Networking Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 30, 2010

Matthew. Meyer, Ed.
British Telecom
JP. Vasseur, Ed.
Cisco Systems, Inc
July 29, 2009

MPLS Traffic Engineering Soft Preemption
draft-ietf-mpls-soft-preemption-18.txt

Status of this Memo

Copyright Notice

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/ eliminating traffic disruption of preempted Traffic Engineering Label Switched Paths (TE LSPs). Initially MPLS RSVP-TE was defined supporting only immediate TE LSP displacement upon preemption. The utilization of a reroute request notification helps more gracefully mitigate the re-route 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 re-routed. For this reason, the feature is primarily but not exclusively interesting in MPLS enabled IP networks with Differentiated Services and Traffic Engineering capabilities.

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].

Table of Contents

1. Introduction

In an 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 ingress (Label Edge Router) LER automated (Traffic Engineering (TE) reservation adjustments are less palatable when preemption is intrusive and high network stability levels are 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 re-routed in a non disruptive fashion (make-before-break) by the ingress LER. During the period that the tunnel is being re-routed, 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.

3. Motivations

Initially Multiprotocol Label Switching (MPLS) RSVP-TE [RFC3209] was defined supporting 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 [I-D.ietf-mpls-3209-patherr]. Note that such preemption is also referred to as a fatal error in [I-D.ietf-mpls-3209-patherr]. 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 Traffic Engineering Labeled Switched Path (TE LSP) and the preempting TE LSP(s).

Hard preemption may be a requirement to protect numerically lower preemption priority traffic in a non Diff-Serv enabled architecture, but in a Diff-Serv 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 Diff-Serv 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 so 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

[I-D.ietf-mpls-gmpls-lsp-reroute] specifies defines a new reroute-specific error code that allows a mid-point to report a TE LSP reroute request (Error-code=34 - Reroute). This document specifies a new error sub-code value for the case of Soft Preemption (to be confirmed by IANA upon publication of this document).

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

Upon (soft) preemption, the preemting node MUST issue a PathErr message with the error code=34 ("Reroute") and a value=1 ("Reroute request soft preemption"), to be confirmed by IANA.

5. Mode of Operation

Let's consider the following example:

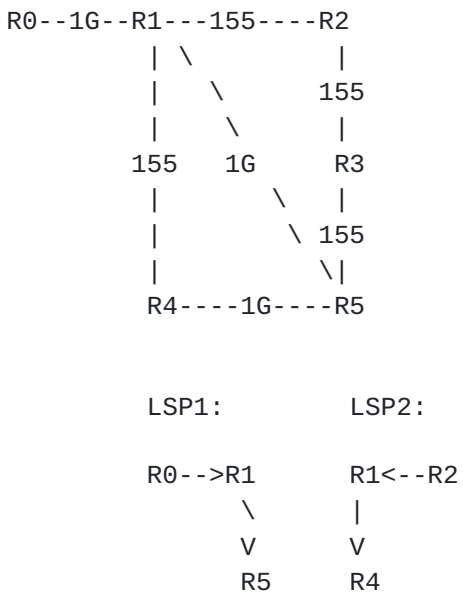

Figure 1: Example of Soft Preemption Operation

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

- Reservable BW on R0-R1, R1-R5 and R4-R5 is 1Gb/sec.
- Reservable BW on R1-R2, R1-R4, R2-R3, R3-R5 is 155 Mb/sec.
- Bandwidths and costs are identical in both directions.
- Each circuit has an IGP metric of 10 and IGP metric is used by CSPF.
- Two TE tunnels are defined: - LSP1: 155 Mb, setup/hold priority 0 tunnel, path R0-R1-R5. - LSP2: 155 Mb, 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 having 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 re-route 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 path tear for LSP2 upon preemption as with hard preemption (which would result in an immediate traffic disruption for LSP2), R1s 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 running calculating a new path for the new LSP. In the case that Diff-Serv [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 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 which requires to preempt a set of TE LSP(s) because not all TE LSPs can be accommodated on a specific interface, a node triggers a preemption action which 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 notifiying a fatal action as documented in [I-D.ietf-mpls-3209-patherr]. Upon receiving the RSVP PathErr message, the head-end LSR sends an RSVP Path Tear message, which 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 error code "Reroute" and an error value "Reroute request soft preemption" SHOULD be sent in reliable mode.

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

Selection of the preempted TE LSP at a preempting mid-point: 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), 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 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 reducing the temporarily elevated under-provisioning ratio on the links where soft preemption occurs and 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 soft preempted state. In the event a local preemption is needed, the relevant priority level LSPs from the cache 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' ('Soft Preemption desired' flag of the corresponding SESSION-ATTRIBUTE object set). Note that such decision MAY also be taken 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 Diff-Serv [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 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 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 operators Diff-Serv (if Diff-Serv 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 total local router (ingress, egress, midpoint)
- List current LSPs and bandwidth in PPend status
- List current sum bandwidth and session count in PPend status per observed ERO hops (ingress, egress views only).
- Cumulative PPend events per observed ERO hops.

9. IANA Considerations

IANA will not need to create a new registry.

9.1. New Session Attribute Object Flag

A new flag of the Session Attribute object is defined (to be confirmed by IANA)

Soft Preemption Desired bit

Bit Flag	Name Flag	Reference
0x40	Soft Preemption Desired	This document

9.2. New error sub-code value

[I-D.ietf-mpls-gmpls-lsp-reroute] defines a new reroute-specific error code that allows a mid-point to report a TE LSP reroute request. This document specifies a new error sub-code value for the case of Soft Preemption (to be confirmed by IANA upon publication of this document).

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.

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. Authors' Addresses

The content of this document was contributed by the editors and the co-authors listed below:

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

Curtis Villamizar
AVICI
curtis@faster-light.net

Amir Birjandi
Juniper Networks
2251 corporate park dr ste
herndon, VA 20171
USA
abirjandi@juniper.net

13. References

13.1. Normative References

[I-D.ietf-mpls-3209-patherr]
Vasseur, J., Swallow, G., and I. Minei, "Node behavior upon originating and receiving Resource ReserVation Protocol (RSVP) Path Error message", draft-ietf-mpls-3209-patherr-04 (work in progress), February 2009.

[I-D.ietf-mpls-gmpls-lsp-reroute]
Berger, L., Papadimitriou, D., and J. Vasseur, "PathErr Message Triggered MPLS and GMPLS LSP Reroute", draft-ietf-mpls-gmpls-lsp-reroute-04 (work in progress), January 2009.

[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.

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.

Authors' Addresses

Matthew R. Meyer (editor)
British Telecom
matthew.meyer@bt.com

Email:

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

Email: jpv@cisco.com