Matthew R. Meyer
Global Crossing
Denver Maddux
Global Crossing
Jean-Philippe Vasseur
Cisco Systems, Inc.

IETF Internet Draft
Expires: August, 2003
February, 2003

<draft-meyer-mpls-soft-preemption-00.txt>

MPLS Traffic Engineering Soft preemption

Status of this Memo

Abstract

This draft documents MPLS TE Soft Preemption, a suite of protocol modifications extending the current concept of preemption with the goal of reducing/eliminating traffic disruption of preempted TE LSPs. Under present RSVP-TE signaling methods, LSPs are immediately displaced upon preemption. The introduction of a new preemption pending flag helps more gracefully mitigate the re-route process of displaced LSPs. For the brief period soft preemption is activated, reservations (though not necessarily traffic levels) are in effect overbooked until the LSP can be re-routed. For this reason, the feature is primarily interesting in packet oriented MPLS networks with Diffserv and TE capabilities.

1. Terminology

LSR - Label Switch Router

HE LSR - Head-End Label Switch Router

LSP - An MPLS Label Switched Path

TE LSP - Traffic Engineering Label Switched Path

Local Repair - Techniques used to repair TE LSP tunnels quickly when a node or link along the TE LSPs path fails.

Preemption Pending flag - This flag is set on an IPv4 or Ipv6 RSVP Resv RRO sub-object to signal to the TE LSP head-end LSR that the TE LSP is about to be preempted and must be re-signaled (in a non disruptive fashion, with make before break) along another path.

PLR - Point of Local Repair. The head-end of a backup tunnel or a detour LSP.

PHB - Per Hop Behavior

PHP - Penultimate Hop Popping

CSPF - Constraint-based Shortest Path First.

2. Motivations

Present MPLS RSVP-TE implementations only support a method of TE LSP preemption which immediately tears down TE LSPs, disregarding the preempted in-transit traffic, in an effort to make way for a higher priority TE LSP if not enough bandwidth is available on the link to accommodate the newly signaled high priority TE LSPs. This process nearly guarantees preempted traffic will be discarded, if only briefly, until the RSVP Path Error message reaches and is processed by the head-end (HE) and a new forwarding path can be established. In cases of actual resource contention this might be helpful, however preemption is triggered by mere reservation contention and reservations may not be entirely accurate up to the moment. The result is that the traffic is often needlessly being discarded.

The current intrusive or 'hard' preemption may be a requirement to protect traffic in a network without Diffserv, 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.

3. Introduction

In an MPLS RSVP-TE and Diffserv enabled network there are currently no defined mechanisms to allow preempted TE LSPs to be handled in a make- before-break fashion: the currently defined preemption scheme proposes a very intrusive method that provokes traffic disruption for potentially a large amount of TE LSPs. Typically, this makes TE LSP dynamic resizing mechanisms less palatable when high network stability is sought. This draft proposes the use of additional signaling and accounting mechanisms to alert the HE LSR of the preemption that is pending and allow for temporary overbooking while the tunnel is re- routed in a non disruptive fashion (make-before-break) by the HE LSR. During the period that the tunnel is being re-routed, link capacity is effectively overbooked on links where soft preemption has occurred.

4. RSVP extensions

4.1. SESSION-ATTRIBUTES Flags

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

Soft preempted desired: 0x40

4.2. RRO IPv4/IPv6 Sub-Object Flags

To report that a soft preemption is pending for an LSP, a new flag is needed for the RSVP Resv RRO object message defined in RFC3209. The RRO is augmented with a preemption pending (PPend) flag. Any LSR compliant with this draft must support the RRO object, as defined in RFC 3209.

RRO IPv4 and IPv6 sub-object address

These two sub-objects currently have the following flags defined in RFC **3209 and [FAST-REROUTE]:**

Local protection available: 0x01
: Indicates that the link downstream of this node is protected via a local repair mechanism, which can be either one-to-one or facility backup.

Local protection in use: 0x02
: Indicates that a local repair mechanism is in use to maintain this tunnel (usually in the face of an outage of the link it was previously routed over, or an outage of the neighboring node).

Bandwidth protection: 0x04
: The PLR will set this when the protected LSP has a backup path which is guaranteed to provide the desired bandwidth specified in the FAST_REROUTE object or the bandwidth of the protected LSP, if no FAST_REROUTE object was included. The PLR may set this whenever the desired bandwidth is guaranteed; the PLR MUST set this flag when the desired bandwidth is guaranteed and the "bandwidth protection desired" flag was set in the SESSION_ATTRIBUTE object. If the requested bandwidth is not

guaranteed, the PLR MUST NOT set this flag.

Node protection: 0x08
The PLR will set this when the protected LSP has a backup path which provides protection against a failure of the next LSR along the protected LSP. The PLR may set this whenever node protection is provided by the protected LSP's backup path; the PLR MUST set this flag when the node protection is provided and the "node protection desired" flag was set in the SESSION_ATTRIBUTE object. If node protection is not provided, the PLR MUST NOT set this flag. Thus, if a PLR could only setup a link-protection backup path, the "Local protection available" bit will be set but the "Node protection" bit will be cleared.

A new flag is added:

Preemption pending: 0x10
The preempting node sets this flag if a pending preemption is in progress for the TE LSP. This indicates to the HE of this LSP that it must be re-routed as soon as possible using a make before break.

5. Mode of operation

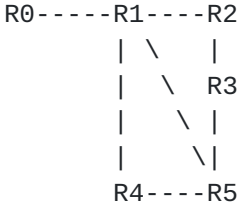

Fig 1.

In the network depicted above in figure 1, let suppose:

-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 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' 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 a Path Error message to all the Head-end having a TE LSP traversing the R1-R5 failed link (R0 is the example above). Upon receiving the link failure notification (RSVP Path Error and/or IGP LSA/LSP update), 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 penultimate hop to the HE LSR - R4 to R1 in this example).
LSP2 is soft preempted at R1 as it has a higher preemption (lower priority) and both bandwidth reservations cannot be satisfied on the R1-R4 link.

Instead of sending a path tear for LSP2 upon preemption as with the current preemption (which would result in an immediate traffic disruption for LSP2), R1s local BW accounting for LSP2 is zeroed and a preemption pending flagged RRO Resv for LSP2 is issued upstream toward the HE LSR, R2. If more than one preempted candidate TE LSP has the same HE, these soft preemption Resv messages MAY be bundled together (see RFC2961)

The preempting node MUST immediately send a Resv message with the 'Preemption pending' RRO flag set for each soft preempted TE LSP.
The node MAY choose for soft preemption to impact the pacing of IGP update and re-flood of the TE-TLV. An implementation MAY provide user configurable exponential back off timers if preemption paced IGP triggering is used.

Should a refresh event for LSP2 arrive before LSP2 is re-routed, soft preempting nodes such as R1 MUST continue to refresh the LSP however the RRO Soft Pending flag MUST be set. This assures that if the initial soft preemption Resv message is somehow dropped, the HE will still receive notification. Resv messages with the RRO 'Preemption pending' flag set should be sent in reliable mode (RFC 2961).

Upon reception of the Resv with the 'Preemption pending' flag set, the HE (of LSP2 in this case) MAY update the working copy of the TE-DB before running CSPF for the new LSP. The preemption pending implies exhausted bandwidth in the affected priority level and greater for the indicated node interface. An implementation MAY choose to reduce or zero available BW for that range until more accurate information is

available (i.e. a new IGP TE update is received).

In the case that reservation availability is restored at the point of preemption (R1) the point of preemption MAY issue a Resv message with the 'Preemption pending' flag unset to signal restoral to the HE. This implies that a HE might have delayed or been unsuccessful in re-signaling.

After the HE has successfully established a new LSP, the old path MUST be torn down.

As a result of this '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 (if deployed) according to the operators Diffserv and queuing scheme.

6. Selection of the preempted TE LSP at a preempting mid-point

When a lower preemption (higher priority) TE LSP is signaled that requires the preemption of a set of lower priority the TE LSPs in order to accommodate the newly signaled high priority TE LSP, the node has to make a decision on the set of TE LSP, candidate for preemption. This decision is a local decision and various algorithms can be used, depending on the objective (minimize the number of preempted LSPs, ...).

As already mentioned, a temporary link overbooking results from the soft preemption, until all the soft preempted TE LSPs are effectively re-routed by their respective HE LSR. In order to reduce this overbooking period of time, the preempting LSR can limit the number of soft preempted TE LSP to the TE LSP that have explicitly requested soft

preemption via signaling, setting their 'Soft Preemption desired' bit in the SESSION-ATTRIBUTE of their RSVP Path messages. This way, the preemption could apply the current 'hard' preemption scheme to the TE LSPs that have not explicitly requested soft preemption, sending a Path Error message to their HE LSR and immediately removing the corresponding local states. This would help reducing the overbooking ratio on the related links. This TE LSP capability (Soft preemption desired) could be reserved to the TE LSP, for which a traffic disruption upon preemption is not unacceptable.

Optionally, a midpoint LSR upstream from a soft preempting node MAY choose to cache the soft preempted LSPs downstream 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 preemption

selection process for the given priority.

7. Interoperability

Backward compatibility is assured since any HE LSR not compliant with this draft that receives a Resv message with the RRO 'Preemption Pending' bit set will simply ignore the flag and treat the Resv message as a regular Resv refresh message. As a consequence, the soft preempted TE LSP will not be re-routed with make before break by the HE LSR. To guard against a situation where bandwidth overbooking will last forever, a local timer (soft preemption expiration timer) MUST be started on the preemption node, upon soft preemption. When this timer expires, the soft preempted TE LSP will be torn down and the preempting node will send a Path Error. This timer should be configurable. The current 'hard' preemption scheme can be emulated with a soft preemption expiration timer set to zero.

8. Management

Both the point of preemption and the HE LSR should provide some form of accounting internally and to the user with regard to what TE LSPs and how much capacity is over-booked due to soft preemption.

9. Security Considerations

The practice described in this draft does not raise specific security issues beyond those of existing TE.

10. Acknowledgment

The authors would like to thank Carol Iturralde, Dave Cooper for their valuable comments.

11. Intellectual Property

The contributor represents that he has disclosed the existence of any proprietary or intellectual property rights in the contribution that are reasonably and personally known to the contributor. The contributor does not represent that he personally knows of all potentially pertinent proprietary and intellectual property rights owned or claimed by the organization he represents (if any) or third

parties.

References

[TE-REQ] Awduche et al, Requirements for Traffic Engineering over MPLS, RFC2702, September 1999.

[OSPF-TE] Katz, Yeung, Traffic Engineering Extensions to OSPF, draft-katz-yeung-ospf-traffic-09.txt, October 2002.

[ISIS-TE] Smit, Li, IS-IS extensions for Traffic Engineering, draft-ietf-isis-traffic-04.txt, December 2002.

[RSVP-TE] Awduche et al, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC3209, December 2001.

[DS-TE] Le Faucheur et al, "Requirements for support of Diff-Serv-aware MPLS Traffic Engineering", draft-ietf-tewg-diff-te-reqts-06.txt, September 2002.

[DS-TE-PROT] Le Faucheur et al, "Protocol extensions for support of Diff-Serv-aware MPLS Traffic Engineering", draft-ietf-tewg-diff-te-proto-02.txt, October 2002

[FAST-REROUTE] Pan, P. et al., "Fast Reroute Techniques in RSVP-TE", Internet Draft, draft-ietf-mpls-rsvp-lsp-fastreroute-01.txt , May, 2003

[REFRESH-REDUCTION] Berger et al, "RSVP Refresh Overhead Reduction Extensions", RFC 2961, April 2001.

Matthew R. Meyer
Global Crossing
14605 S. 50th
Phoenix, AZ 85044
USA
email: mrm@gblx.net

Denver Maddux
Global Crossing
14605 S. 50th
Phoenix, AZ 85044

USA
email: denver@gblx.net

Jean Philippe Vasseur
Cisco Systems, Inc.
300 Apollo Drive
Chelmsford, MA 01824
USA
Email: jpv@cisco.com