

RTGWG
Internet-Draft
Intended status: Informational
Expires: April 14, 2011

C. Villamizar, Ed.
Infinera Corporation
D. McDysan, Ed.
S. Ning
A. Malis
Verizon
L. Yong
Huawei USA
October 11, 2010

Requirements for MPLS Over a Composite Link
draft-ietf-rtgwg-cl-requirement-02

Abstract

There is often a need to provide large aggregates of bandwidth that are best provided using parallel links between routers or MPLS LSR. 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 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 microflow as is current practice.

Current practice related to multipath is described briefly in an appendix.

Status of this Memo

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

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

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

This Internet-Draft will expire on April 14, 2011.

Copyright Notice

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

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	4
1.1. Requirements Language	4
2. Assumptions	4
3. Definitions	4
4. Network Operator Functional Requirements	5
4.1. Availability, Stability and Transient Response	5
4.2. Component Links Provided by Lower Layer Networks	6
4.3. Parallel Component Links with Different Characteristics	7
5. Derived Requirements	9
6. Acknowledgements	10
7. IANA Considerations	10
8. Security Considerations	10
9. References	11
9.1. Normative References	11
9.2. Informative References	11
9.3. Appendix References	12
Appendix A. More Details on Existing Network Operator Practices and Protocol Usage	13
Appendix B. Existing Multipath Standards and Techniques	15
B.1. Common Multipath Load Splitting Techniques	16
B.2. Simple and Adaptive Load Balancing Multipath	17
B.3. Traffic Split over Parallel Links	18
B.4. Traffic Split over Multiple Paths	18
Appendix C. ITU-T G.800 Composite Link Definitions and Terminology	18
Authors' Addresses	19

1. Introduction

The purpose of this document is to describe why network operators require certain functions in order to solve certain business problems (Section 2). The intent is to first describe why things need to be done in terms of functional requirements that are as independent as possible of protocol specifications (Section 4). For certain functional requirements this document describes a set of derived protocol requirements (Section 5). Three appendices provide supporting details as a summary of existing/prior operator approaches (Appendix A), a summary of implementation techniques and relevant protocol standards (Appendix B), and a summary of G.800 terminology used to define a composite link (Appendix C).

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 RFC 2119 [RFC2119].

2. Assumptions

The services supported include L3VPN RFC 4364 [RFC4364], RFC 4797 [RFC4797] L2VPN RFC 4664 [RFC4664] (VPWS, VPLS (RFC 4761 [RFC4761], RFC 4762 [RFC4762]) and VPMS VPMS Framework [I-D.ietf-l2vpn-vpms-frmwk-requirements]), Internet traffic encapsulated by at least one MPLS label, and dynamically signaled MPLS or MPLS-TP LSPs and pseudowires. The MPLS LSPs supporting these services may be pt-pt, pt-mpt, or mpt-mpt.

The locations in a network where these requirements apply are a Label Edge Router (LER) or a Label Switch Router (LSR) as defined in RFC 3031 [RFC3031].

The IP DSCP 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. Definitions

ITU-T G.800 Based Composite and Component Link Definitions:
Section 6.9.2 of ITU-T-G.800 [ITU-T.G.800] defines composite and component links as summarized in Appendix C. The following definitions for composite and component links are derived from and intended to be consistent with the cited ITU-T G.800

terminology.

Composite Link: A composite link is a logical link composed of a set of parallel point-to-point component links, where all links in the set share the same endpoints. A composite link may itself be a component of another composite link, but only a strict hierarchy of links is allowed.

Component Link: A point-to-point physical or 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 specific NPO. Examples of a physical link are: Lambda, Ethernet PHY, and OTN. Examples of a logical link are: MPLS LSP, Ethernet VLAN, and MPLS-TP LSP.

Flow: A sequence of packets that must be transferred in order.

Flow identification: The label stack and other information that uniquely identifies a flow. Other information in flow identification may include an IP header, PW control word, Ethernet MAC address, etc. Note that an LSP may contain one or more Flows or an 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.

Network Performance Objective (NPO): Numerical values for performance measures, principally availability, latency, and delay variation. See Appendix A for more details.

4. Network Operator Functional Requirements

The Functional Requirements in this section are grouped in subsections starting with the highest priority.

4.1. Availability, Stability and Transient Response

Limiting the period of unavailability in response to failures or transient events is extremely important as well as maintaining stability. 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 NPO values. Appendix A provides references and a summary of service types requiring a range of restoration times.

- FR#1 The solution SHALL provide a means to summarize routing advertisements regarding the characteristics of a composite link such that the routing protocol converges within the timeframe needed to meet the network performance objective.
- FR#2 The solution SHALL ensure that all possible restoration operations happen within the timeframe needed to meet the NPO. The solution may need to specify a means for aggregating signaling to meet this requirement.
- FR#3 The solution SHALL provide a mechanism to select a path for a flow across a network that contains a number of paths comprised of pairs of nodes connected by composite links in such a way as to automatically distribute the load over the network nodes connected by composite links while meeting all of the other mandatory requirements stated above. The solution SHOULD work in a manner similar to that of current networks without any composite link protocol enhancements when the characteristics of the individual component links are advertised.
- FR#4 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#5 Any automatic LSP routing and/or load balancing solutions MUST not oscillate such that performance observed by users changes such that an NPO is violated. 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#6 Management and diagnostic protocols MUST be able to operate over composite links.

4.2. Component Links Provided by Lower Layer Networks

Case 3 as defined in [ITU-T.G.800] involves a component link supporting an MPLS layer network over another lower layer network (e.g., circuit switched or another MPLS network (e.g., MPLS-TP)). The lower layer network may change the latency (and/or other performance parameters) seen by the MPLS layer network. Network Operators have NPOs of which some components are based on performance parameters. 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#7 In order to support network NPOs and provide acceptable user experience, the solution SHALL specify a protocol means to allow a lower layer server network to communicate latency to the higher layer client network.
- FR#8 The precision of latency reporting SHOULD be at least 10% of the one way latencies for latency of 1 ms or more.
- FR#9 The solution SHALL provide a means to limit the latency on a per LSP basis between nodes within a network to meet an NPO target when the path between these nodes contains one or more pairs of nodes connected via a composite link.

The NPOs differ across the services, and some services have different NPOs for different QoS classes, for example, one QoS class may have a much larger latency bound than another. Overload can occur which would violate an NPO parameter (e.g., loss) and some remedy to handle this case for a composite link is required.

- FR#10 If the total demand offered by traffic flows exceeds the capacity of the composite link, the solution SHOULD define a means to cause the LSPs for some traffic flows to move to some other point in the network that is not congested. These "preempted LSPs" may not be restored if there is no uncongested path in the network.

4.3. Parallel Component Links with Different Characteristics

Corresponding to Case 1 of [ITU-T.G.800], 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 (See Appendix B.3). When the path for a flow can be chosen from a set of candidate nodes connected via composite links, other techniques have been developed (See Appendix B.4).

- FR#11 The solution SHALL measure traffic on a labeled traffic flow and dynamically select the component link on which to place this flow in order to balance the load so that no component link in the composite link between a pair of nodes is overloaded.

- FR#12 When a traffic flow is moved from one component link to another in the same composite link between a set of nodes (or sites), it MUST be done so in a minimally disruptive manner.

When a flow is moved from a current link to a target link with different latency, reordering can occur if the target link latency is less than that of the current or clumping can occur if target link latency is greater than that of the current. Therefore, some flows (e.g., timing distribution, PW circuit emulation) are quite sensitive to these effects, which may be specified in an NPO or are needed to meet a user experience objective (e.g. jitter buffer under/overflow).

- FR#13 The solution SHALL provide a means to identify flows whose rearrangement frequency needs to be bounded by a configured value.
- FR#14 The solution SHALL provide a means that communicates whether the flows within an LSP can be split across multiple component links. The solution SHOULD provide a means to indicate the flow identification field(s) which can be used along the flow path which can be used to perform this function.
- FR#15 The solution SHALL provide a means to indicate that a traffic flow shall select a component link with the minimum latency value.
- FR#16 The solution SHALL provide a means to indicate that a traffic flow shall select a component link with a maximum acceptable latency value as specified by protocol.
- FR#17 The solution SHALL provide a means to indicate that a traffic flow shall select a component link with a maximum acceptable delay variation value as specified by protocol.
- FR#18 The solution SHALL provide a means local to a node that automatically distributes flows across the component links in the composite link such that NPOs are met.
- FR#19 The solution SHALL provide a means to distribute flows from a single LSP across multiple component links to handle at least the case where the traffic carried in an LSP exceeds that of any component link in the composite link. As defined in section 3, a flow is a sequence of packets that must be transferred on one component link.

- FR#20 The solution SHOULD support the use case where a composite link itself is a component link for a higher order composite link. For example, a composite link comprised of MPLS-TP bi-directional tunnels viewed as logical links could then be used as a component link in yet another composite link that connects MPLS routers.

5. Derived Requirements

This section takes the next step and derives high-level requirements on protocol specification from the functional requirements.

- DR#1 The solution SHOULD attempt to extend existing protocols wherever possible, developing a new protocol only if this adds a significant set of capabilities.

The vast majority of network operators have provisioned L3VPN services over LDP. Many have deployed L2VPN services over LDP as well. TE extensions to IGP and RSVP-TE are viewed as being overly complex by some operators.

- DR#2 A solution SHOULD extend LDP capabilities to meet functional requirements (without using TE methods as decided in [RFC3468]).

- DR#3 Coexistence of LDP and RSVP-TE signaled LSPs MUST be supported on a composite link. Other functional requirements should be supported as independently of signaling protocol as possible.

- DR#4 When the nodes connected via a composite link are in the same MPLS network topology, the solution MAY define extensions to the IGP.

- DR#5 When the nodes are connected via a composite link are in different MPLS network topologies, the solution SHALL NOT rely on extensions to the IGP.

- DR#6 The Solution SHALL support composite link 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 also minimizes the frequency of advertisement updates which may cause IGP convergence to occur.

One solution approach is to summarize the characteristics of the component links in IGP advertisements; however, the intent

of the above requirement is not to specify the form of a solution. Examples of advertisement update triggering events to be considered include: LSP establishment/release, changes in component link characteristics (e.g., latency, up/down state), and/or bandwidth utilization.

DR#7 When a worst case failure scenario occurs, the resulting number of links advertised in the IGP causes IGP convergence to occur, causing a period of unavailability as perceived by users. The convergence time of the solution MUST meet the SLA objective for the duration of unavailability.

DR#8 When a worst case failure scenario occurs, the number of RSVP-TE LSPs to be resigned will cause a period of unavailability as perceived by users. The resigning time of the solution MUST meet the NPO objective for the duration of unavailability. The resigning time of the solution MUST not increase significantly as compared with current methods.

6. Acknowledgements

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

A rewrite of this document occurred after the IETF77 meeting. Dimitri Papadimitriou, Lou Berger, Tony Li, the WG chairs John Scuder and Alex Zinin, and others provided valuable guidance prior to and at the IETF77 RTG WG meeting.

Tony Li and John Drake have made numerous valuable comments on the RTG WG mailing list that are reflected in versions following the IETF77 meeting.

7. IANA Considerations

This memo includes no request to IANA.

8. Security Considerations

This document specifies a set of requirements. The requirements themselves do not pose a security threat. If these requirements are met using MPLS signaling as commonly practiced today with authenticated but unencrypted OSPF-TE, ISIS-TE, and RSVP-TE or LDP, then the requirement to provide additional information in this communication presents additional information that could conceivably

be gathered in a man-in-the-middle confidentiality breach. Such an attack would require a capability to monitor this signaling either through a provider breach or access to provider physical transmission infrastructure. A provider breach already poses a threat of numerous types of attacks which are of far more serious consequence. Encryption of the signaling can prevent or render more difficult any confidentiality breach that otherwise might occur by means of access to provider physical transmission infrastructure.

9. References

9.1. Normative References

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

9.2. Informative References

- [I-D.ietf-l2vpn-vpms-frmwk-requirements]
Kamite, Y., JOUNAY, F., Niven-Jenkins, B., Brungard, D., and L. Jin, "Framework and Requirements for Virtual Private Multicast Service (VPMS)", draft-ietf-l2vpn-vpms-frmwk-requirements-03 (work in progress), July 2010.
- [ITU-T.G.800]
ITU-T, "Unified functional architecture of transport networks", 2007, <<http://www.itu.int/rec/T-REC-G/recommendation.asp?parent=T-REC-G.800>>.
- [RFC2702] Awduche, D., Malcolm, J., Agogbua, J., O'Dell, M., and J. McManus, "Requirements for Traffic Engineering Over MPLS", RFC 2702, September 1999.
- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.
- [RFC3468] Andersson, L. and G. Swallow, "The Multiprotocol Label Switching (MPLS) Working Group decision on MPLS signaling protocols", RFC 3468, February 2003.
- [RFC3809] Nagarajan, A., "Generic Requirements for Provider Provisioned Virtual Private Networks (PPVPN)", RFC 3809, June 2004.
- [RFC4031] Carugi, M. and D. McDysan, "Service Requirements for Layer 3 Provider Provisioned Virtual Private Networks (PPVPNs)",

RFC 4031, April 2005.

- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
- [RFC4664] Andersson, L. and E. Rosen, "Framework for Layer 2 Virtual Private Networks (L2VPNs)", RFC 4664, September 2006.
- [RFC4665] Augustyn, W. and Y. Serbest, "Service Requirements for Layer 2 Provider-Provisioned Virtual Private Networks", RFC 4665, September 2006.
- [RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.
- [RFC4762] Lasserre, M. and V. Kompella, "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.
- [RFC4797] Rekhter, Y., Bonica, R., and E. Rosen, "Use of Provider Edge to Provider Edge (PE-PE) Generic Routing Encapsulation (GRE) or IP in BGP/MPLS IP Virtual Private Networks", RFC 4797, January 2007.
- [RFC5254] Bitar, N., Bocci, M., and L. Martini, "Requirements for Multi-Segment Pseudowire Emulation Edge-to-Edge (PWE3)", RFC 5254, October 2008.

9.3. Appendix References

- [I-D.ietf-pwe3-fat-pw]
Bryant, S., Filsfils, C., Drafz, U., Kompella, V., Regan, J., and S. Amante, "Flow Aware Transport of Pseudowires over an MPLS PSN", draft-ietf-pwe3-fat-pw-03 (work in progress), January 2010.
- [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.Y.1540]
ITU-T, "Internet protocol data communication service - IP packet transfer and availability performance parameters", 2007, <<http://www.itu.int/rec/T-REC-Y.1540/en>>.

- [ITU-T.Y.1541] ITU-T, "Network performance objectives for IP-based services", 2006, <<http://www.itu.int/rec/T-REC-Y.1541/en>>.
- [RFC1717] Sklower, K., Lloyd, B., McGregor, G., and D. Carr, "The PPP Multilink Protocol (MP)", RFC 1717, November 1994.
- [RFC2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, December 1998.
- [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.
- [RFC3260] Grossman, D., "New Terminology and Clarifications for Diffserv", RFC 3260, April 2002.
- [RFC4201] Kompella, K., Rekhter, Y., and L. Berger, "Link Bundling in MPLS Traffic Engineering (TE)", RFC 4201, October 2005.
- [RFC4301] Kent, S. and K. Seo, "Security Architecture for the Internet Protocol", RFC 4301, December 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.

Appendix A. More Details on Existing 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 Service Level Sepcification (SLS) that is for internal use by the operator. See [ITU-T.Y.1540], [ITU-T.Y.1541], RFC3809, Section 4.9 [RFC3809] for examples of the form of such SLA and SLS specifications. In this

document we use the term Network Performance Objective (NPO) as defined in section 5 of [ITU-T.Y.1541] since the SLA and SLS measures have network operator and service specific implications. Note that the numerical NPO values of Y.1540 and Y.1541 span multiple networks and may be looser than network operator SLA or SLS objectives. 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 (e.g., in TCP 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 NPOs have a bound on point-point latency, and as long as this bound is met, the NPO is met -- decreasing the latency is not necessary. In some NPOs, 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 to meet this type of NPO, but this lowers availability since an alternate route that meets the latency NPO 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) this was not known to the MPLS network. This resulted in latency affecting end user experience, sometimes violating NPOs 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 NPO or meet user experience objectives. This results in reduced flexibility and increased cost for network operators. Using lower layer networks to provide restoration and grooming is expected to be more efficient, 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 NPOs for pt-pt services are often tied closely to geographic locations, while latency for multipoint services may be based upon a worst case within a region.

Section 7 of [ITU-T.Y.1540] defines availability for an IP service in terms of loss exceeding a threshold for a period on the order of 5 minutes. However, the timeframes for restoration (i.e., as implemented by pre-determined 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. 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. The approach that has been taken is to aggregate these QoS classes into similar sets on LER-LSR and LSR-LSR links.

Labeled LSPs have been and use of link layer encapsulation have been standardized in order to provide a means to meet these needs.

The IP DSCP cannot be used for flow identification since RFC 4301 Section 5.5 [RFC4301] requires Diffserv transparency, and in general network operators do not rely on the DSCP of Internet packets.

A label is pushed onto Internet packets when they are carried along with L2/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 mpt-pt 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).

Traffic is typically comprised of a few large (some very large) flows and many small flows. In some cases, separate LSPs are established for very large flows. This can occur even if the IP header information is inspected by a router, for example an IPsec tunnel that carries a large amount of traffic. An important example of large flows is that of a L2/L3 VPN customer who has an access line bandwidth comparable to a client-client composite 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, or even BGP, and equal cost LSP, as described in Appendix B.4. Various multipath techniques have strengths and weaknesses.

The term composite link is more general than terms such as link aggregate which is generally considered to be specific to Ethernet and its use here is consistent with the broad definition in [ITU-T.G.800]. The term multipath excludes inverse multiplexing and refers to techniques which only solve the problem of large aggregations of traffic, without addressing the other requirements outlined in this document.

B.1. Common Multipath Load Splitting 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 other LSP as payload, there is no signaling associated with these inner LSP. Therefore even when using RSVP-TE signaling there may be insufficient information provided by signaling to adequately distribute load across a composite link.

Generally a set of label stack entries that is unique across the ordered set of label numbers 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. Such modifications

are outside the scope of this document.

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 flow 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 [I-D.ietf-pwe3-fat-pw]. 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. Simple and Adaptive Load Balancing Multipath

Simple 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. A common simple multipath implementation assumes that all members (component links) are of equal capacity and perform a modulo operation across the hashed value. An alternate simple multipath technique uses a table generally with a power of two size, and distributes the table entries proportionally among members according to the capacity of each member.

Simple 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 simple load balancing.

An adaptive multipath technique is one where the traffic bound to each member (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 delivered out of sequence.

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.

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 LSP that comprise the bundle.

Appendix C. ITU-T G.800 Composite Link Definitions and Terminology

Composite Link:

Section 6.9.2 of ITU-T-G.800 [ITU-T.G.800] defines composite link in terms of three cases, of which the following two are relevant (the one describing inverse (TDM) multiplexing does not apply). Note that these case definitions are taken verbatim from section 6.9, "Layer Relationships".

Case 1: "Multiple parallel links between the same subnetworks can be bundled together into a single composite link. Each component of the composite link is independent in the sense that each component link is supported by a separate server layer trail. The composite link conveys communication information using different server layer trails thus the sequence of symbols crossing this link may not be preserved. This is illustrated in Figure 14."

Case 3: "A link can also be constructed by a concatenation of component links and configured channel forwarding relationships. The forwarding relationships must have a 1:1 correspondence to the link connections that will be provided by the client link. In this case, it is not possible to fully infer the status of the link by observing the server layer trails visible at the ends of the link. This is illustrated in Figure 16."

Subnetwork: A set of one or more nodes (i.e., LER or LSR) and links. As a special case it can represent a site comprised of multiple nodes.

Forwarding Relationship: Configured forwarding between ports on a subnetwork. It may be connectionless (e.g., IP, not considered in this draft), or connection oriented (e.g., MPLS signaled or configured).

Component Link: A topological relationship between subnetworks (i.e., a connection between nodes), which may be a wavelength, circuit, virtual circuit or an MPLS LSP.

Authors' Addresses

Curtis Villamizar (editor)
Infinera Corporation
169 W. Java Drive
Sunnyvale, CA 94089

Email: cvillamizar@infinera.com

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

Email: dave.mcdysan@verizon.com

So Ning
Verizon
2400 N. Glenville Ave.
Richardson, TX 75082

Phone: +1 972-729-7905
Email: ning.so@verizonbusiness.com

Andrew Malis
Verizon
117 West St.
Waltham, MA 02451

Phone: +1 781-466-2362
Email: andrew.g.malis@verizon.com

Lucy Yong
Huawei USA
1700 Alma Dr. Suite 500
Plano, TX 75075

Phone: +1 469-229-5387
Email: lucyyong@huawei.com

OSPF Working Group
Internet-Draft
Intended Status: Standards Track
Expires: April 2011

S. Kini
W. Lu
A. Tian
Ericsson
October 18, 2010

OSPF Fast Notifications
draft-kini-ospf-fast-notification-00.txt

Status of this Memo

Distribution of this memo is unlimited.

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at <http://www.ietf.org/ietf/lid-abstracts.txt>.

The list of Internet-Draft Shadow Directories can be accessed at <http://www.ietf.org/shadow.html>.

This Internet-Draft will expire on April 21, 2011.

Copyright Notice

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

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

Abstract

Several applications could use a mechanism to quickly notify one or more routers about control-protocol events. Current mechanisms to convey such information to routers multiple hops away involves hop-by-hop protocol-specific control plane processing as well as hop-by-hop control plane forwarding. The delay due to control planes involvement in processing/forwarding, adversely affects the application's goal (e.g. fast convergence). This document describes a framework to use data plane forwarding to convey control protocol information multiple hops away. It also defines some sample applications within this framework.

Table of Contents

1. Introduction	4
2. Conventions used in this document	5
3. Scope	5
4. Requirements	5
5. Architecture	5
5. Security Considerations	6
6. IANA Considerations	6
7. References	6
7.1. Normative References	6
8. Acknowledgements	6
Appendix A: OSPF fast convergence on link-down using FN	7
A.1. OSPF procedural changes	7
A.2. FN service using spanning tree	7
Authors' Addresses	9

1. Introduction

There are several applications that could use a mechanism to quickly notify one or more routers in a network about a specific control-protocol event. If the destination router(s) is more than one hop away then the message has to be forwarded by the intermediate routers. This forwarding typically does not exclusively happen via the forwarding plane.

Some applications establish adjacent neighbor relationship with single hop neighbors. Information that needs to be conveyed multiple hops away is first conveyed to adjacent neighbors that are a single hop away. Each neighbor then performs application specific processing and forwards information further. The delay in receiving the information at a router is gated by the processing and forwarding speed of the control plane at each hop along a path from the originating router.

A typical example of an application that sends information to directly connected adjacent neighbors is a link-state routing interior gateway protocol (IGP) such as [OSPF]. When conveying a Link State Advertisement (LSA) to all routers in the area, OSPF's flooding algorithm transmits the LSA to its single hop away adjacent neighbor. The received LSA undergoes processing according to OSPF's processing rules and is then forwarded to OSPF neighbors further away from the router originating the LSA. As explained earlier the delay in receiving a LSA at a router is gated by the processing and forwarding speed of the control plane at each hop along a path from the originating OSPF router.

Some applications need to send information to routers that are multiple hops away even though they do not have adjacency relationship with directly connected neighbors. In such cases the forwarding of application messages depends on the forwarding plane being setup by an underlying protocol that has established adjacent neighbor relationship with routers that are a single hop away. In scenarios where the data plane forwarding is changing due to the underlying protocol, the applications message forwarding speed and reliability is gated by the speed and mechanisms of the underlying protocols hop-by-hop message processing and forwarding by control-plane.

A typical example of an application that could use a mechanism to send information to non-directly connected neighbors is IP FastReroute (IP-FRR). It could use a forwarding mechanism that has been setup by an underlying protocol to trigger (on failure) a non-directly connected neighbor, to switch traffic to an alternate path. To reliably deliver the applications message, the forwarding

mechanism has to be resilient against failures and the changed topology.

2. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

3. Scope

This document describes a framework for quickly delivering notifications from one router to one or more routers using data plane as the main forwarding mechanism. It also defines some solutions under this framework to address the needs of some specific applications.

4. Requirements

Fast notifications (henceforth referred to as FN) must be designed as a set of services that can satisfy the requirements of different control-protocol applications. FN should avoid introducing new protocols and should re-use existing, commonly used protocols as much as possible.

Deploying FN must not introduce new encapsulation requirements for routers unless those encapsulations are already available in the data plane for those applications. Notifying multiple routers should use multicast whenever possible.

5. Architecture

A choice of protocol to realize FN must be based on the set of commonly deployed protocols. These protocols must preferably have applicability in a wide set of network architectures such as IP-routing, L3VPN, L2VPN etc. Also, the knowledge of the network topology would be particularly useful for path computation purposes. The logical candidate for these requirements would be a link-state interior gateway protocol (IGP) such as OSPF or IS-IS.

To implement a specific FN service, a router must convey its capability to the set of routers that setup forwarding to one or more routers in that set for specific packets in a way such that data-plane forwarding of notifications. It must also convey its share of the information that is needed to implement that FN service.

To convey this information via OSPF, an opaque LSA is used. An Opaque type field "FN" is defined. The type specific ID indicates a

particular FN service. The content of the LSA is a variable list of TLVs that include information required to implement that FN service. Different FN services will have different sets of TLVs. A specific instance of a FN service and how an application might use it is specified in Appendix A.

5. Security Considerations

Security considerations of the application also apply when FN service is used by the application. If additional security considerations arise due to the way in which FN is used by the application, then those should be resolved in the document that explains how an application uses FN.

6. IANA Considerations

IANA needs to allocate a OSPF opaque type field for FN. Within that LSID values for different FN services will have to be allocated. Also a TLV type field will have to be allocated.

7. References

7.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [OSPF] Moy, J., "OSPF Version 2", RFC 2328, April 1998.

8. Acknowledgements

The authors would like to thank Joel Halpern for his comments.

Appendix A: OSPF fast convergence on link-down using FN

OSPF fast convergence is gated by how quickly the flooding algorithm can propagate the LSA throughout the area. This requires hop-by-hop processing and forwarding by control plane. If a FN service can transmit the link-down notification to all routers in the area then OSPF's fast convergence can be improved in the link-down scenario.

A.1. OSPF procedural changes OSPF's procedures must be modified to use the FN service as follows. OSPF transmits a copy of the updated Router LSA (on link-down) using a FN service in addition to the normal processing and flooding done by OSPF. The destination IP address of the link-state update (LSU) packet is set to the one dictated by the FN service. If Cryptographic authentication is required, a shared secret key must be configured for the area. The Cryptographic sequence number in the LSU must be set to zero. On receiving a LSU via FN, the router accepts it if authentication succeeds. There must be no acknowledgement for such an LSU. If the received LSA is older than the one in the LSDB, the received LSA is discarded. If the received LSA is newer, the LSA is stored alongside the older copy and a timer T-discard-FN-LSA is started. A flag FN-LSA-present is used in the LSDB entry to indicate that a newer version of the LSA (received via FN) is present. SPF is triggered. During SPF, if the FN-LSA-present flag is true then the LSA received via FN is used instead. When a LSA is received via the flooding procedure of [OSPF], and is determined to be newer, it is compared with the LSA copy received via FN (if one exists). If the two copies are the same, the LSA received via FN becomes the only entry in the LSDB. If the two copies are different, the LSA received through the flooding procedure of [OSPF] becomes the only copy in the LSDB and SPF is triggered. In both cases the flag FN-LSA-present is cleared and the timer T-discard-FN-LSA is canceled. When the timer T-discard-FN-LSA expires, the corresponding LSA copy received via FN is discarded (FN-LSA-present flag is cleared) and SPF is triggered.

A.2. FN service using spanning tree

One way to provide the FN service for this application is as follows. A multicast spanning tree (with a specially allocated multicast destination IP address) is used to send the link-down notification message. The tree must be consistently computed at all routers. It must be computed as a shortest path tree rooted at the highest router-id. During tree computation only routers that are capable of this FN service are picked. When multiple paths are available the neighboring node in the graph with highest LSID is picked. When multiple paths are available through multiple interfaces to a neighboring node, a numbered interface is preferred over an

unnumbered interface. A higher IP address is preferred among numbered interfaces and a higher ifIndex is preferred among unnumbered interfaces. Multicast forwarding state is installed using such a tree as a bi-directional tree. Each router on the tree can send packets to all other routers on that tree. Even when the topology changes such that the tree breaks, the link-down notification is delivered to all routers.

Authors' Addresses

Sriganesh Kini
Ericsson
300 Holger Way, San Jose, CA 95134
EMail: sriganesh.kini@ericsson.com

Wenhu Lu
Ericsson
300 Holger Way, San Jose, CA 95134
EMail: wenhu.lu@ericsson.com

Albert Tian
Ericsson
300 Holger Way, San Jose, CA 95134
EMail: albert.tian@ericsson.com

RTGWG
Internet-Draft
Intended status: Standards Track
Expires: April 21, 2011

W. Lu
A. Tian
S. Kini
Ericsson
October 18, 2010

Fast Notification Framework
draft-lu-fast-notification-framework-00

Abstract

This document describes an architectural work that competes with the IP Fast Re-Route (IPFRR) solution which aims to minimize the network down time in the event of equipments failure. The work provides a layered framework based upon which applications such as the domain-wide fast convergence may be achieved through the transport layer fast delivery of failure notifications.

Status of this Memo

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

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

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

This Internet-Draft will expire on April 21, 2011.

Copyright Notice

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

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of

the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
1.1. Requirements Language	4
1.2. Acronyms	4
2. Event Framework	4
3. Layered Structure	5
4. Operation	6
4.1. Failure detection	6
4.2. Notification Origination	6
4.2.1. IGP PDU	7
4.2.2. Uniform Message	7
4.3. Fast Flooding	7
4.4. Notification Receiving and Handling	8
4.5. Routing/Forwarding Table Update	8
5. Convergence Analyses	8
5.1. Definition of Convergence Time	8
5.2. Domain Wide Convergence	8
5.3. Micro-looping	9
5.4. Packet Reordering	10
6. Scalability Analyses	10
7. Traffic Analyses	10
8. Acknowledgements	10
9. IANA Considerations	10
10. Security Considerations	11
11. References	11
11.1. Normative References	11
11.2. Informative References	11
Authors' Addresses	11

1. Introduction

The ability to recover rapidly from network failures is one of the most sought network characteristics. Few solutions address this issue to the satisfactory.

IPFRR [RFC5714] is one such solution. It mimics MPLS-FRR [RFC4090] solution. The difference is that the MPLS-FRR is path based, or source routing based in other words. This implies that the re-route decision can be carried out by the PLR (point-of-local-repair) router alone, with no need of cooperation of other LSRs in the network.

Unfortunately, IP based FRR is by nature not source routing based. Its re-route decision may not be honored by other routers in the network. The consequence can be very severe, either traffic outage or even routing loops.

Many methods were proposed around IPFRR concept but none is close to be satisfactory. Some methods such as LFA described in [RFC5286] require lot of computation and have coverage issue. Some others such as Not-Via [I-D.ietf-rtgwg-ipfrr-notvia-addresses] are extremely complicated and are prohibitive to be useful.

The primary reason for such difficulties can be understood from the following passage which is quoted from [RFC5714] first paragraph of section 1:

However, there is an alternative approach, which is to compute backup routes that allow the failure to be repaired locally by the router(s) detecting the failure without the immediate need to inform other routers of the failure.

The phrase "without the immediate need to inform other routers of the failure" is against the very nature of the IP network in which the domain-wide synchronization is the key.

In this document we propose a method which directly addresses the rapid network synchronization needs. It is not IPFRR based. However it can achieve the same or better result without much complexity and compromise.

The method lays out a framework which decouples the improvement in the forwarding plane from the control plane. The design also allows and promotes future innovations based upon the framework.

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 RFC 2119 [RFC2119].

1.2. Acronyms

FRR	-	Fast Re-Route
IPFRR	-	IP Fast Re-Route
MPLS	-	Multi-Protocol Label Switch
LFA	-	Loop Free Alternative
TLV	-	Type Length Value tuple
IGP	-	Interior Gateway protocol
OSPF	-	Open Shortest Path First
IS-IS	-	Intermediate System to Intermediate System
PDU	-	Protocol Data Unit
DoS	-	Denial of Service
FNF	-	Fast Notification Framework

2. Event Framework

An event framework is introduced for the purpose of rapid disseminating of events to all interested receivers in a network.

The framework is application independent. Many applications can generate the events and/or register to receive the events. A TLV based framework is proposed to ensure separation between application and the delivery framework.

The event framework is also independent of the underlying delivery mechanisms. Different delivery mechanisms may be introduced, each with different properties suitable for different requirements. For example, some delivery mechanism is solely optimized for simplicity; while other may improve on reliability.

One of the use cases of this event framework is Fast Failure

Notification, which can be used to improve network convergence time. When a failure occurs in a network, routers adjacent to the failure can detect it and quickly disseminate the failure notifications to other routers throughout the area. Routing protocols on different routers can register and receive such failure notifications, then quickly react to the failure to achieve fast convergence.

The routing protocols discussed in this work are Interior Gateway Protocols (IGP) with the focus on the Link State Routing Protocols such as Open Shortest Path First [RFC2328] and Intermediate System to Intermediate System [RFC1195] [ISO.10589.1992].

The event in the scope of this architecture is specifically the link-down event or node-down event. The up events are not fast flooded for the sake of network stability.

3. Layered Structure

The framework can be viewed as a layered structure in which various routing functions can be rearranged. This arrangement is based on the principle of separation of functions. It will facilitate the innovation in various component building blocks and in the mean while allow them to integrate in a systematic manner.

There are two layers that make the framework. One is for routing protocol specific functionality. The other is the data transport layer. Figure 1 depicts this concept.

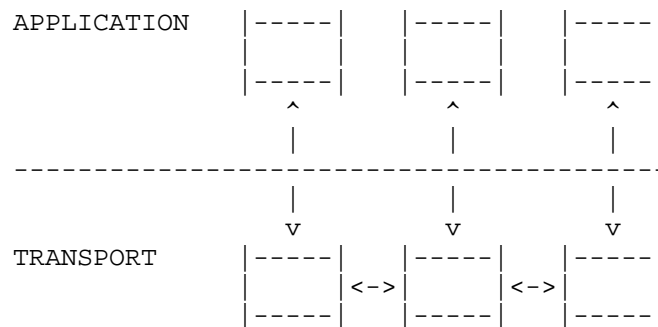


Figure 1: Fast Notification Architecture

Regular routing protocol performs the flooding in store-and-forward manner. While this is reliable (retransmission) and secure (adjacency check), it involves control plane operation and the control plane to data plane communication. It inevitably drags the

network-wide convergence.

With the fast notification architecture, the delivery function is detached from the application layer and moved onto the transport layer. More precisely, the transport layer provides domain-wide fast delivery platform. The normal flooding function is still kept in the application layer to ensure ultimate synchronization in case the fast flooding does not reach some intended routers for whatever reasons.

The speed of the fast flooding needs not to be faster than the data traffic. As long as the messenger travels at the same speed of the data traffic, it always gives the next-hop router the same amount of time for processing as it gives the previous router.

4. Operation

Fast failure notification operates on following steps:

1. Failure detection;
2. Notification composing and dispatching;
3. Notification flooding;
4. Notification receiving;
5. Routing/forwarding table update.

4.1. Failure detection

This can be made in many ways. But it has to be fast and light-weight. Layer-2 link-event monitoring and signaling is obvious an option. Bidirectional Forwarding detection (BFD) is also a good candidate. There may be more, or combinations of them.

The fast notification architecture encourages the innovation in this area which can be pursued freely and independently.

4.2. Notification Origination

This part involves the message format. This document does not specify or endorse a particular format. It is open to any format as long as it fulfills the fast flooding purpose. The detecting router is responsible for the initiation of the fast notification process. Its action is the starting point of the fast flooding.

There are two packet formats worth of mentioning.

4.2.1. IGP PDU

The simplest approach is to use the IGP packet format directly. For example, the OSPF Router-LSA packet which reflects a broken adjacency (one fewer router link) can be fast-flooded to all routers without special modification.

The benefit is that the receivers can process the packet as usual. Moreover since the packet is no different than the one in normal flooding, it guarantees the seamless transition when the "slow" flooding catches up. Plus, there will be no duplicate effort of fast and slow convergence. Flooding stops wherever a router is updated (already fast flooded).

The drawback is that the message cannot be made uniform for multiple protocols. Other protocol such as IS-IS will have to devise a different format. In addition, since IS-IS PDU is not IP based, it may require encapsulation in some cases.

Another drawback is that the normal IGP flooding uses adjacency check to prevent DoS attack or PDU replay from un-trusted parties. The check has to be bypassed for the fast-flooded packets to be accepted. This opens door to the DoS or some other attacks. Domain-wide authentication may be adopted for protection.

4.2.2. Uniform Message

This format must include essential and sufficient information about the broken link. The message will be treated on the receiver router as a local event. The uniformed messaging provides freedom for future expansion. The format thus is recommended TLV-based.

Cautions must be taken in case the message is mistakenly flooded due to bugs or some error conditions. Timeout machinery may be used to protect against such issues.

The detecting router is responsible for the initiation of the fast notification process. Its action is the starting point of the fast flooding.

4.3. Fast Flooding

The fast flooding does not specify the fast flooding mechanism. It is up to the routing society to figure out and single out good solutions. The requirement is that the flooding has to be

- a. Reliable in that it reaches all participants even after failures occur;
- b. Loop-free;
- c. Simple;
- d. Can be authenticated.

4.4. Notification Receiving and Handling

This involves upon the arrival of the notification message, how it is forwarded to the routing protocol for further processing. If the fast-flooding scheme uses specific IP destination addresses or MAC addresses, the receiving router has to recognize it.

When the message reaches the protocol process, it may have to relax its acceptance criteria.

If in the future, some algorithm is developed that the notification handling takes very few CPU cycles, this process may be performed in real-time. Therefore it is worthy of considering move the notification handling into the data plane. This will cut a large chunk of delay and may lead to hitless domain-wide convergence.

4.5. Routing/Forwarding Table Update

This should be the same as normal IGP decision process. It is also possible to pre-download the changes to the data plane if the complexity can be limited. This will improve the overall convergence time dramatically.

5. Convergence Analyses

5.1. Definition of Convergence Time

The convergence time is measured by dividing the number of lost packets with the traffic flow rate between any two routers in the domain. This SHOULD equal to the domain wide network convergence time if all individual routers have the same computing power and the same convergence time.

5.2. Domain Wide Convergence

Due to the propagation delay, all routers do not converge at the same time. The traffic loss, however, stops immediately after the first router repairs.

This is because the data traffic has to go through the same propagation delay, which exactly compensate the late starting of the convergence at remote routers.

Take a ring topology for example, as shown in Figure 2.

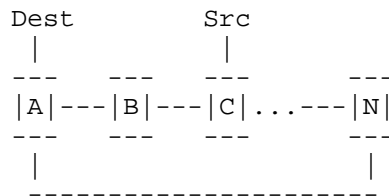


Figure 2: Ring Topology

Assume all routers have same convergence time 50 milliseconds.
Assume the transmission delay over each hop is 20 milliseconds.

Upon link A-B failure, B floods its Link State Update to C. Table 1 shows the convergence timeline.

Node	Converge Starts	converge Completes
B	0	50ms
C	20ms	70ms

During the first 50 milliseconds, packets from B to A are dropped. Right after 50th milliseconds, B re-routes packets toward C. Those packets, after traveling 20 milliseconds, arrive C at 70th milliseconds when C is just repaired. Since C and all downstream routers will correct themselves one by one right before those packets arrive, they will arrive at the destination via the corrected path successfully. The overall convergence time is thus same as B's.

5.3. Micro-looping

If routers' convergence time is different, micro looping may form, although packets will still be delivered after several loops. Still use Figure 2 for example. Assume C needs 90 milliseconds to converge. When B re-routes packets back to C at 70th milliseconds, C has not finished its updating yet. It continues to use its old forwarding table and bounces packets back to B. B in turn re-route packets again to C. This time packets arrive at C at 110th milliseconds. C has done updating and will forward packets

correctly. The packets are looped once.

The micro-looping does not form easily with Fast Flooding method. The routers have to differ in computing speed and differ significantly.

5.4. Packet Reordering

Due to the different convergence timeline, packets may be temporarily forwarded in wrong direction before being placed on the right track. This will not cause packet loss, but will result in packet reordering.

Packet reordering affects TCP communication adversely in that new sequence numbered packets may arrive ahead of the older ones.

This problem is common in IPFRR solutions, and remains an open issue. Not-Via for example, may have packets reordered when it switches to use the final stable routes from the temporary LFAs. On the other hand, the connectionless network by nature never promises ordered packet delivery. This type of problem deserves a separate topic and is beyond the scope of this document.

6. Scalability Analyses

Fast Flooding scales with networks of any size and any topology. At least it scales no inferior to the normal IGP flooding.

7. Traffic Analyses

Traffics that did not route through the broken link are intact. Traffics that did will be successfully re-routed as soon as the affected router converges (as opposed to all routers converge).

Upon the convergence of the affected router, Fast Flooding guarantees correct routes for all affected traffics.

8. Acknowledgements

TBD

9. IANA Considerations

This memo includes no request to IANA.

10. Security Considerations

TBD

11. References

11.1. Normative References

- [ISO.10589.1992]
International Organization for Standardization,
"Intermediate system to intermediate system intra-domain-
routing routine information exchange protocol for use in
conjunction with the protocol for providing the
connectionless-mode Network Service (ISO 8473)",
ISO Standard 10589, 1992.
- [RFC1195] Callon, R., "Use of OSI IS-IS for routing in TCP/IP and
dual environments", RFC 1195, December 1990.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate
Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.

11.2. Informative References

- [I-D.ietf-rtgwg-ipfrr-notvia-addresses]
Shand, M., Bryant, S., and S. Previdi, "IP Fast Reroute
Using Not-via Addresses",
draft-ietf-rtgwg-ipfrr-notvia-addresses-05 (work in
progress), March 2010.
- [RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute
Extensions to RSVP-TE for LSP Tunnels", RFC 4090,
May 2005.
- [RFC5286] Atlas, A. and A. Zinin, "Basic Specification for IP Fast
Reroute: Loop-Free Alternates", RFC 5286, September 2008.
- [RFC5714] Shand, M. and S. Bryant, "IP Fast Reroute Framework",
RFC 5714, January 2010.

Authors' Addresses

Wenhu Lu
Ericsson
300 Holger Way
San Jose, California 95134
USA

Phone: 408 750-5436
Email: wenhu.lu@ericsson.com

Albert Tian
Ericsson
300 Holger Way
San Jose, California 95134
USA

Phone: 408 750-8739
Email: albert.tian@ericsson.com

Sriganesh Kini
Ericsson
300 Holger Way
San Jose, California 95134
USA

Phone: 408 750-5210
Email: sriganesh.kini@ericsson.com

Network Working Group
Internet Draft
Intended Status: Informational
Expires: April 2011

N. So
A. Malis
D. McDysan
Verizon
L. Yong
Huawei
F. Jounay
France Telecom
Y. Kamite
NTT
October 20, 2010

Composite Link Framework in Multi Protocol Label Switching (MPLS)
draft-so-yong-rtgwg-cl-framework-02

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79. This document may not be modified, and derivative works of it may not be created, and it may not be published except as an Internet-Draft.

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79. This document may not be modified, and derivative works of it may not be created, except to publish it as an RFC and to translate it into languages other than English.

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at <http://www.ietf.org/ietf/lid-abstracts.txt>

The list of Internet-Draft Shadow Directories can be accessed at <http://www.ietf.org/shadow.html>

This Internet-Draft will expire on April 20, 2011.

Copyright Notice

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

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

Abstract

This document specifies a composite link framework in MPLS network. A composite link consists of a group of homogenous or non-homogenous links that have the same forward adjacency and can be considered as a single TE link or an IP link in routing. The composite link relies on its component links to carry the traffic over composite link. The document specifies composite link model. Applicability is described for a single pair of MPLS-capable nodes, a sequence of MPLS-capable nodes, or a set of layer networks connecting MPLS-capable nodes.

Table of Contents

1. Introduction.....	3
2. Conventions used in this document.....	3
2.1. Terminology.....	3
3. Composite Link Framework.....	4
4. Composite Link in Control Plane.....	5
5. Composite Link in Data Plane.....	6
6. Security Considerations.....	7
7. IANA Considerations.....	7
8. References.....	7
8.1. Normative References.....	7
8.2. Informative References.....	7
9. Acknowledgments.....	7

1. Introduction

Composite link functional requirement are specified in [CL-REQ]. This document specifies a framework of Composite Link in IP/MPLS network to meet the requirements. Single link and link bundle [RFC4201] have been widely used in today's IP/MPLS networks. A link bundle bundles a group of homogeneous links as a TE link to make routing approach more scalable. A composite link allows bundling non-homogenous links together as a single logical link. The motivations for using a composite link are described in the document [CL-REQ]. This document describes composite link framework in the context of MPLS network with MPLS control plane.

A composite link is a single logical link in MPLS network that contains multiple parallel component links between two routers. Unlike a link bundle [RFC4201], the component links in a composite link can have different properties such as cost or capacity. A composite link can transport aggregated traffic as other physical links from the network perspective and use its component links to carry the traffic internally.

Specific protocol solutions are outside the scope of this document.

2. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

2.1. Terminology

Composite Link: a group of component links, which can be considered as a single MPLS TE link or as a single IP link used for MPLS.

Component Link: a physical link (e.g., Lambda, Ethernet PHY, SONET/SDH, OTN, etc.) with packet transport capability, or a logical link (e.g., MPLS LSP, Ethernet VLAN, MPLS-TP LSP, etc.)

Traffic Flow: A set of packets with a common identifier and characteristics that is used by Composite link interior functions to place the set of packets on the same component link. Identifiers can be an MPLS label stack or any combination of IP addresses and protocol types for routing, signaling, and management packets.

Virtual Interface: Composite link is advertised as an interface in IGP

3. Composite Link Framework

A Composite Link in the context of MPLS network is a set of parallel links between two routers that form a single logical link within the network. Composite link model is illustrated in Figure 1, where a composite link is configured between routers R1 and R2. The composite link has three component links. Individual component links in a composite link may be supported by different transport technologies such as wavelength, Ethernet VLAN. Even if the transport technology implementing the component links is identical, the characteristics (e.g., bandwidth, latency) of the component links may differ.

As shown in Figure 1, the composite link may carry LSP traffic flows and control plane packets. A LSP may be established over the link by either RSVP-TE or LDP signaling protocols. All component links in a composite link have the same forwarding adjacency. The composite link forms one routing interface at the composite link end points for MPLS control plane. In other words, two routers connected via a composite link have forwarding adjacency and routing adjacency. Each component link only has significance to the composite link, i.e. it does not appear as a link in the control plane.

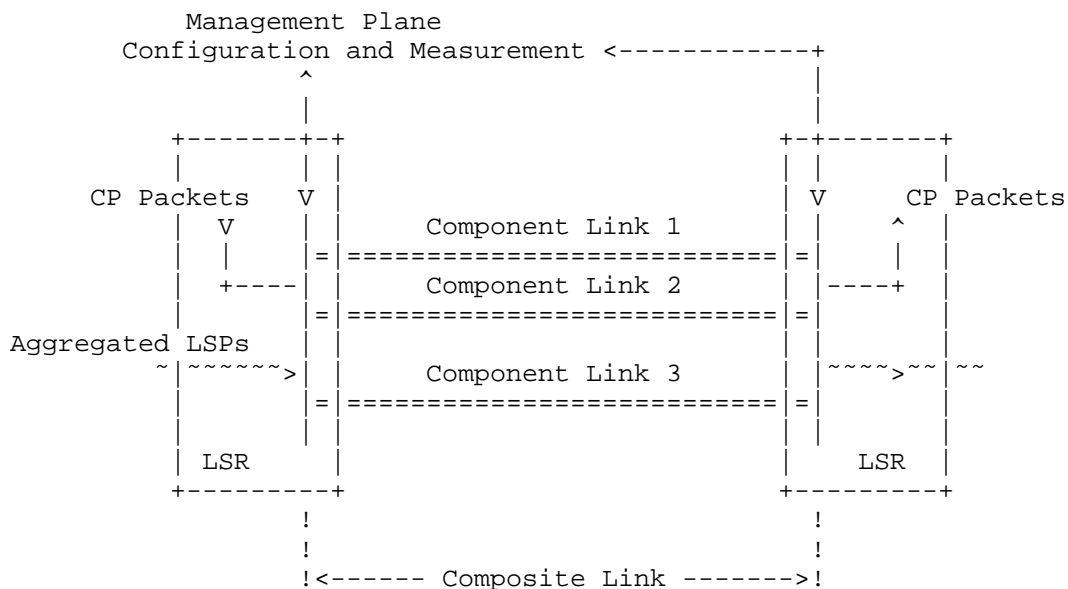


Figure 1 Composite Link Architecture Model

A component link in a composite link may be constructed in different ways.[CL-REQ] Figure 2 shows three common ways that may be deployed in a network.

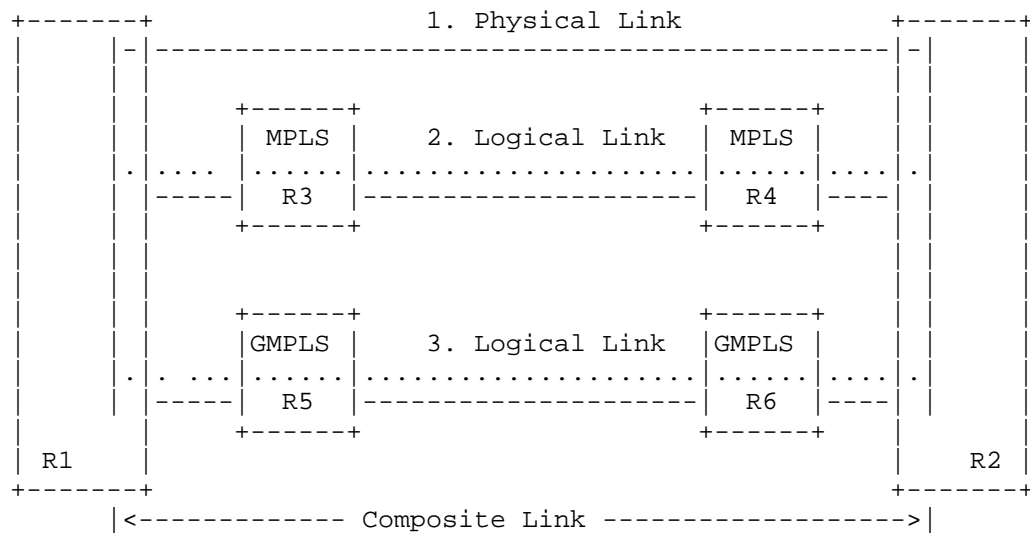


Figure 2 Illustration of Component Link Variances

As shown, the first component link is configured with direct physical media wire. The second component link is a TE tunnel that traverses R3 and R4. Both R3 and R4 are the nodes in the MPLS. The third component link is formed by lower layer network that has GMPLS enabled. In this case, R5 and R6 are not the nodes controlled by the MPLS but provide the connectivity for the component link.

Composite link forms one logical link between connected routers and is used to carry aggregated traffic.[CL-REQ] Composite link relies on its component links to carry the traffic over the composite link. This means that a composite link maps incoming traffic into component links. At the transmitting end (R1 in Figure 1), composite link maps a set of traffic flows including control plane packets to a specific component link. At the receiving end (R2 in Figure 1), composite link receives the packets from its component links and sends them to MPLS forwarding engine like a regular link.

Traffic mapping to component links may be done by control plane, management plane, or data plane.[CL-REQ] The objective is to keep the individual flow packets in sequence and do not overload any component link.[CL-REQ] Operator may have other objectives such as load balance over component links. A flow may be a LSP, or sub-LSP [MLSP], PW, a flow within PW [FAT-PW], entropy flow in LSP [Entropy].

4. Composite Link in Control Plane

A composite Link is advertised as a single logical interface between two connected routers, which forms routing and forwarding adjacency between the routers in IGP. The interface parameters for the composite link can be pre-configured by operator or be derived from

its component links. Composite link advertisement requirements are specified in [CL-REQ].

In IGP-TE, a composite link is advertised as a single TE link between two connected routers. This is similar to a link bundle [RFC4201]. Link bundle applies to a set of homogenous component links. Composite link allows homogenous and non-homogenous component links. The link bundle protocol extension for composite link advertisement is for further study.

Both LDP [RFC5036] and RSVP-TE [RFC3209] can be used to signal a LSP over a composite link. Since composite link capacity is aggregated capacity and is often larger than individual component link capacity, it is possible to signal a LSP whose BW is larger than individual component link capacity. [CL-REQ] Assumption is such LSP carrying an aggregated traffic.

A composite link may contain the set of component links. A component link may be configured by operator or signaled by the control plane. In both cases, it is necessary to convey component link parameters to the composite link. [CL-REQ]

When a component link is supported by lower layer network (third component link in figure 2), the control plane that the composite link resides is able to interoperate with the GMPLS or MPLS-TP control plane that lower layer network uses for component link addition and deletion. [CL-REQ]

5. Composite Link in Data Plane

Composite link may appear as one single interface or multiple interfaces in data plane.

The traffic over composite link is distributed over individual component links. Traffic dissemination may be determined by control plane, management plane, or data plane, and may be changed due to component link status change. [CL-REQ]

A component link in a composite link may fail independently. The composite link functions are able to recognize component link failure and re-assign impacted flows to other active component links in minimal disruptive manner. When a composite link can't recover some impacted flows, it notifies control plane about the flows. When a composite link is not able to transport all flows, it preempts some flows based upon local management configuration and informs the control plane on these preempted flows. This action ensures the remaining traffic is transported properly.

The composite link functions provide component link fault notification and composite link fault notification. Component link fault notification is sent to the management plane. Composite link fault notification is sent to the control plane and management

plane. Composite link allows operator to trace which component link a LSP is assigned to.

6. Security Considerations

For further study.

7. IANA Considerations

IANA actions to provide solutions are for further study.

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.

[RFC3209] D. Awduche, L. Berger, D. Gan, T. Li, V. Srinivasan, G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels," December 2001

[RFC4201] Kompella, K., "Link Bundle in MPLS Traffic Engineering", RFC 4201, March 2005.

[RFC5036] Andersson, L., "LDP Specification", RFC 5036 , October 2007.

8.2. Informative References

[CL-REQ] Villamizar, C. and McDysan, D, "Requirements for MPLS Over Composite Link", Oct. 2010, Work in Progress

[Entropy Label] Kompella, K. and S. Amante, "The Use of Entropy Labels in MPLS Forwarding", November 2008, Work in Progress

[FAT-PW] Bryan, S., et. Al, "Flow Aware Transport of Pseudowire over an MPLS PSN", draft-ietf-pwe3-fat-pw-04, Work in progress

[MLSP] Kompella, K. "Multi-path Label Switched Paths Signaled Using RSVP-TE", draft-kompella-mpls-rsvp-ecmp-00.txt, July 2010, Work in Progress

9. Acknowledgments

Authors would like to thank Adrian Farrel for his extensive comments and suggestions, Ron Bonica, Nabil Bitar, Eric Gray, Lou Berger, and Kireeti Kompella for their reviews and great suggestions.

Authors' Addresses

So Ning
Verizon
2400 N. Glem Ave.,
Richerdson, TX 75082
Phone: +1 972-729-7905
Email: ning.so@verizonbusiness.com

Andrew Malis
Verizon
117 West St.
Waltham, MA 02451
Phone: +1 781-466-2362
Email: andrew.g.malis@verizon.com

Dave McDysan
Verizon
22001 Loudoun County PKWY
Ashburn, VA 20147
Email: dave.mcdysan@verizon.com

Lucy Yong
Huawei USA
1700 Alma Dr. Suite 500
Plano, TX 75075
Phone: +1 469-229-5387
Email: lucyyong@huawei.com

Frederic Jounay
France Telecom
2, avenue Pierre-Marzin
22307 Lannion Cedex,
FRANCE
Email: frederic.jounay@orange-ftgroup.com

Yuji Kamite
NTT Communications Corporation
Granpark Tower
3-4-1 Shibaura, Minato-ku
Tokyo 108-8118
Japan
Email: y.kamite@ntt.com

