MPLS Internet-Draft Intended status: Informational Expires: July 29, 2013

Use of Multipath with MPLS-TP and MPLS draft-villamizar-mpls-multipath-use-02

Abstract

Many MPLS implementations have supported multipath techniques and many MPLS deployments have used multipath techniques, particularly in very high bandwidth applications, such as provider IP/MPLS core networks. MPLS-TP has strongly discouraged the use of multipath techniques. Some degradation of MPLS-TP OAM performance cannot be avoided when operating over many types of multipath implementations.

Using MPLS Entropy label, MPLS LSPs can be carried over multipath links while also providing a fully MPLS-TP compliant server layer for MPLS-TP LSPs. This document describes the means of supporting MPLS as a server layer for MPLS-TP. The use of MPLS-TP LSPs as a server layer for MPLS LSPs is also discussed.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of <u>BCP 78</u> and <u>BCP 79</u>.

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 <u>http://datatracker.ietf.org/drafts/current/</u>.

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 July 29, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to $\frac{\text{BCP}\ 78}{\text{Provisions}}$ and the IETF Trust's Legal Provisions Relating to IETF Documents

(<u>http://trustee.ietf.org/license-info</u>) 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

3
<u>3</u>
5
<u>6</u>
7
0
1
1
1
2
2
2
2
<u>3</u>

1. Introduction

Today the requirement to handle large aggregations of traffic, can be handled by a number of techniques which we will collectively call multipath. Multipath applied to parallel links between the same set of nodes includes Ethernet Link Aggregation [IEEE-802.1AX], link bundling [RFC4201], or other aggregation techniques some of which may be vendor specific. Multipath applied to diverse paths rather than parallel links includes Equal Cost MultiPath (ECMP) as applied to OSPF, ISIS, or BGP, and equal cost LSPs. Some vendors support load split across equal cost MPLS LSPs where the load is split proportionally to the reserved bandwidth of the set of LSPs.

<u>RFC 5654</u> requirement 33 requires the capability to carry a client MPLS-TP or MPLS layer over a server MPLS-TP or MPLS layer [<u>RFC5654</u>]. This is possible in all cases with one exception. When an MPLS LSP exceeds the capacity of any single component link it may be carried by a network using multipath techniques, but may not be carried by a single MPLS-TP LSP due to the inherent MPLS-TP capacity limitation imposed by MPLS-TP OAM fate sharing constraints and MPLS-TP LM OAM packet ordering constraints (see <u>Section 3.1</u>).

The term composite link is more general than terms such as link aggregation (which is specific to Ethernet) or ECMP (which implies equal cost paths within a routing protocol). The use of the term composite link here is consistent with the broad definition in [ITU-T.G.800]. Multipath is very similar to composite link as defined by ITU, but specifically excludes inverse multiplexing.

2. Definitions

Multipath

The term multipath includes all techniques in which

- 1. Traffic can take more than one path from one node to a destination.
- 2. Individual packets take one path only. Packets are not subdivided and reassembled at the receiving end.
- 3. Packets are not resequenced at the receiving end.
- 4. The paths may be:
 - a. parallel links between two nodes, or

[Page 3]

- b. may be specific paths across a network to a destination node, or
- c. may be links or paths to an intermediate node used to reach a common destination.

Link Bundle

Link bundling is a multipath technique specific to MPLS [RFC4201]. Link bundling supports two modes of operations. Either an LSP can be placed on one component link of a link bundle, or an LSP can be load split across all members of the bundle. There is no signaling defined which allows a per LSP preference regarding load split, therefore whether to load split is generally configured per bundle and applied to all LSPs across the bundle.

Link Aggregation

The term "link aggregation" generally refers to Ethernet Link Aggregation [IEEE-802.1AX] as defined by the IEEE. Ethernet Link Aggregation defines a Link Aggregation Control Protocol (LACP) which coordinates inclusion of LAG members in the LAG.

Link Aggregation Group (LAG)

A group of physical Ethernet interfaces that are treated as a logical link when using Ethernet Link Aggregation is referred to as a Link Aggregation Group (LAG).

Equal Cost Multipath (ECMP)

Equal Cost Multipath (ECMP) is a specific form of multipath in which the costs of the links or paths must be equal in a given routing protocol. The load may be split equally across all available links (or available paths), or the load may be split proportionally to the capacity of each link (or path).

Loop Free Alternate Paths

"Loop-free alternate paths" (LFA) are defined in <u>RFC 5714</u>, Section 5.2 [RFC5714] as follows. "Such a path exists when a direct neighbor of the router adjacent to the failure has a path to the destination that can be guaranteed not to traverse the failure." Further detail can be found in [RFC5286]. LFA as defined for IPFRR can be used to load balance by relaxing the equal cost criteria of ECMP, though IPFRR defined LFA for use in selecting protection paths. When used with IP, proportional split is generally not used. LFA use in load balancing is implemented by some vendors though it may be rare or non-existent in deployments.

[Page 4]

Composite Link

The term Composite Link had been a registered trademark of Avici Systems, but was abandoned in 2007. The term composite link is now defined by the ITU in [ITU-T.G.800]. The ITU definition includes multipath as defined here, plus inverse multiplexing which is explicitly excluded from the definition of multipath.

Inverse Multiplexing

Inverse multiplexing either transmits whole packets and resequences the packets at the receiving end or subdivides packets and reassembles the packets at the receiving end. Inverse multiplexing requires that all packets be handled by a common egress packet processing element and is therefore not useful for very high bandwidth applications.

Component Link

The ITU definition of composite link in [ITU-T.G.800] and the IETF definition of link bundling in [RFC4201] both refer to an individual link in the composite link or link bundle as a component link. The term component link is applicable to all multipath.

IAG Member

Ethernet Link Aggregation as defined in [IEEE-802.1AX] refers to an individual link in a LAG as a LAG member. A LAG member is a component link. An Ethernet LAG is a composite link. IEEE does not use the terms composite link or component link.

load split

Load split, load balance, or load distribution refers to subdividing traffic over a set of component links such that load is fairly evenly distributed over the set of component links and certain packet ordering requirements are met. Some existing techniques better acheive these objectives than others.

A small set of requirements are discussed. These requirements make use of keywords such as MUST and SHOULD as described in [RFC2119].

3. MPLS as a Server Layer for MPLS-TP

An MPLS LSP may be used as a server layer for MPLS-TP LSPs as long as all MPLS-TP requirements are met. Section 3.1 reviews the basis for requirements of a server layer that supports MPLS-TP as a client layer. Key requirements include OAM "fate-sharing" the the requirement that packets within an MPLS-TP LSP are not reordered, including both payload and OAM packets. <u>Section 3.2</u> discusses implied requirements where MPLS is the server layer for MPLS-TP

[Page 5]

client LSPs, and describes a set of solutions using existing MPLS mechanisms.

3.1. MPLS-TP Forwarding and Server Layer Requirements

[RFC5960] defines the date plane requirements for MPLS-TP. Two very relevant paragraphs in "Section 3.1.1 LSP Packet Encapsulation and Forwarding" are the following.

```
RFC5960, Section 3.1.1, Paragraph 3
```

Except for transient packet reordering that may occur, for example, during fault conditions, packets are delivered in order on L-LSPs, and on E-LSPs within a specific ordered aggregate.

RFC5960, Section 3.1.1, Paragraph 6

Equal-Cost Multi-Path (ECMP) load-balancing MUST NOT be performed on an MPLS-TP LSP. MPLS-TP LSPs as defined in this document MAY operate over a server layer that supports load-balancing, but this load-balancing MUST operate in such a manner that it is transparent to MPLS-TP. This does not preclude the future definition of new MPLS-TP LSP types that have different requirements regarding the use of ECMP in the server layer.

[RFC5960] paragraph 3 requires that packets within a specific ordered aggregate be delivered in order. This same requirement is already specified by Differentiated Services [<u>RFC2475</u>]. [<u>RFC5960</u>] paragraph 6 explicitly allows a server layer to use ECMP provided that it is transparent to the MPLS-TP client layer.

[RFC6371] adds a requirement for data traffic and OAM traffic "fatesharing". The following paragraph in "Section 1 Introduction" summarizes this requirement.

RFC6371, Section 1, Paragraph 7

OAM packets that instrument a particular direction of a transport path are subject to the same forwarding treatment (i.e., fateshare) as the user data packets and in some cases, where Explicitly TC-encoded-PSC LSPs (E-LSPs) are employed, may be required to have common per-hop behavior (PHB) Scheduling Class (PSC) End-to-End (E2E) with the class of traffic monitored. In case of Label-Only-Inferred-PSC LSP (L-LSP), only one class of traffic needs to be monitored, and therefore the OAM packets have common PSC with the monitored traffic class.

[RFC6371] does not prohibit multilink techniques in "Section 4.6 Fate-Sharing Considerations for Multilink", where multilink is defined as Ethernet Link Aggregation and the use of Link Bundling for MPLS, but does declare that such a network would be only partially

[Page 6]

MPLS-TP compliant. The characteristic that is to be avoided is contained in the following sentence in this section.

RFC6371, Section 4.6, Paragraph 1, last sentence

These techniques frequently share the characteristic that an LSP may be spread over a set of component links and therefore be reordered, but no flow within the LSP is reordered (except when very infrequent and minimally disruptive load rebalancing occurs).

A declaration that implies that Link Bundling for MPLS yields a partially MPLS-TP compliant network, is perhaps overstated since only the Link Bundling all-ones component link has this characteristic.

[RFC6374] defines a direct Loss Measurement (LM) where LM OAM packets cannot be reordered with respect to payload packets. This will require that payload packets themselves not be reordered. The following paragraph in "Section 2.9.4 Equal Cost Multipath" gives the reason for this restriction.

<u>RFC6374, Section 2.9.4</u>, Paragraph 2

The effects of ECMP on loss measurement will depend on the LM mode. In the case of direct LM, the measurement will account for any packets lost between the sender and the receiver, regardless of how many paths exist between them. However, the presence of ECMP increases the likelihood of misordering both of LM messages relative to data packets and of the LM messages themselves. Such misorderings tend to create unmeasurable intervals and thus degrade the accuracy of loss measurement. The effects of ECMP are similar for inferred LM, with the additional caveat that, unless the test packets are specially constructed so as to probe all available paths, the loss characteristics of one or more of the alternate paths cannot be accounted for.

3.2. Methods of Supporting MPLS-TP client LSPs over MPLS

Supporting MPLS-TP LSPs over a fully MPLS-TP conformant MPLS LSP server layer where the MPLS LSPs are making use of multipath, requires special treatment of the MPLS-TP LSPs such that those LSPs meet MPLS-TP forwarding requirements (see <u>Section 3.1</u>). This implies the following brief set of requirements.

MP#1 It MUST be possible for a midpoint MPLS-TP LSR which is serving as ingress to a server layer MPLS LSP to identify MPLS-TP LSPs, so that MPLS-TP forwarding requirements can be applied, or to otherwise accommodate the MPLS-TP forwarding requirements.

[Page 7]

- MP#2 It SHOULD be possible to completely exclude MPLS-TP LSPs from the multipath hash and load split. If the selected component link no longer meets requirements, an LSP is considered down which may trigger protection and/or may require that the ingress LSR select a new path and signal a new LSP.
- MP#3 It SHOULD be possible to insure that MPLS-TP LSPs will not be moved to another component link as a result of a composite link load rebalancing operation. If the selected component link no longer meets requirements, another component link may be selected, however a change in path should not occur solely for load balancing.
- MP#4 Where an RSVP-TE control plane is used, it MUST be possible for an ingress LSR which is setting up an MPLS-TP or an MPLS LSP to determine at path selection time whether a link or Forwarding Adjacency (FA, see [RFC4206]) within the topology can support the MPLS-TP requirements of the LSP.

The reason for requirement MP#1 may not be obvious. A MPLS-TP LSP may be aggregated along with other client LSP by a midpoint LSR into a very large MPLS server layer LSP, as would be the case in a core node to core node MPLS LSP between major cities. In this case the ingress of the MPLS LSP cannot through any existing signaling mechanism determine which client LSP contained within it as MPLS-TP or not MPLS-TP. Multipath load splitting can be avoided for MPLS-TP LSP if at the MPLS server layer LSP ingress LSR an Entropy Label Indicator (ELI) and Entropy Label (EL) are added to the label stack [RFC6790]. For those client LSP that are MPLS-TP LSP, a single EL value must be chosen. For those client LSP that are MPLS LSP, per packet entropy below the top label must, for practical reasons, be used to determine the entropy label value. Requirement MP#1 simply states that there must be a means to make this decision.

There is currently no signaling mechanism defined to support requirement MP#1, though that does not preclude a new extension being defined later. In the absense of a signaling extension, MPLS-TP can be identified through some form of configuration, such as configuration which provides an MPLS-TP compatible server layer to all LSP arriving on a specific interface or originating from a specific set of ingress LSR.

Alternately, the need for requirement MP#1 can be eliminated if evey MPLS-TP LSP can be created by the MPLS-TP ingress makes use of an Entropy Label Indicator (ELI) and Entropy Label (EL) below the MPLS-TP label [RFC6790]. This would require that all MPLS-TP LSR in a deployment support Entropy Label, which may render it impractical in many deployments.

[Page 8]

Some hardware which exists today can support requirement MP#2. Signaling in the absense of MPLS Entropy Label can make use of link bundling with the path pinned to a specific component for MPLS-TP LSP and link bundling using the all-ones component for MPLS LSP. This prevents MPLS-TP LSP from being carried within MPLS LSP but does allow the co-existance of MPLS-TP and very large MPLS LSP.

MPLS-TP LSPs can be carried as client LSPs within an MPLS server LSP if an Entropy Label Indicator (ELI) and Entropy Label (EL) is added after the server layer LSP label(s) in the label stack, just above the MPLS-TP LSP label entry [<u>RFC6790</u>]. The value of EL can be randomly selected at the client MPLS-TP LSP setup time and the same EL value used for all packets of that MPLS-TP LSP. This allows MPLS-TP LSP to be carried as client LSP within MPLS LSP and satisfies MPLS-TP forwarding requirements but requires that MPLS LSR be able to identify MPLS-TP LSP (requirement MP#1).

MPLS-TP traffic can be protected from an degraded performance due to an imperfect load split if the MPLS-TP traffic is given gueuing priority (using strict priority and policing or shaping at ingress or locally or weighted queuing locally). This can be accomplished using the Traffic Class field and Diffserv treatment of traffic [RFC5462][RFC2475]. In the event of congestion due to load imbalance, other traffic will suffer as long as there is a minority of MPLS-TP traffic.

If MPLS-TP LSP are carried within MPLS LSP and ELI and EL are used, requirement MP#3 is satisfied only for uncongested links where load balancing is not required, or if MPLS-TP LSP use TC and Diffserv and the load rebalancing implementation rebalances only the less preferred traffic. Load rebalance is generally needed only when congestion occurs, therefore restricting MPLS-TP to be carried only over MPLS LSP that are known to traverse only links which are expected to be uncongested can satisfy requirement MP#3.

An MPLS-TP LSP can be pinned to a Link Bundle component link if the behavior of requirement MP#2 is preferred. An MPLS-TP LSP can be assigned to a Link Bundle but not pinned if the behavior of requirement MP#3 is preferred. In both of these cases, the MPLS-TP LSP must be the top level LSP, except as noted above.

If MPLS-TP LSP can be moved among component links, then the Link Bundle all-ones component link can be used or server layer MPLS LSPs can be used with no restrictions on the server layer MPLS use of multipath except that Entropy Label must be supported along the entire path. An Entropy Label must be used to insure that all of the MPLS-TP payload and OAM traffic are carried on the same component, except during very infrequent transitions due to load balancing.

[Page 9]

An MPLS-TP LSP may not traverse multipath links on the path where MPLS-TP forwarding requirements cannot be met. Such links include any using pre-RFC6790 Ethernet Link Aggregation, pre-RFC6790 Link Bundling using the all-ones component link, or other form of multipath not supporting termination of the entropy search at the EL label as called for in [RFC6790]. An MPLS-TP LSP must not traverse a server layer MPLS LSP which traverses any form of multipath not supporting termination of the entropy search at the EL label. For this to occur, the MPLS-TP ingress LSR must be aware of these links. This is the reason for requirement MP#4.

Requirement MP#4 can be supported using administrative attributes. Administrative attributes are defined in [<u>RFC3209</u>]. Some configuration is required to support this.

4. MPLS-TP as a Server Layer for MPLS

Carrying MPLS LSP which are larger than a component link over a MPLS-TP server layer requires that the large MPLS client layer LSP be accommodated by multiple MPLS-TP server layer LSPs. MPLS multipath can be used in the client layer MPLS.

Creating multiple MPLS-TP server layer LSP places a greater Incoming Label Map (ILM) scaling burden on the LSR. High bandwidth MPLS cores with a smaller amount of nodes have the greatest tendency to require LSP in excess of component links, therefore the reduction in number of nodes offsets the impact of increasing the number of server layer LSP in parallel. Today, only in cases where deployed LSR ILM are small would this be an issue.

The most significant disadvantage of MPLS-TP as a Server Layer for MPLS is that the use MPLS-TP server layer LSP reduces the efficiency of carrying the MPLS client layer. The service which provides by far the largest offered load in provider networks is Internet, for which the LSP capacity reservations are predictions of expected load. Many of these MPLS LSP may be smaller than component link capacity. Using MPLS-TP as a server layer results in bin packing problems for these smaller LSP. For those LSP that are larger than component link capacity, their capacity are not increments of convenient capacity increments such as 10Gb/s. Using MPLS-TP as an underlying server layer greatly reduces the ability of the client layer MPLS LSP to share capacity. For example, when one MPLS LSP is underutilizing its predicted capacity, the fixed allocation of MPLS-TP to component links may not allow another LSP to exceed its predicted capacity. Using MPLS-TP as a server layer may result in less efficient use of resources and may result in a less cost effective network.

No additional requirements beyond MPLS-TP as it is now currently defined are required to support MPLS-TP as a Server Layer for MPLS. It is therefore viable but has some undesirable characteristics discussed above.

5. Acknowledgements

Carlos Pignataro, Dave Allan, and Mach Chen provided valuable comments and suggestions. Carlos suggested that MPLS-TP requirements in RFC 5960 be explicitly referenced or guoted. An email conversation with Dave led to the inclusion of references and quotes from <u>RFC 6371</u> and <u>RFC 6374</u>. Mach made suggestions to improve clarity of the document.

6. Implementation Status

Note: this section is temporary and supports the experiment called for in draft-sheffer-running-code.

This is an informational document which describes usage of MPLS and MPLS-TP. No new protocol extensions or forwarding behavior are specified. Ethernet Link Aggregation and MPLS Link Bundling are widely implemented and deployed.

Entropy Label is not yet widely implemented and deployed, but both implementation and deployment are expected soon. At least a few existing high end commodity packet processing chips are capable of supporting Entropy Label. It would be helpful if a few LSR suppliers would state their intentions to support <u>RFC 6790</u> on the mpls mailing list.

Dynamic multipath (multipath load split adjustment in response to observed load) is referred to but not a requirement of the usage recommendations made in this document. Dynamic multipath has been implemented and deployed, however (afaik) the only core LSR vendor supporting dynamic multipath is no longer in the router business (Avici Systems). At least a few existing high end commodity packet processing chips are capable of supporting dynamic multipath.

7. IANA Considerations

This memo includes no request to IANA.

<u>8</u>. Security Considerations

This document specifies requirements with discussion of framework for solutions using existing MPLS and MPLS-TP mechanisms. The requirements and framework are related to the coexistence of MPLS/GMPLS (without MPLS-TP) when used over a packet network, MPLS-TP, and multipath. The combination of MPLS, MPLS-TP, and multipath does not introduce any new security threats. The security considerations for MPLS/GMPLS and for MPLS-TP are documented in [RFC5920] and [I-D.ietf-mpls-tp-security-framework].

9. References

9.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", <u>BCP 14</u>, <u>RFC 2119</u>, March 1997.
- [RFC5654] Niven-Jenkins, B., Brungard, D., Betts, M., Sprecher, N., and S. Ueno, "Requirements of an MPLS Transport Profile", <u>RFC 5654</u>, September 2009.
- [RFC5960] Frost, D., Bryant, S., and M. Bocci, "MPLS Transport Profile Data Plane Architecture", <u>RFC 5960</u>, August 2010.
- [RFC6371] Busi, I. and D. Allan, "Operations, Administration, and Maintenance Framework for MPLS-Based Transport Networks", <u>RFC 6371</u>, September 2011.
- [RFC6374] Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks", <u>RFC 6374</u>, September 2011.
- [RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", <u>RFC 6790</u>, November 2012.

<u>9.2</u>. Informative References

[I-D.ietf-mpls-tp-security-framework] Fang, L., Niven-Jenkins, B., Mansfield, S., and R. Graveman, "MPLS-TP Security Framework", <u>draft-ietf-mpls-tp-security-framework-05</u> (work in progress), October 2012.

[IEEE-802.1AX]

IEEE Standards Association, "IEEE Std 802.1AX-2008 IEEE Standard for Local and Metropolitan Area Networks - Link

Aggregation", 2006, <<u>http://standards.ieee.org/getieee802/</u> download/802.1AX-2008.pdf>.

[ITU-T.G.800]

ITU-T, "Unified functional architecture of transport networks", 2007, <<u>http://www.itu.int/rec/T-REC-G/</u> recommendation.asp?parent=T-REC-G.800>.

- [RFC2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, December 1998.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", <u>RFC 3209</u>, December 2001.
- [RFC4201] Kompella, K., Rekhter, Y., and L. Berger, "Link Bundling in MPLS Traffic Engineering (TE)", <u>RFC 4201</u>, October 2005.
- [RFC4206] Kompella, K. and Y. Rekhter, "Label Switched Paths (LSP) Hierarchy with Generalized Multi-Protocol Label Switching (GMPLS) Traffic Engineering (TE)", <u>RFC 4206</u>, October 2005.
- [RFC5286] Atlas, A. and A. Zinin, "Basic Specification for IP Fast Reroute: Loop-Free Alternates", <u>RFC 5286</u>, September 2008.
- [RFC5462] Andersson, L. and R. Asati, "Multiprotocol Label Switching (MPLS) Label Stack Entry: "EXP" Field Renamed to "Traffic Class" Field", <u>RFC 5462</u>, February 2009.
- [RFC5714] Shand, M. and S. Bryant, "IP Fast Reroute Framework", RFC 5714, January 2010.
- [RFC5920] Fang, L., "Security Framework for MPLS and GMPLS Networks", RFC 5920, July 2010.

Author's Address

Curtis Villamizar (editor) Outer Cape Cod Network Consulting

Email: curtis@occnc.com