CCAMP Working Group	L. Jin
Internet-Draft	ZTE Corporation
Intended status: Standards Track	F. Jounay
Expires: August 22, 2011	France Telecom
	M.B. Bhatia
	Alcatel-Lucent
	February 18, 2011

Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Hub and Spoke Multipoint Label Switched Paths (LSPs)
draft-jjb-ccamp-rsvp-te-hsmp-lsp-00

Abstract

In Multiprotocol Label Switching (MPLS) and Generalized MPLS (GMPLS) environment, the RSVP-TE based Point-to-Multipoint (P2MP) LSP allows traffic to transmit from root to leaf node, but there is no co-routed reverse path for traffic from leaf to root node. This draft introduces a Hub and Spoke Multipoint (HSMP) LSP, which allows traffic from both the root to the leaves through a P2MP LSP and also the leaves to the root along a co-routed reverse path.

Status of this Memo

Copyright Notice

Table of Contents

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 RFC2119 [RFC2119]. When used in lower case, these words convey their typical use in common language, and are not to be interpreted as described in RFC2119 [RFC2119].

1. Application

The proposed technique targets one-to-many applications that require reverse one-to-one traffic flow (thus many one-to-one in the reverse direction).
There are a few applications that could use such kind of Resource Reservation Protocol - Traffic Engineering (RSVP-TE) based Hub and Spoke Multipoint LSPs.

1.1. Time Synchronization

The delivery of time synchronization to end equipments, such as base stations, can be achieved using a time protocol as [IEEE] (also known as PTP). This protocol defines Transparent Clock (TC) function, which can be used in transport nodes to improve the accuracy of time synchronization. Two types of TCs exist in [IEEE]: End-to-end Transparent Clock (E2E TC) and Peer-to-peer Transparent Clock (P2P TC). P2P TCs assume that the link delays between the different nodes are calculated
Assuming that a chain of P2P TCs is used between a PTP master and a PTP slave, time synchronization can be delivered to the PTP slave by sending timestamps only in the direction master to slave (one way mode), via PTP Sync messages. This is possible because of the link delay calculation performed locally by each node, which enables it to calculate the propagation delay over the path. This scenario permits that the same PTP Sync messages would be sent by the PTP master to all the PTP slaves.
In this scenario (chain of P2P TCs), the PTP slave might have to send also messages (not carrying timestamps) back to the PTP master in some cases. For instance, PTP Signaling messages could be sent back to the PTP master. These PTP Signaling messages are not intended to be received by the other PTP slaves.
By using Point-to-Multipoint (P2MP) technology to transmit PTP Sync messages will greatly improve the bandwidth usage for above applications. This will also be useful for monitoring performance metrics for two-way delay and related metrics such as delay variation and loopback measurement. Current RSVP-TE based Point-to-Multipoint LSP mechanism only provides unidirectional path from the root to the leaf nodes, which cannot fulfill the above new requirement (i.e. need for a reverse path for the PTP Signaling messages).
This draft attempts to solve this problem. RSVP-TE based Hub and Spoke P2MP LSP described in this draft provides a co-routed reverse path from

the leaf to the root based on current unidirectional Point-to-Multipoint LSP.

1.2. P2MP Pseudowire based L2VPNs (VPMS and VPLS)

Point-to-Multipoint (P2MP) Pseudowires (PW) described in [I-D.ietf-pwe3-p2mp-pw] requires an additional reverse LSP to be set up from the leaf node (referred as egress PE) to root node (referred as ingress PE). Instead, if HSMP LSP is used to multiplex P2MP PW, the reverse path can also be multiplexed to HSMP upstream path to avoid setting up an independent reverse path. In that case, the operational cost will be reduced for maintaining only one HSMP LSP, instead of P2MP LSP and n (number of leaf nodes) P2P reverse LSPs

The VPMS defined in [I-D.ietf-l2vpn-vpms-frmwk-requirements] requires reverse path from the leaf to the root node. The P2MP PW multiplexed to HSMP LSP can provide VPMS with reverse path, without introducing independent reverse paths from each leaf to the root.

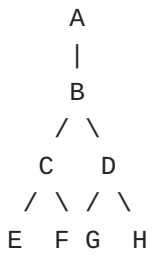
The P2MP PW multiplexed to HSMP LSP can also be used for VPLS [RFC4672], which will reduce the overall broadcast/multicast utilization for VPLS. In current VPLS implementations with a full mesh of P2P LSPs between PEs, broadcast, unknown and multicast (BUM) traffic is efficiently distributed over the physical links between Provider (P) and Provider Edge (PE) routers. [I-D.ietf-l2vpn-vpls-mcast] and [I-D.ietf-l2vpn-ldp-vpls-broadcast-exten] leverages this constraint by introducing the usage of P2MP PW and/or P2MP LSP. But a specific P2P PW over P2P LSP is still needed for unicast traffic between the PEs.

In the VPLS implementation scenario with P2MP PW multiplexed to HSMP LSPs, each PE signals a P2MP PW with itself as a root to all other PEs in the VPLS. Thereafter, all BUM traffic from this PE will use this P2MP PW. Unicast (learnt) traffic from a particular PE (e.g. PE1) to another PE (e.g. PE2) will be sent from leaf to root using the reverse path of P2MP PW where PE2 is the root.

This simplifies the VPLS implementation by reducing (a) link utilization for the BUM traffic and (b) the total number of LSPs maintained by each PE (i.e. instead of requiring a full mesh of LSPs, PEs now only require one HSMP LSP). It also helps in avoiding the unnecessary MAC learning that happens on the hub PE routers in case of H-VPLS.

2. Comparing Hub-Spoke MP LSP with P2MP and Unidirectional Reverse LSP

An HSMP LSP provides a Point-to-Multipoint reachability from the root node to the leaf nodes and a unicast reachability from all the leaf nodes back to the root node. An obvious question that comes up is that how is this better than setting up a P2MP LSP from a root node and Unidirectional reverse LSPs back from the leaves to the root node. This section compares the two mechanisms and demonstrates how establishing one HSMP LSP is better than establishing a P2MP LSP with reverse LSPs from the leaves back to the root.



Consider the topology as shown in Figure 1. Router A wants to establish a Point-to-Multipoint connectivity to Routers E, F, G and H and also wants a Unicast path back from these routers to itself. There are two ways to accomplish this. In the first, we set up a HSMP LSP between A, E, F, G and H. In the second, we set up a P2MP LSP between A, E, F, G and H and establish regular LSPs back from these routers to A.

2.1. Number of Path and Resv State Blocks

When an RSVP-capable router receives an initial Path message, it creates a path state block (PSB) for that particular session. Each PSB consists of parameters derived from the received Path message such as SESSION, SENDER_TEMPLATE, SENDER_TSPEC, RSVP_HOP objects, and the outgoing interface provided by the IGP routing. Similarly, as a Resv message travels upstream toward the sender, it creates a reservation state block (RSB) in each RSVP-capable node along the way which stores information derived from the objects in the received Resv message, such as SESSION, RSVP_HOP, FLOWSPEC, FILTERSPEC, STYLE, etc objects. The PSB and the RSB need to be periodically refreshed by the Path and the Resv messages.
In case of HSMP LSP, the number of PSBs and the RSBs is the same as that for establishing a single P2MP LSP and is a function of how the P2MP LSP is signaled. It is equal to the number of S2L sub-LSPs of the P2MP LSP if each S2L sub-lsp is signaled independently. It is one, if an aggregated mode is used where multiple sub-lsps of the P2MP LSP are signaled togethar.
In the second case routers need to maintain this state for the P2MP LSP and all the Unidirectional LSPs that go via it.
Lets look at the state that branch node B needs to maintain. In case of HSMP LSP it is the same as a P2MP LSP. In the other approach it needs to maintain state for the following LSPs:

1. P2MP LSP from A and E, F, G and H

2. Reverse LSP ECBA

3. Reverse LSP FCBA

4. Reverse LSP GDBA

5. Reverse LSP HDBA

We can thus clearly see that the amount of state that routers need to maintain in the second approach is much more than the HSMP LSP. It becomes all the more pronounced when the P2MP LSP is signalled using the aggregated approach described in [RFC4875] where a single Path and Resv message is used to signal the entire P2MP LSP. In such cases the amount of state that such branch nodes need to maintain increase linearly with the leaf nodes that get added to the P2MP LSP.

2.2. Hardware Programming and Label Utilization

In the HSMP LSP the LSR B advertises the same (upstream) label to C and D, thus consumes only one label and needs to only program one entry in the ILM table.
In the second approach, LSR B needs to advertise two different labels to LSRs C and D and will thus consume 2 ILM entries in HW.
We can clearly see that the number of labels consumed in the second approach will increase linearly with the amount of branching that happens on that LSR. It will further aggravate as the number of P2MP LSPs increase.

2.3. RSVP Control Traffic

In the second approach, LSR B will have RSVP control traffic for the P2MP LSP and all the Unidirectional reverse LSPs that pass through it. In case of HSMP LSR B will only have the RSVP traffic for the P2MP LSP.

3. Setting up a Hub and Spoke Multipoint LSP with RSVP-TE

The Hub and Spoke Multipoint LSP comprises of one downstream unidirectional P2MP LSP from ingress LSR to each of egress LSR, and a co-routed upstream path from each of egress LSR to ingress LSR. [RFC3473] describes a point-to-point bidirectional LSP mechanism for the GMPLS architecture, where a bidirectional LSP setup is indicated by the presence of an Upstream_Label object in the Path message. The Upstream_Label object has the same format as the generalized label, and uses Class-Number 35 (of form 0bbbbbbb) and the C-Type of the label being used. Hub and Spoke Multipoint LSP describe in this draft will use similar mechanism, and reuse the Upstream_Label object defined in [RFC3473]. Note: the downstream label assignment is still applied, and upstream direction is based on the h&s topology (hub = upstream, spoke= downstream), rather on forwarding direction.

3.1. Hub and Spoke Multipoint LSP and Path Messages

[RFC4875] allows a P2MP LSP to be signaled using one or more Path messages . Each Path message may signal one or more source to leaf (S2L) sub-LSPs. This document assumes that a unique Path message is being used to signal each individual sub-LSP of the HSMP LSP. Later

versions of this document can describe mechanisms to use a single Path message to describe each component sub LSP of the HSMP LSP.

3.2. Procedures for Hub and Spoke Multipoint LSP

The process of establishing a Hub and Spoke Multipoint LSP follows the establishment of a unidirectional P2MP LSP define in [RFC4875] with some additions. To support Hub and Spoke Multipoint LSPs an Upstream_Label object is added to the Path message. The Upstream_Label object MUST indicate a label that is valid for forwarding at the time the Path message is sent. When a Path message containing an Upstream_Label object is received, the receiver first verifies that the upstream label is acceptable. If the label is not acceptable, the receiver MUST issue a PathErr message with a "Routing problem/ Unacceptable label value" indication.
The generated PathErr message MAY include an Acceptable Label Set defined in [RFC3473] section 4.1.
The transit node must also allocate one label for the co-routed upstream path before propagating the Path message to all downstream nodes. If a transit node is unable to allocate a label or internal resources, then it MUST issue a PathErr message with a "Routing problem/MPLS label allocation failure" indication. With regards to the co-routed return path from the leafs to the root, the forwarding table on transit node will have one incoming labels allocated for all of the outgoing interfaces, and one outgoing label received from Upstream_Label object in Path message sent by upstream node. That means the traffic from different egress LSRs will be merged at each transit node, and will be sent together to upstream node, see section 3.3 for more detail of bandwidth guarantee in this case.
The Path messages sending downstream with same [P2MP ID, Tunnel ID, Extended Tunnel ID] tuple as part of the SESSION object and the [Tunnel Sender Address, LSP ID] tuple as part of the SENDER_TEMPLATE object, but may different [Sub-Group Originator ID, Sub-Group ID] MUST use same allocated label value for Upstream_Label object.
Leaf nodes process Path messages as usual, with the exception that the upstream label should be used to transport data traffic associated with the Hub and Spoke Multipoint LSP upstream towards the root node.
When a Hub and Spoke Multipoint LSP is removed, both upstream and downstream labels are invalidated and it is no longer valid to send data using the associated labels.

3.3. Bandwidth Allocation

The bandwidth allocation for upstream path from leaf to root could be same as the downstream path from root to leaf node [RFC3473], and the bandwidth will be guarantee only when there is no traffic merging happened on transit node. If there are cases where leaf nodes send traffic to root node at the same time which may cause traffic to be merged on one physical link at transit node, then traffic overload may

happen on these links. There are several ways to avoid this kind of traffic overload. One way is to let the application to do some delay at each leaf node to avoid traffic merging on some links of transit node. Some applications may not require bandwidth guarantee for the upstream path from leaf to root, then it is not necessary to allocate bandwidth for the upstream path. The mechanism described in [I-D.ietf-ccamp-asymm-bw-bidir-lsps-bis] can be used to allocate zero bandwidth for the upstream path.
The mechanism of providing asymmetric bandwidth allocation (non-zero bandwidth of upstream path) for HSMP LSP is out of the draft scope.

4. Setting up the Hub Spoke Multipoint LSP

The Following is an example of establishing a HSMP LSP using the procedures described in the previous sections.

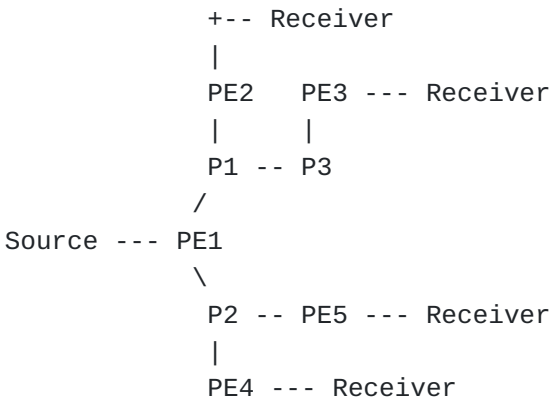


Figure 2

The mechanism is explained using Figure 1. PE1 is a root LER (head end) node. PE2, PE3, PE4 and PE5 are the leaf LER nodes. P1 and P2 are branch LSR nodes and P3 is a plain LSR node.

1. PE1 learns that PE2, PE3, PE4 and PE5 are interested in joining a HSMP tree with a P2MP ID of P2MP ID1. We assume that PE1 learns of the egress LERs at different points in time.

2. PE1 computes the P2P path to reach PE3 and sends a Path message with ERO [PE1, P1, P3, PE3]. It also provides an Upstream Label UL1 in the Upstream_Label object that P1 should use when forwarding packets to PE1.

3. The Path message traverses hop-by-hop and finally reaches PE3. Assume that the Path message from P1 to P3 uses upstream label of UL3, in which case P1 must program the ILM to swap UL3 with UL1. The Path message from P3 to PE3 uses upstream label UL4, and thus P3 programs the ILM to swap UL4 with UL3.

4. PE3 responds with a Resv message that contains label L4, that P3 should use when forwarding packets to PE3. Similarly, the Resv from P3 to P1 contains label L3, that P1 should use when forwarding packets to P3.

5. Similarly when setting up the component sub-LSP from PE1 to PE2, PE1 will use the same Upstream label UL1 as it knows that this sub-LSP belongs to the same HSMP LSP because of the same P2MP session object that both sub-LSPs carry.

6. The Path message, thus for this component sub-LSP goes with ERO [PE1, P1, PE2] along with the Upstream label UL1 that P1 should use when forwarding packets to PE1.

7. P1 forwards the Path message with a new Upstream label UL2. Finally, PE2 sends a Resv message containing label L2, that P1 should use when forwarding packets to PE2. P1 also understands that the Resv messages from PE2 and PE3 refer to the same HSMP LSP, because of the P2MP Session Object carried in each. [

8. P1 sends a separate Resv message to PE1 corresponding to each of the sub-LSPs, but uses the same label L1 since the two sub LSPs belong to the same HSMP LSP.

9. The other component sub LSPs are set up in a similar way as described above.

5. Grafting

The operation of adding leaf LER(s) to an existing HSMP LSP is termed grafting. This operation allows leaf nodes to join a HSMP LSP at different points in time.
The leaf LER(s) can be added by signaling only the impacted component sub- LSPs in a new Path message. Hence, the existing component sub-LSPs do not have to be re-signaled.

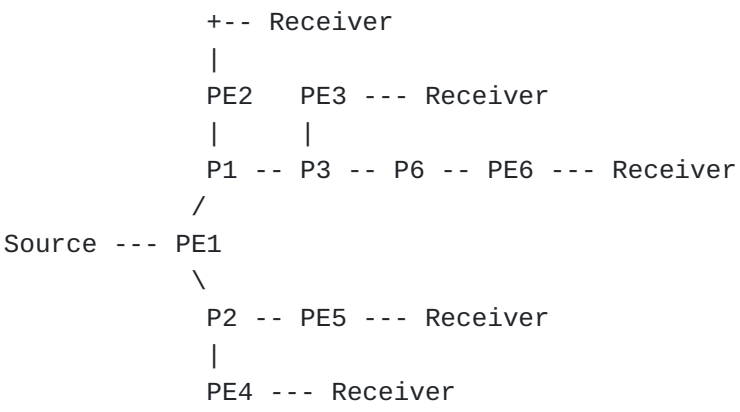


Figure 3

Assume PE1 needs to set up another sub-LSP from PE1 to PE6. Being a part of the same HSMP LSP, PE1 MUST advertise the same Upstream Label to P1 in its Path message. P1 advertises the same Upstream Label to P3. P3 when sending the Path message to P6 would advertise a fresh Upstream label and similarly P6 would use a new upstream label when forwarding the Path message to PE6.
PE6 sends a Resv message with a label back to P6. P6, would send a new label back to P3. P3 because of this new component sub-LSP (PE1-PE6) is now a branch LSR node that performs MPLS multicast replication.

6. Pruning

The operation of removing egress LER nodes from an existing HSMP LSP is termed as pruning. This operation allows leaf nodes to be removed from a HSMP LSP at different points in time. This section describes the mechanisms to perform pruning.
Assume that the LER PE6 wants to be removed from the HSMP LSP. Since we used a unique Path message for each component sub LSP, the teardown will rely on generating a PathTear message for the corresponding Path message. PE6 will send a Path Tear message with the SESSION and SENDER_TEMPLATE objects corresponding to the HSMP LSP and the [Sub-Group Originator ID, Sub-Group ID] tuple corresponding to the Path message. P3 upon receiving the PathTear message would prune the MPLS multicast replication list and will become a normal RSVP LSR node.
In the P2MP and HSMP context the PathTear is used for a specific component sub LSP teardown. This does not necessarily mean the whole path's breakdown from upstream; hence the LSRs MUST retain the Upstream label until all the component sub LSPs of the HSMP LSP are torn down. When a HSMP LSP is removed by the root, a PathTear message MUST be generated for each Path message used to signal the HS Multipoint LSP.

7. Refresh Reduction

The refresh reduction procedures described in [RFC2961] are equally applicable to HS Multipoint LSPs described in this document. Refresh reduction applies to individual messages and the state they install/ maintain, and that continues to be the case for HS Multipoint LSPs.

8. Fast Reroute

[RFC4090] extensions can be used to perform fast reroute for the mechanism described in this document when applied within packet networks.
This is still TBD.

9. Acknowledgements

We would like to thank Dimitri Papadimitriou, Yuji Kamite, Sebastien Jobert for their comments and feedback on the document.

10. Security Considerations

The same security considerations apply as for the RSVP-TE P2MP LSP specification, as described in [RFC4875].

11. IANA Considerations

No requests for IANA at this point of time.

12. References

12.1. Normative References

, "

[IEEE]	IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems ", 2008.
[RFC2119]	Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
[RFC3473]	Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.
[RFC4875]	Aggarwal, R., Papadimitriou, D. and S. Yasukawa, "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, May 2007.
[I-D.ietf-ccamp-asymm-bw-bidir-lsps-bis]	Takacs, A, Berger, L, Caviglia, D, Fedyk, D and J Meuric, "GMPLS Asymmetric Bandwidth Bidirectional Label Switched Paths (LSPs)", Internet-Draft draft-ietf-ccamp-asymm-bw-bidir-lsps-bis-03, August 2011.

12.2. Informative References

[RFC4090]	Pan, P., Swallow, G. and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.
[RFC2961]	Berger, L., Gan, D., Swallow, G., Pan, P., Tommasi, F. and S. Molendini, "RSVP Refresh Overhead Reduction Extensions", RFC 2961, April 2001.
[RFC4672]	De Cnodder, S., Jonnala, N. and M. Chiba, "RADIUS Dynamic Authorization Client MIB", RFC 4672, September 2006.

[I-D.ietf-pwe3-p2mp-pw]	Sivabalan, S, Boutros, S, Martini, L, Konstantynowicz, M, Vecchio, G, Kamite, Y and L Jin, "Signaling Root-Initiated Point-to-Multipoint Pseudowire using LDP", Internet-Draft draft-ietf-pwe3-p2mp-pw-03, October 2011.
[I-D.ietf-l2vpn-vpms-frmwk-requirements]	Kamite, Y, JOUNAY, F, Niven-Jenkins, B, Brungard, D and L Jin, "Framework and Requirements for Virtual Private Multicast Service (VPMS)", Internet-Draft draft-ietf-l2vpn-vpms-frmwk-requirements-04, July 2011.
[I-D.ietf-l2vpn-ldp-vpls-broadcast-exten]	Delord, S, Key, R, JOUNAY, F, Kamite, Y, Liu, Z, Paul, M, Kunze, R, Chen, M and L Jin, "Extension to LDP-VPLS for Ethernet Broadcast and Multicast", Internet-Draft draft-ietf-l2vpn-ldp-vpls-broadcast-exten-02, June 2011.
[I-D.ietf-l2vpn-vpls-mcast]	Aggarwal, R, Kamite, Y and L Fang, "Multicast in VPLS", Internet-Draft draft-ietf-l2vpn-vpls-mcast-09, July 2011.

Authors' Addresses

Lizhong Jin Jin ZTE Corporation Bibo Road, Shanghai, 201203 China
EMail: lizhong.jin@zte.com.cn

Frederic Jounay Jounay France Telecom Lannion Cedex, 95134 France
EMail: frederic.jounay@orange-ftgroup.com

Manav Bhatia Bhatia Alcatel-Lucent Bangalore, India EMail:
manav.bhatia@alcatel-lucent.com