Internet Engineering Task Force (IETF)
Request for Comments: 5712
Category: Standards Track
ISSN: 2070-1721

M. Meyer, Ed.
British Telecom
JP. Vasseur, Ed.
Cisco Systems, Inc.

January 2010

MPLS Traffic Engineering Soft Preemption

Abstract

This document specifies Multiprotocol Label Switching (MPLS) Traffic Engineering Soft Preemption, a suite of protocol modifications extending the concept of preemption with the goal of reducing or eliminating traffic disruption of preempted Traffic Engineering Label Switched Paths (TE LSPs). Initially, MPLS RSVP-TE was defined with support for only immediate TE LSP displacement upon preemption. The utilization of a reroute request notification helps more gracefully mitigate the reroute process of preempted TE LSP. For the brief period soft preemption is activated, reservations (though not necessarily traffic levels) are in effect under-provisioned until the TE LSP(s) can be rerouted. For this reason, the feature is primarily, but not exclusively, interesting in MPLS-enabled IP networks with Differentiated Services and Traffic Engineering capabilities.

Status of This Memo

This is an Internet Standards Track document.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Further information on Internet Standards is available in Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc5712.

Table of Contents

1. Introduction

In a Multiprotocol Label Switching (MPLS) Resource Reservation Protocol Traffic Engineering (RSVP-TE) (see [RFC3209]) enabled IP network, hard preemption is the default behavior. Hard preemption provides no mechanism to allow preempted Traffic Engineering Label Switched Paths (TE LSPs) to be handled in a make-before-break fashion: the hard preemption scheme instead utilizes a very intrusive method that can cause traffic disruption for a potentially large amount of TE LSPs. Without an alternative, network operators either accept this limitation, or remove functionality by using only one preemption priority or using invalid bandwidth reservation values. Understandably desirable features like TE reservation adjustments that are automated by the ingress Label Edge Router (LER) are less palatable when preemption is intrusive and maintaining high levels of network stability levels is a concern.

This document defines the use of additional signaling and maintenance mechanisms to alert the ingress LER of the preemption that is pending and allow for temporary control-plane under-provisioning while the preempted tunnel is rerouted in a non-disruptive fashion (make-before-break) by the ingress LER. During the period that the tunnel is being rerouted, link capacity is under-provisioned on the midpoint where preemption initiated and potentially one or more links upstream along the path where other soft preemptions may have occurred.

2. Terminology

This document follows the nomenclature of the MPLS Architecture defined in [RFC3031].

2.1. Acronyms and Abbreviations

CSPF: Constrained Shortest Path First.

DS: Differentiated Services.

LER: Label Edge Router.

LSR: Label Switching Router.

LSP: Label Switched Path.

MPLS: MultiProtocol Label Switching.

RSVP: Resource ReSerVation Protocol.

TE LSP: Traffic Engineering Label Switched Path.

2.2. Nomenclature

Point of Preemption - the midpoint or ingress LSR which due to RSVP provisioning levels is forced to either hard preempt or under-provision and signal soft preemption.

Hard Preemption - The (typically default) preemption process in which higher numeric priority TE LSPs are intrusively displaced at the point of preemption by lower numeric priority TE LSPs. In hard preemption, the TE LSP is torn down before reestablishment.

2.3. Requirements Language

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].

3. Motivations

Initially, MPLS RSVP-TE [RFC3209] was defined with support for only one method of TE LSP preemption, which immediately tears down TE LSPs, disregarding the preempted in-transit traffic. This simple but abrupt process nearly guarantees preempted traffic will be discarded, if only briefly, until the RSVP Path Error message reaches and is processed by the ingress LER and a new data path can be established. The Error Code and Error Values carried within the RSVP Path Error message to report a preemption action are documented in [RFC5711]. Note that such preemption is also referred to as a fatal error in [RFC5711]. In cases of actual resource contention this might be helpful; however, preemption may be triggered by mere reservation contention, and reservations may not reflect data-plane contention up to the moment. The result is that when conditions that promote preemption exist and hard preemption is the default behavior, inferior priority preempted traffic may be needlessly discarded when sufficient bandwidth exists for both the preempted TE LSP and the preempting TE LSP(s).

Hard preemption may be a requirement to protect numerically lower preemption priority traffic in a non-Diffserv-enabled architecture, but in a Diffserv-enabled-architecture, one need not rely exclusively upon preemption to enforce a preference for the most valued traffic since the marking and queuing disciplines should already be aligned for those purposes. Moreover, even in non-Diffserv-aware networks, depending on the TE LSP sizing rules (imagine all LSPs are sized at double their observed traffic level), reservation contention may not accurately reflect the potential for data-plane congestion.

4. RSVP Extensions

4.1. SESSION-ATTRIBUTE Flags

To explicitly signal the desire for a TE LSP to benefit from the soft preemption mechanism (and thus not to be hard preempted if the soft preemption mechanism is available), the following flag of the SESSION-ATTRIBUTE object (for both the C-Type 1 and 7) is defined:

Soft Preemption Desired bit

Bit Flag	Name Flag
0x40	Soft Preemption Desired

4.2. Path Error - "Reroute Request Soft Preemption" Error Value

[RFC5710] specifies defines a new reroute-specific error code that allows a midpoint to report a TE LSP reroute request (Error Code=34 - Reroute). This document specifies a new Error Value sub-code for the case of soft preemption.

Error-value	Meaning	Reference
1	Reroute Request Soft Preemption	This document

Upon (soft) preemption, the preempting node MUST issue a PathErr message with the Error Code=34 ("Reroute") and a value=1 ("Reroute Request Soft Preemption").

5. Mode of Operation

Let's consider the following example:

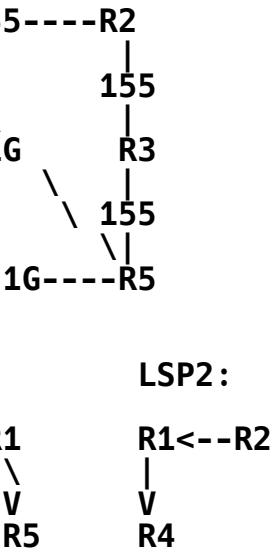

Figure 1: Example of Soft Preemption Operation

In the network depicted above in Figure 1, consider the following conditions:

- Reservable BW on R0-R1, R1-R5, and R4-R5 is 1 Gbit/s.
- Reservable BW on R1-R2, R1-R4, R2-R3, and R3-R5 is 155 Mbit/s.
- Bandwidths and costs are identical in both directions.
- Each circuit has an IGP metric of 10, and the IGP metric is used by CSPF.
- Two TE tunnels are defined:
  - LSP1: 155 Mbit/s, setup/hold priority 0 tunnel, path R0-R1-R5.
  - LSP2: 155 Mbit/s, setup/hold priority 7 tunnel, path R2-R1-R4.

  Both TE LSPs are signaled with the "Soft Preemption Desired" bit of their SESSION-ATTRIBUTE object set.
- Circuit R1-R5 fails.
- Soft Preemption is functional.

When the circuit R1-R5 fails, R1 detects the failure and sends an updated IGP LSA/LSP and Path Error message to all the head-end LSRs that have a TE LSP traversing the failed link (R0 in the example above). Either form of notification may arrive at the head-end LSRs first. Upon receiving the link failure notification, R0 triggers a TE LSP reroute of LSP1, and re-signals LSP1 along shortest path available satisfying the TE LSP constraints: R0-R1-R4-R5 path. The Resv messages for LSP1 travel in the upstream direction (from the destination to the head-end LSR -- R5 to R0 in this example). LSP2 is soft preempted at R1 as it has a numerically lower priority value, and both bandwidth reservations cannot be satisfied on the R1-R4 link.

Instead of sending a PathTear message for LSP2 upon preemption as with hard preemption (which would result in an immediate traffic disruption for LSP2), R1's local bandwidth accounting for LSP2 is zeroed, and a PathErr message with error code "Reroute" and a value "Reroute Request Soft Preemption" for LSP2 is issued.

Upon reception of the PathErr message for LSP2, R2 may update the working copy of the TE-DB before calculating a new path for the new LSP. In the case that Diffserv [RFC3270] and TE [RFC3209] are deployed, receiving a "preemption pending" notification may imply to a head-end LSR that the available bandwidth for the affected priority level and numerically greater priority levels has been exhausted for the indicated node interface. R2 may choose to reduce or zero the available bandwidth for the implied priority range until more accurate information is available (i.e., a new IGP TE update is received). It follows that R2 re-computes a new path and performs a non-traffic-disruptive rerouting of the new TE LSP T2 by means of the make-before-break procedure. The old path is then torn down.

6. Elements Of Procedures

6.1. On a Soft Preempting LSR

When a new TE LSP is signaled that requires a set of TE LSP(s) to be preempted because not all TE LSPs can be accommodated on a specific interface, a node triggers a preemption action that consists of selecting the set of TE LSPs that must be preempted so as to free up some bandwidth in order to satisfy the newly signaled numerically lower preemption TE LSP.

With hard preemption, when a TE LSP is preempted, the preempting node sends an RSVP PathErr message that serves as notification of a fatal action as documented in [RFC5711]. Upon receiving the RSVP PathErr message, the head-end LSR sends an RSVP PathTear message, that would result in an immediate traffic disruption for the preempted TE LSP.

By contrast, the mode of operation with soft preemption is as follows: the preempting node's local bandwidth accounting for the preempted TE LSP is zeroed and a PathErr with error code "Reroute", and a error value "Reroute Request Soft Preemption" for that TE LSP is issued upstream toward the head-end LSR.

If more than one soft preempted TE LSP has the same head-end LSR, these soft preemption PathErr notification messages may be bundled together.

The preempting node MUST immediately send a PathErr with error code "Reroute" and a error value "Reroute Request Soft Preemption" for each soft preempted TE LSP. The node MAY use the occurrence of soft preemption to trigger an immediate IGP update or influence the scheduling of an IGP update.

To guard against a situation where bandwidth under-provisioning will last forever, a local timer (named the "Soft preemption timer") MUST be started on the preemption node upon soft preemption. If this timer expires, the preempting node SHOULD send an RSVP PathTear and either a ResvTear message or a PathErr with the 'Path_State_Removed' flag set.

Should a refresh event for a soft preempted TE LSP arrive before the soft preemption timer expires, the soft preempting node MUST continue to refresh the TE LSP.

When the MESSAGE-ID extensions defined in [RFC2961] are available and enabled, PathErr messages with the error code "Reroute" and error value "Reroute Request Soft Preemption" SHOULD be sent in reliable mode.

The preempting node MAY preempt TE LSPs that have a numerically higher Holding priority than the Setup priority of the newly admitted LSP. Within the same priority, first it SHOULD attempt to preempt LSPs with the "Soft Preemption Desired" bit of the SESSION ATTRIBUTE object cleared, i.e., the TE LSPs that are considered as Hard Preemptable.

Selection of the preempted TE LSP at a preempting midpoint: when a numerically lower priority TE LSP is signaled that requires the preemption of a set of numerically higher priority LSPs, the node where preemption is to occur has to make a decision on the set of TE LSP(s) that are candidates for preemption. This decision is a local decision and various algorithms can be used, depending on the objective (e.g, see [RFC4829]). As already mentioned, soft preemption causes a temporary link under-provisioning condition while the soft preempted TE LSPs are rerouted by their respective head-end

LSRs. In order to reduce this under-provisioning exposure, a soft preempting LSR MAY check first if there exists soft preemptable TE LSP bandwidth that is flagged by another node but still available for soft preemption locally. If sufficient overlap bandwidth exists, the LSR MAY attempt to soft preempt the same TE LSP. This would help reduce the temporarily elevated under-provisioning ratio on the links where soft preemption occurs and reduce the number of preempted TE LSPs. Optionally, a midpoint LSR upstream or downstream from a soft preempting node MAY choose to flag the TE LSPs in soft preempted state. In the event a local preemption is needed, the LSPs that are in the cache and of the relevant priority level are soft preempted first, followed by the normal soft and hard preemption selection process for the given priority.

Under specific circumstances such as unacceptable link congestion, a node MAY decide to hard preempt a TE LSP (by sending a fatal Path Error message, a PathTear, and either a ResvTear or a Path Error message with the 'Path_State_Removed' flag set) even if its head-end LSR explicitly requested soft preemption (by setting the "Soft Preemption Desired" flag of the corresponding SESSION-ATTRIBUTE object). Note that such a decision MAY also be made for TE LSPs under soft preemption state.

6.2. On Head-end LSR of a Soft Preempted TE LSP

Upon reception of a PathErr message with error code "Reroute" and an error value "Reroute request soft preemption", the head-end LSR MAY first update the working copy of the TE-DB before computing a new path (e.g., by running CSPF) for the new LSP. In the case that Diffserv [RFC3270] and MPLS Traffic Engineering [RFC3209] are deployed, receiving "preemption pending" may imply to a head-end LSR that the available bandwidth for the affected priority level and numerically greater priority levels has been exhausted for the indicated node interface. A head-end LSR MAY choose to reduce or zero the available bandwidth for the implied priority range until more accurate information is available (i.e., a new IGP TE update is received).

Once a new path has been computed, the soft preempted TE LSP is rerouted using the non-traffic-disruptive make-before-break procedure. The amount of time the head-end node avoids using the node interface identified by the IP address contained in the PathErr is based on a local decision at the head-end node.

As a result of soft preemption, no traffic will be needlessly black-holed due to mere reservation contention. If loss is to occur, it will be due only to an actual traffic congestion scenario and according to the operator's Diffserv (if Diffserv is deployed) and queuing scheme.

7. Interoperability

Backward compatibility should be assured as long as the implementation followed the recommendations set forth in [RFC3209].

As mentioned previously, to guard against a situation where bandwidth under-provisioning will last forever, a local timer (soft preemption timer) MUST be started on the preemption node upon soft preemption. When this timer expires, the soft preempted TE LSP SHOULD be hard preempted by sending a fatal Path Error message, a PathTear message, and either a ResvTear message or a PathErr message with the 'Path_State_Removed' flag set. This timer SHOULD be configurable, and a default value of 30 seconds is RECOMMENDED.

It is RECOMMENDED that configuring the default preemption timer to 0 will cause the implementation to use hard-preemption.

Soft preemption as defined in this document is designed for use in MPLS RSVP-TE enabled IP networks and may not functionally translate to some GMPLS technologies. As with backward compatibility, if a device does not recognize a flag, it should pass the subobject transparently.

8. Management

Both the point of preemption and the ingress LER SHOULD provide some form of accounting internally and to the network operator interface with regard to which TE LSPs and how much capacity is under-provisioned due to soft preemption. Displays of under-provisioning are recommended for the following midpoint, ingress, and egress views:

- Sum of current bandwidth per preemption priority per local interface
- Sum of current bandwidth total per local interface
- Sum of current bandwidth per local router (ingress, egress, midpoint)
- List of current LSPs and bandwidth in PPend (preemption pending) status

- List of current sum bandwidth and session count in PPend status per observed Explicit Route Object (ERO) hops (ingress and egress views only).

- Cumulative PPend events per observed ERO hop.

9. IANA Considerations

9.1. New Session Attribute Object Flag

A new flag of the Session Attribute Object has been registered by IANA.

Soft Preemption Desired bit

Bit Flag	Name	Reference
0x40	Soft Preemption Desired	This document

9.2. New Error Sub-Code Value

[RFC5710] defines a new reroute-specific error code that allows a midpoint to report a TE LSP reroute request. This document specifies a new error sub-code value for the case of Soft Preemption.

Error-value	Meaning	Reference
1	Reroute Request Soft Preemption	This document

10. Security Considerations

This document does not introduce new security issues. The security considerations pertaining to the original RSVP protocol [RFC3209] remain relevant. Further details about MPLS security considerations can be found in [SEC_FMWK].

As noted in Section 6.1, soft preemption may result in temporary link under provisioning condition while the soft preempted TE LSPs are rerouted by their respective head-end LSRs. Although this is a less serious condition than false hard preemption, and despite the mitigation procedures described in Section 6.1, network operators should be aware of the risk to their network in the case that the soft preemption processes are subverted, and should apply the relevant MPLS control plane security techniques to protect against attacks.

11. Acknowledgements

The authors would like to thank Carol Iturralde, Dave Cooper, Loa Andersson, Arthi Ayyangar, Ina Minei, George Swallow, Adrian Farrel, and Mustapha Aissaoui for their valuable comments.

12. Contributors

Denver Maddux
Limelight Networks
USA
EMail: denver@nitrous.net

Curtis Villamizar
AVICI
EMail:curtis@faster-light.net

Amir Birjandi
Juniper Networks
2251 Corporate Park Dr., Ste. 100
Herndon, VA 20171
USA
EMail: abirjandi@juniper.net

13. References

13.1. Normative References

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

[RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.

[RFC5710] Berger, L., Papadimitriou, D., and JP. Vasseur, "PathErr Message Triggered MPLS and GMPLS LSP Reroutes", RFC 5710, January 2010.

[RFC5711] Vasseur, JP., Swallow, G., and I. Minei, "Node Behavior upon Originating and Receiving Resource Reservation Protocol (RSVP) Path Error Messages", RFC 5711, January 2010.

13.2. Informative References

[RFC2961] Berger, L., Gan, D., Swallow, G., Pan, P., Tommasi, F., and S. Molendini, "RSVP Refresh Overhead Reduction Extensions", RFC 2961, April 2001.

[RFC3270] Le Faucheur, F., Wu, L., Davie, B., Davari, S., Vaananen, P., Krishnan, R., Cheval, P., and J. Heinanen, "Multi-Protocol Label Switching (MPLS) Support of Differentiated Services", RFC 3270, May 2002.

[RFC4829] de Oliveira, J., Vasseur, JP., Chen, L., and C. Scoglio, "Label Switched Path (LSP) Preemption Policies for MPLS Traffic Engineering", RFC 4829, April 2007.

[SEC_FMWK] Fang, L., Ed., "Security Framework for MPLS and GMPLS Networks", Work in Progress, October 2009.

Authors' Addresses

Matthew R. Meyer (editor)
British Telecom

EMail: matthew.meyer@bt.com

JP Vasseur (editor)
Cisco Systems, Inc.
11, Rue Camille Desmoulins
Issy Les Moulineaux, 92782
France

EMail: jpv@cisco.com