Network Working Group
Internet Draft
Expiration Date: January 2001

Kireeti Kompella
Juniper Networks
Yakov Rekhter
Cisco Systems

LSP Hierarchy with MPLS TE

draft-ietf-mpls-lsp-hierarchy-00.txt

1. Status of this Memo

2. Abstract

To improve scalability of MPLS TE it may be useful to aggregate TE LSPs. The aggregation is accomplished by (a) an LSR creating a TE LSP, (b) the LSR forming a forwarding adjacency out of that LSP (advertising this LSP as a link into ISIS/OSPF), (c) allowing other LSRs to use forwarding adjacencies 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. Overview

An LSR uses MPLS TE procedures to create and maintain an LSP. The LSR then may (under its local configuration control) announce this LSP as a link into ISIS/OSPF. When this link is advertised into the same instance of ISIS/OSPF as the one that determines the route taken by the LSP, we call such a link a "forwarding adjacency". We refer to the LSP as the "forwarding adjacency LSP", or just FA-LSP. Note that since the advertised entity is a link in ISIS/OSPF, both the end point LSRs of the FA-LSP must belong to the same ISIS level/OSPF area.

In general, creation/termination of a forwarding adjacency and its FA-LSP could be driven either by mechanisms outside of MPLS (e.g., via configuration control on the LSR at the head-end of the adjacency), or by mechanisms within MPLS (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 forwarding adjacencies just as it floods the information about any other links. As a result of this flooding, an LSR has in its link state database the information about not just conventional links, but forwarding adjacencies as well.

An LSR, when performing path computation, uses not just conventional links, but forwarding adjacencies as well. Once a path is computed, the LSR uses RSVP/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.

4. Routing aspects

In this section we describe procedures for constructing forwarding adjacencies out of LSPs, and handling of forwarding adjacencies 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 5.

Forwarding adjacencies may be represented as either unnumbered or numbered links. In the former case the link IP addresses of forwarding adjacencies are the router IDs on the two ends of the link. In the latter case the link IP addresses of forwarding adjacencies could be addresses assigned to some "virtual" interfaces

on a router (it is assumed that a router may have multiple virtual interfaces).

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 forwarding adjacency 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 a forwarding adjacency is created under administrative control (static provisioning), the attributes of this adjacency 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 must default to the head-end LSR's Router ID); the tail-end address (this MUST be configured, and must be either the Router ID of the tail-end LSR of the forwarding adjacency, or an interface address on the tail-end LSR); bandwidth and resource colors constraints. The path taken by the FA-LSP may be either computed by the by the LSR at the head-end of the FA-LSP, or specified by explicit configuration; this choice is determined by configuration.

When a forwarding adjacency is created dynamically, its attributes 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 forwarding adjacencies, a policy-based mechanism may be needed to associate attributes to forwarding adjacencies.

This document restricts the holding priority of the FA-LSP to 0, regardless of how the FA-LSP is created.

4.1. Traffic Engineering parameters

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

4.1.1. Link type (OSPF only)

The Link Type of a forwarding adjacency is set to "point-to-point".

4.1.2. Link ID (OSPF only)

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

4.1.3. Local and remote interface IP address

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.

4.1.4. Traffic Engineering metric

By default the TE metric on the forwarding adjacency 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 forwarding adjacency.

4.1.5. Maximum bandwidth

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

4.1.6. Unreserved bandwidth

By default, the initial unreserved bandwidth for all priority levels of the forwarding adjacency is set to the bandwidth of the FA-LSP.

4.1.7. Resource class/color

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

4.1.8. Link Mux Capability

The Link Mux Capability (see Section 4.3.1) associated with the forwarding adjacency is the Link Mux Capability of the last link in the FA-LSP.

4.1.9. Path information

A forwarding adjacency advertisement could contain the information about the path taken by the FA-LSP associated with that forwarding adjacency. This information may be used for path calculation by other LSRs. This information is carried in the Path sub-TLV, which is a sub-TLV of the Link Mux Capability TLV. In both IS-IS and OSPF, this sub-TLV is encoded as follows: the type is 1, the length is 4 times the path length, and the value is a list of 4 octet IPv4 addresses identifying the links in the order that they form the path of the forwarding adjacency.

It is possible that the underlying Path sub-TLV might change over time, via configuration updates, or dynamic route modifications. If forwarding adjacencies are bundled (via link bundling), and if the resulting bundled link carries a Path sub-TLV, it MUST be the case that the underlying path followed by each of the FA-LSPs that form the component links is the same.

4.2. Other considerations

It is expected that forwarding adjacencies 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 forwarding adjacencies 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.

Since LSPs are in general unidirectional, it follows that forwarding adjacencies are (by definition) unidirectional links. Therefore, the TE path computation procedures should not perform two-way connectivity check on the links used by the procedures.

4.3. Controlling FA-LSPs boundaries

To facilitate controlling the boundaries of FA-LSPs this document introduces two new mechanisms: Link Mux Capability, and "LSP region" (or just "region").

4.3.1. Link Mux Capability TLV

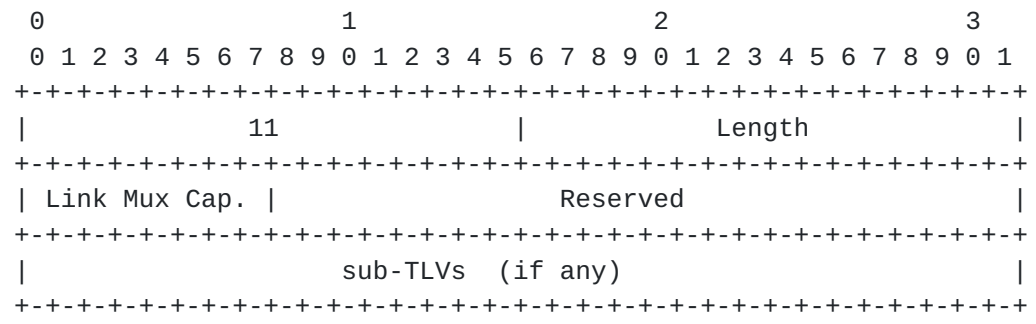
Associated with each link (including forwarding adjacencies) is a new attribute - Link Mux Capability. In this section we define the Link Mux Capability TLV and describe the various values it can take.

A network may have links with different multiplexing/demultiplexing capabilities. For example, a node may not be able to demultiplex individual packets on a given link, but it may be able to multiplex/demultiplex channels within a SONET payload. The Link Mux Capability TLV identifies the associated multiplexing/demultiplexing capability of a link. At present, the Link Mux Capability TLV has one defined sub-TLV, the Path TLV, described in section 4.1.9.

In ISIS the Link Mux Capability is a sub-TLV of the extended IS reachability TLV (type 22) as defined in [ISIS-TE]. The type of the Link Mux Capability TLV is 19. The length of the TLV is one octet plus the length of sub-TLVs of the Link Mux Capability TLV. The value field of the TLV contains the Link Mux Capability, encoded as follows:

Value	Link Mux Capabilities
1	Packet-Switch Capable-1 (PSC-1)
2	Packet-Switch Capable-2 (PSC-2)
3	Packet-Switch Capable-3 (PSC-3)
4	Packet-Switch Capable-4 (PSC-4)
50	Time-Division-Multiplex Capable (TDM)
100	Lambda-Switch Capable (LSC)
150	Fiber-Switch Capable (FSC)

In the OSPF Traffic Engineering LSA, the Link Mux Capability TLV is a sub-TLV of the Link TLV as defined in [OSPF-TE], with type 11 and length of four octets plus the length of the sub-TLVs of the Link Mux Capability TLV. The value field is taken from the above list. The format of the Link Mux Capability sub-TLV is as shown below:



If a link is of type PSC-1 through PSC-4, that means that the node receiving data over this link can demultiplex (switch) the received data on a packet-by-packet basis. The various levels of PSC establish a hierarchy of LSPs tunneled within LSPs.

If a link is of type TDM, that means that the node receiving data over this link can multiplex or demultiplex channels within a SONET/SDH payload.

If a link is of type LSC, that means that the node receiving data over this link can recognize and switch individual lambdas within the link (fiber).

If a link is of type FSC, that means that the node receiving data over this link (fiber) can switch the entire contents to another link (fiber) (without distinguishing lambdas, channels or packets).

Note that the node that is advertising a given link (i.e., the node that is transmitting) needs to know the multiplex/demultiplex capacbilities at the other end of the link (i.e., the receiving end of the link). This is accomplished through coordinated configuration between the nodes, at each end of the link.

4.3.2. LSP regions

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

Define an ordering among link mux capabilities as follows: PSC-1 < PSC-2 < PSC-3 < PSC-4 < TDM < LSC < FSC. Given two links link-1 and link-2 with link types lmc-1 and lmc-2 respectively, say that link-1 < link-2 iff lmc-1 < lmc-2 or lmc-1 == lmc-2 == TDM, and link-1's bandwidth is less than link-2's switching granularity.

Furthermore, say that link-1 is compatible with link-2 iff:

lmc-1 equals lmc-2 and neither is of type TDM; OR

lmc-1 equals lmc-2 equals TDM and both links have the same speed;
OR
lmc-1 and lmc-2 are both packet-switch capable.

Suppose an LSP's path is as follows: node-0, link-1, node-1, link-2, node-2, ..., link-n, node-n. If link-i < link-(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+1 such that link-k is compatible with link-i.

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 Link Mux Capability is PSC-1.

5. Signalling aspects

In this section we describe procedures that an LSR at the head-end of a forwarding adjacency 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 MPLS (e.g., via administrative control), or by mechanisms within MPLS (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 5.1 applied to both cases. Procedures described in Section 5.2 apply only to the latter case.

5.1. Common procedures

For the purpose of processing the ERO in a Path/Request message of an LSP that is to be tunneled over a forwarding adjacency, 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).

If the LSR at the tail of the FA-LSP is capable of packet switching, the Path/Request message for the tunneled LSP can itself be tunneled over the FA-LSP. If the encapsulation on the carrier LSP can distinguish IP from MPLS, the Path/Request message is sent as a plain IP packet. Otherwise, the Path/Request message is sent with a label of 0, meaning "pop the label and treat as IP".

If the LSR at the tail of the FA-LSP is not capable of packet

switching, the Path message is unicast over the control plane to the tail of the carrier LSP, without the Router Alert option. The whole Path message, including IP header, MUST also be encapsulated in another IP header whose destination IP address matches the tail's IP address.

The Resv/Mapping message back to the head-end of the FA-LSP (PHOP) cannot be sent over the same FA-LSP as it is unidirectional. The Resv/Mapping message can either take any LSP whose end-point is the head-end of the FA- LSP, or be unicast over the control plane to the head-end. RSVP Resv Messages SHOULD be encapsulated in another IP header whose destination IP address matches the head-end's IP address.

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 forwarding adjacency. In the presence of link bundling (when link bundling is applied to forwarding adjacencies), 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 forwarding adjacency.

5.2. 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 link types of the previous hop and the next hop in its IGP database, and comparing them using the relation defined in Section 4.3.2. 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 the strict hop subsequence from itself to the other end of the region.

The LSR then compares the strict hop subsequence 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.

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

The unreserved bandwidth of the forwarding adjacency 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 forwarding adjacency associated with the FA-LSP is no longer advertised into IS-IS/OSPF.

6. Security Considerations

Security issues are not discussed in this document.

7. 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.

8. References

[BUNDLE] Kompella, K., Rekhter, Y., Berger, L., "Link Bundling in MPLS Traffic Engineering", draft-kompella-mpls-bundle-02.txt (work in progress)

[ISIS-TE] Smit, H., Li, T., "IS-IS extensions for Traffic Engineering", draft-ietf-isis-traffic-01.txt (work in progress)

[OSPF-TE] Katz, D., Yeung, D., "Traffic Engineering Extensions to OSPF", draft-katz-yeung-ospf-traffic-01.txt (work in progress)

9. Author Information

Kireeti Kompella
Juniper Networks, Inc.
385 Ravendale Drive
Mountain View, CA 94043
e-mail: kireeti@juniper.net

Yakov Rekhter
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134
e-mail: yakov@cisco.com