RTGWG S. Ning
Internet-Draft Tata Communications
Intended status: Informational A. Malis
Expires: November 15, 2014 Consultant
D. McDysan
Verizon
L. Yong
Huawei USA
C. Villamizar
Outer Cape Cod Network Consulting
May 14, 2014

Advanced Multipath Use Cases and Design Considerations
draft-ietf-rtgwg-cl-use-cases-06

Abstract

Advanced Multipath is a formalization of multipath techniques currently in use in IP and MPLS networks and a set of extensions to existing multipath techniques.

This document provides a set of use cases and design considerations for Advanced Multipath. Existing practices are described. Use cases made possible through Advanced Multipath extensions are described.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

Advanced Multipath requirements are specified in [RFC7226]. An Advanced Multipath framework is defined in [I-D.ietf-rtgwg-cl-framework].

Multipath techniques have been widely used in IP networks for over two decades. The use of MPLS began more than a decade ago. Multipath has been widely used in IP/MPLS networks for over a decade with very little protocol support dedicated to effective use of multipath.

The state of the art in multipath prior to Advanced Multipath is documented in Appendix B.

Both Ethernet Link Aggregation [IEEE-802.1AX] and MPLS link bundling [RFC4201] have been widely used in today’s MPLS networks. Advanced Multipath differs in the following characteristics.

1. Advanced Multipath allows bundling of non-homogenous links together as a single logical link.

2. Advanced Multipath provides more information in the TE-LSDB and supports more explicit control over placement of LSP.

2. Assumptions

The supported services are, but not limited to, pseudowire (PW) based services ([RFC3985]), including Virtual Private Network (VPN) services, Internet traffic encapsulated by at least one MPLS label ([RFC3032]), and dynamically signaled MPLS ([RFC3209] or [RFC5036]) or MPLS-TP Label Switched Paths (LSPs) ([RFC5921]).

The MPLS LSPs supporting these services may be point-to-point, point-to-multipoint, or multipoint-to-multipoint. The MPLS LSPs may be signaled using RSVP-TE [RFC3209] or LDP [RFC5036]. With RSVP-TE, extensions to Interior Gateway Protocols (IGPs) may be used, specifically to OSPF-TE [RFC3630] or ISIS-TE [RFC5305].

The locations in a network where these requirements apply are a Label Edge Router (LER) or a Label Switch Router (LSR) as defined in [RFC3031].

The IP DSCP field [RFC2474] [RFC2475] cannot be used for flow identification since L3VPN requires Diffserv transparency (see RFC 4031 5.5.2 [RFC4031]), and in general network operators do not rely on the DSCP of Internet packets.

3. Terminology

Terminology defined in [RFC7226] and [RFC7190] is used in this document.

In addition, the following terms are used:

classic multipath:
: Classic multipath refers to the most common current practice in implementation and deployment of multipath (see Appendix B). The most common current practice when applied to MPLS traffic makes use of a hash on the MPLS label stack, and if IPv4 or IPv6 are

indicated under the label stack, makes use of the IP source and destination addresses [RFC4385] [RFC4928].

classic link bundling:
Classic link bundling refers to the use of [RFC4201] where the "all ones" component is not used. Where the "all ones" component is used, link bundling behaves as classic multipath does. Classic link bundling selects a single component link to carry all of the traffic for a given LSP.

Among the important distinctions between classic multipath or classic link bundling and Advanced Multipath are:

1. Classic multipath has no provision to retain packet order within any specific LSP. Classic link bundling retains packet order among any given LSP but as a result does a poor job of splitting load among components and therefore is rarely (if ever) deployed. Advanced Multipath allows per LSP control of load split characteristics.

2. Classic multipath and classic link bundling do not provide a means to put some LSP on component links with lower delay. Advanced Multipath does.

3. Classic multipath will provide a load balance for IP and LDP traffic. Classic link bundling will not. Neither classic multipath or classic link bundling will measure IP and LDP traffic and reduce the RSVP-TE advertised "Available Bandwidth" as a result of that measurement. Advanced Multipath better supports RSVP-TE used with significant traffic levels of native IP and native LDP.

4. Classic link bundling cannot support an LSP that is greater in capacity than any single component link. Classic multipath supports this capability but may reorder traffic on such an LSP. Advanced Multipath can retain order of an LSP that is carried within an LSP that is greater in capacity than any single component link if the contained LSP has such a requirement.

None of these techniques, classic multipath, classic link bundling, or Advanced Multipath, will reorder traffic among IP microflows. None of these techniques will reorder traffic among PW, if a PWE3 Control Word is used [RFC4385].

4. Multipath Foundation Use Cases

A simple multipath composed entirely of physical links is illustrated in Figure 1, where an multipath is configured between LSR1 and LSR2. This multipath has three component links. Individual component links in a multipath may be supported by different transport technologies such as SONET, OTN, Ethernet, etc. Even if the transport technology implementing the component links is identical, the characteristics (e.g., bandwidth, latency) of the component links may differ.

The multipath in Figure 1 may carry LSP traffic flows and control plane packets. Control plane packets may appear as IP packets or may be carried within a generic associated channel (G-Ach) [RFC5586]. A LSP may be established over the link by either RSVP-TE [RFC3209] or LDP [RFC5036] signaling protocols. All component links in a multipath are summarized in the same forwarding adjacency LSP (FA-LSP) routing advertisement [RFC3945]. The multipath is summarized as one TE-Link advertised into the IGP by the multipath end points (the LER if the multipath is MPLS based). This information is used in path computation when a full MPLS control plane is in use.

If Advanced Multipath techniques are used, then the individual component links or groups of component links may optionally be advertised into the IGP as sub-TLV of the multipath FA advertisement to indicate capacity available with various characteristics, such as a delay range.

```
               Management Plane
           Configuration and Measurement <------------+
                      ^                               |
                      |                               |
              +-------+-+                           +-+-------+
              |       | |                           | |       |
         CP Packets   V |                           | V     CP Packets
              |  V    | |     Component Link 1      | |    ^  |
              |  |    |=|===========================|=|    |  |
              |  +----| |     Component Link 2      | |----+  |
              |       |=|===========================|=|       |
    Aggregated LSPs   | |                           | |       |
             ~|~~~~~~>| |     Component Link 3      | |~~~~>~~|~~
              |       |=|===========================|=|       |
              |       | |                           | |       |
              | LSR1    |                           |    LSR2 |
              +---------+                           +---------+
                      !                               !
                      !                               !
                      !<-------- Multipath ---------->!
```

Figure 1: a multipath constructed with multiple physical links between two LSR

[RFC7226] specifies that component links may themselves be multipath. This is true for most implementations even prior to the Advanced Multipath work in [RFC7226]. For example, a component of a pre-Advanced Multipath MPLS Link Bundle or ISIS or OSPF ECMP could be an Ethernet LAG. In some implementations many other combinations or even arbitrary combinations could be supported. Figure 2 shows three three forms of component links which may be deployed in a network.

```
    +-------+                 1. Physical Link             +-------+
    |     |-|----------------------------------------------|-|     |
    |     | |                                              | |     |
    |     | |     +------+                     +------+    | |     |
    |     | |     | MPLS |    2. Logical Link  | MPLS |    | |     |
    |     |.|.... |......|.....................|......|....|.|     |
    |     | |-----| LSR3 |---------------------| LSR4 |----| |     |
    |     | |     +------+                     +------+    | |     |
    |     | |                                              | |     |
    |     | |                                              | |     |
    |     | |     +------+                     +------+    | |     |
    |     | |     |GMPLS |    3. Logical Link  |GMPLS |    | |     |
    |     |.|. ...|......|.....................|......|....|.|     |
    |     | |-----| LSR5 |---------------------| LSR6 |----| |     |
    |       |     +------+                     +------+    |       |
    | LSR1  |                                              |  LSR2 |
    +-------+                                              +-------+
          |<---------------- Multipath --------------------->|
```

Figure 2: Illustration of Various Component Link Types

The three forms of component link shown in Figure 2 are:

1. The first component link is configured with direct physical media plus a link layer protocol. This case also includes emulated physical links, for example using pseudowire emulation.

2. The second component link is a TE tunnel that traverses LSR3 and LSR4, where LSR3 and LSR4 are the nodes supporting MPLS, but supporting few or no GMPLS extensions.

3. The third component link is formed by lower layer network that has GMPLS enabled. In this case, LSR5 and LSR6 are not the nodes controlled by the MPLS but provide the connectivity for the component link.

A multipath forms one logical link between connected LSR (LSR1 and LSR2 in Figure 1 and Figure 2) and is used to carry aggregated traffic. Multipath relies on its component links to carry the traffic but must distribute or load balance the traffic. The endpoints of the multipath maps incoming traffic into the set of component links.

For example, LSR1 in Figure 1 distributes the set of traffic flows including control plane packets among the set of component links. LSR2 in Figure 1 receives the packets from its component links and sends them to MPLS forwarding engine with no attempt to reorder packets arriving on different component links. The traffic in the opposite direction, from LSR2 to LSR1, is distributed across the set of component links by the LSR2.

These three forms of component link are a limited set of very simple examples. Many other examples are possible. A component link may itself be a multipath. A segment of an LSP (single hop for that LSP) may be a multipath.

5. Advanced Multipath Use Cases

The following subsections provide some uses of the Advanced Multipath extensions. These are not the only uses, simply a set of examples.

5.1. Delay Sensitive Applications

Most applications benefit from lower delay. Some types of applications are far more sensitive than others. For example, real time bidirectional applications such as voice communication or two way video conferencing are far more sensitive to delay than unidirectional streaming audio or video. Non-interactive bulk transfer is almost insensitive to delay if a large enough TCP window is used.

Some applications are sensitive to delay but users of those applications are unwilling to pay extra to insure lower delay. For example, many SIP end users are willing to accept the delay offered to best effort services as long as call quality is good most of the time.

Other applications are sensitive to delay and willing to pay extra to insure lower delay. For example, financial trading applications are extremely sensitive to delay and with a lot at stake are willing to go to great lengths to reduce delay.

Among the requirements of Advanced Multipath are requirements to support non-homogeneous links. One solution in support of lower delay links is to advertise capacity available within configured ranges of delay within a given multipath and then support the ability to place an LSP only on component links that meeting that LSP’s delay requirements.

The Advanced Multipath requirements to accommodate delay sensitive applications are analogous to Diffserv requirements to accommodate applications requiring higher quality of service on the same infrastructure as applications with less demanding requirements. The ability to share capacity with less demanding applications, with best effort applications generally being the least demanding, can greatly reduce the cost of delivering service to the more demanding applications.

5.2. Large Volume of IP and LDP Traffic

IP and LDP do not support traffic engineering. Both make use of a shortest (lowest routing metric) path, with an option to use equal cost multipath (ECMP). Note that though ECMP is prohibited in LDP specifications, it is widely implemented. Where implemented for LDP, ECMP is generally disabled by default for standards compliance, but often enabled in LDP deployments.

Without traffic engineering capability, there must be sufficient capacity to accommodate the IP and LDP traffic. If not, persistent queuing delay and loss will occur. Unlike RSVP-TE, a subset of traffic cannot be routed using constraint based routing to avoid a congested portion of an infrastructure.

In existing networks which accommodate IP and/or LDP with RSVP-TE, either the IP and LDP can be carried over RSVP-TE, or where the traffic contribution of IP and LDP is small, IP and LDP can be carried native and the effect on RSVP-TE can be ignored. Ignoring the traffic contribution of IP is valid on high capacity networks where a very low volume of native IP is used primarily for control and network management and customer IP is carried within RSVP-TE.

Where it is desirable to carry native IP and/or LDP and IP and/or LDP traffic volumes are not negligible, RSVP-TE needs improvement. An enhancement offered by Advanced Multipath is an ability to measure the IP and LDP, filter the measurements, and reduce the capacity available to RSVP-TE to avoid congestion. The treatment given to the IP or LDP traffic is similar to the treatment when using the "auto-bandwidth" feature in some RSVP-TE implementations on that same traffic, and giving a higher priority (numerically lower setup priority and holding priority value) to the "auto-bandwidth" LSP. The difference is that the measurement is made at each hop and the reduction in advertised bandwidth is made more directly.

5.3. Multipath and Packet Ordering

A strong motivation for multipath is the need to provide LSP capacity in IP backbones that exceeds the capacity of single wavelengths provided by transport equipment and exceeds the practical capacity limits achievable through inverse multiplexing. Appendix C describes characteristics and limitations of transport systems today. Section 3 defines the terms "classic multipath" and "classic link bundling" used in this section.

For purpose of discussion, consider two very large cities, city A and city Z. For example, in the US high traffic cities might be New York and Los Angeles and in Europe high traffic cities might be London and Amsterdam. Two other high volume cities, city B and city Y may share common provider core network infrastructure. Using the same examples, the city B and Y may Washington DC and San Francisco or Paris and Stockholm. In the US, the common infrastructure may span Denver, Chicago, Detroit, and Cleveland. Other major traffic contributors on either US coast include Boston, northern Virginia on the east coast, and Seattle, and San Diego on the west coast. The capacity of IP/MPLS links within the shared infrastructure, for example city to city links in the Denver, Chicago, Detroit, and Cleveland path in the US example, have capacities for most of the 2000s decade that greatly exceeded single circuits available in transport networks.

For a case with four large traffic sources on either side of the shared infrastructure, up to sixteen core city to core city traffic flows in excess of transport circuit capacity may be accommodated on the shared infrastructure.

Today the most common IP/MPLS core network design makes use of very large links which consist of many smaller component links, but use classic multipath techniques. A component link typically corresponds to the largest circuit that the transport system is capable of providing (or the largest cost effective circuit). IP source and destination address hashing is used to distribute flows across the set of component links as described in Appendix B.3.

Classic multipath can handle large LSP up to the total capacity of the multipath (within limits, see Appendix B.2). A disadvantage of classic multipath is the reordering among traffic within a given core city to core city LSP. While there is no reordering within any microflow and therefore no customer visible issue, MPLS-TP cannot be used across an infrastructure where classic multipath is in use, except within pseudowires.

Capacity issues force the use of classic multipath today. Classic multipath excludes a direct use of MPLS-TP. The desire for OAM, offered by MPLS-TP, is in conflict with the use of classic multipath. There are a number of alternatives that satisfy both requirements. Some alternatives are described below.

MPLS-TP in network edges only

A simple approach which requires no change to the core is to disallow MPLS-TP across the core unless carried within a pseudowire (PW). MPLS-TP may be used within edge domains where classic multipath is not used. PW may be signaled end to end using single segment PW (SS-PW), or stitched across domains using multisegment PW (MS-PW). The PW and anything carried within the PW may use OAM as long as fat-PW [RFC6391] load splitting is not used by the PW.

Advanced Multipath at core LSP ingress/egress

The interior of the core network may use classic link bundling, with the limitation that no LSP can exceed the capacity of a single circuit. Larger non-MPLS-TP LSP can be configured using multiple ingress to egress component MPLS-TP LSP. This can be accomplished using existing IP source and destination address hashing configured at LSP ingress and egress. Each component LSP, if constrained to be no larger than the capacity of a single circuit, can make use of MPLS-TP and offer OAM for all top level LSP across the core.

MPLS-TP as a MPLS client

A third approach involves making use of Entropy Labels [RFC6790] on all MPLS-TP LSP such that the entire MPLS-TP LSP is treated as a microflow by midpoint LSR, even if further encapsulated in very large server layer MPLS LSP.

The above list of alternatives allow packet ordering within an LSP to be maintained in some circumstances and allow very large LSP capacities. Each of these alternatives are discussed further in the following subsections.

5.3.1. MPLS-TP in network edges only

Classic MPLS link bundling is defined in [RFC4201] and has existed since early in the 2000s decade. Classic MPLS link bundling place any given LSP entirely on a single component link. Classic MPLS link bundling is not in widespread use as the means to accommodate large link capacities in core networks due to the simplicity and better multiplexing gain, and therefore lower network cost of classic multipath.

If MPLS-TP OAM capability in the IP/MPLS network core LSP is not required, then there is no need to change existing network designs which use classic multipath and both label stack and IP source and destination address based hashing as a basis for load splitting.

If MPLS-TP is needed for a subset of LSP, then those LSP can be carried within pseudowires. The pseudowires adds a thin layer of encapsulation and therefore a small overhead. If only a subset of

LSP need MPLS-TP OAM, then some LSP must make use of the pseudowires and other LSP avoid them. A straightforward way to accomplish this is with administrative attributes [RFC3209].

5.3.2. Multipath at core LSP ingress/egress

Multipath can be configured for large LSP that are made of smaller MPLS-TP component LSP. Some implementations already support this capability, though until Advanced Multipath no IETF document required it. This approach is capable of supporting MPLS-TP OAM over the entire set of component link LSP and therefore the entire set of top level LSP traversing the core.

There are two primary disadvantage of this approach. One is the number of top level LSP traversing the core can be dramatically increased. The other disadvantage is the loss of multiplexing gain that results from use of classic link bundling within the interior of the core network.

If component LSP use MPLS-TP, then no component LSP can exceed the capacity of a single circuit. For a given multipath LSP there can either be a number of equal capacity component LSP or some number of full capacity component links plus one LSP carrying the excess. For example, a 350 Gb/s multipath LSP over a 100 Gb/s infrastructure may use five 70 Gb/s component LSP or three 100 Gb/s LSP plus one 50 Gb/s LSP. Classic MPLS link bundling is needed to support MPLS-TP and suffers from a bin packing problem even if LSP traffic is completely predictable, which it never is in practice.

The common means of setting very large LSP link bandwidth parameters uses long term statistical measures. For example, at one time many providers based their LSP bandwidth parameters on the 95th percentile of carried traffic as measured over the prior one week period. It is common to add 10-30% to the 95th percentile value measured over the prior week and adjust bandwidth parameters of LSP weekly. It is also possible to measure traffic flow at the LSR and adjust bandwidth parameters somewhat more dynamically. This is less common in deployments and where deployed, makes use of filtering to track very long term trends in traffic levels. In either case, short term variation of traffic levels relative to signaled LSP capacity are common. Allowing a large over allocation of LSP bandwidth parameters (ie: adding 30% or more) avoids over utilization of any given LSP, but increases unused network capacity and increases network cost. Allowing a small over allocation of LSP bandwidth parameters (ie: 10-20% or less) results in both underutilization and over utilization but statistically results in a total utilization within the core that is under capacity most or all of the time.

The classic multipath solution accommodates the situation in which some very large LSP are under utilizing their signaled capacity and others are over utilizing their capacity with the need for far less unused network capacity to accommodate variation in actual traffic levels. If the actual traffic levels of LSP can be described by a probability distribution, the variation of the sum of LSP is less than the variation of any given LSP for all but a constant traffic level (where the variation of the sum and the variation of the components are both zero).

Splitting very large LSP at the ingress and carrying those large LSP within smaller MPLS-TP component LSP and then using classic link bundling to carry the MPLS-TP LSP is a viable approach. However this approach loses the statistical gain discussed in the prior paragraphs. Losing this statistical gain drives up network costs necessary to acheive the same very low probability of only mild congestion that is expected of provider networks.

There are two situations which can motivate the use of this approach. This design is favored if the provider values MPLS-TP OAM across the core more than efficiency (or is unaware of the efficiency issue). This design can also make sense if transport equipment or very low cost core LSR are available which support only classic link bundling and regardless of loss of multiplexing gain, are more cost effective at carrying transit traffic than using equipment which supports IP source and destination address hashing.

5.3.3. MPLS-TP as a MPLS client

Accommodating MPLS-TP as a MPLS client requires the small change to forwarding behavior necessary to support [RFC6790] and is therefore most applicable to major network overbuilds or new deployments. This approach is described in [RFC7190] and makes use of Entropy Labels [RFC6790] to prevent reordering of MPLS-TP LSP or any other LSP which requires that its traffic not be reordered for OAM or other reasons.

The advantage of this approach is an ability to accommodate MPLS-TP as a client LSP but retain the high multiplexing gain and therefore efficiency and low network cost of a pure MPLS deployment. The disadvantage is the need for a small change in forwarding to support [RFC6790].

6. IANA Considerations

This memo includes no request to IANA.

7. Security Considerations

This document is a use cases document. Existing protocols are referenced such as MPLS. Existing techniques such as MPLS link bundling and multipath techniques are referenced. These protocols and techniques are documented elsewhere and contain security considerations which are unchanged by this document.

This document also describes use cases for multipath and Advanced Multipath. Advanced Multipath requirements are defined in [RFC7226]. [I-D.ietf-rtgwg-cl-framework] defines a framework for Advanced Multipath. Advanced Multipath bears many similarities to MPLS link bundling and multipath techniques used with MPLS. Additional security considerations, if any, beyond those already identified for MPLS, MPLS link bundling and multipath techniques, will be documented in the framework document if specific to the overall framework of Advanced Multipath, or in protocol extensions if specific to a given protocol extension defined later to support Advanced Multipath.

8. Acknowledgments

In the interest of full disclosure of affiliation and in the interest of acknowledging sponsorship, past affiliations of authors are noted. Much of the work done by Ning So occurred while Ning was at Verizon. Much of the work done by Curtis Villamizar occurred while at Infinera. Much of the work done by Andy Malis occurred while Andy was at Verizon.

9. Informative References

[I-D.ietf-rtgwg-cl-framework]
Ning, S., McDysan, D., Osborne, E., Yong, L., and C. Villamizar, "Advanced Multipath Framework in MPLS", draft-ietf-rtgwg-cl-framework-04 (work in progress), July 2013.

[IEEE-802.1AX]
IEEE Standards Association, "IEEE Std 802.1AX-2008 IEEE Standard for Local and Metropolitan Area Networks - Link Aggregation", 2006, <http://standards.ieee.org/getieee802/download/802.1AX-2008.pdf>.

[ITU-T.G.694.2]
ITU-T, "Spectral grids for WDM applications: CWDM wavelength grid", 2003, <http://www.itu.int/rec/T-REC-G.694.2-200312-I>.

[RFC1717] Sklower, K., Lloyd, B., McGregor, G., and D. Carr, "The PPP Multilink Protocol (MP)", RFC 1717, November 1994.

[RFC2474] Nichols, K., Blake, S., Baker, F., and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474, December 1998.

[RFC2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, December 1998.

[RFC2597] Heinanen, J., Baker, F., Weiss, W., and J. Wroclawski, "Assured Forwarding PHB Group", RFC 2597, June 1999.

[RFC2615] Malis, A. and W. Simpson, "PPP over SONET/SDH", RFC 2615, June 1999.

[RFC2991] Thaler, D. and C. Hopps, "Multipath Issues in Unicast and Multicast Next-Hop Selection", RFC 2991, November 2000.

[RFC2992] Hopps, C., "Analysis of an Equal-Cost Multi-Path Algorithm", RFC 2992, November 2000.

[RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.

[RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, January 2001.

[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.

[RFC3260] Grossman, D., "New Terminology and Clarifications for Diffserv", RFC 3260, April 2002.

[RFC3270] Le Faucheur, F., Wu, L., Davie, B., Davari, S., Vaananen, P., Krishnan, R., Cheval, P., and J. Heinanen, "Multi-Protocol Label Switching (MPLS) Support of Differentiated Services", RFC 3270, May 2002.

[RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, September 2003.

[RFC3809] Nagarajan, A., "Generic Requirements for Provider Provisioned Virtual Private Networks (PPVPN)", RFC 3809, June 2004.

[RFC3945] Mannie, E., "Generalized Multi-Protocol Label Switching (GMPLS) Architecture", RFC 3945, October 2004.

[RFC3985] Bryant, S. and P. Pate, "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, March 2005.

[RFC4031] Carugi, M. and D. McDysan, "Service Requirements for Layer 3 Provider Provisioned Virtual Private Networks (PPVPNs)", RFC 4031, April 2005.

[RFC4124] Le Faucheur, F., "Protocol Extensions for Support of Diffserv-aware MPLS Traffic Engineering", RFC 4124, June 2005.

[RFC4201] Kompella, K., Rekhter, Y., and L. Berger, "Link Bundling in MPLS Traffic Engineering (TE)", RFC 4201, October 2005.

[RFC4385] Bryant, S., Swallow, G., Martini, L., and D. McPherson, "Pseudowire Emulation Edge-to-Edge (PWE3) Control Word for Use over an MPLS PSN", RFC 4385, February 2006.

[RFC4928] Swallow, G., Bryant, S., and L. Andersson, "Avoiding Equal Cost Multipath Treatment in MPLS Networks", BCP 128, RFC 4928, June 2007.

[RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.

[RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, October 2008.

[RFC5586] Bocci, M., Vigoureux, M., and S. Bryant, "MPLS Generic Associated Channel", RFC 5586, June 2009.

[RFC5921] Bocci, M., Bryant, S., Frost, D., Levrau, L., and L. Berger, "A Framework for MPLS in Transport Networks", RFC 5921, July 2010.

[RFC6391] Bryant, S., Filsfils, C., Drafz, U., Kompella, V., Regan, J., and S. Amante, "Flow-Aware Transport of Pseudowires over an MPLS Packet Switched Network", RFC 6391, November 2011.

[RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, November 2012.

[RFC7190] Villamizar, C., "Use of Multipath with MPLS and MPLS Transport Profile (MPLS-TP)", RFC 7190, March 2014.

[RFC7226] Villamizar, C., McDysan, D., Ning, S., Malis, A., and L. Yong, "Requirements for Advanced Multipath in MPLS Networks", RFC 7226, May 2014.

Appendix A. Network Operator Practices and Protocol Usage

Often, network operators have a contractual Service Level Agreement (SLA) with customers for services that are comprised of numerical values for performance measures, principally availability, latency, delay variation. Additionally, network operators may have performance objectives for internal use by the operator. See RFC3809, Section 4.9 [RFC3809] for examples of the form of such SLA and performance objective specifications. In this document we use the term Performance Objective as defined in [RFC7226]. Applications and acceptable user experience have an important relationship to these performance parameters.

Consider latency as an example. In some cases, minimizing latency relates directly to the best customer experience (for example, in interactive applications closer is faster). In other cases, user experience is relatively insensitive to latency, up to a specific limit at which point user perception of quality degrades significantly (e.g., interactive human voice and multimedia conferencing). A number of Performance Objectives have a bound on point-to-point latency and as long as this bound is met the Performance Objective is met; decreasing the latency is not necessary. In some Performance Objectives, if the specified latency is not met, the user considers the service as unavailable. An unprotected LSP can be manually provisioned on a set of links to meet this type of Performance Objective, but this lowers availability since an alternate route that meets the latency Performance Objective cannot be determined.

Historically, when an IP/MPLS network was operated over a lower layer circuit switched network (e.g., SONET rings), a change in latency caused by the lower layer network (e.g., due to a maintenance action or failure) was not known to the MPLS network. This resulted in latency affecting end user experience, sometimes violating Performance Objectives or resulting in user complaints.

A response to this problem was to provision IP/MPLS networks over unprotected circuits and set the metric and/or TE-metric proportional to latency. This resulted in traffic being directed over the least latency path, even if this was not needed to meet an Performance Objective or meet user experience objectives. This results in reduced flexibility and increased cost for network operators. Some providers perfer to use lower layer networks to provide restoration and grooming, but the inability to communicate performance parameters, in particular latency, from the lower layer network to the higher layer network is an important problem to be solved before this can be done.

Latency Performance Objectives for point-to-point services are often tied closely to geographic locations, while latency for multipoint services may be based upon a worst case within a region.

The time frames for restoration (i.e., as implemented by predetermined protection, convergence of routing protocols and/or signaling) for services range from on the order of 100 ms or less (e.g., for VPWS to emulate classical SDH/SONET protection switching), to several minutes (e.g., to allow BGP to reconverge for L3VPN) and may differ among the set of customers within a single service.

The presence of only three Traffic Class (TC) bits (previously known as EXP bits) in the MPLS shim header is limiting when a network operator needs to support QoS classes for multiple services (e.g., L2VPN VPWS, VPLS, L3VPN and Internet), each of which has a set of QoS classes that need to be supported and where the operator prefers to use only E-LSP [RFC3270]. In some cases one bit is used to indicate conformance to some ingress traffic classification, leaving only two bits for indicating the service QoS classes. One approach that has been taken is to aggregate these QoS classes into similar sets on LER-LSR and LSR-LSR links and continue to use only E-LSP. Another approach is to use L-LSP as defined in [RFC3270] or use the Class-Type as defined in [RFC4124] to support up to eight mappings of TC into Per-Hop Behavior (PHB).

The IP DSCP cannot be used for flow identification. The use of IP DSCP for flow identification is incompatible with Assured Forwarding services [RFC2597] or any other service which may use more than one DSCP code point to carry traffic for a given microflow. In general network operators do not rely on the DSCP of Internet packets in core networks but must preserve DSCP values for use closer to network edges.

A label is pushed onto Internet packets when they are carried along with L2VPN or L3VPN packets on the same link or lower layer network provides a mean to distinguish between the QoS class for these packets.

Operating an MPLS-TE network involves a different paradigm from operating an IGP metric-based LDP signaled MPLS network. The multipoint-to-point LDP signaled MPLS LSPs occur automatically, and balancing across parallel links occurs if the IGP metrics are set "equally" (with equality a locally definable relation) and if ECMP is enabled for LDP, which network operators generally do in large networks.

Traffic is typically comprised of large (some very large) flows and a much larger number of small flows. In some cases, separate LSPs are established for very large flows. Very large microflows can occur even if the IP header information is inspected by a LSR. For example an IPsec tunnel that carries a large amount of traffic must be carried as a single large flow. An important example of large flows is that of a L2VPN or L3VPN customer who has an access line bandwidth comparable to a client-client component link bandwidth -- there could be flows that are on the order of the access line bandwidth.

Appendix B. Existing Multipath Standards and Techniques

Today the requirement to handle large aggregations of traffic, much larger than a single component link, 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, LDP, or even BGP, and equal cost LSP, as described in Appendix B.4. Various multipath techniques have strengths and weaknesses.

Existing multipath techniques solve the problem of large aggregations of traffic, without addressing the other requirements outlined in this document, particularly those described in Section 5.

B.1. Common Multpath Load Spliting Techniques

Identical load balancing techniques are used for multipath both over parallel links and over diverse paths.

Large aggregates of IP traffic do not provide explicit signaling to indicate the expected traffic loads. Large aggregates of MPLS traffic are carried in MPLS tunnels supported by MPLS LSP. LSP which are signaled using RSVP-TE extensions do provide explicit signaling which includes the expected traffic load for the aggregate. LSP which are signaled using LDP do not provide an expected traffic load.

MPLS LSP may contain other MPLS LSP arranged hierarchically. When an MPLS LSR serves as a midpoint LSR in an LSP carrying client LSP as payload, there is no signaling associated with these client LSP. Therefore even when using RSVP-TE signaling there may be insufficient information provided by signaling to adequately distribute load based solely on signaling.

Generally a set of label stack entries that is unique across the ordered set of label numbers in the label stack can safely be assumed to contain a group of flows. The reordering of traffic can therefore be considered to be acceptable unless reordering occurs within traffic containing a common unique set of label stack entries. Existing load splitting techniques take advantage of this property in addition to looking beyond the bottom of the label stack and determining if the payload is IPv4 or IPv6 to load balance traffic accordingly.

MPLS-TP OAM violates the assumption that it is safe to reorder traffic within an LSP. If MPLS-TP OAM is to be accommodated, then existing multipath techniques must be modified. [RFC6790] and [RFC7190] provide a solution but require a small forwarding change.

For example, a large aggregate of IP traffic may be subdivided into a large number of groups of flows using a hash on the IP source and destination addresses. This is as described in [RFC2475] and clarified in [RFC3260]. For MPLS traffic carrying IP, a similar hash can be performed on the set of labels in the label stack. These techniques are both examples of means to subdivide traffic into groups of flows for the purpose of load balancing traffic across aggregated link capacity. The means of identifying a group of flows should not be confused with the definition of a flow.

Discussion of whether a hash based approach provides a sufficiently even load balance using any particular hashing algorithm or method of distributing traffic across a set of component links is outside of the scope of this document.

The current load balancing techniques are referenced in [RFC4385] and [RFC4928]. The use of three hash based approaches are described in [RFC2991] and [RFC2992]. A mechanism to identify flows within PW is described in [RFC6391]. The use of hash based approaches is mentioned as an example of an existing set of techniques to distribute traffic over a set of component links. Other techniques are not precluded.

B.2. Static and Dynamic Load Balancing Multipath

Static multipath generally relies on the mathematical probability that given a very large number of small microflows, these microflows will tend to be distributed evenly across a hash space. Early very static multipath implementations assumed that all component links are of equal capacity and perform a modulo operation across the hashed value. An alternate static multipath technique uses a table generally with a power of two size, and distributes the table entries proportionally among component links according to the capacity of each component link.

Static load balancing works well if there are a very large number of small microflows (i.e., microflow rate is much less than component link capacity). However, the case where there are even a few large microflows is not handled well by static load balancing.

A dynamic load balancing multipath technique is one where the traffic bound to each component link is measured and the load split is adjusted accordingly. As long as the adjustment is done within a single network element, then no protocol extensions are required and there are no interoperability issues.

Note that if the load balancing algorithm and/or its parameters is adjusted, then packets in some flows may be briefly delivered out of sequence, however in practice such adjustments can be made very infrequent.

B.3. Traffic Split over Parallel Links

The load splitting techniques defined in Appendix B.1 and Appendix B.2 are both used in splitting traffic over parallel links between the same pair of nodes. The best known technique, though far from being the first, is Ethernet Link Aggregation [IEEE-802.1AX]. This same technique had been applied much earlier using OSPF or ISIS Equal Cost MultiPath (ECMP) over parallel links between the same nodes. Multilink PPP [RFC1717] uses a technique that provides inverse multiplexing, however a number of vendors had provided proprietary extensions to PPP over SONET/SDH [RFC2615] that predated Ethernet Link Aggregation but are no longer used.

Link bundling [RFC4201] provides yet another means of handling parallel LSP. RFC4201 explicitly allow a special value of all ones to indicate a split across all members of the bundle. This "all ones" component link is signaled in the MPLS RESV to indicate that the link bundle is making use of classic multipath techniques.

B.4. Traffic Split over Multiple Paths

OSPF or ISIS Equal Cost MultiPath (ECMP) is a well known form of traffic split over multiple paths that may traverse intermediate nodes. ECMP is often incorrectly equated to only this case, and multipath over multiple diverse paths is often incorrectly equated to ECMP.

Many implementations are able to create more than one LSP between a pair of nodes, where these LSP are routed diversely to better make use of available capacity. The load on these LSP can be distributed proportionally to the reserved bandwidth of the LSP. These multiple LSP may be advertised as a single PSC FA and any LSP making use of the FA may be split over these multiple LSP.

Link bundling [RFC4201] component links may themselves be LSP. When this technique is used, any LSP which specifies the link bundle may be split across the multiple paths of the component LSP that comprise the bundle.

Appendix C. Characteristics of Transport in Core Networks

The characteristics of primary interest are the capacity of a single circuit and the use of wave division multiplexing (WDM) to provide a large number of parallel circuits.

Wave division multiplexing (WDM) supports multiple independent channels (independent ignoring crosstalk noise) at slightly different wavelengths of light, multiplexed onto a single fiber. Typical in the early 2000s was 40 wavelengths of 10 Gb/s capacity per wavelength. These wavelengths are in the C-band range, which is about 1530-1565 nm, though some work has been done using the L-band 1565-1625 nm.

The C-band has been carved up using a 100 GHz spacing from 191.7 THz to 196.1 THz by [ITU-T.G.694.2]. This yields 44 channels. If the outermost channels are not used, due to poorer transmission characteristics, then typically 40 are used. For practical reasons, a 50 GhZ or 25 GHz spacing is used by more recent equipment, yielding. 80 or 160 channels in practice.

The early optical modulation techniques used within a single channel yielded 2.5Gb/s and 10 Gb/s capacity per channel. As modulation techniques have improved 40 Gb/s and 100 Gb/s per channel have been achieved.

The 40 channels of 10 Gb/s common in the mid 2000s yields a total of 400 Gb/s. Tighter spacing and better modulations are yielding up to 8 Tb/s or more in more recent systems.

Over the optical modulation is an electrical encoding. In the 1990s this was typically Synchronous Optical Networking (SONET) or Synchronous Digital Hierarchy (SDH), with a maximum defined circuit capacity of 40 Gb/s (OC-768), though the 10 Gb/s OC-192 is more common. More recently the low level electrical encoding has been Optical Transport Network (OTN) defined by ITU-T. OTN currently defines circuit capacities up to a nominal 100 Gb/s (ODU4). Both SONET/SDH and OTN make use of time division multiplexing (TDM) where the a higher capacity circuit such as a 100 Gb/s ODU4 in OTN may be subdivided into lower fixed capacity circuits such as ten 10 Gb/s ODU2.

In the 1990s, all IP and later IP/MPLS networks either used a fraction of maximum circuit capacity, or at most the full circuit capacity toward the end of the decade, when full circuit capacity was 2.5 Gb/s or 10 Gb/s. Beyond 2000, the TDM circuit multiplexing capability of SONET/SDH or OTN was rarely used.

Early in the 2000s both transport equipment and core LSR offered 40 Gb/s SONET OC-768. However 10 Gb/s transport equipment was predominantly deployed throughout the decade, partially because LSR 10GbE ports were far more cost effective than either OC-192 or OC-768 and 10GbE became practical in the second half of the decade.

Entering the 2010 decade, LSR 40GbE and 100GbE are expected to become widely available and cost effective. Slightly preceding this transport equipment making use of 40 Gb/s and 100 Gb/s modulations are becoming available. This transport equipment is capable or carrying 40 Gb/s ODU3 and 100 Gb/s ODU4 circuits.

Early in the 2000s decade IP/MPLS core networks were making use of single 10 Gb/s circuits. Capacity grew quickly in the first half of the decade but more IP/MPLS core networks had only a small number of IP/MPLS links requiring 4-8 parallel 10 Gb/s circuits. However, the use of multipath was necessary, was deemed the simplest and most cost effective alternative, and became thoroughly entrenched. By the end of the 2000s decade nearly all major IP/MPLS core service provider networks and a few content provider networks had IP/MPLS links which exceeded 100 Gb/s, long before 40GbE was available and 40 Gb/s transport in widespread use.

It is less clear when IP/MPLS LSP exceeded 10 Gb/s, 40 Gb/s, and 100 Gb/s. By 2010, many service providers have LSP in excess of 100 Gb/s, but few are willing to disclose how many LSP have reached this capacity.

By 2012 40GbE and 100GbE LSR products had become available, but were mostly still being evaluated or in trial use by service providers and contect providers. The cost of components required to deliver 100GbE products remained high making these products less cost effective. This is expected to change within years.

The important point is that IP/MPLS core network links have long ago exceeded 100 Gb/s and some may have already exceeded a Tb/s and a small number of IP/MPLS LSP exceed 100 Gb/s. By the time 100 Gb/s circuits are widely deployed, many IP/MPLS core network links are likely to exceed 1 Tb/s and many IP/MPLS LSP capacities are likely to exceed 100 Gb/s. The growth in service provider traffic has consistently outpaced growth in DWDM channel capacities and the growth in capacity of single interfaces and is expected to continue to do so. Therefore multipath techniques are likely here to stay.

Authors' Addresses

So Ning
Tata Communications

Email: ning.so@tatacommunications.com

Andrew Malis
Consultant

Email: agmalis@gmail.com

Dave McDysan
Verizon
22001 Loudoun County PKWY
Ashburn, VA 20147
USA

Email: dave.mcdysan@verizon.com

Lucy Yong
Huawei USA
5340 Legacy Dr.
Plano, TX 75025
USA

Phone: +1 469-277-5837
Email: lucy.yong@huawei.com

Curtis Villamizar
Outer Cape Cod Network Consulting

Email: curtis@occnc.com