

PCE Working Group
Internet Draft
Intended status: Standard Track
Expires: April 20, 2014

Zafar Ali
Antonello Bonfanti
Cisco Systems
October 21, 2013

Resource ReserVation Protocol-Traffic Engineering (RSVP-TE)
Extension for Additional Signal Types in G.709 OTN
draft-ali-ccamp-additional-signal-type-g709v3-00.txt

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 20, 2014.

Copyright Notice

Copyright (c) 2013 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.

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.

Abstract

[I-D.draft-ietf-ccamp-gmpls-signaling-g709v3] provides the extensions to the Generalized Multi-Protocol Label Switching (GMPLS) signaling to control the full set of OTN features including ODU0, ODU4, ODU2e and ODUflex. However, it does not cover additional signal types mentioned in [G.Sup43] (ODU1e, ODU3e1, ODU3e2) or (ODU1f, ODU2f). This draft provides GMPLS signaling extension for these additional signal types.

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 RFC 2119 [RFC2119].

Table of Contents

1. Introduction	2
2. RSVP-TE extension for Additional Signal Types	3
3. Security Considerations	3
4. IANA Considerations	3
5. Acknowledgments	3
6. References	3
6.1. Normative References	3
6.2. Informative References	4

1. Introduction

[I-D.draft-ietf-ccamp-gmpls-signaling-g709v3] updates the ODU-related portions of [RFC4328] to provide Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) extensions to support control for [G.709-v3]. However, it does not cover additional signal types mentioned in [G.Sup43] (ODU1e, ODU3e1, ODU3e2) or (ODU1f and ODU2f).

With the evolution and deployment of Optical Transport Network (OTN) technology, it is necessary to support additional signal types mentioned in [G.Sup43] and (ODU1f and ODU2f). [I-D.draft-khuzema-ccamp-gmpls-signaling-g709] had support for signal types mentioned in [G.Sup43] but the signal types values collides with values defined in [I-D.draft-ietf-ccamp-gmpls-signaling-g709v3]. The draft has expired and also does not support ODU1f and ODU2f signal type.

Internet-Draft draft-ali-ccamp-additional-signal-type-g709v3-00.txt

This draft provides GMPLS signaling extension to support additional signal types mentioned in [G.Sup43] and (ODU1f and ODU2f).

2. RSVP-TE extension for Additional Signal Types

[I-D.draft-ietf-ccamp-gmpls-signaling-g709v3] defines the format of Traffic Parameters in OTN-TDM SENDER_TSPEC and OTN-TDM FLOWSPEC objects. The said traffic parameters have a signal type field. This document defines the signal type for ODU1e, ODU3e1, ODU3e2, ODU1f and ODU2f, as follows:

Value	Type
----	----
23	ODU1e (10Gbps Ethernet [GSUP.43])
24	ODU1f (10Gbps Fiber Channel)
25	ODU2f (10Gbps Fiber Channel)
26	ODU3e1 (40Gbps Ethernet [GSUP.43])
27	ODU3e2 (40Gbps Ethernet [GSUP.43])

3. Security Considerations

This document does not introduce any additional security issues above those identified in [I-D.draft-ietf-ccamp-gmpls-signaling-g709v3].

4. IANA Considerations

This document defines signal type for ODU1e, ODU3e1, ODU3e2, ODU1f and ODU2f to be carried in Traffic Parameters in OTN-TDM SENDER_TSPEC and OTN-TDM FLOWSPEC objects [I-D. draft-ietf-ccamp-gmpls-signaling-g709v3].

5. Acknowledgments

The authors would like to thank Sudip Shukla for comments.

6. References

6.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4328] Papadimitriou, D., Ed., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Extensions for G.709 Optical Transport Networks Control", RFC 4328, January 2006.

Internet-Draft draft-ali-ccamp-additional-signal-type-g709v3-00.txt

[G.709-v3] ITU-T, "Interface for the Optical Transport Network (OTN)", G.709/Y.1331 Recommendation, December 2009.

[GSUP.43] ITU-T, "Proposed revision of G.sup43 (for agreement)", December 2008.

[I-D.draft-ietf-ccamp-gmpls-signaling-g709v3] F.Zhang, G.Zhang, S.Belotti, D.Ceccarelli, K.Pithewan, "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Extensions for the evolving G.709 Optical Transport Networks Control, draft-ietf-ccamp-gmpls-signaling-g709v3, work in progress.

6.2. Informative References

[I-D.draft-khuzema-ccamp-gmpls-signaling-g709] Pithewan, K., et al, "Signaling Extensions for Generalized MPLS (GMPLS) Control of G.709 Optical Transport Networks", expired draft.

Authors' Addresses

Zafar Ali
Cisco Systems
Email: zali@cisco.com

Antonello Bonfanti
Cisco Systems
abonfant@cisco.com

CCAMP Working Group
Internet Draft
Intended status: Standard Track
Expires: April 18, 2014

Zafar Ali
George Swallow
Clarence Filsfils
Luyuan Fang
Cisco Systems
Kenji Kumaki
KDDI Corporation
Ruediger Kunze
Deutsche Telekom AG
Daniele Ceccarelli
Ericsson
Xian Zhang
Huawei
October 19, 2013

Resource ReserVation Protocol-Traffic Engineering (RSVP-TE)
Extension for Signaling Objective Function and Metric Bound
draft-ali-ccamp-rc-objective-function-metric-bound-04.txt

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 18, 2014.

Copyright Notice

Copyright (c) 2013 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.

Ali, Swallow, Filsfils

Expires April 2014

[Page 1]

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.

Abstract

In particular networks such as those used by financial institutions, network performance criteria such as latency are becoming critical to data path selection. However cost is still an important consideration. This leads to a situation where path calculation involves multiple metrics and more complex objective functions.

When using GMPLS control plane, there are many scenarios in which a node may need to request a remote node to perform path computation or expansion, like for example multi-domain LSP setup, Generalized Multi-Protocol Label Switching (GMPLS) User-Network Interface (UNI) or simply the utilization of a loose ERO in intra domain signaling. In such cases, the node requesting for the setup of an LSP needs to convey the required objective function to the remote node, to enable it to perform route computation in the desired fashion. Similarly, there are cases the ingress needs to indicate a TE metric bound for a loose segment that is expanded by a remote node.

This document defines extensions to the RSVP-TE Protocol to allow an ingress node to request the required objective function for the route computation, as well as a metric bound to influence route computation decisions at a remote node(s).

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 RFC 2119 [RFC2119].

Table of Contents

Copyright Notice.....	1
1. Introduction.....	3
2. RSVP-TE signaling extensions.....	4
2.1. Objective Function (OF) Subobject.....	4
2.1.1. Minimum TE Metric Cost Path Objective Function.....	6
2.1.2. Minimum IGP Metric Cost Path Objective Function.....	6
2.1.3. Minimum Latency Path Objective Function.....	6
2.1.4. Minimum Latency Variation Path Objective Function....	7
2.2. Metric subobject.....	7
2.3. Processing Rules for the OF Subobjects.....	8
2.4. Processing Rules for the Metric subobject.....	9
3. Security Considerations.....	11
4. IANA Considerations.....	11
5. Acknowledgments.....	12
6. References.....	12
6.1. Normative References.....	12
6.2. Informative References.....	12

1. Introduction

As noted in [OSPF-TE-METRIC] and [ISIS-TE-METRIC], in certain networks such as financial information networks (e.g. stock market data providers), performance criteria such as latency are becoming critical to data path selection along with other metrics. Such networks may require selection of a path that minimizes end-to-end latency. Or a path may need to be found that minimized some other TE metric(s), while subject to a latency bound. Thus there is a requirement to be able to find end-to-end paths with different optimization criteria.

When the entire route for an LSP is computed at the ingress node, this requirement can be met by a local decision at that node. However, there are scenarios where partial or full route computations are performed by remote nodes. The scenarios include (but are not limited to):

- . LSPs with loose hops in the Explicit Route Object (ERO), including intra-domain LSPs.
- . GMPLS-UNI where route computation may be performed by the UNI-Network (server) node [RFC4208];

- . Multi domain LSP setup with per domain path computation;

In these scenarios, there is a need for the ingress node to convey the optimization criteria (e.g., IGP cost, TE cost, hop counts, latency, etc.) to be used for the path computation to the node performing route computation or expansion. Similarly, there is a need for the ingress node to indicate a TE metric bound for the loose segment being expanded by a remote node.

[RFC5541] defines extensions to the Path Computation Element communication Protocol (PCEP) to allow a Path Computation Client (PCC) indicate in a path computation request the desired objective function. [RFC5440] and [ID-SERVICE-AWARE] defines extension to the PCEP to allow a PCC indicate in a path computation request a bound on given TE metric(s). This draft defines similar mechanisms for the RSVP-TE protocol allowing an ingress node to indicate in a Path request the desired objective function along with any associated TE metric bound(s). The nodes performing route expansion use this information to find the "best" candidate route.

2. RSVP-TE signaling extensions

This section defines RSVP-TE signaling extensions required to address the above-mentioned requirements. Two new ERO subobject types, Objective Function (OF) and Metric, are defined. Their purpose is as follows.

- . OF subobject conveys a set of one or more specific optimization criteria that needs be followed in expanding route of a TE-LSP in MultiProtocol Label Switching (MPLS) and GMPLS networks.
- . Metric Bound subobject indicates the bound on the path metric that needs to be observed for the loose segment to be considered as acceptable by the ingress node.

The scope of the Metric and OF subobjects is the node performing the expansion for loose ERO and the subsequent ERO subobject that identifies an abstract node. The following subsection provides the details.

2.1. Objective Function (OF) Subobject

A new ERO subobject type Objective Function (OF) is defined in order for the ingress node to indicate the required objective function on a loose hop. The ERO subobject type OF is optional. It MAY be carried within an ERO object of RSVP-TE Path message

and its scope is limited to previous ERO subobject that identifies an abstract node. For more details please refer to the Processing Rules for the OF Subobjects section.

The OF subobject has the following format:

0										1										2										3																			
0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1																		
+-----+										+-----+										+-----+										+-----+																			
L										Type										Length										OF Code										Reserved									
+-----+										+-----+										+-----+										+-----+										+-----+									

The fields of OF subobject are defined as follows:

L bit: The L bit MUST be set to represent a loose hop in the explicit route.

Type: The Type is to be assigned by IANA (suggested value: 66).

Length: The Length contains the total length of the subobject in bytes, including the Type field, the Length field. The Length of the subobject is 4.

OF Code (1 byte): The identifier of the objective function. The following OF code values are suggested. These values are to be assigned by IANA.

* OF code value 0 is reserved.

* OF code value 1 (to be assigned by IANA) is for Minimum TE Metric Cost Path (MTMCP) OF defined in this document. See definition of MTCP OF in the following.

* OF code value 2 (to be assigned by IANA) is for Minimum Interior Gateway Protocol (IGP) Metric Cost Path (MIMCP) OF defined in the following.

* OF code value 3 (to be assigned by IANA) is for Minimum Load Path (MLP) OF as defined in RFC5541.

* OF code value 4 (to be assigned by IANA) is for Maximum Residual Bandwidth Path (MBP) OF as defined in RFC5541.

* OF code value 5 (to be assigned by IANA) is for Minimize Aggregate Bandwidth Consumption (MBC) OF as defined in RFC5541.

* OF code value 6 (to be assigned by IANA) is for Minimize the Load of the most loaded Link (MLL) OF as defined in RFC5541.

* OF code value 7 is skipped (to keep the objective function code values consistent between [RFC5541] and this draft.

* OF code value 8 (to be assigned by IANA) is for Minimum Latency Path (MLP) OF defined in this document. See definition of MLP OF in the following.

* OF code value 9 (to be assigned by IANA) is for Minimum Latency Variation Path (MLVP) OF defined in this document. See definition of MLVP OF in the following.

Other objective functions may be defined in future.

Reserved (5 bytes): This field MUST be set to zero on transmission and MUST be ignored on receipt.

2.1.1. Minimum TE Metric Cost Path Objective Function

Minimum TE Metric Cost Path (MTMCP) OF is defined as an Objective Function where a path is computed such that the sum of the TE metric of the links along the path is minimized. In the context of loose hop expansion, the ERO expanding node MUST try to find a route such that the sum of the TE metric of the links along the route is minimized.

2.1.2. Minimum IGP Metric Cost Path Objective Function

Minimum IGP Metric Cost Path (MIMCP) OF is defined as an Objective Function where a path is computed such that the sum of the IGP metric of the links along the path is minimized. In the context of loose hop expansion, the ERO expanding node MUST try to find a route such that the sum of the IGP metric of the links along the route is minimized.

2.1.3. Minimum Latency Path Objective Function

Minimum Latency Path (MLP) OF is defined as an Objective Function where a path is computed such that latency of the path is minimized. In the context of loose hop expansion, the ERO expanding node MUST try to find a route such that overall latency of the loose hop is minimized.

2.1.4. Minimum Latency Variation Path Objective Function

Minimum Latency Variation Path (MLVP) OF is defined as an Objective Function where a path is computed such that latency variation in the path is minimized. In the context of loose hop expansion, the ERO expanding node MUST try to find a route such that overall latency variation of the loose hop is minimized.

2.2. Metric Bound subobject

The ERO subobject type Metric Bound (MB) is optional. It MAY be carried within an ERO object of RSVP-TE Path message and its scope is limited to previous ERO subobject that identifies an abstract node. It is possible to identify different Metric Bound subobjects for different hops of the ERO to be expanded. For more details please refer to the Processing Rules for the Metric Bound Subobjects section.

This subobject has the following format:

```

0                               1                               2                               3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|L|      Type      |      Length      | metric-type |B|      Reserved      |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     metric-bound                                     |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

The fields of the Metric subobject are defined as follows:

L bit: The L bit is set if the subobject represents a loose hop in the explicit route. If the bit is not set, the subobject represents a strict hop in the explicit route. Please note that use of MB subobject is also applicable to strict hops, e.g., in selecting a component link within a heterogeneous bundled TE link.

Type: The Type is to be assigned by IANA (suggested value: 67).

Length: The Length is 8.

Metric-type (8 bits): Specifies the metric type associated with the partial route expended by the node processing the loose ERO. The following values are currently defined:

- * T=1: cumulative IGP cost
- * T=2: cumulative TE cost
- * T=3: Hop Counts
- * T=4: Cumulative Latency
- * T=5: Cumulative Latency Variation

B bit: Best-effort bit. When the best-effort (B) bit is set, it means that the ingress allows for the set up of an LSP that does not meeting the MB requirement. When the best-effort (B) bit is not set, it means that the MB needs to be observed.

Reserved: This field MUST be set to zero on transmission and MUST be ignored on receipt.

Metric-bound (32 bits): The metric-bound indicates an upper bound for the path metric that MUST NOT be exceeded for the ERO expending node to consider the computed path as acceptable. The metric bound is encoded in 32 bits using IEEE floating point format as defined in [IEEE.754.1985]). When it indicates a time value (i.e. Latency or Latency Variation) it is expressed in milliseconds.

2.3. Processing rules

A single OF subobjects SHOULD be used between a pair of abstract nodes in ERO. Multiple Metric Bound subobjects MAY be indicated for each hop to be expanded and MUST be placed after each abstract node subobject. Different Metric Bounds MAY be identified for each hop expansion.

2.3.1. Processing Rules for the OF Subobjects

The basic processing rules of an ERO are not altered. Please refer to [RFC3209] for details.

The scope of the OF subobject is the previous ERO subobject that identifies an abstract node, and the subsequent ERO subobject that identifies an abstract node. Multiple OF subobjects may be present between any pair of abstract nodes. However, only first OF subobject is analyzed and others are ignored.

The following conditions SHOULD result in Path Error with error code "Routing Problem" and error subcode "Bad EXPLICIT_ROUTE object":

- . If the first OF subobject is not preceded by an ERO subobject identifying the next hop.
- . If the OF subobject follows an ERO subobject identifying the next hop that does not have the L-bit set.

If the processing node does not understand the OF subobject, it SHOULD send a PathErr with the error code "Routing Error" and error value of "Bad Explicit Route Object" toward the sender [RFC3209].

If the processing node understands the OF subobject and the ERO passes the above mentioned sanity check and any other sanity checks associated with other ERO subobjects local to the node, the node takes the following actions:

- . If the node supports the requested OF, the node expands the loose hop using the requested OF as optimization criterion for computing the route to the next abstract node. After processing, the OF subobjects are removed from the ERO. The rest of the steps for the loose ERO processing follow procedures outlined in [RFC3209].
- . If the node understands the OF subobject but does not support the requested OF, it SHOULD send a Path Error with error code "Routing Problem" and a new error subcode "Unsupported Objective Function". The error subcode "Unsupported Objective Function" for Path Error code "Routing Problem" is to be assigned by IANA.
- . If the OF is supported but policy does not permit applying it, the processing node SHOULD send a Path Error with error code "Policy control failure" (value 2) and subcode "objective function not allowed". The error subcode "objective function not allowed" for Path Error code "Policy control failure" is to be assigned by IANA.

2.3.2. Processing Rules for the MB subobject

The basic processing rules of an ERO are not altered. Please refer to [RFC3209] for details.

The scope of the MB subobject is between the previous ERO subobject that identifies an abstract node, and the subsequent ERO subobject that identifies an abstract node. Multiple MB subobjects may be present between any pair of abstract nodes.

If the processing node does not understand the MB subobject, it SHOULD send a PathErr with the error code "Routing Error" and error value of "Bad Explicit Route Object" toward the sender [RFC3209].

If the processing node understands the MB subobject and the ERO passes the above mentioned sanity check and any other sanity checks associated with other ERO subobjects local to the node, the node takes the following actions:

- . For all the MB subobject(s), the node expands the ERO such that the requested metric bound(s) are met for the route between the two abstract nodes in the ERO. After processing, the Metric subobjects are removed from the ERO. The rest of the steps for the ERO processing follow procedure outlined in [RFC3209].
- . If the node understands the MB subobject but cannot find a route to the next abstract node such that the requested metric bound(s) can be satisfied and the best-effort (B) bit is not set, it SHOULD send a Path Error with error code "Routing Problem" and a new error subcode "No route available toward destination with the requested metric bounds". The error subcode "No route available toward destination with the requested metric bounds" for Path Error code "Routing Problem" is to be assigned by IANA (See IANA section for details).
- . If the node understands the Metric subobject but cannot find a route to the next abstract node such that the requested metric bound(s) can be satisfied and the best-effort (B) bit is set, it SHOULD send a Path Error message with error code "Notify Error" and a new error subcode "Route not matching the requested metric bounds" is to be assigned by IANA (See IANA section for details).
- . The ERO expanding node SHOULD respect the Metric Bound constraints in realizing any segment recovery procedure to change the route of the segment expanded by the said node. If

best-effort (B) bit is set and the new recovery segment violates the Metric Bound constraints, the ERO expanding SHOULD send a Path Error message with error code "Notify Error" and a new error subcode "Route not matching the requested metric bounds" is to be assigned by IANA (See IANA section for details).

3. Security Considerations

This document does not introduce any additional security issues above those identified in [RFC5920], [RFC2205], [RFC3209], and [RFC3473].

4. IANA Considerations

4.1. ERO Subobject

This document adds the following two new subobject of the existing entry for ERO (20, EXPLICIT_ROUTE):

Value	Description
-----	-----
TBA (suggest value: 66)	Objective Function (OF) subobject
TBA (suggest value: 67)	Metric subobject

These subobject may be present in the Explicit Route Object, but not in the Route Record Object.

OF Code values carried in OF subobject requires an IANA entry with suggested values as defined in section 2.1.

4.2. New RSVP error sub-code

For Error Code = 24 "Routing Problem" (see [RFC2205]) the following sub-code is defined.

Sub-code	Value
-----	-----
No route available toward destination with the requested metric bounds	To be assigned by IANA. Suggested Value: TBA.

For Error Code = 25 "Notify Error" (see [RFC2205]) the following sub-code is defined.

ID draft-ali-ccamp-rc-objective-function-metric-bound-04.txt

Sub-code -----	Value -----
Route not matching the requested metric bounds	To be assigned by IANA. Suggested Value: TBA.

5. Acknowledgments

Authors would like to thank Matt Hartley, Ori Gerstel, Gabriele Maria Galimberti, Luyuan Fang and Walid Wakim for their review comments.

6. References

6.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC3473] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.
- [IEEE.754.1985] IEEE Standard 754, "Standard for Binary Floating-Point Arithmetic", August 1985.
- [RFC4208] Swallow, G., Drake, J., Ishimatsu, H., and Y. Rekhter, "Generalized Multiprotocol Label Switching (GMPLS) User-Network Interface (UNI): Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Support for the Overlay Model", RFC 4208, October 2005.

6.2. Informative References

- [RFC5920] Fang, L., Ed., "Security Framework for MPLS and GMPLS Networks", RFC 5920, July 2010.
- [RFC5440] Vasseur, JP., Ed., and JL. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.
- [RFC5541] Le Roux, JL., Vasseur, JP., and Y. Lee, "Encoding of Objective Functions in the Path Computation Element Communication Protocol (PCEP)", RFC 5541, June 2009.
- [ID-SERVICE-AWARE] D. Dhody, V. Manral, Z. Ali, G. Swallow, K. Kumaki, " Extensions to the Path Computation Element Communication Protocol (PCEP) to compute service aware Label Switched Path (LSP)", draft-ietf-pce-pcep-service-aware, work in progress.
- [OSPF-TE-METRIC] S. Giacalone, D. Ward, J. Drake, A. Atlas, S. Previdi, "OSPF Traffic Engineering (TE) Metric Extensions", draft-ietf-ospf-te-metric-extensions, work in progress.
- [ISIS-TE-METRIC] S. Previdi, S. Giacalone, D. Ward, J. Drake, A. Atlas, C. Filsfils, "IS-IS Traffic Engineering (TE) Metric Extensions", draft-previdi-isis-te-metric-extensions, work in progress.

Author's Addresses

Zafar Ali
Cisco Systems.
Email: zali@cisco.com

George Swallow
Cisco Systems.
swallow@cisco.com

Clarence Filsfils
Cisco Systems.
cfilsfil@cisco.com

Luyuan Fang
Cisco Systems.
lufang@cisco.com

ID draft-ali-ccamp-rc-objective-function-metric-bound-04.txt

Kenji Kumaki
KDDI Corporation
Email: ke-kumaki@kddi.com

Rudiger Kunze
Deutsche Telekom AG
Ruediger.Kunze@telekom.de

Daniele Ceccarelli
Ericsson
Email: daniele.ceccarelli@ericsson.com

Xian Zhang
Huawei Technologies
Email: zhang.xian@huawei.com

INTERNET-DRAFT
Intended Status: Informational
Expires: April 17, 2014

Snigdho Bardalai
Khuzema Pithewan
Rajan Rao
Infinera Corp.
October 14, 2013

Overlay Network - Path Computation Approaches
draft-bardalai-ccamp-overlay-path-comp-02

Abstract

This document discusses various path computations approaches which are applicable to overlay networks [framework doc ref]. It discusses how the customer edge nodes uses the information advertised by the provider network to compute a path between two customer edge nodes or how it can request the provider network to compute a path and setup an end-2-end LSP.

Status of this Memo

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/lid-abstracts.html>

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

Copyright and License Notice

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

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

Table of Contents

1	Introduction	3
1.1	Terminology	3
2.	Network Configuration	3
3.	Network Configuration Usecases	4
3.1	ONI is located between CE-PE nodes.	4
3.2	ONI is located between CE and PE nodes.	4
3.3	Nested ONIs	5
4.	Path Computation Use-cases	6
5.	Path Computation Approaches	7
5.1	Virtual Topology Approach	8
5.2	PCE Approach	9
5.3	Hybrid Approach	11
6.	CE-PE / PE-PE Interface	12
7	Security Considerations	12
8	IANA Considerations	12
9	References	12
9.1	Normative References	12
9.2	Informative References	12
	Authors' Addresses	13

1 Introduction

This document attempts to describe possible ways to advertise information required for customer network CE nodes to compute a path for LSPs between two points in two customer network islands connected by a provider network, so as to adhere a set of constraints in provider network without knowledge of the detailed provider network topology. These constraints could be, but not limited to, diversity, latency, jitter, skew etc. Connectivity between customer network islands is presumed to be an "overlay" over provider network.

1.1 Terminology

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. Network Configuration

Multi-layer, multi-domain network typically involve overlay boundaries, where routing information sharing is restricted in nature. These are typically administrative boundaries coupled with technology boundaries.

Overlay network boundaries can be envisioned on two axes.

a. Technology Boundary : This typically involves different types of switching technologies i.e. Packet, OTN, DWDM. These technologies are also known as client or server technologies. Client technologies are typically enabled by Packet, OTN switching, while server technologies are enabled by OTN, DWDM technologies.

b. Administrative Boundary: This boundaries are enforced by administrative contracts that bars exchange of routing information for operational reasons, hence creating a need for special mechanism that facilitates circuit provisioning in such environment.

Customer and Provider domains are the examples of distinct administrative domains.

Intersecting a and b will give us following unique network configurations

UseCase i : Tech boundary coincides with administrative boundary
UseCase ii : Tech boundary is part of provider domain
UseCase iii : stacking of UNI interfaces in provider domain.

following section discuss these usecases in more detail.

3. Network Configuration Usecases

In this section, ONI, overlay network interface terminology is used to indicate the administrative boundary that imposes restriction on routing information exchange. Client layer is assumed to be using packet/OTN technologies while server layer could be Packet, OTN, DWDM etc. the technology transition could be in customer or provider network.

C is referred to as customer network node and P is referred to as provider network node. CE is referred to as Customer Edge and PE is referred to as Provider edge.

3.1 ONI is located between CE-PE nodes.

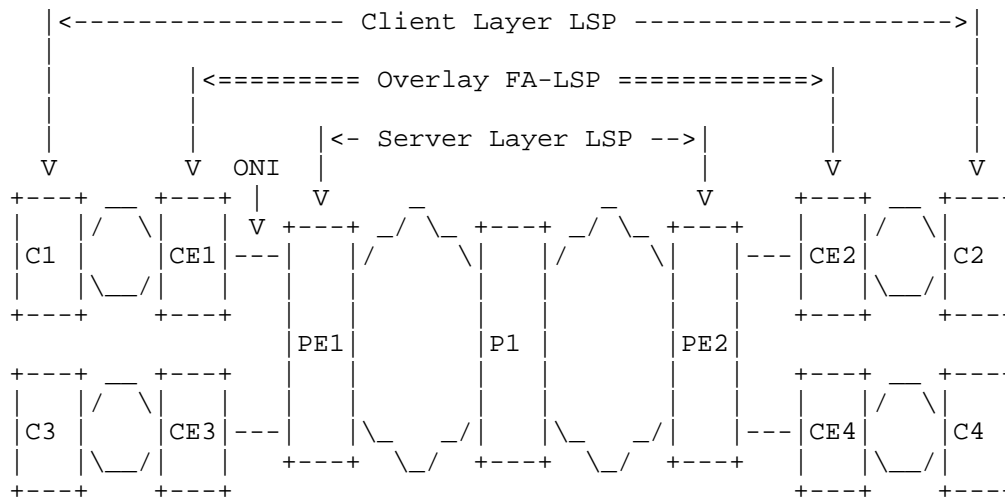


Figure. 1

Here server layer is assumed to be OTN/DWDM. There are couple of scenarios possible here :

- i. CE-PE link could be Packet Link, so layer transition from Packet to OTN/DWDM will happen in PE node
- ii. CE-PE link could be OTN/DWDM link, so layer transition from packet to OTN/DWDM will happen in CE node

3.2 ONI is located between CE and PE nodes.

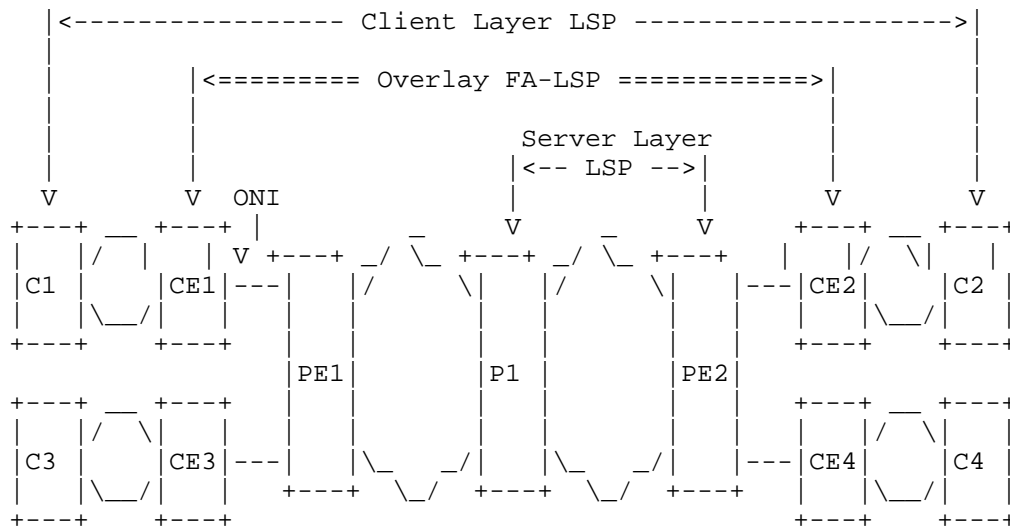
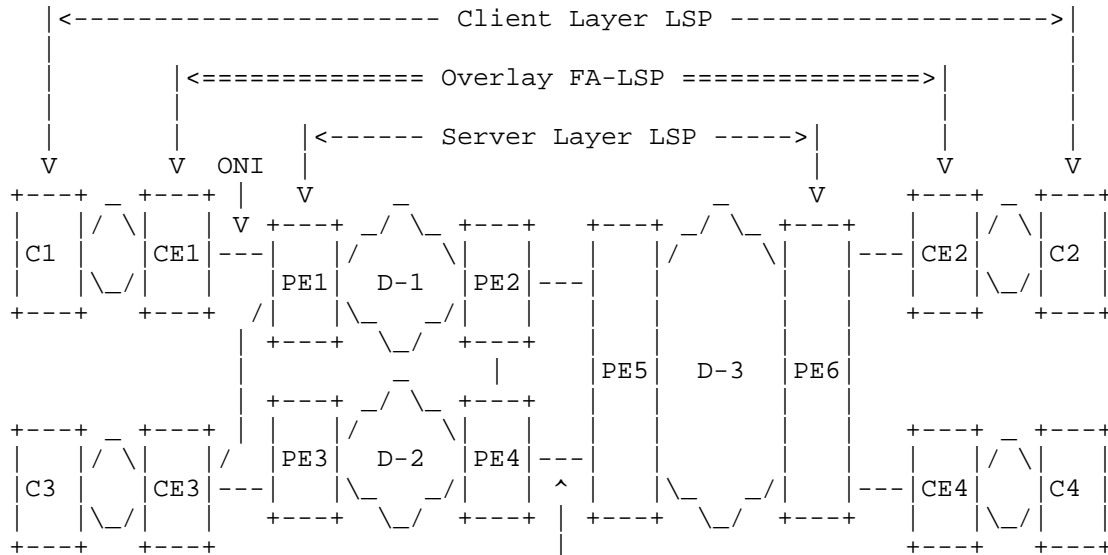


Figure. 2 In Figure 2, the Packet switching continues from customer to provider network and transitions to OTN/DWDM at P1. This kind of configuration is possible in multi-party client and server network, where the provider operates multi-layer network and provide services to its customers.

3.3 Nested ONIs

This is multi-layer network having ONIs between CE and PE, and also between PE and PE (PE2/4 - PE5)



ONI
Figure. 3

Because of multiple server layer technologies, it is possible that a layer closer to packet layer is digital (OTN), which is supported by pure optical layer (DWDM) to achieve better aggregation and improved restoration and protection capabilities.

In this configuration it is assumed that digital layer is playing dual role of customer to provider of optical layer and provider to customer that operates packet layer. In figure 3, domains D-1 and D-2 can be assumed to be digital layer, which is interfacing with packet layer through ONI between PE and CE. Digital domains D-1 and D-2 are also interfacing with optical D-3, again through ONI. If OTN and DWDM multi-layer network belongs to same IGP, then this becomes a multi-layer path-computation and signaling case, and it is out of scope of this document.

4. Path Computation Use-cases

In case of overlay networks it is required to compute a path between the customer edge nodes for the overlay FA-LSP as shown in the figure 4.

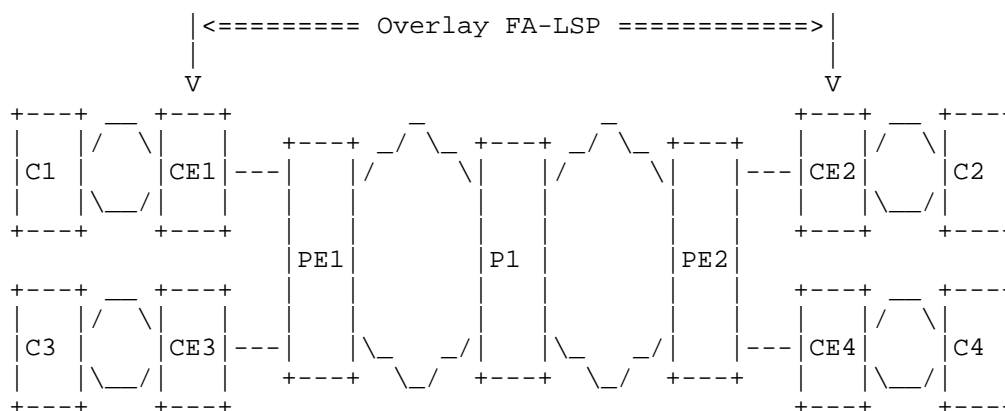


Figure. 4

The typical path computation use-cases are the following:

1. Point-to-point overlay path.
2. Multiple point-to-point diverse overlay paths sharing common LSP head and tail ends.

3. Multiple point-to-point diverse overlay paths that do not share common LSP head and tail ends.
4. Point-to-multipoint overlay paths.
5. Overlay paths over multi-domain (i.e. Multi-area or multi-AS) provider networks.

The typical TE constraints are:

1. Bandwidth or resource (this is technology specific).
2. Include or exclude nodes/links/SRLG or paths identified by path-keys.
3. Latency, jitter, max-hop requirements.
4. Optimization options - minimize cost, minimize latency etc.

5. Path Computation Approaches

There are three path computation approaches

1. Virtual-topology approach
2. PCE approach
3. Hybrid approach - combined virtual topology and PCE approach

5.1 Virtual Topology Approach

This path computation approach uses a virtual topology that is advertised by the provider network by the customer edge nodes.

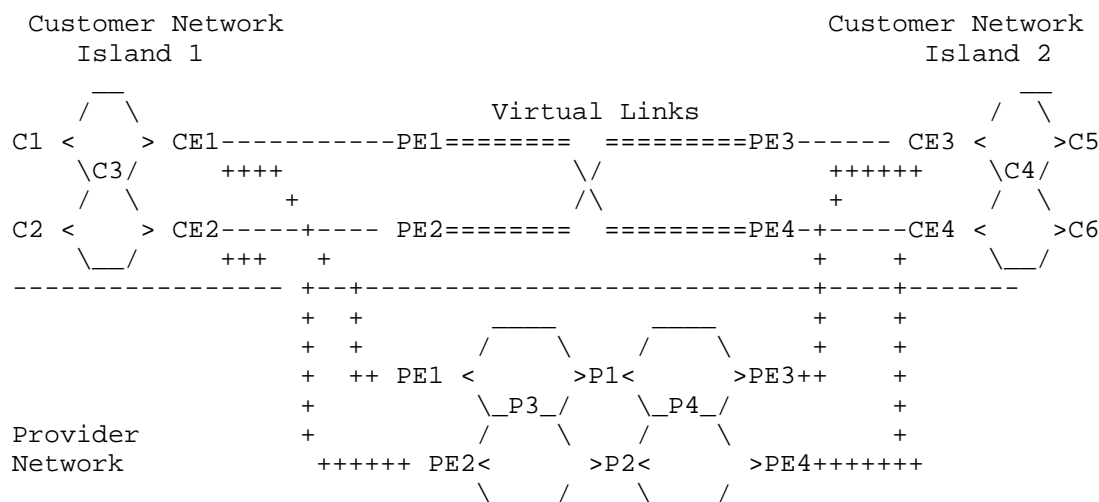


Figure. 5

In Figure 5, Provider Network has 4 interconnected rings supports full node diversity to connect any 2 Provider Edge Nodes.

PE1, PE2, PE3, PE4 are provider edge nodes.

P1, P2, P3, P4 are internal provider Network nodes, that must not be known to the customer network.

Customer Network has two islands connected by provider network.

C1, C2, C3, C4, C5, C6 are internal customer network nodes.

CE1, CE2, CE3, CE4 are customer network edge nodes connected to provider network edge nodes PE1, PE2, PE3, PE4.

Virtual Link Set : Virtual Link set is defined as set of one or more virtual links between any two provider edge nodes. The virtual links in the virtual link set, when realized may take different paths within provider domain, having different SRLGs and other TE metrics.

Above example topology has following Virtual Link Sets

a/ [PE1, PE2]

b/ [PE1, PE3]

c/ [PE1, PE4]

d/ [PE2, PE3]

e/ [PE2, PE4]
f/ [PE3, PE4]

The PEs in provider network do full peering with its attached CEs for virtual topology. So provider network virtual Links along with its SRLG IDs and other TE metrics are advertised into customer network.

Customer network internal Nodes C1..C6 can see provider network virtual TE Links and can compute paths between two points in customer network islands across provider network satisfying required diversity and TE metrics.

5.2 PCE Approach

An alternative approach for a CE node to obtain a path to another remote CE node would be by making a request to a provider network PCE. This approach requires either provider network PE nodes to advertise the PCE's IP address to CE nodes or CE Nodes should be configured with Provider Network PCE IP address. CE nodes needs to advertise the TE link-state of the CE-PE interface. This allows the PCE to build the overlay network topology link-state data-base.

In Figure. 1 above, the example depicted shows the provider network with a single IGP area and the provider network PCE has visibility to the detailed topology and TE information representing the server layer forwarding plane plus the CE-PE interface link-states that have been learned from the CE nodes. The server layer topology in addition to the CE-PE interface link-states constitutes the overlay network topology.

Figure. 2 above shows the case in which the provider network is a multi-layer network and the server layer boundary does not coincide with the provider network boundary. Again, the provider network PCE can have visibility to a single IGP area as described for MLN or alternatively there could be multiple IGP instances as described in [RFC6107], one instance for the overlay network and another instance for the server layer.

Figure. 3 above shows a multi-area or multi-AS provider network (generalized as a multi-domain provider network in this document). For multi-domain networks a hierarchical PCE could be deployed and the IP address of the hierarchical PCE is advertised to the CE nodes. The hierarchical PCE could maintain a multi-domain virtual topology instead of detailed topology of each domain.

In all three cases the head-end CE node is assumed to be aware of the address in the remote CE node for which the path is to be computed.

The exact manner by which this knowledge becomes available is beyond the scope of this document. The head-end CE node then makes a request to the provider network PCE with the remote address and the required set of TE constraints that need to be satisfied by the computed path.

In each case of the provider networks PCE uses the overlay network topology to compute the path. In case of the provider network example shown in Figure. 4 the hierarchical PCE computes the domain-level or inter-domain path first and then computes the intra-domain paths. The exact mechanism could be using the BRPC procedure in order to compute optimal intra-domain paths.

Once the computation is complete the PCE responds back with the path. The path generated by the PCE is expected to contain both real and virtual links and nodes. In case there is a need to maintain confidentiality with respect to the details of the provider network topology from the customer network then the response can include a path-key. In case there is a need to compute diverse paths one of two approaches could be followed - simultaneous computation approach in which case the response will have multiple paths or path-keys or the request could include the exclude hops or exclude path-key.

In the example below the procedure of computing a set of diverse paths using the PCE approach is explained.

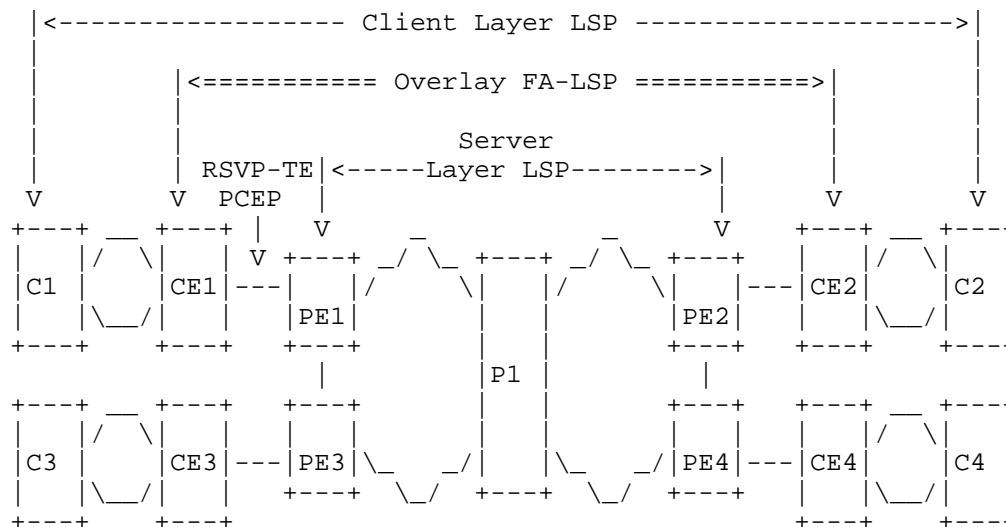


Figure. 6

Step-1: CE1 requests computation of diverse paths between PE1-PE2 and PE3-PE4.

Step-2: PE1 responds with 2 sets of EROs or 2 path-keys.

Step-3: CE1 initiates signaling of LSP PE1-PE2 with ERO or path-key.

Step-4: ERO or path-key is transferred to CE3.

Step-5: CE3 initiates signaling of LSP PE3-PE4 with ERO or path-key.

This approach uses a single computation for a pair of diverse paths. An alternative approach is by computing diverse paths separately as follows:

Step-1: CE1 requests computation of a path between PE1-PE2.

Step-2: PE1 responds with a set of EROs or a path-key.

Step-3: CE1 initiates signaling of LSP PE1-PE2 with ERO or path-key.

Step-4: PE1-PE2 path ERO or path-key is transferred to CE3.

Step-5: CE3 request computation of a path between PE3-PE4 with XRO(= PE1-PE2 ERO or path-key).

Step-6: PE3 responds with a set of EROs or path-key.

Step-7: CE3 initiates signaling of LSP PE3-PE4 with ERO or path-key.

5.3 Hybrid Approach

In the absence of a hierarchical PCE for a multi-domain provider network, it is possible a CE node learns of multiple PCE IP addresses from multiple PE nodes. This is possible in case each PE node lies in separate areas or ASs and with PCEs deployed per-area or per-AS. In such a situation it will be necessary for the CE node to pick one of the PCEs to send the path computation request. One way to select the appropriate PCE would be to advertise a virtual-topology associated with each PCE IP address to provide sufficient information for the CE node to determine whether a path to the remote CE address can be computed by the specific PCE.

In Figure. 4 above, CE3 has a dual-homed connectivity with the multi-domain provider network (i.e. CE3 to D-1 and D-2 via PE1 and PE3

respectively). In the absence of a hierarchical PCE, PE1 can advertise a virtual topology with connectivity to a set of CE nodes. Similarly PE3 advertises a virtual topology with connectivity to another set of CE nodes. This can happen in cases when there is no available bandwidth to a specific CE node via a specific domain. CE3 can determine using the virtual topologies which PCE should it send the path computation request.

6. CE-PE / PE-PE Interface

The CE-PE or PE-PE interface requires a routing interface in order to be able to exchange topology information and a path-computation interface in order to be able to send path computation requests and responses. For signaling the overlay LSP a signaling interface is required as well.

7 Security Considerations

TBD

8 IANA Considerations

TBD

9 References

9.1 Normative References

- [KEYWORDS] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC1776] Crocker, S., "The Address is the Message", RFC 1776, April 1 1995.
- [TRUTHS] Callon, R., "The Twelve Networking Truths", RFC 1925, April 1 1996.

9.2 Informative References

- [EVILBIT] Bellovin, S., "The Security Flag in the IPv4 Header", RFC 3514, April 1 2003.
- [RFC5513] Farrel, A., "IANA Considerations for Three Letter Acronyms", RFC 5513, April 1 2009.
- [RFC5514] Vyncke, E., "IPv6 over Social Networks", RFC 5514, April 1

2009.

Authors' Addresses

Snigdho Bardalai
sbardalai@infinera.com

Rajan Rao
rrao@infinera.com

Khuzema Pithewan
kpithewan@infinera.com

CCAMP Working Group
Internet Draft
Intended status: Standards Track

Igor Bryskin (Ed)
Wes Doonan
ADVA Optical Networking
Vishnu Pavan Beeram (Ed)
John Drake (Ed)
Gert Grammel
Juniper Networks
Manuel Paul
Ruediger Kunze
Deutsche Telekom
Friedrich Armbruster
Cyril Margaria
Coriant GmbH
Oscar Gonzalez de Dios
Telefonica
Daniele Ceccarelli
Ericsson

Expires: March 12, 2014

September 12, 2013

Generalized Multiprotocol Label Switching (GMPLS) External Network
Network Interface (E-NNI): Virtual Link Enhancements for the
Overlay Model
draft-beeram-ccamp-gmpls-enni-03.txt

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), 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 March 12, 2014.

Copyright Notice

Copyright (c) 2012 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.

Abstract

This memo is a companion document to [RFC4208]. It describes how the client domain networking in the overlay model can be enhanced via presenting to the client the network domain as an overlay topology made of Virtual TE Links.

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 RFC-2119 [RFC2119].

Table of Contents

1. Introduction.....	3
2. Hybrid Topology.....	3
3. Traffic Engineering.....	7
3.1. Augmenting the Client layer Topology.....	11
3.1.1. Virtual TE Links.....	13
3.2. Macro SRLGs.....	15
3.3. MELGs.....	17
3.4. Switching Constraints.....	18
4. GMPLS ENNI and Multiple Server Network Domains.....	19
5. Path computation aspects.....	21
6. Access and Virtual TE link addressing.....	22
7. Use cases.....	22
7.1. Service Optimization and Restoration in Multi-layer networks.....	22

7.2. IP/MPLS Offloading with ENNI automation.....	23
7.3. Use of PCE and VNTM in Multi-layer Network Operation.....	24
8. Security Considerations.....	25
9. IANA Considerations.....	25
10. References.....	25
10.1. Normative References.....	25
10.2. Informative References.....	25
11. Acknowledgments.....	26

1. Introduction

[RFC4208] discusses how GMPLS can be applied to the overlay model, which it defines to be a client network that uses a server network to dynamically instantiate LSPs between the client network's nodes. In the client network such an LSP is a link between two adjacent client nodes, while in the server network the LSP may transit multiple links and nodes; the client network is unaware of the server network topology.

While the client network is unaware of the server network topology, [RFC4208] does suggest that there may be an exchange of routing information between the server network and the client network. Building on this premise, this memo describes how introducing a representation of server network domain resources into a client network domain topology enhances client networking in the overlay model

This document is designed to be a companion document to [RFC4208], but because routing is generally not considered to be part of the definition of a UNI, this document uses the term 'External Network Network Interface (E-NNI)'. 'E-NNI' is generally used to indicate a control plane (routing and signaling) reference point for exchange of information between two control plane instances. In this document, the term 'ENNI' is specifically used to describe the interface between two network domains that allows the exchange of routing and signaling information.

2. Hybrid Topology

Two adjacent domains in the overlay model represent, generally speaking, regions of dissimilar transport technology. When an end-to-end service crosses a boundary between the domains, it is necessary to execute distinct forms of service activation within each domain.

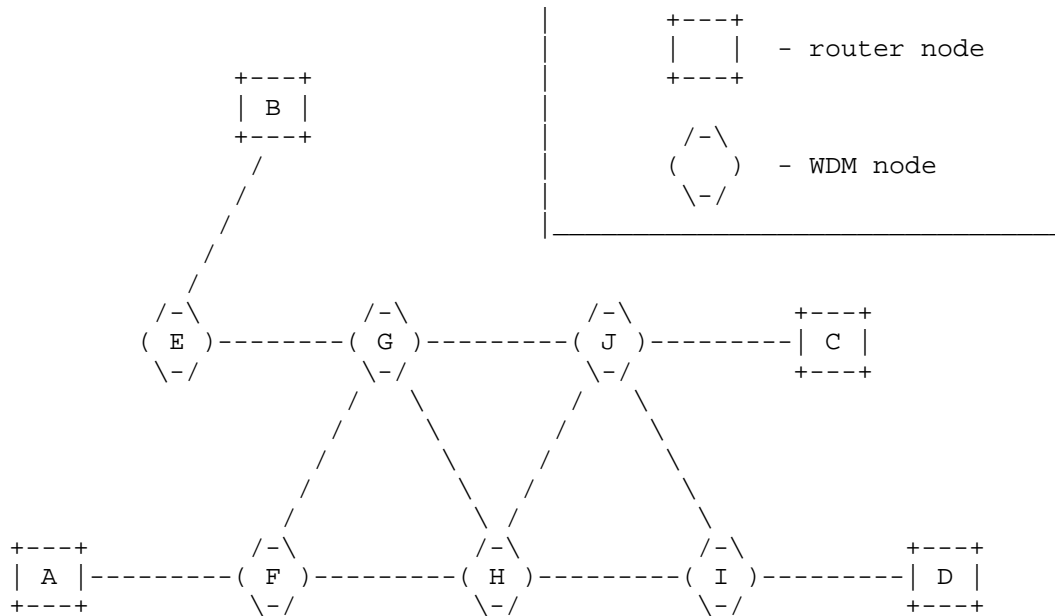


Figure 1: Sample hybrid topology

For example, in the hybrid network illustrated in Fig 1, provisioning a transport service between two GMPLS-enabled IP routers (clients) on either side of the optical WDM transport topology (server network domain) requires operations in two distinct layer networks; the client layer network interconnecting the routers themselves, and the server layer network interconnecting the optical transport elements in between the routers.

The activation of the end-to-end service begins with a path determination process, followed by the initiation of a signaling process from the ingress client network element along the determined path, per the example illustrated in Fig 2a-c.

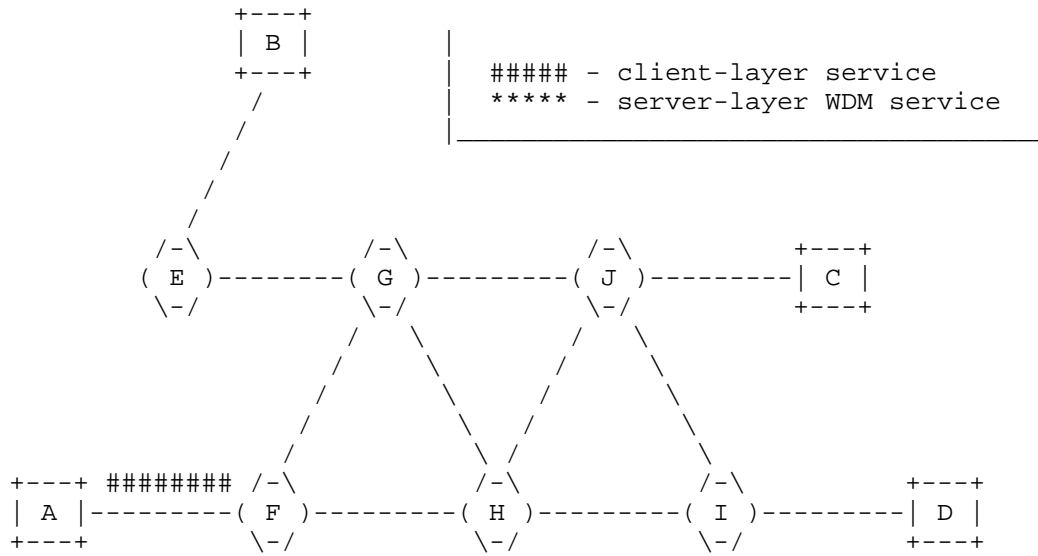


Figure 2a: Hierarchical service activation -
Client-layer service setup is initiated.

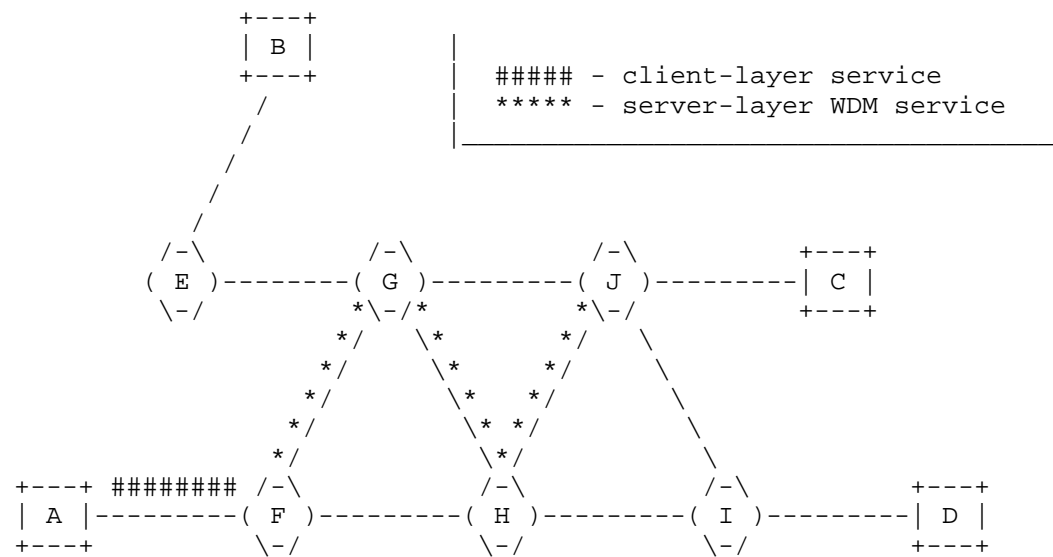


Figure 2b: Hierarchical service activation -
Server-layer WDM service that caters to the
client-layer service is established within the
core.

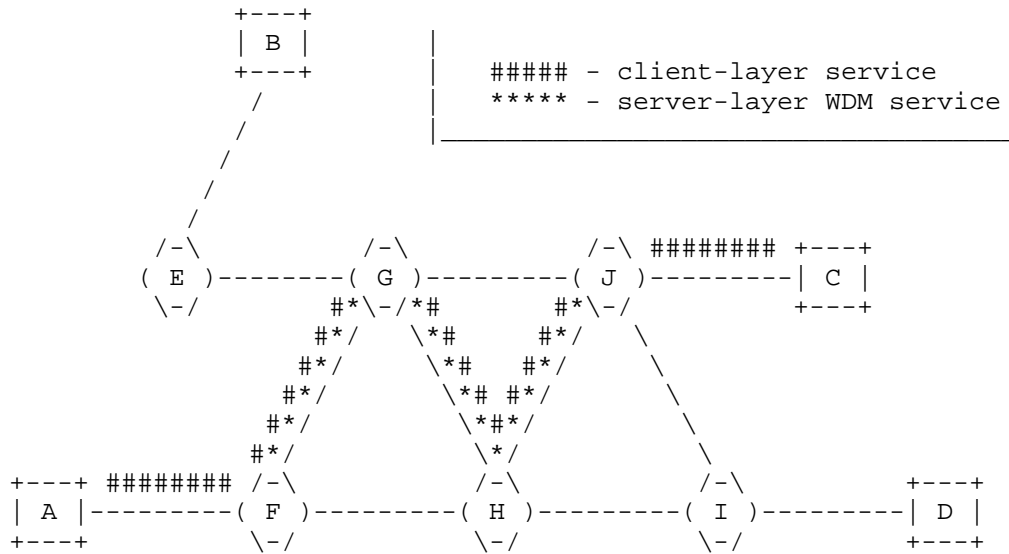


Figure 2c: Hierarchical service activation -
Client-layer service setup is resumed and
the end-to-end connection is established.

3. Traffic Engineering

The previous section outlines the basic method for activating end-to-end services across a multi-domain/multi-layer network. As a necessary part of that process an initial path selection process is to be performed, whereby an appropriate path between the desired endpoints is to be determined through some means. Further, per expectations set through current practices with regard to service provisioning in homogeneous networks, operators expect that the underlying control plane system provides automated mechanisms for computing the desired path(s) between network endpoints.

In particular, operators do not expect under normal circumstances to be required to explicitly specify the end-to-end path; rather, they expect to be able to specify just the endpoints of the path and rely on an automated computational process to identify and qualify all the elements and links on the path between them. Hence when operating a hybrid multi-layer network such as that described in Fig 1, it is necessary to extend existing traffic engineering and path computation mechanisms to operate in a similar manner.

Path computation and qualification operations occur at the path computation element (PCE - RFC4655) selected by ingress network element of an end-to-end service. In order to be able to compute and qualify paths, the PCE should be provided with information regarding the traffic engineering capabilities of the layer network to which it is associated with, in particular, the topology of the layer network and what layer-specific transport capabilities exist at the various nodes and links in that topology.

It is important to note that topology information is layer-specific; e.g. path computation and qualification operations occur within a given layer, and hence information about topology and resource availability are required for the specific layer to which the connection belongs. The topology and resource availability information required by a path computation element in the client layer is quite distinct from that required by a path computation element in the server layer network. Hence, the actual server layer traffic engineering links are of no importance for the client layer network. In fact, it can be desirable to block their advertisements into the client TE domain by the border nodes.

For example, in the sample hybrid network (Fig 1) there are multiple transport elements supporting client the connection (in this memo terms "connection" and "LSP" are used interchangeably) between the GMPLS-enabled clients A and C, the server layer topology between them includes several nodes and links. However, in this example the optical network elements are not capable of switching traffic with the client layer granularity (i.e. IP/MPLS packets), as the optical network elements are lambda switches, not IP/MPLS switches. Hence, while the intervening server layer network elements may physically exist along the path, they are not a part of the topology required by the client layer nodes for the purposes of traffic engineering in the client layer network.

An example of what the client layer Traffic Engineering topology would look like for the sample hybrid network is shown in the top half of Fig 3.

In this example, the TE topology associated with the client layer network is indicated by the links marked with '+' and nodes marked without brackets, whereas the TE topology associated with the server layer network is indicated by the links marked with '~' and nodes marked in '{}'. The nodes at the edge of the server layer network are visible in both the topologies. The client topology is capable of switching traffic within the client layer, whereas the server topology is capable of switching traffic within the server layer.

In this example, if the "B" router attempts to determine a path to the "D" router it will be unable to do so, as the client topology to which the B and D routers is connected does not include a full path made of just client layer links between them. The only way to setup an end-to-end path in this case is to use an ERO with a "loose hop" across the server layer domain as illustrated in Fig 3. This would cause the server layer to create the necessary link in the client layer topology on the fly. However, this approach has a few drawbacks - [a] the necessity for the operator to specify the ERO with the "loose" hop; [b] potential sub-optimal usage of server layer network resources; [c] unpredictability with regard to the fate-sharing of the new link (that is created on the fly) with other links of the client layer topology.

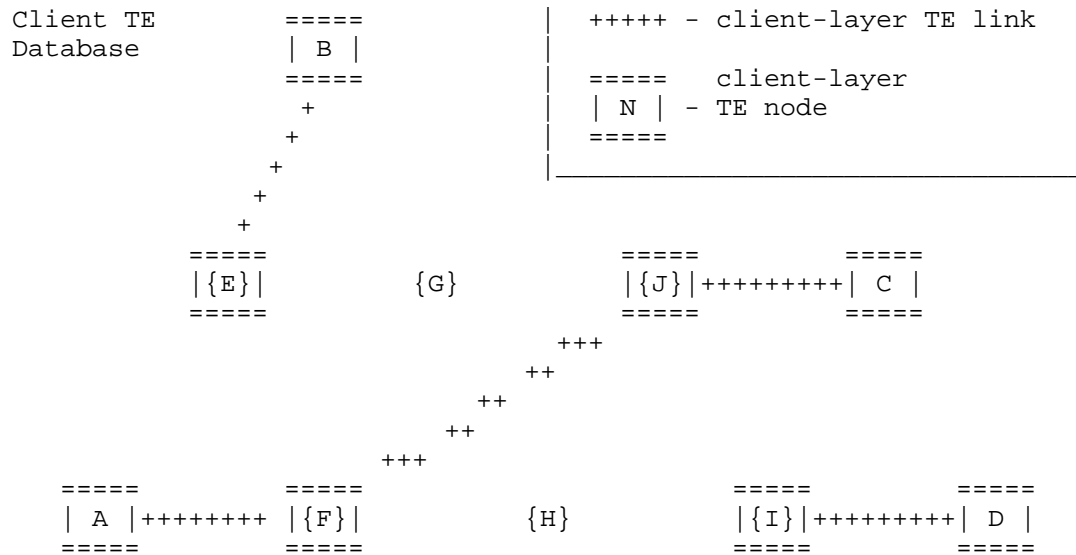
In order to be able to compute an end-to-end path between the two client layer endpoints, the client topology must be sufficiently augmented to indicate where there are paths through the server topology, which can provide connectivity between nodes in the client topology. In other words, in order for a client to compute path(s) across the server layer network to other clients, the feasible paths across the server layer network should be made available (in terms of TE links and nodes that exist in the client layer network) to all the clients. This is discussed in detail in the next section.

As it is mentioned already, in the overlay model the client and network domains, generally speaking, exist in separate layer-networks. One important use case, however, is when the client and network topologies belong to the same layer network. For example, IP routers, connected via GMPLS ENNI to a WDM network, could be capable of terminating optical trails being lambda switched by the network. The method described in the following sections allows also partitioning a single layer network into domains. Those domains do not need to leak the full routing information to their neighboring domains but rather provide sufficient information for a path

computation engine to route connections across a multi-domain network.

3.1. Augmenting the Client layer Topology

In the example hybrid network, shown below in Fig 4, consider a scenario, where each GMPLS-enabled IP router is connected to the optical WDM transport network via a transponder. Further, consider the situation, where the transponder on node F can be connected to the transponder on node J via the optical path F-G-H-J. Suppose, a lambda LSP is provisioned in the server layer along this path and advertised (as a TE link) into the client layer network. With the availability of this TE link, the path computation function at node



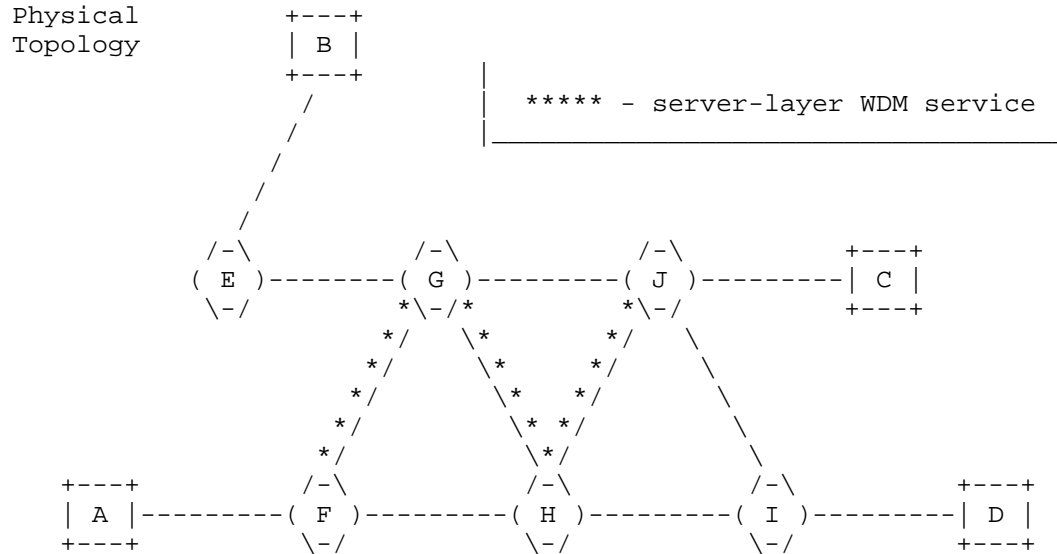


Figure 4: Traffic engineering - end-to-end path computation.[The client layer "TE link" between F and J is produced by creating the underlying server-layer connection; Node A has visibility to end-to-end (A to C) client-layer links and can do CSPF]

A is able to compute an end-to-end path from A to C. In this example, in order for the TE link to be made available in the client layer network topology, the network resources supporting the underlying server layer LSP are fully committed beforehand.

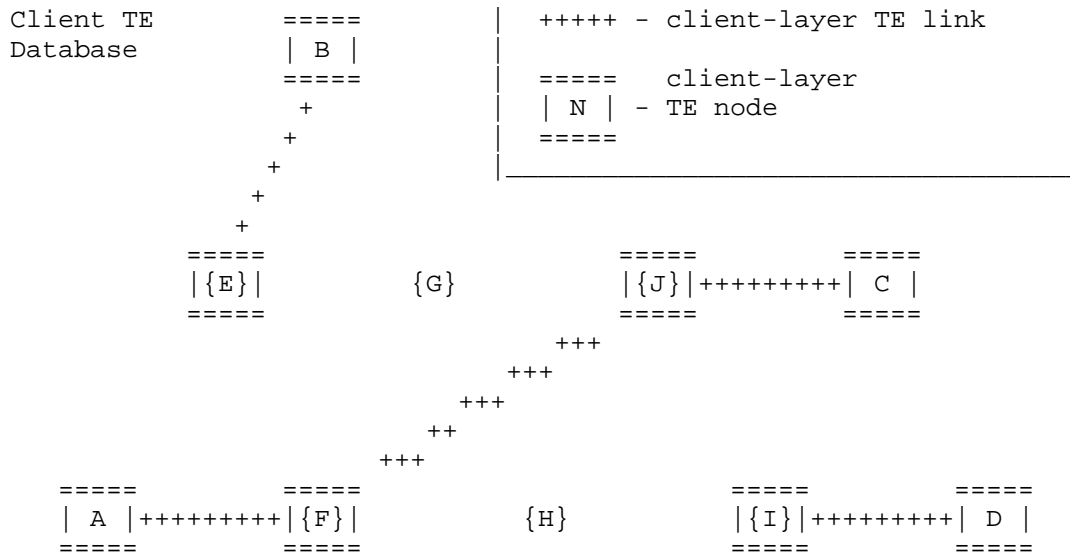
As another scenario, consider a network configuration, where the transponders on nodes E, F, J and I are connected to each other via directionless ROADM technology. In this case it is physically possible to connect any transponder to any other transponder in the server layer network. As there are transport capabilities available in the server layer network between every pair of elements with an adaptation function to the client layer network, the operator in

this case would not wish to commit any network resources in the server layer network until a client LSP is signaled. The next section proposes a method to address this common operational requirement.

3.1.1.1. Virtual TE Links

A "Virtual TE Link" as defined in section 7.3.3 of [RFC4847] is a TE link that is advertised into the client layer network. The advertisement includes information about available but not necessarily reserved/committed resources in the server layer network necessary to support that TE link. In other words, Virtual TE Links represent specific transport capabilities available in the server layer network, which can support the establishment of LSPs in the client layer network.

The two fundamental properties of a Virtual TE Link are: [a] it is advertised just like a real TE link and thus contributes to the buildup of the client layer network topology; and [b] it does not require allocation of resources at the server layer until used, thus allowing the mutually exclusive sharing of server layer network resources with other Virtual TE Links.



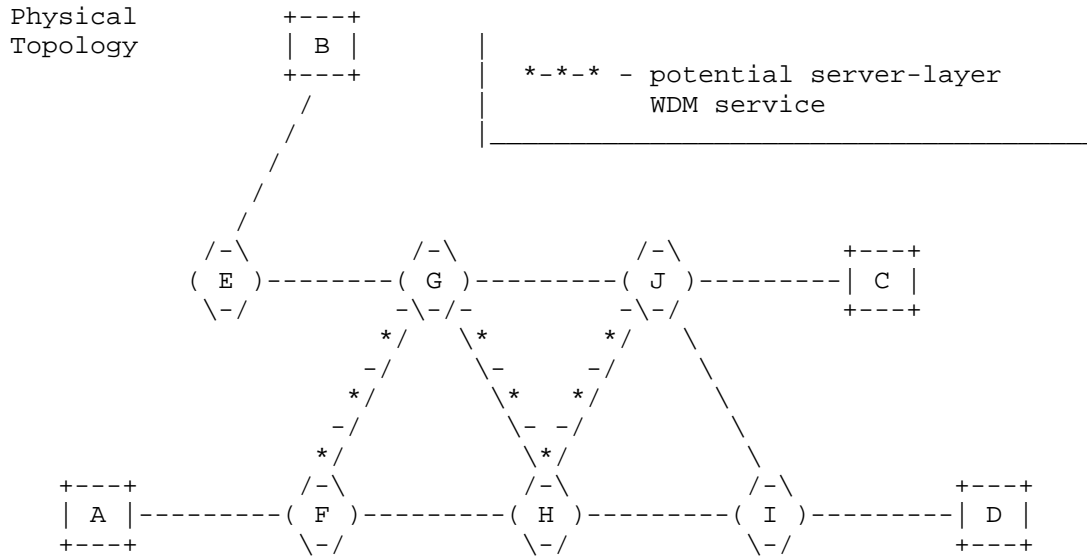


Figure 5: Traffic engineering - end-to-end path computation with "Virtual TE Links". [The "Virtual TE link" between F and J is created in the client layer without actually instantiating the underlying server-layer connection; Node A has visibility to end-to-end client-layer links and can do CSPF]

In the example shown in Fig 5, the availability of a lambda channel along the path F-G-H-J results in the advertisement by nodes F and J of a Virtual TE Link between F and J into the client layer network topology (+++ line). With the advertisement of this Virtual TE Link, the path computation function at node A is able to compute an end-to-end path from A to C.

Whenever a Virtual TE Link gets selected and signaled in the ERO of a client layer LSP, it ceases temporarily to be "virtual" and transforms into a regular TE link. When this transformation takes place, the clients will notice the change in the advertised available bandwidth of this TE link. Also, all other Virtual TE Links that share in a mutual exclusive way some of server layer resources with the TE link in question SHOULD start advertising "zero" available bandwidth. Likewise, the TE network image reverts back to the original form as soon as the last client layer LSP, going through the TE link in question, is released, i.e. Virtual TE Link becomes "virtual" again.

The overlay topology, advertised into the client domain as a set of Virtual TE Links, along with access TE links (the TE links interconnecting client network elements with the network domain) makes up the topology that in the overlay model allows for the client domain path computation function to compute end-to-end paths interconnecting client network elements across the network domain.

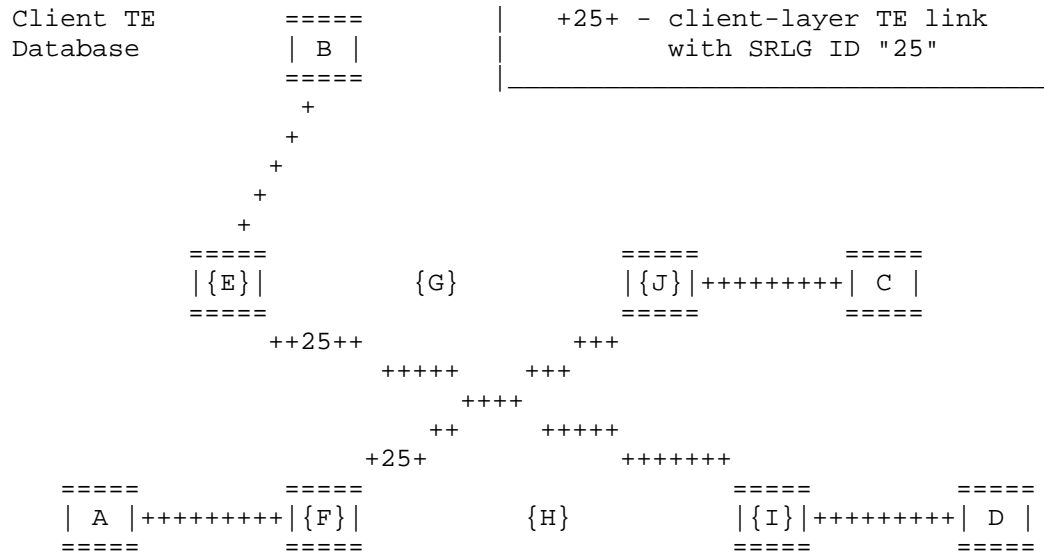
3.2. Macro SRLGs

The Virtual TE Links, which are advertised into the client layer network topology, cannot be assumed to be independent. It is quite possible for a given Virtual TE Link to share fate with one or more other Virtual TE Link(s). This is because the underlying server layer LSPs (established or potential) can traverse the same server layer network link and/or node, and failure of any such shared link/node would make all such LSPs inoperable (along with the Virtual TE Links supported by the LSPs). If diverse end-to-end paths for client layer LSPs are to be computed, the fate sharing information of the Virtual TE Links needs to be taken into account. The standard way of addressing this problem is via the concept of Shared Risk Link Group (SRLG). Specifically, a network resource shared by two or more TE links is identified via a network scope unique number (SRLG ID) and advertised within each such TE link advertisement.

A "traditional" SRLG (per [RFC4202]) represents a shared physical network resource, upon which normal function of a link depends. Such SRLGs can also be referred to as physical SRLGs. Zero, one or more physical SRLGs could be identified and advertised for every TE link in a given layer network. There is a scalability issue with physical SRLGs in multi-layer environments. For example, if a server layer LSP serves a client layer link, every server layer link and node traversed by the LSP must be considered as a separate SRLG. The number of server layer SRLGs to be advertised to client layer per

TE link is directly proportional to the number of hops traversed by the underlying server layer LSP.

This document introduces a notion of Macro SRLGs, which addresses this scaling problem. Macro SRLGs have the same protocol format as their physical counterparts and can be assigned automatically for each TE link that is advertised into the client layer network supported by an underlying server layer LSP (instantiated or otherwise). A Macro SRLG represents a shared path segment that is traversed by two or more of the underlying server layer LSPs. Each shared path segment can be viewed as a set of shared server layer resources. The actual procedure for deriving the Macro SRLGs is beyond the scope of this document.



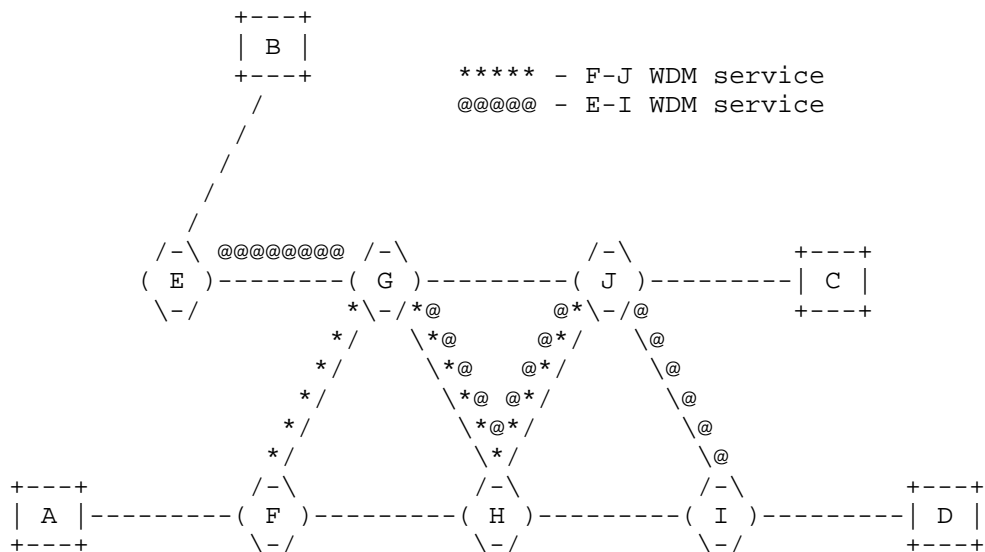


Figure 6: Macro SRLGs - ["TE links" E-I and F-J share fate since the underlying server-layer connections traverse the same path segments [G-H][H-I]. Macro SRLG-ID "25" is assigned to both "TE links"]

3.3. MELGS

If two or more Virtual TE Links share fate, it means that the links could be concurrently activated and used by client LSPs with a caveat that the links could be taken out of service by a single network failure, and, thus, cannot be used in the same protection scheme. There could be a stronger (than fate sharing) relationship between two or more Virtual TE Links. Because a set of Virtual TE Links can depend on the same uncommitted network resources, the situation can arise, when only one Virtual TE Link from the set could be activated at any given time. In other words, two or more Virtual TE Links can be mutually exclusive.

One example of the mutually exclusive relationship of Virtual TE Links is when the paths for the server layer network LSPs supporting the Virtual TE Links not only intersect, but also require usage of the same resource (e.g. lambda channel) on the intersection. Another example is when the said paths depend on a common physical resource

(e.g. transponder, regenerator, wavelength converter, etc.) that could be used only by one LSP at a time.

For a client path computation function (especially a centralized one capable of concurrent computation of multiple paths) it is important to know about such mutually exclusive relationship between Virtual TE Links. This document recommends the use of the extensions defined in [MELG] to address this requirement.

3.4. Switching Constraints

Generally speaking, it SHOULD NOT be assumed that a Virtual TE Link advertised by a given network domain border node can be cross-connected within a client LSP with every access TE link advertised by the said node. This circumstance necessitates the specification of connectivity constraints by network domain border nodes. If such information is not available for client domain path computers, there is a significant risk of provisioning failures of client LSPs, if/when they are signaled with the computed paths (see, Figure 7). This document recommends the use of the advertisements specified in [GEN_CNSTR] and [OSPF_GEN_CNSTR] to address the network element switching limitations problem.

```

+---+a1-----b1--/-\--b3-----c1--/-\--c3-----d1-+---+
| A |           ( B )           ( C )           | D |
+---+a2-----b2--/-\--b4-----c2--/-\--c4-----d2-+---+

```

Access TE-links:	TE links served By the server domain:	Valid paths:
a1-b1, c3-d1	b3-c1	[a1-b1][b3-c1][c3-d1]
a2-b2, c4-d2	b4-c2	[a2-b2][b4-c2][c4-d2]
Binding constraints:		Invalid paths:
b1<->b3		[a1-b1][b4-c2]...
b2<->b4		[a2-b2][b3-c1]...
c1<->c3		[a1-b1][b3-c1][c4-d2]
c2<->c4		[a2-b2][b4-c2][c3-d1]

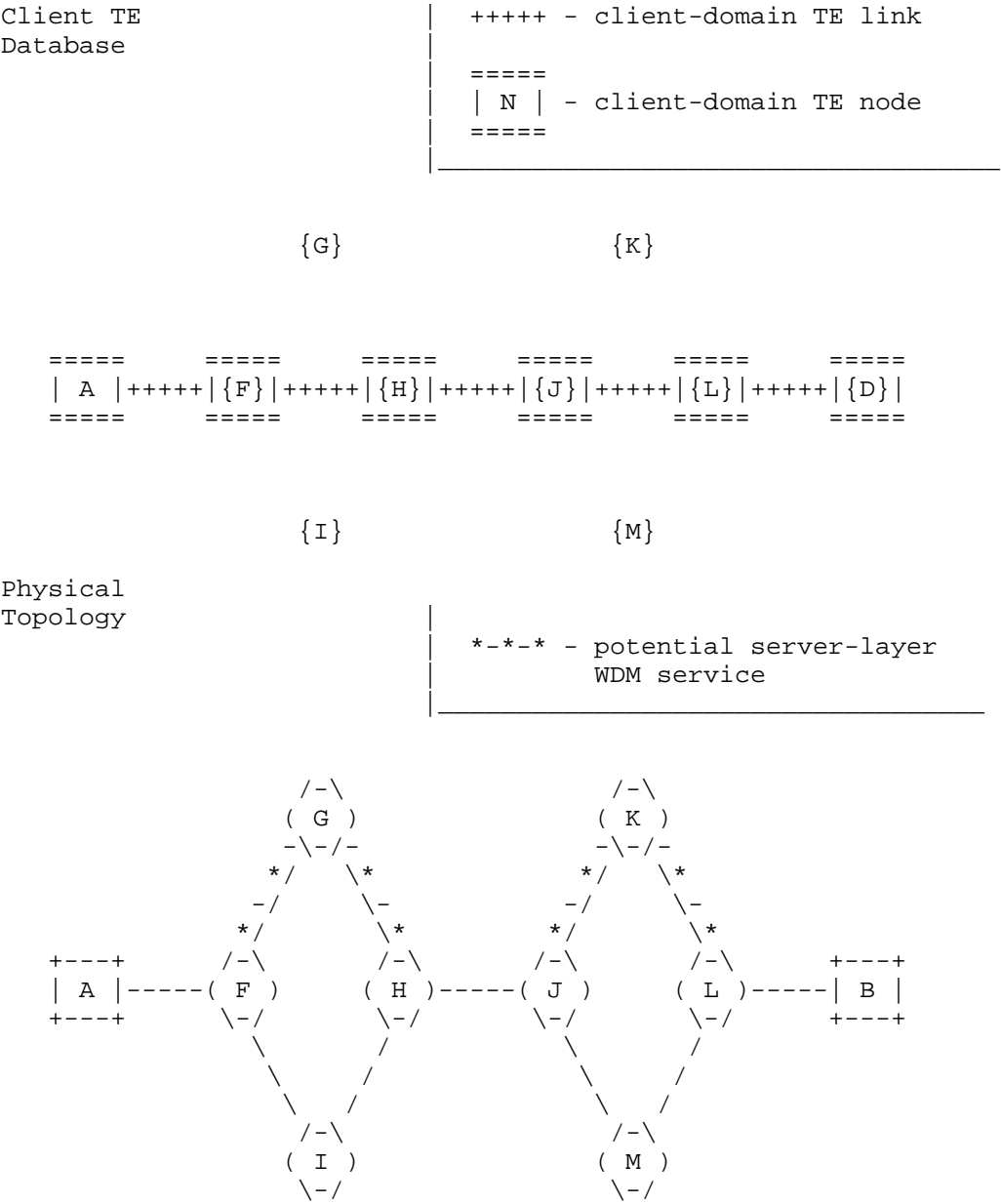
Figure 7: Switching Constraints

4. GMPLS ENNI and Multiple Server Network Domains

In the previous sections a single server network domain GMPLS ENNI configuration was considered. The said configuration is modeled as a set of client nodes, belonging to one or more client domains, connected to a single server network domain. The connectivity is realized via access links in the data plane and GMPLS ENNI interfaces in the control plane. The server domain is independent from the client domain(s) (administratively and from the Traffic Engineering and control/management plane point of view). The network domain exposes its resources to the clients in a form of Virtual TE Links, thus, enabling the clients to influence the way their LSPs are routed across the network domain.

There are important use cases that require client LSPs to traverse more than one server network domains. In such use cases the server domains, generally speaking, are independent from each other and from each of the client domains. In such configurations the clients would still want to control the routing of their LSPs in each of the server domains, the LSPs are going through, for the same reasons it is necessary to do so in the single server domain configuration (e.g. diversity, fate sharing, MELG considerations, etc.). Fortunately, the Virtual TE Link approach allows for exposing of the resources of multiple network domains in the same way as in the single server domain case, and, thus, provides the same tools for dynamic provisioning of client LSPs across either single or multiple server network domains.

Multiple server network domains are modeled as separate independent networks interconnected with each other on their respective border nodes via inter-domain links in the data plane and GMPLS ENNI interfaces in the control plane. A network border node sees no difference between an access link and an inter-domain link terminated on the node (nor can it tell whether it is connected via a given link to a client node or a border node of a neighboring server network domain). Just like in the single-domain case, each server domain exposes its resources to other server and client network domains via Virtual TE Links configured in accordance with local domain policies. It is responsibility of server domain border nodes to advertise into the neighboring domains all access, inter-domain and Virtual TE Links it locally terminates, as well as imposed on them switching limitations. The said advertisements are flooded into the client domains and populate the client path computer's TEDs. Successful path computations produce end-to-end paths in the form of access, Virtual and inter-domain TE link chains.



5. Path computation aspects

It is assumed that a client domain path computation function makes use of advertised access TE links as well as Virtual TE Links, while computing end-to-end paths for client LSPs. The said path computation function could be local (i.e. located on client LSP ingress nodes, as stipulated by [RFC4655] Composite PCE node) or remote (i.e. on network PCEs). Path computations could be triggered by client nodes or NMS. Generally speaking, the responsibility of the client domain path computation function is to (concurrently) compute one or several paths for each source-destination pair (potential client LSP termination points) specified in a single path computation request. The path computation SHOULD be subject to one or more path optimization criterions (such as minimal cost, minimal latency, etc.) and a set of path computation constraints (such as link unreserved bandwidth, link colors, layer-specific constraints, explicit inclusions and exclusions, etc.)

As the overlay topology hides actual server domain/layer links and nodes, it is RECOMMENDED to support SRLG diverse computation of two or more paths.

Furthermore, the path computation SHOULD consider the connectivity/switching limitation constraint (when available) in addition to all other path computation constraints.

The use of the PCE architecture and PCEP protocol is governed by [RFC5440], [RFC5521] and [RFC5541].

As described in section 3.3., two or more Virtual TE Links may not only share risk, but also may exclusively depend on the same server layer resources. Therefore, paths, computed on network topologies containing Virtual TE Links, have an increased probability of LSP setup failures (two LSPs, for example, could be routed over two Virtual TE Links that exclusively depend on the same server layer resource). In such cases concurrent path computation, taking in consideration MELG information, will address this problem. PCEP supports concurrent path computation per [RFC5440]. Specifying MELG diversity constraint in path computation requests is out of scope of this document.

In addition MELG may carry information on the establishment of server-layer resources. A Path computation request MAY constraint the path computation to TE-Links that are fully provisioned only. This information MAY also be used in PCE path computation policies.

6. Access and Virtual TE link addressing

[RFC4208] implies that access TE links could be named from the same address space as network domain TE links or from a separate address space. This memo requires the following:

- It MAY be possible to assign addresses for access TE links from the same address space as the one used for naming network internal TE links (i.e. TE links interconnecting network domain devices);
- It MUST be possible to assign addresses for access TE links from a separate address space, independent from the space used for addressing network internal TE links;
- Virtual TE Links MUST share the address space with any access TE links they are allowed to be cross-connected within a client LSP.

7. Use cases

7.1. Service Optimization and Restoration in Multi-layer networks

Multi-layer networks are a reality today, and they are operated by different groups of people, following different operational procedures. This requires an independent optimization of the client and server layer networks. Such independence may cause a situation, where the re-routing of a client layer LSP fails, because some of resources on the selected alternate path share fate with some of resources on the LSP's failed path. This usually happens due to lack of knowledge of the server layer network by a client layer path computation function at the time when the alternative path is selected.

The high volume and importance of IP traffic in provider networks today requires the client and server layer networks to share sufficient information in order to enable an optimized transport for IP/MPLS services and address existing inefficiencies. From the carrier perspective it is very important that the SRLG information is provided by the server layer TE application and is used by the client layer path computation.

In a typical multi-layer network, where IP/MPLS is the client layer network and WDM/OTN is the server layer network, the client layer network is responsible for the protection of the IP/MPLS traffic from networks failures. This is normally achieved via using

protection schemes, such as FRR and/or LFA. Regardless of the used mechanism, the SRLG information, provided by the server layer network, helps to optimize the client layer network with respect to reduced link utilization and reliable and efficient protection of the user traffic.

Today the SRLGs information is used mainly when calculating diverse alternative paths for the IP/MPLS LSPs. Therefore, the following procedures are performed periodically:

- Building traffic matrix for the server layer network (based on IP links)
- Solving traffic engineering problems in the server layer network
- (Re-)Calculating SRLGs to be propagated into the client layer network
- Simulating failure scenarios
- Making sure that the affected IP/MPLS LSPs function properly after they are replaced onto SRLG diverse alternative paths

GMPLS ENNI reduces the OPEX costs of performing these procedures via the automation as follows:

- server layer network automatically discovers and advertises the SRLG information into client layer network via a common routing protocol;
- client layer network path computer uses the SRLG information when selecting diverse paths.

7.2. IP/MPLS Offloading with ENNI automation

A typical application in multi-layer (IP/MPLS over optical) networks is termed 'IP Offloading', in which the network responds to the increase in traffic of a particular service or across a segment in the IP network by dynamically creating additional IP/MPLS links served by GMPLS LSPs provisioned in the server layer network, and placing the extra IP/MPLS traffic onto said links. Likewise, when the IP/MPLS traffic decreases to a normal pattern, the said GMPLS LSPs are torn down, and the extra IP/MPLS links are removed from the client layer network TE domain. The increase in traffic is typically caused by an elevated number of high traffic flows/services traversing an IP network segment.

The decision process driving IP offloading is complex, and is governed by a set of rules. These rules reduce the cost of running the multi-layer network, while ensuring that it remains stable.

Automation of IP Offloading poses a number of challenges. It includes dynamic provisioning, release and maintenance of GMPLS LSPs in the server layer (e.g. WDM) network as well as automatic advertising/withdrawing them as (numbered or/and unnumbered) TE links into/from the client layer network. In order to pre-plan and manage properly the said dynamic IP/MPLS TE links, it is important to know in advance (and also in real time) the capabilities and resource availability of server layer network. The network domain/layer virtualization procedures described in this document helps to solve this complex operational issue.

7.3. Use of PCE and VNTM in Multi-layer Network Operation

Two key elements have been proposed to help in the management and coordination of multi-layer networks: the Path Computation Element (PCE) and the Virtual Network Topology Manager (VNTM). PCE is responsible for the calculation of paths between endpoints, particularly in complex scenarios involving, for example, WDM layer physical impairments. VNTM is in charge of maintaining the topology of the client layer network by instantiating virtual links, in the server layer network. I.e., it can be used to provide TE links to the client layer network dynamically.

Several cooperation modes between PCE, VNTM and the NMS have been proposed in [RFC5623]. For instance, the operator can request a new MPLS tunnel via the NMS, which communicates with a PCE with information of the multi-layer network. The PCE, in case there are enough resources in the IP/MPLS layer, normally returns a path for the tunnel made of real TE links. On the other hand, if there is a lack of resources in the IP/MPLS layer, the response may contain a path with one or more Virtual TE Links. In this case, the NMS can cooperate with the VNTM to suggest the set-up of a GMPLS LSP(s) in the server layer network. The VNTM, based on the local policies, can accept the suggestion and cause the set-up of the GMPLS LSPs in the server layer network.

In order for the computation to be effective, the PCE needs knowledge of the overlay topology (SRLGs, MELGs, TE metrics of the Virtual TE links), which can be provided via GMPLS ENNI.

8. Security Considerations

TBD

9. IANA Considerations

TBD.

10. References

10.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4202] K. Kompella, Y.Rekhter
"Routing Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4202, October 2005.
- [RFC4208] G. Swallow, J.Drake, H. Ishimatsu, and Y. Rekhter,
"GMPLS UNI: RSVP-TE Support for the Overlay Model", RFC 4208, October 2005.
- [GEN_CNSTR] G.Bernstein, Y.Lee, D.Li, W.Imajuku, "General Network Element Constraint Encoding for GMPLS Controlled Networks"
[draft-general-constraint-encode-10.txt]
- [OSPF_GEN_CNSTR] F.Zhang, J.Han, Y.Lee, D.Li, G.Bernstein, Y.Hu
"OSPF-TE Extensions for General Network Element Constraints"
[draft-general-constraints-ospf-te-04.txt]
- [MELG] V.Beeram, I.Bryskin, et al, "Mutually Exclusive Link Groups", [draft-beeram-ccamp-melg-01.txt]
- [TE INFO XCHG] A.Farrel, N.Bitart, G.Swallow, D.Ceccarelli,
"Problem Statement and Architecture for Information Exchange Between Interconnected Traffic Engineered Networks", [draft-farrel-interconnected-te-info-exchange-01.txt]

10.2. Informative References

- [RFC4847] T. Takeda, "Framework and Requirements for Layer 1

VPNs", RFC 4847, April 2007.

[RFC4655] A. Farrel, J.-P. Vasseur, J. Ash, "A Path
Computation Element (PCE)-Based Architecture", RFC
4655, August 2006.

11. Acknowledgments

Chris Bowers [cbowers@juniper.net]

Authors' Addresses

Igor Bryskin
ADVA Optical Networking

Email: ibryskin@advaoptical.com

Wes Doonan
ADVA Optical Networking

Email: wdoonan@advaoptical.com

Vishnu Pavan Beeram
Juniper Networks

Email: vbeeram@juniper.net

John Drake
Juniper Networks

Email: jdrake@juniper.net

Gert Grammel
Juniper Networks

Email: ggrammel@juniper.net

Manuel Paul
Deutsche Telekom

Email: Manuel.Paul@telekom.de

Ruediger Kunze
Deutsche Telekom

Email: Ruediger.Kunze@telekom.de

Oscar Gonzalez de Dios
Telefonica

Email: ogondio@tid.es

Cyril Margaria
Coriant GmbH

Email: cyril.margaria@coriant.com

Friedrich Armbruster
Coriant GmbH

Email: friedrich.armbruster@coriant.com

Daniele Ceccarelli
Ericsson

Email: daniele.ceccarelli@ericsson.com

CCAMP Working Group
Internet Draft
Intended status: Standards Track

Vishnu Pavan Beeram (Ed)
Juniper Networks
Igor Bryskin (Ed)
ADVA Optical Networking

Expires: April 21, 2014

October 21, 2013

Mutually Exclusive Link Group (MELG)
draft-beeram-ccamp-melg-02.txt

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), 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, 2014.

Copyright Notice

Copyright (c) 2013 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.

Abstract

This document introduces the concept of MELG ("Mutually Exclusive Link Group") and discusses its usage in the context of mutually exclusive Virtual TE Links.

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 RFC-2119 [RFC2119].

Table of Contents

1. Introduction.....	2
2. Virtual TE Link - Semantics.....	3
3. Mutually Exclusive Virtual TE Links.....	3
3.1. Static vs Dynamic.....	4
4. Static Mutual Exclusivity.....	4
5. Mutually Exclusive Link Group.....	7
6. Protocol Extensions.....	8
6.1. OSPF.....	8
6.2. ISIS.....	9
7. Security Considerations.....	10
8. IANA Considerations.....	10
8.1. OSPF.....	10
8.2. ISIS.....	10
9. Normative References.....	10
10. Acknowledgments.....	11

1. Introduction

A Virtual TE Link (as defined in [RFC6001]) advertised into a Client Network Domain represents a potentiality to setup an LSP in the Server Network Domain to support the advertised TE link. The Virtual TE Link gets advertised like any other TE link and follows the same rules that are defined for the advertising, processing and use of regular TE links [RFC4202]. However, "mutual exclusivity" is one attribute that is specific to Virtual TE links. This document discusses the different types of mutual exclusivity (Static vs Dynamic) that come into play and explains the need to advertise this

information into the Client TEDB. It then goes onto introduce a new TE construct (MELG) to carry static mutual exclusivity information.

2. Virtual TE Link - Semantics

A Virtual TE Link (as per existing definitions) represents the potentiality to setup a server layer LSP, but there are currently no strict guidelines imposed on how the underlying server layer LSP would need to get set up. The characteristics of the underlying server-path are not necessarily pinned down until the Virtual TE Link gets actually committed. This means that some important characteristics of the Virtual TE Link like shared-risk and delay (and mutual exclusivity information) may not be known until the corresponding server layer LSP is set up. This makes resource planning (for example - pre-configuring network failure recovery schemes) in a multi-layer network that includes Virtual TE Links a very hard problem.

This document uses a slightly enhanced view of a Virtual TE Link. In the context of this document, the Virtual TE Link (even when it is uncommitted) is always aware of the key characteristics of the underlying server-path. The creation and maintenance of this Virtual TE Link is strictly driven by policy. Policy not only determines which Virtual TE Link to create (What termination points?), but it may also constrain how the corresponding underlying server layer LSP (What path?) needs to get set up. The basic idea behind this "enhanced view" is that it makes the "Virtual TE Link" get as close as it can to representing a "Real TE Link".

Also, as per this document, a Virtual TE Link remains a Virtual TE Link through-out its life-time (until it gets deleted by the user/policy). It may get committed (underlying server LSP gets set up) and uncommitted (underlying server LSP gets deleted) from time to time, but it never really loses its "Virtual" property.

3. Mutually Exclusive Virtual TE Links

Mutual Exclusivity comes into play when multiple Virtual TE Links are dependent on the usage of the same underlying server resource. Since not all of these Virtual TE Links can get committed at the same time, they are deemed to be mutually exclusive.

The existence of this "mutual exclusivity" property would need to be advertised into the Client TE Domain. This is of relevance to Client Path Computation engines; especially those that are capable of doing concurrent computations. If this information is absent, there exists

the risk of the Computation engine yielding erroneous concurrent path computation results where only a subset of the computed paths get successfully provisioned.

The "Mutual Exclusivity" property of a Virtual TE Link can be either static or dynamic in nature.

3.1. Static vs Dynamic

Static Mutual Exclusivity: This type of mutual exclusivity exists permanently within a given network configuration. It comes into play when two or more Virtual TE Links depend on the usage of the same non-shareable underlying server network domain resource. This resource gets used up in its entirety by a single Virtual TE Link when committed. Such resources exist only in the WDM layer.

Dynamic Mutual Exclusivity: This type of mutual exclusivity exists temporarily within a given network configuration. It comes into play when two or more Virtual TE Links depend on the usage of the same shareable underlying server network domain resource. Mutual Exclusivity exists when the amount of the server resource that is available for sharing is limited; it ceases to exist when sufficient amount of the resource is available for accommodating all corresponding Virtual TE Links. Such resources exist in all layers.

Because of their inherent difference, the advertisement paradigm of the TE construct required to carry static mutual exclusivity information is quite different from that of the TE construct required to carry dynamic mutual exclusivity information. Static mutual exclusivity Information can get advertised per TE-Link using a simple sub-TLV construct. There wouldn't be any scaling issues with this approach because of the static nature of the information that gets advertised. On the contrary, advertising dynamic mutual exclusivity information per TE-Link poses serious scaling concerns and hence requires a different type of construct/paradigm.

This document introduces a new TE construct for carrying static mutual exclusivity information. The mechanisms to address dynamic mutual exclusivity are discussed in a separate document [SRcLG].

4. Static Mutual Exclusivity

Consider the network topology depicted in Figure 1a. This is a typical packet optical transport deployment scenario where the WDM layer network domain serves as a Server Network Domain providing

transport connectivity to the packet layer network Domain (Client Network Domain).

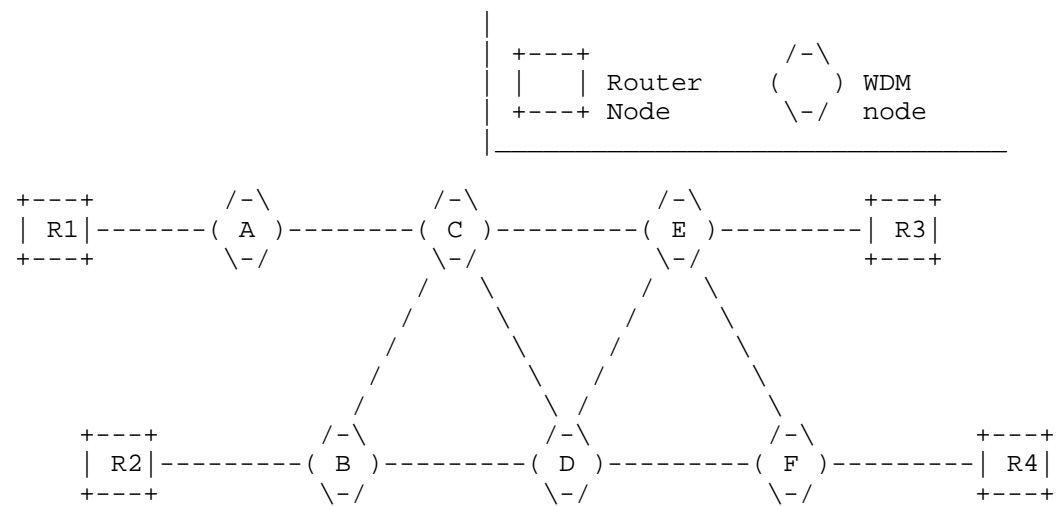


Figure 1a: Sample topology



Figure 1b: Client TE Database

Nodes R1, R2, R3 and R4 are IP routers that are connected to an Optical WDM transport network. A, B, C, D, E and F are WDM nodes that constitute the Server Network Domain. The border nodes (A, B, E

and F) operate in both the server and client domains. Figure 1b depicts how the Client Network Domain TE topology looks like when there are no Client TE Links provisioned across the optical domain.

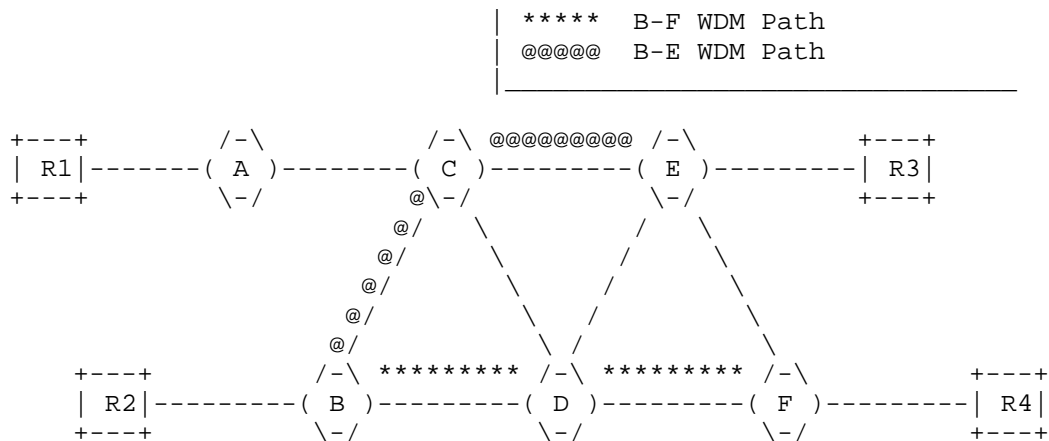


Figure 2a: Mutually Exclusive potential WDM paths

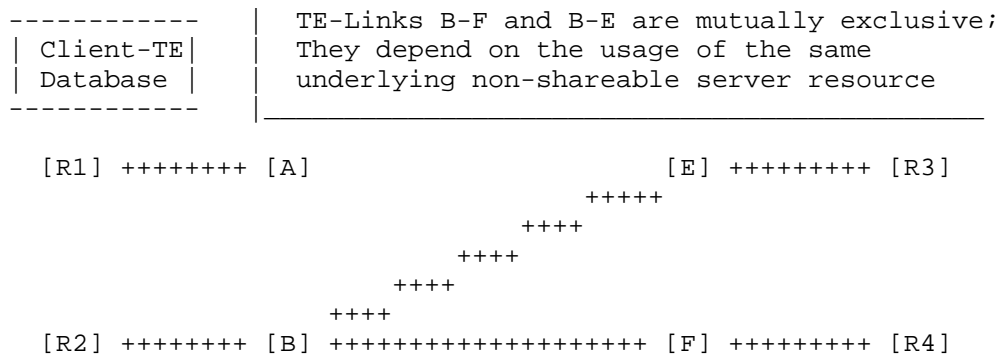


Figure 2b: Client TE Database - Mutually Exclusive Virtual TE Links

Now consider augmenting the Client TE topology by creating a couple of Virtual TE Links across the optical domain. The potential paths in the WDM network catering to these two virtual TE links are as shown in Fig 2a and the corresponding augmented Client TE topology is as illustrated in Fig 2b.

In this particular example, the potential paths in the WDM layer network supporting the Virtual TE Links require the usage of the same source transponder (on "Node B"). Because the Virtual TE Links depend on the same uncommitted network resource, only one of them could get activated at any given time. In other words they are mutually exclusive. This scenario is encountered when the potential paths depend on any common physical resource (e.g. transponder, regenerator, wavelength converter, etc.) that could be used by only one Server Network Domain LSP at a time.

This document proposes the use of "Mutually Exclusive Link Group (MELG)" for catering to this scenario.

5. Mutually Exclusive Link Group

The Mutually Exclusive Link Group (MELG) construct defined in this document has 2 purposes

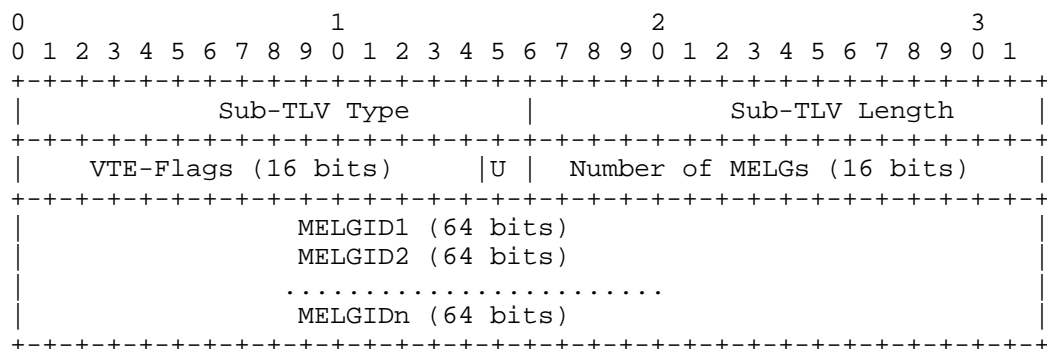
- To indicate via a separate network unique number (MELG ID) an element or a situation that makes the advertised Virtual TE Link belong to one or more Mutually Exclusive Link Groups. Path computing element will be able to decide on whether two or more Virtual TE Links are mutually exclusive or not by finding an overlap of advertised MELGs (similar to deciding on whether two or more TE links share fate or not by finding common SRLGs)
- To indicate whether the advertised Virtual TE Link is committed or not at the moment of the advertising. Such information is important for a path computation element: Committing new Virtual TE links (vs. re-using already committed ones) has a consequence of allocating more server layer resources and disabling other Virtual TE Links that have common MELGs with newly committed Virtual TE Links; Committing a new Virtual TE Link also means a longer setup time for the Client LSP and higher risk of setup-failure.

6. Protocol Extensions

6.1. OSPF

The MELG is a sub-TLV of the top level TE Link TLV. It may occur at most once within the Link TLV. The format of the MELGs sub-TLV is defined as follows:

Name: MELG
 Type: TBD
 Length: Variable



Number of MELGs: number of MELGS advertised for the Virtual TE Link;
 VTE-Flags: Virtual TE Link specific flags;
 MELGID1,MELGID2,...,MELGIDn: 64-bit network domain unique numbers associated with each of the advertised MELGs

Currently defined Virtual TE Link specific flags are:

U bit (bit 1): Uncommitted - if set, the Virtual TE Link is uncommitted at the time of the advertising (i.e. the server layer network LSP is not set up); if cleared, the Virtual TE Link is committed (i.e. the server layer LSP is fully provisioned and functioning). All other bits of the "VTE-Flags" field are reserved for future use and MUST be cleared.

Note: A Virtual TE Link advertisement MAY include MELGs sub-TLV with zero MELGs for the purpose of communicating to the TE domain whether the Virtual TE Link is currently committed or not.

6.2. ISIS

The MELG TLV (of type TBD) contains a data structure consisting of:

```

6      octets of System ID
1      octet of Pseudonode Number
1      octet Flag
4      octets of IPv4 interface address or 4 octets of a Link
      Local Identifier
4      octets of IPv4 neighbor address or 4 octets of a Link
      Remote Identifier
2      octets MELG-Flags
2      octets - Number of MELGs
variable List of MELG values, where each element in the list
      has 8 octets

```

The following illustrates encoding of the value field of the MELG TLV.

```

      0              1              2              3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     System ID                                     |
+-----+-----+-----+-----+-----+-----+-----+-----+
|      System ID (cont.)      |Pseudonode num |      Flags      |
+-----+-----+-----+-----+-----+-----+-----+-----+
|      Ipv4 interface address/Link Local Identifier      |
+-----+-----+-----+-----+-----+-----+-----+-----+
|      Ipv4 neighbor address/Link Remote Identifier      |
+-----+-----+-----+-----+-----+-----+-----+-----+
|      VTE-Flags (16 bits)      |U |      Number of MELGs (16 bits)      |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     MELGID1 (64 bits)                                     |
|                                     MELGID2 (64 bits)                                     |
|                                     .....                                     |
|                                     MELGIDn (64 bits)                                     |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

The neighbor is identified by its System ID (6 octets), plus one octet to indicate the pseudonode number if the neighbor is on a LAN interface.

The least significant bit of the Flag octet indicates whether the interface is numbered (set to 1) or unnumbered (set to 0). All other bits are reserved and should be set to 0.

The length of the TLV is $20 + 8 * (\text{number of MELG values})$.

The semantics of "VTE-Flags", "Number of MELGs" and "MELGID Values" are the same as the ones defined under OSPF extensions.

The MELG TLV MAY occur more than once within the IS-IS Link State Protocol Data Units.

7. Security Considerations

TBD

8. IANA Considerations

8.1. OSPF

IANA is requested to allocate a new sub-TLV type for MELG (as defined in Section 6.1) under the top-level TE Link TLV.

8.2. ISIS

IANA is requested to allocate a new IS-IS TLV type for MELG (as defined in Section 6.2).

9. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4202] K.Kompella, Y.Rekhter, "Routing Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC4202, October 2005.
- [RFC6001] D.Papadimitriou, M.Vigoureaux, K.Shiomoto, D.Brungard and JL. Le Roux, "GMPLS Protocol Extensions for Multi-Layer and Multi-Region Networks", RFC 6001, October 2010.
- [SRcLG] Beeram, V., "Shared Resource Link Group", draft-beeram-ccamp-srclg, October 2013

10. Acknowledgments

Chris Bowers [cbowers@juniper.net]

Authors' Addresses

Vishnu Pavan Beeram
Juniper Networks
Email: vbeeram@juniper.net

Igor Bryskin
ADVA Optical Networking
Email: ibryskin@advaoptical.com

John Drake
Juniper Networks
Email: jdrake@juniper.net

Gert Grammel
Juniper Networks
Email: ggrammel@juniper.net

Wes Doonan
Email: wddlists@gmail.com

Manuel Paul
Deutsche Telekom
Email: Manuel.Paul@telekom.de

Ruediger Kunze
Deutsche Telekom
Email: Ruediger.Kunze@telekom.de

Oscar Gonzalez de Dios
Telefonica
Email: ogondio@tid.es

Cyril Margaria
Email: cyril.margaria@gmail.com

Friedrich Armbruster

Coriant GmbH
Email: friedrich.armbruster@coriant.com

Daniele Ceccarelli
Ericsson
Email: daniele.ceccarelli@ericsson.com

Fatai Zhang
Huawei Technologies
Email: zhangfatai@huawei.com

CCAMP Working Group
Internet Draft
Intended status: Standards Track

Vishnu Pavan Beeram (Ed)
Juniper Networks
Igor Bryskin (Ed)
ADVA Optical Networking

Expires: April 20, 2014

October 20, 2013

Network Assigned Upstream-Label
draft-beeram-ccamp-network-assigned-upstream-label-00.txt

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), 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, 2014.

Copyright Notice

Copyright (c) 2013 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.

Abstract

This document discusses the GMPLS RSVP-TE extensions that are needed to let the network assign an upstream-label for a given LSP. This is useful in scenarios where a given node does not have sufficient information to assign the correct upstream-label on its own. This document also specifies the extensions required for manipulating Label-Symmetric Bidirectional GMPLS LSPs.

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 RFC-2119 [RFC2119].

Table of Contents

1. Introduction.....	2
2. Label Symmetry.....	3
2.1. Processing Rules.....	3
3. Unassigned Upstream Label.....	4
3.1. Processing Rules.....	4
4. Upstream Label Set / Acceptable Upstream Label Set.....	5
4.1. Object Formats.....	6
4.2. Processing Rules.....	6
5. Use-Cases.....	7
5.1. Alien-Wavelength Setup.....	7
5.1.1. Setup Procedure - Example.....	8
6. Security Considerations.....	10
7. IANA Considerations.....	11
8. Normative References.....	11
9. Acknowledgments.....	11

1. Introduction

The GMPLS RSVP-TE extensions for setting up a Bidirectional LSP are discussed in [RFC3473]. The Bidirectional LSP setup is indicated by the presence of an UPSTREAM_LABEL Object in the PATH message. As per the existing setup procedure outlined for a Bidirectional LSP, each upstream-node must allocate a valid upstream-label on the outgoing interface before sending the initial PATH message downstream.

However, there are certain scenarios (Section 5) where it is not desirable for a given node to pick the upstream-label on its own. This document discusses the protocol extensions that are required in such cases to let the network assign an upstream-label for a given LSP.

As per [RFC3471], the upstream-label and the downstream-label for an LSP at a given hop need not be the same. However, most practical scenarios require these two labels to be the same. This document proposes a mechanism for the ingress to request "Label Symmetry" at each hop of the LSP. It also discusses how the request to have "Label Symmetry" gets processed in conjunction with the request to have "a network assigned upstream-label".

2. Label Symmetry

In order to request "Label Symmetry", this document defines a new flag (Label_Symmetry Required) in the Attributes Flags TLV [RFC5420]. The position of this flag in the TLV is TBA.

If the upstream-label and the downstream-label are required to be the same at each hop of the LSP, then the PATH sent out by the ingress would have this flag set in the Attributes Flags TLV of the LSP_REQUIRED_ATTRIBUTES object.

2.1. Processing Rules

The presence of the "Label Symmetry Required" flag in the PATH message indicates that the LSP is bidirectional and that the labels are symmetric in both directions at each hop. Since this flag gets carried in the LSP_REQUIRED_ATTRIBUTES object, a downstream node that does not recognize/support this flag would reject the LSP setup request (indicating that the requested attributes are not supported).

When this flag is set in the PATH message, the upstream node may or may not add the UPSTREAM_LABEL object in the initial setup request sent to the downstream node. If the UPSTREAM_LABEL does get specified in the PATH, the downstream nodes MUST ignore it. If the upstream node desires to pick the symmetric label on its own, it MUST encode this in the LABEL_SET object and send it downstream.

The downstream-node picks an appropriate symmetric label and sends this via the LABEL object in the RESV message. The upstream-node would then start using this symmetric label for both directions of the LSP.

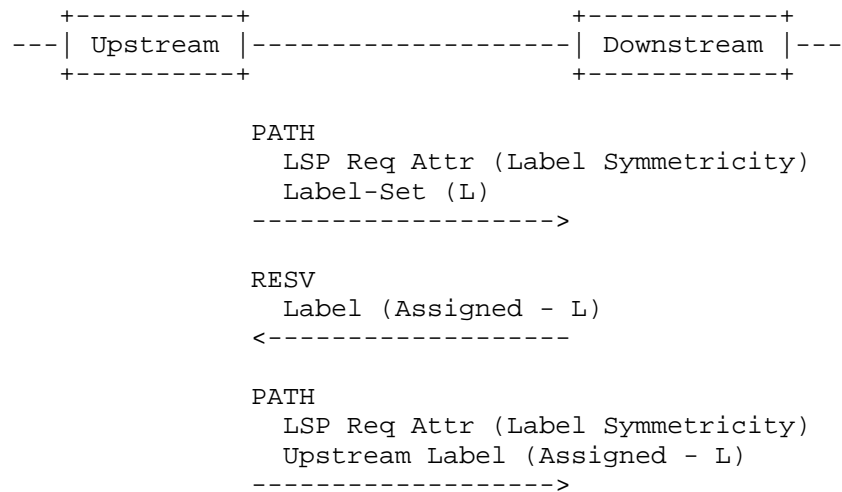


Figure 1: Label Symmetry

The remaining extensions discussed in this document are not relevant for LSPs that require "Label Symmetry".

3. Unassigned Upstream Label

This document proposes the use of a special label value - "0xFFFFFFFF" - to indicate an Unassigned Label. This would get used by a node if it does not have any input on what upstream-label needs to get picked. This special label is filled in the UPSTREAM_LABEL object of the PATH message that is sent downstream.

3.1. Processing Rules

In the ideal scenario, the network responds by filling in a valid UPSTREAM_LABEL in the corresponding RESV message. If the network is not in a position to assign the UPSTREAM_LABEL (or if it doesn't know what to do with an Unassigned UPSTREAM_LABEL), it MUST issue a PATH-ERR message with a "Routing Problem/Unacceptable Label Value" indication. If the RESV comes in without an assigned UPSTREAM_LABEL, then an error with a "Routing Problem/Label Allocation Failure" indication MUST be issued.

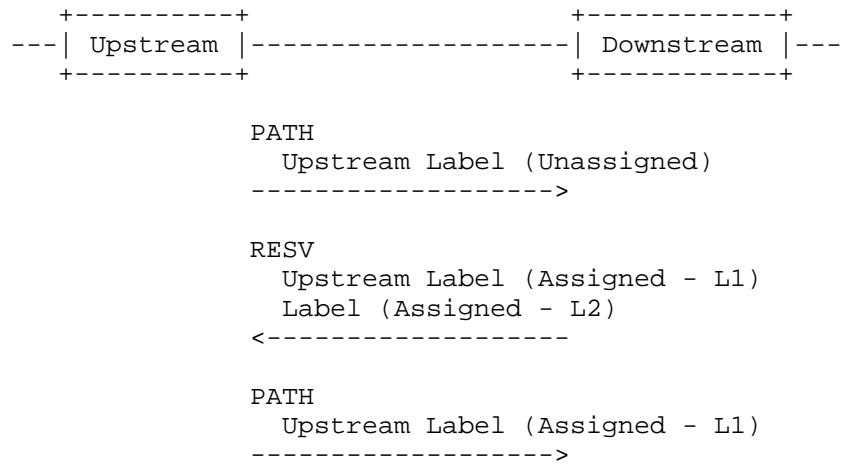


Figure 2: Unassigned UPSTREAM_LABEL

The above processing rules do not apply if an "Unassigned UPSTREAM_LABEL" is included in a PATH message that also has the "Label_Symmetry_Required" bit set. In that case, the downstream node would ignore the presence of the "UPSTREAM_LABEL" (and the rules specified in Section 2.1 come into play).

4. Upstream Label Set / Acceptable Upstream Label Set

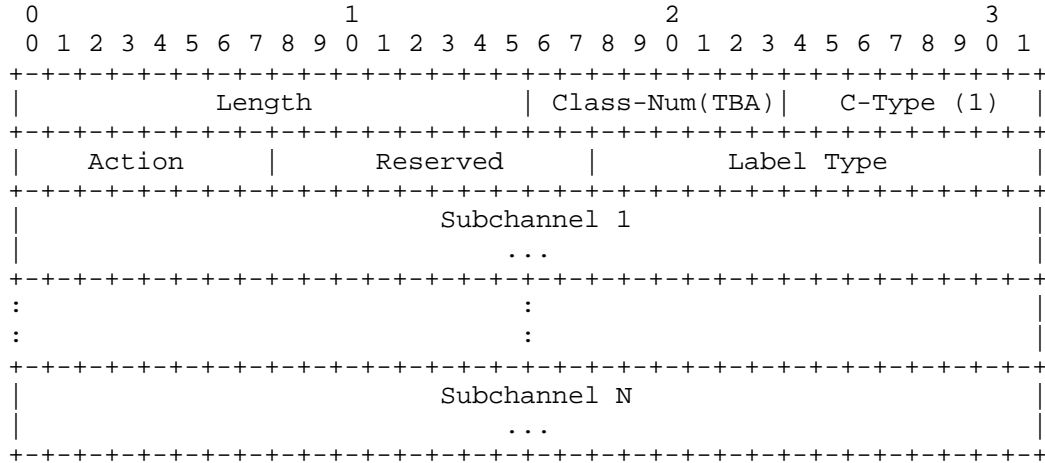
This document proposes the use of UPSTREAM_LABEL_SET and ACCEPTABLE_UPSTREAM_LABEL_SET for scenarios where a given node desires to give the network some choices when picking a valid UPSTREAM_LABEL. The UPSTREAM_LABEL_SET object is the upstream equivalent of the LABEL_SET object. The UPSTREAM_LABEL_SET object carries a list of acceptable upstream labels and gets signaled in the PATH message that is sent downstream. The network responds by picking a valid UPSTREAM_LABEL from the given list and signals it back in the corresponding RESV message.

The ACCEPTABLE_LABEL_SET is currently used to specify both upstream and downstream label-sets. However, in scenarios where there is no label symmetry, it becomes necessary to have constructs that can specify both an acceptable upstream label-set and an acceptable downstream label-set at the same time. The ACCEPTABLE_UPSTREAM_LABEL_SET construct introduced in this document helps fill that void.

4.1. Object Formats

The UPSTREAM_LABEL_SET object uses Class-Number TBA (of form 0bbbbbbb) and the C-Type of 1.

The format of UPSTREAM_LABEL_SET:



The parameters are similar to ones defined for LABEL_SET. See [RFC3471] for their description.

The ACCEPTABLE_UPSTREAM_LABEL_SET object uses class-number TBA (of form 10bbbbbb) and C-Type 1. The format/parameters of this object are identical to that of the UPSTREAM_LABEL_SET.

4.2. Processing Rules

The inclusion of the optional UPSTREAM_LABEL_SET object in the PATH message indicates that the LSP is bidirectional.

In the ideal case, the network picks a valid upstream-label from the specified list and fills this in the UPSTREAM_LABEL object of the corresponding RESV message. If the network is not able to pick a valid upstream-label from the list specified in the UPSTREAM_LABEL_SET, it MUST generate a PATH-ERR message with a "Routing Problem/Unacceptable Label value" indication. The PATH-ERR message may optionally include the ACCEPTABLE_UPSTREAM_LABEL_SET

object to indicate a list of acceptable labels supported by the network at that instant.

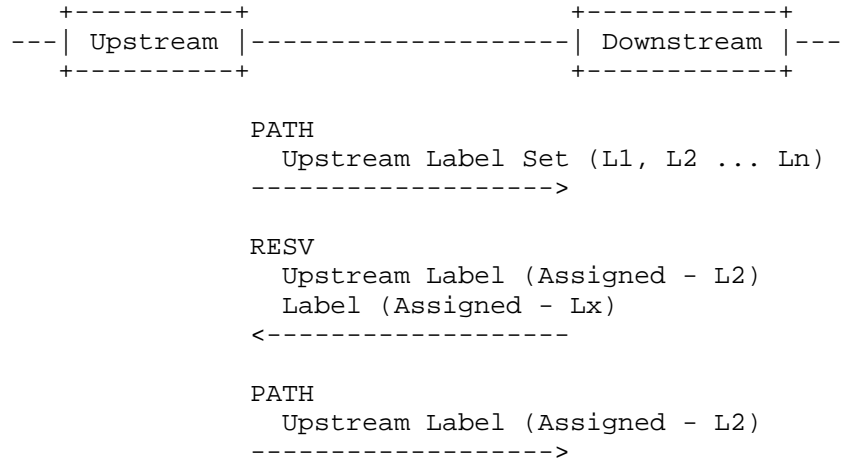


Figure 3: UPSTREAM_LABEL_SET

The UPSTREAM_LABEL object and the UPSTREAM_LABEL_SET object may both be included in a PATH message. The rules of processing when both objects are included are as follows:

- If the UPSTREAM_LABEL carries a valid assigned value, then the UPSTREAM_LABEL_SET object (if present) MUST be ignored.
- If the UPSTREAM_LABEL carries an unassigned value, then the Unassigned UPSTREAM_LABEL MUST be ignored. The UPSTREAM_LABEL_SET gets processed instead in such cases.

The above processing rules do not apply if an "UPSTREAM_LABEL_SET" is included in a PATH message that also has the "Label_Symmetry_Required" bit set. In that case, the downstream node would ignore the presence of the "UPSTREAM_LABEL_SET" (and the rules specified in Section 2.1 come into play).

5. Use-Cases

5.1. Alien-Wavelength Setup

Consider the network topology depicted in Figure 3. Nodes A and B are client IP routers that are connected to an optical WDM transport

network. F, H and I represent WDM nodes. The transponder sits on the router and is directly connected to the add-drop port on a WDM node.

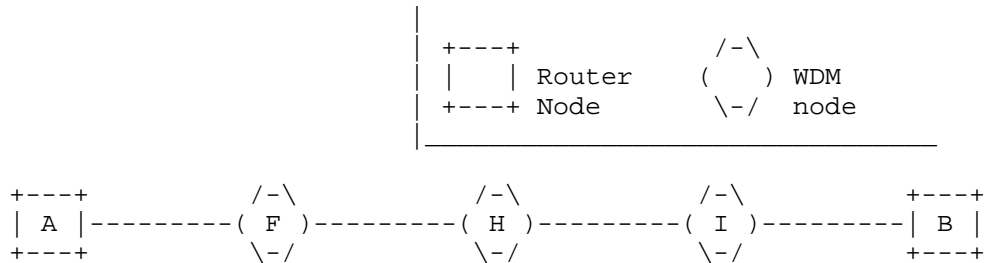


Figure 4: Sample topology

The optical signal originating on "Router A" is tuned to a particular wavelength. On "WDM-Node F", it gets multiplexed with optical signals at other wavelengths via an optical-filter. Depending on the implementation of this multiplexing function, it may not be acceptable to have the router send signal into the optical network unless it is at the correct wavelength. In particular, for some tunable filter implementations, multiplexing of signals with the same wavelength will result in an unreadable signal on that wavelength. Hence, having the router send signal with wrong wavelength may adversely impact existing optical trails. If the clients do not have full visibility into the optical network, they are not in a position to pick the correct wavelength up-front. The mechanisms proposed in this document allow the optical network specify the correct wavelength for such clients.

5.1.1. Setup Procedure - Example

The following is an illustration of gracefully setting up ([GR-SETUP]) a Lambda LSP using "Unassigned Upstream Label". "Label Symmetricity" is not requested for the LSP in this particular example.

```

+---+           /-\           /-\           +---+
| A |----- ( F ) ~~~~~ ( I )-----| B |
+---+           \-/           \-/           +---+

```

Step 1:

```

PATH
  Admin Status (A, R)
  Upstream Label (Unassigned)
----->
      -- ~~ -- ~~ -->
                                PATH
                                Admin Status (A, R)
                                ----->
                                RESV
                                Admin Status (A)
                                <-----
                                <-- ~~ -- ~~ --
RESV
  Admin Status (A)
  Upstream Label (Assigned)
<-----

```

Step 2:

```

PATH
  Admin Status (R),
  Upstream Label (Assigned)
----->
      -- ~~ -- ~~ -->
                                PATH
                                Admin Status (R)
                                ----->
                                RESV
                                Admin Status
                                <-----
                                <-- ~~ -- ~~ --
RESV
  Admin Status
  Upstream Label (Assigned)
<-----

```

Figure 5: Alien Wavelength Setup

Step 1:

- "Router A" does not have enough information to pick the correct client wavelength. It sends a PATH downstream requesting the network to assign an appropriate UPSTREAM_LABEL for it to use. As per the graceful setup procedure outlined in [GR-SETUP], the PATH is sent out with the "A" bit set in the ADMIN_STATUS. This indicates that the LSP is not operational and that the laser is turned off at the ingress client.
- The network receives the PATH, chooses the correct wavelength values and forwards them in appropriate label fields to the egress client ("Router B")
- "Router B" receives the PATH, turns the laser ON and tunes it to the correct wavelength (received in the LABEL_SET of the PATH) and sends out a RESV upstream. The RESV is sent out with the "A" bit set in the ADMIN_STATUS - indicating that the LSP is still not operational.
- The RESV received by the ingress client carries a valid assigned UPSTREAM label. "Router A" turns on the laser and tunes it to the wavelength specified in the network assigned UPSTREAM_LABEL. This completes Step-1.

Step 2:

- "Router A" sends out a PATH trigger with the "A" bit cleared in the ADMIN_STATUS. This indicates the ingress client's desire to make the LSP operational
- The network receives the PATH, adjusts the power-levels appropriately (also takes care of any other applicable provisioning operations) and then forwards the PATH with the "A" bit cleared to the egress client.
- The egress client sends out a RESV trigger in response with the "A" bit cleared in the ADMIN_STATUS. From this point on, the LSP is deemed "ready for use" by the egress client.
- The RESV with the "A" bit cleared in the ADMIN_STATUS makes its way to the ingress client. From this point on, the LSP is deemed fully operational by the ingress client.

6. Security Considerations

TBD

7. IANA Considerations

TBD

8. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3471] Berger, L., "Generalized Multi-Protocol Label Switching Signaling Functional Description", RFC 3471, January 2003
- [RFC3473] Berger, L., "Generalized Multi-Protocol Label Switching Signaling Resource Reservation Protocol-Traffic Engineering Extensions", RFC 3473, January 2003.
- [RFC5420] Farrel, A., "Encoding of Attributes for MPLS LSP Establishment Using Resource Reservation Protocol Traffic Engineering (RSVP-TE)", RFC5420, February 2009.
- [UP-LBL-SET] Oki, E., "Upstream Label Set Support in RSVP-TE extensions", <draft-oki-ccamp-upstream-labelset>, June 2002.
- [GR-SETUP] Beeram, V., "RSVP Graceful Setup", <draft-beeram-ccamp-rsvp-graceful-setup>, October 2013

9. Acknowledgments

We would like to acknowledge the authors of <draft-oki-ccamp-upstream-labelset> for introducing the notion of an UPSTREAM_LABEL_SET.

Authors' Addresses

Vishnu Pavan Beeram
Juniper Networks
Email: vbeeram@juniper.net

John Drake
Juniper Networks
Email: jdrake@juniper.net

Gert Grammel

Juniper Networks
Email: ggrammel@juniper.net

Igor Bryskin
ADVA Optical Networking
Email: ibryskin@advaoptical.com

Pawel Brzozowski
ADVA Optical Networking
Email: pbrzozowski@advaoptical.com

Daniele Ceccarelli
Ericsson
Email: daniele.ceccarelli@ericsson.com

CCAMP Working Group
Internet Draft
Intended status: Standards Track

Vishnu Pavan Beeram (Ed)
Juniper Networks
Igor Bryskin (Ed)
ADVA Optical Networking

Expires: April 20, 2014

October 20, 2013

RSVP Graceful Setup Procedure
draft-beeram-ccamp-rsvp-graceful-setup-00.txt

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), 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, 2014.

Copyright Notice

Copyright (c) 2013 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.

Abstract

The GMPLS RSVP-TE setup procedure for transport LSPs outlined in [RFC3473] involves a single iteration signaling sequence. However there are certain scenarios, where it is not feasible to make an LSP fully operational and ready for use via the existing single-step setup procedure. This document proposes a 2-iteration setup procedure for gracefully bringing up transport LSPs in such cases.

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 RFC-2119 [RFC2119].

Table of Contents

1. Introduction.....	2
2. Graceful Setup Procedure.....	3
3. Use Case.....	4
3.1. Lambda LSP setup.....	4
4. Security Considerations.....	4
5. IANA Considerations.....	4
6. Normative References.....	4
7. Acknowledgments.....	5

1. Introduction

The GMPLS RSVP-TE extensions required for setting up transport LSPs are discussed in [RFC3473]. As per the existing setup procedure, the signaling sequence commences with the ingress sending a PATH message downstream. The PATH message traverses through all the intermediate nodes and reaches the egress. The egress responds to the setup request by sending a RESV message upstream. The setup iteration is completed when the RESV reaches the ingress. At the end of this iteration, the LSP is deemed operational and ready for use at the ingress. Optionally, if the egress desires a confirmation, the ingress would send a RESV-CONFIRM message downstream. The LSP is deemed operational at the egress as soon as it receives the RESV-CONFIRM.

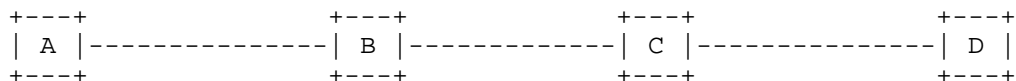
However, in certain cases (Section 3) there is no guarantee that the LSP is operational and ready for use at the end of this first iteration. This document proposes the use of a 2-iteration setup procedure to cater to those cases. By the end of this Graceful Setup Procedure, the end-points are guaranteed that the LSP is operational and ready for immediate use.

2. Graceful Setup Procedure

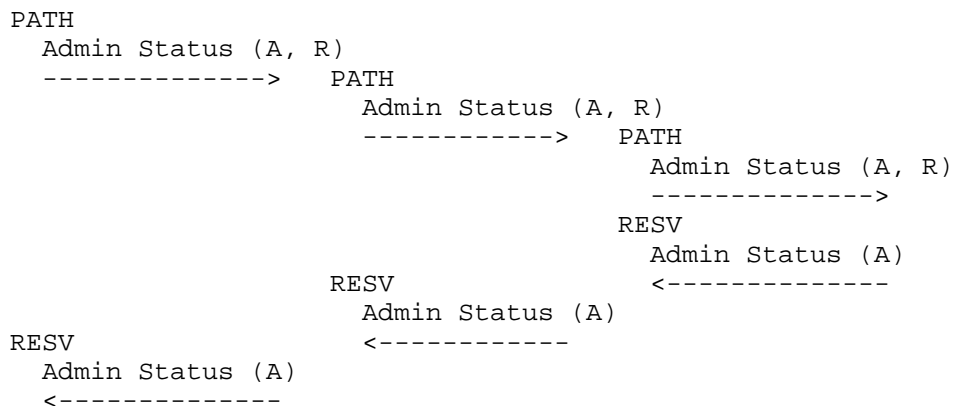
The RSVP graceful setup procedure (illustrated in Figure 1) proposed by this document involves two distinct steps. This setup procedure is similar in spirit to the 2-step RSVP graceful deletion procedure outlined in [RFC3473].

In the first step, the LSP is signaled as "non-operational" - the PATH is sent out with the "A" (Administratively Down) bit set in the ADMIN_STATUS object. All the resources along the path of the LSP are allocated and bound during this iteration.

The LSP is made "operational" only in the second step - the PATH is sent out with the "A" bit cleared in the ADMIN_STATUS. The LSP is deemed fully operational by the egress when it receives the PATH with the "A" bit cleared in the ADMIN_STATUS. Similarly, the LSP is deemed ready for immediate use by the ingress when it receives the RESV with the "A" bit cleared in the ADMIN_STATUS.



Step 1: Prepare the resources along the path of the LSP



Step 2: Make the LSP operational

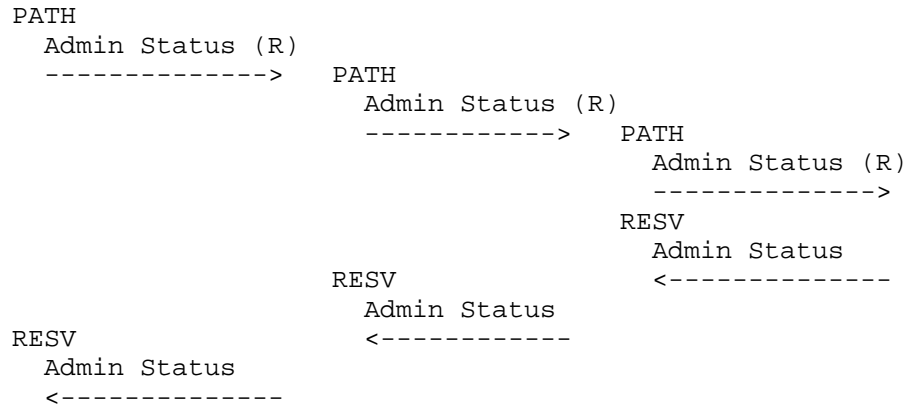


Figure 1: Graceful Setup Procedure - Signaling Sequence

3. Use Case

3.1. Lambda LSP setup

After all the cross-connects are set up in both directions at each node along the path of the LSP and the lasers are turned on at both the ends, the Lambda LSP may still not be ready for immediate use. Certain provisioning operations would need to be performed at each node along the path of the LSP before it is deemed operational. By adopting the Graceful Setup Procedure for Lambda LSPs, operations like "enabling alarm monitoring" and "equalizing power-levels" can get executed in the second step.

4. Security Considerations

TBD

5. IANA Considerations

None.

6. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate

Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC3473] Berger, L., "Generalized Multi-Protocol Label Switching Signaling Resource Reservation Protocol-Traffic Engineering Extensions", RFC 3473, January 2003

7. Acknowledgments

TBD

Authors' Addresses

Vishnu Pavan Beeram
Juniper Networks
Email: vbeeram@juniper.net

John Drake
Juniper Networks
Email: jdrake@juniper.net

Gert Grammel
Juniper Networks
Email: ggrammel@juniper.net

Igor Bryskin
ADVA Optical Networking
Email: ibryskin@advaoptical.com

Pawel Brzozowski
ADVA Optical Networking
Email: pbrzozowski@advaoptical.com

Daniele Ceccarelli
Ericsson
Email: daniele.ceccarelli@ericsson.com

CCAMP Working Group
Internet Draft
Intended status: Standards Track

Vishnu Pavan Beeram (Ed)
Juniper Networks
Igor Bryskin (Ed)
ADVA Optical Networking

Expires: April 21, 2014

October 21, 2013

Shared Resource Link Group (SRcLG)
draft-beeram-ccamp-srclg-00.txt

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), 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, 2014.

Copyright Notice

Copyright (c) 2013 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.

Abstract

This document introduces the concept of SRcLG ("Shared Resource Link Group") and discusses its usage in the context of mutually exclusive Virtual TE Links.

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 RFC-2119 [RFC2119].

Table of Contents

1. Introduction.....	2
2. Dynamic Mutual Exclusivity.....	3
3. Shared Resource Link Group (SRcLG).....	5
3.1. Construct.....	6
3.2. Advertising Rules.....	7
3.3. Processing Rules.....	7
4. Security Considerations.....	7
5. IANA Considerations.....	7
6. Normative References.....	7
7. Acknowledgments.....	8

1. Introduction

A Virtual TE Link (as defined in [RFC6001]) advertised into a Client Network Domain represents a potentiality to setup an LSP in the Server Network Domain to support the advertised TE link. The Virtual TE Link gets advertised like any other TE link and follows the same rules that are defined for the advertising, processing and use of regular TE links [RFC4202]. However, "mutual exclusivity" is one attribute that is specific to Virtual TE Links.

[DRAFT-MELG] discusses the different types of mutual exclusivity (Static vs Dynamic) that come into play, explains the need to advertise this information into the Client TE Domain and introduces a new TE construct (MELG) to carry static mutual exclusivity information.

This document is a companion document to [DRAFT-MELG]. It discusses "Dynamic Mutual Exclusivity" in detail and introduces a new TE construct (SRcLG) to carry dynamic mutual exclusivity information.

2. Dynamic Mutual Exclusivity

As discussed in [DRAFT-MELG], this type of mutual exclusivity exists temporarily within a given network configuration. It comes into play when two or more Virtual TE Links depend on the usage of the same shareable underlying server network domain resource. Mutual Exclusivity exists when the amount of the said server resource that is available for sharing is limited temporarily; it ceases to exist when sufficient amount of the resource is available for accommodating all corresponding Virtual TE Links.

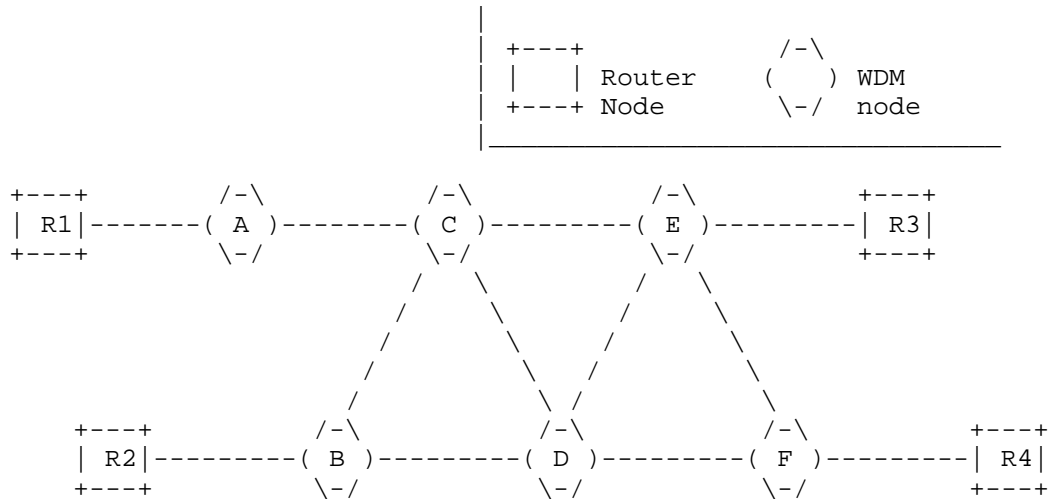


Figure 1a: Sample topology

Consider the network topology depicted in Figure 1a. This is a typical packet optical transport deployment scenario where the WDM layer network domain serves as a Server Network Domain providing transport connectivity to the packet layer network Domain (Client Network Domain).

Nodes R1, R2, R3 and R4 are IP routers that are connected to an Optical WDM transport network. A, B, C, D, E and F are WDM nodes

that constitute the Server Network Domain. The border nodes (A, B, E and F) operate in both the server and client domains. Figure 1b depicts how the Client Network Domain TE topology looks like when there are no Client TE Links provisioned across the optical domain.



Figure 1b: Client TE Database

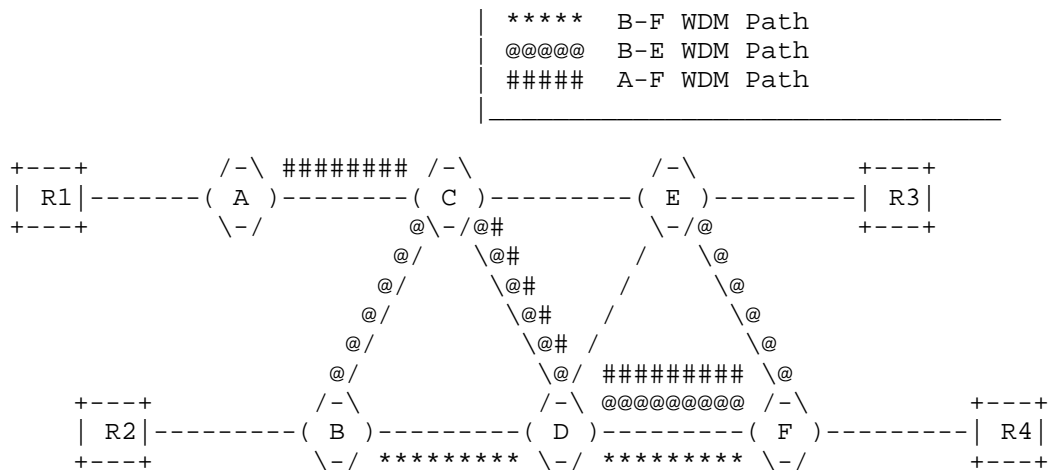


Figure 2a: Mutually Exclusive potential WDM paths

Now consider augmenting the Client TE topology by creating three Virtual TE Links across the optical domain. The potential paths in the WDM network catering to these three virtual TE links are as

shown in Fig 2a and the corresponding augmented Client TE topology is as illustrated in Fig 2b.

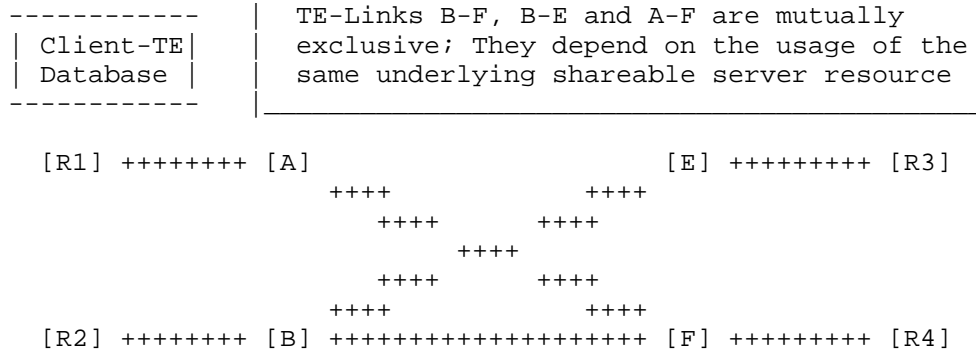


Figure 2b: Client TE Database - Mutually Exclusive Virtual TE Links

In this particular example, all three potential paths traverse through the WDM-Link {D-F}. Now assume that this link has only 2 lambda channels available. Also assume that any available lambda can get picked for each of these 3 corresponding underlying server LSPs. This means that only two out of the three Virtual TE Links can get committed at the moment. This dynamic mutual exclusivity ceases to exist when a third lambda channel becomes available on the WDM-link {D-F}.

This document proposes the use of "Shared Resource Link Group (SRcLG)" for catering to this scenario.

3. Shared Resource Link Group (SRcLG)

SRLG (Shared Risk Link Group - [RFC4202]) represents a set of links that share a resource whose failure may affect all links in the set. Since dynamic mutual exclusivity comes into play when the underlying server resource is shareable, all corresponding Virtual TE-Links would belong to the same SRLG. This document introduces the notion of a "Shared Resource Link Group (SRcLG)", which is meaningful only in the context of Virtual TE Links. SRcLG represents a set of Virtual TE-links that depend on the usage of a shared server-layer

resource that has a variable bandwidth capacity and as a result may sometimes not be able to simultaneously accommodate all corresponding Virtual TE-Links in the set. As is the case with SRLGs, a given Virtual TE Link may belong to multiple SRcLGs.

3.1. Construct

In terms of the TE construct that gets advertised, an SRcLG is nothing but an SRLG with some additional information to help determine which and how many of the corresponding virtual TE Links can get committed simultaneously. This additional information is the per-priority available shared resource bandwidth associated with a given SRLG. Since an SRcLG cannot exist without the presence of a corresponding SRLG, the SRcLG is identified by the corresponding 32-bit SRLG-ID. In other words, the SRcLG-ID is the same as the identifier of the SRLG it represents.

```

      0               1               2               3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     Shared Risk Link Group ID                                     |
+-----+-----+-----+-----+-----+-----+-----+-----+
| Available Shared Resource Bandwidth at Priority 0 |
+-----+-----+-----+-----+-----+-----+-----+-----+
| Available Shared Resource Bandwidth at Priority 1 |
+-----+-----+-----+-----+-----+-----+-----+-----+
| Available Shared Resource Bandwidth at Priority 2 |
+-----+-----+-----+-----+-----+-----+-----+-----+
| Available Shared Resource Bandwidth at Priority 3 |
+-----+-----+-----+-----+-----+-----+-----+-----+
| Available Shared Resource Bandwidth at Priority 4 |
+-----+-----+-----+-----+-----+-----+-----+-----+
| Available Shared Resource Bandwidth at Priority 5 |
+-----+-----+-----+-----+-----+-----+-----+-----+
| Available Shared Resource Bandwidth at Priority 6 |
+-----+-----+-----+-----+-----+-----+-----+-----+
| Available Shared Resource Bandwidth at Priority 7 |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

The SRcLG information advertised into the Client TE Domain is an unordered list of SRcLGs present in a given Virtual Topology. Unlike the SRLG construct or the MELG construct, the SRcLG construct does not get advertised per TE-Link. This is because the information carried in this construct is quite dynamic in nature and advertising it per TE-Link poses serious scaling concerns.

3.2. Advertising Rules

As far as the advertisement of a Virtual TE-Link is concerned, there is no perceived difference between SRLG and SRcLG. The 32-bit IDs of all SRcLGs that a Virtual TE-Link belongs to are advertised via the SRLG construct. Additionally, all SRcLG information associated with a given Virtual Topology is advertised into the Client TE Domain by the provider of the Virtual Topology. It is the responsibility of this provider to keep the bandwidth availability information for each SRcLG current with timely updates. The draft envisions that one or more server domain OSPF/ISIS TE speakers will be tasked to provide these timely updates. This TE speaker may advertise all SRcLG information (that it is responsible for) in the same OSPF-LSA/ISIS-LSP or advertise each SRcLG TLV separately - one in each OSPF-LSA/ISIS-LSP.

3.3. Processing Rules

The intended consumer of this SRcLG information is the PCE in the Client TE Domain. The Client PCE should take this advertised information into account when performing path selection for services over the Virtual Topology provided by the network domain. In particular, this information should be used when deciding how many Virtual TE links could be accommodated simultaneously on a given SRcLG at a given priority level.

4. Security Considerations

TBD

5. IANA Considerations

TBD

6. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4202] K.Kompella, Y.Rekhter, "Routing Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC4202, October 2005.
- [RFC6001] D.Papadimitriou, M.Vigoureux, K.Shiimoto, D.Brunyard

and JL. Le Roux, "GMPLS Protocol Extensions for Multi-Layer and Multi-Region Networks", RFC 6001, October 2010.

[DRAFT-MELG] Beeram, V., "Mutual Exclusive Shared Link Group", draft-beeram-ccamp-melg, October 2013

7. Acknowledgments

TBD

Authors' Addresses

Vishnu Pavan Beeram
Juniper Networks
Email: vbeeram@juniper.net

Igor Bryskin
ADVA Optical Networking
Email: ibryskin@advaoptical.com

Cyril Margaria
Email: cyril.margaria@gmail.com

Fatai Zhang
Huawei Technologies
Email: zhangfatai@huawei.com

CCAMP Working Group
Internet-Draft
Intended status: Informational
Expires: April 24, 2014

D. Ceccarelli
Ericsson
O. Gonzalez de Dios
Telefonica I+D
F. Zhang
X. Zhang
Huawei Technologies
October 21, 2013

Use cases for operating networks in the overlay model context
draft-ceccadedios-ccamp-overlay-use-cases-03

Abstract

This document defines a set of use cases for operating networks in the overlay model context through the Generalized Multiprotocol Label Switching (GMPLS) overlay interfaces.

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 24, 2014.

Copyright Notice

Copyright (c) 2013 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
2. Terminology	3
3. Background and Assumptions	4
4. Use Cases	6
4.1. UC 1 - Provisioning	6
4.2. UC 2 - Provisioning with optimization	7
4.3. UC 3 - Provisioning with constraints	7
4.4. UC 4 - Provisioning with diversity	8
4.5. UC 5 - Concurrent provisioning	9
4.6. UC 6 - Reoptimization	10
4.7. UC 7 - Query	10
4.8. UC 8 - Availability check	11
4.9. UC 9 - P2MP services	11
4.10. UC 10 - Privacy	11
4.11. UC 11 - Colored overlay	11
4.12. UC 12 - Stacking of overlay interfaces	13
5. Security Considerations	14
6. IANA Considerations	14
7. Contributors	14
8. References	15
8.1. Normative References	15
8.2. Informative References	15
Authors' Addresses	15

1. Introduction

The GMPLS overlay model [RFC 4208] specifies a client-server relationship between networks where client and server layers are managed as separate domains because of trustiness, scalability and operational issue. By means of procedures from the GMPLS protocol suite it is possible to build a topology in the client (overlay) network from Traffic Engineering paths in the server network. In this context, the UNI (User to Network Interface) is the demarcation point between networks. It is a boundary where policies, administrative and confidentiality issues apply that limit the exchange of information.

This GMPLS overlay model supports a wide variety of network scenarios. The packet over optical scenario is probably the most popular example where the overlay model applies.

In order to exploit the full potential of client/server network interworking in the overlay model, it may be desirable to know in advance whether is it feasible or not to connect two client network nodes [INTERCON-TE]. This requires to have a certain amount of TE information of the server network in the client network. This need not be the full set of TE information available within each network, but does need to express the potential of providing TE connectivity. This subset of TE information is called TE reachability information.

The goal of this document is to define a set of solution independent use cases applicable to the overlay model. In particular it focuses on the network scenarios where the overlay model applies and analyzes the most interesting aspects of provisioning, recovery and path computation.

2. Terminology

The following terms are used within the document:

- Edge node [RFC4208]: node of the client domain belonging to the overlay network, i.e. nodes with at least one interface connected to the server domain.
- Core node [RFC4208]: node of the server domain.
- Access link: link between core node and edge node. It is the link where the UNI is usually implemented.
- Remote node: node in the client domain which has no direct access to the server domain but can reach it through an edge node

in its same administrative domain.

- Local trigger: LSP setup request issued to an edge node. It triggers the setup of a client layer FA through the server domain via a UNI interface.

- Remote trigger: LSP setup request issued to a remote node. It triggers the setup of a client layer LSP which, upon reaching an edge node, will use connectivity in the server domain dynamically provided via an UNI interface.

3. Background and Assumptions

All the use cases listed in the sections below can be applied to any combination of, unless otherwise specified:

- * Local trigger or remote signaling
- * Grey interface or colored interface

With local trigger we mean the case in which a trigger for the provisioning of a service over the overlay interface is issued to one of the edge nodes belonging to the overlay network, i.e. directly connected to the UNI.

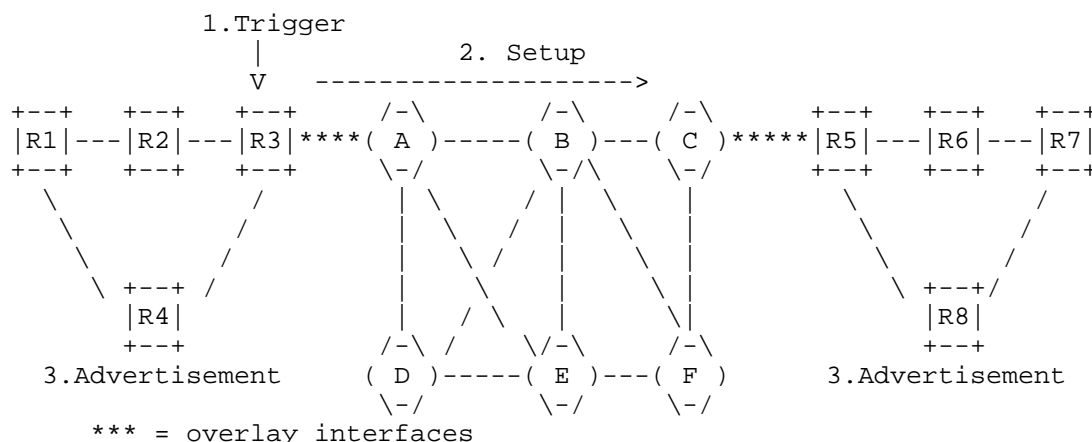


Figure 1: Local trigger

As it is possible to see in the figure above, a trigger is issued on R3 (edge node) for starting the setup request procedure over the

overlay interface (R3-A). Once the LSP in the server domain is setup and an adjacency in the packet layer between R3 and R5 is created, it can be advertised in the rest of the client domain and used by the signaling protocol (e.g. LDP) for setting up end-to-end (e.g. from R1 to R7) client layer LSPs.

On the other hand, the remote signaling consists on the utilization of a connection oriented signaling protocol in the client domain that allows issuing the end to end service setup trigger directly on the end nodes of the client domain. The signaling message, upon reaching the edge node (R3), will trigger the setup of the service in the server layer via the overlay interface.

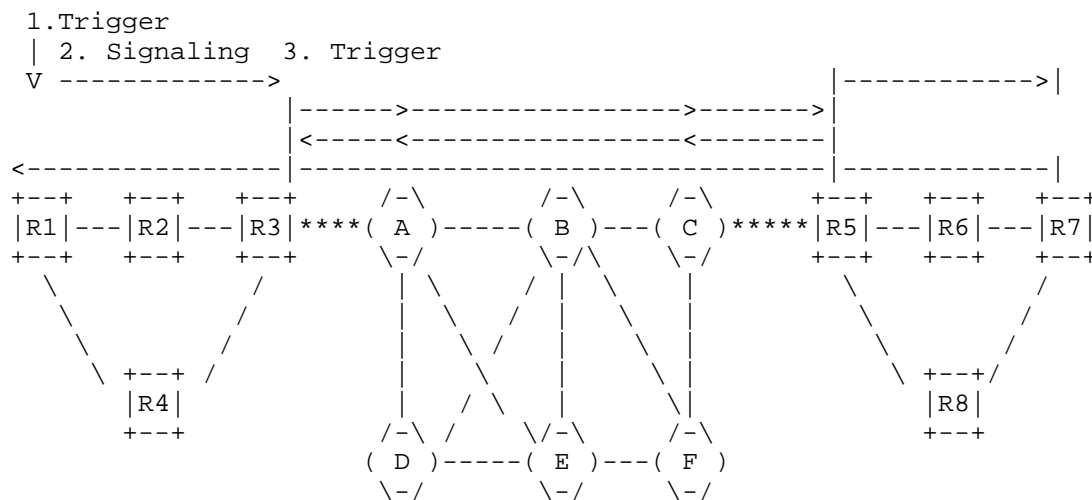


Figure 2: Remote Signaling

The utilization of the remote trigger allows for a strict control of the resources that will be used for the setup of the end to end service. In order to have a correct setup of the end to end service the trigger issued to R1 must include the overlay nodes to be used for the setup of the service in the server layer (R3 and R5). The network operator is supposed to know that the edge nodes to be used are R3 and R5.

When operating an IP over WDM network in the overlay context, a further distinction between grey and colored interfaces needs to be taken into account. In other words in the former case the transponder is hosted on the core border nodes, while in the latter in the edge node. The physical impairments to be considered are

different in the two cases (for further details please see Section 4.11) but the behaviour of the interface does not change and all use cases depicted below can be applied both to the grey and colored interfaces.

The particular case of grey and colored interfaces can be generalized introducing two further differentiation criteria for the characterization of overlay interfaces:

- * Administrative boundary or administrative plus technological boundary

Since the overlay is an administrative boundary between a client and a server layer, it is possible to configure it between a client and a server domain with the same switching capabilities (e.g., IP over IP) or between domains with different switching capabilities (e.g., OTN over WDM). In the former case the boundary is referred to as administrative domain, while in the latter, it is referred to as both administrative and technological boundary.

The second differentiation mentioned above refers to technological boundaries and in particular to:

- * Layer transition on edge node or on core node

When layer transition occurs on the edge node, the edge node is equipped with at least one interface with the switching capability of the client domain and one interface with the switching capability of the server domain. Referring to the IP over WDM this is the case of colored overlay interface with transponder hosted in the edge node. Viceversa, when layer transition occurs on the core node, it is the core node the one with at least two different interfaces with different switching capabilities and we speak about grey interfaces in the IP over WDM context.

Editor note: Actually path computation is assumed to be performed typically at the server layer. The client layer can request the server layer for computing a path or select among a set of paths computed by the server layer and exported to the client layer as virtual/abstract topology.

4. Use Cases

4.1. UC 1 - Provisioning

Requirement: The network operator must be allowed to setup an unprotected end to end service between two client layer nodes.

This use case simply consists on providing an operator with the capability of setting up a service in the client layer either by means of local trigger or remote signaling. The operator does not put any constraint over the path computation in the server layer.

4.2. UC 2 - Provisioning with optimization

Requirement: The network operator must be allowed to setup a service expressing which parameter must be optimized when computing the path.

This use case applies both to the local trigger and the remote signaling scenarios. In both cases the path computation element in the server layer (being it centralized or distributed) is demanded to provide a path between R3 and R5 which minimizes a given parameter (e.g. delay, jitter, TE metric).

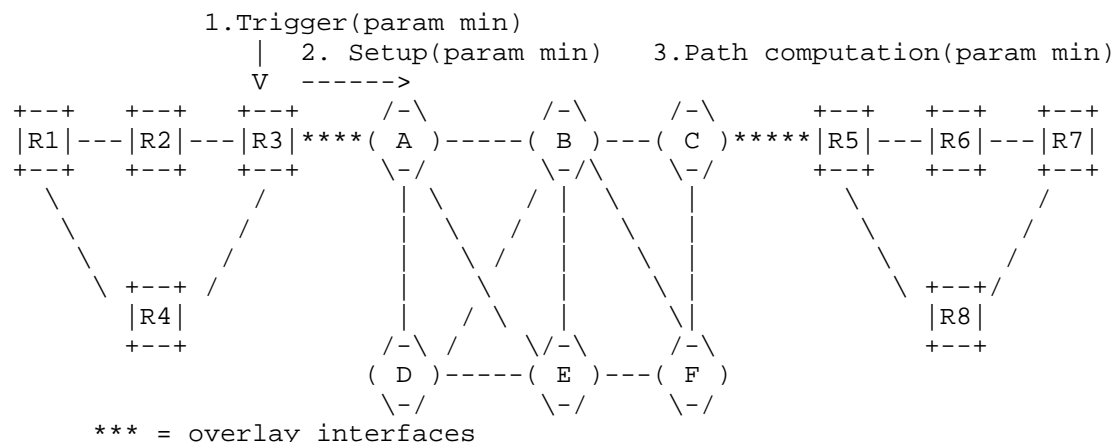


Figure 3: Provisioning with optimization

In the figure above the case of local trigger with specified parameter to be minimized is depicted, but same considerations apply to the remote signaling (trigger on R1). In that case the parameter to be minimized needs to be conveyed from R1 to R3 so that the setup request over the overlay interface can be issued taking into account the OF.

4.3. UC 3 - Provisioning with constraints

Requirement: The network operator must be allowed to setup a service imposing upper bounds for a set of parameters during the path computation.

This use case is extremely similar to the provisioning with Optimization one. This time, instead of/in addition to giving the possibility of specifying which parameter needs to be optimized during the path computation, the network operator is also able to indicate an upper bound for a set of parameters which is not being minimized in the path computation.

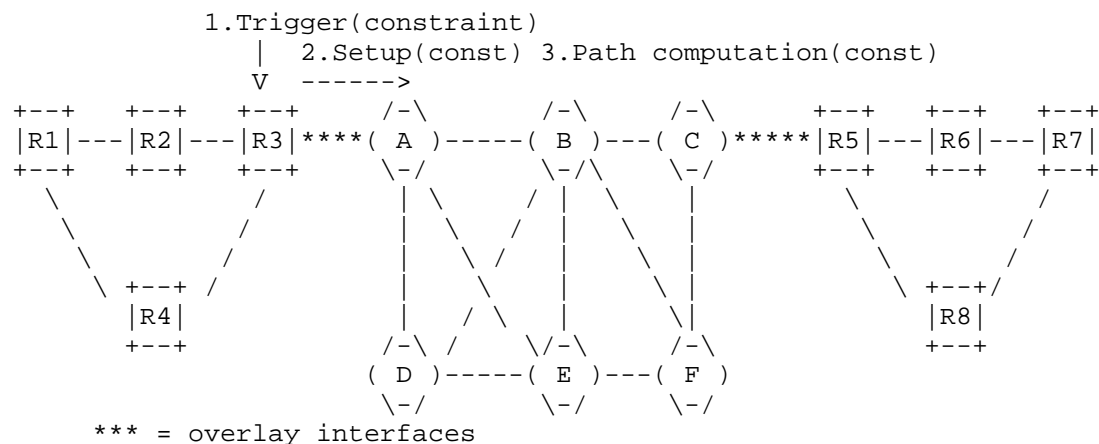


Figure 4: Provisioning with constraints

It is possible for example to ask for a path between R3 and R5 which, in addition to minimizing a given OF, does not introduce a delay higher than 10ms or where the jitter is not more than 3ms.

As per the optimization use case, when remote signaling is used (trigger on R1) a mean to convey the path computation constraints till the edge node (R3) is needed.

4.4. UC 4 - Provisioning with diversity

Requirement: The network operator must be allowed to setup a services in the server layer in diversity with respect to server layer resources or not sharing the same fate with other server layer services.

This scenario is extremely common in those cases where different services in the server domain are used to provision protected services in the client layer. The services in the server layer can be computed/provisioned sequentially or in parallel but in both cases the requirement is to have them totally disjoint, so that a single failure in the server layer does not impact two or more services in

the client layer which are supposed to be in a protection relationship between each other (e.g. 1+1 protection).

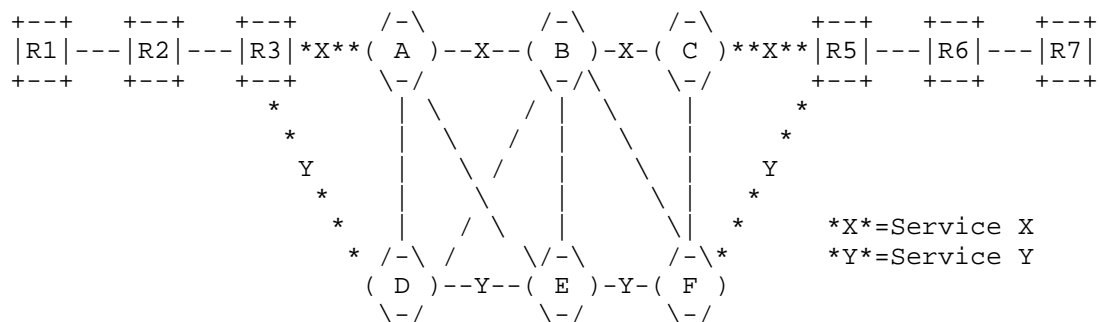


Figure 5: provisioning with diversity

In a scenario like the one depicted above, it is possible to use Service X and Service Y for the setup of a protected service in the client domain as a fault in the server domain would not impact both of them. In the case of parallel request, R3 asks the path computation in the server domain to provide two totally disjoint paths. On the other side, when sequential requests are issued, and identifiers for Service X (or a set of identifiers indicating its resources) is needed so that the request for the setup of Service Y can be issued with the constraint of avoiding the resources related to such identifier.

Another case of provisioning with diversity is the one where the operator in the client domains wants the server domain PCE to exclude some resources from the path computation because of e.g. trustness reasons. In such a case, supposing that such resources are known to the operator, it must be possible to indicate them as path computation constraint in the service setup request.

4.5. UC 5 - Concurrent provisioning

Requirement: The network operator must be allowed to setup a plurality of services not necessarily between the same pair of edge nodes.

Here is another case particularly interesting from a protection point of view. In the case above the same edge node was asking for different services in the server layer, but in order to have end to end diversity (i.e. from R1 to R8 in figure below), there is the need to be able to provide disjoint services between different pairs of

edge nodes.

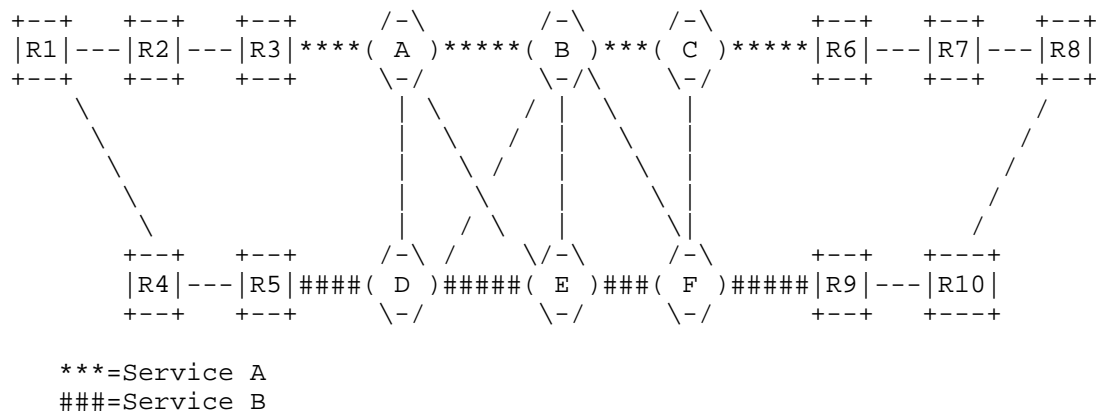


Figure 6: Concurrent provisioning

In this example Service A is provided between R3 and R6 and Service B between R5 and R9. Some sort of coordination is needed between R3 and R5 (directly between them or via R1) so that the requests to the server layer can be conveniently issued.

4.6. UC 6 - Reoptimization

Requirement: The network operator must be allowed to setup a plurality of services so that the overall cost of the network is minimized and not the cost of a single service.

TBD

4.7. UC 7 - Query

Requirement: The server network must be able to tell the network operator the actual parameters characterizing an existing service.

The capability of retrieving from the server domain some parameters qualifying a service can be extremely useful in different cases. One of them is the case of sequential provisioning with diversity requirements. In the case the operator wants to set-up a service in diversity from an existing one, hence it must be possible for the server domain to export some parameters univocally identifying the resources (e.g. SRLGs).

4.8. UC 8 - Availability check

Requirement: The network operator must be allowed to check if in the server layer there are enough resources to setup a service with given parameters.

TBD

4.9. UC 9 - P2MP services

Requirement: If allowed by the technology, the network operator must be allowed to setup a P2MP service with given parameters.

TBD

4.10. UC 10 - Privacy

Requirement: The network operator must be allowed to provision different groups of users with independent addressing spaces.

This is a particularly useful functionality for those cases where the resources of the service provider are leased and shared among several other service providers or customers.

4.11. UC 11 - Colored overlay

Requirement: The network operator must be allowed to provision a service in the server layer through a colored overlay interface.

This use case applies to networks where the server domain is a WDM network. In those cases it is possible to either have a grey interface between client and server domains (i.e. transponder on the border core node) or a colored interface between them (i.e. transponder on the edge node).

All the previous use cases assume the case of grey interface, but there are particular network scenarios in which it is possible to move the transponders from the core to the edge nodes and hence save on expensive pieces of hardware.

The issue with this solution is that the PCE in the server layer, being either centralized or distributed, has only visibility of what is inside the server domain and hence has not all the info needed to perform the validation of a path. The edge node must provide the PCE in the server domain with a set of info needed for a correct path computation and path validation from transponder to transponder (i.e. between edge nodes) all along the server domain.

The type of information needed for this scenario can be classified into three categories:

- Feasibility: Parameters like the output power of the transponder are needed in order to state e.g. the amount of km that can be reached without regeneration.
- Compatibility: The egress transponder must be compatible with the ingress one. Parameters that influence the level of compatibility can be for example the type of FEC (Forward Error Correction) used or the modulation format (which also impacts the feasibility together with the bit rate).
- Availability: Transponders can be tunable within a range of lambdas or even locked to a single lambda. This impacts the path computation as not every path in the network might have such lambda(s) supported or available at the time the path computation is performed.

In figure below it is possible to see that the PCE is aware of all the info between A and C (i.e. within the server domain scope) but what is missing is info related to the transponders on R1 and on R2 and of the access links. (i.e. R1-A and C-R2).

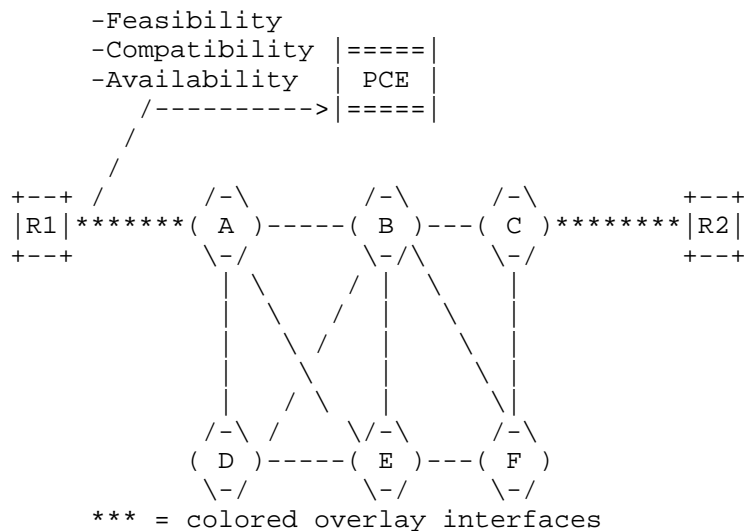


Figure 7: PCE feeding for colored UNI

There is not yet a standard set of parameters that is needed for path

computation in WDM networks but an example of some of them is provided in the following list:

- o Modulation format
- o FEC (type or gain)
- o Minimum transponder output power
- o Bitrate
- o Dispersion tolerance
- o OSNR (minimum required)

4.12. UC 12 - Stacking of overlay interfaces

Requirement: The network operator must be allowed manage a network with an arbitrarily high number of administrative boundaries (i.e., >2).

Operators might want to split their overlay networks in a number of administrative domains for several reasons, among which simplifying network operations and improving scalability. In order to do so it must be possible to create a stack of overlay interfaces between the different domains as shown in figure below:

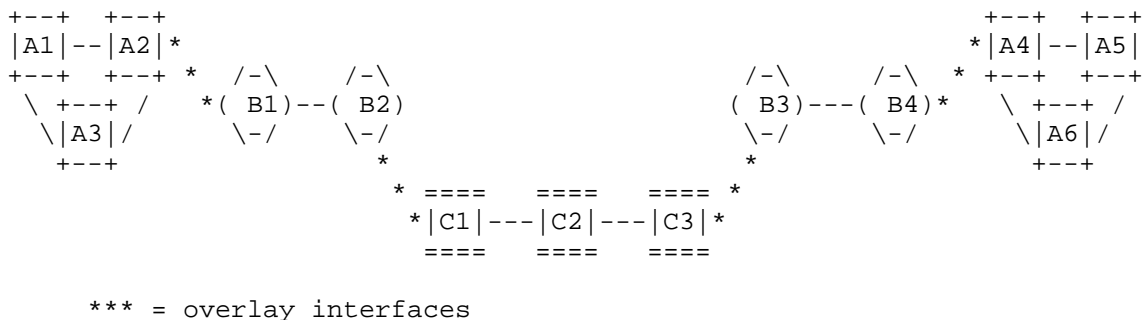


Figure 8: Stacking of interfaces

Nodes "Ax" belong to a domain which is client to the domain composed by nodes "Bx". The domain composed by nodes Bx is hence server layer to the "Ax" nodes domain but client to the "Cx" nodes domain.

A pretty common deployment of this scenario consists of IP over OTN

over WDM layers, where the OTN digital layer is used for the grooming of IP traffic over high bit rate lambdas. In figure 8, Node Bx can be assumed to be digital layer, which is interfacing with packet layer nodes (Ax) across overlay interface. Digital layer nodes Bx are interfacing with DWDM layer nodes Cx. If OTN (Bx) and DWDM (Cx) node belong to same IGP, then this becomes multi-layer path computation and signaling case, and it is out of scope of this document.

However, as already shown in the intro of this memo, the three different domains of the example could have the same switching capability (e.g., IP) and be kept separate just for administrative reasons.

5. Security Considerations

TBD

6. IANA Considerations

TBD

7. Contributors

Diego Caviglia, Ericsson

Via E.Melen, 77 - Genova - Italy

Email: diego.caviglia@ericsson.com

Jeff Tantsura, Ericsson

300 Holger Way, San Jose, CA 95134 - USA

Email: jeff.tantsura@ericsson.com

Khuzema Pithewan, Infinera Corporation

140 Caspian CT., Sunnyvale - CA - USA

Email: kpithewan@infinera.com

Cyril Margaria, Wandl

Email: cyril.margaria@googlemail.com

John Drake, Juniper

Email: jdrake@juniper.net

Sergio Belotti, Alcatel-Lucent

Email: sergio.belotti@alcatel-lucent.com

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

Authors' Addresses

Daniele Ceccarelli
Ericsson
Via E. Melen 77
Genova - Erzelli
Italy

Email: daniele.ceccarelli@ericsson.com

Oscar Gonzalez de Dios
Telefonica I+D
Don Ramon de la Cruz 82-84
Madrid 28045
Spain

Email: ogondio@tid.es

Fatai Zhang
Huawei Technologies
F3-5-B R&D Center, Huawei Base
Shenzhen 518129 P.R.China Bantian, Longgang District
Phone: +86-755-28972912

Email: zhangfatai@huawei.com

Xian Zhang
Huawei Technologies
F3-5-B R&D Center, Huawei Base
Shenzhen 518129 P.R.China Bantian, Longgang District
Phone: +86-755-28972913

Email: zhang.xian@huawei.com

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: April 23, 2014

D. Hiremagalur, Ed.
G. Grammel, Ed.
J. Drake, Ed.
Juniper
G. Galimberti, Ed.
Z. Ali, Ed.
Cisco
R. Kunze, Ed.
Deutsche Telekom
October 20, 2013

Extension to the Link Management Protocol (LMP/DWDM -rfc4209) for Dense
Wavelength Division Multiplexing (DWDM) Optical Line Systems to manage
application code of optical interface parameters in DWDM application
draft-dharinigert-ccamp-g-698-2-lmp-04

Abstract

This memo defines extensions to LMP(rfc4209) for managing Optical parameters associated with Wavelength Division Multiplexing (WDM) systems or characterized by the Optical Transport Network (OTN) in accordance with the Interface Application Code approach defined in ITU-T Recommendation G.698.2.[ITU.G698.2], G.694.1.[ITU.G694.1] and its extensions./>

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

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 23, 2014.

Copyright Notice

Copyright (c) 2013 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	2
2. Extensions to LMP-WDM Protocol	3
3. Black Link General Parameters - BL_General	4
4. Black Link ApplicationCode - BL_ApplicationCode	4
5. Black Link Vendor Transceiver Class - BL_ApplicationCode . .	5
6. Black Link - BL_Ss	6
7. Black Link - BL_Rs	7
8. Security Considerations	8
9. IANA Considerations	8
10. References	9
10.1. Normative References	9
10.2. Informative References	9
Authors' Addresses	10

1. Introduction

This extension is based on "draft-galikunze-ccamp-g-698-2-snmp-mib-03" and "draft-kunze-g-698-2-management-control-framework-02", for the relevant interface optical parameters described in recommendations like ITU-T G.698.2 [ITU.G698.2]. The LMP Model from RFC4902 is extended to provide link property correlation between a client and an OLS device. By using LMP, the capabilities of either end of this link are exchanged where the term 'link' refers to the attachment link between OXC and OLS (see Figure 1). By performing link property correlation, both ends of the link can agree on a common parameter window that can be supported and supervised by each device. The actual selection of a specific parameter value within the parameter window is outside the scope of LMP. In GMPLS the parameter selection (e.g. wavelength) is performed by RSVP-TE and Wavelength routing by IGP.

Figure 1 Extended LMP Model (from [RFC4209])

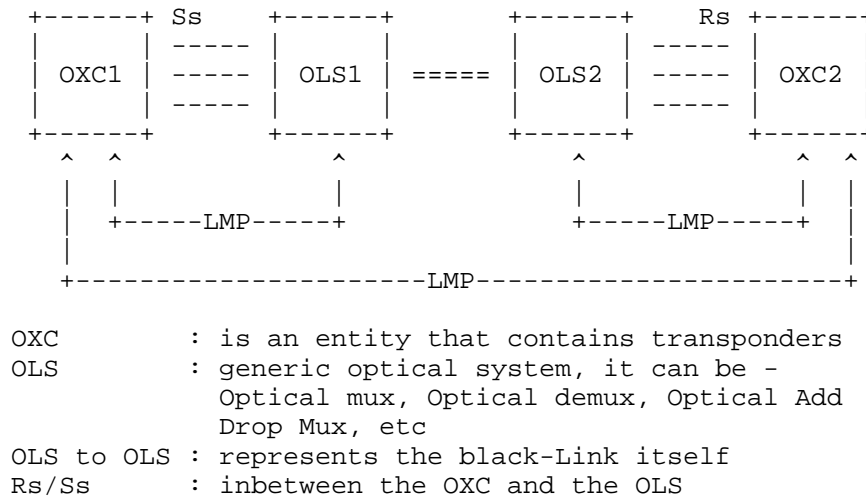


Figure 1: Extended LMP Model

2. Extensions to LMP-WDM Protocol

This document defines extensions to [RFC4209] to allow the Black Link (BL) parameters of G.698.2, as described in the draft draft-kunze-g-698-2-management-control-framework-02, to be exchanged between a router or optical switch and the optical line system to which it is attached. In particular, this document defines additional Data Link sub-objects to be carried in the LinkSummary message defined in [RFC4204] and [RFC6205]. The OXC and OLS systems may be managed by different Network management systems and hence may not know the capability and status of their peer. The intent of this draft is to enable the OXC and OLS systems to exchange this information. These messages and their usage are defined in subsequent sections of this document.

The following new messages are defined for the WDM extension for
ITU-T G.698.2 [ITU.G698.2]/ITU-T G.698.1 [ITU.G698.1]/
ITU-T G.959.1 [ITU.G959.1]

- BL_General (sub-object Type = TBA)
- BL_ApplicationCode (sub-object Type = TBA)
- BL_VendorTransceiverClass (sub-object Type = TBA)
- BL_Ss (sub-object Type = TBA)
- BL_Rs (sub-object Type = TBA)

3. Black Link General Parameters - BL_General

These are the general parameters as described in [G698.2] and [G.694.1]. Please refer to the "draft-galikunze-ccamp-g-698-2-snmp-mib-04" for more details about these parameters and the [RFC6205] for the wavelength definition.

The general parameters are

1. Bit-Rate/line coding of optical tributary signals
2. Wavelength - (Tera Hertz) 4 bytes (see RFC6205 sec.3.2)
3. Number of Application Codes Supported
4. Number of Vendor Transceiver Classes Supported

Figure 2: The format of the this sub-object (Type = TBA, Length = TBA) is as follows:

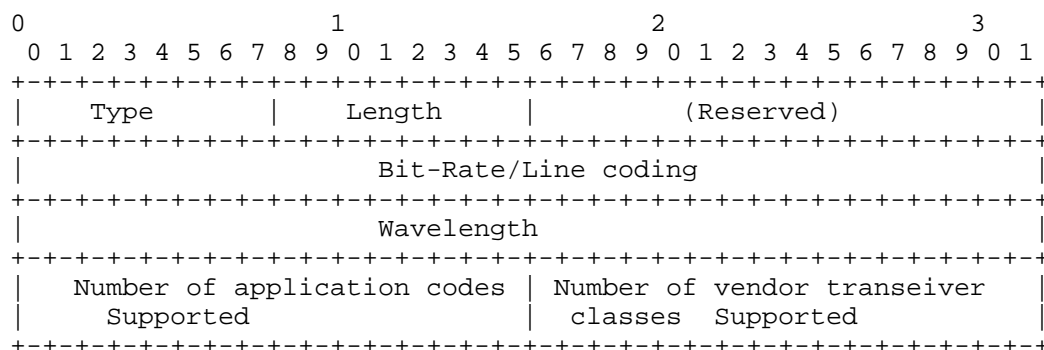


Figure 2: BL_General

4. Black Link ApplicationCode - BL_ApplicationCode

This message is to exchange the application code supported as described in [G698.2]. Please refer to the "draft-galikunze-ccamp-g-698-2-snmp-mib-04". for more details about these parameters. There can be more than one Application Code supported by the OXC/OLS.

The number of application codes supported is exchanged in the "BL_General" message. (from [G698.1]/[G698.2]/[G959.1])

The parameters are

1. Single-channel application code identifier - 8 bits
2. Single-channel application codes -- 32 bytes
(from [G698.1]/[G698.2]/[G959.1] - this parameter can have multiple instances as the transceiver can support multiple application codes.

Figure 3: The format of the this sub-object (Type = TBA, Length = TBA) is as follows:

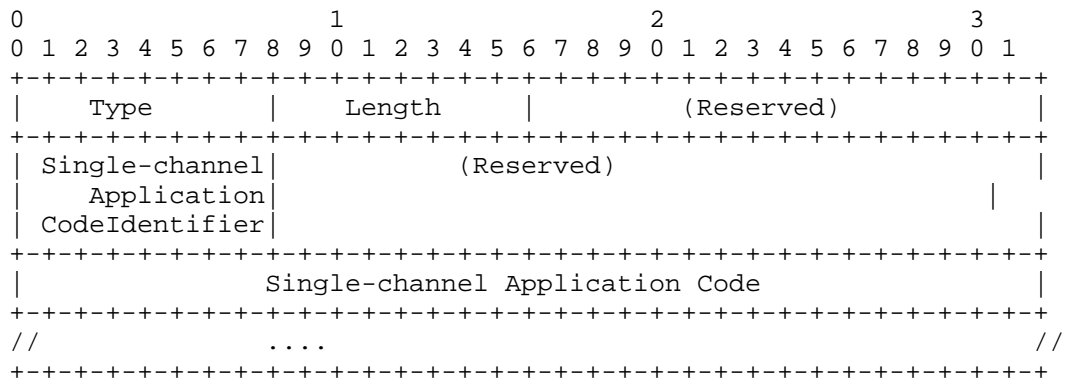


Figure 3: BL_ApplicationCode

5. Black Link Vendor Transceiver Class - BL_ApplicationCode

This message is to exchange the application code supported as described in [G698.2]. Please refer to the "draft-galikunze-ccamp-g-698-2-snmp-mib-04". for more details about these parameters. There can be more than one Vendor Transceiver Class supported by the OXC/OLS. The number of Vendor Transceiver Classes supported is exchanged in the "BL_General" message. (from [G698.1]/[G698.2]/[G959.1])

The parameters are

1. Single-channel Transceiver Class identifier - 8 bits

- 2. Vendor Transceiver Class -- 32 bytes
(from [G698.1]/[G698.2]/[G959.1] - this parameter can have multiple instances as the transceiver can support multiple application codes.

Figure 4: The format of the this sub-object (Type = TBA, Length = TBA) is as follows:

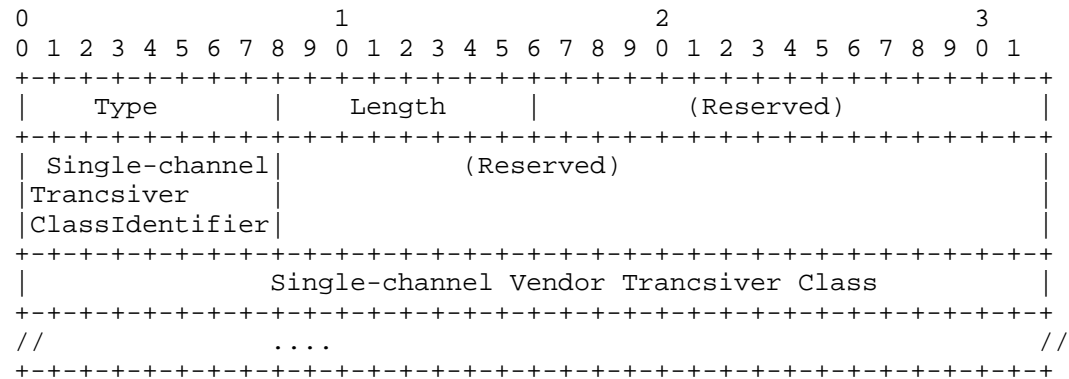


Figure 4: BL_VendorTransceiverClass

6. Black Link - BL_Ss

These are the G.698.2 parameters at the Source(Ss reference points). Please refer to "draft-galikunze-ccamp-g-698-2-snmp-mib-03" for more details about these parameters.

- 1. Output power
- 2. Minimum Mean Channel Output Power -(0.1 dbm) 4 bytes
- 3. Maximum Mean Channel Output Power -(0.1 dbm) 4 bytes

Figure 5: The format of the Black link sub-object (Type = TBA, Length = TBA) is as follows:

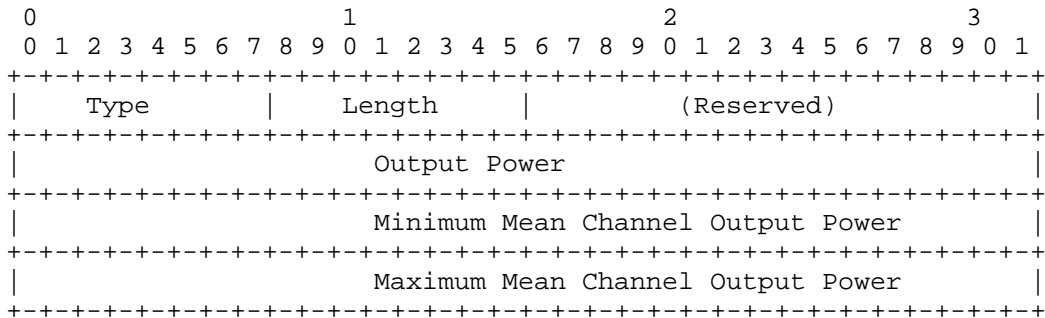


Figure 5: Black Link - BL_Ss

7. Black Link - BL_Rs

These are the G.698.2 parameters at the Sink (Rs reference points). Please refer to the "draft-galikunze-ccamp-g-698-2-snmp-mib-02" for more details about these parameters.

- 1. Current Input Power - (0.1dbm) 4bytes
- 2. Minimum Mean Input Power - (0.1dbm) 4bytes
- 3. Maximum Mean Input Power - (0.1dbm) 4bytes
- 4. Minimum OSNR - (0.1dB) 4bytes
- 5. OSNR Tolerance - (0.1dB) 4bytes

Figure 6: The format of the Black link sub-object (Type = TBA, Length = TBA) is as follows:

The format of the Black Link/OLS Sink sub-object (Type = TBA, Length = TBA) is as follows:

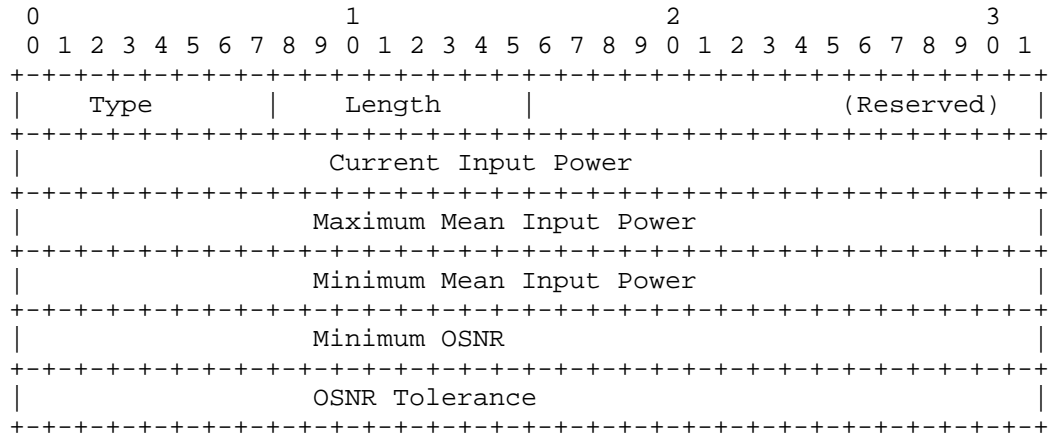


Figure 6: Black Link - BL_Rs

8. Security Considerations

LMP message security uses IPsec, as described in [RFC4204]. This document only defines new LMP objects that are carried in existing LMP messages, similar to the LMP objects in [RFC:4209]. This document does not introduce new security considerations.

9. IANA Considerations

LMP <xref target="RFC4204"/> defines the following name spaces and the ways in which IANA can make assignments to these namespaces:

- LMP Message Type
 - LMP Object Class
 - LMP Object Class type (C-Type) unique within the Object Class
 - LMP Sub-object Class type (Type) unique within the Object Class
- This memo introduces the following new assignments:

LMP Sub-Object Class names:

under DATA_LINK Class name (as defined in <xref target="RFC4204"/>)

- BL_General (sub-object Type = TBA)
- BL_ApplicationCode (sub-object Type = TBA)
- BL_VendorTransceiverClass (sub-object Type = TBA)
- BL_Ss (sub-object Type = TBA)
- BL_Rs (sub-object Type = TBA)

10. References

10.1. Normative References

- [RFC4204] Lang, J., "Link Management Protocol (LMP)", RFC 4204, October 2005.
- [RFC4209] Fredette, A. and J. Lang, "Link Management Protocol (LMP) for Dense Wavelength Division Multiplexing (DWDM) Optical Line Systems", RFC 4209, October 2005.
- [RFC6205] Otani, T. and D. Li, "Generalized Labels for Lambda-Switch-Capable (LSC) Label Switching Routers", RFC 6205, March 2011.
- [RFC4054] Strand, J. and A. Chiu, "Impairments and Other Constraints on Optical Layer Routing", RFC 4054, May 2005.
- [ITU.G698.2] International Telecommunications Union, "Amplified multichannel dense wavelength division multiplexing applications with single channel optical interfaces ", ITU-T Recommendation G.698.2, November 2009.
- [ITU.G694.1] International Telecommunications Union, "'Spectral grids for WDM applications: DWDM frequency grid" ", ITU-T Recommendation G.698.2, February 2012.
- [ITU.G709] International Telecommunications Union, "Interface for the Optical Transport Network (OTN) ", ITU-T Recommendation G.709, March 2003.
- [ITU.G872] International Telecommunications Union, "Architecture of optical transport networks ", ITU-T Recommendation G.872, November 2001.

10.2. Informative References

- [I-D.kunze-g-698-2-management-control-framework] Kunze, R., "A framework for Management and Control of optical interfaces supporting G.698.2", draft-kunze-g-698-2-management-control-framework-00 (work in progress), July 2011.
- [I-D.galimbe-kunze-g-698-2-snmp-mib]

Kunze, R. and D. Hiremagalur, "A SNMP MIB to manage black-link optical interface parameters of DWDM applications", draft-galimbe-kunze-g-698-2-snmp-mib-02 (work in progress), March 2012.

Authors' Addresses

Dharini Hiremagalur (editor)
Juniper
1194 N Mathilda Avenue
Sunnyvale - 94089 California
USA

Phone: +1408
Email: dharinih@juniper.net

Gert Grammel (editor)
Juniper
1194 N Mathilda Avenue
Sunnyvale - 94089 California
USA

Phone: +1408
Email: ggrammel@juniper.net

John E. Drake (editor)
Juniper
1194 N Mathilda Avenue
HW-US, Pennsylvania
USA

Phone: +1408
Email: jdrake@juniper.net

Gabriele Galimberti (editor)
Cisco
Via Philips,12
20052 - Monza
Italy

Phone: +390392091462
Email: ggalimbe@cisco.com

Zafar Ali (editor)
Cisco
3000 Innovation Drive
KANATA
ONTARIO K2K 3E8

Email: zali@cisco.com

Ruediger Kunze (editor)
Deutsche Telekom
Dddd, xx
Berlin
Germany

Phone: +49xxxxxxxxxx
Email: RKunze@telekom.de

Network Working Group

Iftekhar Hussain

Rajan Rao

Marco Sosa

Infinera

Abinder Dhillon

Fujitsu

October 16, 2013

Internet Draft

Intended status: Standard Track

Expires: Apr 16, 2014

OSPFTE extension to support GMPLS for Flex Grid
draft-dhillon-ccamp-super-channel-ospfte-ext-06.txt

Abstract

This document specifies the extension to TELINK LSA of OSPF routing protocol [RFC4203] [3] in support of GMPLS [1] for flex-grid networks [2].

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), 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 Apr 16, 2014.

Copyright Notice

Copyright (c) 2012 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.....	2
2. Terminology.....	3
3. Interface Switching Capability Descriptor.....	3
3.1. Switch Capability Specific Information	5
3.2. BW sub TLV: Bit Map format.....	5
3.2.1. Meaning of sub TLV fields.....	5
3.3. BW sub TLV: List and Rage format.....	7
3.3.1. Meaning of sub TLV fields.....	7
3.4. BW advertisement procedure.....	8
4. Examples.....	8
4.1. Example: BW advertisement without any service present.....	8
4.2. Example: How to use advertized Bandwidth.....	9
5. Security Considerations.....	10
6. IANA Considerations.....	10
7. References.....	10
7.1. Normative References.....	10
7.2. Informative References.....	10
8. Acknowledgments.....	11

1. Introduction

To enable scaling of existing transport systems to ultra-high data rates of 1 Tbps and beyond, next generation systems providing super-channel[2] switching capability are currently being developed. To allow efficient allocation of optical spectral bandwidth for such high bit rate systems, International Telecommunication Union Telecommunication Standardization Sector (ITU-T) is extending the

G.694.1 grid standard (termed ''Fixed-Grid'') to include flexible grid (termed ''Flex-Grid'') support [10].

This document defines OSPF-TE extensions in support of flex-grid networks.

Figure-1 shows a network capable of switching in Flexible-Grid[10]. The physical media/Fiber is modeled as a TE-Link to advertise spectrum (bandwidth) availability. This information is used during Flex-grid LSP[10] creation (also called super-channel LSPs[2]). This draft defines extensions to ISCD in support of Flexible-Grid.

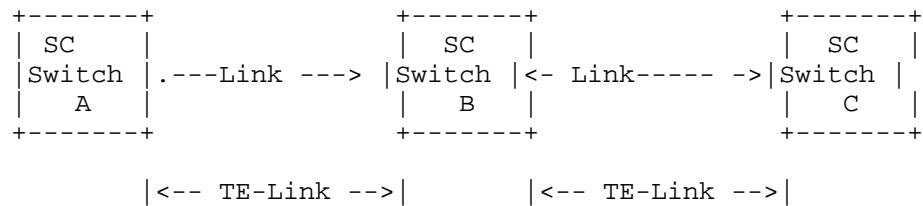


Figure 1: TE-Links

2. Terminology

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].

3. Interface Switching Capability Descriptor

The Interface Switching Capability Descriptor describes switching capability of an interface [RFC 4203]. This document defines a new Switching Capability value for Flex Grid [FLEX-GRID] as follows:

Value	Type
-----	----
102 (TBA by IANA)	Super-Channel-Switch-Capable (SCSC)

Switching Capability and Encoding values MUST be used as follows:

Switching Capability = SCSC

Encoding Type = Lambda [as defined in RFC3471]

The Interface Switching Capability Descriptor is a sub-TLV (of type 15) of the Link TLV. The length is the length of value field in Octets. The format of the value field is as shown below:

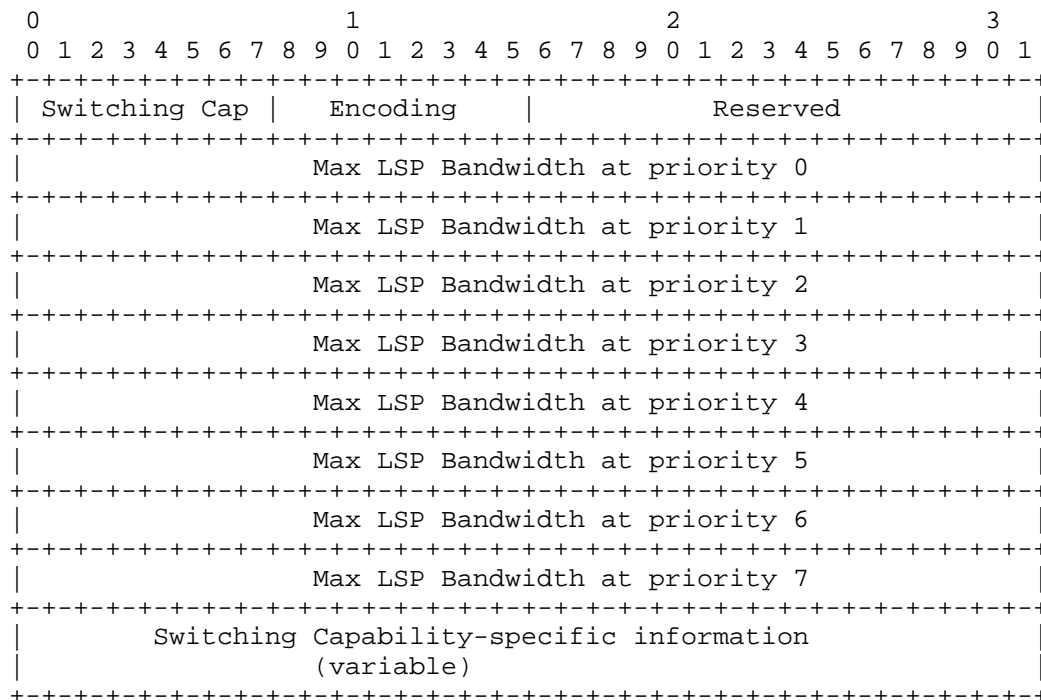


Figure 2: ISCD & SCSI

Max LSP Bandwidth will be based on Max Slot Width field in BW-sub-TLV (Ref to section 3.1 for details on BW sub-TLV) and the modulation format used.

3.1. Switch Capability Specific Information

The technology specific part of the ISCD can include a variable number of sub-TLVs. We propose to encode Slice Information in Bandwidth sub-TLVs under SCSI field. The format of BW sub-TLVs is as shown below.

[Editor's note: To provide options similar to Label set field defined in [9], we have included 2 variants to advertise slice level information. These are bit-format and list/range format].

3.2. BW sub TLV: Bit Map format

The figure below shows format of Type=1 sub-TLV for encoding slice information in bit-map format. This sub-TLV must be repeated for each priority that is supported on the Te-link.

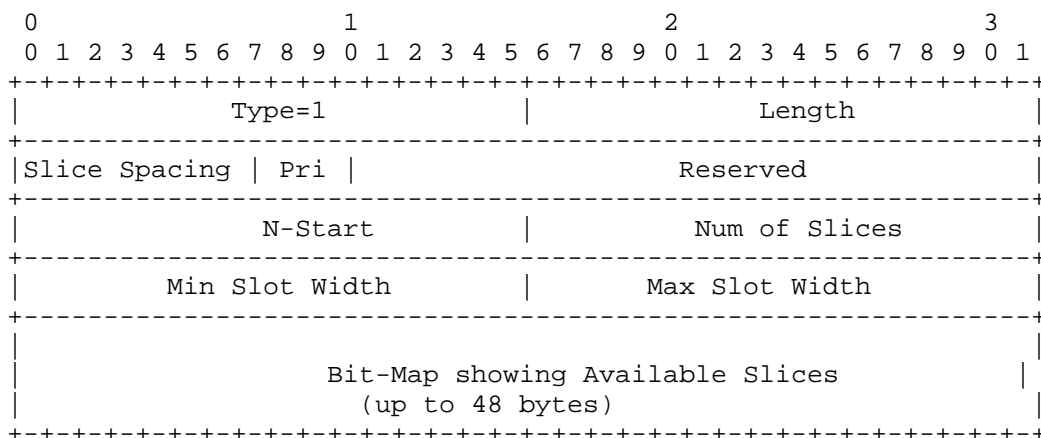


Figure 3: Type=1 BW sub TLV in Bit-Map format

3.2.1. Meaning of sub TLV fields

- o Slice Spacing: 8-bit field (S.S) which can take one of the values as shown in table below.
 - o For e.g., the 12.5GHz spacing is specified by setting this field to value 4.

S.S. (GHz)	Value
Reserved	0
100	1
50	2
25	3
12.5	4
Future use	5 - 15

Table 1: Slice Spacing Values

- o Priority: 3-bit field
 - o 3-bit field to identify one of the 8 priorities for which Slice information (BW) is advertised.
- o N-Start: 16-bit field
 - o Is a two's complement integer to specify start of the grid
 - o Use center freq formula to determine start of spectrum
- o Number of slices: 16-bit field
 - o Total number of slices advertised for the link. This includes (available plus consumed).
- o Minimum Slot Width: 16-bit field
 - o This is a positive integer value
 - o This field is similar to Min LSP BW field. The value in this field is used to determine the smallest frequency slot width that the advertising node can allocate for an LSP. This is defined by the following equation:

$$\text{Smallest Frequency slot width} = \text{Slice Spacing} * \text{integer value in 'Minimum Slot Width' field}$$
- o Maximum Slot Width: 16-bit field
 - o This is a positive integer value
 - o This field is used to determine the Maximum contiguous frequency slot width that the advertising node can allocate for an LSP. This is defined by the following equation:

$$\text{Largest Contiguous Frequency slot width} = \text{Slice Spacing} * \text{integer value in 'Maximum Slot Width' field}$$
- o Available slices encoded as bit-map
 - o Each bit represents availability of one slice of width identified by S.S field
 - o Zero: Available ; One: occupied
 - o Padding MUST be used to align with 32 bit boundary.

3.3. BW sub TLV: List and Range format

The figure below shows format of Type=2 sub-TLV for encoding slice information in list/range format. This sub-TLV must be repeated for each priority that is supported on the Te-Link.

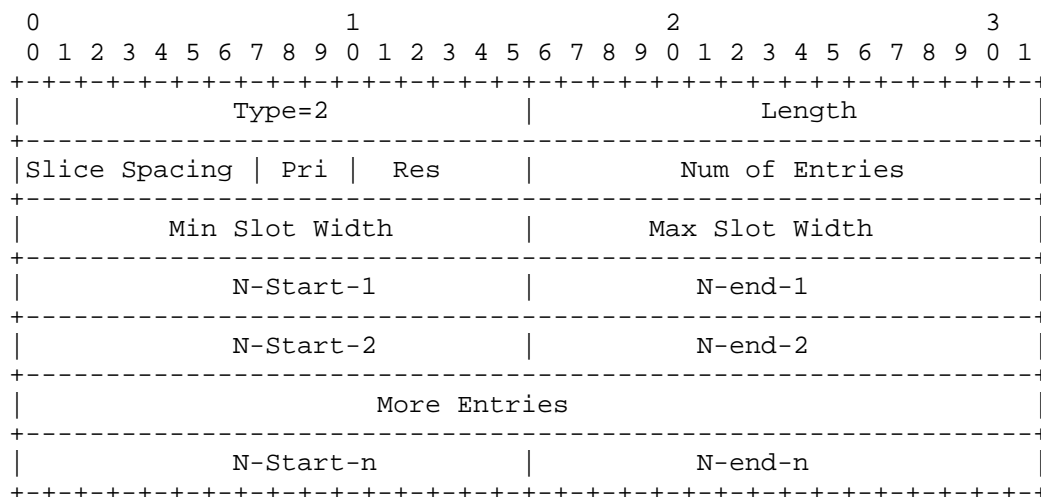


Figure 4: Type=2 BW sub TLV in List/Range format

3.3.1. Meaning of sub TLV fields

- o The meaning of above fields is same as in Type=1 BW-sub-TLV. For details refer to section 3.2.1.
 - o Slice Spacing,
 - o Priority,
 - o Maximum Slot Width &
 - o Minimum Slot Width
- o Number of Entries: 16-bit field
 - o Is a positive integer value.
 - o Total number of N-start & N-End rows advertised for the link.
- o N-Start-x: 16-bit field
 - o Is a two's complement integer value (+ve, -ve or zero) to specify start of the grid.

- o Use center freq formula to determine start of spectrum
- o N-end-x: 16-bit field
 - o Is a two's complement integer value (+ve, -ve or zero) to specify end of the list/range.
 - o Use center freq formula to determine end of spectrum

3.4. BW advertisement procedure

This section describes bandwidth advertisement for Te-Links capable SCSC.

- o Optical nodes capable of Super Channel Switching advertise slices of certain width available based on the frequency spectrum supported by the node (e.g. C band, extended C-band). For example, node(s) supporting extended C-band will advertise 384 slices.
- o The BW advertisement involves an ISCD containing
 - o Slice information in bit-map format (Type=1 BW-sub-TLV) where each bit corresponds to a single slice of width as identified by S.S field. OR
 - o Slice information in list/range format (Type=2 BW-sub-TLV) where each 32-bit entry represents an individual slice or list or range.
- o The slice position/numbering in Type=1 sub-TLV is identified based on N-start field. The N-start field is derived based on ITU center frequency formula.
- o The advertising node MUST also set Number of Slices field.
- o Minimum & Maximum slot width fields are included to allow for any restrictions on the link for carrying super channel LSPs.
- o The BW advertisement is priority based and up to 8 priority levels are allowed.
- o The node capable of supporting one or more priorities MUST set the priority field and include BW-sub TLV for each of the priority supported.

4. Examples

4.1. Example: BW advertisement without any service present

Figure 5 shows an example of BW sub-TLV for a te-link which has no service established over it yet. Attributes of BW sub-TLV in the te-link are:

- o N-start=-142 for extended C-band (2's complement should be included in this field)
- o Total number of slices available on the link = 384 (based on Slice spacing = 12.5GHz)
- o Min SW field shows min consumption of 4 Slices per LSP (=50GHz)
- o Max SW field shows up to 400GHz BW allowed per LSP (32x12.5GHz)
- o 48 bytes showing that all 384 slices are available.

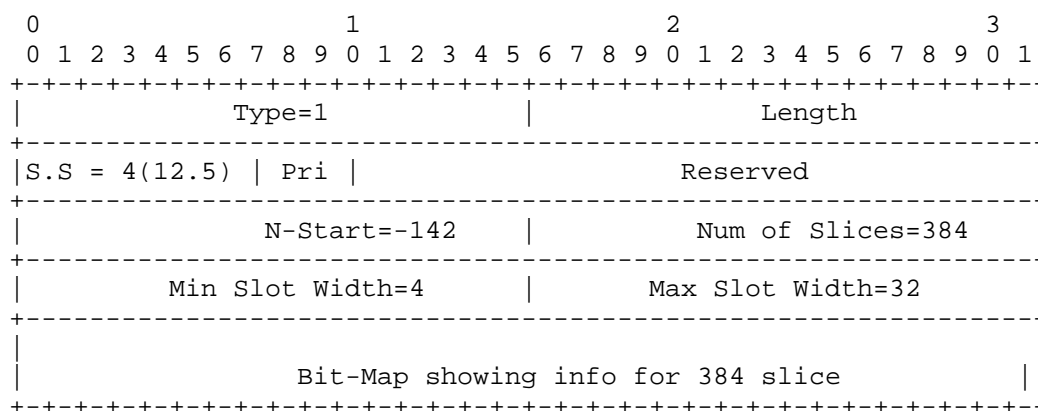


Figure 5: Type=1 BW sub-TLV without any service present

4.2. Example: How to use advertise Bandwidth

Assume user wants to setup Super Channel LSP over a single Flex-Grid link with BW requirement = 200GHz and transponder fully tunable.

- o The path computing node performs the following:
 - o Determine the number of slices required for the LSP ($200/S.S = 16$)
 - o Look for contiguous spectrum availability on each link from BW advertisement (both dir)
 - o Look for 16 contiguous bits in the BW advertisement TLV
 - o If available select the link for LSP creation.
 - o Signal for LSP creation. Once LSP is created, update BW available via new advertisement using the same Bandwidth sub-TLV.

5. Security Considerations

<Add any security considerations>

6. IANA Considerations

IANA needs to assign a new Grid field value to represent ITU-T Flex-Grid.

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.

7.2. Informative References

- [1] Berger, L., Ed., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, January 2003
- [2] Iftekhhar H, Abinder , Zhong , Marco , ''Generalized Label for Super-Channel Assignment on Flexible Grid'', draft-hussain-ccamp-super-channel-label-04.txt, July 2011.
- [3] K. Kompella, Y., " OSPF Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4203, Oct 2005
- [4] Lee, Y., Ed., "Framework for GMPLS and Path Computation Element (PCE) Control of Wavelength Switched Optical Networks (WSOs)", RFC 6163, April 2011
- [5] M. Jinno et. al., ''Spectrum-Efficient and Scalable Elastic Optical Path Network: Architecture, Benefits and Enabling Technologies'', IEEE Comm. Mag., Nov. 2009, pp. 66-73.
- [6] S. Chandrasekhar and X. Liu, ''Terabit Super-Channels for High Spectral Efficiency Transmission ''',in Proc. ECOC 2010, paper Tu.3.C.5, Torino (Italy), September 2010.
- [7] ITU-T Recommendation G.694.1, "Spectral grids for WDM applications: DWDM frequency grid", June 2002

- [8] Oscar G, et al., ''Framework and Requirements for GMPLS based control of Flexi-grid DWDM networks'', draft-ietf-ccamp-flexi-grid-fwk-00, work in progress.
- [9] G. Bernstein, Y. Lee, D. Li, W. Imajuku, " General Network Element Constraint Encoding for GMPLS Controlled Networks", work in progress: draft-ietf-ccamp-general-constraint-encode-05, May 2011
- [10] [FLEX-GRID] "ITU-T Recommendation G.694.1, Spectral grids for WDM applications: DWDM frequency grid", November 2012.

8. Acknowledgments

The authors would like to thank Khuzema Pithewan, Ashok Kunjidhapatham & Mohit Misra for their valuable comments.

Authors' Addresses

Abinder Dhillon
Fujitsu
Richardson, TX
Email: Abinder.Dhillon@us.fujitsu.com

Iftekhar Hussain
Infinera
140 Caspian Ct., Sunnyvale, CA 94089
Email: ihussain@infinera.com

Rajan Rao
Infinera
140 Caspian Ct., Sunnyvale, CA 94089
Email: rrao@infinera.com

Marco Sosa
Infinera
140 Caspian Ct., Sunnyvale, CA 94089
Email: msosa@infinera.com

Contributor's Addresses

Biao Lu
Email: blu@infinera.com

Subhendu Chattopadhyay
Email: schattopadhyay@infinera.com

Harpreet Uppal
Email: harpreet.uppal@infinera.com
Infinera
140 Caspian Ct., Sunnyvale, CA 94089

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 21, 2014

A. Farrel
J. Drake
Juniper Networks

N. Bitar
Verizon Networks

G. Swallow
Cisco Systems, Inc.

D. Ceccarelli
Ericsson
October 21, 2013

Problem Statement and Architecture for Information Exchange
Between Interconnected Traffic Engineered Networks

draft-farrel-interconnected-te-info-exchange-02.txt

Abstract

In Traffic Engineered (TE) systems, it is sometimes desirable to establish an end-to-end TE path with a set of constraints (such as bandwidth) across one or more network from a source to a destination. TE information is the data relating to nodes and TE links that is used in the process of selecting a TE path. The availability of TE information is usually limited to within a network (such as an IGP area) often referred to as a domain.

In order to determine the potential to establish a TE path through a series of connected networks, it is necessary to have available a certain amount of TE information about each network. This need not be the full set of TE information available within each network, but does need to express the potential of providing TE connectivity. This subset of TE information is called TE reachability information.

This document sets out the problem statement and architecture for the exchange of TE information between interconnected TE networks in support of end-to-end TE path establishment. For reasons that are explained in the document, this work is limited to simple TE constraints and information that determine TE reachability.

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."

Copyright Notice

Copyright (c) 2013 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	5
1.1. What is TE Reachability?	6
2. Overview of Use Cases	6
2.1. Peer Networks	6
2.1.1. Where is the Destination?	7
2.2. Client-Server Networks	8
2.3. Dual-Homing	10
3. Problem Statement	11
3.1. Use of Existing Protocol Mechanisms	12
3.2. Policy and Filters	12
3.3. Confidentiality	13
3.4. Information Overload	13
3.5. Issues of Information Churn	14
3.6. Issues of Aggregation	15
3.7. Virtual Network Topology	15
4. Existing Work	17
4.1. Per-Domain Path Computation	17
4.2. Crankback	18
4.3. Path Computation Element	18
4.4. GMPLS UNI and Overlay Networks	20
4.5. Layer One VPN	20
4.6. VNT Manager and Link Advertisement	21
4.7. What Else is Needed and Why?	22
5. Architectural Concepts	22
5.1. Basic Components	22
5.1.1. Peer Interconnection	22
5.1.2. Client-Server Interconnection	23
5.2. TE Reachability	24
5.3. Abstraction not Aggregation	25
5.3.1. Abstract Links	25
5.3.2. The Abstraction Layer Network	26
5.3.3. Abstraction in Client-Server Networks	27
5.3.4. Abstraction in Peer Networks	32
5.4. Considerations for Dynamic Abstraction	34
5.5. Requirements for Advertising Abstracted Links and Nodes	34
6. Building on Existing Protocols	34
6.1. BGP-LS	34
6.2. IGPs	34
6.3. RSVP-TE	35
7. Applicability to Optical Domains and Networks	35
8. Abstraction in L3VPN Multi-AS Environments	39
9. Scoping Future Work	39
9.1. Not Solving the Internet	39
9.2. Working With "Related" Domains	39
9.3. Not Breaking Existing Protocols	39
9.4. Sanity and Scaling	39

10. Manageability Considerations	40
11. IANA Considerations	40
12. Security Considerations	40
13. Acknowledgements	40
14. References	40
14.1. Informative References	40
Authors' Addresses	44
Appendix A. Editor's Notes	44

1. Introduction

Traffic Engineered (TE) systems such as MPLS-TE [RFC2702] and GMPLS [RFC3945] offer a way to establish paths through a network in a controlled way that reserves network resources on specified links. TE paths are computed by examining the Traffic Engineering Database (TED) and selecting a sequence of links and nodes that are capable of meeting the requirements of the path to be established. The TED is constructed from information distributed by the IGP running in the network, for example OSPF-TE [RFC3630] or ISIS-TE [RFC5305].

It is sometimes desirable to establish an end-to-end TE path that crosses more than one network or administrative domain as described in [RFC4105] and [RFC4216]. In these cases, the availability of TE information is usually limited to within each network. Such networks are often referred to as Domains [RFC4726] and we adopt that definition in this document: viz.

For the purposes of this document, a domain is considered to be any collection of network elements within a common sphere of address management or path computational responsibility. Examples of such domains include IGP areas and Autonomous Systems.

In order to determine the potential to establish a TE path through a series of connected domains and to choose the appropriate domain connection points through which to route a path, it is necessary to have available a certain amount of TE information about each domain. This need not be the full set of TE information available within each domain, but does need to express the potential of providing TE connectivity. This subset of TE information is called TE reachability information. The TE reachability information can be exchanged between domains based on the information gathered from the local routing protocol, filtered by configured policy, or statically configured.

This document sets out the problem statement and architecture for the exchange of TE information between interconnected TE domains in support of end-to-end TE path establishment. The scope of this document is limited to the simple TE constraints and information (TE metrics, hop count, bandwidth, delay, shared risk) necessary to determine TE reachability: discussion of multiple additional constraints that might qualify the reachability can significantly complicate aggregation of information and the stability of the mechanism used to present potential connectivity as is explained in the body of this document.

1.1. What is TE Reachability?

In an IP network, reachability is the ability to deliver a packet to a specific address or prefix. That is, the existence of an IP path to that address or prefix. TE reachability is the ability to reach a specific address along a TE path.

TE reachability may be unqualified (there is a TE path, but no information about available resources or other constraints is supplied) which is helpful especially in determining a path to a destination that lies in an unknown domain, or may be qualified by TE attributes such as TE metrics, hop count, available bandwidth, delay, shared risk, etc.

2. Overview of Use Cases

2.1. Peer Networks

The peer network use case can be most simply illustrated by the example in Figure 1. A TE path is required between the source (Src) and destination (Dst), that are located in different domains. There are two points of interconnection between the domains, and selecting the wrong point of interconnection can lead to a sub-optimal path, or even fail to make a path available.

For example, when Domain A attempts to select a path, it may determine that adequate bandwidth is available on from Src through both interconnection points x1 and x2. It may pick the path through x1 for local policy reasons: perhaps the TE metric is smaller. However, if there is no connectivity in Domain Z from x1 to Dst, the path cannot be established. Techniques such as crankback (see Section 4.2) may be used to alleviate this situation, but do not lead to rapid setup or guaranteed optimality. Furthermore RSVP signalling creates state in the network that is immediately removed by the crankback procedure. Frequent events of such a kind impact scalability in a non-deterministic manner.

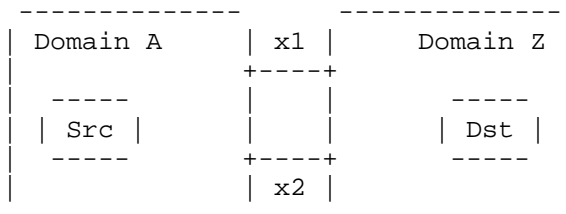


Figure 1 : Peer Networks

There are countless more complicated examples of the problem of peer networks. Figure 2 shows the case where there is a simple mesh of domains. Clearly, to find a TE path from Src to Dst, Domain A must not select a path leaving through interconnect x1 since Domain B has no connectivity to Domain Z. Furthermore, in deciding whether to select interconnection x2 (through Domain C) or interconnection x3 through Domain D, Domain A must be sensitive to the TE connectivity available through each of Domains C and D, as well the TE connectivity from each of interconnections x4 and x5 to Dst within Domain Z.

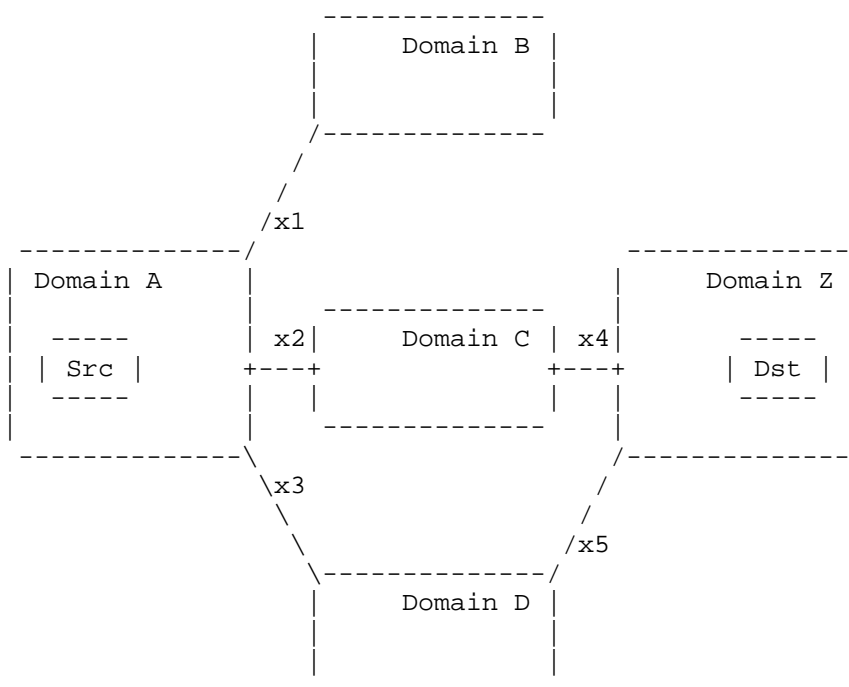


Figure 2 : Peer Networks in a Mesh

Of course, many network interconnection scenarios are going to be a combination of the situations expressed in these two examples. There may be a mesh of domains, and the domains may have multiple points of interconnection.

2.1.1. Where is the Destination?

A variation of the problems expressed in Section 2.1 arises when the source domain (Domain A in both figures) does not know where the

destination is located. That is, when the domain in which the destination node is located is not known to the source domain.

This is most easily seen in consideration of Figure 2 where the decision about which interconnection to select needs to be based on building a path toward the destination domain. Yet this can only be achieved if it is known in which domain the destination node lies, or at least if there is some indication in which direction the destination lies. This function is obviously provided in IP networks by inter-domain routing [RFC4271].

2.2. Client-Server Networks

Two specific use cases relate to the client-server relationship between networks. These use cases have sometimes been referred to as overlay networks.

The first case, shown in Figure 3, occurs when domains belonging to one network are connected by a domain belonging to another network. In this scenario, once connections (or tunnels) are formed across the lower layer network, the domains of the upper layer network can be merged into a single domain by running IGP adjacencies over the tunnels, and treating the tunnels as links in the higher layer network. The TE relationship between the domains (higher and lower layer) in this case is reduced to determining which tunnels to set up, how to trigger them, how to route them, and what capacity to assign them. As the demands in the higher layer network vary, these tunnels may need to be modified.

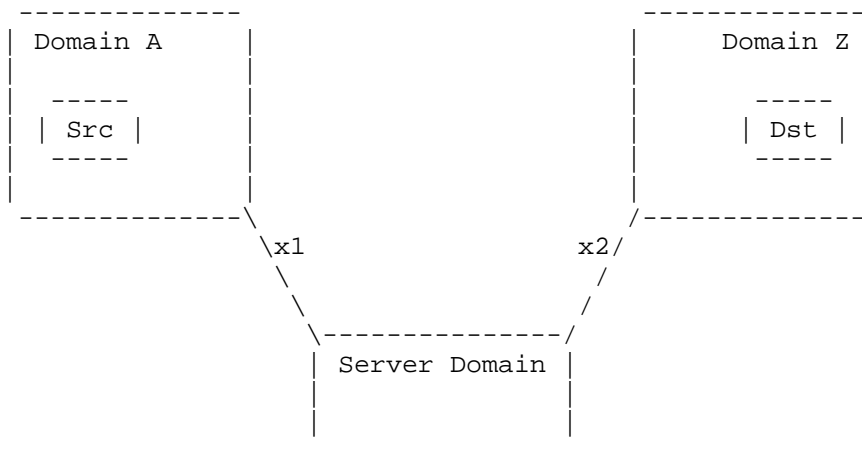


Figure 3 : Client-Server Networks

The second use case relating to client-server networking is for Virtual Private Networks (VPNs). In this case, as opposed to the former one, it is assumed that the client network has a different address space than that of the server layer where non-overlapping IP addresses between the client and the server networks cannot be guaranteed. A simple example is shown in Figure 4. The VPN sites comprise a set of domains that are interconnected over a core domain, the provider network.

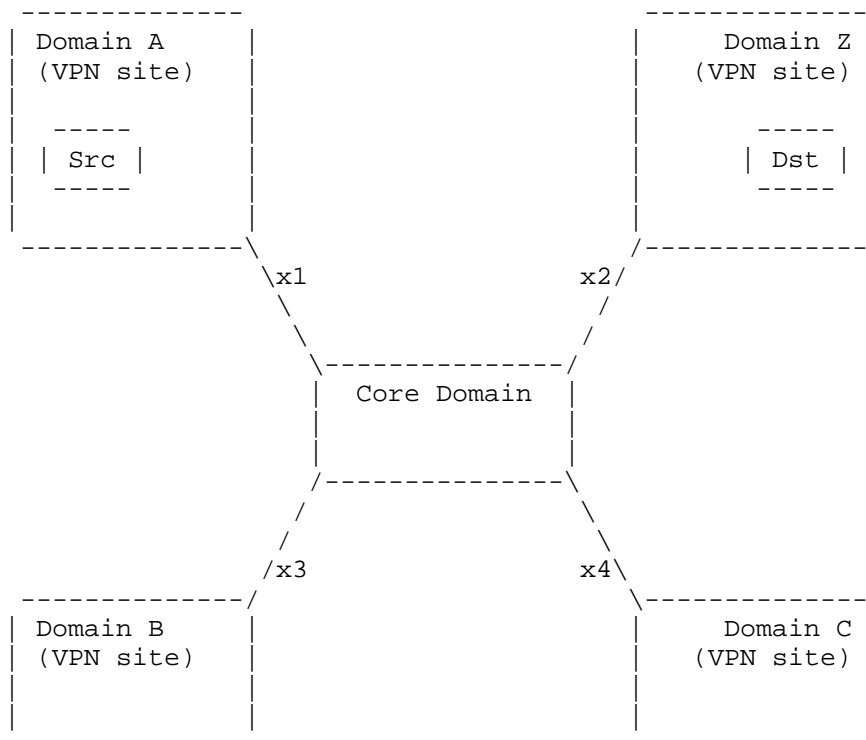


Figure 4 : A Virtual Private Network

Note that in the use cases shown in Figures 3 and 4 the client layer domains may (and, in fact, probably do) operate as a single connected network.

Both use cases in this section become "more interesting" when combined with the use case in Section 2.1. That is, when the connectivity between higher layer domains or VPN sites is provided by a sequence or mesh of lower layer domains. Figure 5 shows how this might look in the case of a VPN.

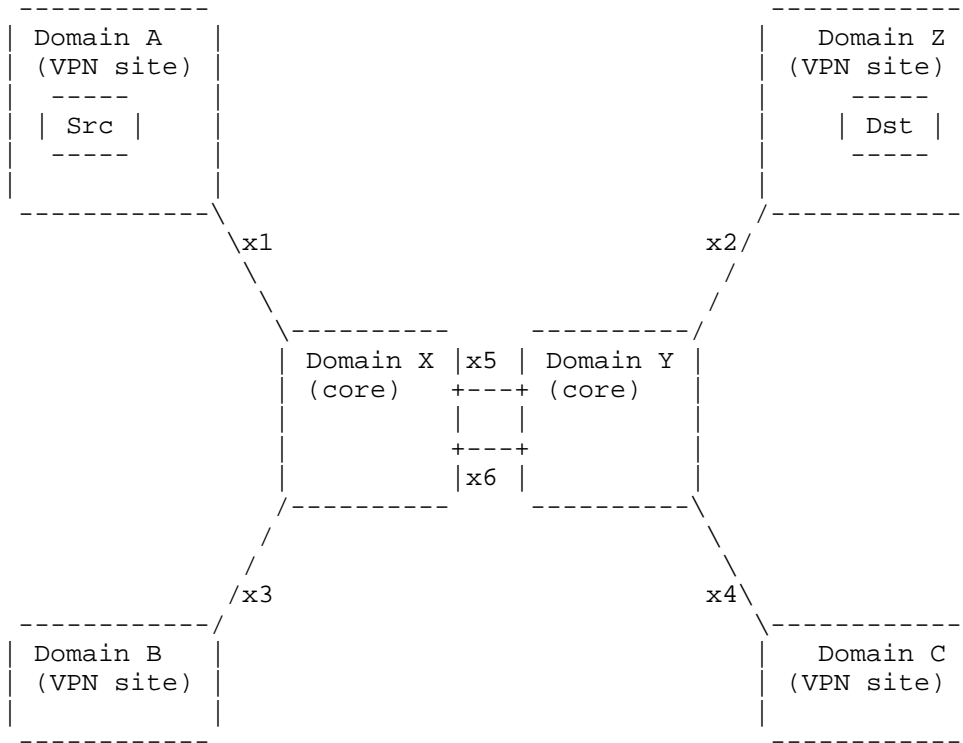


Figure 5 : A VPN Supported Over Multiple Server Domains

2.3. Dual-Homing

A further complication may be added to the client-server relationship described in Section 2.2 by considering what happens when a client domain is attached to more than one server domain, or has two points of attachment to a server domain. Figure 6 shows an example of this for a VPN.

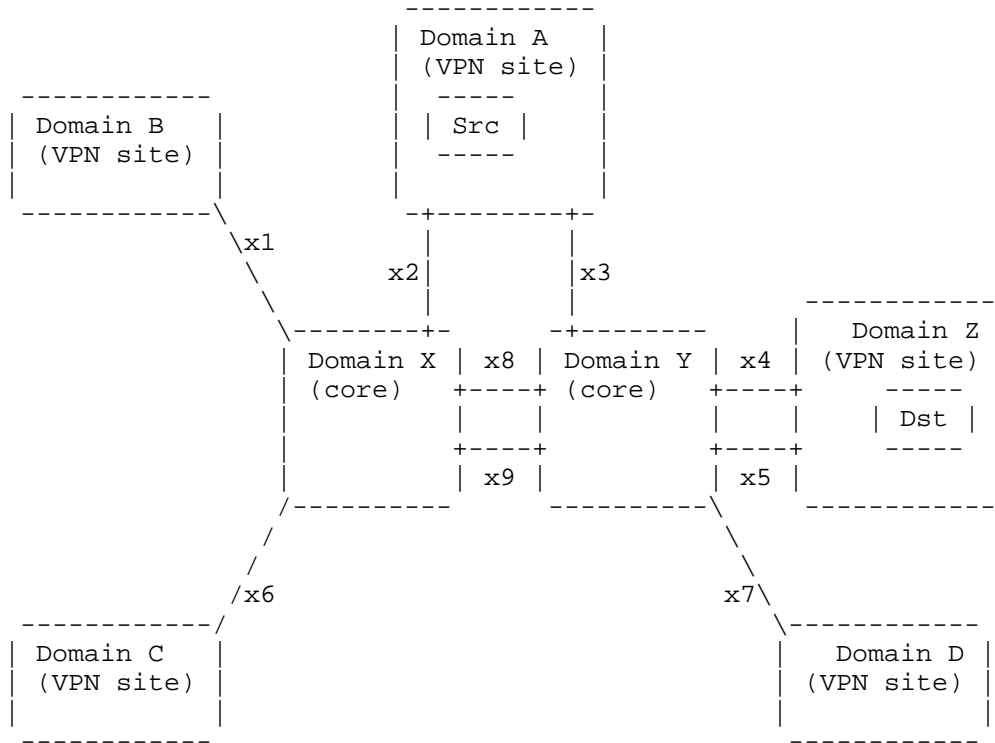


Figure 6 : Dual-Homing in a Virtual Private Network

3. Problem Statement

The problem statement presented in this section is as much about the issues that may arise in any solution (and so have to be avoided) and the features that are desirable within a solution, as it is about the actual problem to be solved.

The problem can be stated very simply and with reference to the use cases presented in the previous section.

A mechanism is required that allows TE-path computation in one domain to make informed choices about the TE-capabilities and exit point from the domain when signaling an end-to-end TE path that will extend across multiple domains.

Thus, the problem is one of information collection and presentation, not about signaling. Indeed, the existing signaling mechanisms for

TE LSP establishment are likely to prove adequate [RFC4726] with the possibility of minor extensions.

An interesting annex to the problem is how the path is made available for use. For example, in the case of a client-server network, the path established in the server network needs to be made available as a TE link to provide connectivity in the client network.

3.1. Use of Existing Protocol Mechanisms

TE information may currently be distributed in a domain by TE extensions to one of the two IGPs as described in OSPF-TE [RFC3630] and ISIS-TE [RFC5305]. TE information may be exported from a domain (for example, northbound) using link state extensions to BGP [I-D.ietf-idr-ls-distribution].

It is desirable that a solution to the problem described in this document does not require the implementation of a new, network-wide protocol. Instead, it would be advantageous to make use of an existing protocol that is commonly implemented on routers and is currently deployed, or to use existing computational elements such as Path Computation Elements (PCEs). This has many benefits in network stability, time to deployment, and operator training.

It is recognized, however, that existing protocols are unlikely to be immediately suitable to this problem space without some protocol extensions. Extending protocols must be done with care and with consideration for the stability of existing deployments. In extreme cases, a new protocol can be preferable to a messy hack of an existing protocol.

3.2. Policy and Filters

A solution must be amenable to the application of policy and filters. That is, the operator of a domain that is sharing information with another domain must be able to apply controls to what information is shared. Furthermore, the operator of a domain that has information shared with it must be able to apply policies and filters to the received information.

Additionally, the path computation within a domain must be able to weight the information received from other domains according to local policy such that the resultant computed path meets the local operator's needs and policies rather than those of the operators of other domains.

3.3. Confidentiality

A feature of the policy described in Section 3.3 is that an operator of a domain may desire to keep confidential the details about its internal network topology and loading. This information could be construed as commercially sensitive.

Although it is possible that TE information exchange will take place only between parties that have significant trust, there are also use cases (such as the VPN supported over multiple server domains described in Section 2.4) where information will be shared between domains that have a commercial relationship, but a low level of trust.

Thus, it must be possible for a domain to limit the information share to just that which the computing domain needs to know with the understanding that less information that is made available the more likely it is that the result will be a less optimal path and/or more crankback events.

3.4. Information Overload

One reason that networks are partitioned into separate domains is to reduce the set of information that any one router has to handle. This also applies to the volume of information that routing protocols have to distribute.

Over the years routers have become more sophisticated with greater processing capabilities and more storage, the control channels on which routing messages are exchanged have become higher capacity, and the routing protocols (and their implementations) have become more robust. Thus, some of the arguments in favor of dividing a network into domains may have been reduced. Conversely, however, the size of networks continues to grow dramatically with a consequent increase in the total amount of routing-related information available. Additionally, in this case, the problem space spans two or more networks.

Any solution to the problems voiced in this document must be aware of the issues of information overload. If the solution was to simply share all TE information between all domains in the network, the effect from the point of view of the information load would be to create one single flat network domain. Thus the solution must deliver enough information to make the computation practical (i.e., to solve the problem), but not so much as to overload the receiving domain. Furthermore, the solution cannot simply rely on the policies and filters described in Section 3.2 because such filters might not always be enabled.

3.5. Issues of Information Churn

As LSPs are set up and torn down, the available TE resources on links in the network change. In order to reliably compute a TE path through a network, the computation point must have an up-to-date view of the available TE resources. However, collecting this information may result in considerable load on the distribution protocol and churn in the stored information. In order to deal with this problem even in a single domain, updates are sent at periodic intervals or whenever there is a significant change in resources, whichever happens first.

Consider, for example, that a TE LSP may traverse ten links in a network. When the LSP is set up or torn down, the resources available on each link will change resulting in a new advertisement of the link's capabilities and capacity. If the arrival rate of new LSPs is relatively fast, and the hold times relatively short, the network may be in a constant state of flux. Note that the problem here is not limited to churn within a single domain, since the information shared between domains will also be changing. Furthermore, the information that one domain needs to share with another may change as the result of LSPs that are contained within or cross the first domain but which are of no direct relevance to the domain receiving the TE information.

In packet networks, where the capacity of an LSP is often a small fraction of the resources available on any link, this issue is partially addressed by the advertising routers. They can apply a threshold so that they do not bother to update the advertisement of available resources on a link if the change is less than a configured percentage of the total (or alternatively, the remaining) resources. The updated information in that case will be disseminated based on an update interval rather than a resource change event.

In non-packet networks, where link resources are physical switching resources (such as timeslots or wavelengths) the capacity of an LSP may more frequently be a significant percentage of the available link resources. Furthermore, in some switching environments, it is necessary to achieve end-to-end resource continuity (such as using the same wavelength on the whole length of an LSP), so it is far more desirable to keep the TE information held at the computation points up-to-date. Fortunately, non-packet networks tend to be quite a bit smaller than packet networks, the arrival rates of non-packet LSPs are much lower, and the hold times considerably longer. Thus the information churn may be sustainable.

3.6. Issues of Aggregation

One possible solution to the issues raised in other sub-sections of this section is to aggregate the TE information shared between domains. Two aggregation mechanisms are often considered:

- Virtual node model. In this view, the domain is aggregated as if it was a single node (or router / switch). Its links to other domains are presented as real TE links, but the model assumes that any LSP entering the virtual node through a link can be routed to leave the virtual node through any other link.
- Virtual link model. In this model, the domain is reduced to a set of edge-to-edge TE links. Thus, when computing a path for an LSP that crosses the domain, a computation point can see which domain entry points can be connected to which other and with what TE attributes.

It is of the nature of aggregation that information is removed from the system. This can cause inaccuracies and failed path computation. For example, in the virtual node model there might not actually be a TE path available between a pair of domain entry points, but the model lacks the sophistication to represent this "limited cross-connect capability" within the virtual node. On the other hand, in the virtual link model it may prove very hard to aggregate multiple link characteristics: for example, there may be one path available with high bandwidth, and another with low delay, but this does not mean that the connectivity should be assumed or advertised as having both high bandwidth and low delay.

The trick to this multidimensional problem, therefore, is to aggregate in a way that retains as much useful information as possible while removing the data that is not needed. An important part of this trick is a clear understanding of what information is actually needed.

It should also be noted in the context of Section 3.5 that changes in the information within a domain may have a bearing on what aggregated data is shared with another domain. Thus, while the data shared is reduced, the aggregation algorithm (operating on the routers responsible for sharing information) may be heavily exercised.

3.7. Virtual Network Topology

The terms "virtual topology" and "virtual network topology" have become overloaded in a relatively short time. We draw on [RFC5212] and [RFC5623] for inspiration to provide a definition for use in this document. Our definition is based on the fact that a topology at the

client network layer is constructed of nodes and links. Typically, the nodes are routers in the client layer, and the links are data links. However, a layered network provides connectivity through the lower layer as LSPs, and these LSPs can provide links in the client layer. Furthermore, those LSPs may have been established in advance, or might be LSPs that could be set up if required. This leads to the definition:

A Virtual Network Topology (VNT) is made up of links in a network layer. Those links may be realized as direct data links or as multi-hop connections (LSPs) in a lower network layer. Those underlying LSPs may be established in advance or created on demand.

The creation and management of a VNT requires interaction with management and policy. Activity is needed in both the client and server layer:

- In the server layer, LSPs need to be set up either in advance in response to management instructions or in answer to dynamic requests subject to policy considerations.
- In the server layer, evaluation of available TE resources can lead to the announcement of potential connectivity (i.e., LSPs that could be set up on demand).
- In the client layer, connectivity (lower layer LSPs or potential LSPs) needs to be announced in the IGP as a normal TE link. Such links may or may not be made available to IP routing: but, they are never made available to IP until fully instantiated.
- In the client layer, requests to establish lower layer LSPs need to be made either when links supported by potential LSPs are about to be used (i.e., when a higher layer LSP is signalled to cross the link, the setup of the lower layer LSP is triggered), or when the client layer determines it needs more connectivity or capacity.

It is a fundamental of the use of a VNT that there is a policy point at the point of instantiation of a lower-layer LSP. At the moment that the setup of a lower-layer LSP is triggered, whether from a client-layer management tool or from signaling in the client layer, the server layer must be able to apply policy to determine whether to actually set up the LSP. Thus, fears that a micro-flow in the client layer might cause the activation of 100G optical resources in the server layer can be completely controlled by the policy of the server layer network's operator (and could even be subject to commercial terms).

These activities require an architecture and protocol elements as

well as management components and policy elements.

4. Existing Work

This section briefly summarizes relevant existing work that is used to route TE paths across multiple domains.

4.1. Per-Domain Path Computation

The per-domain mechanism of path establishment is described in [RFC5152] and its applicability is discussed in [RFC4726]. In summary, this mechanism assumes that each domain entry point is responsible for computing the path across the domain, but that details of the path in the next domain are left to the next domain entry point. The computation may be performed directly by the entry point or may be delegated to a computation server.

This basic mode of operation can run into many of the issues described alongside the use cases in Section 2. However, in practice it can be used effectively with a little operational guidance.

For example, RSVP-TE [RFC3209] includes the concept of a "loose hop" in the explicit path that is signaled. This allows the original request for an LSP to list the domains or even domain entry points to include on the path. Thus, in the example in Figure 1, the source can be told to use the interconnection x2. Then the source computes the path from itself to x2, and initiates the signaling. When the signaling message reaches Domain Z, the entry point to the domain computes the remaining path to the destination and continues the signaling.

Another alternative suggested in [RFC5152] is to make TE routing attempt to follow inter-domain IP routing. Thus, in the example shown in Figure 2, the source would examine the BGP routing information to determine the correct interconnection point for forwarding IP packets, and would use that to compute and then signal a path for Domain A. Each domain in turn would apply the same approach so that the path is progressively computed and signaled domain by domain.

Although the per-domain approach has many issues and drawbacks in terms of achieving optimal (or, indeed, any) paths, it has been the mainstay of inter-domain LSP set-up to date.

4.2. Crankback

Crankback addresses one of the main issues with per-domain path computation: what happens when an initial path is selected that cannot be completed toward the destination? For example, what happens if, in Figure 2, the source attempts to route the path through interconnection x2, but Domain C does not have the right TE resources or connectivity to route the path further?

Crankback for MPLS-TE and GMPLS networks is described in [RFC4920] and is based on a concept similar to the Acceptable Label Set mechanism described for GMPLS signaling in [RFC3473]. When a node (i.e., a domain entry point) is unable to compute a path further across the domain, it returns an error message in the signaling protocol that states where the blockage occurred (link identifier, node identifier, domain identifier, etc.) and gives some clues about what caused the blockage (bad choice of label, insufficient bandwidth available, etc.). This information allows a previous computation point to select an alternative path, or to aggregate crankback information and return it upstream to a previous computation point.

Crankback is a very powerful mechanism and can be used to find an end-to-end in a multi-domain network if one exists.

On the other hand, crankback can be quite resource-intensive as signaling messages and path setup attempts may "wander around" in the network attempting to find the correct path for a long time. Since RSVP-TE signaling ties up networks resources for partially established LSPs, since network conditions may be in flux, and most particularly since LSP setup within well-known time limits is highly desirable, crankback is not a popular mechanism.

Furthermore, even if crankback can always find an end-to-end path, it does not guarantee to find the optimal path. (Note that there have been some academic proposals to use signaling-like techniques to explore the whole network in order to find optimal paths, but these tend to place even greater burdens on network processing.)

4.3. Path Computation Element

The Path Computation Element (PCE) is introduced in [RFC4655]. It is an abstract functional entity that computes paths. Thus, in the example of per-domain path computation (Section 4.1) the source node and each domain entry point is a PCE. On the other hand, the PCE can also be realized as a separate network element (a server) to which computation requests can be sent using the Path Computation Element Communication Protocol (PCEP) [RFC5440].

Each PCE has responsibility for computations within a domain, and has visibility of the attributes within that domain. This immediately enables per-domain path computation with the opportunity to off-load complex, CPU-intensive, or memory-intensive computation functions from routers in the network. But the use of PCE in this way does not solve any of the problems articulated in Sections 4.1 and 4.2.

Two significant mechanisms for cooperation between PCEs have been described. These mechanisms are intended to specifically address the problems of computing optimal end-to-end paths in multi-domain environments.

- The Backward-Recursive PCE-Based Computation (BRPC) mechanism [RFC5441] involves cooperation between the set of PCEs along the inter-domain path. Each one computes the possible paths from domain entry point (or source node) to domain exit point (or destination node) and shares the information with its upstream neighbor PCE which is able to build a tree of possible paths rooted at the destination. The PCE in the source domain can select the optimal path.

BRPC is sometimes described as "crankback at computation time". It is capable of determining the optimal path in a multi-domain network, but depends on knowing the domain that contains the destination node. Furthermore, the mechanism can become quite complicated and involve a lot of data in a mesh of interconnected domains. Thus, BRPC is most often proposed for a simple mesh of domains and specifically for a path that will cross a known sequence of domains, but where there may be a choice of domain interconnections. In this way, BRPC would only be applied to Figure 2 if a decision had been made (externally) to traverse Domain C rather than Domain D (notwithstanding that it could functionally be used to make that choice itself), but BRPC could be used very effectively to select between interconnections x1 and x2 in Figure 1.

- Hierarchical PCE (H-PCE) [RFC6805] offers a parent PCE that is responsible for navigating a path across the domain mesh and for coordinating intra-domain computations by the child PCEs responsible for each PCE. This approach makes computing an end-to-end path across a mesh of domains far more tractable. However, it still leaves unanswered the issue of determining the location of the destination (i.e., discovering the destination domain) as described in Section 2.1.1. Furthermore, it raises the question of who operates the parent PCE especially in networks where the domains are under different administrative and commercial control.

Further issues and considerations of the use of PCE can be found in

[I-D.farrkingel-pce-questions].

4.4. GMPLS UNI and Overlay Networks

[RFC4208] defines the GMPLS User-to-Network Interface (UNI) to present a routing boundary between an overlay network and the core network, i.e. the client-server interface. In the client network, the nodes connected directly to the core network are known as edge nodes, while the nodes in the server network are called core nodes.

In the overlay model defined by [RFC4208] the core nodes act as a closed system and the edge nodes do not participate in the routing protocol instance that runs among the core nodes. Thus the UNI allows access to and limited control of the core nodes by edge nodes that are unaware of the topology of the core nodes.

[RFC4208] does not define any routing protocol extension for the interaction between core and edge nodes but allows for the exchange of reachability information between them. In terms of a VPN, the client network can be considered as the customer network comprised of a number of disjoint sites, and the edge nodes match the VPN CE nodes. Similarly, the provider network in the VPN model is equivalent to the server network.

[RFC4208] is, therefore, a signaling-only solution that allows edge nodes to request connectivity cross the core network, and leaves the core network to select the paths and set up the core LSPs. This solution is supplemented by a number of signaling extensions such as [RFC5553], [I-D.ietf-ccamp-xro-lsp-subobject], and [I-D.ietf-ccamp-te-metric-recording] to give the edge node more control over the LSP that the core network will set up by exchanging information about core LSPs that have been established and by allowing the edge nodes to supply additional constraints on the core LSPs that are to be set up.

Nevertheless, in this UNI/overlay model, the edge node has limited information of precisely what LSPs could be set up across the core, and what TE services (such as diverse routes for end-to-end protection, end-to-end bandwidth, etc.) can be supported.

4.5. Layer One VPN

A Layer One VPN (L1VPN) is a service offered by a core layer 1 network to provide layer 1 connectivity (TDM, LSC) between two or more customer networks in an overlay service model [RFC4847].

As in the UNI case, the customer edge has some control over the establishment and type of the connectivity. In the L1VPN context

three different service models have been defined classified by the semantics of information exchanged over the customer interface: Management Based, Signaling Based (a.k.a. basic), and Signaling and Routing service model (a.k.a. enhanced).

In the management based model, all edge-to-edge connections are set up using configuration and management tools. This is not a dynamic control plane solution and need not concern us here.

In the signaling based service model [RFC5251] the CE-PE interface allows only for signaling message exchange, and the provider network does not export any routing information about the core network. VPN membership is known a priori (presumably through configuration) or is discovered using a routing protocol [RFC5195], [RFC5252], [RFC5523], as is the relationship between CE nodes and ports on the PE. This service model is much in line with GMPLS UNI as defined in [RFC4208].

In the enhanced model there is an additional limited exchange of routing information over the CE-PE interface between the provider network and the customer network. The enhanced model considers four different types of service models, namely: Overlay Extension, Virtual Node, Virtual Link and Per-VPN service models. All of these represent particular cases of the TE information aggregation and representation.

4.6. VNT Manager and Link Advertisement

As discussed in Section 3.7, operation of a VNT requires policy and management input. In order to handle this, [RFC5623] introduces the concept of the Virtual Network Topology Manager. This is a functional component that applies policy to requests from client networks (or agents of the client network, such as a PCE) for the establishment of LSPs in the server network to provide connectivity in the client network.

The VNT Manager would, in fact, form part of the provisioning path for all server network LSPs whether they are set up ahead of client network demand or triggered by end-to-end client network LSP signaling.

An important companion to this function is determining how the LSP set up across the server network is made available as a TE link in the client network. Obviously, if the LSP is established using management intervention, the subsequent client network TE link can also be configured manually. However, if the LSP is signaled dynamically there is need for the end points to exchange the link properties that they should advertise within the client network, and in the case of a server network that supports more than one client,

it will be necessary to indicate which client or clients can use the link. This capability is provided in [RFC6107].

Note that a potential server network LSP that is advertised as a TE link in the client network might to be determined dynamically by the edge nodes. In this case there will need to be some effort to ensure that both ends of the link have the same view of the available TE resources, or else the advertised link will be asymmetrical.

4.7. What Else is Needed and Why?

As can be seen from Sections 4.1 through 4.6, a lot of effort has focused on client-server networks as described in Figure 3. Far less consideration has been given to network peering or the combination of the two use cases.

Various work has been suggested to extend the definition of the UNI such that routing information can be passed across the interface. However, this approach seems to break the architectural concept of network separation that the UNI facilitates.

Other approaches are working toward a flattening of the network with complete visibility into the server networks being made available in the client network. These approaches, while functional, ignore the main reasons for introducing network separation in the first place.

The remainder of this document introduces a new approach based on network abstraction that allows a server network to use its own knowledge of its resources and topology combined with its own policies to determine what edge-to-edge connectivity capabilities it will inform the client networks about.

5. Architectural Concepts

5.1. Basic Components

This section revisits the use cases from Section 2 to present the basic architectural components that provide connectivity in the peer and client-server cases. These component models can then be used in later sections to enable discussion of a solution architecture.

5.1.1. Peer Interconnection

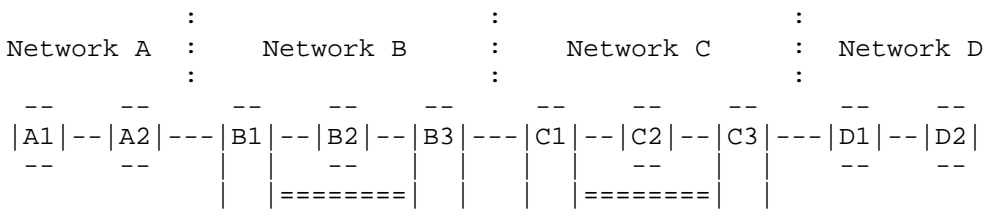
Figure 7 shows the basic architectural concepts for connecting across peer networks. Nodes from four networks are shown: A1 and A2 come from one network; B1, B2, and B3 from another network; etc. The interfaces between the networks (sometimes known as External Network-

to-Network Interfaces - ENNIs) are A2-B1, B3-C1, and C3-D1.

The objective is to be able to support an end-to-end connection A1-to-D2. This connection is for TE connectivity.

As shown in the figure, LSP tunnels that span the transit networks are used to achieve the required connectivity. These transit LSPs form the key building blocks of the end-to-end connectivity.

The transit tunnels can be used as hierarchical LSPs [RFC4206] to carry the end-to-end LSP, or can become stitching segments [RFC5150] of the end-to-end LSP. The transit tunnels B1-B3 and C-C3 can be as an abstract link as discussed in Section 5.3.



Key

```

1
--- Direct connection between two nodes

```

```
=== LSP tunnel across transit network
```

Figure 7 : Architecture for Peering

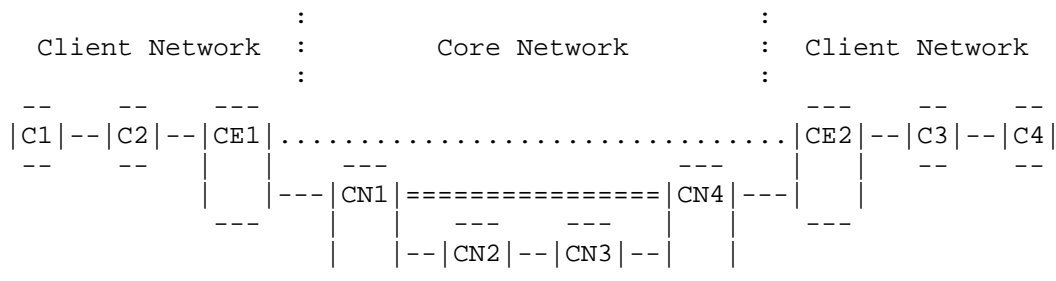
5.1.2. Client-Server Interconnection

Figure 8 shows the basic architectural concepts for a client-server network. The client network nodes are C1, C2, CE1, CE2, C3, and C4. The core network nodes are CN1, CN2, CN3, and CN4. The interfaces CE1-CN1 and CE2-CN2 are the interfaces between the client and core networks.

The objective is to be able to support an end-to-end connection, C1-to-C4, in the client network. This connection may support TE or normal IP forwarding. To achieve this, CE1 is to be connected to CE2 by a link in the client layer that is supported by a core network LSP.

As shown in the figure, two LSPs are used to achieve the required connectivity. One LSP is set up across the core from CN1 to CN2. This core LSP then supports a three-hop LSP from CE1 to CE2 with its middle hop being the core LSP. It is this LSP that is presented as a link in the client network.

The practicalities of how the CE1-CE2 LSP is carried across the core LSP may depend on the switching and signaling options available in the core network. The LSP may be tunneled down the core LSP using the mechanisms of a hierarchical LSP [RFC4206], or the LSP segments CE1-CN1 and CN2-CE2 may be stitched to the core LSP as described in [RFC5150].



Key
 --- Direct connection between two nodes
 ... CE-to-CE LSP tunnel
 === LSP tunnel across the core

Figure 8 : Architecture for Client-Server Network

5.2. TE Reachability

As described in Section 1.1, TE reachability is the ability to reach a specific address along a TE path. The knowledge of TE reachability enables an end-to-end TE path to be computed.

In a single network, TE reachability is derived from the Traffic Engineering Database (TED) that is the collection of all TE information about all TE links in the network. The TED is usually built from the data exchanged by the IGP, although it can be supplemented by configuration and inventory details especially in transport networks.

In multi-network scenarios, TE reachability information can be described as "You can get from node X to node Y with the following TE attributes." For transit cases, nodes X and Y will be edge nodes of the transit network, but it is also important to consider the information about the TE connectivity between an edge node and a specific destination node.

TE reachability may be unqualified (there is a TE path), or may be qualified by TE attributes such as TE metrics, hop count, available bandwidth, delay, shared risk, etc.

TE reachability information can be exchanged between networks so that nodes in one network can determine whether they can establish TE paths across or into another network. Such exchanges are subject to a range of policies imposed by the advertiser (for security and administrative control) and by the receiver (for scalability and stability).

5.3. Abstraction not Aggregation

Aggregation is the process of synthesizing from available information. Thus, the virtual node and virtual link models described in Section 3.6 rely on processing the information available within a network to produce the aggregate representations of links and nodes that are presented to the consumer. As described in Section 3, dynamic aggregation is subject to a number of pitfalls.

In order to distinguish the architecture described in this document from the previous work on aggregation, we use the term "abstraction" in this document. The process of abstraction is one of applying policy to the available TE information within a domain, to produce selective information that represents the potential ability to connect across the domain.

Abstraction does not offer all possible connectivity options (refer to Section 3.6), but does present a general view of potential connectivity. Abstraction may have a dynamic element, but is not intended to keep pace with the changes in TE attribute availability within the network.

Thus, when relying on an abstraction to compute an end-to-end path, the process might not deliver a usable path. That is, there is no actual guarantee that the abstractions are current or feasible.

While abstraction uses available TE information, it is subject to policy and management choices. Thus, not all potential connectivity will be advertised to each client. The filters may depend on commercial relationships, the risk of disclosing confidential information, and concerns about what use is made of the connectivity that is offered.

5.3.1. Abstract Links

An abstract link is a measure of the potential to connect a pair of points with certain TE parameters. An abstract link may be realized by an existing LSP, or may represent the possibility of setting up an LSP.

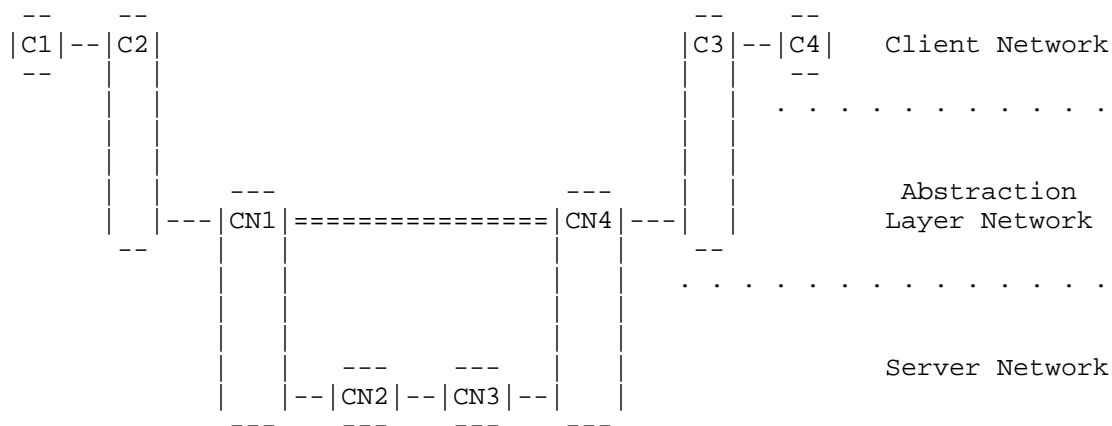
When looking at a network such as that in Figure 8, the link from CN1

to CN4 may be an abstract link. If the LSP has already been set up, it is easy to advertise it as a link with known TE attributes: policy will have been applied in the server network to decide what LSP to set up. If the LSP has not yet been established, the potential for an LSP can be abstracted from the TE information in the core network subject to policy, and the resultant potential LSP can be advertised.

Since the client nodes do not have visibility into the core network, they must rely on abstraction information delivered to them by the core network. That is, the core network will report on the potential for connectivity.

5.3.2. The Abstraction Layer Network

Figure 9 introduces the Abstraction Layer Network. This construct separates the client layer resources (nodes C1, C2, C3, and C4, and the corresponding links), and the server layer resources (nodes CN1, CN2, CN3, and CN4 and the corresponding links). Additionally, the architecture introduces an intermediary layer called the Abstraction Layer. The Abstraction Layer contains the client layer edge nodes (C2 and C3), the server layer edge nodes (CN1 and CN4), the client-server links (C2-CN1 and CN4-C3) and the abstract link CN1-CN4.



Key
 --- Direct connection between two nodes
 === Abstract link

Figure 9 : Architecture for Abstraction Layer Network

The client layer network is able to operate as normal. Connectivity across the network can either be found or not found based on links

that appear in the client layer TED. If connectivity cannot be found, end-to-end LSPs cannot be set up. This failure may be reported but no dynamic action is taken by the client layer.

The server network layer also operates as normal. LSPs across the server layer are set up in response to management commands or in response to signaling requests.

The Abstraction Layer consists of the physical links between the two networks, and also the abstract links. The abstract links are created by the server network according to local policy and represent the potential connectivity that could be created across the server network and which the server network is willing to make available for use by the client network. Thus, in this example, the diameter of the Abstraction Layer Network is only three hops, but an instance of an IGP could easily be run so that all nodes participating in the Abstraction Layer (and in particular the client network edge nodes) can see the TE connectivity in the layer.

When the client layer needs additional connectivity it can make a request to the Abstraction Layer Network. For example, the operator of the client network may want to create a link from C2 to C3. The Abstraction Layer can see the potential path C2-CN1-CN4-C3, and asks the server layer to realise the abstract link CN1-CN4. The server layer provisions the LSP CN1-CN2-CN3-CN4 and makes the LSP available as a hierarchical LSP to turn the abstract link into a link that can be used in the client network. The Abstraction Layer can then set up an LSP C2-CN1-CN4-C3 using stitching or tunneling, and make the LSP available as a virtual link in the client network.

Sections 5.3.3 and 5.3.4 show how this model is used to satisfy the requirements for connectivity in client-server networks and in peer networks.

5.3.3. Abstraction in Client-Server Networks

Section 5.3.2 has already introduced the concept of the Abstraction Layer Network through an example of a simple layered network. But it may be helpful to expand on the example using a slightly more complex network.

Figure 10 shows a multi-layer network comprising client nodes (labeled as C_n for n= 0 to 9) and server nodes (labeled as S_n for n = 1 to 9).

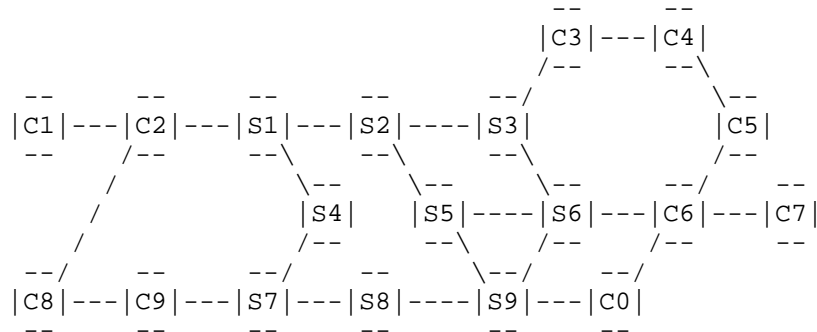


Figure 10 : An example Multi-Layer Network

If the network in Figure 10 is operated as separate client and server networks then the client layer topology will appear as shown in Figure 11. As can be clearly seen, the network is partitioned and there is no way to set up an LSP from a node on the lefthand side (say C1) to a node on the righthand side (say C7).

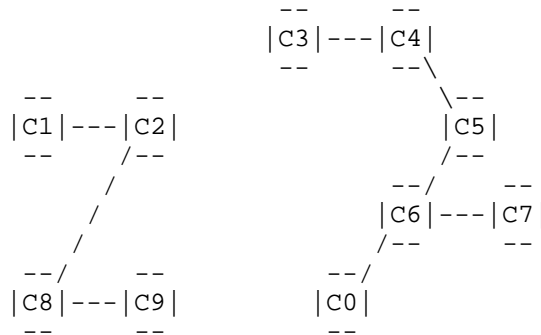


Figure 11 : Client Layer Topology Showing Partitioned Network

For reference, Figure 12 shows the corresponding server layer topology.

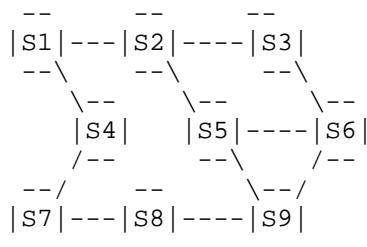


Figure 12 : Server Layer Topology

Operating on the TED for the server layer, a management entity or a software component may apply policy and consider what abstract links it might offer for use by the client layer. To do this it obviously needs to be aware of the connections between the layers (there is no point in offering an abstract link S2-S8 since this could not be of any use in this example).

In our example, after consideration of which LSPs could be set up in the server layer, four abstract links are offered: S1-S3, S3-S6, S1-S9, and S7-S9. These abstract links are shown as double lines on the resulting topology of the Abstraction Layer Network in Figure 13.

The separate IGP instance running in the Abstraction Layer Network mean that this topology is visible at the edge nodes (C2, C3, C6, C9, and C0) as well as at a PCE if one is present.

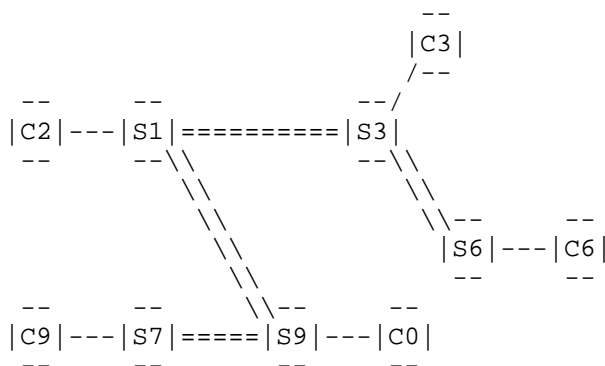


Figure 13 : Abstraction Layer Network with Abstract Links

Now the client layer is able to make requests to the Abstraction Layer Network to provide connectivity. In our example, it requests that C2 is connected to C3 and that C2 is connected to C0. This

results in several actions:

1. The management component for the Abstraction Layer Network asks its PCE to compute the paths necessary to make the connections. This yields C2-S1-S3-C3 and C2-S1-S9-C0.
2. The management component for the Abstraction Layer Network instructs C2 to start the signaling process for the new LSPs in the Abstraction Layer.
3. C2 signals the LSPs for setup using the explicit routes C2-S1-S3-C3 and C2-S1-S9-C0.
4. When the signaling messages reach S1 (in our example, both LSPs traverse S1) the Abstraction Layer Network may find that the necessary underlying LSPs (S1-S2-S3 and S1-S2-S5-S9) have not been established since it is not a requirement that an abstract link be backed up by a real LSP. In this case, S1 computes the paths of the underlying LSPs and signals them.
5. Once the serve layer LSPs have been established, S1 can continue to signal the Abstraction Layer LSPs either using the server layer LSPs as tunnels or as stitching segments.
6. Finally, once the Abstraction Layer LSPs have been set up, the client layer can be informed and can start to advertise the new TE links C2-C3 and C2-C0. The resulting client layer topology is shown in Figure 14.

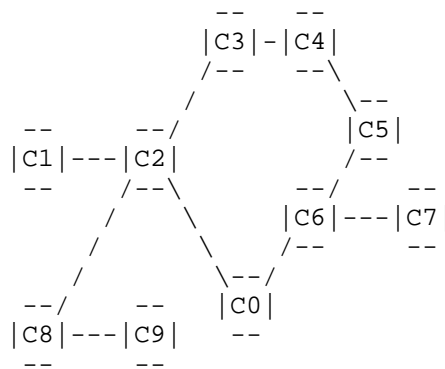


Figure 14 : Connected Client Layer Network with Additional Links

7. Now the client layer can compute an end-to-end path from C1 to C7.

5.3.3.1 Macro Shared Risk Link Groups

Network links often share fate with one or more other links. That is, a scenario that may cause a link to fail could cause one or more other links to fail. This may occur, for example, if the links are supported by the same fiber bundle, or if some links are routed down the same duct or in a common piece of infrastructure such as a bridge. A common way to identify the links that may share fate is to label them as belonging to a Shared Risk Link Group (SRLG) [RFC4202].

TE links created from LSPs in lower layers may also share fate, and it can be hard for a client network to know about this problem because it does not know the topology of the server network or the path of the server layer LSPs that are used to create the links in the client network.

For example, looking at the example used in Section 5.3.3 and considering the two abstract links S1-S3 and S1-S9 there is no way for the client layer to know whether the links C2-C0 and C2-C3 share fate. Clearly, if the client layer uses these links to provide a link-diverse end-to-end protection scheme, it needs to know that the links actually share a piece of network infrastructure (the server layer link S1-S2).

Per [RFC4202], an SRLG represents a shared physical network resource upon which the normal functioning of a link depends. Multiple SRLGs can be identified and advertised for every TE link in a network. However, this can produce a scalability problem in a multi-layer network that equates to advertising in the client layer the server layer route of each TE link.

Macro SRLGs (MSRLGs) address this scaling problem and are a form of abstraction performed at the same time that the abstract links are derived. In this way, only the links that are actually shared need to be advertised rather than every potentially shared link. This saving is possible because the abstract links are formulated on behalf of the server layer by a central management agency that is aware of all of the link abstractions being offered.

It may be noted that a less optimal alternative path for the abstract link S1-S9 exists in the server layer (S1-S4-S7-S8-S9). It would be possible for the client layer request for connectivity C2-C0 to request that the path be maximally disjoint from the path C2-C3. While nothing can be done about the shared link C2-S1, the Abstraction Layer could request that the server layer instantiate the link S1-S9 to be diverse from the link S1-S3, and this request could be honored if the server layer policy allows.

5.3.3.2 A Server with Multiple Clients

A single server network may support multiple client networks. This is not an uncommon state of affairs for example when the server network provides connectivity for multiple customers.

In this case, the abstraction provided by the server layer may vary considerably according to the policies and commercial relationships with each customer. This variance would lead to a separate Abstraction Layer Network maintained to support each client network.

On the other hand, it may be that multiple clients are subject to the same policies and the abstraction can be identical. In this case, a single Abstraction Layer Network can support more than one client.

The choices here are made as an operational issue by the server layer network.

5.3.3.3 A Client with Multiple Servers

A single client network may be supported by multiple server networks. The server networks may provide connectivity between different parts of the client network or may provide parallel (redundant) connectivity for the client network.

In this case the Abstraction Layer Network should contain the abstract links from all server networks so that it can make suitable computations and create the correct TE links in the client network. That is, the relationship between client network and Abstraction Layer Network should be one-to-one.

Note that SRLGs and MSRLGs may be very hard to describe in the case of multiple server layer networks because the abstraction points will not know whether the resources in the various server layers share physical locations.

5.3.4. Abstraction in Peer Networks

Peer networks exist in many situations in the Internet. Packet networks may peer as IGP areas (levels) or as ASes. Transport networks (such as optical networks) may peer to provide concatenations of optical paths through single vendor environments (see Section 7). Figure 15 shows a simple example of three peer networks (A, B, and C) each comprising a few nodes

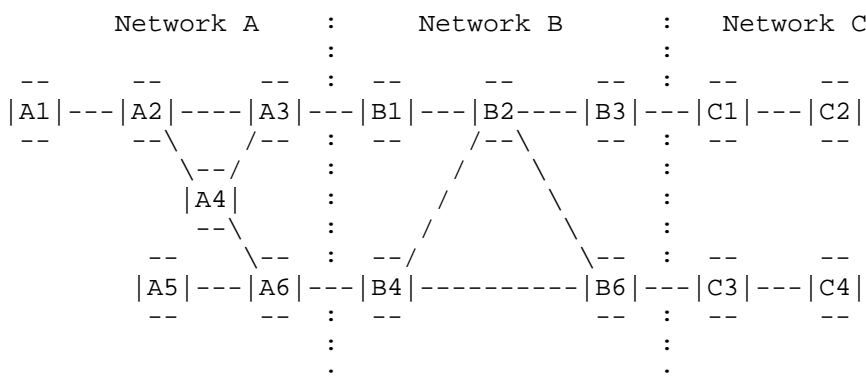


Figure 15 : A Network Comprising Three Peer Networks

As discussed in Section 2, peered networks do not share visibility of their topologies or TE capabilities for scaling and confidentiality reasons. That means, in our example, that computing a path from A1 to C4 can be impossible without the aid of cooperating PCEs or some form of crankback.

But it is possible to produce abstract links for the reachability across transit peer networks and instantiate an Abstraction Layer Network. That network can be enhanced with specific reachability information if a destination network is partitioned as is the case with Network C in Figure 15.

Suppose Network B decides to offer three abstract links B1-B3, B4-B3, and B4-B6. The Abstraction Layer Network could then be constructed to look like the network in Figure 16.

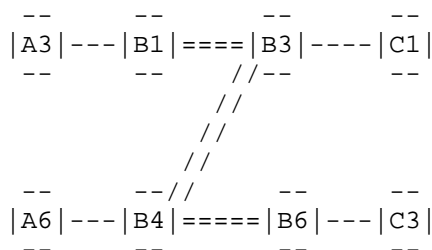


Figure 15 : Abstraction Layer Network for the Peer Network Example

Using a process similar to that described in Section 5.3.3, Network A can request connectivity to Network C and the abstract links can be instantiated as tunnels across the transit network, and edge-to-edge

LSPs can be set up to join the two networks. Furthermore, if Network C is partitioned, reachability information can be exchanged to allow Network A to select the correct edge-to-edge LSP.

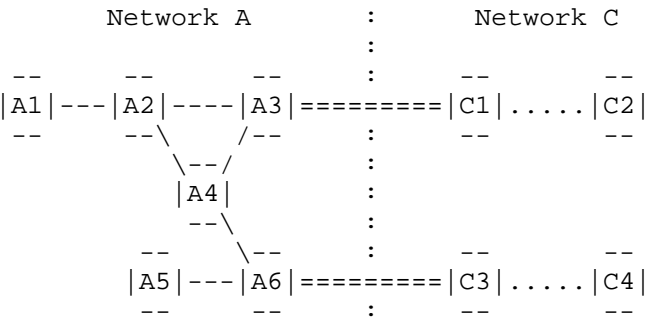


Figure 16 : Tunnel Connections to Network C with TE Reachability

Peer networking cases can be made far more complex by dual homing between network peering nodes (for example, A3 might connect to B1 and B4 in Figure 15) and by the networks themselves being arranged in a mesh (for example, A6 might connect to B4 and C1 in Figure 15). These additional complexities can be handled gracefully by the Abstraction Layer Network model.

Further examples of abstraction in peer networks can be found in Sections 7 and 8.

5.4. Considerations for Dynamic Abstraction

<TBD>

5.5. Requirements for Advertising Abstracted Links and Nodes

<TBD>

6. Building on Existing Protocols

6.1. BGP-LS

<TBD>

6.2. IGPs

<TBD>

6.3. RSVP-TE

<TBD>

7. Applicability to Optical Domains and Networks

Many optical networks are arranged a set of small domains. Each domain is a cluster of nodes, usually from the same equipment vendor and with the same properties. The domain may be constructed as a mesh or a ring, or maybe as an interconnected set of rings.

The network operator seeks to provide end-to-end connectivity across a network constructed from multiple domains, and so (of course) the domains are interconnected. In a network under management control such as through an Operations Support System (OSS), each domain is under the operational control of a Network Management System (NMS). In this way, an end-to-end path may be commissioned by the OSS instructing each NMS, and the NMSes setting up the path fragments across the domains.

However, in a system that uses a control plane, there is a need for integration between the domains.

Consider a simple domain, D1, as shown in Figure 16. In this case, the nodes A through F are arranged in a topological ring. Suppose that there is a control plane in use in this domain, and that OSPF is used as the TE routing protocol.

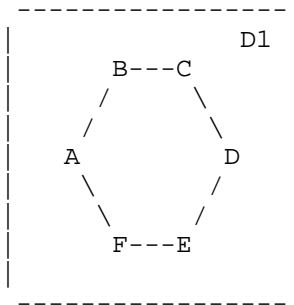


Figure 16 : A Simple Optical Domain

Now consider that the operator's network is built from a mesh of such domains, D1 through D7, as shown in Figure 17. It is possible that these domains share a single, common instance of OSPF in which case there is nothing further to say because that OSPF instance will

distribute sufficient information to build a single TED spanning the whole network, and an end-to-end path can be computed. A more likely scenario is that each domain is running its own OSPF instance. In this case, each is able to handle the peculiarities (or rather, advanced functions) of each vendor's equipment capabilities.

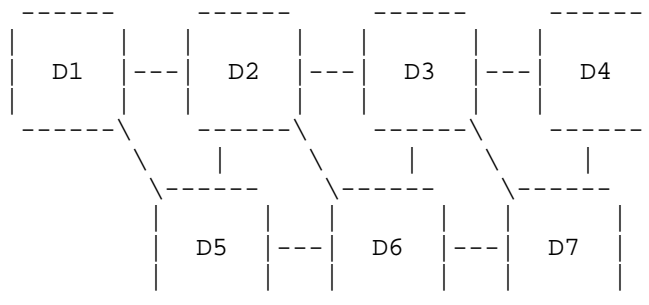


Figure 17 : A Simple Optical Domain

The question now is how to combine the multiple sets of information distributed by the different OSPF instances. Three possible models suggest themselves based on pre-existing routing practices.

- o In the first model (the Area-Based model) each domain is treated as a separate OSPF area. The end-to-end path will be specified to traverse multiple areas, and each area will be left to determine the path across the nodes in the area. The feasibility of an end-to-end path (and, thus, the selection of the sequence of areas and their interconnections) can be derived using hierarchical PCE.

This approach, however, fits poorly with established use of the OSPF area: in this form of optical network, the interconnection points between domains are likely to be links; and the mesh of domains is far more interconnected and unstructured than we are used to seeing in the normal area-based routing paradigm.

Furthermore, while hierarchical PCE may be able to solve this type of network, the effort involved may be considerable for more than a small collection of domains.

- o Another approach (the AS-Based model) treats each domain as a separate Autonomous System (AS). The end-to-end path will be specified to traverse multiple ASes, and each AS will be left to determine the path across the AS.

This model sits more comfortably with the established routing

paradigm, but causes a massive escalation of ASes in the global Internet. It would, in practice, require that the operator used private AS numbers [RFC6996] of which there are plenty.

Then, as suggested in the Area-Based model, hierarchical PCE could be used to determine the feasibility of an end-to-end path and to derive the sequence of domains and the points of interconnection to use. But, just as in that other model, the scalability of the hierarchical PCE approach must be questioned.

Furthermore, determining the mesh of domains (i.e., the inter-AS connections) conventionally requires the use of BGP as an inter-domain routing protocol. However, not only is BGP not normally available on optical equipment, but this approach indicates that the TE properties of the inter-domain links would need to be distributed and updated using BGP: something for which it is not well suited.

- o The third approach (the ASON model) follows the architectural model set out by the ITU-T [G.8080] and uses the routing protocol extensions described in [RFC6827]. In this model the concept of "levels" is introduced to OSPF. Referring back to Figure 17, each OSPF instance running in a domain would be construed as a "lower level" OSPF instance and would leak routes into a "higher level" instance of the protocol that runs across the whole network.

This approach handles the awkwardness of representing the domains as areas or ASes by simply considering them as domains running distinct instances of OSPF. Routing advertisements flow "upward" from the domains to the high level OSPF instance giving it a full view of the whole network and allowing end-to-end paths to be computed. Routing advertisements may also flow "downward" from the network-wide OSPF instance to any one domain so that it has visibility of the connectivity of the whole network.

While architecturally satisfying, this model suffers from having to handle the different characteristics of different equipment vendors. The advertisements coming from each low level domain would be meaningless when distributed into the other domains, and the high level domain would need to be kept up-to-date with the semantics of each new release of each vendor's equipment. Additionally, the scaling issues associated with a well-meshed network of domains each with many entry and exit points and each with network resources that are continually being updated reduces to the same problem as noted in the virtual link model. Furthermore, in the event that the domains are under control of different administrations, the domains would not want to distribute the details of their topologies and TE resources.

Practically, this third model turns out to be very close to the methodology described in this document. As noted in Section 7.1 of [RFC6827], there are policy rules that can be applied to define exactly what information is exported from or imported to a low level OSPF instance. The document even notes that some forms of aggregation may be appropriate. Thus, we can apply the following simplifications to the mechanisms defined in RFC 6827:

- Zero information is imported to low level domains.
- Low level domains export only abstracted links as defined in this document and according to local abstraction policy and with appropriate removal of vendor-specific information.
- There is no need to formally define routing levels within OSPF.
- Export of abstracted links from the domains to the network-wide routing instance (the abstraction routing layer) can take place through any mechanism including BGP-LS or direct interaction between OSPF implementations.

With these simplifications, it can be seen that the framework defined in this document can be constructed from the architecture discussed in RFC 6827, but without needing any of the protocol extensions that that document defines. Thus, using the terminology and concepts already established, the problem may solved as shown in Figure 18. The abstraction layer network is constructed from the inter-domain links, the domain border nodes, and the abstracted (cross-domain) links.

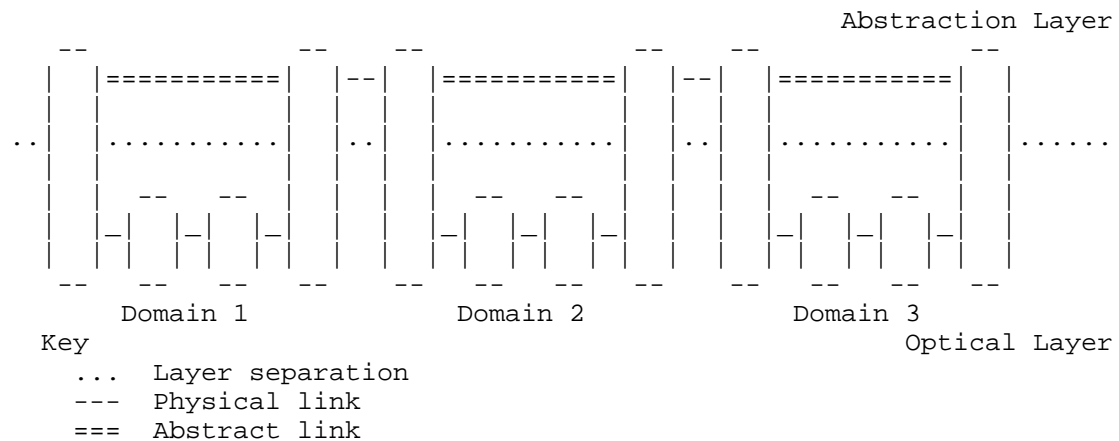


Figure 18 : The Optical Network Implemented Through the Abstraction Layer Network

8. Abstraction in L3VPN Multi-AS Environments

Serving layer-3 VPNs (L3PVNs) across a multi-AS or multi-operator environment currently provides a significant planning challenge. This section shows how the Abstraction Layer Network can address this problem.

<TBD>

9. Scoping Future Work

The section is provided to help guide the work on this problem and to ensure that oceans are not knowingly boiled.

9.1. Not Solving the Internet

The scope of the use cases and problem statement in this document is limited to "some small set of interconnected domains." In particular, it is not the objective of this work to turn the whole Internet into one large, interconnected TE network.

9.2. Working With "Related" Domains

Subsequent to Section 9.1, the intention of this work is to solve the TE interconnectivity for only "related" domains. Such domains may be under common administrative operation (such as IGP areas within a single AS, or ASes belonging to a single operator), or may have a direct commercial arrangement for the sharing of TE information to provide specific services. Thus, in both cases, there is a strong opportunity for the application of policy.

9.3. Not Breaking Existing Protocols

It is a clear objective of this work to not break existing protocols. The Internet relies on the stability of a few key routing protocols, and so it is critical that any new work must not make these protocols brittle or unstable.

9.4. Sanity and Scaling

All of the above points play into a final observation. This work is intended to bite off a small problem for some relatively simple use cases as described in Section 2. It is not intended that this work will be immediately (or even soon) extended to cover many large interconnected domains. Obviously the solution should as far as possible be designed to be extensible and scalable, however, it is also reasonable to make trade-offs in favor of utility and simplicity.

10. Manageability Considerations

<TBD>

11. IANA Considerations

This document makes no requests for IANA action.

12. Security Considerations

<TBD>

13. Acknowledgements

Thanks to Igor Bryskin for useful discussions in the early stages of this work.

Thanks to Gert Grammel for discussions on the extent of aggregation in abstract nodes and links.

Thanks to Deborah Brungard and Dieter Beller for review comments.

Particular thanks to Vishnu Pavan Beeram for detailed discussions and white-board scribbling that made many of the ideas in this document come to life.

Text in Section 5.3.3 is freely adapted from the work of Igo Bryskin, Wes Doonan, Vishnu Pavan Beeram, John Drake, Gert Grammel, Manuel Paul, Ruediger Kunze, Friedrich Armbruster, Cyril Margaria, Oscar Gonzalez de Dios, and Daniele Ceccarelli in [I-D.beeram-ccamp-gmpls-enni] for which the authors of this document express their thanks.

14. References

14.1. Informative References

[G.8080] ITU-T, "Architecture for the automatically switched optical network (ASON)", Recommendation G.8080.

[I-D.beeram-ccamp-gmpls-enni]

Bryskin, I., Beeram, V. P., Drake, J. et al., "Generalized Multiprotocol Label Switching (GMPLS) External Network Network Interface (E-NNI): Virtual Link Enhancements for the Overlay Model", draft-beeram-ccamp-gmpls-enni, work in progress.

- [I-D.farrkingel-pce-questions]
Farrel, A., and D. King, "Unanswered Questions in the Path Computation Element Architecture", draft-farrkingel-pce-questions, work in progress.
- [I-D.ietf-ccamp-xro-lsp-subobject]
Z. Ali, et al., "Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) LSP Route Diversity using Exclude Routes," draft-ali-ccamp-xro-lsp-subobject, work in progress.
- [I-D.ietf-ccamp-te-metric-recording]
Z. Ali, et al., "Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) extension for recording TE Metric of a Label Switched Path," draft-ali-ccamp-te-metric-recording, work in progress.
- [I-D.ietf-idr-ls-distribution]
Gredler, H., Medved, J., Previdi, S., Farrel, A., and Ray, S., "North-Bound Distribution of Link-State and TE Information using BGP", draft-ietf-idr-ls-distribution, work in progress.
- [RFC2702] Awduche, D., Malcolm, J., Agogbua, J., O'Dell, M., and McManus, J., "Requirements for Traffic Engineering Over MPLS", RFC 2702, September 1999.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC3473] L. Berger, "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RC 3473, January 2003.
- [RFC3630] Katz, D., Kompella, and K., Yeung, D., "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, September 2003.
- [RFC3945] Mannie, E., (Ed.), "Generalized Multi-Protocol Label Switching (GMPLS) Architecture", RFC 3945, October 2004.
- [RFC4105] Le Roux, J.-L., Vasseur, J.-P., and Boyle, J., "Requirements for Inter-Area MPLS Traffic Engineering", RFC 4105, June 2005.

- [RFC4202] Kompella, K. and Y. Rekhter, "Routing Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4202, October 2005.
- [RFC4206] Kompella, K. and Y. Rekhter, "Label Switched Paths (LSP) Hierarchy with Generalized Multi-Protocol Label Switching (GMPLS) Traffic Engineering (TE)", RFC 4206, October 2005.
- [RFC4208] Swallow, G., Drake, J., Ishimatsu, H., and Y. Rekhter, "User-Network Interface (UNI): Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Support for the Overlay Model", RFC 4208, October 2005.
- [RFC4216] Zhang, R., and Vasseur, J.-P., "MPLS Inter-Autonomous System (AS) Traffic Engineering (TE) Requirements", RFC 4216, November 2005.
- [RFC4271] Rekhter, Y., Li, T., and Hares, S., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.
- [RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, August 2006.
- [RFC4726] Farrel, A., Vasseur, J.-P., and Ayyangar, A., "A Framework for Inter-Domain Multiprotocol Label Switching Traffic Engineering", RFC 4726, November 2006.
- [RFC4847] T. Takeda (Ed.), "Framework and Requirements for Layer 1 Virtual Private Networks," RFC 4847, April 2007.
- [RFC4920] Farrel, A., Satyanarayana, A., Iwata, A., Fujita, N., and Ash, G., "Crankback Signaling Extensions for MPLS and GMPLS RSVP-TE", RFC 4920, July 2007.
- [RFC5150] Ayyangar, A., Kompella, K., Vasseur, JP., and A. Farrel, "Label Switched Path Stitching with Generalized Multiprotocol Label Switching Traffic Engineering (GMPLS TE)", RFC 5150, February 2008.
- [RFC5152] Vasseur, JP., Ayyangar, A., and Zhang, R., "A Per-Domain Path Computation Method for Establishing Inter-Domain Traffic Engineering (TE) Label Switched Paths (LSPs)", RFC 5152, February 2008.
- [RFC5195] Ould-Brahim, H., Fedyk, D., and Y. Rekhter, "BGP-Based Auto-Discovery for Layer-1 VPNs", RFC 5195, June 2008.

- [RFC5212] Shiimoto, K., Papadimitriou, D., Le Roux, JL., Vigoureux, M., and D. Brungard, "Requirements for GMPLS-Based Multi-Region and Multi-Layer Networks (MRN/MLN)", RFC 5212, July 2008.
- [RFC5251] Fedyk, D., Rekhter, Y., Papadimitriou, D., Rabbat, R., and L. Berger, "Layer 1 VPN Basic Mode", RFC 5251, July 2008.
- [RFC5252] Bryskin, I. and L. Berger, "OSPF-Based Layer 1 VPN Auto-Discovery", RFC 5252, July 2008.
- [RFC5305] Li, T., and Smit, H., "IS-IS Extensions for Traffic Engineering", RFC 5305, October 2008.
- [RFC5440] Vasseur, JP. and Le Roux, JL., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.
- [RFC5441] Vasseur, JP., Zhang, R., Bitar, N., and Le Roux, JL., "A Backward-Recursive PCE-Based Computation (BRPC) Procedure to Compute Shortest Constrained Inter-Domain Traffic Engineering Label Switched Paths", RFC 5441, April 2009.
- [RFC5523] L. Berger, "OSPFv3-Based Layer 1 VPN Auto-Discovery", RFC 5523, April 2009.
- [RFC5553] Farrel, A., Bradford, R., and JP. Vasseur, "Resource Reservation Protocol (RSVP) Extensions for Path Key Support", RFC 5553, May 2009.
- [RFC5623] Oki, E., Takeda, T., Le Roux, JL., and A. Farrel, "Framework for PCE-Based Inter-Layer MPLS and GMPLS Traffic Engineering", RFC 5623, September 2009.
- [RFC6107] Shiimoto, K., and A. Farrel, "Procedures for Dynamically Signaled Hierarchical Label Switched Paths", RFC 6107, February 2011.
- [RFC6805] King, D., and A. Farrel, "The Application of the Path Computation Element Architecture to the Determination of a Sequence of Domains in MPLS and GMPLS", RFC 6805, November 2012.
- [RFC6827] Malis, A., Lindem, A., and D. Papadimitriou, "Automatically Switched Optical Network (ASON) Routing for OSPFv2 Protocols", RFC 6827, January 2013.

[RFC6996] J. Mitchell, "Autonomous System (AS) Reservation for Private Use", BCP 6, RFC 6996, July 2013.

Authors' Addresses

Adrian Farrel
Juniper Networks
EMail: adrian@olddog.co.uk

John Drake
Juniper Networks
EMail: jdrake@juniper.net

Nabil Bitar
Verizon
40 Sylvan Road
Waltham, MA 02145
EMail: nabil.bitar@verizon.com

George Swallow
Cisco Systems, Inc.
1414 Massachusetts Ave
Boxborough, MA 01719
EMail: swallow@cisco.com

Daniele Ceccarelli
Ericsson
Via A. Negrone 1/A
Genova - Sestri Ponente
Italy
EMail: daniele.ceccarelli@ericsson.com

Appendix A. Editor's Notes

[Editor Note: Need to work up some text on addressing to cover the case of each domain having a different (potentially overlapping) address space and the need for inter-domain addressing. In fact, this should be quite simple but needs discussion.]

[Editor Note: Need to explain how the IGP in the Abstraction Layer works to distribute connectivity information when the Abstract Link is not yet up. The answer will be that the DCN needs to exist regardless of the state of the Abstract Link.]

Network Working Group
Internet Draft
Intended status: Standards Track

D. Fedyk
D. Beller
Lieven Levrau
Alcatel-Lucent
D. Ceccarelli
Ericsson
F. Zhang
Huawei Technologies
Y. Tochio
Fujitsu
X. Fu
ZTE

Expires: January 16, 2014

July 15, 2013

UNI Extensions for Diversity and Latency Support
draft-fedyk-ccamp-uni-extensions-02.txt

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), 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>

Copyright Notice

Copyright (c) 2013 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.

Abstract

This document builds on the GMPLS overlay model [RFC4208] and defines extensions to the GMPLS User-Network Interface (UNI) to support route diversity within the core network for sets of LSPs initiated by edge nodes. A particular example where route diversity within the core network is desired, are dual-homed edge nodes. The document also defines GMPLS UNI extensions to deal with latency requirements for edge node initiated LSPs.

This document uses a VPN model that is based on the same premise as L1VPN framework [RFC4847] but may also be applied to other technologies. The extensions are applicable both to VPN and non VPN environments. These extensions move the UNI from basic connectivity to enhanced mode connectivity by including additional constraints while minimizing the exchange of CE to PE information. These extensions are applicable to the overlay extension service model. Route Diversity for customer LSPs are a common requirement applicable to L1VPNs. The UNI mechanisms described in this document are L1VPN compatible and can be applied to achieve diversity for sets of customer LSPs.

The UNI extensions in support of latency constraints can also be applied to the extended overlay service model in order for the customer LSPs to meet certain latency requirements.

Table of Contents

1. Introduction	3
2. Conventions used in this document	4
3. Contributors	4
4. LSP Diversity in the Overlay Extension Service Model	4
4.1. LSP diversity for dual-homed customer edge (CE) devices	5
4.1.1. Exchanging SRLG information between the PEs via the CE device	8
4.1.1.1. Operational Procedures	8
4.1.1.2. Error Handling Procedures	9
4.1.2. Using Path Affinity Set Extension	10
4.1.2.1. Operational Procedures	13
4.1.2.2. Error Handling Procedures	13
4.1.2.3. Distribution of the Path Affinity Set Information	14
5. Latency Signaling Extensions	15
5.1. RSVP-TE Extensions	16
5.2. Operational Procedures	16
5.3. Error Handling Procedures	16
6. Security Considerations	16
7. IANA Considerations	17
8. References	17
8.1. Normative References	17
8.2. Informative References	17
Authors' Addresses	19

1. Introduction

This document builds on the GMPLS overlay model [RFC4208] and defines extensions to the GMPLS User-Network Interface (UNI) to support route diversity within the core network for sets of LSPs initiated by edge nodes. In the following, the term customer edge (CE) device is used synonymously for the term edge node (EN) as in [RFC4208].

Moreover, the VPN terminology (CE and PE) [RFC4026] is used below when the core network is a VPN but is also applicable to UNI interfaces [RFC4208].

This document uses a VPN model that is based on the same premise as L1VPN framework [RFC4847] but may also be applied to other technologies. The extensions are applicable both to VPN and non VPN environments. These extensions move the UNI from basic connectivity to enhanced mode connectivity by including additional constraints while minimizing the exchange of CE to PE information. These extensions are applicable to the overlay extension service model.

The overlay model assumes a UNI interface between the edge nodes of the respective transport domains. Route diversity for LSPs from single homed CE and dual-home CEs is a common requirement in optical transport networks. This document describes two signaling variations that may be used for supporting LSP diversity within the overlay extension service model considering dual-homing. Dual-homing is typically used to avoid a single point of failure (UNI link, PE) or if two disjoint connections are forming a protection group in the CE device, e.g., 1+1 protection. While both methods are similar in that they utilize common mechanisms in the PE network to achieve diversity, they are distinguished according to whether the CE is permitted to retrieve provider SRLG diversity information for an LSP from a PE1 and pass it on to a PE2 (SRLG information is shared with the CE), or whether a new attribute is used that allows the PE2 that receives this attribute to derive the SRLG information for an LSP based on the attribute value. Figure 1 below is depicting the scenario.

The extended overlay service model can support other extensions for VPN signaling, for example, those related to latency. When requesting diverse LSPs, latency may also be an additional requirement.

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 RFC-2119 [RFC2119].

In this document, these words will appear with that interpretation only when in ALL CAPS. Lower case uses of these words are not to be interpreted as carrying RFC-2119 significance.

3. Contributors

The Authors would like to thank Eve Varma and Sergio Belotti for their review and contributions to this document.

4. LSP Diversity in the Overlay Extension Service Model

The L1VPN Framework [RFC4847] (Enhanced Mode) describes the overlay extension service model, which builds upon the UNI Overlay [RFC4208] serving as the interface between the CE edge node and the PE edge node. In this service model, a CE receives a list of CE-PE TE link addresses to which it can request a L1VPN connection (i.e., membership information) and may include additional information concerning these TE links. This document further builds on the overlay extension service model by adding shared constraint information for path diversity in the optical transport network.

While the L1VPN for optical transport is an example specific VPN technology the term VPN is used generically since the extensions can apply to GMPLS UNIs and VPNs for other technologies.

Two signaling variations are outlined here that may be used for supporting LSP diversity within the overlay extension service model considering dual-homing. While both methods utilize common mechanisms in the PE network to achieve diversity, they are distinguished according to whether the CE is permitted to retrieve provider SRLG diversity information for an LSP from a PE1 and pass it on to a PE2 (SRLG information is shared with the CE or whether a new attribute is used that allows the PE2 that receives this attribute to derive the SRLG information for an LSP based on this attribute value. The selection between these methods is governed by both PE-network specific policies and approaches taken (i.e., in terms of how the provider chooses to perform routing internal to their network).

The first method (see 3.1.1) assumes that provider Shared Resource Link Group (SRLG) Identifier information is both available and shareable (policy decision) with the CE. Since SRLG IDs can then be used (passed transparently between PEs via the dual-homed CE) as signaled information on a UNI message, a mechanism supporting LSP diversity for the overlay extension service model can be provided via straightforward signaling extensions.

The second method (see 3.1.2) assumes that provider SRLG IDs are either not available or not shareable (based on provider network operator policy) with the CE. For this case, a mechanism is provided where information signaled to the PE on UNI messages does not require shared knowledge of provider SRLG IDs to support LSP diversity for the overlay extension model.

While both methods could be implemented in the same PE network, it is likely that a GMPLS VPN CE network would use only one mechanism at a time.

4.1. LSP diversity for dual-homed customer edge (CE) devices

Single-homed CE devices are connected to a single PE device via a single UNI link (could be a bundle of parallel links which are typically using the same fiber cable). This single UNI link may constitute a single point of failure. Such a single point of failure can be avoided when the CE device is connected to two PE devices via two UNI interfaces as depicted for CE1 in Figure 1 below.

For the dual-homing case, it is possible to establish two connections from the source CE device to the same destination CE device where one connection is using one UNI link to, for example, PE1 and the other

connection is using the UNI link to PE2. In order to avoid single points of failure within the provider network, it is necessary to also ensure path (LSP) diversity within the provider network in order to achieve end-to-end diversity for the two LSPs between the two CE devices. This document describes how it is possible to enable such path diversity to be achieved within the provider network (which is subject to additional routing constraints). [RFC4202] defines SRLG information that can be used to allow GMPLS to provide path diversity in a GMPLS controlled transport network. As the two connections are entering the provider network at different PE devices, the PE device that receives the connection request for the second connection needs to be capable of determining the additional path computation constraints such that the path of the second LSP is disjoint with respect to the already established first connection entering the network at a different PE device. The methods described in this document allow a PE device to determine the SRLG information for a connection in the provider network that is entering the network on a different PE device.

PE SRLG information can be used directly by a CE if the CE understands the context, and the CE view is limited to its VPN context. In this case, there is a dependency on the provider information and there is a need to be able to query the SRLG in the provider network.

It may, on the other hand, be preferable to avoid this dependency and to decouple the SRLG identifier space used in the provider network from the SRLG space used in the client network. This is possible with both methods detailed below. Even for the method where provider SRLG information is passing through the CE device (note the CE device does not need to process and decode this information) the two SRLG identifier spaces can remain fully decoupled and the operator of the client network is free to assign SRLG identifiers from the client SRLG identifier space to the CE to CE connection that is passing through the provider network.

Referring to Figure 1, the UNI signaling mechanism must support at least one of the two mechanisms described in this document for CE dual homing to achieve LSP diversity in the provider network.

The described mechanisms can also be applied to a scenario where two CE devices are connected to two different PE devices. In this case, the additional information that is exchanged across the UNI interfaces also needs to be exchanged between the two CE devices in order to achieve the desired diversity in the provider network.

This information may be configured or exchanged by some automated

mechanism not described in this document.

In the dual-homing example, CE1 can locally correlate the LSP requests. For the slightly more complicated example involving CE2 and CE3, both requiring a path that shall be diverse to a connection initiated by the other CE device, CE2 and CE3 need to have a common view of the SRLG information to be signaled. In this document, we detail the required diversity information and the signaling of this diversity information; however, the means for distributing this information within the PE domain or the CE domain is out of scope.

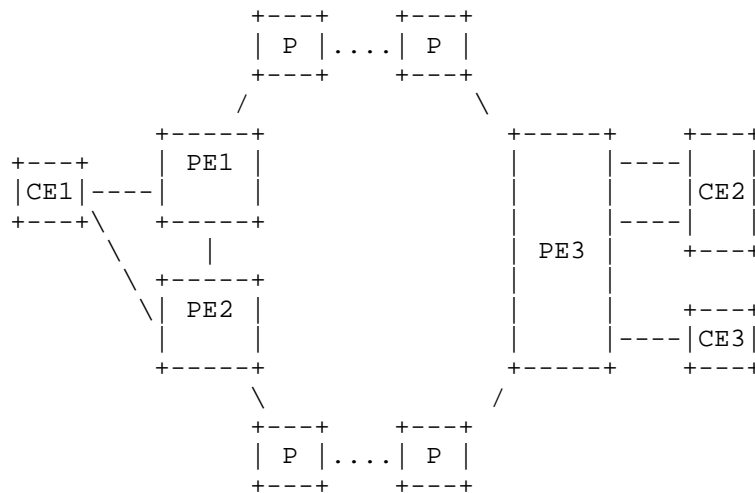


Figure 1 Overlay Reference Diagram

In an overlay model, the information exchanged between the CE and the PE is kept to a minimum.

How diversity is achieved, in terms of configuration, distribution and usage in each part of the transport networks should be kept independent and separate from how diversity is signaled at the UNI between the two transport networks.

Signaling parameters discussed in this document are:

- o SRLG information (see [RFC4202])
- o Path Affinity Set

4.1.1.1. Exchanging SRLG information between the PEs via the CE device

SRLG information is defined in [RFC4202] and if the SRLG information of an LSP is known, it can be used to calculate a path for another LSP that is SRLG diverse with respect to an existing LSP. SRLG information is an unordered list of SRLGs. SRLG information is normally not shared between the transport network and the client network; i.e., not shared with the CEs of a VPN in the VPN context. However, this becomes more challenging when a CE is dual-homed. For example, CE1 in Figure 1 may have requested an LSP1 from CE1 to CE2 via PE1 and PE3. CE1 could subsequently request an LSP2 to CE2 via PE2 and PE3 with the requirement that it should be maximally SRLG disjoint with respect to LSP1. Since PE2 does not have any information about LSP1, PE2 would need to know the SRLG information associated with LSP1. If CE1 could request the SRLG information of LSP1 from PE1, it could then transparently pass this information to PE2 as part of the LSP2 setup request, and PE2 would now be capable of calculating a path for LSP2 that is SRLG disjoint with respect to LSP1.

The exchange of SRLG information is achieved on a per VPN LSP basis using the existing RSVP-TE signaling procedures. It can be exchanged in the PATH (exclusion information) or RESV message in the original request or it can be requested by the CE at any time the path is active.

It shall be noted that SRLG information is an unordered list of SRLG identifiers and the encoding of SRLG information for RSVP signaling is already defined in [SRLG_info]. Even if SRLG information is known for several LSPs it is not possible for the CEs to derive the provider network topology from this information.

4.1.1.1.1. Operational Procedures

Retrieving SRLG information from a PE for an existing LSP:

When a dual-homed CE device intends to establish an LSP to the same destination CE device via another PE node, it can request the SRLG information for an already established LSP by setting the SRLG information flag in the LSP attributes sub-object of the RSVP PATH message (IANA to assign the new SRLG flag). As long as the SRLG information flag is set in the PATH message, the PE node inserts the

SRLG sub-object as defined in [SRLG_info] into the RSVP RESV message that contains the current SRLG information for the LSP. If the provider network's policy has been configured so as not to share SRLG information with the client network, the SRLG sub-object is not inserted in the RESV message even if the SRLG information flag was set in the received PATH message. Note that the SRLG information is expected to be always up-to-date.

Establishment of a new LSP with SRLG diversity constraints:

When a dual-homed CE device sends an LSP setup requests to a PE device for a new LSP that is required to be SRLG diverse with respect to an existing LSP that is entering the network via another PE device, the CE device sets the SRLG diversity flag (note: IANA to assign the new SRLG diversity flag) in the LSP attributes sub-object of the PATH message that initiates the setup of this new LSP. When the PE device receives this request it calculates a path to the given destination and uses the received SRLG information as path computation constraints.

4.1.1.2. Error Handling Procedures

When the CE device receives a RSVP PATH message with the SRLG information flag set and if the provider's network policy does not permit sharing of SRLG information, the PE device shall notify the CE device by sending a RSVP PathErr with a Notify error code (error code to be defined) "Retrieval of SRLG information not permitted". As described above, the PE device must not include the SRLG sub-object with the SRLG information for the LSP in the RSVP RESV message.

If the PE device receives a RSVP PATH message for a new LSP with the SRLG diversity flag set and SRLG information in the SRLG sub-object, the PE device tries to calculate a route to the given destination that is SRLG diverse with respect to the provided SRLG information. If no route can be found, a RSVP PathErr message with an error code (error code to be defined) "No SRLG diverse route available toward destination".

If the PE device receives a RSVP PATH message for a new LSP with the SRLG diversity flag set and SRLG information in the SRLG sub-object and if the PE device does not support the SRLG sub-object, the PE device shall send a PathErr message to the CE device, indicating an "Unknown object class".

Further error handling cases will be added in the next revision of

this document.

4.1.2. Using Path Affinity Set Extension

The Path Affinity Set (PAS) is used to signal diversity in a pure CE context by abstracting SRLG information. There are two types of diversity information in the PAS. The first type of information is a single PAS identifier. The Second part is the optional PATH information, in the form of Source and Destination addresses of an exclude path or set of paths that MAY be specified. The motive behind the PAS information is to have as little exchange of diversity information as possible between the VPN CE and PE elements.

Rather than a detailed CE or PE SRLG list, the Path Affinity Set contains an abstract SRLG identifier that associates the given path as diverse. Logically the identifier is in a VPN context and therefore only unique with respect to a particular VPN.

How the CE determines the PAS identifier is a local matter for the CE administrator. A CE may signal the PAS identifier as a diversity object in the PATH message. This identifier is a suggested identifier and may be overridden by a PE under some conditions.

For example, a PAS identifier can be used with no prior exchange of PAS information between the CE and the PE. Upon reception of the PAS identifier information the PE can infer the CE's requirements. The actual PAS identifier used will be returned in the RESV message. Optionally an empty PAS identifier allows the PE to pick the PAS identifier.

Similar to the section 4.1.1 on SRLG information, a PE can return PAS identifier as the response to a Query allowing flexibility.

A PE interprets the specific PAS identifier, for example, "123" as meaning to exclude the PE SRLG information (or equivalent) that has been allocated by LSPs associated with this Path Affinity Set identifier "123", for any LSPs associated with the resources assigned to the VPN. For example, if a Path exists for the LSP with the identifier "123", the PE would use local knowledge of the PE SRLGs associated with the "123" LSPs and exclude those SRLGs in the path request. In other words, two LSPs that need to be diverse both signal "123" and the PEs interpret this as meaning not to use shared resources. Alternatively, a PE could use the PAS identifier to select from already established LSPs. Once the path is established it becomes the "123" identifier or optionally another PAS identifier for that VPN that replaces "123".

The optional PAS Source and Destination Address tuple represents one or more source addresses and destination addresses associated with the CE Path Affinity Set identifier. These associated address tuples represent paths that use resources that should be excluded for the establishment of the current LSP. The address tuple information gives both finer grain details on the path diversity request and serves as an alternative identifier in the case when the PAS identifier is not known by the PE. The address tuples used in signaling is within a CE context and its interpretation is local to a PE that receives a Path request from a CE. The PE can use the address information to relate to PE Addresses and PE SRLG information. When a PE satisfies a connection setup for a (SRLG) diverse signaled path, the PE may optionally record the PE SRLG information for that connection in terms of PE based parameters and associate that with the CE addresses in the Path message.

Specifically for L1VPNs, Port Information table (PIT) [RFC5251] can be leveraged to translate between CE based addresses and PE based addresses. The Path Affinity Set and associated PE addresses with PE SRLG information can be distributed via the IGP in the provider transport network (or by other means such as configuration); they can be utilized by other PEs when other CE Paths are setup that would require path/connection diversity. This information is distributed on a VPN basis and contains a PAS identifier, PE addresses and SRLG information.

If diversity is not signaled, the assumption is that no diversity is required and the Provider network is free to route the LSP to optimize traffic. No Path affinity set information needs to be recorded for these LSPs. If a diversity object is included in the connection request, the PE in the Provider Network should be able to look-up the existing Provider SRLG information from the provider network and choose an LSP that is maximally diverse from other LSPs.

The mechanisms to achieve this are outside the scope of this document.

A new VPN Diverse LSP LABEL object is specified:

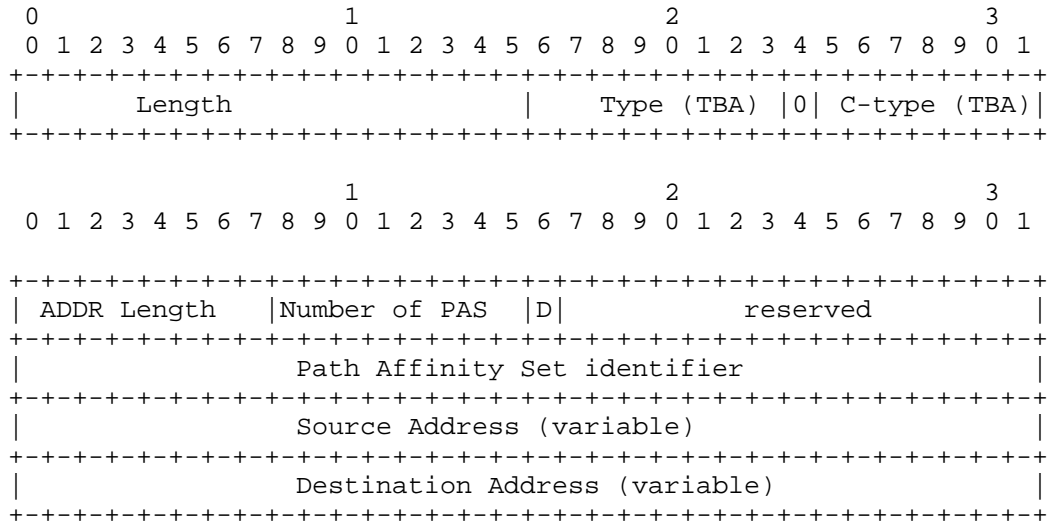


Figure 2 Diverse LSP information

1. The Address Length field (8 bits) is the number of bytes for both the source address and destination address. The address may be in any format from 1 to 32 bytes but the key point is the customers can maintain their existing addresses. A value of zero indicates there are no addresses included.
2. The Number of Path Affinity (8 bits) sets is included in the object. This is typically 1. Addition of other sets is for further study.
3. The Path affinity Set identifier (4 bytes) is a single number that represents a summarized SRLG for this path. Paths with that same Path Affinity set should be set up with diverse paths and associated with the path affinity set. A value of all zeros allows the PE to pick a PAS identifier to return. A PAS identifier of an established path may be different than the requested path identifier.
4. The diversity Bit (D) (one Bit) indicates if the diversity must be satisfied when set as a one. If a PE finds an established path with a Path Affinity set matching the signaled Path Affinity Set or the signaled Address tuple it should attempt find a diverse path.

5. The Diverse Path Source address/destination address tuple is that of an established LSP in the PE network that belongs to the same Path Affinity Set identifier. If the path for these addresses is not established or cannot be determined by the PE edge processing the PATH request then the path is established only with the Path Affinity identifier. If the path(s) for these address tuples are known by the PE the PE uses the SRLG information associated with these addresses. If in any case a diverse path cannot be setup then the Diverse bit controls whether a path is established anyway. The PE must use the PIT to translate CE Addresses into provider addresses when correlating with provider SRLG information. How SRLG information and network address tuples are distributed is for future study.

4.1.2.1. Operational Procedures

When a CE constructs a PATH message it may optionally specify and insert a Path Affinity Set in the PATH message. This Path Affinity Set may optionally include the address of an LSP that that could belong to the same Path Affinity Set. The Path Affinity Set identifier is a value (0 through $2^{32}-255$) that is independent of the mechanism the CE or the PE use for diversity. The Path Affinity Set is a single identifier that can be used to request diversity and associate diversity.

When processing a CE PATH message in a VPN Overlay, the PE first looks up the PE based addresses in the Provider Index Table (PIT). If the Path Affinity Set is included in the PATH message, the PE must look up the SRLG information (or equivalent) in the PE network that has been allocated by LSPs associated with a Path Affinity Set and exclude those resources from the path computation for this LSP if it is a new path. The PE may alternatively choose from an existing path with a disjoint set of resources. If a path that is disjoint cannot be found, the value of the PAS diversity bit determines whether a path should be setup anyway. If the PAS diversity bit is clear, one can still attempt to setup the LSP. A PE should still attempt to minimize shared resources but that is an implementation issue, and is outside the scope of this document.

Optionally the CE may use a value of all zeros in the PAS identifier allowing the PE to select an appropriate PAS identifier. Also the PE may to override the PAS identifier allowing the PE to re-assign the identifier if required. A CE should not assume that the PAS identifier used for setup is the actual PAS identifier.

4.1.2.2. Error Handling Procedures

The PAS object must be understood by the PE device. Otherwise, the CE should not use the PAS object. Path Message processing of the PAS object SHOULD follow CTYPE 0. An Error code of IANA (TBD) indicates that the PAS object is not understood.

When a PAS identifier is not recognized by a PE it must assume this LSP defines that PAS identifier however the PE may override PAS identifier under certain conditions.

If the identifier is recognized but the Source Address-Destination address pair(s) are not recognized, this LSP must be set up using the PAS identifier only.

If the identifier is recognized and the Source Address-Destination address pair(s) are also recognized, then the PE SHOULD use the PE SRLG information associated with the LSPs identified by the address pairs to select a disjoint path.

The Following are the additional error codes:

1. Route Blocked by Exclude Route Value IANA (TBA).

4.1.2.3. Distribution of the Path Affinity Set Information

Information about SRLG is already available in the IGP TE database. A PE network can be designed to have additional opaque records for Provider paths that distribute PE paths and SRLG on a VPN basis. When a PE path is setup, the following information allows a PE to lookup the PE diversity information:

- o L1 VPN Identifier 8 bytes
- o Path Affinity Set Identifier
- o Source PE Address
- o Destination PE Address
- o List of PE SRLG (variable)

The source PE address and destination PE address are the same addresses in the VPN PIT and correspond to the respective CE address identifiers.

Note that all of the information is local to the PE context and is not shared with the CE. The VPN Identifier is associated with a CE. The only value that is signaled from the CE is the Path Affinity Set and optionally the addresses of an existing LSP. The PE stores source and destination PE addresses of the LSP in their native format along with the SRLG information. This information is internal to the PE network and is always known.

PE paths may be setup on demand or they may be pre-established. When paths are pre-established, the Path Affinity Set is set to unassigned 0x0000 and is ignored. When a CE uses a pre-established path the PE may set the Path SRLG Path Affinity Set value if the CE signals one otherwise the Path Affinity Set remains unassigned 0x0000.

5. Latency Signaling Extensions

Some network applications are sensitive to latency (sometimes also called delay) while other applications are sensitive to latency variation (sometimes also called delay variation). Specifically, real time applications typically do have certain latency requirements. It shall be noted that latency variation is typically not an issue for TDM networks including the WDM layer. For these technologies the latency is constant and there is no latency variation added. Latency variation is typically caused in packet networks or when packet based services are encapsulated into a constant bit rate server layer signal, which requires buffering of the arriving packets that may arrive in bursts. An example is an Ethernet VLAN service that is mapped into a constant bit rate server layer such as an ODUk or ODUflex OTN signal.

The GMPLS UNI as defined in [RFC4208] does not support latency as a signaling parameter that would allow a CE device to signal to the PE device that latency and/or latency variation constraints need to be met when a path is calculated for the requested LSP. The path computation function does typically calculate a route to the given destination that has the least TE metric (least cost routing). However, if a CE device requests an LSP via the UNI interface for an application that is sensitive to latency/latency variation, it should be possible to signal to the PE device that the objective function should rather take latency into account instead of the TE metric.

In order to support latency/latency variation as path computation constraint, the network has to support latency/latency variation as TE metric extension as defined in [DRAFT_OSPF TE METRIC EXT] - note that [DRAFT_OSPF TE METRIC EXT] is using the terms delay/delay variation instead of latency/latency variation.

A latency requirement can be added to signaling in the form of a

constraint [DRAFT OBJECTIVE FUNCTION]. The constraint can take the form of:

- o Minimal latency
- o Maximum acceptable latency (upper bound)
- o Minimal latency variation
- o Maximum acceptable latency variation (upper bound), if applicable

While some systems may be able to compute routes based on delay metrics it is usual that minimizing the accumulated TE link metric (link cost) or the number of hops subject to bandwidth reservation are satisfied as the object function and delay is not considered. When considering diversity latency falls after diversity constraints have been satisfied.

Recording the latency of existing paths [DRAFT_TE_METRIC RECORD] to ensure they meet a maximum acceptable latency can be utilized to ensure latency constraint is met.

When a low latency path is required, the minimize latency subject to other constraints criteria should be signaled. A CE device can use the recorded latency to ensure that the maximum acceptable latency has been met.

5.1. RSVP-TE Extensions

At the UNI, the RSVP-TE extensions as defined in [DRAFT OBJECTIVE FUNCTION] SHALL be used for signaling the PE device whether a path with minimal latency is requested or whether certain latency/latency variation upper bound constraints shall be met for the end-to-end connection, i.e., from the source CE device to the destination CE device. The following objective function (OF) code point SHALL be used in the OF sub-object of the ERO to indicate that latency/latency variation constraints SHALL be taken into account when the path computation function that is invoked by the PE node that expands the route from the PE device to the destination CE device:

- o OF code value 8 (to be assigned by IANA) is for the Minimum Latency Path (MLP) OF
- o OF code value 9 (to be assigned by IANA) is for Minimum Latency Variation Path (MLVP) OF

Additionally, an optional OF metric-bound sub-object MAY be carried within an ERO object of the RSVP-TE Path message. The two metric-

bound sub-objects defined in [DRAFT OBJECTIVE FUNCTION] that are corresponding to the two OFs above are:

- o metric bound sub-object of Type T=4: Cumulative Latency
- o metric bound sub-object of Type T=5: Cumulative Latency Variation

The metric-bound indicates an upper bound for the path metric that MUST NOT be exceeded for the ERO expending node to consider the computed path as acceptable. It shall be noted that the metric bound included in the RSVP-TE Path message at the UNI has end-to-end significance, which means that the upper bound metric constraint MUST be met for the path from the source CE device to the destination CE device.

5.2. Operational Procedures

The processing rules as defined in [DRAFT OBJECTIVE FUNCTION] for the OF sub-object and the optional OF metric-bound sub-object SHALL be applied at the ingress PE device when the source CE device requests an LSP (It shall be noted that [DRAFT OBJECTIVE FUNCTION] has a wider scope and may also apply to inter-domain interfaces, i.e., when the provider network is composed of multiple separate domains.).

5.3. Error Handling Procedures

The error handling rules as defined in [DRAFT OBJECTIVE FUNCTION] for the OF sub-object and the optional OF metric-bound sub-object SHALL be applied.

6. Security Considerations

Security for L1VPNs is covered in [RFC4847], [RFC5251] and [RFC5253]. In this document, the model follows a generic GMPLS VPN based on the L1VPN control plane model where CE addresses are completely distinct from the PE addresses.

The use of a private network assumes that entities outside the network cannot spoof or modify control plane communications between CE and PE. Furthermore, all entities in the private network are assumed to be trusted. Thus, no security mechanisms are required by the protocol exchanges described in this document.

However, an operator that is concerned about the security of their private control plane network may use the authentication and

integrity functions available in RSVP-TE [RFC3473] or utilize IPsec ([RFC4301], [RFC4302], [RFC4835], [RFC5996], and [RFC6071]) for the point-to-point signaling between PE and CE. See [RFC5920] for a full discussion of the security options available for the GMPLS control plane.

7. IANA Considerations

TBD

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.
- [RFC4202] Kompella, K., Rekhter, Y., "Routing Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4202, October 2005.
- [RFC4208] Swallow, G., Drake, J., Ishimatsu, H., and Y. Rekhter, "Generalized Multiprotocol Label Switching (GMPLS) User-Network Interface (UNI): Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Support for the Overlay Model", RFC 4208, October 2005.
- [RFC5251] Fedyk, D., Rekhter, Y., Editors "Layer 1 VPN Basic Mode", RFC 5251, July 2008.
- [SRLG_info] Zhang, F., Li, D., Gonzalez de Dios, O., Margaria, C., Hartley, M., "RSVP-TE Extensions for Collecting SRLG Information", draft-ietf-ccamp-rsvp-te-srlg-collect-02.txt, February 2013.
- [DRAFT OBJECTIVE FUNCTION] Ali, Z., Swallow, G., Filsfils, C., Fang, L., Kumaki, K., Kunze, R., "Resource ReserVation Protocol - Traffic Engineering (RSVP-TE) extension for signaling Objective Function and Metric Bound", draft-ali-ccamp-rc-

objective-function-metric-bound-02.txt, July 2012.

8.2. Informative References

- [RFC4026] Andersson, L. and T. Madsen, "Provider Provisioned Virtual Private Network (VPN) Terminology", RFC 4026, March 2005.
- [RFC6071] Frankel, S. and S. Krishnan, "IP Security (IPsec) and Internet Key Exchange (IKE) Document Roadmap", RFC 6071, February 2011.
- [RFC3473] Berger, L. (editor), "Generalized MPLS Signaling - RSVP-TE Extensions", RFC 3473, January 2003.
- [RFC4301] Kent, S. and K. Seo, "Security Architecture for the Internet Protocol", RFC 4301, December 2005.
- [RFC4302] Kent, S., "IP Authentication Header", RFC 4302, December 2005.
- [RFC5996] Kaufman, C., Hoffman, P., Nir, Y., and P. Eronen, "Internet Key Exchange Protocol Version 2 (IKEv2)", RFC 5996, September 2010.
- [RFC4835] Manral, V., "Cryptographic Algorithm Implementation Requirements for Encapsulating Security Payload (ESP) and Authentication Header (AH)", RFC 4835, April 2007.
- [RFC4847] Takeda, T., Editor "Framework and Requirements for Layer Virtual Private Networks", RFC 4847, April 2007.
- [RFC5253] Takeda, T., Ed., "Applicability Statement for Layer 1 Virtual Private Network (L1VPN) Basic Mode", RFC 5253, July 2008.
- [RFC5920] Fang, L., Ed., "Security Framework for MPLS and GMPLS Networks", RFC 5920, July 2010.

[DRAFT_TE_METRIC RECORD] Ali, Z., Swallow, G., Filsfils, C., Hartley, M., Kumaki, K., Kunze, R., "Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) extension for recording TE Metric of a Label Switched Path", draft-ietf-ccamp-te-metric-recording-02.txt, July 2013.

[DRAFT_OSPF TE METRIC EXT] Giacalone, S., Ward, D., Drake, J., Atlas, A., Previdi, S., "OSPF Traffic Engineering (TE) Metric Extensions", draft-ietf-ospf-te-metric-extensions-04.txt, June 2013.

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

Redistribution and use in source and binary forms, with or without modification, is permitted pursuant to, and subject to the license terms contained in, the Simplified BSD License set forth in Section 4.c of the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>).

Authors' Addresses

Don Fedyk
Alcatel-Lucent
Groton, MA, 01450
Email: donald.fedyk@alcatel-lucent.com

Dieter Beller
Alcatel-Lucent
Email: Dieter.Beller@alcatel-lucent.com

Lieven Levrau
Alcatel-Lucent
Email: Lieven.Levrau@alcatel-lucent.com

Daniele Ceccarelli
Ericsson
Email: Daniele.Ceccarelli@ericsson.com

Fatai Zhang
Huawei Technologies
Email: zhangfatai@huawei.com

Yuji Tochio
Fujitsu
Email: tochio@jp.fujitsu.com

Xihua Fu
ZTE
Email: fu.xihua@zte.com.cn

CCAMP Working Group
Internet-Draft
Intended status: Informational
Expires: October 25, 2014

Rakesh Gandhi, Ed.
Zafar Ali
Gabriele Maria Galimberti
Cisco Systems, Inc.
Xian Zhang
Huawei
April 23, 2014

RSVP-TE Signaling For GMPLS Restoration LSP
draft-gandhi-ccamp-gmpls-restoration-lsp-04

Abstract

In transport networks, there are requirements where Generalized Multi-Protocol Label Switching (GMPLS) end-to-end recovery scheme needs to employ restoration Label Switched Path (LSP) while keeping resources for the working and/or protecting LSPs reserved in the network after the failure.

This document reviews how the LSP association is to be provided using Resource Reservation Protocol - Traffic Engineering (RSVP-TE) signaling in the context of GMPLS end-to-end recovery when using restoration LSP where failed LSP is not torn down. No new procedures or mechanisms are defined by this document, and it is strictly informative in nature.

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."

Copyright Notice

Copyright (c) 2014 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
2. Signaling Restoration LSP Association	5
3. IANA Considerations	5
4. Security Considerations	5
5. Acknowledgement	5
6. References	6
6.1. Normative References	6
6.2. Informative References	6
Authors' Addresses	7

1. Introduction

Generalized Multi-Protocol Label Switching (GMPLS) [RFC3473] extends Multi-Protocol Label Switching (MPLS) to include support for different switching technologies. These switching technologies provide several protection schemes [RFC4426][RFC4427] (e.g., 1+1, 1:N and M:N). Resource Reservation Protocol - Traffic Engineering (RSVP-TE) signaling has been extended to support various GMPLS recovery schemes [RFC4872][RFC4873], to establish Label Switched Paths (LSPs), typically for working LSP and protecting LSP. [RFC4427] Section 7 specifies various schemes for GMPLS recovery.

In GMPLS recovery schemes generally considered, restoration LSP is signaled after the failure has been detected and notified on the working LSP. In non-revertive recovery mode, working LSP is assumed to be removed from the network before restoration LSP is signaled. For revertive recovery mode, a restoration LSP is signaled while working LSP and/or protecting LSP are not torn down in control plane due to a failure. In transport networks, as working LSPs are typically signaled over a nominal path, service providers would like to keep resources associated with the working LSPs reserved. This is to make sure that the service (working LSP) can use the nominal path when the failure is repaired to provide deterministic behaviour and guaranteed Service Level Agreement (SLA). Consequently, revertive recovery mode is usually preferred by recovery schemes used in transport networks.

As defined in [RFC4872] and being considered in this document, "fully dynamic rerouting switches normal traffic to an alternate LSP that is not even partially established only after the working LSP failure occurs. The new alternate route is selected at the LSP head-end node, it may reuse resources of the failed LSP at intermediate nodes and may include additional intermediate nodes and/or links."

One example of the recovery scheme considered in this document is 1+R recovery. The 1+R recovery is exemplified in Figure 1. In this example, working LSP on path A-B-C-Z is pre-established. Typically after a failure detection and notification on the working LSP, a second LSP on path A-H-I-J-Z is established as a restoration LSP. Unlike protection LSP, restoration LSP is signaled per need basis.

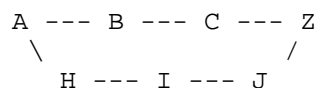


Figure 1: An example of 1+R recovery scheme

During failure switchover with 1+R recovery scheme, in general, working LSP resources are not released and working and restoration LSPs coexist in the network. Nonetheless, working and restoration LSPs can share network resources. Typically when failure is recovered on the working LSP, restoration LSP is no longer required and torn down (e.g., revertive mode).

Another example of the recovery scheme considered in this document is 1+1+R. In 1+1+R, a restoration LSP is signaled for the working LSP and/or the protecting LSP after the failure has been detected and notified on the working LSP or the protecting LSP. The 1+1+R recovery is exemplified in Figure 2. In this example, working LSP on path A-B-C-Z and protecting LSP on path A-D-E-F-Z are pre-established. After a failure detection and notification on a working LSP or protecting LSP, a third LSP on path A-H-I-J-Z is established as a restoration LSP. The restoration LSP in this case provides protection against a second order failure. Restoration LSP is torn down when the failure on the working or protecting LSP is repaired.

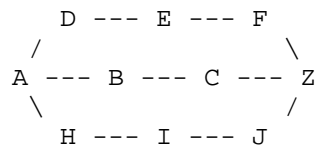


Figure 2: An example of 1+1+R recovery scheme

[RFC4872] Section 14 defines PROTECTION object for GMPLS recovery signaling. As defined, the PROTECTION object is used to identify primary and secondary LSPs using S bit and protecting and working LSPs using P bit. Furthermore, [RFC4872] defines the usage of ASSOCIATION object for associating GMPLS working and protecting LSPs.

[RFC6689] Section 2.2 reviews the procedure for providing LSP associations for GMPLS end-to-end recovery and covers the schemes where the failed working LSP and/or protecting LSP are torn down.

This document reviews how the LSP association is to be provided for GMPLS end-to-end recovery when using restoration LSP where working and protecting LSP resources are kept reserved in the network after the failure.

2. Signaling Restoration LSP Association

Where GMPLS end-to-end recovery scheme needs to employ restoration LSP while keeping resources for the working and/or protecting LSPs reserved in the network after the failure, restoration LSP is signaled with ASSOCIATION object with the association ID set to the LSP ID of the LSP it is restoring. For example, when a restoration LSP is signaled for a working LSP, the ASSOCIATION object in the restoration LSP contains the association ID set to the LSP ID of the working LSP. Similarly, when a restoration LSP is signaled for a protecting LSP, the ASSOCIATION object in the restoration LSP contains the association ID set to the LSP ID of the protecting LSP.

The procedure for signaling the PROTECTION object is specified in [RFC4872]. Specifically, restoration LSP being used as a working LSP is signaled with P bit cleared and being used as a protecting LSP is signaled with P bit set.

As discussed in Section 1 of this document, [RFC6689] Section 2.2 reviews the procedure for providing LSP associations for the GMPLS end-to-end recovery scheme using restoration LSP where the failed working LSP and/or protecting LSP are torn down.

3. IANA Considerations

This document makes no request for IANA action.

4. Security Considerations

This document reviews procedures defined in [RFC4872] and [RFC6689] and does not define any new procedure. As such, no new security considerations are introduced in this document.

5. Acknowledgement

The authors would like to thank George Swallow for the discussions on the GMPLS restoration.

6. References

6.1. Normative References

- [RFC3473] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.
- [RFC4872] Lang, J., Rekhter, Y., and Papadimitriou, D., "RSVP-TE Extensions in Support of End-to-End Generalized Multi-Protocol Label Switching (GMPLS) Recovery", RFC 4872, May 2007.
- [RFC6689] Berger, L., "Usage of the RSVP ASSOCIATION Object", RFC 6689, July 2012.

6.2. Informative References

- [RFC4426] Lang, J., Rajagopalan, B., and Papadimitriou, D., "Generalized Multiprotocol Label Switching (GMPLS) Recovery Functional Specification", RFC 4426, March 2006.
- [RFC4427] Mannie, E., and Papadimitriou, D., "Recovery (Protection and Restoration) Terminology for Generalized Multi-Protocol Label Switching", RFC 4427, March 2006.
- [RFC4873] Berger, L., Bryskin, I., Papadimitriou, D., and Farrel, A., "GMPLS Segment Recovery", RFC 4873, May 2007.

Authors' Addresses

Rakesh Gandhi (editor)
Cisco Systems, Inc.

Email: rgandhi@cisco.com

Zafar Ali
Cisco Systems, Inc.

Email: zali@cisco.com

Gabriele Maria Galimberti
Cisco Systems, Inc.

Email: ggalimbe@cisco.com

Xian Zhang
Huawei Technologies
Research Area F3-1B,
Huawei Industrial Base,
Shenzhen, 518129, China

Email: zhang.xian@huawei.com

CCAMP Working Group
Internet Draft
Intended status: Standards Track
Expires: April 20, 2014

Matt Hartley
Zafar Ali
Cisco Systems
O. Gonzalez de Dios
Telefonica I+D
C. Margaria
Coriant R&D GmbH
October 21, 2013

RSVP-TE extensions for RRO editing
draft-hartley-ccamp-rro-editing-00.txt

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 20, 2014.

Copyright Notice

Copyright (c) 2013 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.

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.

Hartley, Ali, at al

Expires April 21, 2014

[Page 1]

Abstract

This document provides extensions for the Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) to allow the communication of changes made by a node to the information provided by other nodes in a ROUTE_RECORD Object (RRO) in Path and Resv messages, or to indicate that it has itself provided incomplete information.

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 [RFC-2119].

Table of Contents

1. Introduction.....	2
1.1. Use Cases.....	3
1.1.1. Overlay and inter-domain networks.....	3
1.1.2. RRO reduction.....	3
2. RSVP-TE signaling extensions.....	3
2.1. IPv4 RRO-edit RRO sub-object.....	3
2.2. IPv6 RRO-edit RRO sub-object.....	4
2.3. RRO-edit sub-object Processing Rules.....	4
3. Security Considerations.....	5
4. IANA Considerations.....	5
4.1. ROUTE_RECORD Object.....	5
5. Acknowledgments.....	6
5.1. Normative References.....	7
5.2. Informative References.....	7
Author's Addresses.....	8
Disclaimer of Validity.....	8

1. Introduction

The signaling process of a Label-Switched Path (LSP) may require gathering information of the actual path traversed by the LSP. The procedure for collecting this information includes the hop-by-hop construction of a Record Route Object (RRO) in the Path and Resv messages, containing information on the path traversed by the LSP ([RFC-3209], [RFC-3473], [RFC-4873], [RFC-5420], [RFC-5553], [DRAFT-SRLG], [DRAFT-METRIC]). There are several use cases, described in this document, in which one or more nodes on the path of an LSP may require that the RRO in the Path and/or Resv be edited to remove or summarize data contained in the RRO. However, it is important for the ingress or egress nodes to know what RRO subobjects have been edited by intermediate nodes. This document addresses this requirement.

1.1. Use Cases

Use cases where RRO editing can take place are described in this subsection.

1.1.1. Overlay and inter-domain networks

In the GMPLS overlay model there is a client-server relationship [RFC4208]. GMPLS User-Network Interface (UNI) is the reference point where policies can be applied. In this cases policy at the server network boundary may require that some or all information related to the server network be edited, summarized or removed when communicating with the client nodes. Similar policy requirements exist for inter-domain LSPs and in E-NNI use case.

1.1.2. RRO reduction

If an LSP with many hops is signaled and a great deal of information is collected at each hop, it is possible that the RRO may grow to the point where it reaches its maximum possible size or RSVP packet fragmentation becomes a problem. In this case a node may summarize or remove information from the RRO to reduce its size.

2. RSVP-TE signaling extensions

This section describes the signaling extensions required to address the aforementioned requirements. Specifically, the requirements are addressed by defining a new RRO sub-object that can be used to reference what information in RRO has been edited, as detailed in the following.

2.1. IPv4 RRO-edit RRO sub-object

A new RRO sub-object is defined in order to indicate that another RRO sub-object within the same hop has been edited.

[illegible]

+-----+

The sub-object fields are defined as follows:

Type: The sub-object type, to be assigned by IANA (suggested value: TBD).

Length: the total length of the TLV, in bytes. It MUST be 8.

Edited type: the type of the sub-object within the same hop to which the flags in this sub-object apply.

E (Edited) bit: When set, this bit indicates that the specified RRO sub-object has been edited in some way.

P (Partial) bit: When set, this bit indicates that the data contained in the specified RRO sub-object is incomplete.

S (Summary) bit: When set, this bit indicates that the data contained in the specified RRO sub-object has been summarized.

R (Removed) bit. When set, this bit indicates that the specified RRO sub-object has been removed entirely.

Reserved: This field SHOULD be set to zero on transmission, and MUST be ignored on receipt.

Editing node address: an IPv4 address unique to the node that has edited the RRO and inserted this sub-object.

2.2. IPv6 RRO-edit RRO sub-object

To be added in future revision.

2.3. RRO-edit sub-object Processing Rules

The processing rules in this section apply to the processing of both Path and Resv RROs.

Normal RRO processing involves a node simply adding data related to the local hop to the RRO received from the prior node to RRO, and placing the new RRO in the message to be transmitted. In this case the transmitted RRO contains all data that was present in the received RRO.

If a node edits the data in the received RRO such that the same data is not present in the transmitted RRO, or if it is supplying

incomplete or summarized data on its own behalf, then the following rules apply at the processing node.

- . For each sub-object type that has been edited within a hop, a RRO-edited sub-object SHOULD be inserted into the same hop in the RRO. The RRO-edited sub-object MAY be omitted entirely if the processing node's policy prevents communication of this information.
- . Multiple RRO-edited sub-objects describing edits to the same type of sub-object (i.e. with the same "Edited type" field) SHOULD NOT be added in the same hop.
- . Multiple RRO-edited sub-objects describing edits to the same type of sub-object (i.e. with the same "Edited type" field) MAY be added to different hops if appropriate.
- . The node SHOULD add its own local address to the "editing node address" field of the RRO-edited sub-object. This field MAY be set to zero if the processing node's policy prevents self-identification.
- . The node SHOULD set the appropriate bits in the flags field to indicate the changes that have been made to the subsequent RRO sub-object.
- . A node SHOULD NOT insert a RRO-edited sub-object with all flags set to zero.
- . Unassigned flag bits are considered reserved. They SHOULD be set to zero.

The following rules apply at a node processing a received RRO-edited sub-object:

- . Any set flag whose meaning is either unassigned or not understood SHOULD be ignored.
- . If an RRO is received with multiple RRO-edited sub-objects describing edits to the same type of sub-object within the same hop, the second and subsequent RRO-edited sub-objects SHOULD be ignored.

3. Security Considerations

To be added in a future version.

4. IANA Considerations

4.1. ROUTE_RECORD Object

IANA has made the following assignments in the "Class Names, Class Numbers, and Class Types" section of the "RSVP PARAMETERS" registry located at <http://www.iana.org/assignments/rsvp-parameters>. It is

requested that IANA make assignments from the ROUTE_RECORD RFC 3209 [RFC3209] portions of this registry.

This document introduces a new RRO sub-object:

Type	Name	Reference
-----	-----	-----
TBD	RRO-edited sub-object	This I-D

5. Acknowledgments

The authors would like to thank Lou Berger for suggesting the core idea described in this draft. The authors would also like to thank George Swallow for his input.

References

5.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.

5.2. Informative References

- [RFC3473] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.
- [RFC4873] Berger, L., Bryskin, I., Papadimitriou, D., and A. Farrel, "GMPLS Segment Recovery", RFC 4873, May 2007.
- [RFC5420] Farrel, A., Ed., Papadimitriou, D., Vasseur, JP., and A. Ayyangarps, "Encoding of Attributes for MPLS LSP Establishment Using Resource Reservation Protocol Traffic Engineering (RSVP-TE)", RFC 5420, February 2009.
- [RFC5553] Farrel, A., Ed., Bradford, R., Vasseur, JP., "Resource Reservation Protocol (RSVP) Extensions for Path Key Support", RFC 5553, May 2009.
- [DRAFT-SRLG] Zhang, F., Li, D., Gonzalez de Dios, O., Margaria, C., Hartley, M., "RSVP-TE Extensions for Collecting SRLG Information", draft-ietf-ccamp-rsvp-te-srlg-collect-03, October 2013.
- [DRAFT-METRIC] Ali, Z., Swallow, G., Filsfils, C., Hartley, M., Kumaki, K., Kunze, R., "Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) extension for recording TE Metric of a Label Switched Path", draft-ietf-ccamp-te-metric-recording-02, July 2013.

Author's Addresses

Matt Hartley
Cisco Systems
Email: mhartley@cisco.com

Zafar Ali
Cisco Systems
Email: zali@cisco.com

Oscar Gonzalez de Dios
Telefonica I+D
Email: ogondio@tid.es

Cyril Margaria
Email: cyril.margaria@gmail.com

Network Working Group
Internet Draft
Intended status: Standard Track
Expires: April 2014

Iftekhar Hussain
Zhong Pan
Marco Sosa
Infinera

Bert Basch
Steve Liu
Andrew G. Malis
Verizon Communications

Abinder Dhillon
Fujitsu Network Communications

October 8, 2013

Generalized Label for Super-Channel Assignment on Flexible Grid
draft-hussain-ccamp-super-channel-label-06.txt

Abstract

To enable scaling of existing transport systems to ultra high data rates of 1 Tbps and beyond, next generation systems providing super-channel switching capability are currently being developed. To allow efficient allocation of optical spectral bandwidth for such high bit rate systems, International Telecommunication Union Telecommunication Standardization Sector (ITU-T) is extending the G.694.1 grid standard (termed "Fixed-Grid") to include flexible grid (termed "Flex-Grid") support (draft revised ITU-T G.694.1, revision 1.4, Oct 2011). This necessitates definition of new label format for the Flex-Grid. This document defines a super-channel label as a Super-Channel Identifier and an associated list of 12.5 GHz slices representing the optical spectrum of the super-channel. The label information can be encoded using a fixed length or variable length format. This label format can be used in GMPLS signaling and routing protocol to establish super-channel based optical label switched paths (LSPs).

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), 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 8, 2014.

Copyright Notice

Copyright (c) 2013 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
2. Terminology.....	6
3. Motivation for Super-Channel Label.....	6
3.1. Flex-Grid Slice Numbering.....	6
3.2. Super-Channel Label.....	7
3.2.1. Super-Channel Label Encoding Format.....	8
3.2.2. LSP Encoding Type, Switching Type, and Generalized-PID (G-PID) in Generalized Label Request.....	11
4. Security Considerations.....	11

5. IANA Considerations.....	12
6. References.....	12
6.1. Normative References.....	12
6.2. Informative References.....	12
7. Acknowledgments.....	13
Appendix A. Super-Channel Label Format Example.....	14

1. Introduction

Future transport systems are expected to support service upgrades to data rates of 1 Tbps and beyond. To scale networks beyond 100Gbps, multi-carrier super-channels coupled with advanced multi-level modulation formats and flexible channel spectrum bandwidth allocation schemes have become pivotal for future spectral efficient transport network architectures [1,2].

A super-channel represents an ultra high aggregate capacity channel containing multiple carriers which are co-routed through the network as a single entity from the source transceiver to the sink transceiver [3,7]. By multiplexing multiple carriers, modulating each carrier with multi-level advanced modulation formats (such as PM-QPSK, PM-8QAM, PM-16QAM), allocating an appropriate-sized flexible channel spectral bandwidth slot, and using a coherent receiver for detecting closely packed sub-carriers, a super-channel can support ultra high data rates in a spectrally efficient manner while maintaining required system reach. Figure 1 contrasts channel spectrum bandwidth allocation schemes for various bit rate optical paths on fixed-grid and flex-grid. ITU-T fixed-grid permits allocation of channel spectrum bandwidth in "single" fixed-sized slots (e.g., 50GHz, 100GHz etc) independent of the channel bit rate. In contrast, a flex-grid can allocate "arbitrary" size channel spectral bandwidth as an integer multiple of 12.5 GHz fine granularity slices. This means, a flex-grid can support multiple data rates channels (optical paths) in a spectrally efficient manner as it allocates appropriate-sized spectrum bandwidth slots, as opposed to fixed-sized slots. As in the examples in the figure, the optical spectrum slices assigned will be to a given super-channel in a contiguous manner. However, for flexibility in finding available optical spectrum on fragmented fibers and to reduce signaling message overhead, the two schemes proposed in this document also allow for identification of a split-spectrum super-channel with optical spectral slices that are non-contiguous, spread across multiple slots. Note that the channel capacity available on a given number of optical spectral slices depends on (among other factors) how many contiguous optical slots are used. The definition of the channel capacity available for a split-spectrum super-channel split

across multiple slots of different widths is outside the scope of this document.

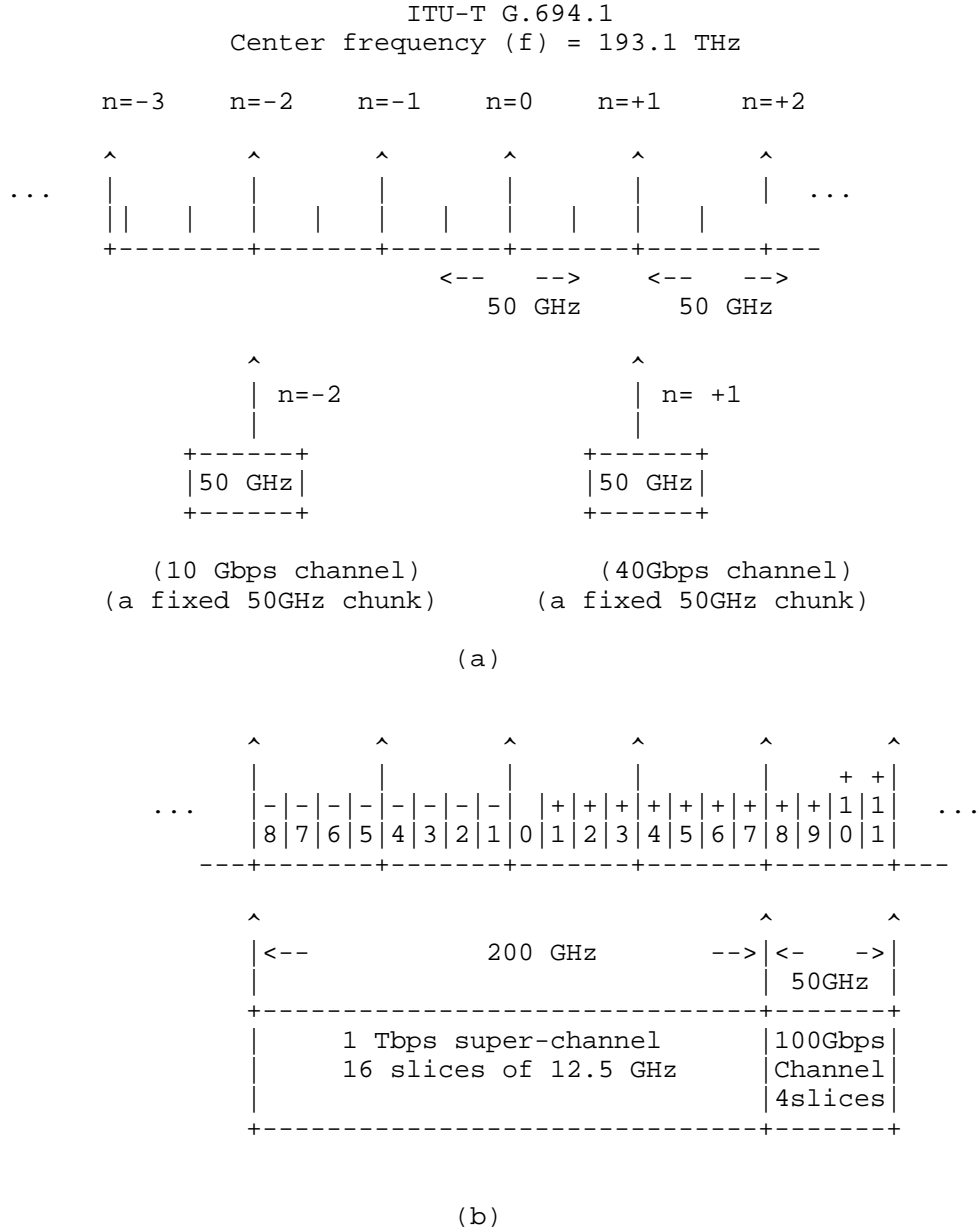


Figure 1 ITU-T (a) 50 GHz fixed-grid (G.694.1) (b) 12.5 GHz granular flex-grid

2. Terminology

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].

3. Motivation for Super-Channel Label

[RFC3471] defines new forms of MPLS "label" for the optical domain that are collectively referred to as a "generalized label". [RFC6205] defines a standard wavelength label based on ITU-T fixed-grids ([G.694.1] and [G.694.2]) for use by Lambda-Switch-Capable (LSC) LSRs.

A new label format for super-channels assignment on flex-grid is needed because the existing label formats (such as the waveband switching label defined in RFC3471 and the wavelength label defined in RFC6205) either lack necessary fields to carry required flex-grid related information (e.g., channel spacing) or do not allow signaling of arbitrary flexible-size optical spectral bandwidth in an efficient manner (e.g., in terms of integer multiple of fine granularity 12.5GHz slices). For example,

- o Waveband switching label format (defined in section 3.3.1 of RFC3471) lacks fields to carry necessary information to support flex-grid.
- o Wavelength label allows signaling of single fixed-size optical spectrum bandwidth slot only.
- o Wavelength label does not allow signaling of arbitrary flexible-size optical spectrum bandwidth needed for super-channels assignment on flex-grid.

3.1. Flex-Grid Slice Numbering

Given a slice spacing value (e.g., 0.0125 THz) and a slice number "n", the slice left edge frequency can be calculated as follows:

Slice Left Edge Frequency(THz)= 193.1 THz + n*slice spacing (THz).

Where "n" is a two's-complement integer (i.e., positive, negative, or 0) and "slice spacing" is 0.0125 THz conforming to ITU-T Flex-

3.2.1. Super-Channel Label Encoding Format

This section describes two options (option A and B) for encoding the super-channel label by making extensions to the waveband switching label[RFC3471] and wavelength label[RFC6205] formats. (Editor's Note: the term super-channel is a placeholder until a new term is defined for this entity).

- o Option A: Encode a super-channel label containing N frequency slots as a list of N entries in the form of (n, m) , where n is an integer that defines the nominal central frequency of the frequency slot and m is a positive integer that defines the slot width in accordance with the G.694.1. Other than the encoding of frequency slots (i.e., list of (n, m) in option A vs. list of (start, end) in option B) all other fields are identical in Option A and B.
- o Option B: Encode super-channel label as a list of start and end slice numbers corresponding to N slots, each consisting of contiguous slices with each slot denoted by its starting and ending slice number (e.g., "n_start_1" and "n_end_1" represent contiguous slices in slot#1, "n_start 2" and "n_end 2" in slot#2, ..., "n_start N" and "n_end N" in slot#N).

```

      0               1               2               3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|      Super-Channel Id(16-bit) |Grid | S.S.  | Reserved (9-bit)|
+-----+-----+-----+-----+-----+-----+-----+-----+
| Reserved (16-bit)           | Number of Entries(16-bit)      |
+-----+-----+-----+-----+-----+-----+-----+-----+
|n_start_1(contiguous slot #1) | n_end_1(contiguous slot #1) |
+-----+-----+-----+-----+-----+-----+-----+-----+
|n_start_2(contiguous slot#2)  | n_end_2(contiguous slot#2)  |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                               |                               |
|                               |                               |
|                               |                               |
+-----+-----+-----+-----+-----+-----+-----+-----+
|n_start_N (contiguous slot#N) | n_end_N (contiguous slot #N |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

Super-Channel Id: 16 bits

This field represents a logical identifier for a super-channel or split-spectrum super-channel. To disambiguate waveband switching and super-channel label applications, we propose to rename the Waveband Identifier (32-bit) as a Super-Channel Identifier (16-bit).

Grid: 3 bits

This field indicates the Grid type. The value for Grid should be set to xx (to be assigned by IANA) for the ITU-T flex-grid.

Grid	Value
Reserved	0
ITU-T DWDM	1
ITU-T CWDM	2
ITU-T Flex-Grid	xx (TBD)
Future use	3 - 7

S.S. (slice spacing): 4 bits

This field should be set to a value of 4 to indicate 12.5 GHz in both labels.

S.S. (GHz)	Value
Reserved	0
100	1
50	2
25	3
12.5	4
Future use	5 - 15

Number of Entries: 16-bit

This field represents the number of 32-bit entries in the super-channel label (i.e., number of slots with contiguous slices). For example, in the case of a super-channel with contiguous optical spectrum, this field should have a value of 1 (indicating one slot of contiguous slices).

n_start_i ($i=1,2,\dots,N$): 16 bits

n_end_i ($i=1,2,\dots,N$): 16 bits

A super-channel with contiguous spectrum or a split-spectrum super-channel with non-contiguous optical spectrum can be represented by N slots of slices where two adjacent slots can be contiguous or non-contiguous, however each slot contains contiguous slices. Each slot is denoted by n_start_i (which indicates the lowest or starting 12.5 GHz slice number of the slot) and n_end_i (which indicates the highest or ending 12.5 GHz slice number of the slot). " n_start_i " and " n_end_i " are two's-complement integers that can take either a positive, negative, or zero value.

- o Option C: Encode super-channel label as a first slice number of the grid (denoted as " n_start of Grid") plus the entire list of slices in the grid as a Bitmap

0																1																2																3															
0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9																								
Super-Channel Id (16-bit)																Grid																S.S.																Reserved (9-bit)															
n_start of Grid (16-bit)																Num of Slices in Grid (16-bit)																																															
Bitmap Word #1(first set of 32 slices from the left most edge)																																																															
Bitmap Word #2 (next set of 32 contiguous slice numbers)																																																															
...																																																															
Bitmap Word #N(last set of 32 contiguous slice numbers)																																																															

Where:

Super-Channel Id, Grid, and S.S fields are same as described earlier in option B.

n_start of Grid: 16-bit

This field indicates the first slice number in Grid for the band being referenced (i.e., the start of the left most edge of the Grid).

Num of Slices in Grid: 16-bit

This field represents the total number of slices in the band. The value in this field determines the number of 32-bitmap words required for the grid.

Bit map (Word): 32-bit

Each bit in the 32-bitmap word represents a particular slice with a value of 1 or 0 to indicate whether for that slice reservation is required (1) or not (0). Bit position zero in the first word represents the first slice in the band (Grid) and corresponds to the value indicated in the "n_start of Grid" field.

All three options allow efficient encoding of a super-channel label with contiguous and non-contiguous slices. Option C yields a fixed length format while option A and B, a variable length format. Option C is relatively simpler, more flexible, however, might be less compact than option A and B for encoding a single super-channel with contiguous optical spectrum. In contrast, option A and B provide a very compact representation for super-channels with contiguous optical spectrum, however, might be less flexible in encoding split-spectrum super-channels with arbitrary non-contiguous set of slices.

3.2.2. LSP Encoding Type, Switching Type, and Generalized-PID (G-PID) in Generalized Label Request

For requesting a super-channel label in a Generalized Label Request defined in section 3.1.1 of RFC3471, this document proposes to use LSP Encoding Type = Lambda (as defined in RFC4328), Switching Type = Super-Channel-Switch-Capable(SCSC) (as defined in [6]), and a new G-PID type = OTUadaptand a new G-PID value (similar to as defined in section 3.1.3 of RFC4328) to be assigned by IANA.

4. Security Considerations

<Add any security considerations>

5. IANA Considerations

IANA needs to assign a new Grid field value to represent ITU-T Flex-Grid and a new G-PID value.

6. References

6.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3471] Berger, L., Ed., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, January 2003.
- [RFC6205] Otani, T., Ed., "Generalized Labels for Lambda-Switch-Capable (LSC) Label Switching Routers", RFC 6205, March 2011.
- [RFC6163] Lee, Y., Ed., "Framework for GMPLS and Path Computation Element (PCE) Control of Wavelength Switched Optical Networks (WSNs)", RFC 6163, April 2011

6.2. Informative References

- [1] Gringeri, S., Basch, B. Shukla, V. Egorov, R. and Tiejun J. Xia, "Flexible Architectures for Optical Transport Nodes and Networks", IEEE Communications Magazine, July 2010, pp. 40-50
- [2] M. Jinno et. al., "Spectrum-Efficient and Scalable Elastic Optical Path Network: Architecture, Benefits and Enabling Technologies", IEEE Comm. Mag., Nov. 2009, pp. 66-73.
- [3] S. Chandrasekhar and X. Liu, "Terabit Super-Channels for High Spectral Efficiency Transmission", in Proc. ECOC 2010, paper Tu.3.C.5, Torino (Italy), September 2010.
- [4] ITU-T Recommendation G.694.1, "Spectral grids for WDM applications: DWDM frequency grid", June 2002
- [5] [4] "Finisar to Demonstrate Flexgrid(TM) WSS Technology at ECOC 2010", press release.
- [6] Abinder D., et. al., "OSPFTE extension to support GMPLS for Flex Grid", draft-dhillon-ccamp-super-channel-ospfte-ext, work in progress, October 2011.

- [7] Sharfuddin S., et. al., "A Framework for control of Flex Grid Networks", draft-syed-ccamp-flexgrid-framework-ext, work in progress, March 2012.

7. Acknowledgments

<Add any acknowledgements>

Appendix A.

Super-Channel Label Format Example

Suppose node A and Node Z are super-channel switching capable and node A receives a request for establishing a 1 Tbps optical LSP from itself to node Z. Assume the super-channel requires a "contiguous" spectral bandwidth of 200 GHz with left-edge frequency of 191.475 THz for the left-most 12.5 GHz slice and left-edge frequency of 191.6625 THz for the right-most slice. This means $n_start = (191.475 - 193.1)/0.0125 = -130$ and $n_end = (191.6625 - 193.1)/0.0125 = -115$ (i.e. we need 16 slices of 12.5 GHz starting from slice number -130 and ending at slice number -115).

Node A signals the LSP via a Path message including a super-channel label format encoding option B defined in section 3.3:

0																1																2																3															
0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1																																
Super-Channel Id (16-bit)																Grid				S.S.				Reserved (9-bit)																																							
Reserved (16-bit)																Number of Entries (16-bit)																																															
n_start_1 (contiguous slot #1)																n_end_1(contiguous slot#1)																																															

Where:

Super-Channel Id = 1 :super-channel number 1

Number of Entries: 1

Grid = xx : ITU-T Flex-Grid

S.S. = 4 : 12.5 GHz Slice Spacing

n_start_1 = -130 : left-most 12.5 GHz slice number for slot 1

n_end_1= -115 : Right-most 12.5 GHz slice number for slot 1

Authors' Addresses

Iftekhar Hussain
Infinera
140 Caspian Ct., Sunnyvale, CA 94089

Email: ihussain@infinera.com

Zhong Pan
Infinera
140 Caspian Ct., Sunnyvale, CA 94089

Email: zpan@infinera.com

Marco Sosa
Infinera
140 Caspian Ct., Sunnyvale, CA 94089

Email: msosa@infinera.com

BertBasch
Verizon Communications
60Sylvan Rd., Waltham, MA02451

Email: bert.e.basch@verizon.com

SteveLiu
Verizon Communications
60Sylvan Rd., Waltham, MA02451

Email: steve.liu@verizon.com

Andrew G. Malis
Verizon Communications
60Sylvan Rd., Waltham, MA02451

Email: andrew.g.malis@verizon.com

Abinder Dhillon
Fujitsu Network Communications
2801 Telecom Parkway, Richardson, TX 75082

Email: abinder.dhillon@us.fujitsu.com

Contributor's Addresses

Rajan Rao
Infinera
140 Caspian Ct., Sunnyvale, CA 94089

Email: rrao@infinera.com

Biao Lu
Infinera
140 Caspian Ct., Sunnyvale, CA 94089

Email: blu@infinera.com

Subhendu Chattopadhyay
Infinera
140 Caspian Ct., Sunnyvale, CA 94089

Email: schattopadhyay@infinera.com

Harpreet Uppal
Infinera
140 Caspian Ct., Sunnyvale, CA 94089

Email: harpreet.uppal@infinera.com

Network Working Group
Internet Draft
Intended status: Standard Track
Expires: April 2014

Iftekhar Hussain
Marco Sosa
Infinera

Abinder Dhillon
Fujitsu Network Communications

October 8, 2013

Super-Channel Optical Parameters GMPLS Routing Extensions
draft-hussain-ccamp-super-channel-param-ospfte-03.txt

Abstract

This document builds on [6][7] and defines GMPLS routing extensions to allow added CSPF constraints for efficient super-channel spectrum assignment on flexible grid networks.

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), 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 8, 2014.

Copyright Notice

Copyright (c) 2013 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.....	23
2. Terminology.....	3
3. GMPLS Routing Extensions for Super-Channel Optical Parameters.....	34
3.1. Super-Channel In-Use Slices sub-TLV.....	45
3.2. Super-Channel Carriers sub-TLV.....	5
4. Procedure for OSPF-TE Advertisement.....	56
5. Possible Applications.....	6
6. TLV Encoding Examples.....	6
7. Security Considerations.....	67
8. IANA Considerations.....	67
9. References.....	7
9.1. Normative References.....	7
9.2. Informative References.....	7
10. Acknowledgments.....	8

1. Introduction

Future transport systems are expected to support service upgrades to data rates of 1 Tbps and beyond. To scale networks beyond 100Gbps, multi-carrier super-channels coupled with advanced multi-level modulation formats and flexible channel spectrum bandwidth

allocation schemes have become pivotal for future spectral efficient transport network architectures [1,2].

The coexistence of super-channels using different modulation formats on the same optical fiber network infrastructure may have a detrimental effect on the Optical Signal to Noise Ratio (OSNR) of adjacent super-channels due to interference such as cross-phase modulation. Therefore, it may be highly desirable to be able to evaluate the mutual impact of the existing and new super-channels on each other's quality of transmission (e.g., bit error rate) before establishing new super-channels.

The document [9] defines GMPLS signaling extensions to convey super-channel optical parameters. This document defines GMPLS routing extensions to advertise the above mentioned super-channel parameters via OSPF-TE link LSA using new Super-Channel sub-TLV. This sub-TLV is carried under the Switching Capability-specific information (SCSI) field of the Interface Switching Capability Descriptor (ISCD) with the Super-Channel-Switch-Capable (SCSC) value defined in [6]. This information allows each source node across the network to apply added CSPF constraints and assign new super-channels spectrum efficiently by considering not only the availability of the required number of slices but also the optical signal compatibility of the existing and the new super-channels along the desired path.

2. Terminology

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].

3. GMPLS Routing Extensions for Super-Channel Optical Parameters

This document defines OSPF-TE extensions for advertising following information using the Super-Channel sub-TLV depicted in Figure 1. For each super-channel this sub-TLV advertises following information:

- o Super-Channel In-Use Slices sub-TLV
- o Super-Channel Carriers sub-TLV

0									1									2									3				
0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1

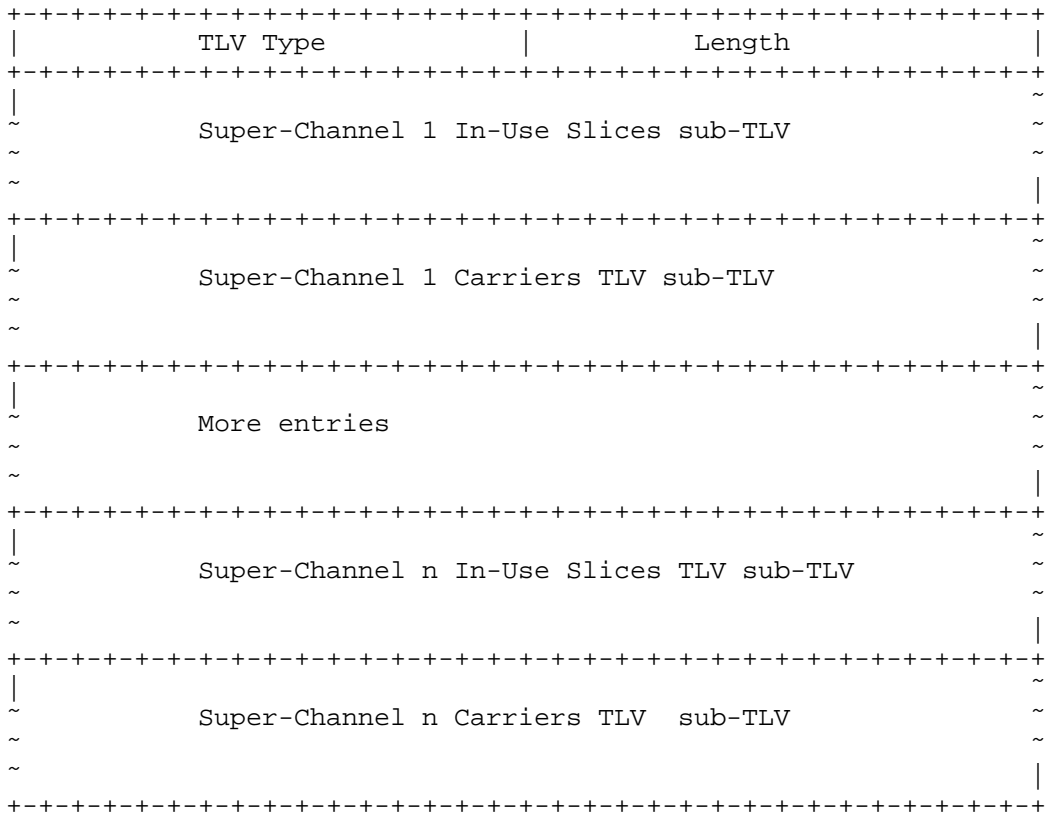
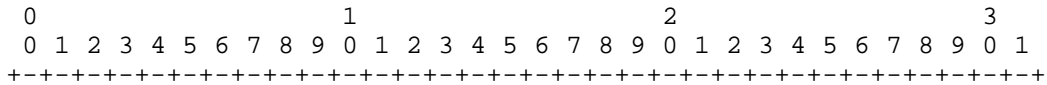


Figure 1: Super-Channel TLV Format.

The Super-Channel sub-TLV is advertised in the OSPF-TE link LSA under the under the SCSI field of the ISCD using Super-Channel-Switch-Capable (SCSC) value defined in [6]

3.1. Super-Channel In-Use Slices sub-TLV

This sub-TLV contains the in-use slices information of a super-channel. For further information about various fields in this sub-TLV refer to [6][7].



TLV Type	Length
Super-Channel Id	Grid S.S. PRI Reserved
n_start_1 (spectral slot 1)	n_end_1 (spectral slot 1)
n_start_2 (spectral slot 2)	n_end_2 (spectral slot 2)
~	~
More entries	~
~	~
n_start_n (spectral slot n)	n_end_n (spectral slot n)

Figure 2: Super-Channel In-Use Slices sub-TLV Format.

[Editor's Note: encoding of in-use slices in bitmap format is left for a possible future revision]

3.2. Super-Channel Carriers sub-TLV

The format of the Super-Channel Carriers sub-TLV is defined in [9]. In summary, this sub-TLV contains following information.

- o Number of Carriers in the Super-Channel
- o Carrier sub-TLV
 - o Carrier Center Frequency sub-sub-TLV
 - o Carrier Modulation sub-sub-TLV
 - o Carrier FEC sub-sub-TLV

4. Procedure for OSPF-TE Advertisement

This section describes procedure for advertising the aforementioned information in the OSPF-TE link LSAs.

- o The optical parameters of the super-channel are signaled when new super-channels are established (see [9]).

- o Over time change in the status of in-use slices occurs when new super-channels are setup (or when established super-channels are released).
- o Each node along the path traversed by the super-channels advertises the current status of the in-use slices for each super-channel in the OSPF-TE link LSA using sub-TLVs described earlier.
- o Through OSPF-TE LSAs flooding other nodes in the routing domain learn about the current status of in-use slices on each TE link.

5. Possible Applications

- o The presence of this information across the network topology enables source nodes in the network to apply added CSPF constraints for example to:
 - o Group super-channels with different modulation formats in different bands (slice ranges)
 - o Group super-channels with same bit-rate in a band while separating with guard band from super-channels with different bit-rate.
- o Allows efficient network utilization (e.g., reduces new requests blocking probability) by avoiding excessive worst-case OSNR penalty while preserving desired quality of transmission of the existing super-channels

6. TLV Encoding Examples

To be added later.

7. Security Considerations

<Add any security considerations>

8. IANA Considerations

IANA needs to assign a new Grid field value to represent ITU-T Flex-Grid.

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.
- [RFC3471] Berger, L., Ed., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, January 2003.
- [RFC6205] Otani, T., Ed., "Generalized Labels for Lambda-Switch-Capable (LSC) Label Switching Routers", RFC 6205, March 2011.
- [RFC6163] Lee, Y., Ed., "Framework for GMPLS and Path Computation Element (PCE) Control of Wavelength Switched Optical Networks (WSNs)", RFC 6163, April 2011

9.2. Informative References

- [1] Gringeri, S., Basch, B. Shukla,V. Egorov, R. and Tiejun J. Xia, "Flexible Architectures for Optical Transport Nodes and Networks", IEEE Communications Magazine, July 2010, pp. 40-50
- [2] M. Jinno et. al., "Spectrum-Efficient and Scalable Elastic Optical Path Network: Architecture, Benefits and Enabling Technologies", IEEE Comm. Mag., Nov. 2009, pp. 66-73.
- [3] S. Chandrasekhar and X. Liu, "Terabit Super-Channels for High Spectral Efficiency Transmission",in Proc. ECOC 2010, paper Tu.3.C.5, Torino (Italy), September 2010.
- [4] ITU-T Recommendation G.694.1, "Spectral grids for WDM applications: DWDM frequency grid", June 2002
- [5] [4] "Finisar to Demonstrate Flexgrid(TM) WSS Technology at ECOC 2010", press release.
- [6] Abinder D., et.al., "OSPFTE extension to support GMPLS for Flex Grid", draft-dhillon-ccamp-super-channel-ospfte-ext, work in progress, November 2011.
- [7] Iftekhar H., et.al., "Generalized Label for Super-Channel Assignment on Flexible Grid", draft-hussain-ccamp-super-channel-label, work in progress, October 2011.

- [8] G. Bernstein, et.al., "Routing and Wavelength Assignment Information Encoding for Wavelength Switched Optical Networks", draft-ietf-ccamp-rwa-wson-encode, work in progress, October 2011.
- [9] Iftekhar H., et.al., "Super-Channel Optical Parameters GMPLS Signaling Extensions", draft-hussain-ccamp-super-channel-param-sig, work in progress, March 2012.

10. Acknowledgments

<Add any acknowledgements>

Authors' Addresses

Iftekhhar Hussain
Infinera
140 Caspian Ct., Sunnyvale, CA 94089

Email: ihussain@infinera.com

Marco Sosa
Infinera
140 Caspian Ct., Sunnyvale, CA 94089

Email: msosa@infinera.com

Abinder Dhillon
Fujitsu Network Communications
2801 Telecom Parkway, Richardson, TX 75082

Email: abinder.dhillon@us.fujitsu.com

Contributor's Addresses

Rajan Rao
Infinera
140 Caspian Ct., Sunnyvale, CA 94089

Email: rrao@infinera.com

Biao Lu
Infinera
140 Caspian Ct., Sunnyvale, CA 94089

Email: blu@infinera.com

Subhendu Chattopadhyay
Infinera
140 Caspian Ct., Sunnyvale, CA 94089

Email: schattopadhyay@infinera.com

Harpreet Uppal
Infinera
140 Caspian Ct., Sunnyvale, CA 94089

Email: harpreet.uppal@infinera.com

Vinayak Dangui
Infinera
140 Caspian Ct., Sunnyvale, CA 94089

Email: vdangui@infinera.com

Michael VanLeeuwen
Infinera
140 Caspian Ct., Sunnyvale, CA 94089

Email: MVanleeuwen@infinera.com

Zhong Pan
Infinera
140 Caspian Ct., Sunnyvale, CA 94089

Email: zpan@infinera.com

Network Working Group
Internet Draft
Intended status: Standard Track
Expires: April 2014

Iftekhar Hussain
Vinayak Dangui
Michael VanLeeuwen
Marco Sosa
Infinera

October 8, 2013

Super-Channel Optical Parameters GMPLS Signaling Extensions
draft-hussain-ccamp-super-channel-param-sig-03.txt

Abstract

This document builds on [6][7] and defines GMPLS signaling extensions to carry super-channel optical parameters for efficient spectrum assignment on flexible grid networks.

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), 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 8, 2014.

Copyright Notice

Copyright (c) 2013 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.....	2
2. Terminology.....	3
3. GMPLS Signaling Extensions for Super-Channel Optical Parameters3	
3.1. Option 1: Encode Super-Channel Optical Parameters in the	
RSVP FLOWSPEC or TSPEC Object.....	4
3.2. Option 2: Encode the Aforementioned Information along with	
the Super-Channel Label.....	6
4. Procedure for Signaling Super-Channel Optical Parameters.....	6
5. TLV Encoding Examples.....	6
6. Security Considerations.....	6
7. IANA Considerations.....	6
8. References.....	6
8.1. Normative References.....	6
8.2. Informative References.....	7
9. Acknowledgments.....	8

1. Introduction

Future transport systems are expected to support service upgrades to data rates of 1 Tbps and beyond. To scale networks beyond 100Gbps, multi-carrier super-channels coupled with advanced multi-level modulation formats and flexible channel spectrum bandwidth allocation schemes have become pivotal for future spectral efficient transport network architectures [1,2].

The coexistence of super-channels using different modulation formats on the same optical fiber network infrastructure may have a detrimental effect on the Optical Signal to Noise Ratio (OSNR) of adjacent super-channels due to interference such as cross-phase modulation. Therefore, it may be highly desirable to be able to evaluate the mutual impact of the existing and new super-channels on each other's quality of transmission (e.g., bit error rate) before establishing new super-channels.

This document defines GMPLS signaling extensions to convey super-channel optical parameters such as number of carriers, each carrier's center frequency, modulation, and FEC type in the RSVP message. This allows nodes along the super-channel path to learn the aforementioned super-channel optical characteristics and in turn advertise this information to other nodes in the network using GMPLS routing extensions defined in [9].

2. Terminology

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].

3. GMPLS Signaling Extensions for Super-Channel Optical Parameters

This document defines extensions for signaling super-channel optical parameters including:

- o Number of Carriers
- o Carrier Center Frequency (THz)
- o Carrier Modulation
- o Carrier Baudrate (Gbit/s)
- o Carrier FEC Type

This document defines two options for encoding this information.

[Editor's note: to allow full flexibility we have included two encoding options]

3.1. Option 1: Encode Super-Channel Optical Parameters in the RSVP FLOWSPEC or TSPEC Object

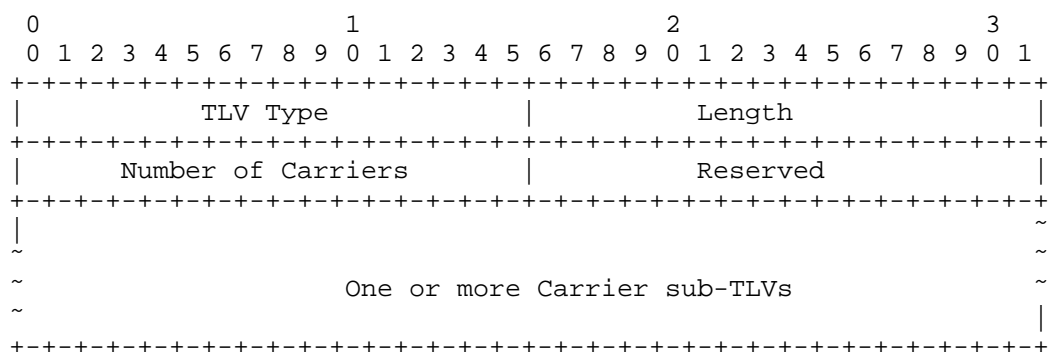


Figure 1: Super-Channel Carriers TLV Format

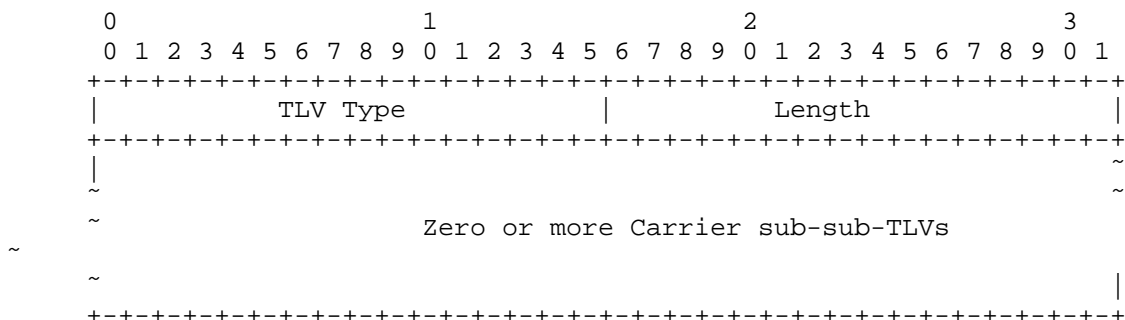


Figure 2: Carrier sub-TLV Format.

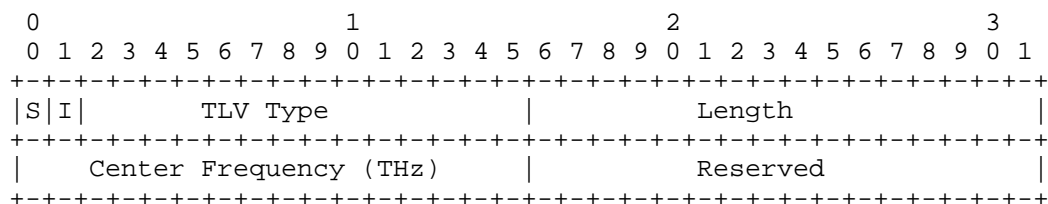


Figure 3: Carrier Center Frequency sub-sub-TLV.

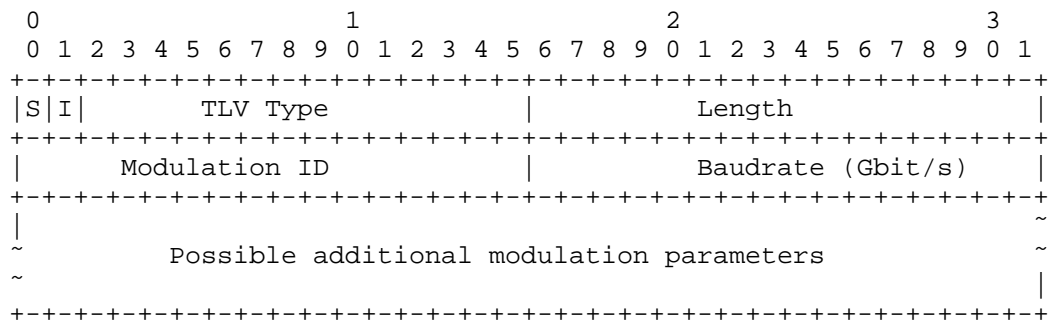


Figure 4: Carrier Modulation sub-sub-TLV.

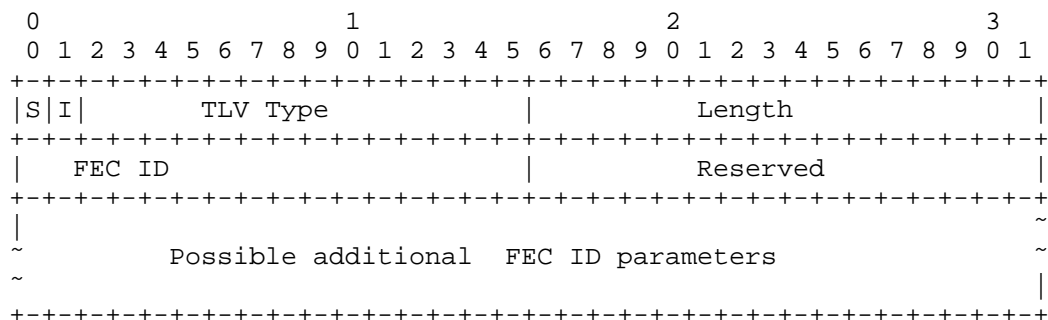


Figure 5: Carrier FEC sub-sub-TLV.

Where:

- o When the S bit in a TLV is set to 1 it indicates that the TLV contains standardized fields (e.g., Modulation, FEC Type) and when the S bit is set to 0 in a TLV it indicates a vendor specific TLV (see [8])
- o Modulation ID, FEC ID, and I fields are similar to as defined in [8]
- o The Length field in the super-channel Carriers TLV specifies the length in octets of the complete set of TLVs including the set of sub-TLVs that follow.

3.2. Option 2: Encode the Aforementioned Information along with the Super-Channel Label

For example use Super-Channel Label defined in [7] to also encode Super-Channel Carriers TLV, the Carrier sub-TLVs, and the associated set of sub-sub-TLVs defined in the previous section.

4. Procedure for Signaling Super-Channel Optical Parameters

- o The optical parameters of the super-channel are signaled in the RSVP message using encoding option 1 (or option 2).
- o During a new super-channel establishment, each node along the new super-channel setup path allocates the required number of slices and also learns the associated set of signaled super-channel optical parameters.

5. TLV Encoding Examples

To be added later.

6. Security Considerations

<Add any security considerations>

7. IANA Considerations

IANA needs to assign a new Grid field value to represent ITU-T Flex-Grid.

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.

[RFC3471] Berger, L., Ed., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, January 2003.

[RFC6205] Otani, T., Ed., "Generalized Labels for Lambda-Switch-Capable (LSC) Label Switching Routers", RFC 6205, March 2011.

[RFC6163] Lee, Y., Ed., "Framework for GMPLS and Path Computation Element (PCE) Control of Wavelength Switched Optical Networks (WSONs)", RFC 6163, April 2011

8.2. Informative References

- [1] Gringeri, S., Basch, B. Shukla,V. Egorov, R. and Tiejun J. Xia, "Flexible Architectures for Optical Transport Nodes and Networks", IEEE Communications Magazine, July 2010, pp. 40-50
- [2] M. Jinno et. al., "Spectrum-Efficient and Scalable Elastic Optical Path Network: Architecture, Benefits and Enabling Technologies", IEEE Comm. Mag., Nov. 2009, pp. 66-73.
- [3] S. Chandrasekhar and X. Liu, "Terabit Super-Channels for High Spectral Efficiency Transmission",in Proc. ECOC 2010, paper Tu.3.C.5, Torino (Italy), September 2010.
- [4] ITU-T Recommendation G.694.1, "Spectral grids for WDM applications: DWDM frequency grid", June 2002
- [5] [4] "Finisar to Demonstrate Flexgrid(TM) WSS Technology at ECOC 2010", press release.
- [6] Abinder D., et.al., "OSPFTE extension to support GMPLS for Flex Grid", draft-dhillon-ccamp-super-channel-ospfte-ext, work in progress, work in progress, November 2011.
- [7] Iftekhar H., et.al., "Generalized Label for Super-Channel Assignment on Flexible Grid", draft-hussain-ccamp-super-channel-label, work in progress, October 2011.
- [8] G. Bernstein, et.al., "Routing and Wavelength Assignment Information Encoding for Wavelength Switched Optical Networks", draft-ietf-ccamp-rwa-wson-encode, work in progress, October 2011.
- [9] Iftekhar H., et.al., "Super-Channel Optical Parameters GMPLS Routing Extensions", draft-hussain-ccamp-super-channel-param-ospfte, work in progress, March 2012.

9. Acknowledgments

<Add any acknowledgements>

Authors' Addresses

Iftekhar Hussain
Infinera
140 Caspian Ct., Sunnyvale, CA 94089

Email: ihussain@infinera.com

Vinayak Dangui
Infinera
140 Caspian Ct., Sunnyvale, CA 94089

Email: vdangui@infinera.com

Michael VanLeeuwen
Infinera
140 Caspian Ct., Sunnyvale, CA 94089

Email: MVanleeuwen@infinera.com

Marco Sosa
Infinera
140 Caspian Ct., Sunnyvale, CA 94089

Email: msosa@infinera.com

Contributor's Addresses

Abinder Dhillon
Infinera
140 Caspian Ct., Sunnyvale, CA 94089

Email: adhillon@infinera.com

Rajan Rao
Infinera
140 Caspian Ct., Sunnyvale, CA 94089

Email: rrao@infinera.com

Biao Lu
Infinera
140 Caspian Ct., Sunnyvale, CA 94089

Email: blu@infinera.com

Subhendu Chattopadhyay
Infinera
140 Caspian Ct., Sunnyvale, CA 94089

Email: schattopadhyay@infinera.com

Harpreet Uppal
Infinera
140 Caspian Ct., Sunnyvale, CA 94089

Email: harpreet.uppal@infinera.com

Zhong Pan
Infinera
140 Caspian Ct., Sunnyvale, CA 94089

Email: zpan@infinera.com

Network Working Group
Internet Draft
Intended status: Standards Track

H. Long
M.Ye
Huawei Technologies Co., Ltd
G. Mirsky
Ericsson
A Alessandro
Telecom Italia S.p.A
October 18, 2013

Expires: April 2014

OSPF Routing Extension for Links with Variable Discrete Bandwidth
draft-long-ccamp-ospf-availability-extension-01.txt

Abstract

Packet switching network may contain links with variable discrete bandwidth, e.g., copper, radio, etc. The bandwidth of such link may change discretely in reaction to changing external environment. Availability is typically used for describing such links during network planning. This document describes an extension for OSPF routing for route computation in a Packet Switched Network (PSN) which contains links with variable discrete bandwidth by introducing an optional availability sub-TLV.

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), 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 18, 2014.

Copyright Notice

Copyright (c) 2013 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
2. Overview	3
3. Extension to OSPF Routing Protocol.....	4
3.1. Interface Switching Capacity Descriptor.....	4
3.2. ISCD Availability sub-TLV.....	5
3.3. Signaling Process.....	6
4. Security Considerations.....	6
5. IANA Considerations	6
6. References	6
6.1. Normative References.....	6
6.2. Informative References.....	7
7. Acknowledgments	7

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 RFC-2119 [RFC2119].

The following acronyms are used in this draft:

OSPF	Open Shortest Path First
PSN	Packet Switched Network
SNR	Signal-to-noise Ratio
LSP	Label Switched Path
ISCD	Interface Switching Capacity Descriptor

PE Provider Edge

LSA Link State Advertisement

1. Introduction

Some data communication technologies allow seamless change of maximum physical bandwidth through a set of known discrete values. For example, in mobile backhaul network, microwave links are very popular for providing connection of last hops. In case of heavy rain, to maintain the link connectivity, the microwave link may lower the modulation level since demodulating lower modulation level need lower signal-to-noise ratio (SNR). This is called adaptive modulation technology [EN 302 217]. However, lower modulation level also means lower link bandwidth. When link bandwidth reduced because of modulation down-shifting, high priority traffic can be maintained, while lower priority traffic is dropped. Similarly the cooper links may change their effective link bandwidth due to external interference.

The parameter, availability [G.827, F.1703, P.530], is often used to describe the link capacity during network planning. Assigning different availability classes to different types of service over such kind of links provides more efficient planning of link capacity. To set up an LSP across these links, availability information is required for the nodes to verify bandwidth satisfaction and make bandwidth reservation. The availability information should be inherited from the availability requirements of the services expected to be carried on the LSP, voice service usually needs "five nines" availability, while non-real time services may adequately perform at four or three nines availability.

For the route computation, the availability information should be provided along with bandwidth resource information. In this document, an extension on Interface Switching Capacity Descriptor (ISCD) [RFC4202] for availability information is defined to support in routing signaling. The extension reuses the reserved field in the ISCD and also introduces an optional availability sub-TLV.

If there is a hop that cannot support the availability sub-TLV, the availability sub-TLV is ignored.

2. Overview

A node which has link(s) with variable bandwidth attached SHOULD contain a <bandwidth, availability> information list in its OSPF TE LSA messages. The list provides the information that how much

bandwidth a link can support for a specified availability. This information is used for path calculation by the PE node(s).

To setup an label switching path (LSP), a PE node may collect link information which is spread in OSPF TE LSA message by network nodes to get know about the network topology, and calculate out an LSP route based on the network topology, and send the calculated LSP route to signaling to initiate a PATH/RESV message for setting up the LSP.

3. Extension to OSPF Routing Protocol

3.1. Interface Switching Capacity Descriptor

The Interface Switching Capacity Descriptor (ISCD) sub-TLV [RFC 4203] has the following format:

0	1	2	3																
0 1 2 3 4 5 6 7 8 9	0 1 2 3 4 5 6 7 8 9	0 1 2 3 4 5 6 7 8 9	0 1																
Type										Length									
Switching Cap										Encoding									
Max LSP Bandwidth at priority 0										AI									
Max LSP Bandwidth at priority 1										Reserved									
Max LSP Bandwidth at priority 2																			
Max LSP Bandwidth at priority 3																			
Max LSP Bandwidth at priority 4																			
Max LSP Bandwidth at priority 5																			
Max LSP Bandwidth at priority 6																			
Max LSP Bandwidth at priority 7																			
Switching Capacity-specific Information																			
(variable)																			

Type: 0x15, 16 bits;

Length: 16 bits;

Switching Cap: 8 bits

Encoding: 8 bits

See [RFC4203] section 1.4.

AI: ISCD Availability sub-TLV index, 8 bits

This new field is the index of availability sub-TLV for this ISCD sub-TLV.

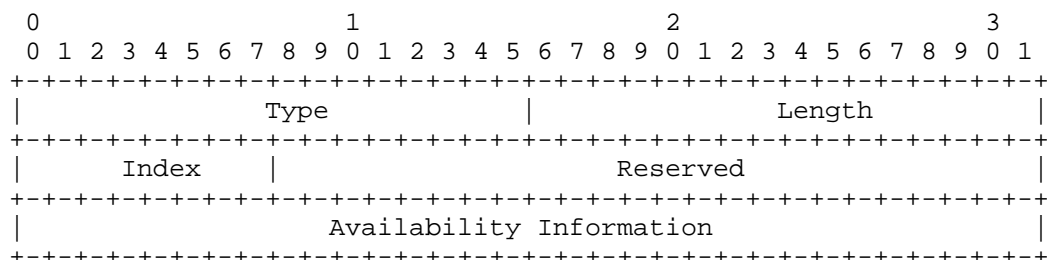
Maximum LSP Bandwidth: 32 bits

Switching Capability specific information: variable

See [RFC4203] section 1.4.

3.2. ISCD Availability sub-TLV

The availability sub-TLV has the following format:



Type: 0x01, 16 bits;

Length: 16 bits;

Index: 8 bits

This field is the index of this availability sub-TLV, referred by the AI field of the ISCD sub-TLV.

Availability Information: 32 bits

This field is a 32-bit IEEE floating point number which describes the availability guarantee of the switching capacity in the ISCD object which has the AI value equal to Index of this sub-TLV. The value must be less than 1.

3.3. Signaling Process

A node which has link(s) with variable bandwidth attached SHOULD contain one or more ISCD Availability sub-TLVs in its OSPF TE LSA messages. Each ISCD Availability sub-TLV provides the information that how much bandwidth a link can support for a specified availability. This information is used for path calculation by the PE node(s).

4. Security Considerations

This document does not introduce new security considerations to the existing OSPF protocol.

5. IANA Considerations

This document introduces an Availability sub-TLV of the ISCD sub-TLV of the TE Link TLV in the TE Opaque LSA for OSPF v2. This document proposes a suggested value for the Availability sub-TLV; it is recommended that the suggested value be granted by IANA. Initial values are as follows:

Type	Length	Format	Description
---	----	-----	-----
0	-	Reserved	Reserved value
0x01	8	see Section 3.2	Availability sub-TLV

6. References

6.1. Normative References

- [RFC2210] Wroclawski, J., "The Use of RSVP with IETF Integrated Services", RFC 2210, September 1997.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC3473] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.

- [RFC4202] Kompella, K. and Rekhter, Y. (Editors), "Routing Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4202, October 2005.
- [RFC4203] Kompella, K., Ed., and Y. Rekhter, Ed., "OSPF Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4203, October 2005.
- [G.827] ITU-T Recommendation, "Availability performance parameters and objectives for end-to-end international constant bit-rate digital paths", September, 2003.
- [F.1703] ITU-R Recommendation, "Availability objectives for real digital fixed wireless links used in 27 500 km hypothetical reference paths and connections", January, 2005.
- [P.530] ITU-R Recommendation, "Propagation data and prediction methods required for the design of terrestrial line-of-sight systems", February, 2012
- [EN 302 217] ETSI standard, "Fixed Radio Systems; Characteristics and requirements for point-to-point equipment and antennas", April, 2009

6.2. Informative References

- [MCOS] Minei, I., Gan, D., Kompella, K., and X. Li, "Extensions for Differentiated Services-aware Traffic Engineered LSPs", Work in Progress, June 2006.

7. Acknowledgments

Authors' Addresses

Hao Long
Huawei Technologies Co., Ltd.
No.1899, Xiyuan Avenue, Hi-tech Western District
Chengdu 611731, P.R.China

Phone: +86-18615778750
Email: longhao@huawei.com

Min Ye
Huawei Technologies Co., Ltd.
No.1899, Xiyuan Avenue, Hi-tech Western District
Chengdu 611731, P.R.China

Email: amy.yemin@huawei.com

Greg Mirsky
Ericsson

Email: gregory.mirsky@ericsson.com

Alessandro D'Alessandro
Telecom Italia S.p.A

Email: alessandro.dalessandro@telecomitalia.it

Network Working Group
Internet Draft
Intended status: Standards Track

H. Long
M.Ye
Huawei Technologies Co., Ltd
G. Mirsky
Ericsson
A Alessandro
Telecom Italia S.p.A
October 18, 2013

Expires: April 2014

RSVP-TE Signaling Extension for Links with Variable Discrete
Bandwidth
draft-long-ccamp-rsvp-te-bandwidth-availability-02.txt

Abstract

Packet switching network may contain links with variable bandwidth, e.g., copper, radio, etc. The bandwidth of such link is sensitive to external environment. Availability is typically used for describing the link during network planning. This document describes an extension for RSVP-TE signaling for setting up a label switching path (LSP) in a Packet Switched Network (PSN) network which contains links with discretely variable bandwidth by introducing an optional availability field in RSVP-TE signaling.

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), 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 18, 2014.

Copyright Notice

Copyright (c) 2013 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
2. Overview	4
3. Extension to RSVP-TE Signaling.....	4
3.1. SENDER_TSPEC Object.....	4
3.1.1. Bandwidth Profile TLV.....	5
3.2. FLOWSPEC Object.....	6
3.3. Signaling Process.....	6
4. Security Considerations.....	7
5. IANA Considerations	7
5.1 RSVP Objects Class Types.....	7
5.2 Ethernet Bandwidth Profile TLV	8
6. References	9
6.1. Normative References.....	9
6.2. Informative References.....	9
7. Acknowledgments	9

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 RFC-2119 [RFC2119].

The following acronyms are used in this draft:

RSVP-TE Resource Reservation Protocol-Traffic Engineering

LSP Label Switched Path

PSN Packet Switched Network

SNR	Signal-to-noise Ratio
TLV	Type Length Value
PE	Provider Edge
LSA	Link State Advertisement

1. Introduction

The RSVP-TE specification [RFC3209] and GMPLS extensions [RFC3473] specify the signaling message including the bandwidth request for setting up a label switching path in a PSN network.

Some data communication technologies allow seamless change of maximum physical bandwidth through a set of known discrete values. For example, in mobile backhaul network, microwave links are very popular for providing connection of last hops. In case of heavy rain, to maintain the link connectivity, the microwave link may lower the modulation level since demodulating lower modulation level need lower signal-to-noise ratio (SNR). This is called adaptive modulation technology [EN 302 217]. However, lower modulation level also means lower link bandwidth. When link bandwidth reduced because of modulation down-shifting, high priority traffic can be maintained, while lower priority traffic is dropped. Similarly the cooper links may change their link bandwidth due to external interference.

The parameter, availability [G.827, F.1703, P.530], is often used to describe the link capacity during network planning. Assigning different availability classes to different types of service over such kind of links provides more efficient planning of link capacity. To set up a LSP across these links, availability information is required for the nodes to verify bandwidth satisfaction and make bandwidth reservation. The availability information should be inherited from the availability requirements of the services expected to be carried on the LSP, voice service usually needs "five nines" availability, while non-real time services may adequately perform at four or three nines availability. Since different service types may need different availabilities guarantee, multiple <availability, bandwidth> pairs may be required when signaling.

To fulfill LSP setup by signaling in these scenarios, this document specifies a new availability sub-TLV as the sub-TLV of Ethernet bandwidth profiles [RFC6003]. Multiple bandwidth profiles with different availability can be carried in the SENDER_TSPEC object.

2. Overview

A PSN tunnel may span one or more links in a network. To setup a label switching path (LSP), a PE node may collect link information which is spread in routing message, e.g., OSPF TE LSA message, by network nodes to get to know about the network topology, and calculate out an LSP route based on the network topology, and send the calculated LSP route to signaling to initiate a PATH/RESV message for setting up the LSP.

In case that there is(are) link(s) with variable discrete bandwidth in a network, a <bandwidth, availability> requirement list should be specified for an LSP. Each <bandwidth, availability> pair in the list means that listed bandwidth with specified availability is required. The list could be inherited from the results of service planning for the LSP.

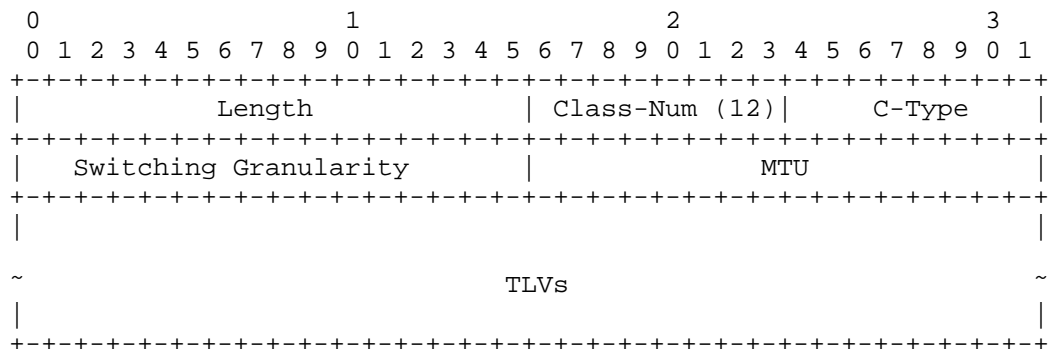
When a PE node initiates a PATH/RESV signaling to set up an LSP, the PATH message SHOULD carry the <bandwidth, availability> requirement list as bandwidth request. Intermediate node(s) will allocate the bandwidth resource for each availability requirement from the remaining bandwidth with corresponding availability. An error message may be returned if any <bandwidth, availability> request cannot be satisfied.

If there is a hop that cannot support the availability sub-TLV, the availability sub-TLV is ignored, and the requirement will be treated as the highest availability.

3. Extension to RSVP-TE Signaling

3.1. SENDER_TSPEC Object

The SENDER_TSPEC object (Class-Num = 12) has the following format:



Switching Granularity (SG): 16 bits

See [RFC6003] section 4.

MTU: 16bits

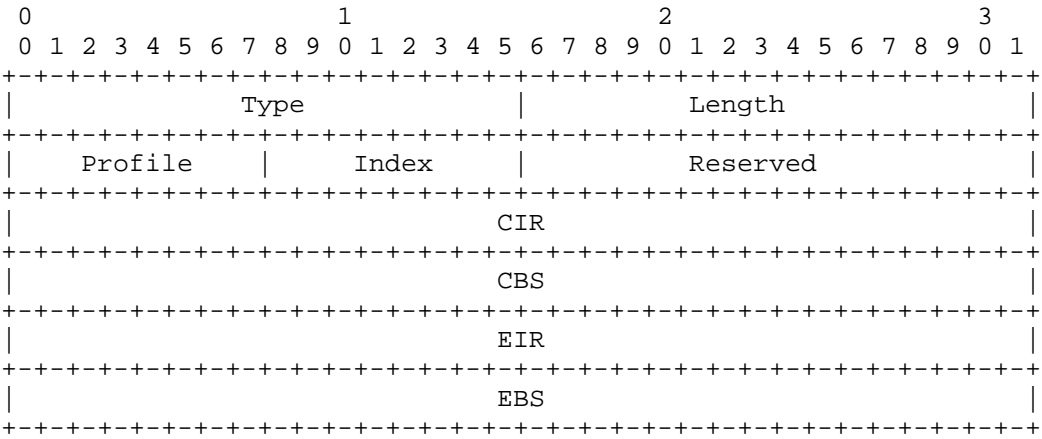
See [RFC6003] section 4.

TLV (Type-Length-Value):

The SENDER_TSPEC object MUST include at least one TLV and MAY include more than one TLV.

3.1.1.1. Bandwidth Profile TLV

The Bandwidth Profile TLV has the following format.



Type: 0x02, 16 bits;

Length: 16 bits;

Profile: 8 bits

This field is defined as a bit vector of binary flags. In RFC 6003, the following flags are defined:

Flag 1 (bit 0): Coupling Flag (CF)

Flag 2 (bit 1): Color Mode (CM)

A new flag is defined in this document:

Flag 3 (bit 2): Availability Flag (AF)

Index: 8 bits

CIR (Committed Information Rate): 32 bits

CBS (Committed Burst Size): 32 bits

EIR (Excess Information Rate): 32 bits

EBS (Excess Burst Size): 32 bits

See [RFC6003] section 4.1.

When the Flag 3 is set to value 1, there is an availability sub-TLV included in this Bandwidth Profile TLV. When the Flag 3 is set to value 0, there won't be an availability sub-TLV. The availability sub-TLV has the following format:

0																1																2																3															
0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9																								
Type																Length																Availability																															

Type (2 octets): TBD

Length (2 octets): 4

Availability (4 octets): a 32-bit floating number describes availability requirement for this bandwidth request. The value must be less than 1.

3.2. FLOWSPEC Object

The FLOWSPEC object (Class-Num = 9, Class-Type = TBD) has the same format as the Ethernet SENDER_TSPEC object.

3.3. Signaling Process

The source node initiates PATH messages including one or more Bandwidth Profile TLVs with different availability value in the

SENDER_TSPEC object. Each Bandwidth Profile TLV specifies the portion of bandwidth request with referred availability requirement.

The destination node checks whether it can satisfy the bandwidth requirements by comparing each bandwidth requirement inside the SENDER_TSPEC objects with the remaining link sub-bandwidth resource with respective availability guarantee when received the PATH message.

- o If all bandwidth requirements can be satisfied, it should reserve the bandwidth resource from each remaining sub-bandwidth portion to set up this LSP. Optionally, the higher availability bandwidth can be allocated to lower availability request when the lower availability bandwidth cannot satisfy the request.
- o If at least one bandwidth requirement cannot be satisfied, it should generate PathErr message with the error code "Admission Control Error" and the error value "Requested Bandwidth Unavailable" (see [RFC2205]).

4. Security Considerations

This document does not introduce new security considerations to the existing RSVP-TE signaling protocol.

5. IANA Considerations

IANA maintains registries and sub-registries for RSVP-TE used by GMPLS. IANA is requested to make allocations from these registries as set out in the following sections.

5.1 RSVP Objects Class Types

This document introduces two new Class Types for existing RSVP objects. IANA is requested to make allocations from the "Resource ReSerVation Protocol (RSVP) Parameters" registry using the "Class Names, Class Numbers, and Class Types" sub-registry.

Class Number	Class Name	Reference
-----	-----	-----
9	FLOWSPEC	[RFC2205]
Class Type (C-Type):		

6 Ethernet SENDER_TSPEC [RFC6003]

Class Number	Class Name	Reference
-----	-----	-----
12	SENDER_TSPEC	[RFC2205]
Class Type (C-Type):		
6	Ethernet SENDER_TSPEC	[RFC6003]

5.2 Ethernet Bandwidth Profile TLV

IANA maintains a registry of GMPLS parameters called ''Generalized Multi-Protocol Label Switching (GMPLS) Signaling Parameters''.

IANA has created a new sub-registry called ''Ethernet Bandwidth Profiles'' to contain bit flags carried in the Ethernet Bandwidth Profile TLV of the Ethernet SENDER_TSPEC object.

Bits are to be allocated by IETF Standards Action. Bits are numbered from bit 0 as the low order bit. A new bit flag is as follow:

Bit	Hex	Description	Reference
---	----	-----	-----
2	0x03	Availability Flag (AF)	[This ID]

Sub-TLV types for Ethernet Bandwidth Profiles are to be allocated by IETF Standard Action. Initial values are as follows:

Type	Length	Format	Description
---	----	-----	-----
0	-	Reserved	Reserved value
TBD	4	see Section 3.1	Availability sub-TLV

6. References

6.1. Normative References

- [RFC2210] Wroclawski, J., "The Use of RSVP with IETF Integrated Services", RFC 2210, September 1997.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC3473] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource Reservation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.
- [RFC6003] Papadimitriou, D. "Ethernet Traffic Parameters", RFC 6003, October 2010.
- [G.827] ITU-T Recommendation, "Availability performance parameters and objectives for end-to-end international constant bit-rate digital paths", September, 2003.
- [F.1703] ITU-R Recommendation, "Availability objectives for real digital fixed wireless links used in 27 500 km hypothetical reference paths and connections", January, 2005.
- [P.530] ITU-R Recommendation, "Propagation data and prediction methods required for the design of terrestrial line-of-sight systems", February, 2012
- [EN 302 217] ETSI standard, "Fixed Radio Systems; Characteristics and requirements for point-to-point equipment and antennas", April, 2009

6.2. Informative References

- [MCOS] Minei, I., Gan, D., Kompella, K., and X. Li, "Extensions for Differentiated Services-aware Traffic Engineered LSPs", Work in Progress, June 2006.

7. Acknowledgments

The authors would like to thank Khuzema Pithewan, Lou Berger, Yuji Tochio, Dieter Beller, and Autumn Liu for their comments on the document.

Authors' Addresses

Hao Long
Huawei Technologies Co., Ltd.
No.1899, Xiyuan Avenue, Hi-tech Western District
Chengdu 611731, P.R.China

Phone: +86-18615778750
Email: longhao@huawei.com

Min Ye
Huawei Technologies Co., Ltd.
No.1899, Xiyuan Avenue, Hi-tech Western District
Chengdu 611731, P.R.China

Email: amy.yemin@huawei.com

Greg Mirsky
Ericsson

Email: gregory.mirsky@ericsson.com

Alessandro D'Alessandro
Telecom Italia S.p.A

Email: alessandro.dalessandro@telecomitalia.it

INTERNET-DRAFT
Intended Status: Standard Track
Expires: April 20, 2014

Khuzema Pithewan
Rajan Rao
Infinera
October 17, 2013

OSPF-TE extensions for MLNMRN based on OTN
draft-rao-ccamp-mlnmrn-otn-ospfte-ext-03.txt

Abstract

This document specifies OSPF extensions for multi-layer/multi-region where one of the regions is multi-layer e.g. OTN, SONET/SDH.

Status of this Memo

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/lid-abstracts.html>

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

Copyright and License Notice

Copyright (c) 2013 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
2 Layer Identification	3
3 OTN Layer ID	4
4 SONET/SDH Layer Identification	6
5 Procedure	6
6 Examples	6
6.1. Ethernet and OTN	7
6.2. OTN and FlexGrid	7
6.3. OTN and SONET/SDH	8
6.4. OTN and OTN	8
7 IANA Considerations	8
8 Security Considerations	9
9 References	9
10. Authors' Addresses	9

1 Introduction

In order to do end-to-end path computation, where a path may involve more than one region and part of single routing domain, TE Links connecting the two regions need to have bandwidth capacity advertised for the switch that connects the two regions. This document specifies the OSPF extensions that are required if any of the region is a multi-layer network. The specification is based on the requirement as specified in RFC 5212. As per the said RFC, ISCD characterizes the information associated to one or more network layers. Same RFC also says that the information about the adjustment capabilities of the nodes in the network allow the path computation process to select an end-to-end multi-layer or multi-region path that includes links with different switching capabilities joined by LSRs that can adapt (i.e., adjust) the signal between the links. By inference, information about the adjustment capabilities should be able to identify a layer in ISCD, if ISCD specifies more than one layer.

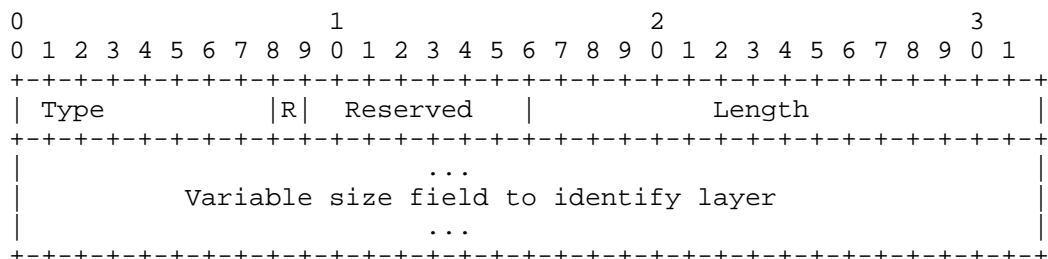
RFC6001 specifies how to advertise adjustment capabilities between two switching regions. IACD definition has provision to extend it for a specific technology through Adjustment Capability Specific information (ACSI) field, if required. ACSI field can be used to identify a layer in the multi-layer ISCD.

While OTN multi-layer technology is a primary driver for this extension, the extensions in this document does cover specifications for multi-layer technologies in general. To make sure the extensions are extensible to other multi-layer technologies as well, this document covers SDH/SONET as well.

2 Layer Identification

Multi-region path computation requires to identify a layer in the multi-layer region. This mandates layer identification along with identification of technology in the region. The technology identification is done via Switching capability and Encoding type.

IACD needs to be extended to be able to carry layer identification. the layer Identification is OPTIONAL and used only when interface supports layer multiplexing and hence creating a need to identify a layer. A new Layer ID Sub-TLV has been defined to carry layer identification.



Type : Type field is used to identify a particular structure of variable size field, which is specific to the particular Switching Capability and Encoding type combination

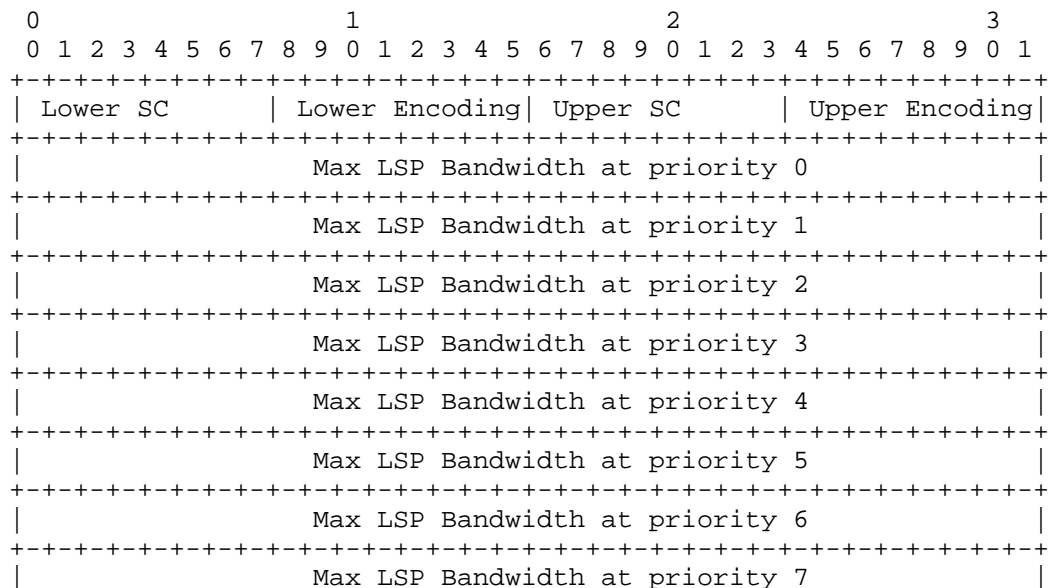
R : This bit is used to make sense whether the Layer ID is for Lower region or upper region. 1 means upper region and 0 means lower.

IACD can have at-most 2 Layer ID TLVs, if both the regions are multi-layer.

Next two sections specifies Layer ID for two multi-layer technologies namely, OTN and SONET/SDH

3 OTN Layer ID

RFC6001 defines IACD sub-TLV as follows. Please refer to the RFC for definition of individual fields of the sub-TLV.



```

+-----+
|               Adjustment Capability-specific information               |
|               (variable)                                             |
+-----+

```

[GMPLS-OTN-OSPF] defines attributes that identifies a layer in multi-layer OTN ISCD. These attributes are part of Bandwidth sub-TLV in Switch capability specific information of ISCD. These attributes are reproduced here for completeness sake.

- * Signal Type: Layer for which bandwidth is being advertised.
- * Hierarchy : also called as multiplexing branch that specifies all the layers between server layer and signal type.
- * TSG : Time Slot Granularity

Adjustment Capability-specific information abbreviated as ACSI henceforth for OTN G.709v3 carries LayerID Sub-TLV which is defined as follows

```

0               1               2               3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+
| Type = 1 | R | Reserved | Length |
+-----+-----+-----+-----+
| Signal type | Num of stages | TSG | Res | Stage#1 |
+-----+-----+-----+-----+
| Stage#2 | ... | Stage#N | Padding |
+-----+-----+-----+-----+

```

This LayerID sub-TLV is applicable only when one of the regions is OTN, which means either lower or upper SC and Encoding type MUST have Switch Cap as OTN-TDM and encoding type as G.709 ODUk.

R bit is used to make sense whether the Layer ID is for Lower region or upper region. 1 means upper region and 0 means lower.

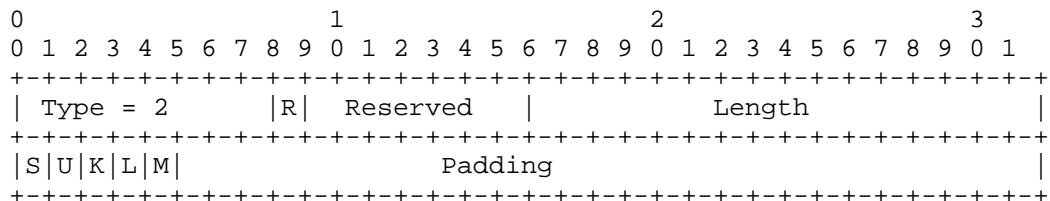
The 8 priorities of the BW as defined in main IACD structure, is adjustment capability between the two regions where one of the region is identifies by LayerID sub-TLV.

Absence of this sub-TLV for OTN means that the OTN ISCD doesn't support multiplexing.

4 SONET/SDH Layer Identification

G.707 defines the structure of SDH multiplexing hierarchy and RFC 4606 defines generalized label structure needed to fully specify SONET/SDH multiplexing hierarchy. This Label structure also referred as SUKLM structure identifies all the layers of the multiplexing hierarchy along with time slots. For the purpose of this draft, only layer identification is needed, hence each layer can be identified by a bit. Bit value 1 signifies presence of the layer and 0, its absence. 5 Bits, each representing one layer is sufficient to fully identify the SONET/SDH multiplexing hierarchy.

Layer ID sub TLV for SONET/SDH is defined as follows



SUKLM bits signifies the presence of SONET/SDH layers and these bits together fully specifies the multiplexing hierarchy. Refer to Section 3 of RFC 4606 for full specification of SUKLM bits.

Absence of sub-TLV means that the SONET/SDH ISCD doesn't support multiplexing and needs only transparent mapping to other Interface.

5 Procedure

A node advertising IACD for the bandwidth between regions where one or both of them are hierarchical i.e. OTN or SONET/SDH, MUST include the Layer ID sub-TLV as part of ACSI as defined above.

For multi-region path computation, the path computing node MUST look at the LayerID sub-TLV (in ACSI part of IACD) if lower/upper {SC,Enc] is {OTN-TDM,G.709ODUK} or {TDM,SONET/SDH} to identify the layer for correct layer for BW check.

6 Examples

This section exemplifies TLV values for various technology region combinations, where one of the region is OTN

6.1. Ethernet and OTN When upper region is Ethernet and lower region is OTN

0										1										2										3										
0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1									
PSC-1										Ethernet										OTN-TDM										G.709 ODUk										
										Max LSP Bandwidth at priority 0																														
										/ / / / / / / / / /																														
										Max LSP Bandwidth at priority 7																														
Type = 1										0	Reserved										Length																			
Signal type										Num of stages										TSG	Res										Stage#1									
Stage#2										...										Stage#N										Padding										

6.2. OTN and FlexGrid

0										1										2										3																				
0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1																			
+-----+										+-----+										+-----+										+-----+																				
OTN-TDM										G.709 ODUk										SCSC										Lambda																				
+-----+										+-----+										+-----+										+-----+																				
										Max LSP Bandwidth at priority 0																																								
										/ / / / / / / / / /																																								
										Max LSP Bandwidth at priority 7																																								
+-----+										+-----+										+-----+										+-----+																				
Type = 1										1	Reserved										Length																													
+-----+										+-----+										+-----+										+-----+																				
Signal type										Num of stages										TSG	Res										Stage#1																			
+-----+										+-----+										+-----+										+-----+																				
Stage#2										...										Stage#N										Padding																				
+-----+										+-----+										+-----+										+-----+																				

TBD

8 Security Considerations

TBD

9 References

- [RFC5212] K. Shiimoto, Papadimitriou, D., JL. Le Roux, Vigoureux, M., Brungard, D., "Requirements for GMPLS-Based Multi-Layer and Multi-Region Networks (MLN/ MRN)", RFC 5212, July 2008.
- [RFC6001] Papadimitriou, D., Vigoureux, M., Shiimoto, K., Brungard, D., and JL. Le Roux, "Generalized MPLS (GMPLS) Protocol Extensions for Multi-Layer and Multi-Region Networks (MLN/MRN)", RFC 6001, October 2010.
- [RFC4606] E. Mannie, Perceval, D. Papadimitriou, "Generalized Multi-Protocol Label Switching (GMPLS) Extensions for Synchronous Optical Network (SONET) and Synchronous Digital Hierarchy (SDH) Control", RFC 4606, Aug 2006
- [GMPLS-OTN-OSPF] Traffic Engineering Extensions to OSPF for Generalized MPLS (GMPLS) Control of Evolving G.709 OTN Networks

10. Authors' Addresses

Khuzema Pithewan
Infinera
140 Caspian Ct., Sunnyvale, CA 94089
Email: kpithewan@infinera.com

Rajan Rao
Infinera
140 Caspian Ct., Sunnyvale, CA 94089
Email: rrao@infinera.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 20, 2014

A. Takacs
F. Fondelli
B. Tremblay
Ericsson
Z. Ali
Cisco Systems
October 21, 2013

RSVP-TE Recovery Extension for data plane initiated reversion,
and protection timer signaling
draft-takacs-ccamp-revertive-ps-09

Abstract

RSVP-TE recovery extensions are specified in [RFC4872] and [RFC4873]. Currently recovery signaling does not support the request for revertive protection and recovery timers values. This document extends the PROTECTION Object format allowing sub-TLVs, and defines two sub-TLVs to carry wait-to-restore and hold-off intervals.

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

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 20, 2014.

Copyright Notice

Copyright (c) 2013 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
2. Updated PROTECTION Object format and sub-TLVs	6
3. Error handling	9
4. IANA Considerations	10
5. Security Considerations	11
6. References	12
Authors' Addresses	13

1. Introduction

Generalized MPLS (GMPLS) extends MPLS to include support for different switching technologies [RFC3471]. These switching technologies provide several protection schemes [RFC4426][RFC4427] (e.g. 1+1, 1:N, M:N). Many characteristics of those protection schemes are common regardless of the switching technology (e.g. TDM, LSC, etc). GMPLS RSVP-TE signaling has been extended to support the various protection schemes and establish Label Switched Paths (LSPs) configuring its specific protection characteristics [RFC4426][RFC4872].

Currently RSVP-TE extensions do not address the configuration of protection switching timers. It also does not provide information on the protection switching operation mode (i.e., revertive or non-revertive).

The Hold-off time (HOFF) is defined as the time between the reporting of signal fail or degrade, and the initialization of the recovery switching operation [RFC4427]. This timer is useful to limit the number of switch actions when multiple layers of recovery are being used, or in case of 1+1 unidirectional protection scheme [G.808.1] to prevent too early switching due to the differential delay between the short and long path.

The Wait-to-Restore time (WTR) is defined as a period of time that must elapse after a recovered fault before an LSP can be used again to transport the normal traffic and/or to select the normal traffic from the LSP [RFC4427]. The WTR time is fundamental in revertive mode of operation, to prevent frequent operation of the protection switch due to an intermittent defect [G.808.1].

Reversion refers to the process of moving normal traffic back to the original working LSP after the failure is cleared and the path is repaired [RFC4426][RFC4427][RFC4872]. In transport networks reversion is desirable since the protection path may not be optimal from a routing and resource consumption point of view, additionally, moving traffic back to the working LSP allows the protection resources to be used to protect other LSPs.

WTR and HOFF timers must be accurately configured at both ends of the LSP. Operators may need to tune WTR and HOFF timers on a per LSP basis to ensure best protection switching performance (e.g., account for differential delays between

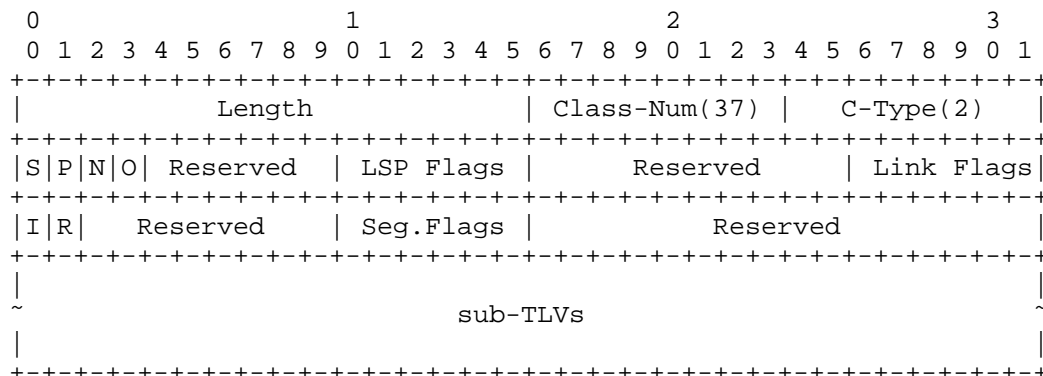
worker and protection paths). Currently these values are either pre-configured to a default value (and so may be suboptimal for some of the LSPs) or need to be manually set/tuned after the connections have been established. Since these parameters are important for recovery in transport networks, it is desirable that GMPLS RSVP-TE protection signaling carries the necessary information.

This document extends the PROTECTION Object format allowing sub-TLVs, and defines three sub-TLVs to carry WTR and HOFF timer values.

2. Updated PROTECTION Object format and sub-TLVs

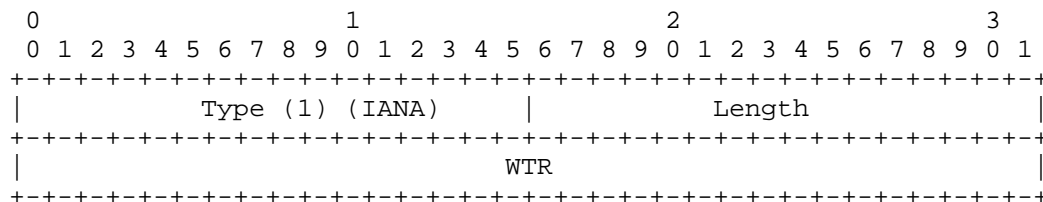
In [RFC4872] and [RFC4873] the PROTECTION object is specified to support end-to-end and segment recovery. In order to ease addition of protection attributes the PROTECTION Object is extended to carry sub-TLVs. The new format updates the PROTECTION Object format of C-Type 2. The updated format is depicted below. IANA is requested to maintain the TLV space for the PROTECTION Object.

We retained C-Type to ensure that nodes not capable of interpreting the new format (sub-TLVs) will still be able to process the object without being required to generate an error; while nodes recognising the new format will process the TLVs accordingly. The processed sub-TLV MUST be included in the PROTECTION Object sent in the Resv message upstream, to ensure that the sender can maintain a consistent view of the actual protection configuration of the LSP.

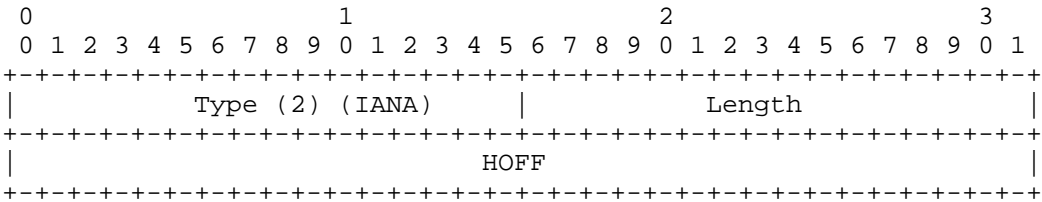


This document specifies three new sub-TLVs.

WTR - Wait-to-Restore time sub-TLV specifies the WTR time. If the WTR field is 0 the protection switching operation mode is non-revertive, otherwise revertive operation with the signalled timer (in milliseconds) is requested. The value 0xffffffff is reserved, and refers to a locally pre-configured WTR value.



HOFF - Hold-off time sub-TLV specifies the HOFF time. The values are in milliseconds. The value 0xffffffff is reserved, and refers to a locally pre-configured HOFF value.



In the case of end-to-end protection the PROTECTION Object is inserted at the top level in the Path message, the WTR and HOFF

options sub-TLVs correspond to the end-to-end protection. In the case when a segment of the LSP is to be protected and the WTR and HOFF timers for the protection segment are to be set by signaling, explicit segment recovery control has to be used, i.e., the PROTECTION Object with the desired timers set must be inserted in the appropriate Secondary Explicit Route Object (SERO).

3. Error handling

In the case a specific configuration of the timers is not supported the corresponding error should be generated and sent in the PathErr message: "Routing Problem/Unsupported WTR value" or "Routing Problem/Unsupported HOFF value".

4. IANA Considerations

4.1. New TLV space for the PROTECTION object

A new TLV space needs to be opened and maintained for the PROTECTION Object in the "Class Names, Class Numbers, and Class Types " Registry.

4.3. New RSVP error sub-code

For Error Code = 24 "Routing Problem" (see [RFC3209]) the following sub-codes are defined.

Sub-code -----	Value -----
Unsupported WTR value	To be assigned by IANA (suggest value: 80)
Unsupported HOFF value	To be assigned by IANA (suggest value: 81)

5. Security Considerations

This document introduces no new security issues. The considerations in [RFC4872] and [RFC4873] apply.

6. References

- [G.808.1] "Generic protection switching -- Linear trail and subnetwork protection", ITU-T Recommendation G.808.1, March 2006.
- [IEEE-PBBTE]
"IEEE 802.1Qay Draft Standard for Provider Backbone Bridging Traffic Engineering", work in progress.
- [RFC3471] "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, January 2003.
- [RFC4426] "Generalized Multi-Protocol Label Switching (GMPLS) Recovery Functional Specification", RFC 4426, March 2006.
- [RFC4427] "Recovery (Protection and Restoration) Terminology for Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4427, March 2006.
- [RFC4872] "RSVP-TE Extensions in Support of End-to-End Generalized Multi-Protocol Label Switching (GMPLS) Recovery", RFC 4872, May 2007.
- [RFC4873] "GMPLS Segment Recovery", RFC 4873, May 2007.

Authors' Addresses

Attila Takacs
Ericsson
Laborc u. 1.
Budapest, 1037
Hungary

Email: attila.takacs@ericsson.com

Francesco Fondelli
Ericsson
Via Negrone
Genova, 16153
Italy

Email: francesco.fondelli.ericsson.com

Benoit Tremblay
Ericsson
8400 Decarie.
Montreal, Quebec H4P 2N2
Canada

Email: benoit.c.tremblay@ericsson.com

Zafar Ali
Cisco Systems
Email: zali@cisco.com

CCAMP Working Group
Internet-Draft
Intended Status: Standards Track
Expires: April 24, 2014

Mike Taillon
Tarek Saad
Rakesh Gandhi
Zafar Ali
(Cisco Systems, Inc)
Manav Bhatia
(Alcatel-Lucent)
Lizhong Jin
()
Frederic Jounay
(Orange CH)
October 21, 2013

Extensions to Resource Reservation Protocol For Fast Reroute of
Bidirectional Co-routed Traffic Engineering LSPs
draft-tsaad-ccamp-rsvpte-bidir-lsp-fastreroute-02

Abstract

This document defines RSVP-TE signaling extensions to support Fast Reroute (FRR) of bidirectional co-routed Traffic Engineering (TE) LSPs. These extensions enable the re-direction of bi-directional traffic and signaling onto bypass tunnels that ensure co-routedness of data and signaling paths in the forward and reverse directions after FRR. In addition, the RSVP-TE signaling extensions allow the coordination of bypass tunnel assignment protecting a common facility in both forward and reverse directions prior to or post failure occurrence.

Status of this Memo

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/lid-abstracts.html>

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

Copyright and License Notice

Copyright (c) 2013 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
2. Terminology	3
3. Link Failure With Node-protection Bypass Tunnels	5
3.1. Behavior Before Local Repair	5
3.1.1. Downstream Merge Point Label Discovery	6
3.1.2. Upstream Merge Point Label Discovery	6
3.2. Behavior Post Link Failure After FRR	6
3.3. Behavior Post Link Failure To Re-coroute	6
4. Bypass Tunnel Assignment Coordination	8
4.1. DOWNSTREAM_BYPASS_ASSIGNMENT Subobject	8
4.2. Bypass Tunnel Assignment Signaling Procedure	10
5. Compatibility	11
6. Security Considerations	11
7. IANA Considerations	11
8. Acknowledgements	11
9. References	11
9.1. Normative References	11
Authors' Addresses	13

1. Introduction

Co-routed bidirectional tunnels are signaled using GMPLS signaling procedures specified in [RFC3473] and [RFC3471]. Existing procedures defined in [RFC4090] describe the behavior of the Point of Local Repair (PLR) to reroute traffic and signaling onto the bypass tunnel in the event of a failure for unidirectional LSPs. These procedures are applicable to unidirectional protected LSPs, and don't address issues that arise when employing FRR for bidirectional co-routed Label Switched Paths (LSPs).

When using current FRR procedures with bidirectional co-routed LSPs, it is possible in some cases (e.g. when using node-protecting bypass tunnels post a link failure event and when RSVP signaling is sent in-fiber and in-band with data), the RSVP signaling refreshes may stop reaching some nodes along the primary bidirectional LSP path after the PLRs complete rerouting traffic and signaling onto the bypass tunnels. This is caused by the asymmetry of paths that may be taken by the bidirectional LSP's signaling in the forward and reverse directions after FRR reroute. In such cases, the RSVP soft-state timeout eventually causes the protected bidirectional LSP to be destroyed, and consequently impacts protected traffic flow after FRR. This problem exists when using either unidirectional or bidirectional bypass tunnels to protect the primary co-routed bidirectional LSP.

When co-routed bidirectional bypass tunnels are used to locally protect bidirectional LSPs, the upstream and downstream PLRs may independently assign different bidirectional bypass tunnels in the forward and reverse direction. Currently, there is no means to coordinate the bypass tunnel selection between the downstream and upstream PLRs. In case of mismatch and after FRR, data traffic and signaling may flow over asymmetric paths in the forward and reverse directions which may be undesirable for certain applications.

This document proposes solutions to the above problems by providing corrective actions in the control plane to complement FRR procedures of [RFC4090] in order to maintain the RSVP soft-state for bidirectional protected LSPs and achieve symmetry in the paths followed by data and signaling in the forward and reverse directions post FRR. The document also extends RSVP signaling so it is possible that the bypass tunnel selected by the upstream PLR matches the one selected by the downstream PLR.

2. Terminology

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].

The reader is assumed to be familiar with the terminology in [RSVP] and [RSVP-TE].

LSR: Label-Switch Router.

LSP: An MPLS Label-Switched Path. In this document, an LSP will always be explicitly routed.

Local Repair: Techniques used to repair LSP tunnels quickly when a node or link along the LSP's path fails.

PLR: Point of Local Repair. The head-end LSR of a bypass tunnel or a detour LSP.

Facility Backup: A local repair method in which a bypass tunnel is used to protect one or more protected LSPs that traverse the PLR, the resource being protected, and the Merge Point in that order.

Protected LSP: An LSP is said to be protected at a given hop if it has one or multiple associated bypass tunnels originating at that hop.

Bypass Tunnel: An LSP that is used to protect a set of LSPs passing over a common facility.

NHOP Bypass Tunnel: Next-Hop Bypass Tunnel. A bypass tunnel that bypasses a single link of the protected LSP.

NNHOP Bypass Tunnel: Next-Next-Hop Bypass Tunnel. A bypass tunnel that bypasses a single node of the protected LSP.

MP: Merge Point. The LSR where one or more bypass tunnels rejoin the path of the protected LSP downstream of the potential failure. The same LSR may be both an MP and a PLR simultaneously.

CSPF: Constraint-based Shortest Path First.

Downstream PLR: A PLR that locally detects a fault and reroutes traffic in the same direction of the protected bidirectional LSP RSVP Path signaling.

Upstream PLR: A PLR that locally detects a fault and reroutes traffic in the opposite direction of the protected bidirectional LSP RSVP Path signaling.

Point of Remote Repair (PRR): an upstream PLR that triggers reroute

of traffic and signaling based on procedures described in this document.

3. Link Failure With Node-protection Bypass Tunnels

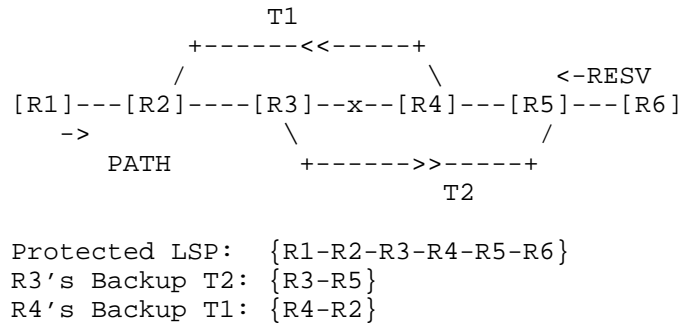


Figure 1: Flow of RSVP signaling post FRR after failure

Consider the Traffic Engineered (TE) network shown in Figure 1. Assume every link in the network is protected with a node- protection bypass tunnel. For the protected bidirectional co-routed LSP whose active/head is on router R1 and passive/tail is on router R6, each traversed router (a potential PLR) independently assigns a node- protection bypass tunnel. Consider a link R3-R4 on the LSP path fails.

The proposed solution introduces two phases to invoking FRR procedures by the PLR post the link failure. The first phase comprises of FRR procedures to fast reroute data traffic onto bypass tunnels in the forward and reverse direction. The second phase re-coroutes the data and signaling in cases where they go over asymmetric paths in the forward and reverse directions after the first phase.

3.1. Behavior Before Local Repair

To correctly reroute data traffic over a node-protection tunnel, the downstream and upstream PLRs have to know, in advance, the downstream and upstream Merge Point (MP) labels so that data in the forward and reverse directions can be tunneled through the bypass tunnel post FRR respectively.

3.1.1. Downstream Merge Point Label Discovery

For unidirectional primary LSPs, [RFC4090] defines procedures for the downstream PLR to obtain the downstream MP label from recorded labels of the RSVP Resv message received at the downstream PLR.

3.1.2. Upstream Merge Point Label Discovery

To obtain the upstream MP label, existing methods to record upstream MP label in the RRO of the RSVP Path message are used. The upstream PLR can obtain the upstream MP label from the recorded label in the RRO of the received RSVP Path message.

3.2. Behavior Post Link Failure After FRR

The downstream PLR R3 and upstream PLR R4 independently trigger fast reroute procedures to redirect traffic onto respective bypass tunnels T2 and T1 in the forward and reverse direction. The downstream PLR R3 also reroutes RSVP Path state onto the bypass tunnel T2 using procedures described in [RFC4090]. Note, at this point, router R4 stops receiving RSVP Path refreshes for the protected bidirectional LSP while primary protected traffic continues to flow over bypass tunnels.

3.3. Behavior Post Link Failure To Re-coroute

The downstream Merge Point (MP) R5 that receives rerouted protected LSP RSVP Path message through the bypass tunnel, in addition to the regular MP processing defined in RF4090, gets promoted to a Point of Remote Repair (PRR role) and performs the following actions to re-coroute signaling and data traffic over the same path in both directions:

For unidirectional bypass tunnels:

- Checks for presence of a bypass tunnel in the reverse direction that terminates on the Downstream PLR R3. Note: the Downstream PLR R3's address is extracted from the "IPV4 tunnel sender address" in the SENDER_TEMPLATE object.
- If present, checks whether the primary LSP traffic and signaling is already rerouted over the found bypass tunnel. If not, PRR R5 activates FRR reroute procedures to direct traffic and signaling (RSVP Resv) over the found bypass tunnel T3 in reverse direction.
- If not present, PRR R5 attempts to auto-provision a bypass tunnel that terminates on the downstream PLR R3. For unidirectional bypass tunnels, if co-routedness in forward and

reverse direction is desired, the reverse path bypass tunnel can be inferred from the forward bypass tunnel path (e.g. by reflecting the RRO recorded in the forward direction as ERO for the reverse direction).

- If PRR R5 is unable to successfully provision a bypass tunnel that terminates on the downstream PLR, it may send an immediate RSVP Notify message back to the head-end. The head-end may tear and re-setup the LSP immediately.

For bidirectional bypass tunnels:

- The PRR follows similar procedures described in the solution to second problem in order to identify the bypass tunnel, and reroute traffic and signaling in the reverse path.

If MP R5 receives multiple RSVP Path messages through multiple bypass tunnels (e.g. as a result of multiple failures), the PRR SHOULD identify/provision a bypass tunnel that terminates on the farthest downstream PLR along the protected LSP path (closest to the bidirectional tunnel headend) and activate the reroute procedures mentioned above.

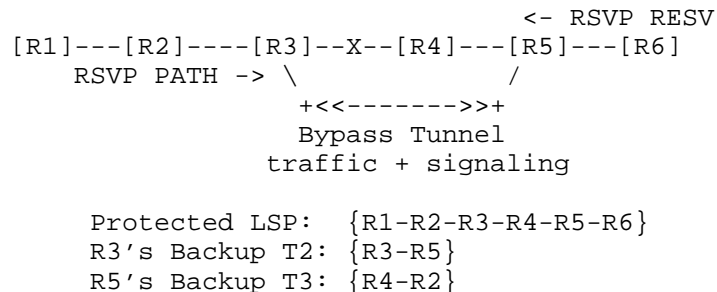


Figure 2: Flow of RSVP signaling post FRR after re-coroute

Figure 2 describes the path taken by traffic and signaling after completing re-coroute of data and signaling in the forward and reverse paths described earlier.

The MP MAY optionally support handling in data plane as follows. If the MP is preconfigured with bidirectional bypass tunnel (by DOWNSTREAM_BYPASS_ASSIGNMENT Subobject in Section 4), as soon as the MP node receives the primary tunnel packets on this bypass tunnel, it MAY switch the upstream traffic on to this bypass tunnel. In order to identify the primary tunnel packets through this bypass tunnel, PHP of the bypass tunnel MUST be disabled. The signaling procedure

described above in this Section will still apply, and MP checks whether the primary tunnel traffic and signaling is already rerouted over the found bypass tunnel, if not, perform the signaling procedure.

4. Bypass Tunnel Assignment Coordination

This document defines a new subobject in RSVP RECORD_ROUTE object, DOWNSTREAM_BYPASS_ASSIGNMENT, to extend RSVP-TE for fast-reroute signaling. This object is backward compatible with LSRs that do not recognize it (see section 3.10 in [RSVP]).

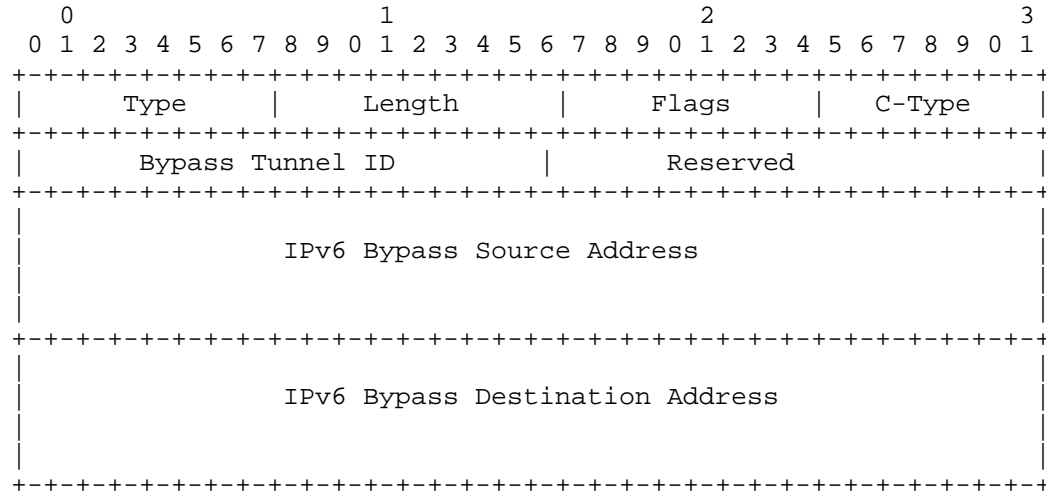
4.1. DOWNSTREAM_BYPASS_ASSIGNMENT Subobject

The DOWNSTREAM_BYPASS_ASSIGNMENT subobject is used to inform the MP of the backup being used by the PLR. This can be used to coordinate the backup used for the protected LSP by the downstream and upstream PLRs in the forward and reverse direction respectively prior or post the failure occurrence. This subobject MUST only be inserted into the Path message by the downstream PLR and MUST NOT be changed by downstream LSRs. The DOWNSTREAM_BYPASS_ASSIGNMENT subobject has the following format:

The IPv4 DOWNSTREAM_BYPASS_ASSIGNMENT subobject has the following format:

0										1										2										3									
0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1								
+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-								
Type										Length										Flags										C-Type									
+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-								
Bypass Tunnel ID										Reserved																													
+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-								
IPv4 Bypass Source Address																																							
+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-								
IPv4 Bypass Destination Address																																							
+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-								

The IPv6 DOWNSTREAM_BYPASS_ASSIGNMENT subobject has the following format:



Type

0x04 (TBD) Downstream Bypass Assignment

Length

The Length contains the total length of the subobject in bytes, including the Type and Length fields.

Flags

TBD.

C-Type

The C-Type of the Downstream Bypass Assignment subobject

Bypass Source Address

The bypass tunnel source IPV4 or IPV6 address.

Bypass Destination Address

The bypass tunnel destination IPV4 or IPV6 address.

Bypass Tunnel ID

The bypass tunnel identifier.

4.2. Bypass Tunnel Assignment Signaling Procedure

In cases where bidirectional bypass tunnels are used for FRR Local Repair for a bidirectional co-routed LSP, it is desirable to coordinate the bypass tunnel selected at the downstream and upstream PLRs so that rerouted traffic and signaling flows on symmetrical paths post FRR. To achieve this, a new RSVP subobject is defined for RECORD_ROUTE object (RRO) that identifies a bidirectional bypass tunnel that is assigned at a downstream PLR to protect a bidirectional LSP.

The DOWNSTREAM_BYPASS_ASSIGNMENT subobject is added by each downstream PLR in the RSVP Path RECORD_ROUTE message of the primary LSP to record the downstream bidirectional bypass tunnel assignment. This subobject is sent in the RSVP Path RECORD_ROUTE message every time the downstream PLR assigns or updates the bypass tunnel assignment so the upstream PLR may reflect the assignment too. The DOWNSTREAM_BYPASS_ASSIGNMENT subobject is added in the RECORD_ROUTE object prior to adding the node's IP address. A node MUST NOT add a DOWNSTREAM_BYPASS_ASSIGNMENT subobject without also adding an IPv4 or IPv6 subobject.

The upstream PLR (downstream MP) that detects a DOWNSTREAM_BYPASS_ASSIGNMENT subobject whose bypass tunnel destination matching its own address assigns the matching bidirectional bypass tunnel in the reverse direction, and forwards the RSVP Path message downstream. Otherwise, the bypass tunnel assignment subobject is simply forwarded downstream along in the RSVP Path message.

In absence of DOWNSTREAM_BYPASS_ASSIGNMENT subobject, the downstream MP can independently assign a bypass tunnel in the reverse direction. In the case of downstream MP receiving multiple DOWNSTREAM_BYPASS_ASSIGNMENT subobjects from multiple downstream PLRs, the decision of selecting a bypass tunnel in the reverse direction can be based on local policy, for example, prefer link protection vs. node protection bypass, or prefer the most upstream vs. least upstream node protection bypass tunnel. Note, the bypass tunnel selection will be corrected after FRR based on the PRR behavior after failure.

5. Compatibility

The DOWNSTREAM_BYPASS_ASSIGNMENT subobject to be defined for RSVP RECORD_ROUTE object with class numbers in the form 1lbbbbbb, which ensures compatibility with non-supporting nodes. Per [RSVP], nodes not supporting this extension will ignore the subobject but forward it, unexamined and unmodified, in all messages resulting from this message.

6. Security Considerations

This document introduces one new RSVP subobject. Thus in the event of the interception of a signaling message, slightly more could be deduced about the state of the network than was previously the case, but this is judged to be a very minor security risk as this information is available by other means.

Otherwise, this document introduces no additional security considerations. For general discussion on MPLS and GMPLS related security issues, see the MPLS/GMPLS security framework [RFC5920].

7. IANA Considerations

A new type for the new DOWNSTREAM_BYPASS_ASSIGNMENT subobject for RECORD_ROUTE object is required.

8. Acknowledgements

Authors would like to thank George Swallow for his detailed and useful comments and suggestions.

9. References

9.1. Normative References

- [RSVP] Braden, R., Ed., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, September 1997.
- [RSVP-TE] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC4090] Pan, P., Ed., Swallow, G., Ed., and A. Atlas, Ed., "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.

- [RFC3473] Berger, L., Ed., "Generalized Multi-Protocol Label
Switching (GMPLS) Signaling Resource ReserVation Protocol-
Traffic Engineering (RSVP-TE) Extensions", RFC 3473,
January 2003.
- [RFC3471] Berger, L., Ed., "Generalized Multi-Protocol Label
Switching (GMPLS) Signaling Functional Description", RFC
3471, January 2003.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate
Requirement Levels", BCP 14, RFC 2119, March 1997.

9.1. Informative References

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

Authors' Addresses

Mike Taillon
Cisco Systems, Inc.

EMail: mtaillon@cisco.com

Tarek Saad
Cisco Systems, Inc.

EMail: tsaad@cisco.com

Rakesh Gandhi
Cisco Systems, Inc.

EMail: rgandhi@cisco.com

Zafar Ali
Cisco Systems, Inc.

EMail: zali@cisco.com

Manav Bhatia
Alcatel-Lucent
India

Email: manav.bhatia@alcatel-lucent.com

Lizhong Jin
Shanghai, China

Email: lizho.jin@gmail.com

Frederic Jounay
Orange CH

Email: frederic.jounay@orange.ch

Network Working Group
Internet Draft
Category: Standards Track

Fatai Zhang
Xian Zhang
Huawei
O. Gonzalez de Dios
Telefonica Investigacion y Desarrollo
C. Margaria. C
Coriant
July 11, 2013

Expires: January 10, 2014

GMPLS-based Hierarchy LSP Creation
in Multi-Region and Multi-Layer Networks

draft-zhang-ccamp-gmpls-h-lsp-mln-05.txt

Abstract

This specification describes the hierarchical LSP creation models in the Multi-Region and Multi-Layer Networks (MRN/MLN), and provides the extensions to the existing protocol mechanisms described in [RFC4206], [RFC6107] and [RFC6001] to create a hierarchical LSP in multiple layer networks.

Status of this Memo

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 January 10, 2014.

Table of Contents

1. Introduction	2
1.1. Conventions used in this document	3
2. Provisioning of FA-LSP in Server Layer Network	3
2.1. Selection of Switching Layers.....	3
2.2. Selection of Switching Granularity Levels	4
2.3. Selection of Adaptation Capabilities	6
3. Signaling Requirements for Server Layer Selection	7
3.1. Model 1: Pre-provisioning of FA-LSP	8
3.2. Model 2: Signaling triggered server layer path computation and setup	9
3.3. Model 3: Signaling triggered server layer path, with explicit server path	9
4. Signaling Extensions ERO Sub-Object	10
4.1. SERVER_LAYER_INFO ERO Subobject	10
4.2. Processing of SERVER_LAYER_INFO sub-object	12
4.3. Alternative Encoding Solutions	12
5. Security Considerations.....	13
6. IANA Considerations	13
7. Acknowledgments	13
8. References	13
8.1. Normative References.....	13
8.2. Informative Reference.....	14
9. Authors' Addresses	15

1. Introduction

Networks may comprise multiple layers which have different switching technologies or different switching granularity levels. The GMPLS technology is required to support control of such network.

[RFC5212] defines the concept of MRN/MLN and describes the framework and requirements of GMPLS controlled MRN/MLN. The GMPLS extension for MRN/MLN, including routing and signaling aspects, is described in [RFC6001].

[RFC4206] and [RFC6107] describe how to set up a hierarchical LSP passing through multi-layer networks and how to advertise the forwarding adjacency LSP (FA-LSP) created in the server layer network as a TE link via GMPLS signaling and routing protocols.

Based on these existing standards, this document further describes the provisioning of a FA-LSP when the region-edge nodes support

multiple interface switching capabilities and/or multiple switching granularities and/or adaptation functions, and then provides the extensions to the RSVP-TE protocol in order to set up a hierarchical LSP according to the modes of hierarchical LSP provisioning.

1.1. 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. Provisioning of FA-LSP in Server Layer Network

2.1. Selection of Switching Layers

As described in [RFC5212], the edge node of a region always has multiple Interface Switching Capabilities (ISCs), i.e., it contains multiple matrices which may be connected to each other by internal links. Nodes with multiple ISCs are further classified as "simplex" or "hybrid" nodes by [RFC5212] and [RFC5339], where the simplex node advertises several TE links each with a single ISC value carried in its ISCD sub-TLV, while the hybrid node advertises a single TE link containing more than one ISCD each with a different ISC value. An example of a hybrid node with a link having multiple ISCs is shown in Figure 1, copied from [RFC5339].

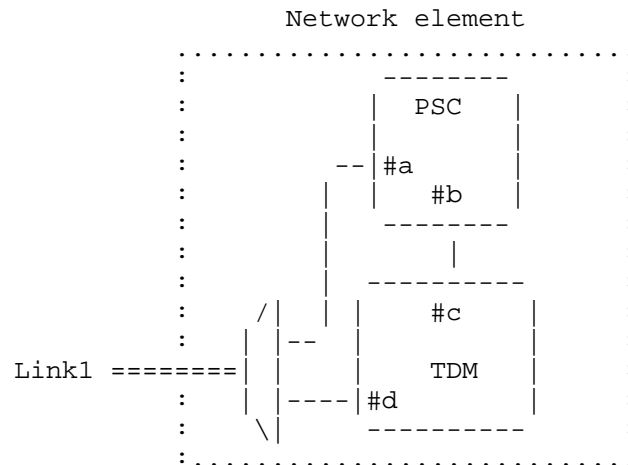


Figure 1 - Hybrid node (Copied from [RFC5339])

In the case where a edge node of a region is a hybrid node, selection of which server layer to create the FA-LSP is necessary.

Figure 2 shows an multi-layer network, where node B and C are region edge nodes having three switching matrices which support, for instance, PSC, TDM and WDM switching, respectively. The three switching matrices are connected to each other by the internal links. Both the link between B and E and the link between E and C support TDM and WDM switching capabilities.

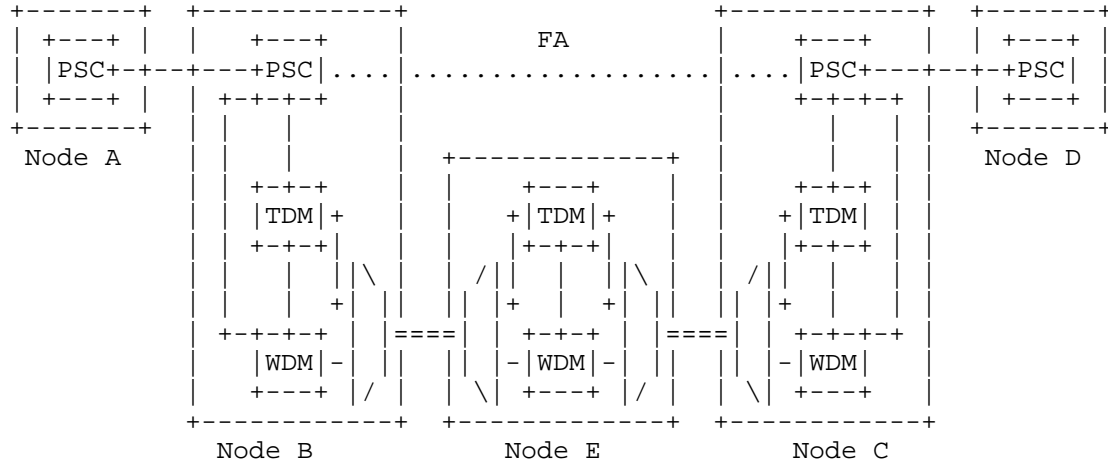


Figure 2 - MLN with multiple ISCs at edge node

As can be seen in Figure 2, there are two choices when providing FA in the PSC layer network between node B and C: one is creating a FA-LSP with TDM switching matrix through node B, E and C, the other is creating a FA-LSP with WDM switching matrix through node B, E and C.

[RFC6001] introduces a new SC (Switching Capability) sub-object into the XRO (ref. to [RFC4874]). This sub-object is used to indicate which switching capability is not expected to be used. When one of the switching capabilities is selected, the SC sub-object can be included in the message to exclude all other SCs.

2.2. Selection of Switching Granularity Levels

Even in the case where the edge node only has one switching capability in the server layer, there may be still multiple choices for the server layer network to set up a FA-LSP to provide new FA in the client layer network. This is because the server layer network may have the capability of providing different switching granularity levels for the FA-LSP.

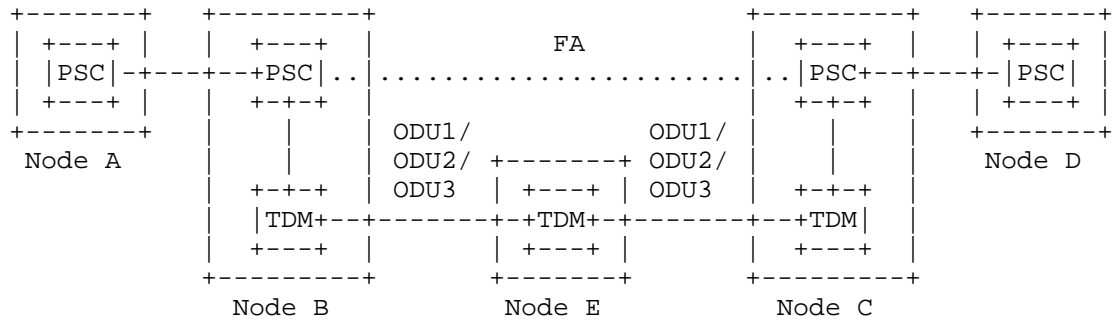


Figure 3a - Multiple switching granularities in server layer

Figure 3a shows an example multi-region network, where the edge node B and C have PSC and TDM switching matrices, and where the TDM switching matrix supports ODU1, ODU2 and ODU3 switching levels. Therefore, when an FA between node B and C in the PSC layer network is needed, either of ODU1, ODU2 or ODU3 connection (FA-LSP) can be created in the TDM layer network.

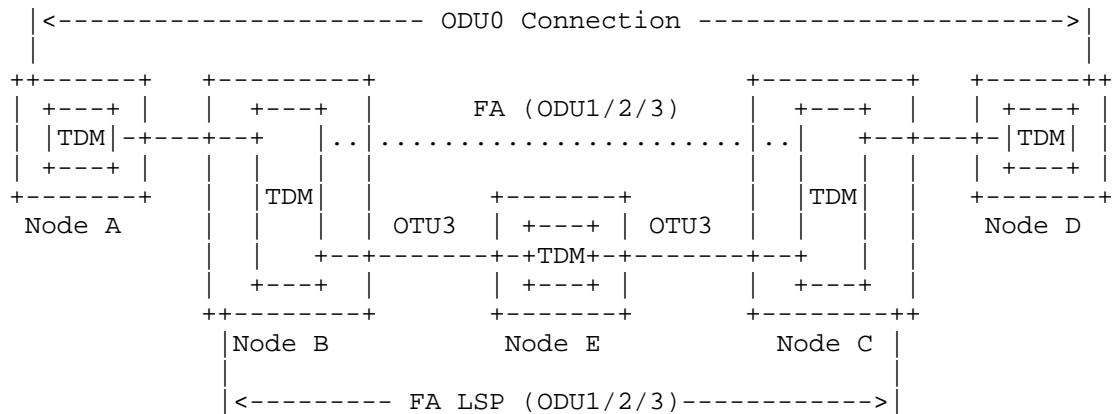


Figure 3b - TDM nested LSP provisioning

Figure 3b is another example multi-layer network within the same region. When there is a need to set up an FA between node B and C for the client layer ODU0 connection, the server layer has multiple

choices, e.g., ODU1 or ODU2 or ODU3, for the FA-LSP if the multi-stage multiplexing is supported at node B and C.

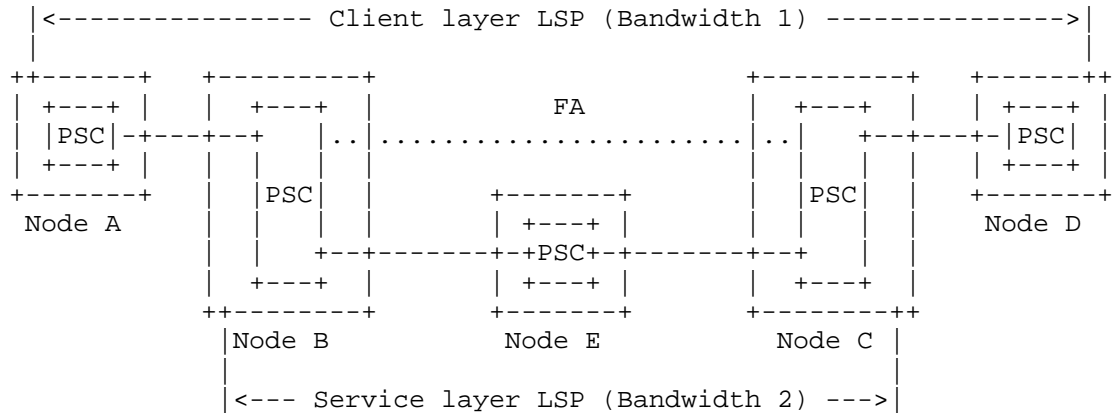


Figure 3c - PSC nested LSP provisioning

Figure 3c is a third example showing an LSP nesting scenario in a PSC signal-layer network (e.g., an MPLS-TP network). A PSC tunnel passing through node B, E and C is requested to carry the client layer LSP. There are multiple choices of the bandwidth of the tunnel, on the premise that the bandwidth of the FA-LSP is equal to or larger than the client layer LSP.

The selection of server layer switching matrix and switching granularity is based on both policy and bandwidth resources. The selection can be performed by a planning tool and/or NMS/PCE/VNTM (Virtual Network Topology Manager, see [RFC5623]) and/or the network node.

2.3. Selection of Adaptation Capabilities

Adaptation function also needs to be selected when creating the server layer connection. This is because the edge nodes may support multiple adaptation functions.

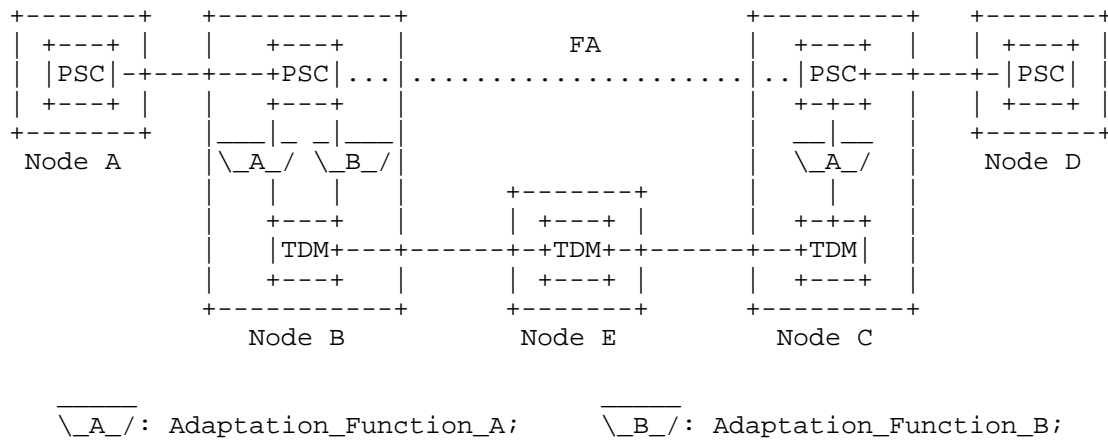


Figure 4 - Selection of adaptation function

For example, in Figure 4, edge node B supports two adaptation functions, i.e., adaptation_function_A and adaptation_function_B, while edge node C only supports adaptation_function_A. In this case, only adaptation_function_A can be used for the server layer connection.

The Call procedure ([RFC4974]) may be used between edge node B and C to negotiate and determine the adaptation function for the server layer if the Call function is supported.

3. Signaling Requirements for Server Layer Selection

[RFC5623], the framework of PCE-based MLN, provides the models of cross-layer LSP path computation and creation, which are listed below:

- Inter-Layer Path Computation Models:
 - o Single PCE
 - o Multiple PCE with inter-PCE
 - o Multiple PCE without inter-PCE
- Inter-Layer Path Control Models:
 - o PCE-VNTM cooperation

- o Higher-layer signaling trigger
- o NMS-VNTM cooperation (integrated flavor)
- o NMS-VNTM cooperation (separate flavor)

This section keeps alignment with [RFC5623] except that the restriction of using a PCE for path computation is not necessary (i.e., other element, such as a network node, may also have path computation capability).

In this document, those models in [RFC4206] are mapped into 3 models on the viewpoint of signaling:

- Model 1: Pre-provisioning of FA-LSP
- Model 2: Signaling triggered server layer path computation and setup
- Model 3: Signaling triggered server layer path, with explicit server path.

3.1. Model 1: Pre-provisioning of FA-LSP

In this model, the FA-LSP in the server layer is created before initiating the signaling of the client layer LSP. Two typical scenarios using this model are:

- Network planning and building at the stage of client network initialization.
- NMS/VNTM triggering the creation of FA-LSP when computing the path of client layer LSP. The path control models of PCE-VNTM cooperation and NMS-VNTM cooperation (both integrated and separate flavor) in [RFC5623] belong to this scenario.

In such case, the server layer selection and path computation is performed by planning tool or NMS/PCE/VNTM or the edge node. The signaling of client layer LSP and server layer FA-LSP are separated. The normal LSP creation procedures ([RFC3471] and [RFC3473]) are followed to set up these two LSPs and no new extension is required.

3.2. Model 2: Signaling triggered server layer path computation and setup

In this model, the source node of client layer LSP only computes the route within its own layer network. When the signaling of the client layer LSP reaches at the region edge node, the edge node performs server layer FA-LSP path computation and then creates the FA-LSP. When a PCE is introduced to perform path computation in each layer of the multi-layer network, this model is the same as the model of "higher-layer signaling trigger with Multiple PCE without inter-PCE" in [RFC5623].

In such case, the edge node will receive the client layer PATH message with a loose ERO indicating an FA is requested, and may perform the server layer selection (e.g., through the server layer PCE or the VNTM) and then compute and set up the FA-LSP. The signaling procedure of client layer LSP and server layer FA-LSP is described in detail in [RFC4206] and [RFC6107].

It's possible that the source node of the client layer LSP selects the server layer SC and/or granularity and/or adaptation function when performing path computation in the client layer, and requests or suggests the edge node to use an appointed server layer to create the FA-LSP.

In this case, the XRO including SC sub-object ([RFC6001]) is adopted for the server layer SC exclusion, which can be used indirectly to select server layer SC. Such solution is not straightforward enough. Furthermore it cannot be used for the selection of server layer granularity and adaptation function. Therefore, new extensions for the selection of server layer SC, switching granularity and adaptation function are required.

3.3. Model 3: Signaling triggered server layer path, with explicit server path

In this model, the source node of the client layer LSP performs a full path computation including the client layer and the server layer routes. The server layer FA-LSP creation is triggered at the edge node by the client layer LSP signaling. When a PCE is introduced to perform path computation in the multi-layer network, this model is the same as the model of "Higher-layer signaling trigger with Single PCE" or "Higher-layer signaling trigger with Multiple PCE with inter-PCE" in [RFC5623].

In such case, the server layer selection and server layer path computation is performed at the source node of the client layer LSP (e.g., through VNTM or PCE), but not at the edge node.

In [RFC4206], the ERO which contains the list of nodes and links (including the client layer and server layer) along the path is used in the client layer PATH message. The edge node can find out the tail end of the FA-LSP based on the switching capability of the node using the IGP database (see session 6.2 of [RFC 4206]).

Similar to the problem of model 2, the edge node is not aware of which switching granularity and which adaptation function to be selected for the FA-LSP because the ERO and/or XRO do not contain such information. Therefore, the edge node may not be able to create the FA-LSP, or may select another switching granularity by itself which is different from the one selected previously at the source node, which makes the creation of hierarchy LSP out of control.

Therefore, new extensions for the selection of server layer SC, switching granularity and adaptation function are also required in this model.

4. Signaling Extensions ERO Sub-Object

4.1. SERVER_LAYER_INFO ERO Subobject

In order to solve the problems described in the previous sections, a new sub-object named SERVER_LAYER_INFO sub-object is introduced in this document, which is carried in the ERO and is used to explicitly indicate which server layer to create the FA-LSP.

The SERVER_LAYER_INFO sub-object is put immediately after the node or link (interface) address sub-object, indicating the related node is a region edge node on the LSP in the ERO.

The format of the SERVER_LAYER_INFO sub-object is shown below:

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|L|      Type      |      Length      |M|      Reserved      |
+-----+-----+-----+-----+-----+-----+-----+-----+
| LSP Enc. Type |Switching Type |      G-PID      |
+-----+-----+-----+-----+-----+-----+-----+-----+
|      Traffic Spec Length      | TSpec Type      | Reserved      |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     Traffic Parameters                                     |
~                                                                 ~
+-----+-----+-----+-----+-----+-----+-----+-----+

```

[Editor's note: the encoding is still under discussion.]

- L bit: MUST be zero and MUST be ignored when received.
- Type: The SERVER_LAYER_INFO sub-object has a type of xx (TBD).
- Length: The total length of the sub-object in bytes, including the Type and Length fields. The value of this field is always a multiple of 4.
- M (Mandatory) bit: When set, it means the edge node MUST set up the FA-LSP in the appointed server layer; otherwise, the appointed server layer is suggested and the edge node may select other server layer by local policy.
- LSP Encoding Type, Switching Type and G-PID: These 3 fields are used to point out which switching layer is requested to set up the FA-LSP. The values of these 3 fields are inherited from the Generalized Label Request Object in GMPLS signaling, referring to [RFC3471], [RFC3473] and other related standards and drafts. Note that G-PID can be used to indicate the payload type of the server layer (i.e., the client signal) as well as the adaptation function for adapting the client signal into the server layer FA-LSP.
- Traffic Spec Length, TSpec Type, Traffic Parameters: The traffic parameters field is used to indicate the switching granularity of the FA-LSP. The format of this field depends on the TSpec Type Traffic Spec Length and is consistent with the existing standards and drafts. For example, the traffic parameters of Ethernet, SONET/SDH and OTN are defined in [RFC6003], [RFC4606] and [OTN-ctrl] respectively.

4.2. Processing of SERVER_LAYER_INFO sub-object

As described in RFC3209 and RFC3473 the ERO is managed as a sub-object list. The SERVER_LAYER_INFO sub-object MUST be appended after the existing sub-object defined in [RFC3209], [RFC3473], [RFC3477], [RFC4873], [RFC4874], [RFC5520] and [RFC5553] TBD:extensions.

When a node receives a PATH message containing ERO and finds that there is a SERVER_LAYER_INFO sub-object immediately after the node or link address sub-object related to itself, the node determines that it's a region edge node. Then, the edge node finds out the server layer selection information from the sub-object:

- Determine the switching layer by the LSP Encoding Type and Switching Type fields;
- Determine the switching granularity of the FA-LSP by the Traffic Parameters field;
- Determine the adaptation function for adapting the client signal into the server layer FA-LSP by the G-PID field.

The edge node MUST then determine the other edge of the region, i.e., the tail end of the FA-LSP, with respect to the subsequence of hops of the ERO. The node that satisfies the following conditions will be treated as the tail end of the FA-LSP:

- There is a SERVER_LAYER_INFO sub-object that immediately follows the node or link address sub-object which is related to that node;
- The LSP Encoding Type, Switching Type, G-PID and the Traffic Parameters fields of this SERVER_LAYER_INFO sub-object is the same as the SERVER_LAYER_INFO sub-object corresponding to the head end;
- The node is the first one that satisfies the two conditions above in the subsequence of hops of the ERO.

If a match of tail end is found, the head end now has the clear server layer information of the FA-LSP and then initiates an RSVP-TE session to create the FA-LSP in the appointed server layer between the head end and the tail end.

4.3. Alternative Encoding Solutions

[Editor's note: the section is still under discussion.]

A first alternative solution is to use the mechanism defined in [LSP-RO], i.e., create an ERO HOP attribute TLV.

The content and procedure are not changed from the previous section.

5. A second alternative solution aims to simplify the SERVER_LAYER_INFO processing by using the SERO mechanisms. This can be a new requirements to the SERO or to the ERO Hop attribute. This alternative is not further described here but mentioned for discussions.

6. Security Considerations

TBD.

7. IANA Considerations

TBD.

8. Acknowledgments

TBD.

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.
- [RFC3945] Mannie, E., "Generalized Multi-Protocol Label Switching (GMPLS) Architecture", RFC 3945, October 2004.
- [RFC3209] D. Awduche et al, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC3209, December 2001.
- [RFC3471] Berger, L., Ed., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, January 2003.
- [RFC3473] L. Berger, Ed., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.

- [RFC5212] K. Shiimoto et al, "Requirements for GMPLS-Based Multi-Region and Multi-Layer Networks (MRN/MLN)", RFC5212, July 2008.
- [RFC5339] JL. Le Roux et al, "Evaluation of Existing GMPLS Protocols against Multi-Layer and Multi-Region Networks (MLN/MRN)", RFC5339, September 2008.
- [RFC4206] K. Kompella et al, "Label Switched Paths (LSP) Hierarchy with Generalized Multi-Protocol Label Switching (GMPLS) Traffic Engineering (TE)", RFC4206, October 2005.
- [RFC6107] K. Shiimoto, A. Farrel, "Procedures for Dynamically Signaled Hierarchical Label Switched Paths", RFC6107, February 2011.
- [RFC6001] Dimitri Papadimitriou et al, "Generalized Multi-Protocol Label Switching (GMPLS) Protocol Extensions for Multi-Layer and Multi-Region Networks (MLN/MRN)", RFC6001, October, 2010.

9.2. Informative Reference

- [RFC4974] D. Papadimitriou and A. Farrel, "Generalized MPLS (GMPLS) RSVP-TE Signaling Extensions in Support of Calls", RFC4974, August 2007.
- [RFC5623] E. Oki et al, "Framework for PCE-Based Inter-Layer MPLS and GMPLS Traffic Engineering", RFC 5623, September 2009.
- [RFC4606] E. Mannie, D. Papadimitriou, "Generalized Multi-Protocol Label Switching (GMPLS) Extensions for Synchronous Optical Network (SONET) and Synchronous Digital Hierarchy (SDH) Control", RFC 4606, August 2006.
- [OTN-ctrl] Fatai Zhang et al, "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Extensions for the evolving G.709 Optical Transport Networks Control", draft-ietf-ccamp-gmpls-signaling-g709v3-08.txt, April, 2013.
- [RFC6003] D. Papadimitriou, "Ethernet Traffic Parameters", RFC6003, October, 2010.
- [LSP-RO] Margaria, C., Giovanni, G., et al, "draft-ietf-ccamp-lsp-attribute-ro", draft-ietf-ccamp-lsp-attribute-ro-01.txt, work in progress;

10. Authors' Addresses

Fatai Zhang
Huawei Technologies
F3-1B R&D Center, Huawei Base
Bantian, Longgang District
Shenzhen 518129 P.R.China

Phone: +86-755-28972603
Email: zhangfatai@huawei.com

Xian Zhang
Huawei Technologies
F3-1B R&D Center, Huawei Base
Bantian, Longgang District
Shenzhen 518129 P.R.China

Phone: +86-755-28972645
Email: huawei.danli@huawei.com

Yi Lin
Huawei Technologies Co., Ltd.
F3-1B R&D Center, Huawei Base
Bantian, Longgang District
Shenzhen 518129 P.R.China

Phone: +86-755-28972597
Email: yi.lin@huawei.com

Oscar Gonzalez de Dios
Telefonica Investigacion y Desarrollo
Emilio Vargas 6
Madrid, 28045 Spain

Phone: +34 913374013
Email: ogondio@tid.es

Cyril Margaria
Coriant GmbH
St Martin Strasse 76
Munich, 81541
Germany

Phone: +49 89 5159 16934
Email: cyril.margaria@coriant.com

Intellectual Property

The IETF Trust takes no position regarding the validity or scope of any Intellectual Property Rights or other rights that might be claimed to pertain to the implementation or use of the technology described in any IETF Document or the extent to which any license under such rights might or might not be available; nor does it represent that it has made any independent effort to identify any such rights.

Copies of Intellectual Property disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the IETF on-line IPR repository at <http://www.ietf.org/ipr>

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement any standard or specification contained in an IETF Document. Please address the information to the IETF at ietf-ipr@ietf.org.

The definitive version of an IETF Document is that published by, or under the auspices of, the IETF. Versions of IETF Documents that are published by third parties, including those that are translated into other languages, should not be considered to be definitive versions of IETF Documents. The definitive version of these Legal Provisions is that published by, or under the auspices of, the IETF. Versions of these Legal Provisions that are published by third parties, including those that are translated into other languages, should not be considered to be definitive versions of these Legal Provisions.

For the avoidance of doubt, each Contributor to the IETF Standards Process licenses each Contribution that he or she makes as part of

the IETF Standards Process to the IETF Trust pursuant to the provisions of RFC 5378. No language to the contrary, or terms, conditions or rights that differ from or are inconsistent with the rights and licenses granted under RFC 5378, shall have any effect and shall be null and void, whether published or posted by such Contributor, or included with or in such Contribution.

Disclaimer of Validity

All IETF Documents and the information contained therein are provided on an "AS IS" basis and THE CONTRIBUTOR, THE ORGANIZATION HE/SHE REPRESENTS OR IS SPONSORED BY (IF ANY), THE INTERNET SOCIETY, THE IETF TRUST AND THE INTERNET ENGINEERING TASK FORCE DISCLAIM ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION THEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Copyright Notice

Copyright (c) 2013 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.

Network Working Group
Internet Draft
Category: Informational

Fatai Zhang
Huawei
O. Gonzalez de Dios
Telefonica Investigacion y Desarrollo
A. Farrel
Old Dog Consulting
Xian Zhang
Huawei
D. Ceccarelli
Ericsson
July 09, 2013

Expires: January 08, 2014

Applicability of Generalized Multiprotocol Label Switching (GMPLS)
User-Network Interface (UNI)

draft-zhang-ccamp-gmpls-uni-app-04.txt

Status of this Memo

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 January 09, 2014.

Abstract

Generalized Multiprotocol Label Switching (GMPLS) defines a set of protocols for the creation of Label Switched Paths (LSPs) in various switching technologies. The GMPLS User-Network Interface (UNI) was developed in RFC4208 in order to be applied to an overlay network architectural model.

This document examines a number of GMPLS UNI application scenarios. It shows how techniques developed after the GMPLS UNI can be applied to automate or enable critical processes for these applications. This document also suggests simple extensions that could be made to existing technologies to further enable the UNI and points out some unresolved issues.

Table of Contents

1. Introduction	3
2. UNI Addressing	5
3. UNI Auto Discovery	6
4. UNI Path Computation.....	7
4.1. UNI Link Selection.....	8
5. Additional Parameters across UNI.....	10
5.1. Constrained Path Computation.....	10
5.2. Collection Requests over UNI.....	11
6. UNI Path Provisioning Models.....	11
6.1. Flat Model	12
6.2. Stitching Model.....	12
6.3. Session Shuffling Model.....	13
6.4. Hierarchal Model.....	13
7. UNI Recovery	14
7.1. End-to-end Recovery.....	15
7.1.1. Serial Provisioning of Working and Protection Paths	15
7.1.2. Concurrent Computation of Working and Protection Path	16
7.2. Segment Recovery.....	17
8. UNI Call	17
8.1. Exchange of UNI Link Information.....	18
8.2. Control of Call Route.....	18
9. UNI Multicast	19
9.1. UNI Multicast Connection Model	19
9.2. UNI Multicast Connection Provisioning	21
10. Security Considerations.....	22
11. IANA Considerations.....	22
12. Acknowledgments	22
13. References	23
13.1. Normative References.....	23
13.2. Informative References.....	25
14. Contributors' Address.....	26
15. Authors' Addresses	27

1. Introduction

Generalized Multiprotocol Label Switching (GMPLS) [RFC3945] defines a set of protocols, including Open Shortest Path First - Traffic Engineering (OSPF-TE) [RFC4203] and Resource ReserVation Protocol - Traffic Engineering (RSVP-TE) [RFC3473], which can be used to create Label Switched Paths (LSPs) in a number of deployment scenarios with various transport technologies.

The User-Network Interface (UNI) reference point is defined in the Automatically Switched Optical Network (ASON) [G.8080]. According to [G.8080], the UNI may be implemented as a peering between a client-side entity (UNI-C) and a network-side entity (UNI-N). End-to-end connectivity between UNI-C nodes is achieved across the core network by three components: a UNI request from source UNI-C to source UNI-N; a core network connection from source UNI-N to destination UNI-N; and a UNI request from destination UNI-N to destination UNI-C.

The GMPLS overlay model, as per [RFC4208], can be applied at the UNI, as shown in Figure 1.

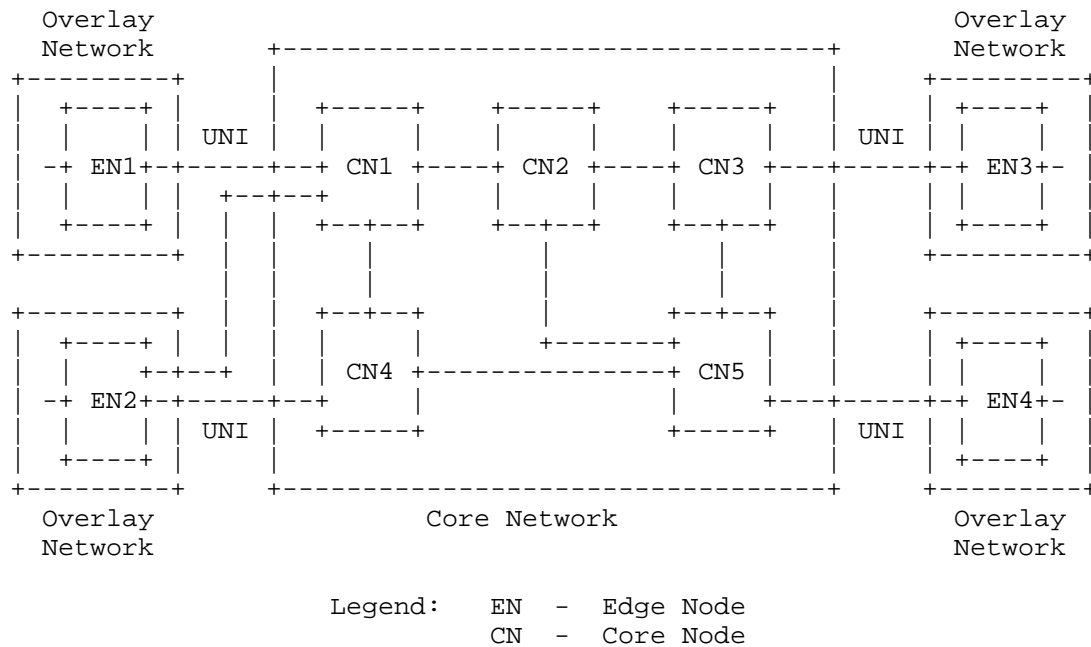


Figure 1 - Applying GMPLS overlay model at UNI

In Figure 1, assume that there is an end-to-end UNI connection passing through EN1-CN1-CN2-CN3-EN3. For convenience, some terms used in this document are defined below:

- "source EN" refers to the edge-node which initiates the connection (i.e., EN1);
- "destination EN" refers to the edge-node where the connection is terminated (i.e., EN3);
- "ingress CN" refers to the core-node to which the source EN is attached (i.e., CN1);
- "egress CN" refers to the core-node to which the destination EN is attached (i.e., CN3).

[RFC4208] provides mechanisms for UNI signaling, which are compatible with GMPLS RSVP-TE signaling ([RFC3471] and [RFC3473]). A single end-to-end RSVP session between source EN and destination EN is used for the user connection, just as it would be for connection creation between two core nodes. However, when considering the isolation of topology information between the core network and the overlay network, additional processing of the RSVP-TE Explicit Route Object (ERO) and Record Route Object (RRO) is required. For example, the ingress CN should verify the ERO it receives against its topology database and may enhance it with additional path information before forwarding the PATH message. And the ingress/egress CN may edit or remove the RRO in order to hide the path segment used inside the core network from the EN.

The GMPLS UNI can be used in many application scenarios. For example, in a multi-layer network [RFC6001] the interface between client layer node and server layer node can be seen as a UNI. Or, when deploying VPN services such as Layer One Virtual Private Networks (L1VPNs) [RFC4847], [RFC5253], users can connect to a service provider network via a UNI.

This document examines a number of current and future GMPLS application scenarios. It shows how techniques developed after the GMPLS UNI can be used to automate or enable critical aspects of these application scenarios. It points out some potential technology extensions that could improve UNI operation, and highlights some unresolved issues.

2. UNI Addressing

In [RFC4208], the GMPLS overlay model is applied at the UNI reference point, and it is required that the edge-node and its attached core-node of the overlay network share the same address space that is used by GMPLS to signal between the edge-nodes across the core network. Under this condition, the user connection can be created using a single end-to-end RSVP session, which is consistent with the RSVP model. Therefore, RSVP-TE defined in [RFC3473] can be used for support GMPLS UNI without any extensions.

However, in some deployments of the GMPLS UNI, it is not practical for the EN and its attached CN to share the same address space. This can arise if the core and overlay networks were designed and deployed separately or belong to different carriers. For example, the core network may use IPv6 addresses, while the overlay network uses IPv4 addresses. Or, since the core network is a closed system, the assignment of the IP addresses of the CNs may be independent of other IP addresses outside the core network. This implies that the nodes in the core network may use addresses which could collide with the edge nodes in the overlay network.

[RFC4208] does not state how to ensure that an edge-node and its attached core-node share the same address space. This document analyses the addressing deployment scenarios as follows:

1. Overlay network and core network share a common addressing policy. This might be quite feasible in a multi-layer network operated by a single carrier.

In this scenario, end-to-end UNI connectivity may use a single RSVP session, and the core routing information (assuming it is shared and not stripped for confidentiality reasons) will be meaningful to the ENs. Note, however, that the overlay model examined by this document assumes that there is some separation between the overlay and core networks, and this might mean that the overlay network is not able to see the topology or routing information of the core network even when they share a common address space.

2. ENs have visibility into the core network, but overlay and core networks have different address spaces. This is the more common model envisaged by [RFC4208] and for basic mode L1VPN deployments [RFC5251]. The previous scenario can be seen to be a special case of this scenario where the two address spaces are complementary. In this deployment the ENs each have two addresses: one in the overlay network and one in the core network. The source EN is

aware of the addresses for itself, the ingress CN, the egress CN, and the destination EN in the address space of the core network. It may also have full visibility into the core network, but this is not a requirement.

In this scenario, the ENs are responsible for performing address mapping between the overlay network's addresses for the ENs, and the core network's addresses for the same nodes and/or its TE links. A typical deployment may assign addresses in the core network address space for the EN and/or its TE links at the EN side, so that EN can use these addresses to communicate with the core network for UNI connection provisioning.

In this deployment, a single end-to-end RSVP-TE session can still be utilized from the source EN to the destination EN using addressing and naming from the core network's address space.

3. ENs do not have any knowledge of the core address space, or do not support the address space the core network uses (e.g., ENs do not support IPv6 that is used by the core network). ENs will have no visibility into the core network.

In this scenario, the ingress CN is responsible for mapping addresses to the core address space and filling in any additional routing information. A typical deployment is to assign addresses in the overlay address space for the ingress CN and/or its TE links at the CN side, so that the EN can use overlay addresses to reach the ingress CN and to identify the destination EN.

In this deployment the end-to-end connectivity must be created either using "session stitching" (see Section 6.2) or "session shuffling" (see Section 6.3).

3. UNI Auto Discovery

When the end-to-end connection is set up across the core network, it must be targeted at the destination CN so that it can be extended to the destination EN. This means that either the source EN must know the identity of the destination CN to which the destination EN is attached, or the source CN must know this information. This requires some form of "discovery" (possibly including configuration), and depending on the addressing scheme in use (see Section 2), address mapping needs to be performed by the source EN or the source CN.

The discovery problem may be exacerbated when a variety of services are requested since the source EN will need to know the capabilities and available resources on the link between the destination CN and the destination EN. It could discover this by attempting to set up a connection and by drawing conclusions from connection setup failures, but this is not efficient. Furthermore, in the case of a dual-homed destination EN (such as EN2 in Figure 1), a choice of destination CN must be made, and that choice may be influenced by the capabilities and available resources on the CN-EN links leading to the destination EN.

If the UNI is applied in an L1VPN scenario, two mechanisms for auto discovery have been defined. Auto discovery of UNI using OSPFv2 is provided in [RFC5252] using an L1VPN LSA to advertise the L1VPN information via the L1VPN info TLV and the TE information of the CE-PE link (in the language of UNI, it's the EN-CN link) via the TE link TLV. Auto discovery of UNI using BGP is provided in [RFC5195] by having each edge CN advertise to other edge CN the following information, at a minimum: its own IP address and the list of <private address, provider address> tuples local to that PE. Once that information is received, the remote PEs will identify the list of VPN members they have in common with the advertising PE, and use the information carried within the discovery mechanism to perform address resolution during the signaling phase of Layer-1 VPN connections.

4. UNI Path Computation

End-to-end UNI path computation includes three parts: the selection of the source UNI link, the path computation inside the core network and the selection of the destination UNI link.

The selection of UNI links may not be necessary in all scenarios. One example is in the case of single-homing with only one UNI link between EN and CN. Another example is manual selection of the UNI link when the service is requested (i.e., as a function of the service request such as the port mapping used in a L1VPN). In such cases, the CN to which the source EN is attached, or the path Computation Element (PCE) ([RFC4655]) which is responsible for the core network, can perform the path computation across the core network when the UNI signaling request is sent from the source EN to the source CN.

4.1. UNI Link Selection

This document is specific to the overlay architectural model, and that means that the source EN does not have the topology and TE information of the core network. Therefore, in the case of multi-homing (i.e., the source EN is connected to more than one CN), the source EN does not have enough information to make a correct choice among all the UNI links between itself and the core network for an optimal end-to-end connection.

In this case, a PCE whose computation domain covers both the core network and the ENs attached to it can be used. Note that the GMPLS UNI predates PCE and hence a PCE was not available in early GMPLS UNI deployments. A PCE that has the topology and TE information of the core network can use the UNI discovery mechanism described in Section 3 to learn the EN-CN relationship and the TE information of the UNI links, and therefore has the ability to select the optimal UNI link for the connection.

Figure 2 shows the procedures for UNI path computation using a single PCE with visibility into the core network and information about all of the CN-EN links. When the UNI path computation request is received, the PCE can help the source EN to compute the end-to-end route of the UNI connection based on routing information it has access to, so that the source EN can create the UNI connection using the optimal UNI link. As shown in Figure 2, the following steps are carried out:

Step 1: EN1 requests a path from EN1 to EN2 by sending a PCReq message to the PCE;

Step 2: The PCE computes a path based on its view of the core network and knowledge of all the EN-CN links. In this case, it returns the path EN1-CN4-CN5-CN6-EN2 to the EN1 node;

Step 3: EN1 starts the signaling process to set up the LSP by using a standard RSVP signaling process, using the path information as computed.

If confidentiality of the topology within the core network needs to be preserved, the Path Key Subobject (PKS) can be used for either approach outlined here (see [RFC5520] and [RFC5553]). In the PCRep message returned to EN1, the Confidential Path Segment (CPS) (i.e., CN4-CN5-CN6) is encoded as a PKS by the PCE. Therefore, EN1 only learns the selected UNI link from the PCE. When CN4 receives the UNI signaling message from EN1 carrying the PKS, CN4 asks the PCE to decode the PKS and then continues to signal the LSP.

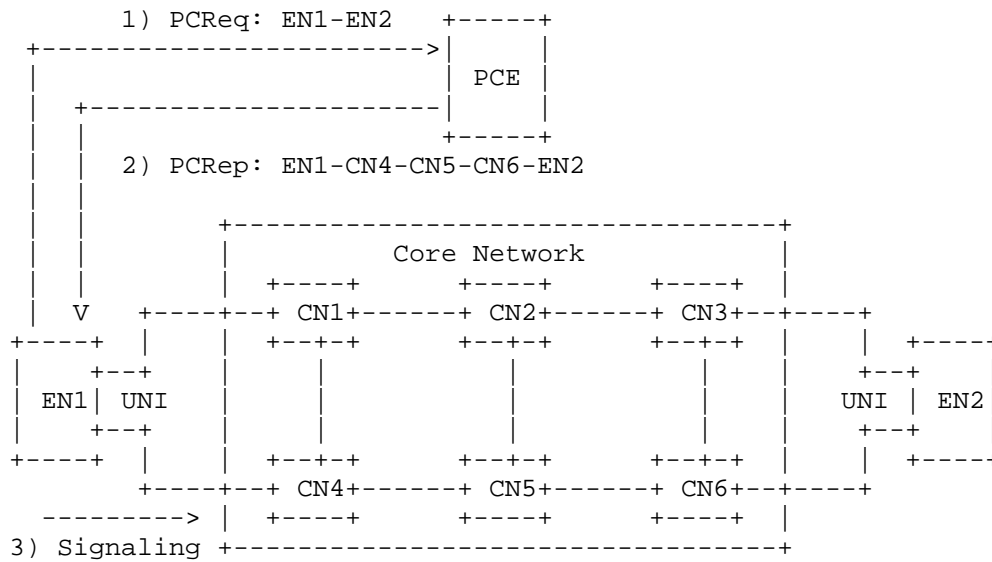


Figure 2 - Procedure using a PCE for UNI path computation

Note that the case described in this section, the PCE needs to be visible to the ENs, and there also needs to be a control channel between the PCE and the ENs for the exchange of PCE Protocol (PCEP) messages. An alternative implementation could be that a PCE is located inside each CN to which the source EN is attached, so that the source EN can use the UNI control channel to send and receive the PCEP messages.

The node requesting for a LSP, crossing UNI, may not be an EN node, as depicted in Figure 3. The procedure described above still applies.

In this case, if an explicit route is desired there is an additional requirement that the PCE needs to have visibility into the overlay networks. Otherwise, the PCE can only provide the route between two EN nodes as illustrated in Figure 3.

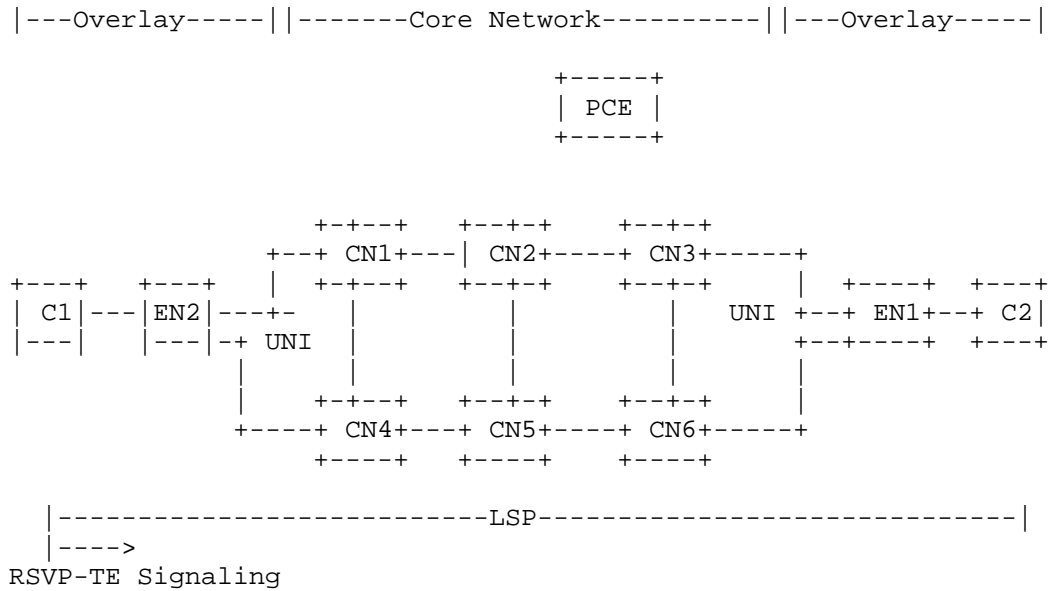


Figure 3 - Procedure of non-EN Node in the Overlay Path Computation

5. Additional Parameters across UNI

With new extensions currently proposed for RSVP-TE protocol, new parameters/functions can also be applicable to UNI.

5.1. Constrained Path Computation

Constraints that can be applied to the path computation in the core network are:

- + Diversity: it is possible to indicate the resources must or should be avoided during the path computation by means of the Exclude Router Object (XRO) [RFC4874], the Explicit Exclusion Route Subobject (EXRS) [RFC4874] and the LSP subobject [LSP-DIV]. Such resources can consist of:

- IPv4 prefix, IPv6 prefix, Unnumbered Interface ID, AS Number and SRLG [RFC4874]

- IPv4 P2P subobject and IPv6 P2P subobject [LSP-DIV]

- + Latency, Latency Variation and Cost: max delay/delay variation and cost allowed by the server layer LSP [UNI PLUS]

The overlay Edge Node can include into the RSVP-TE Path message an arbitrary number of path computation constraints for the provisioning of the LSP in the server domain. For example, in Figure 2, EN1 can request a path with a constraint: max latency should be 200ms.

If the path computation in the core network is able to provide an LSP meeting the requirements (at least those requirements which must be met) such LSP is established and a RESV message is returned to the Edge node; otherwise an error message (PathErr) is returned.

Use cases described in Section 7 can be viewed as a special use case of diversity.

5.2. Collection Requests over UNI

In addition to the path request with path computation constraints, the overlay nodes can also request for the collection in the core network of the effective values of the parameters indicated as path computation constraints. The collection of such parameters is indicated via dedicated flags in the LSP_ATTRIBUTES and LSP_REQUIRED_ATTRIBUTES in Path Message. Flags defined are:

- Cost collection flag [TE-REC]
- Latency collection flag [TE-REC]
- Latency Variation collection flag [TE-REC]
- SRLG collection flag [SRLG-FA]

In the scenario depicted in Figure 2 a request with constraints on max latency might be issued together with the request of collecting e.g. the effective SRLGs of the provided path, in order to set up a SRLG-disjoint recovery path, as explained in Section 7. Collected parameters are returned to the overlay edge node via the Record Route Object (RRO) in the RESV message.

6. UNI Path Provisioning Models

The basic GMPLS UNI application is to provide end-to-end connections between edge-nodes through a core network via the overlay model. This section briefly describes four ways in which the end-to-end LSP can be created and operated across the core network.

6.1. Flat Model

In this model, the edge-nodes have the same switching capability as the nodes in the core network. In this case, one single end-to-end RSVP session through the edge-nodes and a series of core-nodes can be used to create the connection, which forms a flat LSP model, as shown in Figure 4.

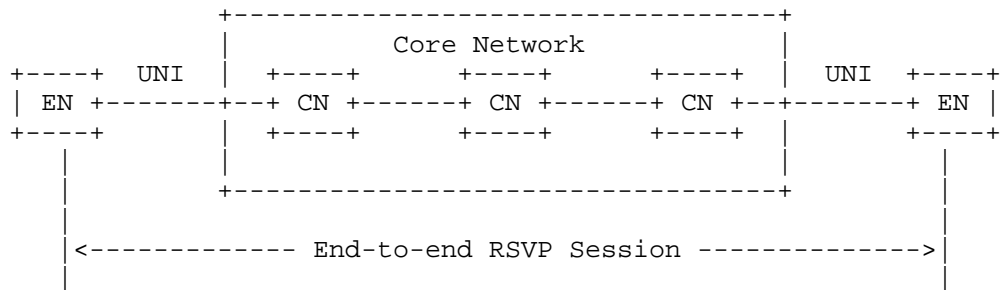


Figure 4 - The Flat Model

If the edge-nodes and their attached core-nodes share the same address space, or the ENs can perform address mapping into the core network address space, the GMPLS signaling described in [RFC3471], [RFC3473] and other related specifications, with special ERO and RRO processing as described in [RFC4208], can be used to create a connection. Note the procedures mentioned still apply in the scenarios where the source node of a connection is not an edge-node but rather nodes within the same domain as EN.

6.2. Stitching Model

The stitching mechanism described in [RFC5150] can be used to create an LSP segment (S-LSP) between the ingress and the egress CN, and to stitch the end-to-end UNI connection to the created S-LSP, as shown in Figure 5.

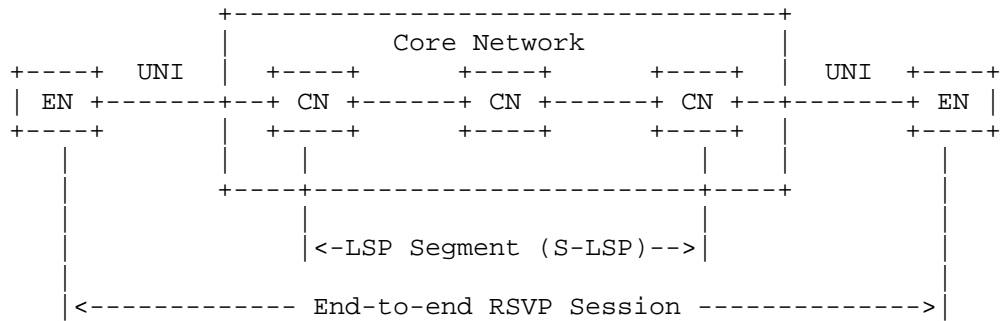


Figure 5 - The Stitching Model

This model allows the core network a degree of independence so that the S-LSP can be set up and modified without the knowledge of the overlay network. Remember that stitching is a data plane function, so that the EN-CN LSP segments are cross-connected to the S-LSP at the edge CNs. This means that, just as in Section 6.1, the overlay and core networks must have the same switching capabilities. However, the control plane for the stitching model operates just as the hierarchical model described in Section 6.4, so the S-LSP appears as a single hop in the overlay network.

6.3. Session Shuffling Model

The session shuffling approach ([RFC5251]) is a modification of the flat model described in Section 6.1. In this approach a single end-to-end session is established, but as the signaling messages pass through the ingress and egress CNs, address mapping is performed on all addresses carried by the messages to replace the addresses with values from the correct address space. The ERO and RRO are stripped from the messages as previously discussed, so there is no need for the CNs to examine those objects to map addresses. However, all other addresses must be mapped including the important session identifiers (the source and destination addresses). Viewed from the outside (perhaps through an NMS) this gives the impression of session stitching because the session has different identifiers as it crosses the core network. An NMS might, therefore, present the shuffling model as the stitching model, or it might operate the same address shuffling/mapping as is used by CNs.

6.4. Hierarchical Model

If the ENs and CNs have the same switching capability, a tunnel between the ingress and egress core-nodes can be provisioned to carry the end-to-end connection. The tunnel may have a larger capacity than

the end-to-end UNI connection, depending on the policies configured at the ingress CN of the core network. The end-to-end connection can be nested into a tunnel, which forms the LSP hierarchy [RFC4206] as shown in Figure 6. If the tunnel has a larger capacity, other LSPs can also be nested within the same tunnel.

Alternatively, if the ENs and CNs have different switching capabilities the LSP hierarchical model can also be used exactly as described in [RFC4206].

In the hierarchal model, the end-to-end connection can be divided into three hops: one for each UNI link and one hop across the core network. The core network tunnel can be pre-provisioned via network planning, or triggered by the UNI signalling. For the latter case, [RFC5212], [RFC6001] and other multi-layer network related specifications can be used to create the hierarchical LSP.

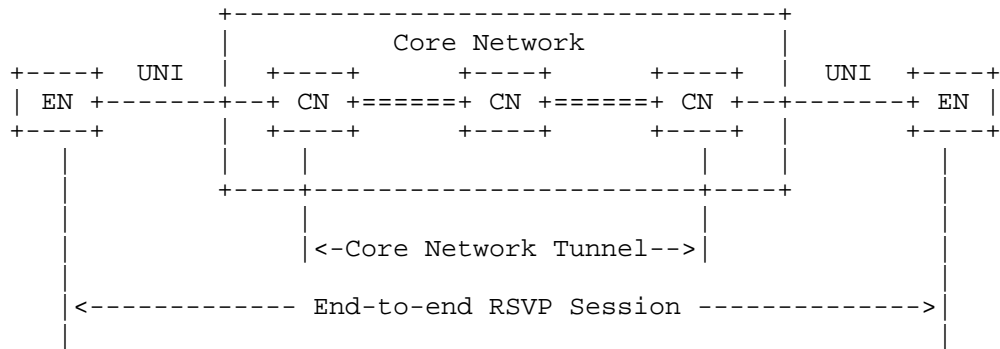


Figure 6 - The Hierarchal Model

7. UNI Recovery

One of the significant uses of GMPLS is to provide recovery mechanisms for connections. Recovery and protection mechanisms are also needed in many UNI scenarios, and the relationship between the overlay and core network provide obvious places at which to operate the recovery techniques.

7.1. End-to-end Recovery

In the case of multi-homing, UNI end-to-end recovery is possible. As shown in Figure 7, the working path (W) and the protection path (P) are disjoint from each other not only inside the core network, but also at both the source and destination sides of the UNI. Mechanisms need to be provided to ensure the selection of disjoint working and backup paths as discussed in the following subsections.

It should be noted that end-to-end recovery can be operated even when the ENs are single-homed. However, obviously, in this case there is no protection against the failure of an EN-CN link, or of the edge CN itself.

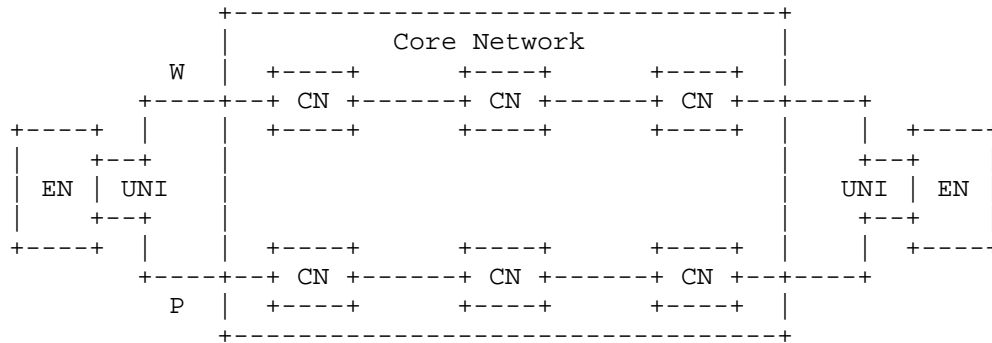


Figure 7 - UNI End-to-End Recovery

7.1.1. Serial Provisioning of Working and Protection Paths

In serial provisioning, one path is computed before another and the associated LSP may even be set up before the second path is computed. In the case where the working path is computed and created before the protection path, path computation for the protection path needs to select a (maximally) disjoint path given this existing working path.

If the EN is allowed to see details of the core network, the EN can use the RRO to collect the route of the working path. It can then use the Exclude Route Object (XRO) to exclude the working path when signaling the protection path, as described in [RFC4874].

But in most cases, in order to preserve the confidentiality of topology within the core network, the route of the working path as it

traverses the core network will be hidden from the EN. In such cases, the RRO and XRO mechanism cannot be used. Alternative includes:

- Only collect the Shared Risk Group (SRG) information, but not the full path information [SLRG-FA]. This is because the SRG information is normally less confidential than the information of node ID and link ID.

- Another possible solution is encrypted the SRG information and provide it to the EN nodes, so that the EN nodes can using this information to convey the diversity constraint, as the method specified in [UNIExt].

- In an application scenario where a PCE is involved inside the core network, then the Path Key mechanism can be used. The confidential path segment, i.e., the route of the working path as it traverses the core network, is encoded as a PKS by the PCE when computing the working path [RFC5520]. This PKS can be used by the EN when it requests the PCE to compute a protection path, to exclude the nodes and links used by the working path. As previously described, the PKS is also used in signaling [RFC5553] so that the EN can indicate to the CN what path to use across the core network.

In order to specify the diversity requirement, it is required that the PKS should be carried in the XRO in both PCEP message and RSVP-TE signaling.

7.1.2. Concurrent Computation of Working and Protection Path

The working and protection path can be computed at the same time (e.g., by PCE or by one of the CNs to which the source EN is attached).

[PCE-GMPLS] adds support for an end node to request a protected service using the protection types defined in [RFC4872]. Therefore, it's possible that the source EN requests the edge CN or PCE to compute both the working and the protection path at the same time. At this time, the disjunction requirement can be resolved inside the path computation server.

Same as described in the previous section, the path segment traversing the core network can be encoded as a PKS if confidentiality is requested.

7.2. Segment Recovery

The UNI connection may request protection only inside the core network, especially in case of single-homing. A UNI segment protection example is shown in Figure 8. In this case, the core network provides a "recovery domain".

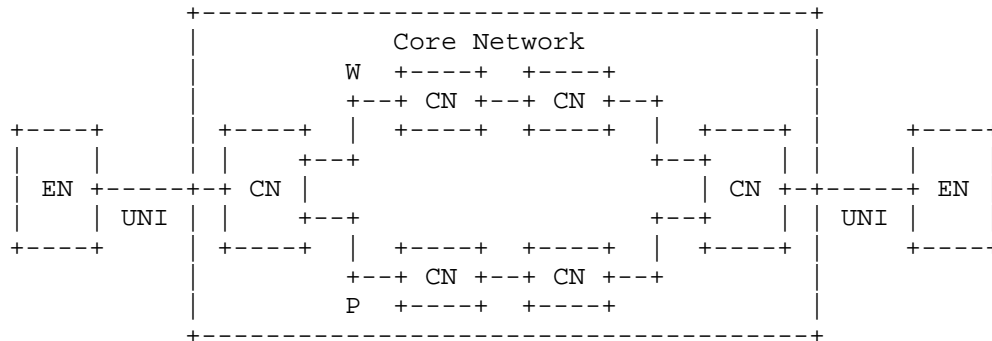


Figure 8 - UNI Segment Recovery

[RFC4873] provides a mechanism for segment recovery, in which the PROTECTION Object is extended to indicate segment recovery, and the Secondary ERO (SERO) is introduced for the explicit control of the protection LSP between the branch node and the merge node.

However, in the overlay model, the mechanisms of segment recovery described in [RFC4873] may not be appropriate. In particular, the source EN might not know the CN to which the destination EN is attached. That means that the source EN knows the branch for the protection segment, but does not know the merge node.

But the model shown in Figure 8 is particularly important because it places the responsibility for service delivery with the edge CNs. This will be a common operational model in overlay networks. Fortunately the stitching model (Section 6.2) and the hierarchical model (Section 6.4) are good at providing the necessary protection within the core network without the ENs having to be aware of the paths in the core network.

8. UNI Call

The Call is a fundamental component of the ASON model [G.8080]. It is used to maintain the association between one or more user

applications and the network, and to control the set-up, release, modification, and maintenance of sets of Connections (LSPs). In simple cases, the Call and Connection can be established at the same time and in a strict one-to-one ratio. In this case, Call signaling requires only minor extensions to connection signaling. However, if Calls are handled separately from Connections, or if more than one Connection can be associated with a single Call, additional Call signaling is required.

The GMPLS Call, defined in [RFC4974], provides a mechanism to negotiate agreement between endpoints possibly in cooperation with the nodes that provide access to the network. Typically the GMPLS Call can be applied in the UNI scenario for access link capability exchange, policy, authorization, security, and so on.

8.1. Exchange of UNI Link Information

It is possible that the TE attributes of the access link (i.e., the UNI link) are not shared across the core network. So the source EN may not have the TE information of the destination access link as well as the capability of the destination EN. For example, in case of TDM network, the Virtual Concatenation (VCAT) and Link Capacity Adjustment Scheme (LCAS) capability of the destination EN may not be known.

In this case, the source EN can raise a Call carrying the LINK_CAPABILITY object to have a capability exchange with the destination EN, as described in [RFC4974].

8.2. Control of Call Route

When applying the Call, it's possible that there are multiple core network domains between the source EN (Call initiator) and the destination EN (Call terminator), or there is more than one Call manager in the core network (e.g., in the multi-homing scenario where the CNS to which the ENs are attached act as the Call managers).

In the both cases, when establishing the Call, there may be multiple alternative routes for the Call message to reach the destination EN. One can simply use the hop-by-hop manner (i.e., each Call manager determines the next Call manager to which the Call message will be sent by itself) to control the path of the Call.

However, in the practical deployment of UNI Call, commercial and policy motivations normally play an important role in selecting the Call route, especially in the multi-domain scenario. In this case, the hop-by-hop manner is not practical because the route of the Call

needs to be pre-determined in consideration of commercial and policy factors before establishing the Call.

Therefore, it is desirable to allow full control of the Call by the source EN. That is, the source EN can identify the full Call route and signal it explicitly, so that the Call message can be forwarded along the desired route. Moreover, the management plane needs to be able to identify the Call route explicitly as an instruction to the source EN.

9. UNI Multicast

Data plane multicasting is supported in existing Traffic-Engineering networks. GMPLS provides extensions to RSVP-TE to support provisioning of point-to-multipoint (P2MP) TE LSPs via the control plane, as described in [RFC4461] and [RFC4875].

In the scenarios where P2MP is supported using the overlay architectural model, it is a requirement to transport signals from one source EN to multiple destination ENs. One could create a point-to-point (P2P) connection between the source EN and each destination EN, but it will likely be a waste of bandwidth resource both of the UNI link and in the core network.

Therefore, there are some scenarios required to support point-to-multipoint (P2MP) TE LSPs from one source EN to multiple leaf ENs.

9.1. UNI Multicast Connection Model

There are two cases for the UNI multicast. For the first case, only the ingress and egress CNs in the core network support P2MP. The core network has to provide multiple P2P connections between ingress CN and each egress CN for the end-to-end UNI multicast, as shown in Figure 9. This relieves the pressure on the source UNI link, but does not help the over use of the core links such as CN1-CN2.

For the session shuffling model, one end-to-end P2MP session can be used to create the P2MP LSP, with an address mapping performed at both ingress and egress CNs.

For the hierarchical model, multiple P2P LSP tunnels or one P2MP LSP tunnel between the ingress CN and each egress CNs needs be triggered by the UNI signaling for creating the P2MP LSP. GMPLS UNI signaling should have the capability to convey the multicast information by using the hierarchical model.

10. Security Considerations

[RFC5920] provides an overview of security vulnerabilities and protection mechanisms for the GMPLS control plane, which is applicable to this document.

The details of the specific security measures of the overlay network architectural model are provided in [RFC4208], which permits the core network to filter out specific RSVP objects to hide its topology from the EN.

Furthermore, if PCE is used, the security issues described in [RFC4655] should also be considered.

Additionally, when the PKS mechanism is applied, the security issues can be dealt with using [RFC5520] and [RFC5553].

11. IANA Considerations

This informational document does not make any requests for IANA action.

12. Acknowledgments

The authors would like to thank Zafar Ali for his comments.

13. References

13.1. Normative References

- [RFC3209] D. Awduche et al, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC3209, December 2001.
- [RFC3471] Berger, L., Ed., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, January 2003.
- [RFC3473] L. Berger, Ed., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.
- [RFC3945] Mannie, E., "Generalized Multi-Protocol Label Switching (GMPLS) Architecture", RFC 3945, October 2004.
- [RFC4203] Kompella, K., and Rekhter, Y., "OSPF Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4203, October 2005.
- [RFC4206] K. Kompella et al, "Label Switched Paths (LSP) Hierarchy with Generalized Multi-Protocol Label Switching (GMPLS) Traffic Engineering (TE)", RFC4206, October 2005.
- [RFC4208] G. Swallow et al, "Generalized Multiprotocol Label Switching (GMPLS) User-Network Interface (UNI): Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Support for the Overlay Model", RFC4208, October 2005.
- [RFC4655] A. Farrel et al, "A Path Computation Element (PCE)-Based Architecture", RFC4655, August 2006.
- [RFC4847] T. Takeda, Ed., "Framework and Requirements for Layer 1 Virtual Private Networks", RFC4847, April 2007.
- [RFC4872] J.P. Lang et al, "RSVP-TE Extensions in Support of End-to-End Generalized Multi-Protocol Label Switching (GMPLS) Recovery", RFC4872, May 2007.
- [RFC4873] L. Berger et al, "GMPLS Segment Recovery", RFC4873, May 2007.

- [RFC4874] CY. Lee et al, "Exclude Routes - Extension to Resource Reservation Protocol-Traffic Engineering (RSVP-TE)", RFC4874, April 2007.
- [RFC4875] R. Aggarwal et al, "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC4875, May 2007.
- [RFC4974] D. Papadimitriou and A. Farrel, Ed., "Generalized MPLS (GMPLS) RSVP-TE Signaling Extensions in Support of Calls", RFC4974, August 2007.
- [RFC5150] A. Ayyangar et al, "Label Switched Path Stitching with Generalized Multiprotocol Label Switching Traffic Engineering (GMPLS TE)", RFC5150, February 2008.
- [RFC5195] Ould-Brahim, H., Fedyk, D., and Y. Rekhter, "BGP-Based Auto-Discovery for Layer-1 VPNs", RFC 5195, June 2008.
- [RFC5251] D. Fedyk and Y. Rekhter, Ed., "Layer 1 VPN Basic Mode", RFC5251, July 2008.
- [RFC5252] I. Bryskin and L. Berger Ed., "OSPF-Based Layer 1 VPN Auto-Discovery", RFC5252, July 2008.
- [RFC5520] R. Bradford, Ed., "Preserving Topology Confidentiality in Inter-Domain Path Computation Using a Path-Key-Based Mechanism", RFC5520, April 2009.
- [RFC5553] A. Farrel, Ed., "Resource Reservation Protocol (RSVP) Extensions for Path Key Support", RFC5553, May 2009.
- [RFC6001] Dimitri Papadimitriou et al, "Generalized Multi-Protocol Label Switching (GMPLS) Protocol Extensions for Multi-Layer and Multi-Region Networks (MLN/MRN)", RFC6001, October, 2010.
- [RFC6107] K. Shiimoto, A. Farrel, "Procedures for Dynamically Signaled Hierarchical Label Switched Paths", RFC6107, February 2011.
- [G.8080] ITU-T Rec. G.8080/Y.1304, "Architecture for the Automatically Switched Optical Network (ASON)," June 2006 (and Amend.2, September 2010).

13.2. Informative References

- [RFC4461] S. Yasukawa, Ed., "Signaling Requirements for Point-to-Multipoint Traffic-Engineered MPLS Label Switched Paths (LSPs)", RFC4461, April 2006.
- [RFC5212] K. Shiomoto et al, "Requirements for GMPLS-Based Multi-Region and Multi-Layer Networks (MRN/MLN)", RFC5212, July 2008.
- [RFC5253] T. Takeda, Ed., "Applicability Statement for Layer 1 Virtual Private Network (L1VPN) Basic Mode", RFC 5253, July 2008.
- [RFC5339] J.L. Le Roux et al, "Evaluation of Existing GMPLS Protocols against Multi-Layer and Multi-Region Networks (MLN/MRN)", RFC5339, September 2008.
- [RFC5441] J.P. Vasseur et al, "A Backward-Recursive PCE-Based Computation (BRPC) Procedure to Compute Shortest Constrained Inter-Domain Traffic Engineering Label Switched Paths", RFC5441, April 2009.
- [RFC5623] Oki, E., Takeda, T., Le Roux, J.L., and Farrel, A., "Framework for PCE-Based Inter-Layer MPLS and GMPLS Traffic Engineering", RFC 5623, September 2009.
- [RFC5920] L. Fang, Ed., "Security Framework for MPLS and GMPLS Networks", RFC5920, July 2010.
- [Call-ext] Fatai Zhang et al, "RSVP-TE extensions to GMPLS Calls", draft-zhang-ccamp-gmpls-call-extensions-01.txt, July 08, 2009.
- [PCE-GMPLS] C. Margaria et al, "PCEP extensions for GMPLS", draft-ietf-pce-gmpls-pcep-extensions-07.txt, October 21, 2012.
- [SRLG-FA] Fatai Zhang et al, "RSVP-TE Extensions for Configuration SRLG of an FA", draft-ietf-ccamp-rsvp-te-srlg-collect-02.txt, work in progress.
- [RFC6344] G. Bernstein et al, "Operating Virtual Concatenation (VCAT) and the Link Capacity Adjustment Scheme (LCAS) with Generalized Multi-Protocol Label Switching (GMPLS)", RFC6344, August 2011.

- [UNIExt] D. Fedyk, D. Beller, Lieven Levrau, D. Ceccarelli, F. Zhang, et al, "UNI Extensions for Diversity and Latency Support", draft-fedyk-ccamp-uni-extensions-00.txt, Feb. 2013;
- [LSP-DIV] A., Zafar, G., Swallow et al, "Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Path Diversity using Exclude Routes", draft-ietf-ccamp-lsp-diversity-01.txt, work in progress;
- [UNI-PLUS] A., Zafar, G., Swallow et al, "Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) extension for signaling Objective Function and Metric Bound", draft-ali-ccamp-rc-objective-function-metric-bound-02.txt, work in progress;
- [TE-REC] A., Zafar, G., Swallow et al, "Resource ReserVation Protocol-Traffic Engineering (RSVP-TE)extension for recording TE Metric of a Label Switched Path", draft-ietf-ccamp-te-metric-recording-01.txt, work in progress;

14. Contributors' Address

Yi Lin
Huawei Technologies
F3-5-B R&D Center, Huawei Base
Bantian, Longgang District
Shenzhen 518129 P.R.China

Phone: +86-755-28972914
Email: yi.lin@huawei.com

Young Lee
Huawei Technologies
1700 Alma Drive, Suite 100
Plano, TX 75075
USA

Phone: (972) 509-5599 (x2240)
Email: leeyoung@huawei.com

Dan Li
Huawei Technologies

F3-5-B R&D Center, Huawei Base
Bantian, Longgang District
Shenzhen 518129 P.R.China

Phone: +86-755-28973237
Email: huawei.danli@huawei.com

15. Authors' Addresses

Fatai Zhang
Huawei Technologies
F3-5-B R&D Center, Huawei Base
Bantian, Longgang District
Shenzhen 518129 P.R.China

Phone: +86-755-28972912
Email: zhangfatai@huawei.com

Oscar Gonzalez de Dios
Telefonica Investigacion y Desarrollo
Emilio Vargas 6
Madrid, 28045
Spain

Phone: +34 913374013
Email: ogondio@tid.es

Adrian Farrel
Old Dog Consulting

EMail: adrian@olddog.co.uk

Xian Zhang
Huawei Technologies
F3-5-B R&D Center, Huawei Base
Bantian, Longgang District
Shenzhen 518129 P.R.China

Phone: +86-755-28972913
Email: zhang.xian@huawei.com

Daniele Ceccarelli

Ericsson
Via A. Negrone 1/A
Genova - Sestri Ponente
Italy

Email: daniele.ceccarelli@ericsson.com

Intellectual Property

The IETF Trust takes no position regarding the validity or scope of any Intellectual Property Rights or other rights that might be claimed to pertain to the implementation or use of the technology described in any IETF Document or the extent to which any license under such rights might or might not be available; nor does it represent that it has made any independent effort to identify any such rights.

Copies of Intellectual Property disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the IETF on-line IPR repository at <http://www.ietf.org/ipr>

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement any standard or specification contained in an IETF Document. Please address the information to the IETF at ietf-ipr@ietf.org.

The definitive version of an IETF Document is that published by, or under the auspices of, the IETF. Versions of IETF Documents that are published by third parties, including those that are translated into other languages, should not be considered to be definitive versions of IETF Documents. The definitive version of these Legal Provisions is that published by, or under the auspices of, the IETF. Versions of these Legal Provisions that are published by third parties, including those that are translated into other languages, should not be considered to be definitive versions of these Legal Provisions.

For the avoidance of doubt, each Contributor to the IETF Standards Process licenses each Contribution that he or she makes as part of the IETF Standards Process to the IETF Trust pursuant to the provisions of RFC 5378. No language to the contrary, or terms,

conditions or rights that differ from or are inconsistent with the rights and licenses granted under RFC 5378, shall have any effect and shall be null and void, whether published or posted by such Contributor, or included with or in such Contribution.

Disclaimer of Validity

All IETF Documents and the information contained therein are provided on an "AS IS" basis and THE CONTRIBUTOR, THE ORGANIZATION HE/SHE REPRESENTS OR IS SPONSORED BY (IF ANY), THE INTERNET SOCIETY, THE IETF TRUST AND THE INTERNET ENGINEERING TASK FORCE DISCLAIM ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION THEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Copyright Notice

Copyright (c) 2013 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.

CCAMP Working Group
Internet Draft
Category: Standards track

Xian Zhang
Fatai Zhang
Huawei
O. Gonzalez de Dios
Telefonica I+D
Igor Bryskin
ADVA Optical Networking

Expires: March 31, 2014

September 29, 2013

Extensions to Resource ReSerVation Protocol-Traffic Engineering (RSVP-TE) to Support Route Exclusion Using Path Key Subobject

draft-zhang-ccamp-route-exclusion-pathkey-00.txt

Status of this Memo

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 March 31, 2014.

Abstract

This document extends the Resource ReSerVation Protocol-Traffic Engineering (RSVP-TE) eXclude Route Object (XRO) and Explicit Route Object (ERO) to support specifying route exclusion requirement using Path Key Subobject (PKS).

Table of Contents

1. Introduction	2
1.1. Example Use	3
2. RSVP-TE Extensions.....	4
2.1. Path Key Subobject (PKS)	4
2.2. PKS Processing Rules	4
3. Security Considerations.....	5
4. IANA Considerations.....	5
4.1. New Subobject Type.....	5
4.2. New Error Code.....	6
5. Acknowledgments	6
6. References	6
6.1. Normative References	6
6.2. Informative References.....	6
7. Authors' Addresses.....	7

1. Introduction

[RFC5520] defines the concept of a Path Key. This object can be used by a Path Computation Element (PCE) in place of a segment of a path that it wishes to keep confidential. The Path Key can be signaled in Resource ReSerVation Protocol-Traffic Engineering (RSVP-TE) protocol by placing it in an Explicit Route Object (ERO) as described in [RFC5553].

When establishing a set of LSPs to provide protection services [RFC4427], it is often desirable that the LSPs should take different paths through the network. This can be achieved by path computation entities that have full end-to-end visibility, but it is more complicated in multi-domain environments when segments of the path may be hidden so that they are not visible outside the domain they traverse.

This document describes how the Path Key object can be used in the RSVP-TE eXclude Route Object (XRO), and the Explicit eXclusion Route subobject (EXRS) of the ERO in order to facilitate path hiding, but allow diverse end-to-end paths to be established in multi-domain environments.

1.1. Example Use

Figure 1 shows a simple network with two domains. It is desired to set up a pair of path-disjoint LSPs from the source in Domain 1 to the destination in Domain 2, but the domains keep strict confidentiality about all path and topology information.

The first LSP will be signaled by the source with ERO {A, B, loose Dst} and will be set up with the path {Src, A, B, U, V, W, Dst}. But when sending the RRO out of Domain 2, node U would normally strip the path and replace it with a loose hop to the destination. With this limited information, the source is unable to include enough detail in the ERO of the second LSP to avoid it taking, for example, the path {Src, C, D, X, V, W, Dst} which is not path-disjoint.

In order to improve the outcome, node U can replace the path segment {U, V, W} in the RRO with a Path Key. The Path Key Object assigns an identifier to the key and also indicates that it was node U that made the replacement.

With this additional information, the source is able to signal the second LSP with ERO set to {C, D, exclude Path Key, loose Dst}. When the signaling message reaches node X, it can consult node U to expand the Path Key and so know to avoid the path of the first LSP. Alternatively, the source could use an ERO of {C, D, loose Dst} and include an XRO containing the Path Key.

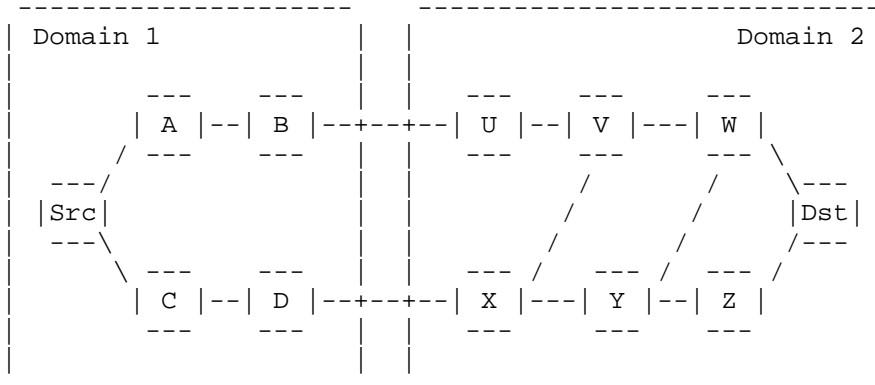


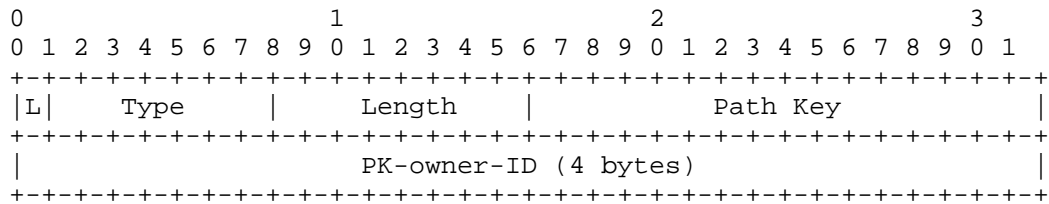
Figure 1: A Simple Multi-Domain Network

2. RSVP-TE Extensions

This section defines the subobject that can be either in the XRO object or Explicit eXclusion Route subobject (EXRS) as defined in [RFC4874].

2.1. Path Key Subobject (PKS)

The IPv4 PKS has the following format:

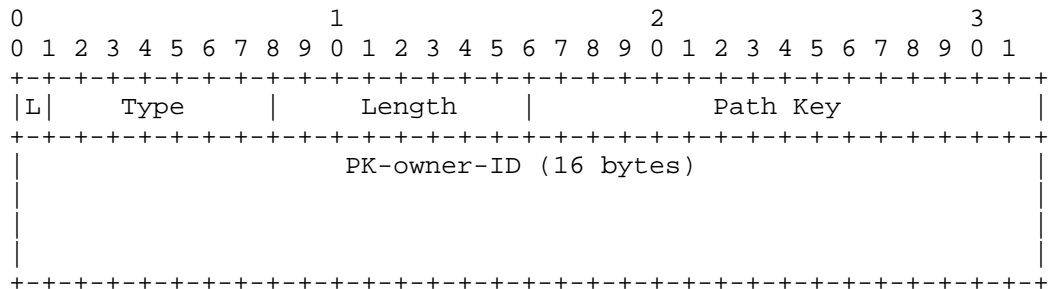


The meaning of the field L bit, Length, Path Key is defined in [RFC4874].

Type: sub-object type for XRO Path Key; TBD.

PK-owner-ID: The IPv4 address of a node that assigned the Path Key identifier and that can return an expansion of the Path Key or use the Path Key as an exclusion in a path computation.

Similarly, the format of IPv6 PKS is as follows:



2.2. PKS Processing Rules

The exclude route list is encoded as a series of subobjects contained in an EXCLUDE_ROUTE object or an EXRS of the ERO. The procedure defined in [RFC4874] for processing XRO and EXRS is not changed by this document.

Irrespective of the L flag, if the node, receiving the PKS, cannot recognize the subobject, it will react according to [RFC4874] and SHOULD ignore the constraint.

Otherwise, if it cannot find a route/route segment meeting the constraint:

- if L flag is set to 0, it will react according to [RFC4874] and SHOULD send a PathErr message with the error code/value combination ''Routing Problem'' / ''Route Blocked by Exclude Route''.

- if L flag is set to 1, which means the node SHOULD try to be as much diversified as possible with the specified resource. If it cannot fully support the constraint, it SHOULD send a PathErr message with the error code/value combination "Notify Error" / "Fail to find diversified path" (TBD).

This mechanism can work together with the presence of a Path Computation Element (PCE) or if the local node generates the PK itself. Note that other mechanisms to use or expand the PK are out of scope of this document.

3. Security Considerations

The use of path keys proposed in this draft allows nodes to hide parts of the path as it is signaled. This can be used to improve the confidentiality of the LSP setup. Moreover, it may serve to improve security of the control plane for the LSP as well as data plane traffic carried on this LSP. However, the benefits of using path key are lost unless there is an appropriate access control of any tool that allows expansion of the path key.

4. IANA Considerations

4.1. New Subobject Type

IANA registry: RSVP PARAMETERS

Subsection: Class Names, Class Numbers, and Class Types

This document introduces two new subobjects for the EXCLUDE_ROUTE object [RFC4874], C-Type 1.

Subobject Type	Subobject Description
-----	-----

64(TBD by IANA)

IPv4 Path Key Subobject

65(TBD By IANA)

IPv6 Path Key Subobject

Note well: [RFC5520] defines the PKS for use in PCEP. The above number suggestions for use in RSVP-TE follow that assigned for the PKS in PCEP [RFC5520].

4.2. New Error Code

IANA registry: RSVP PARAMETERS

Subsection: Error Codes and Globally-Defined Error Value Sub-Codes

New Error Values sub-codes have been registered for the Error Code 'Notify Error' (25).

TBD = "Fail to find diversified path"

5. Acknowledgments

TBD.

6. References

6.1. Normative References

- [RFC3209] D. Awduche et al, ''RSVP-TE: Extensions to RSVP for LSP Tunnels'', RFC3209, December 2001.
- [RFC4874] CY. Lee, A. Farrel, S. De Cnodder, ''Exclude Routes - Extension to Resource ReserVation Protocol-Traffic Engineering (RSVP-TE), RFC4874, April 2007.
- [RFC5553] A. Farrel, Ed., ''Resource Reservation Protocol (RSVP) Extensions for Path Key Support'', RFC5553, May 2009.

6.2. Informative References

- [RFC5520] R. Bradford, Ed., ''Preserving Topology Confidentiality in Inter-Domain Path Computation Using a Path-Key-Based Mechanism'', RFC5520, April 2009.
- [RFC4427] E. Mannie, Ed., ''Recovery (Protection and Restoration) Terminology for Generalized Multi-Protocol Label Switching (GMPLS)'', RFC4427, March 2006.

7. Authors' Addresses

Xian Zhang
Huawei Technologies

Email: zhang.xian@huawei.com

Fatai Zhang
Huawei Technologies

Email: zhangfatai@huawei.com

Oscar Gonzalez de Dios
Telefonica I+D
Don Ramon de la Cruz
Madrid, 28006
Spain

Phone: +34 913328832
Email: ogondio@tid.es

Igor Bryskin
ADVA Optical Networking

Email: ibryskin@advaoptical.com

Intellectual Property

The IETF Trust takes no position regarding the validity or scope of any Intellectual Property Rights or other rights that might be claimed to pertain to the implementation or use of the technology described in any IETF Document or the extent to which any license under such rights might or might not be available; nor does it represent that it has made any independent effort to identify any such rights.

Copies of Intellectual Property disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or

users of this specification can be obtained from the IETF on-line IPR repository at <http://www.ietf.org/ipr>

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement any standard or specification contained in an IETF Document. Please address the information to the IETF at ietf-ipr@ietf.org.

The definitive version of an IETF Document is that published by, or under the auspices of, the IETF. Versions of IETF Documents that are published by third parties, including those that are translated into other languages, should not be considered to be definitive versions of IETF Documents. The definitive version of these Legal Provisions is that published by, or under the auspices of, the IETF. Versions of these Legal Provisions that are published by third parties, including those that are translated into other languages, should not be considered to be definitive versions of these Legal Provisions.

For the avoidance of doubt, each Contributor to the IETF Standards Process licenses each Contribution that he or she makes as part of the IETF Standards Process to the IETF Trust pursuant to the provisions of RFC 5378. No language to the contrary, or terms, conditions or rights that differ from or are inconsistent with the rights and licenses granted under RFC 5378, shall have any effect and shall be null and void, whether published or posted by such Contributor, or included with or in such Contribution.

Disclaimer of Validity

All IETF Documents and the information contained therein are provided on an "AS IS" basis and THE CONTRIBUTOR, THE ORGANIZATION HE/SHE REPRESENTS OR IS SPONSORED BY (IF ANY), THE INTERNET SOCIETY, THE IETF TRUST AND THE INTERNET ENGINEERING TASK FORCE DISCLAIM ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION THEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Copyright Notice

Copyright (c) 2013 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.

