3, for an LSP setup from CE1 to CE2, the FA-LSPs/LSP segments may be setup between ASBR4-ASBR7 and potentially ASBR9-R7. The Path message in this case traverses along CE1-R0-ASBR1-ASBR4-ASBR7-ASBR9-R7-CE2. In the RRO sent in the Resv message, the ASBRs which are ingress into the AS (like ASBR4, ASBR9, ASBR3, ASBR10) 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 ASBR1 and ASBR4) are more than one hop away, then instead of creating RSVP state for every inter-AS LSP traversing ASBR1 and ASBR4, one MAY decided to aggregate these requests by setting up a FA-LSP between the ASBRs to nest the inter-AS LSP requests. The boundary LSR ASBR1, by default is not a candidate to initiate a FA-LSP or LSP segment setup. But this behavior MAY be overridden by configuration.

5.3. Signaling specifics with TE LSP stitching for packet LSPs

This section only applies to an inter-area/AS packet LSP being stitched to another intra-area/AS packet LSP. If a boundary LSR (ABR/ASBR) desires to perform LSP stitching, then it MUST indicate this in the Path message for the intra-area/AS 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 onto the LSP segment. This will allow it to allocate the correct label(s) as explained below. Also, so that the head-end LSR can ensure that correct stitching actions were carried out at the egress LSR, a new flag is 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 local intra-area/AS LSP segment by the head-end LSR of the LSP segment (boundary LSR) that desires LSP stitching. This flag SHOULD not be modified by any other LSRs in that area/AS.

An intra-area/AS LSP segment can only be used for stitching if appropriate label actions were carried out at the egress LSR of the LSP segment. In order to indicate this to the head-end 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, and also recognizes the ææLSP stitching desiredÆÆ 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 head-end LSR of the intra-area/AS LSP segment.

If an egress LSR receiving a Path message with the ææLSP stitching desiredÆÆ 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 head-end 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 ææLSP segment stitching readyÆÆ bit defined in the RRO sub-object. Also, when the egress LSR for

the LSP segment receives a Path message for an inter-area/AS LSP using this LSP segment, it SHOULD first check if it is also the egress for

the inter-area/AS TE LSP. If the egress LSR is the egress for both the intra-area/AS LSP segment as well as the inter-area/AS TE LSP, and it requires Penultimate Hop Popping (PHP), then the LSR MUST send back a Resv refresh for the intra-area/AS 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-area/AS TE LSP.

Finally, if the egress LSR for the intra-area/AS 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-area/AS 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 ææLSP segment stitching readyÆÆ flag is cleared.

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

Example: In case of inter-AS TE LSP setup from CE1 to CE2 as described in the example, let us assume that ASBR4 is doing one-to-one LSP stitching. When ASBR4 receives the inter-AS TE LSP Path message, it will first initiate the setup of an intra-AS LSP segment to ASBR7, if not already present. In the Path message for this LSP segment, ASBR4 will set the "LSP stitching desired" flag in the Attributes Flags TLV of the LSP_ATTRIBUTES object. When ASBR7 receives this Path message, it will allocate a non-NULL label (real label for swap action) in the Resv message for this LSP segment. Also, it will set the "LSP segment stitching ready" flag in the RRO subobject in the Resv message. Once the LSP segment is signaled successfully, ASBR4 will then forward the Path message for the inter-AS TE LSP to ASBR7, which propagates it further. Eventually as the Resv message for the inter-AS TE LSP traverses back from ASBR9 to ASBR7 and reaches ASBR7, ASBR7 will remember to swap the LSP segment label with the label received for the inter-AS LSP from ASBR9. Also, ASBR7 will itself allocate a NULL label in the Resv message for the inter-AS TE LSP and sends the Resv message to ASBR4. ASBR4 ignores the Label object in the Resv message received from ASBR7 for the inter-AS TE LSP and remembers to swap the label that it allocates in the inter-AS Resv message sent to ASBR1 with the label that it had received from say, LSR R4 for the intra-AS LSP segment. In

this manner, the inter-AS TE LSP is stitched to an intra-area/AS LSP segment in AS2. In this example, if the LSP destination for the inter-AS LSP had been ASBR7, if this is a packet-switched LSP and if ASBR7 requires PHP, then on receiving the Path message for the inter-AS LSP, ASBR7 will re-send a Resv message for the intra-area/AS LSP segment to say R4, by changing the Label to a NULL label.

6. Scenario 2: end to end shortest path computation

6.1. Introduction and definition of an optimal path

Qualifying a path as optimal requires clarification. Indeed, a globally optimal TE LSP placement usually refers to a set of TE LSP whose placements optimize the network resources (i.e a placement that reduces the maximum or average network load for instance). By contrast, a optimal path for a TE LSP, is the shortest path that obeys the set of required constraints (bandwidth, affinities,à), minimizing either the IGP or TE metric cost (See [SECOND-METRIC] and [MULTIPLE-METRICS]). In this document, an optimal inter-AS TE path is defined as the optimal path that would be obtained in the absence of AS/Areas, in a totally flat network between the source and destination of the TE LSP.

6.2. Notion of PCE (Path Computation Element)

An LSR is said to be a PCE (Path Computation Element) when it has the ability to compute an inter-area/AS TE LSP path for a TE LSP it is not the head-end of. Ideal candidates to support a PCE function are ABRs in the context of inter-area TE (since each ABR has the view of two of more areas in its TED) and ASBR in the context of inter-AS TE. Note that in this document an LSR supporting the function of PCE is simply referred to as a PCE. As in the case of intra-area TE, it is not made any assumption on the actual path computation algorithm in use by the PCE (it can be any variant of CSPF, algorithm based on linear-programming to solve multi-constraints optimization problems,à).

6.3. Dynamic PCE discovery

PCE(s) can either be statically configured on each LSR requesting an inter-area/AS TE LSP path computation or dynamically discovered by means of IGP extensions defined in [OSPF-CAP], [OSPF-TE-CAP], [ISIS-CAP] and [ISIS-TE-CAP]. This allows an Operator to elect a subset of ABRs/ASBRs to act as PCEs.

Note that if the AS is made of multiple areas/levels, [OSPF-CAP] and

[ISIS-CAP] support the capabilities announcements across the entire routing domain (making use of TLV leaking procedure for IS-IS and OSPF opaque LSA type 11 for OSPF).

6.4. PCE selection

It belongs to an LSR informed of the existence of multiple PCEs having the capability to serve an inter-area/AS TE LSP path computation request to select the preferred PCE. For instance, an LSR may select the closest PCE based on the IGP metric or may just randomly select one of the PCE. In case of multiple PCEs, the selected PCE should be such that the requests are balanced across multiple PCEs. An LSR MUST be able to select another PCE if its preferred PCE does not answer to its request. Note that the PCE may or not be along the TE LSP Path. This implies that the PCE is just responsible for the TE LSP path computation, not for its maintenance. Moreover, the PCE may compute

just a path segment, not the whole path end to end; in this case, the returned computed path will contain loose hops.

6.5. LSR-PCE signaling protocol

Any LSR can send an RSVP path computation request to a PCE that will in turn compute a set of TE LSP(s) path and return the corresponding path parameters via an RSVP path computation reply message. The format of the RSVP path computation requests and reply messages are defined in [PATH-COMP] as well as the set of optional objects characterizing the constraints:

REQUEST-ID object: must be present in any RSVP Path computation request and reply message and specifies the request-ID-number, several requests characteristics.

METRIC-TYPE object: allows the PCC to indicate to the PCE the metric to be used to compute the shortest path (currently two metrics are defined: the IGP or TE metric).

PATH-COST object: object inserted in the RSVP path computation reply message to indicate the cost of a computed TE LSP in addition to the path. This object is mandatory if the cost has been explicitly requested in the RSVP path computation request and optional in any other case.

The protocol state machine is also defined in [PATH-COMP].

6.6. Computation of an optimal end to end TE LSP path

This section details the set of mechanisms allowing to compute an optimal (shortest) inter-area/AS TE LSP path obeying a set of specified constraints.

Each step of the mechanism is illustrated with the example of an inter-AS TE LSP obeying a set of specified constraints: the shortest path of an inter-AS TE LSP T1 originated at R0 in AS1 and terminated at R6 in AS3 is computed). The case of inter-area TE LSP optimal path computation is very similar.

1) Step 1: discovery by the head-end LSR of a PCE capable of serving its path computation request. The PCE will either be an ABR (inter-area TE) or an ASBR (Inter-AS TE). In the case of inter-AS TE, the PCE must be able to serve the source AS and can compute inter-AS TE LSP path terminating in the destination ASn. As mentioned above, the PCE can either be statically configured or dynamically discovered via IGP extensions. If multiple PCEs are discovered, the head-end LSR selects one PCE based on some local policies/heuristics.

Ex: R1 selects ASBR1 as the PCE serving its request for the T1 path computation.

2) Step 2: an RSVP Path computation request is sent to the selected PCE.

Case of inter-area TE: the head-end LSR sends its path computation requests to the selected PCE (ABR).

Case of inter-AS TE: the RSVP path computation request can be sent either (1) to a PCE in the same AS which will in turn relay the request to a PCE of the next hop AS (Ex: R0 sends an RSVP path computation request to ASBR1 which relays the request to say ASBR4) or (2) to the PCE in the next hop AS if the head-end LSR has a complete topology and TE view up to the next hop PCE (Ex: R0 sends an RSVP path computation request to ASBR4). It is expected that (1) will be the most common inter-AS TE deployment scenario for some security issues.

Note that it may be desirable to set up some policies on the PCE to limit the access to specific LSRs. Moreover, the usual RSVP authentication process may be used when sending a request to a PCE.

Step i: the PCE of ASi relays the path computation request to the

selected PCE peer in AS(i+1) located in the next hop AS. Note that the address of the list of PCE(es) in the next hop AS must be statically configured on the PCE. This implies some static configuration on the PCE only.

Ex: ASBR1 sends an RSVP path computation request to ASBR4

If the TE LSP destination is in ASi, then the PCE provides the list of shortest paths (with their corresponding ERO (potentially partial ERO) + Path-cost) from every ASBR to the inter-AS TE LSP destination. See a detailed example below.

If the TE LSP destination is not in ASi, the PCE relays the RSVP path computation request to the targeted PCE peer in AS(i+2) in the next hop AS.

The process is iterated until the destination AS is reached.

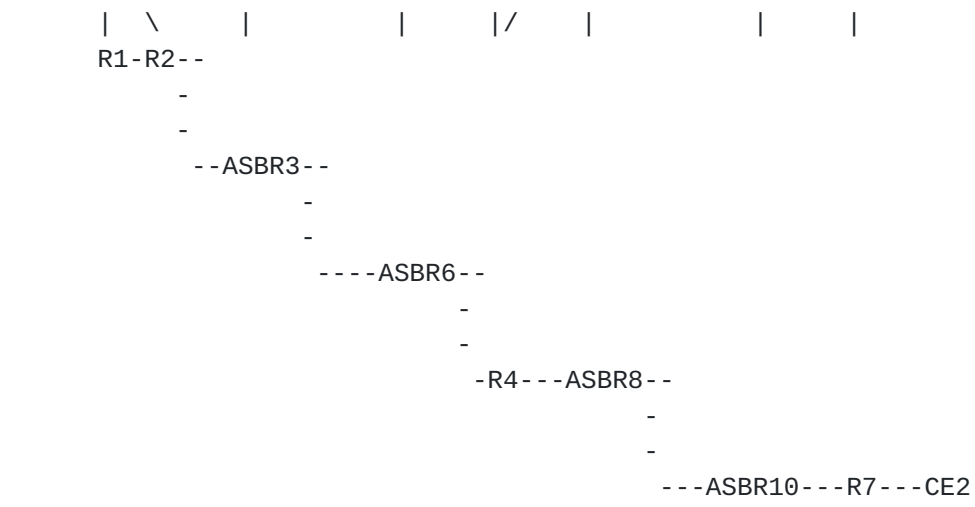
Ex: ASBR4 relays the RSVP path computation request to ASBR9 which determines that the T1Æs destination address belongs to its AS. ASBR9 will in turn return a path computation reply to the requesting PCE, e.g ASBR4. The RSVP path computation reply contains two paths (provided that two paths obeying the set of constraints exist):

- ERO 1: ASBR9-R6, Cost=c1,
- ERO 2: ASBR10-R7-R6, Cost=c2

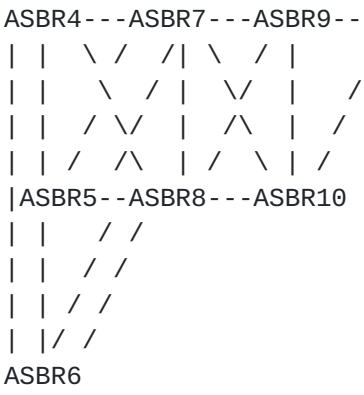
Note that the ERO object might be made of loose hop to preserve confidentiality.

Step 3: once a requesting PCEi receives an RSVP Path computation reply, the shortest path is computed from ASi to ASi+1 by route concatenation.

This is done by running a virtual SPT (Shortest Path Tree) computation using CSPF where the nodes are the ABSRs connected by virtual link whose costs are equal to the shortest path connecting the ASBRs.



Resulting Virtual SPT computed by ASBR 4



Within ASi, the cost of each ASBR-ASBR virtual link is equal to the shortest path cost. This information is known by PCEi.

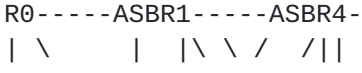
The cost of the ASBR-ASBR link between ASBR of different ASes is also known by the PCEi (see section 6.2).

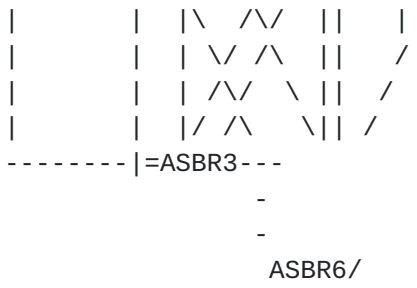
Within ASi+1, the cost of the ASBR-ASBR virtual link is provided in the RSVP path computation reply of the PCEi+1.

Ex: ASBR4 will then compute the shortest path for the TE LSP traversing AS2 and AS3.

Then the process is reiterated recursively until the optimal end-to-end Path computation is completed. The whole path may not be seen by each PCE for confidentiality reason but this process will ensure that the shortest path is selected.

Example: the resulting computed virtual SPT computed by ASBR1 will finally be:





Then once the optimal end to end path is computed, the head-end LSR sets up the inter-AS TE using a complete list of ERO if the various PCEs have provided the list of ERO or some loose hops in the contrary.

An implementation MAY decide to support local caching of path computation in order to optimize the path computation process. The flip side of path caching is the potential increase of call set up failure. When caching is in use, it must be flushed upon TE LSP failure provided that the PCE is along the inter-area/AS TE LSP path.

As opposed to scenario 1, an end-to-end shortest path obeying the set of required constraint is computed. In that sense, the path is optimal.

Some other variants of such an algorithm relying on the same principles are also possible.

Note also that in case of ECMP paths, more than one path could be returned to the requesting LSR.

6.7. Path optimality

In the case of inter-area TE, it is a common usage to adopt the same policy for the IGP/TE metric (based on the link-speed, propagation delay or some other combination of constraints). Hence, the proposed set of mechanism always computes the shortest path across multiple areas obeying the required set of constraints. In the case of Inter-AS TE, in order for this path computation to be meaningful, a metric normalization between ASes is required. One solution to avoid IGP

metric modification would be for the SPs to agree on a TE metric normalization and use the TE metric for TE LSP path computation (in that case, this must be requested in the RSVP Path computation request via the METRIC-COST object defined in [PATH-COMP]).

6.8. Diverse end to end path computation

The RSVP signaling protocol defined in [PATH-COMP] allows an LSR to request the computation of a set of N diversely routed TE LSPs. Then in this scenario, a set of diversely routed TE LSP between two LSRs can be computed since both paths are simultaneously computed with a minimal required amount of steps.

7. Mode of operation of MPLS Traffic Engineering Fast Reroute for inter-area/AS TE LSPs

MPLS Traffic Engineering Fast Reroute ([FAST-REROUTE]) defines local protection schemes that provide fast recovery (in 10s of msecs) of protected TE LSPs upon link/SRLG/Node failure. A backup TE LSP path is either statically configured or dynamically computed and then the backup TE LSP is signaled at each hop. Upon detecting a network element failure (via link failure detection mechanisms provided via layer 2 protocol, or IGP/BFD/RSVP fast hellos), the node immediately upstream to the failure (called the PLR (Point of Local Repair)) reroutes the set of protected TE LSPs onto the appropriate backup tunnel(s) around the failed resource. In the context of inter-area/AS TE, one must consider various failure scenarios and analyze for each of them the potential required extensions for MPLS TE FRR. [FAST-REROUTE] specifies two modes referred to as the one to one mode and facility backup mode. While this section specifies the use of MPLS TE Fast Reroute for the facility backup mode, similar procedures also apply for the one-to-one backup mode.

The failure scenarios specific to inter-area/AS TE are the following:

- Failure of an ABR or an ASBR node
- Failure of an inter-ASBRs link

Because the cases of a contiguous LSP significantly differ from the one of a stitched/nested TE LSP, they will be treated separately.

The current set of mechanisms defined in [FAST-REROUTE] applies without any restriction to any link/SRLG/Node failure within an area or an AS. In other words, a protected inter-area/AS TE LSP (an LSP signaled with

the "local protection desired" bit set in the SESSION-ATTRIBUTE object or with a FAST-REROUTE object) can be protected via the MPLS TE Fast Reroute mechanism regardless of whether the TE LSP is an intra-area/AS or inter-AS TE LSP in case of link/SRLG/node failure within the AS. This is true for contiguous, nested and stitched inter-area/AS TE LSP.

However, MPLS TE Fast Reroute is a temporary local protection mechanism. Upon a link/SRLG/node failure, the PLR triggers Fast Reroute and for each rerouted TE LSP, the PLR MUST send a notification of the local repair by sending an RSVP Path Error message with error code of "Notify"(Error code =25) and an error value field of ss00 cccc cccc cccc where ss=00 and the sub-code = 3 ("Tunnel locally repaired") (see [RSVP-TE]). The receipt of such a Notify Path Error is used by the head-end LSR to trigger a reoptimization such that the TE LSP follows a more optimal path.

Case of a contiguous inter-area/AS TE LSP

- The case of a contiguous inter-area/AS TE LSP is identical to an intra-area TE LSP.

Case of a stitched/nested TE LSP

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-area/AS FA-LSP/LSP segment in that area. 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-area/AS LSPs using the FA-LSP/LSP segment will start taking the new path in that area/AS. However, in case the head-end LSR in that area/AS is unable to find a path around the failure to re-route the intra-area FA-LSP/LSP segment, then a failure and repair notification stated above (PathErr) for all the affected inter-area/AS TE LSPs MAY be propagated to the upstream area/AS towards the head-end LSR for the inter-area/AS TE LSP. This could be a local policy decision. Other area/AS boundary LSRs along the way could intercept the error message to do some kind of crankback if crankback is allowed for the inter-area/AS TE LSP. This two-phase approach tries to handle the failure first locally within an area/AS as far as possible by intercepting the error notification at the area/AS boundary LSR and re-routing the intra-area/AS 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 head-end LSR for the inter-area/AS TE LSP in the originating area/AS. 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: back to the example of the inter-AS TE LSP setup, let us assume that the FA-LSP/LSP segment traverses R4 in AS2, and is node-protected against the failure of R4. In that case, when R4 or the corresponding link to R4 fails, then the traffic will be locally protected by the corresponding backup path LSP associated with the protected FA-LSP/LSP segment. When the PathErr/Notify message "Tunnel Locally Repaired" reaches ASBR4, it 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-AS traffic is sent along the FA-LSP/LSP segment, that traffic has been protected as well.

Note: in the context of inter-AS TE LSP, if the failure occurs in an area/AS different from the head-end LSR, the head-end LSR exclusively relies on the Path Error message to get informed that a local repair has been performed in order to potentially perform a reoptimization. Hence, the RSVP Path Error message SHOULD be sent in reliable mode ([REFRESH-REDUCTION]).

7.1. Support of MPLS TE Fast Reroute for a contiguous inter-area/AS TE LSP

7.1.1. Failure of a network element within an area/AS

The mode of operation of MPLS TE Fast Reroute to protect a contiguous, stitched or nested TE LSP within an area or AS is identical as the single area/AS case.

7.1.2. Failure of an inter-AS link

To protect an inter-ASBR link with MPLS TE Fast Reroute, the following actions are required:

- A set of backup tunnels must be configured or dynamically computed between the two ASBRs diversely routed from the protected inter-ASBRs link. Mechanisms like ææauto-discoveryÆÆ 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.

Notes:

- Typically, the region connecting two ASes is not TE enabled. So an implementation will have to support the set up of TE LSP over a non-TE enabled region. The backup tunnel path will be configured on each ASBR as a set of strict hops and then signaled via the RSVP-TE procedure defined in RFC3209.

- The reason why a set of NHOP backup tunnels might be required is in case of requirement for bandwidth protection if a single backup tunnel satisfying the bandwidth requirement cannot be found (see [BANDWIDTH-PROTECTION]).

- For each protected inter-AS TE LSP traversing the protected link, a NHOP backup must be selected by a PLR (i.e ASBR), when the TE LSP is first set up. This requires for the PLR to select a backup tunnel terminating at the NHOP. Finding the NHOP backup tunnel of an inter-AS LSP can be achieved by analyzing the content of the RRO object received in the RSVP Resv message of both the backup tunnel and the protected TE LSP(s). As defined in [RSVP-TE], the addresses specified in the RRO IPv4 subobjects can be node-ids and/or interface addresses (with specific recommendation to use the interface address of the outgoing Path messages). Within a single area, the PLR can easily find whether the backup tunnel intersects the protected TE LSP regardless of whether the node-id or the interfaces are specified in the RRO object since it has the complete topology knowledge in its IGP database. This is not the case when the MP resides in a different AS. [NODEID] proposes a solution to this issue, defining an additional RRO IPv4 suboject that specifies a node-id address.

Example: The ASBR1-ASBR4 link is protected by the backup tunnel B1 that follows the ASBR1-ASBR2-ASBR4 path.

7.1.3. Failure of an ABR or an ASBR node

To protect a contiguous inter-area/AS TE LSP from an ABR/ASBR node failure using MPLS TE Fast Reroute, the following actions are required:

Case of inter-AS TE:

- A set of backup tunnel(s) must be configured from the penultimate hop in AS1 (penultimate node directly connected to the last ASBR in AS1) to the first ASBR in AS2 to protect against the failure of the last ASBR in AS1.

Ex: B1 from X1 to ASBR4 follows the X1-ASBR2-ASBR4 path and protects against the failure of the ASBR1 node.

- A set of backup tunnel(s) must be configured from the last ASBR in AS1 to the next hop of the first ASBR in AS2 to protect against the failure of the first ASBR in AS2.

Ex: B3 from ASBR1 to R3 follows the ASBR1-ASBR2-ASBR3-ASBR6-ASBR5-R3 path and protects against the failure of the ASBR4 node.

Case of inter-area TE:

- A set of NHOP backup tunnel(s) must be configured from the ABRÆs upstream LSR to the ABRÆs downstream LSR.

Example: B1 from X1 (upstream neighbor of ABR1 in area 1) to Y1 (downstream neighbor of ABR1 in area 0).

For each protected inter-AS TE LSP traversing the protected link/node, a NNHOP backup must be selected by a PLR (i.e LSR/ASBR). This requires for the PLR to select a backup tunnel terminating at the NNHOP.

Finding the NNHOP backup tunnel of an inter-AS LSP can be achieved by analyzing the content of the RRO object received in the RSVP Resv message of both the backup tunnel and the protected TE LSP(s) (see [NODE-ID]).

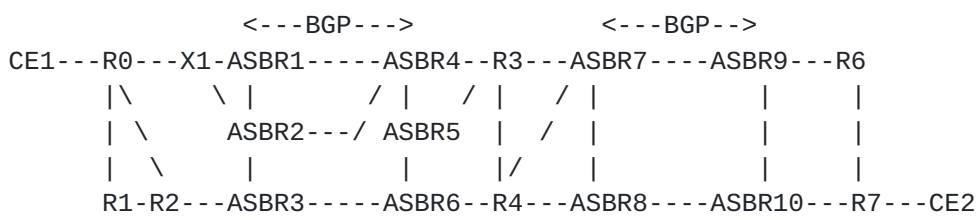
7.1.4. Procedure during MPLS TE Fast Reroute

In addition to the rules defined in [FAST-REROUTE], in the context of inter-area/AS TE LSP, there is a specific action that must be performed when protecting the first ASBR of the next AS via a NNHOP backup tunnel (see 5.6.3 (1)).

The ASBR acting as a PLR (Point of Local Repair) MUST:

- Identify the MP address in the RRO received in the corresponding Resv message,
- Remove all the sub-objects preceding the first address belonging to the MP,
- Replace this first MP address with the IP address of the MP (its node-id address).

Example with inter-AS TE:



- T1: a protected inter-AS TE LSP from R0 to R6, whose path is defined on R0 as a set of loose hops: R0-ASBR1(loose)-ASBR4(loose)-ASBR9(loose)-R6

- B3: a backup tunnel from ASBR1 to R3 following the ASBR1-ASBR2-ASBR3-ASBR6-ASBR5-R3 path and protecting against a failure of the ASBR4 node.

- The ERO subobject content signaled in the rerouted RSVP Path message of T1 over B3 by ASBR1 (PLR) must content the MPs as the next hop address (R3). Otherwise, R3 will receive an incorrect ERO.

A similar mechanism is required when rerouting an inter-AS TE LSP from the failure of the last ASBR of an AS.

- The RRO object may need to be updated by inserting an IPv4 or IPv6 subobject corresponding to the outbound interface address the rerouted traffic is forwarded onto (both the "Local protection in use" and "Local Protection Available" flags must be set).

7.2. Support of MPLS TE Fast Reroute for a stitched/nested TE LSP

7.2.1. Failure of an inter-AS link

The case of inter-ASBR link protection for stitched/nested TE LSPs is identical as with contiguous TE LSPs.

7.2.2. Failure of an ABR or an ASBR node

The major difference with contiguous inter-area/AS TE LSP is that with stitched/nested inter-area/AS TE, the MP for the inter-area/AS LSP MUST always be an area/AS boundary LSR (ABR/ASBR). This is because the FA-LSP/LSP segment is a different LSP (different session) from the inter-area/AS LSP, so the inter-area/AS LSP backup can only intersect the protected LSP path at the area/AS boundary LSRs.

Node protection of an exit ABR/ASBR

Let us consider the inter-AS TE example where the objective is to protect the fast re-routable inter-AS TE LSP from a failure of ASBR7 by means of MPLS TE Fast Reroute.

Considering the FA-LSP/LSP segment terminating at ASBR7, this is the last hop for the FA-LSP/LSP segment, so there can be no node-protection for ASBR7 via the FA-LSP/LSP segment. However, as far as the inter-AS LSP is concerned, its path is along R0-ASBR1-ASBR4-ASBR7-ASBR9-R7 and the FA-LSP/LSP segment between ASBR4 and ASBR7 is a link. So for protecting against ASBR7's failure, ASBR4 is the PLR and ASBR4 will setup a bypass tunnel to the NNHOP for this LSP, which is ASBR9. Again the NNHOP is determined by examining the received RRO for the inter-area/AS LSP. So one or more bypass tunnels following ASBR4-ASBR8-ASBR10-ASBR9 must be set up on ASBR4 to protect against node ASBR7's failure.

It is worth mentioning that this adds some additional constraints on the backup path since the bypass tunnel path needs to be diverse from the ASBR4-ASBR7-ASBR9 path instead of just being diverse from the X-ASBR7-ASBR9 path where X is the upstream neighbor of ASBR7.

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.

Node protection of an entry ABR/ASBR

Let us now consider the protection of an entry ASBR: for instance ASBR4.

Again, in this case, the FA-LSP/LSP segment offers no protection; so one or more backups MUST be set up from the previous hop LSR, i.e. ASBR1, to the NNHOP with respect to the inter-AS TE LSP, which is in this case ASBR7. A bypass tunnel ASBR1-ASBR3-ASBR6-ASBR7 would protect against ASBR4'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 resulting in the same consequences in term of backup tunnel path sub-optimality.

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-area/AS 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].

7.3. Failure handling of inter-AS TE LSP

In the context of MPLS Inter-area and inter-AS Traffic Engineering, if a link/SRLG/Node failure occurs in an area/AS different from the head-end LSR, the head-end LSR exclusively relies on the receipt of an RSVP Path Error message to get informed that the TE LSP has suffered a failure in a downstream AS (a ææNotifyÆÆ Path Error ææNotifyÆÆ message if the inter-AS TE has been locally repaired via MPLS TE Fast Reroute. For those reasons, as already mentioned, the Path Error message SHOULD be sent in reliable mode ([REFRESH-REDUCTION]). Note that this requires to configure the reliable messaging mechanisms proposed in [REFRESH-REDUCTION] between every pair of LSRs in the network (more precisely between every PLR and any potential head-end LSRs).

Upon receiving an RSVP Path Error message, a head-end LSR must perform a TE reroute (new route computation) in a make before break fashion.

It is worth highlighting that the set up of inter-AS TE LSP might be significantly slower than in the case of intra-area TE LSP:

- In scenario 1, the process may involve several ASBRs performing policy control, partial route computation (ERO expansions), à In case of set up failure, the number of trials can be significant, which even more increases the set up time.

  Furthermore, in case of dynamic loose hop computation, both the IGP and BGP reachability solutions have drawbacks in term of convergence upon failure. This is due to the slow convergence property of BGP. With BGP redistribution within ASes, the convergence might be even slower especially when BGP Route Reflectors are in use with no multi-paths load balancing.

- In scenario 2, some signaling exchange between several PCC and PCEs must be performed prior to setting up the TE LSP. Note that in scenario 2, the probability of TE LSP set up failure is limited to some lack of synchronization of the TE databases and

as such is significantly lower than in the case of scenario 1.

Moreover, in case of a large amount of inter-AS TE LSP set up, some non negligible extra signaling and routing computation load will be required on the loose hops (scenario 1) and loose hops/PCE (scenario 2). Some implementation may implement some pacing of inter-AS TE LSP set up rate. Typically a link/node/SRLG failure may impact a large number of TE LSPs. Relying on a local repair mechanism like MPLS TE Fast Reroute allows to relax the load on ASBR/PCE and reduces the need for urgent inter-AS TE LSP reroute. This is the recommended approach.

8. Reoptimization of an inter-area/AS TE LSP

The ability to reoptimize an existing inter-area/AS TE LSP path is of course a requirement. The reoptimization process significantly differs based upon the nature of the TE LSP and the mechanism in use for the TE LSP path computation.

If the head-end LSR uses a dynamic and distributed path computation technique such as the PCE-based path computation (described in section 6), then the head-end LSR can leverage this to send re-optimization requests to the PCE to obtain an optimal end-to-end path. On the other hand, in the absence of such a mechanism, the following mechanisms can be used for re-optimization, which are dependent on the nature of the inter-area/AS TE LSP.

8.1. Contiguous TE LSPs

8.1.1. Per-area/AS path computation (scenario 1)

After an inter-AS TE LSP has been set up, a more optimal route might appear in the various traversed ASes. Then in this case, it is desirable to get the ability to reroute an inter-AS TE LSP in a non-disruptive fashion (making use of the so called Make Before Break procedure) to follow this more optimal path. This is a known as a TE LSP reoptimization procedure.

[LOOSE-REOPT] proposes a mechanisms allowing:

- The head-end LSR to trigger on every LSR whose next hop is a loose hop the re evaluation of the current path in order to detect a potentially more optimal path. This is done via explicit signaling request: the head-end LSR sets the ææERO

Expansion requestÆÆ bit of the SESSION-ATTRIBUTE object carried in the RSVP Path message.

- An LSR whose next hop is a loose-hop to signal to the head-end LSR that a better path exists. This is performed by sending an RSVP Path Error Notify message (ERROR-CODE = 25), sub-code 6 (Better path exists).

This indication may be sent either:

- In response to a query sent by the head-end LSR,
- Spontaneously by any LSR having detected a more optimal path

Such a mechanism allows to reoptimize a TE LSP if and only if a better path is some downstream area/AS is detected.

The reoptimization event can either be timer or event-driven based (a link UP event for instance).

Note that the reoptimization MUST always be performed in a non-disruptive fashion.

Once the head-end LSR is informed of the existence of a more optimal path either in its head-end area/AS or in another AS, the inter-AS TE Path computation is triggered using the same set of mechanisms as when the TE LSP is first set up (per-AS path computation as in scenario 1 or involving some PCE(s) in scenario 2). Then the inter-AS TE LSP is set up following the more optimal path, making use of the make before break procedure. In case of a contiguous LSP, the reoptimization process is strictly controlled by the head-end LSR which triggers the make-before-break procedure, regardless of the location where the more optimal path is.

8.1.2. End to end shortest path computation (scenario 2)

[PATH-COMP] provides the ability to request a path reoptimization. In order to avoid double bandwidth accounting which could result in falsely triggered call set up failure the requesting LSR just provides the current path of the inter-area/AS TE LSP path to be reoptimized.

8.2. Stitched or nested (non-contiguous) TE LSPs

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

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

If, however, an operator desires to manually trigger reoptimization at the head-end LSR for the inter-area/AS LSP, then this solution does not prevent that. A manual trigger for reoptimization at the head-end LSR SHOULD force a reoptimization 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-area/AS TE LSP is being stitched to the intra-area/AS 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 head-end LSR does not know a priori whether a more optimal path exists. Such a manual trigger from the head-end LSR of the inter-area/AS TE LSP is, however, not considered to be a frequent occurrence.

Note that because stitching or nesting rely on local optimization, the reoptimization process allows to locally reoptimize each TE LSP segment or FA-LSP: hence, the reoptimization is not global and cannot guarantee that the optimal path end to end is found.

9. Routing traffic onto inter-area/AS TE LSPs

Once an inter-area/AS TE LSP has been set up, the head-end LSR has to determine the set of traffic to be routed onto the TE LSP.

In the case of intra-area/AS TE LSP, various options are available:

(1) modify the IGP SPF such that shortest path calculation can be performed taking into account existing TE LSP, with some path preference,

(2) make use of static routing. Note that the recursive route resolution of BGP allows routing any traffic to a particular (MP)BGP peer making use of a unique static route pointing the BGP peer address to the TE LSP. So any routes advertised by the BGP peer (IPv4/VPNv4 routes) will be reached using the TE LSP.

With an inter-area/AS TE LSP, just the mode (2) is available, as the TE LSP head-end does not have any topology information related to the destination area/AS.

10. Evaluation criteria and applicability

The aim of this section is to evaluate each proposed set of mechanisms described above with respects to the set of requirements listed in [INTER-AS-TE-REQS] and [INTER-AREA-TE-REQS].

10.1. Path optimality

Inter-area/AS TE LSP path optimality is one of the major differences between the various path computation techniques. In scenario 1, the

path is computed on a per-area/AS basis (making use of mechanisms like auto-discovery based on IGP/BGP information) cannot guarantee to compute an optimal (shortest) path across multiple areas/ASes. The resulting TE LSP path is the first path obeying the required set of constraints. This gets particularly true as TE LSP gets rerouted due the network element failures. On the other hand, a path computation mechanism like PCE (described in section 6) relies on a distributed path computation algorithm involving multiple ABR/ASBRs acting as PCEs

(Path Computation Elements) which guarantees to compute the shortest path end to end. Hence the PCE-based path computation method fully complies with the requirements states in [INTER-AS-TE-REQS] to be able to compute a shortest path end to end.

10.2. Reoptimization

In the absence of a distributed path computation method like the PCE-based, both the contiguous LSP and non-contiguous TE LSP (stitching/nesting) solution allows for reoptimization but they significantly differ in term of reoptimization process. A stitched or nested TE LSP is reoptimized on a per-area/AS basis. Each ABR/ASBR which is also the head-end LSR of an LSP segment or FA-LSP is responsible for the local reoptimization of that LSP segment or FA-LSP in the corresponding area/AS: in other words, the reoptimization process is contained within an area/AS. The reoptimization criteria and frequency is individually controlled by each head-end LSR (ABR/ASBR) of the LSP segment/FA-LSP independently of other segments and is transparent to the inter-area/AS TE LSP head-end LSR. The head-end LSR of the inter-area/AS TE LSP could still enforce a reoptimization but without knowing in advance whether a more optimal path actually exist in some downstream area/AS. Note also that each reoptimization is performed in a non-disruptive fashion (Make before break procedure). XX Indeed, each reoptimization implies some jitter and potentially some packet reordering usually undesirable for sensitive traffic. The use of contiguous inter-area/AS TE LSP used in conjunction with [LOOSE-PATH-REOPT] allows the head-end LSR to exert a strict control on the reoptimization process and perform a reoptimization if and only if a better path exists in some downstream area/AS. It relies on both a polling mechanism upon which an inter-area/AS TE LSP head-end LSR can poll the downstream nodes involved in partial path computation to learn whether a better (shorter) path exists. In addition, a downstream node can explicitly notify the head-end LSR of the existence of a better path (such a notification can be governed by local policy: timer-based, event-driven, à). In any case, the decision is led to the head-end LSR to perform an end to end reoptimization: it is expected that the head-end LSR will make use of some dampening mechanism to control the reoptimization frequency based on the inter-area/AS attributes. Note that the inter-area/AS TE LSP is reoptimized in every area/AS and may follow an identical path in some area(s)/AS(es).

In scenario 2, when a distributed path computation mechanism like PCE is used by the head-end LSR, an inter-area/AS TE LSP reoptimization is

similar to a path computation with the exception that the path of the inter-area/AS TE LSP is provided to the PCE to avoid any double bandwidth accounting. The reoptimization procedure is entirely controlled by the head-end LSR of the inter-area/AS TE LSP: this includes the path computation request frequency, decision to trigger an actual reoptimization (for example, the Head-end LSR may decide the perform a reoptimization if and only if the new more optimal path meets some specific requirements like a gain of x% in term of path cost compared to the TE LSP in place).

10.3. Support of MPLS Traffic Engineering Fast Reroute

As stated in [INTER-AS-TE-REQS] and [INTER-AREA-TE-REQS], the support of MPLS Traffic Engineering Fast Reroute is a strong requirement for inter-area/AS TE LSP and MUST cover the case of an inter-area/AS link/SRLG/Node failure, an inter-ASBRs link failure and an ABR/ASBR node failure.

The various solutions proposed in this document are equivalent in term of recovery time but significantly differ in term of backup tunnel path optimality. In the case of a fast reroutable contiguous TE LSP, the backup tunnel computed to protect against an ABR/ASBR node failure starts on the node immediately upstream to the ABR/ASBR and terminates on the node immediately downstream to the ABR/ASBR. By contrast, the computed backup tunnel to protect an inter-area/AS TE LSP making use of the stitching/nesting method MUST start and terminate on a boundary LSR (ABR/ASBR).

Hence, in the case of inter-AS TE for example, in order to protect against the failure of an exit ASBR, the backup tunnel must start on the entry upstream ASBR in the AS and terminate on the entry ASBR in the next-hop AS. In order the protect against the failure of entry ASBR, the backup tunnel starts on the node immediately upstream to the ASBR (exit ASBR on the upstream AS) and terminates on an exit ASBR in the AS (on the tail-end LSR of the FA-LSP/segment).

Such an additional constraint has the consequences for the backup tunnel path to be potentially sub-optimal compared to the backup tunnel path for a contiguous inter-area/AS TE LSP, hence implying more jitter during Fast Reroute. Moreover, this potentially reduces the capability to provide bandwidth protection and perform some efficient bandwidth sharing between backup tunnels protecting independent resources. Finally, this may increase the number of TE LSPs per mid-point LSR.

10.4. Support of diversely routed paths

There are several circumstances where the ability to set up a set of diversely routed TE LSP paths between two LSRs might be desirable:

(1) Load balancing

When a single TE LSP path satisfying the required constraints cannot be found between two LSRs, an alternative may consist in setting up N TE LSP such that the sum of their bandwidth is equal to the total required bandwidth. In addition, having diverse paths allows to limit the traffic disruption in case of network element failure between the two nodes to the set of affected TE LSPs.

(2) Path Protection

In some networks, Path protection is used to protect TE LSP from link/SRLG/node failure. This requires setting up, for each TE LSP, a set of diversely routed TE LSPs. In case of failure along the primary TE LSP path, the node directly attached to the failed resources signals to the head-end LSR that the TE LSP has failed, sending an RSVP Path Error. The head-end LSR can also detect that the TE LSP has suffered a failure when receiving an IGP update reflecting the failed resource. Note that the head-end LSR cannot rely on the IGP topology database to detect the failure if the failure does not occur in the Head-End area/AS in the case of inter-area/AS TE. Once the head-end LSR learns the failure, the traffic is switched onto the pre-established backup TE LSP. Note that a set of TE LSP can potentially share a single backup TE LSP (1:N protection).

Scenario 1: per-AS path computation

In the case of scenario 1, the set up of N diversely routed TE LSP paths can be done using the following scenario:

- Set up the first TE LSP among the set of N TE LSPs and include an RSVP RRO object in the RSVP Resv message to record the Path,

- For i=2 to N
  Set up the TE LSPi, excluding the elements traversed by the already set up TE LSP1, à, TE LSP i-1. The exclusion of a set of resources from a TE LSP path can be performed on the head-end LSR by CSPF and in other ASes by the loose hops along the path, each of them performing the computation of a part of the TE LSP. This requires from the head-end to pass the ææexcludeÆÆ constraints (see [EXCLUDE-ROUTE]).

Important note: such an algorithm does not guarantee that diverse paths can be found for the successive TE LSPs since the TE LSP path are not simultaneously computed, even if a possible solution exists. Also this simple algorithm does not allow finding two paths such that the sum of their cost is minimal. In case of an inter-area/AS 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.

Scenario 2: end to end shortest path computation: since both the primary and secondary TE LSP paths are simultaneously computed by the

distributed PCEs, it is possible to compute N diversely routed TE LSPs if such paths exist, with possibly and/or if necessary different constraints for both the primary and secondary inter-area TE LSPs. [PATH-COMP] proposes some RSVP extensions to signal such a requirement.

10.5. Diffserv-aware MPLS TE

There are no restrictions as far as Diffserv-aware MPLS TE is concerned introduced by the mechanisms proposed in the document.

10.6. Hierarchical LSP support

The non-contiguous TE LSP signaling for both nested TE LSPs is based on LSP Hierarchy signaling. Furthermore, the nesting of multiple inter-area/AS TE LSPs into an intra-area/AS FA LSP provides the Hierarchical LSP support in both control and forwarding planes.

10.7. Policy Control at the AS boundaries

Policy control essentially applies to TE LSPs spanning multiple ASes where each AS belongs to a different Operator. As stated in [INTER-AS-TE-REQS], a set of configurable options may be made available upon which ingress control policies can be implemented governing or honoring inter-AS TE agreements made by two interconnect SPs. During the path computation process, the inter-AS RSVP path message sent to its downstream loose hop (ASBR also) in a different AS can be firstly passed through an inter-AS TE policy control process on the downstream ASBR prior to its ERO expansion. The inter-AS RSVP path setup request will get rejected resulting in an path-error message which will be sent to the head-end LSR should it fail the control policy: for example, requesting bandwidth reservation more than agreed upon or wrong preemption priorities. Another approach consists in performing some constraint mapping. In the case of a contiguous TE LSPs, the local policy can dictate some constraints rewrite in order to make a TE LSP compliant with the agreements between the SPs. In the case of LSP stitching or nesting, the operation is eased by the fact that a different LSP segment or FA-LSP is established within the AS; consequently, other constraints can be applied to this intra-area/AS

LSP segment or FA-LSP like different affinities, preemption,etc.

10.8. Inter-AS MPLS TE Management

[LSPPING] proposes a solution which can be adopted for inter-AS TE LSPs whereby a head-end LSR sends MPLS echo requests over the LSP being tested. When the destination LSR receives the message, it needs to acknowledge the source LSR by sending an LSP_ECHO object in a RSVP Resv message.

The TTL processing over inter-area/AS TE primary or local backup LSPs will be supported as per [MPLS-TTL].

10.9. Confidentiality

Confidentiality issues essentially apply to the case of a TE LSP spanning multiple ASes, where each AS belongs to a different Operator. That being said, this can also apply to other scenarios where confidentially must be preserved outside of some specific domain. As mentioned in [INTER-AS-REQS], the solution should allow preserving AS confidentiality, by hiding the set of hops followed by the inter-AS TE LSP within an AS.

In scenario 1, as far as TE LSP signaling using RSVP is concerned, this requirement can be met via some proper RRO filtering at the AS boundaries (this applies to the RRO object carried in both the Path and Resv message). Note that, if MPLS TE Fast Reroute is required to protect inter-AS TE LSP against the failure of an ASBR, the RRO object carried in the Resv message of an inter-AS TE LSP must not be completely filtered, as mentioned in section 8. As least, the information (label, IPv4 or IPv6 subobject (node-id subobject)) pertaining to the next-hop ASBRs must be preserved.

In scenario 2, the RSVP Path computation reply can be filtered to provide a partial ERO (an ERO containing loose hops). If the agreement between SPs at AS boundary is such that confidentiality must be guaranteed, just a partial EROs be returned PCEs.

For the sake of illustration, [PATH-COMP] proposes some signaling extensions whereby the requesting LSR in ASx sends a Path computation request to the PCE in AS y, with the "Partial" flag of the REQUEST-ID object set. The PCE controls that this flag is appropriately set; if not, the PCE might decide either to provide a partial ERO or to drop the request.

Note that even when the returned ERO is partial, the PCE should provide the cost of the computed path.

Again for illustration, [PATH-COMP] proposes that the path computation reply includes a PATH-COST object in the RSVP Path computation reply message. If the agreement between SPs at AS boundaries is such that path cost might be provided, then the requesting LSR in ASx might send a Path computation request to the PCE in ASy, with the "cost" flag of the REQUEST-ID object set. The PCE controls that this flag is appropriately set; if set, the PCE MUST include a PATH-COST object in its RSVP Path Computation reply message. This is required to compute end to end shortest path.

11. Scalability and extensibility

All the features related to intra-area TE LSP are also applicable to inter-AS TE LSP, without any restriction.

12. Security Considerations

When signaling an inter-AS TE, an Operator may make use of the already defined security features related to RSVP (authentication). This may require some coordination between SPs to share the keys (see RFC 2747 and RFC 3097).

13. Intellectual Property Considerations

The IETF takes no position regarding the validity or scope of any intellectual property or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; neither does it represent that it has made any effort to identify any such rights. Information on the IETF's procedures with respect to rights in standards-track and standards-related documentation can be found in BCP-11. Copies of claims of rights made available for publication and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementors or users of this specification can be obtained from the IETF Secretariat.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights which may cover technology that may be required to practice this standard. Please address the information to the IETF Executive Director.

14. Acknowledgments

We would like to acknowledge input and helpful comments from Adrian Farrel.

Normative References

[RSVP] Braden, et al, " Resource ReSerVation Protocol (RSVP) - - Version 1, Functional SpecificationÆÆ, RFC 2205, September 1997.

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

[REFRESH-REDUCTION] Berger et al, ææRSVP Refresh Overhead Reduction ExtensionsÆÆ, RFC2961, April 2001.

[FAST-REROUTE] Ping Pan, et al, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", draft-ietf-mpls-rsvp-lsp-fastreroute-03.txt, December 2003.

[OSPF-TE] Katz, D., Yeung, D., Kompella, K., "Traffic Engineering Extensions to OSPF Version 2", draft-katz-yeung-ospf-traffic-09.txt(work in progress).

[ISIS-TE] Li, T., Smit, H., "IS-IS extensions for Traffic Engineering", draft-ietf-isis-traffic-04.txt (work in progress)

Informative references

[BANDWIDTH-PROTECTION] Vasseur et all, ææMPLS Traffic Engineering Fast reroute: bypass tunnel path computation for bandwidth protection ©, draft-vasseur-mpls-backup-computation-01.txt, October 2002, Work in progress.

[SECOND-METRIC] Le faucheur, ææUse of IGP Metric as a second TE MetricÆÆ, Internet draft, draft-lefaucheur-te-metric-igp-02.txt.

[MULTIPLE-METRICS] Fedyk D., Ghanwani A., Ash J., Vedrenne A. ææMultiple Metrics for Traffic Engineering with IS-IS and OSPFÆÆ, Internet draft, draft-fedyk-isis-ospf-te-metrics-01.txt

[PATH-COMP] Vasseur et al, ææRSVP Path computation request and reply messagesÆÆ, draft-vasseur-mpls-computation-rsvp-04.txt, work in progress.

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

[OSPF-TE-CAP] Vasseur et al."OSPF TE TLV capabilities", draft-ccamp-mpls-ospf-te-cap-00.txt, work in Progress.

[OSPF-CAP] Lindem et all ææ Extensions to OSPF for Advertising Optional Router CapabilitiesÆÆ, draft-ietf-ospf-cap-01.txt, work in progress.

[ISIS-CAP] Aggarwal et all, ææExtensions to IS-IS for Advertising Optional Router CapabilitiesÆÆ, work in progress

[ISIS-CAPA] Vasseur et al, ½ IS-IS extensions for advertising optional router capabilities ©, draft-vasseur-isis-cap-00.txt, work in progress.

[ISIS-TE-CAP] Vasseur et al, ææIS-IS TE TLV capabilitiesÆÆ, draft-vasseur-ccamp-isis-te-cap-00.txt, work in progress.

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

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

[EXCLUDE-ROUTE] Lee et all , Exclude Routes - Extension to RSVP-TE, draft-ietf-ccamp-rsvp-te-exclude-route-00.txt, work in progress.

[INTER-AREA-TE] Kompella et all,ÆÆMulti-area MPLS Traffic EngineeringÆÆ,

draft-kompella-mpls-multiarea-te-04.txt, work in progress.

[LSPPING] Kompella, K., Pan, P., Sheth, N., Cooper, D.,Swallow, G., Wadhwa, S., Bonica, R., " Detecting Data Plane Liveliness in MPLS", Internet Draft <draft-ietf-mpls-lsp-ping-02.txt>, October 2002. (Work in Progress)

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

[INTER-AS-TE-REQS] Zhang et al, ææMPLS Inter-AS Traffic Engineering requirementsÆÆ, draft-ietf-tewg-interas-mpls-te-req-06.txt, work in progress.

[INTER-AREA-TE-REQTS-1] Boyle J., "Requirements for support of Inter-Area and Inter-AS MPLS Traffic Engineering", (work in progress).

[INTER-AREA-TE-REQTS-2] Leroux et al, ææRequirements for Inter-area MPLS Traffic EngineeringÆÆ, draft-leroux-tewg-interarea-mpls-te-req-00.txt, work in progress.

[LOOSE-PATH-REOPT] Vasseur and Ikejiri, ææReoptimization of an explicit loosely routed MPLS TE pathsÆÆ, draft-vasseur-ccamp-loose-path-reopt-00.txt, June 2003, Work in Progress.

[NODE-ID] Vasseur, Ali and Sivabalan, ææDefinition of an RRO node-id subobjectÆÆ, draft-ietf-mpls-nodeid-subobject-02.txt, work in progress.

[LSP-HIER] Kompella K., Rekhter Y., "LSP Hierarchy with Generalized MPLS TE", draft-ietf-mpls-lsp-hierarchy-08.txt, March 2002.

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

Authors' Address:

Jean-Philippe Vasseur
Cisco Systems, Inc.
300 Beaver Brook Road
Boxborough , MA - 01719
USA
Email: jpv@cisco.com

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

Full Copyright Statement