

Network Working Group
Internet-Draft
Updates: [3209](#), [3473](#)
(if approved)
Intended status: Standards Track
Expires: January 6, 2011

K. Kompella
Juniper Networks
July 5, 2010

Multi-path Label Switched Paths Signaled Using RSVP-TE
draft-kompella-mp-ls-rsvp-ecmp-00.txt

Abstract

This document describes extensions to Resource ReSerVation Protocol - Traffic Engineering for the set up of multi-path Traffic Engineered Label Switched Paths (LSPs) in Multi Protocol Label Switching and Generalized MPLS networks, i.e., LSPs that conform to traffic engineering constraints, but follow multiple independent paths from the source to the destination that allow load balancing.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of [BCP 78](#) and [BCP 79](#).

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on January 6, 2011.

Copyright Notice

Copyright (c) 2010 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to [BCP 78](#) and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must

include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1.	Introduction	3
1.1.	Terminology	3
1.2.	Conventions used in this document	4
2.	Theory of Operation	5
2.1.	Multi-path Label Switched Paths	5
2.2.	ECMP	6
2.3.	Discussion	8
2.4.	The Capabilities of TE-based Load Balancing	9
3.	Operation of MLSPs	10
3.1.	Signaling MLSPs	10
3.1.1.	MLSP_TUNNEL Sender Template	10
3.1.2.	MLSP_TUNNEL Filter Specification	11
3.2.	Label Allocation	11
3.3.	Bandwidth Accounting	11
3.4.	MLSP Data Plane Actions	13
4.	Security Considerations	14
5.	Acknowledgments	15
6.	IANA Considerations	16
7.	References	17
7.1.	Normative References	17
7.2.	Informative References	17
	Author's Address	19

1. Introduction

In selecting a protocol for setting up and signaling "tunnel" Labeled Switched Paths (LSPs) in Multi Protocol Label Switching (MPLS) and Generalized MPLS (GMPLS) networks, one first chooses whether one wants Equal Cost Multi-Path (ECMP) load balancing or Traffic Engineering (TE). For the former, one uses the Label Distribution Protocol (LDP) ([[RFC5036](#)]); for the latter, the Resource ReSerVation Protocol - Traffic Engineering (RSVP-TE) ([[RFC3209](#)]). [Two other criteria, the need for fast protection and the desire for less configuration, are no longer the deciding factors they used to be, thanks to "IP fast reroute" ([[RFC5286](#)]) and "RSVP-TE automesh" ([[RFC4972](#)])].

This document describes how one can set up a tunnel LSP that has both ECMP and TE characteristics using RSVP-TE. The techniques described in this document can be used to create a single overall "ECMP TE LSP" to a single destination that consists of several "sub-LSPs", each taking a different path through the network to the same destination. The techniques can also be used to create a single ECMP TE LSP to multiple equivalent destinations (such as equidistant BGP nexthops announcing a common set of reachable addresses), such that each destination is served by one or more sub-LSPs. The techniques described here borrow the notion of sub-LSPs from [[RFC4875](#)].

Several options are available for ECMP TE LSPs. One is that the ingress Label Switching Router (LSR) computes (or otherwise obtains) all sub-LSP paths; alternatively, LSRs along the various paths can compute paths further downstream (using techniques such as "loose hop expansion", as in [[RFC5152](#)]). Another is that an RSVP Path message can contain information about exactly one path through the network (or sub-LSP); alternately, a Path message can contain information about more than one such path. A third option is that the various paths that make up the multi-path LSP have equal cost (or distance) from ingress to egress (i.e., ECMP), as opposed to paths that may have differing costs. Another option (mentioned above) is to terminate a multi-path LSP on a single egress or on several equivalent egresses. For now, the first of each of these alternatives is assumed; future work can explore other choices.

1.1. Terminology

The terms "tunnel", "tunnel LSP" and "LSP" all refer to a container LSP from an ingress LSR to egress LSR(s). An LSP is the unit of configuration, signaling and management.

An ECMP (or generally, a multi-path) TE LSP is called a Multi-path Label Switched Path (MLSP), and consists of one or more sub-LSPs.

A sub-LSP consists of a single path from the ingress to one egress. A "regular" point-to-point TE LSP is equivalent to an MLSP with a single sub-LSP.

The "downstream links" of an LSR X with respect to an MLSP Z is the set of all links adjacent to X traversed after X by at least one sub-LSP of MLSP Z.

1.2. 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 [[RFC2119](#)].

2. Theory of Operation

2.1. Multi-path Label Switched Paths

An MLSP is configured at the ingress with various constraints typically associated with TE LSPs, such as destination LSR(s), bandwidth (on a per-class basis, if desired), link colors, Shared Risk Link Groups, etc. [Auto-mesh techniques ([\[RFC4972\]](#)) can be used to reduce configuration; this is not described further here.] In addition, parameters specifically related to MLSPs, such as how many (or the maximum number of) sub-LSPs to create, whether traffic should be split equally across sub-LSPs or not, etc. may also be specified.

The ingress LSR can use the configuration parameters to decide how many sub-LSPs to compute for this MLSP and what paths they should take. Each sub-LSP MUST meet all the constraints of the MLSP (except the bandwidth). The bandwidths (per-class, if applicable) of all the sub-LSPs MUST add up to the bandwidth of the MLSP. If a Path Computation Element (PCE; [\[RFC4655\]](#)) that is multi-path LSP-aware is used, the PCE is subject to these same requirements; how MLSP requirements are signaled to a PCE is beyond the scope of this document.

Having computed (or otherwise obtained) the paths of all the sub-LSPs, the ingress A then signals the MLSP by signaling all the individual sub-LSPs across the MPLS/GMPLS network. If multiple sub-LSPs of the same MLSP pass through LSR Y, and Y has downstream links YP, YQ and YR for the various sub-LSPs, then Y has to load balance incoming traffic for the MLSP across the three downstream links in proportion to the sum of the bandwidths of the sub-LSPs going to each downstream (see Figure 1).

One must distinguish carefully between the (signaled) bandwidth of a sub-LSP, a static value capturing the expected or maximum traffic on the sub-LSP, and the instantaneous traffic received on a sub-LSP, a constantly varying quantity. Suppose there are three sub-LSPs traversing Y, with bandwidths 10Gbps, 20Gbps and 30Gbps, going to P, Q and R respectively. Suppose further Y receives some traffic over each of these sub-LSPs. Y must balance this received traffic over the three downstream links YP, YQ and YR in the ratio 1:2:3.

Staying with this example, A has one sub-LSP of bandwidth 30Gbps to M and four sub-LSPs of total bandwidth 90Gbps to X. Thus, A should load

If a set of EB sub-LSPs of the same MLSP traverse an LSR S, say to downstream links SP, SQ and SR, then S MUST attempt to load balance traffic received on these EB sub-LSPs equally among the links SP, SQ and SR, independent of how many sub-LSPs go over each of these links. Furthermore, S MUST redistribute traffic received from each of its upstream LSRs, and SHOULD redistribute all traffic received from upstream as a whole. One can do the former by signaling the same label to each of its upstream LSRs; one can do the latter by signaling the same label to all upstream LSRs (see [Section 3.2](#)). For example, in Figure 2, if L sends 12Gbps of traffic to S and M sends 18Gbps to S, S can redistribute L's traffic by sending 4Gbps to each of P, Q and R; and can similarly send 6Gbps of M's traffic to each of P, Q and R. Alternatively, S can load balance the aggregate 30Gbps of traffic received from L and M to each of P, Q and R, thus sending 10Gbps to each. EB sub-LSPs have an added benefit of not requiring

unequal load balancing across links, which may pose problems for some hardware.

Given the notion of EB sub-LSPs and EB MLSPs, A can signal an EB MLSP Z comprised of five EB sub-LSPs E1 through E5 with the following paths: ALSPTUB, AMSQTVB, ALSRTWB, AMSPTXB and ALSQTYB (respectively). Then, A has two downstream links for the five sub-LSPs, AL and AM, between which A will load balance equally. Similarly, S has three downstream links, SP, SQ and SR; and T has five downstreams, TU, TV, TW, TX and TY. Thus the load balancing behavior of the MLSP will replicate IP load balancing. The state required for an EB MLSP to achieve IP-style load balancing is somewhat greater than for LDP LSPs, but significantly less than that for multiple "regular" TE LSPs, or for a non-EB MLSP.

2.3. Discussion

Some of the power of TE-based ECMP was illustrated in the above examples. Another is ability to request that all sub-LSPs avoid links colored red. If in the example network in Figure 1, the QT link is colored red but all other links are not, then there are four ECMP paths that satisfy these constraints, and the traffic distribution among them will naturally be different than it would without the link color constraint.

One can also ask whether an MLSP with sub-LSPs is any better than N "regular" LSPs from the same ingress to the same egress. Here are some benefits of an MLSP:

1. With an MLSP, there is a single entity to provision, manage and monitor, versus N separate entities in the case of LSPs. A consequence of this is that with an MLSP, changes in topology can be dealt with easily and autonomously by the ingress LSR, by adding, changing or removing sub-LSPs to rebalance traffic, while maintaining the same TE constraints. With individual LSPs, such changes would require changes in configuration, and thus are harder to automate.
2. An ingress LSR, knowing that an MLSP is for load balancing, can decide on an optimum number of sub-LSPs, and place them appropriately across the network to optimize load balancing. On the other hand, an ingress LSR asked to create N independent LSPs will do so without regard to whether N is a good number of equal cost paths, and, more importantly, may place several of the N LSPs on the same path, defeating the purpose of load balancing.
3. The EB sub-LSP mechanism will, in many cases, result in far fewer sub-LSPs than independent LSPs and thus less control plane state.

4. Finally, an MLSP will usually have less data plane state than N independent LSPs: whenever multiple sub-LSPs traverse a link, a single label will be used for all of them, whereas if multiple LSPs traverse a link, each will need a separate label.

2.4. The Capabilities of TE-based Load Balancing

Definition: Let $G=(V, E)$ be a directed graph (or network), and let A and B in V be two nodes in G . Let T be the traffic arriving at A destined for B . T is said to be "IP-style" load balanced if for every node X on a shortest path from A to B , the portion of T arriving at X is split equally among all nodes Y_i that are adjacent to X and are on a shortest path from X to B .

Theorem: An MLSP can accurately mimic IP-style load balancing between any two nodes in any network.

Proof: left to the reader.

Corollary: MLSPs provide a strictly more powerful load balancing mechanism than IP-style load balancing.

3. Operation of MLSPs

3.1. Signaling MLSPs

An MLSP is identified by an LSP_TUNNEL SESSION object defined in [RFC3209]. All sub-LSPs of an MLSP have the same field values in their LSP_TUNNEL SESSION object.

A sub-LSP of an MLSP is identified by the LSP_TUNNEL SESSION object plus a new Sender Template object called the MLSP_TUNNEL Sender Template. The MLSP_TUNNEL Sender Template comes in two flavors, IPv4 and IPv6, shown below. The 15-bit Sub-LSP ID uniquely identifies a sub-LSP of an MLSP, and stays the same during the lifetime of the sub-LSP. The LSP ID may change as in [RFC3209] to let a sub-LSP share resources with another incarnation of the sub-LSP, for example to reroute and/or change bandwidths of the sub-LSP. The "E" bit defines whether a sub-LSP is an equi-bandwidth sub-LSP (E=1) or not (E=0). The equi-bandwidth character of a sub-LSP (i.e., the value of the E bit) MUST remain the same from ingress to egress as well as during the lifetime of a sub-LSP.

3.1.1. MLSP_TUNNEL Sender Template

Class = SENDER_TEMPLATE, MLSP_TUNNEL_IPv4 C-Type = TBD

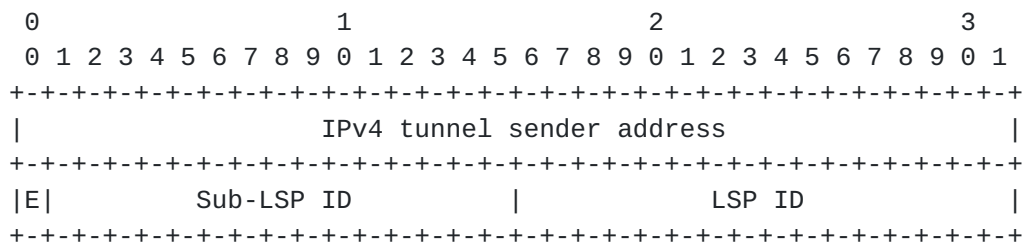


Figure 3: MLSP_TUNNEL_IPv4 Sender Template

MLSP Z is divided equally among all the downstream links of X which at least one of the EB sub-LSPs traverses. Second, LSR X MUST ensure that, for each upstream link of X, there is sufficient bandwidth to accommodate all EB sub-LSPs of MLSP Z that traverse that link.

Let's take the example of Figure 2, with MLSP Z having five EB sub-LSPs E1 to E5, and say that MLSP Z is configured with a bandwidth of 30Gbps. Here are some of the steps involved.

1. LSR A, being the ingress, has no upstream links. A has two downstream links, AL and AM. Three EB sub-LSPs of MLSP Z traverse AL, and two traverse AM. A MUST signal a total of 15Gbps for the sub-LSPs to L, and a total of 15Gbps for the sub-LSPs to M. The required bandwidth may be divided up among the sub-LSPs to L (similarly, to M) in any manner so long as the total is 15Gbps. For example, A can signal sub-LSP E1 with 15Gbps, and sub-LSPs E3 and E5 with 0 bandwidth.
2. LSR L has one upstream link AL with three EB sub-LSPs with a total bandwidth of 15Gbps. L MUST ensure that 15Gbps is available for the AL link. If this bandwidth is not available, L MUST send a PathErr on ALL of the EB sub-LSPs on the AL link. Let's assume that the AL link has sufficient bandwidth.
3. Next, it is up to L to decide how to divide the incoming 15Gbps among the three downstream EB sub-LSPs to S. Say L signals sub-LSP E1 with 15Gbps, and the others with 0 bandwidth.
4. LSR S has two upstream links: LS with three EB sub-LSPs with a total bandwidth of 15Gbps, and MS with two EB sub-LSPs with a total bandwidth of 15Gbps. S MUST ensure that 15Gbps is available for each of the LS and MS links. S has thus a total incoming bandwidth of 30Gbps on MLSP Z. S has to divide this equally among its downstream links SP, SQ and SR, yielding 10Gbps each. S MUST ensure that the total bandwidth requested on the SP link for sub-LSPs E1 and E4 is 10Gbps. S may choose to signal these sub-LSPs with 5Gbps each. Similarly for the SQ and SR links.

There are two important points to note here. One is that the bandwidth reservation (TSpec) for a given EB sub-LSP can (and usually will) change hop-by-hop. The second is that as new EB sub-LSPs are signaled for an MLSP, the bandwidth reservations for existing EB sub-LSPs belonging to the same MLSP may have to be updated. To minimize these updates, it is RECOMMENDED that the first EB sub-LSP on a link be signaled with the total required bandwidth (as far as is known), and later sub-LSPs on the same link be signaled with 0 bandwidth.

3.4. MLSP Data Plane Actions

Traffic intended to be sent over an MLSP is determined at the ingress LSR by means outside the scope of this document, and at transit LSRs by the label(s) assigned by the transit LSR to its upstream LSRs. In the case of non-EB sub-LSPs, this traffic is load balanced across downstream links in the ratio of the bandwidths of the sub-LSPs that comprise the MLSP. In the case of EB sub-LSPs, the traffic belonging to an MLSP from an upstream LSR (or better still, the aggregate traffic for the MLSP from all upstream LSRs) is load balanced equally among all downstream links.

As noted above, the overriding concern is that flows are mapped to the same downstream link (except when the MLSP or some constituent sub-LSPs are changing); this is typically done by hashing fields that define a flow, and mapping hash results to different downstream LSRs. Hash-based load balancing typically assumes that the numbers of flows is sufficiently large and the bandwidth per flow is reasonably well-balanced so that the results of hashing yields reasonable traffic distribution.

Entropy labels ([[I-D.kompella-mpls-entropy-label](#)] and [[I-D.ietf-pwe3-fat-pw](#)]) can be used to improve load balancing at intermediate nodes.

4. Security Considerations

This document introduces no new security concerns in the setup and signaling of LSPs using RSVP-TE, or in the use of the RSVP protocol. [RFC2205] specifies the message integrity mechanisms for RSVP signaling. These mechanisms apply to RSVP-TE signaling of MLSPs described in this document, and are highly recommended pending newer mechanisms for RSVP.

5. Acknowledgments

The author would like to thank the Routing Protocol group at Juniper Networks for their questions, comments and encouragement for this proposal. While many participated, special thanks go to Yakov Rekhter, John Drake and Rahul Aggarwal.

6. IANA Considerations

IANA is requested to assign two new C-Types for the Class "Sender Template", one for the "MLSP_TUNNEL_IPv4" Sender Template and one for the "MLSP_TUNNEL_IPv6" Sender Template.

IANA is also requested to assign two new C-Types for the Class "Filter Specification", one for the "MLSP_TUNNEL_IPv4" Filter Specification and one for the "MLSP_TUNNEL_IPv6" Filter Specification.

7. References

7.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", [BCP 14](#), [RFC 2119](#), March 1997.
- [RFC2205] Braden, B., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", [RFC 2205](#), September 1997.
- [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.
- [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.

7.2. Informative References

- [RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", [RFC 4655](#), August 2006.
- [RFC4972] Vasseur, JP., Leroux, JL., Yasukawa, S., Previdi, S., Psenak, P., and P. Mabbey, "Routing Extensions for Discovery of Multiprotocol (MPLS) Label Switch Router (LSR) Traffic Engineering (TE) Mesh Membership", [RFC 4972](#), July 2007.
- [RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", [RFC 5036](#), October 2007.
- [RFC5152] Vasseur, JP., Ayyangar, A., and R. Zhang, "A Per-Domain Path Computation Method for Establishing Inter-Domain Traffic Engineering (TE) Label Switched Paths (LSPs)", [RFC 5152](#), February 2008.
- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", [BCP 26](#), [RFC 5226](#), May 2008.
- [RFC5286] Atlas, A. and A. Zinin, "Basic Specification for IP Fast Reroute: Loop-Free Alternates", [RFC 5286](#), September 2008.
- [I-D.ietf-pwe3-fat-pw] Bryant, S., Filsfils, C., Drafz, U., Kompella, V., Regan,

J., and S. Amante, "Flow Aware Transport of Pseudowires over an MPLS PSN", [draft-ietf-pwe3-fat-pw-03](#) (work in progress), January 2010.

[I-D.kompella-mpls-entropy-label]

Kompella, K. and S. Amante, "The Use of Entropy Labels in MPLS Forwarding", [draft-kompella-mpls-entropy-label-00](#) (work in progress), July 2008.

Author's Address

Kireeti Kompella
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: kireeti@juniper.net