Internet Engineering Task Force (IETF)
Request for Comments: 7226
Category: Informational
ISSN: 2070-1721

C. Villamizar, Ed.
OCCNC, LLC
D. McDysan, Ed.
Verizon
S. Ning
Tata Communications
A. Malis
Huawei
L. Yong
Huawei USA

May 2014

Requirements for Advanced Multipath in MPLS Networks

Abstract

This document provides a set of requirements for Advanced Multipath in MPLS networks.

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.

Status of This Memo

This document is not an Internet Standards Track specification; it is published for informational purposes.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Not all documents approved by the IESG are a candidate for any level of Internet Standard; see Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc7226.

Table of Contents

1. Introduction

There is often a need to provide large aggregates of bandwidth that are best provided using parallel links between routers or carrying traffic over multiple MPLS Label Switched Paths (LSPs). In core networks, there is often no alternative since the aggregate capacities of core networks today far exceed the capacity of a single physical link or a single packet-processing element.

The presence of parallel links, with each link potentially comprised of multiple layers, has resulted in additional requirements. Certain services may benefit from being restricted to a subset of the component links or a specific component link, where component link characteristics, such as latency, differ. Certain services require that an LSP be treated as atomic and avoid reordering. Other services will continue to require only that reordering not occur within a flow as is current practice.

Numerous forms of multipath exist today, including MPLS Link Bundling [RFC4201], Ethernet Link Aggregation [IEEE-802.1AX], and various forms of Equal Cost Multipath (ECMP) such as for OSPF ECMP, IS-IS ECMP, and BGP ECMP. Refer to the appendices in [USE-CASES] for a description of existing techniques and a set of references.

The purpose of this document is to clearly enumerate a set of requirements related to the protocols and mechanisms that provide MPLS-based Advanced Multipath. The intent is to first provide a set of functional requirements, in Section 3, that are as independent as possible of protocol specifications. A set of general protocol requirements are defined in Section 4. A set of network management requirements are defined in Section 5.

1.1. Requirements Language

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].

Any statement that requires the solution to support some new functionality through use of [RFC2119] keywords should be interpreted as follows. The implementation either MUST or SHOULD support the new functionality, depending on the use of either MUST or SHOULD in the requirements statement. The implementation SHOULD, in most or all cases, allow any new functionality to be individually enabled or disabled through configuration. A service provider or other deployment MAY enable or disable any feature in their network, subject to implementation limitations on sets of features that can be disabled.

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,

   b. specific paths across a network to a destination node, or

   c. links or paths to an intermediate node used to reach a common destination.

The paths need not have equal capacity. The paths may or may not have equal cost in a routing protocol.

Advanced Multipath
: Advanced Multipath is a formalization of multipath techniques that meets the requirements defined in this document. A key capability of Advanced Multipath is the support of non-homogeneous component links.

Advanced Multipath Group (AMG)
: An AMG is a collection of component links where Advanced Multipath techniques are applied.

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

Inverse Multiplexing
: Inverse multiplexing is another method of sending traffic over multiple links. 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-T 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 forms of multipath. The IEEE uses the term "member" rather than "component link" in Ethernet Link Aggregation [IEEE-802.1AX].

Client Layer
: A client layer is the layer immediately above a server layer.

Server Layer
: A server layer is the layer immediately below a client layer.

Higher Layers
: Relative to a particular layer, a client layer and any layer above that is considered a higher layer. Upper layer is synonymous with higher layer.

Lower Layers
: Relative to a particular layer, a server layer and any layer below that is considered a lower layer.

Client LSP
: A client LSP is an LSP that has been set up over one or more lower layers. In the context of this discussion, one type of client LSP is an LSP that has been set up over an AMG.

Flow
: A sequence of packets that should be transferred in order on one component link of a multipath.

Flow Identification
: The label stack and other information that uniquely identifies a flow. Other information in flow identification may include an IP header, pseudowire (PW) control word, Ethernet Media Access Control (MAC) address, etc. Note that a client LSP may contain one or more flows, or a client LSP may be equivalent to a flow. Flow identification is used to locally select a component link or a path through the network toward the destination.

Load Balance
: 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 achieve these objectives than others.

Performance Objective
: Numerical values for performance measures: principally availability, latency, and delay variation. Performance objectives may be related to Service Level Agreements (SLAs) as defined in [RFC2475] or may be strictly internal. Performance objectives may span links from edge to edge or from end to end. Performance objectives may span one provider or multiple providers.

A component link may be a point-to-point physical link (where a "physical link" includes one or more link layers, plus a physical layer) or a logical link that preserves ordering in the steady state. A component link may have transient out-of-order events, but such events must not exceed the network's performance objectives. For example, a component link may be comprised of any supportable combination of link layers over a physical layer or over logical sub-layers -- including those providing physical-layer emulation -- or over MPLS server-layer LSP.

The ingress and egress of a multipath may be midpoint LSRs with respect to a given client LSP. A midpoint LSR does not participate in the signaling of any clients of the client LSP. Therefore, in general, multipath endpoints cannot determine requirements of clients of a client LSP through participation in the signaling of the clients of the client LSP.

This document makes no statement on whether Advanced Multipath is itself a layer or whether an instance of AMG is itself a layer. This is to avoid engaging in long and pointless discussions about what constitutes a proper layer.

The term "Advanced Multipath" is intended to be used within the context described in this document and related documents, for example, [USE-CASES] and [FRAMEWORK]. Other Advanced Multipath techniques may arise in the future. If the capabilities defined in this document become commonplace, they would no longer be considered "advanced". Use of the term "advanced multipath" outside this document, if referring to the term as defined here, should indicate Advanced Multipath as defined by this document, citing the current document name. If using another definition of "advanced multipath", documents may optionally clarify that they are not using the term "advanced multipath" as defined by this document if clarification is deemed helpful.

3. Functional Requirements

The functional requirements in this section are grouped in subsections, starting with the highest priority.

3.1. Availability, Stability, and Transient Response

In addition to maintaining stability, limiting the period of unavailability in response to failures or transient events is extremely important.

FR#1 The transient period between some service disrupting event and the convergence of the routing and/or signaling protocols MUST occur within a time frame specified by performance objective values.

FR#2 An AMG MAY be announced in conjunction with detailed parameters about its component links, such as bandwidth and latency. The AMG SHALL behave as a single IGP adjacency.

FR#3 The solution SHALL provide a means to summarize some routing advertisements regarding the characteristics of an AMG such that the updated protocol mechanisms maintain convergence times within the time frame needed to meet or not significantly exceed existing performance objectives for convergence on the same network or convergence on a network with a similar topology.

FR#4 The solution SHALL ensure that restoration operations happen within the time frame needed to meet existing performance objectives for restoration time on the same network or restoration time on a network with a similar topology.

FR#5 The solution shall provide a mechanism to select a set of paths for an LSP across a network in such a way that flows within the LSP are distributed across the set of paths, while meeting all of the other requirements stated above. The solution SHOULD work in a manner similar to existing multipath techniques, except as necessary to accommodate Advanced Multipath requirements.

FR#6 If extensions to existing protocols are specified and/or new protocols are defined, then the solution SHOULD provide a means for a network operator to migrate an existing deployment in a minimally disruptive manner.

FR#7 Any load-balancing solutions MUST NOT oscillate. Some change in path MAY occur. The solution MUST ensure that path stability and traffic reordering continue to meet performance objectives on the same network or on a network with a similar topology. Since oscillation may cause reordering, there MUST be means to control the frequency of changing the component link over which a flow is placed.

FR#8 Management and diagnostic protocols MUST be able to operate over AMGs.

Existing scaling techniques used in MPLS networks apply to MPLS networks that support Advanced Multipath. Scalability and stability are covered in more detail in [FRAMEWORK].

3.2. Component Links Provided by Lower-Layer Networks

A component link may be supported by a lower-layer network. For example, the lower layer may be a circuit-switched network or another MPLS network (e.g., MPLS Transport Profile (MPLS-TP)). The lower-layer network may change the latency (and/or other performance parameters) seen by the client layer. Currently, there is no protocol for the lower-layer network to inform the higher-layer network of a change in a performance parameter. Communication of the latency performance parameter is a very important requirement. Communication of other performance parameters (e.g., delay variation) is desirable.

FR#9 The solution SHALL specify a protocol means to allow a server-layer network to communicate latency to the client-layer network.

FR#10 The precision of latency reporting SHOULD be configurable. A reasonable default SHOULD be provided. Implementations SHOULD support precision of at least 10% of the one-way latencies for latency of 1 msec or more.

The intent is to measure the predominant latency in uncongested service-provider networks, where geographic delay dominates and is on the order of milliseconds or more. The argument for including queuing delay is that it reflects the delay experienced by applications. The argument against including queuing delay is that if used in routing decisions, it can result in routing instability. This trade-off is discussed in detail in [FRAMEWORK].

3.3. Component Links with Different Characteristics

As one means to provide high availability, network operators deploy a topology in the MPLS network using lower-layer networks that have a certain degree of diversity at the lower layer(s). Many techniques have been developed to balance the distribution of flows across component links that connect the same pair of nodes or ultimately lead to a common destination.

FR#11 In the requirements that follow in this document, the word "indicate" is used where information may be provided by either the combination of link state IGP advertisement and MPLS LSP signaling or via management plane protocols. In later documents, providing framework and protocol definitions, both signaling and management plane mechanisms, MUST be defined.

FR#12 The solution SHALL provide a means for the client layer to indicate a requirement that a client LSP will traverse a component link with the minimum-latency value. This will provide a means by which minimum latency performance objectives of flows within the client LSP can be supported.

FR#13 The solution SHALL provide a means for the client layer to indicate a requirement that a client LSP will traverse a component link with a maximum acceptable latency value as specified by protocol. This will provide a means by which bounded latency performance objectives of flows within the client LSP can be supported.

FR#14 The solution SHALL provide a means for the client layer to indicate a requirement that a client LSP will traverse a component link with a maximum acceptable delay variation value as specified by protocol.

The above set of requirements applies to component links with different characteristics, regardless of whether those component links are provided by parallel physical links between nodes or by sets of paths across a network provided by a server-layer LSP.

Allowing multipath to contain component links with different characteristics can improve the overall load balance and can be accomplished while still accommodating the more strict requirements of a subset of client LSP.

3.4. Considerations for Bidirectional Client LSP

Some client LSPs MAY require a path bound to a specific set of component links. This case is most likely to occur in a bidirectional client LSP where time synchronization protocols such as the Precision Time Protocol (PTP) or the Network Time Protocol (NTP) are carried or in any other case where symmetric delay is highly desirable. There may be other uses of this capability.

Other client LSPs may only require that the LSP serve the same set of nodes in both directions. This is necessary if protocols are carried that make use of the reverse direction of the LSP as a back channel in cases such Operations, Administration, and Maintenance (OAM) protocols using IPv4 Time to Live (TTL) or IPv4 Hop Limit to monitor or diagnose the underlying path. There may be other uses of this capability.

FR#15 The solution SHALL provide a means for the client layer to indicate a requirement that a client LSP be bound to a particular component link within an AMG. If this option is not exercised, then a client LSP that is carried over an AMG may be bound to any component link or set of component links matching all other signaled requirements, and different directions of a bidirectional client LSP can be bound to different component links.

FR#16 The solution MUST support a means for the client layer to indicate a requirement that for a specific co-routed bidirectional client LSP, both directions of the co-routed bidirectional client LSP MUST be bound to the same set of nodes.

FR#17 A client LSP that is bound to a specific component link SHOULD NOT exceed the capacity of a single component link. This is inherent in the assumption that a network SHOULD NOT operate in a congested state if congestion is avoidable.

For some large bidirectional client LSPs, it may not be necessary (or possible due to the client LSP capacity) to bind the LSP to a common set of component links, but it may be necessary or desirable to constrain the path taken by the LSP to the same set of nodes in both directions. Without an entirely new and highly dynamic protocol, it is not feasible to constrain such a bidirectional client LSP from taking multiple paths and coordinating load balance on each side in order to keep both directions of flows within such an LSP on common paths.

3.5. Multipath Load-Balancing Dynamics

Multipath load balancing attempts to keep traffic levels on all component links below congestion levels if possible and preferably well balanced. Load balancing is minimally disruptive (see the discussion below this section's list of requirements). The sensitivity to these minimal disruptions of traffic flows within a specific client LSP needs to be considered.

FR#18 The solution SHALL provide a means for the client layer to indicate a requirement that a specific client LSP MUST NOT be split across multiple component links.

FR#19 The solution SHALL provide a means local to a node that automatically distributes flows across the component links in the AMG such that performance objectives are met, as described in the prior requirements in Section 3.3.

FR#20 The solution SHALL measure traffic flows or groups of traffic flows and dynamically select the component link on which to place this traffic in order to balance the load so that no component link in the AMG between a pair of nodes is overloaded.

FR#21 When a traffic flow is moved from one component link to another in the same AMG between a set of nodes, it MUST be done so in a minimally disruptive manner.

FR#22 Load balancing MAY be used during sustained low-traffic periods to reduce the number of active component links for the purpose of power reduction.

FR#23 The solution SHALL provide a means for the client layer to indicate a requirement that a specific client LSP contains traffic whose frequency of component link change due to load balancing needs to be bounded by a specific value. The solution MUST provide a means to bound the frequency of a component link change due to load balancing for subsets of traffic flow on AMGs.

FR#24 The solution SHALL provide a means to distribute traffic flows from a single client LSP across multiple component links to handle at least the case where the traffic carried in a client LSP exceeds that of any component link in the AMG.

FR#25 The solution SHOULD support the use case where an AMG itself is a component link for a higher order AMG. For example, an AMG comprised of MPLS-TP bidirectional tunnels viewed as logical links could then be used as a component link in yet another AMG that connects MPLS routers.

FR#26 If the total demand offered by traffic flows exceeds the capacity of the AMG, the solution SHOULD define a means to cause some client LSPs to move to an alternate set of paths that are not congested. These "preempted LSPs" may not be restored if there is no uncongested path in the network.

A minimally disruptive change implies that as little disruption as is practical occurs. Such a change can be achieved with zero packet loss. A delay discontinuity may occur, which is considered to be a minimally disruptive event for most services if this type of event is sufficiently rare. A delay discontinuity is an example of a minimally disruptive behavior corresponding to current techniques.

A delay discontinuity is an isolated event that may greatly exceed the normal delay variation (jitter). A delay discontinuity has the following effect. When a flow is moved from a current link to a target link with lower latency, reordering can occur. When a flow is moved from a current link to a target link with a higher latency, a time gap can occur. Some flows (e.g., timing distribution and PW circuit emulation) are quite sensitive to these effects. A delay discontinuity can also cause a jitter buffer underrun or overrun, affecting user experience in real-time voice services (causing an audible click). These sensitivities may be specified in a performance objective.

As with any load-balancing change, a change initiated for the purpose of power reduction may be minimally disruptive. Typically, the disruption is limited to a change in delay characteristics and the potential for a very brief period with traffic reordering. When

configuring a network for power reduction, the network operator should weigh the benefit of power reduction against the disadvantage of a minimal disruption.

4. General Requirements for Protocol Solutions

This section defines requirements for protocol specifications used to meet the functional requirements specified in Section 3.

GR#1 The solution SHOULD extend existing protocols wherever possible, developing a new protocol only where doing so adds a significant set of capabilities.

GR#2 A solution SHOULD extend LDP capabilities to meet functional requirements. This MUST be accomplished without defining LDP Traffic Engineering (TE) methods as decided in [RFC3468].

GR#3 Coexistence of LDP- and RSVP-TE-signaled LSPs MUST be supported on an AMG. Function requirements SHOULD, where possible, be accommodated in a manner that supports LDP-signaled LSP, RSVP-signaled LSP, and LSP setup using management plane mechanisms.

GR#4 When the nodes connected via an AMG are in the same routing domain, the solution MAY define extensions to the IGP.

GR#5 When the nodes are connected via an AMG are in different MPLS network topologies, the solution SHALL NOT rely on extensions to the IGP.

GR#6 The solution SHOULD support AMG IGP advertisement that results in convergence time better than that of advertising the individual component links. The solution SHALL be designed so that it represents the range of capabilities of the individual component links such that functional requirements are met, and it also minimizes the frequency of advertisement updates that may cause IGP convergence to occur.

Examples of advertisement-update-triggering events to be considered include: client LSP establishment/release, changes in component-link characteristics (e.g., latency and up/down state), and/or bandwidth utilization.

GR#7 When a worst-case failure scenario occurs, the number of RSVP-TE client LSPs to be resignaled will cause a period of unavailability as perceived by users. The resignaling time of the solution MUST support protocol mechanisms meeting existing provider performance objectives for the duration of unavailability without significantly relaxing those existing performance objectives for the same network or for networks with similar topology. For example, the processing load due to IGP readvertisement MUST NOT increase significantly, and the resignaling time of the solution MUST NOT increase significantly as compared with current methods.

5. Management Requirements

MR#1 The Management Plane MUST support polling of the status and configuration of an AMG and its individual component links and support notification of status change.

MR#2 The Management Plane MUST be able to activate or deactivate any component link in an AMG in order to facilitate operation maintenance tasks. The routers at each end of an AMG MUST redistribute traffic to move traffic from a deactivated link to other component links based on the traffic flow TE criteria.

MR#3 The Management Plane MUST be able to configure a client LSP over an AMG and be able to select a component link for the client LSP.

MR#4 The Management Plane MUST be able to trace which component link a client LSP is assigned to and monitor individual component link and AMG performance.

MR#5 The Management Plane MUST be able to verify connectivity over each individual component link within an AMG.

MR#6 Component link fault notification MUST be sent to the management plane.

MR#7 AMG fault notification MUST be sent to the management plane and MUST be distributed via a link state message in the IGP.

MR#8 The Management Plane SHOULD provide the means for an operator to initiate an optimization process.

MR#9 An operator-initiated optimization MUST be performed in a minimally disruptive manner, as described in Section 3.5.

6. Acknowledgements

Frederic Jounay of France Telecom and Yuji Kamite of NTT Communications Corporation coauthored a version of this document.

A rewrite of this document occurred after the IETF 77 meeting. Dimitri Papadimitriou, Lou Berger, Tony Li, the former WG Chairs John Scuder and Alex Zinin, the current WG Chair Alia Atlas, and others provided valuable guidance prior to and at the IETF 77 RTGWG meeting.

Tony Li and John Drake have made numerous valuable comments on the RTGWG mailing list that are reflected in versions following the IETF 77 meeting.

Iftekhar Hussain and Kireeti Kompella made comments on the RTGWG mailing list after the IETF 82 meeting that identified a new requirement. Iftekhar Hussain made numerous valuable comments on the RTGWG mailing list that resulted in improvements to the document's clarity.

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

Tom Yu and Francis Dupont provided the SecDir and GenArt reviews, respectively. Both reviews provided useful comments. The current wording of the security section is based on suggested wording from Tom Yu. Lou Berger provided the RtgDir review, which resulted in the document being renamed and the substantial clarification of terminology and document wording, particularly in the Abstract, Introduction, and Definitions sections.

7. Security Considerations

The security considerations for MPLS/GMPLS and for MPLS-TP are documented in [RFC5920] and [RFC6941]. This document does not impact the security of MPLS, GMPLS, or MPLS-TP.

The additional information that this document requires does not provide significant additional value to an attacker beyond the information already typically available from attacking a routing or signaling protocol. If the requirements of this document are met by extending an existing routing or signaling protocol, the security considerations of the protocol being extended apply. If the requirements of this document are met by specifying a new protocol, the security considerations of that new protocol should include an

evaluation of what level of protection is required by the additional information specified in this document, such as data origin authentication.

8. References

8.1. Normative References

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

8.2. Informative References

[FRAMEWORK] Ning, S., McDysan, D., Osborne, E., Yong, L., and C. Villamizar, "Advanced Multipath Framework in MPLS", 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.800] ITU-T, "Unified functional architecture of transport networks", ITU-T Recommendation G.800, February 2012, <http://www.itu.int/rec/T-REC-G/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.

[RFC3468] Andersson, L. and G. Swallow, "The Multiprotocol Label Switching (MPLS) Working Group decision on MPLS signaling protocols", RFC 3468, February 2003.

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

[RFC5920] Fang, L., "Security Framework for MPLS and GMPLS Networks", RFC 5920, July 2010.

[RFC6941] Fang, L., Niven-Jenkins, B., Mansfield, S., and R. Graveman, "MPLS Transport Profile (MPLS-TP) Security Framework", RFC 6941, April 2013.

[USE-CASES]
Ning, S., Malis, A., McDysan, D., Yong, L., and C. Villamizar, "Advanced Multipath Use Cases and Design Considerations", Work in Progress, November 2013.

Authors' Addresses

Curtis Villamizar (editor)
OCCNC, LLC

EMail: curtis@occnc.com

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

EMail: dave.mcdysan@verizon.com

So Ning
Tata Communications

EMail: ning.so@tatacommunications.com

Andrew G. Malis
Huawei Technologies
2330 Central Expressway
Santa Clara, CA 95050
USA

EMail: agmalis@gmail.com

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

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