Network Working Group
Internet Draft
Category: Standard Track
Expires: August 2008

J.L. Le Roux (Ed.)
France Telecom

R. Aggarwal
Juniper Networks

J.P. Vasseur
Cisco Systems, Inc.

M. Vigoureux
Alcatel-Lucent

March 2008

P2MP MPLS-TE Fast Reroute with P2MP Bypass Tunnels

draft-ietf-mpls-p2mp-te-bypass-02.txt

Status of this Memo

Abstract

This document defines procedures for fast reroute protection of Point-To-MultiPoint (P2MP) Traffic Engineering Label Switched Paths (TE-LSP) in MultiProtocol Label Switching (MPLS) networks, based upon Point-To-MultiPoint Bypass Tunnels. The motivation for using P2MP Bypass Tunnels is to avoid potentially expensive data duplication along the backup path that could occur if Point-To-Point Bypass Tunnels were used, i.e., to optimize the bandwidth usage, during fast reroute protection of a link or a node. During link or node failure the traffic carried onto a protected P2MP TE-LSP is tunnelled within one or several P2MP Bypass Tunnels towards a set of Merge Points. To avoid data duplication, backup labels (i.e., inner labels) are assigned by the Point of Local Repair (PLR) according to the RSVP-TE upstream label assignment procedure.

Conventions used in this document

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.

Table of Contents

1. Terminology

This document uses terminologies defined in [RFC3031], [RFC3209], [RFC4090] and [RFC4461]. It defines the following new terms:

P2MP Bypass Tunnel: Point-to-Multipoint Bypass Tunnel. A P2MP TE-LSP that is used to protect a set of P2MP TE-LSPs traversing a common facility (link or node).

P2MP Facility Backup: A local repair method in which a P2MP Bypass Tunnel is used to protect one or more P2MP TE-LSPs that traverse the Point of Local Repair (P2MP Bypass Ingress) and the resource being protected.

Backup P2MP LSP: The LSP that is used to backup up one of the many protected P2MP LSPs in P2MP Facility Backup.

Backup label: Label of a backup P2MP LSP.

Backup S2L sub-LSP: A S2L sub-LSP of a backup P2MP LSP.

PLR: Point of Local Repair: Head-end LSR of the bypass tunnel

MP: Merge Point: LSR where a primary LSP and its backup LSP merge.

.

2. Introduction

[RFC4090] defines Fast ReRoute (FRR) extensions to RSVP-TE [RFC3209] for local protection of Point-To-Point (P2P) Traffic Engineered Label Switched Paths (TE LSP) in MultiProtocol Label Switching (MPLS) networks. Two techniques are defined: the one-to-one backup method, which creates a detour LSP for each protected LSP at each point of local repair (PLR), and the facility backup method, which creates a bypass tunnel that can be used to protect a set of TE LSPs by taking advantage of MPLS label stacking.

[RFC4875] defines extensions to RSVP-TE for setting up Point-To-Multipoint (P2MP) TE LSPs. It specifies extensions to one-to-one and facility backup Fast Reroute procedures defined in [RFC4090] so as to support fast reroute protection of P2MP TE LSPs.
The facility backup solution defined in [RFC4875] only relies on P2P Bypass Tunnels for link and node protection. This faces the following limitations:

- The protection of a downstream link of a P2MP TE LSP on a branch LSR may require a P2P Bypass LSP that uses another downstream link of the P2MP LSP, and this leads to twice the traffic on that link during failure, which is inefficient.

Finding a bypass path that avoids all downstream links on the P2MP LSP would be a solution but this is often not achievable in lowly meshed topologies.

- The protection of a P2MP TE LSP against node failures requires, when the protected node is a Branch LSR, a set of P2P Next-Next-Hop (NNHOP) Bypass Tunnels toward all LSRs downstream to the protected node. During failure the PLR has to replicate traffic on each P2P NNHOP Bypass Tunnel. If there are K next-next-hops, this may lead to K times the traffic on some links, which is not acceptable.

- Similarly the protection of a P2MP TE LSP against the failure of a LAN interface that connects a branch LSR and a set of K downstream LSRs requires one P2P Bypass Tunnel per downstream LSR, which may lead to K times the traffic on some links during failure.

To overcome these limitations it is highly desirable to define extensions to the fast reroute facility backup solution, so as to support P2MP Bypass Tunnels. This retains the scalability advantages of MPLS label stacking and avoids sending multiple copies of a packet on some links during failure.

This draft specifies extensions to the Fast ReRoute (FRR) procedures defined in [RFC4090] and [RFC4875] to support local repair of P2MP TE LSP with P2MP Bypass Tunnels.

Procedures defined in [RFC3209], [RFC4090] and [RFC4875] MUST be followed unless specified below.

3. Solution overview

The P2MP Facility Backup method defined in this document relies on the use of P2MP Bypass Tunnels. Similarly to the P2P case, the same P2MP Bypass Tunnel can be used to protect a set of P2MP TE LSPs, by taking advantage of MPLS label stacking.

A P2MP Bypass Tunnel can be used to protect a P2MP TE-LSP against downstream link or node failures.

There are various options for the protection of a downstream link or node of a P2MP TE-LSP:

- Option 1: Rely on a single P2MP Bypass Tunnel whose set of leaf LSRs exactly matches the set of Merge Points (MP). Merge points are transit or egress LSRs on the protected P2MP LSP downstream to the PLR or downstream to the protected element (link or node).
- Option 2: Rely on a single P2MP Bypass Tunnel whose set of leaf LSRs is a superset of the set of MPs. Leaf LSRs which

are not MP have to drop the traffic.

- Option 3: Rely on a combination of P2MP Bypass LSPs whose leaf LSRs include a subset of the set of MPs but their

combination encompass all MPs (and this combination may be a superset of the set of MPs).

These three options differ in terms of bandwidth optimization and control plane state minimization. Option 1 increases the number of states compared to option 2, and, in some cases, to option 3(it implies more P2MP Bypass LSPs), but is less expensive in terms of bandwidth (traffic only sent to MPs). With point-to-multipoint hierarchy there is always a tension between minimizing the amount of control plane state and minimizing bandwidth consumption. Choosing one of these options is a decision local to the PLR. The choice depends on the desired trade-off between control plane and data plane optimization, and the operational complexity associated with the different options.

When the P2MP Facility Backup method is used, during failure the PLR MUST send data for each protected P2MP LSP into the set of one or more P2MP Bypass Tunnels. Label stacking is used: the inner label is the backup label for the backup P2MP LSP, that will be used on the MP to forward traffic to the corresponding protected P2MP LSP, and the outer label is the P2MP Bypass Tunnel label.

To avoid data replication at the PLR and to avoid traffic mis-routing in Merge Points, the same backup label MUST be used for all S2L sub-LSPs of a given backup P2MP LSP, tunneled within the same P2MP Bypass Tunnel. This backup label will indicate to the Merge Points that packets received with that label should be switched along the protected P2MP LSP.

For that purpose upstream label assignment procedures defined in [MPLS-UPSTREAM] and RSVP-TE extensions for upstream label assignment defined in [RSVP-UP] MUST be used. To signal a backup P2MP LSP, the same backup label, is distributed by the PLR to all MPs belonging to a same P2MP Bypass Tunnel, in the context of this P2MP Bypass Tunnel. This requires the backup P2MP LSP to be signaled prior to the failure

At the MP, backup S2L sub-LSPs (i.e., S2L sub-LSPs of the Backup P2MP LSP) are merged with protected S2L sub-LSPs. A MP (i.e., the bypass tunnel leaf LSRs), maintains a context specific Incoming Label Map (ILM) for the P2MP Bypass Tunnel. This can be implemented by maintaining a different context specific ILM for each LSR that is the root of a P2MP Bypass Tunnel (per-neighbor), or by maintaining a different context specific ILM for each P2MP Bypass Tunnel (per-tunnel). The context of an inner label (i.e., a backup label) is determined by the underlying P2MP Bypass Tunnel on which it is received (as described in section 5). This requires deactivating Penultimate Hop Popping (PHP) on the P2MP Bypass Tunnel. A backup

label, in a given P2MP Bypass Tunnel specific ILM, is mapped to the outgoing interface(s) and label(s) of the corresponding protected P2MP LSP.

The way in which the MP determines whether the PLR assigns upstream-assigned labels from a per-tunnel, or per-platform pool (a.k.a label space) is outside the scope of this document.

4. PLR procedures

4.1. Before failure

4.1.1. P2MP Bypass Tunnel(s) Selection

To protect a P2MP TE LSP against a downstream link or node failure, using P2MP Facility Backup, a PLR has to select a set of one or more P2MP Bypass Tunnel(s), denoted {B1.Bm}, as follows:

- The P2MP Bypass Tunnel(s) MUST NOT traverse the protected link/node/SRLG.

- The set of leaf LSRs of P2MP Bypass Tunnels {B1.Bm}, denoted {LSR1.LSRn} must include a set of Merge Points (MP), on the protected P2MP LSP. These Merge Points are transit or egress LSRs on the protected P2MP LSP downstream to the PLR or downstream to the protected element (link or node). We will denote this set of Merge Points as {MP1.MPq}. Note that the case where some MPs are LSRs downstream to the PLR but not downstream to the failed element allows avoiding sending twice the traffic on downstream links during failure.

-The set of Merge Points {MP1.MPq} is such that in the event of failure of the protected link or node, traffic received on the protected P2MP LSP by the PLR, can be delivered to ALL the leaf LSRs of the protected P2MP LSP downstream to the PLR, if it is tunnelled to {MP1.MPq} over the set of one or more P2MP Bypass Tunnel(s) {B1.Bm}.
Note: This condition, which provides full protection, does not apply when partial protection mode is used (as described in Section 7).

The PLR will assign backup labels to Merge Points {MP1.MPq} for the backup P2MP LSP. The same label will be assigned to all Merge Points belonging to the same P2MP Bypass Tunnel, as defined in [MPLS-UPSTREAM] and [RSVP-UP].

A MP may actually be the leaf LSR of multiple P2MP Bypass Tunnels used by the PLR to protect a given LSP (using Option 3 as described in Section 3), but will be associated to only one P2MP Bypass Tunnel (at a given time for a given protected P2MP LSP). That is, a PLR will signal the P2MP Backup LSP to that MP, for a single P2MP Bypass Tunnel context. When a MP is a leaf LSR of multiple P2MP Bypass

Tunnels, and if the PLR assigns the same backup label (i.e., inner upstream-assigned label) for the backup P2MP LSP on several different

P2MP Backup Tunnels, the MP MUST maintain a per-tunnel ILM (and not a per-neighbor ILM) to perform contextual lookup of the backup label.

{LSR1.LSRn} may be a superset of {MP1.MPq}, that is some leaf LSRs of a given P2MP Bypass Tunnel, noted {LSRx.LSRy}, may not belong to {MP1.MPq}. The PLR will not distribute the backup label for the backup P2MP LSP to these LSRs {LSRx.LSRy}.
However due to the nature of the P2MP Bypass Tunnel, during failure, packets with the backup label will also be delivered onto the P2MP Bypass Tunnel to {LSRx.LSRy}. {LSRx.LSRy} MUST discard these packets based on the absence of an entry for this label in the context specific ILM referred to that P2MP Bypass Tunnel. This requires that {LSRx.LSRy} create a context specific ILM, per-tunnel or per-neighbor for that P2MP Bypass Tunnel label.

PHP MUST be deactivated on the P2MP Bypass Tunnel, in order to allow MPs to determine the context for the backup labels assigned by the PLR. Hence the P2MP Bypass Tunnel will be signaled with the "non PHP behavior desired" bit set in the Attribute Flags TLV as specified in [NO-PHP].

Note that P2MP Bypass Tunnels may be signaled in advance, prior to the establishment of any protected P2MP LSP, either automatically or via configuration, or may be dynamically setup upon signaling of a protected P2MP LSP. Such procedures rely on local implementation issues and are beyond the scope of this document.

4.1.2. P2MP Backup LSP Signaling over a P2MP Bypass Tunnel

The same backup label (i.e., the inner label) MUST be used for all backup S2L sub-LSPs which are tunneled within the same P2MP Bypass Tunnel, so as to avoid traffic replication at the PLR, and to avoid traffic misrouting in the MPs. This label MUST be assigned by the PLR using upstream label assignment procedures as specified in [MPLS-UPSTREAM] and [RSVP-UP].

Backup P2MP LSPs MUST be signaled prior to the failure. To signal the backup P2MP LSP, the PLR will send one or more Path messages, referred to as a backup LSP's Path message, to each MP, as specified in [RFC4875]. A backup LSP's Path message to a given MP comprises one or more backup S2L sub-LSPs that transit through this MP.
A backup Path message MUST be sent to the MP using directed signaling, i.e., it is addressed to the MP, without Router Alert option.

As specified in [RFC4875] it is RECOMMENDED that the PLR use the sender template specific method to identify a backup LSP's Path

message, that is, the PLR will set the source address in the sender template to a local PLR address.

The backup label MUST be assigned by the PLR, in the context of the underlying P2MP Bypass Tunnel, following upstream label assignment [MPLS-UPSTREAM] and P2MP RSVP-TE context identification procedures defined in [RSVP-UP]. Hence, a backup LSP's Path message sent to a given MP MUST include an Upstream Assigned Label object carrying the value of the backup label. It MUST also include an RSVP-TE P2MP LSP TLV within an IF_ID_RSVP_HOP object, that carries the session object of the underlying P2MP Bypass Tunnel. This allows the MP to identify the label space of the backup label assigned by the PLR. The same backup label MUST be sent to all MPs belonging to a given P2MP Bypass Tunnel.

Note that the PLR MUST continue to refresh Path messages for the protected P2MP TE LSP along the nominal route.

The processing of backup S2L sub-LSP SEROs/SRROs MUST follow backup LSP ERO/RRO processing described in [RFC4090].

4.2. During failure

When the PLR detects a link or/and node failure condition, it has to reroute a protected P2MP LSP onto a set of one or more P2MP Bypass Tunnels protecting the failed element, using as inner label(s) the backup label(s) assigned for this P2MP LSP.

The PLR needs to localize the failed elements in order to activate the P2MP Bypass Tunnel(s) protecting this element. Mechanisms through which this location is retrieved are out of the scope of this document.

Note that when some MPs are LSRs downstream to the PLR but not downstream to the failed element, the PLR MUST stop sending traffic directly within the protected P2MP TE LSP towards these MPs. This allows avoiding sending twice the traffic on downstream links during failure.

The processing of backup S2L sub-LSP SEROs/SRROs MUST follow backup tunnel ERO/RRO processing described in [RFC4090].

4.3. After failure

Reversion procedures for restoring the P2MP TE LSP to a full working path after failure MUST follow procedures defined in section 6.5.2 of [RFC4090], that is there are two basic strategies for restoring the P2MP TE-LSP: The global revertive mode and the local revertive mode.

5. MP Procedures

A MP receives one or more Path messages for the protected P2MP TE LSP and one or more Path messages for the backup P2MP LSP.

Note that, as specified in [RFC4090], the reception of a backup LSP's Path message does not indicate that a failure has occurred or that the incoming protected LSP will no longer be used.

A S2L sub-LSP is received within a Path message for the protected P2MP LSP and within a Path message for the backup P2MP LSP. These two Path messages are distinguished thanks to the sender-template specific method. As specified in [RFC4090], each of these Path messages will have a different sender address. The protected LSP can be recognized because it will include the FAST_REROUTE object or have the "local protection desired" flag set in the SESSION_ATTRIBUTE object, or both.

A MP MUST maintain one context specific ILM table per PLR or per P2MP Bypass Tunnel for which it is a leaf LSR.

A MP MUST install the upstream assigned label received in a backup LSP's Path message (i.e., the backup label), within an ILM either specific to the underlying P2MP Bypass Tunnel or specific to the PLR, which is the ingress node of the underlying P2MP Bypass Tunnel. The underlying P2MP bypass tunnel is identified by its session object, carried within the IF_ID_RSVP_HOP object of the backup LSP's Path message. An upstream assigned label for a backup P2MP LSP MUST be mapped to the outgoing interface(s) and label(s) of the corresponding protected P2MP LSP.

As specified in [RSVP-UP], the Resv message sent by a MP to the PLR, does not carry any Label Object.

The processing of backup S2L sub-LSP SEROs/SRROs MUST follow backup tunnel ERO/RRO processing described in [RFC4090].

6. Combination of P2P and P2MP Bypass tunnels

The P2MP Facility Backup method defined in this document and the P2P Facility Backup method defined in [RFC4090] may be used in conjunction. That is, a P2MP LSP may be protected on a PLR using a combination of a set of P2P and P2MP Bypass Tunnels.

In this case S2L sub-LSPs protected by a P2P Bypass Tunnel are signalled using procedures defined in [RFC4875] with the backup label downstream assigned by the MP, while S2L sub-LSPs protected by a P2MP Bypass Tunnel are signaled using procedures defined in this

document, with the backup label upstream assigned by the PLR.

This allows for backward compatibility with LSRs that do not support

upstream assigned labels. A P2P Bypass Tunnel MUST be used to tunnel traffic to a MP that do not support upstream assigned labels.

7. Partial Protection

In some cases, in particular in some networks where bandwidth is a scarce resource, the PLR may not be able to find a set of P2P and/or P2MP Bypass Tunnels that cover all merge points. That is, only a subset of merge points are covered, and upon failure, traffic will not be delivered to all leaf LSRs downstream to the failed element. Such a situation is referred to as partial protection as opposed to full protection where all merge points are covered.

Hence, on a given PLR, a P2MP LSP may be fully, partially or non protected. The default behavior on a PLR is that when a P2MP LSP cannot be fully protected it is not protected at all. But the ingress LSR may request 'P2MP Partial Protection' such that if a P2MP LSP cannot be fully protected it is partially protected.

In order to request and record "P2MP Partial Protection" behavior, this document defines a new bit in the Attributes Flags TLV of the LSP_ATTRIBUTES object and the RRO Attributes sub-object defined in [RFC4420]:

Bit Number 8 (TBD): P2MP Partial Protection

This bit SHOULD be set by the Ingress node in the Attributes Flags TLV of the LSP_ATTRIBUTES object in the Path message for the LSP for which P2MP Partial Protection is desired when full protection cannot be provided. This bit MUST NOT be modified by any other nodes along the LSP.

If a PLR supports the P2MP Partial Protection bit, and the bit is set in a Path message, then it SHOULD setup a partial protection when a full protection cannot be provided. In all other cases, the default procedure applies, that is, the PLR MUST not setup any protection if a full protection cannot be provided.

A PLR that recognizes the partial protection flag, but does not support it, MUST ignore the request and apply default procedure.

When this bit is set in an RRO Attributes Subobject, this means that the P2MP LSP is only partially protected on the node. In this case the local protection available bit in the RRO flags MUST also be set. A PLR that supports this flag MUST set it in the RRO Attributes subobject, if it has setup a partial protection for a P2MP LSP.

The way in which a PLR chooses which set of MPs to target, when it has to setup a partial protection, is out of the scope of this document.

8. Location of the PLR

The PLR may be directly upstream to the protected link or node or may also be two or more hops upstream.

In case the PLR is not directly upstream to the failure, rerouting within the Bypass Tunnel(s) may be triggered by the following events:

- -Failure of a BFD session between the PLR and the protected Element.
- -A PathErr message, that indicates the location of the failed Element.

9. Security Considerations

No new security issues are raised in this document.

10. IANA Considerations

10.1. LSP Attributes Flags

IANA has been asked to manage the space of flags in the Attributes Flags TLV carried in the LSP_ATTRIBUTES object [RFC4420].

This document defines a new flag as follows:

Bit Number:	8 (suggested value)
Meaning:	P2MP Partial Protection Desired
Used in Attributes Flags on Path:	Yes
Used in Attributes Flags on Resv:	No
Used in Attributes Flags on RRO:	Yes
Referenced Section of this Doc:	7

11. Acknowledgments

We would like to thank Kireeti Kompella, Venu Hemige, Laurent Ciavaglia, and Yannick Brehon, for the useful comments and discussions.

12. References

12.1. Normative references

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

[RFC3031] E. Rosen, A. Viswanathan, R. Callon, "MPLS Architecture",

RFC 3031.

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

[RFC4461] S. Yasukawa et al., "Signaling Requirements for Point-to-Multipoint Traffic-Engineered MPLS Label Switched Paths (LSPs)", RFC4461.

[RFC4090] Pan, Swallow, Atlas, et al., "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC4090.

[RFC4875] Aggarwal, Papadimitriou, Yasukawa et al. "Extensions to RSVP-TE for Point to Multipoint TE LSPs", RFC4875, May 2007.

[MPLS-UPSTREAM] Aggarwal, Rekhter, Rosen, "MPLS Upstream Label Assignment and Context Specific Label Space", draft-ietf-mpls-upstream-label, work in progress.

[RSVP-UP] Aggarwal, Le Roux, "MPLS Upstream Label Assignment for RSVP-TE", draft-ietf-mpls-rsvp-upstream, work in progress.

[RFC4420] Farrel, A., Ed., et al."Encoding of Attributes for Multiprotocol Label Switching (MPLS) Label Switched Path (LSP) Establishment Using Resource ReserVation Protocol-Traffic Engineering (RSVP-TE)", RFC 4420, February 2006.

12.2. Informational references

[NO-PHP] Ali, Swallow, Aggarwal, "Non PHP Behavior and out-of-band mapping for RSVP-TE LSPs", draft-ietf-mpls-rsvp-te-no-php-oob-mapping, work in progress

13. Authors' Addresses:

Jean-Louis Le Roux
France Telecom
2, avenue Pierre-Marzin
22307 Lannion Cedex
FRANCE
Email: jeanlouis.leroux@orange-ftgroup.com

Rahul Aggarwal
Juniper Networks
1194 North Mathilda Ave.
Sunnyvale, CA 94089
USA

Email: rahul@juniper.net

Jean-Philippe Vasseur
Cisco Systems, Inc.
1414 Massachusetts avenue
Boxborough , MA - 01719
USA
Email: jpv@cisco.com

M. Vigoureux
Alcatel-Lucent France
Route de Villejust
91620 Nozay
FRANCE
Email: martin.vigoureux@alcatel-lucent.fr

14. Intellectual Property Statement

The IETF takes no position regarding the validity or scope of any Intellectual Property Rights or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; nor does it represent that it has made any independent effort to identify any such rights. Information on the procedures with respect to rights in RFC documents can be found in BCP 78 and BCP 79.

Copies of IPR disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the IETF on-line IPR repository at http://www.ietf.org/ipr.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement this standard. Please address the information to the IETF at ietf-ipr@ietf.org.

Disclaimer of Validity

This document and the information contained herein are provided on an "AS IS" basis and THE CONTRIBUTOR, THE ORGANIZATION HE/SHE REPRESENTS OR IS SPONSORED BY (IF ANY), THE INTERNET SOCIETY, THE IETF TRUST AND THE INTERNET ENGINEERING TASK FORCE DISCLAIM ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.