Network Working Group
Internet Draft
Expiration Date: March 2002

Kireeti Kompella
Juniper Networks
Yakov Rekhter
Juniper Networks

LSP Hierarchy with Generalized MPLS TE

draft-ietf-mpls-lsp-hierarchy-08.txt

1. Status of this Memo

2. Abstract

To improve scalability of Generalized Multi-Protocol Label Switching (GMPLS) it may be useful to aggregate Label Switched Paths (LSPs) by creating a hierarchy of such LSPs. A way to create such a hierarchy is by (a) a Label Switching Router (LSR) creating a Traffic Engineering Label Switched Path (TE LSP), (b) the LSR forming a forwarding adjacency (FA) out of that LSP (by advertising this LSP as a Traffic Engineering (TE) link into the same instance of ISIS/OSPF as the one that was used to create the LSP), (c) allowing other LSRs to use FAs for their path computation, and (d) nesting of LSPs originated by other LSRs into that LSP (by using the label stack construct).

This document describes the mechanisms to accomplish this.

3. Specification of Requirements

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].

4. Overview

An LSR uses Generalized MPLS (GMPLS) TE procedures to create and maintain an LSP. The LSR then may (under local configuration control) announce this LSP as a Traffic Engineering (TE) link into the same instance of the GMPLS control plane (or more precisely its ISIS/OSPF component) as the one that was used to create the LSP. We call such a link a "forwarding adjacency" (FA). We refer to the LSP as the "forwarding adjacency LSP", or just FA-LSP. Note that an FA-LSP is both created and used as a TE link by exactly the same instance of the GMPLS control plane. Thus the concept of an FA is applicable only when an LSP is both created and used as a TE link by exactly the same instance of the GMPLS control plane. Note also that an FA is a TE link between two GMPLS nodes whose path transits zero or more (G)MPLS nodes in the same instance of the GMPLS control plane.

The nodes connected by a 'basic' TE link may have a routing adjacency; however, the nodes connected by an FA would not usually have a routing adjacency. A TE link of any kind (either 'basic' or FA) would have to have a signaling adjacency in order for it to be used to establish an LSP across it.

In general, the creation/termination of an FA and its FA-LSP could be driven either by mechanisms outside of GMPLS (e.g., via configuration control on the LSR at the head-end of the adjacency), or by mechanisms within GMPLS (e.g., as a result of the LSR at the head-end of the adjacency receiving LSP setup requests originated by some other LSRs).

ISIS/OSPF floods the information about FAs just as it floods the information about any other links. As a result of this flooding, an LSR has in its TE link state database the information about not just basic TE links, but FAs as well.

An LSR, when performing path computation, uses not just basic TE links, but FAs as well. Once a path is computed, the LSR uses RSVP/CR-LDP [RSVP-TE, CR-LDP] for establishing label binding along the path.

In this document we define mechanisms/procedures to accomplish the above. These mechanisms/procedures cover both the routing

(ISIS/OSPF) and the signalling (RSVP/CR-LDP) aspects.

Note that an LSP may be advertised as a point-to-point link into ISIS or OSPF, to be used in normal SPF by nodes other than the head end. While this is similar in spirit to an FA, this is beyond the scope of this document.

Scenarios where an LSP is created (and maintained) by one instance of the GMPLS control plane, and is used as a (TE) link by another instance of the GMPLS control plane are outside the scope of this document.

5. Routing aspects

In this section we describe procedures for constructing FAs out of LSPs, and handling of FAs by ISIS/OSPF. Specifically, this section describes how to construct the information needed to advertise LSPs as links into ISIS/OSPF. Procedures for creation/termination of such LSPs are defined in Section "Controlling FA-LSPs boundaries".

FAs may be represented as either unnumbered or numbered links. If FAs are numbered with IPv4 addresses, the local and remote IPv4 addresses come out of a /31 that is allocated by the LSR that originates the FA-LSP; the head-end address of the FA-LSP is the one specified as the IPv4 tunnel sender address; the remote (tail-end) address can then be inferred. If the LSP is bidirectional, the tail-end can thus know the addresses to assign to the reverse FA.

If there are multiple LSPs that all originate on one LSR and all terminate on another LSR, then at one end of the spectrum all these LSPs could be merged (under control of the head-end LSR) into a single FA using the concept of Link Bundling (see [BUNDLE]), while at the other end of the spectrum each such LSP could be advertised as its own adjacency.

When an FA is created under administrative control (static provisioning), the attributes of the FA-LSP have to be provided via configuration. Specifically, the following attributes may be configured for the FA-LSP: the head-end address (if left unconfigured, this defaults to the head-end LSR's Router ID); the tail-end address; bandwidth and resource colors constraints. The path taken by the FA-LSP may be either computed by the LSR at the head-end of the FA-LSP, or specified by explicit configuration; this choice is determined by configuration.

When an FA is created dynamically, the attributes of its FA-LSP are inherited from the LSP which induced its creation. Note that the

bandwidth of the FA-LSP must be at least as big as the LSP that induced it, but may be bigger if only discrete bandwidths are available for the FA-LSP. In general, for dynamically provisioned FAs, a policy-based mechanism may be needed to associate attributes to the FA-LSPs.

5.1. Traffic Engineering parameters

In this section, the Traffic Engineering parameters (see [OSPF-TE] and [ISIS-TE]) for FAs are described.

5.1.1. Link type (OSPF only)

The Link Type of an FA is set to "point-to-point".

5.1.2. Link ID (OSPF only)

The Link ID is set to the Router ID of the tail end of FA-LSP.

5.1.3. Local and remote interface IP address

If the FA is to be numbered, the local interface IP address (OSPF) or IPv4 interface address (ISIS) is set to the head-end address of the FA-LSP. The remote interface IP address (OSPF) or IPv4 neighbor address (ISIS) is set to the tail-end address of the FA-LSP.

5.1.4. Local and Remote Link Identifiers

For an unnumbered FA the assignment and handling of the local and remote link identifiers is specified in [UNNUM-RSVP], [UNNUM-CRLDP].

5.1.5. Traffic Engineering metric

By default the TE metric on the FA is set to max(1, (the TE metric of the FA-LSP path) - 1) so that it attracts traffic in preference to setting up a new LSP. This may be overridden via configuration at the head-end of the FA.

5.1.6. Maximum bandwidth

By default the Maximum Reservable Bandwidth and the initial Maximum LSP Bandwidth for all priorities of the FA is set to the bandwidth of the FA-LSP. These may be overridden via configuration at the head-end of the FA (note that the Maximum LSP Bandwidth at any one priority should be no more than the bandwidth of the FA-LSP).

5.1.7. Unreserved bandwidth

The initial unreserved bandwidth for all priority levels of the FA is set to the bandwidth of the FA-LSP.

5.1.8. Resource class/color

By default, a FA does not have resource colors (administrative groups). This may be overridden by configuration at the head-end of the FA.

5.1.9. Interface Switching Capability

The (near end) Interface Switching Capability associated with the FA is the (near end) Interface Switching Capability of the first link in the FA-LSP.

When the (near end) Interface Switching Capability field is PSC-1, PSC-2, PSC-3, or PSC-4, the specific information includes Interface MTU and Minimum LSP Bandwidth. The Interface MTU is the minimum MTU along the path of the FA-LSP; the Minimum LSP Bandwidth is the bandwidth of the LSP.

5.1.10. SRLG information

An FA advertisement could contain the information about the Shared Risk Link Groups (SRLG) for the path taken by the FA-LSP associated with that FA. This information may be used for path calculation by other LSRs. The information carried is the union of the SRLGs of the underlying TE links that make up the FA-LSP path; it is carried in the SRLG TLV in IS-IS or the SRLG sub-TLV of the TE Link TLV in OSPF. See [GMPLS-ISIS, GMPLS-OSPF] for details on the format of this information.

It is possible that the underlying path information might change over time, via configuration updates, or dynamic route modifications,

resulting in the change of the SRLG TLV.

If FAs are bundled (via link bundling), and if the resulting bundled link carries a SRLG TLV, it MUST be the case that the list of SRLGs in the underlying path followed by each of the FA-LSPs that form the component links is the same (note that the exact paths need not be the same).

6. Other considerations

It is expected that FAs will not be used for establishing ISIS/OSPF peering relation between the routers at the ends of the adjacency.

It may be desired in some cases that FAs only be used in Traffic Engineering path computations. In IS-IS, this can be accomplished by setting the default metric of the extended IS reachability TLV for the FA to the maximum link metric (2^24 - 1). In OSPF, this can be accomplished by not advertising the link as a regular LSA, but only as a TE opaque LSA.

7. Controlling FA-LSPs boundaries

To facilitate controlling the boundaries of FA-LSPs this document introduces two new mechanisms: Interface Switching Capability (see [GMPLS-ISIS, GMPLS-OSPF], and "LSP region" (or just "region").

7.1. LSP regions

The information carried in the Interface Switching Capabilities is used to construct LSP regions, and determine regions' boundaries as follows.

Define an ordering among interface switching capabilities as follows: PSC-1 < PSC-2 < PSC-3 < PSC-4 < TDM < LSC < FSC. Given two interfaces if-1 and if-2 with interface switching capabilities isc-1 and isc-2 respectively, say that if-1 < if-2 iff isc-1 < isc-2 or isc-1 == isc-2 == TDM, and if-1's max LSP bandwidth is less than if-2's max LSP bandwidth.

Suppose an LSP's path is as follows: node-0, link-1, node-1, link-2, node-2, ..., link-n, node-n. Moreover, for link-i denote by [link-i, node-(i-1)] the interface that connects link-i to node-(i-1), and by [link-i, node-i] the interface that connects link-i to node-i.

If [link-(i+1), node-i)] < [link-(i+1), node-(i+1)], we say that the

LSP has crossed a region boundary at node-i; with respect to that LSP path, the LSR at node-i is an edge LSR. The 'other edge' of the region with respect to the LSP path is node-k, where k is the smallest number greater than i such that [link-(i+1), node-(i+1)] equal [link-k, node-(k-1)], and [link-k, node-(k-1)] > [link-k, node-k].

Path computation may take into account region boundaries when computing a path for an LSP. For example, path computation may restrict the path taken by an LSP to only the links whose Interface Switching Capability is PSC-1.

Note that an interface may have multiple Interface Switching Capabilities. In such a case, the test whether if-i < if-j depends on the Interface Switching Capabilities chosen for if-i and if-j, which in turn determines whether or not there is a region boundary at node-i.

8. Signalling aspects

In this section we describe procedures that an LSR at the head-end of an FA uses for handling LSP setup originated by other LSR.

As we mentioned before, establishment/termination of FA-LSPs may triggered either by mechanisms outside of GMPLS (e.g., via administrative control), or by mechanisms within GMPLS (e.g., as a result of the LSR at the edge of an aggregate LSP receiving LSP setup requests originated by some other LSRs beyond LSP aggregate and its edges). Procedures described in Section "Common procedures" applied to both cases. Procedures described in Section "Specific procedures" apply only to the latter case.

8.1. Common procedures

For the purpose of processing the ERO in a Path/Request message of an LSP that is to be tunneled over an FA , an LSR at the head-end of the FA-LSP views the LSR at the tail of that FA-LSP as adjacent (one IP hop away).

How this is to be achieved for RSVP-TE and CR-LDP is described in the following subsections.

In either case (RSVP-TE or CR-LDP), when an LSP is tunneled through an FA-LSP, the LSR at the head-end of the FA-LSP subtracts the LSP's bandwidth from the unreserved bandwidth of the FA.

In the presence of link bundling (when link bundling is applied to FAs), when an LSP is tunneled through an FA-LSP, the LSR at the head-end of the FA-LSP also need to adjust Max LSP bandwidth of the FA.

8.1.1. RSVP-TE

If one uses RSVP-TE to signal an LSP to be tunneled over an FA-LSP, then the Path message MUST contain an IF_ID RSVP_HOP object [GRSVP-TE, GSIG] instead of an RSVP_HOP object; and the data interface identification MUST identify the FA-LSP.

The preferred method of sending the Path message is to set the destination IP address of the Path message to the computed NHOP for that Path message. This NHOP address must be a routable address; in the case of separate control and data planes, this must be a control plane address.

Furthermore, the IP header for the Path message MUST NOT have the Router Alert option. The Path message is intended to be IP-routed to the tail end of the FA-LSP without being intercepted and processed as an RSVP message by any of the intermediate nodes.

Finally, the IP TTL vs. RSVP TTL check MUST NOT be made. In general, if the IF_ID RSVP_HOP object is used, this check must be disabled, as the number of hops over the control plane may be greater than one. Instead, the following check is done by the receiver Y of the IF_ID RSVP_HOP object:

1. Make sure that the data interface identified in the IF_ID RSVP_HOP object actually terminates on Y.
2. Find the "other end" of the above data interface, say X. Make sure that the PHOP in the IF_ID RSVP_HOP object is a control channel address that belongs to the same node as X.

How check #2 is carried out is beyond the scope of this document; suffice it to say that it may require a Traffic Engineering Database, or the use of LMP [LMP] or yet other means.

An alternative method is to encapsulate the Path message in an IP tunnel (or, in the case that the Interface Switching Capability of the FA-LSP is PSC[1-4], in the FA-LSP itself), and unicast the message to the tail end of the FA-LSP, without the Router Alert option. This option may be needed if intermediate nodes process RSVP messages regardless of whether the Router Alert option is present.

A PathErr sent in response to a Path message with an IF_ID RSVP_HOP object SHOULD contain an IF_ID HOP object. (Note: a PathErr does not

normally carry an RSVP_HOP object, but in the case of separated control and data, it is necessary to identify the data channel in the PathErr message.)

The Resv message back to the head-end of the FA-LSP (PHOP) is IP-routed to the PHOP in the Path message. If necessary, Resv Messages MAY be encapsulated in another IP header whose destination IP address is the PHOP of the received Path message.

8.1.2. CR-LDP

If one uses CR-LDP to signal an LSP to be tunneled over an FA-LSP, then the Request message MUST contain an IF_ID TLV [GCR-LDP] object; and the data interface identification MUST identify the FA-LSP.

Furthermore, the head end LSR must create a targetted LDP session with the tail end LSR. The Request (Mapping) message is unicast from the head end (tail end) to the tail end (head end).

8.2. Specific procedures

When an LSR receives a Path/Request message, the LSR determines whether it is at the edge of a region with respect to the ERO carried in the message. The LSR does this by looking up the interface switching capabilities of the previous hop and the next hop in its IGP database, and comparing them using the relation defined in Section "Specific procedures". If the LSR is not at the edge of a region, the procedures in this section do not apply.

If the LSR is at the edge of a region, it must then determine the other edge of the region with respect to the ERO, again using the IGP database. The LSR then extracts from the ERO the subsequence of hops from itself to the other end of the region.

The LSR then compares the subsequence of hops with all existing FA-LSPs originated by the LSR; if a match is found, that FA-LSP has enough unreserved bandwidth for the LSP being signaled, and the L3PID of the FA-LSP is compatible with the L3PID of the LSP being signaled, the LSR uses that FA-LSP as follows. The Path/Request message for the original LSP is sent to the egress of the FA-LSP, not to the next hop along the FA-LSP's path. The PHOP in the message is the address of the LSR at the head-end of the FA-LSP. Before sending the Path/Request message, the ERO in that message is adjusted by removing the subsequence of the ERO that lies in the FA-LSP, and replacing it with just the end point of the FA-LSP.

Otherwise (if no existing FA-LSP is found), the LSR sets up a new FA-LSP. That is, it initiates a new LSP setup just for the FA-LSP. Note that the new LSP may traverse either 'basic' TE links or FAs.

After the LSR establishes the new FA-LSP, the LSR announces this LSP into IS-IS/OSPF as an FA.

The unreserved bandwidth of the FA is computed by subtracting the bandwidth of sessions pending the establishment of the FA-LSP associated from the bandwidth of the FA-LSP.

An FA-LSP could be torn down by the LSR at the head-end of the FA-LSP as a matter of policy local to the LSR. It is expected that the FA-LSP would be torn down once there are no more LSPs carried by the FA-LSP. When the FA-LSP is torn down, the FA associated with the FA-LSP is no longer advertised into IS-IS/OSPF.

8.3. FA-LSP Holding Priority

The value of the holding priority of an FA-LSP must be the minimum of the configured holding priority of the FA-LSP and the holding priorities of the LSPs tunneling through the FA-LSP (note that smaller priority values denote higher priority). Thus, if an LSP of higher priority than the FA-LSP tunnels through the FA-LSP, the FA-LSP is itself promoted to the higher priority. However, if the tunneled LSP is torn down, the FA-LSP need not drop its priority to its old value right away; it may be advisable to apply hysteresis in this case.

If the holding priority of an FA-LSP is configured, this document restricts it to 0.

9. Security Considerations

From a security point of view, the primary change introduced in this document is that the implicit assumption of a binding between data interfaces and the interface over which a control message is sent is no longer valid.

This means that the "sending interface" or "receiving interface" is no longer well defined, as the interface over which an RSVP message is sent may change as routing changes; therefore, mechanisms that depend on these concepts (for example, the definition of a security association) need a clearer definition.

[RFC2747] provides a solution: in section 2.1, under "Key

Identifier", an IP address is a valid identifier for the sending (and by analogy, receiving) interface. Since RSVP messages for a given LSP are sent to an IP address that identifies the next/previous hop for the LSP, one can replace all occurrences of 'sending [receiving] interface' with 'receiver's [sender's] IP address' (respectively). For example, in Section 4, third paragraph, instead of:

> "Each sender SHOULD have distinct security associations (and keys) per secured sending interface (or LIH). ... At the sender, security association selection is based on the interface through which the message is sent."

read:

> "Each sender SHOULD have distinct security associations (and keys) per secured receiver's IP address. ... At the sender, security association selection is based on the IP address to which the message is sent."

Note that CR-LDP does not have this issue, as CR-LDP messages are sent over TCP sessions, and no assumption is made that these sessions are to direct neighbors. The recommended mechanism for authentication and integrity of LDP message exchange is to use the TCP MD5 option [LDP].

Another consequence (relevant to RSVP) of the changes proposed in this document is that IP destination address of Path messages be set to the receiver's address, not to the session destination. Thus, the objections raised in section 1.2 of [RFC2747] should be revisited to see if IPSec AH is now a viable means of securing RSVP-TE messages.

10. Acknowledgements

Many thanks to Alan Hannan, whose early discussions with Yakov Rekhter contributed greatly to the notion of Forwarding Adjacencies. We would also like to thank George Swallow, Quaizar Vohra and Ayan Banerjee.

11. Normative References

[GCR-LDP] Ashwood-Smith, Berger et al, "Generalized MPLS - CR-LDP Extensions" (work in progress).

[GMPLS-ISIS] Kompella, K., Rekhter, Y., Banerjee, A. et al, "IS-IS Extensions in Support of Generalized MPLS", (work in progress).

[GMPLS-OSPF] Kompella, K., Rekhter, Y., Banerjee, A. et al, "OSPF Extensions in Support of Generalized MPLS", (work in progress).

[GRSVP-TE] Ashwood-Smith, Berger et al, "Generalized MPLS - RSVP-TE Extensions" (work in progress).

[GSIG] Ashwood-Smith, Berger et al, "Generalized MPLS - Signaling Functional Description" (work in progress).

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

[LDP] "Label Distribution Protocol", RFC3036

[OSPF-TE] Katz, D., Yeung, D., Kompella, K., "Traffic Engineering Extensions to OSPF", (work in progress).

[UNNUM-CRLDP] Kompella, K., Rekhter, Y., Kullberg, A., "Signalling Unnumbered Links in CR-LDP", (work in progress).

[UNNUM-RSVP] Kompella, K., Rekhter, Y., "Signalling Unnumbered Links in RSVP", (work in progress).

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC2747] Baker, F., Lindell, B., and M. Talwar, "RSVP Cryptographic Authentication", RFC 2747, January 2000.

12. Non-normative References

[BUNDLE] Kompella, K., Rekhter, Y., Berger, L., "Link Bundling in MPLS Traffic Engineering", (work in progress).

[LMP] "Link Management Protocol (LMP)", draft-ietf-ccamp-lmp-02.txt, (work in progress)

13. Author Information

Kireeti Kompella
Juniper Networks, Inc.
1194 N. Mathilda Ave
Sunnyvale, CA 94089
e-mail: kireeti@juniper.net

Yakov Rekhter
Juniper Networks, Inc.
1194 N. Mathilda Ave
Sunnyvale, CA 94089
e-mail: yakov@juniper.net