Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: May 2, 2012

H. Chen
Huawei Technologies
C. Margaria
Nokia Siemens Networks

October 30, 2011

Extensions to the Path Computation Element Communication Protocol (PCEP) for Backup Egress of a Traffic Engineering Label Switched Path

draft-chen-pce-compute-backup-egress-03.txt

Abstract

This document presents extensions to the Path Computation Element Communication Protocol (PCEP) for a PCC to send a request for computing a backup egress for an MPLS TE P2MP LSP or an MPLS TE P2P LSP to a PCE and for a PCE to compute the backup egress and reply to the PCC with a computation result for the backup egress.

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). 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 May 2, 2012.

Table of Contents

1. Introduction

RFC 4655 "A Path Computation Element-(PCE) Based Architecture" describes a set of building blocks for constructing solutions to compute Point-to-Point (P2P) Traffic Engineering (TE) label switched paths across multiple areas or Autonomous System (AS) domains. A typical PCE-based system comprises one or more path computation servers, traffic engineering databases (TED), and a number of path computation clients (PCC). A routing protocol is used to exchange traffic engineering information from which the TED is constructed. A PCC sends a Point-to-Point traffic engineering Label Switched Path (LSP) computation request to the path computation server, which uses the TED to compute the path and responses to the PCC with the computed path. A path computation server is named as a PCE. The communications between a PCE and a PCC for Point-to-Point label switched path computations follow the PCE communication protocol (PCEP).

RFC6006 "Extensions to PCEP for Point-to-Multipoint Traffic Engineering Label Switched Paths" describes extensions to PCEP to handle requests and responses for the computation of paths for P2MP TE LSPs.

This document defines extensions to the Path Computation Element Communication Protocol (PCEP) for a PCC to send a request for computing a backup egress node for an MPLS TE P2MP LSP or an MPLS TE P2P LSP to a PCE and for a PCE to compute the backup egress node and reply to the PCC with a computation result for the backup egress node.

2. Terminology

This document uses terminologies defined in RFC5440, and RFC4875.

3. 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.

4. Extensions to PCEP

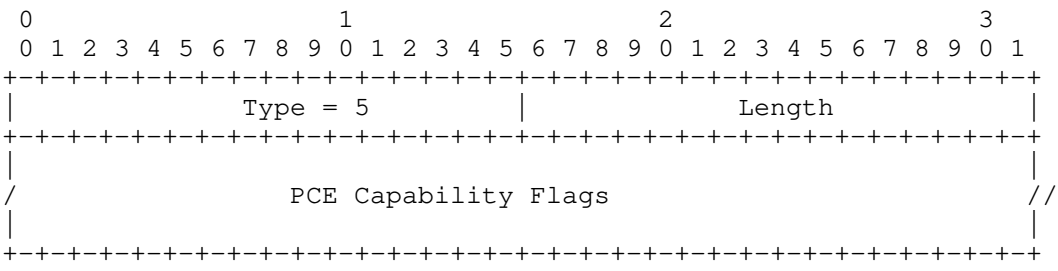
This section describes the extensions to PCEP for computing a backup egress of an MPLS TE P2MP LSP and an MPLS TE P2P LSP.

4.1. Backup Egress Capability Advertisement

4.1.1. Capability TLV in Existing PCE Discovery Protocol

An option for advertising a PCE capability for computing a backup egress for an MPLS TE P2MP LSP or an MPLS TE P2P LSP is to define two new flags. One new flag in the OSPF and IS-IS PCE Capability Flags indicates the capability that a PCE is capable to compute a backup egress for an MPLS TE P2MP LSP; and another new flag in the OSPF and IS-IS PCE Capability Flags indicates the capability that a PCE is capable to compute a backup egress for an MPLS TE P2P LSP..

The format of the PCE-CAP-FLAGS sub-TLV is as follows:



Type: 5
Length: Multiple of 4 octets
Value: This contains an array of units of 32-bit flags numbered from the most significant as bit zero, where each bit represents one PCE capability.

The following capability bits have been assigned by IANA:

Bit	Capabilities
0	Path computation with GMPLS link constraints
1	Bidirectional path computation
2	Diverse path computation
3	Load-balanced path computation
4	Synchronized path computation
5	Support for multiple objective functions
6	Support for additive path constraints (max hop count, etc.)
7	Support for request prioritization
8	Support for multiple requests per message
9	Global Concurrent Optimization (GCO)
10	P2MP path computation
11-31	Reserved for future assignments by IANA.

Reserved bits SHOULD be set to zero on transmission and MUST be ignored on receipt.

For the backup egress capabilities, one bit such as bit 13 may be assigned to indicate that a PCE is capable to compute a backup egress for an MPLS TE P2MP LSP and another bit such as bit 14 may be assigned to indicate that a PCE is capable to compute a backup egress for an MPLS TE P2P LSP as follows.

Bit	Capabilities
13	Backup egress computation for P2MP LSP
14	Backup egress computation for P2P LSP
15-31	Reserved for future assignments by IANA.

4.1.2. Open Message Extension

If a PCE does not advertise its backup egress compution capability during discovery, PCEP should be used to allow a PCC to discover, during the Open Message Exchange, which PCEs are capable of supporting backup egress computation.

To achieve this, we extend the PCEP OPEN object by defining a new optional TLV to indicate the PCE's capability to perform backup egress computation for an MPLS TE P2MP LSP and an MPLS TE P2P LSP.

We request IANA to allocate a value such as 8 from the "PCEP TLV Type Indicators" subregistry, as documented in Section below ("Backup Egress Capability TLV"). The description is "backup egress capable", and the length value is 2 bytes. The value field is set to indicate the capability of a PCE for backup egress computation for an MPLS TE LSP in details.

We can use flag bits in the value field in the same way as the PCE Capability Flags described in the previous section.

The inclusion of this TLV in an OPEN object indicates that the sender can perform backup egress computation for an MPLS TE P2MP LSP or an MPLS TE P2P LSP.

The capability TLV is meaningful only for a PCE, so it will typically appear only in one of the two Open messages during PCE session establishment. However, in case of PCE cooperation (e.g., inter-domain), when a PCE behaving as a PCC initiates a PCE session it SHOULD also indicate its path computation capabilities.

4.2. Request and Reply Message Extension

This section describes extensions to the existing RP (Request Parameters) object to allow a PCC to request a PCE for computing a backup egress of an MPLS TE P2MP LSP or an MPLS TE P2P LSP when the PCE receives the PCEP request.

4.2.1. RP Object Extension

The following flags are added into the RP Object:

The T bit is added in the flag bits field of the RP object to tell the receiver of the message that the request/reply is for computing a bcakup egress of an MPLS TE P2MP LSP and an MPLS TE P2P LSP.

- T (Backup Egress bit - 1 bit):

  0: This indicates that this is not PCReq/PCRep for backup egress.

  1: This indicates that this is PCReq or PCRep message for backup egress.

The IANA request is referenced in Section below (Request Parameter Bit Flags) of this document.

This T bit with the N bit defined in RFC 6006 can indicate whether a request/reply is for a bcakup egress of an MPLS TE P2MP LSP or an MPLS TE P2P LSP.

- T = 1 and N = 1: This indicates that this is a PCReq/PCRep message for backup egress of an MPLS TE P2MP LSP.

- T = 1 and N = 0: This indicates that this is a PCReq/PCRep message for backup egress of an MPLS TE P2P LSP.

4.2.2. External Destination Nodes

In addition to the information about the path that an MPLS TE P2MP LSP or an MPLS TE P2P LSP traverses, a request message may comprise other information that may be used for computing the backup egress for the P2MP LSP or P2P LSP. For example, the information about an external destination node, to which data traffic is delivered from an

egress node of the P2MP LSP or P2P LSP, is useful for computing a backup egress node.

4.2.2.1. External Destination Nodes Object

The PCC can specify an external destination nodes (EDN) Object. In order to represent the external destination nodes efficiently, we define two types of encodes for the external destination nodes in the object.

One encode indicates that the EDN object contains an external destination node for every egress node of an MPLS TE P2MP LSP or an MPLS TE P2P LSP. The order of the external destination nodes in the object is the same as the egress node(s) of the P2MP LSP or P2P LSP contained in the PCE messages.

Another encode indicates that the EDN object contains a list of egress node and external destination node pairs. For an egress node and external destination node pair, the data traffic is delivered to the external destination node from the egress node of the LSP.

The format of the external destination nodes (EDN) object boby for IPv4 with the first type of encodes is illustrated as follows:

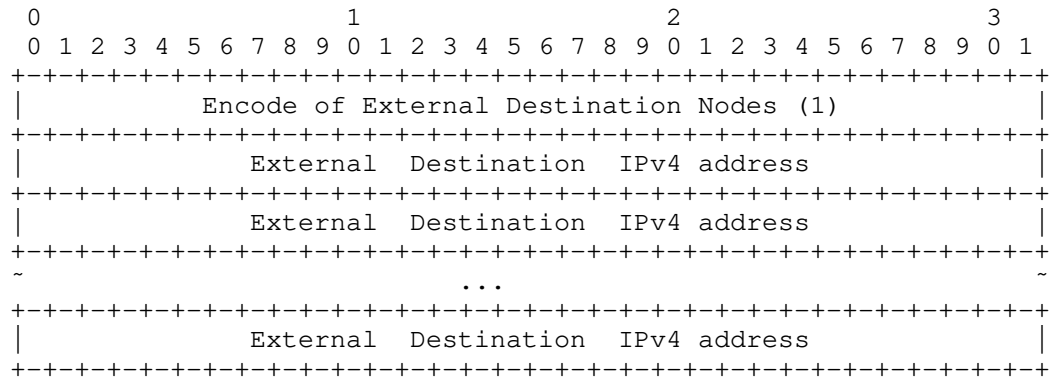


Figure 1: Format of EDN Object with one Encode for IPv4

The format of the external destination nodes (EDN) object boby for IPv4 with the second type of encodes is illustrated below:

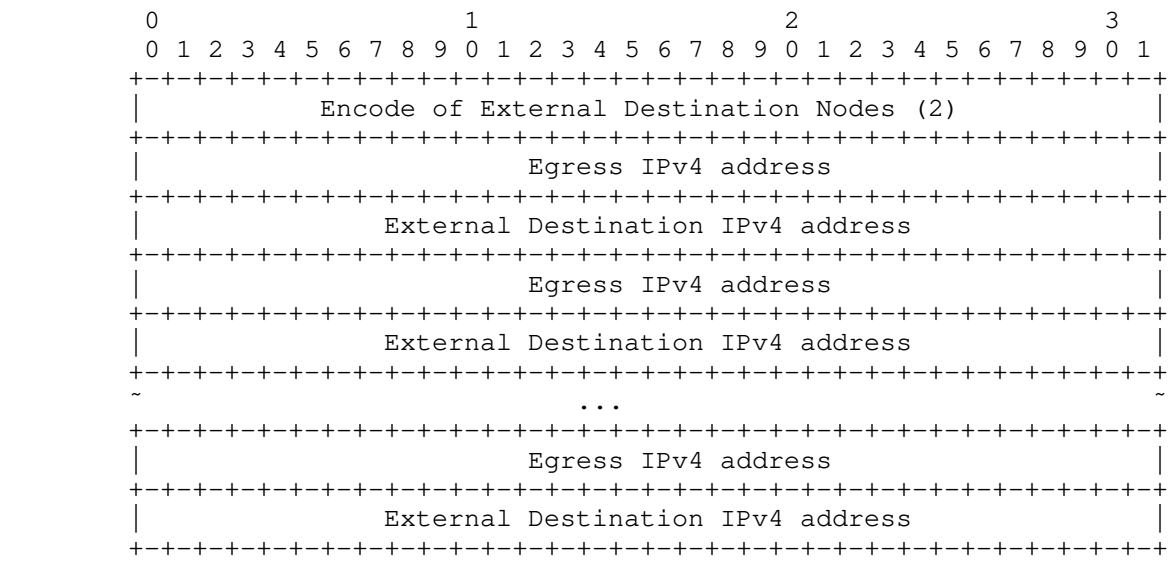


Figure 2: Format of EDN Object with another Encode for IPv4

The format of the external destination nodes (EDN) object boby for IPv6 with the first type of encodes is illustrated as follows:

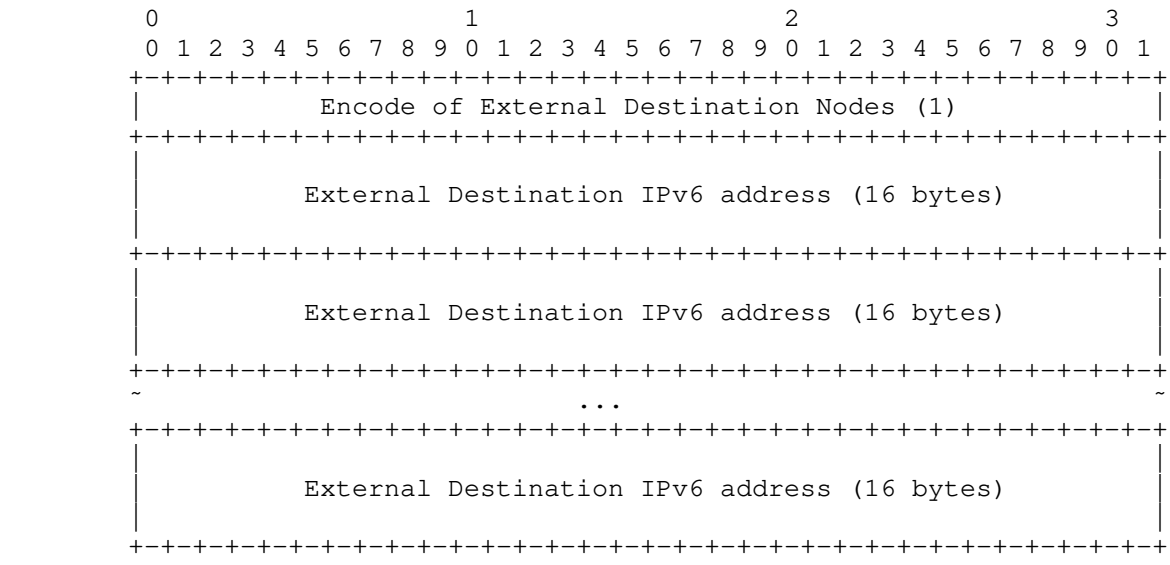


Figure 3: Format of EDN Object with one Encode for IPv6

The format of the external destination nodes (EDN) object boby for IPv6 with the second type of encodes is illustrated below:

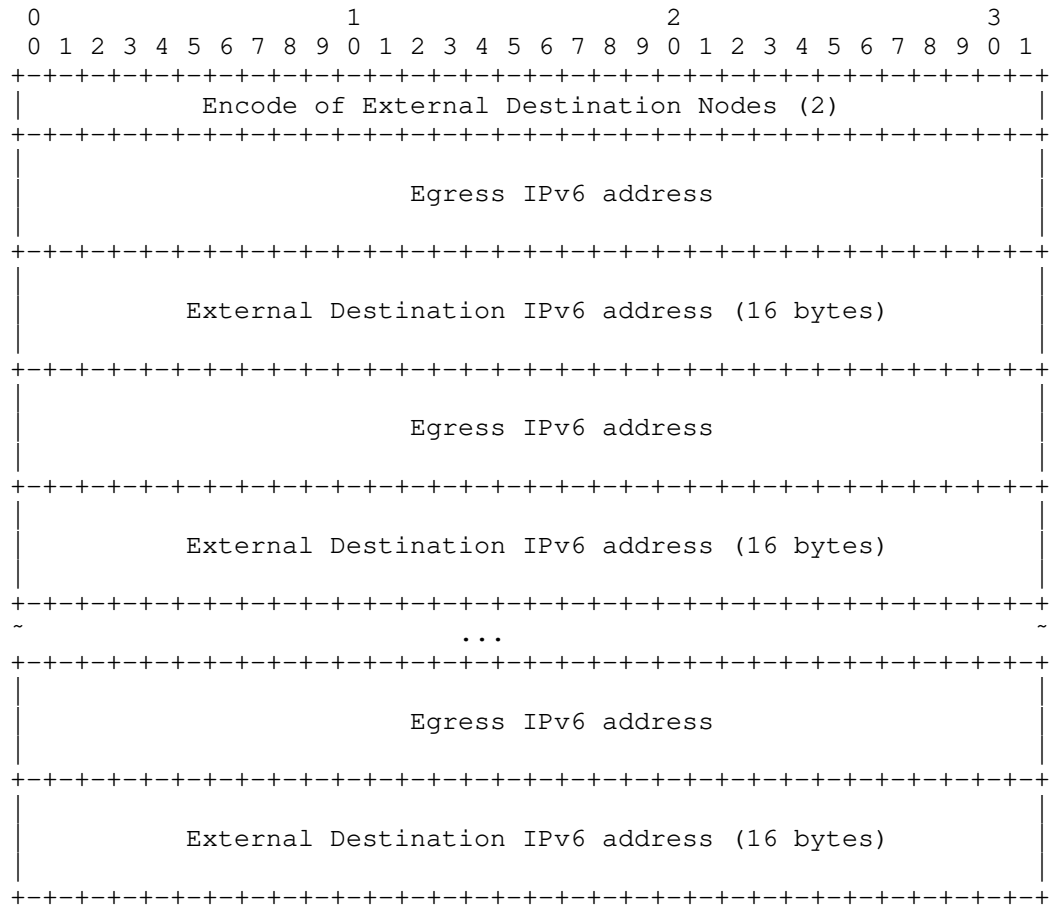


Figure 4: Format of EDN Object with another Encode for IPv6

The object can only be carried in a PCReq message. A Path Request may carry at most one external destination nodes Object.

The Object-Class and Object-types will need to be allocated by IANA. The IANA request is documented in Section below (PCEP Objects).

4.2.2.2. New Type of END-POINTS Objects

Alternatively, we may use END-POINTS to represent external destination nodes in a request message for computing backup egress

nodes of MPLS LSP.

The format of the external destination nodes (EDN) END-POINTS object boby for IPv4 with the first type of encodes is illustrated as follows:

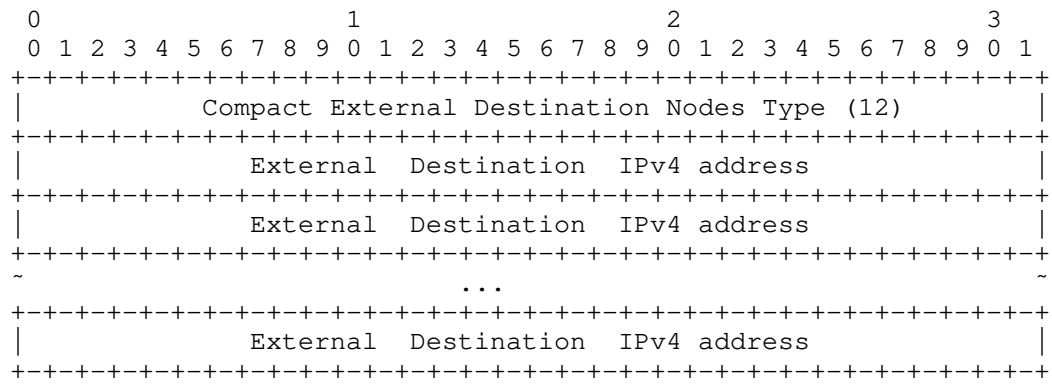


Figure 5: EDN END-POINTS Body with one Encode for IPv4

The new type of END-POINTS is Compact External Destination Nodes Type (12). The final value for the type will be assigned by IANA. The EDN END-POINTS object of type 12 contains an external destination node for every egress node of an MPLS TE P2MP LSP or an MPLS TE P2P LSP. The order of the external destination nodes in the object is the same as the egress node(s) of the P2MP LSP or P2P LSP contained in the PCE messages.

The format of the external destination nodes END-POINTS object boby for IPv4 with the second type of encodes is illustrated below:

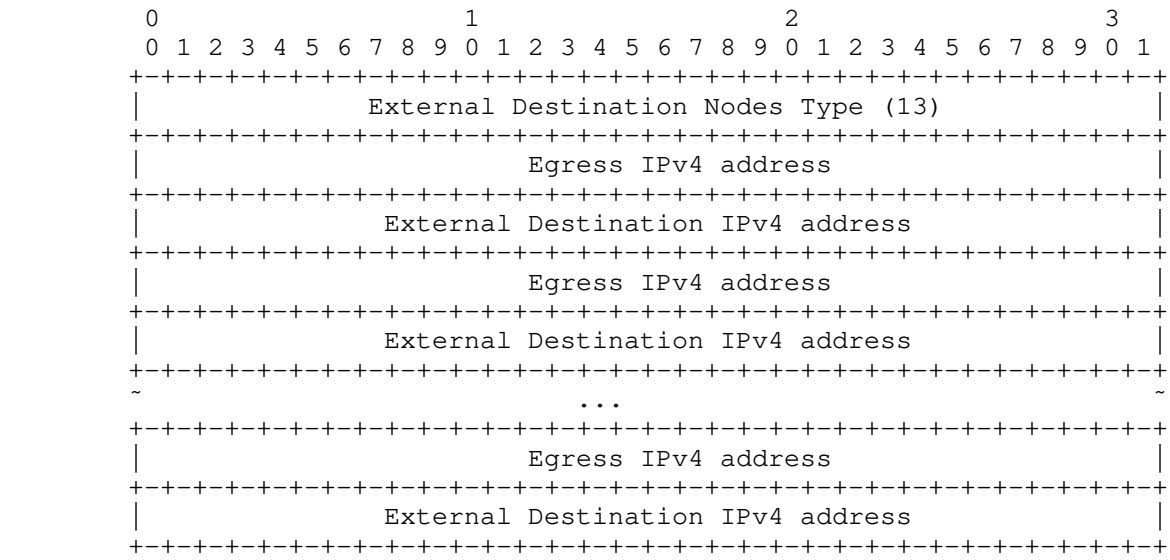


Figure 6: EDN END-POINTS Body with another Encode for IPv4

The new type of END-POINTS is External Destination Nodes Type (13). The final value for the type will be assigned by IANA. The EDN END-POINTS object of type 13 contains a list of egress node and external destination node pairs. For an egress node and external destination node pair, the data traffic is delivered to the external destination node from the egress node of the LSP.

4.2.3. Constraints between Egress and Backup Egress

A request message sent to a PCE from a PCC for computing a backup egress of an MPLS TE P2MP LSP or an MPLS TE P2P LSP may comprise a constraint indicating that there must be a path from the backup egress node to be computed to the egress node of the P2MP LSP or P2P LSP and that the length of the path is within a given hop limit such as one hop.

This constraint can be considered as default by a PCE or explicitly sent to the PCE by a PCC [TBD].

4.2.4. Constraints for Backup Path

A request message sent to a PCE from a PCC for computing a backup egress of a P2MP LSP or P2P LSP may comprise a constraint indicating that the backup egress node to be computed may not be a node on the P2MP LSP or P2P LSP. In addition, the request message may comprise a

list of nodes, each of which is a candidate for the backup egress node.

A request message sent to a PCE from a PCC for computing a backup egress of a P2MP LSP or P2P LSP may comprise a constraint indicating that there must be a path from the previous hop node of the egress node of the P2MP LSP or P2P LSP to the backup egress node to be computed and that there is not an internal node of the path from the previous hop node of the egress node of the P2MP LSP or P2P LSP to the backup egress that is on the path of the P2MP LSP or P2P LSP.

Most of these constraints for the backup path can be considered as default by a PCE. The constraints for the backup path may be explicitly sent to the PCE by a PCC [TBD].

4.2.5. Backup Egress Node

The PCE may send a reply message to the PCC in return to the request message for computing a new backup egress node or a number of backup egress nodes. The reply message may comprise information about the computed backup egress node(s), which is contained in the path(s) from the previous-hop node of the egress node of the P2MP LSP or P2P LSP to the backup egress node(s) computed.

4.2.6. Backup Egress PCEP Error Objects and Types

In some cases, the PCE may not complete the backup egress computation as requested, for example based on a set of constraints. As such, the PCE may send a reply message to the PCC that indicates an unsuccessful backup egress computation attempt. The reply message may comprise a PCEP-error object, which may comprise an error-value, error-type and some detail information.

4.2.7. Request Message Format

The PCReq message is encoded as follows using RBNF as defined in [RFC5511].

Below is the message format for a request message:

```
<PCReq Message>::= <Common Header>
                   [<svec-list>]
                   <request>
<request>::= <RP>
             <end-point-rro-pair-list>
             [<OF>]
             [<LSPA>]
             [<BANDWIDTH>]
             [<metric-list>]
             [<EDNO>]
             [<IRO>]
             [<LOAD-BALANCING>]
where:
       <EDNO> is an external destination nodes object.
```

Figure 7: The Format for a Request Message

The definitions for svec-list, RP, end-point-rro-pair-list, OF, LSPA, BANDWIDTH, metric-list, IRO, and LOAD-BALANCING are described in RFC5440 and RFC6006.

4.2.8. Reply Message Format

The PCRep message is encoded as follows using RBNF as defined in [RFC5511].

Below is the message format for a reply message:

```
<PCRep Message>::= <Common Header>
                   <response>
<response>::= <RP>
              <end-point-path-pair-list>
              [<NO-PATH>]
              [<attribute-list>]
```

where:

```
<end-point-path-pair-list>::=
        [<END-POINTS>]<path>[<end-point-path-pair-list>]


<path> ::= (<ERO>|<SERO>) [<path>]

<attribute-list>::= [<OF>]
                    [<LSPA>]
                    [<BANDWIDTH>]
                    [<metric-list>]
                    [<IRO>]
```

Figure 8: The Format for a Reply Message

The definitions for RP, NO-PATH, END-POINTS, OF, LSPA, BANDWIDTH, metric-list, IRO, and SERO are described in RFC5440, RFC6006 and RFC4875.

5. Security Considerations

The mechanism described in this document does not raise any new security issues for the PCEP, OSPF or IS-IS protocols.

6. IANA Considerations

This section specifies requests for IANA allocation.

6.1. Backup Egress Capability Flag

Two new OSPF Capability Flags are defined in this document to indicate the capabilities for computing a backup egress for an MPLS TE P2MP LSP and an MPLS TE P2P LSP. IANA is requested to make the assignment from the "OSPF Parameters Path Computation Element (PCE) Capability Flags" registry:

Bit	Description	Reference
13	Backup egress for P2MP LSP	This I-D
14	Backup egress for P2P LSP	This I-D

6.2. Backup Egress Capability TLV

A new backup egress capability TLV is defined in this document to allow a PCE to advertize its backup egress computation capability. IANA is requested to make the following allocation from the "PCEP TLV Type Indicators" sub-registry.

Value	Description	Reference
8	Backup egress capable	This I-D

6.3. Request Parameter Bit Flags

A new RP Object Flag has been defined in this document. IANA is requested to make the following allocation from the "PCEP RP Object Flag Field" Sub-Registry:

Bit	Description	Reference
15	Backup egress (T-bit)	This I-D

6.4. PCEP Objects

An External Destination Nodes Object-Type is defined in this document. IANA is requested to make the following Object-Type allocation from the "PCEP Objects" sub-registry:

Object-Class Value	34
Name	External Destination Nodes
Object-Type	1: IPv4 2: IPv6 3-15: Unassigned
Reference	This I-D

7. Acknowledgement

The author would like to thank Ramon Casellas, Dhruv Dhody and Quintin Zhao for their valuable comments on this draft.

8. References

8.1. Normative References

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

[RFC3209] 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., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.

[RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.

[RFC4875] Aggarwal, R., Papadimitriou, D., and S. Yasukawa, "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, May 2007.

[RFC6006] Zhao, Q., King, D., Verhaeghe, F., Takeda, T., Ali, Z., and J. Meuric, "Extensions to the Path Computation Element Communication Protocol (PCEP) for Point-to-Multipoint Traffic Engineering Label Switched Paths", RFC 6006, September 2010.

8.2. Informative References

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

[RFC5862] Yasukawa, S. and A. Farrel, "Path Computation Clients (PCC) - Path Computation Element (PCE) Requirements for Point-to-Multipoint MPLS-TE", RFC 5862, June 2010.

Authors' Addresses

Huaimo Chen
Huawei Technologies
Boston, MA
USA

Email: Huaimochen@huawei.com

Cyril Margaria
Nokia Siemens Networks
St Martin Strasse 76, Munich, 81541
Germany

Email: cyril.margaria@nsn.com

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: January 8, 2020

H. Chen
Futurewei

July 7, 2019

Extensions to the Path Computation Element Communication Protocol (PCEP) for Backup Egress of a Traffic Engineering Label Switched Path

draft-chen-pce-compute-backup-egress-14.txt

Abstract

This document presents extensions to the Path Computation Element Communication Protocol (PCEP) for a PCC to send a request for computing a backup egress for an MPLS TE P2MP LSP or an MPLS TE P2P LSP to a PCE and for a PCE to compute the backup egress and reply to the PCC with a computation result for the backup egress.

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). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://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 January 8, 2020.

Table of Contents

1. Introduction

RFC 4655 "A Path Computation Element-(PCE) Based Architecture" describes a set of building blocks for constructing solutions to compute Point-to-Point (P2P) Traffic Engineering (TE) label switched paths across multiple areas or Autonomous System (AS) domains. A typical PCE-based system comprises one or more path computation servers, traffic engineering databases (TED), and a number of path computation clients (PCC). A routing protocol is used to exchange traffic engineering information from which the TED is constructed. A PCC sends a Point-to-Point traffic engineering Label Switched Path (LSP) computation request to the path computation server, which uses the TED to compute the path and responses to the PCC with the computed path. A path computation server is named as a PCE. The communications between a PCE and a PCC for Point-to-Point label switched path computations follow the PCE communication protocol (PCEP).

RFC6006 "Extensions to PCEP for Point-to-Multipoint Traffic Engineering Label Switched Paths" describes extensions to PCEP to handle requests and responses for the computation of paths for P2MP TE LSPs.

This document defines extensions to the Path Computation Element Communication Protocol (PCEP) for a PCC to send a request for computing a backup egress node for an MPLS TE P2MP LSP or an MPLS TE P2P LSP to a PCE and for a PCE to compute the backup egress node and reply to the PCC with a computation result for the backup egress node.

2. Terminology

This document uses terminologies defined in RFC5440, and RFC4875.

3. 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.

4. Extensions to PCEP

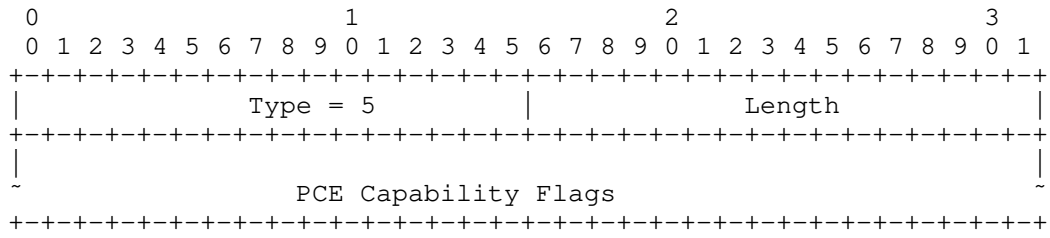
This section describes the extensions to PCEP for computing a backup egress of an MPLS TE P2MP LSP and an MPLS TE P2P LSP.

4.1. Backup Egress Capability Advertisement

4.1.1. Capability TLV in Existing PCE Discovery Protocol

An option for advertising a PCE capability for computing a backup egress for an MPLS TE P2MP LSP or an MPLS TE P2P LSP is to define two new flags. One new flag in the OSPF and IS-IS PCE Capability Flags indicates the capability that a PCE is capable to compute a backup egress for an MPLS TE P2MP LSP; and another new flag in the OSPF and IS-IS PCE Capability Flags indicates the capability that a PCE is capable to compute a backup egress for an MPLS TE P2P LSP.

The format of the PCE-CAP-FLAGS sub-TLV is as follows:



Type: 5
Length: Multiple of 4 octets
Value: This contains an array of units of 32-bit flags numbered from the most significant as bit zero, where each bit represents one PCE capability.

The following capability bits have been assigned by IANA:

Bit	Capabilities
0	Path computation with GMPLS link constraints
1	Bidirectional path computation
2	Diverse path computation
3	Load-balanced path computation
4	Synchronized path computation
5	Support for multiple objective functions
6	Support for additive path constraints (max hop count, etc.)
7	Support for request prioritization
8	Support for multiple requests per message
9	Global Concurrent Optimization (GCO)
10	P2MP path computation
11-31	Reserved for future assignments by IANA.

Reserved bits SHOULD be set to zero on transmission and MUST be ignored on receipt.

For the backup egress capabilities, one bit such as bit 13 may be assigned to indicate that a PCE is capable to compute a backup egress for an MPLS TE P2MP LSP and another bit such as bit 14 may be assigned to indicate that a PCE is capable to compute a backup egress for an MPLS TE P2P LSP as follows.

Bit	Capabilities
13	Backup egress computation for P2MP LSP
14	Backup egress computation for P2P LSP
15-31	Reserved for future assignments by IANA.

4.1.2. Open Message Extension

If a PCE does not advertise its backup egress compution capability during discovery, PCEP should be used to allow a PCC to discover, during the Open Message Exchange, which PCEs are capable of supporting backup egress computation.

To achieve this, we extend the PCEP OPEN object by defining a new optional TLV to indicate the PCE's capability to perform backup egress computation for an MPLS TE P2MP LSP and an MPLS TE P2P LSP.

We request IANA to allocate a value such as 8 from the "PCEP TLV Type Indicators" subregistry, as documented in Section below ("Backup Egress Capability TLV"). The description is "backup egress capable", and the length value is 2 bytes. The value field is set to indicate the capability of a PCE for backup egress computation for an MPLS TE LSP in details.

We can use flag bits in the value field in the same way as the PCE Capability Flags described in the previous section.

The inclusion of this TLV in an OPEN object indicates that the sender can perform backup egress computation for an MPLS TE P2MP LSP or an MPLS TE P2P LSP.

The capability TLV is meaningful only for a PCE, so it will typically appear only in one of the two Open messages during PCE session establishment. However, in case of PCE cooperation (e.g., inter-domain), when a PCE behaving as a PCC initiates a PCE session it SHOULD also indicate its path computation capabilities.

4.2. Request and Reply Message Extension

This section describes extensions to the existing RP (Request Parameters) object to allow a PCC to request a PCE for computing a backup egress of an MPLS TE P2MP LSP or an MPLS TE P2P LSP when the PCE receives the PCEP request.

4.2.1. RP Object Extension

The following flags are added into the RP Object:

The T bit is added in the flag bits field of the RP object to tell the receiver of the message that the request/reply is for computing a bcakup egress of an MPLS TE P2MP LSP and an MPLS TE P2P LSP.

- o T (Backup Egress bit - 1 bit):
  - 0: This indicates that this is not PCReq/PCRep for backup egress.
  - 1: This indicates that this is PCReq or PCRep message for backup egress.

The IANA request is referenced in Section below (Request Parameter Bit Flags) of this document.

This T bit with the N bit defined in RFC 6006 can indicate whether a request/reply is for a bcakup egress of an MPLS TE P2MP LSP or an MPLS TE P2P LSP.

- o T = 1 and N = 1: This indicates that this is a PCReq/PCRep message for backup egress of an MPLS TE P2MP LSP.
- o T = 1 and N = 0: This indicates that this is a PCReq/PCRep message for backup egress of an MPLS TE P2P LSP.

4.2.2. External Destination Nodes

In addition to the information about the path that an MPLS TE P2MP LSP or an MPLS TE P2P LSP traverses, a request message may comprise other information that may be used for computing the backup egress for the P2MP LSP or P2P LSP. For example, the information about an external destination node, to which data traffic is delivered from an egress node of the P2MP LSP or P2P LSP, is useful for computing a backup egress node.

4.2.2.1. External Destination Nodes Object

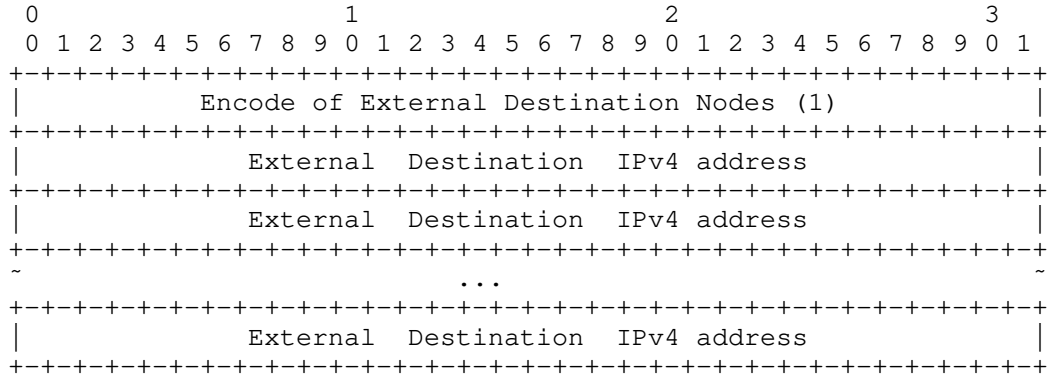
The PCC can specify an external destination nodes (EDN) Object. In order to represent the external destination nodes efficiently, we define two types of encodes for the external destination nodes in the object.

One encode indicates that the EDN object contains an external destination node for every egress node of an MPLS TE P2MP LSP or an MPLS TE P2P LSP. The order of the external destination nodes in the object is the same as the egress node(s) of the P2MP LSP or P2P LSP contained in the PCE messages.

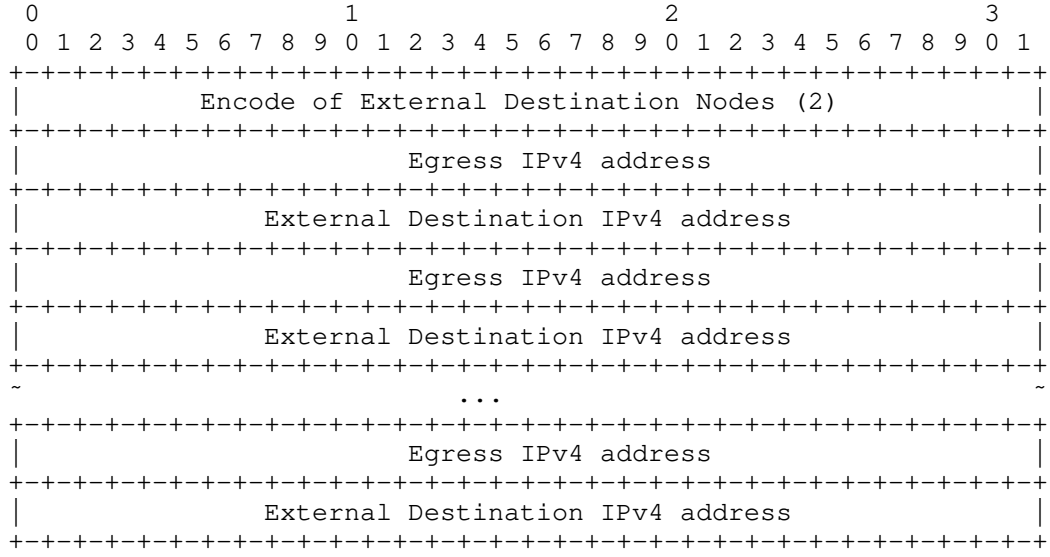
Another encode indicates that the EDN object contains a list of egress node and external destination node pairs. For an egress node

and external destination node pair, the data traffic is delivered to the external destination node from the egress node of the LSP.

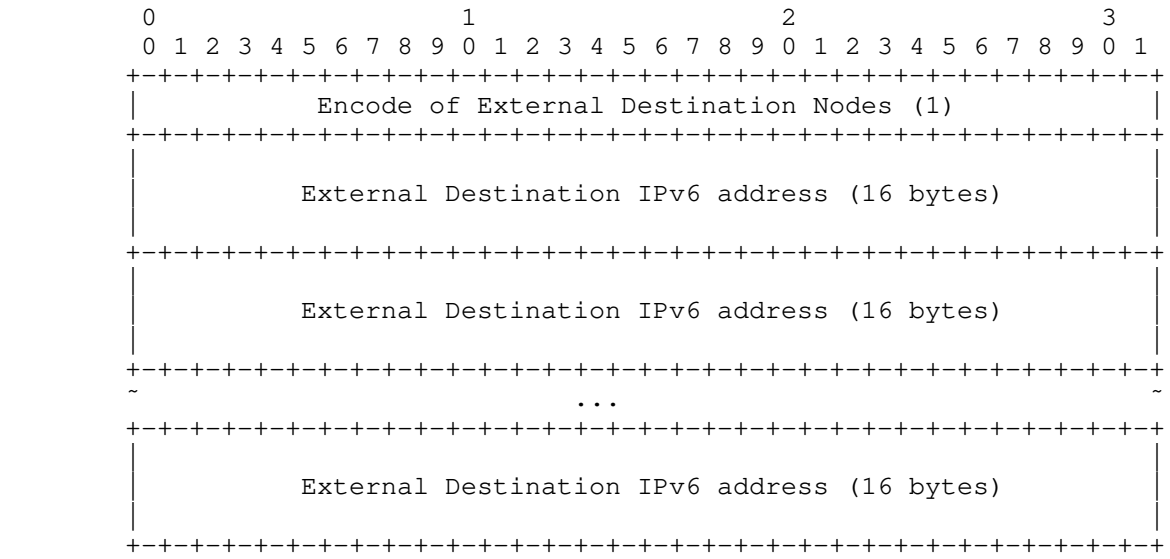
The format of the external destination nodes (EDN) object boby for IPv4 with the first type of encodes is illustrated as follows:



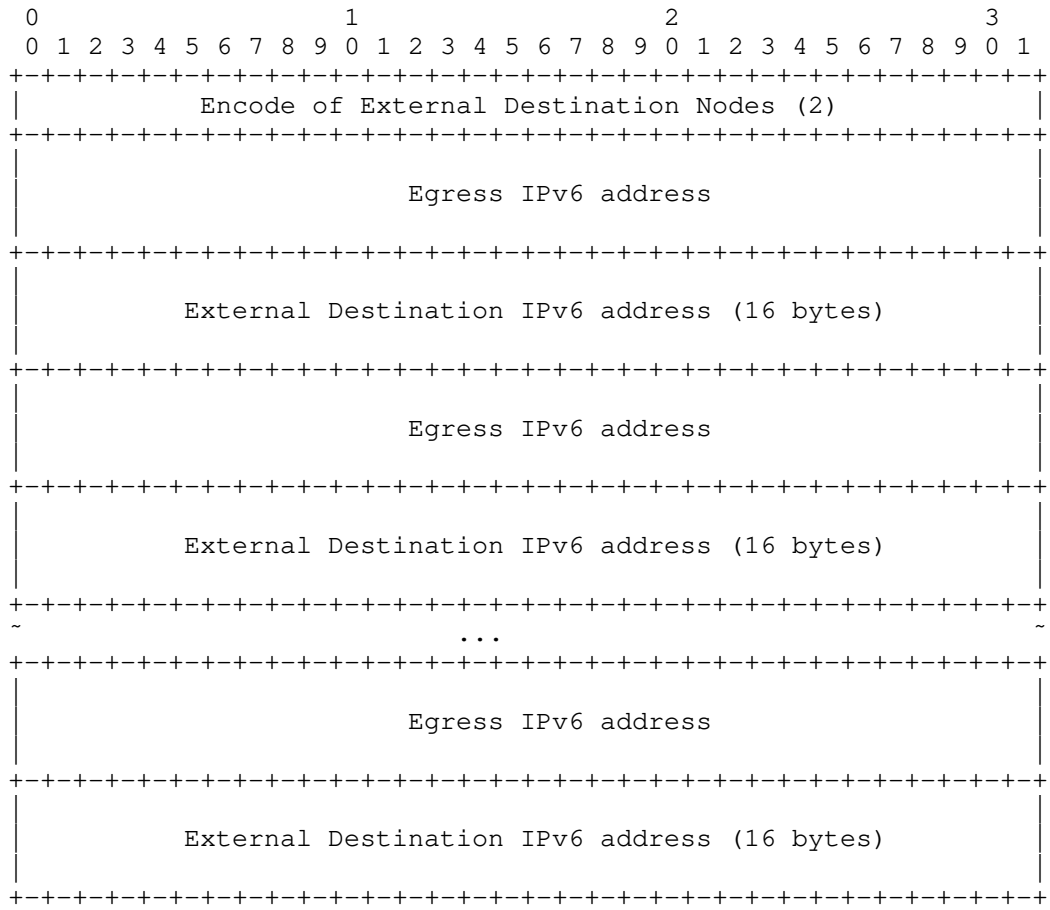
The format of the external destination nodes (EDN) object boby for IPv4 with the second type of encodes is illustrated below:



The format of the external destination nodes (EDN) object boby for IPv6 with the first type of encodes is illustrated as follows:



The format of the external destination nodes (EDN) object boby for IPv6 with the second type of encodes is illustrated below:



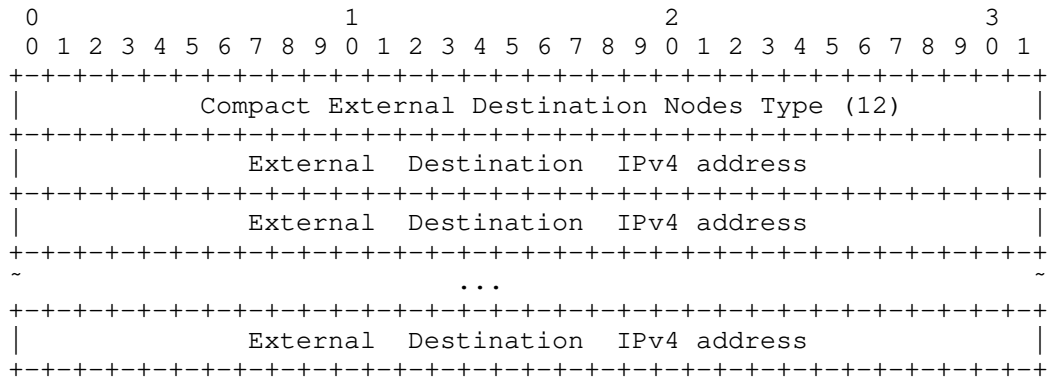
The object can only be carried in a PCReq message. A Path Request may carry at most one external destination nodes Object.

The Object-Class and Object-types will need to be allocated by IANA. The IANA request is documented in Section below (PCEP Objects).

4.2.2.2. New Type of END-POINTS Objects

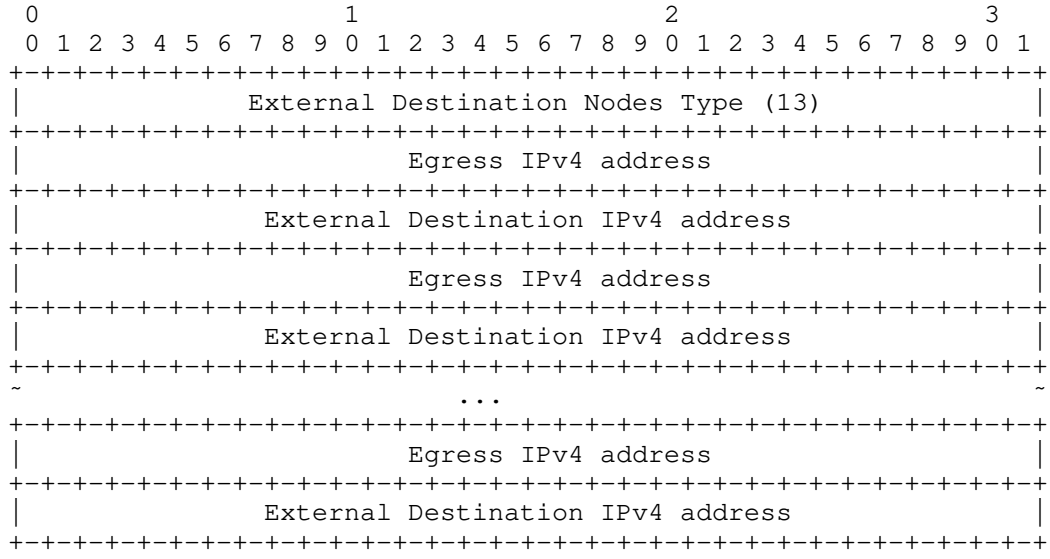
Alternatively, we may use END-POINTS to represent external destination nodes in a request message for computing backup egress nodes of MPLS LSP.

The format of the external destination nodes (EDN) END-POINTS object boby for IPv4 with the first type of encodes is illustrated as follows:



The new type of END-POINTS is Compact External Destination Nodes Type (12). The final value for the type will be assigned by IANA. The EDN END-POINTS object of type 12 contains an external destination node for every egress node of an MPLS TE P2MP LSP or an MPLS TE P2P LSP. The order of the external destination nodes in the object is the same as the egress node(s) of the P2MP LSP or P2P LSP contained in the PCE messages.

The format of the external destination nodes END-POINTS object boby for IPv4 with the second type of encodes is illustrated below:



The new type of END-POINTS is External Destination Nodes Type (13). The final value for the type will be assigned by IANA. The EDN END-POINTS object of type 13 contains a list of egress node and external destination node pairs. For an egress node and external destination node pair, the data traffic is delivered to the external destination node from the egress node of the LSP.

4.2.3. Constraints between Egress and Backup Egress

A request message sent to a PCE from a PCC for computing a backup egress of an MPLS TE P2MP LSP or an MPLS TE P2P LSP may comprise a constraint indicating that there must be a path from the backup egress node to be computed to the egress node of the P2MP LSP or P2P LSP and that the length of the path is within a given hop limit such as one hop.

This constraint can be considered as default by a PCE or explicitly sent to the PCE by a PCC [TBD].

4.2.4. Constraints for Backup Path

A request message sent to a PCE from a PCC for computing a backup egress of a P2MP LSP or P2P LSP may comprise a constraint indicating that the backup egress node to be computed may not be a node on the P2MP LSP or P2P LSP. In addition, the request message may comprise a list of nodes, each of which is a candidate for the backup egress node.

A request message sent to a PCE from a PCC for computing a backup egress of a P2MP LSP or P2P LSP may comprise a constraint indicating that there must be a path from the previous hop node of the egress node of the P2MP LSP or P2P LSP to the backup egress node to be computed and that there is not an internal node of the path from the previous hop node of the egress node of the P2MP LSP or P2P LSP to the backup egress that is on the path of the P2MP LSP or P2P LSP.

Most of these constraints for the backup path can be considered as default by a PCE. The constraints for the backup path may be explicitly sent to the PCE by a PCC [TBD].

4.2.5. Backup Egress Node

The PCE may send a reply message to the PCC in return to the request message for computing a new backup egress node or a number of backup egress nodes. The reply message may comprise information about the computed backup egress node(s), which is contained in the path(s) from the previous-hop node of the egress node of the P2MP LSP or P2P LSP to the backup egress node(s) computed.

4.2.6. Backup Egress PCEP Error Objects and Types

In some cases, the PCE may not complete the backup egress computation as requested, for example based on a set of constraints. As such, the PCE may send a reply message to the PCC that indicates an unsuccessful backup egress computation attempt. The reply message may comprise a PCEP-error object, which may comprise an error-value, error-type and some detail information.

4.2.7. Request Message Format

The PCReq message is encoded as follows using RBNF as defined in [RFC5511].

Below is the message format for a request message:

```
            <PCReq Message>::= <Common Header>
                               [<svec-list>]
                               <request>
            <request>::= <RP> <end-point-rro-pair-list>
                         [<OF>] [<LSPA>] [<BANDWIDTH>]
                         [<metric-list>]
                         [<EDNO>]
                         [<IRO>]
                         [<LOAD-BALANCING>]
        where:
               <EDNO> is an external destination nodes object.
```

The definitions for svec-list, RP, end-point-rro-pair-list, OF, LSPA, BANDWIDTH, metric-list, IRO, and LOAD-BALANCING are described in RFC5440 and RFC6006.

4.2.8. Reply Message Format

The PCRep message is encoded as follows using RBNF as defined in [RFC5511].

Below is the message format for a reply message:

```
<PCRep Message>::= <Common Header>
                   <response>
<response>::= <RP>
              <end-point-path-pair-list>
              [<NO-PATH>]
              [<attribute-list>]
where:
      <end-point-path-pair-list>::=
             [<END-POINTS>]<path>[<end-point-path-pair-list>]

      <path> ::= (<ERO>|<SERO>) [<path>]

      <attribute-list>::= [<OF>]
                          [<LSPA>]
                          [<BANDWIDTH>]
                          [<metric-list>]
                          [<IRO>]
```

The definitions for RP, NO-PATH, END-POINTS, OF, LSPA, BANDWIDTH, metric-list, IRO, and SERO are described in RFC5440, RFC6006 and RFC4875.

5. Security Considerations

The mechanism described in this document does not raise any new security issues for the PCEP, OSPF or IS-IS protocols.

6. IANA Considerations

This section specifies requests for IANA allocation.

6.1. Backup Egress Capability Flag

Two new OSPF Capability Flags are defined in this document to indicate the capabilities for computing a backup egress for an MPLS TE P2MP LSP and an MPLS TE P2P LSP. IANA is requested to make the assignment from the "OSPF Parameters Path Computation Element (PCE) Capability Flags" registry:

Bit	Description	Reference
13	Backup egress for P2MP LSP	This I-D
14	Backup egress for P2P LSP	This I-D

6.2. Backup Egress Capability TLV

A new backup egress capability TLV is defined in this document to allow a PCE to advertize its backup egress computation capability. IANA is requested to make the following allocation from the "PCEP TLV Type Indicators" sub-registry.

Value	Description	Reference
8	Backup egress capable	This I-D

6.3. Request Parameter Bit Flags

A new RP Object Flag has been defined in this document. IANA is requested to make the following allocation from the "PCEP RP Object Flag Field" Sub-Registry:

Bit	Description	Reference
15	Backup egress (T-bit)	This I-D

6.4. PCEP Objects

An External Destination Nodes Object-Type is defined in this document. IANA is requested to make the following Object-Type allocation from the "PCEP Objects" sub-registry:

Object-Class Value	34
Name	External Destination Nodes
Object-Type	1: IPv4 2: IPv6 3-15: Unassigned
Reference	This I-D

7. Acknowledgement

The author would like to thank Ramon Casellas, Dhruv Dhody and Quintin Zhao for their valuable comments on this draft.

8. References

8.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, DOI 10.17487/RFC3209, December 2001, <https://www.rfc-editor.org/info/rfc3209>.

[RFC4090] Pan, P., Ed., Swallow, G., Ed., and A. Atlas, Ed., "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, DOI 10.17487/RFC4090, May 2005, <https://www.rfc-editor.org/info/rfc4090>.

[RFC4875] Aggarwal, R., Ed., Papadimitriou, D., Ed., and S. Yasukawa, Ed., "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, DOI 10.17487/RFC4875, May 2007, <https://www.rfc-editor.org/info/rfc4875>.

[RFC5440] Vasseur, JP., Ed. and JL. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, DOI 10.17487/RFC5440, March 2009, <https://www.rfc-editor.org/info/rfc5440>.

[RFC6006] Zhao, Q., Ed., King, D., Ed., Verhaeghe, F., Takeda, T., Ali, Z., and J. Meuric, "Extensions to the Path Computation Element Communication Protocol (PCEP) for Point-to-Multipoint Traffic Engineering Label Switched Paths", RFC 6006, DOI 10.17487/RFC6006, September 2010, <https://www.rfc-editor.org/info/rfc6006>.

8.2. Informative References

[RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, DOI 10.17487/RFC4655, August 2006, <https://www.rfc-editor.org/info/rfc4655>.

[RFC5862] Yasukawa, S. and A. Farrel, "Path Computation Clients (PCC) - Path Computation Element (PCE) Requirements for Point-to-Multipoint MPLS-TE", RFC 5862, DOI 10.17487/RFC5862, June 2010, <https://www.rfc-editor.org/info/rfc5862>.

Author's Address

Huaimo Chen
Futurewei
Boston, MA
USA

Email: Huaimo.chen@futurewei.com

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: May 2, 2012

H. Chen
Huawei Technologies

October 30, 2011

Extensions to Path Computation Element Communication Protocol (PCEP) for Backup Ingress of a Traffic Engineering Label Switched Path

draft-chen-pce-compute-backup-ingress-03.txt

Abstract

This document presents extensions to the Path Computation Element Communication Protocol (PCEP) for a PCC to send a request for computing a backup ingress for an MPLS TE P2MP LSP or an MPLS TE P2P LSP to a PCE and for a PCE to compute the backup ingress and reply to the PCC with a computation result for the backup ingress.

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). 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 May 2, 2012.

Copyright Notice

Table of Contents

1. Introduction

RFC4090 "Fast Reroute Extensions to RSVP-TE for LSP Tunnels" describes two methods to protect P2P LSP tunnels or paths at local repair points. The local repair points may comprise a number of intermediate nodes between an ingress node and an egress node along the path. The first method is a one-to-one backup method, where a detour backup P2P LSP for each protected P2P LSP is created at each potential point of local repair. The second method is a facility bypass backup protection method, where a bypass backup P2P LSP tunnel is created using MPLS label stacking to protect a potential failure point for a set of P2P LSP tunnels. The bypass backup tunnel can protect a set of P2P LSPs that have similar backup constraints.

RFC4875 "Extensions to RSVP-TE for P2MP TE LSPs" describes how to use the one-to-one backup method and facility bypass backup method to protect a link or intermediate node failure on the path of a P2MP LSP.

However, there is no mention of locally protecting an ingress node failure in a protected P2MP LSP or P2P LSP.

The methods for protecting an ingress node of a P2MP LSP or P2P LSP may be classified into two categories.

A first category uses a backup P2MP LSP that is from a backup ingress node to the number of destination nodes for the P2MP LSP, and a backup P2P LSP that is from a backup ingress node to the destination node for the P2P LSP. The disadvantages of this class of methods include more network resource such as computer power and link bandwidth consumption since the backup P2MP LSP or P2P LSP is from the backup ingress node to the number of destination nodes or the destination respectively.

A second category uses a local P2MP LSP or P2P LSP for protecting the ingress of a P2MP LSP or P2P LSP locally. The local P2MP LSP is from a backup ingress node to the next hop nodes of the ingress of the P2MP LSP. The local P2P LSP is from a backup ingress node to the next hop node of the ingress of the P2P LSP. It is desirable to let PCE compute these backup ingress nodes.

This document defines extensions to the Path Computation Element Communication Protocol (PCEP) for a PCC to send a request for computing a backup ingress node for an MPLS TE P2MP LSP or an MPLS TE P2P LSP to a PCE and for a PCE to compute the backup ingress node and reply to the PCC with a computation result for the backup ingress node.

2. Terminology

This document uses terminologies defined in RFC5440, RFC4090, and RFC4875.

3. 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.

4. Extensions to PCEP

This section describes the extensions to PCEP for computing a backup ingress of an MPLS TE P2MP LSP and an MPLS TE P2P LSP.

4.1. Backup Ingress Capability Advertisement

4.1.1. Capability TLV in Existing PCE Discovery Protocol

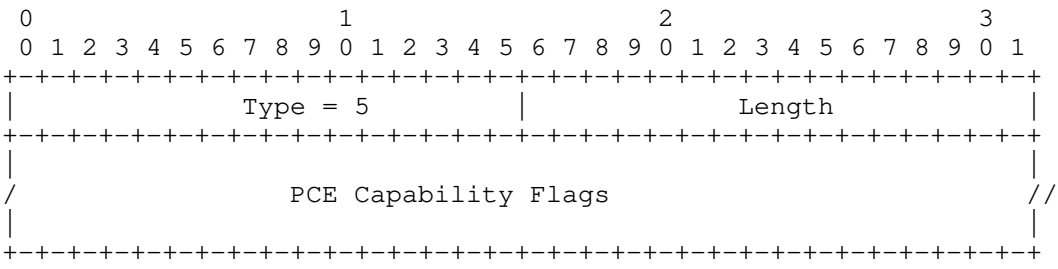
There are a couple of options for advertising a PCE capability for computing a backup ingress for an MPLS TE P2MP LSP or an MPLS TE P2P LSP.

The first option is to define a new flag in the OSPF and ISIS PCE Capability Flags to indicate the capability that a PCE is capable to compute both a backup ingress for an MPLS TE P2MP LSP and a backup ingress for an MPLS TE P2P LSP.

The second option is to define two new flags. One new flag in the OSPF and ISIS PCE Capability Flags indicates the capability that a PCE is capable to compute a backup ingress for an MPLS TE P2MP LSP; and another new flag in the OSPF and ISIS PCE Capability Flags indicates the capability that a PCE is capable to compute a backup ingress for an MPLS TE P2P LSP.

This second option is preferred now.

The format of the PCE-CAP-FLAGS sub-TLV is as follows:



Type: 5
Length: Multiple of 4 octets
Value: This contains an array of units of 32-bit flags numbered from the most significant as bit zero, where each bit represents one PCE capability.

The following capability bits have been assigned by IANA:

Bit	Capabilities
0	Path computation with GMPLS link constraints
1	Bidirectional path computation
2	Diverse path computation
3	Load-balanced path computation
4	Synchronized path computation
5	Support for multiple objective functions
6	Support for additive path constraints (max hop count, etc.)
7	Support for request prioritization
8	Support for multiple requests per message
9	Global Concurrent Optimization (GCO)
10	P2MP path computation
11-31	Reserved for future assignments by IANA.

Reserved bits SHOULD be set to zero on transmission and MUST be ignored on receipt.

For the second option, one bit such as bit 11 may be assigned to indicate that a PCE is capable to compute a backup ingress for an MPLS TE P2MP LSP and another bit such as bit 12 may be assigned to indicate that a PCE is capable to compute a backup ingress for an MPLS TE P2P LSP.

Bit	Capabilities
11	Backup ingress computation for P2MP LSP
12	Backup ingress computation for P2P LSP
13-31	Reserved for future assignments by IANA.

4.1.2. Open Message Extension

If a PCE does not advertise its backup ingress compution capability during discovery, PCEP should be used to allow a PCC to discover, during the Open Message Exchange, which PCEs are capable of supporting backup ingress computation.

To achieve this, we extend the PCEP OPEN object by defining a new optional TLV to indicate the PCE's capability to perform backup ingress compution for an MPLS TE P2MP LSP and an MPLS TE P2P LSP.

We request IANA to allocate a value such as 8 from the "PCEP TLV Type Indicators" subregistry, as documented in Section below ("Backup Ingress Capability TLV"). The description is "backup ingress capable", and the length value is 2 bytes. The value field is set to indicate the capability of a PCE for backup ingress compution for an MPLS TE LSP in details.

We can use flag bits in the value field in the same way as the PCE Capability Flags described in the previous section.

The inclusion of this TLV in an OPEN object indicates that the sender can perform backup ingress compution for an MPLS TE P2MP LSP or an MPLS TE P2P LSP.

The capability TLV is meaningful only for a PCE, so it will typically appear only in one of the two Open messages during PCE session establishment. However, in case of PCE cooperation (e.g., inter-domain), when a PCE behaving as a PCC initiates a PCE session it SHOULD also indicate its path computation capabilities.

4.2. Request and Reply Message Extension

This section describes extensions to the existing RP (Request Parameters) object to allow a PCC to request a PCE for computing a backup ingress of an MPLS TE P2MP LSP or an MPLS TE P2P LSP when the PCE receives the PCEP request.

4.2.1. RP Object Extension

The following flags are added into the RP Object:

The I bit is added in the flag bits field of the RP object to tell the receiver of the message that the request/reply is for computing a bcakup ingress of an MPLS TE P2MP LSP and an MPLS TE P2P LSP.

- o I (Backup Ingress bit - 1 bit):

  0: This indicates that this is not PCReq/PCRep for backup ingress.

  1: This indicates that this is PCReq or PCRep message for backup ingress.

The IANA request is referenced in Section below (Request Parameter Bit Flags) of this document.

This I bit with the N bit defined in RFC6006 can indicate whether the request/reply is for a bcakup ingress of an MPLS TE P2MP LSP or an MPLS TE P2P LSP.

- o I = 1 and N = 1: This indicates that this is a PCReq/PCRep message for backup ingress of an MPLS TE P2MP LSP.

- o I = 1 and N = 0: This indicates that this is a PCReq/PCRep message for backup ingress of an MPLS TE P2P LSP.

4.2.2. External Source Node

In addition to the information about the path that an MPLS TE P2MP LSP or an MPLS TE P2P LSP traverses, a request message may comprise other information that may be used for computing the backup ingress for the P2MP LSP or P2P LSP. For example, the information about an external source node, from which data traffic is delivered to the ingress node of the P2MP LSP or P2P LSP and transported to the egress node(s) via the P2MP LSP or P2P LSP, is useful for computing a backup ingress node.

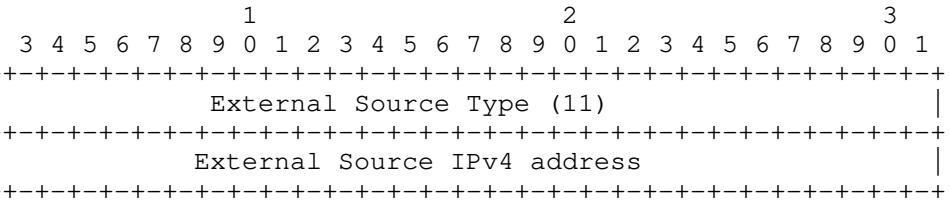
The PCC can specify an external source node (ESN) Object. The ESN Object has the same format as the IRO object defined in [RFC5440] except that it only supports IPv4 and IPv6 prefix sub-objects.

The object can only be carried in a PCReq message. A Path Request may carry at most one external source node Object.

The Object-Class and Object-types will need to be allocated by IANA. The IANA request is documented in Section below. (PCEP Objects).

Alternatively, we may use END-POINTS to represent an external source node in a request message for computing a backup ingress node of MPLS LSP.

To represent an external source node efficiently, we define a new type of END-POINTS objects for computing a backup ingress node of MPLS LSP. The format of the new END-POINTS object body for IPv4 (Object-Type 3) is as follows:



The new type of END-POINTS is External Source Node Type (11). The final value for the type will be assigned by IANA. This new type of END-POINTS object contains an external source node IPv4 address.

4.2.3. Constraints between Ingress and Backup Ingress

A request message sent to a PCE from a PCC for computing a backup ingress of an MPLS TE P2MP LSP or an MPLS TE P2P LSP may comprise a constraint indicating that there must be a path from the backup ingress node to be computed to the ingress node of the P2MP LSP or P2P LSP and that the length of the path is within a given hop limit such as one hop.

This constraint can be considered as default by a PCE or explicitly sent to the PCE by a PCC [TBD].

4.2.4. Constraints for Backup Path

A request message sent to a PCE from a PCC for computing a backup ingress of a P2MP LSP or P2P LSP may comprise a constraint indicating that the backup ingress node to be computed may not be a node on the P2MP LSP or P2P LSP. In addition, the request message may comprise a list of nodes, each of which is a candidate for the backup ingress node.

A request message sent to a PCE from a PCC for computing a backup ingress of a P2MP LSP or P2P LSP may comprise a constraint indicating that there must be a path from the backup ingress node to be computed to the next-hop nodes of the ingress node of the P2MP LSP or P2P LSP and that there is not an internal node of the path from the backup ingress to the next-hop nodes on the P2MP LSP or P2P LSP .

Most of these constraints for the backup path can be considered as default by a PCE. The constraints for the backup path may be explicitly sent to the PCE by a PCC [TBD].

4.2.5. Backup Ingress Node

The PCE may send a reply message to the PCC in return to the request message for computing a new backup ingress node. The reply message may comprise information about the computed backup ingress node, which is contained in the path from the backup ingress computed to the next-hop node(s) of the ingress node of the P2MP LSP or P2P LSP.

The backup ingress node is the root or head node of the backup path computed.

4.2.6. Backup Ingress PCEP Error Objects and Types

In some cases, the PCE may not complete the backup ingress computation as requested, for example based on a set of constraints. As such, the PCE may send a reply message to the PCC that indicates an unsuccessful backup ingress computation attempt. The reply message may comprise a PCEP-error object, which may comprise an error-value, error-type and some detail information.

4.2.7. Request Message Format

The PCReq message is encoded as follows using RBNF as defined in [RFC5511].

Below is the message format for a request message:

```
<PCReq Message>::= <Common Header>
                   [<svec-list>]
                   <request>
<request>::= <RP>
             <end-point-rro-pair-list>
             [<OF>]
             [<LSPA>]
             [<BANDWIDTH>]
             [<metric-list>]
             [<ESNO>]
             [<IRO>]
             [<LOAD-BALANCING>]
where:
      <ESNO> is an external source node object.
```

Figure 1: The Format for a Request Message

The definitions for svec-list, RP, end-point-rro-pair-list, OF, LSPA, BANDWIDTH, metric-list, IRO, and LOAD-BALANCING are described in RFC5440 and RFC6006.

4.2.8. Reply Message Format

The PCRep message is encoded as follows using RBNF as defined in [RFC5511].

Below is the message format for a reply message:

```
<PCRep Message>::= <Common Header>
                   <response>
<response>::= <RP>
              <end-point-path-pair-list>
              [<NO-PATH>]
              [<attribute-list>]
```

where:

```
<end-point-path-pair-list>::=
        [<END-POINTS>]<path>[<end-point-path-pair-list>]


<path> ::= (<ERO>|<SERO>) [<path>]

<attribute-list>::= [<OF>]
                    [<LSPA>]
                    [<BANDWIDTH>]
                    [<metric-list>]
                    [<IRO>]
```

Figure 2: The Format for a Reply Message

The definitions for RP, NO-PATH, END-POINTS, OF, LSPA, BANDWIDTH, metric-list, IRO, and SERO are described in RFC5440, RFC6006 and RFC4875.

5. Security Considerations

The mechanism described in this document does not raise any new security issues for the PCEP, OSPF and IS-IS protocols.

6. IANA Considerations

This section specifies requests for IANA allocation.

6.1. Backup Ingress Capability Flag

Two new OSPF Capability Flags are defined in this document to indicate the capabilities for computing a backup ingress for an MPLS TE P2MP LSP and an MPLS TE P2P LSP. IANA is requested to make the assignment from the "OSPF Parameters Path Computation Element (PCE) Capability Flags" registry:

Bit	Description	Reference
11	Backup ingress for P2MP LSP	This I-D
12	Backup ingress for P2P LSP	This I-D

6.2. Backup Ingress Capability TLV

A new backup ingress capability TLV is defined in this document to allow a PCE to advertize its backup ingress computation capability. IANA is requested to make the following allocation from the "PCEP TLV Type Indicators" sub-registry.

Value	Description	Reference
8	Backup ingress capable	This I-D

6.3. Request Parameter Bit Flags

A new RP Object Flag has been defined in this document. IANA is requested to make the following allocation from the "PCEP RP Object Flag Field" Sub-Registry:

Bit	Description	Reference
16	Backup ingress (I-bit)	This I-D

6.4. PCEP Objects

An External Source Node Object-Type is defined in this document. IANA is requested to make the following Object-Type allocation from the "PCEP Objects" sub-registry:

Object-Class Value	33
Name	External Source Node
Object-Type	1: IPv4 2: IPv6 3-15: Unassigned
Reference	This I-D

7. Acknowledgement

The author would like to thank Cyril Margaria, Ramon Casellas, Dhruv Dhody and Quintin Zhao for their valuable comments and suggestions on this draft.

8. References

8.1. Normative References

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

[RFC3209] 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.

[RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.

[RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.

[RFC4875] Aggarwal, R., Papadimitriou, D., and S. Yasukawa, "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, May 2007.

[RFC6006] Zhao, Q., King, D., Verhaeghe, F., Takeda, T., Ali, Z., and J. Meuric, "Extensions to the Path Computation Element Communication Protocol (PCEP) for Point-to-Multipoint Traffic Engineering Label Switched Paths", RFC 6006, September 2010.

8.2. Informative References

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

[RFC5862] Yasukawa, S. and A. Farrel, "Path Computation Clients (PCC) - Path Computation Element (PCE) Requirements for Point-to-Multipoint MPLS-TE", RFC 5862, June 2010.

Author's Address

Huaimo Chen
Huawei Technologies
Boston, MA
USA

Email: Huaimochen@huawei.com

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: January 8, 2020

H. Chen
Futurewei

July 7, 2019

Extensions to Path Computation Element Communication Protocol (PCEP) for Backup Ingress of a Traffic Engineering Label Switched Path draft-chen-pce-compute-backup-ingress-14.txt

Abstract

This document presents extensions to the Path Computation Element Communication Protocol (PCEP) for a PCC to send a request for computing a backup ingress for an MPLS TE P2MP LSP or an MPLS TE P2P LSP to a PCE and for a PCE to compute the backup ingress and reply to the PCC with a computation result for the backup ingress.

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). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://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 January 8, 2020.

Table of Contents

1. Introduction

RFC4090 "Fast Reroute Extensions to RSVP-TE for LSP Tunnels" describes two methods to protect P2P LSP tunnels or paths at local repair points. The local repair points may comprise a number of intermediate nodes between an ingress node and an egress node along the path. The first method is a one-to-one backup method, where a detour backup P2P LSP for each protected P2P LSP is created at each potential point of local repair. The second method is a facility bypass backup protection method, where a bypass backup P2P LSP tunnel is created using MPLS label stacking to protect a potential failure point for a set of P2P LSP tunnels. The bypass backup tunnel can protect a set of P2P LSPs that have similar backup constraints.

RFC4875 "Extensions to RSVP-TE for P2MP TE LSPs" describes how to use the one-to-one backup method and facility bypass backup method to protect a link or intermediate node failure on the path of a P2MP LSP.

However, there is no mention of locally protecting an ingress node failure in a protected P2MP LSP or P2P LSP.

The methods for protecting an ingress node of a P2MP LSP or P2P LSP may be classified into two categories.

A first category uses a backup P2MP LSP that is from a backup ingress node to the number of destination nodes for the P2MP LSP, and a backup P2P LSP that is from a backup ingress node to the destination node for the P2P LSP. The disadvantages of this class of methods include more network resource such as computer power and link bandwidth consumption since the backup P2MP LSP or P2P LSP is from the backup ingress node to the number of destination nodes or the destination respectively.

A second category uses a local P2MP LSP or P2P LSP for protecting the ingress of a P2MP LSP or P2P LSP locally. The local P2MP LSP is from a backup ingress node to the next hop nodes of the ingress of the P2MP LSP. The local P2P LSP is from a backup ingress node to the next hop node of the ingress of the P2P LSP. It is desirable to let PCE compute these backup ingress nodes.

This document defines extensions to the Path Computation Element Communication Protocol (PCEP) for a PCC to send a request for computing a backup ingress node for an MPLS TE P2MP LSP or an MPLS TE P2P LSP to a PCE and for a PCE to compute the backup ingress node and reply to the PCC with a computation result for the backup ingress node.

2. Terminology

This document uses terminologies defined in RFC5440, RFC4090, and RFC4875.

3. 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.

4. Extensions to PCEP

This section describes the extensions to PCEP for computing a backup ingress of an MPLS TE P2MP LSP and an MPLS TE P2P LSP.

4.1. Backup Ingress Capability Advertisement

4.1.1. Capability TLV in Existing PCE Discovery Protocol

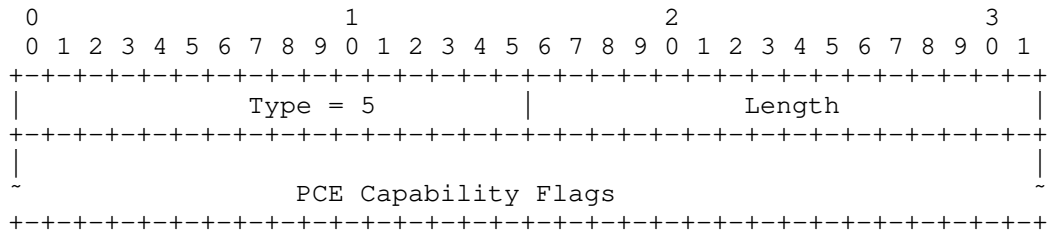
There are a couple of options for advertising a PCE capability for computing a backup ingress for an MPLS TE P2MP LSP or an MPLS TE P2P LSP.

The first option is to define a new flag in the OSPF and ISIS PCE Capability Flags to indicate the capability that a PCE is capable to compute both a backup ingress for an MPLS TE P2MP LSP and a backup ingress for an MPLS TE P2P LSP.

The second option is to define two new flags. One new flag in the OSPF and ISIS PCE Capability Flags indicates the capability that a PCE is capable to compute a backup ingress for an MPLS TE P2MP LSP; and another new flag in the OSPF and ISIS PCE Capability Flags indicates the capability that a PCE is capable to compute a backup ingress for an MPLS TE P2P LSP.

This second option is preferred now.

The format of the PCE-CAP-FLAGS sub-TLV is as follows:



Type: 5
Length: Multiple of 4 octets
Value: This contains an array of units of 32-bit flags numbered from the most significant as bit zero, where each bit represents one PCE capability.

The following capability bits have been assigned by IANA:

Bit	Capabilities
0	Path computation with GMPLS link constraints
1	Bidirectional path computation
2	Diverse path computation
3	Load-balanced path computation
4	Synchronized path computation
5	Support for multiple objective functions
6	Support for additive path constraints (max hop count, etc.)
7	Support for request prioritization
8	Support for multiple requests per message
9	Global Concurrent Optimization (GCO)
10	P2MP path computation
11-31	Reserved for future assignments by IANA.

Reserved bits SHOULD be set to zero on transmission and MUST be ignored on receipt.

For the second option, one bit such as bit 11 may be assigned to indicate that a PCE is capable to compute a backup ingress for an MPLS TE P2MP LSP and another bit such as bit 12 may be assigned to indicate that a PCE is capable to compute a backup ingress for an MPLS TE P2P LSP.

Bit	Capabilities
11	Backup ingress computation for P2MP LSP
12	Backup ingress computation for P2P LSP
13-31	Reserved for future assignments by IANA.

4.1.2. Open Message Extension

If a PCE does not advertise its backup ingress compution capability during discovery, PCEP should be used to allow a PCC to discover, during the Open Message Exchange, which PCEs are capable of supporting backup ingress computation.

To achieve this, we extend the PCEP OPEN object by defining a new optional TLV to indicate the PCE's capability to perform backup ingress compution for an MPLS TE P2MP LSP and an MPLS TE P2P LSP.

We request IANA to allocate a value such as 8 from the "PCEP TLV Type Indicators" subregistry, as documented in Section below ("Backup Ingress Capability TLV"). The description is "backup ingress capable", and the length value is 2 bytes. The value field is set to indicate the capability of a PCE for backup ingress compution for an MPLS TE LSP in details.

We can use flag bits in the value field in the same way as the PCE Capability Flags described in the previous section.

The inclusion of this TLV in an OPEN object indicates that the sender can perform backup ingress compution for an MPLS TE P2MP LSP or an MPLS TE P2P LSP.

The capability TLV is meaningful only for a PCE, so it will typically appear only in one of the two Open messages during PCE session establishment. However, in case of PCE cooperation (e.g., inter-domain), when a PCE behaving as a PCC initiates a PCE session it SHOULD also indicate its path computation capabilities.

4.2. Request and Reply Message Extension

This section describes extensions to the existing RP (Request Parameters) object to allow a PCC to request a PCE for computing a backup ingress of an MPLS TE P2MP LSP or an MPLS TE P2P LSP when the PCE receives the PCEP request.

4.2.1. RP Object Extension

The following flags are added into the RP Object:

The I bit is added in the flag bits field of the RP object to tell the receiver of the message that the request/reply is for computing a bcakup ingress of an MPLS TE P2MP LSP and an MPLS TE P2P LSP.

- o I (Backup Ingress bit - 1 bit):
  - 0: This indicates that this is not PCReq/PCRep for backup ingress.
  - 1: This indicates that this is PCReq or PCRep message for backup ingress.

The IANA request is referenced in Section below (Request Parameter Bit Flags) of this document.

This I bit with the N bit defined in RFC6006 can indicate whether the request/reply is for a bcakup ingress of an MPLS TE P2MP LSP or an MPLS TE P2P LSP.

- o I = 1 and N = 1: This indicates that this is a PCReq/PCRep message for backup ingress of an MPLS TE P2MP LSP.
- o I = 1 and N = 0: This indicates that this is a PCReq/PCRep message for backup ingress of an MPLS TE P2P LSP.

4.2.2. External Source Node

In addition to the information about the path that an MPLS TE P2MP LSP or an MPLS TE P2P LSP traverses, a request message may comprise other information that may be used for computing the backup ingress for the P2MP LSP or P2P LSP. For example, the information about an external source node, from which data traffic is delivered to the ingress node of the P2MP LSP or P2P LSP and transported to the egress node(s) via the P2MP LSP or P2P LSP, is useful for computing a backup ingress node.

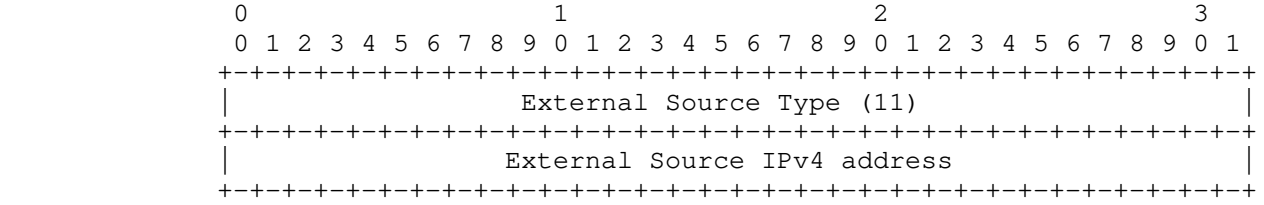
The PCC can specify an external source node (ESN) Object. The ESN Object has the same format as the IRO object defined in [RFC5440] except that it only supports IPv4 and IPv6 prefix sub-objects.

The object can only be carried in a PCReq message. A Path Request may carry at most one external source node Object.

The Object-Class and Object-types will need to be allocated by IANA. The IANA request is documented in Section below. (PCEP Objects).

Alternatively, we may use END-POINTS to represent an external source node in a request message for computing a backup ingress node of MPLS LSP.

To represent an external source node efficiently, we define a new type of END-POINTS objects for computing a backup ingress node of MPLS LSP. The format of the new END-POINTS object body for IPv4 (Object-Type 3) is as follows:



The new type of END-POINTS is External Source Node Type (11). The final value for the type will be assigned by IANA. This new type of END-POINTS object contains an external source node IPv4 address.

4.2.3. Constraints between Ingress and Backup Ingress

A request message sent to a PCE from a PCC for computing a backup ingress of an MPLS TE P2MP LSP or an MPLS TE P2P LSP may comprise a constraint indicating that there must be a path from the backup ingress node to be computed to the ingress node of the P2MP LSP or P2P LSP and that the length of the path is within a given hop limit such as one hop.

This constraint can be considered as default by a PCE or explicitly sent to the PCE by a PCC [TBD].

4.2.4. Constraints for Backup Path

A request message sent to a PCE from a PCC for computing a backup ingress of a P2MP LSP or P2P LSP may comprise a constraint indicating that the backup ingress node to be computed may not be a node on the P2MP LSP or P2P LSP. In addition, the request message may comprise a list of nodes, each of which is a candidate for the backup ingress node.

A request message sent to a PCE from a PCC for computing a backup ingress of a P2MP LSP or P2P LSP may comprise a constraint indicating that there must be a path from the backup ingress node to be computed to the next-hop nodes of the ingress node of the P2MP LSP or P2P LSP and that there is not an internal node of the path from the backup ingress to the next-hop nodes on the P2MP LSP or P2P LSP .

Most of these constraints for the backup path can be considered as default by a PCE. The constraints for the backup path may be explicitly sent to the PCE by a PCC [TBD].

4.2.5. Backup Ingress Node

The PCE may send a reply message to the PCC in return to the request message for computing a new backup ingress node. The reply message may comprise information about the computed backup ingress node, which is contained in the path from the backup ingress computed to the next-hop node(s) of the ingress node of the P2MP LSP or P2P LSP.

The backup ingress node is the root or head node of the backup path computed.

4.2.6. Backup Ingress PCEP Error Objects and Types

In some cases, the PCE may not complete the backup ingress computation as requested, for example based on a set of constraints. As such, the PCE may send a reply message to the PCC that indicates an unsuccessful backup ingress computation attempt. The reply message may comprise a PCEP-error object, which may comprise an error-value, error-type and some detail information.

4.2.7. Request Message Format

The PCReq message is encoded as follows using RBNF as defined in [RFC5511].

Below is the message format for a request message:

```
            <PCReq Message>::= <Common Header>
                               [<svec-list>]
                               <request>
            <request>::= <RP> <end-point-rro-pair-list> [<OF>]
                         [<LSPA>] [<BANDWIDTH>] [<metric-list>]
                         [<ESNO>]
                         [<IRO>]
                         [<LOAD-BALANCING>]
       where:
             <ESNO> is an external source node object.
```

The definitions for svec-list, RP, end-point-rro-pair-list, OF, LSPA, BANDWIDTH, metric-list, IRO, and LOAD-BALANCING are described in RFC5440 and RFC6006.

4.2.8. Reply Message Format

The PCRep message is encoded as follows using RBNF as defined in [RFC5511].

Below is the message format for a reply message:

```
            <PCRep Message>::= <Common Header>
                               <response>
            <response>::= <RP> <end-point-path-pair-list>
                          [<NO-PATH>]
                          [<attribute-list>]
      where:
            <end-point-path-pair-list>::=
                    [<END-POINTS>]<path>[<end-point-path-pair-list>]

            <path> ::= (<ERO>|<SERO>) [<path>]

            <attribute-list>::= [<OF>] [<LSPA>] [<BANDWIDTH>]
                                [<metric-list>] [<IRO>]
```

The definitions for RP, NO-PATH, END-POINTS, OF, LSPA, BANDWIDTH, metric-list, IRO, and SERO are described in RFC5440, RFC6006 and RFC4875.

5. Security Considerations

The mechanism described in this document does not raise any new security issues for the PCEP, OSPF and IS-IS protocols.

6. IANA Considerations

This section specifies requests for IANA allocation.

6.1. Backup Ingress Capability Flag

Two new OSPF Capability Flags are defined in this document to indicate the capabilities for computing a backup ingress for an MPLS TE P2MP LSP and an MPLS TE P2P LSP. IANA is requested to make the assignment from the "OSPF Parameters Path Computation Element (PCE) Capability Flags" registry:

Bit	Description	Reference
11	Backup ingress for P2MP LSP	This I-D
12	Backup ingress for P2P LSP	This I-D

6.2. Backup Ingress Capability TLV

A new backup ingress capability TLV is defined in this document to allow a PCE to advertize its backup ingress computation capability. IANA is requested to make the following allocation from the "PCEP TLV Type Indicators" sub-registry.

Value	Description	Reference
8	Backup ingress capable	This I-D

6.3. Request Parameter Bit Flags

A new RP Object Flag has been defined in this document. IANA is requested to make the following allocation from the "PCEP RP Object Flag Field" Sub-Registry:

Bit	Description	Reference
16	Backup ingress (I-bit)	This I-D

6.4. PCEP Objects

An External Source Node Object-Type is defined in this document. IANA is requested to make the following Object-Type allocation from the "PCEP Objects" sub-registry:

Object-Class Value	33
Name	External Source Node
Object-Type	1: IPv4 2: IPv6 3-15: Unassigned
Reference	This I-D

7. Acknowledgement

The author would like to thank Cyril Margaria, Ramon Casellas, Dhruv Dhody and Quintin Zhao for their valuable comments and suggestions on this draft.

8. References

8.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, DOI 10.17487/RFC3209, December 2001, <https://www.rfc-editor.org/info/rfc3209>.

[RFC4090] Pan, P., Ed., Swallow, G., Ed., and A. Atlas, Ed., "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, DOI 10.17487/RFC4090, May 2005, <https://www.rfc-editor.org/info/rfc4090>.

[RFC4875] Aggarwal, R., Ed., Papadimitriou, D., Ed., and S. Yasukawa, Ed., "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, DOI 10.17487/RFC4875, May 2007, <https://www.rfc-editor.org/info/rfc4875>.

[RFC5440] Vasseur, JP., Ed. and JL. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, DOI 10.17487/RFC5440, March 2009, <https://www.rfc-editor.org/info/rfc5440>.

[RFC6006] Zhao, Q., Ed., King, D., Ed., Verhaeghe, F., Takeda, T., Ali, Z., and J. Meuric, "Extensions to the Path Computation Element Communication Protocol (PCEP) for Point-to-Multipoint Traffic Engineering Label Switched Paths", RFC 6006, DOI 10.17487/RFC6006, September 2010, <https://www.rfc-editor.org/info/rfc6006>.

8.2. Informative References

[RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, DOI 10.17487/RFC4655, August 2006, <https://www.rfc-editor.org/info/rfc4655>.

[RFC5862] Yasukawa, S. and A. Farrel, "Path Computation Clients (PCC) - Path Computation Element (PCE) Requirements for Point-to-Multipoint MPLS-TE", RFC 5862, DOI 10.17487/RFC5862, June 2010, <https://www.rfc-editor.org/info/rfc5862>.

Author's Address

Huaimo Chen
Futurewei
Boston, MA
USA

Email: Huaimo.chen@futurewei.com

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: May 2, 2012

H. Chen
Huawei Technologies
O. Gonzalez de Dios
Telefonica I+D

October 30, 2011

A Forward-Search P2MP TE LSP Inter-Domain Path Computation
draft-chen-pce-forward-search-p2mp-path-02.txt

Abstract

This document presents a forward search procedure for computing a path for a Point-to-MultiPoint (P2MP) Traffic Engineering (TE) Label Switched Path (LSP) acrossing a number of domains through using multiple Path Computation Elements (PCEs). In addition, extensions to the Path Computation Element Communication Protocol (PCEP) for supporting the forward search procedure are described.

Table of Contents

1. Introduction

RFC 4105 "Requirements for Inter-Area MPLS TE" lists the requirements for computing a shortest path for a TE LSP acrossing multiple IGP areas; and RFC 4216 "MPLS Inter-Autonomous System (AS) TE Requirements" describes the requirements for computing a shortest path for a TE LSP acrossing multiple ASes. RFC 5671 "Applicability of PCE to P2MP MPLS and GMPLS TE" examines the applicability of PCE to path computation for P2MP TE LSPs in MPLS and GMPLS networks.

This document presents a forward search procedure to address these requirements for computing a path for a P2MP TE LSP acrossing domains through using multiple Path Computation Elements (PCEs).

The procedure is called "Forward Search Shortest P2MP LSP Path Crossing Domains". The major characteristics of this procedure for computing a path for a P2MP TE LSP from a source node to a number of destination nodes acrossing multiple domains include the following three ones.

1. It guarantees that the path computed from the source node to the destination nodes is shortest.

2. It does not depend on any domain path tree or domain sequences from the source node to the destination nodes.

3. Navigating a mesh of domains is simple and efficient.

2. Terminology

ABR: Area Border Router. Routers used to connect two IGP areas (areas in OSPF or levels in IS-IS).

ASBR: Autonomous System Border Router. Routers used to connect together ASes of the same or different service providers via one or more inter-AS links.

Boundary Node (BN): a boundary node is either an ABR in the context of inter-area Traffic Engineering or an ASBR in the context of inter-AS Traffic Engineering.

Entry BN of domain(n): a BN connecting domain(n-1) to domain(n) along a determined sequence of domains.

Exit BN of domain(n): a BN connecting domain(n) to domain(n+1) along a determined sequence of domains.

Inter-area TE LSP: A TE LSP that crosses an IGP area boundary.

Inter-AS TE LSP: A TE LSP that crosses an AS boundary.

LSP: Label Switched Path.

LSR: Label Switching Router.

PCC: Path Computation Client. Any client application requesting a path computation to be performed by a Path Computation Element.

PCE: Path Computation Element. An entity (component, application, or network node) that is capable of computing a network path or route based on a network graph and applying computational constraints.

PCE(i) is a PCE with the scope of domain(i).

TED: Traffic Engineering Database.

This document uses terminologies defined in RFC5440.

3. 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.

4. Requirements

This section summarizes the requirements specific for computing a path for a Traffic Engineering (TE) LSP acrossing multiple domains (areas or ASes). More requirements for Inter-Area and Inter-AS MPLS Traffic Engineering are described in RFC 4105 and RFC 4216.

A number of requirements specific for a solution to compute a path for a TE LSP acrossing multiple domains is listed as follows:

1. The solution SHOULD provide the capability to compute a shortest path dynamically, satisfying a set of specified constraints across multiple IGP areas.

2. The solution MUST provide the ability to reoptimize in a minimally disruptive manner (make before break) an inter-area TE LSP, should a more optimal path appear in any traversed IGP area.

3. The solution SHOULD provide mechanism(s) to compute a shortest end-to-end path for a TE LSP acrossing multiple ASes and satisfying a set of specified constraints dynamically.

4. Once an inter-AS TE LSP has been established, and should there be any resource or other changes inside anyone of the ASes, the solution MUST be able to re-optimize the LSP accordingly and non-disruptively, either upon expiration of a configurable timer or upon being triggered by a network event or a manual request at the TE tunnel Head-End.

5. Forward Search P2MP Path Computation

This section gives an overview of the forward search path computation procedure to satisfy the requirements for computing a path for a P2MP TE LSP acrossing multiple domains described above and describes the procedure in details.

5.1. Overview of Procedure

Simply speaking, the idea of the Forward Search P2MP inter-domain path computation method for computing a path for an MPLS TE P2MP LSP crossing multiple domains from a source node to a number of destination nodes includes:

Start from the source node and the source domain.

Consider the optimal path segment from the source node to every exit boundary node of the source domain as a special link;

Consider the optimal path segment from an entry boundary node to every exit boundary node of a domain as a special link; and the optimal path segment is computed as needed.

The whole topology consisting of many domains can be considered as a special topology, which contains those special links, the normal links in the destination domains and the inter-domain links.

Compute a shortest path in this special topology from the source node to the multiple destination nodes using CSPF.

Forward Search P2MP inter-domain path computation method running at any PCE just grows the result path list/tree in the same way as normal CSPF on the special topology. When the result path list/tree reaches all the destination nodes, the shortest path from the source node to the destination nodes is found and a PCRep message with the shortest path is sent to the PCE/PCC that sends the PCReq message

eventually.

5.2. Description of Procedure

Suppose that we have the following variables:

A current PCE named as CurrentPCE which is currently computing the path.

A number of rest destination nodes named as RestDestinationNodes, which is the number of destination nodes to which shortest paths are to be found. RestDestinationNodes is initially the number of all the destination nodes of an MPLS TE P2MP LSP.

A candidate node list named as CandidateNodeList, which contains the nodes through which the shortest path from the source node to a destination node may be. Each node C in CandidateNodeList has the following information:

the cost of the path from the source node to node C,

the previous hop node P and the link between P and C,

the PCE responsible for C, and

the flags for C. The flags include:

bit D indicating that node C is a Destination node if it is set;

bit S indicating that C is the Source node if it is set;

bit E indicating that C is an Exit boundary node if it is set;

bit I indicating that C is an entry boundary node if it is set; and

bit N indicating that C is a Node in a destination domain if it is set.

The nodes in CandidateNodeList are ordered by path cost. Initially, CandidateNodeList contains only a Source Node, with path cost 0, PCE responsible for the source domain, and flags with S bit set.

A result path list or tree named as ResultPathTree, which contains the shortest paths from the source node to the boundary nodes or the nodes in the destination domains. Initially, ResultPathTree is empty.

The Forward Search path computation method for computing a path for

an MPLS TE P2MP LSP crossing a number of domains from a source node to a number of destination nodes can be described as follows:

Initially, a PCC sets RestDestinationNodes to the number of all the destination nodes of the MPLS TE P2MP LSP, ResultPathTree to empty and CandidateNodeList to contain the source node and sends a PCE responsible for the source domain a request with the source node, RestDestinationNodes, CandidateNodeList and ResultPathTree.

When the PCE responsible for a domain (called current domain) receives a request for computing the path for the MPLS TE P2MP LSP, it checks whether the current PCE is the PCE responsible for the node C with the minimum cost in the CandidateNodeList. If it is, then remove C from CandidateNodeList and graft it into ResultPathTree; otherwise, a PCReq message is sent to the PCE for node C.

Suppose that node C has Flags. The ResultPathTree is built from C in the following steps.

If node C is a destination node (i.e., the Destination Node (D) bit in the Flags is set), then RestDestinationNodes is decreased by one. If RestDestinationNodes is zero (i.e., all the destinations are on the result path tree), then the shortest path is found and a PCRep message with the path is sent to the PCE/PCC which sends the request to the current PCE.

If node C is in the destination domain (i.e., the Node in Destination domain (N) bit in the Flags is set), then for every node N connected to node C and not on ResultPathTree, it is merged into CandidateNodeList. The cost to node N is the sum of the cost to node C and the cost of the link between C and N. The PCE for node N is the current PCE.

If node C is an Entry Boundary Node or Source Node (i.e., the Entry/ Incoming Boundary Node (I) bit or the Source Node (S) bit is set), then path segments from node C to every exit boundary node of the current domain that is not on the result path tree are computed through using CSPF and as special links. For every node N connected to node C through a special link (i.e., a path segment), it is merged into CandidateNodeList. The cost to node N is the sum of the cost to node C and the cost of the special link (i.e., path segment) between C and N. The PCE for node N is the current PCE.

If node C is an Exit Boundary Node (i.e., the Exit Boundary Node (E) bit is set) and there exist inter-domain links connected to it, then for every node N connected to C and not on the result path tree, it is merged into the candidate node list. The cost to node N is the sum of the cost to node C and the cost of the link between C and N.

The PCE for node N is the PCE responsible for node N.

If the CurrentPCE is the same as the PCE of the node with the minimum cost in CandidateNodeList, then the node is removed from CandidateNodeList, grafted to ResultPathTree, and the above steps are repeated; otherwise, the CurrentPCE sends the PCE a request with the source node, RestDestinationNodes, CandidateNodeList and ResultPathTree.

5.3. Comparing to BRPC

RFC 5441 describes the Backward Recursive Path Computation (BRPC) algorithm or procedure for computing an MPLS TE P2P LSP path from a source node to a destination node crossing multiple domains. Comparing to BRPC, there are a number of differences between BRPC and the Forward-Search P2MP TE LSP Inter-Domain Path Computation. Some of the differences are briefed below.

At first, BRPC is for computing a shortest path from a source node to a destination node crossing multiple domains. The Forward-Search P2MP TE LSP Inter-Domain Path Computation is for computing a shortest path from a source node to a number of destination nodes crossing multiple domains.

Secondly, for BRPC to compute a shortest path from a source node to a destination node crossing multiple domains, we MUST provide a sequence of domains from the source node to the destination node to BRPC in advance. The Forward-Search P2MP TE LSP Inter-Domain Path Computation does not need any sequence of domains for computing a shortest inter-domain P2MP path.

Moreover, for a given sequence of domains domain(1), domain(2), ... , domain(n), BRPC searches the shortest path from domain(n), to domain(n-1), until domain(1). Thus it is hard for BRPC to be extended for computing a shortest path from a source node to a number of destination nodes crossing multiple domains. The Forward-Search P2MP TE LSP Inter-Domain Path Computation calculates a shortest path in a special topology from the source node to the destination nodes using CSPF.

6. Extensions to PCEP

The extensions to PCEP for Forward Search P2MP Inter-domain Path Computation include the definition of a new flag in the RP object, a result path list/tree and a candidate node list in a request message.

6.1. RP Object Extension

The following flag is added into the RP Object:

The F bit is added in the flag bits field of the RP object to tell the receiver of the message that the request/reply is for Forward Search Path Computation.

- F (Forward search Path Computation bit - 1 bit):

  0: This indicates that this is not PCReq/PCRep for Forward Search Path Computation.

  1: This indicates that this is PCReq or PCRep message for Forward Search Path Computation.

The IANA request is referenced in Section below (Request Parameter Bit Flags) of this document.

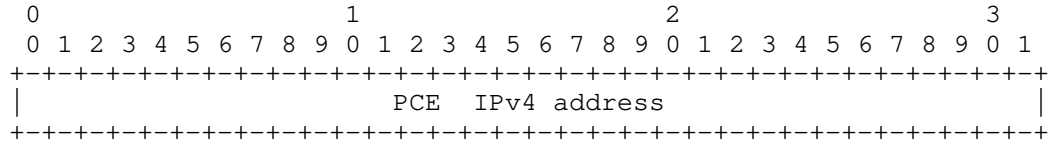
This F bit with the N bit defined in RFC6006 can indicate whether the request/reply is for Forward Search Path Computation of an MPLS TE P2MP LSP or an MPLS TE P2P LSP.

- F = 1 and N = 1: This indicates that this is a PCReq/PCRep message for Forward Search Path Computation of an MPLS TE P2MP LSP.

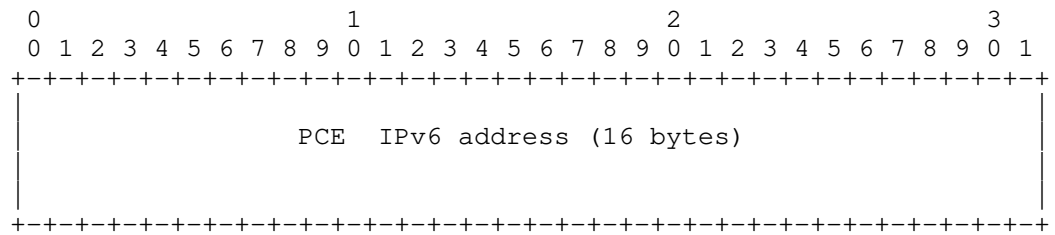
- F = 1 and N = 0: This indicates that this is a PCReq/PCRep message for Forward Search Path Computation of an MPLS TE P2P LSP.

6.2. PCE Object

The figure below illustrates a PCE IPv4 object body (Object-Type=2), which comprises a PCE IPv4 address. The PCE IPv4 address object indicates the IPv4 address of a PCE , with which a PCE session may be established and to which a request message may be sent.

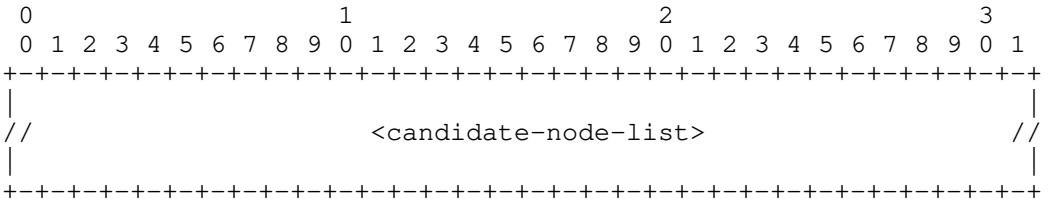


The format of the PCE object body for IPv6 (Object-Type=2) is as follows:



6.3. Candidate Node List Object

The candidate-node-list-obj object contains a list of candidate nodes. A new PCEP object class and type are requested for it. The format of the candidate-node-list-obj object body is as follows:



The following is the definition of the candidate node list.

```
<candidate-node-list>::= <candidate-node>
                         [<candidate-node-list>]
<candidate-node>::= <ERO>
                    <candidate-attribute-list>

<candidate-attribute-list>::= [<attribute-list>]
                              [<PCE>]
                              [<Node-Flags>]
```

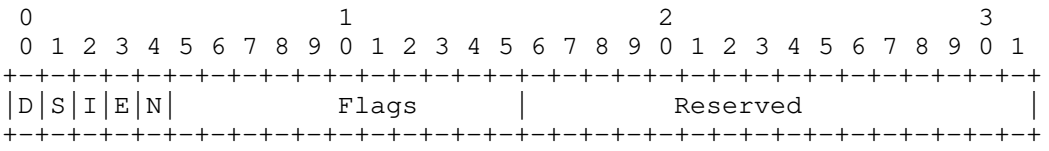
The ERO in a candidate node contain just the path segment of the last link of the path, which is from the previous hop node of the tail end node of the path to the tail end node. With this information, we can graft the candidate node into the existing result path list or tree.

Simply speaking, a candidate node has the same or similar format of a path defined in RFC 5440, but the ERO in the candidate node just contain the tail end node of the path and its previous hop, and the candidate node may contain two new objects PCE and node flags.

6.4. Node Flags Object

The Node Flags object is used to indicate the characteristics of the node in a candidate node list in a request or reply message for Forward Search Inter-domain Path Computation. The Node Flags object comprises a Reserved field, and a number of Flags.

The format of the Node Flags object body is as follows:

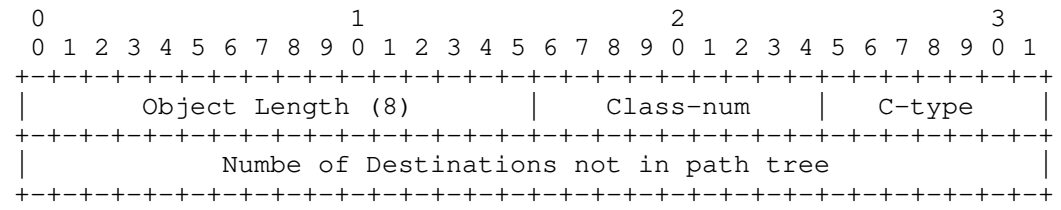


where

- o D = 1: The node is a destination node.
- o S = 1: The node is a source node.
- o I = 1: The node is an entry boundary node.
- o E = 1: The node is an exit boundary node.
- o N = 1: The node is a node in a destination domain.

6.5. Rest Destination Nodes Object

The figure below is an illustration of an object called a number of destinations not in path tree/list , which comprises an Object Length field, a Class-num field, a C-type filed, and a number of destinations not in path. As shown, the value of Object Length field in the object may be 8, which is a length of the object in bytes; the value of Class-num field and the value of C-type field will be assigned by Internet Assigned Numbers Authority (IANA); and the value of the number of destinations not in path tree field comprises a number, which is the number of destinations that are not in the final path computed yet.



6.6. Request Message Extension

Below is the message format for a request message with the extension of a result path list and a candidate node list:

```
<PCReq Message>::= <Common Header>
                   [<svec-list>]
                   <request-list>
<request-list>::=<request>[<request-list>]
<request>::= <RP>
             <END-POINTS>
             [<OF>]
             [<LSPA>]
             [<BANDWIDTH>]
             [<metric-list>]
             [<RRO>[<BANDWIDTH>]]
             [<IRO>]
             [<LOAD-BALANCING>]
             [<result-path-list>]
             [<candidate-node-list-obj>]
             [<rest-destination-nodes>]

where:
<result-path-list>::=<path>[<result-path-list>]
<path>::= <ERO><attribute-list>
<attribute-list>::=[<LSPA>]
                   [<BANDWIDTH>]
                   [<metric-list>]
                   [<IRO>]

<candidate-node-list-obj> contains a <candidate-node-list>

<candidate-node-list>::= <candidate-node>
                         [<candidate-node-list>]
<candidate-node>::= <ERO>
                    <candidate-attribute-list>

<candidate-attribute-list>::= [<attribute-list>]
                              [<PCE>]
                              [<Node-Flags>]
```

Figure 1: The Format for a Request Message

The definition for the result path list that may be added into a request message is the same as that for the path list in a reply message that is described in RFC5440.

7. Security Considerations

The mechanism described in this document does not raise any new

security issues for the PCEP protocols.

8. IANA Considerations

This section specifies requests for IANA allocation.

8.1. Request Parameter Bit Flags

A new RP Object Flag has been defined in this document. IANA is requested to make the following allocation from the "PCEP RP Object Flag Field" Sub-Registry:

Bit	Description	Reference
18	Forward Path Computation (F-bit)	This I-D

9. Acknowledgement

The author would like to thank Julien Meuric, Daniel King, Cyril Margaria, Ramon Casellas, Olivier Dugeon and Dhruv Dhody for their valuable comments on this draft.

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.

[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.

[RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.

[RFC6006] Zhao, Q., King, D., Verhaeghe, F., Takeda, T., Ali, Z., and J. Meuric, "Extensions to the Path Computation Element Communication Protocol (PCEP) for Point-to-Multipoint Traffic Engineering Label Switched Paths", RFC 6006, September 2010.

10.2. Informative References

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

[RFC5862] Yasukawa, S. and A. Farrel, "Path Computation Clients (PCC) - Path Computation Element (PCE) Requirements for Point-to-Multipoint MPLS-TE", RFC 5862, June 2010.

Authors' Addresses

Huaimo Chen
Huawei Technologies
Boston, MA
USA

Email: Huaimochen@huawei.com

Oscar Gonzalez de Dios
Telefonica I+D
Emilio Vargas 6, Madrid
Spain

Email: ogondio@tid.es

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 8, 2020

H. Chen
Futurewei

July 7, 2019

A Forward-Search P2MP TE LSP Inter-Domain Path Computation draft-chen-pce-forward-search-p2mp-path-16

Abstract

This document presents a forward search procedure for computing a path for a Point-to-MultiPoint (P2MP) Traffic Engineering (TE) Label Switched Path (LSP) crossing a number of domains through using multiple Path Computation Elements (PCEs). In addition, extensions to the Path Computation Element Communication Protocol (PCEP) for supporting the forward search procedure are described.

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 https://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 January 8, 2020.

Table of Contents

1. Introduction

RFC 4105 "Requirements for Inter-Area MPLS TE" lists the requirements for computing a shortest path for a TE LSP crossing multiple IGP areas; and RFC 4216 "MPLS Inter-Autonomous System (AS) TE Requirements" describes the requirements for computing a shortest path for a TE LSP crossing multiple ASes. RFC 5671 "Applicability of PCE to P2MP MPLS and GMPLS TE" examines the applicability of PCE to path computation for P2MP TE LSPs in MPLS and GMPLS networks.

This document presents a forward search procedure to address these requirements for computing a path for a P2MP TE LSP crossing domains through using multiple Path Computation Elements (PCEs).

The procedure is called "Forward Search Shortest P2MP LSP Path Crossing Domains" or FSPC for short. The major characteristics of this procedure for computing a path for a P2MP TE LSP from a source

node to a number of destination nodes crossing multiple domains include the following three ones.

1. It guarantees that the path computed from the source node to the destination nodes is shortest.

2. It does not depend on any domain path tree or domain sequences from the source node to the destination nodes.

3. Navigating a mesh of domains is simple and efficient.

2. Terminology

ABR: Area Border Router. Routers used to connect two IGP areas (areas in OSPF or levels in IS-IS).

ASBR: Autonomous System Border Router. Routers used to connect together ASes of the same or different service providers via one or more inter-AS links.

Boundary Node (BN): a boundary node is either an ABR in the context of inter-area Traffic Engineering or an ASBR in the context of inter-AS Traffic Engineering.

Entry BN of domain(n): a BN connecting domain(n-1) to domain(n) along the path found from the source node to the BN, where domain(n-1) is the previous hop domain of domain(n).

Exit BN of domain(n): a BN connecting domain(n) to domain(n+1) along the path found from the source node to the BN, where domain(n+1) is the next hop domain of domain(n).

Inter-area TE LSP: A TE LSP that crosses an IGP area boundary.

Inter-AS TE LSP: A TE LSP that crosses an AS boundary.

LSP: Label Switched Path.

LSR: Label Switching Router.

PCC: Path Computation Client. Any client application requesting a path computation to be performed by a Path Computation Element.

PCE: Path Computation Element. An entity (component, application, or network node) that is capable of computing a network path or route based on a network graph and applying computational constraints.

PCE(i) is a PCE with the scope of domain(i).

TED: Traffic Engineering Database.

This document uses terminologies defined in RFC5440.

3. 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.

4. Requirements

This section summarizes the requirements specific for computing a path for a Traffic Engineering (TE) LSP crossing multiple domains (areas or ASes). More requirements for Inter-Area and Inter-AS MPLS Traffic Engineering are described in RFC 4105 and RFC 4216.

A number of requirements specific for a solution to compute a path for a TE LSP crossing multiple domains is listed as follows:

1. The solution SHOULD provide the capability to compute a shortest path dynamically, satisfying a set of specified constraints across multiple IGP areas.

2. The solution MUST provide the ability to reoptimize in a minimally disruptive manner (make before break) an inter-area TE LSP, should a more optimal path appear in any traversed IGP area.

3. The solution SHOULD provide mechanism(s) to compute a shortest end-to-end path for a TE LSP crossing multiple ASes and satisfying a set of specified constraints dynamically.

4. Once an inter-AS TE LSP has been established, and should there be any resource or other changes inside anyone of the ASes, the solution MUST be able to re-optimize the LSP accordingly and non-disruptively, either upon expiration of a configurable timer or upon being triggered by a network event or a manual request at the TE tunnel Head-End.

5. Forward Search P2MP Path Computation

This section gives an overview of the forward search path computation procedure (FSPC) to satisfy the requirements for computing a path for a P2MP TE LSP crossing multiple domains described above and describes the procedure in details.

5.1. Overview of Procedure

Simply speaking, the idea of FSPC for computing a path for an MPLS TE P2MP LSP crossing multiple domains from a source node to a number of destination nodes includes:

Start from the source node and the source domain.

Consider the optimal path segment from the source node to every exit boundary and destination node of the source domain as a special link;

Consider the optimal path segment from an entry boundary node to every exit boundary node and destination node of a domain as a special link; and the optimal path segment is computed as needed.

The whole topology consisting of many domains can be considered as a special virtual topology, which contains those special links and the inter-domain links.

Compute a shortest path in this special topology from the source node to the multiple destination nodes using CSPF.

FSPC running at any PCE just grows the result path list/tree in the same way as normal CSPF on the special virtual topology. When the result path list/tree reaches all the destination nodes, the shortest path from the source node to the destination nodes is found and a PCRep message with the path is sent to the PCE/PCC that sends the PCReq message eventually.

5.2. Description of Procedure

Suppose that we have the following variables:

A current PCE named as CurrentPCE which is currently computing the path.

A candidate node list named as CandidateNodeList, which contains the nodes to each of which the temporary optimal path from the source node is currently found and satisfies a set of given constraints. Each node C in CandidateNodeList has the following information:

- the cost of the path from the source node to node C,
- the previous hop node P and the link between P and C,
- the PCE responsible for C (i.e., the PCE responsible for the domain containing C. Alternatively, we may use the domain instead of the PCE.), and

- the flags for C.

The flags include:

- bit D indicating that C is a Destination node if it is set;
- bit S indicating that C is the Source node if it is set;
- bit E indicating that C is an Exit boundary node if it is set;
- bit I indicating that C is an entry boundary node if it is set; and
- bit T indicating that C is on result path Tree if it is set.

The nodes in CandidateNodeList are ordered by path cost.

Initially, CandidateNodeList contains a Source Node, with path cost 0, PCE responsible for the source domain, and flags with S bit set. It also contains every destination node, with path cost infinity and flags with D bit set.

A result path list or tree named as ResultPathTree, which contains the shortest paths from the source node to the boundary nodes and destination nodes. Initially, ResultPathTree is empty.

Alternatively, the result path list or tree can be combined into the candidate node list. We may set bit T to one in the node flags for the candidate node when grafting it into the existing result path list or tree. Thus all the candidate nodes with bit T set to one in the candidate list constitute the result path tree or list.

FSPC for computing a path for an MPLS TE P2MP LSP crossing a number of domains from a source node to a number of destination nodes can be described as follows:

Initially, a PCC sends a PCE responsible for the source domain a request with CandidateNodeList and ResultPathTree initialized.

When the PCE responsible for a domain (called current domain) receives a request for computing the path for the MPLS TE P2MP LSP, it checks whether the current PCE is the PCE responsible for the node C with the minimum cost in the CandidateNodeList. If it is, then remove C from CandidateNodeList and graft it into ResultPathTree (i.e., set flag bit T of node C to one); otherwise, a PCReq message is sent to the PCE for node C.

Suppose that node C is in the current domain. The ResultPathTree is built from C in the following steps.

If node C is a destination node (i.e., the Destination Node (D) bit in the Flags is set), then check whether the path cost to node C is infinity. If it is, then we can not find any path for the P2MP LSP, and a repply message with failure reasons is sent; otherwise, if all the destinations are on the result path tree, then the shortest path is found and a PCRep message with the path is sent to the PCE/PCC which sends the request to the current PCE.

If node C is an entry boundary node or the source node, then the optimal path segments from node C to every destination node and every exit boundary node of the current domain that is not on the result path tree and satisfies the given constraints are computed through using CSPF and as special links.

For every node N connected to node C through a special link (i.e., the optimal path segment satisfying the given constraints), it is merged into CandidateNodeList. The cost to node N is the sum of the cost to node C and the cost of the special link (i.e., the path segment) between C and N. If node N is not in the candidate node list, then node N is added into the list with the cost to node N, node C as its previous hop node and a PCE for node N. The PCE for node N is the current PCE if node N is an ASBR; otherwise (node N is an ABR, an exit boundary node of the current domain and an entry boundary node of the domain next to the current domain) the PCE for node N is the PCE for the next domain. If node N is in the candidate node list and the cost to node N through node C is less than the cost to node N in the list, then replace the cost to node N in the list with the cost to node N through node C and the previous hop to node N in the list with node C.

If node C is an exit boundary node and there are inter-domain links connecting to it (i.e., node C is an ASBR) and satisfying the constraints, then for every node N connecting to C, satisfying the constraints and not on the result path tree, it is merged into the candidate node list. The cost to node N is the sum of the cost to node C and the cost of the link between C and N. If node N is not in the candidate node list, then node N is added into the list with the cost to node N, node C as its previous hop node and the PCE for node N. If node N is in the candidate node list and the cost to node N through node C is less than the cost to node N in the list, then replace the cost to node N in the list with the cost to node N through node C and the previous hop to node N in the list with node C.

If the CurrentPCE is the same as the PCE for the node D with the minimum cost in CandidateNodeList, then the node D is removed from CandidateNodeList and grafted to ResultPathTree (i.e., set flag bit T of node D to one), and the above steps are repeated; otherwise, a request message is to be sent to the PCE for node D.

5.3. Processing Request and Reply Messages

In this section, we describe the processing of the request and reply messages with Forward search bit set for FSPC. Each of the request and reply messages mentioned below has its Forward search bit set even though we do not indicate this explicitly.

In the case that a reply message is a final reply, which contains the optimal path from the source to the destination, the reply message is sent toward the PCC along the path that the request message goes from the PCC to the current PCE in reverse direction.

In the case that a request message is to be sent to the PCE for node D with the minimum cost in the candidate node list and there is a PCE session between the current domain and the next domain containing node D, the current PCE sends the PCE for node D through the session a request message with the source node, the destination node, CandidateNodeList and ResultPathTree.

In the case that a request message is to be sent to the PCE for node D and there is not any PCE session between the current PCE and the PCE for node D, a reply message is sent toward a branch point on the result path tree from the current domain along the path that the request message goes from the PCC to the current PCE in reverse direction. From the branch point, there is a downward path to the domain containing the previous hop node of node D on the result path tree and to the domain containing node D. At this branch point, the request message is sent to the PCE for node D along the downward path.

Suppose that node D has the minimum cost in CandidateNodeList when a PCE receives a request message or a reply message containing CandidateNodeList.

When a PCE (current PCE) for a domain (current domain) receives a reply message PCRep, it checks whether the reply is a final reply with the optimal path from the source to the destination. If the reply is the final reply, the current PCE sends the reply to the PCE that sends the request to the current PCE; otherwise, it checks whether there is a path from the current domain to the domain containing the previous hop node of node D on ResultPathTree and to the domain containing node D. If there is a path, the PCE sends a

request PCReq to the PCE responsible for the next domain along the path; otherwise, it sends a reply PCRep to the PCE that sends the request to the current PCE.

When a PCE receives a request PCReq, it checks whether the current domain contains node D. If it does, then node D is removed from CandidateNodeList and grafted to ResultPathTree (i.e., set flag bit T of node D to one), and the above steps in the previous sub section are repeated; otherwise, the PCE sends a request PCReq to the PCE responsible for the next domain along the path from the current domain to the domain containing the previous hop node of node D on ResultPathTree and to the domain containing node D.

6. Comparing to BRPC

RFC 5441 describes the Backward Recursive Path Computation (BRPC) algorithm or procedure for computing an MPLS TE P2P LSP path from a source node to a destination node crossing multiple domains. Comparing to BRPC, there are a number of differences between BRPC and the Forward-Search P2MP TE LSP Inter-Domain Path Computation (FSPC). Some of the differences are briefed below.

At first, BRPC is for computing a shortest path from a source node to a destination node crossing multiple domains. FSPC is for computing a shortest path from a source node to a number of destination nodes crossing multiple domains.

Secondly, for BRPC to compute a shortest path from a source node to a destination node crossing multiple domains, we MUST provide a sequence of domains from the source node to the destination node to BRPC in advance. FSPC does not need any sequence of domains for computing a shortest inter-domain P2MP path.

Moreover, for a given sequence of domains domain(1), domain(2), ... , domain(n), BRPC searches the shortest path from domain(n), to domain(n-1), until domain(1). Thus it is hard for BRPC to be extended for computing a shortest path from a source node to a number of destination nodes crossing multiple domains. FSPC calculates a shortest path in a special topology from the source node to the destination nodes using CSPF.

7. Extensions to PCEP

The extensions to PCEP for FSPC include the definition of a new flag in the RP object, a result path list/tree and a candidate node list in a request message.

7.1. RP Object Extension

The following flag is added into the RP Object:

The F bit is added in the flag bits field of the RP object to tell the receiver of the message that the request/reply is for FSPC.

- F (FSPC bit - 1 bit):
  0: This indicates that this is not PCReq/PCRep for FSPC.
  1: This indicates that this is PCReq or PCRep message for FSPC.

The T bit is added in the flag bits field of the RP object to tell the receiver of the message that the reply is for transferring a request message to the domain containing the node with minimum cost in the candidate list.

- T (Transfer request bit - 1 bit):
  0: This indicates that this is not a PCRep for transferring a request message.
  1: This indicates that this is a PCRep message for transferring a request message.

Setting Transfer request T-bit in a RP Object to one indicates that a reply message containing the RP Object is for transferring a request message to the domain containing the node with minimum cost in the candidate list.

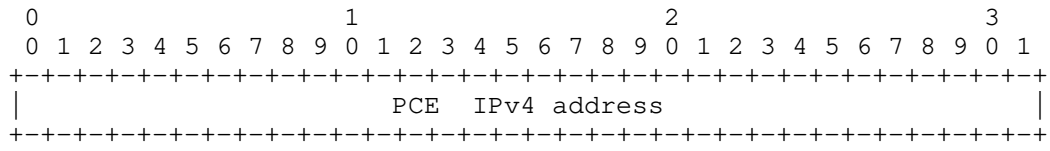
The IANA request is referenced in Section below (Request Parameter Bit Flags) of this document.

This F bit with the N bit defined in RFC6006 can indicate whether the request/reply is for FSPC of an MPLS TE P2MP LSP or an MPLS TE P2P LSP.

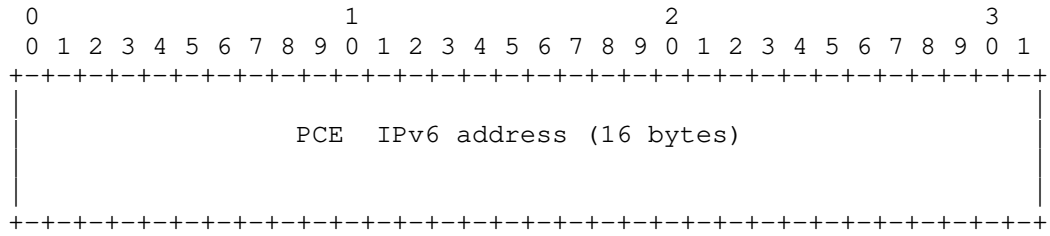
- F = 1 and N = 1: This indicates that this is a PCReq/PCRep message for FSPC of an MPLS TE P2MP LSP.
- F = 1 and N = 0: This indicates that this is a PCReq/PCRep message for FSPC of an MPLS TE P2P LSP.

7.2. PCE Object

The figure below illustrates a PCE IPv4 object body (Object-Type=1), which comprises a PCE IPv4 address. The PCE IPv4 address object indicates the IPv4 address of a PCE , with which a PCE session may be established and to which a request message may be sent.

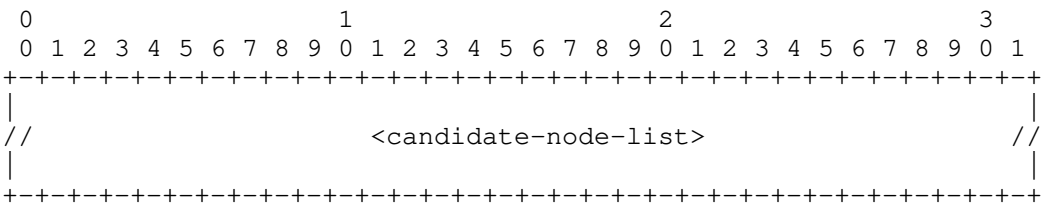


The format of the PCE object body for IPv6 (Object-Type=2) is as follows:



7.3. Candidate Node List Object

The candidate-node-list-obj object contains a list of candidate nodes. A new PCEP object class and type are requested for it. The format of the candidate-node-list-obj object body is as follows:



The following is the definition of the candidate node list.

```
<candidate-node-list>::= <candidate-node>
                         [<candidate-node-list>]
<candidate-node>::= <ERO>
                    <candidate-attribute-list>

<candidate-attribute-list>::= [<attribute-list>]
                              [<PCE>]
                              [<Node-Flags>]
```

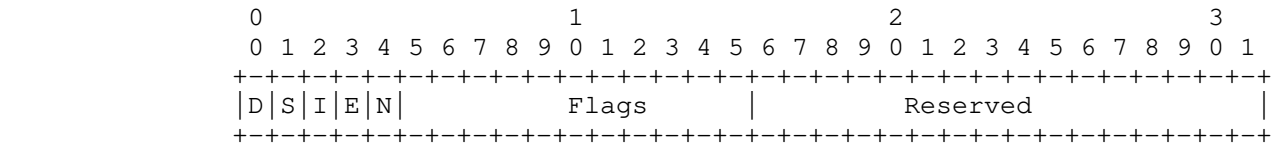
The ERO in a candidate node contain just the path segment of the last link of the path, which is from the previous hop node of the tail end node of the path to the tail end node. With this information, we can graft the candidate node into the existing result path list or tree.

Simply speaking, a candidate node has the same or similar format of a path defined in RFC 5440, but the ERO in the candidate node just contain the tail end node of the path and its previous hop, and the candidate node may contain two new objects PCE and node flags.

7.4. Node Flags Object

The Node Flags object is used to indicate the characteristics of the node in a candidate node list in a request or reply message for FSPC. The Node Flags object comprises a Reserved field, and a number of Flags.

The format of the Node Flags object body is as follows:



where

- o D = 1: The node is a destination node.
- o S = 1: The node is a source node.
- o I = 1: The node is an entry boundary node.
- o E = 1: The node is an exit boundary node.
- o T = 1: The node is on the result path tree.

7.5. Request Message Extension

Below is the message format for a request message with the extension of a result path list and a candidate node list:

```
<PCReq Message>::= <Common Header>
                   <request>
<request>::= <RP> <END-POINT-RRO-PAIR-LIST>  [<OF>]  [<LSPA>]
             [<BANDWIDTH>]  [<metric-list>]  [<RRO>[<BANDWIDTH>]]
             [<IRO>] [<LOAD-BALANCING>]
             [<result-path-list>]
             [<candidate-node-list-obj>]
where:
<result-path-list>::=<path>[<result-path-list>]
<path>::= <ERO><attribute-list>
<attribute-list>::=[<LSPA>]  [<BANDWIDTH>]  [<metric-list>]
                   [<IRO>]

<candidate-node-list-obj> contains a <candidate-node-list>

<candidate-node-list>::= <candidate-node>
                         [<candidate-node-list>]
<candidate-node>::= <ERO>
                    <candidate-attribute-list>

<candidate-attribute-list>::= [<attribute-list>]
                              [<PCE>]
                              [<Node-Flags>]
```

Figure 1: The Format for a Request Message

The definition for the result path list that may be added into a request message is the same as that for the path list in a reply message that is described in RFC5440.

7.6. Reply Message Extension

Below is the message format for a reply message with the extension of a result path list and a candidate node list:

```
<PCRep Message> ::= <Common Header>
                    <response>
<response> ::= <RP>  [<END-POINT-PATH-PAIR-LIST>]
               [<NO-PATH>] [<attribute-list>]
               [<result-path-list>]
               [<candidate-node-list-obj>]
where:
<candidate-node-list-obj> contains a <candidate-node-list>
```

If the path from the source to the destinations is not found yet and there are still chances to find a path (i.e., the candidate list is

not empty), the reply message contains candidate-node-list-obj consisting of the information of the candidate list, which is encoded. In this case, the Transfer request T-bit in the RP Object is set to one.

If the path from the source to the destination is found, the reply message contains path-list comprising the information of the path.

8. Security Considerations

The mechanism described in this document does not raise any new security issues for the PCEP protocols.

9. IANA Considerations

This section specifies requests for IANA allocation.

9.1. Request Parameter Bit Flags

A new RP Object Flag has been defined in this document. IANA is requested to make the following allocation from the "PCEP RP Object Flag Field" Sub-Registry:

Bit	Description		Reference
18	FSPC (F-bit)		This I-D
19	Transfer Request	(T-bit)	This I-D

10. Acknowledgement

The author would like to appreciate Dhruv Dhody for his great contributions and to thank Julien Meuric, Daniel King, Cyril Margaria, Ramon Casellas, Olivier Dugeon and Oscar Gonzalez de Dios for their valuable comments on this draft.

11. References

11.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, DOI 10.17487/RFC3209, December 2001, <https://www.rfc-editor.org/info/rfc3209>.

[RFC5440] Vasseur, JP., Ed. and JL. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, DOI 10.17487/RFC5440, March 2009, <https://www.rfc-editor.org/info/rfc5440>.

[RFC6006] Zhao, Q., Ed., King, D., Ed., Verhaeghe, F., Takeda, T., Ali, Z., and J. Meuric, "Extensions to the Path Computation Element Communication Protocol (PCEP) for Point-to-Multipoint Traffic Engineering Label Switched Paths", RFC 6006, DOI 10.17487/RFC6006, September 2010, <https://www.rfc-editor.org/info/rfc6006>.

11.2. Informative References

[RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, DOI 10.17487/RFC4655, August 2006, <https://www.rfc-editor.org/info/rfc4655>.

[RFC5862] Yasukawa, S. and A. Farrel, "Path Computation Clients (PCC) - Path Computation Element (PCE) Requirements for Point-to-Multipoint MPLS-TE", RFC 5862, DOI 10.17487/RFC5862, June 2010, <https://www.rfc-editor.org/info/rfc5862>.

Author's Address

Huaimo Chen
Futurewei
Boston, MA
USA

EMail: Huaimo.chen@futurewei.com

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: May 2, 2012

H. Chen
Huawei Technologies
O. Gonzalez de Dios
Telefonica I+D

October 30, 2011

A Forward-Search P2P TE LSP Inter-Domain Path Computation
draft-chen-pce-forward-search-p2p-path-computation-02.txt

Abstract

This document presents a forward search procedure for computing paths for Point-to-Point (P2P) Traffic Engineering (TE) Label Switched Paths (LSPs) crossing a number of domains through using multiple Path Computation Elements (PCEs). In addition, extensions to the Path Computation Element Communication Protocol (PCEP) for supporting the forward search procedure are described.

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 May 2, 2012.

Copyright Notice

Table of Contents

1. Introduction

It would be useful to extend MPLS TE capabilities across multiple domains (i.e., IGP areas or Autonomous Systems) to support inter-domain resources optimization, to provide strict QoS guarantees between two edge routers located within distinct domains.

RFC 4105 "Requirements for Inter-Area MPLS TE" lists the requirements for computing a shortest path for a TE LSP acrossing multiple IGP areas; and RFC 4216 "MPLS Inter-Autonomous System (AS) TE Requirements" describes the requirements for computing a shortest path for a TE LSP acrossing multiple ASes. RFC 4655 "A PCE-Based Architecture" discusses centralized and distributed computation models for the computation of a path for a TE LSP acrossing multiple domains.

This document presents a forward search procedure to address these requirements through using multiple Path Computation Elements (PCEs). This procedure guarantees that the path found from the source to the destination is shortest. It does not depend on any sequence of domains from the source node to the destination node. Navigating a mesh of domains is simple and efficient.

2. Terminology

ABR: Area Border Router. Routers used to connect two IGP areas (areas in OSPF or levels in IS-IS).

ASBR: Autonomous System Border Router. Routers used to connect together ASes of the same or different service providers via one or more inter-AS links.

Boundary Node (BN): a boundary node is either an ABR in the context of inter-area Traffic Engineering or an ASBR in the context of inter-AS Traffic Engineering.

Entry BN of domain(n): a BN connecting domain(n-1) to domain(n) along a determined sequence of domains.

Exit BN of domain(n): a BN connecting domain(n) to domain(n+1) along a determined sequence of domains.

Inter-area TE LSP: A TE LSP that crosses an IGP area boundary.

Inter-AS TE LSP: A TE LSP that crosses an AS boundary.

LSP: Label Switched Path.

LSR: Label Switching Router.

PCC: Path Computation Client. Any client application requesting a path computation to be performed by a Path Computation Element.

PCE: Path Computation Element. An entity (component, application, or network node) that is capable of computing a network path or route based on a network graph and applying computational constraints.

PCE(i) is a PCE with the scope of domain(i).

TED: Traffic Engineering Database.

This document uses terminologies defined in RFC 5440.

3. 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.

4. Requirements

This section summarizes the requirements specific for computing a path for a P2P Traffic Engineering (TE) LSP acrossing multiple domains (areas or ASes). More requirements for Inter-Area and Inter-AS MPLS Traffic Engineering are described in RFC 4105 and RFC 4216.

A number of requirements specific for a solution to compute a path for a P2P TE LSP acrossing multiple domains is listed as follows:

1. The solution SHOULD provide the capability to compute a shortest path dynamically, satisfying a set of specified constraints across multiple IGP areas.

2. The solution MUST provide the ability to reoptimize in a minimally disruptive manner (make before break) an inter-area TE LSP, should a more optimal path appear in any traversed IGP area.

3. The solution SHOULD provide mechanism(s) to compute a shortest end-to-end path for a TE LSP acrossing multiple ASes and satisfying a set of specified constraints dynamically.

4. Once an inter-AS TE LSP has been established, and should there be any resource or other changes inside anyone of the ASes, the

solution MUST be able to re-optimize the LSP accordingly and non-disruptively, either upon expiration of a configurable timer or upon being triggered by a network event or a manual request at the TE tunnel Head-End.

5. Forward Search Path Computation

This section gives an overview of the forward search path computation procedure to satisfy the requirements described above and describes the procedure in details.

5.1. Overview of Procedure

Simply speaking, the idea of the forward search path computation procedure for computing a path for an MPLS TE P2P LSP acrossing multiple domains from a source node to a destination node includes:

Start from the source node and the source domain.

Consider the optimal path segment from the source node to every exit boundary node of the source domain as a special link;

Consider the optimal path segment from an entry boundary node to every exit boundary node of a domain as a special link; and the optimal path segment is computed as needed.

The whole topology consisting of many domains can be considered as a special topology, which contains those special links, the normal links in the destination domain and the inter-domain links.

Compute an optimal path in this special topology from the source node to the destination node using CSPF.

The forward search path computation procedure for computing a path for an MPLS TE P2P LSP starts at the source domain, in which the source (or ingress) node of the MPLS TE LSP locates. When a PCE in the source domain receives a PCReq for the path for the MPLS TE LSP, it computes the optimal path from the source node to every exit boundary node of the domain towards the destination node.

There are two lists involved in the path computation. One list is called candidate node list, which contains the nodes with brief information about the temporary optimal paths from the source node to each of these nodes currently found. The nodes in the candidate list are ordered by the cost of the path. Initially, the candidate node list contains only source node with cost 0.

The other is called result path list or tree, which contains the final optimal paths from the source node to the boundary nodes or the nodes in the destination domain. Initially, the result path list is empty.

When a PCE responsible for a domain (called current domain) receives a PCReq for computing the path for the MPLS TE LSP, it removes the node with the minimum cost from the candidate node list and put or graft the node to the result path list or tree.

If the destination node is in the current domain, the PCE tries to compute the optimal path from the source node to the destination node and sends a PCRep with the optimal path to the PCE or PCC from which the PCReq is received.

Otherwise (i.e., if the destination is not in the domain), the PCE computes the optimal path from the source node to every exit boundary node of the current domain towards the destination node and further to the entry boundary nodes of the domain connected to the current domain, puts the new node into the candidate list in order by path cost, updates the existing node in the candidate node list with the new node with lower cost, and then sends a PCReq with the new candidate node list to the PCE that is responsible for the domain with the first node in the candidate node list.

5.2. Description of Procedure

Suppose that we have the following variables:

A current PCE named as CurrentPCE which is currently computing the path.

A candidate node list named as CandidateNodeList, which contains the nodes to each of which the temporary optimal path from the source node is currently found. The information about each node C in CandidateNodeList consists of

the cost of the path from the source node to node C,

the previous hop node P and the link between P and C,

the PCE responsible for C, and

the flags for C. The flags include

one bit D indicating that node C is a Destination node if it is set;

one bit S indicating that C is the Source node if it is set;

one bit E indicating that C is an Exit boundary node if it is set;

one bit I indicating that C is an entry boundary node if it is set; and

one bit N indicating that C is a Node in the destination domain if it is set.

The nodes in CandidateNodeList are ordered by path cost. Initially, CandidateNodeList contains only a Source Node, with path cost 0, PCE responsible for the source domain, and flags with S bit set.

A result path list or tree named as ResultPathTree, which contains the final optimal paths from the source node to the boundary nodes or the nodes in the destination domain. Initially, ResultPathTree is empty.

The Forward Search Path Computation procedure for computing the path for the MPLS TE P2P LSP is described as follows:

Initially, a PCC sets ResultPathTree to empty and CandidateNodeList to contain the source node and sends PCE responsible for the source domain a PCReq with the source node, the destination node, CandidateNodeList and ResultPathTree.

When the PCE responsible for a domain (called current domain) receives a request for computing the path for the MPLS TE P2MP LSP, it checks whether the current PCE is the PCE responsible for the node C with the minimum cost in the CandidateNodeList. If it is, then remove C from CandidateNodeList and graft it into ResultPathTree; otherwise, a PCReq message is sent to the PCE for node C.

Suppose that node C has Flags. The ResultPathTree is built from C in the following steps.

If the D (Destination Node) bit in the Flags is set, then the optimal path from the source node to the destination node is found, and a PCRep message with the path is sent to the PCE/PCC which sends the request to the current PCE.

If the N (Node in Destination domain) bit in the Flags is set, then for every node N connected to node C and not on ResultPathTree, it is merged into CandidateNodeList. The cost to node N is the sum of the cost to node C and the cost of the link between C and N. The PCE for node N is the current PCE.

If the Entry/Incoming Boundary Node (I) bit or the Source Node (S) bit is set), then path segments from node C to every exit boundary

node of the current domain that is not on the result path tree are computed through using CSPF and as special links. For every node N connected to node C through a special link (i.e., a path segment), it is merged into CandidateNodeList. The cost to node N is the sum of the cost to node C and the cost of the special link (i.e., path segment) between C and N. The PCE for node N is the current PCE.

If the Exit Boundary Node (E) bit is set and there exist inter-domain links connected to it, then for every node N connected to C and not on the result path tree, it is merged into the candidate node list. The cost to node N is the sum of the cost to node C and the cost of the link between C and N. The PCE for node N is the PCE responsible for node N.

If the CurrentPCE is the same as the PCE of the node with the minimum cost in CandidateNodeList, then the node is removed from CandidateNodeList, grafted to ResultPathTree, and the above steps are repeated; otherwise, the CurrentPCE sends the PCE a request with the source node, CandidateNodeList and ResultPathTree.

5.3. Comparing to BRPC

RFC 5441 describes the Backward Recursive Path Computation (BRPC) algorithm or procedure for computing an MPLS TE P2P LSP path from a source node to a destination node crossing multiple domains. Comparing to BRPC, there are a number of differences between BRPC and the Forward-Search P2P TE LSP Inter-Domain Path Computation. Some of the differences are briefed below.

First, for BRPC to compute a shortest path from a source node to a destination node crossing multiple domains, we MUST provide a sequence of domains from the source node to the destination node to BRPC in advance. The Forward-Search P2P TE LSP Inter-Domain Path Computation does not need any sequence of domains for computing a shortest path.

Secondly, for a given sequence of domains domain(1), domain(2), ... , domain(n), BRPC searches the shortest path from domain(n), to domain(n-1), until domain(1) along the reverse order of the given sequence of domain. It will get the shortest path within the given domain sesuence. The Forward-Search P2P TE LSP Inter-Domain Path Computation calculates an optimal path in a special topology from the source node to the destination node using CSPF. It will find the shortest path within all the domains.

Moreover, if the sequence of domains from the source node to the destination node provided to BRPC does not contain the shortest path from the source to the destionation, then the path computed by BRPC

is not optimal. The Forward-Search P2P TE LSP Inter-Domain Path Computation guarantees that the path found is optimal.

6. Extensions to PCEP

This section describes the extensions to PCEP for Forward Search Path Computation. The extensions include the definition of a new flag in the RP object, a result path list and a candidate node list in the PCReq message.

6.1. RP Object Extension

The following flag is added into the RP Object:

The F bit is added in the flag bits field of the RP object to tell the receiver of the message that the request/reply is for Forward Search Path Computation.

- F (Forward search Path Computation bit - 1 bit):

  0: This indicates that this is not PCReq/PCRep for Forward Search Path Computation.

  1: This indicates that this is PCReq or PCRep message for Forward Search Path Computation.

The IANA request is referenced in Section below (Request Parameter Bit Flags) of this document.

This F bit with the N bit defined in RFC6006 can indicate whether the request/reply is for Forward Search Path Computation of an MPLS TE P2P LSP or an MPLS TE P2MP LSP.

- F = 1 and N = 0: This indicates that this is a PCReq/PCRep message for Forward Search Path Computation of an MPLS TE P2P LSP.

- F = 1 and N = 1: This indicates that this is a PCReq/PCRep message for Forward Search Path Computation of an MPLS TE P2MP LSP.

6.2. PCE Object

The figure below illustrates a PCE IPv4 object body (Object-Type=2), which comprises a PCE IPv4 address. The PCE IPv4 address object indicates the IPv4 address of a PCE , with which a PCE session may be established and to which a request message may be sent.

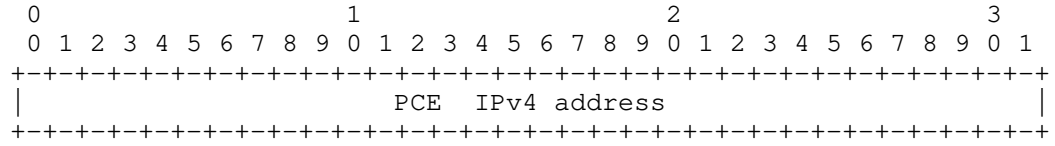


Figure 1: PCE Object Body for IPv4

The format of the PCE object body for IPv6 (Object-Type=2) is as follows:

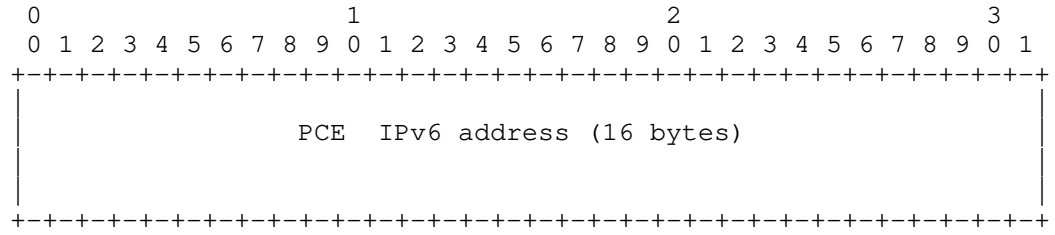


Figure 2: PCE Object Body for IPv6

6.3. Node Flags Object

The Node Flags object is used to indicate the characteristics of the node in a candidate node list in a request or reply message for Forward Search Inter-domain Path Computation. The Node Flags object comprises a Reserved field, and a number of Flags.

The format of the Node Flags object body is as follows:

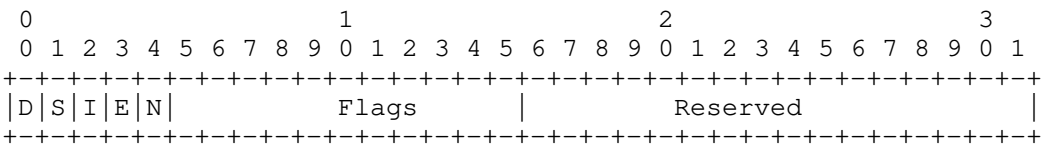


Figure 3: Node Flags Object Body

where

- o D = 1: The node is a destination node.
- o S = 1: The node is a source node.
- o I = 1: The node is an entry boundary node.
- o E = 1: The node is an exit boundary node.
- o N = 1: The node is a node in a destination domain.

6.4. Candidate Node List Object

The candidate-node-list-obj object contains the nodes in the candidate node list. A new PCEP object class and type are requested for it. The format of the candidate-node-list-obj object body is as follows:

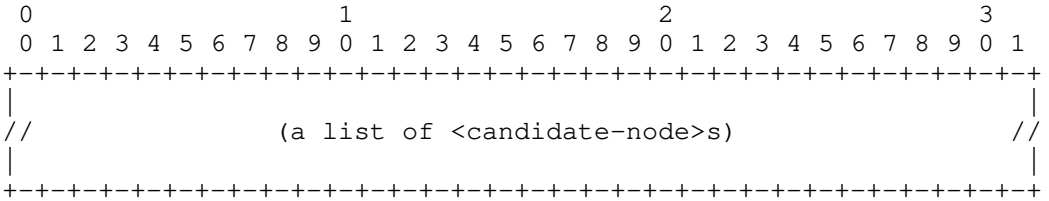


Figure 4: Candidate Node List Object

The following is the definition of candidate node list, which may contain Node Flags.

```
<candidate-node-list>::= <candidate-node>
                         [<candidate-node-list>]
<candidate-node>::= <ERO>
                    <candidate-attribute-list>

<candidate-attribute-list>::= [<attribute-list>]
                              [<PCE>]
                              [<Node-Flags>]
```

The ERO in a candidate node contain just the path segment of the last link of the path, which is from the previous hop node of the tail end node of the path to the tail end node. With this information, we can graft the candidate node into the existing result path list or tree.

Simply speaking, a candidate node has the same or similar format of a path defined in RFC 5440, but the ERO in the candidate node just

contain the tail end node of the path and its previous hop, and the candidate path may contain two new objects PCE and node flags.

6.5. Request Message Extension

Below is the message format for a request message with the extension of a result path list and a candidate node list:

```
        <PCReq Message>::= <Common Header>
                           [<svec-list>]
                           <request-list>
        <request-list>::=<request>[<request-list>]
        <request>::= <RP>
                     <END-POINTS>
                     [<OF>]
                     [<LSPA>]
                     [<BANDWIDTH>]
                     [<metric-list>]
                     [<RRO>[<BANDWIDTH>]]
                     [<IRO>]
                     [<LOAD-BALANCING>]
                     [<result-path-list>]
                     [<candidate-node-list-obj>]

     where:
        <result-path-list>::=<path>[<result-path-list>]
        <path>::= <ERO><attribute-list>
        <attribute-list>::=[<LSPA>]
                           [<BANDWIDTH>]
                           [<metric-list>]
                           [<IRO>]

        <candidate-node-list-obj> contains a <candidate-node-list>
```

The definition for the result path list that may be added into a request message is the same as that for the path list in a reply message that is described in RFC5440.

7. Security Considerations

The mechanism described in this document does not raise any new security issues for the PCEP protocols.

8. IANA Considerations

This section specifies requests for IANA allocation.

8.1. Request Parameter Bit Flags

A new RP Object Flag has been defined in this document. IANA is requested to make the following allocation from the "PCEP RP Object Flag Field" Sub-Registry:

Bit	Description	Reference
18	Forward Path Computation (F-bit)	This I-D

9. Acknowledgement

The authors would like to thank Julien Meuric, Daniel King, Cyril Margaria, Ramon Casellas, Olivier Dugeon and Dhruv Dhody for their valuable comments on this draft.

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.

[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.

[RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.

[RFC6006] Zhao, Q., King, D., Verhaeghe, F., Takeda, T., Ali, Z., and J. Meuric, "Extensions to the Path Computation Element Communication Protocol (PCEP) for Point-to-Multipoint Traffic Engineering Label Switched Paths", RFC 6006, September 2010.

10.2. Informative References

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

[RFC5862] Yasukawa, S. and A. Farrel, "Path Computation Clients (PCC) - Path Computation Element (PCE) Requirements for Point-to-Multipoint MPLS-TE", RFC 5862, June 2010.

Authors' Addresses

Huaimo Chen
Huawei Technologies
Boston, MA
USA

Email: Huaimochen@huawei.com

Oscar Gonzalez de Dios
Telefonica I+D
Emilio Vargas 6, Madrid
Spain

Email: ogondio@tid.es

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 8, 2020

H. Chen
Futurewei

July 7, 2019

A Forward-Search P2P TE LSP Inter-Domain Path Computation draft-chen-pce-forward-search-p2p-path-computation-17

Abstract

This document presents a forward search procedure for computing paths for Point-to-Point (P2P) Traffic Engineering (TE) Label Switched Paths (LSPs) crossing a number of domains using multiple Path Computation Elements (PCEs). In addition, extensions to the Path Computation Element Communication Protocol (PCEP) for supporting the forward search procedure are described.

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 https://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 January 8, 2020.

Copyright Notice

Table of Contents

1. Introduction

It would be useful to extend MPLS TE capabilities across multiple domains (i.e., IGP areas or Autonomous Systems) to support inter-domain resources optimization, to provide strict QoS guarantees between two edge routers located within distinct domains.

[RFC4105] "Requirements for Inter-Area MPLS TE" lists the requirements for computing a shortest path for a TE LSP crossing multiple IGP areas; and [RFC4216] "MPLS Inter-Autonomous System (AS) TE Requirements" describes the requirements for computing a shortest path for a TE LSP crossing multiple ASes. [RFC4655] "A PCE-Based Architecture" discusses centralized and distributed computation models for the computation of a path for a TE LSP crossing multiple domains.

This document presents a forward search procedure to address these requirements using multiple Path Computation Elements (PCEs). This procedure guarantees that the path found from the source to the destination is shortest. It does not depend on any sequence of domains from the source node to the destination node. Navigating a mesh of domains is simple and efficient.

2. Terminology

The following terminology is used in this document.

ABR: Area Border Router. Router used to connect two IGP areas (Areas in OSPF or levels in IS-IS).

ASBR: Autonomous System Border Router. Router used to connect together ASes of the same or different service providers via one or more inter-AS links.

BN: Boundary Node. A boundary node is either an ABR in the context of inter-area Traffic Engineering or an ASBR in the context of inter-AS Traffic Engineering.

Entry BN of domain(n): a BN connecting domain(n-1) to domain(n) along the path found from the source node to the BN, where domain(n-1) is the previous hop domain of domain(n).

Exit BN of domain(n): a BN connecting domain(n) to domain(n+1) along the path found from the source node to the BN, where domain(n+1) is the next hop domain of domain(n).

Inter-area TE LSP: a TE LSP that crosses an IGP area boundary.

Inter-AS TE LSP: a TE LSP that crosses an AS boundary.

LSP: Label Switched Path

LSR: Label Switching Router

PCC: Path Computation Client. Any client application requesting a path computation to be performed by a Path Computation Element.

PCE: Path Computation Element. An entity (component, application, or network node) that is capable of computing a network path or route based on a network graph and applying computational constraints.

PCE(i): a PCE with the scope of domain(i).

TED: Traffic Engineering Database.

This document uses terminology defined in [RFC5440].

3. 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].

4. Requirements

This section summarizes the requirements specific for computing a path for a P2P Traffic Engineering (TE) LSP crossing multiple domains (areas or ASes). More requirements for Inter-Area and Inter-AS MPLS Traffic Engineering are described in [RFC4105] and [RFC4216].

A number of requirements specific for a solution to compute a path for a P2P TE LSP crossing multiple domains is listed as follows:

1. The solution SHOULD provide the capability to compute a shortest path dynamically, satisfying a set of specified constraints across multiple IGP areas.

2. The solution MUST provide the ability to reoptimize in a minimally disruptive manner (make before break) an inter-area TE LSP, should a more optimal path appear in any traversed IGP area.

3. The solution SHOULD provide mechanism(s) to compute a shortest end-to-end path for a TE LSP crossing multiple ASes and satisfying a set of specified constraints dynamically.

4. Once an inter-AS TE LSP has been established, and should there be any resource or other changes inside anyone of the ASes, the solution MUST be able to re-optimize the LSP accordingly and non-disruptively, either upon expiration of a configurable timer or upon being triggered by a network event or a manual request at the TE tunnel Head-End.

5. Forward Search Path Computation

This section gives an overview of the forward search path computation procedure (FSPC for short) to satisfy the requirements described above and describes the procedure in detail.

5.1. Overview of Procedure

Simply speaking, the idea of FSPC for computing a path for an MPLS TE P2P LSP crossing multiple domains from a source node to a destination node includes:

Start from the source node and the source domain.

Consider the optimal path segment from the source node to every exit boundary node of the source domain as a special link;

Consider the optimal path segment from an entry boundary node to every exit boundary node and the destination node of a domain as a special link; and the optimal path segment is computed as needed.

The whole topology consisting of many domains can be considered as a special topology, which contains those special links and the inter-domain links.

Compute an optimal path in this special topology from the source node to the destination node using CSPF.

5.2. Description of Procedure

Suppose that we have the following variables:

A current PCE named as CurrentPCE which is currently computing the path.

A candidate node list named as CandidateNodeList, which contains the nodes to each of which the temporary optimal path from the source node is currently found and satisfies a set of given constraints. The information about each node C in CandidateNodeList consists of:

o the cost of the path from the source node to node C,

- the hopcount of the path from the source node to node C,
- the previous hop node P and the link between P and C,
- the domain list of C (ABR or ASBR) where C has visibility to multiple domains and clearly mark the domain by which C is added to CandidateNodeList,
- the PCE responsible for C (i.e., the PCE responsible for the domain containing C. Alternatively, we may use the above mentioned domain instead of the PCE.), and
- the flags for C.

The flags include:

- bit D indicating that C is a Destination node if it is set,
- bit S indicating that C is the Source node if it is set,
- bit T indicating that C is on result path Tree if it is set.

The nodes in CandidateNodeList are ordered by path cost. Initially, CandidateNodeList contains only a Source Node, with path cost 0, PCE responsible for the source domain.

A result path list or tree named as ResultPathTree, which contains the final optimal paths from the source node to the boundary nodes or the destination node. Initially, ResultPathTree is empty.

Alternatively, the result path list or tree can be combined into the CandidateNodeList. We may set bit T to one in the NODE-FLAGS object for the candidate node when grafting it into the existing result path list or tree. Thus all the candidate nodes with bit T set to one in the CandidateNodeList constitute the result path tree or list.

FSPC for computing the path for the MPLS TE P2P LSP is described as follows:

Initially, a PCC sets ResultPathTree to empty and CandidateNodeList to contain the source node and sends PCE responsible for the source domain a PCReq with the source node, the destination node, CandidateNodeList and ResultPathTree.

When the PCE responsible for a domain (called current domain) receives a request for computing the path for the MPLS TE P2P LSP, it obtains node Cm with the minimum path cost in the CandidateNodeList. The node Cm is the first node in the CandidateNodeList, which is

sorted by path cost. It checks whether the CurrentPCE is the PCE responsible for the node Cm(always expand node Cm only if it is an entry boundary node). If it is, then remove Cm from CandidateNodeList and graft it into ResultPathTree (i.e., set flag bit T of node Cm to one); otherwise, a PCReq message is sent to the PCE for node Cm (see Section 5.3 for the case that there is not any direct session between the CurrentPCE and the PCE for node Cm).

Suppose that node Cm is in the current domain. The ResultPathTree is built from Cm in the following steps.

If node Cm is the destination node, then the optimal path from the source node to the destination node is found, and a PCRep message with the path is sent to the PCE/PCC which sends the request to the CurrentPCE.

If node Cm is an entry boundary node or the source node, then the optimal path segments from node Cm to the destination node (if it is in the current domain) and every exit boundary node of the current domain that is not on the result path tree and satisfies the given constraints are computed through using CSPF and as special links.

For every node N connected to node Cm through a special link (i.e., the optimal path segment satisfying the given constraints), it is merged into CandidateNodeList. The cost to node N is the sum of the cost to node Cm and the cost of the special link (i.e., the path segment) between Cm and N. If node N is not in the CandidateNodeList, then node N is added into the list with the cost to node N, node Cm as its previous hop node and the PCE for node N. The PCE for node N is the CurrentPCE if node N is an ASBR; otherwise (node N is an ABR, an exit boundary node of the current domain and an entry boundary node of the domain next to the current domain) the PCE for node N is the PCE for the next domain. If node N is in the CandidateNodeList and the cost to node N through node Cm is less than the cost to node N in the list, then replace the cost to node N in the list with the cost to node N through node Cm and the previous hop to node N in the list with node Cm.

If node Cm is an exit boundary node and there are inter-domain links connecting to it (i.e., node Cm is an ASBR) and satisfying the constraints, then for every node N connecting to Cm, satisfying the constraints and not on the result path tree, it is merged into the CandidateNodeList. The cost to node N is the sum of the cost to node Cm and the cost of the link between Cm and N. If node N is not in the CandidateNodeList, then node N is added into the list with the cost to node N, node Cm as its previous hop node and the PCE for node N. If node N is in the CandidateNodeList and the cost to node N through node Cm is less than the cost to node N in the list, then

replace the cost to node N in the list with the cost to node N through node Cm and the previous hop to node N in the list with node Cm.

After the CandidateNodeList is updated, there will be a new node Cm with the minimum cost in the updated CandidateNodeList. If the CurrentPCE is the same as the PCE for the new node Cm, then the node Cm is removed from the CandidateNodeList and grafted to ResultPathTree (i.e., set flag bit T of node Cm to one), and the above steps are repeated; otherwise, a request message is to be sent to the PCE for node Cm.

Note that if node Cm has visibility to multiple domains, do not remove it from the CandidateNodeList until it is expanded in all domains. Also mark in the domain list of node Cm, for which domains it is already expanded.

5.3. Processing Request and Reply Messages

In this section, we describe the processing of the request and reply messages with Forward search bit set for FSPC. Each of the request and reply messages mentioned below has its Forward search bit set even though we do not indicate this explicitly.

In the case that a reply message is a final reply, which contains the optimal path from the source to the destination, the reply message is sent toward the PCC along the path that the request message goes from the PCC to the current PCE in reverse direction.

In the case that a request message is to be sent to the PCE for node Cm with the minimum cost in the CandidateNodeList and there is a PCE session between the current domain and the next domain containing node Cm, the CurrentPCE sends the PCE for node Cm through the session a request message with the source node, the destination node, CandidateNodeList and ResultPathTree.

In the case that a request message is to be sent to the PCE for node Cm and there is not any PCE session between the CurrentPCE and the PCE for node Cm, a request message with T bit set to one in RP is sent toward a branch point on the result path tree from the current domain along the path that the request message goes from the PCC to the CurrentPCE in reverse direction. From the branch point, there is a downward path to the domain containing the previous hop node of node Cm on the result path tree and to the domain containing node Cm. At this branch point, the request message with T bit set to zero is sent to the PCE for node Cm along the downward path.

6. Comparing to BRPC

[RFC5441] describes the Backward Recursive Path Computation (BRPC) algorithm or procedure for computing an MPLS TE P2P LSP path from a source node to a destination node crossing multiple domains. Comparing to BRPC, there are a number of differences between BRPC and the Forward-Search P2P TE LSP Inter-Domain Path Computation (FSPC). Some of the differences are briefed below.

First, for BRPC to compute a shortest path from a source node to a destination node crossing multiple domains, we MUST provide a sequence of domains from the source node to the destination node to BRPC in advance. It is a big burden and very challenging for users to provide a sequence of domains for every LSP path crossing domains in general. In addition, it increases the cost of operation and maintenance of the network. FSPC does not need any sequence of domains for computing a shortest path.

Secondly, for a given sequence of domains domain(1), domain(2), ... ,domain(n), BRPC searches the shortest path from domain(n), to domain(n-1), until domain(1) along the reverse order of the given sequence of domain. It will get the shortest path within the given domain sesuence. FSPC calculates an optimal path in a special topology from the source node to the destination node. It will find the shortest path within all the domains.

Moreover, if the sequence of domains from the source node to the destination node provided to BRPC does not contain the shortest path from the source to the destination, then the path computed by BRPC is not optimal. FSPC guarantees that the path found is optimal.

7. Extensions to PCEP

This section describes the extensions to PCEP for FSPC. The extensions include the definition of a new flag in the RP object, a result path list and a candidate node list in the PCReq and PCRep message.

7.1. RP Object Extension

The following flags are added into the RP Object:

The F bit is added in the flag bits field of the RP object to tell the receiver of the message that the request/reply is for FSPC.

- F (FSPC bit - 1 bit):
  - 0: This indicates that this is not a PCReq/PCRep for FSPC.
  - 1: This indicates that this is a PCReq or PCRep for FSPC.

The T bit is added in the flag bits field of the RP object to tell the receiver of the message that the request is for transferring a request message to the domain containing the node with minimum cost in the candidate list.

- o T (Transfer request bit - 1 bit):
  - 0: This indicates that this is not a PCReq for transferring a request message.
  - 1: This indicates that this is a PCReq message for transferring a request message.

Setting Transfer request T-bit in a RP Object to one indicates that a reqest message containing the RP Object is for transferring a request message to the domain containing the node with minimum cost in the candidate list.

The IANA request is referenced in Section below (Request Parameter Bit Flags) of this document.

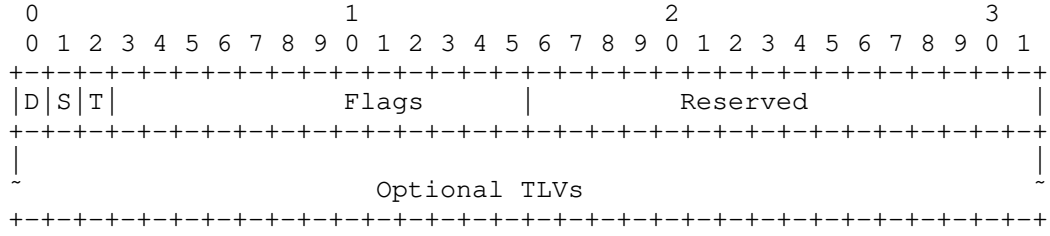
This F bit with the N bit defined in [RFC6006] can indicate whether the request/reply is for FSPC of an MPLS TE P2P LSP or an MPLS TE P2MP LSP.

- o F = 1 and N = 0: This indicates that this is a PCReq/PCRep message for FSPC of an MPLS TE P2P LSP.
- o F = 1 and N = 1: This indicates that this is a PCReq/PCRep message for FSPC of an MPLS TE P2MP LSP.

7.2. NODE-FLAGS Object

The NODE-FLAGS object is used to indicate the characteristics of the node in a Candidate Node List in a request or reply message for FSPC. The NODE-FLAGS object comprises a Reserved field, and a number of Flags. The NODE-FLAGS object may also contain a set of TLVs.

The format of the NODE-FLAGS object body is as follows:



NODE-FLAGS Object Body

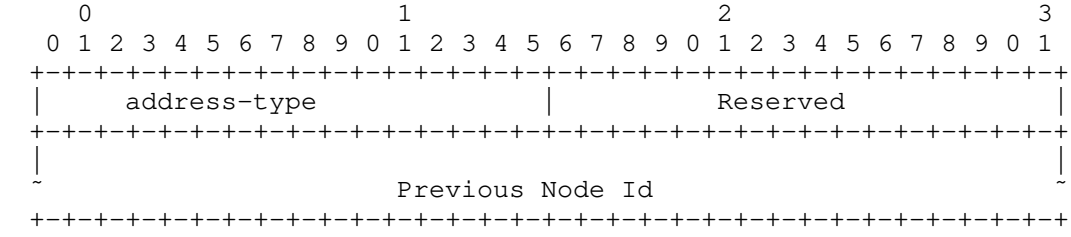
where

- D = 1: The node is a destination node.
- S = 1: The node is a source node.
- T = 1: The node is on the result path tree.

Following are the TLVs defined to convey the characteristics of the candidate node.

7.2.1. PREVIOUS-NODE TLV

The PREVIOUS-NODE TLV contains the Previous Node Id of the candidate node. The PREVIOUS-NODE TLV has the following format:



PREVIOUS-NODE TLV format

The Type of PREVIOUS-NODE TLV is to be assigned by IANA.

The length is 8 if the address type is IPv4 or 20 if the address type is IPV6.

Address Type (16 bits): Indicates the address type of Previous Node Id. 1 means IPv4 address type, 2 means IPv6 address type.

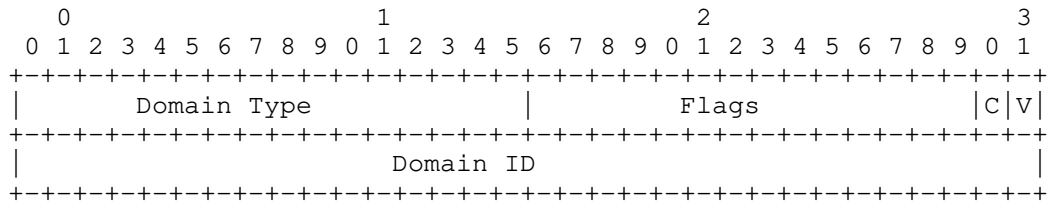
Reserved(16 bits): SHOULD be set to zero on transmission and MUST be ignored on receipt.

Previous Node Id : IP address of the node.

7.2.2. DOMAIN-ID TLV

The DOMAIN-ID TLV contains the domain Id of the candidate node (ABR or ASBR). The NODE-FLAG Object may include multiple DOMAIN-ID TLVs when the candidate node has visibility into multiple Domains.

The DOMAIN-ID TLV has the following format:



DOMAIN-ID TLV format

The Type of DOMAIN-ID TLV is to be assigned by IANA.

The length is 8.

Domain Type (8 bits): Indicates the domain type. There are two types of domain defined currently:

o Type=1: the Domain ID field carries an IGP Area ID.

o Type=2: the Domain ID field carries an AS number.

C Flag (1 bit): If the flag is set to 1, it indicates the candidate node is added into Candidate Node List by this domain.

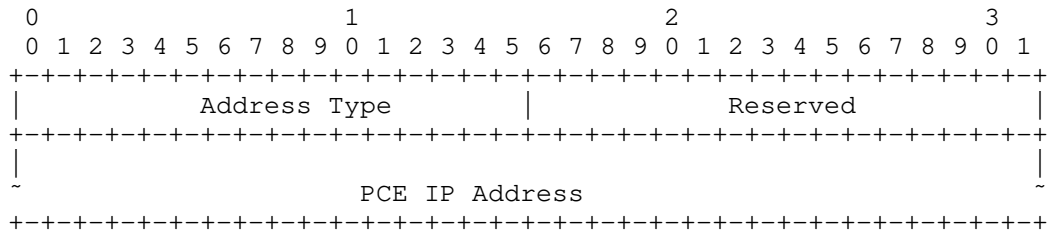
V Flag (1 bit): If the flag is set to 1, it indicates the candidate node has been expanded in this domain.

Domain ID (32 bits): With the Domain Type set to 1, this indicates the 32-bit Area ID of an IGP area where the candidate belongs. With Domain Type set to 2, this indicates an AS number of an AS where the candidate belongs. When the AS number is coded in two octets, the AS Number field MUST have its first two octets set to 0.

[Editor's note: [PCE-HIERARCHY-EXT], section 3.1.3 deals with the encoding of Domain-Id TLV in OPEN Object. Later on DOMAIN-ID TLV defined here can be incorporate with this draft]

7.2.3. PCE-ID TLV

The PCE-ID TLV is used to indicate the PCE that added this node to the CandidateList. The PCE-ID TLV has the following format:



PCE-ID TLV format

The type of PCE-ID TLV is to be assigned by IANA.

The length is 8.

Address Type (16 bits): Indicates the address type of PCE IP Address. 1 means IPv4 address type, 2 means IPv6 address type.

PCE IP Address: Indicates the reachable address of a PCE.

[Editor's note: [PCE-HIERARCHY-EXT], section 3.1.4 deals with the encoding of PCE-Id TLV in OPEN Object. Later on PCE-ID TLV defined here can be incorporate with this draft]

7.3. Candidate Node List

The Candidate Node List has the following format:

```
<candidate-node-list>::= <node>
                         [<candidate-node-list>]
where

<node>::= <ERO>  <NODE-FLAGS>
          <attribute-list>

<attribute-list>::=<metric-list>
                   [<IRO>]

<metric-list>::=<METRIC>[<metric-list>]
```

The ERO in a candidate node contain just the path segment of the last link of the path, which is from the previous hop node of the tail end node of the path to the tail end node. With this information, we can graft the candidate node into the existing result path list or tree.

Simply speaking, a candidate node has the same or similar format of a path defined in [RFC5440], but the ERO in the candidate node just contain the tail end node of the path and its previous hop, and the candidate path may contain a new object NODE-FLAGS along with new TLVs.

7.4. Result Path List

The Result Path List has the following format:

```
<result-path-list>::= <node>
                      [<result-path-list>]
where

<node>::= <ERO>  <NODE-FLAGS>
          <attribute-list>

<attribute-list>::=<metric-list>
                   [<IRO>]

<metric-list>::=<METRIC>[<metric-list>]
```

The usage of ERO, NODE-FLAGS objects etc, is similar to Candidate Node List. The T-bit of NODE-FLAGS Object would be set denoting that the best path to this node is already found.

7.5. Request Message Extension

Below is the message format for a request message with the extension of result-path-list and candidate-node-list:

```
<PCReq Message>::= <Common Header>
                   [<svec-list>]
                   <request-list>

<request-list>::=<request>[<request-list>]

<request>::= <RP> <END-POINTS> [<OF>] [<LSPA>] [<BANDWIDTH>]
             [<metric-list>]  [<RRO>[<BANDWIDTH>]]  [<IRO>]
             [<LOAD-BALANCING>]
             [<result-path-list>]
             [<candidate-node-list>]
      where:
             <result-path-list> and <candidate-node-list>
             are as defined above.
```

7.6. Reply Message Extension

Below is the message format for a reply message with the extension of result-path-list and candidate-node-list:

```
<PCRep Message> ::= <Common Header>
                    <response-list>

<response-list> ::=<response>[<response-list>]

<response> ::= <RP> [<NO-PATH>] [<attribute-list>]
               [<path-list>]
               [<result-path-list>]
               [<candidate-node-list >]
where:
               <result-path-list> and <candidate-node-list>
               are as defined above.
```

If the path from the source to the destination is found, the reply message contains path-list comprising the information of the path.

8. Suggestion to improve performance

To get much better performance all the candidate nodes of current domain can be expanded before moving on to a new domain. Note in this case, after expanding the least cost candidate node, PCE can look for and expand any other candidates in this domain.

9. Manageability Considerations

9.1. Control of Function and Policy

TBD

9.2. Information and Data Models

TBD

9.3. Liveness Detection and Monitoring

TBD

9.4. Verify Correct Operations

TBD

9.5. Requirements On Other Protocols

TBD

9.6. Impact On Network Operations

TBD

10. Security Considerations

The mechanism described in this document does not raise any new security issues for the PCEP protocols.

11. IANA Considerations

This section specifies requests for IANA allocation.

11.1. Request Parameter Bit Flags

Two new RP Object Flags have been defined in this document. IANA is requested to make the following allocation from the "PCEP RP Object Flag Field" Sub-Registry:

Bit	Description		Reference
TBA	FSPC	(F-bit)	This I-D
TBA	Transfer Request	(T-bit)	This I-D

Setting FSPC F-bit in a RP Object to one indicates that a request/ reply message containing the RP Object is for FSPC.

Setting Transfer Request T-bit in a RP Object to one indicates that a request message containing the RP Object is for transferring a request message to the domain containing the node with minimum cost in the candidate list.

11.2. New PCEP Object

11.2.1. NODE-FLAGS Object

The NODE-FLAGS Object-Type and Object-Class has been defined in this document. IANA is requested to make the following allocation:

NODE-FLAGS Object-Type : TBA

NODE-FLAGS Object-Class: TBA

Flag field of the NODE-FLAG Object:

Bit	Description	Reference
0	The node is a destination node	This I-D
1	The node is a source node	This I-D
2	The node is on the result path tree	This I-D

Each bit should be tracked with the following qualities:

- Bit number (counting from bit 0 as the most significant bit)
- Name flag
- Reference

11.3. New PCEP TLV

IANA is requested to make the following allocation:

Value	Meaning	Reference
TBA	DOMAIN-ID TLV	This I-D
TBA	PCE-ID TLV	This I-D
TBA	PREVIOUS-NODE TLV	This I-D

11.3.1. DOMAIN-ID TLV

IANA is requested to make the following allocation:

Flag field of the DOMAIN-ID TLV

Bit	Name	Description	Reference
15	V-bit	Candidate Node has been expanded by the domain	This I-D
14	C-bit	Candidate Node added by the domain	This I-D

Each bit should be tracked with the following qualities:

- Bit number (counting from bit 0 as the most significant bit)
- Name flag
- Reference

12. Acknowledgement

The authors would like to appreciate Dhruv Dhody for his great contributions and to thank Julien Meuric, Daniel King, Ramon Casellas, Cyril Margaria,Olivier Dugeon, Oscar Gonzalez de Dios, Udayasree Palle, Reeja Paul and Sandeep Kumar Boina for their valuable comments on this draft.

13. References

13.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, DOI 10.17487/RFC4655, August 2006, <https://www.rfc-editor.org/info/rfc4655>.

[RFC5440] Vasseur, JP., Ed. and JL. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, DOI 10.17487/RFC5440, March 2009, <https://www.rfc-editor.org/info/rfc5440>.

13.2. Informative References

[RFC4105] Le Roux, J., Ed., Vasseur, J., Ed., and J. Boyle, Ed., "Requirements for Inter-Area MPLS Traffic Engineering", RFC 4105, DOI 10.17487/RFC4105, June 2005, <https://www.rfc-editor.org/info/rfc4105>.

[RFC4216] Zhang, R., Ed. and J. Vasseur, Ed., "MPLS Inter-Autonomous System (AS) Traffic Engineering (TE) Requirements", RFC 4216, DOI 10.17487/RFC4216, November 2005, <https://www.rfc-editor.org/info/rfc4216>.

[RFC5441] Vasseur, JP., Ed., Zhang, R., Bitar, N., and JL. Le Roux, "A Backward-Recursive PCE-Based Computation (BRPC) Procedure to Compute Shortest Constrained Inter-Domain Traffic Engineering Label Switched Paths", RFC 5441, DOI 10.17487/RFC5441, April 2009, <https://www.rfc-editor.org/info/rfc5441>.

[RFC6006] Zhao, Q., Ed., King, D., Ed., Verhaeghe, F., Takeda, T., Ali, Z., and J. Meuric, "Extensions to the Path Computation Element Communication Protocol (PCEP) for Point-to-Multipoint Traffic Engineering Label Switched Paths", RFC 6006, DOI 10.17487/RFC6006, September 2010, <https://www.rfc-editor.org/info/rfc6006>.

[PCE-HIERARCHY-EXT]
Zhang, F., Zhao, Q., King, O., Casellas, R., and D. King, "Extensions to Path Computation Element Communication Protocol (PCEP) for Hierarchical Path Computation Elements (PCE) (draft-zhang-pce-hierarchy-extensions-02)", August 2012.

Author's Address

Huaimo Chen
Futurewei
Boston, MA
USA

EMail: Huaimo.chen@futurewei.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: May 2, 2012

E. Crabbe
Google, Inc.
J. Medved
R. Varga
Juniper Networks, Inc.

October 30, 2011

PCEP Extensions for Stateful PCE
draft-crabbe-pce-stateful-pce-01

Abstract

The Path Computation Element Communication Protocol (PCEP) provides mechanisms for Path Computation Elements (PCEs) to perform path computations in response to Path Computation Clients (PCCs) requests.

Although PCEP explicitly makes no assumptions regarding the information available to the PCE, it also makes no provisions for synchronization or PCE control of timing and sequence of path computations within and across PCEP sessions. This document describes a set of extensions to PCEP to enable this functionality, providing stateful control of Multiprotocol Label Switching (MPLS) Traffic Engineering Label Switched Paths (TE LSP) via PCEP.

Requirements Language

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

Status of this Memo

Table of Contents

1. Introduction

[RFC5440] describes the Path Computation Element Protocol (PCEP. PCEP defines the communication between a Path Computation Client (PCC) and a Path Control Element (PCE), or between PCE and PCE, enabling computation of Multiprotocol Label Switching (MPLS) for Traffic Engineering Label Switched Path (TE LSP) characteristics

This document specifies a set of extensions to PCEP to enable stateful control of TE LSPs between and across PCEP sessions in compliance with [RFC4657]. It includes mechanisms to effect LSP state synchronization between PCCs and PCEs, delegation of control of LSPs to PCEs, and PCE control of timing and sequence of path computations within and across PCEP sessions.

2. Terminology

This document uses the following terms defined in [RFC5440]: PCC, PCE, PCEP Peer.

This document uses the following terms defined in [RFC4090]: MPLS TE Fast Reroute (FRR), FRR One-to-One Backup, FRR Facility Backup.

The following terms are defined in this document:

Passive Stateful PCE: uses LSP state information learned from PCCs to optimize path computations. It does not actively update LSP state. A PCC maintains synchronization with the PCE.

Active Stateful PCE: uses LSP state information learned from PCCs to optimize path computations. Additionally, it actively updates LSP parameters in those PCCs that delegated control over their LSPs to the PCE.

Delegation: An operation to grant a PCE temporary rights to modify a subset of LSPs parameters on one or more PCC's LSPs. LSPs are delegated from a PCC to a PCE.

Delegation Timeout Interval: when a PCEP session is terminated, a PCC waits for this time period before revoking LSP delegation to a PCE.

LSP State Report: an operation to send LSP state (Operational / Admin Status, LSP attributes configured and set by a PCE, etc.) from a PCC to a PCE.

LSP Update Request: an operation where a PCE requests a PCC to update one or more attributes of an LSP and to re-signal the LSP with updated attributes.

LSP Priority: a specific pair of MPLS setup and hold priority values.

Minimum Cut Set: the minimum set of links for a specific source destination pair which, when removed from the network, result in a specific source being completely isolated from specific destination. The summed capacity of these links is equivalent to the maximum capacity from the source to the destination by the max-flow min-cut theorem.

MPLS TE Global Default Restoration: once an LSP failure is detected by some downstream mode, the head-end LSP is notified by means of RSVP. Upon receiving the notification, the headend LSR recomputes the path and signals the LSP along an alternate path. [NET-REC]

MPLS TE Global Path Protection: once an LSP failure is detected by some downstream mode, the head-end LSP is notified by means of RSVP. Upon receiving the notification, the headend LSR reroutes traffic using a pre-signaled backup (secondary) LSP. [NET-REC].

Within this document, when describing PCE-PCE communications, the requesting PCE fills the role of a PCC. This provides a saving in documentation without loss of function.

The message formats in this document are specified using Routing Backus-Naur Format (RBNF) encoding as specified in [RFC5511].

3. Motivation and Objectives

3.1. Motivation

3.1.1. Background

Traffic engineering has been a goal of the MPLS architecture since its inception ([RFC3031], [RFC2702], [RFC3346]). In the traffic engineering system provided by [RFC3630], [RFC5305], and [RFC3209] information about network resources utilization is only available as total reserved capacity by traffic class on a per interface basis; individual LSP state is available only locally on each LER for it's own LSPs. In most cases, this makes good sense, as distribution and retention of total LSP state for all LERs within in the network would be prohibitively costly.

Unfortunately, this visibility in terms of global LSP state may result in a number of issues for some demand patterns, particularly within a common setup and hold priority. This issue affects online traffic engineering systems, and in particular, the widely implemented but seldom deployed auto-bandwidth system.

A sufficiently over-provisioned system will by definition have no issues routing its demand on the shortest path. However, lowering the degree to which network over-provisioning is required in order to run a healthy, functioning network is a clear and explicit promise of MPLS architecture. In particular, it has been a goal of MPLS to provide mechanisms to alleviate congestion scenarios in which "traffic streams are inefficiently mapped onto available resources; causing subsets of network resources to become over-utilized while others remain underutilized" ([RFC2702]).

3.1.2. Why a Stateful PCE?

[RFC4655] defines a stateful PCE to be one in which the PCE maintains "strict synchronization between the PCE and not only the network states (in term of topology and resource information), but also the set of computed paths and reserved resources in use in the network." [RFC4655] also expressed a number of concerns with regard to a stateful PCE, specifically:

- Any reliable synchronization mechanism would result in significant control plane overhead
- Out-of-band ted synchronization would be complex and prone to race conditions
- Path calculations incorporating total network state would be highly complex

In general, stress on the MPLS TE control plane will be directly proportional to the size of the system being controlled and the and the tightness of the control loop, and indirectly proportional to the amount of over-provisioning in terms of both network capacity and reservation overhead.

Despite these concerns in terms of implementation complexity and scalability, several TE algorithms exist today that have been demonstrated to be extremely effective in large TE systems, providing both rapid convergence and significant benefits in terms of optimality of resource usage [MXMN-TE]. All of these systems share at least two common characteristics: the requirement for both global visibility of a flow (or in this case, a TE LSP) state and for ordered control of path reservations across devices within the system

being controlled. While some approaches have been suggested in order to remove the requirements for ordered control (See [MPLS-PC]), these approaches are highly dependent on traffic distribution, and do not allow for multiple simultaneous LSP priorities representing diffserv classes.

The following use cases demonstrate a need for visibility into global inter-PCC LSP state in PCE path computations, and for a PCE control of sequence and timing in altering LSP path characteristics within and across PCEP sessions. Reference topologies for the use cases described later in this section are shown in Figures 1 and 2.

All use cases assume that all LSPs listed exist at the same LSP priority.

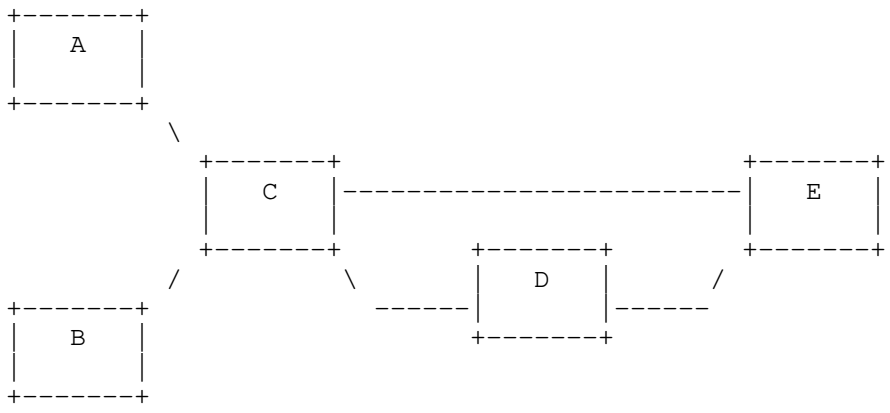


Figure 1: Reference topology 1

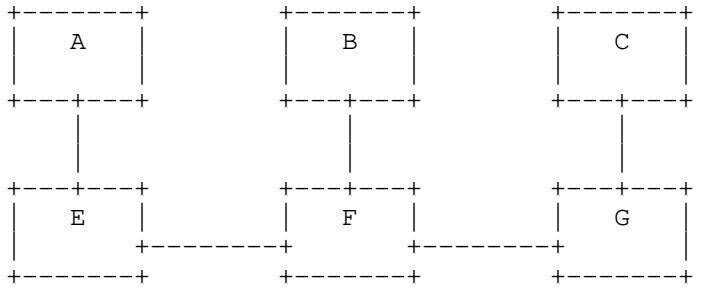


Figure 2: Reference topology 2

3.1.2.1. Throughput Maximization and Bin Packing

Because LSP attribute changes in [RFC5440] are driven by PCReq messages under control of a PCC's local timers, the sequence of RSVP reservation arrivals occurring in the network will be randomized. This, coupled with a lack of global LSP state visibility on the part of a stateless PCE may result in suboptimal throughput in a given network topology.

Reference topology 2 in Figure 2 and Tables 1 and 2 show an example in which throughput is at 50% of optimal as a result of lack of visibility and synchronized control across PCC's. In this scenario, the decision must be made as to whether to route any portion of the E-G demand, as any demand routed for this source and destination will decrease system throughput. This is addressed in Section 3.1.2.2.

Link	Metric	Capacity
A-E	1	10
B-F	1	10
C-G	1	10
E-F	1	10
F-G	1	10

Table 1: Link parameters for Throughput use case

Time	LSP	Src	Dst	Demand	Routable	Path
1	1	E	G	10	Yes	E-F-G
2	2	A	B	10	No	---
3	1	B	C	10	No	---

Table 2: Throughput use case demand time series

In many cases throughput maximization becomes a bin packing problem. While bin packing itself is an NP-hard problem, a number of common heuristics which run in polynomial time can provide significant improvements in throughput over random reservation event distribution, especially when traversing links which are members of the minimum cut set for a large subset of source destination pairs.

Tables 3 and 4 show a simple use case using Reference Topology 1 in Figure 1, where LSP state visibility and control of reservation order across PCCs would result in significant improvement in total

throughput.

Link	Metric	Capacity
A-C	1	10
B-C	1	10
C-E	10	5
C-D	1	10
D-E	1	10

Table 3: Link parameters for Bin Packing use case

Time	LSP	Src	Dst	Demand	Routable	Path
1	1	A	E	5	Yes	A-C-D-E
2	2	B	E	10	No	---

Table 4: Bin Packing use case demand time series

3.1.2.2. Max-Min Fair Allocation

3.1.2.3. Deadlock

Most existing RSVP-TE implementations will not tear down existing, established LSPs in the face of path setup in order to effect bandwidth increase of an existing tunnel [RFC3209]. While this behavior is directly implied to be correct in [RFC3209] it is not desirable from an operator's perspective, because either a) the destination prefixes are not reachable via any means other than MPLS or b) this would result in significant packet loss as demand is shifted to other LSPs in the overlay mesh.

In addition, there are currently few implementations offering ingress admission control at the LSP level. Again, having ingress admission control on a per LSP basis is not necessarily desirable from an operational perspective, as a) one must over-provision tunnels significantly in order to avoid deleterious effects resulting from stacked transport and flow control systems and b) there is currently no efficient commonly available northbound interface for dynamic configuration of per LSP ingress admission control (such an interface could easily be defined using the extensions present in this spec, but it beyond the scope of the current document).

Lack of ingress admission control coupled with the behavior in

[RFC3209] effectively results in mis-signaled LSPs during periods of contention for network capacity between LSPs in a given LSP priority. This in turn causes information loss in the TED with regard to actual network state, resulting in LSPs sharing common network interfaces with mis-signaled LSPs operating in a degraded state for significant periods of time, even when unused network capacity may potentially be available.

Reference Topology 2 in Figure 2 and Tables 5 and 6 show a use case that demonstrates this behavior. The problem could be easily ameliorated by global visibility of LSP state coupled with PCC-external demand measurements.

Link	Metric	Capacity
A-C	1	10
B-C	1	10
C-E	10	5
C-D	1	10
D-E	1	10

Table 5: Link parameters for the 'Deadlock' example

Time	LSP	Src	Dst	Demand	Routable	Path
1	1	A	E	2	Yes	A-C-D-E
2	2	B	E	2	Yes	B-C-D-E
3	1	A	E	20	No	---

Table 6: Deadlock LSP and demand time series

3.1.2.4. Minimal Perturbation Problem

3.1.2.5. Predictability

Randomization of reservation events caused by lack of control over event ordering across PCE sessions results in poor predictability in LSP routing. An offline system applying a consistent optimization method will produce predictable results to within either the boundary of forecast error when reservations are over-provisioned by reasonable margins or to the variability of the signal and the forecast error when applying some hysteresis in order to minimize churn.

Reference Topology 1 and Tables 7, 8 and 9 show the impact of event ordering and predictability of LSP routing.

Link	Metric	Capacity
A-C	1	10
B-C	1	10
C-E	1	10
C-D	1	10
D-E	1	10

Table 7: Link parameters for the 'Predictability' example

Time	LSP	Src	Dst	Demand	Routable	Path
1	1	A	E	7	Yes	A-C-E
2	2	B	E	7	Yes	B-C-D-E

Table 8: Predictability LSP and demand time series 1

Time	LSP	Src	Dst	Demand	Routable	Path
1	2	B	E	7	Yes	B-C-E
2	1	A	E	7	Yes	A-C-D-E

Table 9: Predictability LSP and demand time series 2

3.1.2.6. Global Concurrent Optimization

Global Concurrent Optimization (GCO) defined in [RFC5557] is a network optimization mechanism that is able to simultaneously consider the entire topology of the network and the complete set of existing TE LSPs and their existing constraints, and look to optimize or reoptimize the entire network to satisfy all constraints for all TE LSPs. It allows for bulk path computations in order to avoid blocking problems and to achieve more optimal network-wide solutions.

Global control of LSP operation sequence in [RFC5557] is predicated on the use of what is effectively a stateful (or semi-stateful) NMS. The NMS can be either not local to the switch, in which case another northbound interface is required for LSP attribute changes, or local/ collocated, in which case there are significant issues with

efficiency in resource usage. Stateful PCE adds a few features that:

- Roll the NMS visibility into the PCE and remove the requirement for an additional northbound interface
- Allow the PCE to determine when re-optimization is needed
- Allow the PCE to determine which LSPs should be re-optimized
- Allow a PCE to control the sequence of events across multiple PCCs, allowing for bulk (and truly global) optimization, LSP shuffling etc.

3.1.3. Protocol vs. Configuration

Note that existing configuration tools and protocols can be used to set LSP state. However, this solution has several shortcomings:

- Scale & Performance: configuration operations often require processing of additional configuration portions beyond the state being directly acted upon, with corresponding cost in CPU cycles, negatively impacting both PCC stability LSP update rate capacity.
- Scale & Performance: configuration operations often have transactional semantics which are typically heavyweight and require additional CPU cycles, negatively impacting PCC update rate capacity.
- Security: opening up a configuration channel to a PCE would allow a malicious PCE to take over a PCC. The proposed PCEP extensions only allow a PCE control over a very limited set of LSP attributes.
- Interoperability: each vendor has a proprietary information model for configuring LSP state, which prevents interoperability of a PCE with PCCs from different vendors. The proposed PCEP extensions allow for a common information model for LSP state for all vendors.
- Efficient State Synchronization: configuration channels may be heavyweight and unidirectional, therefore efficient state synchronization between a PCE and a PCE may be a problem.

3.2. Objectives

The objectives for the protocol extensions to support stateful PCE described in this document are as follows:

- o Allow a single PCC to interact with a mix of stateless and stateful PCEs simultaneously using the same PCEP.
- o Support efficient LSP state synchronization between the PCC and one or more active or passive stateful PCEs.
- o Allow a PCC to delegate control of its LSPs to an active stateful PCE such that a single LSP is under the control a single PCE at any given time. A PCC may revoke this delegation at any point during the lifetime of the PCEP session. A PCE may return this delegation at any point during the lifetime of the PCEP session.
- o Allow a PCE to control computation timing and update timing across all LSPs that have been delegated to it.
- o Allow a PCE to specify protection / restoration settings for all LSPs that have been delegated to it.
- o Enable uninterrupted operation of PCC's LSPs in the event PCE failure or while control of LSPs is being transferred between PCEs.

4. New Functions to Support Stateful PCEs

Several new functions will be required in PCEP to support stateful PCEs. A function can be initiated either from a PCC towards a PCE (C-E) or from a PCE towards a PCC (E-C). The new functions are:

Capability negotiation (E-C,C-E): both the PCC and the PCE must announce during PCEP session establishment that they support PCEP Stateful PCE extensions defined in this document.

LSP state synchronization (C-E): after the session between the PCC and a stateful PCE is initialized, the PCE must learn the state of a PCC's LSPs before it can perform path computations or update LSP attributes in a PCC.

LSP Update Request (E-C): A PCE requests modification of attributes on a PCC's LSP.

LSP State Report (C-E): a PCC sends an LSP state report to a PCE whenever the state of an LSP changes.

LSP control delegation (C-E,E-C): a PCC grants to a PCE the right to update LSP attributes on one or more LSPs; the PCE becomes the authoritative source of the LSP's attributes as long as the delegation is in effect (See Section 5.5); the PCC may withdraw

the delegation or the PCE may give up the delegation

In addition to new PCEP functions, stateful capabilities discovery will be required in OSPF ([RFC5088]) and IS-IS ([RFC5089]). Stateful capabilities discovery is not in scope of this document.

5. Architectural Overview of Protocol Extensions

5.1. LSP State Ownership

In the PCEP protocol (defined in [RFC5440]), LSP state is owned by the PCC. While the PCC receives LSP attribute values from an external PCE, it is the PCC that decides when and how to apply received parameters and setup the LSP. With PCEP extensions proposed in this draft, an active stateful PCE may have control of a PCC's LSPs be delegated to it, but the LSP state ownership is retained by the PCC. In particular, in addition to specifying values for (a subset of) LSP's attributes, an active stateful PCE also decides when to make LSP modifications .

Retaining LSP state ownership on the PCC allows for:

- a PCC to interact with both stateless and stateful PCEs at the same time
- a stateful PCE to only modify a small subset of LSP parameters, i.e. to set only a small subset of the overall LSP state; other parameters may be set by the operator through CLI commands
- a PCC to revert delegated LSP to an operator-defined default or to delegate the LSPs to a different PCE, if the PCC get disconnected from a PCE with currently delegated LSPs

5.2. New Messages

In this document, we define the following new PCEP messages:

Path Computation State Report (PCRpt): a PCEP message sent by a PCE to a PCC to report the status of one or more LSPs. Each LSP Status Report in a PCRpt message can contain the actual LSP's path,bandwidth, operational and administrative status, etc. An LSP Status Report carried on a PCRpt message is also used in delegation or revocation of control of an LSP to/from a PCE. The PCRep message is described in Section 6.1.

Path Computation Update Request (PCUpd): a PCEP message sent by a PCE to a PCC to update LSP parameters, on one or more LSPs. Each LSP Update Request on a PCUpd message MUST contain all LSP parameters that a PCE wishes to set for a given LSP. An LSP Update Request carried on a PCUpd message is also used to return LSP delegations if at any point PCE no longer desires control of an LSP. The PCUpd message is described in Section 6.2.

The new functions defined in Section 4 are mapped onto the new messages as shown in the following table.

Function	Message
Capability Negotiation (E-C,C-E)	Open
State Synchronization (C-E)	PCRpt
LSP State Report (C-E)	PCRpt
LSP Control Delegation (C-E,E-C)	PCRp, PCUpd
LSP Update Request (E-C)	PCUpd
ISIS stateful capability advertisement	ISIS PCE-CAP-FLAGS sub-TLV
OSPF stateful capability advertisement	OSPF RI LSA, PCE TLV, PCE-CAP-FLAGS sub-TLV

Table 10: New Function to Message Mapping

5.3. Capability Negotiation

During PCEP Initialization Phase, PCEP Speakers (PCE pr PCC) negotiate the use of stateful PCEP extensions. A PCEP Speaker includes the "Stateful PCE Capability" TLV, described in Section 7.1.1, in the OPEN Object to advertise its support for PCEP stateful extensions. The Stateful Capability TLV includes the 'LSP Update' Flag that indicates whether the PCEP Speaker supports LSP parameter updates.

The presence of the Stateful PCE Capability TLV in PCC's OPEN Object indicates that the PCC is willing to send LSP State Reports whenever LSP parameters or operational status changes.

The presence of the Stateful PCE Capability TLV in PCE's OPEN message indicates that the PCE is interested in receiving LSP State Reports whenever LSP parameters or operational status changes.

The PCEP protocol extensions for stateful PCEs MAY only be used if both sides have included the Stateful PCE Capability TLV in their respective OPEN messages, otherwise a PCErr with code "Stateful PCE

capability not negotiated" (see Section 7.3) will be generated and the PCEP session will be terminated.

LSP delegation and LSP update operations defined in this document MAY only be used if both PCEP Speakers set the 'LSP Update' Flag in the "Stateful Capability" TLV to 'Updates Allowed (U Flag = 1)', otherwise a PCErr with code "Delegation not negotiated" (see Section 7.3) will be generated. Note that even if the update capability has not been negotiated, a PCE can still receive LSP Status Reports from a PCC and build and maintain an up to date view of the state of the PCC's LSPs.

5.4. State Synchronization

The purpose of State Synchronization is to provide a checkpoint-in-time state replica of a PCC's LSP state in a PCE. State Synchronization is performed immediately after the Initialization phase ([RFC5440]).

During State Synchronization, a PCC first takes a snapshot of the state of its LSPs state, then sends the snapshot to a PCE in a sequence of LSP State Reports. The set of LSPs for which state is synchronized with a PCE is determined by negotiated stateful PCEP capabilities and PCC's local configuration (see more details in Section 9.1). A PCC indicates that State Synchronization is complete by setting the 'Sync Done' Flag to 1 on the LSP State Report for the last LSP in the synchronized set.

A PCE SHOULD NOT send PCUpd messages to a PCC before State Synchronization is complete. A PCC SHOULD NOT send PCReq messages to a PCE before State Synchronization is complete. This is to allow the PCE to get the best possible view of the network before it starts computing new paths.

If the PCC encounters a problem which prevents it from completing the state transfer, it MUST send a PCErr message to the PCE and terminate the session using the PCEP session termination procedure.

The PCE does not send positive acknowledgements for properly received synchronization messages. It MUST respond with a PCErr message indicating "PCRpt error" (see) if it encounters a problem with the LSP State Report it received from the PCC. Either the PCE or the PCC MAY terminate the session if the PCE encounters a problem during the synchronization.

The successful State Synchronization sequence is shown in Figure 3.

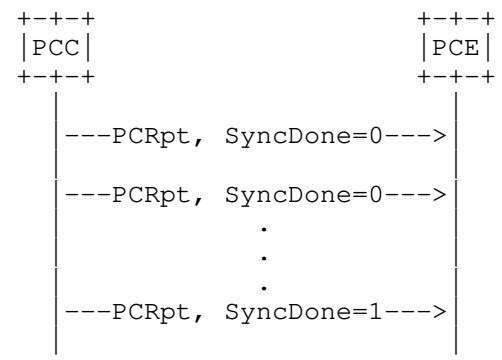


Figure 3: Successful state synchronization

The sequence where the PCE fails during the State Synchronization phase is shown in Figure 4.

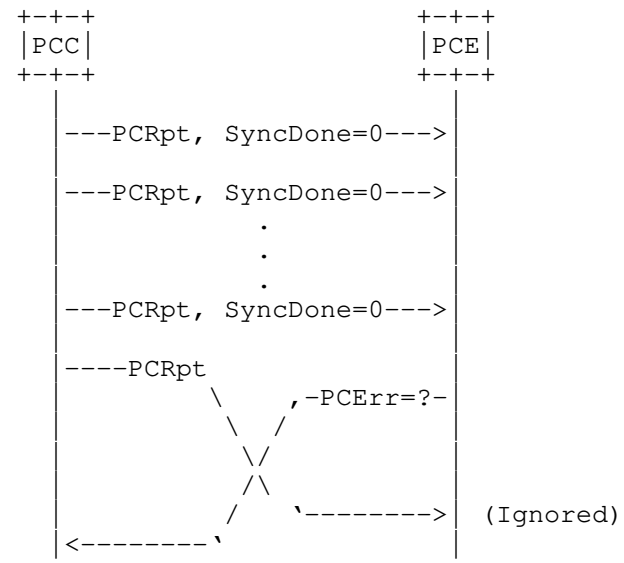


Figure 4: Failed state synchronization (PCE failure)

The sequence where the PCC fails during the State Synchronization phase is shown in Figure 5.

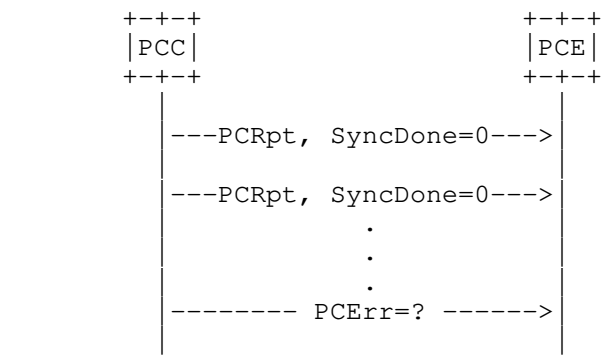


Figure 5: Failed state synchronization (PCC failure)

5.5. LSP Delegation

If during Capability negotiation both the PCE and the PCC have indicated that they support LSP Update, then the PCC may choose to grant the PCE a temporary right to update (a subset of) LSP attributes on one or more LSPs. This is called "LSP Delegation", and it MAY be performed at any time after the Initialization phase.

Delegation occurs on a per LSP basis, and different LSPs may be delegated to different PCEs. Only a single PCE may have control of an LSP and either the PCE or PCC may revoke this delegation at any time. A previously delegated LSP MAY be revoked by the PCC or MAY be given up by the PCE if the PCE no longer wishes to update the LSP's state. Delegation, Revocation, and Return are done individually for each LSP.

5.5.1. Delegating an LSP

A PCC delegates an LSP to a PCE by setting the Delegate flag in LSP State Report to 1. A PCE confirms the delegation when it sends the first LSP Update Request for the delegated LSP to the PCC by setting the Delegate flag to 1. Note that a PCE does not immediately confirm to the PCC the acceptance of LSP Delegation; Delegation acceptance is confirmed when the PCC wishes to update the LSP via the LSP Update Request. If a PCE does not accept the LSP Delegation, it MUST immediately respond with an empty LSP Update Request which has the Delegate flag set to 0.

The delegation sequence is shown in Figure 6.

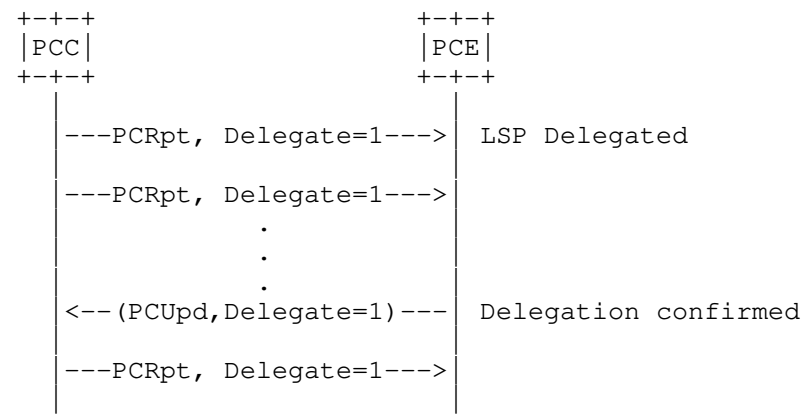


Figure 6: Delegating and LSP

Note that for an LSP to remain delegated to a PCE, the PCC MUST set the Delegate flag to 1 on each LSP Status Report sent to the PCE.

5.5.2. Revoking a Delegation

A PCC revokes an LSP delegation by sending an LSP State Report with the Delegate flag set to 0. A PCC MAY revoke an LSP delegation at any time during the PCEP session life time. After an LSP delegation has been revoked, a PCE can no longer update LSP's parameters, and will result in the PCC sending a PCErr message indicating "LSP is not delegated" (see Section 7.3).

The revocation sequence is shown in Figure 7.

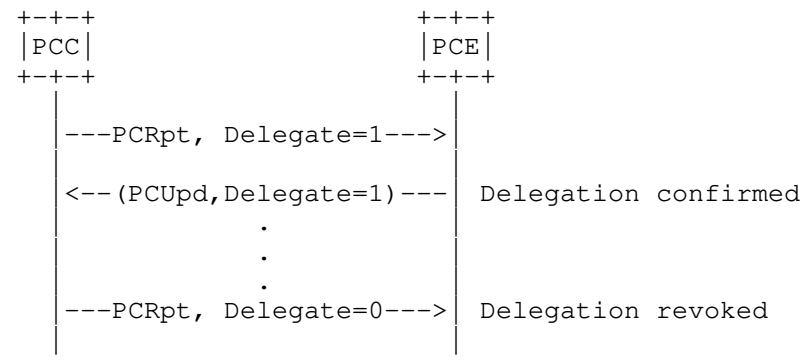


Figure 7: Revoking a Delegation

If a PCC can not delegate an LSP to a PCE (for example, if a PCC is not connected to any active stateful PCE or if no connected active

stateful PCE accepts the delegation), the LSP delegation on the PCC will time out within a configurable Delegation Timeout Interval and the PCC MUST flush any LSP state set by a PCE.

5.5.3. Returning a Delegation

A PCE that no longer wishes to update an LSP's parameters SHALL return the LSP delegation back to the PCC by sending an empty LSP Update Request which has the Delegate flag set to 0. Note that in order to keep a delegation, the PCE MUST set the Delegate flag to 1 on each LSP Update Request sent to the PCC.

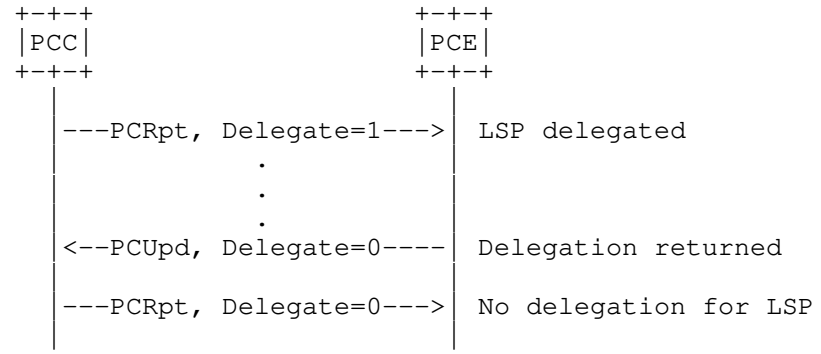


Figure 8: Returning a Delegation

If a PCC can not delegate an LSP to a PCE (for example, if a PCC is not connected to any active stateful PCE or if no connected active stateful PCE accepts the delegation), the LSP delegation on the PCC will time out within a configurable Delegation Timeout Interval and the PCC MUST flush any LSP state set by a PCE.

5.5.4. Redundant Stateful PCEs

Note that a PCE may not have any delegated LSPs: in a redundant configuration where one PCE is backing up another PCE, the backup PCE will not have any delegated LSPs. The backup PCE does not update any LSPs, but it receives all LSP State Reports from a PCC. When the primary PCE fails, a PCC will delegate to the secondary PCE all LSPs that had been previously delegated to the failed PCE.

5.6. LSP Operations

5.6.1. Passive Stateful PCE Path Computation Request/Response

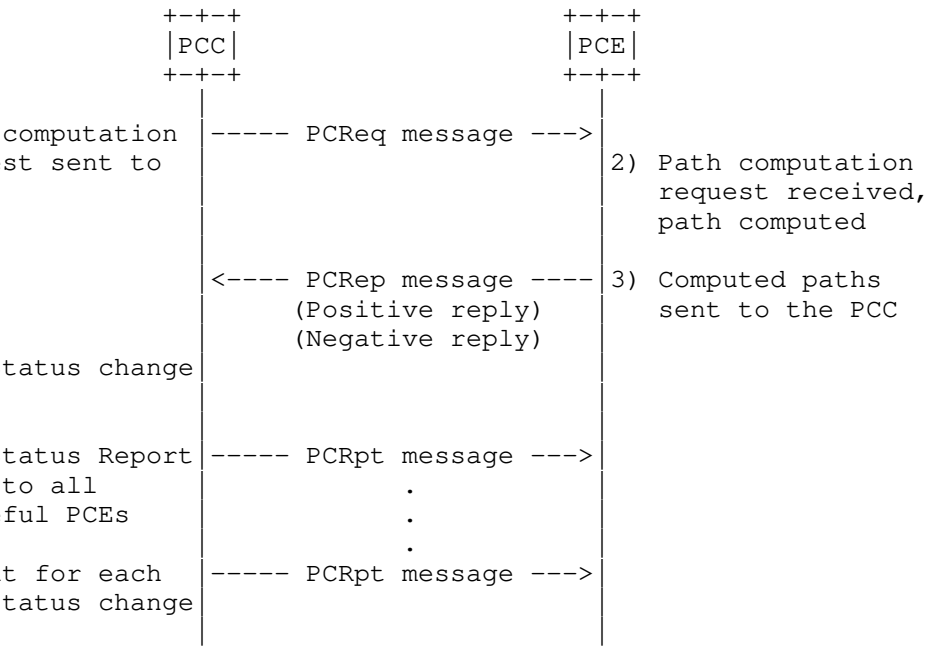


Figure 9: Passive Stateful PCE Path Computation Request/Response

Once a PCC has successfully established a PCEP session with a passive stateful PCE and the PCC's LSP state is synchronized with the PCE (i.e. the PCE knows about all PCC's existing LSPs), if an event is triggered that requires the computation of a set of paths, the PCC sends a path computation request to the PCE ([RFC5440], Section 4.2.3). The PCReq message MAY contain the LSP Object to identify the LSP for which the path computation is requested.

Upon receiving a path computation request from a PCC, the PCE triggers a path computation and returns either a positive or a negative reply to the PCC ([RFC5440], Section 4.2.4).

Upon receiving a positive path computation reply, the PCC receives a set of computed paths and starts to setup the LSPs. For each LSP, it sends an LSP State Report carried on a PCRpt message to the PCE, indicating that the LSP's status is 'Pending'.

Once an LSP is up, the PCC sends an LSP State Report carried on a PCRpt message to the PCE, indicating that the LSP's status is 'Up'. If the LSP could not be set up, the PCC sends an LSP State Report indicating that the LSP is "Down' and stating the cause of the

failure. Note that due to timing constraints, the LSP status may change from 'Pending' to 'Up' (or 'Down') before the PCC has had a chance to send an LSP State Report indicating that the status is 'Pending'. In such cases, the PCC may choose to only send the PCRpt indicating the latest status ('Up' or 'Down').

Upon receiving a negative reply from a PCE, a PCC may decide to resend a modified request or take any other appropriate action. For each requested LSP, it also sends an LSP State Report carried on a PCRpt message to the PCE, indicating that the LSP's status is 'Down'.

There is no direct correlation between PCRep and PCRpt messages. For a given LSP, multiple LSP State Reports will follow a single PC Reply, as a PCC notifies a PCE of the LSP's state changes.

A PCC sends each LSP State Report to each stateful PCE that is connected to the PCC.

Note that a single PCRpt message MAY contain multiple LSP State Reports.

The passive stateful PCE is the model for stateful PCEs is described in [RFC4655], Section 6.8.

5.6.2. Active Stateful PCE LSP Update

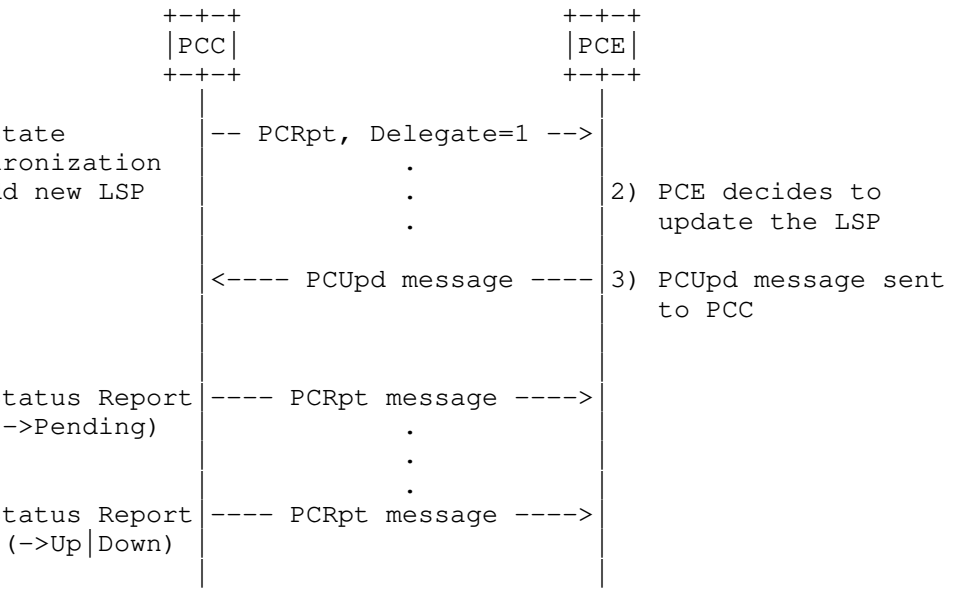


Figure 10: Active Stateful PCE

Once a PCC has successfully established a PCEP session with an active stateful PCE, the PCC's LSP state is synchronized with the PCE (i.e. the PCE knows about all PCC's existing LSPs) and LSPs have been delegated to the PCE, the PCE can modify LSP parameters of delegated LSPs.

A PCE sends an LSP Update Request carried on a PCUpd message to the PCC. The LSP Update Request contains a variety of objects that specify the set of constraints and attributes for the LSP's path. Additionally, the PCC may specify the urgency of such request by assigning a request priority. A single PCUpd message MAY contain multiple LSP Update Requests.

Upon receiving a PCUpd message the PCC starts to setup LSPs specified in LSP Update Requests carried in the message. For each LSP, it sends an LSP State Report carried on a PCRpt message to the PCE, indicating that the LSP's status is 'Pending'.

Once an LSP is up, the PCC sends an LSP State Report (PCRpt message) to the PCE, indicating that the LSP's status is 'Up'. If the LSP could not be set up, the PCC sends an LSP State Report indicating that the LSP is 'Down' and stating the cause of the failure. A PCC may choose to compress LSP State Updates to only reflect the most up to date state, as discussed in the previous section.

A PCC sends each LSP State Report to each stateful PCE that is connected to the PCC.

A PCC MUST NOT send to any PCE a Path Computation Request for a delegated LSP.

5.7. LSP Protection

With a stateless PCE or a passive stateful PCE, LSP protection and restoration settings may be operator-configured locally at a PCC. A PCE may be merely asked to compute the protected (primary) and backup (secondary) paths for the LSP.

An active stateful PCE controls the LSPs that are delegated to it, and must therefore be able to set via PCEP the desired protection / restoration mechanism for each delegated LSP. PCEP extensions for stateful PCEs SHOULD support, at a minimum, the following protection mechanisms:

- MPLS TE Global Default Restoration
- MPLS TE Global Path Protection

- o FRR One-to-One Backup
- o FRR Facility Backup - link protection, node protection, or both

5.8. Transport

A Permanent PCEP session MUST be established between a stateful PCEs and the PCC.

State cleanup after session termination, as well as session setup failures will be described in a later version of this document.

6. PCEP Messages

As defined in [RFC5440], a PCEP message consists of a common header followed by a variable-length body made of a set of objects that can be either mandatory or optional. An object is said to be mandatory in a PCEP message when the object must be included for the message to be considered valid. For each PCEP message type, a set of rules is defined that specify the set of objects that the message can carry. An implementation MUST form the PCEP messages using the object ordering specified in this document.

6.1. The PCRpt Message

A Path Computation LSP State Report message (also referred to as PCRpt message) is a PCEP message sent by a PCC to a PCE to report the current state of an LSP. A PCRpt message can carry more than one LSP State Reports. A PCC can send an LSP State Report either in response to an LSP Update Request from a PCE, or asynchronously when the state of an LSP changes. The Message-Type field of the PCEP common header for the PCRpt message is set to [TBD].

The format of the PCRpt message is as follows:

```
<PCRpt Message> ::= <Common Header>
                    <state-report-list>
```

Where:

```
<state-report-list> ::= <state-report>[<state-report-list>]

<state-report> ::= <LSP>
                   [<primary-path> [<backup-path-list>]]
```

Where:

```
<primary-path>::=<path>

<backup-path-list>::=<path>[<backup-path-list>]

<path>::=<ERO><attribute-list>
```

Where:

```
<attribute-list> ::= [<LSPA>]
                     [<BANDWIDTH>]
                     [<RRO>]
                     [<metric-list>]

<metric-list> ::= <METRIC>[<metric-list>]
```

The LSP object (see Section 7.2) is mandatory, and it MUST be included in each LSP State Report on the PCRpt message. If the LSP object is missing, the receiving PCE MUST send a PCErr message with Error-type=6 (Mandatory Object missing) and Error-value=[TBD] (LSP object missing).

The LSP State Report MAY contain a path descriptor for the primary path and one or more path descriptors for backup paths, if MPLS TE Global Default Restoration or MPLS TE Global Path Protection had been specified on the LSP. A path descriptor MUST contain an ERO object as it was specified by a PCE or an operator. A path descriptor MUST contain the RRO object if a primary or secondary LSP is set up along the path in the network. A path descriptor MAY contain the LSPA, BANDWIDTH, and METRIC objects. The ERO,LSPA, BANDWIDTH, METRIC, and RRO objects are defined in[RFC5440].

6.2. The PCUpd Message

A Path Computation LSP Update Request message (also referred to as PCUpd message) is a PCEP message sent by a PCE to a PCC to update attributes of an LSP. A PCUpd message can carry more than one LSP Update Request. The Message-Type field of the PCEP common header for

the PCRpt message is set to [TBD].

The format of a PCUpd message is as follows:

```
<PCUpd Message> ::= <Common Header>
                    <udpate-request-list>
```

Where:

```
<update-request-list> ::= <update-request>[<update-request-list>]

<update-request> ::= <LSP>
                     [<primary-path> [<backup-path-list>]]
```

Where:

```
<primary-path>::=<path>

<backup-path-list>::=<path>[<backup-path-list>]

<path>::=<ERO><attribute-list>
```

Where:

```
<attribute-list> ::= [<LSPA>]
                     [<BANDWIDTH>]
                     [<metric-list>]
                     [<IRO>]

<metric-list> ::= <METRIC>[<metric-list>]
```

There is one mandatory object that MUST be included within each LSP Update Request in the PCUpd message: the LSP object (see Section 7.2). If the LSP object is missing, the receiving PCE MUST send a PCErr message with Error-type=6 (Mandatory Object missing) and Error-value=[TBD] (LSP object missing).

The LSP State Report MUST contain a path descriptor for the primary path, and MAY contain one or more path descriptors for backup paths, if MPLS TE Global Default Restoration or MPLS TE Global Path Protection is desired on the LSP. A path descriptor MUST contain an ERO object, and MAY contain the LSPA, BANDWIDTH, IRO, and METRIC objects. The ERO, LSPA, BANDWIDTH, METRIC, and IRO objects are defined in [RFC5440].

Each LSP Update Request results in a separate LSP setup operation at a PCC. An LSP Update Request MUST contain all LSP parameters that a PCC wishes to set for the LSP. A PCC MAY set missing parameters from

locally configured defaults. If the LSP specified the Update Request is already up, it will be torn down and re-signaled. The PCC will use make-before-break whenever possible in the re-signaling operation.

A PCC MUST respond with an LSP State Report to each LSP Update Request to indicate the resulting state of the LSP in the network. A PCC MAY respond with multiple LSP State Reports to report LSP setup progress of a single LSP.

If the rate of PCUpd messages sent to a PCC for the same target LSP exceeds the rate at which the PCC can signal LSPs into the network, the PCC MAY perform state compression and only re-signal the last modification in its queue.

Note that a PCC MUST process all LSP Update Requests - for example, an LSP Update Request is sent when a PCE returns delegation or puts an LSP into non-operational state. The protocol relies on TCP for message-level flow control.

Note also that it's up to the PCE to handle inter-LSP dependencies; for example, if ordering of LSP set-ups is required, the PCE has to wait for an LSP State Report for a previous LSP before triggering the LSP setup of a next LSP.

7. Object Formats

The PCEP objects defined in this document are compliant with the PCEP object format defined in [RFC5440]. The P flag and the I flag of the PCEP objects defined in this document MUST always be set to 0 on transmission and MUST be ignored on receipt since these flags are exclusively related to path computation requests.

7.1. OPEN Object

This document defines a new optional TLV for the OPEN Object to support stateful PCE capability negotiation.

7.1.1. Stateful PCE Capability TLV

The format of the Stateful PCE Capability TLV is shown in the following figure:

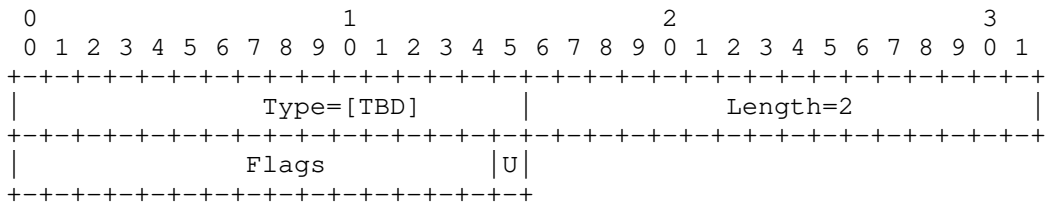


Figure 11: The Stateful PCE Capability TLV format

The type of the TLV is [TBD] and it has a fixed length of 2 octets.

The value comprises a single field - Flags (16 bits):

U (LSP Update Capability - 1 bit): if set to 1 by a PCC, the U Flag indicates that the PCC allows modification of LSP parameters; if set to 1 by a PCE, the U Flag indicates that the PCE wishes to update LSP parameters. The LSP Update capability must be advertised by both a PCC and a PCE for PCUpd messages to be allowed on a PCEP session.

Unassigned bits are considered reserved. They MUST be set to 0 on transmission and MUST be ignored on receipt.

7.2. LSP Object

The LSP object MUST be present within PCRpt and PCUpd messages. The LSP object MAY be carried within PCReq and PCRep messages if the stateful PCE capability has been negotiated on the session. The LSP object contains a set of fields used to specify the target LSP, the operation to be performed on the LSP, and LSP Delegation. It is also contains a flag to indicate to a PCE that the initial LSP state synchronization has been done.

LSP Object-Class is [TBD].

LSP Object-Type is 1.

The format of the LSP object body is shown in Figure 12:

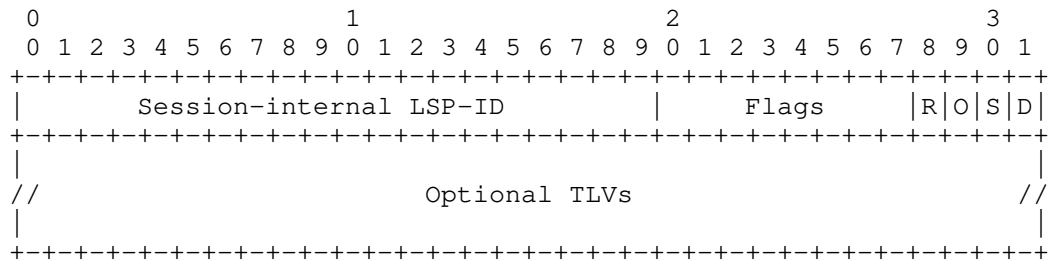


Figure 12: The LSP Object TLV format

The LSP object body has a variable length and may contain additional TLVs.

Session-internal LSP-ID (20 bits): Per-PCEP session identifier for an LSP. In each PCEP session the PCC creates a unique LSP-ID for each LSP that will remain constant for the duration of the session. The mapping of the LSP Symbolic Name to LSP-ID is communicated to the PCE by sending a PCRpt message containing the 'LSP Symbolic Name' TLV. All subsequent PCEP messages then address the LSP by its Session-internal LSP-ID.

Flags (12 bits):

D (Delegate - 1 bit): on a PCRpt message, the D Flag set to 1 indicates that the PCC is delegating the LSP the PCE. On a PCUpd message, the D flag set to 1 indicates that the PCE is confirming the LSP Delegation. To keep an LSP delegated to the PCE, the PCC must set the D flag to 1 on each PCRpt message for the duration of the delegation - the first PCRpt with the D flag set to 0 revokes the delegation. To keep the delegation, the PCE must set the D flag to 1 on each PCUpd message for the duration of the delegation - the first PCUpd with the D flag set to 0 returns the delegation.

S (Sync Done- 1 bit): the S Flag MUST be set to 1 on the LSP State Report for the last LSP in the synchronized set during State Synchronization. The S Flag MUST be set to 0 otherwise.

O (Operational - 1 bit): On PCRpt messages the O Flag indicates the LSP status. Value of '1' means that the LSP is operational, i.e. it is either being signaled or it is active. Value of '0' means that the LSP is not operational, i.e it is de-routed and the PCC is not attempting to set it up. On PCUpd messages the flag indicates the desired status for the LSP. Value of '1' means that the desired LSP state is operational, value of '0' means that the target LSP should be non-operational. Setting the LSP status from the PCE SHALL NOT override the operator: if a pce-controlled LSP

has been configured to be non-operational, setting the LSP's status to '1' from an PCE will not make it operational.

R (Remove - 1 bit): On PCRpt messages the R Flag indicates that the LSP has been removed from the PCC. Upon receiving an LSP State Update with the R Flag set to 1, the PCE SHOULD remove all state related to the LSP from its database.

Unassigned bits are considered reserved. They MUST be set to 0 on transmission and MUST be ignored on receipt.

TLVs that are currently defined for the LSP Object are described in the following sections.

7.2.1. The LSP Symbolic Name TLV

Each LSP MUST have a symbolic name that is unique in the PCC. The LSP Symbolic Name MUST remain constant throughout an LSP's lifetime, which may span across multiple consecutive PCEP sessions and/or PCC restarts. The LSP Symbolic Name MAY be specified by an operator in a PCC's CLI configuration. If the operator does not specify a Symbolic Name for an LSP, the PCC MUST auto-generate one.

The LSP Symbolic Name TLV MUST be included in the LSP State Report when during a given PCEP session an LSP is first reported to a PCE. A PCC sends to a PCE the first LSP State Report either during State Synchronization, or when a new LSP is configured at the PCC. LSP State Report MAY be included in subsequent LSP State Reports for the LSP.

The format of the LSP Symbolic Name TLV is shown in the following figure:

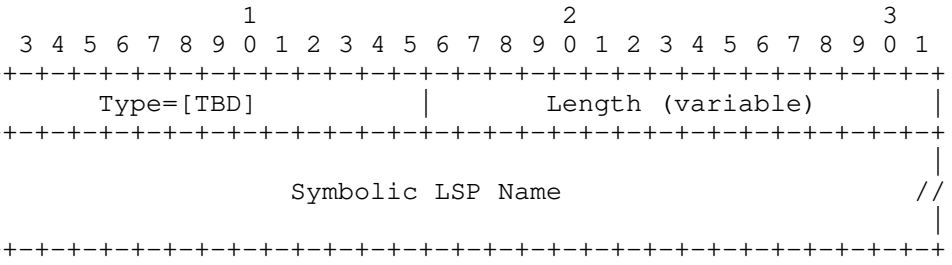


Figure 13: LSP symbolic name TLV format

The type of the TLV is [TBD] and it has a variable length, which MUST be greater than 0.

7.2.2. LSP Identifiers TLVs

Whenever the value of an LSP identifier changes, a PCC MUST send out an LSP State Report, where the LSP Object carries the LSP Identifiers TLV that contains the new value. The LSP Identifiers TLV MUST also be included in the LSP object during state synchronization. There are two LSP Identifiers TLVs, one for IPv4 and one for IPv6.

The format of the IPv4 LSP Identifiers TLV is shown in the following figure:

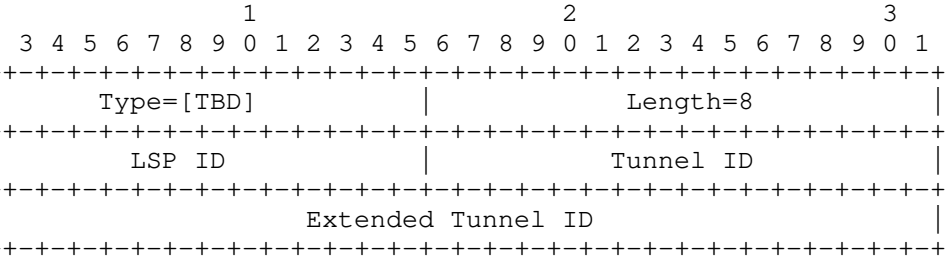


Figure 14: IPv4 LSP Identifiers TLV format

The type of the TLV is [TBD] and it has a fixed length of 8 octets. The value contains two fields:

LSP ID: contains the 16-bit 'LSP ID' identifier defined in [RFC3209], Section 4.6.2.1 for the LSP_TUNNEL_IPv4 Sender Template Object.

Tunnel ID: contains the 16-bit 'Tunnel ID' identifier defined in [RFC3209], Section 4.6.1.1 for the LSP_TUNNEL_IPv4 Session Object. Tunnel ID remains constant over the life time of a tunnel. However, when Global Path Protection or Global Default Restoration is used, both the primary and secondary LSPs have their own Tunnel IDs. A PCC will report a change in Tunnel ID when traffic switches over from primary LSP to secondary LSP (or vice versa).

Extended Tunnel ID: contains the 32-bit 'Extended Tunnel ID' identifier defined in [RFC3209], Section 4.6.1.1 for the LSP_TUNNEL_IPv4 Session Object.

The format of the IPv6 LSP Identifiers TLV is shown in the following figure:

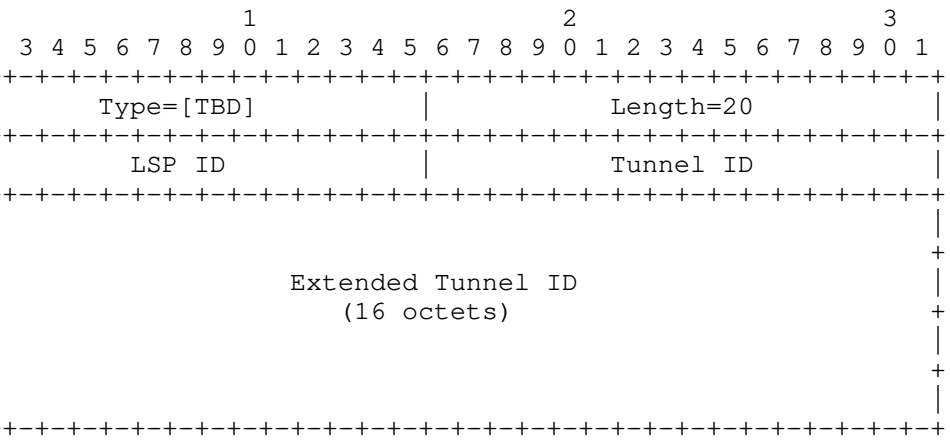


Figure 15: IPv6 LSP Identifiers TLV format

The type of the TLV is [TBD] and it has a fixed length of 20 octets. The value contains two fields:

LSP ID: contains the 16-bit 'LSP ID' identifier defined in [RFC3209], Section 4.6.2.2 for the LSP_TUNNEL_IPv6 Sender Template Object.

Tunnel ID: contains the 16-bit 'Tunnel ID' identifier defined in [RFC3209], Section 4.6.1.2 for the LSP_TUNNEL_IPv6 Session Object. Tunnel ID remains constant over the life time of a tunnel. However, when Global Path Protection or Global Default Restoration is used, both the primary and secondary LSPs have their own Tunnel IDs. A PCC will report a change in Tunnel ID when traffic switches over from primary LSP to secondary LSP (or vice versa).

Extended Tunnel ID: contains the 32-bit 'Extended Tunnel ID' identifier defined in [RFC3209], Section 4.6.1.2 for the LSP_TUNNEL_IPv6 Session Object.

7.2.3. LSP Update Error Code TLV

If an LSP Update Request failed, an LSP State Report MUST be sent to all connected stateful PCEs. LSP State Report MUST contain the LSP Update Error Code TLV, indicating the cause of the failure.

The format of the LSP Update Error Code TLV is shown in the following figure:

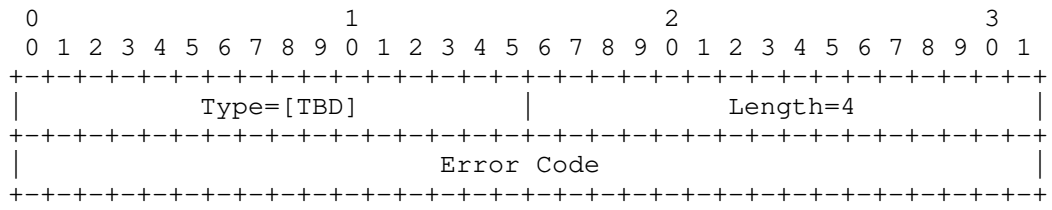


Figure 16: LSP Update Error Code TLV format

The type of the TLV is [TBD] and it has a fixed length of 4 octets. The value contains the error code that indicates the cause of the LSP setup failure. Error codes will be defined in a later revision of this document.

7.2.4. RSVP ERROR_SPEC TLVs

If the set up of an LSP failed at a downstream node which returned an ERROR_SPEC to the PCC, the ERROR_SPEC MUST be included in the LSP State Report. Depending on whether RSVP signaling was performed over IPv4 or IPv6, the LSP Object will contain an IPv4 ERROR_SPEC TLV or an IPv6 ERROR_SPEC TLV.

The format of the IPv4 ERROR_SPEC TLV is shown in the following figure:

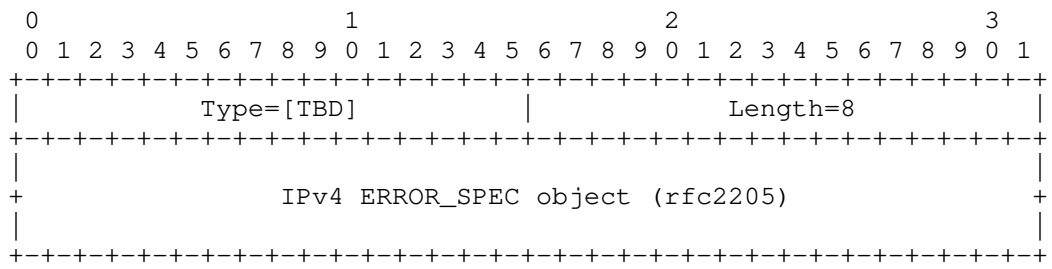


Figure 17: The IPv4 ERROR_SPEC TLV format

The type of the TLV is [TBD] and it has a fixed length of 8 octets. The value contains the RSVP IPv4 ERROR_SPEC object defined in [RFC2205]. Error codes allowed in the ERROR_SPEC object are defined in [RFC2205] and [RFC3209].

The format of the IPv4 ERROR_SPEC TLV is shown in the following figure:

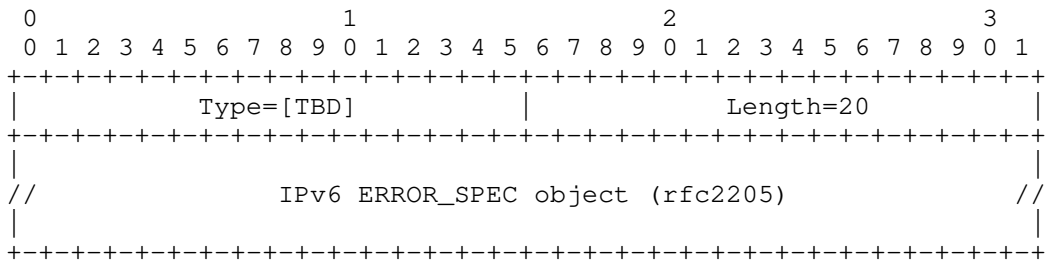


Figure 18: The IPv6 ERROR_SPEC TLV format

The type of the TLV is [TBD] and it has a fixed length of 20 octets. The value contains the RSVP IPv6 ERROR_SPEC object defined in [RFC2205]. Error codes allowed in the ERROR_SPEC object are defined in [RFC2205] and [RFC3209].

7.2.5. Delegation Parameters TLVs

Multiple delegation parameters, such as sub-delegation permissions, authentication parameters, etc. need to be communicated from a PCC to a PCE during the delegation operation. Delegation parameters will be carried in multiple delegation parameter TLVs, which will be defined in future revisions of this document.

7.3. PCEP-Error Object

New error types and values will be defined, among others, for the following errors:

PCRpt Error: encountered an error with the PCRpt message during synchronization; type 10, value 2 (need to double check), and need to add the offending message

LSP not delegated: type tbd, value tbd and need to include the LSP id or the LSP name

Delegation not negotiated: generated on receipt of an PCUpd when the U flag was not set) type tbd, value tbd.

A complete list of new error types will be specified in a later revision of this draft.

8. IANA Considerations

A future revision of this document will request IANA actions to allocate code points for the protocol elements that have been

defined..

9. Manageability Considerations

All manageability requirements and considerations listed in [RFC5440] apply to PCEP protocol extensions defined in this document. In addition, requirements and considerations listed in this section apply.

9.1. Control Function and Policy

In addition to configuring specific PCEP session parameters, as specified in [RFC5440], Section 8.1, a PCE or PCC implementation MUST allow configuring the stateful PCEP capability and the LSP Update capability. A PCC implementation SHOULD allow the operator to specify multiple candidate PCEs for and a delegation preference for each candidate PCE. A PCC SHOULD allow the operator to specify an LSP delegation policy where LSPs are delegated to the most-preferred online PCE. A PCC MAY allow the operator to specify different LSP delegation policies.

A PCC implementation which allows concurrent connections to multiple PCEs SHOULD allow the operator to group the PCEs by administrative domains and it MUST NOT advertise LSP existence and state to a PCE if the LSP is delegated to a PCE in a different group.

A PCC implementation SHOULD allow the operator to specify whether the PCC will advertise LSP existence and state for LSPs that are not controlled by any PCE (for example, LSPs that are statically configured at the PCC).

A PCC implementation SHOULD allow the operator to specify the Delegation Timeout Interval. The default value of the Delegation Timeout Interval SHOULD be set to 30 seconds.

When an LSP can no longer be delegated to a PCE, after the expiration of the Delegation Timeout Interval, the LSP MAY either: 1) retain its current parameters or 2) revert to operator-defined default LSP parameters. This behavior SHOULD be configurable and in the case when (2) is supported, a PCC implementation MUST allow the operator to specify the default LSP parameters.

A PCC implementation SHOULD allow the operator to specify delegation priority for PCEs. This effectively defines the primary PCE and one or more backup PCEs to which primary PCE's LSPs can be delegated when the primary PCE fails.

Policies defined for stateful PCEs and PCCs should eventually fit in the Policy-Enabled Path Computation Framework defined in [RFC5394], and the framework should be extended to support Stateful PCEs.

9.2. Information and Data Models

PCEP session configuration and information in the PCEP MIB module SHOULD be extended to include negotiated stateful capabilities, synchronization status, and delegation status (at the PCC list PCEs with delegated LSPs).

9.3. Liveness Detection and Monitoring

PCEP protocol extensions defined in this document do not require any new mechanisms beyond those already defined in [RFC5440], Section 8.3.

9.4. Verifying Correct Operation

Mechanisms defined in [RFC5440], Section 8.4 also apply to PCEP protocol extensions defined in this document. In addition to monitoring parameters defined in [RFC5440], a stateful PCC-side PCEP implementation SHOULD provide the following parameters:

- Total number of LSP updates
- Number of successful LSP updates
- Number of dropped LSP updates
- Number of LSP updates where LSP setup failed

A PCC implementation SHOULD provide a command to show to which PCEs LSPs are delegated.

A PCC implementation SHOULD allow the operator to manually revoke LSP delegation.

9.5. Requirements on Other Protocols and Functional Components

PCEP protocol extensions defined in this document do not put new requirements on other protocols.

9.6. Impact on Network Operation

Mechanisms defined in [RFC5440], Section 8.6 also apply to PCEP protocol extensions defined in this document.

Additionally, a PCEP implementation SHOULD allow a limit to be placed on the rate PCUpd and PCRpt messages sent by a PCEP speaker and processed from a peer. It SHOULD also allow sending a notification when a rate threshold is reached.

A PCC implementation SHOULD allow a limit to be placed on the rate of LSP Updates to the same LSP to avoid signaling overload discussed in Section 10.3.

10. Security Considerations

10.1. Vulnerability

This document defines extensions to PCEP to enable stateful PCEs. The nature of these extensions and the delegation of path control to PCEs results in more information being available for a hypothetical adversary and a number of additional attack surfaces which must be protected.

The security provisions described in [RFC5440] remain applicable to these extensions. However, because the protocol modifications outlined in this document allow the PCE to control path computation timing and sequence, the PCE defense mechanisms described in [RFC5440] section 7.2 are also now applicable to PCC security.

As a general precaution, it is RECOMMENDED that these PCEP extensions only be activated on authenticated and encrypted sessions across PCEs and PCCs belonging to the same administrative authority.

The following sections identify specific security concerns that may result from the PCEP extensions outlined in this document along with recommended mechanisms to protect PCEP infrastructure against related attacks.

10.2. LSP State Snooping

The stateful nature of this extension explicitly requires LSP status updates to be sent from PCC to PCE. While this gives the PCE the ability to provide more optimal computations to the PCC, it also provides an adversary with the opportunity to eavesdrop on decisions made by network systems external to PCE. This is especially true if the PCC delegates LSPs to multiple PCEs simultaneously.

Adversaries may gain access to this information by eavesdropping on unsecured PCEP sessions, and might then use this information in various ways to target or optimize attacks on network infrastructure. For example by flexibly countering anti-DDoS measures being taken to

protect the network, or by determining choke points in the network where the greatest harm might be caused.

PCC implementations which allow concurrent connections to multiple PCEs SHOULD allow the operator to group the PCEs by administrative domains and they MUST NOT advertise LSP existence and state to a PCE if the LSP is delegated to a PCE in a different group.

10.3. Malicious PCE

The LSP delegation mechanism described in this document allows a PCC to grant effective control of an LSP to the PCE for the duration of a PCEP session. While this enables PCE control of the timing and sequence of path computations within and across PCEP sessions, it also introduces a new attack vector: an attacker may flood the PCC with PCUpd messages at a rate which exceeds either the PCC's ability to process them or the network's ability to signal the changes, either by spoofing messages or by compromising the PCE itself.

A PCC is free to revoke an LSP delegation at any time without needing any justification. A defending PCC can do this by enqueueing the appropriate PCRpt message. As soon as that message is enqueued in the session, the PCC is free to drop any incoming PCUpd messages without additional processing.

10.4. Malicious PCC

A stateful session also result in increased attack surface by placing a requirement for the PCE to keep an LSP state replica for each PCC. It is RECOMMENDED that PCE implementations provide a limit on resources a single PCC can occupy.

Delegation of LSPs can create further strain on PCE resources and a PCE implementation MAY preemptively give back delegations if it finds itself lacking the resources needed to effectively manage the delegation. Since the delegation state is ultimately controlled by the PCC, PCE implementations SHOULD provide throttling mechanisms to prevent strain created by flaps of either a PCEP session or an LSP delegation.

11. Acknowledgements

We would like to thank Adrian Farrel and Ina Minei for their contributions to this document.

We would like to thank Shane Asante, Julien Meuric, Kohei Shiomoto, Paul Schultz and Raveendra Torvi for their helpful comments.

12. References

12.1. Normative References

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

[RFC2205] Braden, B., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, 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.

[RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.

[RFC5088] Le Roux, JL., Vasseur, JP., Ikejiri, Y., and R. Zhang, "OSPF Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5088, January 2008.

[RFC5089] Le Roux, JL., Vasseur, JP., Ikejiri, Y., and R. Zhang, "IS-IS Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5089, January 2008.

[RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.

[RFC5511] Farrel, A., "Routing Backus-Naur Form (RBNF): A Syntax Used to Form Encoding Rules in Various Routing Protocol Specifications", RFC 5511, April 2009.

12.2. Informative References

[MPLS-PC] Chaieb, I., Le Roux, JL., and B. Cousin, "Improved MPLS-TE LSP Path Computation using Preemption", Global Information Infrastructure Symposium, July 2007.

[MXMN-TE] Danna, E., Mandal, S., and A. Singh, "Practical linear programming algorithm for balancing the max-min fairness and throughput objectives in traffic engineering", pre-print, 2011.

[NET-REC] Vasseur, JP., Pickavet, M., and P. Demeester, "Network Recovery: Protection and Restoration of Optical, SONET-

SDH, IP, and MPLS", The Morgan Kaufmann Series in Networking, June 2004.

[RFC2702] Awduche, D., Malcolm, J., Agogbua, J., O'Dell, M., and J. McManus, "Requirements for Traffic Engineering Over MPLS", RFC 2702, September 1999.

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

[RFC3346] Boyle, J., Gill, V., Hannan, A., Cooper, D., Awduche, D., Christian, B., and W. Lai, "Applicability Statement for Traffic Engineering with MPLS", RFC 3346, August 2002.

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

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

[RFC4657] Ash, J. and J. Le Roux, "Path Computation Element (PCE) Communication Protocol Generic Requirements", RFC 4657, September 2006.

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

[RFC5394] Bryskin, I., Papadimitriou, D., Berger, L., and J. Ash, "Policy-Enabled Path Computation Framework", RFC 5394, December 2008.

[RFC5557] Lee, Y., Le Roux, JL., King, D., and E. Oki, "Path Computation Element Communication Protocol (PCEP) Requirements and Protocol Extensions in Support of Global Concurrent Optimization", RFC 5557, July 2009.

Authors' Addresses

Edward Crabbe
Google, Inc.
1600 Amphitheatre Parkway
Mountain View, CA 94043
US

Email: edc@google.com

Jan Medved
Juniper Networks, Inc.
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: jmedved@juniper.net

Robert Varga
Juniper Networks, Inc.
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: rvarga@juniper.net

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: July 25, 2012

E. Crabbe
Google, Inc.
J. Medved
Cisco Systems, Inc.
R. Varga
Pantheon Technologies LLC
I. Minei
Juniper Networks, Inc.

January 22, 2012

PCEP Extensions for Stateful PCE
draft-crabbe-pce-stateful-pce-02

Abstract

The Path Computation Element Communication Protocol (PCEP) provides mechanisms for Path Computation Elements (PCEs) to perform path computations in response to Path Computation Clients (PCCs) requests.

Although PCEP explicitly makes no assumptions regarding the information available to the PCE, it also makes no provisions for synchronization or PCE control of timing and sequence of path computations within and across PCEP sessions. This document describes a set of extensions to PCEP to enable this functionality, providing stateful control of Multiprotocol Label Switching (MPLS) Traffic Engineering Label Switched Paths (TE LSP) via PCEP.

Requirements Language

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

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 July 25, 2012.

Table of Contents

1. Introduction

[RFC5440] describes the Path Computation Element Protocol (PCEP. PCEP defines the communication between a Path Computation Client (PCC) and a Path Control Element (PCE), or between PCE and PCE, enabling computation of Multiprotocol Label Switching (MPLS) for Traffic Engineering Label Switched Path (TE LSP) characteristics

This document specifies a set of extensions to PCEP to enable stateful control of TE LSPs between and across PCEP sessions in compliance with [RFC4657]. It includes mechanisms to effect LSP state synchronization between PCCs and PCEs, delegation of control of LSPs to PCEs, and PCE control of timing and sequence of path computations within and across PCEP sessions.

2. Terminology

This document uses the following terms defined in [RFC5440]: PCC, PCE, PCEP Peer.

This document uses the following terms defined in [RFC4090]: MPLS TE Fast Reroute (FRR), FRR One-to-One Backup, FRR Facility Backup.

The following terms are defined in this document:

Passive Stateful PCE: uses LSP state information learned from PCCs to optimize path computations. It does not actively update LSP state. A PCC maintains synchronization with the PCE.

Active Stateful PCE: uses LSP state information learned from PCCs to optimize path computations. Additionally, it actively updates LSP parameters in those PCCs that delegated control over their LSPs to the PCE.

Delegation: An operation to grant a PCE temporary rights to modify a subset of LSPs parameters on one or more PCC's LSPs. LSPs are delegated from a PCC to a PCE.

Delegation Timeout Interval: when a PCEP session is terminated, a PCC waits for this time period before revoking LSP delegation to a PCE.

LSP State Report: an operation to send LSP state (Operational / Admin Status, LSP attributes configured and set by a PCE, etc.) from a PCC to a PCE.

LSP Update Request: an operation where a PCE requests a PCC to update one or more attributes of an LSP and to re-signal the LSP with updated attributes.

LSP Priority: a specific pair of MPLS setup and hold priority values.

LSP State Database: information about and attributes of all LSPs that are being reported to one or more PCEs via LSP State Reports.

Minimum Cut Set: the minimum set of links for a specific source destination pair which, when removed from the network, result in a specific source being completely isolated from specific destination. The summed capacity of these links is equivalent to the maximum capacity from the source to the destination by the max-flow min-cut theorem.

MPLS TE Global Default Restoration: once an LSP failure is detected by some downstream mode, the head-end LSP is notified by means of RSVP. Upon receiving the notification, the headend Label Switching Router (LSR) recomputes the path and signals the LSP along an alternate path. [NET-REC]

MPLS TE Global Path Protection: once an LSP failure is detected by some downstream mode, the head-end LSP is notified by means of RSVP. Upon receiving the notification, the headend LSR reroutes traffic using a pre-signaled backup (secondary) LSP. [NET-REC].

Within this document, when describing PCE-PCE communications, the requesting PCE fills the role of a PCC. This provides a saving in documentation without loss of function.

The message formats in this document are specified using Routing Backus-Naur Format (RBNF) encoding as specified in [RFC5511].

3. Motivation and Objectives

3.1. Motivation

3.1.1. Background

Traffic engineering has been a goal of the MPLS architecture since its inception ([RFC3031], [RFC2702], [RFC3346]). In the traffic engineering system provided by [RFC3630], [RFC5305], and [RFC3209] information about network resources utilization is only available as total reserved capacity by traffic class on a per interface basis; individual LSP state is available only locally on each LER for it’s

own LSPs. In most cases, this makes good sense, as distribution and retention of total LSP state for all LERs within in the network would be prohibitively costly.

Unfortunately, this visibility in terms of global LSP state may result in a number of issues for some demand patterns, particularly within a common setup and hold priority. This issue affects online traffic engineering systems, and in particular, the widely implemented but seldom deployed auto-bandwidth system.

A sufficiently over-provisioned system will by definition have no issues routing its demand on the shortest path. However, lowering the degree to which network over-provisioning is required in order to run a healthy, functioning network is a clear and explicit promise of MPLS architecture. In particular, it has been a goal of MPLS to provide mechanisms to alleviate congestion scenarios in which "traffic streams are inefficiently mapped onto available resources; causing subsets of network resources to become over-utilized while others remain underutilized" ([RFC2702]).

3.1.2. Why a Stateful PCE?

[RFC4655] defines a stateful PCE to be one in which the PCE maintains "strict synchronization between the PCE and not only the network states (in term of topology and resource information), but also the set of computed paths and reserved resources in use in the network." [RFC4655] also expressed a number of concerns with regard to a stateful PCE, specifically:

- Any reliable synchronization mechanism would result in significant control plane overhead
- Out-of-band ted synchronization would be complex and prone to race conditions
- Path calculations incorporating total network state would be highly complex

In general, stress on the MPLS TE control plane will be directly proportional to the size of the system being controlled and the and the tightness of the control loop, and indirectly proportional to the amount of over-provisioning in terms of both network capacity and reservation overhead.

Despite these concerns in terms of implementation complexity and scalability, several TE algorithms exist today that have been demonstrated to be extremely effective in large TE systems, providing both rapid convergence and significant benefits in terms of

optimality of resource usage [MXMN-TE]. All of these systems share at least two common characteristics: the requirement for both global visibility of a flow (or in this case, a TE LSP) state and for ordered control of path reservations across devices within the system being controlled. While some approaches have been suggested in order to remove the requirements for ordered control (See [MPLS-PC]), these approaches are highly dependent on traffic distribution, and do not allow for multiple simultaneous LSP priorities representing diffserv classes.

The following use cases demonstrate a need for visibility into global inter-PCC LSP state in PCE path computations, and for a PCE control of sequence and timing in altering LSP path characteristics within and across PCEP sessions. Reference topologies for the use cases described later in this section are shown in Figures 1 and 2.

Unless otherwise cited, use cases assume that all LSPs listed exist at the same LSP priority.

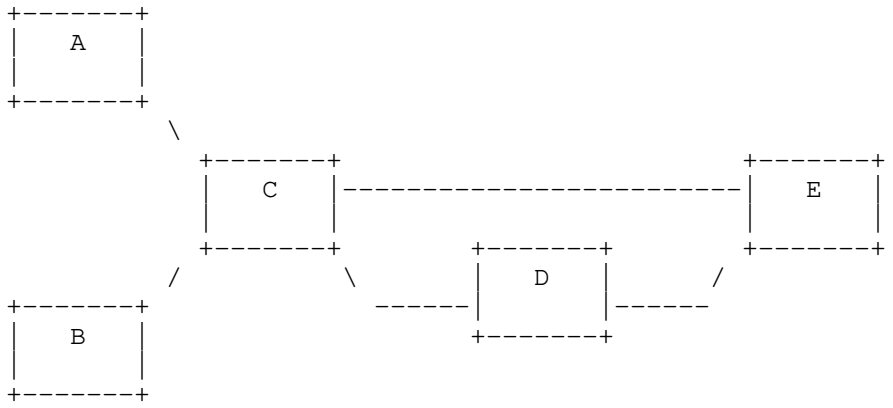


Figure 1: Reference topology 1

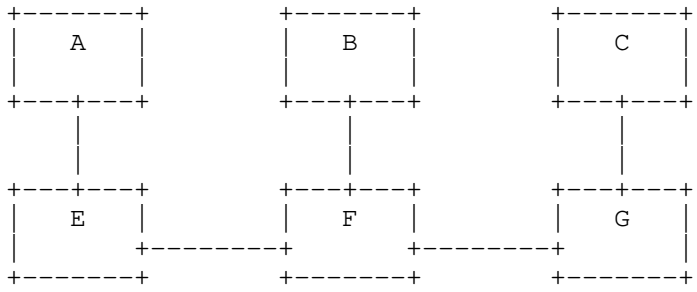


Figure 2: Reference topology 2

3.1.2.1. Throughput Maximization and Bin Packing

Because LSP attribute changes in [RFC5440] are driven by PCReq messages under control of a PCC's local timers, the sequence of RSVP reservation arrivals occurring in the network will be randomized. This, coupled with a lack of global LSP state visibility on the part of a stateless PCE may result in suboptimal throughput in a given network topology.

Reference topology 2 in Figure 2 and Tables 1 and 2 show an example in which throughput is at 50% of optimal as a result of lack of visibility and synchronized control across PCC's. In this scenario, the decision must be made as to whether to route any portion of the E-G demand, as any demand routed for this source and destination will decrease system throughput.

Link	Metric	Capacity
A-E	1	10
B-F	1	10
C-G	1	10
E-F	1	10
F-G	1	10

Table 1: Link parameters for Throughput use case

Time	LSP	Src	Dst	Demand	Routable	Path
1	1	E	G	10	Yes	E-F-G
2	2	A	B	10	No	---
3	1	F	C	10	No	---

Table 2: Throughput use case demand time series

In many cases throughput maximization becomes a bin packing problem. While bin packing itself is an NP-hard problem, a number of common heuristics which run in polynomial time can provide significant improvements in throughput over random reservation event distribution, especially when traversing links which are members of the minimum cut set for a large subset of source destination pairs.

Tables 3 and 4 show a simple use case using Reference Topology 1 in

Figure 1, where LSP state visibility and control of reservation order across PCCs would result in significant improvement in total throughput.

Link	Metric	Capacity
A-C	1	10
B-C	1	10
C-E	10	5
C-D	1	10
D-E	1	10

Table 3: Link parameters for Bin Packing use case

Time	LSP	Src	Dst	Demand	Routable	Path
1	1	A	E	5	Yes	A-C-D-E
2	2	B	E	10	No	---

Table 4: Bin Packing use case demand time series

3.1.2.2. Deadlock

Most existing RSVP-TE implementations will not tear down established LSPs in the event of the failure of the bandwidth increase procedure detailed in [RFC3209]. This behavior is directly implied to be correct in [RFC3209] and is often desirable from an operator's perspective, because either a) the destination prefixes are not reachable via any means other than MPLS or b) this would result in significant packet loss as demand is shifted to other LSPs in the overlay mesh.

In addition, there are currently few implementations offering ingress admission control at the LSP level. Again, having ingress admission control on a per LSP basis is not necessarily desirable from an operational perspective, as a) one must over-provision tunnels significantly in order to avoid deleterious effects resulting from stacked transport and flow control systems and b) there is currently no efficient commonly available northbound interface for dynamic configuration of per LSP ingress admission control (such an interface could easily be defined using the extensions present in this spec, but it beyond the scope of the current document).

Lack of ingress admission control coupled with the behavior in

[RFC3209] effectively results in mis-signaled LSPs during periods of contention for network capacity between LSPs in a given LSP priority. This in turn causes information loss in the TED with regard to actual network state, resulting in LSPs sharing common network interfaces with mis-signaled LSPs operating in a degraded state for significant periods of time, even when unused network capacity may potentially be available.

Reference Topology 1 in Figure 1 and Tables 5 and 6 show a use case that demonstrates this behavior. Two LSPs, LSP 1 and LSP 2 are signaled with demand 2 and routed along paths A-C-D-E and B-C-D-E respectively. At a later time, the demand of LSP 1 increases to 20. Under such a demand, the LSP cannot be resignaled. However, the existing LSP will not be torn down. In the absence of ingress policing, traffic on LSP 1 will cause degradation for traffic of LSP 2 (due to oversubscription on the links C-D and D-E), as well as information loss in the TED with regard to the actual network state.

The problem could be easily ameliorated by global visibility of LSP state coupled with PCC- external demand measurements and placement of two LSPs on disjoint links. Note that while the demand of 20 for LSP 1 could never be satisfied in the given topology, what could be achieved would be isolation from the ill-effects of the (unsatisfiable) increased demand.

Link	Metric	Capacity
A-C	1	10
B-C	1	10
C-E	10	5
C-D	1	10
D-E	1	10

Table 5: Link parameters for the 'Deadlock' example

Time	LSP	Src	Dst	Demand	Routable	Path
1	1	A	E	2	Yes	A-C-D-E
2	2	B	E	2	Yes	B-C-D-E
3	1	A	E	20	No	---

Table 6: Deadlock LSP and demand time series

3.1.2.3. Minimum Perturbation

As a result of both the lack of visibility into global LSP state and the lack of control over event ordering across PCE sessions, unnecessary perturbations may be introduced into the network by a stateless PCE. Tables 7 and 8 show an example of an unnecessary network perturbation using Reference Topology 1 in Figure 1. In this case an unimportant (high LSP priority value) LSP (LSP1) is first set up along the shortest path. At time 2, which is assumed to be relatively close to time 1, a second more important (lower LSP-priority value) LSP is established, preempting LSP 1 and shifting it to the longer A-C-E path.

Link	Metric	Capacity
A-C	1	10
B-C	1	10
C-E	10	10
C-D	1	10
D-E	1	10

Table 7: Link parameters for the 'Minimum-Perturbation' example

Time	LSP	Src	Dst	Demand	LSP Prio	Routable	Path
1	1	A	E	7	7	Yes	A-C-D-E
2	2	B	E	7	0	Yes	B-C-D-E
3	1	A	E	7	7	Yes	A-C-E

Table 8: Minimum-Perturbation LSP and demand time series

3.1.2.4. Predictability

Randomization of reservation events caused by lack of control over event ordering across PCE sessions results in poor predictability in LSP routing. An offline system applying a consistent optimization method will produce predictable results to within either the boundary of forecast error when reservations are over-provisioned by reasonable margins or to the variability of the signal and the forecast error when applying some hysteresis in order to minimize churn.

Reference Topology 1 and Tables 9, 10 and 11 show the impact of event ordering and predictability of LSP routing.

Link	Metric	Capacity
A-C	1	10
B-C	1	10
C-E	1	10
C-D	1	10
D-E	1	10

Table 9: Link parameters for the 'Predictability' example

Time	LSP	Src	Dst	Demand	Routable	Path
1	1	A	E	7	Yes	A-C-E
2	2	B	E	7	Yes	B-C-D-E

Table 10: Predictability LSP and demand time series 1

Time	LSP	Src	Dst	Demand	Routable	Path
1	2	B	E	7	Yes	B-C-E
2	1	A	E	7	Yes	A-C-D-E

Table 11: Predictability LSP and demand time series 2

3.1.2.5. Global Concurrent Optimization

Global Concurrent Optimization (GCO) defined in [RFC5557] is a network optimization mechanism that is able to simultaneously consider the entire topology of the network and the complete set of existing TE LSPs and their existing constraints, and look to optimize or reoptimize the entire network to satisfy all constraints for all TE LSPs. It allows for bulk path computations in order to avoid blocking problems and to achieve more optimal network-wide solutions.

Global control of LSP operation sequence in [RFC5557] is predicated on the use of what is effectively a stateful (or semi-stateful) NMS. The NMS can be either not local to the switch, in which case another northbound interface is required for LSP attribute changes, or local/ collocated, in which case there are significant issues with efficiency in resource usage. Stateful PCE adds a few features that:

- Roll the NMS visibility into the PCE and remove the requirement for an additional northbound interface
- Allow the PCE to determine when re-optimization is needed
- Allow the PCE to determine which LSPs should be re-optimized
- Allow a PCE to control the sequence of events across multiple PCCs, allowing for bulk (and truly global) optimization, LSP shuffling etc.

3.1.3. Protocol vs. Configuration

Note that existing configuration tools and protocols can be used to set LSP state. However, this solution has several shortcomings:

- Scale & Performance: configuration operations often require processing of additional configuration portions beyond the state being directly acted upon, with corresponding cost in CPU cycles, negatively impacting both PCC stability LSP update rate capacity.
- Scale & Performance: configuration operations often have transactional semantics which are typically heavyweight and require additional CPU cycles, negatively impacting PCC update rate capacity.
- Security: opening up a configuration channel to a PCE would allow a malicious PCE to take over a PCC. The PCEP extensions described in this document only allow a PCE control over a very limited set of LSP attributes.
- Interoperability: each vendor has a proprietary information model for configuring LSP state, which prevents interoperability of a PCE with PCCs from different vendors. The PCEP extensions described in this document allow for a common information model for LSP state for all vendors.
- Efficient State Synchronization: configuration channels may be heavyweight and unidirectional, therefore efficient state synchronization between a PCE and a PCE may be a problem.

3.2. Objectives

The objectives for the protocol extensions to support stateful PCE described in this document are as follows:

- Allow a single PCC to interact with a mix of stateless and stateful PCEs simultaneously using the same PCEP.

- o Support efficient LSP state synchronization between the PCC and one or more active or passive stateful PCEs.

- o Allow a PCC to delegate control of its LSPs to an active stateful PCE such that a single LSP is under the control a single PCE at any given time. A PCC may revoke this delegation at any point during the lifetime of the PCEP session. A PCE may return this delegation at any point during the lifetime of the PCEP session.

- o Allow a PCE to control computation timing and update timing across all LSPs that have been delegated to it.

- o Allow a PCE to specify protection / restoration settings for all LSPs that have been delegated to it.

- o Enable uninterrupted operation of PCC's LSPs in the event PCE failure or while control of LSPs is being transferred between PCEs.

4. New Functions to Support Stateful PCEs

Several new functions will be required in PCEP to support stateful PCEs. A function can be initiated either from a PCC towards a PCE (C-E) or from a PCE towards a PCC (E-C). The new functions are:

Capability negotiation (E-C,C-E): both the PCC and the PCE must announce during PCEP session establishment that they support PCEP Stateful PCE extensions defined in this document.

LSP state synchronization (C-E): after the session between the PCC and a stateful PCE is initialized, the PCE must learn the state of a PCC's LSPs before it can perform path computations or update LSP attributes in a PCC.

LSP Update Request (E-C): A PCE requests modification of attributes on a PCC's LSP.

LSP State Report (C-E): a PCC sends an LSP state report to a PCE whenever the state of an LSP changes.

LSP control delegation (C-E,E-C): a PCC grants to a PCE the right to update LSP attributes on one or more LSPs; the PCE becomes the authoritative source of the LSP's attributes as long as the delegation is in effect (See Section 5.5); the PCC may withdraw the delegation or the PCE may give up the delegation

In addition to new PCEP functions, stateful capabilities discovery

will be required in OSPF ([RFC5088]) and IS-IS ([RFC5089]). Stateful capabilities discovery is not in scope of this document.

5. Architectural Overview of Protocol Extensions

5.1. LSP State Ownership

In the PCEP protocol (defined in [RFC5440]), LSP state and operation are under the control of a PCC (a PCC may be an LSR or a management station). Attributes received from a PCE are subject to PCC's local policy. The PCEP protocol extensions described in this document do not change this behavior.

An active stateful PCE may have control of a PCC's LSPs be delegated to it, but the LSP state ownership is retained by the PCC. In particular, in addition to specifying values for LSP's attributes, an active stateful PCE also decides when to make LSP modifications.

Retaining LSP state ownership on the PCC allows for:

- a PCC to interact with both stateless and stateful PCEs at the same time
- a stateful PCE to only modify a small subset of LSP parameters, i.e. to set only a small subset of the overall LSP state; other parameters may be set by the operator through CLI commands
- a PCC to revert delegated LSP to an operator-defined default or to delegate the LSPs to a different PCE, if the PCC get disconnected from a PCE with currently delegated LSPs

5.2. New Messages

In this document, we define the following new PCEP messages:

Path Computation State Report (PCRpt): a PCEP message sent by a PCC to a PCE to report the status of one or more LSPs. Each LSP Status Report in a PCRpt message can contain the actual LSP's path,bandwidth, operational and administrative status, etc. An LSP Status Report carried on a PCRpt message is also used in delegation or revocation of control of an LSP to/from a PCE. The PCRep message is described in Section 6.1.

Path Computation Update Request (PCUpd): a PCEP message sent by a PCE to a PCC to update LSP parameters, on one or more LSPs. Each LSP Update Request on a PCUpd message MUST contain all LSP parameters that a PCE wishes to set for a given LSP. An LSP

Update Request carried on a PCUpd message is also used to return LSP delegations if at any point PCE no longer desires control of an LSP. The PCUpd message is described in Section 6.2.

The new functions defined in Section 4 are mapped onto the new messages as shown in the following table.

Function	Message
Capability Negotiation (E-C,C-E)	Open
State Synchronization (C-E)	PCRpt
LSP State Report (C-E)	PCRpt
LSP Control Delegation (C-E,E-C)	PCRpt, PCUpd
LSP Update Request (E-C)	PCUpd
ISIS stateful capability advertisement	ISIS PCE-CAP-FLAGS sub-TLV
OSPF stateful capability advertisement	OSPF RI LSA, PCE TLV, PCE-CAP-FLAGS sub-TLV

Table 12: New Function to Message Mapping

5.3. Capability Negotiation

During PCEP Initialization Phase, PCEP Speakers (PCE pr PCC) negotiate the use of stateful PCEP extensions. A PCEP Speaker includes the "Stateful PCE Capability" TLV, described in Section 7.1.1, in the OPEN Object to advertise its support for PCEP stateful extensions. The Stateful Capability TLV includes the 'LSP Update' Flag that indicates whether the PCEP Speaker supports LSP parameter updates.

The presence of the Stateful PCE Capability TLV in PCC's OPEN Object indicates that the PCC is willing to send LSP State Reports whenever LSP parameters or operational status changes.

The presence of the Stateful PCE Capability TLV in PCE's OPEN message indicates that the PCE is interested in receiving LSP State Reports whenever LSP parameters or operational status changes.

The PCEP protocol extensions for stateful PCEs MAY only be used if both sides have included the Stateful PCE Capability TLV in their respective OPEN messages, otherwise a PCErr with code "Stateful PCE capability not negotiated" (see Section 8.4) will be generated and the PCEP session will be terminated.

LSP delegation and LSP update operations defined in this document MAY

only be used if both PCEP Speakers set the LSP-UPDATE Flag in the "Stateful Capability" TLV to 'Updates Allowed (U Flag = 1)', otherwise a PCErr with code "Delegation not negotiated" (see Section 8.4) will be generated. Note that even if the update capability has not been negotiated, a PCE can still receive LSP Status Reports from a PCC and build and maintain an up to date view of the state of the PCC's LSPs.

5.4. State Synchronization

The purpose of State Synchronization is to provide a checkpoint-in-time state replica of a PCC's LSP state in a PCE. State Synchronization is performed immediately after the Initialization phase ([RFC5440]).

During State Synchronization, a PCC first takes a snapshot of the state of its LSPs state, then sends the snapshot to a PCE in a sequence of LSP State Reports. Each LSP State Report sent during State Synchronization has the SYNC Flag in the LSP Object set to 1. The set of LSPs for which state is synchronized with a PCE is determined by negotiated stateful PCEP capabilities and PCC's local configuration (see more details in Section 9.1).

A PCE SHOULD NOT send PCUpd messages to a PCC before State Synchronization is complete. A PCC SHOULD NOT send PCReq messages to a PCE before State Synchronization is complete. This is to allow the PCE to get the best possible view of the network before it starts computing new paths.

If the PCC encounters a problem which prevents it from completing the state transfer, it MUST send a PCErr message to the PCE and terminate the session using the PCEP session termination procedure.

The PCE does not send positive acknowledgements for properly received synchronization messages. It MUST respond with a PCErr message indicating "PCRpt error" (see) if it encounters a problem with the LSP State Report it received from the PCC. Either the PCE or the PCC MAY terminate the session if the PCE encounters a problem during the synchronization.

The successful State Synchronization sequence is shown in Figure 3.

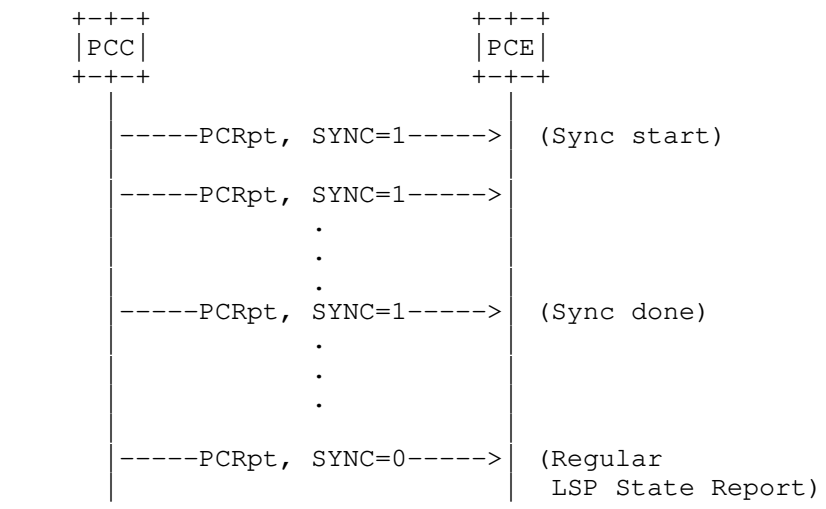


Figure 3: Successful state synchronization

The sequence where the PCE fails during the State Synchronization phase is shown in Figure 4.

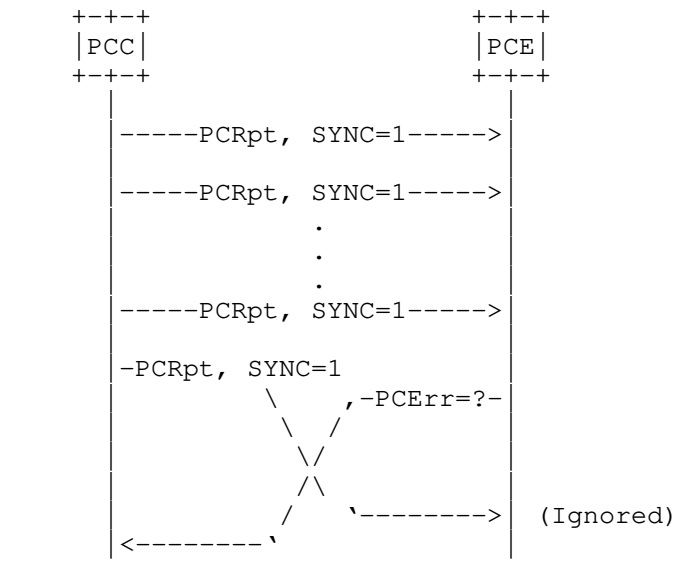


Figure 4: Failed state synchronization (PCE failure)

The sequence where the PCC fails during the State Synchronization phase is shown in Figure 5.

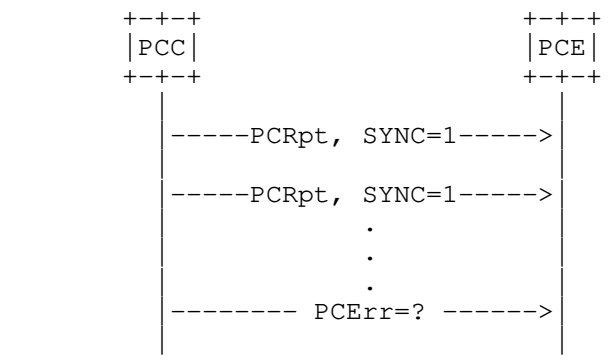


Figure 5: Failed state synchronization (PCC failure)

5.4.1. State Synchronization Avoidance

State Synchronization MAY be skipped following a PCEP session restart if the state of both PCEP peers did not change during the period prior to session re-initialization. To be able to make this determination, state must be exchanged and maintained by both PCE and PCC during normal operation. This is accomplished by keeping track of the changes to the LSP State Database. When State Synchronization avoidance is enabled on a PCEP session, a PCC includes the LSP-DB-VERSION TLV as an optional TLV in the LSP Object on each LSP State Report. The LSP-DB-VERSION TLV contains a PCC's LSP State Database version, which is incremented each time a change is made to the PCC's local LSP State Database. The LSP State Database version is an unsigned 64-bit value that MUST be incremented by 1 for each successive change in the LSP state database. The LSP State Database version MUST start at 1 and may wrap around. Values 0 and 0xFFFFFFFFFFFFFFFF are reserved.

State Synchronization Avoidance is negotiated on a PCEP session during session startup.

If both PCEP speakers set the INCLUDE-DB-VERSION Flag in the OPEN object's STATEFUL-PCE-CAPABILITY TLV to 1, the PCC will include the LSP-DB-VERSION TLV in each LSP Object. The TLV will contain the PCC's latest LSP State Database version.

If a PCE's LSP State Database survived the restart of a PCEP session, the PCE will include the LSP-DB-VERSION TLV in its OPEN object, and the TLV will contain the last LSP State Database version received on an LSP State Update from the PCC in a previous PCEP session. If a PCC's LSP State Database survived the restart, the PCC will include the LSP-DB-VERSION TLV in its OPEN object and the TLV will contain the last LSP State Database version sent on an LSP State Update from

the PCC in the previous PCEP session. If a PCEP Speaker's LSP State Database did not survive the restart of a PCEP session, the PCEP Speaker MUST NOT include the LSP-DB-VERSION TLV in the OPEN Object.

If both PCEP Speakers include the LSP-DB-VERSION TLV in the OPEN Object and the TLV values match, the PCC MAY skip State Synchronization. Otherwise, the PCE MUST purge any remaining LSP state and expect the PCC to perform State Synchronization; if the PCC attempts to skip State Synchronization (i.e. the SYNC Flag = 0 on the first LSP State Report from the PCC), the PCE MUST send back a PCError with Error-type 20 Error-value 2 'LSP Database version mismatch', and close the PCEP session.

Note that a PCE MAY force State Synchronization by not including the LSP-DB-VERSION TLV in its OPEN object.

Figure 6 shows an example sequence where State Synchronization is skipped.

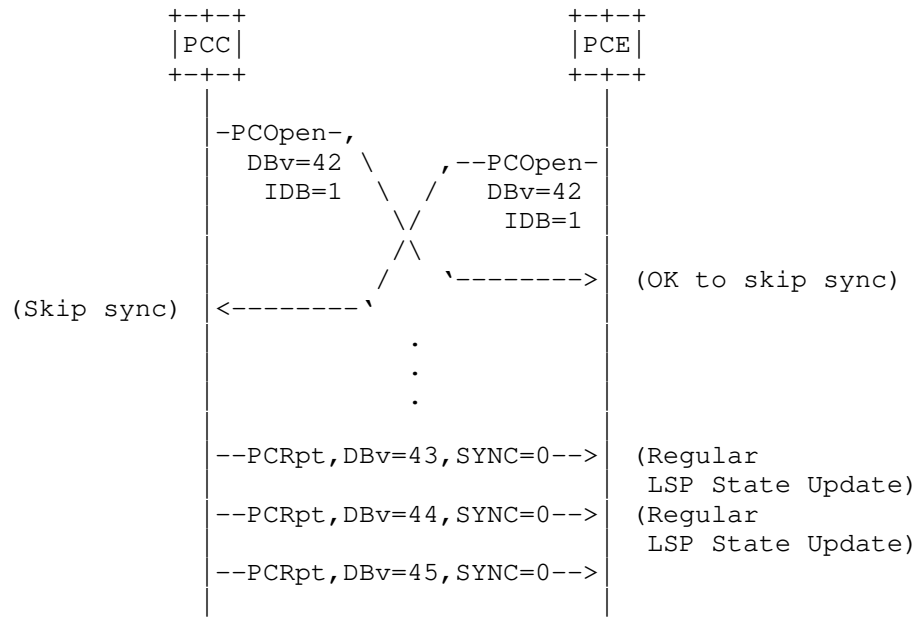


Figure 6: State Synchronization skipped

Figure 7 shows an example sequence where State Synchronization is performed due to LSP State Database version mismatch during the PCEP session setup. Note that the same State Synchronization sequence would happen if either the PCE or the PCE would not include the LSP-DB-VERSION TLV in their respective Open messages.

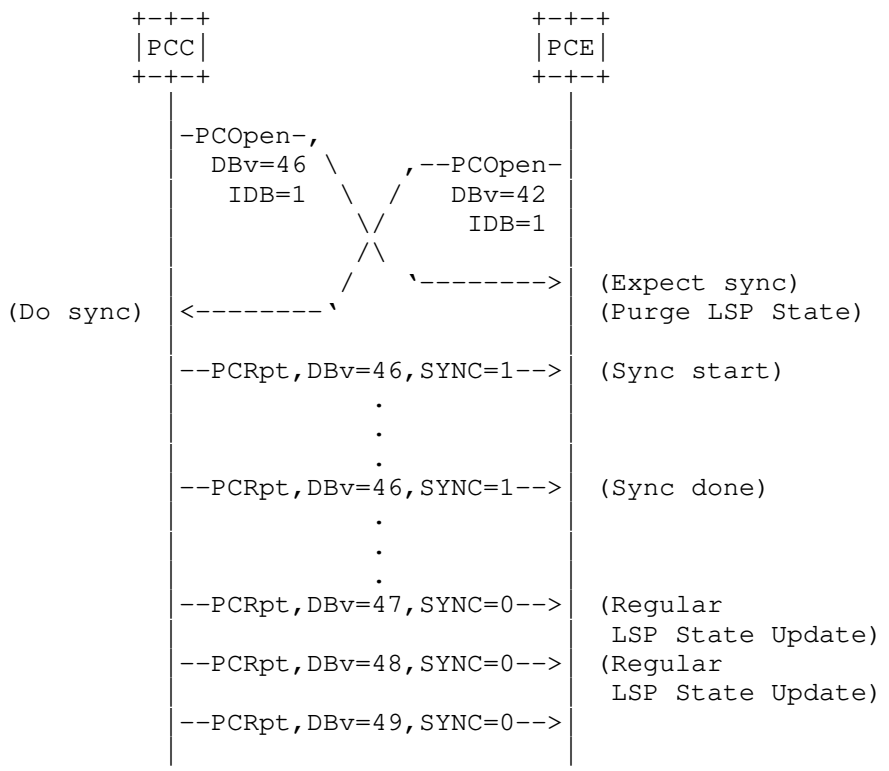


Figure 7: State Synchronization performed

Figure 8 shows an example sequence where State Synchronization is skipped, but because one or both PCEP Speakers set the INCLUDE-DB-VERSION Flag to 0, the PCC does not send LSP-DB-VERSION TLVs to the PCE. If the current PCEP session restarts, the PCEP Speakers will have to perform State Synchronization, since the PCE will not know the PCC's latest LSP State Database version.

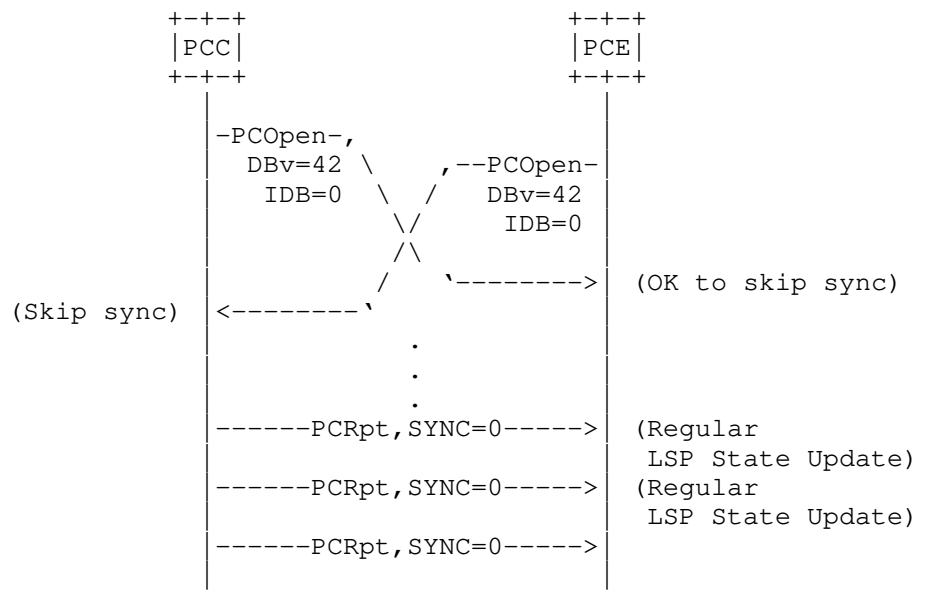


Figure 8: State Synchronization skipped, no LSP-DB-VERSION TLVs sent from PCC

5.5. LSP Delegation

If during Capability negotiation both the PCE and the PCC have indicated that they support LSP Update, then the PCC may choose to grant the PCE a temporary right to update (a subset of) LSP attributes on one or more LSPs. This is called "LSP Delegation", and it MAY be performed at any time after the Initialization phase.

LSP Delegation is controlled by operator-defined policies on a PCC. LSPs are delegated individually - different LSPs may be delegated to different PCEs, and an LSP may be delegated to one or more PCEs. The delegation policy, when all PCC's LSPs are delegated to a single PCE at any given time, SHOULD be supported by all delegation-capable PCCs. Conversely, the policy revoking the delegation for all PCC's LSPs SHOULD also be supported

A PCE may return LSP delegation at any time if it no longer wishes to update the LSP's state. A PCC may revoke LSP delegation at any time. Delegation, Revocation, and Return are done individually for each LSP.

5.5.1. Delegating an LSP

A PCC delegates an LSP to a PCE by setting the Delegate flag in LSP State Report to 1. A PCE confirms the delegation when it sends the first LSP Update Request for the delegated LSP to the PCC by setting the Delegate flag to 1. Note that a PCE does not immediately confirm to the PCC the acceptance of LSP Delegation; Delegation acceptance is confirmed when the PCC wishes to update the LSP via the LSP Update Request. If a PCE does not accept the LSP Delegation, it MUST immediately respond with an empty LSP Update Request which has the Delegate flag set to 0.

The delegation sequence is shown in Figure 9.

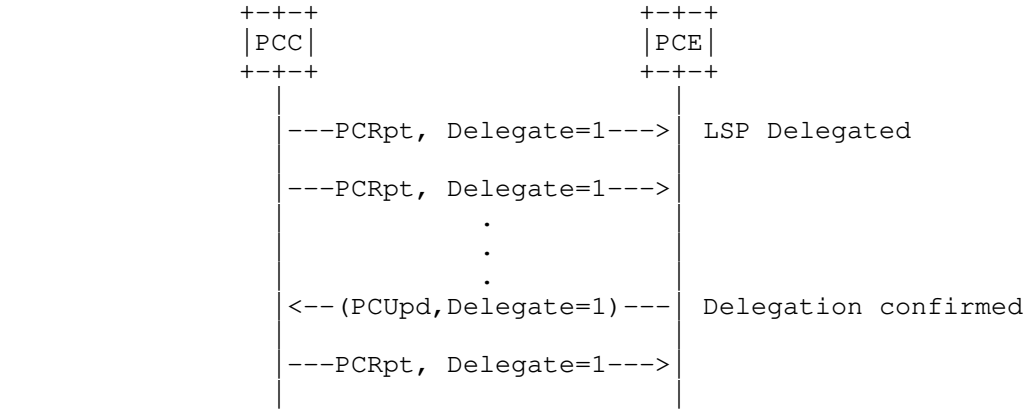


Figure 9: Delegating an LSP

Note that for an LSP to remain delegated to a PCE, the PCC MUST set the Delegate flag to 1 on each LSP Status Report sent to the PCE.

5.5.2. Revoking a Delegation

A PCC revokes an LSP delegation by sending an LSP State Report with the Delegate flag set to 0. A PCC MAY revoke an LSP delegation at any time during the PCEP session life time. When a PCC's PCEP session with the PCE terminates, the PCC SHALL wait a time interval specified in 'Delegation Timeout Interval' and then revoke all LSP delegations to the PCE .

After an LSP delegation has been revoked, a PCE can no longer update LSP's parameters; an attempt to update parameters of a non-delegated LSP will result in the PCC sending a PCErr message indicating "LSP is not delegated" (see Section 8.4).

The revocation sequence is shown in Figure 10.

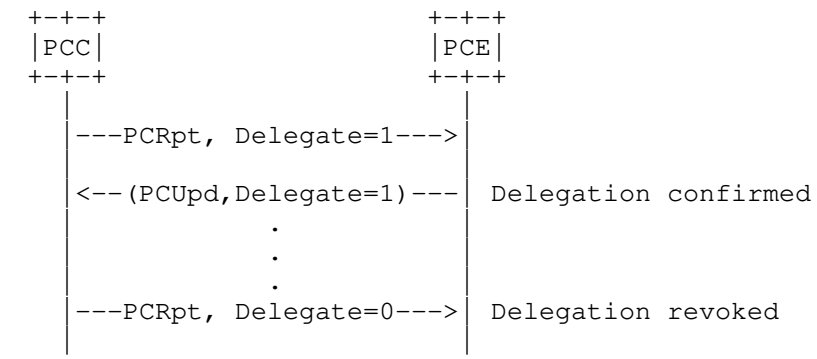


Figure 10: Revoking a Delegation

If a PCC can not delegate an LSP to a PCE (for example, if a PCC is not connected to any active stateful PCE or if no connected active stateful PCE accepts the delegation), the LSP delegation on the PCC will time out within a configurable Delegation Timeout Interval and the PCC SHALL flush any LSP state set by a PCE.

5.5.3. Returning a Delegation

A PCE that no longer wishes to update an LSP's parameters SHOULD return the LSP delegation back to the PCC by sending an empty LSP Update Request which has the Delegate flag set to 0. Note that in order to keep a delegation, the PCE MUST set the Delegate flag to 1 on each LSP Update Request sent to the PCC.

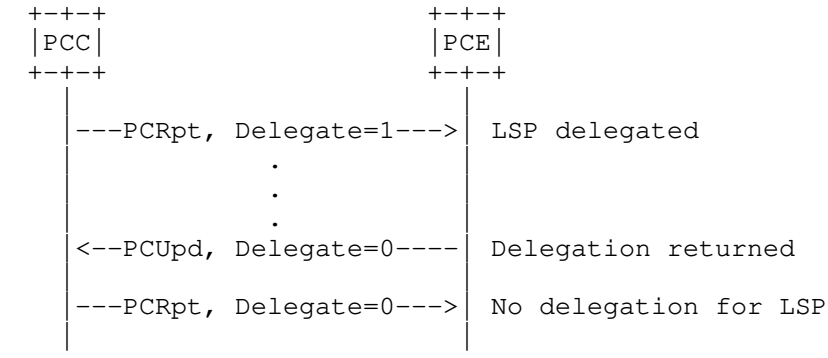


Figure 11: Returning a Delegation

If a PCC can not delegate an LSP to a PCE (for example, if a PCC is not connected to any active stateful PCE or if no connected active

stateful PCE accepts the delegation), the LSP delegation on the PCC will time out within a configurable Delegation Timeout Interval and the PCC MUST flush any LSP state set by a PCE.

5.5.4. Redundant Stateful PCEs

Note that a PCE may not have any delegated LSPs: in a redundant configuration where one PCE is backing up another PCE, the backup PCE will not have any delegated LSPs. The backup PCE does not update any LSPs, but it receives all LSP State Reports from a PCC. When the primary PCE fails, a PCC will delegate to the secondary PCE all LSPs that had been previously delegated to the failed PCE.

5.6. LSP Operations

5.6.1. Passive Stateful PCE Path Computation Request/Response

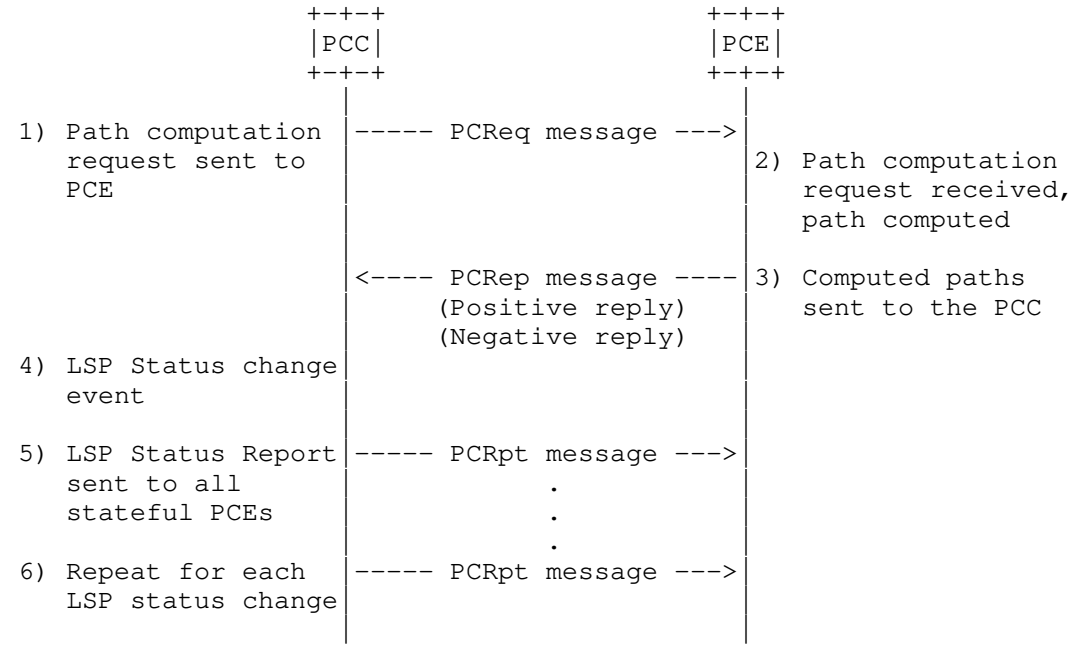


Figure 12: Passive Stateful PCE Path Computation Request/Response

Once a PCC has successfully established a PCEP session with a passive stateful PCE and the PCC's LSP state is synchronized with the PCE (i.e. the PCE knows about all PCC's existing LSPs), if an event is triggered that requires the computation of a set of paths, the PCC sends a path computation request to the PCE ([RFC5440], Section 4.2.3). The PCReq message MAY contain the LSP Object to identify the

LSP for which the path computation is requested.

Upon receiving a path computation request from a PCC, the PCE triggers a path computation and returns either a positive or a negative reply to the PCC ([RFC5440], Section 4.2.4).

Upon receiving a positive path computation reply, the PCC receives a set of computed paths and starts to setup the LSPs. For each LSP, it sends an LSP State Report carried on a PCRpt message to the PCE, indicating that the LSP's status is 'Pending'.

Once an LSP is up, the PCC sends an LSP State Report carried on a PCRpt message to the PCE, indicating that the LSP's status is 'Up'. If the LSP could not be set up, the PCC sends an LSP State Report indicating that the LSP is "Down' and stating the cause of the failure. Note that due to timing constraints, the LSP status may change from 'Pending' to 'Up' (or 'Down') before the PCC has had a chance to send an LSP State Report indicating that the status is 'Pending'. In such cases, the PCC may choose to only send the PCRpt indicating the latest status ('Up' or 'Down').

Upon receiving a negative reply from a PCE, a PCC may decide to resend a modified request or take any other appropriate action. For each requested LSP, it also sends an LSP State Report carried on a PCRpt message to the PCE, indicating that the LSP's status is 'Down'.

There is no direct correlation between PCRep and PCRpt messages. For a given LSP, multiple LSP State Reports will follow a single PC Reply, as a PCC notifies a PCE of the LSP's state changes.

A PCC sends each LSP State Report to each stateful PCE that is connected to the PCC.

Note that a single PCRpt message MAY contain multiple LSP State Reports.

The passive stateful PCE is the model for stateful PCEs is described in [RFC4655], Section 6.8.

5.6.2. Active Stateful PCE LSP Update

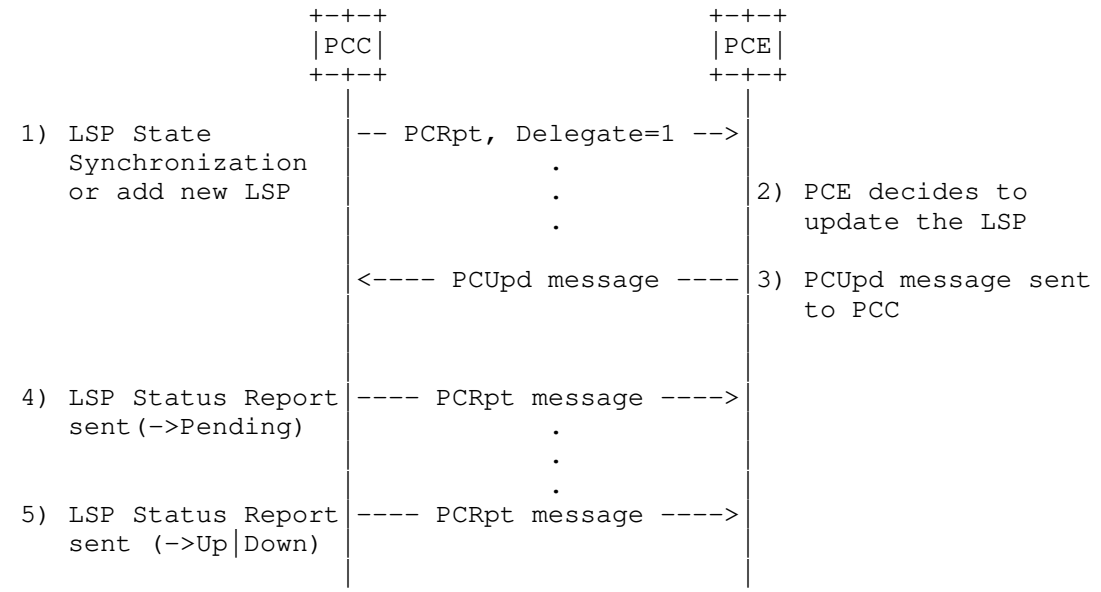


Figure 13: Active Stateful PCE

Once a PCC has successfully established a PCEP session with an active stateful PCE, the PCC's LSP state is synchronized with the PCE (i.e. the PCE knows about all PCC's existing LSPs) and LSPs have been delegated to the PCE, the PCE can modify LSP parameters of delegated LSPs.

A PCE sends an LSP Update Request carried on a PCUpd message to the PCC. The LSP Update Request contains a variety of objects that specify the set of constraints and attributes for the LSP's path. Additionally, the PCC may specify the urgency of such request by assigning a request priority. A single PCUpd message MAY contain multiple LSP Update Requests.

Upon receiving a PCUpd message the PCC starts to setup LSPs specified in LSP Update Requests carried in the message. For each LSP, it sends an LSP State Report carried on a PCRpt message to the PCE, indicating that the LSP's status is 'Pending'.

Once an LSP is up, the PCC sends an LSP State Report (PCRpt message) to the PCE, indicating that the LSP's status is 'Up'. If the LSP could not be set up, the PCC sends an LSP State Report indicating that the LSP is 'Down' and stating the cause of the failure. A PCC may choose to compress LSP State Updates to only reflect the most up

to date state, as discussed in the previous section.

A PCC sends each LSP State Report to each stateful PCE that is connected to the PCC.

A PCC MUST NOT send to any PCE a Path Computation Request for a delegated LSP.

5.7. LSP Protection

With a stateless PCE or a passive stateful PCE, LSP protection and restoration settings may be operator-configured locally at a PCC. A PCE may be merely asked to compute the protected (primary) and backup (secondary) paths for the LSP.

An active stateful PCE controls the LSPs that are delegated to it, and must therefore be able to set via PCEP the desired protection / restoration mechanism for each delegated LSP. PCEP extensions for stateful PCEs SHOULD support, at a minimum, the following protection mechanisms:

- MPLS TE Global Default Restoration
- MPLS TE Global Path Protection
- FRR One-to-One Backup
- FRR Facility Backup - link protection, node protection, or both

5.8. Transport

A Permanent PCEP session MUST be established between a stateful PCEs and the PCC.

State cleanup after session termination, as well as session setup failures will be described in a later version of this document.

6. PCEP Messages

As defined in [RFC5440], a PCEP message consists of a common header followed by a variable-length body made of a set of objects that can be either mandatory or optional. An object is said to be mandatory in a PCEP message when the object must be included for the message to be considered valid. For each PCEP message type, a set of rules is defined that specify the set of objects that the message can carry. An implementation MUST form the PCEP messages using the object ordering specified in this document.

6.1. The PCRpt Message

A Path Computation LSP State Report message (also referred to as PCRpt message) is a PCEP message sent by a PCC to a PCE to report the current state of an LSP. A PCRpt message can carry more than one LSP State Reports. A PCC can send an LSP State Report either in response to an LSP Update Request from a PCE, or asynchronously when the state of an LSP changes. The Message-Type field of the PCEP common header for the PCRpt message is set to [TBD].

The format of the PCRpt message is as follows:

```
<PCRpt Message> ::= <Common Header>
                    <state-report-list>
```

Where:

```
<state-report-list> ::= <state-report>[<state-report-list>]

<state-report> ::= <LSP>
                   [<path-list>]
```

Where:

```
<path-list>::=<path>[<path-list>]

<path>::=<ERO><attribute-list>
```

Where:

```
 <attribute-list> ::= [<LSPA>]
                      [<BANDWIDTH>]
                      [<RRO>]
                      [<metric-list>]

 <metric-list> ::= <METRIC>[<metric-list>]
```

The LSP object (see Section 7.2) is mandatory, and it MUST be included in each LSP State Report on the PCRpt message. If the LSP object is missing, the receiving PCE MUST send a PCErr message with Error-type=6 (Mandatory Object missing) and Error-value=[TBD] (LSP object missing).

The LSP State Report MAY contain a path descriptor for the primary path and one or more path descriptors for backup paths. A path descriptor MUST contain an ERO object as it was specified by a PCE or an operator. A path descriptor MUST contain the RRO object if a primary or secondary LSP is set up along the path in the network. A path descriptor MAY contain the LSPA, BANDWIDTH, and METRIC objects.

The ERO,LSPA, BANDWIDTH, METRIC, and RRO objects are defined in[RFC5440].

6.2. The PCUpd Message

A Path Computation LSP Update Request message (also referred to as PCUpd message) is a PCEP message sent by a PCE to a PCC to update attributes of an LSP. A PCUpd message can carry more than one LSP Update Request. The Message-Type field of the PCEP common header for the PCRpt message is set to [TBD].

The format of a PCUpd message is as follows:

```
<PCUpd Message> ::= <Common Header>
                    <udpate-request-list>
```

Where:

```
<update-request-list> ::= <update-request>[<update-request-list>]

<update-request> ::= <LSP>
                     [<path-list>]
```

Where:

```
<path-list>::=<path>[<path-list>]

<path>::=<ERO><attribute-list>
```

Where:

```
<attribute-list> ::= [<LSPA>]
                     [<BANDWIDTH>]
                     [<metric-list>]
                     [<IRO>]

<metric-list> ::= <METRIC>[<metric-list>]
```

There is one mandatory object that MUST be included within each LSP Update Request in the PCUpd message: the LSP object (see Section 7.2). If the LSP object is missing, the receiving PCE MUST send a PCErr message with Error-type=6 (Mandatory Object missing) and Error-value=[TBD] (LSP object missing).

The LSP Update Request MUST contain a path descriptor for the primary path, and MAY contain one or more path descriptors for backup paths. A path descriptor MUST contain an ERO object. A path descriptor MAY further contain the BANDWIDTH, IRO, and METRIC objects. The ERO,

LSPA, BANDWIDTH, METRIC, and IRO objects are defined in [RFC5440].

Each LSP Update Request results in a separate LSP setup operation at a PCC. An LSP Update Request MUST contain all LSP parameters that a PCC wishes to set for the LSP. A PCC MAY set missing parameters from locally configured defaults. If the LSP specified the Update Request is already up, it will be re-signaled. The PCC will use make-before-break whenever possible in the re-signaling operation.

A PCC MUST respond with an LSP State Report to each LSP Update Request to indicate the resulting state of the LSP in the network. A PCC MAY respond with multiple LSP State Reports to report LSP setup progress of a single LSP.

If the rate of PCUpd messages sent to a PCC for the same target LSP exceeds the rate at which the PCC can signal LSPs into the network, the PCC MAY perform state compression and only re-signal the last modification in its queue.

Note that a PCC MUST process all LSP Update Requests - for example, an LSP Update Request is sent when a PCE returns delegation or puts an LSP into non-operational state. The protocol relies on TCP for message-level flow control.

Note also that it's up to the PCE to handle inter-LSP dependencies; for example, if ordering of LSP set-ups is required, the PCE has to wait for an LSP State Report for a previous LSP before triggering the LSP setup of a next LSP.

7. Object Formats

The PCEP objects defined in this document are compliant with the PCEP object format defined in [RFC5440]. The P flag and the I flag of the PCEP objects defined in this document MUST always be set to 0 on transmission and MUST be ignored on receipt since these flags are exclusively related to path computation requests.

7.1. OPEN Object

This document defines a new optional TLV for the OPEN Object to support stateful PCE capability negotiation.

7.1.1. Stateful PCE Capability TLV

The format of the STATEFUL-PCE-CAPABILITY TLV is shown in the following figure:

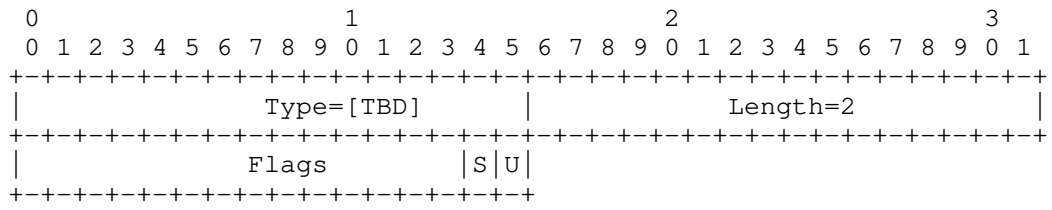


Figure 14: STATEFUL-PCE-CAPABILITY TLV format

The type of the TLV is [TBD] and it has a fixed length of 2 octets.

The value comprises a single field - Flags (16 bits):

U (LSP-UPDATE-CAPABILITY - 1 bit): if set to 1 by a PCC, the U Flag indicates that the PCC allows modification of LSP parameters; if set to 1 by a PCE, the U Flag indicates that the PCE wishes to update LSP parameters. The LSP-UPDATE-CAPABILITY Flag must be advertised by both a PCC and a PCE for PCUpd messages to be allowed on a PCEP session.

S (INCLUDE-DB-VERSION - 1 bit): if set to 1 by both PCEP Speakers, the PCC will include the LSP-DB-VERSION TLV in each LSP Object.

Unassigned bits are considered reserved. They MUST be set to 0 on transmission and MUST be ignored on receipt.

7.1.2. LSP State Database Version TLV

LSP-DB-VERSION is an optional TLV that MAY be included in the OPEN Object when a PCEP Speaker wishes to determine if State Synchronization can be skipped when a PCEP session is restarted. If sent from a PCE, the TLV contains the local LSP State Database version from the last valid LSP State Report received from a PCC. If sent from a PCC, the TLV contains the PCC's local LSP State Database version, which is incremented each time the LSP State Database is updated.

The format of the LSP-DB-VERSION TLV is shown in the following figure:

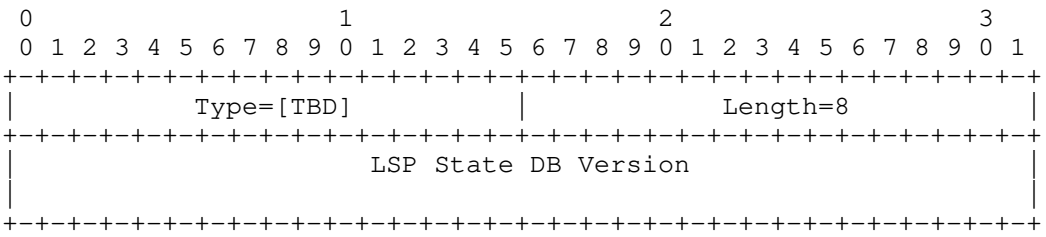


Figure 15: LSP-DB-VERSION TLV format

The type of the TLV is [TBD] and it has a fixed length of 8 octets. The value contains a 64-bit unsigned integer.

7.2. LSP Object

The LSP object MUST be present within PCRpt and PCUpd messages. The LSP object MAY be carried within PCReq and PCRep messages if the stateful PCE capability has been negotiated on the session. The LSP object contains a set of fields used to specify the target LSP, the operation to be performed on the LSP, and LSP Delegation. It is also contains a flag to indicate to a PCE that the initial LSP state synchronization has been done.

LSP Object-Class is [TBD].

LSP Object-Type is 1.

The format of the LSP object body is shown in Figure 16:

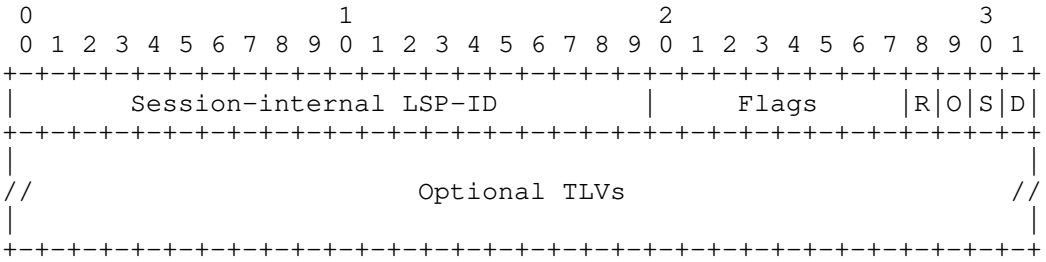


Figure 16: The LSP Object format

The LSP object body has a variable length and may contain additional TLVs.

Session-internal LSP-ID (20 bits): Per-PCEP session identifier for an LSP. In each PCEP session the PCC creates a unique LSP-ID for each LSP that will remain constant for the duration of the session. The

mapping of the LSP Symbolic Name to LSP-ID is communicated to the PCE by sending a PCRpt message containing the 'LSP Symbolic Name' TLV. All subsequent PCEP messages then address the LSP by its Session-internal LSP-ID.

Flags (12 bits):

D (Delegate - 1 bit): on a PCRpt message, the D Flag set to 1 indicates that the PCC is delegating the LSP the PCE. On a PCUpd message, the D flag set to 1 indicates that the PCE is confirming the LSP Delegation. To keep an LSP delegated to the PCE, the PCC must set the D flag to 1 on each PCRpt message for the duration of the delegation - the first PCRpt with the D flag set to 0 revokes the delegation. To keep the delegation, the PCE must set the D flag to 1 on each PCUpd message for the duration of the delegation - the first PCUpd with the D flag set to 0 returns the delegation.

S (SYNC - 1 bit): the S Flag MUST be set to 1 on each LSP State Report sent from a PCC during State Synchronization. The S Flag MUST be set to 0 otherwise.

O (Operational - 1 bit): On PCRpt messages the O Flag indicates the LSP status. Value of '1' means that the LSP is operational, i.e. it is either being signaled or it is active. Value of '0' means that the LSP is not operational, i.e it is de-routed and the PCC is not attempting to set it up. On PCUpd messages the flag indicates the desired status for the LSP. Value of '1' means that the desired LSP state is operational, value of '0' means that the target LSP should be non-operational. Setting the LSP status from the PCE SHALL NOT override the operator: if a pce-controlled LSP has been configured to be non-operational, setting the LSP's status to '1' from an PCE will not make it operational.

R (Remove - 1 bit): On PCRpt messages the R Flag indicates that the LSP has been removed from the PCC. Upon receiving an LSP State Update with the R Flag set to 1, the PCE SHOULD remove all state related to the LSP from its database.

Unassigned bits are considered reserved. They MUST be set to 0 on transmission and MUST be ignored on receipt.

TLVs that are currently defined for the LSP Object are described in the following sections.

7.2.1. The LSP Symbolic Name TLV

Each LSP MUST have a symbolic name that is unique in the PCC. The LSP Symbolic Name MUST remain constant throughout an LSP's lifetime,

which may span across multiple consecutive PCEP sessions and/or PCC restarts. The LSP Symbolic Name MAY be specified by an operator in a PCC's CLI configuration. If the operator does not specify a Symbolic Name for an LSP, the PCC MUST auto-generate one.

The LSP-SYMBOLIC-NAME TLV MUST be included in the LSP State Report when during a given PCEP session an LSP is first reported to a PCE. A PCC sends to a PCE the first LSP State Report either during State Synchronization, or when a new LSP is configured at the PCC. LSP State Report MAY be included in subsequent LSP State Reports for the LSP.

The format of the LSP-SYMBOLIC-NAME TLV is shown in the following figure:

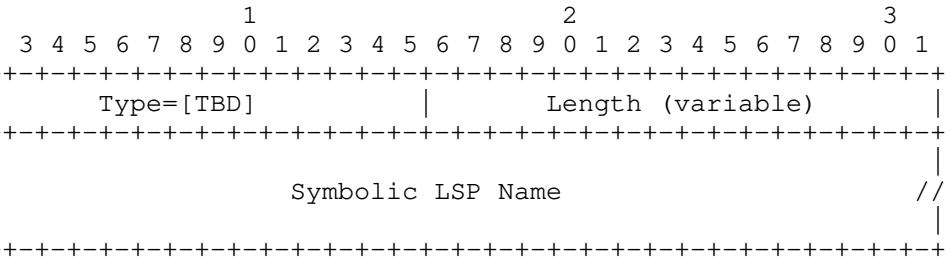


Figure 17: LSP-SYMBOLIC-NAME TLV format

The type of the TLV is [TBD] and it has a variable length, which MUST be greater than 0.

7.2.2. LSP Identifiers TLVs

Whenever the value of an LSP identifier changes, a PCC MUST send out an LSP State Report, where the LSP Object carries the LSP Identifiers TLV that contains the new value. The LSP Identifiers TLV MUST also be included in the LSP object during state synchronization. There are two LSP Identifiers TLVs, one for IPv4 and one for IPv6.

The format of the IPV4-LSP-IDENTIFIERS TLV is shown in the following figure:

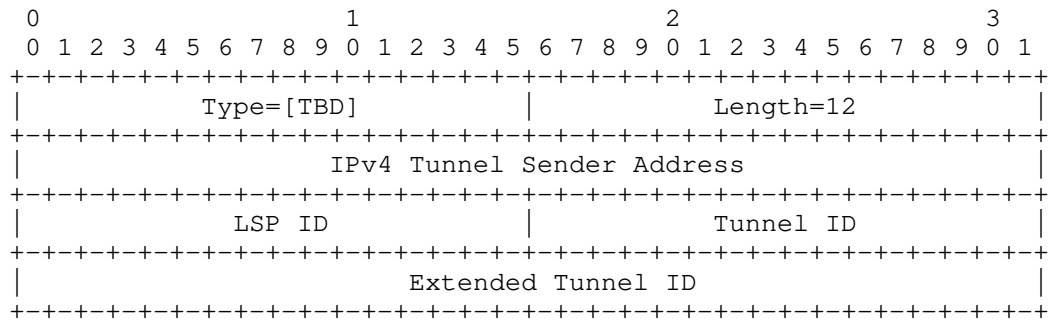


Figure 18: IPV4-LSP-IDENTIFIERS TLV format

The type of the TLV is [TBD] and it has a fixed length of 8 octets. The value contains two fields:

IPv4 Tunnel Sender Address: contains the sender node's IPv4 address, as defined in [RFC3209], Section 4.6.2.1 for the LSP_TUNNEL_IPv4 Sender Template Object.

LSP ID: contains the 16-bit 'LSP ID' identifier defined in [RFC3209], Section 4.6.2.1 for the LSP_TUNNEL_IPv4 Sender Template Object.

Tunnel ID: contains the 16-bit 'Tunnel ID' identifier defined in [RFC3209], Section 4.6.1.1 for the LSP_TUNNEL_IPv4 Session Object. Tunnel ID remains constant over the life time of a tunnel. However, when Global Path Protection or Global Default Restoration is used, both the primary and secondary LSPs have their own Tunnel IDs. A PCC will report a change in Tunnel ID when traffic switches over from primary LSP to secondary LSP (or vice versa).

Extended Tunnel ID: contains the 128-bit 'Extended Tunnel ID' identifier defined in [RFC3209], Section 4.6.1.1 for the LSP_TUNNEL_IPv4 Session Object.

The format of the IPV6-LSP-IDENTIFIERS TLV is shown in the following figure:

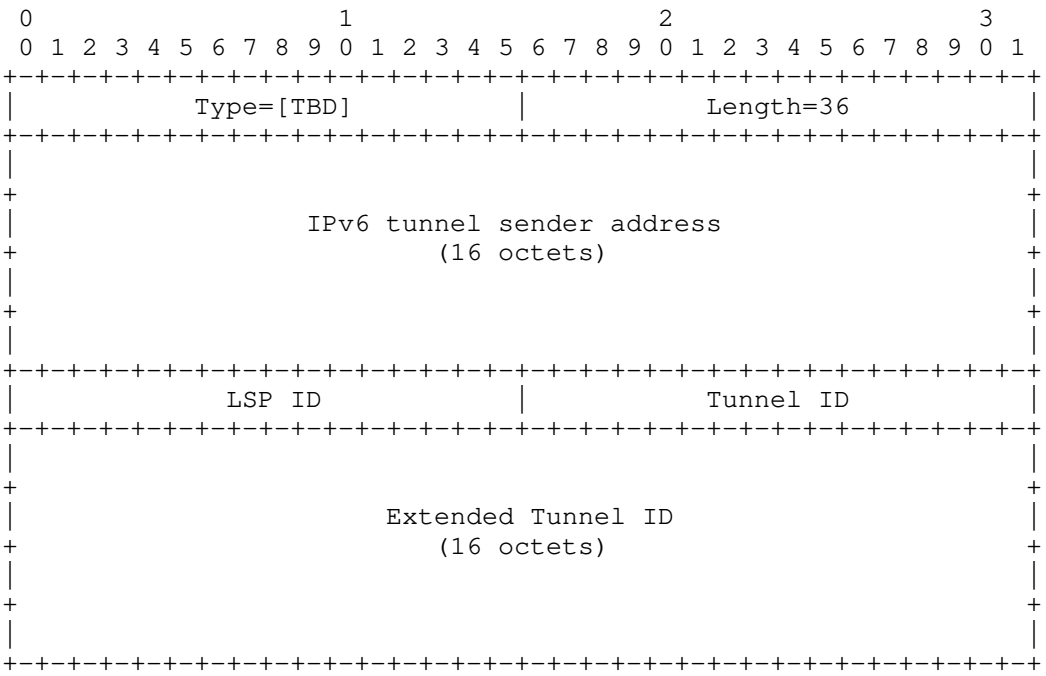


Figure 19: IPV6-LSP-IDENTIFIERS TLV format

The type of the TLV is [TBD] and it has a fixed length of 20 octets. The value contains two fields:

IPv6 Tunnel Sender Address: contains the sender node's IPv6 address, as defined in [RFC3209], Section 4.6.2.2 for the LSP_TUNNEL_IPv6 Sender Template Object.

LSP ID: contains the 16-bit 'LSP ID' identifier defined in [RFC3209], Section 4.6.2.2 for the LSP_TUNNEL_IPv6 Sender Template Object.

Tunnel ID: contains the 16-bit 'Tunnel ID' identifier defined in [RFC3209], Section 4.6.1.2 for the LSP_TUNNEL_IPv6 Session Object. Tunnel ID remains constant over the life time of a tunnel. However, when Global Path Protection or Global Default Restoration is used, both the primary and secondary LSPs have their own Tunnel IDs. A PCC will report a change in Tunnel ID when traffic switches over from primary LSP to secondary LSP (or vice versa).

Extended Tunnel ID: contains the 32-bit 'Extended Tunnel ID' identifier defined in [RFC3209], Section 4.6.1.2 for the LSP_TUNNEL_IPv6 Session Object.

7.2.3. LSP Update Error Code TLV

If an LSP Update Request failed, an LSP State Report MUST be sent to all connected stateful PCEs. LSP State Report MUST contain the LSP Update Error Code TLV, indicating the cause of the failure.

The format of the LSP-UPDATE-ERROR-CODE TLV is shown in the following figure:

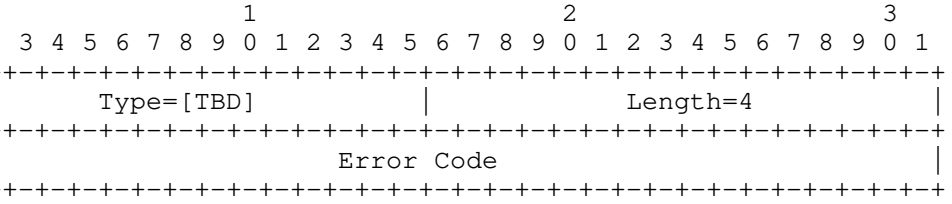


Figure 20: LSP-UPDATE-ERROR-CODE TLV format

The type of the TLV is [TBD] and it has a fixed length of 4 octets. The value contains the error code that indicates the cause of the LSP setup failure. Error codes will be defined in a later revision of this document.

7.2.4. RSVP ERROR_SPEC TLVs

If the set up of an LSP failed at a downstream node which returned an ERROR_SPEC to the PCC, the ERROR_SPEC MUST be included in the LSP State Report. Depending on whether RSVP signaling was performed over IPv4 or IPv6, the LSP Object will contain an IPV4-ERROR_SPEC TLV or an IPV6-ERROR_SPEC TLV.

The format of the IPV4-RSVP-ERROR-SPEC TLV is shown in the following figure:

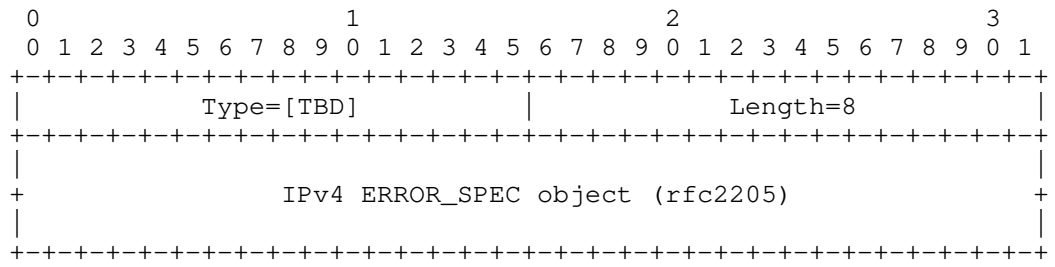


Figure 21: IPV4-RSVP-ERROR-SPEC TLV format

The type of the TLV is [TBD] and it has a fixed length of 8 octets. The value contains the RSVP IPv4 ERROR_SPEC object defined in [RFC2205]. Error codes allowed in the ERROR_SPEC object are defined in [RFC2205] and [RFC3209].

The format of the IPV6-RSVP-ERROR-SPEC TLV is shown in the following figure:

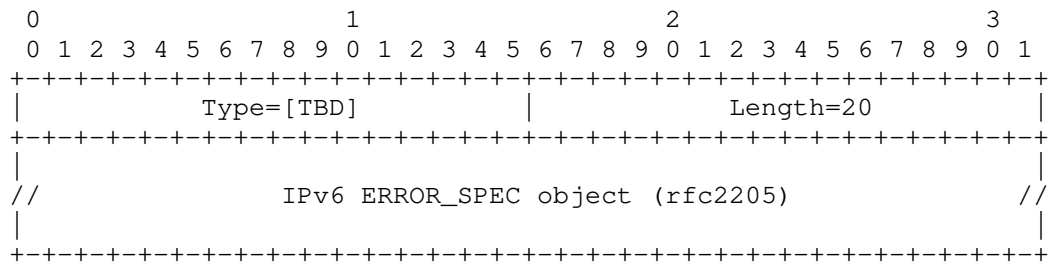


Figure 22: IPV6-RSVP-ERROR-SPEC TLV format

The type of the TLV is [TBD] and it has a fixed length of 20 octets. The value contains the RSVP IPv6 ERROR_SPEC object defined in [RFC2205]. Error codes allowed in the ERROR_SPEC object are defined in [RFC2205] and [RFC3209].

7.2.5. LSP State Database Version TLV

The LSP-DB-VERSION TLV can be included as an optional TLV in the LSP object. The LSP-DB-VERSION TLV is discussed in Section 5.4.1 which covers state synchronization avoidance. The format of the TLV is described in Section 7.1.2, where the details of its use in the OPEN message are listed.

If State Synchronization Avoidance has been enabled on a PCEP session (as described in Section 5.4.1) , a PCC MUST include the LSP-DB-VERSION TLV in each LSP Object sent out on the session. If the TLV

is missing, the PCE will generate an error with error-type 6 (mandatory object missing) and Error Value 12 (LSP-DB-VERSION TLV missing) and close the session. If State Synchronization Avoidance has not been enabled on a PCEP session, the PCC SHOULD NOT include the LSP-DB-VERSION TLV in the LSP Object and the PCE SHOULD ignore it were it to receive one.

Since a PCE does not send LSP updates to a PCC, a PCC should never encounter this TLV. A PCC SHOULD ignore the LSP-DB-VERSION TLV, were it to receive one from a PCE.

7.2.6. Delegation Parameters TLVs

Multiple delegation parameters, such as sub-delegation permissions, authentication parameters, etc. need to be communicated from a PCC to a PCE during the delegation operation. Delegation parameters will be carried in multiple delegation parameter TLVs, which will be defined in future revisions of this document.

8. IANA Considerations

This document requests IANA actions to allocate code points for the protocol elements defined in this document. Values shown here are suggested for use by IANA.

8.1. PCEP Messages

This document defines the following new PCEP messages:

Value	Meaning	Reference
10	Report	This document
11	Update	This document

8.2. PCEP Objects

This document defines the following new PCEP Object-classes and Object-values:

Object-Class Value	Name	Reference
32	LSP Object-Type 1	This document

8.3. LSP Object

This document requests that a registry is created to manage the Flags field of the LSP object. New values are to be assigned by Standards Action [RFC5226]. Each bit should be tracked with the following qualities:

- Bit number (counting from bit 0 as the most significant bit)
- Capability description
- Defining RFC

The following values are defined in this document:

Bit	Description	Reference
28	Remove	This document
29	Operational	This document
30	SYNC	This document
31	Delegate	This document

8.4. PCEP-Error Object

This document defines new Error-Type and Error-Value for the following new error conditions:

Error-Type	Meaning
6	Mandatory Object missing
	Error-value=8: LSP Object missing
	Error-value=9: ERO Object missing for a path in an LSP Update Request where TE-LSP setup is requested
	Error-value=10: BANDWIDTH Object missing for a path in an LSP Update Request where TE-LSP setup is requested
	Error-value=11: LSPA Object missing for a path in an LSP Update Request where TE-LSP setup is requested
	Error-value=12: LSP-DB-VERSION TLV missing
19	Invalid Operation
	Error-value=1: Attempted LSP Update Request for a non-delegated LSP. The PCEP-ERROR Object is followed by the LSP Object that identifies the LSP.

	Error-value=2:	Attempted LSP Update Request if active stateful PCE capability was not negotiated active PCE.
20	LSP State synchronization error.	
	Error-value=1:	A PCE indicates to a PCC that it can not process (an otherwise valid) LSP State Report. The PCEP-ERROR Object is followed by the LSP Object that identifies the LSP.
	Error-value=2:	LSP Database version mismatch.
	Error-value=3:	The LSP-DB-VERSION TLV Missing when State Synchronization Avoidance enabled.

8.5. PCEP TLV Type Indicators

This document defines the following new PCEP TLVs:

Value	Meaning	Reference
16	STATEFUL-PCE-CAPABILITY	This document
17	LSP-SYMBOLIC-NAME	This document
18	IPV4-LSP-IDENTIFIERS	This document
19	IPV6-LSP-IDENTIFIERS	This document
20	LSP-UPDATE-ERROR-CODE	This document
21	IPV4-RSVP-ERROR-SPEC	This document
22	IPV6-RSVP-ERROR-SPEC	This document
23	LSP-DB-VERSION	This document

8.6. STATEFUL-PCE-CAPABILITY TLV

This document requests that a registry is created to manage the Flags field in the STATEFUL-PCE-CAPABILITY TLV in the OPEN object. New values are to be assigned by Standards Action [RFC5226]. Each bit should be tracked with the following qualities:

- Bit number (counting from bit 0 as the most significant bit)
- Capability description
- Defining RFC

The following values are defined in this document:

Bit	Description	Reference
30	INCLUDE-DB-VERSION	This document
31	LSP-UPDATE-CAPABILITY	This document

8.7. LSP-UPDATE-ERROR-CODE TLV

This document requests that a registry is created to manage the Error Codes in the LSP-UPDATE-ERROR-CODE TLV. New values are to be assigned by Standards Action [RFC5226].

The following Error Codes are defined in this document:

Value	Meaning	Reference
1	LSP Setup failed outside of the node. Must be followed by the RSVP-ERROR-SPEC TLV, which indicates the failure cause.	This document
2	LSP not operational	This document

9. Manageability Considerations

All manageability requirements and considerations listed in [RFC5440] apply to PCEP protocol extensions defined in this document. In addition, requirements and considerations listed in this section apply.

9.1. Control Function and Policy

In addition to configuring specific PCEP session parameters, as specified in [RFC5440], Section 8.1, a PCE or PCC implementation MUST allow configuring the stateful PCEP capability and the LSP Update capability. A PCC implementation SHOULD allow the operator to specify multiple candidate PCEs for and a delegation preference for each candidate PCE. A PCC SHOULD allow the operator to specify an LSP delegation policy where LSPs are delegated to the most-preferred online PCE. A PCC MAY allow the operator to specify different LSP delegation policies.

A PCC implementation which allows concurrent connections to multiple PCEs SHOULD allow the operator to group the PCEs by administrative domains and it MUST NOT advertise LSP existence and state to a PCE if the LSP is delegated to a PCE in a different group.

A PCC implementation SHOULD allow the operator to specify whether the PCC will advertise LSP existence and state for LSPs that are not

controlled by any PCE (for example, LSPs that are statically configured at the PCC).

A PCC implementation SHOULD allow the operator to specify the Delegation Timeout Interval. The default value of the Delegation Timeout Interval SHOULD be set to 30 seconds.

When an LSP can no longer be delegated to a PCE, after the expiration of the Delegation Timeout Interval, the LSP MAY either: 1) retain its current parameters or 2) revert to operator-defined default LSP parameters. This behavior SHOULD be configurable and in the case when (2) is supported, a PCC implementation MUST allow the operator to specify the default LSP parameters.

A PCC implementation SHOULD allow the operator to specify delegation priority for PCEs. This effectively defines the primary PCE and one or more backup PCEs to which primary PCE's LSPs can be delegated when the primary PCE fails.

Policies defined for stateful PCEs and PCCs should eventually fit in the Policy-Enabled Path Computation Framework defined in [RFC5394], and the framework should be extended to support Stateful PCEs.

9.2. Information and Data Models

PCEP session configuration and information in the PCEP MIB module SHOULD be extended to include negotiated stateful capabilities, synchronization status, and delegation status (at the PCC list PCEs with delegated LSPs).

9.3. Liveness Detection and Monitoring

PCEP protocol extensions defined in this document do not require any new mechanisms beyond those already defined in [RFC5440], Section 8.3.

9.4. Verifying Correct Operation

Mechanisms defined in [RFC5440], Section 8.4 also apply to PCEP protocol extensions defined in this document. In addition to monitoring parameters defined in [RFC5440], a stateful PCC-side PCEP implementation SHOULD provide the following parameters:

- Total number of LSP updates
- Number of successful LSP updates

- Number of dropped LSP updates
- Number of LSP updates where LSP setup failed

A PCC implementation SHOULD provide a command to show to which PCEs LSPs are delegated.

A PCC implementation SHOULD allow the operator to manually revoke LSP delegation.

9.5. Requirements on Other Protocols and Functional Components

PCEP protocol extensions defined in this document do not put new requirements on other protocols.

9.6. Impact on Network Operation

Mechanisms defined in [RFC5440], Section 8.6 also apply to PCEP protocol extensions defined in this document.

Additionally, a PCEP implementation SHOULD allow a limit to be placed on the rate PCUpd and PCRpt messages sent by a PCEP speaker and processed from a peer. It SHOULD also allow sending a notification when a rate threshold is reached.

A PCC implementation SHOULD allow a limit to be placed on the rate of LSP Updates to the same LSP to avoid signaling overload discussed in Section 10.3.

10. Security Considerations

10.1. Vulnerability

This document defines extensions to PCEP to enable stateful PCEs. The nature of these extensions and the delegation of path control to PCEs results in more information being available for a hypothetical adversary and a number of additional attack surfaces which must be protected.

The security provisions described in [RFC5440] remain applicable to these extensions. However, because the protocol modifications outlined in this document allow the PCE to control path computation timing and sequence, the PCE defense mechanisms described in [RFC5440] section 7.2 are also now applicable to PCC security.

As a general precaution, it is RECOMMENDED that these PCEP extensions only be activated on authenticated and encrypted sessions across PCEs

and PCCs belonging to the same administrative authority.

The following sections identify specific security concerns that may result from the PCEP extensions outlined in this document along with recommended mechanisms to protect PCEP infrastructure against related attacks.

10.2. LSP State Snooping

The stateful nature of this extension explicitly requires LSP status updates to be sent from PCC to PCE. While this gives the PCE the ability to provide more optimal computations to the PCC, it also provides an adversary with the opportunity to eavesdrop on decisions made by network systems external to PCE. This is especially true if the PCC delegates LSPs to multiple PCEs simultaneously.

Adversaries may gain access to this information by eavesdropping on unsecured PCEP sessions, and might then use this information in various ways to target or optimize attacks on network infrastructure. For example by flexibly countering anti-DDoS measures being taken to protect the network, or by determining choke points in the network where the greatest harm might be caused.

PCC implementations which allow concurrent connections to multiple PCEs SHOULD allow the operator to group the PCEs by administrative domains and they MUST NOT advertise LSP existence and state to a PCE if the LSP is delegated to a PCE in a different group.

10.3. Malicious PCE

The LSP delegation mechanism described in this document allows a PCC to grant effective control of an LSP to the PCE for the duration of a PCEP session. While this enables PCE control of the timing and sequence of path computations within and across PCEP sessions, it also introduces a new attack vector: an attacker may flood the PCC with PCUpd messages at a rate which exceeds either the PCC's ability to process them or the network's ability to signal the changes, either by spoofing messages or by compromising the PCE itself.

A PCC is free to revoke an LSP delegation at any time without needing any justification. A defending PCC can do this by enqueueing the appropriate PCRpt message. As soon as that message is enqueued in the session, the PCC is free to drop any incoming PCUpd messages without additional processing.

10.4. Malicious PCC

A stateful session also result in increased attack surface by placing a requirement for the PCE to keep an LSP state replica for each PCC. It is RECOMMENDED that PCE implementations provide a limit on resources a single PCC can occupy.

Delegation of LSPs can create further strain on PCE resources and a PCE implementation MAY preemptively give back delegations if it finds itself lacking the resources needed to effectively manage the delegation. Since the delegation state is ultimately controlled by the PCC, PCE implementations SHOULD provide throttling mechanisms to prevent strain created by flaps of either a PCEP session or an LSP delegation.

11. Acknowledgements

We would like to thank Adrian Farrel and Cyril Margaria for their contributions to this document.

We would like to thank Shane Amante, Julien Meuric, Kohei Shiomoto, Paul Schultz and Raveendra Torvi for their comments and suggestions. Thanks also to Dhruv Dhoddy, Oscar Gonzales de Dios, Tomas Janciga, Stefan Kobza and Kexin Tang for helpful discussions.

12. References

12.1. Normative References

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

[RFC2205] Braden, B., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, 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.

[RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.

[RFC5088] Le Roux, JL., Vasseur, JP., Ikejiri, Y., and R. Zhang, "OSPF Protocol Extensions for Path Computation Element

(PCE) Discovery", RFC 5088, January 2008.

[RFC5089] Le Roux, JL., Vasseur, JP., Ikejiri, Y., and R. Zhang, "IS-IS Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5089, January 2008.

[RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.

[RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.

[RFC5511] Farrel, A., "Routing Backus-Naur Form (RBNF): A Syntax Used to Form Encoding Rules in Various Routing Protocol Specifications", RFC 5511, April 2009.

12.2. Informative References

[MPLS-PC] Chaieb, I., Le Roux, JL., and B. Cousin, "Improved MPLS-TE LSP Path Computation using Preemption", Global Information Infrastructure Symposium, July 2007.

[MXMN-TE] Danna, E., Mandal, S., and A. Singh, "Practical linear programming algorithm for balancing the max-min fairness and throughput objectives in traffic engineering", pre-print, 2011.

[NET-REC] Vasseur, JP., Pickavet, M., and P. Demeester, "Network Recovery: Protection and Restoration of Optical, SONET-SDH, IP, and MPLS", The Morgan Kaufmann Series in Networking, June 2004.

[RFC2702] Awduche, D., Malcolm, J., Agogbua, J., O'Dell, M., and J. McManus, "Requirements for Traffic Engineering Over MPLS", RFC 2702, September 1999.

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

[RFC3346] Boyle, J., Gill, V., Hannan, A., Cooper, D., Awduche, D., Christian, B., and W. Lai, "Applicability Statement for Traffic Engineering with MPLS", RFC 3346, August 2002.

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

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

[RFC4657] Ash, J. and J. Le Roux, "Path Computation Element (PCE) Communication Protocol Generic Requirements", RFC 4657, September 2006.

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

[RFC5394] Bryskin, I., Papadimitriou, D., Berger, L., and J. Ash, "Policy-Enabled Path Computation Framework", RFC 5394, December 2008.

[RFC5557] Lee, Y., Le Roux, JL., King, D., and E. Oki, "Path Computation Element Communication Protocol (PCEP) Requirements and Protocol Extensions in Support of Global Concurrent Optimization", RFC 5557, July 2009.

Authors' Addresses

Edward Crabbe
Google, Inc.
1600 Amphitheatre Parkway
Mountain View, CA 94043
US

Email: edc@google.com

Jan Medved
Cisco Systems, Inc.
170 West Tasman Dr.
San Jose, CA 95134
US

Email: jmedved@cisco.com

Robert Varga
Pantheon Technologies LLC
Mlynske Nivy 56
Bratislava 821 05
Slovakia

Email: robert.varga@pantheon.sk

Ina Minei
Juniper Networks, Inc.
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: ina@juniper.net

PCE Working Group
Internet-Draft
Intended status: Standard
Expires: February 25, 2012

D. Dhody
U. Palle
Huawei Technologies India Pvt Ltd
R. Casellas
CTTC - Centre Tecnologic de
Telecomunicacions de Catalunya
August 24, 2011

Standard Representation Of Domain Sequence
draft-dhody-pce-pcep-domain-sequence-01

Abstract

The ability to compute shortest constrained Traffic Engineering Label Switched Paths (TE LSPs) in Multiprotocol Label Switching (MPLS) and Generalized MPLS (GMPLS) networks across multiple domains has been identified as a key requirement for P2P and P2MP scenarios. In this context, a domain is a collection of network elements within a common sphere of address management or path computational responsibility such as an IGP area or an Autonomous Systems. This document specifies a standard representation of domain sequence that can be utilized in all PCE deployment scenarios.

This Internet-Draft will expire on February 25, 2012.

Table of Contents

1. Introduction

RFC 5441 [A Backward-Recursive PCE-Based Computation (BRPC) Procedure to Compute Shortest Constrained Inter-Domain Traffic Engineering Label Switched Paths] mentions -

"The sequence of domains to be traversed is either administratively predetermined or discovered by some means that is outside of the scope of this document. The PCC MAY indicate the sequence of domains to be traversed using the Include Route Object (IRO) defined in [RFC5440] so that it is available to all PCEs."

This document proposes a standard way to represent domain sequence using IRO in various deployment scenarios.

It further gives examples of various deployment scenario including P2P, P2MP and HPCE.

The domain sequence (the set of domains traversed to reach the destination domain) is either administratively predetermined or discovered by some means (H-PCE) that is outside of the scope of this document. Here the focus is only on a standard representation of the domain sequence in all possible scenarios.

1.1. Requirements Language

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

2. Terminology

The following terminology is used in this document.

ABR: OSPF Area Border Router. Routers used to connect two IGP areas.

AS: Autonomous System.

ASBR: Autonomous System Boundary Router.

BN: Boundary Node, Can be an ABR or ASBR.

BRPC: Backward Recursive Path Computation

Domain: Any collection of network elements within a common sphere of address management or path computational responsibility. Examples of domains include Interior Gateway Protocol (IGP) areas and Autonomous Systems (ASs).

Domain-Seq: The sequence of domains for a path.

ERO: Explicit Route Object

H-PCE: Hierarchical PCE

IGP: Interior Gateway Protocol. Either of the two routing protocols, Open Shortest Path First (OSPF) or Intermediate System to Intermediate System (IS-IS).

IRO: Include Route Object

IS-IS: Intermediate System to Intermediate System.

OSPF: Open Shortest Path First.

PCC: Path Computation Client: any client application requesting a path computation to be performed by a Path Computation Element.

PCE: Path Computation Element. An entity (component, application, or network node) that is capable of computing a network path or route based on a network graph and applying computational constraints.

P2MP: Point-to-Multipoint

P2P: Point-to-Point

TE LSP: Traffic Engineering Label Switched Path.

3. Detail Description

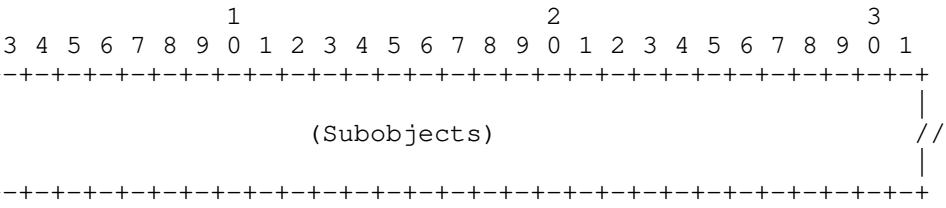
3.1. Domains

A domain is any collection of network elements within a common sphere of address management or path computation responsibility. Examples of domains include IGP areas or Autonomous Systems (ASes). To uniquely identify a domain in the domain sequence both AS and Area-id is important.

3.2. Standard Representation

The IRO (Include Route Object) is used to specify the domain sequence that the computed inter-domain path MUST traverse.

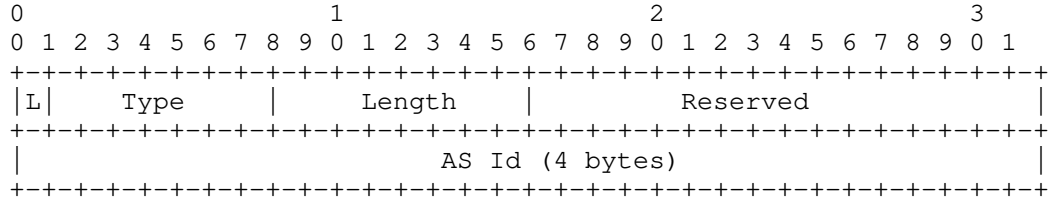
IRO Object-Class is 10.
IRO Object-Type is 1.



Sub-objects: The IRO is made of sub-objects.
The following sub-object types are used.

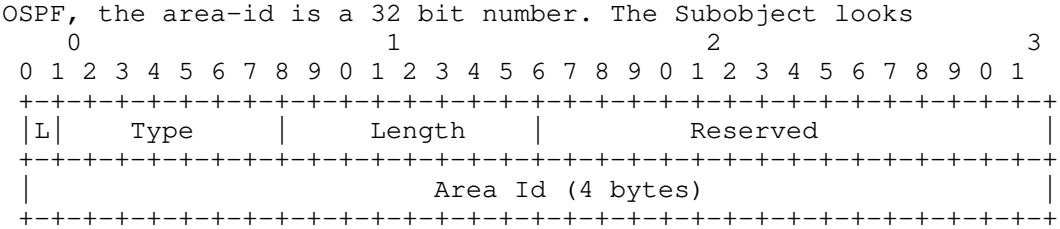
Type	Sub-object
32	Autonomous system number (2 Byte) [RFC 3209]
TBD	Autonomous system number (4 Byte)
TBD	OSPF Area id
TBD	ISIS Area id

[RFC 3209] define 2 octet AS number.
To support 4 octet AS number [RFC4893] following subobject is defined:



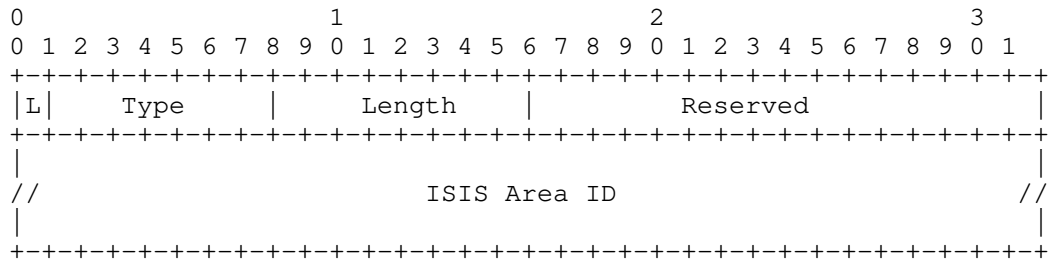
Since the length of Area-id is different for OSPF and ISIS, we propose different sub-objects.

For OSPF, the area-id is a 32 bit number. The Subobject looks



The length if fixed.

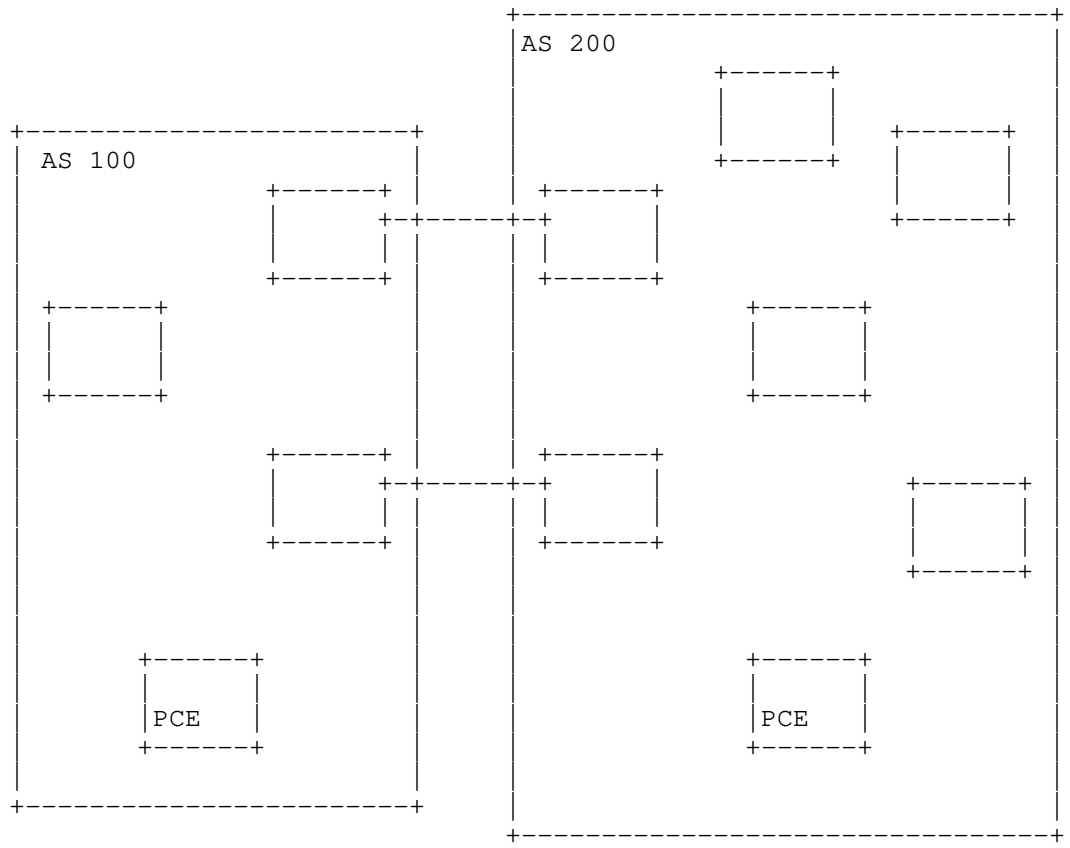
For ISIS, the area-id is of variable length and thus the length of the Subobject is variable. The Area-id is as described in ISIS by ISO standard. The Length MUST be at least 4, and MUST be a multiple of 4.



3.3. Deployment Scenarios

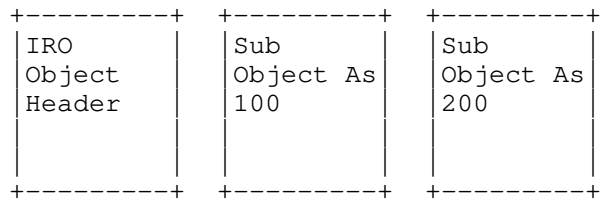
3.3.1. Only AS

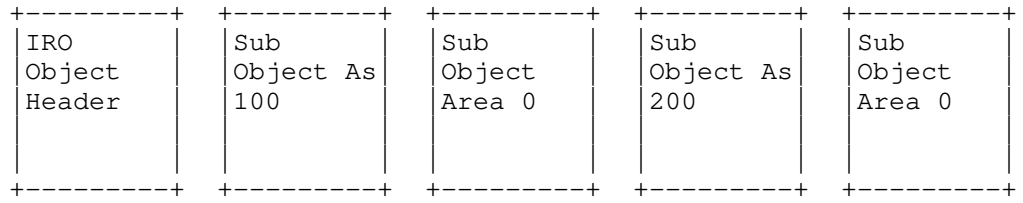
Considering each AS to be made of a single area, in this scenario the area MAY be skipped in the domain sequence. The domain sequence could be represented with just AS numbers.



Both AS are made of Area 0.

This could be represented as <IRO> as:

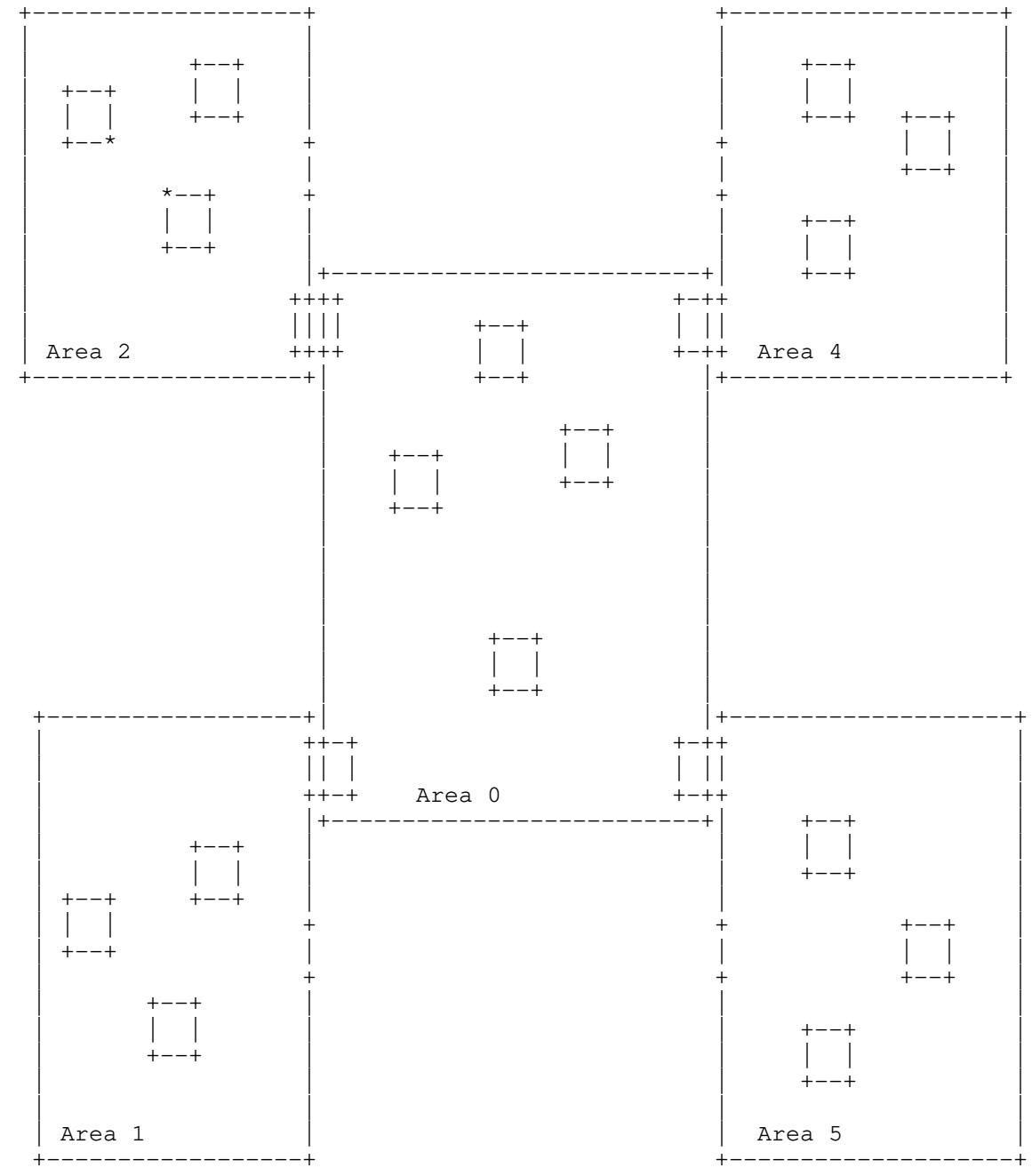




Area is optional and it MAY be skipped. PCE should be able to understand both notations.

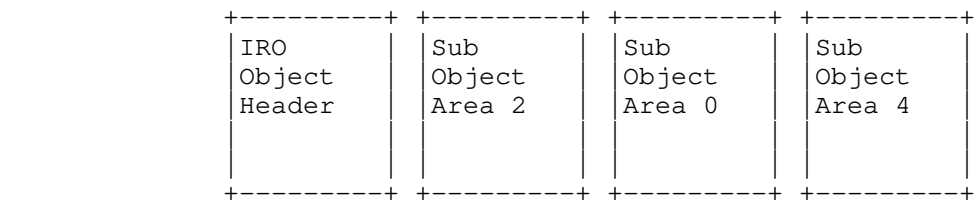
3.3.2. Only Area

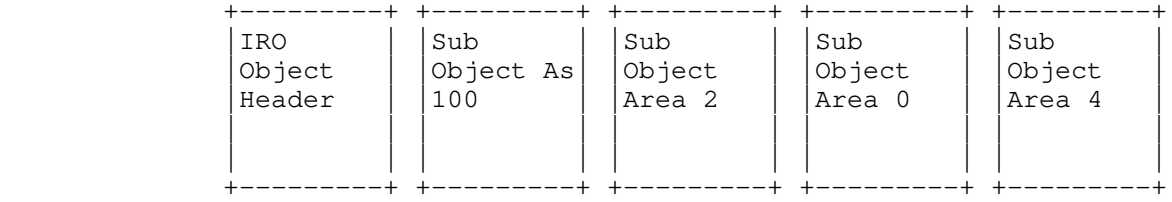
Consider a case where both end of LSP belong to different area but within the same AS, this could be represented in domain sequence using the AREA sub-object. AS number MAYBE skipped.



AS Number is 100.

This could be represented as <IRO> as:

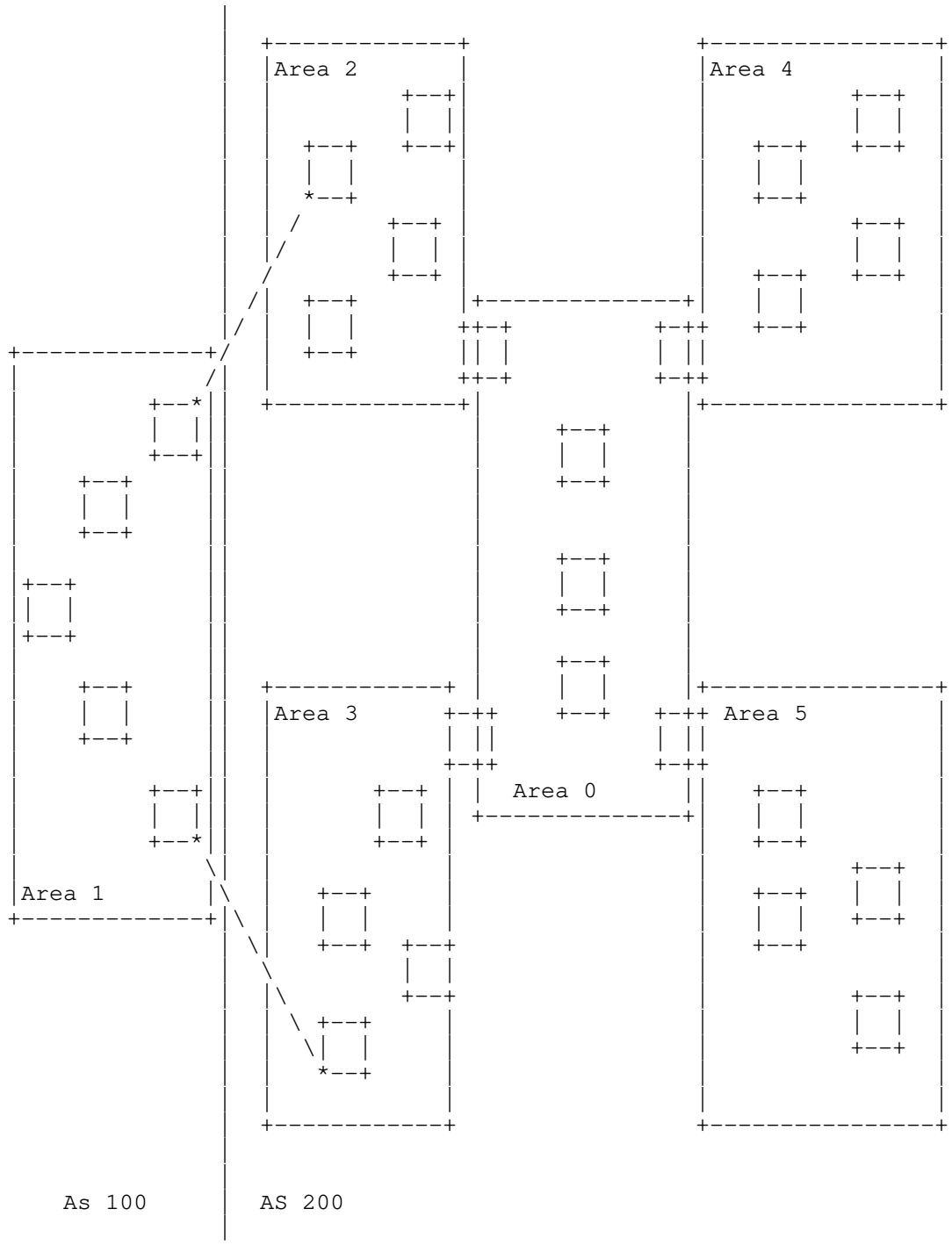




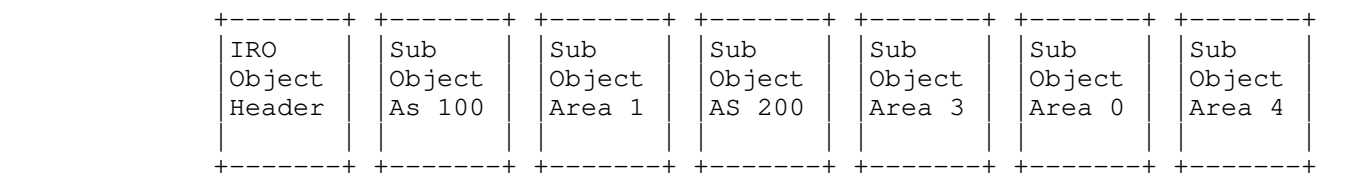
AS is optional and it MAY be skipped. PCE should be able to understand both notations.

3.3.3. Mix of AS and Area

In inter-AS case where an AS is further made up of multiple areas, both AS number and area should be a part of domain sequence.


Area 2
Area 4
Area 3
Area 5
Area 0
Area 1
As 100
AS 200

The domain sequence can be carried in IRO as shown below:



Combination of both AS and Area uniquely identify a domain in the domain sequence.

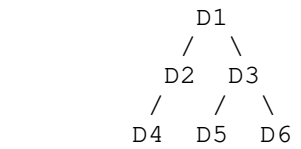
3.3.4. PCE serving multiple domains

A single PCE maybe responsible for multiple domains; for example PCE function deployed on an ABR. Domain sequence should have no impact on this. PCE which can support 2 adjacent domains can internally handle this situation without any impact on the neighboring domains.

3.3.5. P2MP

In case of P2MP the path domain tree is nothing but a series of Domain-Seq, as shown in the below figure:

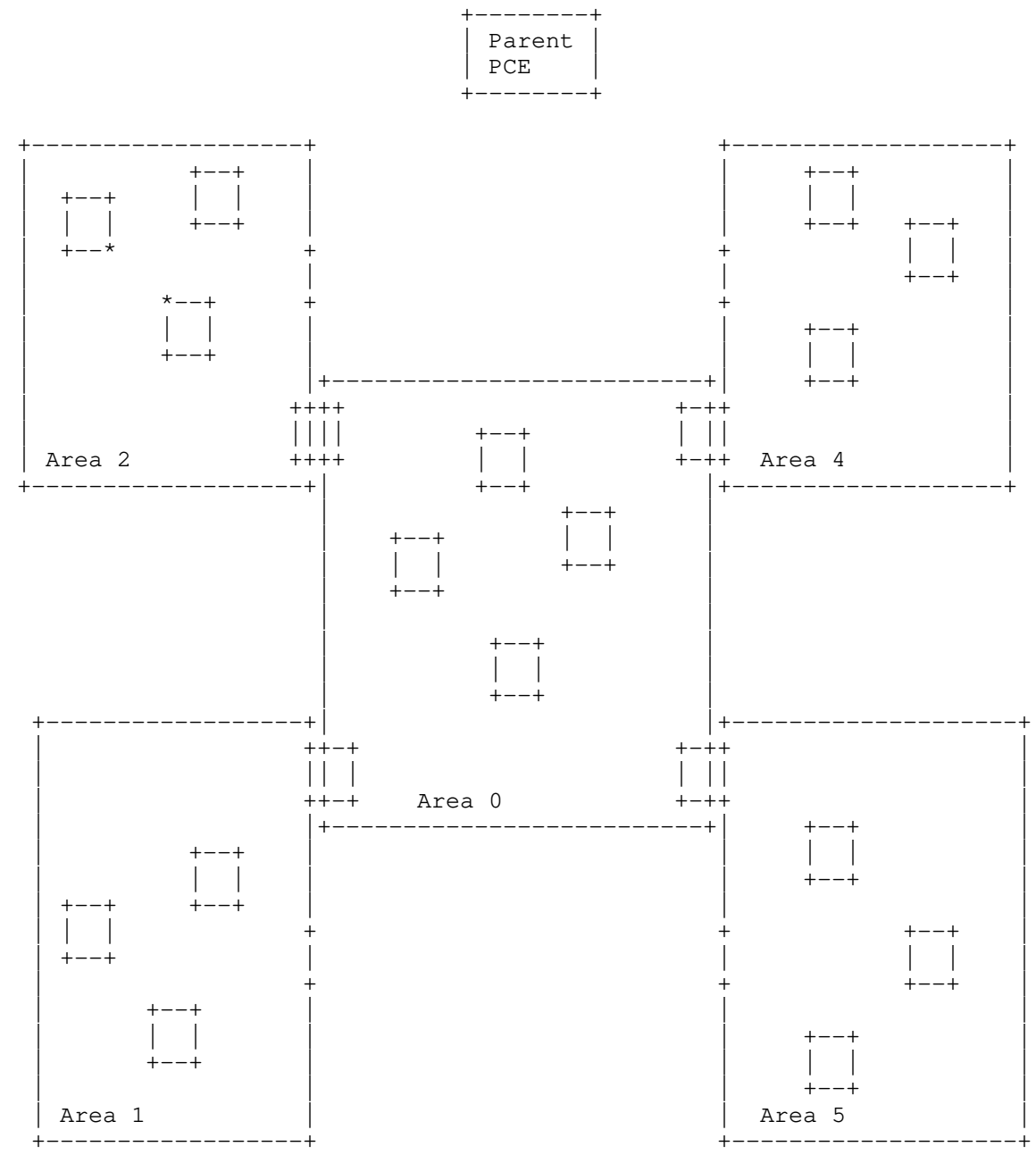
D1-D3-D6, D1-D3-D5 and D1-D2-D4.



The same domain sequence are carried in IRO as explained above.

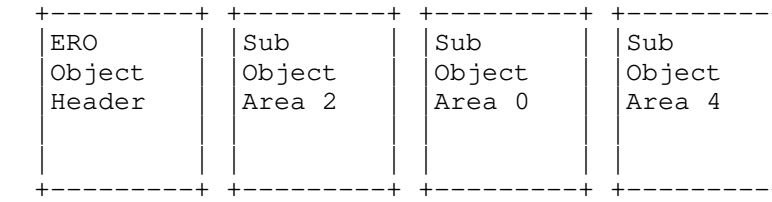
3.3.6. HPCE

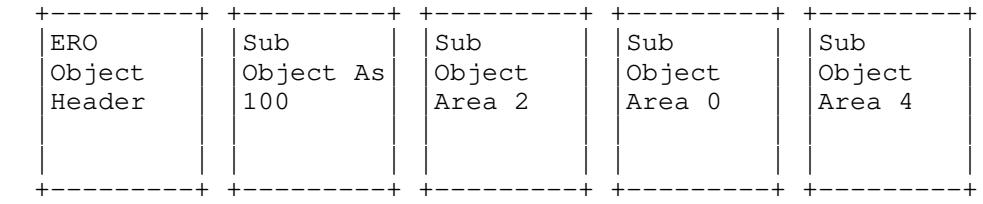
Consider a case as shown below consisting of child and parent PCE.



In HPCE implementation PCE(1) can request the parent PCE to determine the domain path and return in the PCRep in form of ERO. The Subobject would be AS and Area (OSPF/ISIS). So in this case, the

reply would carry the result as





3.3.7. Domain Seq V/s PCE Seq

[PCE-P2MP-PROCEDURES] introduces the concept of PCE-Sequence, where a sequence of PCE based on the domain sequence should be decided and attached in the PCReq at the very beginning of path computation. It is much simpler and advantageous to carry only domain-sequence rather than PCE-Sequence.

Advantages

- All PCE must be aware of all other PCEs in all domain for PCE-Sequence. There is no clear method for this. In domain-sequence PCE should be aware of the domains and not all the PCEs serving the domain. PCE needs to be aware of the neighboring PCEs as done by discovery protocols.

- There maybe multiple PCE in a domain, the selection of PCE shouldn't be made at the PCC/PCE(1). This decision is made only at the neighboring PCE which is completely aware of states of PCE via notification messages.

- Domain sequence would be compatible to P2P inter-domain BRPC method as described in RFC 5441.

There is no need for PCE-Sequence and it doesn't give any benefits over Domain Seq.

4. IANA Considerations

IANA has defined a registry for OSPF and ISIS Area sub-object.

Type	Sub-object
TBD	AS Number (4 Byte)
TBD	OSPF Area id
TBD	ISIS Area id

5. Security Considerations

This document specifies a standard representation of domain sequence, which is used in all inter-domain PCE scenarios as explained in other RFC and drafts. It does not introduce any new security considerations.

6. Manageability Considerations

TBD

7. Acknowledgments

We would like to thank Pradeep Shastry, Suresh babu, Quintin Zhao and Chen Huaimo for their useful comments and suggestions.

8. References

8.1. Normative References

[ISO] "Intermediate System to Intermediate System Intra-Domain Routeing Exchange Protocol for use in Conjunction with the Protocol for Providing the Connectionless-mode Network Service" ISO/IEC 10589:2002 Second Edition

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

8.2. Informative References

[PCE-HIERARCHY-FWK] 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", September 2010.

[PCE-P2MP-PROCEDURES] Zhao, Q., Ali, Z., Saad,, T., and D. King, "PCE-based Computation Procedure To Compute Shortest Constrained P2MP Inter-domain Traffic Engineering Label Switched Paths", January 2011.

[RFC 3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels",

December 2001.

[RFC 4893] Vohra, Q. and E. Chen, "BGP Support for Four-octet AS Number Space", May 2007.

[RFC5440] Ayyangar, A ., Farrel, A ., Oki, E., Atlas, A., Dolganow, A., Ikejiri, Y., Kumaki, K., Vasseur, J., and J. Roux, "Path Computation Element (PCE) communication Protocol (PCEP)", March 2009.

[RFC5441] Vasseur, JP., Zhang, R., Bitar, N., and JL. Le Roux, "A Backward-Recursive PCE-Based Computation (BRPC) Procedure to Compute Shortest Constrained Inter-Domain Traffic Engineering Label Switched Paths", April 2009.

Authors' Addresses

Dhruv Dhody
Huawei Technologies India Pvt Ltd
Leela Palace
Bangalore, Karnataka 560008
INDIA

EMail: dhruvd@huawei.com

Udayasree Palle
Huawei Technologies India Pvt Ltd
Leela Palace
Bangalore, Karnataka 560008
INDIA

EMail: udayasreepalle@huawei.com

Ramon Casellas
CTTC - Centre Tecnologic de Telecomunicacions de Catalunya
Av. Carl Friedrich Gauss n7
Castelldefels, Barcelona 08860
SPAIN

EMail: ramon.casellas@cttc.es

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 12, 2012

D. Dhody
U. Palle
Huawei Technologies India Pvt Ltd
R. Casellas
CTTC - Centre Tecnologic de Telecomunicacions de Catalunya

February 9, 2012

Standard Representation Of Domain Sequence
draft-dhody-pce-pcep-domain-sequence-02

Abstract

The ability to compute shortest constrained Traffic Engineering Label Switched Paths (TE LSPs) in Multiprotocol Label Switching (MPLS) and Generalized MPLS (GMPLS) networks across multiple domains has been identified as a key requirement for P2P and P2MP scenarios. In this context, a domain is a collection of network elements within a common sphere of address management or path computational responsibility such as an IGP area or an Autonomous Systems. This document specifies a standard representation and encoding of a domain sequence, which is defined as an ordered sequence of domains traversed to reach the destination domain. This document also defines new sub-objects to be used to encode domain identifiers.

Table of Contents

1. Introduction

A PCE may be used to compute end-to-end paths across multi-domain environments using a per-domain path computation technique [RFC5152]. The so called backward recursive path computation (BRPC) mechanism [RFC5441] defines a PCE-based path computation procedure to compute inter-domain constrained (G)MPLS TE LSPs. However, both per-domain and BRPC techniques assume that the sequence of domains to be crossed from source to destination is known, either fixed by the network operator or obtained by other means. For inter-domain point-to-multi-point (P2MP) tree, [PCE-P2MP-PROCEDURES] assumes the domain-tree is known.

The list of domains in a point-to-point (P2P) path or a point-to-multi-point (P2MP) tree is usually a constraint in the path computation request. The PCE decouples the domain to determine the next PCE to forward the request.

According to BRPC mechanism the PCC MAY indicate the sequence of domains to be traversed using the Include Route Object (IRO) defined in [RFC5440].

This document proposes a standard way to represent and encode a domain sequence using IRO in various deployment scenarios including P2P, P2MP and Hierarchical PCE (HPCE) [PCE-HIERARCHY-FWK].

The domain sequence (the set of domains traversed to reach the destination domain) is either administratively predetermined or discovered by some means (H-PCE) that is outside of the scope of this document. Here the focus is only on a standard representation of the domain sequence in all possible scenarios.

1.1. Requirements Language

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

2. Terminology

The following terminology is used in this document.

ABR: OSPF Area Border Router. Routers used to connect two IGP areas.

AS: Autonomous System.

ASBR: Autonomous System Boundary Router.

BN: Boundary Node, Can be an ABR or ASBR.

BRPC: Backward Recursive Path Computation

Domain: Any collection of network elements within a common sphere of address management or path computational responsibility. Examples of domains include Interior Gateway Protocol (IGP) areas and Autonomous Systems (ASs).

Domain-Seq: An ordered sequence of domains traversed to reach the destination domain.

ERO: Explicit Route Object

H-PCE: Hierarchical PCE

IGP: Interior Gateway Protocol. Either of the two routing protocols, Open Shortest Path First (OSPF) or Intermediate System to Intermediate System (IS-IS).

IRO: Include Route Object

IS-IS: Intermediate System to Intermediate System.

OSPF: Open Shortest Path First.

PCC: Path Computation Client: any client application requesting a path computation to be performed by a Path Computation Element.

PCE: Path Computation Element. An entity (component, application, or network node) that is capable of computing a network path or route based on a network graph and applying computational constraints.

P2MP: Point-to-Multipoint

P2P: Point-to-Point

TE LSP: Traffic Engineering Label Switched Path.

3. Detail Description

3.1. Domains

A domain can be defined as a separate administrative or geographic environment within the network. A domain may be further defined as a zone of routing or computational ability. Under these definitions a domain might be categorized as an Antonymous System (AS) or an Interior Gateway Protocol (IGP) area (as per [RFC4726] and [RFC4655]). To uniquely identify a domain in the domain sequence both AS and Area-id is important.

3.2. Domain-Sequence

A domain-sequence is an ordered sequence of domains traversed to reach the destination domain. In this context a Domain could be an Autonomous System (AS) or an IGP Area. Note that an AS can be further made of multiple Area.

Domain Sequence can be applied as a constraint and carried in path computation request to PCE(s). In case of HPCE [PCE-HIERARCHY-FWK] Parent PCE MAY send the domain sequence as a result in path computation reply.

In this context, ordered sequence is important, in a P2P path, the domains listed appear in the order that they are crossed. In a P2MP path, the domain tree is represented as list of domain sequences.

One main goal of the Domain-Sequence is to enable a PCE to select the next PCE to forward the path computation request based on the domain information.

A PCC or PCE MAY add an additional constraints covering which Boundary Nodes (ABR or ASBR) or Border links (Inter-AS-link) MUST be traversed while defining a domain sequence.

Thus a Domain-Sequence MAY be made up of one or more of -

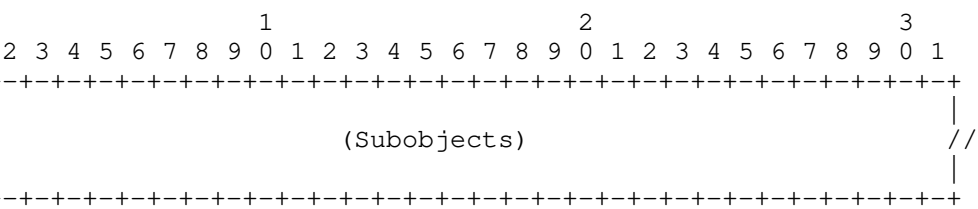
- AS Number
- Area ID
- Boundary Node ID
- Inter-AS-Link Address

3.3. Standard Representation

The IRO (Include Route Object) [RFC5440] is an optional object used to specify a set of specified network elements that the computed path MUST traverse. [RFC5440] in its description of IRO does not constrain the sub-objects to be in a given particular order. When considering a domain sequence, the domain relative ordering is a basic criterion and, as such, this document specifies a new IRO object type.

We define a new type of IRO Object to define Domain Sequence.

IRO Object-Class is 10.
IRO Object-Type is TBD. (2 suggested value to IANA)



Sub-objects: The IRO is made of sub-objects identical to the ones defined in [RFC3209], [RFC3473], and [RFC3477], where the IRO sub-object type is identical to the sub-object type defined in the related documents. Some new sub-objects related to Domain-Sequence are also added in this document.

The following sub-object types are used.

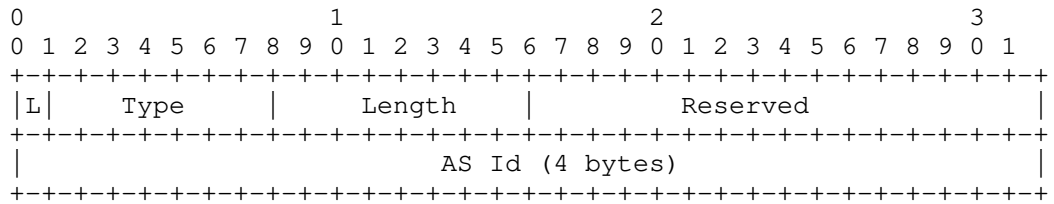
Type	Sub-object
1	IPv4 prefix
2	IPv6 prefix
4	Unnumbered Interface ID
32	Autonomous system number (2 Byte)
TBD	Autonomous system number (4 Byte)
TBD	OSPF Area id
TBD	ISIS Area id

[RFC3209] defines sub-objects for IPv4, IPv6 and unnumbered Interface ID, which in the context of domain-sequence is used to specify Boundary Node (ABR/ASBR) and Inter-AS-Links.

[RFC3209] also defines 2 octet AS number.

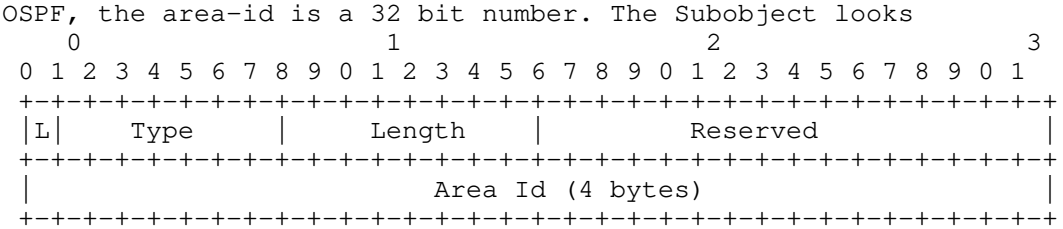
To support 4 octet AS number [RFC4893] following subobject is

defined:



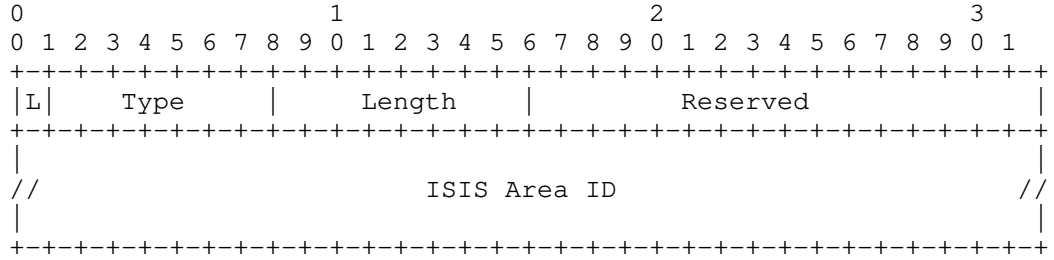
Since the length of Area-id is different for OSPF and ISIS, we propose different sub-objects.

For OSPF, the area-id is a 32 bit number. The Subobject looks



The length if fixed.

For ISIS, the area-id is of variable length and thus the length of the Subobject is variable. The Area-id is as described in ISIS by ISO standard [ISO 10589]. The Length MUST be at least 4, and MUST be a multiple of 4.



The above sub-objects in various combinations can be used to encode the domain-sequence. When the domain-sequence is used as a constraint in path computation request it is carried in IRO Domain Sequence Object Type. The same sub-objects and their encoding can be used in ERO and path reply message when the domain sequence is computed from Parent PCE.

All other rules of PCEP objects and message processing is as per

[RFC5440].

3.4. Mode of Operation

A domain sequence IRO object constraints or defines the domains involved in a muti-domain path computation, typically involving two or more collaborative PCEs.

Consequently, a Domain-Sequence can be used:

1. by a PCE in order to discover or select the next PCE in a collaborative path computation, such as in BRPC [RFC5441];

2. by the Parent PCE to return the domain sequence when unknown, this can further be an input to BRPC procedure;

3. By a PCC (or PCE) to constraint the domains used in a H-PCE path computation, explicitly specifying which domains to be expanded;

A domain sequence can have varying degrees on granularity; it is possible to have a domain sequence composed of, uniquely, AS identifiers. It is also possible to list the involved areas for a given AS.

In any case, the mapping between domains and responsible PCEs is not defined in this document. It is assumed that a PCE that needs to obtain a "next PCE" from a domain sequence is able to do so (e.g. via administrative configuration, or discovery).

The following algorithm can be applied to select the next domain and, if need be, the PCE responsible for that domain. Note the PCC select the PCE(1) based on its own domain information.

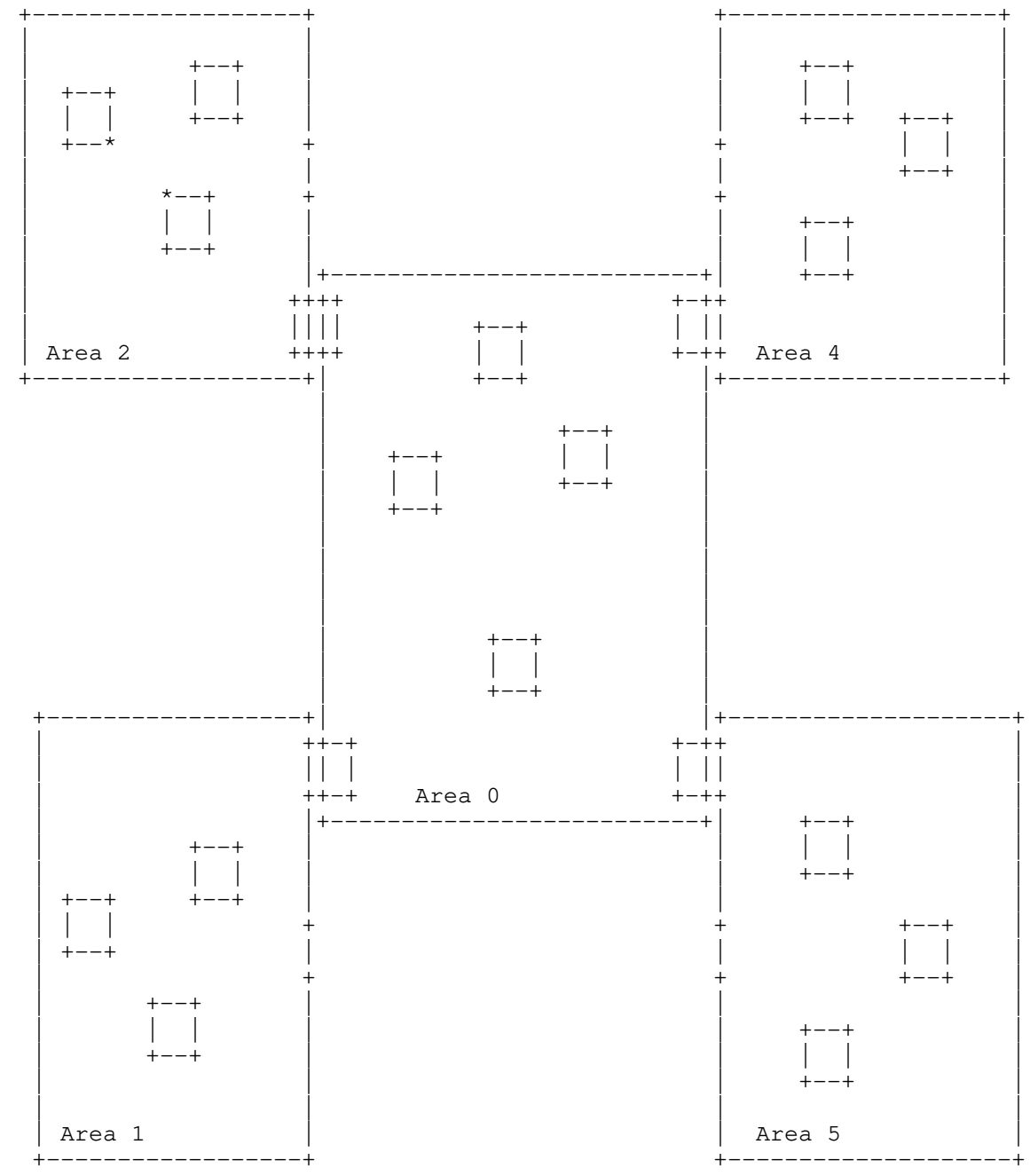
```
START
Get the first Sub-Object S1 from the Domain-Sequence
IF S1's Type is Area (OSPF or ISIS)
   IF S1's Domain is same as current PCE's Area
      Remove S1 from Domain-Sequence and Goto START
   ELSE
      Find the next PCE based on S1's Area within the AS
   ENDIF
ELSEIF S1's Type is AS (2 or 4 Byte)
   IF S1's Domain is same as current PCE's AS
      Remove S1 from Domain-Sequence and Goto START
   ELSE
     Get the next Sub-Object S2 from the Domain-Sequence
     IF the S2 is NULL or S2's type is AS
        Find the next PCE based on S1's Domain (AS) only
     ELSEIF S1's Type is Area
        Find the next PCE based on S1's Domain (AS)
        and S2's Domain (Area)
     ELSE
     ENDIF
   ENDIF
ENDIF
IF Domain-Sequence is empty or next PCE is not found
   Send PCRep with NO-Path
ENDIF
```

If the Sub-Object is of other type representing Boundary Node or Inter-As-Link, it is not used to select the next PCE, but used only while applying BRPC or any other inter-domain procedure.

3.5. Examples

3.5.1. Inter-Area Path Computation

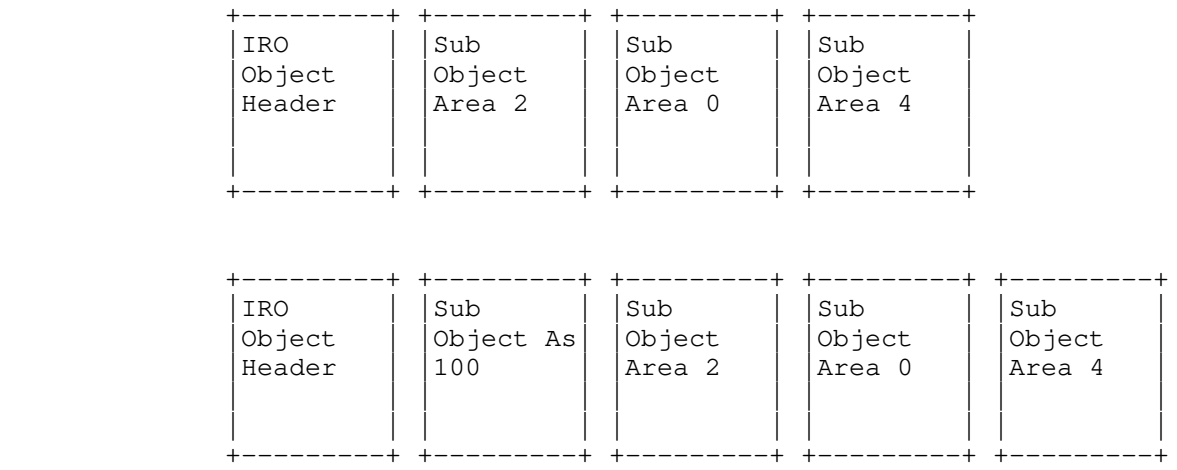
In an inter-area path computation where ingress and egress belong to different IGP area, the domain sequence MAYBE represented using a ordered list of AREA sub-objects. AS number MAYBE skipped, as area information is enough to select the next PCE.



AS Number is 100.

Figure 1: Inter-Area Path Computation

This could be represented as <IRO> as:



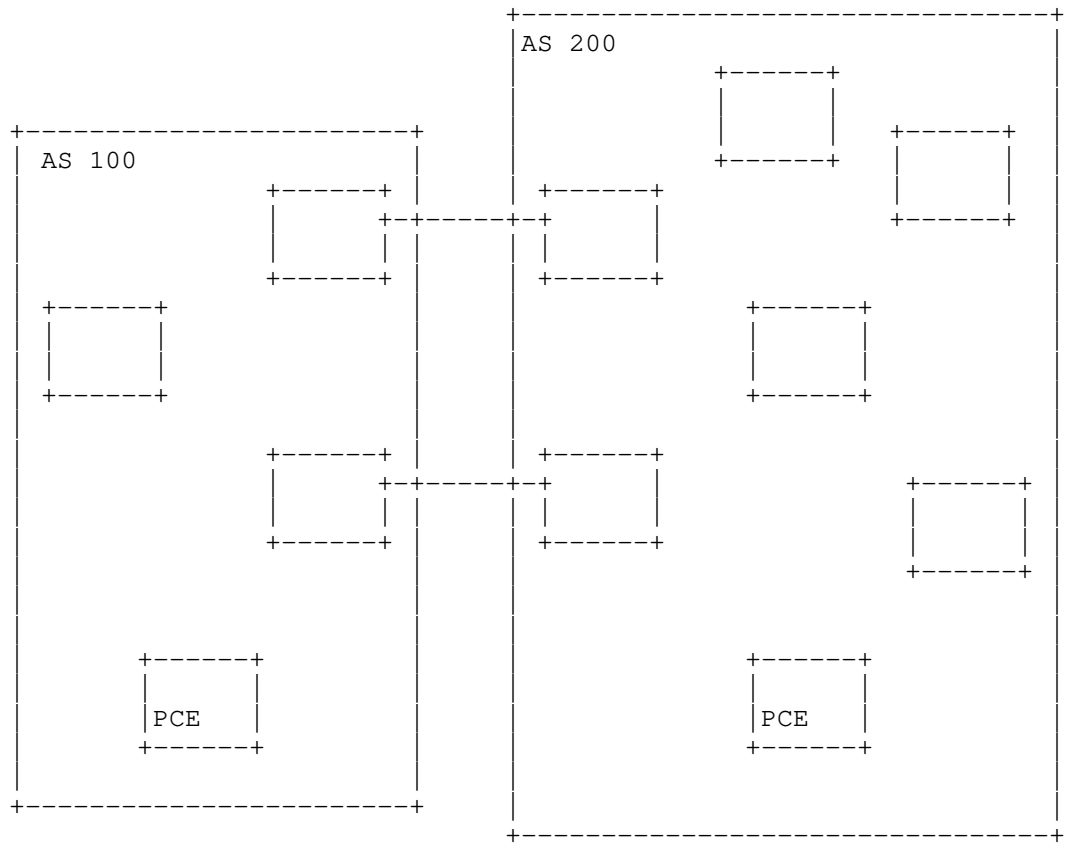
AS is optional and it MAY be skipped. PCE should be able to understand both notations.

3.5.2. Inter-AS Path Computation

In inter-AS path computation, where ingress and egress belong to different AS, the domain sequence is represented using an ordered list of AS sub-objects. The domain sequence MAY further include decomposed area information in AREA sub-objects.

3.5.2.1. Example 1

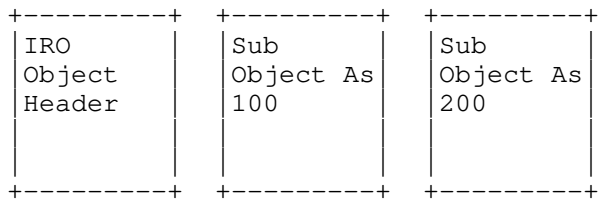
As shown in Figure 2, where AS to be made of a single area, the area subobject MAY be skipped in the domain sequence as AS is enough to uniquely identify the next domain and PCE.

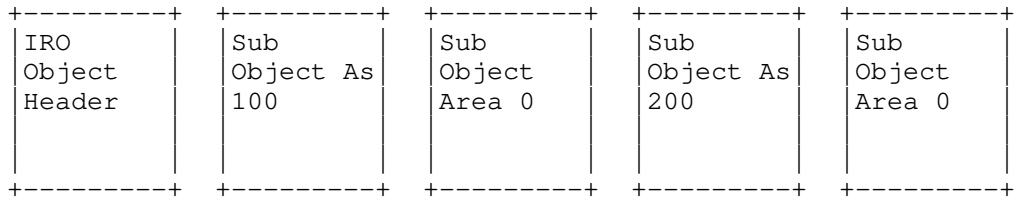


Both AS are made of Area 0.

Figure 2: Inter-AS Path Computation

This could be represented as <IRO> as:





Area is optional and it MAY be skipped. PCE should be able to understand both notations.

3.5.2.2. Example 2

As shown in Figure 3, where AS 200 is made up of multiple areas and multiple domain-sequence exist, PCE MAY include both AS and AREA subobject to uniquely identify the next domain and PCE.

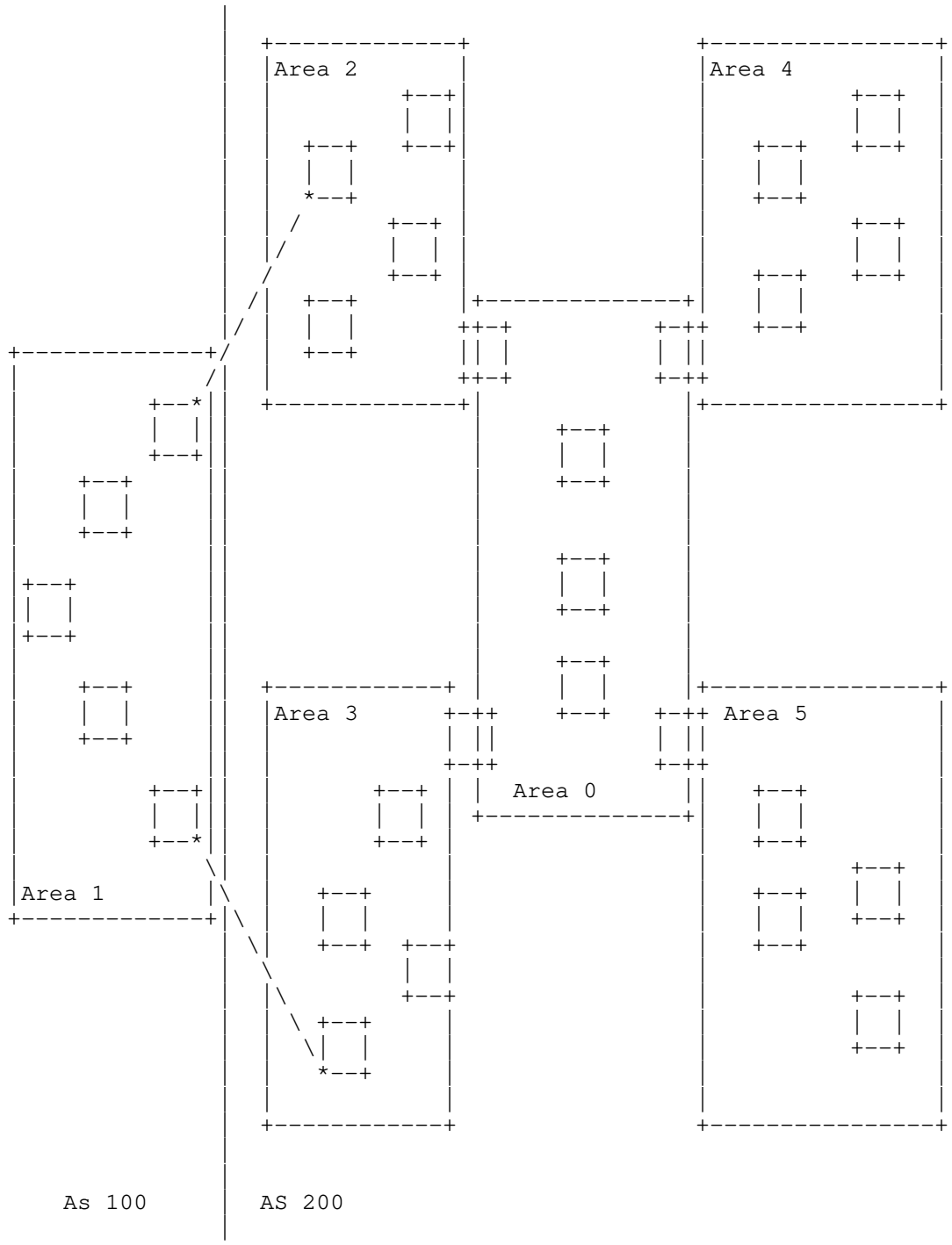
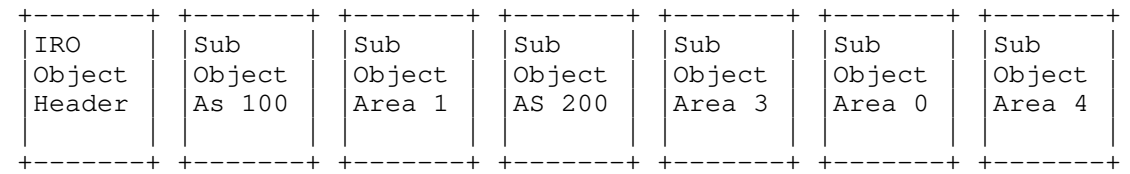

Area 2
Area 4
Area 3
Area 5
Area 0
Area 1
As 100
AS 200

Figure 3: Inter-AS Path Computation

The domain sequence can be carried in IRO as shown below:



Combination of both AS and Area uniquely identify a domain in the domain sequence.

Note that an Area domain identifier always belongs to the previous AS that appear before it or, if no AS sub-objects are present, it is assumed to be the current AS.

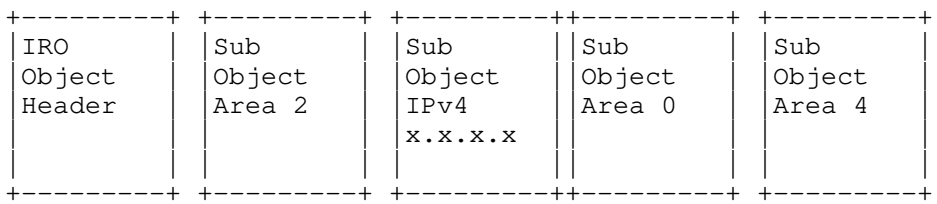
If the area information cannot be provided, PCE MAY forward the path computation request to the next PCE based on AS only. If multiple PCEs of different area domain exist, PCE MAY apply local policy to select the next PCE. Furthermore the domain sequence (list of areas within AS) in the next PCE MAYBE pre-administered or MAYBE discovered via some mechanism (ex. HPCE).

3.5.3. Boundary Node and Inter-AS-Link

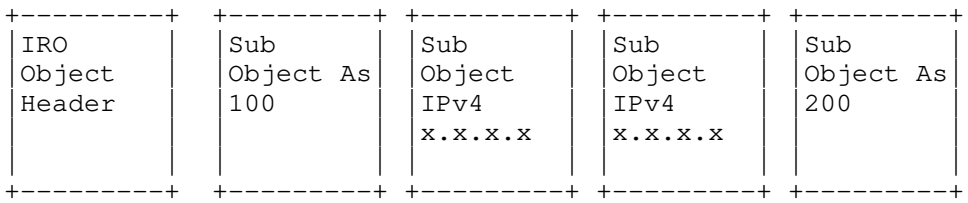
A PCC or PCE MAY add additional constraints covering which Boundary Nodes (ABR or ASBR) or Border links (Inter-AS-link) MUST be traversed while defining a domain sequence. In which case the Boundary Node or Link MAY be encoded as a part of the domain-sequence using the existing sub-objects.

Boundary Node (ABR / ASBR) can be encoded using the IPv4 or IPv6 prefix sub-objects. The Inter-AS link can be encoded using the IPv4 or IPv6 prefix or unnumbered interface sub-objects.

For Figure 1, an ABR to be traversed can be specified as:



For Figure 2, an inter-AS-link to be traversed can be specified as:



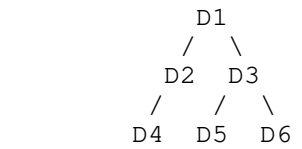
3.5.4. PCE serving multiple domains

A single PCE MAYBE responsible for multiple domains; for example PCE function deployed on an ABR. Domain sequence should have no impact on this. PCE which can support 2 adjacent domains can internally handle this situation without any impact on the neighboring domains.

3.5.5. P2MP

In case of P2MP the path domain tree is nothing but a series of Domain Sequences, as shown in the below figure:

D1-D3-D6, D1-D3-D5 and D1-D2-D4.



3.5.6. HPCE

As per [PCE-HIERARCHY-FWK], consider a case as shown in Figure 4 consisting of multiple child PCEs and a parent PCE.

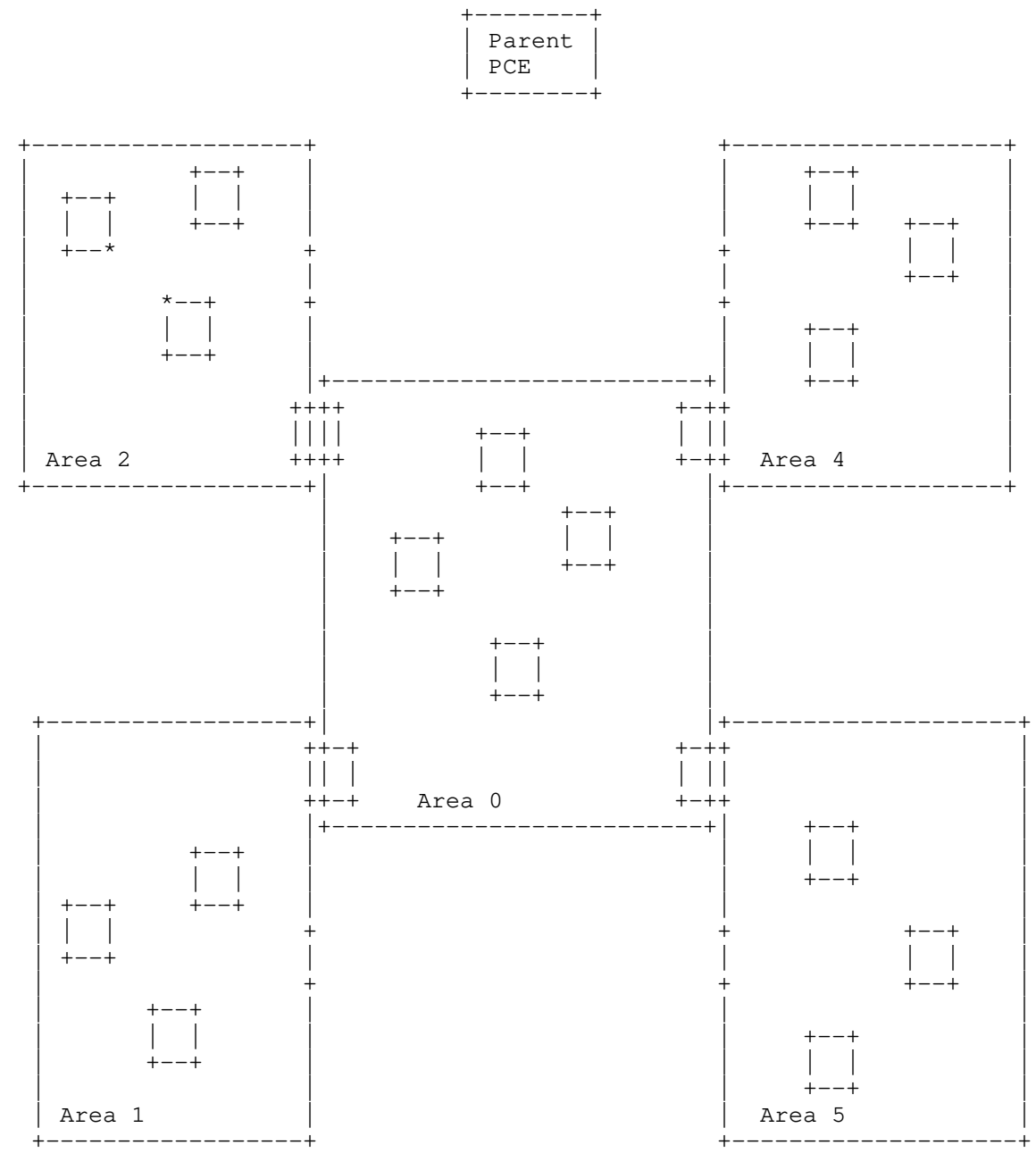
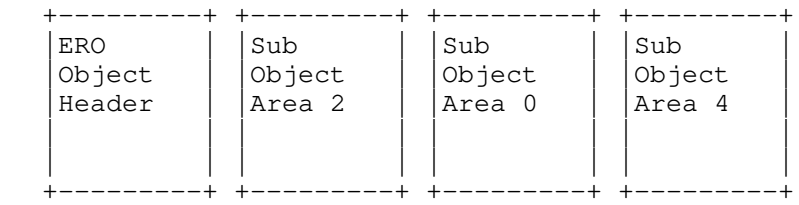
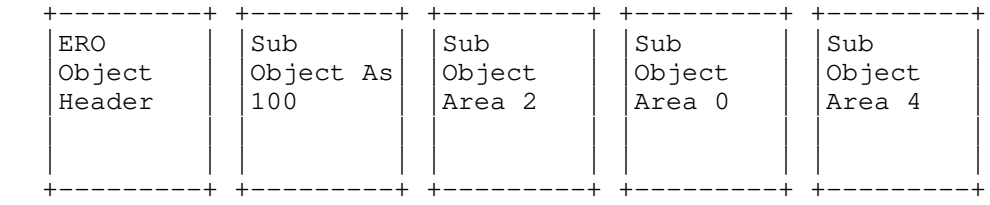


Figure 4: Hierarchical PCE

In HPCE implementation the initiator PCE - PCE(1) can request the

parent PCE to determine the domain sequence and return in the path computation reply message (PCRep), using the ERO Object. The ERO can contain an ordered sequence of sub-object such as AS and Area (OSPF/ ISIS). In this case, the PCRep would carry the domain sequence result as:





Note that, in the case of ERO objects, no new PCEP object type is required since the ordering constraint is assumed.

3.5.7. Relationship to PCE Sequence

[RFC5886] and [PCE-P2MP-PROCEDURES] along with Domain Sequence introduces the concept of PCE-Sequence, where a sequence of PCEs, based on the domain sequence, should be decided and attached in the PCReq at the very beginning of path computation. An alternative would be to use domain sequences, which simplifies as explained below:

Advantages

- All PCE must be aware of all other PCEs in all domain for PCE-Sequence. There is no clear method for this. In domain-sequence PCE should be aware of the domains and not all the PCEs serving the domain. PCE needs to be aware of the neighboring PCEs as done by discovery protocols.

- There maybe multiple PCE in a domain, the selection of PCE should not be made at the PCC/PCE(1). This decision is made only at the neighboring PCE which is aware of state of PCEs via notification

messages.

- Domain sequence would be compatible to P2P inter-domain BRPC method as described in [RFC5441].

4. IANA Considerations

4.1. New IRO Object Type

IANA has defined a registry for Domain-Sequence.

IRO Object-Class 10

IRO Object-Type 2

4.2. Sub-Objects

IANA has defined a registry for following sub-objects.

Type	Sub-object
TBD	AS Number (4 Byte)
TBD	OSPF Area id
TBD	ISIS Area id

5. Security Considerations

This document specifies a standard representation of domain sequence, which is used in all inter-domain PCE scenarios as explained in other RFC and drafts. It does not introduce any new security considerations.

6. Manageability Considerations

TBD

7. Acknowledgments

We would like to thank Pradeep Shastry, Suresh babu, Quintin Zhao, Fatai Zhang, Daniel King, Oscar Gonzalez and Chen Huaimo for their useful comments and suggestions.

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.

[ISO 10589] ISO, "Intermediate system to Intermediate system routeing information exchange protocol for use in conjunction with the Protocol for providing the Connectionless-mode Network Service (ISO 8473)", ISO/IEC 10589:2002.

8.2. Informative References

[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.

[RFC3477] Kompella, K. and Y. Rekhter, "Signalling Unnumbered Links in Resource ReSerVation Protocol - Traffic Engineering (RSVP-TE)", RFC 3477, January 2003.

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

[RFC4726] Farrel, A., Vasseur, J., and A. Ayyangar, "A Framework for Inter-Domain Multiprotocol Label Switching Traffic Engineering", RFC 4726, November 2006.

[RFC4893] Vohra, Q. and E. Chen, "BGP Support for Four-octet AS Number Space", RFC 4893, May 2007.

[RFC5152] Vasseur, JP., Ayyangar, A., and R. Zhang, "A Per-Domain Path Computation Method for Establishing Inter-Domain Traffic Engineering (TE) Label Switched Paths (LSPs)", RFC 5152, February 2008.

[RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.

[RFC5441] Vasseur, JP., Zhang, R., Bitar, N., and JL. Le Roux, "A Backward-Recursive PCE-Based Computation (BRPC) Procedure to Compute

Shortest Constrained Inter-Domain Traffic Engineering Label Switched Paths", RFC 5441, April 2009.

[RFC5886] Vasseur, JP., Le Roux, JL., and Y. Ikejiri, "A Set of Monitoring Tools for Path Computation Element (PCE)-Based Architecture", RFC 5886, June 2010.

[PCE-P2MP-PROCEDURES] Zhao, Q., Dhody, D., Ali, Z., Saad,, T., Sivabalan,, S., and R. Casellas, "PCE-based Computation Procedure To Compute Shortest Constrained P2MP Inter-domain Traffic Engineering Label Switched Paths (draft-ietf-pce-pcep-inter-domain-p2mp-procedures-02)", February 2012.

[PCE-HIERARCHY-FWK] 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. (draft-ietf-pce-hierarchy-fwk-00)", October 2011.

Authors' Addresses

Dhruv Dhody
Huawei Technologies India Pvt Ltd
Leela Palace
Bangalore, Karnataka 560008
INDIA

EMail: dhruv.dhody@huawei.com

Udayasree Palle
Huawei Technologies India Pvt Ltd
Leela Palace
Bangalore, Karnataka 560008
INDIA

EMail: udayasree.palle@huawei.com

Ramon Casellas
CTTC - Centre Tecnologic de Telecomunicacions de Catalunya
Av. Carl Friedrich Gauss n7
Castelldefels, Barcelona 08860
SPAIN

EMail: ramon.casellas@cttc.es

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: March 12, 2012

D. Dhody
Huawei Technologies India Pvt Ltd
V. Manral
Hewlett-Packard Corp.
September 9, 2011

Extensions to the Path Computation Element Communication Protocol (PCEP) to compute service aware Label Switched Path (LSP). draft-dhody-pce-pcep-service-aware-01

Abstract

In certain networks like financial information network (stock/ commodity trading) and enterprises using cloud based applications, Latency (delay), Latency-Variation (jitter) and Packet loss is becoming a key requirement for path computation along with other constraints and metrics. Latency, Latency-Variation and Packet Loss is associated with the Service Level Agreement (SLA) between customers and service providers.

[MPLS-SERVICE] describes MPLS architecture to allow latency, loss and jittering as properties. [OSPF-TE-EXPRESS] describes mechanisms with which network performance information is distributed via OSPF. This document describes the extension to PCEP to carry Latency, Latency-Variation and Loss as constraints for end to end path computation.

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/1id-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, 2012.

Table of Contents

1. Introduction

Real time Network Performance is becoming a critical in the path computation in some networks. There exist mechanism described in [MPLS-LOSS-DELAY] to measure latency, latency-Variation and packet loss after the LSP has been established, which is inefficient. It's important that latency, latency-variation and packet loss are considered during path selection process itself.

TED is populated with network performance information like link latency, latency variation and packet loss through [OSPF-TE-EXPRESS]. Path Computation Client (PCC) can request Path Computation Element (PCE) to provide a path meeting end to end network performance criteria. This document extends Path Computation Element Communication Protocol (PCEP) [RFC 5440] to handle network performance constraint.

1.1. Requirements Language

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

2. Terminology

The following terminology is used in this document.

IGP: Interior Gateway Protocol. Either of the two routing protocols, Open Shortest Path First (OSPF) or Intermediate System to Intermediate System (IS-IS).

IS-IS: Intermediate System to Intermediate System.

OSPF: Open Shortest Path First.

PCC: Path Computation Client: any client application requesting a path computation to be performed by a Path Computation Element.

PCE: Path Computation Element. An entity (component, application, or network node) that is capable of computing a network path or route based on a network graph and applying computational constraints.

TE: Traffic Engineering.

3. Requirement for PCEP

End-to-end service optimization based on latency, latency-variation and packet loss is a key requirement for service provider. Following key requirements associated with latency, latency-variation and loss is identified for PCEP:

1. Path Computation Element (PCE) supporting this draft MUST have the capability to compute end-to-end path with latency, latency-variation and packet loss constraints. It MUST also support the combination of network performance constraint (latency, latency-variation, loss...) with existing constraints (cost, hop-limit...)

2. Path Computation Client (PCC) supporting this draft MUST be able to request for network performance constraint in path request message as the key constraint to be optimized or to suggest boundary condition that should not be crossed.

3. PCEs are not required to support service aware path computation. Therefore, it MUST be possible for a PCE to reject a Path Computation Request message with a reason code that indicates no support for service-aware path computation.

4. PCEP supporting this draft SHOULD provide a means to return end to end network performance information of the computed path in the reply message.

5. PCEP supporting this draft SHOULD provide mechanism to compute multi-domain (e.g., Inter-AS, Inter-Area or Multi-Layer) service aware paths.

It must be understood that such constraints are only meaningful if used consistently: for instance, if the delay of a computed path segment is exchanged between two PCEs residing in different domains, consistent ways of defining the delay must be used.

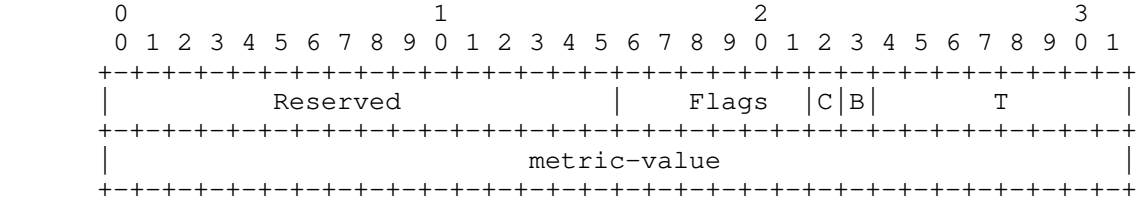
4. Objects

This section defines PCEP extensions (see [RFC5440]) so as to support network performance and service aware path computation.

[RFC5440] defines the optional METRIC object for several purposes. In a PCReq message, a PCC MAY insert one or more METRIC objects to indicate the metric that MUST be optimized or to indicate a bound on the path that MUST NOT be exceeded for the path to be considered as acceptable by the PCC. In a PCRep message, the METRIC object MAY be inserted so as to provide the value for the computed path. It MAY

also be inserted within a PCRep with the NO-PATH object to indicate that the metric constraint could not be satisfied.

As per [RFC5440] the format of the METRIC object body is as follows:



T (Type - 8 bits): Specifies the metric type.
Three values are currently defined:
* T=1: IGP metric
* T=2: TE metric
* T=3: Hop Counts

Based on the section 3, PCEP is extended to define new METRIC types for network performance constraints.

4.1. Latency (Delay) Metric

The end to end Latency (Delay) for the path is represented by this metric.

* T=13(IANA): Latency metric

PCC MAY use this latency metric In PCReq to request a path meeting the end to end latency requirement. In this case B bit MUST be set to suggest a bound (a maximum) for the path latency metric that must not be exceeded for the PCC to consider the computed path as acceptable. The path metric must be less than or equal to the value specified in the metric-value field.

PCC MAY also use this metric to ask PCE to optimize delay during path computation, in this case B flag will be cleared.

PCE MAY use this latency metric In PCRep along with NO-PATH object incase PCE cannot compute a path meeting this constraint. PCE MAY also use this metric to reply the computed end to end latency metric to PCC.

The metric value represents the end to end Latency (delay) measured in milliseconds.

4.2. Latency Variation (Jitter) Metric

The end to end Latency Variation (Jitter) for the path is represented by this metric.

* T=14(IANA): Latency Variation metric

PCC MAY use this latency variation metric In PCReq to request a path meeting the end to end latency variation requirement. In this case B bit MUST be set to suggest a bound (a maximum) for the path latency variation metric that must not be exceeded for the PCC to consider the computed path as acceptable. The path metric must be less than or equal to the value specified in the metric-value field.

PCC MAY also use this metric to ask PCE to optimize jitter during path computation, in this case B flag will be cleared.

PCE MAY use this latency variation metric In PCRep along with NO-PATH object incase PCE cannot compute a path meeting this constraint. PCE MAY also use this metric to reply the computed end to end latency variation metric to PCC.

The metric value represents the end to end Latency variation (jitter) measured in microseconds.

4.3. Packet Loss Metric

The end to end Packet Loss for the path is represented by this metric.

* T=15(IANA): Packet Loss metric

PCC MAY use this packet loss metric In PCReq to request a path meeting the end to end packet loss requirement. In this case B bit MUST be set to suggest a bound (a maximum) for the path packet loss metric that must not be exceeded for the PCC to consider the computed path as acceptable. The path metric must be less than or equal to the value specified in the metric-value field.

PCC MAY also use this metric to ask PCE to optimize packet loss during path computation, in this case B flag will be cleared.

PCE MAY use this packet loss metric In PCRep along with NO-PATH object incase PCE cannot compute a path meeting this constraint. PCE MAY also use this metric to reply the computed end to end packet loss metric to PCC.

The metric value represents the end to end packet loss measured as a

percentage and represented in 32 bit floating point.

4.4. Non-Understanding / Non-Support of Service Aware Path Computation

If the P bit is clear in the object header and PCE doesn't understand or doesn't support service aware path computation it SHOULD simply ignore this METRIC.

If the P Bit is set in the object header and PCE receives new METRIC type in path request and it understands the METRIC type, but the PCE is not capable of service aware path computation, the PCE MUST send a PCErr message with a PCEP-ERROR Object Error-Type = 4 (Not supported object) [RFC5440]. The path computation request MUST then be cancelled.

If the PCE does not understand the new METRIC type, then the PCE MUST send a PCErr message with a PCEP-ERROR Object Error-Type = 3 (Unknown object) [RFC5440].

4.5. Mode of Operation

As explained in [RFC5440], The METRIC object is optional and can be used for several purposes. In a PCReq message, a PCC MAY insert one or more METRIC objects:

- To indicate the metric that MUST be optimized by the path computation algorithm (Latency, Latency-Variation or Loss)

- To indicate a bound on the path METRIC (Latency, Latency-Variation or Loss) that MUST NOT be exceeded for the path to be considered as acceptable by the PCC.

In a PCRep message, the METRIC object MAY be inserted so as to provide the METRIC (Latency, Latency-Variation or Loss) for the computed path. It MAY also be inserted within a PCRep with the NO-PATH object to indicate that the metric constraint could not be satisfied.

The path computation algorithmic aspects used by the PCE to optimize a path with respect to a specific metric are outside the scope of this document.

All the rules of processing METRIC object as explained in [RFC5440] are applicable to the new metric types as well.

4.5.1. Examples

If a PCC sends a path computation request to a PCE where the metric to optimize is the latency and the packet loss must not exceed the value of M, two METRIC objects are inserted in the PCReq message:

- First METRIC object with B=0, T=13, C=1, metric-value=0x0000
- Second METRIC object with B=1, T=15, metric-value=M

If a path satisfying the set of constraints can be found by the PCE and there is no policy that prevents the return of the computed metric, the PCE inserts one METRIC object with B=0, T=13, metric-value= computed end to end latency. Additionally, the PCE may insert a second METRIC object with B=1, T=15, metric-value= computed end to end packet loss.

5. Protocol Consideration

There is no change in the message format of Path Request and Reply Message.

5.1. Inter domain Consideration

[RFC5441] describes the BRPC procedure to compute end to end optimized inter domain path by cooperating PCEs. The network performance constraints can be applied end to end in similar manner as IGP or TE cost.

5.1.1. Inter-AS Link

The IGP in each neighbor domain can advertise its inter-domain TE link capabilities, this has been described in [RFC5316] (ISIS) and [RFC5392] (OSPF). The network performance link properties are described in [OSPF-TE-EXPRESS], the same properties must be advertised using the mechanism described in [RFC5392] (OSPF).

5.1.2. Inter-Layer Consideration

TBD

5.2. Reoptimization Consideration

TBD

5.3. Policy Consideration

TBD

6. IANA Considerations

IANA has defined a registry for new METRIC type.

Type	Meaning
13(TBD)	Latency (delay) metric
14(TBD)	Latency Variation (jitter) metric
15(TBD)	Packet Loss metric

7. Security Considerations

TBD

8. Manageability Considerations

TBD

9. References

9.1. Normative References

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

9.2. Informative References

[MPLS-LOSS-DELAY] Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks [draft-ietf-mpls-loss-delay-04]", July 2011.

[MPLS-SERVICE] Manral, V., "Traffic Engineering architecture for services aware MPLS [draft-manral-mpls-service-01]", July 2011.

[OSPF-TE-EXPRESS] Giacalone, S., Ward, D., Drake, J., Atlas, A., and V. Previdi, "OSPF Traffic Engineering (TE) Express Path [draft-giacalone-ospf-te-express-path-01]", May 2011.

[RFC4655] Ash, J., Vasseur, JP., and A. Farrel, "A Path Computation Element (PCE)-Based Architecture", Aug 2006.

[RFC4657] Ash, J. and JL. Le Roux, "Path Computation Element

(PCE) Communication Protocol Generic Requirements", Sept 2006.

[RFC5316] M Chen, M., Zhang, R., and X. Duan, "ISIS Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", December 2008.

[RFC5392] M Chen, M., Zhang, R., and X. Duan, "OSPF Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", January 2009.

[RFC5440] Ayyangar, A ., Farrel, A ., Oki, E., Atlas, A., Dolganow, A., Ikejiri, Y., Kumaki, K., Vasseur, J., and J. Roux, "Path Computation Element (PCE) communication Protocol (PCEP)", March 2009.

[RFC5441] Vasseur, JP., Zhang, R., Bitar, N., and JL. Le Roux, "A Backward-Recursive PCE-Based Computation (BRPC) Procedure to Compute Shortest Constrained Inter-Domain Traffic Engineering Label Switched Paths", April 2009.

Authors' Addresses

Dhruv Dhody
Huawei Technologies India Pvt Ltd
Leela Palace
Bangalore, Karnataka 560008
INDIA

EMail: dhruv.dhody@huawei.com

Vishwas Manral
Hewlett-Packard Corp.
191111 Pruneridge Ave.
Cupertino, CA 95014
USA

EMail: vishwas.manral@hp.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 24, 2013

D. Dhody
Huawei Technologies India Pvt Ltd
V. Manral
Hewlett-Packard Corp.
Z. Ali
G. Swallow
Cisco Systems
K. Kumaki
KDDI Corporation

February 25, 2013

Extensions to the Path Computation Element Communication Protocol (PCEP) to compute service aware Label Switched Path (LSP). draft-dhody-pce-pcep-service-aware-05

Abstract

In certain networks like financial information network (stock/ commodity trading) and enterprises using cloud based applications, Latency (delay), Latency-Variation (jitter) and Packet loss is becoming a key requirement for path computation along with other constraints and metrics. Latency, Latency-Variation and Packet Loss is associated with the Service Level Agreement (SLA) between customers and service providers.

[MPLS-DELAY-FWK] describes MPLS architecture to allow Latency (delay), Latency-Variation (jitter) and Packet loss as properties. [OSPF-TE-EXPRESS] and [ISIS-TE-EXPRESS] describes mechanisms with which network performance information is distributed via OSPF and ISIS respectively. This document describes the extension to PCEP to carry Latency, Latency-Variation and Loss as constraints for end to end path computation.

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 June 4, 2013.

Table of Contents

1. Introduction

Real time Network Performance is becoming a critical in the path computation in some networks. There exist mechanism described in [RFC6374] to measure latency, latency-Variation and packet loss after the LSP has been established, which is inefficient. It is important that latency, latency-variation and packet loss are considered during path selection process, even before the LSP is setup.

TED is populated with network performance information like link latency, latency variation and packet loss through [OSPF-TE-EXPRESS] or [ISIS-TE-EXPRESS]. Path Computation Client (PCC) can request Path Computation Element (PCE) to provide a path meeting end to end network performance criteria. This document extends Path Computation Element Communication Protocol (PCEP) [RFC5440] to handle network performance constraint.

PCE MAY use mechanism described in [MPLS-TE-EXPRESS-PATH] on how to use the link latency, latency variation and packet loss information for end to end path selection.

1.1. Requirements Language

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

2. Terminology

The following terminology is used in this document.

IGP: Interior Gateway Protocol. Either of the two routing protocols, Open Shortest Path First (OSPF) or Intermediate System to Intermediate System (IS-IS).

IS-IS: Intermediate System to Intermediate System.

OSPF: Open Shortest Path First.

PCC: Path Computation Client: any client application requesting a path computation to be performed by a Path Computation Element.

PCE: Path Computation Element. An entity (component, application, or network node) that is capable of computing a network path or route based on a network graph and applying computational constraints.

TE: Traffic Engineering.

3. PCEP Requirements

End-to-end service optimization based on latency, latency-variation and packet loss is a key requirement for service provider. Following key requirements associated with latency, latency-variation and loss are identified for PCEP:

1. Path Computation Element (PCE) supporting this draft MUST have the capability to compute end-to-end path with latency, latency-variation and packet loss constraints. It MUST also support the combination of network performance constraint (latency, latency-variation, loss...) with existing constraints (cost, hop-limit...)

2. Path Computation Client (PCC) MUST be able to request for network performance constraint in path request message as the key constraint to be optimized or to suggest boundary condition that should not be crossed.

3. PCEs are not required to support service aware path computation. Therefore, it MUST be possible for a PCE to reject a Path Computation Request message with a reason code that indicates no support for service-aware path computation.

4. PCEP SHOULD provide a means to return end to end network performance information of the computed path in the reply message.

5. PCEP SHOULD provide mechanism to compute multi-domain (e.g., Inter-AS, Inter-Area or Multi-Layer) service aware paths.

It is assumed that such constraints are only meaningful if used consistently: for instance, if the delay of a computed path segment is exchanged between two PCEs residing in different domains, consistent ways of defining the delay must be used.

4. PCEP extensions

This section defines PCEP extensions (see [RFC5440]) for requirements outlined in Section 3. The proposed solution is used to support network performance and service aware path computation.

This document defines the following optional types for the METRIC object defined in [RFC5440].

For explanation of these metrics, the following terminology is used

and expanded along the way.

- A network comprises of a set of N links {Li, (i=1...N)}.

- A path P of a P2P LSP is a list of K links {Lpi,(i=1...K)}.

4.1. Latency (Delay) Metric

Link delay metric is defined in [OSPF-TE-EXPRESS] and [ISIS-TE-EXPRESS]. P2P latency metric type of METRIC object in PCEP encodes the sum of the link delay metric of all links along a P2P Path. Specifically, extending on the above mentioned terminology:

- A Link delay metric of link L is denoted D(L).

- A P2P latency metric for the Path P = Sum {D(Lpi), (i=1...K)}.

* T=13(IANA): Latency metric

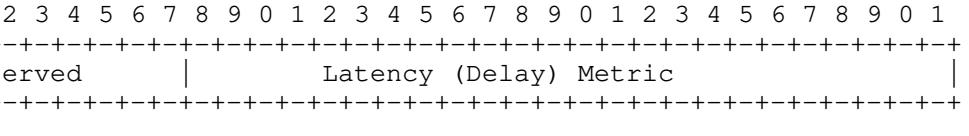
PCC MAY use this latency metric In PCReq to request a path meeting the end to end latency requirement. In this case B bit MUST be set to suggest a bound (a maximum) for the path latency metric that must not be exceeded for the PCC to consider the computed path as acceptable. The path metric must be less than or equal to the value specified in the metric-value field.

PCC MAY also use this metric to ask PCE to optimize delay during path computation, in this case B flag will be cleared.

PCE MAY use this latency metric In PCRep along with NO-PATH object incase PCE cannot compute a path meeting this constraint. PCE MAY also use this metric to reply the computed end to end latency metric to PCC.

4.1.1. Latency (Delay) Metric Value

[OSPF-TE-EXPRESS] and [ISIS-TE-EXPRESS] defines "Unidirectional Link Delay Sub-TLV" in a 24-bit field. [RFC5440] defines the METRIC object with 32-bit metric value. Consequently, encoding for Latency (Delay) Metric Value is defined as follows:



Reserved (8 bits): Reserved field. This field MUST be set to zero on

transmission and MUST be ignored on receipt.

Latency (Delay) Metric (24 bits): Represents the end to end Latency (delay) quantified in units of microseconds and MUST be encoded as integer value. With the maximum value 16,777,215 representing 16.777215 sec.

4.2. Latency Variation (Jitter) Metric

Link delay variation metric is defined in [OSPF-TE-EXPRESS] and [ISIS-TE-EXPRESS]. P2P latency variation metric type of METRIC object in PCEP encodes a function of the link delay variation metric of all links along a P2P Path. Specifically, extending on the above mentioned terminology:

- A Latency variation of link L is denoted DV(L).

- A P2P latency variation metric for the Path P = function {DV(Lpi), (i=1...K)}.

Specification of the "Function" used to drive latency variation metric of a path from latency variation metrics of individual links along the path is beyond the scope of this document.

* T=14(IANA): Latency Variation metric

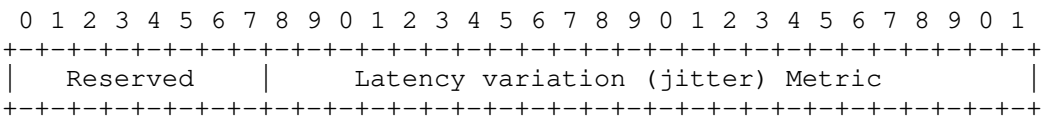
PCC MAY use this latency variation metric In PCReq to request a path meeting the end to end latency variation requirement. In this case B bit MUST be set to suggest a bound (a maximum) for the path latency variation metric that must not be exceeded for the PCC to consider the computed path as acceptable. The path metric must be less than or equal to the value specified in the metric-value field.

PCC MAY also use this metric to ask PCE to optimize jitter during path computation, in this case B flag will be cleared.

PCE MAY use this latency variation metric In PCRep along with NO-PATH object incase PCE cannot compute a path meeting this constraint. PCE MAY also use this metric to reply the computed end to end latency variation metric to PCC.

4.2.1. Latency Variation (Jitter) Metric Value

[OSPF-TE-EXPRESS] and [ISIS-TE-EXPRESS] defines "Unidirectional Delay Variation Sub-TLV" in a 24-bit field. [RFC5440] defines the METRIC object with 32-bit metric value. Consequently, encoding for Latency Variation (Jitter) Metric Value is defined as follows:



Reserved (8 bits): Reserved field. This field MUST be set to zero on transmission and MUST be ignored on receipt.

Latency variation (jitter) Metric (24 bits): Represents the end to end Latency variation (jitter) quantified in units of microseconds and MUST be encoded as integer value. With the maximum value 16,777,215 representing 16.777215 sec.

4.3. Packet Loss Metric

[OSPF-TE-EXPRESS] and [ISIS-TE-EXPRESS] defines "Unidirectional Link Loss". Packet Loss Metric metric type of METRIC object in PCEP encodes a function of the link's unidirectional loss metric of all links along a P2P Path. Specifically, extending on the above mentioned terminology:

The end to end Packet Loss for the path is represented by this metric.

- A Packet loss of link L is denoted PL(L).

- A P2P packet loss metric for the Path P = function {PL(Lpi), (i=1...K)}.

Specification of the "Function" used to drive end to end packet loss metric of a path from packet loss metrics of individual links along the path is beyond the scope of this document.

* T=15(IANA): Packet Loss metric

PCC MAY use this packet loss metric In PCReq to request a path meeting the end to end packet loss requirement. In this case B bit MUST be set to suggest a bound (a maximum) for the path packet loss metric that must not be exceeded for the PCC to consider the computed path as acceptable. The path metric must be less than or equal to the value specified in the metric-value field.

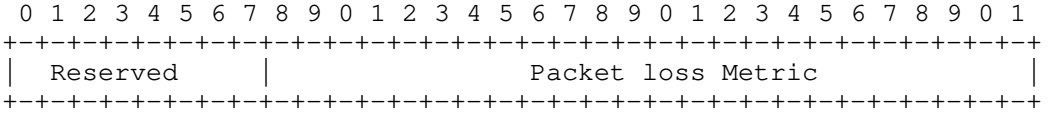
PCC MAY also use this metric to ask PCE to optimize packet loss during path computation, in this case B flag will be cleared.

PCE MAY use this packet loss metric In PCRep along with NO-PATH object incase PCE cannot compute a path meeting this constraint. PCE

MAY also use this metric to reply the computed end to end packet loss metric to PCC.

4.3.1. Packet Loss Metric Value

[OSPF-TE-EXPRESS] and [ISIS-TE-EXPRESS] defines "Unidirectional Link Loss Sub-TLV" in a 24-bit field. [RFC5440] defines the METRIC object with 32-bit metric value. Consequently, encoding for Packet Loss Metric Value is defined as follows:



Reserved (8 bits): Reserved field. This field MUST be set to zero on transmission and MUST be ignored on receipt.

Packet loss Metric (24 bits): Represents the end to end packet loss quantified as a percentage of packets lost and MUST be encoded as integer. The basic unit is 0.000003%, with the maximum value 16,777,215 representing 50.331645% (16,777,215 * 0.000003%). This value is the highest packet loss percentage that can be expressed.

4.4. Non-Understanding / Non-Support of Service Aware Path Computation

If the P bit is clear in the object header and PCE does not understand or does not support service aware path computation it SHOULD simply ignore this METRIC.

If the P Bit is set in the object header and PCE receives new METRIC type in path request and it understands the METRIC type, but the PCE is not capable of service aware path computation, the PCE MUST send a PCErr message with a PCEP-ERROR Object Error-Type = 4 (Not supported object) [RFC5440]. The path computation request MUST then be cancelled.

If the PCE does not understand the new METRIC type, then the PCE MUST send a PCErr message with a PCEP-ERROR Object Error-Type = 3 (Unknown object) [RFC5440].

4.5. Mode of Operation

As explained in [RFC5440], The METRIC object is optional and can be used for several purposes. In a PCReq message, a PCC MAY insert one or more METRIC objects:

- To indicate the metric that MUST be optimized by the path computation algorithm (Latency, Latency-Variation or Loss)

- To indicate a bound on the path METRIC (Latency, Latency-Variation or Loss) that MUST NOT be exceeded for the path to be considered as acceptable by the PCC.

In a PCRep message, the METRIC object MAY be inserted so as to provide the METRIC (Latency, Latency-Variation or Loss) for the computed path. It MAY also be inserted within a PCRep with the NO-PATH object to indicate that the metric constraint could not be satisfied.

The path computation algorithmic aspects used by the PCE to optimize a path with respect to a specific metric are outside the scope of this document.

All the rules of processing METRIC object as explained in [RFC5440] are applicable to the new metric types as well.

In a PCReq message, a PCC MAY insert more than one METRIC object to be optimized, in such a case PCE should find the path that is optimal when both the metrics are considered together.

4.5.1. Examples

Example 1: If a PCC sends a path computation request to a PCE where two metric to optimize are the latency and the packet loss, two METRIC objects are inserted in the PCReq message:

- First METRIC object with B=0, T=13 (TBA - IANA), C=1, metric-value=0x0000

- Second METRIC object with B=0, T=15 (TBA - IANA), C=1, metric-value=0x0000

PCE in such a case should try to optimize both the metrics and find a path with the minimum latency and packet loss, if a path can be found by the PCE and there is no policy that prevents the return of the computed metric, the PCE inserts two METRIC object with B=0, T=13 (TBA - IANA), metric-value= computed end to end latency and second METRIC object with B=1, T=15 (TBA - IANA), metric-value= computed end to end packet loss.

Example 2: If a PCC sends a path computation request to a PCE where the metric to optimize is the latency and the packet loss must not exceed the value of M, two METRIC objects are inserted in the PCReq message:

- First METRIC object with B=0, T=13 (TBA - IANA), C=1, metric-value=0x0000
- Second METRIC object with B=1, T=15 (TBA - IANA), metric-value=M

If a path satisfying the set of constraints can be found by the PCE and there is no policy that prevents the return of the computed metric, the PCE inserts one METRIC object with B=0, T=13 (TBA - IANA), metric-value= computed end to end latency. Additionally, the PCE may insert a second METRIC object with B=1, T=15 (TBA - IANA), metric-value= computed end to end packet loss.

5. Relationship with Objective function

[RFC5541] defines mechanism to specify an optimization criteria, referred to as objective functions. The new metric types specified in this document can continue to use the existing Objective function.

Minimum Cost Path (MCP) is one such objective function.

- A network comprises a set of N links {Li, (i=1...N)}.
- A path P is a list of K links {Lpi,(i=1...K)}.
- Metric of link L is denoted M(L). This can be any metric, including the ones defined in this document.
- The cost of a path P is denoted C(P), where C(P) = sum {M(Lpi),(i=1...K)}.

Name: Minimum Cost Path (MCP)

Description: Find a path P such that C(P) is minimized.

The new metric types for example latency (delay) can continue to use the above objective function to find the minimum cost path where cost is latency (delay). At the same time new objective functions can be defined in future to optimize these new metric types.

6. Protocol Consideration

There is no change in the message format of Path Request and Reply Message.

6.1. Inter domain Consideration

[RFC5441] describes the BRPC procedure to compute end to end optimized inter domain path by cooperating PCEs. The network

performance constraints can be applied end to end in similar manner as IGP or TE cost.

All domains should have the same understanding of the METRIC (Latency-Variation etc) for end-to-end inter-domain path computation to make sense. Otherwise some form of Metric Normalization as described in [RFC5441] MAY need to be applied.

6.1.1. Inter-AS Link

The IGP in each neighbor domain can advertise its inter-domain TE link capabilities, this has been described in [RFC5316] (ISIS) and [RFC5392] (OSPF). The network performance link properties are described in [OSPF-TE-EXPRESS] and [ISIS-TE-EXPRESS], the same properties must be advertised using the mechanism described in [RFC5392] (OSPF) and [RFC5316] (ISIS).

6.1.2. Inter-Layer Consideration

PCEP supporting this draft SHOULD provide mechanism to support different Metric requirements for different Layers. This is important as the network performance metric would be different for Packet and Optical (TDM, LSC etc) Layers. In order to allow different Metric-Value to be applied within different network layers, multiple METRIC objects of the same type MAY be present. In such a case, the first METRIC object specifies an metric for the higher-layer network, and subsequent METRIC objects specify objection functions of the subsequent lower-layer networks.

6.2. Reoptimization Consideration

PCC can monitor the setup LSPs and incase of degradation of network performance constraints, it MAY ask PCE for reoptimization as per [RFC5440].

6.3. Point-to-Multipoint (P2MP)

This document defines the following optional types for the METRIC object defined in [RFC5440] for P2MP TE LSPs. Additional metric types for P2MP TE LSPs are to be added in a future revision

6.3.1. P2MP Latency Metric

P2MP latency metric type of METRIC object in PCEP encodes the path latency metric for destination that observes the worst latency metric among all destination of the P2MP tree. Specifically, extending on the above mentioned terminology:

- A P2MP Tree T comprises of a set of M destinations {Dest_j, (j=1...M)}

- P2P latency metric of the Path to destination Dest_j is denoted by LM(Dest_j).

- P2MP latency metric for the P2MP tree T = Maximum {LM(Dest_j), (j=1...M)}.

Value for P2MP latency metric is to be assigned by IANA

6.3.2. P2MP Latency Variation Metric

P2MP latency variation metric type of METRIC object in PCEP encodes the path latency variation metric for destination that observes the worst latency variation metric among all destination of the P2MP tree. Specifically, extending on the above mentioned terminology:

- A P2MP Tree T comprises of a set of M destinations {Dest_j, (j=1...M)}

- P2P latency variation metric of the Path to destination Dest_j is denoted by LVM(Dest_j).

- P2MP latency variation metric for the P2MP tree T = Maximum {LVM(Dest_j), (j=1...M)}.

Value for P2MP latency variation metric is to be assigned by IANA

7. IANA Considerations

IANA has defined a registry for new METRIC type.

Type	Meaning
13(TBD)	Latency (delay) metric
14(TBD)	Latency Variation (jitter) metric
15(TBD)	Packet Loss metric
16(TBD)	P2MP latency metric
17(TBD)	P2MP latency variation metric

8. Security Considerations

This document defines three new METRIC Types which does not add any new security concerns to PCEP protocol.

9. Manageability Considerations

9.1. Control of Function and Policy

The only configurable item is the support of the new service-aware METRICS on a PCE which MAY be controlled by a policy module. If the new METRIC is not supported/allowed on a PCE, it MUST send a PCErr message as specified in Section 4.4.

9.2. Information and Data Models

[PCEP-MIB] describes the PCEP MIB, there are no new MIB Objects for this document.

9.3. Liveness Detection and Monitoring

Mechanisms defined in this document do not imply any new liveness detection and monitoring requirements in addition to those already listed in [RFC5440].

9.4. Verify Correct Operations

Mechanisms defined in this document do not imply any new operation verification requirements in addition to those already listed in [RFC5440].

9.5. Requirements On Other Protocols

PCE requires the TED to be populated with network performance information like link latency, latency variation and packet loss. This mechanism is described in [OSPF-TE-EXPRESS] or [ISIS-TE-EXPRESS].

9.6. Impact On Network Operations

Mechanisms defined in this document do not have any impact on network operations in addition to those already listed in [RFC5440].

10. Acknowledgments

We would like to thank Young Lee, Venugopal Reddy, Reeja Paul, Sandeep Kumar Boina, Suresh babu, Quintin Zhao and Chen Huaimo for their useful comments and suggestions.

11. References

11.1. Normative References

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

[RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.

11.2. Informative References

[RFC5441] Vasseur, JP., Zhang, R., Bitar, N., and JL. Le Roux, "A Backward-Recursive PCE-Based Computation (BRPC) Procedure to Compute Shortest Constrained Inter-Domain Traffic Engineering Label Switched Paths", RFC 5441, April 2009.

[RFC5316] Chen, M., Zhang, R., and X. Duan, "ISIS Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 5316, December 2008.

[RFC5392] Chen, M., Zhang, R., and X. Duan, "OSPF Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 5392, January 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.

[RFC6374] Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks", RFC 6374, September 2011.

[MPLS-DELAY-FWK] Fu, X., Manral, V., McDysan, D., Malis, A., Giacalone, S., Betts, M., Wang, Q., and J. Drake, "Traffic Engineering architecture for services aware MPLS [draft-fuxh-mpls-delay-loss-te-framework]", Oct 2012.

[OSPF-TE-EXPRESS] Giacalone, S., Ward, D., Drake, J., Atlas, A., and S. Previdi, "OSPF Traffic Engineering (TE) Metric Extensions

[draft-ietf-ospf-te-metric-extensions]", May 2012.

[ISIS-TE-EXPRESS] Previdi, S., Giacalone, S., Ward, D., Drake, J., Atlas, A., and C. Filsfils, "IS-IS Traffic Engineering (TE) Metric Extensions [draft-previdi-isis-te-metric-extensions]", Oct 2012.

[MPLS-TE-EXPRESS-PATH] Atlas, A., Drake, J., Ward, D., Giacalone, S., Previdi, S., and C. Filsfils, "Performance-based Path Selection for Explicitly Routed LSPs [draft-atlas-mpls-te-express-path]", June 2012.

[PCEP-MIB] Kiran Koushik, A S., Stephan, E., Zhao, Q., King, D., and J. Hardwick, "PCE communication protocol(PCEP) Management Information Base [draft-ietf-pce-pcep-mib]", July 2012.

Appendix A. Contributor Addresses

Clarence Filsfils
Cisco Systems
EMail: cfilsfil@cisco.com

Siva Sivabalan
Cisco Systems
EMail: msiva@cisco.com

Stefano Previdi
Cisco Systems
EMail: sprevidi@cisco.com

Udayasree Palle
Huawei Technologies India Pvt Ltd
Leela Palace
Bangalore, Karnataka 560008
INDIA
EMail: udayasree.palle@huawei.com

Authors' Addresses

Dhruv Dhody
Huawei Technologies India Pvt Ltd
Leela Palace
Bangalore, Karnataka 560008
INDIA

EMail: dhruv.dhody@huawei.com

Vishwas Manral
Hewlett-Packard Corp.
191111 Pruneridge Ave.
Cupertino, CA 95014
USA

EMail: vishwas.manral@hp.com

Zafar Ali
Cisco Systems

EMail: zali@cisco.com

George Swallow
Cisco Systems

EMail: swallow@cisco.com

Kenji Kumaki
KDDI Corporation

EMail: ke-kumaki@kddi.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 23, 2012

WJ. He, Ed.
ZTE

October 21, 2011

Extensions to the Path Computation Element Communication Protocol (PCEP) for Associated Bidirectional LSP
draft-he-pce-pcep-associated-lsp-extensions-00

Abstract

The MPLS Transport Profile (MPLS-TP) requirements document[RFC5654], describes that MPLS-TP MUST support associated bidirectional point-to-point LSPs. Path Computation Element (PCE), see [RFC4655], may be used for path computation of an associated bidirectional LSP. This document defines the Path Computation Element Protocol (PCEP)-based [RFC5440] extensions for associated bidirectional LSP.

Table of Contents

1. Introduction

The MPLS Transport Profile (MPLS-TP) requirements [RFC5654] and control plane framework documents[RFC6373]describe that MPLS-TP MUST support associated bidirectional point-to-point LSPs. Path Computation Element (PCE), see [RFC4655], may be used for path computation of a GMPLS LSP, see [I-D.ietf-pce-gmpls-pcep-extensions],and consequently an associated bidirectional LSP, across domains and in a single domain.

Dependent path computations are requests that need to be synchronized in order to meet specific objectives, see [RFC6007]. For associated bidirectional LSP, if the forward LSP and the backward LSP are computed concurrently, the PCE can find the optimum path.

This document defines the Path Computation Element Protocol (PCEP)-based [RFC5440] extensions for associated bidirectional LSP.

2. Conventions used in this document

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

3. Processing

Consider the topology described in Figure 1. (An example of associated bidirectional LSP). The LSP1 [via nodes A,D,B] (from A to B) and LSP2 [via nodes B,D,C,A] (from B to A) need to be established, which can form an associated bidirectional LSP deployed by Single Sided Provisioning model or Double Sided Provisioning model[I-D.ietf-ccamp-mpls-tp-rsvpte-ext-associated-lsp]. Node A, the ingress LSR of LSP1, can play the role of a PCC and request the PCE to compute the LSP1 or the associated bidirectional LSP.

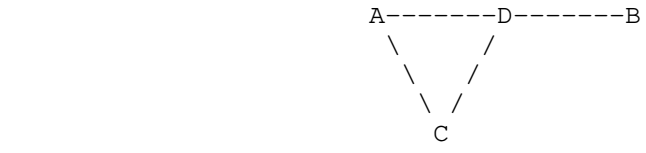


Figure 1 : An example of associated bidirectional LSP

3.1. The Single Sided Provisioning

For the single sided provisioning, the path computation can be realized by the concurrent or successive computation. The concurrent computation means that the head-end submits the computation request for both two directional LSPs concurrently. As to the successive computation, the head-end and the tail-end send the forward LSP and backward LSP computation requests separately.

3.1.1. Concurrent Computation

The PCC sends the PCReq message to PCE for computing an associated bidirectional LSP, whose forward and backward paths are computed concurrently. Concurrent computation can ensure that the paths for the associated bidirectional LSP is optimal [RFC5557].

The basic procedure are as follows:

1. The PCC node sends the PCReq message to the PCE with the A flag of the RP object set, indicates the request is for an associated bidirectional LSP. Except the constraint information about the forward LSP, the REVERSE_LSP object may also be included in the PCReq message to specify the TE parameters of the backward LSP.
2. Once receiving the PCReq message, the PCE will compute the two reverse LSP based on the constraints, and choose the optimal LSP for the associated bidirectional LSP. If the A bit of the RP object set to 1, but the REVERSE_LSP object is not present in the PCReq message, the PCE computes the path of the reverse LSP according to the forward LSP information, such as bandwidth, protection and so on.
3. After the successful computation, the PCE will supply the PCC with a fully computed explicit routes of an associated bidirectional LSP. The explicit path for the forward LSP is carried by the ERO object and the backward LSP by the ERO subobject inserted in the REVERSE_LSP object.

If the PCE does not support the extensions in this document, responses with notification.

3.1.2. Successive Computation

Successive computation means that the forward LSP and the backward LSP are computed separately. The head-end will send request to the PCE for the forward LSP. After receiving the successful computation result, the head-end starts to signal the forward LSP with Extended Association object and Reverse LSP object inserted in the Path message [I-D.ietf-ccamp-mpls-tp-rsvpte-ext-associated-lsp]. Once receiving the Path message, the tail-end will be triggered to create

the backward LSP. The REVERSE_LSP object is extracted from the Path message and may be put into the PCReq message for a path computation. There is no need to extend the PCEP to support the successive computation.

3.2. The Double Sided Provisioning

For the double sided provisioning, the forward and the backward LSP configuration are send to the head-end and the tail-end separately. The head-end and the tail-end will send the PCReq message for the unidirectional LSP computation. After the successfully computation, the head-end and the tail-end start to create the LSP separately.

4. PCEP Extensions

4.1. The Extension of the RP Object

The PCReq and PCRep messages will need the following additional parameters for associated bidirectional LSP.

An A-bit is added to the flag bits of the RP object to indicate the request is about an associated bidirectional LSP or not.

- A (RP Associated bit - 1 bit): when set, the PCC specifies that the path computation request relates to an associated bidirectional TE LSP that may be has the different traffic engineering requirements including fate sharing, protection and restoration, LSRs, TE links, and resource requirements (e.g., latency and jitter) in each direction. When cleared, the TE LSP is not an associated bidirectional TE LSP .

4.2. REVERSE_LSP Object

The REVERSE_LSP object is used in a PCReq message to specify the information of the reverse LSP for which a path computation is requested. This object is optional. The format of the REVERSE_LSP object is as follows:

Object-Class is TBD,Object-Type is TBD.

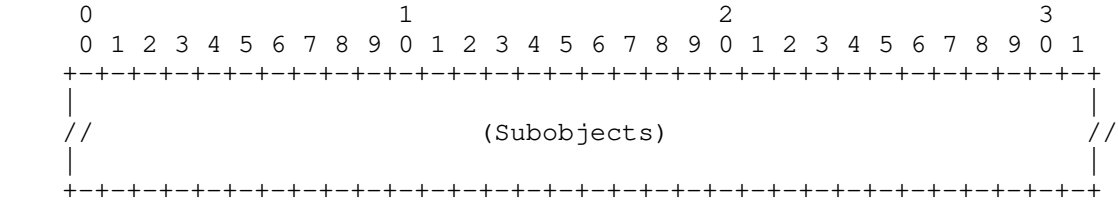


Figure 2 : REVERSE_LSP Object Body Format

This object MUST NOT be used when the A bit of RP object set to 0.

Subojects

The contents of a REVERSE_LSP object are a series of variable-length data items called subobjects, which can be BANDWIDTH, IRO and XRO object, LSPA Object, METRIC Object, etc.

5. IANA Considerations

TBD

6. Security Considerations

TBD

7. Acknowledgement

TBD

8. References

8.1. Normative References

[I-D.ietf-ccamp-mpls-tp-rsvpte-ext-associated-lsp]
Zhang, F. and R. Jing, "RSVP-TE Extensions for Associated Bidirectional LSPs", draft-ietf-ccamp-mpls-tp-rsvpte-ext-associated-lsp-02 (work in progress), October 2011.

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

[RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.

8.2. Informative References

[I-D.ietf-pce-gmpls-pcep-extensions]
Margaria, C., Dios, O., and F. Zhang, "PCEP extensions for GMPLS", draft-ietf-pce-gmpls-pcep-extensions-03 (work in

progress), July 2011.

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

[RFC5557] Lee, Y., Le Roux, JL., King, D., and E. Oki, "Path Computation Element Communication Protocol (PCEP) Requirements and Protocol Extensions in Support of Global Concurrent Optimization", RFC 5557, July 2009.

[RFC5654] Niven-Jenkins, B., Brungard, D., Betts, M., Sprecher, N., and S. Ueno, "Requirements of an MPLS Transport Profile", RFC 5654, September 2009.

[RFC6007] Nishioka, I. and D. King, "Use of the Synchronization VECtor (SVEC) List for Synchronized Dependent Path Computations", RFC 6007, September 2010.

[RFC6373] Andersson, L., Berger, L., Fang, L., Bitar, N., and E. Gray, "MPLS Transport Profile (MPLS-TP) Control Plane Framework", RFC 6373, September 2011.

Author's Address

Wenjuan He (editor)
ZTE

Email: he.wenjuan1@zte.com.cn

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: May 17, 2012

WJ. He, Ed.
ZTE

November 14, 2011

Extensions to the Path Computation Element Communication Protocol (PCEP) for Associated Bidirectional LSP
draft-he-pce-pcep-associated-lsp-extensions-01

Abstract

The MPLS Transport Profile (MPLS-TP) requirements document[RFC5654], describes that MPLS-TP MUST support associated bidirectional point-to-point LSPs. Path Computation Element (PCE), see [RFC4655], may be used for path computation of an associated bidirectional LSP. This document defines the Path Computation Element Protocol (PCEP)-based [RFC5440] extensions for associated bidirectional LSP.

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 May 17, 2012.

Copyright Notice

Table of Contents

1. Introduction

The MPLS Transport Profile (MPLS-TP) requirements [RFC5654] and control plane framework documents[RFC6373]describe that MPLS-TP MUST support associated bidirectional point-to-point LSPs. Path Computation Element (PCE), see [RFC4655], may be used for path computation of a GMPLS LSP, see [I-D.ietf-pce-gmpls-pcep-extensions],and consequently an associated bidirectional LSP, across domains and in a single domain.

Dependent path computations are requests that need to be synchronized in order to meet specific objectives, see [RFC6007]. For associated bidirectional LSP, if the forward LSP and the backward LSP are computed concurrently, the PCE can find the optimum path.

This document defines the Path Computation Element Protocol (PCEP)-based [RFC5440] extensions for associated bidirectional LSP.

2. Conventions used in this document

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

3. Processing

Consider the topology described in Figure 1. (An example of associated bidirectional LSP). The LSP1 [via nodes A,D,B] (from A to B) and LSP2 [via nodes B,D,C,A] (from B to A) need to be established, which can form an associated bidirectional LSP deployed by Single Sided Provisioning model or Double Sided Provisioning model[I-D.ietf-ccamp-mpls-tp-rsvpte-ext-associated-lsp]. Node A, the ingress LSR of LSP1, can play the role of a PCC and request the PCE to compute the LSP1 or the associated bidirectional LSP.

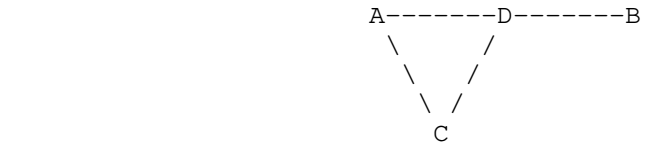


Figure 1 : An example of associated bidirectional LSP

The path computation for associated bidirectional LSP can be realized

by the concurrent or successive computation. The concurrent computation means that the head-end submits the computation request for both two directional LSPs concurrently, which is applicable to the Single Sided Provisioning model. As to the successive computation, the head-end and the tail-end send the forward LSP and backward LSP computation requests separately, which is applicable to both the Single Sided Provisioning model and the Double Sided Provisioning model.

3.1. Concurrent Computation

The PCC sends the PCReq message to PCE for computing an associated bidirectional LSP, whose forward and backward paths are computed concurrently. Concurrent computation can ensure that the paths for the associated bidirectional LSP is optimal [RFC5557].

The SVEC object described in [RFC6007] can be used to synchronize the requests about the forward and backward LSPs, and get the optimal path for the associated bidirectional LSP.

3.2. Successive Computation

For the Successive computation, the PCCs submit the path computation request for the forward LSP and the backward LSP separately, then, the path for the associated bidirectional LSP may be not optimal. So that the two reverse LSPs should be associated, the ASSOCIATION object [I-D.ietf-ccamp-assoc-ext]may be useful. The stateful PCE [RFC4655] can coordinate the two reverse LSPs to get the optimal path for the associated bidirectional LSP through the ASSOCIATION object, if both the head-end and the tail-end PCCs delegate their respective LSPs (forward and backward) to the PCE .

When the PCC submits the path computation request for the forward LSP or the backward LSP, the PCReq message may carry Extended ASSOCIATION object to indicate there is a reverse LSP to be associated, see [I-D.ietf-ccamp-mpls-tp-rsvpte-ext-associated-lsp]. At the same time, both the head-end and the tail-end PCCs delegate their respective LSPs (forward and backward) to the PCE using PCRpt messages[I-D.crabbe-pce-stateful-pce].

Upon receipt of the PCReq message,the PCE will locate the reverse LSP with the same association information. If there is no matched reverse LSP, the PCE will compute the LSP independently. Otherwise, the PCE will coordinate the two reverse LSPs and compute path for the associated bidirectional LSP. After the successful computation, the PCE will trigger the head-end to setup the forward LSP and the tail-end to setup the backward LSP using PCUpd messages. The PCC will use make-before-break whenever possible in the re-signaling

operation,[I-D.crabbe-pce-stateful-pce].

4. PCEP Extensions

4.1. Extended ASSOCIATION Object

The Extended ASSOCIATION Object is used to associate the two reverse LSPs, which form an associated bidirectional LSP. The Extended ASSOCIATION Object is carried within a PCRep message to locate the reverse LSP, and the PCE will coordinate the forward LSP and the backward LSP to get the optimal path for the associated bidirectional LSP.

The contents of this object are identical in encoding to the contents of the RSVP-TE Extended ASSOCIATION Object defined in [I-D.ietf-ccamp-assoc-ext] and [I-D.ietf-ccamp-mpls-tp-rsvpte-ext-associated-lsp].

PCEP Extended ASSOCIATION object types correspond to RSVP-TE Extended ASSOCIATION object types.

Extended ASSOCIATION Object-Class is TBD.

Extended ASSOCIATION Object-Type is TBD.

5. IANA Considerations

TBD

6. Security Considerations

TBD

7. Acknowledgement

The author would like to thank Jan Medved for his valuable comments on the double sided provisioning, Cyril for the discussion of the concurrent computation. At the same time, the author would also like to acknowledge the contributions of Fei Zhang for the discussions.

8. References

8.1. Normative References

[I-D.crabbe-pce-stateful-pce]
Crabbe, E., Medved, J., and R. Varga, "PCEP Extensions for Stateful PCE", draft-crabbe-pce-stateful-pce-01 (work in progress), October 2011.

[I-D.ietf-ccamp-assoc-ext]
Berger, L., Faucheur, F., and A. Narayanan, "RSVP Association Object Extensions", draft-ietf-ccamp-assoc-ext-01 (work in progress), October 2011.

[I-D.ietf-ccamp-mpls-tp-rsvpte-ext-associated-lsp]
Zhang, F. and R. Jing, "RSVP-TE Extensions for Associated Bidirectional LSPs", draft-ietf-ccamp-mpls-tp-rsvpte-ext-associated-lsp-02 (work in progress), October 2011.

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

[RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.

8.2. Informative References

[I-D.ietf-pce-gmpls-pcep-extensions]
Margaria, C., Dios, O., and F. Zhang, "PCEP extensions for GMPLS", draft-ietf-pce-gmpls-pcep-extensions-04 (work in progress), October 2011.

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

[RFC5557] Lee, Y., Le Roux, JL., King, D., and E. Oki, "Path Computation Element Communication Protocol (PCEP) Requirements and Protocol Extensions in Support of Global Concurrent Optimization", RFC 5557, July 2009.

[RFC5654] Niven-Jenkins, B., Brungard, D., Betts, M., Sprecher, N., and S. Ueno, "Requirements of an MPLS Transport Profile", RFC 5654, September 2009.

[RFC6007] Nishioka, I. and D. King, "Use of the Synchronization VECtor (SVEC) List for Synchronized Dependent Path Computations", RFC 6007, September 2010.

[RFC6373] Andersson, L., Berger, L., Fang, L., Bitar, N., and E. Gray, "MPLS Transport Profile (MPLS-TP) Control Plane Framework", RFC 6373, September 2011.

Author's Address

Wenjuan He (editor)
ZTE

Email: he.wenjuan1@zte.com.cn

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 9, 2012

C. Margaria, Ed.
Nokia Siemens Networks
O. Gonzalez de Dios, Ed.
Telefonica Investigacion y Desarrollo
F. Zhang, Ed.
Huawei Technologies
July 08, 2011

PCEP extensions for GMPLS
draft-ietf-pce-gmpls-pcep-extensions-03

Abstract

This memo provides extensions for the Path Computation Element communication Protocol (PCEP) for the support of GMPLS control plane.

Table of Contents

1. Introduction

PCEP RFCs [RFC5440], [RFC5521], [RFC5541], [RFC5520] are focused on path computation requests in MPLS networks. [RFC4655] defines the PCE framework also for GMPLS networks. This document complements these RFCs by providing some consideration of GMPLS applications and routing requests, for example for OTN and WSON networks.

The requirements on PCE extensions to support those characteristics are described in [I-D.ietf-pce-gmpls-aps-req] and [I-D.ietf-pce-wson-routing-wavelength].

1.1. Contributing Authors

Elie Sfeir, Franz Rambach (Nokia Siemens Networks) Francisco Javier Jimenez Chico (Telefonica Investigacion y Desarrollo) Suresh BR, Young Lee, SenthilKumar S, Jun Sun (Huawei Technologies), Ramon Casellas (CTTC)

1.2. PCEP requirements for GMPLS

This section provides a set of PCEP requirements to support GMPLS LSPs and assure signal compatibility in the path. When requesting a path computation (PCReq) to PCE, the PCC should be able to indicate, according to [I-D.ietf-pce-gmpls-aps-req] and to RSVP procedures like explicit label control (ELC), the following additional attributes:

(1) Switching capability: for instance PSC1-4, L2SC, TDM, LSC, FSC

(2) Encoding type: as defined in [RFC4202], [RFC4203], e.g., Ethernet, SONET/SDH, Lambda, etc.

(3) Signal Type: Indicates the type of elementary signal that constitutes the requested LSP. A lot of signal types with different granularity have been defined in SONET/SDH and G.709 ODUk, such as VC11, VC12, VC2, VC3 and VC4 in SDH, and ODU1, ODU2 and ODU3 in G.709 ODUk [RFC4606], [RFC4328] and other signal types like the one defined in [I-D.ceccarelli-ccamp-gmpls-ospf-g709] or [I-D.zhang-ccamp-gmpls-evolving-g709] .

(4) Concatenation Type: In SDH/SONET and G.709 OTN networks, two kinds of concatenation modes are defined: contiguous concatenation which requires co-route for each member signal and requires all the interfaces along the path to support this capability, and virtual concatenation which allows diverse routes for the member signals and only requires the ingress and egress interfaces to support this capability. Note that for the virtual concatenation, it also may specify co-routed or separated-routed. See [RFC4606]

and [RFC4328] about concatenation information.

(5) Concatenation Number: Indicates the number of signals that are requested to be contiguously or virtually concatenated. See also [RFC4606] and [RFC4328].

(6) Technology specific label(s) such as wavelength label as defined in [RFC6205]

(7) e2e Path protection type: as defined in [RFC4872], e.g., 1+1 protection, 1:1 protection, (pre-planned) rerouting, etc.

(8) Link Protection type: as defined in [RFC4203]

(9) Support for unnumbered interfaces: as defined in [RFC3477]

(10) Support for asymmetric bandwidth requests.

(11) Ability to indicate the requested granularity for the path ERO: node, link, label. This is to allow the use of the explicit label control of RSVP.

(12) In order to support the label control the Path computation response should provide label information matching signaling capabilities

(13) The PCC should be able to provide label restrictions similar to RSVP on the requests.

We describe in this document a proposal to fulfill those requirements.

1.3. PCEP existing objects related to GMPLS

PCEP as of [RFC5440], [RFC5521] and [I-D.ietf-pce-inter-layer-ext], supports the following information (in the PCReq and PCRep) related to the described requirements.

From [RFC5440]:

- numbered endpoints
- bandwidth (encoded as IEEE float)
- ERO
- LSP attributes (setup and holding priorities)

- Request attribute (include some LSP attributes)

From [RFC5521],Extensions to PCEP for Route Exclusions, definition of a XRO object and a new semantic (F bit):

- This object also allows to exclude (strict or not) resources; XRO include the diversity level (node, link, SRLG). The requested diversity is expressed in the XRO

- This Object with the F bit set indicates that the existing route is failed and resources present in the RRO can be reused.

From [I-D.ietf-pce-inter-layer-ext]:

- INTER-LAYER : indicates if inter-layer computation is allowed

- SWITCH-LAYER : indicates which layer(s) should be considered, can be used to represent the RSVP-TE generalized label request

- REQ-ADAP-CAP : indicates the adaptation capabilities requested, can also be used for the endpoints in case of mono-layer computation

The shortcomings of the existing PCEP information are:

The BANDWIDTH and LOAD-BALANCING objects do not describe the details of the traffic request (for example NVC, multiplier) in the context of GMPLS networks, for instance TDM or OTN networks.

The END-POINTS object does not allow specifying an unnumbered interface, nor the labels on the interface. Those parameters are of interest in case of switching constraints.

Current attributes do not allow to express the requested link level protection and end-to-end protection attributes.

The covered PCEP extensions are:

New objects are introduced (GENERALIZED-BANDWIDTH and GENERALIZED-LOAD-BALANCING) for flexible bandwidth encoding,

New Objects are introduced (LABEL-SET and SUGGESTED-LABEL-SET) on order to allow the PCC to restrict/influence the range of labels returned

A new object type is introduced for the END-POINTS object (generalized-endpoint),

A new TLV is added to the LSPA object.

In order to indicate the mandatory routing granularity in the response, a new flag in the RP object is added.

1.4. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119.

2. PCEP objects and extensions

This section describes the required PCEP objects and extensions. The PCReq and PCRep messages are defined in [RFC5440]. The format of the request and response messages with the proposed extensions (GENERALIZED-BANDWIDTH, GENERALIZED-LOAD-BALANCING, SUGGESTED-LABEL-SET and LABEL-SET) is as follows:

```
<request>::= <RP>
             <segment-computation>|<path-key-expansion>

<segment-computation> ::=
  <END-POINTS>
  [<LSPA>]
  [<BANDWIDTH>]
  [<GENERALIZED-BANDWIDTH>...]
  [<metric-list>]
  [<OF>]
  [<RRO> [<BANDWIDTH>] [<GENERALIZED-BANDWIDTH>...]]
  [<IRO>]
  [<SUGGESTED-LABEL-SET>]
  [<LABEL-SET>...]
  [<LOAD-BALANCING>]
  [<GENERALIZED-LOAD-BALANCING>...]
  [<XRO>]

<path-key-expansion> ::= <PATH-KEY>

<response>::=<RP>
  [<NO-PATH>]
  [<attribute-list>]
  [<path-list>]

<path-list>::=<path>[<path-list>]
<path>::= <ERO><attribute-list>
<metric-list>::=<METRIC>[<metric-list>]
```

Where:

```
 <attribute-list>::=[<LSPA>]
 [<BANDWIDTH>]
 [<LABEL-SET>...]
 [<SUGGESTED-LABEL-SET>...]
 [<GENERALIZED-BANDWIDTH>...]
 [<GENERALIZED-LOAD-BALANCING>...]
 [<metric-list>]
 [<IRO>]
```

For point-to-multipoint(P2MP) computations, the proposed grammar is:

```
 <segment-computation> ::=
     <end-point-rro-pair-list>
     [<LSPA>]
     [<BANDWIDTH>]
     [<GENERALIZED-BANDWIDTH>...]
     [<metric-list>]
     [<IRO>]
     [<SUGGESTED-LABEL-SET>]
     [<LABEL-SET>]
     [<LOAD-BALANCING>]
     [<GENERALIZED-LOAD-BALANCING>...]
     [<XRO>]

  <end-point-rro-pair-list>::=
           <END-POINTS>[<RRO-List>][<BANDWIDTH>]
          [<GENERALIZED-BANDWIDTH>...]
          [<end-point-rro-pair-list>]

  <RRO-List>::=<RRO>[<BANDWIDTH>]
  [< GENERALIZED-BANDWIDTH>...][<RRO-List>]
```

2.1. RP object extension

Explicit label control (ELC) is a procedure supported by RSVP-TE, where the outgoing label(s) is(are) encoded in the ERO. In consequence, the PCE may be able to provide such label(s) directly in the path ERO. The PCC, depending on policies or switching layer, may be required to use explicit label control or expect explicit link, thus it need to indicate in the PCReq which granularity it is expecting in the ERO. This correspond to requirement 11 of [I-D.ietf-pce-gmpls-aps-req] The possible granularities can be node, link, label. The granularities are inter-dependent, in the sense that link granularity imply the presence of node information in the ERO, similarly a label granularity imply that the ERO contain node, link and label information.

A new 2-bit routing granularity (RG) flag is defined in the RP object. The values are defined as follows

0 : node

1 : link

2 : label

3 : reserved

When the RP object appears in a request within a PCReq message the flag indicates the requested route granularity. The PCE MAY try to follow this granularity and MAY return a NO-PATH if the requested granularity cannot be provided. The PCE MAY return more details on the route based on its policy. The PCC can decide if the ERO is acceptable based on its content.

If a PCE did use the requested routing granularity in a PCReq is MUST indicate the routing granularity in the PCRep. The RG flag is backward-compatible with previous RFCs: the value sent by an implementation not supporting it will indicate a node granularity. This flag is optional for responses. A new capability flag in the PCE-CAP-FLAGS from [RFC5088] and [RFC5089] may be added.

2.2. Traffic parameters encoding, GENERALIZED-BANDWIDTH

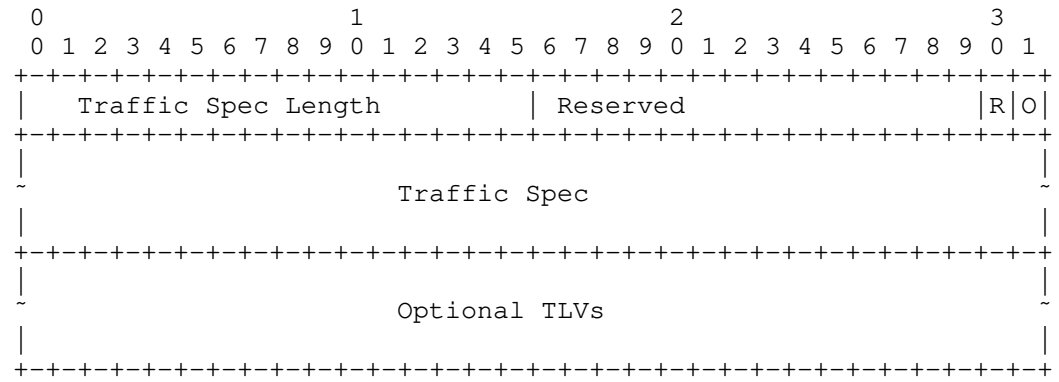
The PCEP BANDWIDTH does not describe the details of the signal (for example NVC, multiplier), hence the bandwidth information should be extended to use the RSVP Tspec object encoding. The PCEP BANDWIDTH object defines two types: 1 and 2. C-Type 2 is representing the existing bandwidth in case of re-optimization.

The following possibilities cannot be represented in the BANDWIDTH object:

- Asymmetric bandwidth (different bandwidth in forward and reverse direction), as described in [RFC5467]
- GMPLS (SDH/SONET, G.709, ATM, MEF etc) parameters are not supported.

This correspond to requirement 3,4,5 and 10 of [I-D.ietf-pce-gmpls-aps-req].

According to [RFC5440] the BANDWIDTH object has no TLV and has a fixed size of 4 bytes. This definition does not allow extending it with the required information. To express this information, a new object named GENERALIZED-BANDWIDTH having the following format is defined:



The GENERALIZED-BANDWIDTH has a variable length. The Traffic spec length field indicates the length of the Traffic spec field. The bits R and O have the following meaning:

O bit : when set the value refers to the previous bandwidth in case of re-optimization

R bit : when set the value refers to the bandwidth of the reverse direction

The Object type determines which type of bandwidth is represented by the object. The following object types are defined:

1. Intserv

2. SONET/SDH

3. G.709

4. Ethernet

The encoding of the field Traffic Spec is the same as in RSVP-TE, it can be found in the following references.

Object Type	Name	Reference
0	Reserved	
1	Reserved	
2	Intserv	[RFC2210]
3	Re served	
4	SONET/SDH	[RFC4606]
5	G.709	[RFC4328]
6	Ethernet	[RFC6003]

Traffic Spec field encoding

The GENERALIZED-BANDWIDTH MAY appear more than once in a PCReq message. If more than one GENERALIZED-BANDWIDTH have the same Object Type, Reserved, R and O values, only the first one is processed, the others are ignored.

a PCE MAY ignore GENERALIZED-BANDWIDTH objects, a PCC that requires a GENERALIZED-BANDWIDTH to be used can set the P (Processing) bit in the object header.

When a PCC needs to get a bi-directional path with asymmetric bandwidth, it SHOULD specify the different bandwidth in forward and reverse directions through two separate GENERALIZED-BANDWIDTH objects. If the PCC set the P bit on both object the PCE MUST compute a path that satisfies the asymmetric bandwidth constraint and return the path to PCC if the path computation is successful. If the P bit on the reverse GENERALIZED-BANDWIDTH object the PCE MAY ignore this constraint.

a PCE MAY include the GENERALIZED-BANDWIDTH objects in the response to indicate the GENERALIZED-BANDWIDTH of the path

Optional TLVs may be included within the object body to specify more specific bandwidth requirements. The specification of such TLVs is outside the scope of this document.

2.3. Traffic parameters encoding, GENERALIZED-LOAD-BALANCING

The LOAD-BALANCING object is used to request a set of maximum Max-LSP TE-LSP having in total the bandwidth specified in BANDWIDTH, each TE-LSP having a minimum of min-bandwidth bandwidth. The LOAD-BALANCING

follows the bandwidth encoding of the BANDWIDTH object, it does not describe enough details for the traffic specification expected by GMPLS. A PCC should be allowed to request a set of TE-LSP also in case of GMPLS traffic specification.

According to [RFC5440] the LOAD-BALANCING object has no TLV and has a fixed size of 8 bytes. This definition does not allows extending it with the required information. To express this information, a new Object named GENERALIZED-LOAD-BALANCING is defined.

The GENERALIZED-LOAD-BALANCING object, as the LOAD-BALANCING object, allows the PCC to request a set of TE-LSP having in total the GENERALIZED-BANDWIDTH traffic specification with potentially Max-Lsp, each TE-LSP having a minimum of Min Traffic spec. The GENERALIZED-LOAD-BALANCING is optional.

GENERALIZED-LOAD-BALANCING Object-Class is to be assigned by IANA. The GENERALIZED-LOAD-BALANCING Object type determines which type of minimum bandwidth is represented by the object. The following object types are defined:

1. Intserv

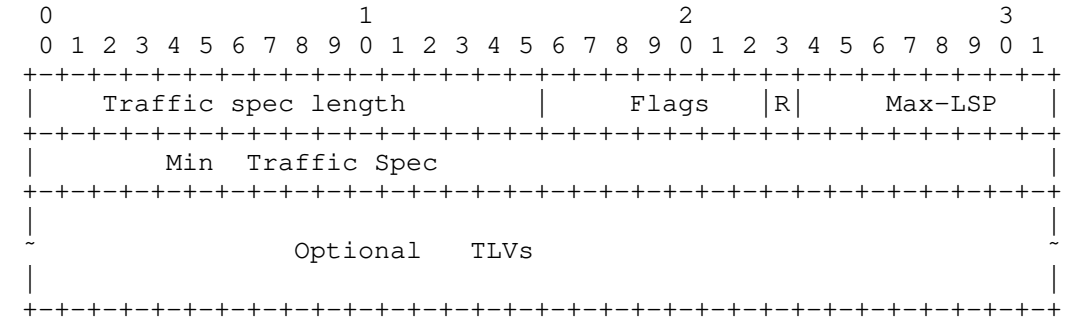
2. SONET/SDH

3. G.709

4. Ethernet

The GENERALIZED-LOAD-BALANCING has a variable length.

The format of the GENERALIZED-LOAD-BALANCING object body is as follows:



Traffic spec length (16 bits): the total length of the min traffic specification. It should be noted that the RSVP traffic

specification may also include TLV different than the PCEP TLVs.

Flags (8 bits): The undefined Flags field MUST be set to zero on transmission and MUST be ignored on receipt. The following flag is defined:

R Flag : (1 bit) set when the value refer to the bandwidth of the reverse direction

Max-LSP (8 bits): maximum number of TE LSPs in the set.

Min-Traffic spec (variable): Specifies the minimum traffic spec of each element of the set of TE LSPs.

The encoding of the field Traffic Spec is the same as in RSVP-TE, it can be found in the following references.

Object Type	Name	Reference
2	Intserv	[RFC2210]
4	SONET/SDH	[RFC4606]
5	G.709	[RFC4328]
6	Ethernet	[RFC6003]

Traffic Spec field encoding

The GENERALIZED-LOAD-BALANCING MAY appear more than once in a PCReq message. If more than one GENERALIZED-LOAD-BALANCING have the same Object Type, and R Flag, only the first one is processed, the others are ignored.

a PCE MAY ignore GENERALIZED-LOAD-BALANCING objects. A PCC that requieres a GENERALIZED-LOAD-BALANCING to be used can set the P (Processing) bit in the object header.

When a PCC needs to get a bi-directional path with asymmetric bandwidth, it SHOULD specify the different bandwidth in forward and reverse directions through two separate GENERALIZED-LOAD-BALANCING objects with different R Flag. If the PCC set the P bit on both object the PCE MUST compute a path that satisfies the asymmetric bandwidth constraint and return the path to PCC if the path computation is successful. If the P bit on the reverse GENERALIZED-LOAD-BALANCING object the PCE MAY ignore this constraint.

Optional TLVs may be included within the object body to specify more

specific bandwidth requirements. The specification of such TLVs is outside the scope of this document.

The GENERALIZED-LOAD-BALANCING object has the same semantic as the LOAD-BALANCING object; If a PCC requests the computation of a set of TE LSPs so that the total of their generalized bandwidth is X, the maximum number of TE LSPs is N, and each TE LSP must at least have a bandwidth of B, it inserts a GENERALIZED-BANDWIDTH object specifying X as the required bandwidth and a GENERALIZED-LOAD-BALANCING object with the Max-LSP and Min-traffic spec fields set to N and B, respectively.

For example a request for one co-signaled n x VC-4 TE-LSP will not use the GENERALIZED-LOAD-BALANCING. In case the V4 components can use different paths, the GENERALIZED-BANDWIDTH will contain a traffic specification indicating the complete n x VC4 traffic specification and the GENERALIZED-LOAD-BALANCING the minimum co-signaled VC4. For a SDH network, a request to have a TE-LSP group with 10 VC4 container, each path using at minimum 2VC4 container, can be represented with a GENERALIZED-BANDWIDTH object with OT=4, the content of the Traffic specification is ST=6,RCC=0,NCC=0,NVC=10,MT=1. The GENERALIZED-LOAD-BALANCING, OT=4,R=0,Max-LSP=5, min Traffic spec is (ST=6,RCC=0,NCC=0,NVC=2,MT=1). The PCE can respond with a response with maximum 5 path, each of then having a GENERALIZED-BANDWIDTH OT=4,R=0, and traffic spec matching the minimum traffic spec from the GENERALIZED-LOAD-BALANCING object of the corresponding request.

2.4. END-POINTS Object extensions

The END-POINTS object is used in a PCReq message to specify the source and destination of the path for which a path computation is requested. From [RFC3471] the source IP address and the destination IP address are used to identify those. A new Object Type is defined to address the following possibilities:

- Different endpoint types.
- Label restrictions on the endpoint.
- Specification of unnumbered endpoints type as seen in GMPLS networks.

The Object encoding is described in the following sections.

2.4.1. Generalized Endpoint Object Type

In GMPLS context the endpoints can:

- Be unnumbered
- Have label(s) associated to them
- May have different switching capabilities

The IPv4 and IPv6 endpoints are used to represent the source and destination IP addresses. The scope of the IP address (Node or Link) is not explicitly stated. It should also be possible to request a Path between a numbered link and an unnumbered link, or a P2MP path between different type of endpoints.

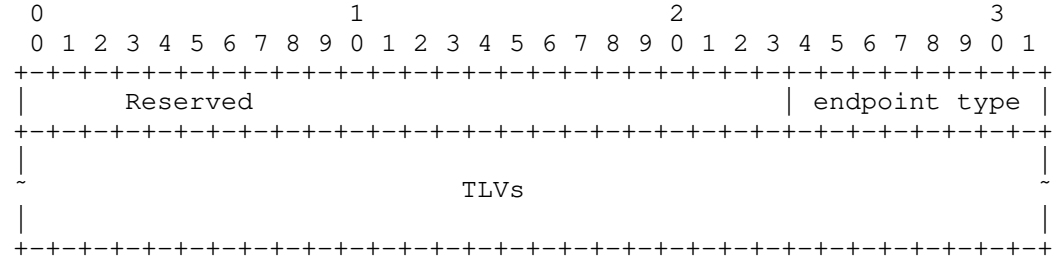
Since the PCEP END-POINTS object only support endpoints of the same type a new C-Type is proposed that support different endpoint types, including unnumbered. This new C-Type also supports the specification of constraints on the endpoint label to be use. The PCE might know the interface restrictions but this is not a requirement. On the path calculation request only the Tspec and switch layer need to be coherent, the endpoint labels could be different (supporting a different Tspec). Hence the label restrictions include a Generalized label request in order to interpret the labels. This correspond to requirement 6 and 9 of [I-D.ietf-pce-gmpls-aps-req].

The proposed object format consists of a body and a list of TLVs, which give the details of the endpoints and are described in Section 2.4.2. For each endpoint type, a different grammar is defined. The TLVs defined to describe an endpoint are:

1. IPv4 address.
2. IPv6 address.
3. Unnumbered endpoint.
4. Label request.
5. Label.
6. Upstream label.
7. Label set.

8. Suggested label set.

The labels TLV are used to restrict the label allocation in the PCE. They follow the set of restrictions provided by signaling with explicit value (label and upstream label), mandatory range restrictions (Label set) and optional range restriction (suggested label set). Single suggested value is using the suggested label set. The label range restriction are valid in GMPLS networks, either by PCC policy or depending on the switching technology used, for instance on given Ethernet or ODU equipment having limited hardware capabilities restricting the label range. Label set restriction also applies to WSON networks where the optical sender and receivers are limited in their frequency tunability ranges, restricting then in GMPLS the possible label ranges on the interface. The END-POINTS Object with Generalized Endpoint object type is encoded as follow:



Reserved bits should be set to 0 when a message is sent and ignored when the message is received

the endpoint type is defined as follow:

Value	Type	Meaning
0	Point-to-Point	
1	Point-to-Multipoint	New leaves to add
2		Old leaves to remove
3		Old leaves whose path can be modified/reoptimized
4		Old leaves whose path must be left unchanged
5-244	Reserved	
245-255	Experimental range	

The endpoint type is used to cover both point-to-point and different point-to-multipoint endpoint semantic. Endpoint type 0 MAY be accepted by the PCE, other endpoint type MAY be supported if the PCE implementation supports P2MP path calculation. A PCE not supporting a given endpoint type MUST respond with a PCErr with error code "Path computation failure", error type "Unsupported endpoint type in END-POINTS Generalized Endpoint object type". The TLVs present in the object body MUST follow the following grammar:

```
<generalized-endpoint-tlvs>::=
  <p2p-endpoints> | <p2mp-endpoints>

<p2p-endpoints> ::=
  <source-endpoint>
  <destination-endpoint>

<source-endpoint> ::=
  <endpoint>
  [<endpoint-restriction-list>]

<destination-endpoint> ::=
  <endpoint>
  [<endpoint-restriction-list>]

<p2mp-endpoints> ::=
  <endpoint> [<endpoint-restriction-list>]
  [<endpoint> [<endpoint-restriction-list>]]...
```

For endpoint type Point-to-Multipoint several endpoint objects may be

present in the message and represent a leave, exact meaning depend on the endpoint type defined of the object.

An endpoint is defined as follows:

```
 <endpoint>::=<IPV4-ADDRESS>|<IPV6-ADDRESS>|<UNNUMBERED-ENDPOINT>
 <endpoint-restriction-list> ::=
              <endpoint-restriction>
              [<endpoint-restriction-list>]

 <endpoint-restriction> ::=
                  <LABEL-REQUEST><label-restriction-list>

 <label-restriction-list> ::= <label-restriction>
                              [<label-restriction-list>]
 <label-restriction> ::= <LABEL>|<UPSTREAM-LABEL>|
                         <LABEL-SET>|
                         <SUGGESTED-LABEL-SET>
```

The different TLVs are described in the following sections. A PCE MAY support IPV4-ADDRESS,IPV6-ADDRESS or UNNUMBERED-ENDPOINT TLV. A PCE not supporting one of those TLV in a PCReq MUST respond with a PCRep with NO-PATH with the bit "Unknown destination" or "Unknown source" in the NO-PATH-VECTOR TLV, the PCRep MUST include the ENDPOINT object in the response with only the TLV it did not understood.

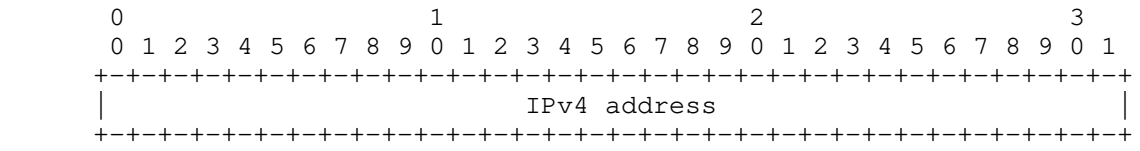
A PCE MAY support LABEL-REQUEST, LABEL, UPSTREAM-LABEL, LABEL-SET or SUGGESTED-LABEL-SET TLV. A PCE not supporting one of those TLV in a PCReq MUST respond with a PCRep with NO-PATH with the bit "No endpoint label resource" or "No endpoint label resource in range" in the NO-PATH-VECTOR TLV, the PCRep MUST include the ENDPOINT object in the response with only the TLV it did not understood or could not met the the constraint.

2.4.2. END-POINTS TLVs extensions

All endpoint TLVs have the standard PCEP TLV header as defined in [RFC5440] section 7.1

2.4.2.1. IPV4-ADDRESS

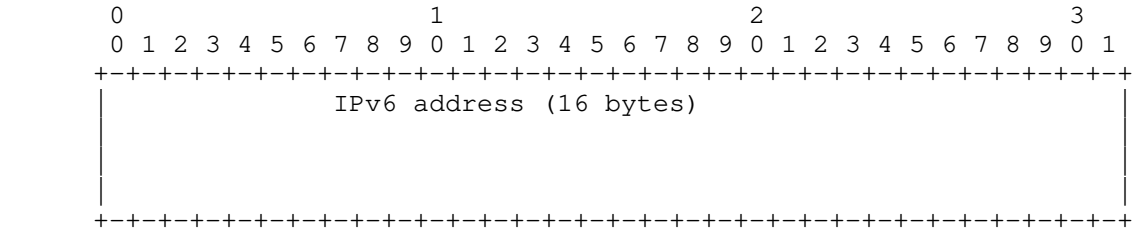
This TLV represent a numbered endpoint using IPv4 numbering, the format of the IPv4-ADDRESS TLV value (TLV-Type=TBA) is as follows:



This TLV MAY be ignored, in which case a PCRep with NO-PATH should be responded, as described in Section 2.4.1.

2.4.2.2. IPV6-ADDRESS TLV

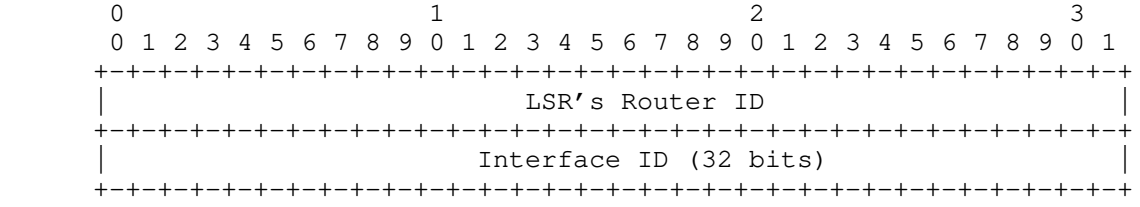
This TLV represent a numbered endpoint using IPV6 numbering, the format of the IPv6-ADDRESS TLV value (TLV-Type=TBA) is as follows:



This TLV MAY be ignored, in which case a PCRep with NO-PATH should be responded, as described in Section 2.4.1.

2.4.2.3. UNNUMBERED-ENDPOINT TLV

This TLV represent an unnumbered interface. This TLV has the same semantic as in [RFC3477] The TLV value is encoded as follow (TLV-Type=TBA)



This TLV MAY be ignored, in which case a PCRep with NO-PATH should be responded, as described in Section 2.4.1.

2.4.2.4. LABEL-REQUEST TLV

The LABEL-REQUEST TLV indicates the switching capability and encoding type of the label restriction list. Its format is the same as described in [RFC3471] Section 3.1 Generalized label request. The

LABEL-REQUEST TLV use TLV-Type=TBA. The fields are encoded as in the RSVP-TE. The Encoding Type indicates the encoding type, e.g., SONET/ SDH/GigE etc., that will be used with the data associated with the LSP. The Switching type indicates the type of switching that is being requested on the link. G-PID identifies the payload of the TE-LSP. This TLV and the following one are introduced to satisfy requirement 13 for the endpoint.

This TLV MAY be ignored, in which case a PCRep with NO-PATH should be responded, as described in Section 2.4.1.

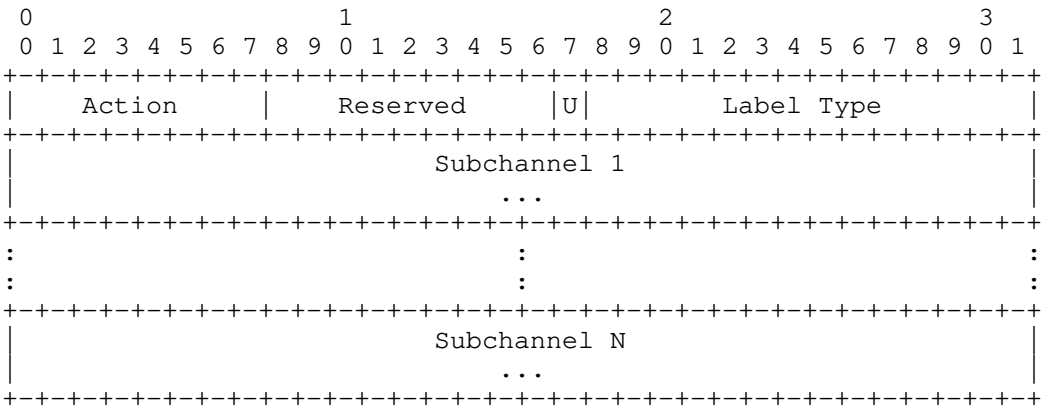
2.4.2.5. Labels TLV

Label or label range restrictions may be specified for the TE-LSP endpoints. Those are encoded in the TLVs. The label value need to be interpreted with a description on the Encoding and switching type. The REQ-ADAP-CAP object from [I-D.ietf-pce-inter-layer-ext] can be used in case of mono-layer request, however in case of multilayer it is possible to have in the future more than one object, so it is better to have a dedicated TLV for the label and label request (the scope is then more clear). Those TLV MAY be ignored, in which case a PCRep with NO-PATH should be responded, as described in Section 2.4.1. TLVs are encoded as follow (following [RFC5440]) :

- LABEL TLV, Type=TBA. The TLV Length is variable, the value is the same as [RFC3471] Section 3.2 Generalized label. This represent the downstream label

- UPSTREAM-LABEL TLV, Type=TBA, The TLV Length is variable, the value is the same as [RFC3471] Section 3.2 Generalized label. This represent the upstream label

- LABEL-SET TLV, Type=TBA. The TLV Length is variable, Encoding follow [RFC3471] Section 3.5 "Label set" with the addition of a U bit : the U bit is set for upstream direction in case of bidirectional LSP.



- o SUGGESTED-LABEL-SET TLV Set, Type=TBA. The TLV length is variable, Encoding is as LABEL-SET TLV.

A LABEL TLV represent the label used on the unnumbered interface, bit U is used to indicate which exact direction is considered. The label type indicates which type of label is carried. A LABEL-SET TLV represents a set of possible labels that can be used on the unnumbered interface. the label allocated on the first link SHOULD be within the label set range. The action parameter in the Label set indicates the type of list provided. Those parameters are described by [RFC3471] section 3.5.1 A SUGGESTED-LABEL-SET TLV has the same encoding as the LABEL-SET TLV, it indicates to the PCE a set of preferred (ordered) set of labels to be used. the PCE MAY use those labels for label allocation.

The U bit has the following meaning:

U: Upstream direction: set when the label or label set is in the reverse direction

2.5. LABEL-SET object

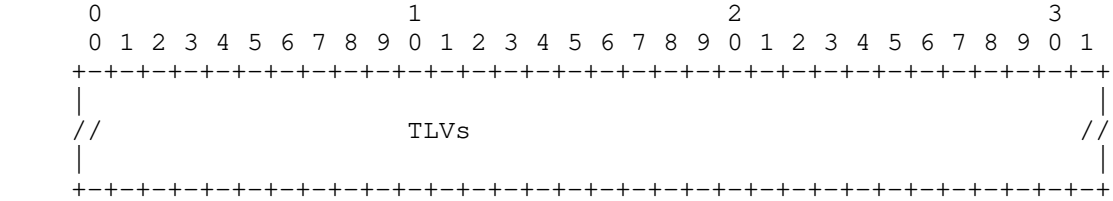
The LABEL-SET object is carried in a request within a PCReq message to restrict the set of labels to be assigned during the path computation. This is introduced to satisfy requirement 13.

When the P bit is set and the object accepted any label allocated by the PCE (and included in the ERO object on the response) MUST be in the range stated in the LABEL-SET. When no path satisfy this constraint a PCRep with a NO-PATH should be responded wit a NO-PATH-VECTOR TLV with the bit "No label resource in range" set and the

LABEL-SET object MAY be included to indicate the set of constraint that could not be satisfied.

When the P bit is not set a PCE MAY consider constraint, the PCC can verify that the constraint was applied by checking the ERO returned

The LABEL-SET Object encoding is defined as following



where TLVs follow the following grammar

<label-set-tlvs> ::= <LABEL-REQUEST><LABEL-SET>[<LABEL-SET>]

The LABEL-REQUEST and LABEL-SET TLVs are as defined in Section 2.4.2.5, See also [RFC3471] and [RFC3473] for the definitions of the fields.

It is allowed to have more than one LABEL-SET object per request within a PCReq message (for example in case of multiple SWITCH-LAYER present).

2.6. SUGGESTED-LABEL-SET object

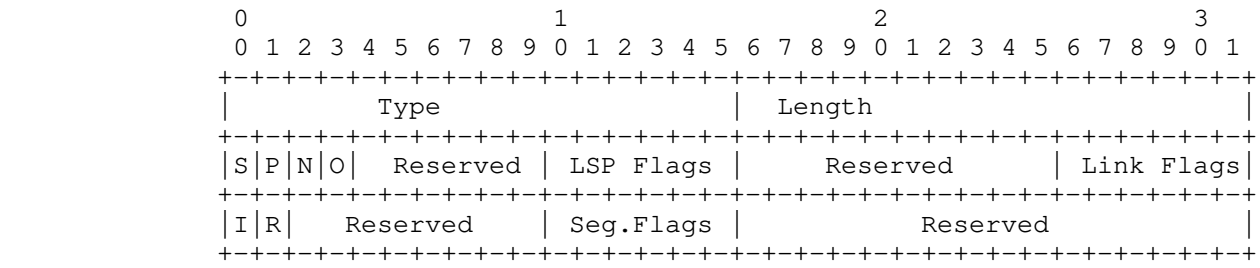
Similar to the endpoint restriction SUGGESTED-LABEL-SET TLV, but with end-to-end scope the SUGGESTED-LABEL-SET object indicate an optional set of label that the PCE MAY use when selecting the labels. The SUGGESTED-LABEL-SET object is carried within a PCReq or PCRep message to indicate the preferred set of label to be assigned during the path computation. The encoding is the same as the LABEL-SET object. It is allowed to have more than one SUGGESTED LABEL-SET object per PCReq (for example in case of multiple SWITCH-LAYER present).

This object is introduced similarly to the LABEl-SET to satisfy the requirement 6 and 13, more specifically the ability to indicate optional preference for the label selection support by RSVP using the SUGGESTED_LABEL.

2.7. LSPA extensions

The LSPA carries the LSP attributes. In the end-to-end protection context this also includes the protection state information. This

object is introduced to fulfill requirement 7 and is used as a policy input for route and label selection. The LSPA object can be extended by a protection TLV type: Type TBA: PROTECTION-ATTRIBUTE



The content is as defined in [RFC4872], [RFC4873].

LSP Flags can be considered for routing policy based on the protection type. The other attributes are only meaningful for a s_ateful PCE.

This TLV is optional and MAY be ignored by the PCE, in which case MUST NOT include the TLV in the LSPA, if present, of the PCRep. When the TLV is used by the PCE, a LSPA object and the PROTECTION-ATTRIBUTE TLV MUST be included in the PCRep. Fields that were not considered MUST be set to 0.

2.8. NO-PATH Object Extension

The NO-PATH object is used in PCRep messages in response to an unsuccessful path computation request (the PCE could not find a path satisfying the set of constraints). In this scenario, PCE MUST include a NO-PATH object in the PCRep message. The NO-PATH object may carries the NO-PATH-VECTOR TLV that specifies more information on the reasons that led to a negative reply. In case of GMPLS networks there could be some more additional constraints that led to the failure like protection mismatch, lack of resources, and so on. Few new flags have been introduced in the 32-bit flag field of the NO-PATH-VECTOR TLV and no modifications have been made in the NO-PATH object.

2.8.1. Extensions to NO-PATH-VECTOR TLV

The modified NO-PATH-VECTOR TLV carrying the additional information is as follows: New fields PM and NR are defined in the 23th and 22th bit of the Flags field respectively.

Bit number TBA - Protection Mismatch (1-bit). Specifies the mismatch of the protection type in the PROTECTION-ATTRIBUTE TLV in

the request.

Bit number TBA - No Resource (1-bit). Specifies that the resources are not currently sufficient to provide the path.

Bit number TBA - Granularity not supported (1-bit). Specifies that the PCE is not able to provide a route with the requested granularity.

Bit number TBA - No endpoint label resource (1-bit). Specifies that the PCE is not able to provide a route because of the endpoint label restriction.

Bit number TBA - No endpoint label resource in range (1-bit). Specifies that the PCE is not able to provide a route because of the endpoint label set restriction.

Bit number TBA - No label resource in range (1-bit). Specifies that the PCE is not able to provide a route because of the label set restriction.

3. Additional Error Type and Error Values Defined

A PCEP-ERROR object is used to report a PCEP error and is characterized by an Error-Type that specifies the type of error while Error-value that provides additional information about the error type. An additional error type and few error values are defined to represent some of the errors related to the newly identified objects related to SDH networks. For each PCEP error, an Error-Type and an Error-value are defined. Error-Type 1 to 10 are already defined in [RFC5440]. Additional Error- values are defined for Error-Type 10 and A new Error-Type is introduced (value TBA).

Error-Type	Error-value	
10	Reception of an invalid object	
	Error-value=TBA:	Bad Generalized Bandwidth Object value.
	Error-value=TBA:	Unsupported LSP Protection Type in PROTECTION-ATTRIBUTE TLV.
	Error-value=TBA:	Unsupported LSP Protection Flags in PROTECTION-ATTRIBUTE TLV.
	Error-value=TBA:	Unsupported Secondary LSP Protection Flags in PROTECTION-ATTRIBUTE TLV.
	Error-value=TBA:	Unsupported Link Protection Type in PROTECTION-ATTRIBUTE TLV.
	Error-value=TBA:	Unsupported Link Protection Type in PROTECTION-ATTRIBUTE TLV.
TBA	Path computation failure	
	Error-value=TBA:	Unacceptable request message.
	Error-value=TBA:	Generalized bandwidth object not supported.
	Error-value=TBA:	Label Set constraint could not be met.
	Error-value=TBA:	Label constraint could not be met.
	Error-value=TBA:	Unsupported endpoint type in END-POINTS Generalized Endpoint object type

Error-value=TBA: Unsupported TLV present in END-POINTS Generalized Endpoint object type

Error-value=TBA: Unsupported granularity in the RP object flags

4. Manageability Considerations

Liveness Detection and Monitoring This document makes no change to the basic operation of PCEP and so there are no changes to the requirements for liveness detection and monitoring set out in [RFC4657] and [RFC5440].

5. IANA Considerations

IANA assigns values to the PCEP protocol objects and TLVs. IANA is requested to make some allocations for the newly defined objects and TLVs introduced in this document. Also, IANA is requested to manage the space of flags that are newly added in the TLVs.

5.1. PCEP Objects

As described in Section 2.2 and Section 2.3new Objects are defined IANA is requested to make the following Object-Type allocations from the "PCEP Objects" sub-registry.

Object Class to be assigned

Name GENERALIZED-BANDWIDTH

Object-Type 0 to 6

Reference This document (section Section 2.2)

Object Class to be assigned

Name GENERALIZED-LOAD-BALANCING

Object-Type 0 to 6

Reference This document (section Section 2.3)

Object Class to be assigned

Name LABEL-SET

Object-Type 0

Reference This document (section Section 2.5)

Object Class to be assigned

Name SUGGESTED-LABEL-SET

Object-Type 0

Reference This document (section Section 2.6)

As described in Section 2.4.1 a new Object type is defined IANA is requested to make the following Object-Type allocations from the "PCEP Objects" sub-registry. The values here are suggested for use by IANA.

Object Class 4

Name END-POINTS

Object-Type 5 : Generalized Endpoint

6-15 : unassigned

Reference This document (section Section 2.2)

5.2. END-POINTS object, Object Type Generalized Endpoint

IANA is requested to create a registry to manage the endpoint type field of the END-POINTS object, Object Type Generalized Endpoint and manage the code space.

New endpoint type in the Reserved range may be allocated by an IETF consensus action. Each endpoint type should be tracked with the following qualities:

- endpoint type
- Description
- Defining RFC

New endpoint type in the Experimental range are for experimental use; these will not be registered with IANA and MUST NOT be mentioned by RFCs.

The following values have been defined by this document. (Section 2.4.1, Table 4):

Value	Type	Meaning
0	Point-to-Point	
1	Point-to-Multipoint	New leaves to add
2		Old leaves to remove
3		Old leaves whose path can be modified/reoptimized
4		Old leaves whose path must be left unchanged
5-244	Reserved	
245-255	Experimental range	

5.3. New PCEP TLVs

IANA manages the PCEP TLV code point registry (see [RFC5440]). This is maintained as the "PCEP TLV Type Indicators" sub-registry of the "Path Computation Element Protocol (PCEP) Numbers" registry. This document defines new PCEP TLVs, to be carried in the END-POINTS object with Generalized Endpoint object Type. IANA is requested to do the following allocation. The values here are suggested for use by IANA.

Value	Meaning	Reference
7	IPv4 endpoint	This document (section Section 2.4.2.1)
8	IPv6 endpoint	This document (section Section 2.4.2.2)
9	Unnumbered endpoint	This document (section Section 2.4.2.3)
10	Label request	This document (section Section 2.4.2.4)
11	Requested GMPLS Label	This document (section Section 2.4.2.5)
12	Requested GMPLS Upstream Label	This document (section Section 2.4.2.5)

13	Requested GMPLS Label Set	This document (section Section 2.4.2.5)
14	Suggested GMPLS Label Set	This document (section Section 2.4.2.5)
15	LSP Protection Information	This document (section Section 2.7)

5.4. RP Object Flag Field

As described in Section 2.1 new flag are defined in the RP Object Flag IANA is requested to make the following Object-Type allocations from the "RP Object Flag Field" sub-registry. The values here are suggested for use by IANA.

Bit	Description	Reference
bit 17-16	routing granularity (RG)	This document, Section 2.1

5.5. New PCEP Error Codes

As described in Section Section 3, new PCEP Error-Type and Error Values are defined. IANA is requested to make the following allocation in the "PCEP-ERROR Object Error Types and Values" registry. The values here are suggested for use by IANA.

Error	name	Reference
Type=10	Reception of an invalid object	[RFC5440]
Value=2:	Bad Generalized Bandwidth Object value.	This Document
Value=3:	Unsupported LSP Protection Type in PROTECTION-ATTRIBUTE TLV.	This Document
Value=4:	Unsupported LSP Protection Flags in PROTECTION-ATTRIBUTE TLV.	This Document
Value=5:	Unsupported Secondary LSP Protection Flags in PROTECTION-ATTRIBUTE TLV.	This Document
Value=6:	Unsupported Link Protection Type in PROTECTION-ATTRIBUTE TLV.	This Document
Value=7:	Unsupported Link Protection Type in PROTECTION-ATTRIBUTE TLV.	This Document
Type=14	Path computation failure	This Document
Value=1:	Unacceptable request message.	This Document
Value=2:	Generalized bandwidth object not supported.	This Document
Value=3:	Label Set constraint could not be met.	This Document
Value=4:	Label constraint could not be met.	This Document
Value=5:	Unsupported endpoint type in END-POINTS Generalized Endpoint object type	This Document
Value=6:	Unsupported TLV present in END-POINTS Generalized Endpoint object type	This Document
Value=7:	Unsupported granularity in the RP object flags	This Document

5.6. New NO-PATH-VECTOR TLV Fields

As described in Section Section 2.8.1, new NO-PATH-VECTOR TLV Flag Fields have been defined. IANA is requested to do the following allocations in the "NO-PATH-VECTOR TLV Flag Field" sub-registry. The values here are suggested for use by IANA.

Bit number 23 - Protection Mismatch (1-bit). Specifies the mismatch of the protection type of the PROTECTION-ATTRIBUTE TLV in the request.

Bit number 22 - No Resource (1-bit). Specifies that the resources are not currently sufficient to provide the path.

Bit number 21 - Granularity not supported (1-bit). Specifies that the PCE is not able to provide a route with the requested granularity.

Bit number 20 - No endpoint label resource (1-bit). Specifies that the PCE is not able to provide a route because of the endpoint label restriction.

Bit number 19 - No endpoint label resource in range (1-bit). Specifies that the PCE is not able to provide a route because of the endpoint label set restriction.

Bit number 18 - No label resource in range (1-bit). Specifies that the PCE is not able to provide a route because of the label set restriction.

6. Security Considerations

None.

7. Contributing Authors

Nokia Siemens Networks:

Elie Sfeir
St Martin Strasse 76
Munich, 81541
Germany

Phone: +49 89 5159 16159
Email: elie.sfeir@nsn.com

Franz Rambach
St Martin Strasse 76
Munich, 81541
Germany

Phone: +49 89 5159 31188
Email: franz.rambach@nsn.com

Francisco Javier Jimenez Chico
Telefonica Investigacion y Desarrollo
C/ Emilio Vargas 6
Madrid, 28043
Spain

Phone: +34 91 3379037
Email: fjjc@tid.es

Huawei Technologies

Suresh BR
Shenzhen
China
Email: sureshbr@huawei.com

Young Lee
1700 Alma Drive, Suite 100
Plano, TX 75075
USA

Phone: (972) 509-5599 (x2240)
Email: ylee@huawei.com

SenthilKumar S
Shenzhen
China
Email: senthilkumars@huawei.com

Jun Sun
Shenzhen
China
Email: johnsun@huawei.com

CTTC - Centre Tecnologic de Telecomunicacions de Catalunya

Ramon Casellas
PMT Ed B4 Av. Carl Friedrich Gauss 7
08860 Castelldefels (Barcelona)
Spain
Phone: (34) 936452916
Email: ramon.casellas@cttc.es

8. Acknowledgments

The research of Ramon Casellas, Francisco Javier Jimenez Chico, Oscar Gonzalez de Dios, Cyril Margaria, and Franz Rambach leading to these results has received funding from the European Community's Seventh Framework Programme FP7/2007-2013 under grant agreement no 247674.

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.

[RFC2210] Wroclawski, J., "The Use of RSVP with IETF Integrated Services", RFC 2210, September 1997.

[RFC3471] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, January 2003.

[RFC3473] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.

[RFC3477] Kompella, K. and Y. Rekhter, "Signalling Unnumbered Links in Resource ReSerVation Protocol - Traffic Engineering (RSVP-TE)", RFC 3477, January 2003.

[RFC4202] Kompella, K. and Y. Rekhter, "Routing Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4202, October 2005.

[RFC4203] Kompella, K. and Y. Rekhter, "OSPF Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4203, October 2005.

[RFC4328] Papadimitriou, D., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Extensions for G.709 Optical Transport Networks Control", RFC 4328, January 2006.

[RFC4606] Mannie, E. and D. Papadimitriou, "Generalized Multi-Protocol Label Switching (GMPLS) Extensions for Synchronous Optical Network (SONET) and Synchronous Digital Hierarchy (SDH) Control", RFC 4606, August 2006.

[RFC4872] Lang, J., Rekhter, Y., and D. Papadimitriou, "RSVP-TE Extensions in Support of End-to-End Generalized Multi-Protocol Label Switching (GMPLS) Recovery", RFC 4872, May 2007.

[RFC4873] Berger, L., Bryskin, I., Papadimitriou, D., and A. Farrel, "GMPLS Segment Recovery", RFC 4873, May 2007.

[RFC5088] Le Roux, JL., Vasseur, JP., Ikejiri, Y., and R. Zhang,

"OSPF Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5088, January 2008.

[RFC5089] Le Roux, JL., Vasseur, JP., Ikejiri, Y., and R. Zhang, "IS-IS Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5089, January 2008.

[RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.

[RFC5520] Bradford, R., Vasseur, JP., and A. Farrel, "Preserving Topology Confidentiality in Inter-Domain Path Computation Using a Path-Key-Based Mechanism", RFC 5520, April 2009.

[RFC5521] Oki, E., Takeda, T., and A. Farrel, "Extensions to the Path Computation Element Communication Protocol (PCEP) for Route Exclusions", RFC 5521, April 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.

[RFC6003] Papadimitriou, D., "Ethernet Traffic Parameters", RFC 6003, October 2010.

[RFC6205] Otani, T. and D. Li, "Generalized Labels for Lambda-Switch-Capable (LSC) Label Switching Routers", RFC 6205, March 2011.

9.2. Informative References

[I-D.ceccarelli-ccamp-gmpls-ospf-g709]
Ceccarelli, D., Caviglia, D., Zhang, F., Li, D., Belotti, S., Grandi, P., Rao, R., Pithewan, K., and J. Drake, "Traffic Engineering Extensions to OSPF for Generalized MPLS (GMPLS) Control of Evolving G.709 OTN Networks", draft-ceccarelli-ccamp-gmpls-ospf-g709-06 (work in progress), June 2011.

[I-D.ietf-pce-gmpls-aps-req]
Otani, T., Ogaki, K., Caviglia, D., and F. Zhang, "Requirements for GMPLS applications of PCE", draft-ietf-pce-gmpls-aps-req-04 (work in progress), June 2011.

[I-D.ietf-pce-inter-layer-ext]
Oki, E., Takeda, T., Roux, J., Farrel, A., and F. Zhang,

"Extensions to the Path Computation Element communication Protocol (PCEP) for Inter-Layer MPLS and GMPLS Traffic Engineering", draft-ietf-pce-inter-layer-ext-05 (work in progress), June 2011.

[I-D.ietf-pce-wson-routing-wavelength]
Lee, Y., Bernstein, G., Martensson, J., Takeda, T., and T. Tsuritani, "PCEP Requirements for WSON Routing and Wavelength Assignment", draft-ietf-pce-wson-routing-wavelength-04 (work in progress), March 2011.

[I-D.zhang-ccamp-gmpls-evolving-g709]
Zhang, F., Zhang, G., Belotti, S., Ceccarelli, D., and K. Pithewan, "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Extensions for the evolving G.709 Optical Transport Networks Control", draft-zhang-ccamp-gmpls-evolving-g709-08 (work in progress), July 2011.

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

[RFC4657] Ash, J. and J. Le Roux, "Path Computation Element (PCE) Communication Protocol Generic Requirements", RFC 4657, September 2006.

[RFC5467] Berger, L., Takacs, A., Caviglia, D., Fedyk, D., and J. Meuric, "GMPLS Asymmetric Bandwidth Bidirectional Label Switched Paths (LSPs)", RFC 5467, March 2009.

Authors' Addresses

Cyril Margaria (editor)
Nokia Siemens Networks
St Martin Strasse 76
Munich, 81541
Germany

Phone: +49 89 5159 16934
Email: cyril.margaria@nsn.com

Oscar Gonzalez de Dios (editor)
Telefonica Investigacion y Desarrollo
C/ Emilio Vargas 6
Madrid, 28043
Spain

Phone: +34 91 3374013
Email: ogondio@tid.es

Fatai Zhang (editor)
Huawei Technologies
F3-5-B R&D Center, Huawei Base
Bantian, Longgang District
Shenzhen, 518129
P.R.China

Email: zhangfatai@huawei.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 17, 2020

C. Margaria, Ed.
Juniper
O. Gonzalez de Dios, Ed.
Telefonica Investigacion y Desarrollo
F. Zhang, Ed.
Huawei Technologies
October 15, 2019

PCEP extensions for GMPLS
draft-ietf-pce-gmpls-pcep-extensions-15

Abstract

A Path Computation Element (PCE) provides path computation functions for Multiprotocol Label Switching (MPLS) and Generalized MPLS (GMPLS) networks. Additional requirements for GMPLS are identified in RFC7025.

This memo provides extensions to the Path Computation Element communication Protocol (PCEP) for the support of the GMPLS control plane to address those requirements.

Table of Contents

1. Introduction

Although [RFC4655] defines the PCE architecture and framework for both MPLS and GMPLS networks, most preexisting PCEP RFCs [RFC5440], [RFC5521], [RFC5541], [RFC5520] are focused on MPLS networks, and do not cover the wide range of GMPLS networks. This document complements these RFCs by addressing the extensions required for GMPLS applications and routing requests, for example for Optical Transport Network (OTN) and Wavelength Switched Optical Network (WSON) networks.

The functional requirements to be addressed by the PCEP extensions to support these applications are fully described in [RFC7025] and [RFC7449].

1.1. Terminology

This document uses terminologies from the PCE architecture document [RFC4655], the PCEP documents including [RFC5440], [RFC5521], [RFC5541], [RFC5520], [RFC7025] and [RFC7449], and the GMPLS documents such as [RFC3471], [RFC3473] and so on. Note that it is expected the reader is familiar with these documents. The following abbreviations are used in this document

ODU ODU Optical Channel Data Unit [G.709-v3]

OTN Optical Transport Network [G.709-v3]

L2SC Layer-2 Switch Capable [RFC3471]

TDM Time-Division Multiplex Capable [RFC3471]

LSC Lambda Switch Capable [RFC3471]

SONET Synchronous Optical Networking

SDH Synchronous Digital Hierarchy

PCC Path Computation Client

RSVP-TE Resource Reservation Protocol - Traffic Engineering

LSP Label Switched Path

TE-LSP Traffic Engineering LSP

IRO Include Route Object

ERO Explicit Route Object

XRO eXclude Route Object

RRO Record Route Object

LSPA LSP Attribute

SRLG Shared Risk Link Group

NVC Number of Virtual Components [RFC4328][RFC4606]

NCC Number of Contiguous Components [RFC4328][RFC4606]

MT Multiplier [RFC4328][RFC4606]

RCC Requested Contiguous Concatenation [RFC4606]

PCReq Path Computation Request [RFC5440]

PCRep Path Computation Reply [RFC5440]

MEF Metro Ethernet Forum

SSON Spectrum-Switched Optical Network

P2MP Point to Multi-Point

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

1.2. PCEP Requirements for GMPLS

The document [RFC7025] describes the set of PCEP requirements to support GMPLS TE-LSPs. This document assumes a significant familiarity with [RFC7025] and existing PCEP extensions. As a short overview, those requirements can be broken down into the following categories.

- Which data flow is switched by the LSP: a combination of Switching type (for instance L2SC or TDM), LSP Encoding type (e.g., Ethernet, SONET/SDH) and sometimes the Signal Type (e.g., in case of TDM/LSC switching capability).

- Data flow specific traffic parameters, which are technology specific. For instance, in SDH/SONET and [G.709-v3] OTN networks the Concatenation Type and the Concatenation Number have an influence on the switched data and on which link it can be supported

- Support for asymmetric bandwidth requests.

- Support for unnumbered interface identifiers, as defined in [RFC3477]

- Label information and technology specific label(s) such as wavelength labels as defined in [RFC6205]. A PCC should also be able to specify a label restriction similar to the one supported by RSVP-TE in [RFC3473].

- Ability to indicate the requested granularity for the path ERO: node, link or label. This is to allow the use of the explicit label control feature of RSVP-TE.

The requirements of [RFC7025] apply to several objects conveyed by PCEP, this is described in Section 1.3. Some of the requirements of [RFC7025] are already supported in existing documents, as described in Section 1.4.

This document describes a set of PCEP extensions, including new object types, TLVs, encodings, error codes and procedures, in order to fulfill the aforementioned requirements not covered in existing RFCs.

1.3. Requirements Applicability

This section follows the organization of [RFC7025] Section 3 and indicates, for each requirement, the affected piece of information carried by PCEP and its scope.

1.3.1. Requirements on Path Computation Request

(1) Switching capability/type: as described in [RFC3471] this piece of information is used with the Encoding Type and Signal Type to fully describe the switching technology and data carried by the TE-LSP. This is applicable to the TE-LSP itself and also to the TE-LSP endpoint (Carried in the END-POINTS object for MPLS networks in [RFC5440]) when considering multiple network layers. Inter-layer path computation requirements are addressed in in [RFC8282] which addressing the TE-LSP itself, but the TE-LSP endpoints are not addressed.

(2) Encoding type: see (1).

(3) Signal type: see (1).

(4) Concatenation type: this parameter and the Concatenation Number (5) are specific to some TDM (SDH and ODU) switching technology. They MUST be described together and are used to derive the requested resource allocation for the TE-LSP. It is scoped to the TE-LSP and is related to the [RFC5440] BANDWIDTH object in MPLS networks. See [RFC4606] and [RFC4328] about concatenation information.

(5) Concatenation number: see (4).

(6) Technology-specific label(s): as described in [RFC3471] the GMPLS Labels are specific to each switching technology. They can be specified on each link and also on the TE-LSP endpoints , in WSON networks for instance, as described in [RFC6163]. The label restriction can apply to endpoints and on each hop, the related PCEP objects are END-POINTS, IRO, XRO and RRO.

(7) End-to-End (E2E) path protection type: as defined in [RFC4872], this is applicable to the TE-LSP. In MPLS networks the related PCEP object is LSPA (carrying local protection information).

(8) Administrative group: as defined in [RFC3630], this information is already carried in the LSPA object.

(9) Link protection type: as defined in [RFC4872], this is applicable to the TE-LSP and is carried in association with the E2E path protection type.

(10) Support for unnumbered interfaces: as defined in [RFC3477]. Its scope and related objects are the same as labels

(11) Support for asymmetric bandwidth requests: as defined [RFC6387], the scope is similar to (4)

(12) Support for explicit label control during the path computation. This affects the TE-LSP and amount of information returned in the ERO.

(13) Support of label restrictions in the requests/responses: This is described in (6).

1.3.2. Requirements on Path Computation Response

(1) Path computation with concatenation: This is related to Path Computation request requirement (4). In addition there is a specific type of concatenation called virtual concatenation that allows different routes to be used between the endpoints. It is similar to the semantic and scope of the LOAD-BALANCING in MPLS networks.

(2) Label constraint: The PCE should be able to include Labels in the path returned to the PCC, the related object is the ERO object.

(3) Roles of the routes: as defined in [RFC4872], this is applicable to the TE-LSP and is carried in association with the E2E path protection type.

1.4. Existing Support for GMPLS in Base PCEP Objects and its Limitations

The support provided by specifications in [RFC8282] and [RFC5440] for the requirements listed in [RFC7025] is summarized in Table 1 and Table 2. In some cases the support may not be complete, as noted, and additional support need to be provided in this specification.

Req.	Name	Support
1	Switching capability/type	SWITCH-LAYER (RFC8282)
2	Encoding type	SWITCH-LAYER (RFC8282)
3	Signal type	SWITCH-LAYER (RFC8282)
4	Concatenation type	No
5	Concatenation number	No
6	Technology-specific label	(Partial) ERO (RFC5440)
7	End-to-End (E2E) path protection type	No
8	Administrative group	LSPA (RFC5440)
9	Link protection type	No
10	Support for unnumbered interfaces	(Partial) ERO (RFC5440)
11	Support for asymmetric bandwidth requests	No
12	Support for explicit label control during the path computation	No
13	Support of label restrictions in the requests/responses	No

Table 1: RFC7025 Section 3.1 requirements support

Req.	Name	Support
1	Path computation with concatenation	No
2	Label constraint	No
3	Roles of the routes	No

Table 2: RFC7025 Section 3.2 requirements support

As described in Section 1.3 PCEP as of [RFC5440], [RFC5521] and [RFC8282], supports the following objects, included in requests and responses, related to the described requirements.

From [RFC5440]:

- END-POINTS: related to requirements (1, 2, 3, 6, 10 and 13). The object only supports numbered endpoints. The context specifies whether they are node identifiers or numbered interfaces.

- BANDWIDTH: related to requirements (4, 5 and 11). The data rate is encoded in the bandwidth object (as IEEE 32 bit float). [RFC5440] does not include the ability to convey an encoding proper to all GMPLS-controlled networks.

- ERO: related to requirements (6, 10, 12 and 13). The ERO content is defined in RSVP in [RFC3209][RFC3473][RFC3477][RFC7570] and supports all the requirements already.

- LSPA: related to requirements (7, 8 and 9). The requirement 8 (setup and holding priorities) is already supported.

From [RFC5521]:

- XRO:

  * This object allows excluding (strict or not) resources and is related to requirements (6, 10 and 13). It also includes the requested diversity (node, link or SRLG).

  * When the F bit is set, the request indicates that the existing path has failed and the resources present in the RRO can be reused.

From [RFC8282]:

- SWITCH-LAYER: addresses requirements (1, 2 and 3) for the TE-LSP and indicates which layer(s) should be considered. The object can be used to represent the RSVP-TE generalized label request. It does not address the endpoints case of requirements (1, 2 and 3).

- REQ-ADAP-CAP: indicates the adaptation capabilities requested, can also be used for the endpoints in case of mono-layer computation

The gaps in functional coverage of the base PCEP objects are:

The BANDWIDTH and LOAD-BALANCING objects do not describe the details of the traffic request (requirements 4 and 5, for example NVC, multiplier) in the context of GMPLS networks, for instance TDM or OTN networks.

The END-POINTS object does not allow specifying an unnumbered interface, nor potential label restrictions on the interface (requirements 6, 10 and 13). Those parameters are of interest in case of switching constraints.

The Include/eXclude Route Objects (IRO/XRO) do not allow the inclusion/exclusion of labels (requirements 6, 10 and 13).

Base attributes do not allow expressing the requested link protection level and/or the end-to-end protection attributes.

The PCEP extensions defined later in this document to cover the gaps are:

Two new object types are defined for the BANDWIDTH object (Generalized bandwidth, Generalized bandwidth of existing TE-LSP for which a reoptimization is requested).

A new object type is defined for the LOAD-BALANCING object (Generalized Load Balancing).

A new object type is defined for the END-POINTS object (Generalized Endpoint).

A new TLV is added to the Open message for capability negotiation.

A new TLV is added to the LSPA object.

The Label TLV is now allowed in the IRO and XRO objects.

In order to indicate the used routing granularity in the response, a new flag in the RP object is added.

2. PCEP Objects and Extensions

This section describes the necessary PCEP objects and extensions. The PCReq and PCRep messages are defined in [RFC5440]. This document does not change the existing grammars.

2.1. GMPLS Capability Advertisement

2.1.1. GMPLS Computation TLV in the Existing PCE Discovery Protocol

IGP-based PCE Discovery (PCED) is defined in [RFC5088] and [RFC5089] for the OSPF and IS-IS protocols. Those documents have defined bit 0 in PCE-CAP-FLAGS Sub-TLV of the PCED TLV as "Path computation with GMPLS link constraints". This capability is optional and can be used to detect GMPLS-capable PCEs. PCEs that set the bit to indicate support of GMPLS path computation MUST follow the procedures in Section 2.1.2 to further qualify the level of support during PCEP session establishment.

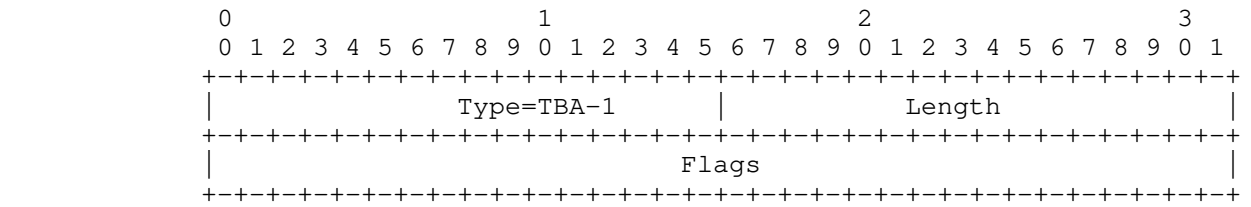
2.1.2. OPEN Object Extension GMPLS-CAPABILITY TLV

In addition to the IGP advertisement, a PCEP speaker MUST be able to discover the other peer GMPLS capabilities during the Open message exchange. This capability is also useful to avoid misconfigurations. This document defines a GMPLS-CAPABILITY TLV for use in the OPEN object to negotiate the GMPLS capability. The inclusion of this TLV

in the Open message indicates that the PCEP speaker support the PCEP extensions defined in the document. A PCEP speaker that is able to support the GMPLS extensions defined in this document MUST include the GMPLS-CAPABILITY TLV on the Open message. If one of the PCEP peers does not include the GMPLS-CAPABILITY TLV in the Open message, the peers MUST NOT make use of the objects and TLVs defined in this document.

If the PCEP speaker supports the extensions of this specification but did not advertise the GMPLS-CAPABILITY capability, upon receipt of a message from the PCE including an extension defined in this document, it MUST generate a PCEP Error (PCErr) with Error-Type=10 (Reception of an invalid object) and Error-value=TBA-42 (Missing GMPLS-CAPABILITY TLV), and it SHOULD terminate the PCEP session.

IANA has allocated value TBA-1 from the "PCEP TLV Type Indicators" sub-registry, as documented in Section 5.3 ("New PCEP TLVs"). The description is "GMPLS-CAPABILITY". Its format is shown in the following figure.



No Flags are defined in this document, they are reserved for future use.

2.2. RP Object Extension

Explicit label control (ELC) is a procedure supported by RSVP-TE, where the outgoing labels are encoded in the ERO. As a consequence, the PCE can provide such labels directly in the path ERO. Depending on policies or switching layer, it can be necessary for the PCC to use explicit label control or explicit link ids, thus it needs to indicate in the PCReq which granularity it is expecting in the ERO. This corresponds to requirement 12 of [RFC7025]. The possible granularities can be node, link or label. The granularities are inter-dependent, in the sense that link granularity implies the presence of node information in the ERO; similarly, a label granularity implies that the ERO contains node, link and label information.

A new 2-bit routing granularity (RG) flag (Bits TBA-13) is defined in the RP object. The values are defined as follows

0: reserved
1: node
2: link
3: label

Table 3: RG flag

The flag in the RP object indicates the requested route granularity. The PCE SHOULD follow this granularity and MAY return a NO-PATH if the requested granularity cannot be provided. The PCE MAY return any granularity on the route based on its policy. The PCC can decide if the ERO is acceptable based on its content.

If a PCE honored the requested routing granularity for a request, it MUST indicate the selected routing granularity in the RP object included in the response. Otherwise, the PCE MUST use the reserved RG to leave the check of the ERO to the PCC. The RG flag is backward-compatible with [RFC5440]: the value sent by an implementation (PCC or PCE) not supporting it will indicate a reserved value.

2.3. BANDWIDTH Object Extensions

From [RFC5440] the object carrying the requested size for the TE-LSP is the BANDWIDTH object. The object types 1 and 2 defined in [RFC5440] do not describe enough information to describe the TE-LSP bandwidth in GMPLS networks. The BANDWIDTH object encoding has to be extended to allow the object to express the bandwidth as described in [RFC7025]. RSVP-TE extensions for GMPLS provide a set of encodings allowing such representation in an unambiguous way, this is encoded in the RSVP-TE TSpec and FlowSpec objects. This document extends the BANDWIDTH object with new object types reusing the RSVP-TE encoding.

The following possibilities are supported by the extended encoding:

- Asymmetric bandwidth (different bandwidth in forward and reverse direction), as described in [RFC6387]

- GMPLS (SDH/SONET, G.709, ATM, MEF, etc.) parameters.

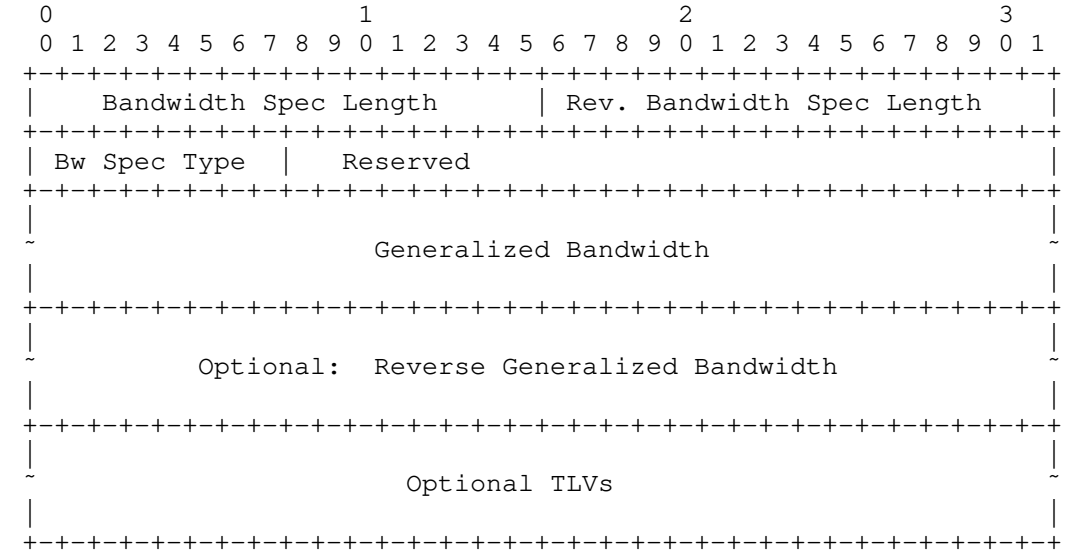
This corresponds to requirements 3, 4, 5 and 11 of [RFC7025] Section 3.1.

This document defines two Object Types for the BANDWIDTH object:

TBA-2 Generalized bandwidth

TBA-3 Generalized bandwidth of an existing TE-LSP for which a reoptimization is requested

The definitions below apply for Object Type TBA-2 and TBA-3. The body is as follows:



The BANDWIDTH object type TBA-2 and TBA-3 have a variable length. The 16-bit Bandwidth Spec Length field indicates the length of the Generalized Bandwidth field. The Bandwidth Spec Length MUST be strictly greater than 0. The 16-bit Reverse Bandwidth Spec Length field indicates the length of the Reverse Generalized Bandwidth field. The Reverse Bandwidth Spec Length MAY be equal to 0.

The Bw Spec Type field determines which type of bandwidth is represented by the object.

The Bw Spec Type corresponds to the RSVP-TE SENDER_TSPEC (Object Class 12) C-Types

The encoding of the fields Generalized Bandwidth and Reverse Generalized Bandwidth is the same as the Traffic Parameters carried in RSVP-TE, it can be found in the following references. It is to be noted that the RSVP-TE traffic specification MAY also include TLVs (e.g., [RFC6003] different from the PCEP TLVs).

Bw Spec Type	Name	Reference
2	Intserv	[RFC2210]
4	SONET/SDH	[RFC4606]
5	G.709	[RFC4328]
6	Ethernet	[RFC6003]
7	OTN-TDM	[RFC7139]
8	SSON	[RFC7792]

Table 4: Generalized Bandwidth and Reverse Generalized Bandwidth field encoding

When a PCC requests a bi-directional path with symmetric bandwidth, it SHOULD only specify the Generalized Bandwidth field, and set the Reverse Bandwidth Spec Length to 0. When a PCC needs to request a bi-directional path with asymmetric bandwidth, it SHOULD specify the different bandwidth in the forward and reverse directions with a Generalized Bandwidth and Reverse Generalized Bandwidth fields.

The procedure described in [RFC5440] for the PCRep is unchanged: a PCE MAY include the BANDWIDTH objects in the response to indicate the BANDWIDTH of the path.

As specified in [RFC5440] in the case of the reoptimization of a TE-LSP, the bandwidth of the existing TE-LSP MUST also be included in addition to the requested bandwidth if and only if the two values differ. The Object Type TBA-3 MAY be used instead of the previously specified object type 2 to indicate the existing TE-LSP bandwidth originally specified with object type TBA-2. A PCC that requested a path with a BANDWIDTH object of object type 1 MUST use object type 2 to represent the existing TE-LSP BANDWIDTH.

OPTIONAL TLVs MAY be included within the object body to specify more specific bandwidth requirements. No TLVs for the Object Type TBA-2 and TBA-3 are defined by this document.

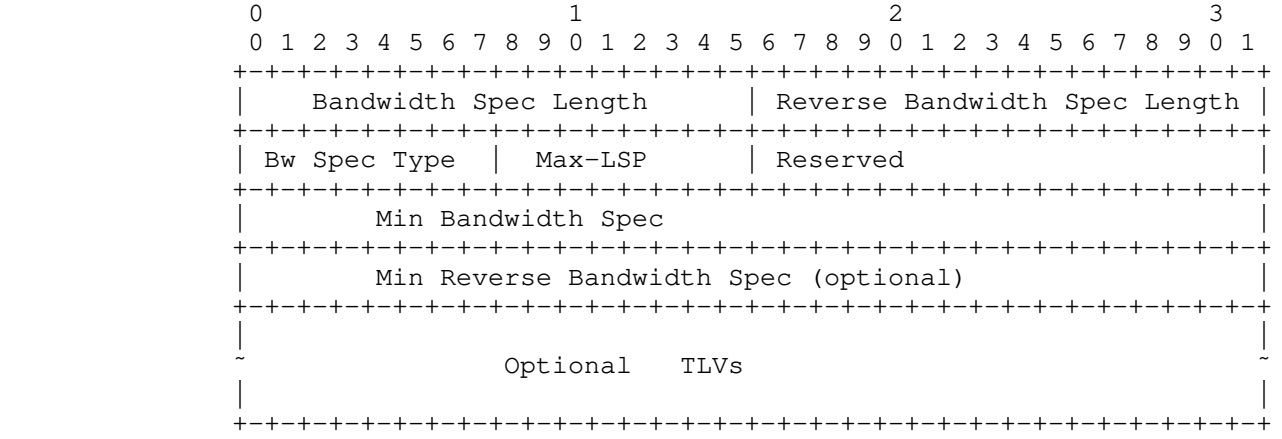
2.4. LOAD-BALANCING Object Extensions

The LOAD-BALANCING object [RFC5440] is used to request a set of at most Max-LSP TE-LSP having in total the bandwidth specified in BANDWIDTH, with each TE-LSP having at least a specified minimum bandwidth. The LOAD-BALANCING follows the bandwidth encoding of the BANDWIDTH object, and thus the existing definition from [RFC5440] does not describe enough details for the bandwidth specification expected by GMPLS.

Similarly to the BANDWIDTH object, a new object type is defined to allow a PCC to represent the bandwidth types supported by GMPLS networks.

This document defines the Generalized Load Balancing object type TBA-4 for the LOAD-BALANCING object. The Generalized Load Balancing object type has a variable length.

The format of the Generalized Load Balancing object type is as follows:



Bandwidth Spec Length (16 bits): the total length of the Min Bandwidth Spec field. The length MUST be strictly greater than 0.

Reverse Bandwidth Spec Length (16 bits): the total length of the Min Reverse Bandwidth Spec field. It MAY be equal to 0.

Bw Spec Type (8 bits): the bandwidth specification type, it corresponds to the RSVP-TE SENDER_TSPEC (Object Class 12) C-Types.

Max-LSP (8 bits): maximum number of TE-LSPs in the set.

Min Bandwidth Spec (variable): specifies the minimum bandwidth specification of each element of the TE-LSP set.

Min Reverse Bandwidth Spec (variable): specifies the minimum reverse bandwidth specification of each element of the TE-LSP set.

The encoding of the fields Min Bandwidth Spec and Min Reverse Bandwidth Spec is the same as in RSVP-TE SENDER_TSPEC object, it can be found in Table 4 from Section 2.3 from this document.

When a PCC requests a bi-directional path with symmetric bandwidth while specifying load balancing constraints it SHOULD specify the Min Bandwidth Spec field, and set the Reverse Bandwidth Spec Length to 0. When a PCC needs to request a bi-directional path with asymmetric bandwidth while specifying load balancing constraints, it MUST specify the different bandwidth in forward and reverse directions through a Min Bandwidth Spec and Min Reverse Bandwidth Spec fields.

OPTIONAL TLVs MAY be included within the object body to specify more specific bandwidth requirements. No TLVs for the Generalized Load Balancing object type are defined by this document.

The semantic of the LOAD-BALANCING object is not changed. If a PCC requests the computation of a set of TE-LSPs with at most N TE-LSPs so that it can carry generalized bandwidth X , each TE-LSP must at least transport bandwidth B, it inserts a BANDWIDTH object specifying X as the required bandwidth and a LOAD-BALANCING object with the Max-LSP and Min Bandwidth Spec fields set to N and B, respectively. When the BANDWIDTH and Min Bandwidth Spec can be summarized as scalars, the sum of all TE-LSPs bandwith in the set is greater than X. The mapping of X over N path with (at least) bandwidth B is technology and possibly node specific. Each standard definition of the transport technology is defining those mappings and are not repeated in this document. A simplified example for SDH is described in Appendix A

In all other cases, including for technologies based on statistical multiplexing (e.g., InterServ, Ethernet), the exact bandwidth management (e.g., Ethernet's Excessive Rate) is left to the PCE's policies, according to the operator's configuration. If required, further documents may introduce a new mechanism to finely express complex load balancing policies within PCEP.

The BANDWITH and LOAD-BALANCING Bw Spec Type can be different depending on the endpoint nodes architecture. When the PCE is not able to handle those two Bw Spec Type, it MUST return a NO-PATH with the bit "LOAD-BALANCING could not be performed with the bandwidth constraits " set in the NO-PATH-VECTOR TLV.

2.5. END-POINTS Object Extensions

The END-POINTS object is used in a PCEP request message to specify the source and the destination of the path for which a path computation is requested. From [RFC5440], the source IP address and the destination IP address are used to identify those. A new Object Type is defined to address the following possibilities:

- Different source and destination endpoint types.

- o Label restrictions on the endpoint.
- o Specification of unnumbered endpoints type as seen in GMPLS networks.

The Object encoding is described in the following sections.

In path computation within a GMPLS context the endpoints can:

- o Be unnumbered as described in [RFC3477].
- o Have labels associated to them, specifying a set of constraints on the allocation of labels.
- o Have different switching capabilities

The IPv4 and IPv6 endpoints are used to represent the source and destination IP addresses. The scope of the IP address (Node or numbered Link) is not explicitly stated. It is also possible to request a Path between a numbered link and an unnumbered link, or a P2MP path between different type of endpoints.

This document defines the Generalized Endpoint object type TBA-5 for the END-POINTS object. This new type also supports the specification of constraints on the endpoint label to be used. The PCE might know the interface restrictions but this is not a requirement. This corresponds to requirements 6 and 10 of [RFC7025].

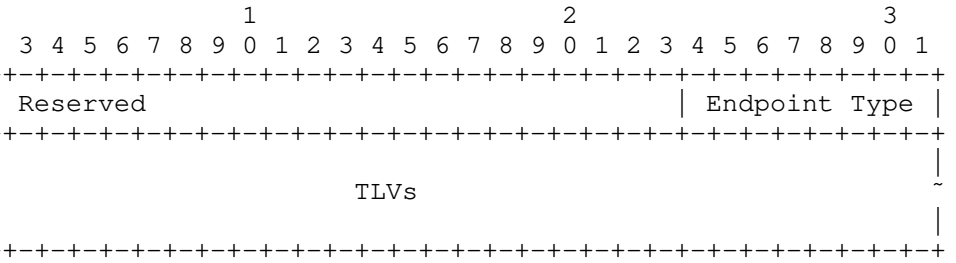
2.5.1. Generalized Endpoint Object Type

The Generalized Endpoint object type format consists of a body and a list of TLVs scoped to this object. The TLVs give the details of the endpoints and are described in Section 2.5.2. For each Endpoint Type, a different grammar is defined. The TLVs defined to describe an endpoint are:

1. IPv4 address endpoint.
2. IPv6 address endpoint.
3. Unnumbered endpoint.
4. Label request.
5. Label set.

The Label set TLV is used to restrict or suggest the label allocation in the PCE. This TLV expresses the set of restrictions which may

apply to signaling. Label restriction support can be an explicit or a suggested value (Label set describing one label, with the L bit respectively cleared or set), mandatory range restrictions (Label set with L bit cleared) and optional range restriction (Label set with L bit set). Endpoints label restriction may not be part of the RRO or IRO. They can be included when following [RFC4003] in signaling for egress endpoint, but ingress endpoint properties can be local to the PCC and not signaled. To support this case the label set allows indication which label are used in case of reoptimization. The label range restrictions are valid in GMPLS-controlled networks, either by PCC policy or depending on the switching technology used, for instance on given Ethernet or ODU equipment having limited hardware capabilities restricting the label range. Label set restriction also applies to WSON networks where the optical senders and receivers are limited in their frequency tunability ranges, consequently restricting the possible label ranges on the interface in GMPLS. The END-POINTS Object with Generalized Endpoint object type is encoded as follow:



Reserved bits SHOULD be set to 0 when a message is sent and ignored when the message is received.

The Endpoint Type is defined as follow:

Value	Type	Meaning
0	Point-to-Point	
1	Point-to-Multipoint	New leaves to add
2		Old leaves to remove
3		Old leaves whose path can be modified/reoptimized
4		Old leaves whose path has to be left unchanged
5-244	Reserved	
245-255	Experimental range	

Table 5: Generalized Endpoint endpoint types

The Endpoint Type is used to cover both point-to-point and different point-to-multipoint endpoints. A PCE may accept only Endpoint Type 0: Endpoint Types 1-4 apply if the PCE implementation supports P2MP path calculation. A PCE not supporting a given Endpoint Type SHOULD respond with a PCErr with Error-Type=4 (Not supported object), Error-value=TBA-15 (Unsupported endpoint type in END-POINTS Generalized Endpoint object type). As per [RFC5440], a PCE unable to process Generalized Endpoints may respond with Error-Type=3 (Unknown Object), Error-value=2 (Unrecognized object Type) or Error-Type=4 (Not supported object), Error-value=2 (Not supported object Type). The TLVs present in the request object body MUST follow the following [RFC5511] grammar:

```
<generalized-endpoint-tlvs>::=
  <p2p-endpoints> | <p2mp-endpoints>

<p2p-endpoints> ::=
  <endpoint> [<endpoint-restriction-list>]
  <endpoint> [<endpoint-restriction-list>]

<p2mp-endpoints> ::=
  <endpoint> [<endpoint-restriction-list>]
  <endpoint> [<endpoint-restriction-list>]
  [<endpoint> [<endpoint-restriction-list>]]...
```

For endpoint type Point-to-Point, 2 endpoint TLVs MUST be present in the message. The first endpoint is the source and the second is the destination.

For endpoint type Point-to-Multipoint, several END-POINT objects MAY be present in the message and the exact meaning depending on the endpoint type defined for the object. The first endpoint TLV is the root and other endpoints TLVs are the leaves. The root endpoint MUST be the same for all END-POINTS objects. If the root endpoint is not the same for all END-POINTS, a PCErr with Error-Type=17 (P2MP END-POINTS Error), Error-value=4 (The PCE cannot satisfy the request due to inconsistent END-POINTS) MUST be returned. The procedure defined in [RFC8306] Section 3.10 also apply to the Generalized Endpoint with Point-to-Multipoint endpoint types.

An endpoint is defined as follows:

```
<endpoint>::=<IPV4-ADDRESS>|<IPV6-ADDRESS>|<UNNUMBERED-ENDPOINT>
<endpoint-restriction-list> ::=          <endpoint-restriction>
            [<endpoint-restriction-list>]

<endpoint-restriction> ::=
                [<LABEL-REQUEST>][<label-restriction-list>]

<label-restriction-list> ::= <label-restriction>
                             [<label-restriction-list>]
<label-restriction> ::= <LABEL-SET>
```

The different TLVs are described in the following sections. A PCE MAY support any or all of IPV4-ADDRESS, IPV6-ADDRESS, and UNNUMBERED-ENDPOINT TLVs. When receiving a PCReq, a PCE unable to resolve the identifier in one of those TLVs MUST respond using a PCRep with NO-PATH and set the bit "Unknown destination" or "Unknown source" in the NO-PATH-VECTOR TLV. The response SHOULD include the END-POINTS object with only the unsupported TLV(s).

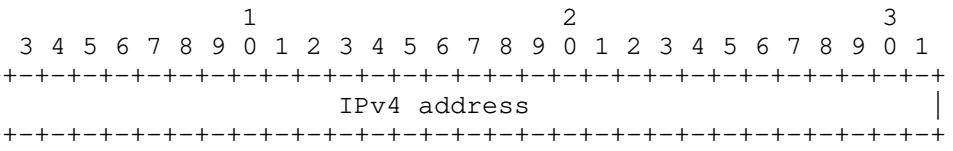
A PCE MAY support either or both of the LABEL-REQUEST and LABEL-SET TLVs. If a PCE finds a non-supported TLV in the END-POINTS the PCE MUST respond with a PCErr message with Error-Type=4 (Not supported object) and Error-value=TBA-15 (Unsupported TLV present in END-POINTS Generalized Endpoint object type) and the message SHOULD include the END-POINTS object in the response with only the endpoint and endpoint restriction TLV it did not understand. A PCE supporting those TLVs but not being able to fulfil the label restriction MUST send a response with a NO-PATH object which has the bit "No endpoint label resource" or "No endpoint label resource in range" set in the NO-PATH-VECTOR TLV. The response SHOULD include an END-POINTS object containing only the TLV(s) related to the constraints the PCE could not meet.

2.5.2. END-POINTS TLV Extensions

All endpoint TLVs have the standard PCEP TLV header as defined in [RFC5440] Section 7.1. For the Generalized Endpoint Object Type the TLVs MUST follow the ordering defined in Section 2.5.1.

2.5.2.1. IPV4-ADDRESS TLV

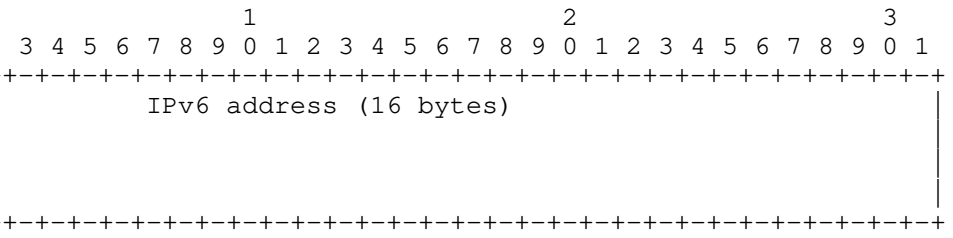
This TLV represents a numbered endpoint using IPv4 numbering, the format of the IPv4-ADDRESS TLV value (TLV-Type=TBA-6) is as follows:



This TLV MAY be ignored, in which case a PCRep with NO-PATH SHOULD be returned, as described in Section 2.5.1.

2.5.2.2. IPV6-ADDRESS TLV

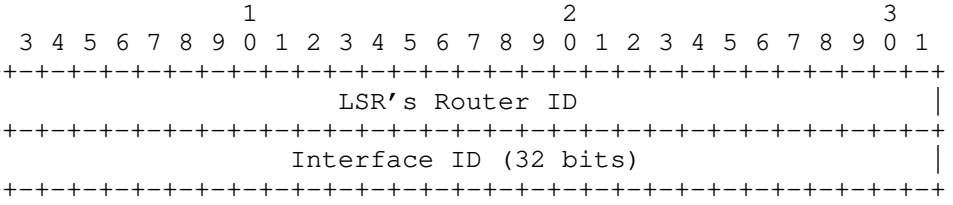
This TLV represents a numbered endpoint using IPV6 numbering, the format of the IPv6-ADDRESS TLV value (TLV-Type=TBA-7) is as follows:



This TLV MAY be ignored, in which case a PCRep with NO-PATH SHOULD be returned, as described in Section 2.5.1.

2.5.2.3. UNNUMBERED-ENDPOINT TLV

This TLV represents an unnumbered interface. This TLV has the same semantic as in [RFC3477]. The TLV value is encoded as follows (TLV-Type=TBA-8)



This TLV MAY be ignored, in which case a PCRep with NO-PATH SHOULD be returned, as described in Section 2.5.1.

2.5.2.4. LABEL-REQUEST TLV

The LABEL-REQUEST TLV indicates the switching capability and encoding type of the following label restriction list for the endpoint. The value format and encoding is the same as described in [RFC3471]

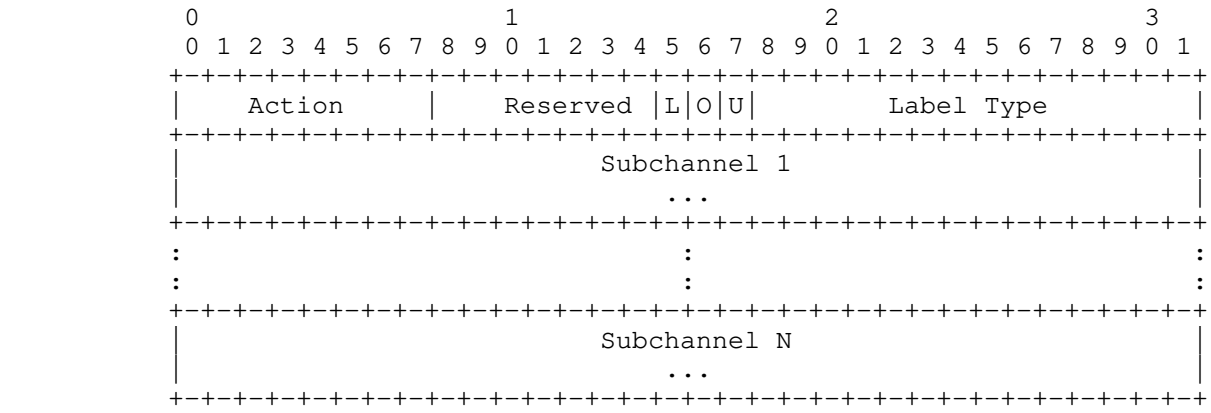
Section 3.1 Generalized label request. The LABEL-REQUEST TLV uses TLV-Type=TBA-9. The Encoding Type indicates the encoding type, e.g., SONET/SDH/GigE etc., of the LSP with which the data is associated. The Switching type indicates the type of switching that is being requested on the endpoint. G-PID identifies the payload. This TLV and the following one are defined to satisfy requirement 13 of [RFC7025] for the endpoint. It is not directly related to the TE-LSP label request, which is expressed by the SWITCH-LAYER object.

On the path calculation request only the GENERALIZED-BANDWIDTH and SWITCH-LAYER need to be coherent, the endpoint labels could be different (supporting a different LABEL-REQUEST). Hence the label restrictions include a Generalized label request in order to interpret the labels. This TLV MAY be ignored, in which case a PCRep with NO-PATH SHOULD be returned, as described in Section 2.5.1.

2.5.2.5. LABEL-SET TLV

Label or label range restrictions can be specified for the TE-LSP endpoints. Those are encoded using the LABEL-SET TLV. The label value need to be interpreted with a description on the Encoding and switching type. The REQ-ADAP-CAP object from [RFC8282] can be used in case of mono-layer request, however in case of multilayer it is possible to have more than one object, so it is better to have a dedicated TLV for the label and label request. These TLVs MAY be ignored, in which case a response with NO-PATH SHOULD be returned, as described in Section 2.5.1. TLVs are encoded as follows (following [RFC5440]):

- LABEL-SET TLV, Type=TBA-10. The TLV Length is variable, Encoding follows [RFC3471] Section 3.5 "Label set" with the addition of a U bit, O bit and L bit. The L bit is used to represent a suggested set of labels, following the semantic of SUGGESTED_LABEL defined by [RFC3471].



A LABEL-SET TLV represents a set of possible labels that can be used on an interface. If the L bit is cleared, the label allocated on the first endpoint MUST be within the label set range. The action parameter in the Label set indicates the type of list provided. These parameters are described by [RFC3471] Section 3.5.1.

The U, O and L bits have the following meaning:

U: Upstream direction: The U bit is set for upstream (revers) direction in case of bidirectional LSP.
O: Old Label: set when the TLV represent the old (previously allocated) label in case of re-optimization. The R bit of the RP object MUST be set to 1. If the L bit is set, this bit SHOULD be set to 0 and ignored on receipt. When this bit is set, the Action field MUST be set to 0 (Inclusive List) and the Label Set MUST contain one subchannel.
L: Loose Label: set when the TLV indicates to the PCE a set of preferred (ordered) labels to be used. The PCE MAY use those labels for label allocation.

Labels TLV bits

Several LABEL_SET TLVs MAY be present with the O bit cleared, LABEL_SET TLVs with L bit set can be combined with a LABEL_SET TLV with L bit cleared. At most 2 LABEL_SET TLVs MAY be present with the O bit set, with at most one of these having the U bit set and at most one of these having the U bit cleared. For a given U bit value, if more than one LABEL_SET TLV with the O bit set is present, the first TLV MUST be processed and the following TLVs with the same U and O bit MUST be ignored.

A LABEL-SET TLV with the O and L bit set MUST trigger a PCErr message with Error-Type=10 (Reception of an invalid object) Error-value=TBA-25 (Wrong LABEL-SET TLV present with O and L bit set).

A LABEL-SET TLV with the O bit set and an Action Field not set to 0 (Inclusive list) or containing more than one subchannel MUST trigger a PCErr message with Error-Type=10 (Reception of an invalid object) Error-value=TBA-26 (Wrong LABEL-SET TLV present with O bit and wrong format).

If a LABEL-SET TLV is present with O bit set, the R bit of the RP object MUST be set, otherwise a PCErr message MUST be sent with Error-Type=10 (Reception of an invalid object) Error-value=TBA-24 (LABEL-SET TLV present with O bit set but without R bit set in RP).

2.6. IRO Extension

The IRO as defined in [RFC5440] is used to include specific objects in the path. RSVP-TE allows the inclusion of a label definition. In order to fulfill requirement 13 of [RFC7025] the IRO needs to support the new subobject type as defined in [RFC3473]:

Type	Sub-object
TBA-38	LABEL

The Label subobject MUST follow a subobject identifying a link, currently an IP address subobject (Type 1 or 2) or an interface ID (type 4) subobject. If an IP address subobject is used, then the given IP address MUST be associated with a link. More than one label subobject MAY follow each link subobject. The procedure associated with this subobject is as follows.

If the PCE is able to allocate labels (e.g., via explicit label control) the PCE MUST allocate one label from within the set of label values for the given link. If the PCE does not assign labels, then it sends a response with a NO-PATH object, containing a NO-PATH-VECTOR TLV with the bit 'No label resource in range' set.

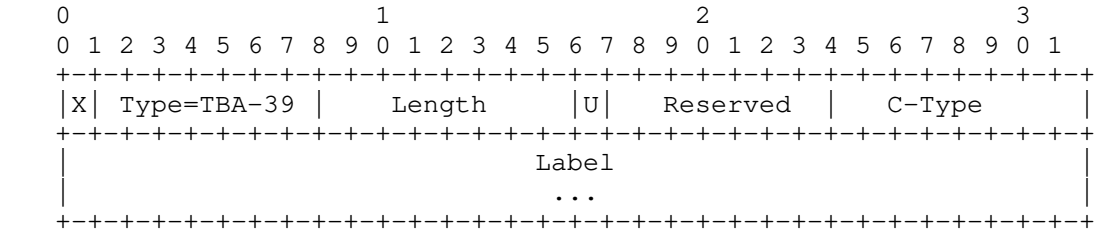
2.7. XRO Extension

The XRO as defined in [RFC5521] is used to exclude specific objects in the path. RSVP-TE allows the exclusion of certain labels ([RFC6001]). In order to fulfill requirement 13 of [RFC7025] Section 3.1, the PCEP's XRO needs to support a new subobject to enable label exclusion.

The encoding of the XRO Label subobject follows the encoding of the Label ERO subobject defined in [RFC3473] and XRO subobject defined in

[RFC5521]. The XRO Label subobject represent one Label and is defined as follows:

XRO Subobject Type TBA-39: Label Subobject.



X (1 bit): as per [RFC5521]. The X-bit indicates whether the exclusion is mandatory or desired. 0 indicates that the resource specified MUST be excluded from the path computed by the PCE. 1 indicates that the resource specified SHOULD be excluded from the path computed by the PCE, but MAY be included subject to PCE policy and the absence of a viable path that meets the other constraints and excludes the resource.

Type (7 bits): The Type of the XRO Label subobject is TBA-39.

Length (8 bits): see [RFC5521], the total length of the subobject in bytes (including the Type and Length fields). The Length is always divisible by 4.

U (1 bit): see [RFC3471] Section 6.1.

C-Type (8 bits): the C-Type of the included Label Object as defined in [RFC3473].

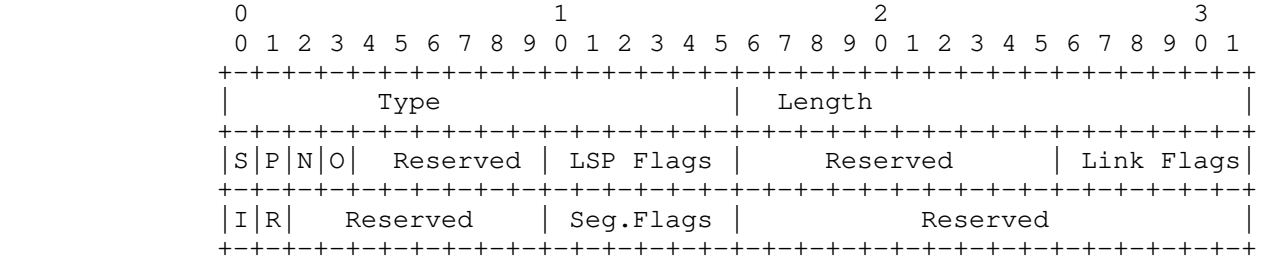
Label: see [RFC3471].

The Label subobject MUST follow a subobject identifying a link, currently an IP address subobject (Type 1 or 2) or an interface ID (type 4) subobject. If an IP address subobject is used, then the given IP address MUST be associated with a link. More than one label subobject MAY follow each link subobject.

Type	Sub-object
3	LABEL

2.8. LSPA Extensions

The LSPA carries the LSP attributes. In the end-to-end recovery context, this also includes the protection state information. A new TLV is defined to fulfil requirement 7 of [RFC7025] Section 3.1 and requirement 3 of [RFC7025] Section 3.2. This TLV contains the information of the PROTECTION object defined by [RFC4872] and can be used as a policy input. The LSPA object MAY carry a PROTECTION-ATTRIBUTE TLV defined as: Type TBA-12: PROTECTION-ATTRIBUTE



The content is as defined in [RFC4872] Section 14, [RFC4873] Section 6.1.

LSP (protection) Flags or Link flags field can be used by a PCE implementation for routing policy input. The other attributes are only meaningful for a stateful PCE.

This TLV is OPTIONAL and MAY be ignored by the PCE. If ignored by the PCE, it MUST NOT include the TLV in the LSPA of the response. When the TLV is used by the PCE, a LSPA object and the PROTECTION-ATTRIBUTE TLV MUST be included in the response. Fields that were not considered MUST be set to 0.

2.9. NO-PATH Object Extension

The NO-PATH object is used in PCRep messages in response to an unsuccessful path computation request (the PCE could not find a path satisfying the set of constraints). In this scenario, PCE MUST include a NO-PATH object in the PCRep message. The NO-PATH object MAY carry the NO-PATH-VECTOR TLV that specifies more information on the reasons that led to a negative reply. In case of GMPLS networks there could be some additional constraints that led to the failure such as protection mismatch, lack of resources, and so on. Several new flags have been defined in the 32-bit flag field of the NO-PATH-VECTOR TLV but no modifications have been made in the NO-PATH object.

2.9.1. Extensions to NO-PATH-VECTOR TLV

The modified NO-PATH-VECTOR TLV carrying the additional information is as follows:

Bit number TBA-32 - Protection Mismatch (1-bit). Specifies the mismatch of the protection type in the PROTECTION-ATTRIBUTE TLV in the request.

Bit number TBA-33 - No Resource (1-bit). Specifies that the resources are not currently sufficient to provide the path.

Bit number TBA-34 - Granularity not supported (1-bit). Specifies that the PCE is not able to provide a path with the requested granularity.

Bit number TBA-35 - No endpoint label resource (1-bit). Specifies that the PCE is not able to provide a path because of the endpoint label restriction.

Bit number TBA-36 - No endpoint label resource in range (1-bit). Specifies that the PCE is not able to provide a path because of the endpoint label set restriction.

Bit number TBA-37 - No label resource in range (1-bit). Specifies that the PCE is not able to provide a path because of the label set restriction.

3. Additional Error-Types and Error-Values Defined

A PCEP-ERROR object is used to report a PCEP error and is characterized by an Error-Type that specifies the type of error while Error-value that provides additional information about the error. An additional error type and several error values are defined to represent some of the errors related to the newly identified objects related to GMPLS networks. For each PCEP error, an Error-Type and an Error-value are defined. Error-Type 1 to 10 are already defined in [RFC5440]. Additional Error-values are defined for Error-Types 4 and 10. A new Error-Type is defined (value TBA-27).

The Error-Type TBA-27 (path computation failure) is used to reflect constraints not understood by the PCE, for instance when the PCE is not able to understand the generalized bandwidth. If the constraints are understood, but the PCE is unable to find with those constraints, the NO-PATH is to be used.

Error-Type	Error-value	
4	Not supported object	
	value=TBA-14:	Bandwidth Object type TBA-2 or TBA-3 not supported
	value=TBA-15:	Unsupported endpoint type in END-POINTS Generalized Endpoint object type
	value=TBA-16:	Unsupported TLV present in END-POINTS Generalized Endpoint object type
	value=TBA-17:	Unsupported granularity in the RP object flags
10	Reception of an invalid object	
	value=TBA-18:	Bad Bandwidth Object type TBA-2(Generalized bandwidth) or TBA-3(Generalized bandwidth of existing TE-LSP for which a reoptimization is requested)
	value=TBA-20:	Unsupported LSP Protection Flags in PROTECTION-ATTRIBUTE TLV
	value=TBA-21:	Unsupported Secondary LSP Protection Flags in PROTECTION-ATTRIBUTE TLV
	value=TBA-22:	Unsupported Link Protection Type in PROTECTION-ATTRIBUTE TLV
	value=TBA-24:	LABEL-SET TLV present with 0 bit set but without R bit set in RP
	value=TBA-25:	Wrong LABEL-SET TLV present with 0 and L bit set
	value=TBA-26:	Wrong LABEL-SET with O bit set and wrong format
	value=TBA-42:	Missing GMPLS-CAPABILITY TLV
TBA-27	Path computation failure	
	value=0:	Unassigned
	value=TBA-28:	Unacceptable request message
	value=TBA-29:	Generalized bandwidth value not supported
	value=TBA-30:	Label Set constraint could not be met
	value=TBA-31:	Label constraint could not be met

4. Manageability Considerations

This section follows the guidance of [RFC6123].

4.1. Control of Function through Configuration and Policy

This document makes no change to the basic operation of PCEP and so the requirements described in [RFC5440] Section 8.1. also apply to this document. In addition to those requirements a PCEP implementation may allow the configuration of the following parameters:

Accepted RG in the RP object.

Default RG to use (overriding the one present in the PCReq)

Accepted BANDWIDTH object type TBA-2 and TBA-3 parameters in request, default mapping to use when not specified in the request

Accepted LOAD-BALANCING object type TBA-4 parameters in request.

Accepted endpoint type and allowed TLVs in object END-POINTS with object type Generalized Endpoint.

Accepted range for label restrictions in label restriction in END-POINTS, or IRO or XRO objects

PROTECTION-ATTRIBUTE TLV acceptance and suppression.

The configuration of the above parameters is applicable to the different sessions as described in [RFC5440] Section 8.1 (by default, per PCEP peer, etc.).

4.2. Information and Data Models

This document makes no change to the basic operation of PCEP and so the requirements described in [RFC5440] Section 8.2. also apply to this document. This document does not introduce any new ERO sub objects, so that the, ERO information model is already covered in [RFC4802].

4.3. Liveness Detection and Monitoring

This document makes no change to the basic operation of PCEP and so there are no changes to the requirements for liveness detection and monitoring set out in [RFC4657] and [RFC5440] Section 8.3.

4.4. Verifying Correct Operation

This document makes no change to the basic operations of PCEP and considerations described in [RFC5440] Section 8.4. New errors defined by this document should satisfy the requirement to log error events.

4.5. Requirements on Other Protocols and Functional Components

No new Requirements on Other Protocols and Functional Components are made by this document. This document does not require ERO object extensions. Any new ERO subobject defined in the TEAS or CCAMP working group can be adopted without modifying the operations defined in this document.

4.6. Impact on Network Operation

This document makes no change to the basic operations of PCEP and considerations described in [RFC5440] Section 8.6. In addition to the limit on the rate of messages sent by a PCEP speaker, a limit MAY be placed on the size of the PCEP messages.

5. IANA Considerations

IANA assigns values to the PCEP objects and TLVs. IANA is requested to make some allocations for the newly defined objects and TLVs defined in this document. Also, IANA is requested to manage the space of flags that are newly added in the TLVs.

5.1. PCEP Objects

As described in Section 2.3, Section 2.4 and Section 2.5.1 new Objects types are defined. IANA is requested to make the following Object-Type allocations from the "PCEP Objects" sub-registry.

Object Class	5
Name	BANDWIDTH
Object-Type	TBA-2: Generalized bandwidth TBA-3: Generalized bandwidth of an existing TE-LSP for which a reoptimization is requested
Reference	This document (Section 2.3)

Object Class	14
Name	LOAD-BALANCING
Object-Type	TBA-4: Generalized Load Balancing
Reference	This document (Section 2.4)

Object Class	4
Name	END-POINTS
Object-Type	TBA-5: Generalized Endpoint
Reference	This document (Section 2.5)

5.2. Endpoint type field in Generalized END-POINTS Object

IANA is requested to create a registry to manage the Endpoint Type field of the END-POINTS object, Object Type Generalized Endpoint and manage the code space.

New endpoint type in the Reserved range are assigned by Standards Action [RFC8126]. Each endpoint type should be tracked with the following attributes:

- Endpoint type
- Description
- Defining RFC

New endpoint type in the Experimental range are for experimental use; these will not be registered with IANA and MUST NOT be mentioned by RFCs.

The following values have been defined by this document. (Section 2.5.1, Table 5):

Value	Type	Meaning
0	Point-to-Point	
1	Point-to-Multipoint	New leaves to add
2		Old leaves to remove
3		Old leaves whose path can be modified/reoptimized
4		Old leaves whose path has to be left unchanged
5-244	Unassigned	
245-255	Experimental range	

5.3. New PCEP TLVs

IANA manages the PCEP TLV code point registry (see [RFC5440]). This is maintained as the "PCEP TLV Type Indicators" sub-registry of the "Path Computation Element Protocol (PCEP) Numbers" registry. IANA is requested to do the following allocation. Note: TBA-11 is not used

Value	Meaning	Reference
TBA-6	IPV4-ADDRESS	This document (Section 2.5.2.1)
TBA-7	IPV6-ADDRESS	This document (Section 2.5.2.2)
TBA-8	UNNUMBERED-ENDPOINT	This document (Section 2.5.2.3)
TBA-9	LABEL-REQUEST	This document (Section 2.5.2.4)
TBA-10	LABEL-SET	This document (Section 2.5.2.5)
TBA-12	PROTECTION-ATTRIBUTE	This document (Section 2.8)
TBA-1	GMPLS-CAPABILITY	This document (Section 2.1.2)

5.4. RP Object Flag Field

As described in Section 2.2 new flag are defined in the RP Object Flag IANA is requested to make the following Object-Type allocations from the "RP Object Flag Field" sub-registry.

Bit	Description	Reference
TBA-13	routing granularity (2 bits) (RG)	This document, Section 2.2

5.5. New PCEP Error Codes

As described in Section 3, new PCEP Error-Types and Error-values are defined. IANA is requested to make the following allocation in the "PCEP-ERROR Object Error Types and Values" registry.

Error	name	Reference
Type=4	Not supported object	[RFC5440]
Value=TBA-14:	Bandwidth Object type TBA-2 or TBA-3 not supported	This Document
Value=TBA-15:	Unsupported endpoint type in END-POINTS Generalized Endpoint object type	This Document
Value=TBA-16:	Unsupported TLV present in END-POINTS Generalized Endpoint object type	This Document
Value=TBA-17:	Unsupported granularity in the RP object flags	This Document
Type=10	Reception of an invalid object	[RFC5440]
Value=TBA-18:	Bad Bandwidth Object type TBA-2(Generalized bandwidth) or TBA-3(Generalized bandwidth of existing TE-LSP for which a reoptimization is requested)	This Document
Value=TBA-20:	Unsupported LSP Protection Flags in PROTECTION-ATTRIBUTE TLV	This Document
Value=TBA-21:	Unsupported Secondary LSP Protection Flags in PROTECTION-ATTRIBUTE TLV	This Document
Value=TBA-22:	Unsupported Link Protection Type in PROTECTION-ATTRIBUTE TLV	This Document
Value=TBA-24:	LABEL-SET TLV present with 0 bit set but without R bit set in RP	This Document
Value=TBA-25:	Wrong LABEL-SET TLV present with 0 and L bit set	This Document
Value=TBA-26:	Wrong LABEL-SET with O bit set and wrong format	This Document
Value=TBA-42:	Missing GMPLS-CAPABILITY TLV	This Document
Type=TBA-27	Path computation failure	This Document
Value=0	Unassigned	This Document
Value=TBA-28:	Unacceptable request message	This Document
Value=TBA-29:	Generalized bandwidth value not supported	This Document
Value=TBA-30:	Label Set constraint could not be met	This Document
Value=TBA-31:	Label constraint could not be met	This Document

5.6. New NO-PATH-VECTOR TLV Fields

As described in Section 2.9.1, new NO-PATH-VECTOR TLV Flag Fields have been defined. IANA is requested to do the following allocations in the "NO-PATH-VECTOR TLV Flag Field" sub-registry.

Bit number TBA-32 - Protection Mismatch (1-bit). Specifies the mismatch of the protection type of the PROTECTION-ATTRIBUTE TLV in the request.

Bit number TBA-33 - No Resource (1-bit). Specifies that the resources are not currently sufficient to provide the path.

Bit number TBA-34 - Granularity not supported (1-bit). Specifies that the PCE is not able to provide a path with the requested granularity.

Bit number TBA-35 - No endpoint label resource (1-bit). Specifies that the PCE is not able to provide a path because of the endpoint label restriction.

Bit number TBA-36 - No endpoint label resource in range (1-bit). Specifies that the PCE is not able to provide a path because of the endpoint label set restriction.

Bit number TBA-37 - No label resource in range (1-bit). Specifies that the PCE is not able to provide a path because of the label set restriction.

Bit number TBA-40 - LOAD-BALANCING could not be performed with the bandwidth constraits (1 bit). Specifies that the PCE is not able to provide a path because it could not map the BANDWIDTH into the parameters specified by the LOAD-BALANCING.

5.7. New Subobject for the Include Route Object

The "PCEP Parameters" registry contains a subregistry "IRO Subobjects" with an entry for the Include Route Object (IRO).

IANA is requested to add a further subobject that can be carried in the IRO as follows:

Subobject type		Reference
TBA-38	Label subobject	This Document

5.8. New Subobject for the Exclude Route Object

The "PCEP Parameters" registry contains a subregistry "XRO Subobjects" with an entry for the XRO object (Exclude Route Object).

IANA is requested to add a further subobject that can be carried in the XRO as follows:

Subobject type		Reference
TBA-39	Label subobject	This Document

5.9. New GMPLS-CAPABILITY TLV Flag Field

IANA is requested to create a sub-registry to manage the Flag field of the GMPLS-CAPABILITY TLV within the "Path Computation Element Protocol (PCEP) Numbers" registry.

New bit numbers are to be assigned by Standards Action [RFC8126]. Each bit should be tracked with the following qualities:

- Bit number (counting from bit 0 as the most significant bit)
- Capability description
- Defining RFC

The initial contents of the sub-registry are empty, with all bits marked unassigned

6. Security Considerations

GMPLS controls multiple technologies and types of network elements. The LSPs that are established using GMPLS, whose paths can be computed using the PCEP extensions to support GMPLS described in this document, can carry a high volume of traffic and can be a critical part of a network infrastructure. The PCE can then play a key role in the use of the resources and in determining the physical paths of the LSPs and thus it is important to ensure the identity of PCE and PCC, as well as the communication channel. In many deployments there will be a completely isolated network where an external attack is of very low probability. However, there are other deployment cases in which the PCC-PCE communication can be more exposed and there could be more security considerations. Three main situations in case of an attack in the GMPLS PCE context could happen:

- PCE Identity theft: A legitimate PCC could request a path for a GMPLS LSP to a malicious PCE, which poses as a legitimate PCE. The answer can make that the LSP traverses some geographical place known to the attacker where confidentiality (sniffing), integrity (traffic modification) or availability (traffic drop) attacks could be performed by use of an attacker-controlled middlebox device. Also, the resulting LSP can omit constraints given in the requests (e.g., excluding certain fibers, avoiding some SRLGs) which could make that the LSP which will be later set-up can look perfectly fine, but will be in a risky situation. Also, the result can lead to the creation of an LSP that does not provide the desired quality and gives less resources than necessary.

- o PCC Identity theft: A malicious PCC, acting as a legitimate PCC, requesting LSP paths to a legitimate PCE can obtain a good knowledge of the physical topology of a critical infrastructure. It could get to know enough details to plan a later physical attack.

- o Message inspection: As in the previous case, knowledge of an infrastructure can be obtained by sniffing PCEP messages.

The security mechanisms can provide authentication and confidentiality for those scenarios where the PCC-PCE communication cannot be completely trusted. [RFC8253] provides origin verification, message integrity and replay protection, and ensures that a third party cannot decipher the contents of a message.

In order to protect against the malicious PCE case the PCC SHOULD have policies in place to accept or not the path provided by the PCE. Those policies can verify if the path follows the provided constraints. In addition, technology specific data plane mechanism can be used (following [RFC5920] Section 5.8) to verify the data plane connectivity and deviation from constraints.

The document [RFC8253] describes the usage of Transport Layer Security (TLS) to enhance PCEP security. The document describes the initiation of the TLS procedures, the TLS handshake mechanisms, the TLS methods for peer authentication, the applicable TLS ciphersuites for data exchange, and the handling of errors in the security checks. PCE and PCC SHOULD use [RFC8253] mechanism to protect against malicious PCC and PCE.

Finally, as mentioned by [RFC7025] the PCEP extensions to support GMPLS should be considered under the same security as current PCE work and this extension will not change the underlying security issues. However, given the critical nature of the network infrastructures under control by GMPLS, the security issues described above should be seriously considered when deploying a GMPLS-PCE based control plane for such networks. For more information on the security considerations on a GMPLS control plane, not only related to PCE/PCEP, [RFC5920] provides an overview of security vulnerabilities of a GMPLS control plane.

7. Contributing Authors

Elie Sfeir
Coriant
St Martin Strasse 76
Munich, 81541
Germany

Email: elie.sfeir@coriant.com

Franz Rambach
Nockherstrasse 2-4,
Munich 81541
Germany

Phone: +49 178 8855738
Email: franz.rambach@cgi.com

Francisco Javier Jimenez Chico
Telefonica Investigacion y Desarrollo
C/ Emilio Vargas 6
Madrid, 28043
Spain

Phone: +34 91 3379037
Email: fjjc@tid.es

Huawei Technologies

Suresh BR
Shenzhen
China
Email: sureshbr@huawei.com

Young Lee
1700 Alma Drive, Suite 100
Plano, TX 75075
USA

Phone: (972) 509-5599 (x2240)
Email: ylee@huawei.com

SenthilKumar S
Shenzhen
China
Email: senthilkumars@huawei.com

Jun Sun
Shenzhen
China
Email: johnsun@huawei.com

CTTC - Centre Tecnologic de Telecomunicacions de Catalunya

Ramon Casellas
PMT Ed B4 Av. Carl Friedrich Gauss 7
08860 Castelldefels (Barcelona)
Spain
Phone: (34) 936452916
Email: ramon.casellas@cttc.es

8. Acknowledgments

The research of Ramon Casellas, Francisco Javier Jimenez Chico, Oscar Gonzalez de Dios, Cyril Margaria, and Franz Rambach leading to these results has received funding from the European Community's Seventh Framework Program FP7/2007-2013 under grant agreement no 247674 and no 317999.

The authors would like to thank Julien Meuric, Lyndon Ong, Giada Lander, Jonathan Hardwick, Diego Lopez, David Sinicrope, Vincent Roca and Tianran Zhou for their review and useful comments to the document.

9. References

9.1. Normative References

[G.709-v3] ITU-T, "Interfaces for the optical transport network, Recommendation G.709/Y.1331", June 2016, <https://www.itu.int/rec/T-REC-G.709-201606-I/en>.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC2210] Wroclawski, J., "The Use of RSVP with IETF Integrated Services", RFC 2210, DOI 10.17487/RFC2210, September 1997, <https://www.rfc-editor.org/info/rfc2210>.

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, DOI 10.17487/RFC3209, December 2001, <https://www.rfc-editor.org/info/rfc3209>.

[RFC3471] Berger, L., Ed., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, DOI 10.17487/RFC3471, January 2003, <https://www.rfc-editor.org/info/rfc3471>.

[RFC3473] Berger, L., Ed., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, DOI 10.17487/RFC3473, January 2003, <https://www.rfc-editor.org/info/rfc3473>.

[RFC3477] Kompella, K. and Y. Rekhter, "Signalling Unnumbered Links in Resource ReSerVation Protocol - Traffic Engineering (RSVP-TE)", RFC 3477, DOI 10.17487/RFC3477, January 2003, <https://www.rfc-editor.org/info/rfc3477>.

[RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, DOI 10.17487/RFC3630, September 2003, <https://www.rfc-editor.org/info/rfc3630>.

[RFC4003] Berger, L., "GMPLS Signaling Procedure for Egress Control", RFC 4003, DOI 10.17487/RFC4003, February 2005, <https://www.rfc-editor.org/info/rfc4003>.

[RFC4328] Papadimitriou, D., Ed., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Extensions for G.709 Optical Transport Networks Control", RFC 4328, DOI 10.17487/RFC4328, January 2006, <https://www.rfc-editor.org/info/rfc4328>.

[RFC4606] Mannie, E. and D. Papadimitriou, "Generalized Multi-Protocol Label Switching (GMPLS) Extensions for Synchronous Optical Network (SONET) and Synchronous Digital Hierarchy (SDH) Control", RFC 4606, DOI 10.17487/RFC4606, August 2006, <https://www.rfc-editor.org/info/rfc4606>.

[RFC4802] Nadeau, T., Ed. and A. Farrel, Ed., "Generalized Multiprotocol Label Switching (GMPLS) Traffic Engineering Management Information Base", RFC 4802, DOI 10.17487/RFC4802, February 2007, <https://www.rfc-editor.org/info/rfc4802>.

[RFC4872] Lang, J., Ed., Rekhter, Y., Ed., and D. Papadimitriou, Ed., "RSVP-TE Extensions in Support of End-to-End Generalized Multi-Protocol Label Switching (GMPLS) Recovery", RFC 4872, DOI 10.17487/RFC4872, May 2007, <https://www.rfc-editor.org/info/rfc4872>.

[RFC4873] Berger, L., Bryskin, I., Papadimitriou, D., and A. Farrel, "GMPLS Segment Recovery", RFC 4873, DOI 10.17487/RFC4873, May 2007, <https://www.rfc-editor.org/info/rfc4873>.

[RFC5088] Le Roux, JL., Ed., Vasseur, JP., Ed., Ikejiri, Y., and R. Zhang, "OSPF Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5088, DOI 10.17487/RFC5088, January 2008, <https://www.rfc-editor.org/info/rfc5088>.

[RFC5089] Le Roux, JL., Ed., Vasseur, JP., Ed., Ikejiri, Y., and R. Zhang, "IS-IS Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5089, DOI 10.17487/RFC5089, January 2008, <https://www.rfc-editor.org/info/rfc5089>.

[RFC5440] Vasseur, JP., Ed. and JL. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, DOI 10.17487/RFC5440, March 2009, <https://www.rfc-editor.org/info/rfc5440>.

[RFC5511] Farrel, A., "Routing Backus-Naur Form (RBNF): A Syntax Used to Form Encoding Rules in Various Routing Protocol Specifications", RFC 5511, DOI 10.17487/RFC5511, April 2009, <https://www.rfc-editor.org/info/rfc5511>.

[RFC5520] Bradford, R., Ed., Vasseur, JP., and A. Farrel, "Preserving Topology Confidentiality in Inter-Domain Path Computation Using a Path-Key-Based Mechanism", RFC 5520, DOI 10.17487/RFC5520, April 2009, <https://www.rfc-editor.org/info/rfc5520>.

[RFC5521] Oki, E., Takeda, T., and A. Farrel, "Extensions to the Path Computation Element Communication Protocol (PCEP) for Route Exclusions", RFC 5521, DOI 10.17487/RFC5521, April 2009, <https://www.rfc-editor.org/info/rfc5521>.

[RFC5541] Le Roux, JL., Vasseur, JP., and Y. Lee, "Encoding of Objective Functions in the Path Computation Element Communication Protocol (PCEP)", RFC 5541, DOI 10.17487/RFC5541, June 2009, <https://www.rfc-editor.org/info/rfc5541>.

[RFC6001] Papadimitriou, D., Vigoureux, M., Shiomoto, K., Brungard, D., and JL. Le Roux, "Generalized MPLS (GMPLS) Protocol Extensions for Multi-Layer and Multi-Region Networks (MLN/MRN)", RFC 6001, DOI 10.17487/RFC6001, October 2010, <https://www.rfc-editor.org/info/rfc6001>.

[RFC6003] Papadimitriou, D., "Ethernet Traffic Parameters", RFC 6003, DOI 10.17487/RFC6003, October 2010, <https://www.rfc-editor.org/info/rfc6003>.

[RFC6205] Otani, T., Ed. and D. Li, Ed., "Generalized Labels for Lambda-Switch-Capable (LSC) Label Switching Routers", RFC 6205, DOI 10.17487/RFC6205, March 2011, <https://www.rfc-editor.org/info/rfc6205>.

[RFC6387] Takacs, A., Berger, L., Caviglia, D., Fedyk, D., and J. Meuric, "GMPLS Asymmetric Bandwidth Bidirectional Label Switched Paths (LSPs)", RFC 6387, DOI 10.17487/RFC6387, September 2011, <https://www.rfc-editor.org/info/rfc6387>.

[RFC7139] Zhang, F., Ed., Zhang, G., Belotti, S., Ceccarelli, D., and K. Pithewan, "GMPLS Signaling Extensions for Control of Evolving G.709 Optical Transport Networks", RFC 7139, DOI 10.17487/RFC7139, March 2014, <https://www.rfc-editor.org/info/rfc7139>.

[RFC7570] Margaria, C., Ed., Martinelli, G., Balls, S., and B. Wright, "Label Switched Path (LSP) Attribute in the Explicit Route Object (ERO)", RFC 7570, DOI 10.17487/RFC7570, July 2015, <https://www.rfc-editor.org/info/rfc7570>.

[RFC7792] Zhang, F., Zhang, X., Farrel, A., Gonzalez de Dios, O., and D. Ceccarelli, "RSVP-TE Signaling Extensions in Support of Flexi-Grid Dense Wavelength Division Multiplexing (DWDM) Networks", RFC 7792, DOI 10.17487/RFC7792, March 2016, <https://www.rfc-editor.org/info/rfc7792>.

[RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 8126, DOI 10.17487/RFC8126, June 2017, <https://www.rfc-editor.org/info/rfc8126>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

[RFC8253] Lopez, D., Gonzalez de Dios, O., Wu, Q., and D. Dhody, "PCEPS: Usage of TLS to Provide a Secure Transport for the Path Computation Element Communication Protocol (PCEP)", RFC 8253, DOI 10.17487/RFC8253, October 2017, <https://www.rfc-editor.org/info/rfc8253>.

[RFC8282] Oki, E., Takeda, T., Farrel, A., and F. Zhang, "Extensions to the Path Computation Element Communication Protocol (PCEP) for Inter-Layer MPLS and GMPLS Traffic Engineering", RFC 8282, DOI 10.17487/RFC8282, December 2017, <https://www.rfc-editor.org/info/rfc8282>.

[RFC8306] Zhao, Q., Dhody, D., Ed., Palleti, R., and D. King, "Extensions to the Path Computation Element Communication Protocol (PCEP) for Point-to-Multipoint Traffic Engineering Label Switched Paths", RFC 8306, DOI 10.17487/RFC8306, November 2017, <https://www.rfc-editor.org/info/rfc8306>.

9.2. Informative References

[RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, DOI 10.17487/RFC4655, August 2006, <https://www.rfc-editor.org/info/rfc4655>.

[RFC4657] Ash, J., Ed. and J. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol Generic Requirements", RFC 4657, DOI 10.17487/RFC4657, September 2006, <https://www.rfc-editor.org/info/rfc4657>.

[RFC5920] Fang, L., Ed., "Security Framework for MPLS and GMPLS Networks", RFC 5920, DOI 10.17487/RFC5920, July 2010, <https://www.rfc-editor.org/info/rfc5920>.

[RFC6123] Farrel, A., "Inclusion of Manageability Sections in Path Computation Element (PCE) Working Group Drafts", RFC 6123, DOI 10.17487/RFC6123, February 2011, <https://www.rfc-editor.org/info/rfc6123>.

[RFC6163] Lee, Y., Ed., Bernstein, G., Ed., and W. Imajuku, "Framework for GMPLS and Path Computation Element (PCE) Control of Wavelength Switched Optical Networks (WSONs)", RFC 6163, DOI 10.17487/RFC6163, April 2011, <https://www.rfc-editor.org/info/rfc6163>.

[RFC7025] Otani, T., Ogaki, K., Caviglia, D., Zhang, F., and C. Margaria, "Requirements for GMPLS Applications of PCE", RFC 7025, DOI 10.17487/RFC7025, September 2013, <https://www.rfc-editor.org/info/rfc7025>.

[RFC7449] Lee, Y., Ed., Bernstein, G., Ed., Martensson, J., Takeda, T., Tsuritani, T., and O. Gonzalez de Dios, "Path Computation Element Communication Protocol (PCEP) Requirements for Wavelength Switched Optical Network (WSON) Routing and Wavelength Assignment", RFC 7449, DOI 10.17487/RFC7449, February 2015, <https://www.rfc-editor.org/info/rfc7449>.

Appendix A. LOAD-BALANCING Usage for SDH Virtual Concatenation

For example a request for one co-signaled n x VC-4 TE-LSP will not use the LOAD-BALANCING. In case the VC-4 components can use different paths, the BANDWIDTH with object type TBA-2 will contain a traffic specification indicating the complete n x VC-4 traffic specification and the LOAD-BALANCING the minimum co-signaled VC-4. For an SDH network, a request to have a TE-LSP group with 10 VC-4 containers, each path using at minimum 2 x VC-4 containers, can be represented with a BANDWIDTH object with OT=TBA-2, Bw Spec Type set to 4, the content of the Generalized Bandwidth is ST=6, RCC=0, NCC=0, NVC=10, MT=1. The LOAD-BALANCING, OT=TBA-4 with Bw Spec Type set to 4, Max-LSP=5, Min Bandwidth Spec is (ST=6, RCC=0, NCC=0, NVC=2, MT=1). The PCE can respond with a response with maximum 5 paths, each of them having a BANDWIDTH OT=TBA-2 and Generalized Bandwidth matching the Min Bandwidth Spec from the LOAD-BALANCING object of the corresponding request.

Authors' Addresses

Cyril Margaria (editor)
Juniper

Email: cmargaria@juniper.net

Oscar Gonzalez de Dios (editor)
Telefonica Investigacion y Desarrollo
C/ Ronda de la Comunicacion
Madrid 28050
Spain

Phone: +34 91 4833441
Email: oscar.gonzalezdedios@telefonica.com

Fatai Zhang (editor)
Huawei Technologies
F3-5-B R&D Center, Huawei Base
Bantian, Longgang District
Shenzhen 518129
P.R.China

Email: zhangfatai@huawei.com

Network Working Group
Internet-Draft
Intended Status: Informational
Expires: April 4, 2012

D. King (Ed.)
Old Dog Consulting
A. Farrel (Ed.)
Old Dog Consulting
October 4, 2011

The Application of the Path Computation Element Architecture to the Determination of a Sequence of Domains in MPLS and GMPLS

draft-ietf-pce-hierarchy-fwk-00.txt

Abstract

Computing optimum routes for Label Switched Paths (LSPs) across multiple domains in MPLS Traffic Engineering (MPLS-TE) and GMPLS networks presents a problem because no single point of path computation is aware of all of the links and resources in each domain. A solution may be achieved using the Path Computation Element (PCE) architecture.

Where the sequence of domains is known a priori, various techniques can be employed to derive an optimum path. If the domains are simply-connected, or if the preferred points of interconnection are also known, the Per-Domain Path Computation technique can be used. Where there are multiple connections between domains and there is no preference for the choice of points of interconnection, the Backward Recursive Path Computation Procedure (BRPC) can be used to derive an optimal path.

This document examines techniques to establish the optimum path when the sequence of domains is not known in advance. The document shows how the PCE architecture can be extended to allow the optimum sequence of domains to be selected, and the optimum end-to-end path to be derived through the use of a hierarchical relationship between domains.

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."

This Internet-Draft will expire on April 4, 2012.

Contents

1. Introduction

The capability to compute the routes of end-to-end inter-domain MPLS Traffic Engineering (TE) and GMPLS Label Switched Paths (LSPs) may be provided by a Path Computation Element (PCE). The PCE architecture is defined in [RFC4655]. The methods for establishing and controlling inter-domain MPLS-TE and GMPLS LSPs are documented in [RFC4726].

A domain can be defined as a separate administrative, geographic, or switching environment within the network. A domain may be further defined as a zone of routing or computational ability. Under these definitions a domain might be categorized as an Antonymous System (AS) or an Interior Gateway Protocol (IGP) area [RFC4726] and [RFC4655]. Domains are connected through ingress and egress boundary nodes (BNs). A more detailed definition is given in Section 1.2.

In a multi-domain environment, the determination of an end-to-end traffic engineered path is a problem because no single point of path computation is aware of all of the links and resources in each domain. PCEs can be used to compute end-to-end paths using a per-domain path computation technique [RFC5152]. Alternatively, the backward recursive path computation (BRPC) mechanism [RFC5441] allows multiple PCEs to collaborate in order to select an optimal end-to-end path that crosses multiple domains. Both mechanisms assume that the sequence of domains to be crossed between ingress and egress in known in advance.

This document examines techniques to establish the optimum path when the sequence of domains is not known in advance. It shows how the PCE architecture can be extended to allow the optimum sequence of domains to be selected, and the optimum end-to-end path to be derived.

The model described in this document introduces a hierarchical relationship between domains. It is applicable to environments with small groups of domains where visibility from the ingress Label Switching Router (LSR) is limited. Applying the hierarchical PCE model to large groups of domains such as the Internet, is not considered feasible or desirable, and is out of scope for this document.

1.1 Problem Statement

Using a PCE to compute a path between nodes within a single domain is relatively straightforward. Computing an end-to-end path when the source and destination nodes are located in different domains requires co-operation between multiple PCEs, each responsible for its own domain.

Techniques for inter-domain path computation described so far ([RFC5152] and [RFC5441]) assume that the sequence of domains to be crossed from source to destination is well known. No explanation is given (for example, in [RFC4655]) of how this sequence is generated or what criteria may be used for the selection of paths between domains. In small clusters of domains, such as simple cooperation between adjacent ISPs, this selection process is not complex. In more advanced deployments (such as optical networks constructed from multiple sub-domains, or multi-AS environments) the choice of domains in the end-to-end domain sequence can be critical to the determination of an optimum end-to-end path.

This document introduces the concept of a hierarchical PCE architecture and shows how to coordinate PCEs in peer domains in order to derive an optimal end-to-end path.

The work is currently scoped to operate with a small group of domains and there is no intent to apply this model to a large group of domains, e.g., to the Internet.

1.2 Definition of a Domain

A domain is defined in [RFC4726] as 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. Wholly or partially overlapping domains are not within the scope of this document.

In the context of GMPLS, a particularly important example of a domain is the Automatically Switched Optical Network (ASON) subnetwork [G-8080]. In this case, computation of an end-to-end path requires the selection of nodes and links within a parent domain where some nodes may, in fact, be subnetworks. Furthermore, a domain might be an ASON routing area [G-7715]. A PCE may perform the path computation function of an ASON routing controller as described in [G-7715-2].

See Section 6.2 for a further discussion of the applicability to the ASON architecture.

This document assumes that the selection of a sequence of domains for an end-to-end path is in some sense a hierarchical path computation problem. That is, where one mechanism is used to determine a path across a domain, a separate mechanism (or at least a separate set of paradigms) is used to determine the sequence of domains.

1.3 Assumptions and Requirements

Networks are often constructed from multiple domains. These domains are often interconnected via multiple interconnect points. Its assumed that the sequence of domains for an end-to-end path is not always well known; that is, an application requesting end-to-end connectivity has no preference for, or no ability to specify, the sequence of domains to be crossed by the path.

The traffic engineering properties of a domain cannot be seen from outside the domain. Traffic engineering aggregation or abstraction, hides information and can lead to failed path setup or the selection of suboptimal end-to-end paths [RFC4726]. The aggregation process may also have significant scaling issues for networks with many possible routes and multiple TE metrics. Flooding TE information breaks confidentiality and does not scale in the routing protocol.

The primary goal of this document is to define how to derive optimal end-to-end, multi-domain paths when the sequence of domains is not known in advance. The solution needs to be scalable and to maintain

internal domain topology confidentiality while providing the optimal end-to-end path. It cannot rely on the exchange of TE information between domains, and it cannot utilise a computation element that has universal knowledge of TE properties and topology of all domains.

The sub-sections that follow set out the primary objectives and requirements to be satisfied by a PCE solution to multi-domain path computation.

1.3.1 Metric Objectives

The definition of optimality is dependent on policy, and is based on a single objective or a group objectives. An objective is expressed as an objective function [RFC5541] and may be specified on a path computation request. The following objective functions are identified in this document. They define how the path metrics and TE link qualities are manipulated during inter-domain path computation. The list is not proscriptive and may be expanded in other documents.

- Minimize the cost of the path [RFC5541]
- Select a path using links with the minimal load [RFC5541]
- Select a path that leaves the maximum residual bandwidth [RFC5541]
- Minimize aggregate bandwidth consumption [RFC5541]
- Minimize the Load of the most loaded Link [RFC5541]
- Minimize the Cumulative Cost of a set of paths [RFC5541]
- Minimize the number of boundary nodes used
- Limit the number of domains crossed
- Disallow domain re-entry

See Section 5.1 for further discussion of objective functions.

1.3.2 Domain Diversity

A pair of paths are domain-diverse if they do not transit any of the same domains. A pair of paths that share a common ingress and egress are domain-diverse if they only share the same domains at the ingress and egress (the ingress and egress domains). Domain diversity may be maximized for a pair of paths by selecting paths that have the smallest number of shared domains. (Note that this is not the same as finding paths with the greatest number of distinct domains!)

Path computation should facilitate the selection of paths that share ingress and egress domains, but do not share any transit domains. This provides a way to reduce the risk of shared failure along any path, and automatically helps to ensure path diversity for most of the route of a pair of LSPs.

Thus, domain path selection should provide the capability to include or exclude specific domains and specific boundary nodes.

1.3.3 Existing Traffic Engineering Constraints

Any solution should take advantage of typical traffic engineering constraints (hop count, bandwidth, lambda continuity, path cost, etc.) to meet the service demands expressed in the path computation request [RFC4655].

1.3.4 Commercial Constraints

The solution should provide the capability to include commercially relevant constraints such as policy, SLAs, security, peering preferences, and dollar costs.

Additionally it may be necessary for the service provider to request that specific domains are included or excluded based on commercial relationships, security implications, and reliability.

1.3.5 Domain Confidentiality

A key requirement is the ability to maintain domain confidentiality when computing inter-domain end-to-end paths. When required by local policy, a PCE should not need to disclose to any other PCE the intra-domain paths it computes or the internal topology of the domain it serves.

1.3.6 Limiting Information Aggregation

It is important to minimise the amount of aggregation within the solution. There should be no associated computation burden or requirement to aggregate and abstract traffic engineering link information.

1.3.7 Domain Interconnection Discovery

To support domain mesh topologies, the solution should allow the discovery and selection of domain inter-connections. Pre-configuration of preferred domain interconnections should also be supported for network operators that have bilateral agreement, and preference for the choice of points of interconnection.

1.4 Terminology

This document uses PCE terminology defined in [RFC4655], [RFC4875], and [RFC5440]. Additional terms are defined below.

Domain Path: The sequence of domains for a path.

Ingress Domain: The domain that includes the ingress LSR of a path.

Transit Domain: A domain that has an upstream and downstream neighbor domain for a specific path.

Egress Domain: The domain that includes the egress LSR of a path.

Boundary Nodes: Each Domain has entry LSRs and exit LSRs that could be Area Border Routers (ABRs) or Autonomous System Border Routers (ASBRs) depending on the type of domain. They are defined here more generically as Boundary Nodes (BNs).

Entry BN of domain(n): a BN connecting domain(n-1) to domain(n) on a path.

Exit BN of domain(n): a BN connecting domain(n) to domain(n+1) on a path.

Parent Domain: A domain higher up in a domain hierarchy such that it contains other domains (child domains) and potentially other links and nodes.

Child Domain: A domain lower in a domain hierarchy such that it has a parent domain.

Parent PCE: A PCE responsible for selecting a path across a parent domain and any number of child domains by coordinating with child PCEs and examining a topology map that shows domain inter-connectivity.

Child PCE: A PCE responsible for computing the path across one or more specific (child) domains. A child PCE maintains a relationship with at least one parent PCE.

OF: Objective Function: A set of one or more optimization criteria used for the computation of a single path (e.g., path cost minimization), or the synchronized computation of a set of paths (e.g., aggregate bandwidth consumption minimization). See [RFC4655] and [RFC5541].

2. Per-Domain Path Computation

The per-domain path computation method for establishing inter-domain TE-LSPs [RFC5152] defines a technique whereby the path is computed during the signalling process on a per-domain basis. The entry BN of each domain is responsible for performing the path computation for the section of the LSP that crosses the domain, or for requesting that a PCE for that domain computes that piece of the path.

During per-domain path computation, each computation results in the best path across the domain to provide connectivity to the next domain in the domain sequence (usually indicated in signalling by an identifier of the next domain or the identity of the next entry BN).

Per-domain path computation may lead to sub-optimal end-to-end paths because the most optimal path in one domain may lead to the choice of an entry BN for the next domain that results in a very poor path across that next domain.

In the case that the domain path (in particular, the sequence of boundary nodes) is not known, the PCE must select an exit BN based on some determination of how to reach the destination that is outside the domain for which the PCE has computational responsibility. [RFC5152] suggest that this might be achieved using the IP shortest path as advertise by BGP. Note, however, that the existence of an IP forwarding path does guarantee the presence of sufficient bandwidth, let alone an optimal TE path. Furthermore, in many GMPLS systems inter-domain IP routing will not be present. Thus, per-domain path computation may require a significant number of crankback routing attempts to establish even a sub-optimal path.

Note also that the PCEs in each domain may have different computation capabilities, may run different path computation algorithms, and may apply different sets of constraints and optimization criteria, etc. This can result in the end-to-end path being inconsistent and sub-optimal.

Per-domain path computation can suit simply-connected domains where the preferred points of interconnection are known.

3. Backward Recursive Path Computation

The Backward Recursive Path Computation (BRPC) [RFC5441] procedure involves cooperation and communication between PCEs in order to compute an optimal end-to-end path across multiple domains. The sequence of domains to be traversed can either be determined before or during the path computation. In the case where the sequence of domains is known, the ingress Path Computation Client (PCC) sends a path computation request to the PCE responsible for the ingress domain. This request is forwarded between PCEs, domain-by-domain, to the PCE responsible for the egress domain. The PCE in the egress domain creates a set of optimal paths from all of the domain entry BNs to the egress LSR. This set is represented as a tree of potential paths called a Virtual Shortest Path Tree (VSPT), and the PCE passes it back to the previous PCE on the domain path. As the VSPT is passed back toward the ingress domain, each PCE computes the optimal paths from its entry BNs to its exit BNs that connect to the rest of the

tree. It adds these paths to the VSPT and passes the VSPT on until the PCE for the ingress domain is reached and computes paths from the ingress LSR to connect to the rest of the tree. The ingress PCE then selects the optimal end-to-end path from the tree, and returns the path to the initiating PCC.

BRPC may suit environments where multiple connections exist between domains and there is no preference for the choice of points of interconnection. It is best suited to scenarios where the domain path is known in advance, but can also be used when the domain path is not known.

3.1. Applicability of BRPC when the Domain Path is Not Known

As described above BRPC can be used to determine an optimal inter-domain path when the sequence is known. Even when the sequence of domains is not known BRPC could be used as follows.

- The PCC sends a request to the PCE for the ingress domain (the ingress PCE).
- The ingress PCE sends the path computation request direct to the PCE responsible for the domain containing the destination node (the egress PCE).
- The egress PCE computes an egress VSPT and passes it to a PCE responsible for each of the adjacent (potentially upstream) domains.
- Each PCE in turn constructs a VSPT and passes it on to all of its neighboring PCEs.
- When the ingress PCE has received a VSPT from each of its neighboring domains it is able to select the optimum path.

Clearly this mechanism (which could be called path computation flooding) has significant scaling issues. It could be improved by the application of policy and filtering, but such mechanisms are not simple and would still leave scaling concerns.

4. Hierarchical PCE

In the hierarchical PCE architecture, a parent PCE maintains a domain topology map that contains the child domains (seen as vertices in the topology) and their interconnections (links in the topology). The parent PCE has no information about the content of the child domains; that is, the parent PCE does not know about the resource availability within the child domains, nor about the availability of connectivity across each domain. The parent PCE is aware of the TE capabilities of the interconnections between child domains as these interconnections are links in its own topology map.

Note that in the case that the domains are IGP areas, there is no link between the domains (the ABRs have a presence in both neighboring areas). The parent domain may choose to represent this in its TED as a virtual link that is unconstrained and has zero cost, but this is entirely an implementation issue.

Each child domain has at least one PCE capable of computing paths across the domain. These PCEs are known as child PCEs and have a relationship with the parent PCE. Each child PCE also knows the identity of the domains that neighbor its own domain. A child PCE only knows the topology of the domain that it serves and does not know the topology of other child domains. Child PCEs are also not aware of the general domain mesh connectivity (i.e., the domain topology map) beyond the connectivity to the immediate neighbor domains of the domain it serves.

The parent PCE builds the domain topology map either from configuration or from information received from each child PCE. This tells it how the domains are interconnected including the TE properties of the domain interconnections. But the parent PCE does not know the contents of the child domains. Discovery of the domain topology and domain interconnections is discussed further in Section 5.3.

When a multi-domain path is needed, the ingress PCE sends a request to the parent PCE (using the path computation element protocol, PCEP [RFC5440]). The parent PCE selects a set of candidate domain paths based on the domain topology and the state of the inter-domain links. It then sends computation requests to the child PCEs responsible for each of the domains on the candidate domain paths.

Each child PCE computes a set of candidate path segments across its domain and sends the results to the parent PCE. The parent PCE uses this information to select path segments and concatenate them to derive the optimal end-to-end inter-domain path. The end-to-end path is then sent to the child PCE which received the initial path request and this passes the path on to the PCC that issues the original request.

5. Hierarchical PCE Procedures

5.1 Objective Functions and Policy

Deriving the optimal end-to-end domain path sequence is dependent on the policy applied during domain path computation. An Objective Function (OF) [RFC5541], or set of OFs, may be applied to define the policy being applied to the domain path computation.

The OF specifies the desired outcome of the computation. It does not describe the algorithm to use. When computing end-to-end inter-domain paths, required OFs may include (see Section 1.3.1):

- o Minimum cost path
- o Minimum load path
- o Maximum residual bandwidth path
- o Minimize aggregate bandwidth consumption
- o Minimize the number of boundary nodes used
- o Minimize the number of transit domains
- o Disallow domain re-entry

The objective function may be requested by the PCC, the ingress domain PCE (according to local policy), or maybe applied by the parent PCE according to inter-domain policy.

5.2 Maintaining Domain Confidentiality

Information about the content of child domains is not shared for scaling and confidentiality reasons. This means that a parent PCE is aware of the domain topology and the nature of the connections between domains, but is not aware of the content of the domains. Similarly, a child PCE cannot know the internal topology of another child domain. Child PCEs also do not know the general domain mesh connectivity, this information is only known by the parent PCE.

As described in the earlier sections of this document, PCEs can exchange path information in order to construct an end-to-end inter-domain path. Each per-domain path fragment reveals information about the topology and resource availability within a domain. Some management domains or ASes will not want to share this information outside of the domain (even with a trusted parent PCE). In order to conceal the information, a PCE may replace a path segment with a path-key [RFC5520]. This mechanism effectively hides the content of a segment of a path.

5.3 PCE Discovery

It is a simple matter for each child PCE to be configured with the address of its parent PCE. Typically, there will only be one or two parents of any child.

The parent PCE also needs to be aware of the child PCEs for all child domains that it can see. This information is most likely to be configured (as part of the administrative definition of each domain).

Discovery of the relationships between parent PCEs and child PCEs does not form part of the H-PCE architecture. Mechanisms that rely on advertising or querying PCE locations across domain or provider boundaries are undesirable for security, scaling, commercial, and confidentiality reasons.

The parent PCE also needs to know the inter-domain connectivity. This information could be configured with suitable policy and commercial rules, or could be learned from the child PCEs as described in Section 4.

In order for the parent PCE to learn about domain interconnection the child PCE will report the identity of its neighbor domains. The IGP in each neighbor domain can advertise its inter-domain TE link capabilities [RFC5316], [RFC5392]. This information can be collected by the child PCEs and forwarded to the parent PCE, or the parent PCE could participate in the IGP in the child domains.

5.4 Parent Domain Traffic Engineering Database

The parent PCE maintains a domain topology map of the child domains and their interconnectivity. Where inter-domain connectivity is provided by TE links the capabilities of those links must also be known to the parent PCE. Furthermore the parent domain may contain nodes and links in its own right. Therefore, the parent PCE maintains a traffic engineering database (TED) for the parent domain in the same way that any PCE does.

The parent domain may just be the collection of child domains and the inter-domain links, or it may contain nodes and links in its own right.

The mechanism for building the parent TED is likely to rely heavily on administrative configuration and commercial issues because the network was probably partitioned into domains specifically to address these issues.

In practice, certain information may be passed from the child domains to the parent PCE to help build the parent TED. In theory, the parent PCE could listen to the routing protocols in the child domains, but this would violate the confidentiality and scaling issues that may be responsible for the partition of the network into domains. So it is much more likely that a suitable solution will involve specific communication from an entity in the child domain (such as the child PCE) to convey the necessary information. As already mentioned, the "necessary information" relates to how the child domains are inter-connected. The topology and available resources within the child domain do not need to be communicated to the parent PCE: doing so would violate the PCE architecture. Mechanisms for reporting this information are described in the examples in Section 5.6 in abstract terms as "a child PCE reports its neighbor domain connectivity to its parent PCE"; the specifics of a solution are out of scope of this document, but the requirements are indicated in Section 5.8.

In models such as ASON (see Section 6.2), it is possible to consider a separate instance of an IGP running within the parent domain where the participating protocol speakers are the nodes directly present in that domain and the PCEs (routing controllers) responsible for each of the child domains.

5.5 Determination of Destination Domain

The PCC asking for an inter-domain path computation is aware of the identity of the destination node by definition. If it knows the egress domain it can supply this information as part of the path computation request. However, if it does not know the egress domain this information must be determined by the parent PCE.

In some specialist topologies the parent PCE could determine the destination domain based on the destination address, for example from configuration. However, this is not appropriate for many multi-domain addressing scenarios. In IP-based multi-domain networks the parent PCE may be able to determine the destination domain by participating in inter-domain routing. Finally, the parent PCE could issue specific requests to the child PCEs to discover if they contain the destination node, but this has scaling implications.

5.6 Hierarchical PCE Examples

The following example describes the hierarchical domain topology. Figure 1 (sample hierarchical domain topology) demonstrates four interconnected domains within a fifth parent domain. Each domain contains a single PCE:

- Domain 1 is the ingress domain and child PCE 1 is able to compute paths within the domain. Its neighbors are Domain 2 and Domain 4. The domain also contains the source LSR (S) and three egress boundary nodes (BN11, BN12, and BN13).

- Domain 2 is served by child PCE 2. Its neighbors are Domain 1 and Domain 3. The domain also contains four boundary nodes (BN21, BN22, BN23, and BN24).

- Domain 3 is the egress domain and is served by child PCE 3. Its neighbors are Domain 2 and Domain 4. The domain also contains the destination LSR (D) and three ingress boundary nodes (BN31, BN32, and BN33).

- Domain 4 is served by child PCE 4. Its neighbors are Domain 2 and Domain 3. The domain also contains two boundary nodes (BN41 and BN42).

All of these domains are encompassed within Domain 5 which is served by the parent PCE (PCE 5).

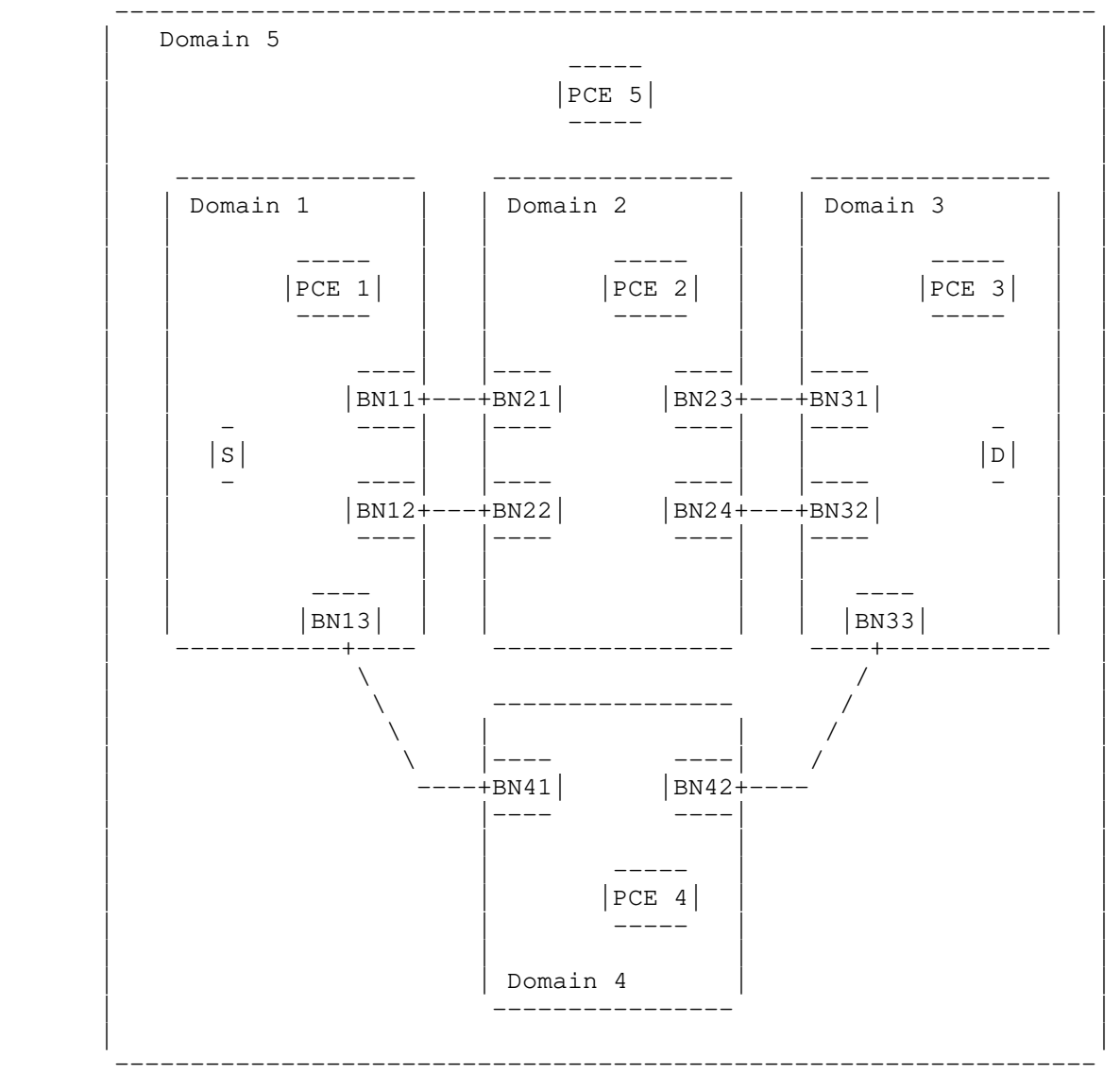


Figure 1 : Sample Hierarchical Domain Topology

Figure 2, shows the view of the domain topology as seen by the parent PCE (PCE 5). This view is an abstracted topology; PCE 5 is aware of domain connectivity, but not of the internal topology within each domain.

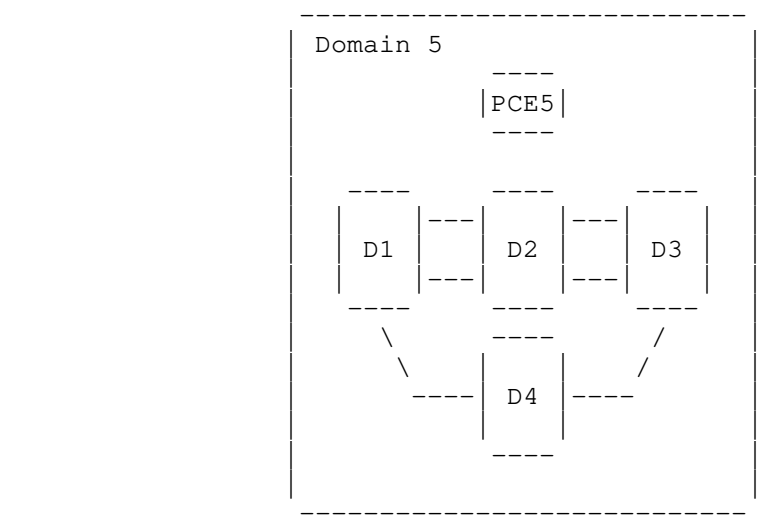


Figure 2 : Abstract Domain Topology as Seen by the Parent PCE

5.6.1 Hierarchical PCE Initial Information Exchange

Based on the Figure 1 topology, the following is an illustration of the initial hierarchical PCE information exchange.

1. Child PCE 1, the PCE responsible for Domain 1, is configured with the location of its parent PCE (PCE5).

2. Child PCE 1 establishes contact with its parent PCE. The parent applies policy to ensure that communication with PCE 1 is allowed.

3. Child PCE 1 listens to the IGP in its domain and learns its inter-domain connectivity. That is, it learns about the links BN11-BN21, BN12-BN22, and BN13-BN41.

4. Child PCE 1 reports its neighbor domain connectivity to its parent PCE.

5. Child PCE 1 reports any change in the resource availability on its inter-domain links to its parent PCE.

Each child PCE performs steps 1 through 5 so that the parent PCE can create a domain topology view as shown in Figure 2.

5.6.2 Hierarchical PCE End-to-End Path Computation Procedure

The procedure below is an example of a source PCC requesting an

end-to-end path in a multi-domain environment. The topology is represented in Figure 1. It is assumed that the each child PCE has connected to its parent PCE and exchanged the initial information required for the parent PCE to create its domain topology view as described in Section 5.6.1.

1. The source PCC (the ingress LSR in our example), sends a request to the PCE responsible for its domain (PCE1) for a path to the destination LSR.

2. PCE 1 determines the destination, is not in domain 1.

3. PCE 1 sends a computation request to its parent PCE (PCE 5).

4. The parent PCE determines that the destination is in Domain 3. (See Section 5.5).

5. PCE 5 determines the likely domain paths according to the domain interconnectivity and TE capabilities between the domains. For example, three domain paths (S-BN11-BN21-D2-BN23-BN31-D, S-BN11-BN21-D2-BN24-BN32-D, and S-BN13-BN41-D4-BN42-BN33-D) are determined (assuming the link BN12-BN22 is not suitable for the requested path).

6. PCE 5 sends edge-to-edge path computation requests to PCE 2 which is responsible for Domain 2 (e.g., BN21-BN23 and BN21-BN24) and to PCE 4 for Domain 4 (e.g., BN41-BN42).

7. PCE 5 sends source-to-edge path computation requests to PCE 1 which is responsible for Domain 1 (e.g., S-BN11 and S-BN13).

8. PCE 5 sends edge-to-egress path computation requests to PCE3 which is responsible for Domain 3 (e.g., BN31-D, BN32-D, and BN33-D).

9. PCE 5 correlates all the computation responses from each child PCE, adds in the information about the inter-domain links, and applies any requested and locally configured policies.

10. PCE 5 then selects the optimal end-to-end multi-domain path that meets the policies and objective functions, and supplies the resulting path to PCE 1.

11. PCE 1 forwards the path to the PCC (the ingress LSR).

5.7 Hierarchical PCE Error Handling

In the event that a child PCE in a domain cannot find a suitible path to the egress. The child PCE should return the relevent error notifying the parent PCE. Depending on the error response the parent PCE can elect to:

- o Cancel the request and send the relevent response back to the intial child PCE requesting an end-to-end path.
- o Relax the contraints associated with the intial path request;
- o Select another candidate domain and send the path request to the child PCE responsible for the domain.

If the parent PCE does not recieve a response from a child PCE within an alloted time period. The parent PCE can either:

- o Send the path request to another child PCE in the same domain, if a secoundary child PCE exists;
- o Select another candidate domain and send the path request to the child PCE responsible for that domain.

5.8 Requirements for Hierarchical PCEP Protocol Extensions

This section lists the high-level requirements for extensions to the PCEP to support the hierachical PCE model.

[Editors Note: This section may be expanded as work progresses.]

5.8.1 PCEP Request Qualifiers

PCEP request (PCReq) messages are used by a PCC or a PCE to make a computation request or enquiry to a PCE. The requests are qualified so that the PCE knows what type of action is required.

Support of the H-PCE architecture will introduce two new qualifications as follows:

- o It must be possible for a child PCE to indicate that the request it sends to a parent PCE shold be satisfied by a domain sequence only, that is, not by a full end-to-end path. This allows the child PCE to initiate per-domain or backward recursive path computation.

- o A parent PCE needs to be able to ask a child PCE whether a particular node address (the destination of an end-to-end path) is present in the domain that the child PCE serves.

In PCEP, such request qualifications are carried as bit-flags in the RP object carried within the PCReq message.

5.8.2 Indication of H-PCE Capability

Although parent/child PCE relationships are likely configured, it assist network operations if the parent PCE is able to indicate to the child that it really is capable of acting as a parent PCE. This will help to trap misconfigurations.

A parent PCE needs a way to indicate that is capable of acting as a parent PCE, and should also be able to indicate the identity of the parent domain. This informaiton is most obviously carried in the Open Object within the Open message.

5.8.3 Intention to Utilize Parent PCE Capabilities

A PCE that is capable of acting as a parent PCE might not be configured or willing to act as the parent for a specific child PCE. This fact could be determined when the child sends a PCReq that requires parental activity (such as querying other child PCEs), and could result in a negative response in a PCEP Error (PCErr) message.

However, the expense of a poorly targetted PCReq can be avoided if the child PCE indicates that it might wish to use the parent as a parent (for example, on the Open message), and if the parent determines at that time whether it is willing to act as a parent to this child.

5.8.4 Communication of Domain Connectivity Information

Section 5.4 describes how the parent PCE needs a parent TED and indicates that the information might be supplied from the child PCEs in each domain. This requires a mechanism whereby information about inter-domain links can be supplied by a child PCE to a parent PCE, for example on a PCEP Notify (PCNtf) message.

The information that would be exchanged includes:

- Identifier of advertising child PCE
- Identifier of PCE's domain
- Identifier of the link
- TE properties of the link (metrics, bandwidth)
- Other properties of the link (technology-specific)
- Identifier of link end-points
- Identifier of adjacent domain

It may be desirable for this information to be periodically updated, for example, when available bandwidth changes. In this case, the parent PCE might be given the ability to configure thresholds in the child PCE to prevent flapping of information.

5.8.5 Domain Identifiers

Domain identifiers are already needed to allow a PCE to indicate which domains it serves, and to allow the representation of domains as abstract nodes in paths. The wider use of domains in the context of this work on H-PCE will require that domains can be identified in more places within objects in PCEP messages. This should pose no problems.

However, more attention may need to be applied to the precision of domian identifier definitions.

6. Hierarchical PCE Applicability

As per [RFC4655], PCE can inherently support inter-domain path computation for any definition of a domain as set out in Section 1.2.

Hierarchical PCE can be applied to inter-domain environments, including Antonymous Systems and IGP areas. The hierarchical PCE procedures make no distinction between, Antonymous Systems and IGP area applications, although it should be noted that the TED maintained by a parent PCE must be able to support the concept of child domains connected by inter-domain links or directly connected at boundary nodes (see Section 4).

This section sets out the applicability of hierarchical PCE to three environments:

- MPLS traffic engineering across multiple Autonomous Systems
- MPLS traffic engineering across multiple IGP areas
- GMPLS traffic engineering in the ASON architecture

6.1 Antonymous Systems and Areas

Networks are comprised of domains. A domain can be considered to be a collection of network elements within an AS or area that has a common sphere of address management or path computational responsibility.

As networks increase in size and complexity it may be required to introduce scaling methods to reduce the amount information flooded within the network and make the network more manageable. An IGP hierarchy is designed to improve IGP scalability by dividing the IGP domain into areas and limiting the flooding scope of topology information to within area boundaries. This restricts visibility of the area to routers in a single area. If a router needs to compute a route to destination located in another AS or area a method is required to compute a path across teh AS and area boundaries.

When an LSR within an AS or area needs to compute a path across an area or AS boundary it must also use an inter-AS computation technique. Hierachical PCE is equally applicable to computing inter-area and inter-AS MPLS and GMPLS paths across domain boundaries.

6.2 ASON Architecture

The International Telecommunications Union (ITU) defines the ASON architecture in [G-8080]. [G-7715] defines the routing architecture for ASON and introduces a hierarchical architecture. In this architecture, the Routing Areas (RAs) have a hierarchical relationship between different routing levels, which means a parent (or higher level) RA can contain multiple child RAs. The interconnectivity of the lower RAs is visible to the higher level RA. Note that the RA hierarchy can be recursive.

In the ASON framework, a path computation request is termed a Route Query. This query is executed before signaling is used to establish an LSP termed a Switched Connection (SC) or a Soft Permanent Connection (SPC). [G-7715-2] defines the requirements and architecture for the functions performed by Routing Controllers (RC) during the operation of remote route queries - an RC is synonymous with a PCE. For an end-to-end connection, the route may be computed by a single RC or multiple RCs in a collaborative manner (i.e., RC federations). In the case of RC federations, [G-7715-2] describes three styles during remote route query operation:

- o Step-by-step remote path computation
- o Hierarchical remote path computation
- o A combination of the above.

In a hierarchical ASON routing environment, a child RC may communicate with its parent RC (at the next higher level of the ASON routing hierarchy) to request the computation of an end-to-end path across several RAs. It does this using a route query message (known as the abstract message RI_QUERY). The corresponding parent RC may communicate with other child RCs that belong to other child RAs at the next lower hierarchical level. Thus, a parent RC can act as either a Route Query Requester or Route Query Responder.

It can be seen that the hierarchical PCE architecture fits the hierarchical ASON routing architecture well. It can be used to provide paths across subnetworks, and to determine end-to-end paths in networks constructed from multiple subnetworks or RAs.

When hierarchical PCE is applied to implement hierarchical remote path computation in [G-7715-2], it is very important for operators to understand the different terminology and implicit consistency between hierarchical PCE and [G-7715-2].

6.2.1 Implicit Consistency Between Hierarchical PCE and G.7715.2

This section highlights the correspondence between features of the hierarchical PCE architecture and the ASON routing architecture.

(1) RC (Routing Controller) and PCE (Path Computation Element)

[G-8080] describes the Routing Controller Component as an abstract entity, which is responsible for responding to requests for path (route) information and topology information. It can be implemented as a single entity, or as a distributed set of entities that make up a cooperative federation.

[RFC4655] describes PCE (Path Computation Element) is an entity (component, application, or network node) that is capable of computing a network path or route based on a network graph and applying computational constraints.

Therefore, in the ASON architecture, a PCE can be regarded as a realizations of the RC.

(2) Route Query Requester/Route Query Responder and PCC/PCE

[G-7715-2] describes the Route Query Requester as a Connection Controller or Routing Controller that sends a route query message to a Routing Controller requesting for one or more paths that satisfy a set of routing constraints. The Route Query Responder is a Routing Controller that performs path computation upon receipt of a route query message from a Route Query Requester, sending a response back at the end of the path computation.

In the context of ASON, a signaling controller initiates and processes signaling messages and closely coupled to a signaling protocol speaker. A routing controller makes routing decisions and is usually coupled to configuration entities and/or routing a protocol speaker.

It can be seen that a PCC corresponds to a Route Query Requester, and a PCE corresponds to a Route Query Responder. A PCE/RC can also act as a Route Query Requester sending requests to another Route Query Responder.

The PCEP path computation request (PCReq) and path computation reply (PCRep) messages between PCC and PCE correspond to the RI_QUERY and RI_UPDATE messages in [G-7715-2].

(3) Routing Area Hierarchy and Hierarchical Domain

The ASON routing hierarchy model is shown in Figure 6 of [G-7715] through an example that illustrates routing area levels. If the hierarchical remote path computation mechanism of [G-7715-2] is applied in this scenario, each routing area should have at least one RC for route query function and there is a parent RC for the child RCs in each routing area.

According to [G-8080], the parent RC has visibility of the structure of the lower level, so it knows the interconnectivity of the RAs in the lower level. Each child RC can compute edge-to-edge paths across its own child RA.

Thus, an RA corresponds to a domain, and the hierarchical relationship between RAs corresponds to the hierarchical relationship between domains. Furthermore, a parent PCE in a parent domain can be regarded as parent RC in a higher routing level, and a child PCE in a child domain can be regarded as child RC in a lower routing level.

6.2.2 Benefits of Hierarchical PCEs in ASON

RCs in an ASON environment can use the hierarchical PCE model to fully match the ASON hierarchical routing model, so the hierarchical PCE mechanisms can be applied to fully satisfy the architecture and requirements of [G-7715-2] without any changes. If the hierarchical PCE mechanism is applied in ASON, it can be used to determine end-to-end optimized paths across sub-networks and RAs before initiating signaling to create the connection. It can also improve the efficiency of connection setup to avoid crankback.

7. Management Considerations

General PCE management considerations are discussed in [RFC4655]. In the case of the hierarchical PCE architecture, there are additional management considerations.

The administrative entity responsible for the management of the parent PCEs must be determined. In the case of multi-domains (e.g., IGP areas or multiple ASes) within a single service provider network, the management responsibility for the parent PCE would most likely be handled by the service provider. In the case of multiple ASes within different service provider networks, it may be necessary for a third-party to manage the parent PCEs according to commercial and policy agreements from each of the participating service providers.

7.1 Control of Function and Policy

7.1.1 Child PCE

Support of the hierarchical procedure will be controlled by the management organization responsible for each child PCE. A child PCE must be configured with the address of its parent PCE in order for it to interact with its parent PCE. The child PCE must also be authorized to peer with the parent PCE.

7.1.2 Parent PCE

The parent PCE must only accept path computation requests from

authorized child PCEs. If a parent PCE receives requests from an unauthorized child PCE, the request should be dropped.

This means that a parent PCE must be configured with the identities and security credentials of all of its child PCEs, or there must be some form of shared secret that allows an unknown child PCE to be authorized by the parent PCE.

7.1.3 Policy Control

It may be necessary to maintain a policy module on the parent PCE [RFC5394]. This would allow the parent PCE to apply commercially relevant constraints such as SLAs, security, peering preferences, and dollar costs.

It may also be necessary for the parent PCE to limit end-to-end path selection by including or excluding specific domains based on commercial relationships, security implications, and reliability.

7.2 Information and Data Models

A PCEP MIB module is defined in [PCEP-MIB] that describes managed objects for modeling of PCEP communication. An additional PCEP MIB will be required to report parent PCE and child PCE information, including:

- Parent PCE configuration and status,
- Child PCE configuration and information,
- Notifications to indicate session changes between parent PCEs and child PCEs.
- Notification of parent PCE TED updates and changes.

7.3 Liveness Detection and Monitoring

The hierarchical procedure requires interaction with multiple PCEs. Once a child PCE requests an end-to-end path, a sequence of events occurs that requires interaction between the parent PCE and each child PCE. If a child PCE is not operational, and an alternate transit domain is not available, then a failure must be reported.

7.4 Verifying Correct Operation

Verifying the correct operation of a parent PCE can be performed by monitoring a set of parameters. The parent PCE implementation should provide the following parameters:

Parameters monitored by the parent PCE:

- Number of child PCE requests.
- Number of successful hierarchical PCE procedures completions on a per-PCE-peer basis.
- Number of hierarchical PCE procedure completion failures on a per-PCE-peer basis.
- Number of hierarchical PCE procedure requests from unauthorized child PCEs.

7.5. Impact on Network Operation

The hierarchical PCE procedure is a multiple-PCE path computation scheme. Subsequent requests to and from the child and parent PCEs do not differ from other path computation requests and should not have any significant impact on network operations.

8. Security Considerations

The hierarchical PCE procedure relies on PCEP and inherits the security requirements defined [RFC5440]. Any multi-domain operation necessarily involves the exchange of information across domain boundaries. This is bound to represent a significant security and confidentiality risk especially when the child domains are controlled by different commercial concerns.

The hierarchical PCE architecture makes use of PCE policy [RFC5394] and the security aspects of the PCE communication protocol documented in [RFC5440]. It is expected that the parent PCE will require all child PCEs to use full security when communicating with the parent and that security will be maintained by not supporting the discovery by a parent of child PCEs.

Confidentiality may be enhanced by the use of Path Keys [RFC5520].

Further considerations of the security issues related to inter-AS path computation see [RFC5376].

9. IANA Considerations

This document makes no requests for IANA action.

10. Acknowledgements

The authors would like to thank David Amzallag, Oscar Gonzalez de Diosm and Franz Rambach for their comments and suggestions.

11. References

11.1 Normative References

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

[RFC5152] Vasseur, JP., Ayyangar, A., and R. Zhang, "A Per-Domain Path Computation Method for Establishing Inter-Domain Traffic Engineering (TE) Label Switched Paths (LSPs)", RFC 5152, February 2008.

[RFC5394] Bryskin, I., Papadimitriou, D., Berger, L., and J. Ash, "Policy-Enabled Path Computation Framework", RFC 5394, December 2008.

[RFC5440] Ayyangar, A., Farrel, A., Oki, E., Atlas, A., Dolganow, A., Ikejiri, Y., Kumaki, K., Vasseur, J., and J. Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.

[RFC5441] Vasseur, J.P., Ed., "A Backward Recursive PCE-based Computation (BRPC) procedure to compute shortest inter-domain Traffic Engineering Label Switched Paths", RFC 5441, April 2009.

[RFC5520] Brandford, R., Vasseur J.P., and Farrel A., "Preserving Topology Confidentiality in Inter-Domain Path Computation Using a Key-Based Mechanism RFC5520, April 2009.

[G-8080] ITU-T Recommendation G.8080/Y.1304, Architecture for the automatically switched optical network (ASON).

[G-7715] ITU-T Recommendation G.7715 (2002), Architecture and Requirements for the Automatically Switched Optical Network (ASON).

[G-7715-2] ITU-T Recommendation G.7715.2 (2007), ASON routing architecture and requirements for remote route query.

11.2. Informative References

[RFC4726] Farrel, A., Vasseur, J., and A. Ayyangar, "A Framework for Inter-Domain Multiprotocol Label Switching Traffic Engineering", RFC 4726, November 2006.

[RFC4875] Aggarwal, R., Papadimitriou, D., and Yasukawa, S., "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, May 2007.

[RFC5152] Vasseur, JP., Ayyangar, A., and R. Zhang, "A Per-Domain Path Computation Method for Establishing Inter-Domain Traffic Engineering (TE) Label Switched Paths (LSPs)", RFC 5152, February 2008.

[RFC5316] Chen, M., Zhang, R., and X. Duan, "ISIS Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 5316, December 2008.

[RFC5376] Bitar, N., et al., "Inter-AS Requirements for the Path Computation Element Communication Protocol (PCECP)", RFC 5376, November 2008.

[RFC5392] Chen, M., Zhang, R., and X. Duan, "OSPF Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 5392, January 2009.

[RFC5541] Roux, J., Vasseur, J., and Y. Lee, "Encoding of Objective Functions in the Path Computation Element Communication Protocol (PCEP)", RFC5541, December 2008.

[PCEP-MIB] Stephan, E., K. Koushik, Q. Zhao, and D. King, "PCE communication protocol (PCEP) Management Information Base", Work in Progress, June 2010

12. Authors' Addresses

Daniel King
Old Dog Consulting
Email: daniel@olddog.co.uk

Adrian Farrel
Old Dog Consulting
Email: adrian@olddog.co.uk

Quintin Zhao
Huawei Technology
125 Nagog Technology Park
Acton, MA 01719
US
Email: qzhao@huawei.com

Fatai Zhang
Huawei Technologies
F3-5-B R&D Center, Huawei Base
Bantian, Longgang District
Shenzhen 518129 P.R.China
Email: zhangfatai@huawei.com

Network Working Group
Internet-Draft
Intended Status: Informational
Expires: 29 January 2013

D. King (Ed.)
Old Dog Consulting
A. Farrel (Ed.)
Old Dog Consulting
29 August 2012

The Application of the Path Computation Element Architecture to the Determination of a Sequence of Domains in MPLS and GMPLS

draft-ietf-pce-hierarchy-fwk-05.txt

Abstract

Computing optimum routes for Label Switched Paths (LSPs) across multiple domains in MPLS Traffic Engineering (MPLS-TE) and GMPLS networks presents a problem because no single point of path computation is aware of all of the links and resources in each domain. A solution may be achieved using the Path Computation Element (PCE) architecture.

Where the sequence of domains is known a priori, various techniques can be employed to derive an optimum path. If the domains are simply-connected, or if the preferred points of interconnection are also known, the Per-Domain Path Computation technique can be used. Where there are multiple connections between domains and there is no preference for the choice of points of interconnection, the Backward Recursive Path Computation Procedure (BRPC) can be used to derive an optimal path.

This document examines techniques to establish the optimum path when the sequence of domains is not known in advance. The document shows how the PCE architecture can be extended to allow the optimum sequence of domains to be selected, and the optimum end-to-end path to be derived through the use of a hierarchical relationship between domains.

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."

This Internet-Draft will expire on 29 August 2012.

Copyright Notice

Contents

1. Introduction

The capability to compute the routes of end-to-end inter-domain MPLS Traffic Engineering (TE) and GMPLS Label Switched Paths (LSPs) is expressed as requirements in [RFC4105] and [RFC4216]. This capability may be realized by a Path Computation Element (PCE). The PCE architecture is defined in [RFC4655]. The methods for establishing and controlling inter-domain MPLS-TE and GMPLS LSPs are documented in [RFC4726].

In this context, a domain can be defined as a separate

administrative, geographic, or switching environment within the network. A domain may be further defined as a zone of routing or computational ability. Under these definitions a domain might be categorized as an autonomous System (AS) or an Interior Gateway Protocol (IGP) area [RFC4726] and [RFC4655]. Domains are connected through ingress and egress boundary nodes (BNs). A more detailed definition is given in Section 1.2.

In a multi-domain environment, the determination of an end-to-end traffic engineered path is a problem because no single point of path computation is aware of all of the links and resources in each domain. PCEs can be used to compute end-to-end paths using a per-domain path computation technique [RFC5152]. Alternatively, the backward recursive path computation (BRPC) mechanism [RFC5441] allows multiple PCEs to collaborate in order to select an optimal end-to-end path that crosses multiple domains. Both mechanisms assume that the sequence of domains to be crossed between ingress and egress is known in advance.

This document examines techniques to establish the optimum path when the sequence of domains is not known in advance. It shows how the PCE architecture can be extended to allow the optimum sequence of domains to be selected, and the optimum end-to-end path to be derived.

The model described in this document introduces a hierarchical relationship between domains. It is applicable to environments with small groups of domains where visibility from the ingress Label Switching Router (LSR) is limited. Applying the hierarchical PCE model to large groups of domains such as the Internet, is not considered feasible or desirable, and is out of scope for this document.

This document does not specify any protocol extensions or enhancements. That work is left for future protocol specification documents. However, some assumptions are made about which protocols will be used to provide specific functions, and guidelines to future protocol developers are made in the form of requirements statements.

1.1 Problem Statement

Using a PCE to compute a path between nodes within a single domain is relatively straightforward. Computing an end-to-end path when the source and destination nodes are located in different domains requires co-operation between multiple PCEs, each responsible for its own domain.

Techniques for inter-domain path computation described so far ([RFC5152] and [RFC5441]) assume that the sequence of domains to be

crossed from source to destination is well known. No explanation is given (for example, in [RFC4655]) of how this sequence is generated or what criteria may be used for the selection of paths between domains. In small clusters of domains, such as simple cooperation between adjacent ISPs, this selection process is not complex. In more advanced deployments (such as optical networks constructed from multiple sub-domains, or in multi-AS environments) the choice of domains in the end-to-end domain sequence can be critical to the determination of an optimum end-to-end path.

1.2 Definition of a Domain

A domain is defined in [RFC4726] as 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. Wholly or partially overlapping domains are not within the scope of this document.

In the context of GMPLS, a particularly important example of a domain is the Automatically Switched Optical Network (ASON) subnetwork [G-8080]. In this case, a domain might be an ASON Routing Area [G-7715]. Furthermore, computation of an end-to-end path requires the selection of nodes and links within a routing area where some nodes may, in fact, be subnetworks. A PCE may perform the path computation function of an ASON Routing Controller as described in [G-7715-2]. See Section 5.2 for a further discussion of the applicability to the ASON architecture.

This document assumes that the selection of a sequence of domains for an end-to-end path is in some sense a hierarchical path computation problem. That is, where one mechanism is used to determine a path across a domain, a separate mechanism (or at least a separate set of paradigms) is used to determine the sequence of domains. The responsibility for the selection of domain interconnection can belong to either or both of the mechanisms.

1.3 Assumptions and Requirements

Networks are often constructed from multiple domains. These domains are often interconnected via multiple interconnect points. Its assumed that the sequence of domains for an end-to-end path is not always well known; that is, an application requesting end-to-end connectivity has no preference for, or no ability to specify, the sequence of domains to be crossed by the path.

The traffic engineering properties of a domain cannot be seen from outside the domain. Traffic engineering aggregation or abstraction,

hides information and can lead to failed path setup or the selection of suboptimal end-to-end paths [RFC4726]. The aggregation process may also have significant scaling issues for networks with many possible routes and multiple TE metrics. Flooding TE information breaks confidentiality and does not scale in the routing protocol. See Section 6 for a discussion of the concept of inter-domain traffic engineering information exchange known as BGP-TE.

The primary goal of this document is to define how to derive optimal end-to-end, multi-domain paths when the sequence of domains is not known in advance. The solution needs to be scalable and to maintain internal domain topology confidentiality while providing the optimal end-to-end path. It cannot rely on the exchange of TE information between domains, and for the confidentiality, scaling, and aggregation reasons just cited, it cannot utilize a computation element that has universal knowledge of TE properties and topology of all domains.

The sub-sections that follow set out the primary objectives and requirements to be satisfied by a PCE solution to multi-domain path computation.

1.3.1 Metric Objectives

The definition of optimality is dependent on policy, and is based on a single objective or a group objectives. An objective is expressed as an objective function [RFC5541] and may be specified on a path computation request. The following objective functions are identified in this document. They define how the path metrics and TE link qualities are manipulated during inter-domain path computation. The list is not proscriptive and may be expanded in other documents.

- o Minimize the cost of the path [RFC5541]
- o Select a path using links with the minimal load [RFC5541]
- o Select a path that leaves the maximum residual bandwidth [RFC5541]
- o Minimize aggregate bandwidth consumption [RFC5541]
- o Minimize the Load of the most loaded Link [RFC5541]
- o Minimize the Cumulative Cost of a set of paths [RFC5541]
- o Minimize or cap the number of domains crossed
- o Disallow domain re-entry

See Section 4.1 for further discussion of objective functions.

1.3.2 Diversity

1.3.2.1 Physical Diversity

Within a Carrier's carrier environment MPLS services may share common underlying equipment and resources, including optical fiber and

nodes. An MPLS service request may require a path for traffic that is physically disjointed from another service. Thus, if a physical link or node fails on one of the disjoint paths, not all traffic is lost.

1.3.2.2 Domain Diversity

A pair of paths are domain-diverse if they do not transit any of the same domains. A pair of paths that share a common ingress and egress are domain-diverse if they only share the same domains at the ingress and egress (the ingress and egress domains). Domain diversity may be maximized for a pair of paths by selecting paths that have the smallest number of shared domains. (Note that this is not the same as finding paths with the greatest number of distinct domains!)

Path computation should facilitate the selection of paths that share ingress and egress domains, but do not share any transit domains. This provides a way to reduce the risk of shared failure along any path, and automatically helps to ensure path diversity for most of the route of a pair of LSPs.

Thus, domain path selection should provide the capability to include or exclude specific domains and specific boundary nodes.

1.3.3 Existing Traffic Engineering Constraints

Any solution should take advantage of typical traffic engineering constraints (hop count, bandwidth, lambda continuity, path cost, etc.) to meet the service demands expressed in the path computation request [RFC4655].

1.3.4 Commercial Constraints

The solution should provide the capability to include commercially relevant constraints such as policy, SLAs, security, peering preferences, and monetary costs.

Additionally it may be necessary for the service provider to request that specific domains are included or excluded based on commercial relationships, security implications, and reliability.

1.3.5 Domain Confidentiality

A key requirement is the ability to maintain domain confidentiality when computing inter-domain end-to-end paths. It should be possible for local policy to require that a PCE not disclose to any other PCE the intra-domain paths it computes or the internal topology of the domain it serves. This requirement should have no impact in the optimality or quality of the end-to-end path that is derived.

1.3.6 Limiting Information Aggregation

In order to reduce processing overhead and to not sacrifice computational detail, there should be no requirement to aggregate or abstract traffic engineering link information.

1.3.7 Domain Interconnection Discovery

To support domain mesh topologies, the solution should allow the discovery and selection of domain inter-connections. Pre-configuration of preferred domain interconnections should also be supported for network operators that have bilateral agreement, and preference for the choice of points of interconnection.

1.4 Terminology

This document uses PCE terminology defined in [RFC4655], [RFC4726], and [RFC5440]. Additional terms are defined below.

Domain Path: The sequence of domains for a path.

Ingress Domain: The domain that includes the ingress LSR of a path.

Transit Domain: A domain that has an upstream and downstream neighbor domain for a specific path.

Egress Domain: The domain that includes the egress LSR of a path.

Boundary Nodes: Each Domain has entry LSRs and exit LSRs that could be Area Border Routers (ABRs) or Autonomous System Border Routers (ASBRs) depending on the type of domain. They are defined here more generically as Boundary Nodes (BNs).

Entry BN of domain(n): a BN connecting domain(n-1) to domain(n) on a path.

Exit BN of domain(n): a BN connecting domain(n) to domain(n+1) on a path.

Parent Domain: A domain higher up in a domain hierarchy such that it contains other domains (child domains) and potentially other links and nodes.

Child Domain: A domain lower in a domain hierarchy such that it has a parent domain.

Parent PCE: A PCE responsible for selecting a path across a parent domain and any number of child domains by coordinating with child PCEs and examining a topology map that shows domain inter-connectivity.

Child PCE: A PCE responsible for computing the path across one or more specific (child) domains. A child PCE maintains a relationship with at least one parent PCE.

OF: Objective Function: A set of one or more optimization criteria used for the computation of a single path (e.g., path cost minimization), or the synchronized computation of a set of paths (e.g., aggregate bandwidth consumption minimization). See [RFC4655] and [RFC5541].

2. Examination of Existing PCE Mechanisms

This section provides a brief overview of two existing PCE cooperation mechanisms called the per-domain path computation method, and the backward recursive path computation method. It describes the applicability of these methods to the multi-domain problem.

2.1 Per-Domain Path Computation

The per-domain path computation method for establishing inter-domain TE-LSPs [RFC5152] defines a technique whereby the path is computed during the signalling process on a per-domain basis. The entry BN of each domain is responsible for performing the path computation for the section of the LSP that crosses the domain, or for requesting that a PCE for that domain computes that piece of the path.

During per-domain path computation, each computation results in the best path across the domain to provide connectivity to the next domain in the domain sequence (usually indicated in signalling by an identifier of the next domain or the identity of the next entry BN).

Per-domain path computation may lead to sub-optimal end-to-end paths because the most optimal path in one domain may lead to the choice of an entry BN for the next domain that results in a very poor path across that next domain.

In the case that the domain path (in particular, the sequence of boundary nodes) is not known, the path computing entity must select an exit BN based on some determination of how to reach the destination that is outside the domain for which the path computing entity has computational responsibility. [RFC5152] suggest that this might be achieved using the IP shortest path as advertise by BGP. Note, however, that the existence of an IP forwarding path does not guarantee the presence of sufficient bandwidth, let alone an optimal TE path. Furthermore, in many GMPLS systems inter-domain IP routing will not be present. Thus, per-domain path computation may require a significant number of crankback routing attempts to establish even a sub-optimal path.

Note also that the path computing entities in each domain may have different computation capabilities, may run different path computation algorithms, and may apply different sets of constraints and optimization criteria, etc.

This can result in the end-to-end path being inconsistent and sub-optimal.

Per-domain path computation can suit simply-connected domains where the preferred points of interconnection are known.

2.2 Backward Recursive Path Computation

The Backward Recursive Path Computation (BRPC) [RFC5441] procedure involves cooperation and communication between PCEs in order to compute an optimal end-to-end path across multiple domains. The sequence of domains to be traversed can either be determined before or during the path computation. In the case where the sequence of domains is known, the ingress Path Computation Client (PCC) sends a path computation request to a PCE responsible for the ingress domain. This request is forwarded between PCEs, domain-by-domain, to a PCE responsible for the egress domain. The PCE in the egress domain creates a set of optimal paths from all of the domain entry BNs to the egress LSR. This set is represented as a tree of potential paths called a Virtual Shortest Path Tree (VSPT), and the PCE passes it back to the previous PCE on the domain path. As the VSPT is passed back toward the ingress domain, each PCE computes the optimal paths from its entry BNs to its exit BNs that connect to the rest of the tree. It adds these paths to the VSPT and passes the VSPT on until the PCE for the ingress domain is reached and computes paths from the ingress LSR to connect to the rest of the tree. The ingress PCE then selects the optimal end-to-end path from the tree, and returns the path to the initiating PCC.

BRPC may suit environments where multiple connections exist between domains and there is no preference for the choice of points of interconnection. It is best suited to scenarios where the domain path is known in advance, but can also be used when the domain path is not known.

2.2.1. Applicability of BRPC when the Domain Path is Not Known

As described above, BRPC can be used to determine an optimal inter-domain path when the domain sequence is known. Even when the sequence of domains is not known BRPC could be used as follows.

- The PCC sends a request to a PCE for the ingress domain (the ingress PCE).

- The ingress PCE sends the path computation request direct to a PCE responsible for the domain containing the destination node (the egress PCE).

- The egress PCE computes an egress VSPT and passes it to a PCE responsible for each of the adjacent (potentially upstream) domains.

- Each PCE in turn constructs a VSPT and passes it on to all of its neighboring PCEs.

- When the ingress PCE has received a VSPT from each of its neighboring domains it is able to select the optimum path.

Clearly this mechanism (which could be called path computation flooding) has significant scaling issues. It could be improved by the application of policy and filtering, but such mechanisms are not simple and would still leave scaling concerns.

3. Hierarchical PCE

In the hierarchical PCE architecture, a parent PCE maintains a domain topology map that contains the child domains (seen as vertices in the topology) and their interconnections (links in the topology). The parent PCE has no information about the content of the child domains; that is, the parent PCE does not know about the resource availability within the child domains, nor about the availability of connectivity across each domain because such knowledge would violate the confidentiality requirement and would either require flooding of full information to the parent (scaling issue) or would necessitate some form of aggregation. The parent PCE is aware of the TE capabilities of the interconnections between child domains as these interconnections are links in its own topology map.

Note that in the case that the domains are IGP areas, there is no link between the domains (the ABRs have a presence in both neighboring areas). The parent domain may choose to represent this in its TED as a virtual link that is unconstrained and has zero cost, but this is entirely an implementation issue.

Each child domain has at least one PCE capable of computing paths across the domain. These PCEs are known as child PCEs and have a relationship with the parent PCE. Each child PCE also knows the identity of the domains that neighbor its own domain. A child PCE only knows the topology of the domain that it serves and does not know the topology of other child domains. Child PCEs are also not aware of the general domain mesh connectivity (i.e., the domain

topology map) beyond the connectivity to the immediate neighbor domains of the domain it serves.

The parent PCE builds the domain topology map either from configuration or from information received from each child PCE. This tells it how the domains are interconnected including the TE properties of the domain interconnections. But the parent PCE does not know the contents of the child domains. Discovery of the domain topology and domain interconnections is discussed further in Section 4.3.

When a multi-domain path is needed, the ingress PCE sends a request to the parent PCE (using the path computation element protocol, PCEP [RFC5440]). The parent PCE selects a set of candidate domain paths based on the domain topology and the state of the inter-domain links. It then sends computation requests to the child PCEs responsible for each of the domains on the candidate domain paths. These requests may be sequential or parallel depending on implementation details.

Each child PCE computes a set of candidate path segments across its domain and sends the results to the parent PCE. The parent PCE uses this information to select path segments and concatenate them to derive the optimal end-to-end inter-domain path. The end-to-end path is then sent to the child PCE which received the initial path request and this child PCE passes the path on to the PCC that issued the original request.

Specific deployment and implementation scenarios are out of scope of this document. However the hierarchical PCE architecture described does support the function of parent PCE and child PCE being implemented as a common PCE.

4. Hierarchical PCE Procedures

4.1 Objective Functions and Policy

Deriving the optimal end-to-end domain path sequence is dependent on the policy applied during domain path computation. An Objective Function (OF) [RFC5541], or set of OFs, may be applied to define the policy being applied to the domain path computation.

The OF specifies the desired outcome of the computation. It does not describe the algorithm to use. When computing end-to-end inter-domain paths, required OFs may include (see Section 1.3.1):

- Minimum cost path
- Minimum load path
- Maximum residual bandwidth path
- Minimize aggregate bandwidth consumption

- Minimize or cap the number of transit domains
- Disallow domain re-entry

The objective function may be requested by the PCC, the ingress domain PCE (according to local policy), or applied by the parent PCE according to inter-domain policy.

More than one OF (or a composite OF) may be chosen to apply to a single computation provided they are not contradictory. Composite OFs may include weightings and preferences for the fulfilment of pieces of the desired outcome.

4.2 Maintaining Domain Confidentiality

Information about the content of child domains is not shared for scaling and confidentiality reasons. This means that a parent PCE is aware of the domain topology and the nature of the connections between domains, but is not aware of the content of the domains. Similarly, a child PCE cannot know the internal topology of another child domain. Child PCEs also do not know the general domain mesh connectivity, this information is only known by the parent PCE.

As described in the earlier sections of this document, PCEs can exchange path information in order to construct an end-to-end inter-domain path. Each per-domain path fragment reveals information about the topology and resource availability within a domain. Some management domains or ASes will not want to share this information outside of the domain (even with a trusted parent PCE). In order to conceal the information, a PCE may replace a path segment with a path-key [RFC5520]. This mechanism effectively hides the content of a segment of a path.

4.3 PCE Discovery

It is a simple matter for each child PCE to be configured with the address of its parent PCE. Typically, there will only be one or two parents of any child.

The parent PCE also needs to be aware of the child PCEs for all child domains that it can see. This information is most likely to be configured (as part of the administrative definition of each domain).

Discovery of the relationships between parent PCEs and child PCEs does not form part of the hierarchical PCE architecture. Mechanisms that rely on advertising or querying PCE locations across domain or provider boundaries are undesirable for security, scaling, commercial, and confidentiality reasons.

The parent PCE also needs to know the inter-domain connectivity. This information could be configured with suitable policy and commercial rules, or could be learned from the child PCEs as described in Section 4.4.

In order for the parent PCE to learn about domain interconnection the child PCE will report the identity of its neighbor domains. The IGP in each neighbor domain can advertise its inter-domain TE link capabilities [RFC5316], [RFC5392]. This information can be collected by the child PCEs and forwarded to the parent PCE, or the parent PCE could participate in the IGP in the child domains.

4.4 Parent Domain Traffic Engineering Database

The parent PCE maintains a domain topology map of the child domains and their interconnectivity. Where inter-domain connectivity is provided by TE links the capabilities of those links may also be known to the parent PCE. The parent PCE maintains a traffic engineering database (TED) for the parent domain in the same way that any PCE does.

The parent domain may just be the collection of child domains and their interconnectivity, may include details of the inter-domain TE links, and may contain nodes and links in its own right.

The mechanism for building the parent TED is likely to rely heavily on administrative configuration and commercial issues because the network was probably partitioned into domains specifically to address these issues.

In practice, certain information may be passed from the child domains to the parent PCE to help build the parent TED. In theory, the parent PCE could listen to the routing protocols in the child domains, but this would violate the confidentiality and scaling issues that may be responsible for the partition of the network into domains. So it is much more likely that a suitable solution will involve specific communication from an entity in the child domain (such as the child PCE) to convey the necessary information. As already mentioned, the "necessary information" relates to how the child domains are inter-connected. The topology and available resources within the child domain do not need to be communicated to the parent PCE: doing so would violate the PCE architecture. Mechanisms for reporting this information are described in the examples in Section 4.6 in abstract terms as "a child PCE reports its neighbor domain connectivity to its parent PCE"; the specifics of a solution are out of scope of this document, but the requirements are indicated in Section 4.8. See Section 6 for a brief discussion of BGP-TE.

In models such as ASON (see Section 5.2), it is possible to consider a separate instance of an IGP running within the parent domain where the participating protocol speakers are the nodes directly present in that domain and the PCEs (Routing Controllers) responsible for each of the child domains.

4.5 Determination of Destination Domain

The PCC asking for an inter-domain path computation is aware of the identity of the destination node by definition. If it knows the egress domain it can supply this information as part of the path computation request. However, if it does not know the egress domain this information must be known by the child PCE or known/determined by the parent PCE.

In some specialist topologies the parent PCE could determine the destination domain based on the destination address, for example from configuration. However, this is not appropriate for many multi-domain addressing scenarios. In IP-based multi-domain networks the parent PCE may be able to determine the destination domain by participating in inter-domain routing. Finally, the parent PCE could issue specific requests to the child PCEs to discover if they contain the destination node, but this has scaling implications.

For the purposes of this document, the precise mechanism of the discovery of the destination domain is left out of scope. Suffice to say that for each multi-domain path computation some mechanism will be required to determine the location of the destination.

4.6 Hierarchical PCE Examples

The following example describes the generic hierarchical domain topology. Figure 1 demonstrates four interconnected domains within a fifth, parent domain. Each domain contains a single PCE:

- Domain 1 is the ingress domain and child PCE 1 is able to compute paths within the domain. Its neighbors are Domain 2 and Domain 4. The domain also contains the source LSR (S) and three egress boundary nodes (BN11, BN12, and BN13).

- Domain 2 is served by child PCE 2. Its neighbors are Domain 1 and Domain 3. The domain also contains four boundary nodes (BN21, BN22, BN23, and BN24).

- Domain 3 is the egress domain and is served by child PCE 3. Its neighbors are Domain 2 and Domain 4. The domain also contains the destination LSR (D) and three ingress boundary nodes (BN31, BN32, and BN33).

- Domain 4 is served by child PCE 4. Its neighbors are Domain 2 and Domain 3. The domain also contains two boundary nodes (BN41 and BN42).

All of these domains are contained within Domain 5 which is served by the parent PCE (PCE 5).

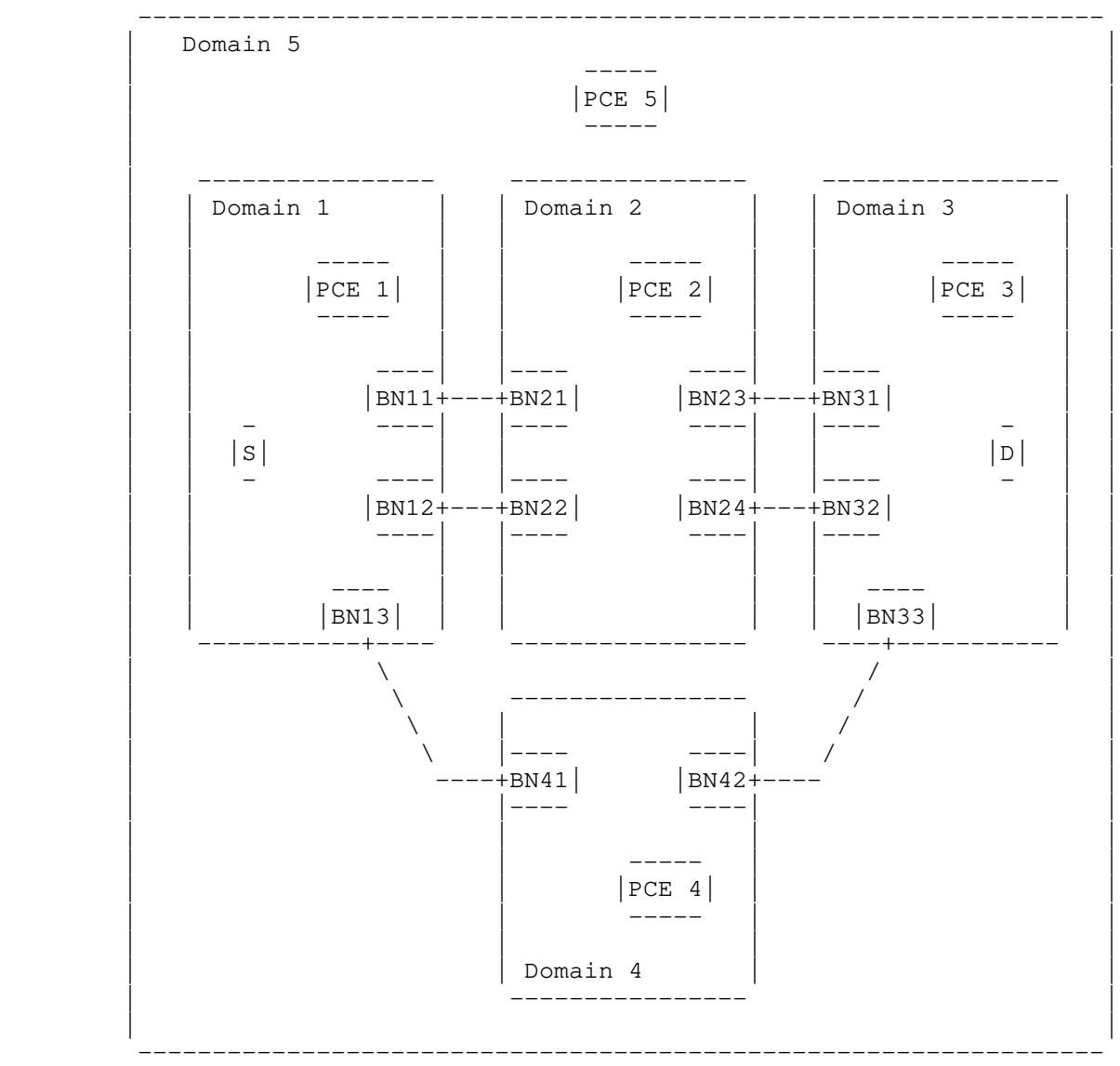


Figure 1 : Sample Hierarchical Domain Topology

Figure 2, shows the view of the domain topology as seen by the parent PCE (PCE 5). This view is an abstracted topology; PCE 5 is aware of domain connectivity, but not of the internal topology within each domain.

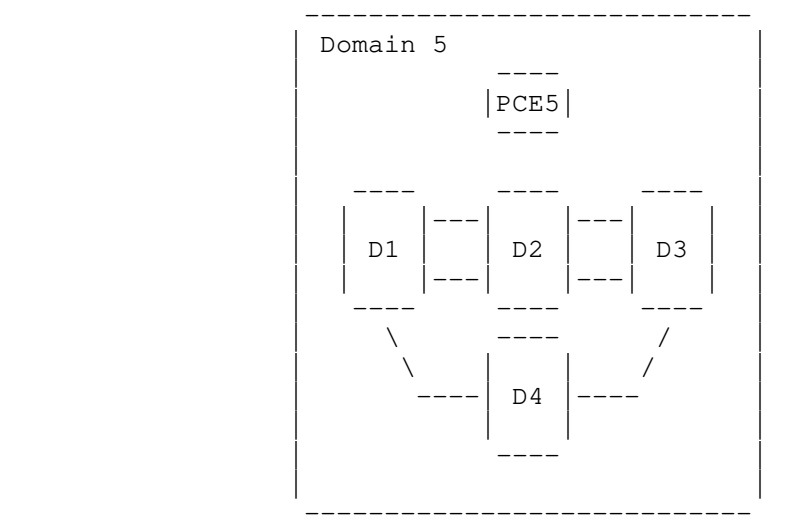


Figure 2 : Abstract Domain Topology as Seen by the Parent PCE

4.6.1 Hierarchical PCE Initial Information Exchange

Based on the Figure 1 topology, the following is an illustration of the initial hierarchical PCE information exchange.

1. Child PCE 1, the PCE responsible for Domain 1, is configured with the location of its parent PCE (PCE5).

2. Child PCE 1 establishes contact with its parent PCE. The parent applies policy to ensure that communication with PCE 1 is allowed.

3. Child PCE 1 listens to the IGP in its domain and learns its inter-domain connectivity. That is, it learns about the links BN11-BN21, BN12-BN22, and BN13-BN41.

4. Child PCE 1 reports its neighbor domain connectivity to its parent PCE.

5. Child PCE 1 reports any change in the resource availability on its inter-domain links to its parent PCE.

Each child PCE performs steps 1 through 5 so that the parent PCE can create a domain topology view as shown in Figure 2.

4.6.2 Hierarchical PCE End-to-End Path Computation Procedure

The procedure below is an example of a source PCC requesting an end-to-end path in a multi-domain environment. The topology is

represented in Figure 1. It is assumed that the each child PCE has connected to its parent PCE and exchanged the initial information required for the parent PCE to create its domain topology view as described in Section 4.6.1.

1. The source PCC (the ingress LSR in our example), sends a request to the PCE responsible for its domain (PCE 1) for a path to the destination LSR (D).

2. PCE 1 determines the destination is not in domain 1.

3. PCE 1 sends a computation request to its parent PCE (PCE 5).

4. The parent PCE determines that the destination is in Domain 3. (See Section 4.5).

5. PCE 5 determines the likely domain paths according to the domain interconnectivity and TE capabilities between the domains. For example, assuming that the link BN12-BN22 is not suitable for the requested path, three domain paths are determined:

   S-BN11-BN21-D2-BN23-BN31-D
   S-BN11-BN21-D2-BN24-BN32-D
   S-BN13-BN41-D4-BN42-BN33-D

6. PCE 5 sends edge-to-edge path computation requests to PCE 2 which is responsible for Domain 2 (i.e., BN21-to-BN23 and BN21-to-BN24), and to PCE 4 for Domain 4 (i.e., BN41-to-BN42).

7. PCE 5 sends source-to-edge path computation requests to PCE 1 which is responsible for Domain 1 (i.e., S-to-BN11 and S-to-BN13).

8. PCE 5 sends edge-to-egress path computation requests to PCE3 which is responsible for Domain 3 (i.e., BN31-to-D, BN32-to-D, and BN33-to-D).

9. PCE 5 correlates all the computation responses from each child PCE, adds in the information about the inter-domain links, and applies any requested and locally configured policies.

10. PCE 5 then selects the optimal end-to-end multi-domain path that meets the policies and objective functions, and supplies the resulting path to PCE 1.

11. PCE 1 forwards the path to the PCC (the ingress LSR).

Note that there is no requirement for steps 6, 7, and 8 to be carried out in parallel or in series. Indeed, they could be overlapped with step 5. This is an implementation issue.

4.7 Hierarchical PCE Error Handling

In the event that a child PCE in a domain cannot find a suitable path to the egress, the child PCE should return the relevant error to notify the parent PCE. Depending on the error response the parent PCE can elect to:

- Cancel the request and send the relevant response back to the initial child PCE that requested an end-to-end path;
- Relax some of the constraints associated with the initial path request;
- Select another candidate domain and send the path request to the child PCE responsible for the domain.

If the parent PCE does not receive a response from a child PCE within an allotted time period. The parent PCE can elect to:

- Cancel the request and send the relevant response back to the initial child PCE that requested an end-to-end path;
- Send the path request to another child PCE in the same domain, if a secondary child PCE exists;
- Select another candidate domain and send the path request to the child PCE responsible for that domain.

The parent PCE may also want to prune any unresponsive child PCE domain paths from the candidate set.

4.8 Requirements for Hierarchical PCEP Protocol Extensions

This section lists the high-level requirements for extensions to the PCEP to support the hierarchical PCE model. It is provided to offer guidance to PCEP protocol developers in designing a solution suitable for use in a hierarchical PCE framework.

4.8.1 PCEP Request Qualifiers

PCEP request (PCReq) messages are used by a PCC or a PCE to make a computation request or enquiry to a PCE. The requests are qualified so that the PCE knows what type of action is required.

Support of the hierarchical PCE architecture will introduce two new qualifications as follows:

- It must be possible for a child PCE to indicate that the response it receives from the parent PCE should consist of a domain sequence only (i.e., not a fully-specified end-to-end path). This allows the child PCE to initiate per-domain or backward recursive path computation.

- A parent PCE may need to be able to ask a child PCE whether a particular node address (the destination of an end-to-end path) is present in the domain that the child PCE serves.

In PCEP, such request qualifications are carried as bit-flags in the RP object within the PCReq message.

4.8.2 Indication of Hierarchical PCE Capability

Although parent/child PCE relationships are likely configured, it will assist network operations if the parent PCE is able to indicate to the child that it really is capable of acting as a parent PCE. This will help to trap misconfigurations.

In PCEP, such capabilities are carried in the Open Object within the Open message.

4.8.3 Intention to Utilize Parent PCE Capabilities

A PCE that is capable of acting as a parent PCE might not be configured or willing to act as the parent for a specific child PCE. This fact could be determined when the child sends a PCReq that requires parental activity (such as querying other child PCEs), and could result in a negative response in a PCEP Error (PCErr) message.

However, the expense of a poorly targeted PCReq can be avoided if the child PCE indicates that it might wish to use the parent-capable as a parent (for example, on the Open message), and if the parent-capable determines at that time whether it is willing to act as a parent to this child.

4.8.4 Communication of Domain Connectivity Information

Section 4.4 describes how the parent PCE needs a parent TED and indicates that the information might be supplied from the child PCEs in each domain. This requires a mechanism whereby information about inter-domain links can be supplied by a child PCE to a parent PCE, for example on a PCEP Notify (PCNtf) message.

The information that would be exchanged includes:

- Identifier of advertising child PCE
- Identifier of PCE's domain
- Identifier of the link
- TE properties of the link (metrics, bandwidth)
- Other properties of the link (technology-specific)
- Identifier of link end-points
- Identifier of adjacent domain

It may be desirable for this information to be periodically updated, for example, when available bandwidth changes. In this case, the parent PCE might be given the ability to configure thresholds in the child PCE to prevent flapping of information.

4.8.5 Domain Identifiers

Domain identifiers are already present in PCEP to allow a PCE to indicate which domains it serves, and to allow the representation of domains as abstract nodes in paths. The wider use of domains in the context of this work on hierarchical PCE will require that domains can be identified in more places within objects in PCEP messages. This should pose no problems.

However, more attention may need to be applied to the precision of domain identifier definitions to ensure that it is always possible to unambiguously identify a domain from its identifier. This work will be necessary in configuration, and also in protocol specifications (for example, an OSPF area identifier is sufficient within an Autonomous System, but becomes ambiguous in a path that crosses multiple Autonomous Systems).

5. Hierarchical PCE Applicability

As per [RFC4655], PCE can inherently support inter-domain path computation for any definition of a domain as set out in Section 1.2 of this document.

Hierarchical PCE can be applied to inter-domain environments, including autonomous Systems and IGP areas. The hierarchical PCE procedures make no distinction between, autonomous Systems and IGP area applications, although it should be noted that the TED maintained by a parent PCE must be able to support the concept of child domains connected by inter-domain links or directly connected at boundary nodes (see Section 3).

This section sets out the applicability of hierarchical PCE to three environments:

- MPLS traffic engineering across multiple Autonomous Systems
- MPLS traffic engineering across multiple IGP areas
- GMPLS traffic engineering in the ASON architecture

5.1 autonomous Systems and Areas

Networks are comprised of domains. A domain can be considered to be a collection of network elements within an AS or area that has a common sphere of address management or path computational responsibility.

As networks increase in size and complexity it may be required to introduce scaling methods to reduce the amount information flooded within the network and make the network more manageable. An IGP

hierarchy is designed to improve IGP scalability by dividing the IGP domain into areas and limiting the flooding scope of topology information to within area boundaries. This restricts a router's visibility to information about links and other routers within the single area. If a router needs to compute a route to destination located in another area, a method is required to compute a path across the area boundary.

When an LSR within an AS or area needs to compute a path across an area or AS boundary it must also use an inter-AS computation technique. Hierarchical PCE is equally applicable to computing inter-area and inter-AS MPLS and GMPLS paths across domain boundaries.

5.2 ASON Architecture

The International Telecommunications Union (ITU) defines the ASON architecture in [G-8080]. [G-7715] defines the routing architecture for ASON and introduces a hierarchical architecture. In this architecture, the Routing Areas (RAs) have a hierarchical relationship between different routing levels, which means a parent (or higher level) RA can contain multiple child RAs. The interconnectivity of the lower RAs is visible to the higher level RA. Note that the RA hierarchy can be recursive.

In the ASON framework, a path computation request is termed a Route Query. This query is executed before signaling is used to establish an LSP termed a Switched Connection (SC) or a Soft Permanent Connection (SPC). [G-7715-2] defines the requirements and architecture for the functions performed by Routing Controllers (RC) during the operation of remote route queries - an RC is synonymous with a PCE. For an end-to-end connection, the route may be computed by a single RC or multiple RCs in a collaborative manner (i.e., RC federations). In the case of RC federations, [G-7715-2] describes three styles during remote route query operation:

- Step-by-step remote path computation
- Hierarchical remote path computation
- A combination of the above.

In a hierarchical ASON routing environment, a child RC may communicate with its parent RC (at the next higher level of the ASON routing hierarchy) to request the computation of an end-to-end path across several RAs. It does this using a route query message (known as the abstract message RI_QUERY). The corresponding parent RC may communicate with other child RCs that belong to other child RAs at the next lower hierarchical level. Thus, a parent RC can act as either a Route Query Requester or Route Query Responder.

It can be seen that the hierarchical PCE architecture fits the hierarchical ASON routing architecture well. It can be used to provide paths across subnetworks, and to determine end-to-end paths in networks constructed from multiple subnetworks or RAs.

When hierarchical PCE is applied to implement hierarchical remote path computation in [G-7715-2], it is very important for operators to understand the different terminology and implicit consistency between hierarchical PCE and [G-7715-2].

5.2.1 Implicit Consistency Between Hierarchical PCE and G.7715.2

This section highlights the correspondence between features of the hierarchical PCE architecture and the ASON routing architecture.

(1) RC (Routing Controller) and PCE (Path Computation Element)

[G-8080] describes the Routing Controller component as an abstract entity, which is responsible for responding to requests for path (route) information and topology information. It can be implemented as a single entity, or as a distributed set of entities that make up a cooperative federation.

[RFC4655] describes PCE (Path Computation Element) is an entity (component, application, or network node) that is capable of computing a network path or route based on a network graph and applying computational constraints.

Therefore, in the ASON architecture, a PCE can be regarded as a realizations of the RC.

(2) Route Query Requester/Route Query Responder and PCC/PCE

[G-7715-2] describes the Route Query Requester as a Connection Controller or Routing Controller that sends a route query message to a Routing Controller requesting one or more paths that satisfy a set of routing constraints. The Route Query Responder is a Routing Controller that performs path computation upon receipt of a route query message from a Route Query Requester, sending a response back at the end of the path computation.

In the context of ASON, a Signaling Controller initiates and processes signaling messages and is closely coupled to a Signaling Protocol Speaker. A Routing Controller makes routing decisions and is usually coupled to configuration entities and/or a Routing Protocol Speaker.

It can be seen that a PCC corresponds to a Route Query Requester, and a PCE corresponds to a Route Query Responder. A PCE/RC can

also act as a Route Query Requester sending requests to another Route Query Responder.

The PCEP path computation request (PCReq) and path computation reply (PCRep) messages between PCC and PCE correspond to the RI_QUERY and RI_UPDATE messages in [G-7715-2].

(3) Routing Area Hierarchy and Hierarchical Domain

The ASON routing hierarchy model is shown in Figure 6 of [G-7715] through an example that illustrates routing area levels. If the hierarchical remote path computation mechanism of [G-7715-2] is applied in this scenario, each routing area should have at least one RC for route query function and there is a parent RC for the child RCs in each routing area.

According to [G-8080], the parent RC has visibility of the structure of the lower level, so it knows the interconnectivity of the RAs in the lower level. Each child RC can compute edge-to-edge paths across its own child RA.

Thus, an RA corresponds to a domain in the PCE architecture, and the hierarchical relationship between RAs corresponds to the hierarchical relationship between domains in the hierarchical PCE architecture. Furthermore, a parent PCE in a parent domain can be regarded as parent RC in a higher routing level, and a child PCE in a child domain can be regarded as child RC in a lower routing level.

5.2.2 Benefits of Hierarchical PCEs in ASON

RCs in an ASON environment can use the hierarchical PCE model to fully match the ASON hierarchical routing model, so the hierarchical PCE mechanisms can be applied to fully satisfy the architecture and requirements of [G-7715-2] without any changes. If the hierarchical PCE mechanism is applied in ASON, it can be used to determine end-to-end optimized paths across sub-networks and RAs before initiating signaling to create the connection. It can also improve the efficiency of connection setup to avoid crankback.

6. A Note on BGP-TE

The concept of exchange of TE information between Autonomous Systems (ASes) is discussed in [BGP-TE]. The information exchanged in this way could be the full TE information from the AS, an aggregation of that information, or a representation of the potential connectivity across the AS. Furthermore, that information could be updated frequently (for example, for every new LSP that is set up across the AS) or only at threshold-crossing events.

There are a number of discussion points associated with the use of [BGP-TE] concerning the volume of information, the rate of churn of information, the confidentiality of information, the accuracy of aggregated or potential-connectivity information, and the processing required to generate aggregated information. The PCE architecture and the architecture enabled by [BGP-TE] make different assumptions about the operational objectives of the networks, and this document does not attempt to make one of the approaches "right" and the other "wrong". Instead, this work assumes that a decision has been made to utilize the PCE architecture.

6.1 Use of BGP for TED Synchronization

Indeed, [BGP-TE] may have some uses within the PCE model. For example, [BGP-TE] could be used as a "northbound" TE advertisement such that a PCE does not need to listen to an IGP in its domain, but has its TED populated by messages received (for example) from a Route Reflector. Furthermore, the inter-domain connectivity and connectivity capabilities that is required information for a parent PCE could be obtained as a filtered subset of the information available in [BGP-TE]. This scenario is discussed further in [PCE-AREA-AS].

7. Management Considerations

General PCE management considerations are discussed in [RFC4655]. In the case of the hierarchical PCE architecture, there are additional management considerations.

The administrative entity responsible for the management of the parent PCEs must be determined. In the case of multi-domains (e.g., IGP areas or multiple ASes) within a single service provider network, the management responsibility for the parent PCE would most likely be handled by the service provider. In the case of multiple ASes within different service provider networks, it may be necessary for a third-party to manage the parent PCEs according to commercial and policy agreements from each of the participating service providers.

7.1 Control of Function and Policy

7.1.1 Child PCE

Support of the hierarchical procedure will be controlled by the management organization responsible for each child PCE. A child PCE must be configured with the address of its parent PCE in order for it to interact with its parent PCE. The child PCE must also be authorized to peer with the parent PCE.

7.1.2 Parent PCE

The parent PCE must only accept path computation requests from authorized child PCEs. If a parent PCE receives requests from an unauthorized child PCE, the request should be dropped.

This means that a parent PCE must be configured with the identities and security credentials of all of its child PCEs, or there must be some form of shared secret that allows an unknown child PCE to be authorized by the parent PCE.

7.1.3 Policy Control

It may be necessary to maintain a policy module on the parent PCE [RFC5394]. This would allow the parent PCE to apply commercially relevant constraints such as SLAs, security, peering preferences, and monetary costs.

It may also be necessary for the parent PCE to limit end-to-end path selection by including or excluding specific domains based on commercial relationships, security implications, and reliability.

7.2 Information and Data Models

A PCEP MIB module is defined in [PCEP-MIB] that describes managed objects for modeling of PCEP communication. An additional PCEP MIB will be required to report parent PCE and child PCE information, including:

- Parent PCE configuration and status,
- Child PCE configuration and information,
- Notifications to indicate session changes between parent PCEs and child PCEs.
- Notification of parent PCE TED updates and changes.

7.3 Liveness Detection and Monitoring

The hierarchical procedure requires interaction with multiple PCEs. Once a child PCE requests an end-to-end path, a sequence of events occurs that requires interaction between the parent PCE and each child PCE. If a child PCE is not operational, and an alternate transit domain is not available, then a failure must be reported.

7.4 Verifying Correct Operation

Verifying the correct operation of a parent PCE can be performed by

monitoring a set of parameters. The parent PCE implementation should provide the following parameters monitored by the parent PCE:

- Number of child PCE requests.
- Number of successful hierarchical PCE procedures completions on a per-PCE-peer basis.
- Number of hierarchical PCE procedure completion failures on a per-PCE-peer basis.
- Number of hierarchical PCE procedure requests from unauthorized child PCEs.

7.5. Impact on Network Operation

The hierarchical PCE procedure is a multiple-PCE path computation scheme. Subsequent requests to and from the child and parent PCEs do not differ from other path computation requests and should not have any significant impact on network operations.

8. Security Considerations

The hierarchical PCE procedure relies on PCEP and inherits the security requirements defined [RFC5440]. As noted in Section 7, there is a security relationship between child and parent PCEs. This relationship, like any PCEP relationship assumes pre-configuration of identities, authority, and keys, or can operate through any key distribution mechanism outside the scope of PCEP. As PCEP operates over TCP, it may make use of any TCP security mechanism.

The hierarchical PCE architecture makes use of PCE policy [RFC5394] and the security aspects of the PCE communication protocol documented in [RFC5440]. It is expected that the parent PCE will require all child PCEs to use full security when communicating with the parent and that security will be maintained by not supporting the discovery by a parent of child PCEs.

PCE operation also relies on information used to build the TED. Attacks on a PCE system may be achieved by falsifying or impeding this flow of information. The child PCE TEDs are constructed as described in [RFC4655] and are unchanged in this document: if the PCE listens to the IGP for this information, then normal IGP security measures may be applied, and it should be noted that an IGP routing system is generally assumed to be a trusted domain such that router subversion is not a risk. The parent PCE TED is constructed as described in this document and may involve:

- multiple parent-child relationships using PCEP (as already described)

- the parent PCE listening to child domain IGPs (with the same security features as a child PCE listening to its IGP)

- an external mechanism (such as [BGP-TE]) which will need to be authorized and secured.

Any multi-domain operation necessarily involves the exchange of information across domain boundaries. This is bound to represent a significant security and confidentiality risk especially when the child domains are controlled by different commercial concerns. PCEP allows individual PCEs to maintain confidentiality of their domain path information using Path Keys [RFC5520], and the hierarchical PCE architecture is specifically designed to enable as much isolation of domain topology and capabilities information as is possible.

Further considerations of the security issues related to inter-AS path computation see [RFC5376].

9. IANA Considerations

This document makes no requests for IANA action.

10. Acknowledgements

The authors would like to thank David Amzallag, Oscar Gonzalez de Dios, Franz Rambach, Ramon Casellas, Olivier Dugeon, Filippo Cugini, Dhruv Dhody and Julien Meuric for their comments and suggestions.

11. References

11.1 Normative References

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

[RFC5152] Vasseur, JP., Ayyangar, A., and R. Zhang, "A Per-Domain Path Computation Method for Establishing Inter-Domain Traffic Engineering (TE) Label Switched Paths (LSPs)", RFC 5152, February 2008.

[RFC5394] Bryskin, I., Papadimitriou, D., Berger, L., and J. Ash, "Policy-Enabled Path Computation Framework", RFC 5394, December 2008.

[RFC5440] Ayyangar, A., Farrel, A., Oki, E., Atlas, A., Dolganow, A., Ikejiri, Y., Kumaki, K., Vasseur, J., and J. Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.

[RFC5441] Vasseur, J.P., Ed., "A Backward Recursive PCE-based Computation (BRPC) procedure to compute shortest inter-domain Traffic Engineering Label Switched Paths", RFC 5441, April 2009.

[RFC5520] Brandford, R., Vasseur J.P., and Farrel A., "Preserving Topology Confidentiality in Inter-Domain Path Computation Using a Key-Based Mechanism RFC5520, April 2009.

11.2. Informative References

[RFC4105] Le Roux, JL., Vasseur, J., Boyle, J., "Requirements for Inter-Area MPLS Traffic Engineering", RFC 4105, June 2005.

[RFC4216] Zhang, R., and Vasseur, J., "MPLS Inter-Autonomous System (AS) Traffic Engineering (TE) Requirements", RFC 4216, November 2005.

[RFC4726] Farrel, A., Vasseur, J., Ayyangar, A., "A Framework for Inter-Domain Multiprotocol Label Switching Traffic Engineering", RFC 4726, November 2006.

[RFC5152] Vasseur, JP., Ayyangar, A., Zhang, R., "A Per-Domain Path Computation Method for Establishing Inter-Domain Traffic Engineering (TE) Label Switched Paths (LSPs)", RFC 5152, February 2008.

[RFC5316] Chen, M., Zhang, R., Duan, X., "ISIS Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 5316, December 2008.

[RFC5376] Bitar, N., et al., "Inter-AS Requirements for the Path Computation Element Communication Protocol (PCECP)", RFC 5376, November 2008.

[RFC5392] Chen, M., Zhang, R., Duan, X., "OSPF Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 5392, January 2009.

[RFC5541] Le Roux, J., Vasseur, J., Lee, Y., "Encoding of Objective Functions in the Path Computation Element Communication Protocol (PCEP)", RFC5541, December 2008.

[G-8080] ITU-T Recommendation G.8080/Y.1304, Architecture for the automatically switched optical network (ASON).

[G-7715] ITU-T Recommendation G.7715 (2002), Architecture and Requirements for the Automatically Switched Optical Network (ASON).

[G-7715-2] ITU-T Recommendation G.7715.2 (2007), ASON routing architecture and requirements for remote route query.

[BGP-TE] Gredler, H., Medved, J, Farrel, A. Previdi, S., "North-Bound Distribution of Link-State and TE Information using BGP", draft-gredler-idr-ls-distribution, work in progress.

[PCE-AREA-AS] King, D., Meuric, J., Dugeon, O., Zhao, Q., Gonzalez de Dios, O., "Applicability of the Path Computation Element to Inter-Area and Inter-AS MPLS and GMPLS Traffic Engineering", draft-ietf-pce-inter-area-as-applicability, work in progress.

[PCEP-MIB] Stephan, E., Koushik, K., Zhao, Q., King, D., "PCE Communication Protocol (PCEP) Management Information Base", work in progress.

12. Authors' Addresses

Daniel King
Old Dog Consulting
UK

Email: daniel@olddog.co.uk

Adrian Farrel
Old Dog Consulting
UK

Email: adrian@olddog.co.uk

Quintin Zhao
Huawei Technology
125 Nagog Technology Park
Acton, MA 01719
US

Email: qzhao@huawei.com

Fatai Zhang
Huawei Technologies
F3-5-B R&D Center, Huawei Base
Bantian, Longgang District
Shenzhen 518129 P.R.China

Email: zhangfatai@huawei.com

Pce Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 30, 2012

Q. Zhao
D. Dhody
Huawei Technology
Z. Ali
T. Saad
S. Sivabalan
Cisco Systems, Inc.
D. King
Old Dog Consulting
R. Casellas
CTTC - Centre Tecnologic de
Telecomunicacions de Catalunya

October 28, 2011

PCE-based Computation Procedure To Compute Shortest Constrained P2MP Inter-domain Traffic Engineering Label Switched Paths
draft-ietf-pce-pcep-inter-domain-p2mp-procedures-01

Abstract

The ability to compute paths for constrained point-to-multipoint (P2MP) Traffic Engineering Label Switched Paths (TE LSPs) across multiple domains has been identified as a key requirement for the deployment of P2MP services in MPLS and GMPLS networks. The Path Computation Element (PCE) has been recognized as an appropriate technology for the determination of inter-domain paths of P2MP TE LSPs.

This document describes the procedures and extensions to the PCE communication Protocol (PCEP) to handle requests and responses for the computation of inter-domain paths for P2MP TE 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). 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 30, 2012.

Copyright Notice

Table of Contents

1. Introduction

Multicast services are increasingly in demand for high-capacity applications such as multicast Virtual Private Networks (VPNs), IP-television (IPTV) which may be on-demand or streamed, and content-rich media distribution (for example, software distribution, financial streaming, or data-sharing). The ability to compute constrained Traffic Engineering Label Switched Paths (TE LSPs) for point-to-multipoint (P2MP) LSPs in Multiprotocol Label Switching (MPLS) and Generalized MPLS (GMPLS) networks across multiple domains is becoming important. A domain can be defined as a collection of network elements within a common sphere of address management or path computational responsibility such as an IGP area or an Autonomous Systems.

The applicability of the Path Computation Element (PCE) [RFC4655] for the computation of such paths is discussed in [RFC5671], and the requirements placed on the PCE communications Protocol (PCEP) for this are given in [RFC5862].

This document describes how multiple PCE techniques can be combined to address the requirements. These mechanisms include the use of the per-domain path computation technique specified in [RFC5152], extensions to the backward recursive path computation (BRPC) technique specified in [RFC5441] for P2MP LSP path computation in an inter-domain environment, and a new procedure for core-tree based path computation defined in this document. These three mechanisms are suitable for different environments (topologies, administrative domains, policies, service requirements, etc.) and can also be effectively combined.

1.1. Requirements Language

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

2. Terminology

Terminology used in this document is consistent with the related MPLS/GMPLS and PCE documents [RFC4461], [RFC4655], [RFC4875], [RFC5376], [RFC5440], [RFC5441], [RFC5671] and [RFC5862].

ABR: Area Border Router. Router used to connect two IGP domains (areas in OSPF or levels in IS-IS).

ASBR: Autonomous System Border Router. Router used to connect together ASes of the same or different Service Providers via one or

more Inter-AS links.

Boundary Node (BN): a boundary node is either an ABR in the context of inter-area Traffic Engineering or an ASBR in the context of inter-AS Traffic Engineering.

Core Tree: the core tree is a P2MP tree where the root is the ingress LSR, and the leaf nodes are the entry BNs of the leaf domains.

Destination: The lead Nodes can be in Root Domain, Transit Domain and Leaf Domain.

Entry BN of domain(n): a BN connecting domain(n-1) to domain(n) along a determined sequence of domains.

Exit BN of domain(n): a BN connecting domain(n) to domain(n+1) along a determined sequence of domains.

Inter-AS TE LSP: a TE LSP that crosses an AS boundary.

Inter-area TE LSP: a TE LSP that crosses an IGP area boundary.

Leaf Domain: a domain that does not have a downstream neighbor domain. Note that, with this definition, a domain with one or more leaf nodes is not necessarily a leaf domain.

Leaf Boundary Nodes: the entry boundary node in the leaf domain.

Leaf Nodes: the LSR which is the P2MP LSP's final.

LSR: Label Switching Router.

LSP: Label Switched Path.

OF: Objective Function. A set of one or more optimization criterion (criteria) used for the computation of paths either for single or for synchronized requests (e.g. path cost minimization), or the synchronized computation of a set of paths (e.g. aggregate bandwidth consumption minimization, etc.). See [RFC4655] and [RFC5441].

P2MP LSP Path Tree: A set of LSRs and TE links that comprise the path of a P2MP TE LSP from its ingress LSR to all of its egress LSRs.

Path Domain Sequence: The known sequence of domains for a path between root and leaf.

Path Domain Tree: The tree formed by the domains that the P2MP path crosses, where the source (ingress) domain is the root domain.

PCC: Path Computation Client. Any client application requesting a path computation to be performed by the Path Computation Element.

PCE (Path Computation Element): an entity (component, application or network node) that is capable of computing a network path or route based on a network graph and applying computational constraints.

P2MP LSP Path Tree: A set of LSRs and TE links that comprise the path of a P2MP TE LSP from its ingress LSR to all of its egress LSRs.

Path Domain Sequence: the known sequence of domains for a path between the root node and a leaf node.

PCE Sequence: the known sequence of PCEs for calculating a path between the root node and a leaf node.

PCE Topology Tree: a list of PCE Sequences which has all the PCE Sequence for each path of the P2MP LSP path tree.

PCE(i): a PCE that performs path computations for domain(i).

Root Boundary Node: the egress LSR from the root domain on the path of the P2MP LSP.

Root Domain: the domain that includes the ingress (root) LSR.

TED: Traffic Engineering Database.

Transit/branch Domain: a domain that has an upstream and one or more downstream neighbour domain.

VSPT: Virtual Shortest Path Tree [RFC5441].

3. Problem Statement

The Path Computation Element (PCE) defined in [RFC4655] is an entity that is capable of computing a network path or route based on a network graph, and applying computational constraints. A Path Computation Client (PCC) may make requests to a PCE for paths to be computed.

[RFC4875] describes how to set up P2MP TE LSPs for use in MPLS and GMPLS networks. The PCE is identified as a suitable application for the computation of paths for P2MP TE LSPs [RFC5671].

[RFC5441] specifies a procedure relying on the use of multiple PCEs to compute (P2P) inter-domain constrained shortest paths across a predetermined sequence of domains, using a backward recursive path

computation technique. The technique can be combined with the use of path keys [RFC5520] to preserve confidentiality across domains, which is sometimes required when domains are managed by different Service Providers.

The PCE communication Protocol (PCEP) [RFC5440] is extended for point-to-multipoint(P2MP) path computation requests and in [RFC6006]. However, that specification does not provide all the necessary mechanisms to request the computation of inter-domain P2MP TE LSPs.

As discussed in [RFC4461], a P2MP tree is a graphical representation of all TE links that are committed for a particular P2MP LSP. In other words, a P2MP tree is a representation of the corresponding P2MP tunnel on the TE network topology. A sub-tree is a part of the P2MP tree describing how the root or an intermediate P2MP LSPs minimizes packet duplication when P2P TE sub-LSPs traverse common links. As described in [RFC5671] the computation of a P2MP tree requires three major pieces of information. The first is the path from the ingress LSR of a P2MP LSP to each of the egress LSRs, the second is the traffic engineering related parameters, and the third is the branch capability information.

Generally, an inter-domain P2MP tree (i.e., a P2MP tree with source and at least one destination residing in different domains) is particularly difficult to compute even for a distributed PCE architecture. For instance, while the BRPC recursive path computation may be well-suited for P2P paths, P2MP path computation involves multiple branching path segments from the source to the multiple destinations. As such, inter-domain P2MP path computation may result in a plurality of per-domain path options that may be difficult to coordinate efficiently and effectively between domains. That is, when one or more domains have multiple ingress and/or egress border nodes, there is currently no known technique for one domain to determine which border routers another domain will utilize for the inter-domain P2MP tree, and no way to limit the computation of the P2MP tree to those utilized border nodes.

A trivial solution to the computation of inter-domain P2MP tree would be to compute shortest inter-domain P2P paths from source to each destination and then combine them to generate an inter-domain, shortest-path-to-destination P2MP tree. This solution, however, cannot be used to trade cost to destination for overall tree cost (i.e., it cannot produce a MCT tree) and in the context of inter-domain P2MP LSPs it cannot be used to reduce the number of domain border nodes that are transited.

Computing P2P LSPs individually is not an acceptable solution for computing a P2MP tree. Even per domain path computation [RFC5152]

can be used to compute P2P multi-domain paths, but it does not guarantee to find the optimal path which crosses multiple domains. Furthermore, constructing a P2MP tree from individual source to leaf P2P LSPs does not guarantee to produce a least-cost tree. This approach may also be considered to have scaling issues during LSP setup. That is, the LSP to each leaf is signaled separately, and each border node must perform path computation for each leaf.

P2MP Minimum Cost Tree (MCT), i.e. one which guarantees the least cost resulting tree, is an NP-complete problem. Moreover, adding and/or removing a single destination to/from the tree may result in an entirely different tree. In this case, frequent MCT path computation requests may prove computationally intensive, and the resulting frequent tunnel reconfiguration may even cause network destabilization. There are several heuristic algorithms presented in the literature that approximate the result within polynomial time that are applicable within the context of a single-domain.

This document presents a solution, and procedures and extensions to PCEP to support P2MP inter-domain path computation.

4. Assumptions

It is assumed that, due to deployment and commercial limitations (e.g., inter-AS peering agreements), the sequence of domains for a path (the path domain tree) will be known in advance.

[DOMAIN-SEQ] descibes the use of domain path tree in P2MP scenarios. In the figure below, the P2MP tree spans 6 domains, with D1 being the root domain. The corresponding domain sequences which are assumed known would be: D1-D3-D6, D1-D3-D5 and D1-D2-D4.

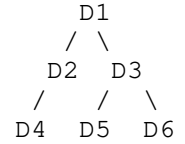


Figure 1: Domain Sequence Tree

The examples and scenarios used in this document are also based on the following assumptions:

- PCC is either aware of the domain sequence for each of the P2MP destination as described in [DOMAIN-SEQ] or PCE sequence (i.e. PCE that serves each domain in the path domain tree). The set of PCEs and their relationships is propagated to each PCE during the

first exchange of path computation requests; [Editors note - this assumption needs to be more explicit.]

- Each PCE knows about any leaf LSRs in the domain it serves;

- The boundary nodes to use on the LSP are pre-determined. [Editors Note - In this version of the document we do not consider multi-homed domains.]

Additional assumptions are documented in [RFC5441] and will not be repeated here.

5. Requirements

This section summarizes the requirements specific to computing inter-domain P2MP paths. In these requirements we note that the actual computation times by any PCE implementation are outside the scope of this document, but we observe that reducing the complexity of the required computations has a beneficial effect on the computation time regardless of implementation. Additionally, reducing the number of message exchanges and the amount of information exchanged will reduce the overall computation time for the entire P2MP tree. We refer to the "Complexity of the computation" as the impact on these aspects of path computation time as various parameters of the topology and the P2MP LSP are changed.

Its also important that the solution preserves confidentiality across domains, which is required when domains are managed by different Service Providers.

Other than the requirements specified in [RFC5376], a number of requirements specific to P2MP are detailed below:

1. The computed P2MP LSP should be optimal when only considering the paths among the BNs.

2. Grafting and pruning of multicast destinations in a domain should have no impact on other domains and on the paths among BNs.

3. The complexity of the computation for each sub-tree within each domain should be dependent only on the topology of the domain and it should be independent of the domain sequence.

4. The number of PCEP request and reply messages should be independent of the number of multicast destinations in each domain.

5. Specifying the domain entry and exit nodes.

6. Specifying which nodes should be used as branch nodes.

7. Reoptimization of existing sub-trees.

8. Computation of P2MP paths that need to be diverse from existing P2MP paths.

6. Objective Functions

For the computation of a single or a set of P2MP TE LSPs, a request to meet specific optimization criteria, called an Objective Function (OF) may be indicated.

The computation of one or more P2MP TE-LSPs may be subject to an OF in order to select the "best" candidate paths. A variety of objective functions have been identified as being important during the computation of inter-domain P2MP LSPs. These include:

1. The sub-tree within each domain should be optimized, which can be either the Minimum cost tree [RFC5862] or Shortest path tree [RFC5862].

2. The P2MP LSP path, formed by considering only the entry and exit nodes of the domains (the Core Tree) should be optimal.

3. It should be possible to limit the number of entry points to a domain.

4. It should be possible to force the branches for all leaves within a domain to be in that domain.

7. P2MP Path Computation Procedures

The following sections describe the Core Tree based procedures to satisfy the requirements specified in the previous section. A core tree based solution provides an optimal inter-domain P2MP TE LSP.

7.1. Core Trees

A Core Tree is defined as a tree, which satisfies the following conditions:

- The root of the core tree is the ingress LSR in the root domain;

- The leaves of the core tree are the entry nodes in the leaf domains;

Note that Path-Key Mechanism [RFC5520] MAY be used to hide internal nodes.

An optimal core-tree [based on the OF] will be computed with analyzing the nodes and links within the domains. To support confidentiality the same nodes and links can be hidden via a path-key but they must be computed and be a part of core-tree.

For example, consider the Domain Tree from the figure below, representing a domain tree of 6 domains, and part of the resulting Core Tree which satisfies the aforementioned conditions.

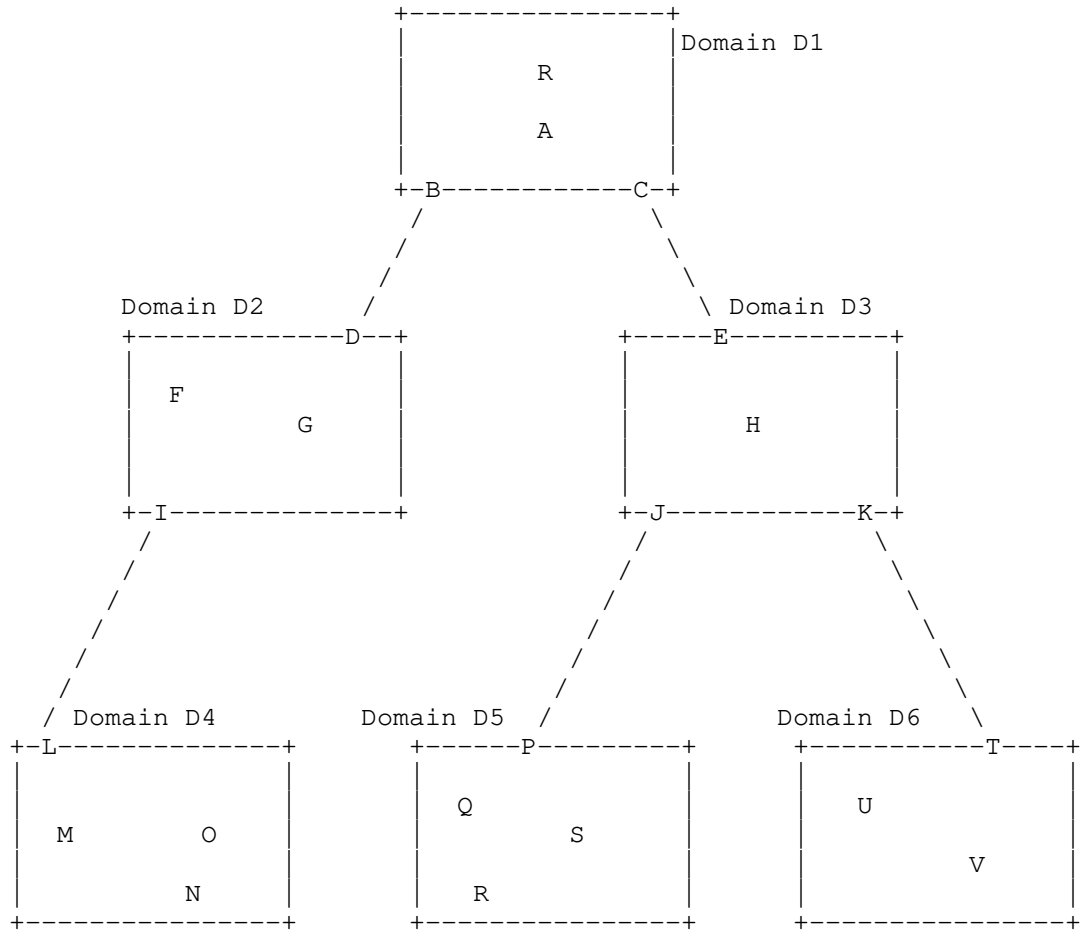


Figure 2: Domain Tree Example

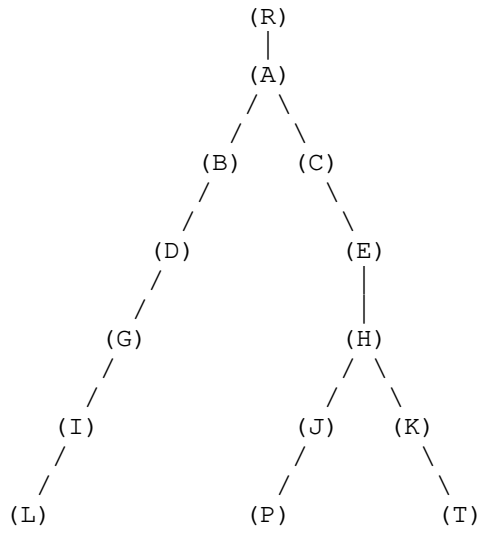


Figure 3: Core Tree

A core tree is computed such that root of the tree is R and the leaf node are the entry nodes of the destination domains (L, P and T). Path-key Mechanism can be used to hide the internal nodes and links in the final core tree.

7.2. Core Tree Computation Procedures

The algorithms to compute the optimal large core tree are outside scope of this document. The following extended BRPC based procedure can be used to compute the core tree.

BRPC Based Core Tree Path Computation Procedure:

1. Using the BRPC procedures to compute the VSPT(i) for each leaf BN(i), i=1 to n, where n is the total number of entry nodes for all the leaf domains. In each VSPT(i), there are a number of P(i) paths.

2. When the root PCE has computed all the VSPT(i), i=1 to n, take one path from each VSPT and form a set of paths, we call it a PathSet(j), j=1 to M, where M=P(1)xP(2)...xP(n);

3. For each PathSet(j), there are n S2L (Source to Leaf BN) paths and form these n paths into a Core Tree(j);

4. There will be M number of Core Trees computed from step3. Apply the OF to each of these M Core Trees and find the optimal Core Tree.

Note that the application of BRPC in the aforementioned procedure differs from the typical one since paths returned from a downstream PCE are not necessary pruned from the solution set by intermediate PCEs.

The reason for this is that if the PCE in a downstream domain does the prunning and returns the single optimal sub-path to its parent PCE, BRPC insures that the ingress PCE will get all the best optimal sub-paths for each LN (Leaf Border Nodes), but the combination of these single optimal sub-paths into a P2MP tree is not necessarily optimal even if each S2L (Source-to-Leaf) sub-path is optimal.

Without trimming, the ingress PCE will get all the possible S2L sub-paths set for LN, and eventually by looking through all the combinations, and taking one sub-path from each set to built one P2MP tree it finds the optimal tree.

Note that if the OF is SPT, VSPT is enough for computing core-tree and downstream PCE can continue to do the prunning. One way to address this would be that a transit PCE in core-tree computation can decide the numbers of paths sent upstream based on the configuration and/or OF. In case of SPT, the number of path sent will be 1.

The proposed method may present a scalability problem for the dynamic computation of the Core Tree (by iterative checking of all combinations of the solution space), specially with dense/meshed domains. Considering a domain sequence D1, D2, D3, D4, where the Leaf border node is at domain D4, PCE(4) will return 1 path. PCE(3) will return N paths, where N is E(3) x X(3), where E(k) x X(k) denotes the number of entry nodes times the number of exit nodes for that domain. PCE(2) will return M paths, where M = E(2) x X(2) x N = E(2) x X(2) x E(3) x X(3) x 1, etc. Generally speaking the number of potential paths at the ingress PCE Q = \prod E(k) x X(k).

Consequently, it is expected that the Core Path will be typically computed offline, without precluding the use of dynamic, online mechanisms such as the one presented here, in which case it SHOULD be possible to configure transit PCEs to control the number of paths sent upstream during BRPC (trading trimming for optimality at the point of trimming and downwards).

7.3. Sub Tree Computation Procedures

Once the core tree is built, the grafting of all the leaf nodes from each domain to the core tree can be achieved by a number of algorithms. One algorithm for doing this phase is that the root PCE will send the request with C bit set for the path computation to the destination(s) directly to the PCE where the destination(s) belong(s) along with the core tree computed from the phase 1.

This approach requires that the root PCE manage a potentially large number of adjacencies (either in persistent or non-persistent mode), including PCEP adjacencies to PCEs that are not within neighboring domains.

A first alternative would involve establishing PCEP adjacencies that correspond to the PCE domain tree. This would require that branch PCEs forward requests and responses from the root PCE towards the leaf PCEs and vice-versa.

Finally, another alternative would use a hierarchical PCE [H-PCE] architecture. The "hierarchically" parent would request sub tree path computations.

The algorithms to compute the optimal large sub tree are outside scope of this document. In the case that the number of destinations and the number of BNs within a domain are not big, the incremental procedure based on p2p path computation using the OSPF can be used.

7.4. PCEP Protocol Extensions

7.4.1. The Extension of RP Object

The RP Object is defined in [RFC5440] as -

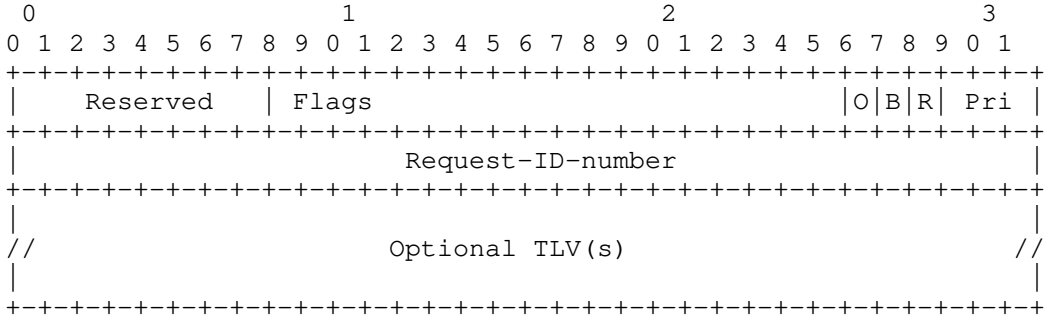


Figure 4: RP Object Body Format

The extended format of the RP object body to include the C bit is as follows:

The C bit is added in the flag bits field of the RP object to signal the receiver of the message that the request/reply is for inter-domain P2MP Core Tree or not.

The following flag is added in this draft:

C bit (P2MP Core Tree bit - 1 bit):

0: This indicates that this is normal PCReq/PCRrep for P2MP.

1: This indicates that this is PCReq or PCRep message for inter-domain Core Tree P2MP. When the C bit is set, then the request message should have the Core Tree passed along with the destinations which and then graphed to the tree.

7.4.2. Domain or PCE Sequence

[DOMAIN-SEQ] mentions the benefit of using domain-sequence over PCE-Sequence. The domain-seq can be used in P2MP scenarios. [PER-DEST] provides a mechanims to encode Domain-Sequence (in form of IRO) per destination.

But if the administrator wants to control PCE rather than domain then PCE-SEQUENCE Object can be used.

The PCE Sequence Object is added to the existing PCE protocol. A list of this objects will represent the PCE topology tree. A list of Sequence Objects can be exchanged between PCEs during the PCE capability exchange or on the first path computation request message between PCEs. In this case, the request message format needs to be changed to include the list of PCE Sequence Objects for the PCE inter-domain P2MP calculation request.

Each PCE Sequence can be obtained from the domain sequence for a specific path. All the PCE sequences for all the paths of P2MP inter-domain form the PCE Topology Tree of the P2MP LSP.

Object Class for the PCE Sequence Object: To be assigned by IANA.

The format of the new PCE Sequence Object for IPv4 (Object-Type 3) is as follows:

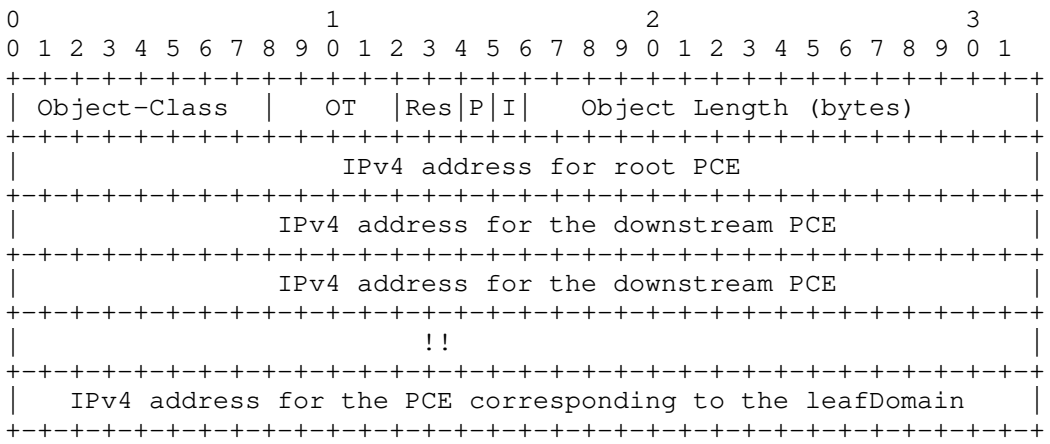


Figure 5: The New PCE Sequence Object Body Format for IPv4

The format of the new PCE Sequence Object for IPv6 (Object-Type 3) is as follows:

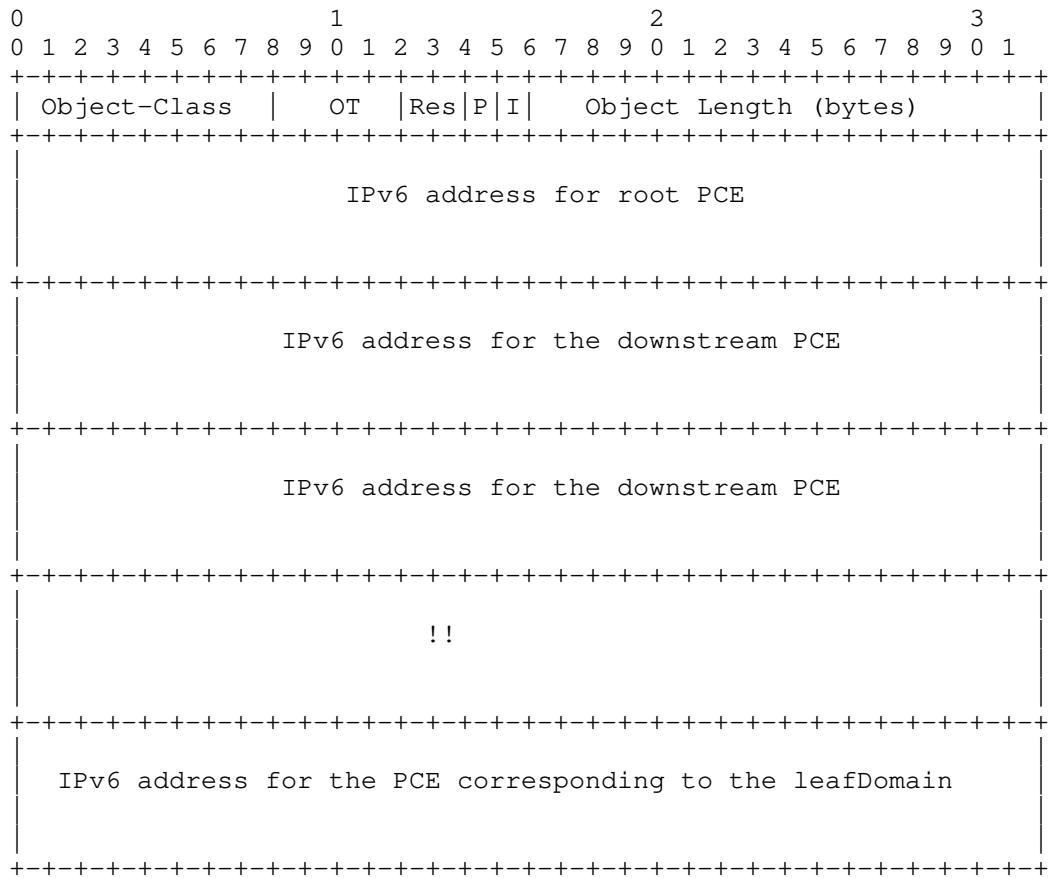


Figure 6: The New PCE Sequence Object Body Format for IPv6

7.5. Relationship with Hierarchical PCE

The actual grafting of subtrees into the Multi-Domain tree needs to be carried out by the source node. This means that the source node needs to get the computed sub-trees from all the involved domains. This requires that the source node either has a PCEP session with all the PCEs, or PCEP messages are routed via the PCEP sessions. This may mean an excessive number of sessions or an added complexity in implementations.

Alternatively, one may use an architecture based on the concept of hierarchical PCE [H-PCE]. The parent PCE would be responsible to request Intra-domain subtrees to the PCEs, combine them and return the overall P2MP tree.

7.6. Parallelism

In order to minimize latency in path computation in multi-domain networks, intra-domain path segments and intra-domain sub-trees SHOULD be computed in parallel when possible. The proposed procedures in this draft present opportunities for parallelism:

1. The BRPC procedure for each leaf node can be launched in parallel by the ingress/root PCE if the dynamic computation of the Core Tree is enabled.

2. Intra-domain P2MP paths can also be computed in parallel by the PCEs once the entry and exit nodes within a domain are known

One of the potential issues of parallelism is that the ingress PCE would require a potentially high number of PCEP adjacencies to "remote" PCEs and that may not be desirable, but a given PCE would only receive requests for the destinations that are in its domain (+ the core nodes), without PCEs forwarding requests.

8. Protection

It is envisaged that protection may be required when deploying and using inter-domain P2MP LSPs. The procedures and mechanisms defined in this document do not prohibit the use of existing and proposed types of protection, including: end-to-end protection [RFC4875] and domain protection schemes.

Segment or facility (link and node) protection is problematic in inter-domain environment due to the limit of Fast-reroute (FRR) [RFC4875] requiring knowledge of its next-hop across domain boundaries whilst maintaining domain confidentiality. Although the FRR protection might be implemented if manually provisioned if next-hop information was known in advance.

8.1. End-to-end Protection

End-to-end protection (Node and Link Protection) principle can be applied for computing backup P2MP LSP. During computaion of Core-Tree and Sub-Tree, end-to-end protection can be taken into consideration. PCE may compute the Primary and backup P2MP LSP together or sequentially.

8.2. Domain Protection

In this protection scheme, backup P2MP Tree can be computed which excludes the transit/branch domain completely. A backup domain path tree is needed with the same source domain and destinations domains

and a new set of transit domains. The backup domain path tree can be applied to the above procedure to obtion the backup P2MP LSP with disjoint transit domains.

9. Manageability Considerations

[RFC5862] describes various manageability requirements in support of P2MP path computation when applying PCEP. This section describes how manageability requirements mentioned in [RFC5862] are supported in the context of PCEP extensions specified in this document.

Note that [RFC5440] describes various manageability considerations in PCEP, and most of manageability requirements mentioned in [PCE-P2MP] are already covered there.

9.1. Control of Function and Policy

In addition to PCE configuration parameters listed in [RFC5440], the following additional parameters might be required:

- The ability to enable or disable single domain P2MP path computations on the PCE.
- The ability to enable or disable multi-domain P2MP path computations on the PCE.
- The PCE may be configured to enable or disable the advertisement of its single domain and multi-domain P2MP path computation capability.

9.2. Information and Data Models

A number of MIB objects have been defined for general PCEP control and monitoring of P2P computations in [PCEP-MIB]. [RFC5862] specifies that MIB objects will be required to support the control and monitoring of the protocol extensions defined in this document. [PCEP-P2MP-MIB] describes managed objects for modeling of PCEP communications between a PCC and PCE, and PCE to PCE, P2MP path computation requests and responses.

In case of offline Core tree computation and configuration MAYBE stored.

9.3. Liveness Detection and Monitoring

No changes are necessary to the liveness detection and monitoring requirements as already embodied in [RFC4657].

It should be noted that multi-domain P2MP computations are likely to take longer than P2P computations, and single domain P2MP computatoins. The liveness detection and monitoring features of the PCECP SHOULD take this into account.

9.4. Verifying Correct Operation

There are no additional requirements beyond those expressed in [RFC4657] for verifying the correct operation of the PCECP. Note that verification of the correct operation of the PCE and its algorithms is out of scope for the protocol requirements, but a PCC MAY send the same request to more than one PCE and compare the results.

9.5. Requirements on Other Protocols and Functional Components

A PCE operates on a topology graph that may be built using information distributed by TE extensions to the routing protocol operating within the network. In order that the PCE can select a suitable path for the signaling protocol to use to install the P2MP LSP, the topology graph must include information about the P2MP signaling and branching capabilities of each LSR in the network.

Mechanisms for the knowledge of other domains, the discovery of corresponding PCEs and their capabilities should be provided and that this information MAY be collected by other mechanisms.

Whatever means is used to collect the information to build the topology graph, the graph MUST include the requisite information. If the TE extensions to the routing protocol are used, these SHOULD be as described in [RFC5073].

9.6. Impact on Network Operation

The use of a PCE to compute P2MP paths is not expected to have significant impact on network operations. However, it should be noted that the introduction of P2MP support to a PCE that already provides P2P path computation might change the loading of the PCE significantly, and that might have an impact on the network behavior, especially during recovery periods immediately after a network failure.

The dynamic computation of Core Trees might also have an impact on the load of the involved PCEs as well as path computation times.

9.7. Policy Control

[RFC5394] provides additional details on policy within the PCE architecture and also provides context for the support of PCE Policy. The are also applicable to Interdomain P2MP Path computation via Core Tree Mechanism.

10. Security Considerations

As described in [RFC5862], P2MP path computation requests are more CPU-intensive and also utilize more link bandwidth. In the event of an unauthorized P2MP path computation request, or a denial of service attack, the subsequent PCEP requests and processing may be disruptive to the network. Consequently, it is important that implementations conform to the relevant security requirements of [RFC5440] that specifically help to minimize or negate unauthorized P2MP path computation requests and denial of service attacks. These mechanisms include:

- Securing the PCEP session requests and responses using TCP security techniques (Section 10.2 of [RFC5440]).

- Authenticating the PCEP requests and responses to ensure the message is intact and sent from an authorized node (Section 10.3 of [RFC5440]).

- Providing policy control by explicitly defining which PCCs, via IP access-lists, are allowed to send P2MP path requests to the PCE (Section 10.6 of [RFC5440]).

PCEP operates over TCP, so it is also important to secure the PCE and PCC against TCP denial of service attacks. Section 10.7.1 of [RFC5440] outlines a number of mechanisms for minimizing the risk of TCP based denial of service attacks against PCEs and PCCs.

PCEP implementations SHOULD also consider the additional security provided by the TCP Authentication Option (TCP-AO) [RFC5925].

11. IANA Considerations

11.1. New Flag of the RP Object

A new flag of the RP object (specified in [RFC5440]) is defined in this document. IANA maintains a registry of RP object flags in the "RP Object Flag Field" sub-registry of the "Path Computation Element Protocol (PCEP) Numbers" registry.

IANA has allocated the following value:

Bit	Description	Reference
TBA	P2MP Core Tree bit	[This.I-D]

11.2. New PCEP Object

IANA is requested to assign a new object class in the registry of PCEP Objects as follows.

Object Class	Name	Object Type	Name	Reference
TBA	Pce-Seq	1	Pce Sequence [IPv4]	[This.I-D]
		2	Pce Sequence [IPv6]	[This.I-D]

12. Acknowledgements

The authors would like to thank Adrian Farrel, Dan Tappan and Olufemi Komolafe for their valuable comments on this document.

13. References

13.1. Normative References

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

[RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.

[RFC6006] Zhao, Q., King, D., Verhaeghe, F., Takeda, T., Ali, Z., and J. Meuric, "Extensions to the Path Computation Element Communication Protocol (PCEP) for Point-to-Multipoint Traffic Engineering Label Switched Paths", RFC 6006, September 2010.

13.2. Informative References

[RFC4461] Yasukawa, S., "Signaling Requirements for Point-to-Multipoint Traffic-Engineered MPLS Label Switched Paths (LSPs)", RFC 4461, April 2006.

[RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture",

RFC 4655, August 2006.

[RFC4657] Ash, J. and J. Le Roux, "Path Computation Element (PCE) Communication Protocol Generic Requirements", RFC 4657, September 2006.

[RFC4875] Aggarwal, R., Papadimitriou, D., and S. Yasukawa, "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, May 2007.

[RFC5073] Vasseur, J. and J. Le Roux, "IGP Routing Protocol Extensions for Discovery of Traffic Engineering Node Capabilities", RFC 5073, December 2007.

[RFC5152] Vasseur, JP., Ayyangar, A., and R. Zhang, "A Per-Domain Path Computation Method for Establishing Inter-Domain Traffic Engineering (TE) Label Switched Paths (LSPs)", RFC 5152, February 2008.

[RFC5376] Bitar, N., Zhang, R., and K. Kumaki, "Inter-AS Requirements for the Path Computation Element Communication Protocol (PCECP)", RFC 5376, November 2008.

[RFC5394] Bryskin, I., Papadimitriou, D., Berger, L., and J. Ash, "Policy-Enabled Path Computation Framework", RFC 5394, December 2008.

[RFC5441] Vasseur, JP., Zhang, R., Bitar, N., and JL. Le Roux, "A Backward-Recursive PCE-Based Computation (BRPC) Procedure to Compute Shortest Constrained Inter-Domain Traffic Engineering Label Switched Paths", RFC 5441, April 2009.

[RFC5520] Bradford, R., Vasseur, JP., and A. Farrel, "Preserving Topology Confidentiality in Inter-Domain Path Computation Using a Path-Key-Based Mechanism", RFC 5520, April 2009.

[RFC5671] Yasukawa, S. and A. Farrel, "Applicability of the Path Computation Element (PCE) to Point-to-Multipoint (P2MP) MPLS and GMPLS Traffic Engineering (TE)", RFC 5671, October 2009.

[RFC5862] Yasukawa, S. and A. Farrel, "Path Computation Clients (PCC) - Path Computation Element (PCE)

Requirements for Point-to-Multipoint MPLS-TE", RFC 5862, June 2010.

[RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, June 2010.

[H-PCE] 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", October 2011.

[PCEP-MIB] Koushik, K., Stephan, E., Zhao, Q., and D. King, "PCE communication protocol (PCEP) Management Information Base (Work in Progress)", July 2010.

[PCEP-P2MP-MIB] Zhao, Q., Dhody, D., Palle, U., and D. King, "Management Information Base for the PCE Communications Protocol (PCEP) When Requesting Point-to-Multipoint Services (Work in Progress)", Sept 2011.

[DOMAIN-SEQ] Dhody, D., Palle, U., and R. Casellas, "Standard Representation Of Domain Sequence (Work in Progress)", Aug 2011.

[PER-DEST] Dhody, D. and U. Palle, "Supporting explicit-path per destination in Path Computation Element Communication Protocol (PCEP) - P2MP Path Request. (Work in Progress)", June 2011.

Authors' Addresses

Quintin Zhao
Huawei Technology
125 Nagog Technology Park
Acton, MA 01719
US

EMail: quintin.zhao@huawei.com

Dhruv Dhody
Huawei Technology
Leela Palace
Bangalore, Karnataka 560008
INDIA

EMail: dhruv.dhody@huawei.com

Zafar Ali
Cisco Systems, Inc.
2000 Innovation Drive
Kanata, Ontario K2K 3E8
CANADA

EMail: zali@cisco.com

Tarek Saad
Cisco Systems, Inc.
US

EMail: tsaad@cisco.com

Siva Sivabalan
Cisco Systems, Inc.
2000 Innovation Drive
Kanata, Ontario K2K 3E8
CANADA

EMail: msiva@cisco.com

Daniel King
Old Dog Consulting
UK

EMail: daniel@olddog.co.uk

Ramon Casellas
CTTC - Centre Tecnologic de Telecomunicacions de Catalunya
Av. Carl Friedrich Gauss n7
Castelldefels, Barcelona 08860
SPAIN

EMail: ramon.casellas@cttc.e

PCE Working Group
Internet-Draft
Intended status: Experimental
Expires: December 17, 2014

Q. Zhao
D. Dhody
Huawei Technology
D. King
Old Dog Consulting
Z. Ali
Cisco Systems
R. Casellas
CTTC

June 17, 2014

PCE-based Computation Procedure To Compute Shortest Constrained P2MP Inter-domain Traffic Engineering Label Switched Paths draft-ietf-pce-pcep-inter-domain-p2mp-procedures-08

Abstract

The ability to compute paths for constrained point-to-multipoint (P2MP) Traffic Engineering Label Switched Paths (TE LSPs) across multiple domains has been identified as a key requirement for the deployment of P2MP services in MPLS and GMPLS-controlled networks. The Path Computation Element (PCE) has been recognized as an appropriate technology for the determination of inter-domain paths of P2MP TE LSPs.

This document describes an experiment to provide procedures and extensions to the PCE communication Protocol (PCEP) for the computation of inter-domain paths for P2MP TE 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). 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 December 17, 2014.

Copyright Notice

Table of Contents

1. Introduction

Multicast services are increasingly in demand for high-capacity applications such as multicast Virtual Private Networks (VPNs), IP-television (IPTV) which may be on-demand or streamed, and content-rich media distribution (for example, software distribution, financial streaming, or database-replication). The ability to compute constrained Traffic Engineering Label Switched Paths (TE LSPs) for point-to-multipoint (P2MP) LSPs in Multiprotocol Label Switching (MPLS) and Generalized MPLS (GMPLS) networks across multiple domains are therefore required.

The applicability of the PCE [RFC4655] for the computation of such paths is discussed in [RFC5671], and the requirements placed on the PCE communications Protocol (PCEP) for this are given in [RFC5862].

This document details the requirements for inter-domain P2MP path computation, it then describes the experimental procedure "core-tree" path computation, developed to address the requirements and objectives for inter-domain P2MP path computation.

When results of implementation and deployment are available, this document will be updated and refined, and then moved from Experimental status to Standards Track.

1.2. Scope

The inter-domain P2MP path computation procedures described in this document is experimental. The experiment is intended to enable research for the usage of the PCE to support inter-domain P2MP path computation.

This document is not intended to replace the intra-domain P2MP path computation approach defined by [RFC6006], and will not impact existing PCE procedures and operations.

1.3. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

2. Terminology

Terminology used in this document is consistent with the related MPLS/GMPLS and PCE documents [RFC4461], [RFC4655], [RFC4875], [RFC5376], [RFC5440], [RFC5441], [RFC5671] and [RFC5862].

The additional terms Core-Tree, Leaf Domain, Path Tree, Path Domain Sequence, Path Domain Tree, Root Domain, Sub-Tree and Transit/branch Domain are further defined below.

Core-Tree: a P2MP tree where the root is the ingress Label Switching Router (LSR), and the leaf nodes are the entry BNs of the leaf domains.

Entry BN of domain(n): a Boundary Node (BN) connecting domain(n-1) to domain(n) along a determined sequence of domains.

Exit BN of domain(n): a BN connecting domain(n) to domain(n+1) along a determined sequence of domains.

H-PCE: Hierarchical PCE (as per [RFC6805]).

Leaf Domain: a domain with one or more leaf nodes.

Path Tree: a set of LSRs and TE links that comprise the path of a P2MP TE LSP from the ingress LSR to all egress LSRs (the leaf nodes).

Path Domain Sequence: the known sequence of domains for a path between the root domain and a leaf domain.

Path Domain Tree: the tree formed by the domains that the P2MP path crosses, where the source (ingress) domain is the root domain.

PCE(i): a PCE that performs path computations for domain(i).

Root Domain: the domain that includes the ingress (root) LSR.

Sub-tree: a P2MP tree where the root is the selected entry BN of the leaf domain and the leaf nodes are the destinations (leaves) in that domain. The sub-trees are grafted to the core-tree.

Transit/branch Domain: a domain that has an upstream and one or more downstream neighbor domain.

3. Examination of Existing Mechanisms

The Path Computation Element (PCE) defined in [RFC4655] is an entity that is capable of computing a network path or route based on a network graph, and applying computational constraints. A Path

Computation Client (PCC) may make requests to a PCE for paths to be computed.

[RFC4875] describes how to set up P2MP TE LSPs for use in MPLS and GMPLS-controlled networks. The PCE is identified as a suitable application for the computation of paths for P2MP TE LSPs [RFC5671].

[RFC5441] specifies a procedure relying on the use of multiple PCEs to compute Point to Point (P2P) inter-domain constrained shortest paths across a predetermined sequence of domains, using a Backward Recursive Path Computation (BRPC) technique. The technique can be combined with the use of Path-Keys [RFC5520] to preserve confidentiality across domains, which is sometimes required when domains are managed by different Service Providers.

PCEP [RFC5440] was extended for point-to-multipoint (P2MP) path computation requests in [RFC6006].

As discussed in [RFC4461], a P2MP tree is the ordered set of LSRs and TE links that comprise the path of a P2MP TE LSP from its ingress LSR to all of its egress LSRs. A P2MP LSP is set up with TE constraints and allows efficient packet or data replication at various branching points in the network. As per [RFC5671] branch point selection is fundamental to the determination of the paths for a P2MP TE LSP. Not only is this selection constrained by the network topology and available network resources, but it is determined by the objective functions (OF) that may be applied to path computation.

Generally, an inter-domain P2MP tree (i.e., a P2MP tree with source and at least one destination residing in different domains) is particularly difficult to compute even for a distributed PCE architecture. For instance, while the BRPC may be well-suited for P2P paths, P2MP path computation involves multiple branching path segments from the source to the multiple destinations. As such, inter-domain P2MP path computation may result in a plurality of per-domain path options that may be difficult to coordinate efficiently and effectively between domains. That is, when one or more domains have multiple ingress and/or egress boundary nodes (i.e., when the domains are multiply inter-connected), existing techniques may be convoluted when used to determine which boundary node of another domain will be utilized for the inter-domain P2MP tree, and no way to limit the computation of the P2MP tree to those utilized boundary nodes.

A trivial solution to the computation of inter-domain P2MP tree would be to compute shortest inter-domain P2P paths from source to each destination and then combine them to generate an inter-domain, shortest-path-to-destination P2MP tree. This solution, however, cannot be used to trade cost to destination for overall tree cost

(i.e., it cannot produce a Minimum Cost Tree (MCT)) and in the context of inter-domain P2MP TE LSPs it cannot be used to reduce the number of domain boundary nodes that are transited. Computing P2P TE LSPs individually does not guarantee the generation of an optimal P2MP tree for every definition of "optimal" in every topology.

Per Domain path computation [RFC5152] may be used to compute P2MP multi-domain paths, but may encounter the issues previously described. Furthermore, this approach may also be considered to have scaling issues during LSP setup. That is, the LSP to each leaf is signaled separately, and each boundary node needs to perform path computation for each leaf.

P2MP Minimum Cost Tree (MCT), i.e. a computation which guarantees the least cost resulting tree, typically is an NP-complete problem. Moreover, adding and/or removing a single destination to/from the tree may result in an entirely different tree. In this case, frequent MCT path computation requests may prove computationally intensive, and the resulting frequent tunnel reconfiguration may even cause network destabilization.

This document presents a solution, procedures and extensions to PCEP to support P2MP inter-domain path computation.

4. Assumptions

Within this document we make the following assumptions:

- Due to deployment and commercial limitations (e.g., inter-AS (Autonomous System) peering agreements), the path domain tree will be known in advance;
- Each PCE knows about any leaf LSRs in the domain it serves;

Additional assumptions are documented in [RFC5441] and are not repeated here.

5. Requirements

This section summarizes the requirements specific to computing inter-domain P2MP paths. In these requirements we note that the actual computation time taken by any PCE implementation is outside the scope of this document, but we observe that reducing the complexity of the required computations has a beneficial effect on the computation time regardless of implementation. Additionally, reducing the number of message exchanges and the amount of information exchanged will reduce the overall computation time for the entire P2MP tree. We refer to

the "complexity of the computation" as the impact on these aspects of path computation time as various parameters of the topology and the P2MP TE LSP are changed.

It is also important that the solution can preserve confidentiality across domains, which is required when domains are managed by different Service Providers via Path-Key mechanism [RFC5520].

Other than the requirements specified in [RFC5862], a number of requirements specific to inter-domain P2MP are detailed below:

1. The complexity of the computation for each sub-tree within each domain SHOULD be dependent only on the topology of the domain and it SHOULD be independent of the domain sequence.

2. The number of PCReq (Path Computation Request) and PCRep (Path Computation Reply) messages SHOULD be independent of the number of multicast destinations in each domain.

3. It SHOULD be possible to specify the domain entry and exit nodes in the PCReq.

4. Specifying which nodes are be used as branch nodes SHOULD be supported in the PCReq.

5. Reoptimization of existing sub-trees SHOULD be supported.

6. It SHOULD be possible to compute diverse P2MP paths from existing P2MP paths.

6. Objective Functions and Constraints

For the computation of a single or a set of P2MP TE LSPs, a request to meet specific optimization criteria, called an Objective Function (OF), MAY be used. Using an OF to select the "best" candidate path, include:

- The sub-tree within each domain SHOULD be optimized using minimum cost tree [RFC5862], or shortest path tree [RFC5862].

In addition to the OFs, the following constraints MAY also be beneficial for inter-domain P2MP path computation:

1. The computed P2MP "core-tree" SHOULD be optimal when only considering the paths to the leaf domain entry BNs.

2. Grafting and pruning of multicast destinations (sub-tree) within a leaf domain SHOULD ensure minimal impact on other domains

and on the core-tree.

3. It SHOULD be possible to choose to optimize the core-tree.

4. It SHOULD be possible to choose optimize the entire tree (P2MP LSP).

5. It SHOULD be possible to combine the aforementioned OFs and constraints for P2MP path computation.

When implementing and operating P2MP LSPs, following needs to be taken into consideration:

- The complexity of computation.
- The optimality of the tree (core-tree as well as full P2MP LSP tree).
- The stability of the core-tree.

The solution SHOULD allow these trade-offs to be made at computation time.

The algorithms used to compute optimal paths using a combination of OFs and multiple constraints is out of scope of this document.

7. P2MP Path Computation Procedures

7.1. General

A P2MP path computation can be broken down into two steps of core-tree computation and grafting of sub-trees. Breaking the procedure into these specific steps has the following impact:

- The core-tree and sub-tree are smaller in comparison to the full P2MP Tree and are thus easier to compute.
- An implementation MAY choose to keep the core-tree fairly static or computed offline (trade-off with optimality).
- Adding/Pruning of leaves which require changes to sub-tree in leaf-domain only.
- The PCEP message size is smaller in comparison.

Allowing the core-tree based solution to provide an optimal inter-domain P2MP TE LSP.

The following sub-sections describe the core-tree based mechanism, including procedures and PCEP extensions, that satisfy the requirements and objectives specified in Section 5 and Section 6 of this document.

7.2. Core-Trees

A core-tree is defined as a tree that satisfies the following conditions:

- The root of the core-tree is the ingress LSR in the root domain;
- The leaves of the core-tree are the entry boundary nodes in the leaf domains.

To support confidentiality these nodes and links MAY be hidden using the path-key mechanism [RFC5520], but they MUST be computed and be a part of core-tree.

For example, consider the Domain Tree in Figure 1 below, representing a domain tree of 6 domains, and part of the resulting core-tree which satisfies the aforementioned conditions.

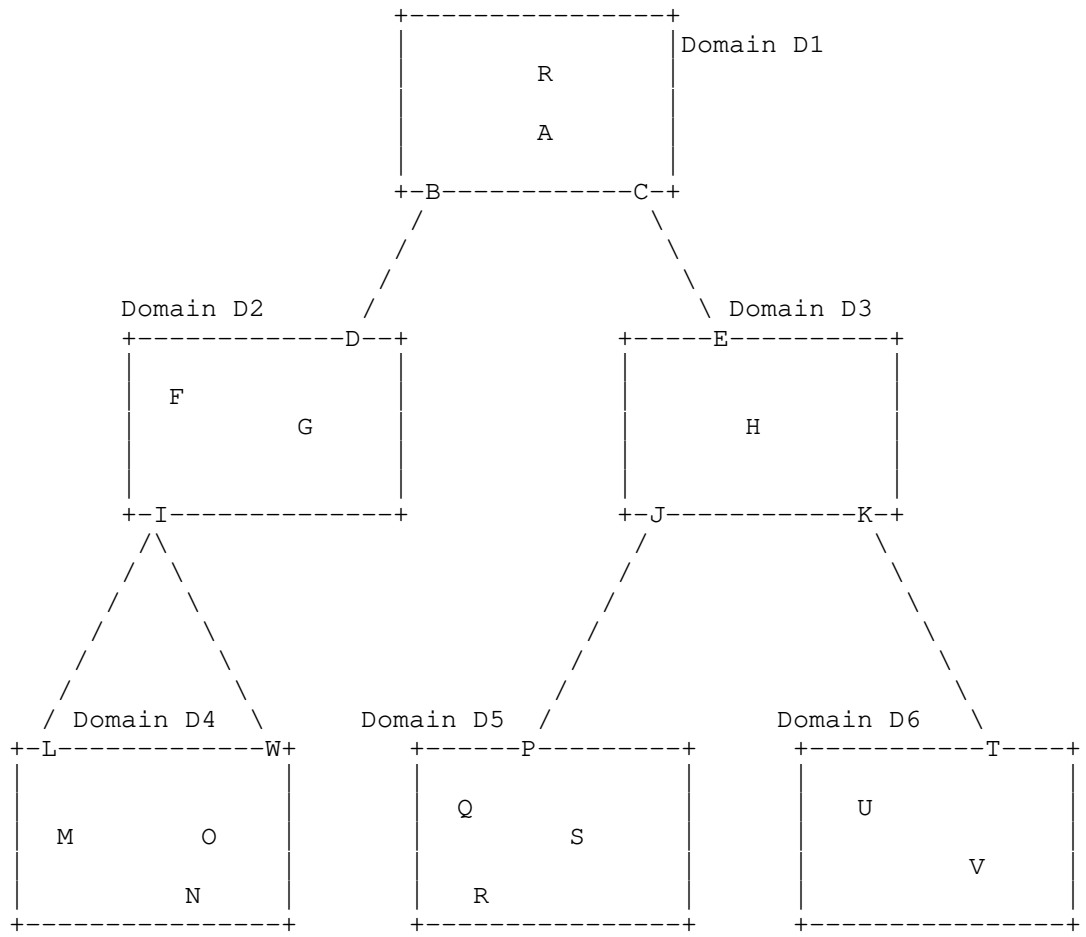


Figure 1: Domain Tree Example

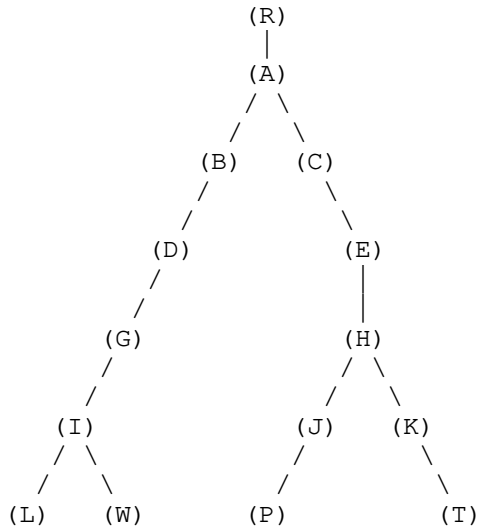


Figure 2: Core-Tree

A core-tree is computed such that root of the tree is R and the leaf node are the entry nodes of the destination domains (L, W, P and T). Path-key mechanism can be used to hide the internal nodes and links (node G and H are hidden via Path-Key PK1 and PK2 respectively) in the final core-tree as shown below for domain D2 and D3.

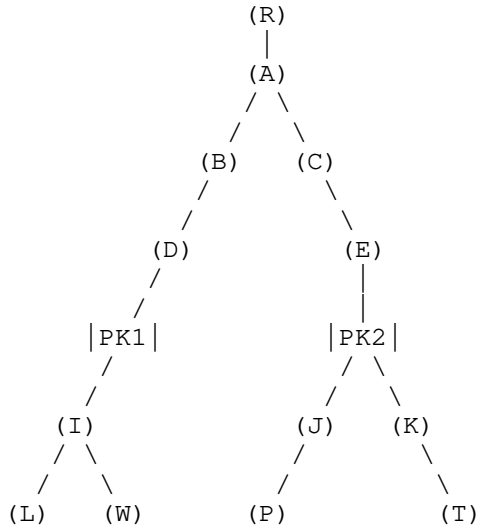


Figure 3: Core-Tree with Path-Key

7.3. Optimal Core-Tree Computation Procedure

Applying the core-tree procedure to large groups of domains, such as the Internet, is not considered feasible or desirable, and is out of scope for this document.

The following extended BRPC-based procedure can be used to compute the core-tree. Note that a root PCE MAY further use its own enhanced optimization techniques in future to compute the core-tree.

A BRPC-based core-tree path computation procedure is described below:

1. Using the BRPC procedures to compute the VSPT(i) (Virtual Shortest Path Tree) for each leaf BN(i), i=1 to n, where n is the total number of entry nodes for all the leaf domains. In each VSPT(i), there are a number of P(i) paths.

2. When the root PCE has computed all the VSPT(i), i=1 to n, take one path from each VSPT and form all possible sets of paths, we call them PathSet(j), j=1 to M, where M=P(1)xP(2)...xP(n);

3. For each PathSet(j), there are n S2L (Source-to-Leaf) BN paths and form these n paths into a core-tree(j);

4. There will be M number core-trees computed from step 3. An optimal core-tree is selected based on the OF and constraints.

Note that, since point to point BRPC procedure is used to compute VSPT, the path request and response message format defined in [RFC5440] are used.

Also note that the application of BRPC in the aforementioned procedure differs from the typical one since paths returned from a downstream PCE are not necessarily pruned from the solution set (extended VSPT) by intermediate PCEs. The reason for this is that if the PCE in a downstream domain does the pruning and returns the single optimal sub-path to the upstream PCE, the combination of these single optimal sub-paths into a core-tree is not necessarily optimal even if each S2L (Source-to-Leaf) sub-path is optimal.

Without trimming, the ingress PCE will obtain all the possible S2L sub-paths set for the entry boundary nodes of the leaf domain. The PCE will then, by looking through all the combinations and taking one sub-path from each set to build one tree, can select the optimal core-tree.

A PCE MAY add equal cost paths within the domain while constructing an extended VSPT. This will provide the ingress PCE more candidate paths for an optimal core-tree.

The proposed method may present a scalability problem for the dynamic computation of the core-tree (by iterative checking of all combinations of the solution space), specially with dense/meshed domains. Considering a domain sequence D1, D2, D3, D4, where the Leaf Boundary Node is at domain D4, PCE(4) will return 1 path. PCE(3) will return N paths, where N is E(3) x X(3), where E(k) x X(k) denotes the number of entry nodes times the number of exit nodes for that domain. PCE(2) will return M paths, where M = E(2) x X(2) x N = E(2) x X(2) x E(3) x X(3) x 1, etc. Generally speaking the number of potential paths at the ingress PCE Q = prod E(k) x X(k).

Consequently, it is expected that the core-tree will be typically computed offline, without precluding the use of dynamic, online mechanisms such as the one presented here, in which case it SHOULD be possible to configure transit PCEs to control the number of paths sent upstream during BRPC (trading trimming for optimality at the point of trimming and downwards).

7.4. Sub-tree Computation Procedures

Once the core-tree is built, the grafting of all the leaf nodes from each domain to the core-tree can be achieved by a number of algorithms. One algorithm for doing this phase is that the root PCE will send the request with C bit set (as defined in section 7.4.1 of this document) for the path computation to the destination(s) directly to the PCE where the destination(s) belong(s) along with the core-tree computed from section 7.2.

This approach requires that the root PCE manage a potentially large number of adjacencies (either in persistent or non-persistent mode), including PCEP adjacencies to PCEs that are not within neighbor domains.

An alternative would involve establishing PCEP adjacencies that correspond to the PCE domain tree. This would require that branch PCEs forward requests and responses from the root PCE towards the leaf PCEs and vice-versa.

Note that the P2MP path request and response format is as per [RFC6006], where Record Route Object (RRO) are used to carry the core-tree paths in the P2MP grafting request.

The algorithms to compute the optimal large sub-tree are outside scope of this document.

7.5. PCEP Protocol Extensions

7.5.1. The Extension of RP Object

This experiment will be carried out by extending the RP (Request Parameters) object (defined in [RFC5440]) used in PCEP requests and responses.

The extended format of the RP object body to include the C bit is as follows:

The C bit is added in the flag bits field of the RP object to signal the receiver of the message that the request/reply is for inter-domain P2MP core-tree or not.

The following flag is added in this draft:

Bit Number	Name Flag
TBA	Core-tree computation (C-bit)

C bit (Core-Tree bit - 1 bit):

0: This indicates that this is not for an inter-domain P2MP core-tree.

1: This indicates that this is a PCEP request or a response for the computation of a inter-domain core-tree or for the grafting of a sub-tree to a inter-domain core-tree.

7.5.2. Domain and PCE Sequence

The procedure described in this document requires the domain-tree to be known in advance. This information MAY be either administratively predetermined or dynamically discovered by some means such as Hierarchical PCE (H-PCE) [RFC6805] framework, or derived through the IGP/BGP routing information.

Examples of ways to encode the domain path tree include [RFC5886] using PCE-ID Object and [DOMAIN-SEQ].

7.6. Using H-PCE for Scalability

The ingress/root PCE is responsible for the core-tree computation as well as grafting of sub-trees into the multi-domain tree. Therefore, the ingress/root PCE will receive all computed path segments from all the involved domains. When the ingress/root PCE chooses to have a PCEP session with all involved PCEs, this may cause an excessive number of sessions or added complexity in implementations.

The use of the H-PCE framework [RFC6805] may be used to establish a dedicated PCE with the capability (memory and CPU) and knowledge to maintain the necessary PCEP sessions. The parent PCE would be responsible to request intra-domain path computation request to the

PCEs, combine them and return the overall P2MP tree.

7.7. Parallelism

In order to minimize latency in path computation in multi-domain networks, intra-domain path segments and intra-domain sub-trees can be computed in parallel when possible. The proposed procedures in this draft present opportunities for parallelism:

1. The BRPC procedure for each leaf boundary node can be launched in parallel by the ingress/root PCE for dynamic computation of core-tree.

2. The grafting of sub-trees can be triggered in parallel once the core-tree is computed.

One of the potential issues of parallelism is that the ingress PCE would require a potentially high number of PCEP adjacencies to "remote" PCEs at the same time and that may not be desirable.

8. Protection

It is envisaged that protection may be required when deploying and using inter-domain P2MP TE LSPs. The procedures and mechanisms defined in this document do not prohibit the use of existing and proposed types of protection, including: end-to-end protection [RFC4875] and domain protection schemes.

Segment or facility (link and node) protection is problematic in inter-domain environment due to the limit of Fast-reroute (FRR) [RFC4875] requiring knowledge of its next-hop across domain boundaries whilst maintaining domain confidentiality. Although the FRR protection might be implemented if next-hop information was known in advance.

8.1. End-to-end Protection

An end-to-end protection (for nodes and links) principle can be applied for computing backup P2MP TE LSPs. During computation of the core-tree and sub-trees, may also be taken into consideration. A PCE may compute the primary and backup P2MP TE LSP together or sequentially.

8.2. Domain Protection

In this protection scheme, backup P2MP Tree can be computed which excludes the transit/branch domain completely. A backup domain path tree is needed with the same source domain and destinations domains

and a new set of transit domains. The backup path tree can be applied to the above procedure to obtain the backup P2MP TE LSP with disjoint transit domains.

9. Manageability Considerations

[RFC5862] describes various manageability requirements in support of P2MP path computation when applying PCEP. This section describes how manageability requirements mentioned in [RFC5862] are supported in the context of PCEP extensions specified in this document.

Note that [RFC5440] describes various manageability considerations in PCEP, and most of manageability requirements mentioned in [RFC6006] are already covered there.

9.1. Control of Function and Policy

In addition to PCE configuration parameters listed in [RFC5440] and [RFC6006], the following additional parameters might be required:

- The ability to enable or disable multi-domain P2MP path computations on the PCE.
- The PCE may be configured to enable or disable the advertisement of its multi-domain P2MP path computation capability.

9.2. Information and Data Models

A number of MIB objects have been defined for general PCEP control and monitoring of P2P computations in [PCEP-MIB]. [RFC5862] specifies that MIB objects will be required to support the control and monitoring of the protocol extensions defined in this document. [PCEP-P2MP-MIB] describes managed objects for modeling of PCEP communications between a PCC and PCE, and PCE to PCE, P2MP path computation requests and responses.

9.3. Liveness Detection and Monitoring

No changes are necessary to the liveness detection and monitoring requirements as already embodied in [RFC4657].

It should be noted that multi-domain P2MP computations are likely to take longer than P2P computations, and single domain P2MP computations. The liveness detection and monitoring features of the PCEP SHOULD take this into account.

9.4. Verifying Correct Operation

There are no additional requirements beyond those expressed in [RFC4657] for verifying the correct operation of the PCEP. Note that verification of the correct operation of the PCE and its algorithms is out of scope for the protocol requirements, but a PCC MAY send the same request to more than one PCE and compare the results.

9.5. Requirements on Other Protocols and Functional Components

A PCE operates on a topology graph that may be built using information distributed by TE extensions to the routing protocol operating within the network. In order that the PCE can select a suitable path for the signaling protocol to use to install the P2MP TE LSP, the topology graph MUST include information about the P2MP signaling and branching capabilities of each LSR in the network.

Mechanisms for the knowledge of other domains, the discovery of corresponding PCEs and their capabilities SHOULD be provided and that this information MAY be collected by other mechanisms.

Whatever means is used to collect the information to build the topology graph, the graph MUST include the requisite information. If the TE extensions to the routing protocol are used, these SHOULD be as described in [RFC5073].

9.6. Impact on Network Operation

The use of a PCE to compute P2MP paths is not expected to have significant impact on network operations. However, it should be noted that the introduction of P2MP support to a PCE that already provides P2P path computation might change the loading of the PCE significantly, and that might have an impact on the network behavior, especially during recovery periods immediately after a network failure.

The dynamic computation of core-trees might also have an impact on the load of the involved PCEs as well as path computation times.

It should be noted that pre-computing and maintaining domain-trees might be a considerable administration effort on the operator.

9.7. Policy Control

[RFC5394] provides additional details on policy within the PCE architecture and also provides context for the support of PCE Policy. They are also applicable to Inter-domain P2MP Path computation via the core-tree mechanism.

10. Security Considerations

As described in [RFC5862], P2MP path computation requests are more CPU-intensive and also utilize more link bandwidth. In the event of an unauthorized P2MP path computation request, or a denial of service attack, the subsequent PCEP requests and processing may be disruptive to the network. Consequently, it is important that implementations conform to the relevant security requirements of [RFC5440] that specifically help to minimize or negate unauthorized P2MP path computation requests and denial of service attacks. These mechanisms include:

- Securing the PCEP session requests and responses using TCP security techniques (Section 10.2 of [RFC5440]).

- Authenticating the PCEP requests and responses to ensure the message is intact and sent from an authorized node (Section 10.3 of [RFC5440]).

- Providing policy control by explicitly defining which PCCs, via IP access-lists, are allowed to send P2MP path requests to the PCE (Section 10.6 of [RFC5440]).

PCEP operates over TCP, so it is also important to secure the PCE and PCC against TCP denial of service attacks. Section 10.7.1 of [RFC5440] outlines a number of mechanisms for minimizing the risk of TCP-based denial of service attacks against PCEs and PCCs.

PCEP implementations SHOULD also consider the additional security provided by the TCP Authentication Option (TCP-AO) [RFC5925].

Finally, any multi-domain operation necessarily involves the exchange of information across domain boundaries. This may represent a significant security and confidentiality risk especially when the domains are controlled by different commercial entities. PCEP allows individual PCEs to maintain confidentiality of their domain path information by using path-keys [RFC5520] and would allow for securing of domain path information when performing core-tree based path computations.

11. IANA Considerations

IANA maintains the "Path Computation Element Protocol (PCEP) Numbers" registry with the "RP Object Flag Field" sub-registry.

IANA is requested to allocate a new bit from this registry as follows:

Bit	Description	Reference

12. Acknowledgements

The authors would like to thank Adrian Farrel, Dan Tappan, Olufemi Komolafe, Oscar Gonzalez de Dios and Julien Meuric for their valuable comments on this document.

13. References

13.1. Normative References

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

[RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.

[RFC5441] Vasseur, JP., Zhang, R., Bitar, N., and JL. Le Roux, "A Backward-Recursive PCE-Based Computation (BRPC) Procedure to Compute Shortest Constrained Inter-Domain Traffic Engineering Label Switched Paths", RFC 5441, April 2009.

[RFC6006] Zhao, Q., King, D., Verhaeghe, F., Takeda, T., Ali, Z., and J. Meuric, "Extensions to the Path Computation Element Communication Protocol (PCEP) for Point-to-Multipoint Traffic Engineering Label Switched Paths", RFC 6006, September 2010.

13.2. Informative References

[RFC4461] Yasukawa, S., "Signaling Requirements for Point-to-Multipoint Traffic-Engineered MPLS Label Switched Paths (LSPs)", RFC 4461, April 2006.

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

[RFC4657] Ash, J. and J. Le Roux, "Path Computation Element (PCE) Communication Protocol Generic Requirements", RFC 4657, September 2006.

[RFC4875] Aggarwal, R., Papadimitriou, D., and S. Yasukawa, "Extensions to Resource Reservation Protocol -

Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, May 2007.

[RFC5073] Vasseur, J. and J. Le Roux, "IGP Routing Protocol Extensions for Discovery of Traffic Engineering Node Capabilities", RFC 5073, December 2007.

[RFC5152] Vasseur, JP., Ayyangar, A., and R. Zhang, "A Per-Domain Path Computation Method for Establishing Inter-Domain Traffic Engineering (TE) Label Switched Paths (LSPs)", RFC 5152, February 2008.

[RFC5376] Bitar, N., Zhang, R., and K. Kumaki, "Inter-AS Requirements for the Path Computation Element Communication Protocol (PCECP)", RFC 5376, November 2008.

[RFC5394] Bryskin, I., Papadimitriou, D., Berger, L., and J. Ash, "Policy-Enabled Path Computation Framework", RFC 5394, December 2008.

[RFC5520] Bradford, R., Vasseur, JP., and A. Farrel, "Preserving Topology Confidentiality in Inter-Domain Path Computation Using a Path-Key-Based Mechanism", RFC 5520, April 2009.

[RFC5671] Yasukawa, S. and A. Farrel, "Applicability of the Path Computation Element (PCE) to Point-to-Multipoint (P2MP) MPLS and GMPLS Traffic Engineering (TE)", RFC 5671, October 2009.

[RFC5862] Yasukawa, S. and A. Farrel, "Path Computation Clients (PCC) - Path Computation Element (PCE) Requirements for Point-to-Multipoint MPLS-TE", RFC 5862, June 2010.

[RFC5886] Vasseur, JP., Le Roux, JL., and Y. Ikejiri, "A Set of Monitoring Tools for Path Computation Element (PCE)-Based Architecture", RFC 5886, June 2010.

[RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, June 2010.

[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.

[PCEP-MIB] Koushik, K., Stephan, E., Zhao, Q., King, D., and J. Hardwick, "PCE communication protocol (PCEP) Management Information Base (Work in Progress)", April 2014.

[PCEP-P2MP-MIB] Zhao, Q., Dhody, D., Palle, U., and D. King, "Management Information Base for the PCE Communications Protocol (PCEP) When Requesting Point-to-Multipoint Services (Work in Progress)", Aug 2012.

[DOMAIN-SEQ] Dhody, D., Palle, U., and R. Casellas, "Standard Representation Of Domain Sequence (Work in Progress)", July 2014.

14. Contributor Addresses

Siva Sivabalan
Cisco Systems
2000 Innovation Drive
Kanata, Ontario K2K 3E8
CANADA

EMail: msiva@cisco.com

Tarek Saad
Cisco Systems, Inc.
2000 Innovation Drive
Kanata, Ontario K2K 3E8
CANADA

EMail: tsaad@cisco.com

15. Authors' Addresses

Quintin Zhao
Huawei Technology
125 Nagog Technology Park
Acton, MA 01719
US

EMail: quintin.zhao@huawei.com

Dhruv Dhody
Huawei Technology
Leela Palace
Bangalore, Karnataka 560008

INDIA

EMail: dhruv.dhody@huawei.com

Zafar Ali
Cisco Systems
2000 Innovation Drive
Kanata, Ontario K2K 3E8
CANADA

EMail: zali@cisco.com

Daniel King
Old Dog Consulting
UK

EMail: daniel@olddog.co.uk

Ramon Casellas
CTTC
Av. Carl Friedrich Gauss n7
Castelldefels, Barcelona 08860
SPAIN

EMail: ramon.casellas@cttc.es

Network Working Group
Internet Draft
Intended status: Informational
Expires: January 2012

Y. Lee
Huawei

G. Bernstein
Grotto Networking

Jonas Martensson
Acreo

T. Takeda
NTT

T. Tsuritani
KDDI

July 6, 2011

PCEP Requirements for WSON Routing and Wavelength Assignment

draft-ietf-pce-wson-routing-wavelength-05.txt

Status of this Memo

Copyright Notice

Abstract

This memo provides application-specific requirements for the Path Computation Element communication Protocol (PCEP) for the support of Wavelength Switched Optical Networks (WSON). Lightpath provisioning in WSONs requires a routing and wavelength assignment (RWA) process. From a path computation perspective, wavelength assignment is the process of determining which wavelength can be used on each hop of a path and forms an additional routing constraint to optical light path computation. Requirements for Optical impairments will be addressed in a separate document.

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 0.

Table of Contents

1. Introduction

[RFC4655] defines the PCE based Architecture and explains how a Path Computation Element (PCE) may compute Label Switched Paths (LSP) in Multiprotocol Label Switching Traffic Engineering (MPLS-TE) and Generalized MPLS (GMPLS) networks at the request of Path Computation Clients (PCCs). A PCC is shown to be any network component that makes such a request and may be for instance an Optical Switching Element within a Wavelength Division Multiplexing (WDM) network. The PCE, itself, can be located anywhere within the network, and may be within an optical switching element, a Network Management System (NMS) or Operational Support System (OSS), or may be an independent network server.

The PCE communications Protocol (PCEP) is the communication protocol used between PCC and PCE, and may also be used between cooperating PCEs. [RFC4657] sets out the common protocol requirements for PCEP. Additional application-specific requirements for PCEP are deferred to separate documents.

This document provides a set of application-specific PCEP requirements for support of path computation in Wavelength Switched Optical Networks (WSON). WSON refers to WDM based optical networks in which switching is performed selectively based on the wavelength of an optical signal.

The path in WSON is referred to as a lightpath. A lightpath may span multiple fiber links and the path should be assigned a wavelength for each link. A transparent optical network is made up of optical devices that can switch but not convert from one wavelength to

another. In a transparent optical network, a lightpath operates on the same wavelength across all fiber links that it traverses. In such case, the lightpath is said to satisfy the wavelength-continuity constraint. Two lightpaths that share a common fiber link can not be assigned the same wavelength. To do otherwise would result in both signals interfering with each other. Note that advanced additional multiplexing techniques such as polarization based multiplexing are not addressed in this document since the physical layer aspects are not currently standardized. Therefore, assigning the proper wavelength on a lightpath is an essential requirement in the optical path computation process.

When a switching node has the ability to perform wavelength conversion the wavelength-continuity constraint can be relaxed, and a lightpath may use different wavelengths on different links along its route from origin to destination. It is, however, to be noted that wavelength converters may be limited due to their relatively high cost, while the number of WDM channels that can be supported in a fiber is also limited. As a WSON can be composed of network nodes that cannot perform wavelength conversion, nodes with limited wavelength conversion, and nodes with full wavelength conversion abilities, wavelength assignment is an additional routing constraint to be considered in all lightpath computation.

In this document we first review the processes for routing and wavelength assignment (RWA) used when wavelength continuity constraints are present and then specify requirements for PCEP to support RWA.

The remainder of this document uses terminology from [RFC4655].

1.1. WSON RWA Processes

In [WSON-Frame] three alternative process architectures were given for performing routing and wavelength assignment. These are shown schematically in Figure 1.

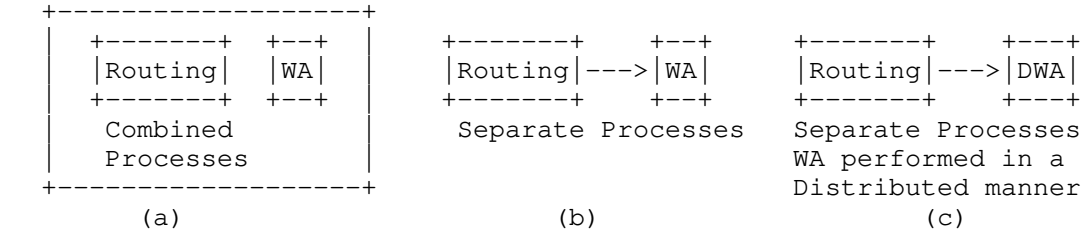


Figure 1 RWA process alternatives.

These alternatives have the following properties and impact on PCEP requirements in this document.

1. Combined Processes (R&WA) - Here path selection and wavelength assignment are performed as a single process. The requirements for PCC-PCE interaction with such a combined RWA process PCE is addressed in this document.

2. Routing separate from Wavelength Assignment (R+WA) - Here the routing process furnishes one or more potential paths to the wavelength assignment process that then performs final path selection and wavelength assignment. The requirements for PCE-PCE interaction with one PCE implementing the routing process and another implementing the wavelength assignment process are not addressed in this document.

3. Routing and distributed Wavelength Assignment (R+DWA) - Here a standard path computation (unaware of detailed wavelength availability) takes place, then wavelength assignment is performed along this path in a distributed manner via signaling (RSVP-TE). This alternative should be covered by existing or emerging GMPLS PCEP extensions and does not present new WSON specific requirements.

2. WSON PCE Architectures and Requirements

In the previous section various process architectures for implementing RWA have been reviewed. Figure 2 shows one typical PCE based implementation, which is referred to as Combined Process (R&WA). With this architecture, the two processes of routing and wavelength assignment are accessed via a single PCE. This architecture is the base architecture from which the requirements are specified in this document.

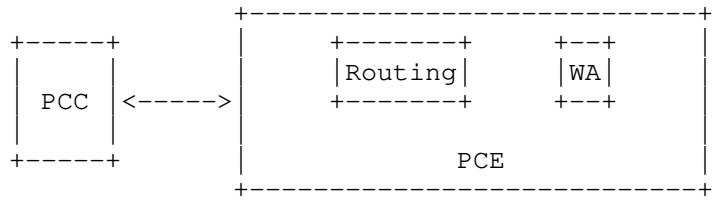


Figure 2 Combined Process (R&WA) architecture

2.1. RWA PCC to PCE Interface

The requirements for the PCC to PCE interface of Figure 2 are specified in this section.

2.1.1. RWA Computation Type and Wavelength Assignment Option

1. The PCReq Message MUST include the path computation type. This can be:

   (i) Both Routing and Wavelength Assignment (RWA), or

   (ii) Routing only.

   This requirement is needed to differentiate between the currently supported routing with distributed wavelength assignment option and combined RWA. In case of distributed wavelength assignment option, wavelength assignment will be performed at each node of the route.

2. When the PCReq Message is RWA path computation type, the PCReq Message MUST further include the wavelength assignment options. At the minimum, the following option should be supported:

   (i) Explicit Label Control (ELC) [RFC4003]

   (ii) Non-Explicit labels in the form of Label Sets (This will allow Distributed WA at a node level where each node would select the wavelength from the Label Sets)

3. The PCRep Message MUST include the route, wavelengths assigned to the route and indication of which wavelength assignment option has been applied (ELC or Label Sets).

4. In the case where a valid path is not found, the PCRep Message MUST include why the path is not found (e.g., no route, wavelength not found, optical quality check failed, etc.)

2.1.2. Bulk RWA path request/reply

1. The PCReq Message MUST be able to specify an option for bulk RWA path request. Bulk path request is an ability to request a number of simultaneous RWA path requests.

2. The PCRep Message MUST include the route, wavelength assigned to the route for each RWA path request specified in the original bulk PCReq Message.

2.1.3. An RWA path re-optimization request/reply

1. For a re-optimization request, the PCReq Message MUST provide the path to be re-optimized and include the following options:

   a. Re-optimize the path keeping the same wavelength(s)

   b. Re-optimize wavelength(s) keeping the same path

   c. Re-optimize allowing both wavelength and the path to change

2. The corresponding PCRep Message for the re-optimized request MUST provide the Re-optimized path and wavelengths.

3. In case that the path is not found, the PCRep Message MUST include why the path is not found (e.g., no route, wavelength not found, both route and wavelength not found, etc.)

2.1.4. Wavelength Range Constraint

For any PCReq Message that is associated with a request for wavelength assignment the requester (PCC) MUST be able to specify a restriction on the wavelengths to be used.

Note that the requestor (PCC) is NOT required to furnish any range restrictions. This restriction is to be interpreted by the PCE as a constraint on the tuning ability of the origination laser transmitter.

2.1.5. Wavelength Policy Constraint

The PCReq Message May include specific operator’s policy information for WA (E.g., random assignment, descending order, ascending order, etc.)

2.1.6. Signal Processing Capability Restriction

The PCReq Message MUST be able to specify restrictions for signal compatibility either on the endpoint or any given link. The following signal processing capability should be supported at a minimum:

- Modulation Type List
- FEC Type List

3. Manageability Considerations

Manageability of WSON Routing and Wavelength Assignment (RWA) with PCE must address the following considerations:

3.1. Control of Function and Policy

In addition to the parameters already listed in Section 8.1 of [RFC5440], a PCEP implementation SHOULD allow configuring the following PCEP session parameters on a PCC:

- The ability to send a WSON RWA request.

In addition to the parameters already listed in Section 8.1 of [RFC5440], a PCEP implementation SHOULD allow configuring the following PCEP session parameters on a PCE:

- The support for WSON RWA.
- The maximum number of bulk path requests associated with WSON RWA per request message.

These parameters may be configured as default parameters for any PCEP session the PCEP speaker participates in, or may apply to a specific session with a given PCEP peer or a specific group of sessions with a specific group of PCEP peers.

3.2. Information and Data Models, e.g. MIB module

As this document only concerns the requirements to support WSON RWA, no additional MIB module is defined in this document. However, the corresponding solution draft will list the information that should be added to the PCE MIB module defined in [PCEP-MIB].

3.3. Liveness Detection and Monitoring

Mechanisms defined in this document do not imply any new liveness detection and monitoring requirements in addition to those already listed in section 8.3 of [RFC5440].

3.4. Verifying Correct Operation

Mechanisms defined in this document do not imply any new verification requirements in addition to those already listed in section 8.4 of [RFC5440]

3.5. Requirements on Other Protocols and Functional Components

The PCE Discovery mechanisms ([RFC5089] and [RFC5088]) may be used to advertise WSON RWA path computation capabilities to PCCs.

3.6. Impact on Network Operation

Mechanisms defined in this document do not imply any new network operation requirements in addition to those already listed in section 8.6 of [RFC5440].

4. Security Considerations

This document has no requirement for a change to the security models within PCEP [RFC5440]. However the additional information distributed in order to address the RWA problem represents a disclosure of network capabilities that an operator may wish to keep private. Consideration should be given to securing this information.

5. IANA Considerations

A future revision of this document will present requests to IANA for codepoint allocation.

6. Acknowledgments

The authors would like to thank Adrian Farrel for many helpful comments that greatly improved the contents of this draft.

This document was prepared using 2-Word-v2.0.template.dot.

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.

[RFC3471] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, January 2003.

[RFC3473] Berger, L., Ed., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.

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

[RFC4657] Ash, J. and J. Le Roux, "Path Computation Element (PCE) Communication Protocol Generic Requirements", RFC 4657, September 2006.

[RFC5440] Vasseur, JP., Ed. and JL. Le Roux, Ed., "Path Computation Element (PCE) communication Protocol", RFC 5440, March 2009.

[PCEP-MIB] Koushik, K, et al., "PCE communication protocol(PCEP) Management Information Base", draft-ietf-pce-pcep-mib, work in progress.

7.2. Informative References

[WSON-IMP] Lee, Y. and Bernstein, G. (Editors), D. Li and G. Martinelli "A Framework for the Control and Measurement of Wavelength Switched Optical Networks (WSON) with Impairments, draft-ietf-ccamp-wson-impairments, work in progress.

[WSON-Frame] Lee, Y. and Bernstein, G. (Editors), and W. Imajuku, "Framework for GMPLS and PCE Control of Wavelength Switched Optical Networks", draft-ietf-ccamp-rwa-wson-framework, work in progress.

[RFC5088] Le Roux, JL., Ed., Vasseur, JP., Ed., Ikejiri, Y., and R. Zhang, "OSPF Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5088, January 2008.

[RFC5089] Le Roux, JL., Ed., Vasseur, JP., Ed., Ikejiri, Y., and R. Zhang, "IS-IS Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5089, January 2008.

Authors' Addresses

Young Lee (Ed.)
Huawei Technologies
1700 Alma Drive, Suite 100
Plano, TX 75075, USA
Phone: (972) 509-5599 (x2240)
Email: ylee@huawei.com

Greg Bernstein (Ed.)
Grotto Networking
Fremont, CA, USA
Phone: (510) 573-2237
Email: gregb@grotto-networking.com

Jonas Martensson
Acreo
Email:Jonas.Martensson@acreo.se

Tomonori Takeda
NTT Corporation
3-9-11, Midori-Cho
Musashino-Shi, Tokyo 180-8585, Japan
Email: takeda.tomonori@lab.ntt.co.jp

Takehiro Tsuritani
KDDI R&D Laboratories, Inc.
2-1-15 Ohara Kamifukuoka Saitama, 356-8502. Japan
Phone: +81-49-278-7357
Email: tsuri@kddilabs.jp

Intellectual Property Statement

Disclaimer of Validity

Acknowledgment

Funding for the RFC Editor function is currently provided by the Internet Society.

Network Working Group
Internet Draft
Intended status: Informational
Expires: April 2015

Y. Lee
Huawei
G. Bernstein
Grotto Networking
Jonas Martensson
Acreo
T. Takeda
NTT
T. Tsuritani
KDDI
O. G. de Dios
Telefonica

October 28, 2014

PCEP Requirements for WSON Routing and Wavelength Assignment

draft-ietf-pce-wson-routing-wavelength-15.txt

Abstract

This memo provides application-specific requirements for the Path Computation Element communication Protocol (PCEP) for the support of Wavelength Switched Optical Networks (WSON). Lightpath provisioning in WSONs requires a routing and wavelength assignment (RWA) process. From a path computation perspective, wavelength assignment is the process of determining which wavelength can be used on each hop of a path and forms an additional routing constraint to optical light path computation. Requirements for PCEP extensions in support of optical impairments will be addressed in a separate document.

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."

Copyright Notice

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].

Table of Contents

1. Introduction

[RFC4655] defines the PCE-based architecture and explains how a Path Computation Element (PCE) may compute Label Switched Paths (LSP) in Multiprotocol Label Switching Traffic Engineering (MPLS-TE) and Generalized MPLS (GMPLS)-controlled networks at the request of Path Computation Clients (PCCs). A PCC is shown to be any network component that makes such a request and may be for instance an optical switching element within a Wavelength Division Multiplexing (WDM) network. The PCE, itself, can be located anywhere within the network, and may be within an optical switching element, a Network Management System (NMS) or Operational Support System (OSS), or may be an independent network server.

The PCE communication Protocol (PCEP) is the communication protocol used between PCC and PCE, and may also be used between cooperating PCEs. [RFC4657] sets out the common protocol requirements for PCEP. Additional application-specific requirements for PCEP are deferred to separate documents.

This document provides a set of application-specific PCEP requirements for support of path computation in Wavelength Switched Optical Networks (WSON). WSON refers to WDM-based optical networks in which switching is performed selectively based on the wavelength of an optical signal.

The path in WSON is referred to as a lightpath. A lightpath may span multiple fiber links and the path should be assigned a wavelength for each link.

A transparent optical network is made up of optical devices that can switch but not convert from one wavelength to another. In a transparent optical network, a lightpath operates on the same wavelength across all fiber links that it traverses. In such case, the lightpath is said to satisfy the wavelength-continuity

constraint. Two lightpaths that share a common fiber link cannot be assigned the same wavelength. To do otherwise would result in both signals interfering with each other. Note that advanced additional multiplexing techniques such as polarization based multiplexing are not addressed in this document since the physical layer aspects are not currently standardized. Therefore, assigning the proper wavelength on a lightpath is an essential requirement in the optical path computation process.

When a switching node has the ability to perform wavelength conversion the wavelength-continuity constraint can be relaxed, and a lightpath may use different wavelengths on different links along its path from origin to destination. It is, however, to be noted that wavelength converters may be limited for cost reasons, while the number of WDM channels that can be supported in a fiber is also limited. As a WSON can be composed of network nodes that cannot perform wavelength conversion, nodes with limited wavelength conversion, and nodes with full wavelength conversion abilities, wavelength assignment is an additional routing constraint to be considered in all lightpath computations.

In this document we first review the processes for routing and wavelength assignment (RWA) used when wavelength continuity constraints are present and then specify requirements for PCEP to support RWA. Requirements for optical impairments will be addressed in a separate document.

The remainder of this document uses terminology from [RFC4655].

2. WSON RWA Processes & Architecture

In [RFC6163] three alternative process architectures were given for performing routing and wavelength assignment. These are shown schematically in Figure 1. R stands for Routing, WA for Wavelength Assignment, and DWA for Distributed Wavelength Assignment.

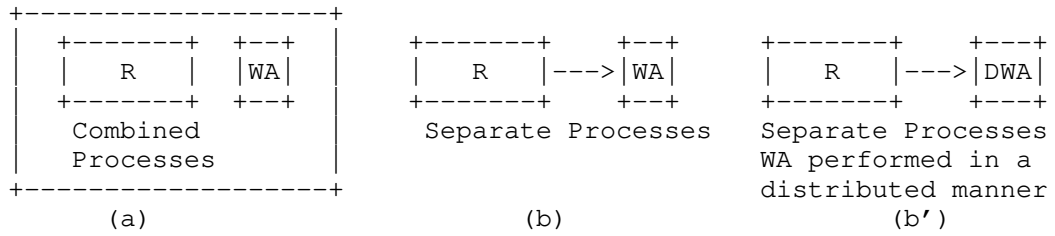


Figure 1. RWA process alternatives

These alternatives have the following properties and impact on PCEP requirements in this document.

(a) Combined Processes (R&WA)

Here path selection and wavelength assignment are performed as a single process. The requirements for PCC-PCE interaction with such a combined RWA process PCE is addressed in this document.

(b) Routing separate from Wavelength Assignment (R+WA)

Here the routing process furnishes one or more potential paths to the wavelength assignment process that then performs final path selection and wavelength assignment. The requirements for PCE-PCE interaction with one PCE implementing the routing process and another implementing the wavelength assignment process are not addressed in this document.

(b') Routing and distributed Wavelength Assignment (R+DWA)

Here a standard path computation (unaware of detailed wavelength availability) takes place, then wavelength assignment is performed along this path in a distributed manner via signaling (RSVP-TE). This alternative is a particular case of R+WA and it should be covered by GMPLS PCEP extensions and does not present new WSON-specific requirements.

In the previous section various process architectures for implementing RWA have been reviewed. Figure 2 shows one typical PCE-based implementation, which is referred to as Combined Process (R&WA). With this architecture, the two processes of routing and wavelength assignment are accessed via a single PCE. This architecture is the base architecture from which the requirements are specified in this document.

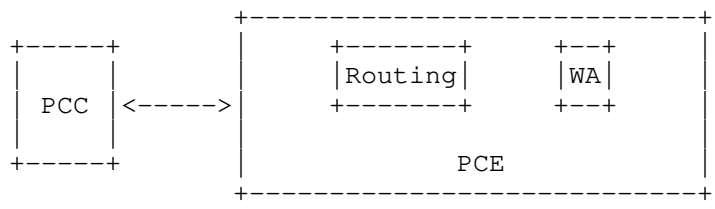


Figure 2. Combined Process (R&WA) architecture

3. Requirements

The requirements for the PCC to PCE interface of Figure 2 are specified in this section.

3.1. Path Computation Type Option

A PCEP request MAY include the path computation type. This can be:

(i) Both Routing and Wavelength Assignment (RWA),

(ii) Routing only.

This requirement is needed to differentiate between the currently supported routing with distributed wavelength assignment option and combined RWA. In case of distributed wavelength assignment option, wavelength assignment will be performed at each node of the route.

3.2. RWA Processing

(a) When the request is a RWA path computation type, the request MUST further include the wavelength assignment options. At the minimum, the following option should be supported:

(i) Explicit Label Control (ELC) [RFC3473]

(ii) A set of recommended labels for each hop. The PCC can select the label based on local policy.

Note that option (ii) may also be used in R+WA or R+DWA.

(b) In case of a RWA computation type, the response MUST include the wavelength(s) assigned to the path and an indication of which label assignment option has been applied (ELC or label set).

(c) In the case where a valid path is not found, the response MUST include why the path is not found (e.g., network disconnected, wavelength not found, or both, etc.). Note that 'wavelength not found' may include several sub-cases such as wavelength continuity not met, unsupported FEC/Modulation type, etc.

3.3. Bulk RWA Path Request/Reply

Sending simultaneous path requests for "routing only" computation is supported by PCEP specification [RFC5440]. To remain consistent the following requirements are added.

(a) A PCEP request MUST be able to specify an option for bulk RWA path request. Bulk path request is an ability to request a number of simultaneous RWA path requests.

(b) The PCEP response MUST include the path and the assigned wavelength assigned for each RWA path request specified in the original bulk request.

3.4. RWA Path Re-optimization Request/Reply

1. For a re-optimization request, the request MUST provide both the path and current wavelength to be re-optimized and MAY include the following options:

    a. Re-optimize the path keeping the same wavelength(s)

    b. Re-optimize wavelength(s) keeping the same path

    c. Re-optimize allowing both the wavelength and the path to change

2. The corresponding response to the re-optimized request MUST provide the re-optimized path and wavelengths even when the request asked for the path or the wavelength to remain unchanged.

3. In case that the new path is not found, the response MUST include why the path is not found (e.g., network disconnected, wavelength not found, or both, etc.). Note that 'wavelength not found' may include several sub-cases such as wavelength continuity not met, unsupported FEC/Modulation type, etc.

3.5. Wavelength Range Constraint

For any RWA computation type request, the requester (PCC) MUST be allowed to specify a restriction on the wavelengths to be used. The requester MAY use this option to restrict the assigned wavelength for explicit label or label set. This restriction may for example come from the tuning ability of a laser transmitter, any optical element, or a policy-based restriction.

Note that the requester (e.g., PCC) is not required to furnish any range restrictions.

3.6. Wavelength Assignment Preference

1. A RWA computation type request MAY include the requester preference for, e.g., random assignment, descending order, ascending order, etc. A response SHOULD follow the requestor preference unless it conflicts with operator's policy.

2. A request for two or more paths MUST allow the requester to include an option constraining the paths to have the same wavelength(s) assigned. This is useful in the case of protection with single transponder (e.g., 1+1 link disjoint paths).

In a network with wavelength conversion capabilities (e.g. sparse 3R regenerators), a request SHOULD be able to indicate whether a single, continuous wavelength should be allocated or not. In other words, the requesting PCC SHOULD be able to specify the precedence of wavelength continuity even if wavelength conversion is available.

3.7. Signal Processing Capability Restriction

Signal processing compatibility is an important constraint for optical path computation. The signal type for an end-to-end optical path must match at source and at destination.

The PCC MUST be allowed to specify the signal type at the endpoints (i.e., at source and at destination). The following signal processing capabilities should be supported at a minimum:

- Modulation Type List
- FEC Type List

The PCC MUST also be allowed to state whether transit modification is acceptable for the above signal processing capabilities.

4. Manageability Considerations

Manageability of WSON Routing and Wavelength Assignment (RWA) with PCE must address the following considerations:

4.1. Control of Function and Policy

In addition to the parameters already listed in Section 8.1 of [RFC5440], a PCEP implementation SHOULD allow configuring the following PCEP session parameters on a PCC:

- The ability to send a WSON RWA request.

In addition to the parameters already listed in Section 8.1 of [RFC5440], a PCEP implementation SHOULD allow configuring the following PCEP session parameters on a PCE:

- The support for WSON RWA.
- The maximum number of bulk path requests associated with WSON RWA per request message.

These parameters may be configured as default parameters for any PCEP session the PCEP speaker participates in, or may apply to a specific session with a given PCEP peer or a specific group of sessions with a specific group of PCEP peers.

4.2. Information and Data Models, e.g. MIB module

As this document only concerns the requirements to support WSON RWA, no additional MIB module is defined in this document. However, the corresponding solution draft will list the information that should be added to the PCE MIB module defined in [PCEP-MIB].

4.3. Liveness Detection and Monitoring

No new mechanism is defined in this document that implies any new liveness detection and monitoring requirements in addition to those already listed in section 8.3 of [RFC5440].

4.4. Verifying Correct Operation

No new mechanism is defined in this document that implies any new verification requirements in addition to those already listed in section 8.4 of [RFC5440]

4.5. Requirements on Other Protocols and Functional Components

If PCE discovery mechanisms ([RFC5089] and [RFC5088]) were to be extended for technology-specific capabilities, advertising WSON RWA path computation capability should be considered.

4.6. Impact on Network Operation

No new mechanism is defined in this document that implies any new network operation requirements in addition to those already listed in section 8.6 of [RFC5440].

5. Security Considerations

This document has no requirement for a change to the security models within PCEP [RFC5440]. However the additional information distributed in order to address the RWA problem represents a disclosure of network capabilities that an operator may wish to keep private. Consideration should be given to securing this information.

Solutions that address the requirements in this document need to verify that existing PCEP security mechanisms adequately protect the additional network capabilities and must include new mechanisms as necessary.

6. IANA Considerations

This informational document does not make any requests for IANA action.

7. Acknowledgments

The authors would like to thank Adrian Farrel, Cycil Margaria and Ramon Casellas for many helpful comments that greatly improved the contents of this draft.

This document was prepared using 2-Word-v2.0.template.dot.

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.

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

[RFC5440] Vasseur, JP., Ed. and JL. Le Roux, Ed., "Path Computation Element (PCE) communication Protocol", RFC 5440, March 2009.

8.2. Informative References

[RFC3473] L. Berger, "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.

[RFC4657] Ash, J. and J. Le Roux, "Path Computation Element (PCE) Communication Protocol Generic Requirements", RFC 4657, September 2006.

[RFC6163] Y. Lee, G. Bernstein, W. Imajuku, "Framework for GMPLS and PCE Control of Wavelength Switched Optical Networks", RFC 6163, April 2011.

[RFC5088] Le Roux, JL., Ed., Vasseur, JP., Ed., Ikejiri, Y., and R. Zhang, "OSPF Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5088, January 2008.

[RFC5089] Le Roux, JL., Ed., Vasseur, JP., Ed., Ikejiri, Y., and R. Zhang, "IS-IS Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5089, January 2008.

[PCEP-MIB] Koushik, K, et al., "PCE communication protocol(PCEP) Management Information Base", draft-ietf-pce-pcep-mib, work in progress.

Authors' Addresses

Young Lee (Ed.)
Huawei Technologies
5340 Legacy Drive, Building 3
Plano, TX 75245, USA
Phone: (469)277-5838
Email: leeyoung@huawei.com

Greg Bernstein (Ed.)
Grotto Networking
Fremont, CA, USA
Phone: (510) 573-2237
Email: gregb@grotto-networking.com

Jonas Martensson
Acreo
Email:Jonas.Martensson@acreo.se

Tomonori Takeda
NTT Corporation
3-9-11, Midori-Cho
Musashino-Shi, Tokyo 180-8585, Japan
Email: takeda.tomonori@lab.ntt.co.jp

Takehiro Tsuritani
KDDI R&D Laboratories, Inc.
2-1-15 Ohara Kamifukuoka Saitama, 356-8502. Japan
Phone: +81-49-278-7357
Email: tsuri@kddilabs.jp

Oscar Gonzalez de Dios
Telefonica Investigacion y Desarrollo
C/ Emilio Vargas 6
Madrid, 28043
Spain
Phone: +34 91 3374013
Email: ogondio@tid.es

Network Working Group
Internet Draft

Y. Lee, Ed.
Huawei Technologies

Intended status: Standard
Expires: April 2012

R. Casellas, Ed.
CTTC

October 31, 2011

PCEP Extension for WSON Routing and Wavelength Assignment

draft-lee-pce-wson-rwa-ext-03.txt

Abstract

This draft provides the Path Computation Element communication Protocol (PCEP) extensions for the support of Routing and Wavelength Assignment (RWA) in Wavelength Switched Optical Networks (WSON). Lightpath provisioning in WSONs requires a routing and wavelength assignment (RWA) process. From a path computation perspective, wavelength assignment is the process of determining which wavelength can be used on each hop of a path and forms an additional routing constraint to optical light path computation.

Table of Contents

1. Terminology

This document uses the terminology defined in [RFC4655], and [RFC5440].

2. Requirements Language

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

3. Introduction

[RFC4655] defines the PCE based Architecture and explains how a Path Computation Element (PCE) may compute Label Switched Paths (LSP) in Multiprotocol Label Switching Traffic Engineering (MPLS-TE) and Generalized MPLS (GMPLS) networks at the request of Path Computation Clients (PCCs). A PCC is shown to be any network component that makes such a request and may be for instance an Optical Switching Element within a Wavelength Division Multiplexing (WDM) network. The PCE, itself, can be located anywhere within the network, and may be within an optical switching element, a Network Management System (NMS) or Operational Support System (OSS), or may be an independent network server.

The PCE communications Protocol (PCEP) is the communication protocol used between PCC and PCE, and may also be used between cooperating PCEs. [RFC4657] sets out the common protocol requirements for PCEP. Additional application-specific requirements for PCEP are deferred to separate documents.

This document provides the PCEP extension for the support of Routing and Wavelength Assignment (RWA) in Wavelength Switched Optical Networks (WSON) based on the requirements specified in [PCE-RWA].

WSON refers to WDM based optical networks in which switching is performed selectively based on the wavelength of an optical signal. In this document, it is assumed that wavelength converters require electrical signal regeneration. Consequently, WSONs can be transparent (A transparent optical network is made up of optical devices that can switch but not convert from one wavelength to another, all within the optical domain) or translucent (3R regenerators are sparsely placed in the network).

A LSC Label Switched Path (LSP) may span one or several transparent segments, which are delimited by 3R regenerators (typically with electronic regenerator and optional wavelength conversion). Each transparent segment or path in WSON is referred to as an optical path. An optical path may span multiple fiber links and the path should be assigned the same wavelength for each link. In such case, the optical path is said to satisfy the wavelength-continuity constraint. Figure 1 illustrates the relationship between a LSC LSP and transparent segments (optical paths).

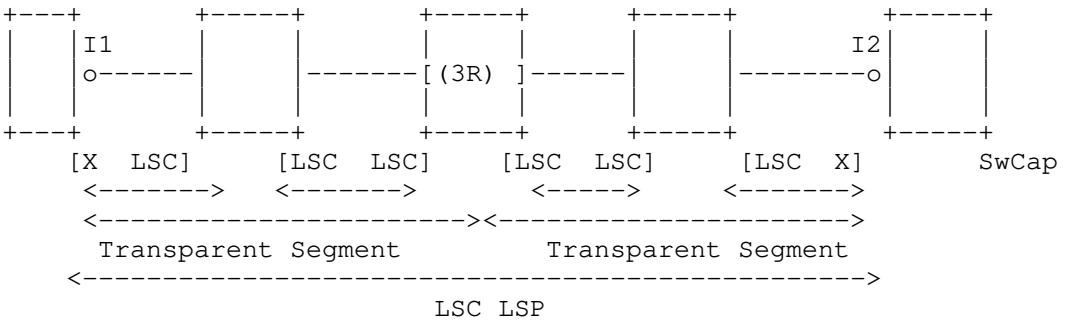


Figure 1 Illustration of a LSC LSP and transparent segments

Note that two optical paths within a WSON LSP need not operate on the same wavelength (due to the wavelength conversion capabilities). Two optical paths that share a common fiber link cannot be assigned the same wavelength. To do otherwise would result in both signals interfering with each other. Note that advanced additional multiplexing techniques such as polarization based multiplexing are not addressed in this document since the physical layer aspects are

not currently standardized. Therefore, assigning the proper wavelength on a lightpath is an essential requirement in the optical path computation process.

When a switching node has the ability to perform wavelength conversion, the wavelength-continuity constraint can be relaxed, and a LSC Label Switched Path (LSP) may use different wavelengths on different links along its route from origin to destination. It is, however, to be noted that wavelength converters may be limited due to their relatively high cost, while the number of WDM channels that can be supported in a fiber is also limited. As a WSON can be composed of network nodes that cannot perform wavelength conversion, nodes with limited wavelength conversion, and nodes with full wavelength conversion abilities, wavelength assignment is an additional routing constraint to be considered in all lightpath computation.

For example, within a translucent WSON, a LSC LSP may be established between interfaces I1 and I2, spanning 2 transparent segments (optical paths) where the wavelength continuity constraint applies (i.e. the same unique wavelength MUST be assigned to the LSP at each TE link of the segment). If the LSC LSP induced a Forwarding Adjacency / TE link, the switching capabilities of the TE link would be [X X] where X < LSC (PSC, TDM, ...).

[Ed note: in general, WSON LSC may not be the only switching layer with switching constraints. From a GMPLS/PCEP perspective, wavelength assignment corresponds to label allocation. This document should align with GMPLS extensions for PCEP. Wavelength restrictions and constraints should be formulated in terms of labels (i.e. LABEL_SET, SUGGESTED_LABEL, UPSTREAM_LABEL, etc.)In addition to those label switching constraints, each optical path is constrained by the optical signal quality. The optical signal quality depends first on the optical sender and receiver capabilities. Second contributors to optical signal constraints are the optical elements used on the path (optical fibers, amplifiers, boosters, optical components). All those elements have an impact on the optical signal quality that limits the ability of the optical path to carry traffic. In order to improve the signal quality and limit some optical effects several advanced modulation processing are used. Those modulation properties contribute not only to optical signal quality checks but also constrain the selection of sender and receiver, as they should have matching signal processing capabilities.

The optical modulation properties, also referred to as signal compatibility, are already considered in signaling in [RWA-Encode] and [WSON-OSPF].

This document includes signal compatibility constraint as part of RWA path computation. That is, the signal processing capabilities (e.g., modulation and FEC) must be compatible between the sender and the receiver of the optical path across all optical elements.

This document, however, does not address optical impairments as part of RWA path computation. See [WSON-Imp] and [PSVP-Imp] for more information on optical impairments and GMPLS.

4. Encoding of a RWA Path Request

Figure 2 shows one typical PCE based implementation, which is referred to as Combined Process (R&WA). With this architecture, the two processes of routing and wavelength assignment are accessed via a single PCE. This architecture is the base architecture from which the requirements have been specified in [PCE-RWA] and the PCEP extensions that are going to be specified in this document based on this architecture.

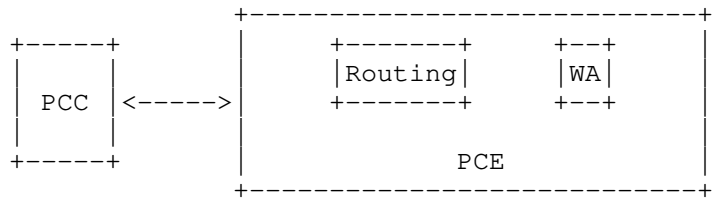


Figure 2 Combined Process (R&WA) architecture

4.1. Wavelength Assignment (WA) Object

The current RP object is used to indicate routing related information in a new path request per [RFC5440]. Since a new RWA path request involves both routing and wavelength assignment, the wavelength assignment related information in the request SHOULD be coupled in the path request.

Wavelength allocation can be performed by the PCE by different means:

(a) By means of Explicit Label Control, in the sense that one (or two) allocated labels MAY appear after an interface route subobject.

(b) By means of a Label Set, containing one or more allocated Labels, provided by the PCE.

Option (b) allows distributed label allocation (performed during signaling) to complete wavelength assignment.

Additionally, given a range of potential labels to allocate, the request SHOULD convey the heuristic / mechanism to the allocation, including vendor-specific approaches.

The format of a PCReq message after incorporating the WA object is as follows:

<PCReq Message> ::= <Common Header>

[<svec-list>]

<request-list>

Where:

<request-list>::=<request>[<request-list>]

<request>::= <RP>

<ENDPOINTS>

<WA>

[other optional objects...]

If WA object is present in the request, the WA object MUST be encoded after the ENDPOINTS object.

The format of the Wavelength Assignment (WA) object body is as follows:

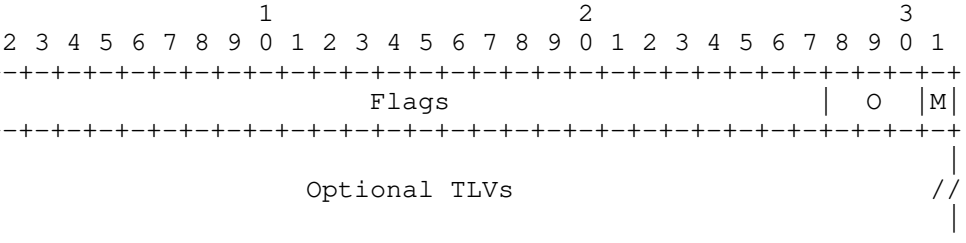


Figure 3 WA Object

o Flags (32 bits)

The following new flags SHOULD be set

. M (Mode - 1 bit): M bit is used to indicate the mode of wavelength assignment. When M bit is set to 1, this indicates that the label assigned by the PCE must be explicit. That is, the selected way to convey the allocated wavelength is by means of Explicit Label Control (ELC) [RFC4003] for each hop of a computed LSP. Otherwise, the label assigned by the PCE needs not be explicit (i.e., it can be suggested in the form of label set objects in the corresponding response, to allow distributed WA. In such case, the PCE MUST return a Label_Set object in the response if the path is found.

(Ed note: When the distributed WA is applied, some specific wavelength range and/or the maximum number of wavelengths to be returned in the Label Set might be additionally indicated. The optional TLV field will be used for conveying this additional request. The details of this encoding will be provided in a later revision.)

. O (Order - 3 bits): O bit is used to indicate the wavelength assignment constraint in regard to the order of wavelength assignment to be returned by the PCE. This case is only applied when M bit is set to "explicit." The following indicators should be defined:

000 - Reserved

001 - Random Assignment

010 - First Fit (FF) in descending Order

011 - First Fit (FF) in ascending Order

100 - Last Fit (LF) in ascending Order

101 - Last Fit (LF) in descending Order

110 - Vendor Specific/Private

111 - Reserved

When the Order bit is set for "Vendor Specific/Private", the optional TLV field will be used to indicate specifics of the order algorithm applied by the PCE.

4.2. Wavelength Restriction Constraint TLV

For any request that contains a wavelength assignment, the requester (PCC) MUST be able to specify a restriction on the wavelengths to be used. This restriction is to be interpreted by the PCE as a constraint on the tuning ability of the origination laser transmitter or on any other maintenance related constraints. Note that if the LSP LSC spans different segments, the PCE MUST have mechanisms to know the tunability restrictions of the involved wavelength converters / regenerators, e.g. by means of the TED either via IGP or NMS. Even if the PCE knows the tunability of the transmitter, the PCC MUST be able to apply additional constraints to the request.

[Ed note: Which PCEP Object will home this TLV is yet to be determined. Since this involves the end-point, The END-POINTS Object might be a good candidate to encode this TLV, which will be provided in a later revision.]

[Ed note: The current encoding assumes that tunability restriction applied to link-level.]

The TLV type is TBD, recommended value is TBD. This TLV MAY appear more than once to be able to specify multiple restrictions.

The TLV data is defined as follows:

<Wavelength Restriction Constraint> ::=

<Action> <Format> <Reserved>

(<Link Identifiers> <Wavelength Restriction>)...

Where

<Link Identifiers> ::=

<Unnumbered IF ID> | <IPV4 Address> | <IPV6 Address>

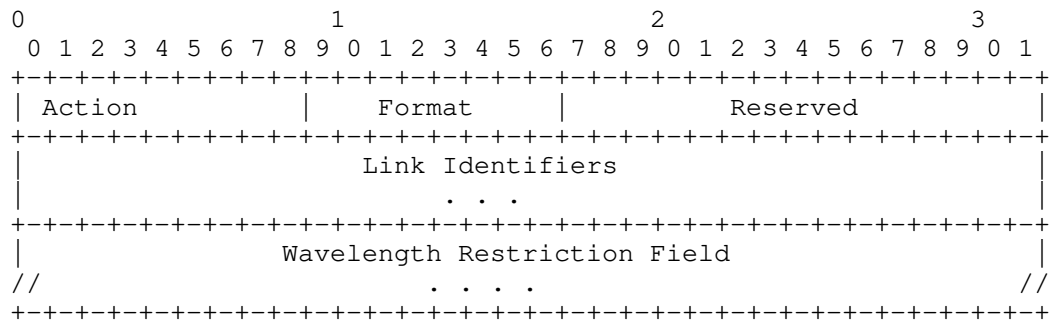


Figure 4 Wavelength Restriction

o Action: 8 bits

. 0 - Inclusive List indicates that one or more link identifiers are included in the Link Set. Each identifies a separate link that is part of the set.

. 1 - Inclusive Range indicates that the Link Set defines a range of links. It contains two link identifiers. The first identifier indicates the start of the range (inclusive). The second identifier indicates the end of the range (inclusive). All links with numeric values between the bounds are considered to be part of the set. A value of zero in either position indicates that there is no bound on the corresponding portion of the range. Note that the Action field can be set to 0 when unnumbered link identifier is used.

Note that "interfaces" such as those discussed in the Interfaces MIB [RFC2863] are assumed to be bidirectional.

o Format: The format of the link identifier (8 bits)

. 0 -- Unnumbered Link Identifier
. 1 -- Local Interface IPv4 Address
. 2 -- Local Interface IPv6 Address
. Others TBD.

Note that all link identifiers in the same list must be of the same type.

o Reserved: Reserved for future use (16 bits)

o Link Identifiers: Identifies each link ID for which restriction is applied. The length is dependent on the link format. See the following section for Link Identifier encoding.

4.2.1. Link Identifier sub-TLV

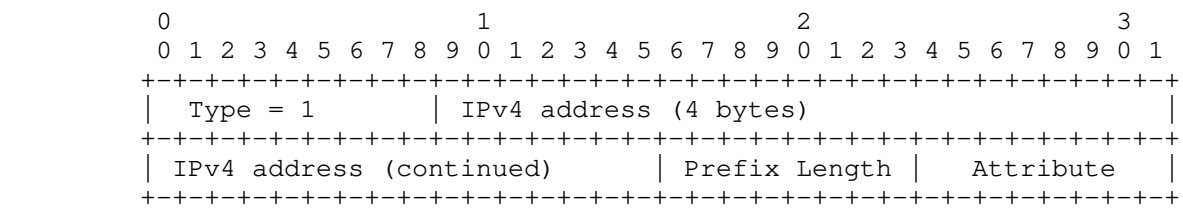
The link identifier field can be an IPv4, IPv6 or unnumbered interface ID.

<Link Identifier> ::=

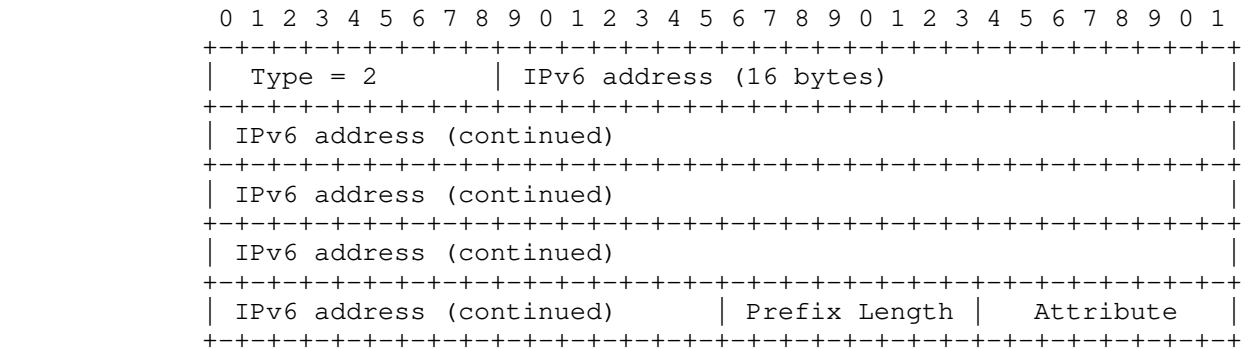
<IPV4 Address> | <IPV6 Address> | <Unnumbered IF ID>

The encoding of each case is as follows:

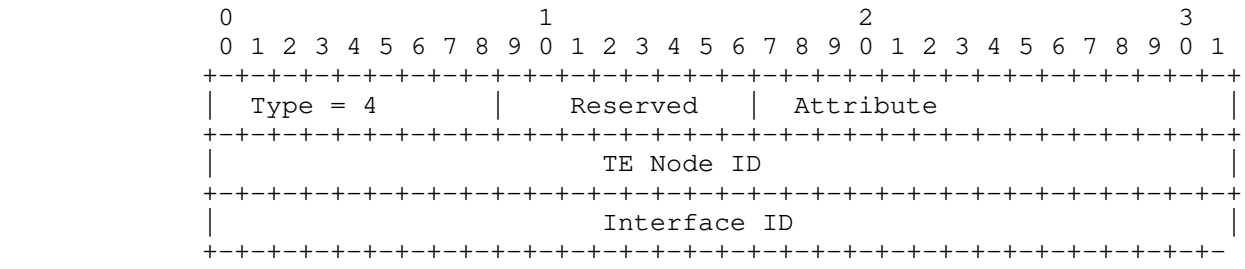
IPv4 prefix Sub-TLV



IPv6 prefix Sub-TLV

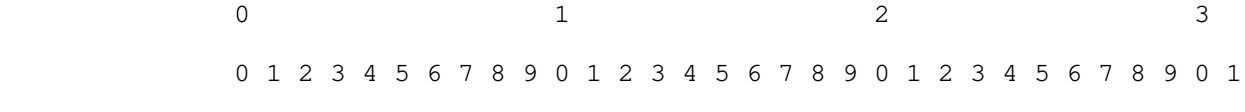


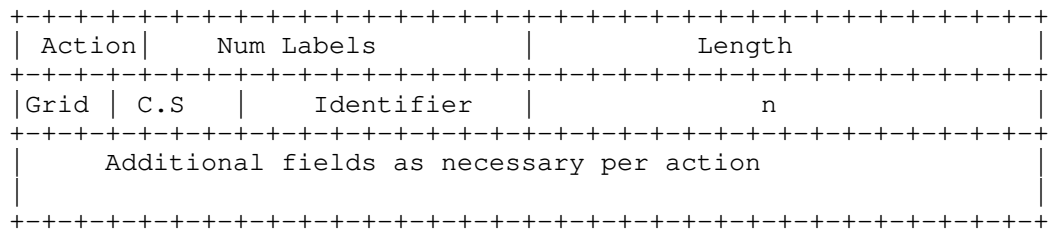
Unnumbered Interface ID Sub-TLV



4.2.2. Wavelength Restriction Field sub-TLV

The Wavelength Restriction Field of the wavelength restriction TLV is encoded as a Label Set field as specified in [GEN-Encode] section 2.2, as shown below, with base label encoded as a 32 bit LSC label, defined in [RFC6205]. See [RFC6205] for a description of Grid, C.S, Identifier and n, as well as [GEN-Encode] for the details of each action.





4.3. Signal processing capability restrictions

Path computation for WSON include the check of signal processing capabilities, those capability MAY be provided by the IGP, however this is not a MUST. Moreover, a PCC should be able to indicate additional restrictions for those signal compatibility, either on the endpoint or any given link.

The supported signal processing capabilities are the one described in [RWA-Info]:

- Modulation Type List
- FEC Type List
- Bit rate
- Client signal

The Bit-rate restriction is already expressed in [PCEP-GMPLS] in the GENERALIZED-BANDWIDTH object.

The client signal information can be expressed using the REQ-ADAP-CAP object from the [PCEP-Layer].

In order to support the Modulation and FEC information two new TLV are introduced as endpoint-restriction in the END-POINTS type Generalized endpoint:

- Modulation restriction TLV
- FEC restriction TLV.

The END-POINTS type generalized endpoint is extended as follow:

<endpoint-restrictions> ::= <LABEL-REQUEST>

<label-restriction-list>

[<signal-compatibility-restriction>...]

Where

signal-compatibility-restriction ::=

<MODULATION-FORMAT-LIST>|<FEC-LIST>

The MODULATION-FORMAT-LIST and FEC-LIST TLV are described in the following sections.

4.3.1. MODULATION-FORMAT-LIST Restriction TLV

This optional TLV represents a modulation format restriction. The TLV type is TBD, recommended value 17.

The TLV data is defined as follow:

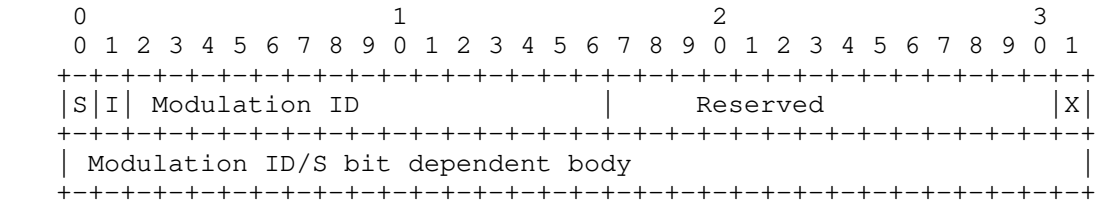


Figure 5 Modulation-Format Field

The format follows the definition from [WSON-Encode] section 5.2. The X bit is set to 1 to exclude the Modulation format, the X bit is set to 0 to include the modulation format.

4.3.2. FEC-LIST Restriction TLV

This optional TLV represents a FEC restriction. The TLV type is TBD, recommended value 18.

The TLV data is defined as follow:

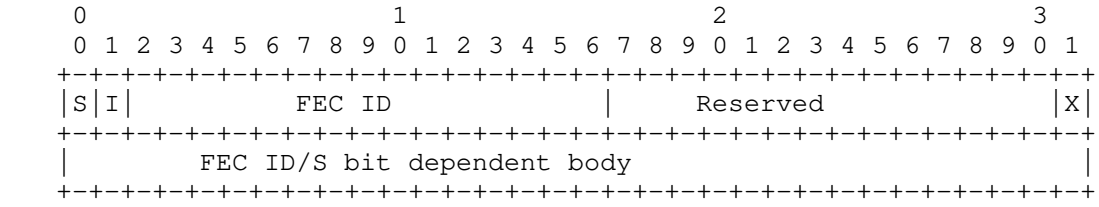


Figure 6 FEC Field

The format follows the definition from [WSON-Encode] section 5.3. The X bit is set to 1 to exclude the FEC; the X bit is set to 0 to include the FEC.

4.3.3. Signal Processing Exclusion XRO Sub-Object

The PCC/PCE should be able to exclude particular types of signal processing along the path in order to handle client restriction or multi-domain path computation.

In order to support the exclusion a new XRO sub-object is defined: the signal processing exclusion:

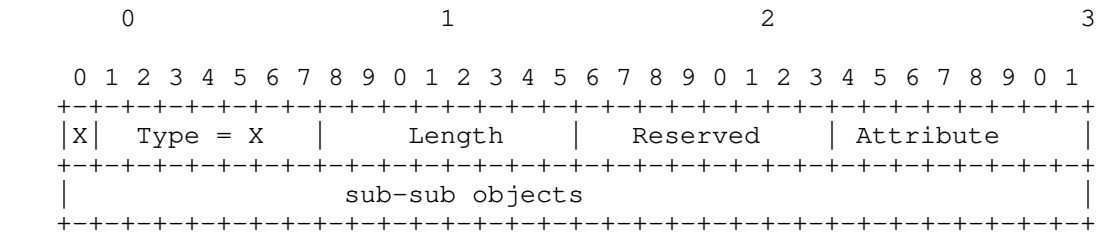


Figure 7 Signaling Processing XRO Sub-Object

The Attribute field indicates how the exclusion sub-object is to be interpreted. The Attribute can only be 0 (Interface) or 1 (Node).

The sub-sub objects are encoded as in RSVP signaling definition [WSON-Sign].

4.3.4. IRO sub-object: signal processing inclusion

Similar to the XRO sub-object the PCC/PCE should be able to include particular types of signal processing along the path in order to handle client restriction or multi-domain path computation.

This is supported by adding the sub-object "processing" defined for ERO in [WSON-Sign] to the PCEP IRO object.

4.3.5. Objective Functions

TBD. [Ed Note: consider a separate draft]

In WSON and DWDM networks the signal is not descrete but has a given spectrum. The spectrum depends in current standard on the channel spacing. In the context of RWA it is important to take into accountthis aspects, as optical effect can cause some cross-talk between the different signals, so it can be desired to reduce it by having sufficient space between the signals. In networks where the channel spacing is constant this can be expressed by the distance between two frequency but also by considering the free spectrum between signals, which is more general and is more important in networks where different system work on different channel spacing.

In order to have a generic expression of this aspect new objective functions are introduced in order to indicate on which criteria the path should be chosen. The following quantities are defined:

- smallest free spectrum on link L is denoted s(L).
- biggest free spectrum on link L is denoted S(L).
- total spectrum width remaining on link L is denoted ST(L)

The Following new objective functions are defined (Objective Function Codes: TBA)

- Maximum residual spectrum

Description: Find a Path such that (Max { (s(Li))), i=1...N}) is minimized.

. Minimize the free spectrum

Description: Find a Path such that (Max { (S(Li)), i=1...N}) is maximized.

. Maximize the remaining total spectrum.

Description: Find a Path such that (Max { (ST(Li)), i=1...N}) is maximized.

5. Encoding of a RWA Path Reply

The ERO is used to encode the path of a TE LSP through the network. The ERO is carried within a given path of a PCEP response, which is in turn carried in a PCRep message to provide the computed TE LSP if the path computation was successful. The preferred way to convey the allocated wavelength is by means of Explicit Label Control (ELC) [RFC4003].

In order to encode wavelength assignment, the Wavelength Assignment (WA) Object needs to be employed to be able to specify wavelength assignment. Since each segment of the computed optical path is associated with wavelength assignment, the WA Object should be aligned with the ERO object.

Encoding details will be provided further revisions and will be aligned as much as possible with [WSON-Sign].

5.1. Error Indicator

To indicate errors associated with the RWA request, a new Error Type (TDB) and subsequent error-values are defined as follows for inclusion in the PCEP-ERROR Object:

A new Error-Type (TDB) and subsequent error-values are defined as follows:

. Error-Type=TBD; Error-value=1: if a PCE receives a RWA request and the PCE is not capable of processing the request due to insufficient memory, the PCE MUST send a PCErr message with a PCEP-ERROR Object (Error-Type=TDB) and an Error-value(Error-value=1). The PCE stops processing the request. The corresponding RWA request MUST be cancelled at the PCC.

. Error-Type=TBD; Error-value=2: if a PCE receives a RWA request and the PCE is not capable of RWA computation, the PCE MUST send a PCErr message with a PCEP-ERROR Object (Error-Type=15) and an Error-value (Error-value=2). The PCE stops processing the request. The corresponding RWA computation MUST be cancelled at the PCC.

5.2. NO-PATH Indicator

To communicate the reason(s) for not being able to find RWA for the path request, the NO-PATH object can be used in the PCRep message. The format of the NO-PATH object body is defined in [RFC5440]. The object may contain a NO-PATH-VECTOR TLV to provide additional information about why a path computation has failed.

Two new bit flags are defined to be carried in the Flags field in the NO-PATH-VECTOR TLV carried in the NO-PATH Object.

. Bit TDB: When set, the PCE indicates no feasible route was found that meets all the constraints associated with RWA.

. Bit TDB: When set, the PCE indicates that no wavelength was assigned to at least one hop of the route in the response.

. Bit TDB: When set, the PCE indicate that no path was found satisfying the signal compatibility constraints.

6. Manageability Considerations

Manageability of WSON Routing and Wavelength Assignment (RWA) with PCE must address the following considerations:

6.1. Control of Function and Policy

In addition to the parameters already listed in Section 8.1 of [PCEP], a PCEP implementation SHOULD allow configuring the following PCEP session parameters on a PCC:

. The ability to send a WSON RWA request.

In addition to the parameters already listed in Section 8.1 of [PCEP], a PCEP implementation SHOULD allow configuring the following PCEP session parameters on a PCE:

- The support for WSON RWA.
- A set of WSON RWA specific policies (authorized sender, request rate limiter, etc).

These parameters may be configured as default parameters for any PCEP session the PCEP speaker participates in, or may apply to a specific session with a given PCEP peer or a specific group of sessions with a specific group of PCEP peers.

6.2. Information and Data Models, e.g. MIB module

Extensions to the PCEP MIB module defined in [PCEP-MIB] should be defined, so as to cover the WSON RWA information introduced in this document. A future revision of this document will list the information that should be added to the MIB module.

6.3. Liveness Detection and Monitoring

Mechanisms defined in this document do not imply any new liveness detection and monitoring requirements in addition to those already listed in section 8.3 of [RFC5440].

6.4. Verifying Correct Operation

Mechanisms defined in this document do not imply any new verification requirements in addition to those already listed in section 8.4 of [RFC5440]

6.5. Requirements on Other Protocols and Functional Components

The PCE Discovery mechanisms ([RFC5089] and [RFC5088]) may be used to advertise WSON RWA path computation capabilities to PCCs.

6.6. Impact on Network Operation

Mechanisms defined in this document do not imply any new network operation requirements in addition to those already listed in section 8.6 of [RFC5440].

7. Security Considerations

This document has no requirement for a change to the security models within PCEP [PCEP]. However the additional information distributed in order to address the RWA problem represents a disclosure of network capabilities that an operator may wish to keep private. Consideration should be given to securing this information.

8. IANA Considerations

A future revision of this document will present requests to IANA for codepoint allocation.

9. Acknowledgments

The authors would like to thank Adrian Farrel for many helpful comments that greatly improved the contents of this draft.

This document was prepared using 2-Word-v2.0.template.dot.

10. References

10.1. Informative 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 (GMPLS) Signaling Functional Description", RFC 3471, January 2003.

[RFC3473] Berger, L., Ed., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.

[RFC3477] Kompella, K. and Y. Rekhter, "Signalling Unnumbered Links in Resource ReSerVation Protocol - Traffic Engineering (RSVP-TE)", RFC 3477, January 2003.

[RFC4003] Berger, L., "GMPLS Signaling Procedure for Egress Control", RFC 4003, February 2005.

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

[RFC4657] Ash, J. and J. Le Roux, "Path Computation Element (PCE) Communication Protocol Generic Requirements", RFC 4657, September 2006.

[RFC5440] Vasseur, JP., Ed. and JL. Le Roux, Ed., "Path Computation Element (PCE) communication Protocol", RFC 5440, March 2009.

[PCEP-GMPLS] Margaria, et al., "PCEP extensions for GMPLS", draft-ietf-pce-gmpls-pcep-extensions, work in progress.

[PCEP-Layer] Oki, Takeda, Le Roux, and Farrel, "Extensions to the Path Computation Element communication Protocol (PCEP) for Inter-Layer MPLS and GMPLS Traffic Engineering", draft-ietf-pce-inter-layer-ext, work in progress.

[RFC6163] Lee, Y. and Bernstein, G. (Editors), and W. Imajuku, "Framework for GMPLS and PCE Control of Wavelength Switched Optical Networks", RFC 6163, March 2011.

[PCE-RWA] Lee, Y., et. al., "PCEP Requirements for WSON Routing and Wavelength Assignment", draft-ietf-pce-wson-routing-wavelength, work in progress.

[RFC6205] Tomohiro, O. and D. Li, "Generalized Labels for Lambda-Switching Capable Label Switching Routers", RFC 6205, January, 2011.

[WSON-Sign] Bernstein et al,"Signaling Extensions for Wavelength Switched Optical Networks", draft-ietf-ccamp-wson-signaling, work in progress.

[WSON-OSPF] Lee and Bernstein,"OSPF Enhancement for Signal and Network Element Compatibility for Wavelength Switched Optical Networks",draft-ietf-ccamp-wson-signal-compatibility-ospf,work in progress.

[RWA-Info] Bernstein and Lee, "Routing and Wavelength Assignment Information Model for Wavelength Switched Optical Networks",draft-ietf-ccamp-rwa-info, work in progress.

[RWA-Encode]Bernstein and Lee, "Routing and Wavelength Assignment Information Encoding for Wavelength Switched Optical Networks",draft-ietf-ccamp-rwa-wson-encode, work in progress.

[GEN-Encode] Bernstein and Lee, "General Network Element Constraint Encoding for GMPLS Controlled Networks",draft-ietf-ccamp-general-constraint-encode, work in progress.

[WSON-Imp] Y. Lee, G. Bernstein, D. Li, G. Martinelli, "A Framework for the Control of Wavelength Switched Optical Networks (WSON) with Impairments", draft-ietf-ccamp-wson-impairments, work in progress.

[RSVP-Imp] agraz, "RSVP-TE Extensions in Support of Impairment Aware Routing and Wavelength Assignment in Wavelength Switched Optical Networks WSONs)", draft-agraz-ccamp-wson-impairment-rsvp, work in progress.

[OSPF-Imp] Bellagamba, et al., "OSPF Extensions for Wavelength Switched Optical Networks (WSON) with Impairments",draft-eb-ccamp-ospf-wson-impairments, work in progress.

11. Contributors

Authors' Addresses

Young Lee, Editor
Huawei Technologies
1700 Alma Drive, Suite 100
Plano, TX 75075, USA
Phone: (972) 509-5599 (x2240)
Email: leeyoung@huawei.com

Ramon Casellas, Editor
CTTC PMT Ed B4 Av. Carl Friedrich Gauss 7
08860 Castelldefels (Barcelona)
Spain
Phone: (34) 936452916
Email: ramon.casellas@cttc.es

Fatai Zhang
Huawei Technologies
Email: zhangfatai@huawei.com

Cyril Margaria
Nokia Siemens Networks
St Martin Strasse 76
Munich, 81541
Germany
Phone: +49 89 5159 16934
Email: cyril.margaria@nsn.com

Oscar Gonzalez de Dios
Telefonica Investigacion y Desarrollo
C/ Emilio Vargas 6
Madrid, 28043
Spain
Phone: +34 91 3374013
Email: ogondio@tid.es

Greg Bernstein
Grotto Networking
Fremont, CA, USA
Phone: (510) 573-2237
Email: gregb@grotto-networking.com

Intellectual Property Statement

Disclaimer of Validity

Acknowledgment

Funding for the RFC Editor function is currently provided by the Internet Society.

Network Working Group
Internet Draft

Y. Lee, Ed.
Huawei Technologies

Intended status: Standard
Expires: August 2013

R. Casellas, Ed.
CTTC

February 6, 2013

PCEP Extension for WSON Routing and Wavelength Assignment

draft-lee-pce-wson-rwa-ext-05.txt

Abstract

This draft provides the Path Computation Element communication Protocol (PCEP) extensions for the support of Routing and Wavelength Assignment (RWA) in Wavelength Switched Optical Networks (WSON). Lightpath provisioning in WSONs requires a routing and wavelength assignment (RWA) process. From a path computation perspective, wavelength assignment is the process of determining which wavelength can be used on each hop of a path and forms an additional routing constraint to optical light path computation.

Table of Contents

1. Terminology

This document uses the terminology defined in [RFC4655], and [RFC5440].

2. Requirements Language

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

3. Introduction

[RFC4655] defines the PCE based Architecture and explains how a Path Computation Element (PCE) may compute Label Switched Paths (LSP) in Multiprotocol Label Switching Traffic Engineering (MPLS-TE) and Generalized MPLS (GMPLS) networks at the request of Path Computation Clients (PCCs). A PCC is said to be any network component that makes such a request and may be, for instance, an Optical Switching Element within a Wavelength Division Multiplexing (WDM) network. The PCE, itself, can be located anywhere within the network, and may be within an optical switching element, a Network Management System (NMS) or Operational Support System (OSS), or may be an independent network server.

The PCE communications Protocol (PCEP) is the communication protocol used between PCC and PCE, and may also be used between cooperating PCEs. [RFC4657] sets out the common protocol requirements for PCEP. Additional application-specific requirements for PCEP are deferred to separate documents.

This document provides the PCEP extensions for the support of Routing and Wavelength Assignment (RWA) in Wavelength Switched

Optical Networks (WSON) based on the requirements specified in [PCE-RWA].

WSON refers to WDM based optical networks in which switching is performed selectively based on the wavelength of an optical signal. In this document, it is assumed that wavelength converters require electrical signal regeneration. Consequently, WSONs can be transparent (A transparent optical network is made up of optical devices that can switch but not convert from one wavelength to another, all within the optical domain) or translucent (3R regenerators are sparsely placed in the network).

A LSC Label Switched Path (LSP) may span one or several transparent segments, which are delimited by 3R regenerators (typically with electronic regenerator and optional wavelength conversion). Each transparent segment or path in WSON is referred to as an optical path. An optical path may span multiple fiber links and the path should be assigned the same wavelength for each link. In such case, the optical path is said to satisfy the wavelength-continuity constraint. Figure 1 illustrates the relationship between a LSC LSP and transparent segments (optical paths).

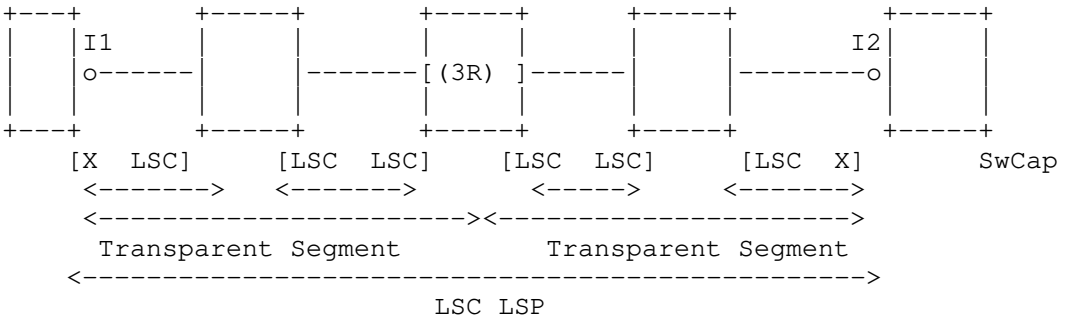


Figure 1 Illustration of a LSC LSP and transparent segments

Note that two optical paths within a WSON LSP need not operate on the same wavelength (due to the wavelength conversion capabilities). Two optical paths that share a common fiber link cannot be assigned the same wavelength. To do otherwise would result in both signals interfering with each other. Note that advanced additional multiplexing techniques such as polarization based multiplexing are not addressed in this document since the physical layer aspects are not currently standardized. Therefore, assigning the proper

wavelength on a lightpath is an essential requirement in the optical path computation process.

When a switching node has the ability to perform wavelength conversion, the wavelength-continuity constraint can be relaxed, and a LSC Label Switched Path (LSP) may use different wavelengths on different links along its route from origin to destination. It is, however, to be noted that wavelength converters may be limited due to their relatively high cost, while the number of WDM channels that can be supported in a fiber is also limited. As a WSON can be composed of network nodes that cannot perform wavelength conversion, nodes with limited wavelength conversion, and nodes with full wavelength conversion abilities, wavelength assignment is an additional routing constraint to be considered in all lightpath computation.

For example, within a translucent WSON, a LSC LSP may be established between interfaces I1 and I2, spanning 2 transparent segments (optical paths) where the wavelength continuity constraint applies (i.e. the same unique wavelength MUST be assigned to the LSP at each TE link of the segment). If the LSC LSP induced a Forwarding Adjacency / TE link, the switching capabilities of the TE link would be [X X] where X < LSC (PSC, TDM, ...).

This document aligns with GMPLS extensions for PCEP [PCEP-GMPLS] for generic property such as label, label-set and label assignment noting that wavelength is a type of label. Wavelength restrictions and constraints are also formulated in terms of labels per [GEN-ENCODE].

The optical modulation properties, which are also referred to as signal compatibility, are already considered in signaling in [RWA-Encode] and [WSON-OSPF]. In order to improve the signal quality and limit some optical effects several advanced modulation processing are used. Those modulation properties contribute not only to optical signal quality checks but also constrain the selection of sender and receiver, as they should have matching signal processing capabilities. This document includes signal compatibility constraint as part of RWA path computation. That is, the signal processing capabilities (e.g., modulation and FEC) must be compatible between the sender and the receiver of the optical path across all optical elements.

This document, however, does not address optical impairments as part of RWA path computation. See [WSON-Imp] and [RSVP-Imp] for more information on optical impairments and GMPLS.

4. Encoding of a RWA Path Request

Figure 2 shows one typical PCE based implementation, which is referred to as Combined Process (R&WA). With this architecture, the two processes of routing and wavelength assignment are accessed via a single PCE. This architecture is the base architecture from which the requirements have been specified in [PCE-RWA] and the PCEP extensions that are going to be specified in this document based on this architecture.

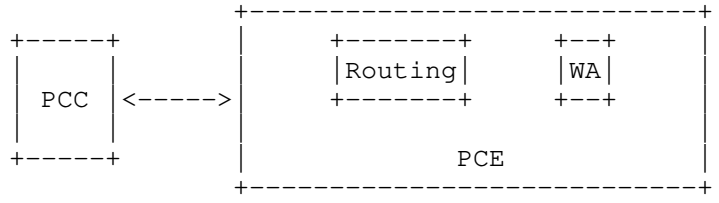


Figure 2 Combined Process (R&WA) architecture

4.1. Wavelength Assignment (WA) Object

The current RP object is used to indicate routing related information in a new path request per [RFC5440]. Since a new RWA path request involves both routing and wavelength assignment, the wavelength assignment related information in the request SHOULD be coupled in the path request.

Wavelength allocation can be performed by the PCE by different means:

(a) By means of Explicit Label Control, in the sense that one (or two) allocated labels MAY appear after an interface route subobject.
(b) By means of a Label Set, containing one or more allocated Labels, provided by the PCE.

Option (b) allows distributed label allocation (performed during signaling) to complete wavelength assignment.

Additionally, given a range of potential labels to allocate, the request SHOULD convey the heuristic / mechanism to the allocation.

The format of a PCReq message after incorporating the WA object is as follows:

<PCReq Message> ::= <Common Header>

[<svec-list>]

<request-list>

Where:

<request-list>::=<request>[<request-list>]

<request>::= <RP>

<ENDPOINTS>

<WA>

[other optional objects...]

If WA object is present in the request, the WA object MUST be encoded after the ENDPOINTS object.

The format of the Wavelength Assignment (WA) object body is as follows:

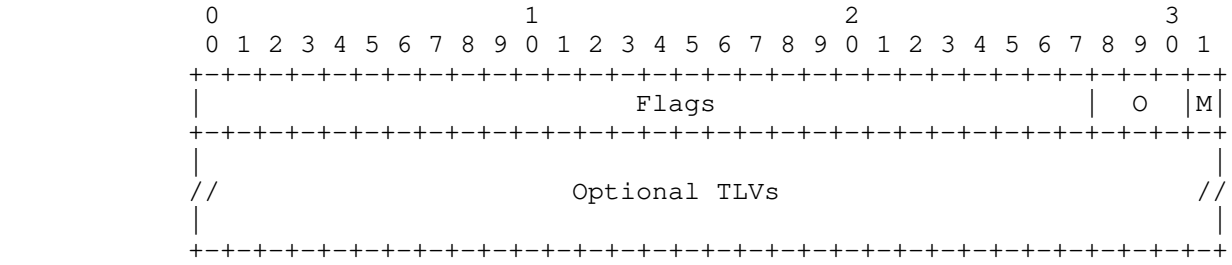


Figure 3 WA Object

o Flags (32 bits)

The following new flags SHOULD be set

M (Mode - 1 bit): M bit is used to indicate the mode of wavelength assignment. When M bit is set to 1, this indicates that the label assigned by the PCE must be explicit. That is, the selected way to convey the allocated wavelength is by means of Explicit Label Control (ELC) [RFC4003] for each hop of a computed LSP. Otherwise, the label assigned by the PCE needs not be explicit (i.e., it can be suggested in the form of label set objects in the corresponding response, to allow distributed WA. In such case, the PCE MUST return a Label Set object as described in Section 2.2 of [Gen-Encode] in the response.

O (Order - 3 bits): O bit is used to indicate the wavelength assignment constraint in regard to the order of wavelength assignment to be returned by the PCE. This case is only applied when M bit is set to "explicit." The following indicators should be defined:

000 - Reserved

001 - Random Assignment

010 - First Fit (FF) in descending Order

011 - First Fit (FF) in ascending Order

100 - Last Fit (LF) in ascending Order

101 - Last Fit (LF) in descending Order

110 - Unspecified

111 - Reserved

4.2. Wavelength Restriction Constraint TLV

For any request that contains a wavelength assignment, the requester (PCC) MUST be able to specify a restriction on the wavelengths to be used. This restriction is to be interpreted by the PCE as a constraint on the tuning ability of the origination laser transmitter or on any other maintenance related constraints. Note that if the LSP LSC spans different segments, the PCE MUST have mechanisms to know the tunability restrictions of the involved wavelength converters / regenerators, e.g. by means of the TED either via IGP or NMS. Even if the PCE knows the tunability of the transmitter, the PCC MUST be able to apply additional constraints to the request.

[Ed note: Which PCEP Object will home this TLV is yet to be determined. Since this involves the end-point, The END-POINTS Object might be a good candidate to encode this TLV, which will be provided in a later revision.]

[Ed note: The current encoding assumes that tunability restriction applied to link-level.]

The TLV type is TBD, recommended value is TBD. This TLV MAY appear more than once to be able to specify multiple restrictions.

The TLV data is defined as follows:

<Wavelength Restriction Constraint> ::=

<Action> <Format> <Reserved>

(<Link Identifiers> <Wavelength Restriction>)...

Where

<Link Identifiers> ::=

<Unnumbered IF ID> | <IPV4 Address> | <IPV6 Address>

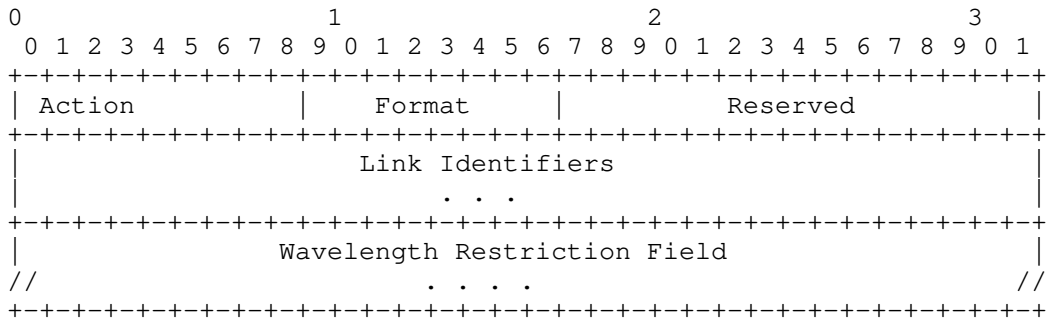


Figure 4 Wavelength Restriction

- Action: 8 bits

  0 - Inclusive List indicates that one or more link identifiers are included in the Link Set. Each identifies a separate link that is part of the set.

  1 - Inclusive Range indicates that the Link Set defines a range of links. It contains two link identifiers. The first identifier indicates the start of the range (inclusive). The second identifier indicates the end of the range (inclusive). All links with numeric values between the bounds are considered to be part of the set. A value of zero in either position indicates that there is no bound on the corresponding portion of the range. Note that the Action field can be set to 0 when unnumbered link identifier is used.

Note that "interfaces" such as those discussed in the Interfaces MIB [RFC2863] are assumed to be bidirectional.

- Format: The format of the link identifier (8 bits)

  0 -- Unnumbered Link Identifier
  1 -- Local Interface IPv4 Address
  2 -- Local Interface IPv6 Address
  Others TBD.

Note that all link identifiers in the same list must be of the same type.

- Reserved: Reserved for future use (16 bits)

- Link Identifiers: Identifies each link ID for which restriction is applied. The length is dependent on the link format. See the following section for Link Identifier encoding.

4.2.1. Link Identifier sub-TLV

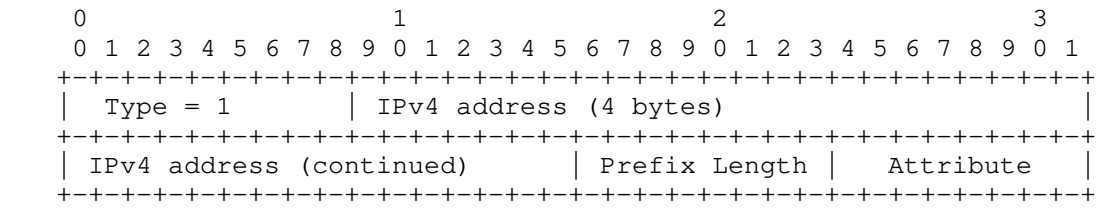
The link identifier field can be an IPv4, IPv6 or unnumbered interface ID.

<Link Identifier> ::=

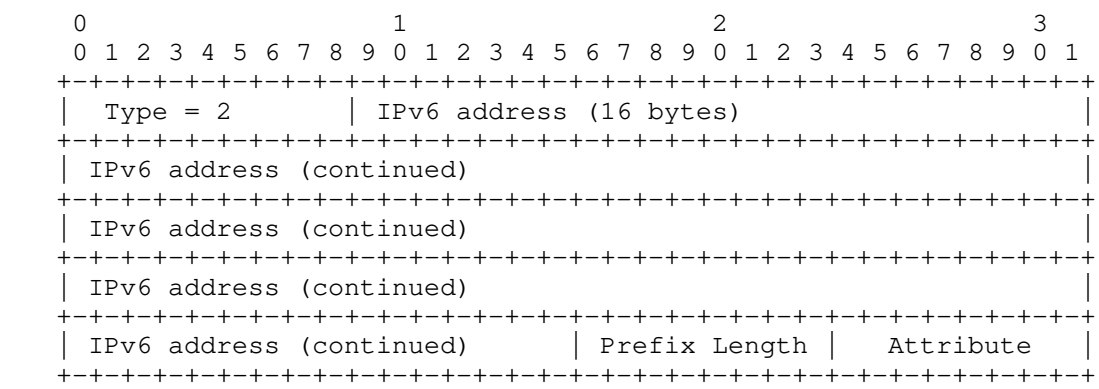
<IPV4 Address> | <IPV6 Address> | <Unnumbered IF ID>

The encoding of each case is as follows:

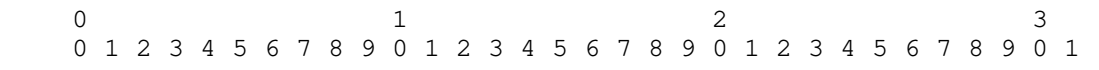
IPv4 prefix Sub-TLV

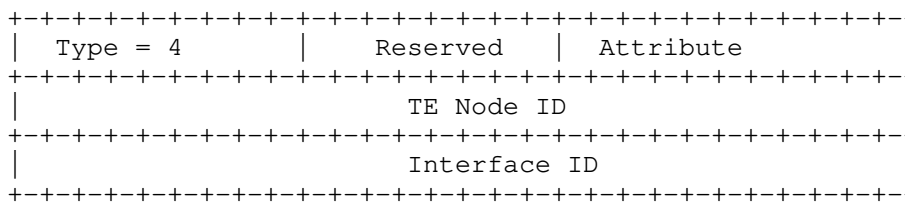


IPv6 prefix Sub-TLV



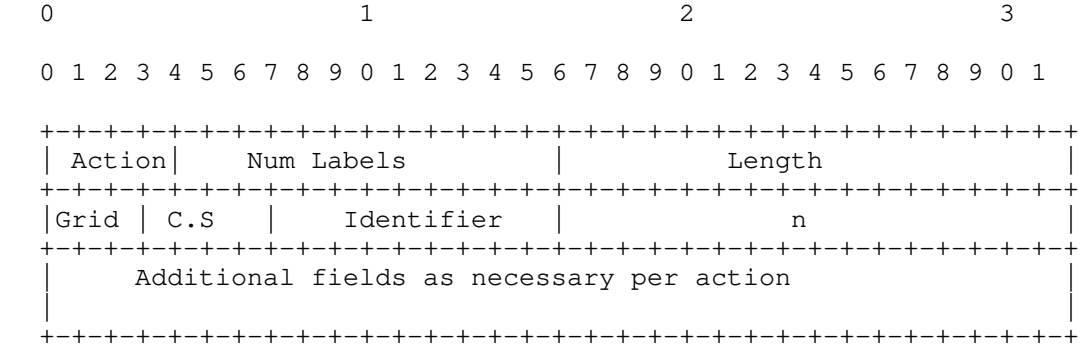
Unnumbered Interface ID Sub-TLV





4.2.2. Wavelength Restriction Field sub-TLV

The Wavelength Restriction Field of the wavelength restriction TLV is encoded as a Label Set field as specified in [GEN-Encode] section 2.2, as shown below, with base label encoded as a 32 bit LSC label, defined in [RFC6205]. See [RFC6205] for a description of Grid, C.S, Identifier and n, as well as [GEN-Encode] for the details of each action.



4.3. Signal processing capability restrictions

Path computation for WSON include the check of signal processing capabilities, those capability MAY be provided by the IGP, however this is not a MUST. Moreover, a PCC should be able to indicate additional restrictions for those signal compatibility, either on the endpoint or any given link.

The supported signal processing capabilities are the one described in [RWA-Info]:

Optical Interface Class List

Bit rate

Client signal

The Bit-rate restriction is already expressed in [PCEP-GMPLS] in the GENERALIZED-BANDWIDTH object.

The client signal information can be expressed using the REQ-ADAP-CAP object from the [PCEP-Layer].

In order to support the Optical Interface Class information a new TLV are introduced as endpoint-restriction in the END-POINTS type Generalized endpoint:

Optical Interface Class List TLV

The END-POINTS type generalized endpoint is extended as follow:

<endpoint-restrictions> ::= <LABEL-REQUEST>

<label-restriction-list>

[<signal-compatibility-restriction>...]

Where

signal-compatibility-restriction ::=

<Optical Interface Class List>

The encoding for Optical Interface Class List is described in Section 5.2 of [RWA-Encode].

4.3.1. Signal Processing Exclusion XRO Sub-Object

The PCC/PCE should be able to exclude particular types of signal processing along the path in order to handle client restriction or multi-domain path computation.

In order to support the exclusion a new XRO sub-object is defined: the signal processing exclusion:

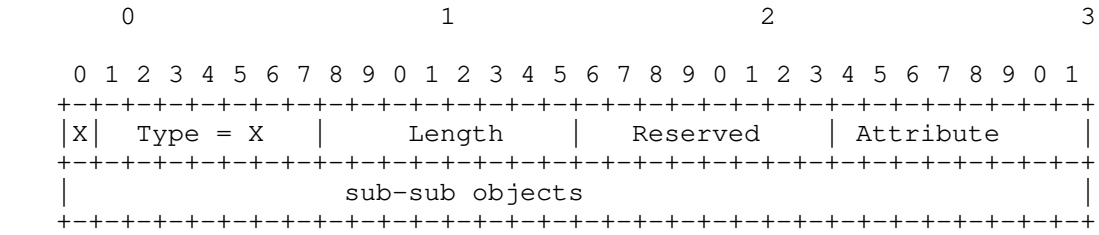


Figure 5 Signaling Processing XRO Sub-Object

The Attribute field indicates how the exclusion sub-object is to be interpreted. The Attribute can only be 0 (Interface) or 1 (Node).

The sub-sub objects are encoded as in RSVP signaling definition [WSON-Sign].

4.3.2. IRO sub-object: signal processing inclusion

Similar to the XRO sub-object the PCC/PCE should be able to include particular types of signal processing along the path in order to handle client restriction or multi-domain path computation.

This is supported by adding the sub-object "processing" defined for ERO in [WSON-Sign] to the PCEP IRO object.

5. Encoding of a RWA Path Reply

The ERO is used to encode the path of a TE LSP through the network. The ERO is carried within a given path of a PCEP response, which is in turn carried in a PCRep message to provide the computed TE LSP if the path computation was successful. The preferred way to convey the allocated wavelength is by means of Explicit Label Control (ELC) [RFC4003].

In order to encode wavelength assignment, the Wavelength Assignment (WA) Object needs to be employed to be able to specify wavelength assignment. Since each segment of the computed optical path is associated with wavelength assignment, the WA Object should be aligned with the ERO object.

Encoding details will be provided further revisions and will be aligned as much as possible with [WSON-Sign] and [LSPA-ERO]

5.1. Error Indicator

To indicate errors associated with the RWA request, a new Error Type (TDB) and subsequent error-values are defined as follows for inclusion in the PCEP-ERROR Object:

A new Error-Type (TDB) and subsequent error-values are defined as follows:

Error-Type=TBD; Error-value=1: if a PCE receives a RWA request and the PCE is not capable of processing the request due to insufficient memory, the PCE MUST send a PCErr message with a PCEP-ERROR Object (Error-Type=TDB) and an Error-value(Error-value=1). The PCE stops processing the request. The corresponding RWA request MUST be cancelled at the PCC.

Error-Type=TBD; Error-value=2: if a PCE receives a RWA request and the PCE is not capable of RWA computation, the PCE MUST send a PCErr message with a PCEP-ERROR Object (Error-Type=15) and an Error-value (Error-value=2). The PCE stops processing the request. The corresponding RWA computation MUST be cancelled at the PCC.

5.2. NO-PATH Indicator

To communicate the reason(s) for not being able to find RWA for the path request, the NO-PATH object can be used in the PCRep message. The format of the NO-PATH object body is defined in [RFC5440]. The object may contain a NO-PATH-VECTOR TLV to provide additional information about why a path computation has failed.

Two new bit flags are defined to be carried in the Flags field in the NO-PATH-VECTOR TLV carried in the NO-PATH Object.

Bit TDB: When set, the PCE indicates no feasible route was found that meets all the constraints associated with RWA.

Bit TDB: When set, the PCE indicates that no wavelength was assigned to at least one hop of the route in the response.

Bit TDB: When set, the PCE indicate that no path was found satisfying the signal compatibility constraints.

6. Manageability Considerations

Manageability of WSON Routing and Wavelength Assignment (RWA) with PCE must address the following considerations:

6.1. Control of Function and Policy

In addition to the parameters already listed in Section 8.1 of [PCEP], a PCEP implementation SHOULD allow configuring the following PCEP session parameters on a PCC:

The ability to send a WSON RWA request.

In addition to the parameters already listed in Section 8.1 of [PCEP], a PCEP implementation SHOULD allow configuring the following PCEP session parameters on a PCE:

The support for WSON RWA.

A set of WSON RWA specific policies (authorized sender, request rate limiter, etc).

These parameters may be configured as default parameters for any PCEP session the PCEP speaker participates in, or may apply to a specific session with a given PCEP peer or a specific group of sessions with a specific group of PCEP peers.

6.2. Information and Data Models, e.g. MIB module

Extensions to the PCEP MIB module defined in [PCEP-MIB] should be defined, so as to cover the WSON RWA information introduced in this document. A future revision of this document will list the information that should be added to the MIB module.

6.3. Liveness Detection and Monitoring

Mechanisms defined in this document do not imply any new liveness detection and monitoring requirements in addition to those already listed in section 8.3 of [RFC5440].

6.4. Verifying Correct Operation

Mechanisms defined in this document do not imply any new verification requirements in addition to those already listed in section 8.4 of [RFC5440]

6.5. Requirements on Other Protocols and Functional Components

The PCE Discovery mechanisms ([RFC5089] and [RFC5088]) may be used to advertise WSON RWA path computation capabilities to PCCs.

6.6. Impact on Network Operation

Mechanisms defined in this document do not imply any new network operation requirements in addition to those already listed in section 8.6 of [RFC5440].

7. Security Considerations

This document has no requirement for a change to the security models within PCEP [PCEP]. However the additional information distributed in order to address the RWA problem represents a disclosure of network capabilities that an operator may wish to keep private. Consideration should be given to securing this information.

8. IANA Considerations

A future revision of this document will present requests to IANA for codepoint allocation.

9. Acknowledgments

The authors would like to thank Adrian Farrel for many helpful comments that greatly improved the contents of this draft.

This document was prepared using 2-Word-v2.0.template.dot.

10. References

10.1. Informative 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 (GMPLS) Signaling Functional Description", RFC 3471, January 2003.

[RFC3473] Berger, L., Ed., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.

[RFC3477] Kompella, K. and Y. Rekhter, "Signalling Unnumbered Links in Resource ReSerVation Protocol - Traffic Engineering (RSVP-TE)", RFC 3477, January 2003.

[RFC4003] Berger, L., "GMPLS Signaling Procedure for Egress Control", RFC 4003, February 2005.

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

[RFC4657] Ash, J. and J. Le Roux, "Path Computation Element (PCE) Communication Protocol Generic Requirements", RFC 4657, September 2006.

[RFC5440] Vasseur, JP., Ed. and JL. Le Roux, Ed., "Path Computation Element (PCE) communication Protocol", RFC 5440, March 2009.

[PCEP-GMPLS] Margaria, et al., "PCEP extensions for GMPLS", draft-ietf-pce-gmpls-pcep-extensions, work in progress.

[LSPA-ERO] Margaria, et al., "LSP Attribute in ERO", draft-margaria-ccamp-lsp-attribute-ero, work in progress.

[PCEP-Layer] Oki, Takeda, Le Roux, and Farrel, "Extensions to the Path Computation Element communication Protocol (PCEP) for Inter-Layer MPLS and GMPLS Traffic Engineering", draft-ietf-pce-inter-layer-ext, work in progress.

[RFC6163] Lee, Y. and Bernstein, G. (Editors), and W. Imajuku, "Framework for GMPLS and PCE Control of Wavelength Switched Optical Networks", RFC 6163, March 2011.

[PCE-RWA] Lee, Y., et. al., "PCEP Requirements for WSON Routing and Wavelength Assignment", draft-ietf-pce-wson-routing-wavelength, work in progress.

[RFC6205] Tomohiro, O. and D. Li, "Generalized Labels for Lambda-Switching Capable Label Switching Routers", RFC 6205, January, 2011.

[WSON-Sign] Bernstein et al,"Signaling Extensions for Wavelength Switched Optical Networks", draft-ietf-ccamp-wson-signaling, work in progress.

[WSON-OSPF] Lee and Bernstein,"OSPF Enhancement for Signal and Network Element Compatibility for Wavelength Switched Optical Networks",draft-ietf-ccamp-wson-signal-compatibility-ospf, work in progress.

[RWA-Info] Bernstein and Lee, "Routing and Wavelength Assignment Information Model for Wavelength Switched Optical Networks",draft-ietf-ccamp-rwa-info, work in progress.

[RWA-Encode]Bernstein and Lee, "Routing and Wavelength Assignment Information Encoding for Wavelength Switched Optical Networks",draft-ietf-ccamp-rwa-wson-encode, work in progress.

[GEN-Encode] Bernstein and Lee, "General Network Element Constraint Encoding for GMPLS Controlled Networks",draft-ietf-ccamp-general-constraint-encode, work in progress.

[WSON-Imp] Y. Lee, G. Bernstein, D. Li, G. Martinelli, "A Framework for the Control of Wavelength Switched Optical Networks (WSON) with Impairments", draft-ietf-ccamp-wson-impairments, work in progress.

[RSVP-Imp] agraz, "RSVP-TE Extensions in Support of Impairment Aware Routing and Wavelength Assignment in Wavelength Switched Optical Networks WSONs)", draft-agraz-ccamp-wson-impairment-rsvp, work in progress.

[OSPF-Imp] Bellagamba, et al., "OSPF Extensions for Wavelength Switched Optical Networks (WSON) with Impairments",draft-eb-ccamp-ospf-wson-impairments, work in progress.

11. Contributors

Authors' Addresses

Young Lee, Editor
Huawei Technologies
1700 Alma Drive, Suite 100
Plano, TX 75075, USA
Phone: (972) 509-5599 (x2240)
Email: leeyoung@huawei.com

Ramon Casellas, Editor
CTTC PMT Ed B4 Av. Carl Friedrich Gauss 7
08860 Castelldefels (Barcelona)
Spain
Phone: (34) 936452916
Email: ramon.casellas@cttc.es

Fatai Zhang
Huawei Technologies
Email: zhangfatai@huawei.com

Cyril Margaria
Nokia Siemens Networks
St Martin Strasse 76
Munich, 81541
Germany
Phone: +49 89 5159 16934
Email: cyril.margaria@nsn.com

Oscar Gonzalez de Dios
Telefonica Investigacion y Desarrollo
C/ Emilio Vargas 6
Madrid, 28043
Spain
Phone: +34 91 3374013
Email: ogondio@tid.es

Greg Bernstein
Grotto Networking
Fremont, CA, USA
Phone: (510) 573-2237
Email: gregb@grotto-networking.com

Intellectual Property Statement

Disclaimer of Validity

Acknowledgment

Funding for the RFC Editor function is currently provided by the Internet Society.

Network Working Group
Internet-Draft
Intended status: Informational
Expires: May 3, 2012

Kexin Tang
Xuerong Wang
Xuping Cao
ZTE Corporation
Oct 31, 2011

Stateful PCE
draft-tang-pce-stateful-pce-02.txt

Abstract

A PCE can be either stateful or stateless. The information carried in a stateful PCE is more detailed than that of a stateless PCE. This draft focus on stateful PCE, describes the problems without stateful PCE, and gives the IGP and PCEP extensions to realize stateful PCE.

Table of Contents

1. Introduction

As defined in section 6.8 of [RFC4655], a PCE can be either stateful or stateless. For stateful PCE, there is a strict synchronization between the PCEs not only the network states (in term of topology and resource information), but also the set of computed paths and reserved resources in use in the network. Since stateful PCE has more network information, it can be used to do some complicated work.

However, the existing PCE discovery protocol ([RFC5088], [RFC5089]) and PCEP ([RFC5440]) do not support stateful PCE. And there is no effective synchronization mechanism defined to realize stateful PCE yet. For these sufficient reasons, this document focus on stateful PCE, describes the applicability of stateful PCE and gives the IGP and PCEP extensions to support stateful PCE.

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. Terminology

- PCC: Path Computation Client. A client application requesting a path computation to be performed by the Path Computation Element.
- PCE: Path Computation Element. An entity that is capable of computing a network path or route based on a network graph, and of applying computational constraints during the computation.
- PCED: PCE Discovery.
- PCEP: Path Computation Element communication Protocol.
- TED: Traffic Engineering Database, which contains the topology and resource information of the domain. The TED may be fed by Interior Gateway Protocol (IGP) extensions or potentially by other means.

3. Problems

This section lists the typical problems without stateful PCE. As described in [RFC4655], stateful PCE uses information not only from the TED, but also information about existing paths (for example, TE LSPs) in the network when processing new requests, so the information

carried in stateful PCE are more detailed than that of stateless PCE.

Without stateful PCE, there would not be a strict synchronization between PCE and LSP state. Consequently,the PCE may not know the existing path in the network in time,and suppose the network resource taken by the existing path is unoccupied, so it would consider these resource in the other path computations. Resource conflict occurs under this condition.

A PCE may received several PCReq messages, this may occurs when a PCE is required to compute a large number of path for several LSPs,for example when fiber cutting happened or in a sudden accident,which may result in a large number of links down simultaneity,and need to be recovery in a short time.

Figure 1 shows a demonstration topology for the subsequent context.

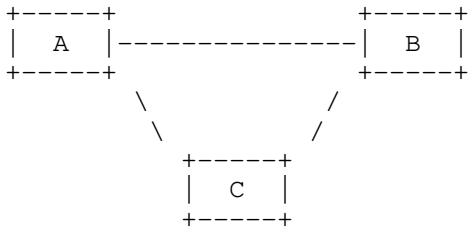


Figure 1: demonstration topology

In Figure1,we make the assumptions as follows:the link capacity between all these three nodes is 100M,and the node that hosts the PCE function received a PCReq massage,which required it to compute paths for two LSP respectively:LSP1 and LSP2. LSP1 and LSP2 share the same original and terminal node pair,that is from node A to node B,and require the same link bandwidth:80M.

For LSP 1,to get a shortest path,the PCE may return a path:A--B,while the computation result is not synchronized in its TED in time,the PCE assume the unreserved bandwidth of the link between A and B is still 100M.

Then for LSP2,which require 80M link capacity either,the PCE may still consider the shortest path,that is A--B as before,and it seems that in the TED currently,this shortest path has enough capacity for LSP2,so the PCE returns the optimal path A--B for LSP2.

Once the LSP setup procedure by signaling for LSP1 and LSP2 is running,the setup procedure for the first LSP would successful, while

the latter one would encounter resource conflict:there would not be enough link capacity for the LSP to be setup,and it would return a signaling failure notification message.

As for Inter-area or Inter-AS path computation for a LSP, it is often involved in multiple PCEs cooperation to compute an end-to-end path, for example, BRPC ([RFC5441]) and H-PCE ([H-PCE-FWK]). Resource conflict would even worse in this situation because it takes extra time for communication between the cooperative PCEs.

4. Protocol Extensions

4.1. PCED Extensions

[RFC5088] defines extensions to OSPFv2 ([RFC2328]) and OSPFv3 ([RFC5340]) to allow a PCE in an OSPF routing domain to advertise some information useful to a PCC for PCE selection. It defines a new TLV (named the PCE Discovery TLV (PCED TLV)) to be carried within the OSPF Router Information LSA ([RFC4970]). The type 5 sub-TLV of PCED TLV, which named CE-CAP-FLAGS sub-TLV, used to indicate the capabilities of a PCE. It contains eight capabilities currently, but not includes the stateful capability of a PCE. So the PCE in an OSPF routing domain cannot advertise its state capability information to a PCC，which would help the PCC to make advanced and informed choices in PCE selection.

To discover stateful PCE, a PCC SHOULD know whether PCE is stateful. Therefore, the PCE discovery message SHOULD indicate whether the PCE advertising this message is a stateful PCE. Since PCE-CAP-FLAGS Sub-TLV ([RFC5088] for OSPF, [RFC5089] for IS-IS) contains PCE Capability Flags, this document defines a new flag, Stateful PCE Capability Flag, as follows (need to be assigned by IANA):

Bit	Capabilities
TBD	Stateful PCE

4.2. LSP State Synchronization

With respect to the definition of stateful PCE defined in [RFC4655], a stateful PCE utilizes information from the TED as well as information about existing paths (for example, TE LSPs) in the network when processing new path computation requests. Therefore, a stateful PCE SHOULD know the status (created or deleted) of a LSP. For this reason, there SHOULD be a timely synchronization of LSP state between PCC and PCE, including the path computation result, and the LSP setup or deletion result. For this purpose, this section

makes an extension to the PCNtf message defined in [RFC5440], defines a new Notification-type as follows (need to be assigned by IANA):

- Notification-type=TBD: LSP Status
  - Notification-value=1: end-to-end path computation success(This value is available for distributed path computation only). When a optimal end-to-end path is computed successfully, the head-end PCE SHOULD send a notification message with Notification-type=TBD and Notification-value=1 to the other PCEs collaboratively in the PCE chain which computed the other path segments successfully.
  - Notification-value=2: end-to-end path computation failure(This value is available for distributed path computation only). If an end-to-end path computation is failed,the head-end PCE SHOULD send a notification message with Notification-type=TBD and Notification-value=2 to the other PCEs collaboratively in the PCE chain which computed the other path segments successfully.
  - Notification-value=3: LSP setup success. When a LSP is created successfully by signaling, the PCC SHOULD send a notification message with Notification-type=TBD and Notification-value=3 to all the PCEs.
  - Notification-value=4: LSP setup failure. When a LSP is failed to setup by signaling, the PCC SHOULD send a notification message with Notification-type=TBD and Notification-value=4 to all the PCEs.
  - Notification-value=5: LSP delete success. When a LSP is delete in the network successfully, the PCC SHOULD send a notification message with Notification-type= TBD and Notification-value=5 to all the PCEs.
  - Notification-value=6: LSP delete failure . When a LSP path is failed to delete, the PCC SHOULD send a notification message with Notification-type= TBD and Notification-value=6 to all the PCEs.

This draft extended the PCNtf message, added the Path object which defined for PCRep message [RFC5440] to the PCNtf message ,to identify the ERO and the attribute of a LSP.

The format of the extended PCNtf message is as follows:

```
<PCNtf Message>::=<Common Header>
                  <notify-list>

<notify-list>::=<notify> [<notify-list>]

<notify>::= [<request-id-list>]
             <notification-list>

<request-id-list>::=<RP><PATH> [<request-id-list>]

<PATH>::= <ERO><attribute-list>
```

where:

```
<attribute-list>::=[<LSPA>]
                   [<BANDWIDTH>]
                   [<metric-list>]
                   [<IRO>]

<notification-list>::=<NOTIFICATION>[<notification-list>]
```

With the newly added Path object which carries the ERO object and the attribute of a LSP,a PCE can identifies the status of which LSP changes,and the information of the LSP, so as to identifies the resource that taken by the LSP.

5. Security Considerations

The extensions of this draft are based on PCEP and OSPF, only some optional protocol elements are added which will not change the security of existing network.

6. IANA Consideration

The extensions of this draft is baed on PCEP and OSPF, only some optional protocol elements are added which will not change the security of existing network.

7. Normative References

[H-PCE-FWK]
Daniel King, Adrian Farrel, Quintin Zhao, and Fatai Zhang,

"The Application of the Path Computation Element Architecture to the Determination of a Sequence of Domains in MPLS and GMPLS", draft-ietf-pce-hierarchy-fwk-00.txt .

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

[RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.

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

[RFC4970] Lindem, A., Shen, N., Vasseur, JP., Aggarwal, R., and S. Shaffer, "Extensions to OSPF for Advertising Optional Router Capabilities", RFC 4970, July 2007.

[RFC5088] Le Roux, JL., Vasseur, JP., Ikejiri, Y., and R. Zhang, "OSPF Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5088, January 2008.

[RFC5089] Le Roux, JL., Vasseur, JP., Ikejiri, Y., and R. Zhang, "IS-IS Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5089, January 2008.

[RFC5340] Coltun, R., Ferguson, D., Moy, J., and A. Lindem, "OSPF for IPv6", RFC 5340, July 2008.

[RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.

[RFC5441] Vasseur, JP., Zhang, R., Bitar, N., and JL. Le Roux, "A Backward-Recursive PCE-Based Computation (BRPC) Procedure to Compute Shortest Constrained Inter-Domain Traffic Engineering Label Switched Paths", RFC 5441, April 2009.

Authors' Addresses

Kexin Tang
ZTE Corporation
No.50 Software Avenue, Yuhuatai District
Nanjing, Jiangsu 210012
P.R.China

Phone: +86-025-88014225
Email: tang.kexin@zte.com.cn

Xuerong Wang
ZTE Corporation
R&D Building 3, ZTE Industrial Zone, Liuxian Road,Nanshan District
Shenzhen 518055
P.R.China

Phone: +86-755-26773926
Email: wang.xuerong@zte.com.cn

Xuping Cao
ZTE Corporation
21F, ZTE Plaza, No.19 East Huayuan Road,Haidian District
Beijing 100191
P.R.China

Email: cao.xuping@zte.com.cn

PCE Working Group
Internet-Draft
Intended status: Informational
Expires: April 13, 2012

F. Zhang
Q. Zhao
Huawei Technologies
O. Gonzalez de Dios
Telefonica I+D
R. Casellas
CTTC
D. King
Old Dog Consulting
October 13, 2011

Extensions to Path Computation Element Communication Protocol (PCEP) for Hierarchical Path Computation Elements (PCE) draft-zhang-pce-hierarchy-extensions-01

Abstract

The hierarchical Path Computation Element (PCE) architecture, defined in the companion framework document [I-D.ietf-pce-hierarchy-fwk], allows the selection of an optimum domain sequence and the optimum end-to-end path, to be derived through the use of a hierarchical relationship between domains.

This document defines the Path Computation Element Protocol (PCEP) extensions for the purpose of implementing hierarchical PCE procedures which are described the aforementioned document.

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 13, 2012.

Copyright Notice

Table of Contents

1. Introduction

[I-D.ietf-pce-hierarchy-fwk] describes a hierarchical PCE architecture which can be used for computing end-to-end paths for inter-domain MPLS Traffic Engineering (TE) and GMPLS Label Switched Paths (LSPs). In the hierarchical PCE architecture, the parent PCE can compute a multi-domain path based on the domain connectivity information and each child PCE is able to compute the intra-domain path based on its domain topology information. The end-to-end domain path computing procedures can be abstracted as follows:

- o A path computation client (PCC) requests its own child PCE the computation of an inter-domain path.
- o The child PCE forwards the request to the parent PCE.
- o The parent PCE computes one or multiple domain paths from the ingress domain to the egress domain.
- o The parent PCE sends the intra-domain path computation requests (between the domain border nodes) to the child PCEs which are responsible for the domains along the domain path(s).
- o The child PCEs return the intra-domain paths to the parent PCE.
- o The parent PCE constructs the end-to-end inter-domain path based on the intra-domain paths and returns the inter-domain path to the child PCE.
- o The child PCE forwards the inter-domain path to the PCC.

This document defines the PCEP extensions for the purpose of implementing hierarchical PCE procedures, which are described in [I-D.ietf-pce-hierarchy-fwk].

The document also uses a number of editor notes to describe options and alternative solutions. These options and notes will be removed before publication.

1.1. Terminology

This document uses the terminology defined in [RFC4655], [RFC5440] and [I-D.ietf-pce-hierarchy-fwk].

1.2. Requirements Language

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

2. PCEP Extension Requirements

2.1. Building of parent topology

As stated in section 5.4 of [I-D.ietf-pce-hierarchy-fwk] a parent PCE maintains a domain topology map and may contain nodes and links on its own right, maintaining a traffic engineering database (TED) for the parent domain.

The parent PCE TED may be configured or learnt by the child PCEs.

Editors note. A child PCE could forward the topology within PCNtf messages or any other mechanisms, without an IGP adjacency. Further discussion of the discovery mechanism and scope will be discussed in later versions of this document.

2.2. New Objective Functions

For inter-domain path computation, there are three new objective functions which are defined in section 1.3.1 of [I-D.ietf-pce-hierarchy-fwk].

- Minimize the number of boundary nodes used.
- Limit the number of domains crossed.
- Disallow domain re-entry.

During the PCEP session establishment procedure, the parent PCE needs to be capable of indicating the objective functions (OF) capability in the Open message. This information can be, in turn, announced by child PCEs and used for selecting the PCE when a PCC want a path that satisfies a certain inter-domain objective function.

When a PCC requests a PCE to compute an inter-domain path, the PCC needs also to be capable of indicating the new objective functions for inter-domain path. Note that a given PCE may act as a regular PCE and as a parent PCE.

For the reasons described above, new OF codes need to be defined for the new inter-domain objective functions. Then the PCE can notify its new inter-domain objective functions to the PCC by carrying them in the OF-list TLV which is carried in the OPEN object. The PCC can specify which objective function code to use, which is carried in the OF object when requesting a PCE to compute an inter-domain path.

The proposed solutions may need to differentiate between the OF code that is requested at the parent level and the OF code that is requested at the intra-domain (child) level.

A parent PCE needs to be able to insure homogeneity when applying OF codes for the intra-domain requests.

2.3. PCEP Request Qualifiers

As described in section 5.8.1 of [I-D.ietf-pce-hierarchy-fwk], the H-PCE architecture will introduce new request qualifications as follows:

- It MUST be possible for a child PCE to indicate that a request it sends to a parent PCE should be satisfied by a domain sequence only, that is, not by a full end-to-end path. This allows the child PCE to initiate a per-domain [RFC5152] or a backward recursive path computation (BRPC) [RFC5441].
- A parent PCE needs to be able to ask a child PCE whether a particular node address (the destination of an end-to-end path) is present in the domain that the child PCE serves.
- As stated in [I-D.ietf-pce-hierarchy-fwk], section 5.5, if a PCC knows the egress domain, it can suppy this information as the path computation request. It SHOULD be possible to specify the destination domain information in a PCEP request, if it is known.

To meet the above requirements, the PCEP PCReq message should be extended.

2.4. Discovery Between Parent and Child PCEs

In the H-PCE architecture, the parent PCE does not need to be aware of each child domain topology. Therefore, it is possible that the parent PCE does not join the IGP instance of the child PCE domain, i.e. there is no IGP discovery mechanism between the parent PCE and child PCE.

Therefore there must be a discovery mechanism for basic PCE information between the parent and child PCEs. In this case, PCEP needs to provide discovery mechanisms that do not rely on IGP announcement/discovery procedures. A simple discovery mechanism relies on the static configuration / provisioning of the parent PCE id and address, which is configured at each child PCE.

2.4.1. Parent PCE Capability Discovery

As described in [I-D.ietf-pce-hierarchy-fwk], during the PCEP session establishment procedure, the child PCE needs to be capable of indicating to the parent PCE whether it requests the parent PCE capability or not. The parent PCE needs also to be capable of indicating whether its parent capability can be provided to the child PCE or not.

2.4.2. PCE Domain and PCE ID Discovery

A PCE domain is a single domain with an associated PCE. it is possible for a PCE to manage multiple domains. The PCE domain may be an IGP area or AS.

The PCE ID is an IPv4 and/or IPv6 address that is used to reach the parent/child PCE. It is RECOMMENDED to use an address that is always reachable if there is any connectivity to the PCE.

The PCE ID information and PCE domain identifiers may be provided during the PCEP session establishment procedure or the domain connectivity information collection procedure.

2.5. Domain Connectivity Information Collection

As described in [I-D.ietf-pce-hierarchy-fwk], the parent PCE builds the domain topology map either from configuration or from information received from each child PCE. A child PCE may report its neighbor domain connectivity to its parent PCE. It is reasonable to use PCEP PCNtf message to do this procedure. If an IGP adjacency is established between parent and children, it could be used for this purpose.

There are two types of domain border for providing the domain connectivity information:

- Domain border is a TE link, e.g. the inter-AS TE link which connects two ASs.
- Domain border is a node, e.g. the IGP ABR which connects two IGP areas.

For the inter-AS TE links, the following information needs to be notified to the parent PCE:

- Identifier of advertising child PCE.
- Identifier of PCE's domain.
- Identifier of the link.
- TE properties of the link (metrics, bandwidth).
- Other properties of the link (technology-specific).
- Identifier of link end-points.
- Identifier of adjacent domain.

For the ABR, the following information needs to be notified to the parent PCE:

- Identifier of the ABR.
- Identifier of the IGP Area IDs.

2.6. Error Case Handling

A PCE that is capable of acting as a parent PCE might not be configured or willing to act as the parent for a specific child PCE. This fact could be determined when the child sends a PCReq that requires parental activity (such as querying other child PCEs), and could result in a negative response in a PCEP Error (PCErr) message and indicate the hierarchy PCE error types.

3. PCEP Extensions

3.1. Extensions to OPEN object

3.1.1. OF Codes

There are three new OF codes defined here for H-PCE:

- MBN
  - Name: Minimize the number of Boundary Nodes used
  - Objective Function Code: (to be assigned by IANA, recommended 11)
  - Description: Find a path P such that passes through the least boundary nodes.
- MTD
  - Name: Minimize the number of Transit Domains
  - Objective Function Code: (to be assigned by IANA, recommended 12)
  - Description: Find a path P such that passes through the least transit domains.
- DDR
  - Name: Disallow Domain Re-entry (DDR)
  - Objective Function Code: (to be assigned by IANA, recommended 13)
  - Description: Find a path P such that does not entry a domain more than once.

3.1.2. OPEN Object Flags

There are two OPEN object flags defined here for H-PCE:

- Parent PCE request bit (to be assigned by IANA, recommended bit 0): if set it means the child PCE wishes to use the peer PCE as a parent PCE.

- o Parent PCE indication bit (to be assigned by IANA, recommended bit 1): if set it means the PCE can be used as a parent PCE by the peer PCE.
- o Editors Note. It is possible that a parent PCE will also act as a child PCE.

3.1.3. Domain-ID TLV

The type of Domain-ID TLV is to be assigned by IANA (recommended 7). The length is 8 octets. The format of this TLV is defined below:

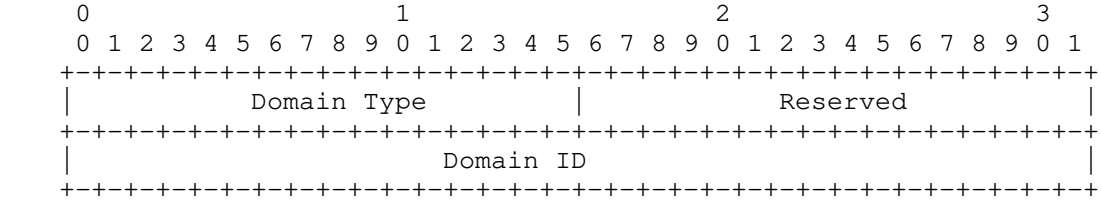


Figure 1: Domain-ID TLV

Domain Type (8 bits): Indicates the domain type. There are two types of domain defined currently:

- o Type=1: the Domain ID field carries an IGP Area ID.
- o Type=2: the Domain ID field carries an AS number.

Domain ID (32 bits): Indicates an IGP Area ID or AS number.

An AS number may be 2 or 4 bytes long. For 2-byte AS numbers, the AS value is left-padded with 0.

Editor's note: it may be necessary to support 64 bit domain IDs.

Editor's note: draft-dhody-pce-pcep-domain-sequence, section 3.2 deals with the encoding of domain sequences, using ERO-subobjects. Work is ongoing to define domain identifiers for OSPF-TE areas, IS-IS area (which are variable sized), 2-byte and 4-byte AS number, and any other domain that may be defined in the future. It uses RSVP-TE subobject discriminators, rather than new type 1/ type 2. A domain sequence may be encoded as a route object. The "VALUE" part of the TLV could follow common RSVP-TE subobject format:

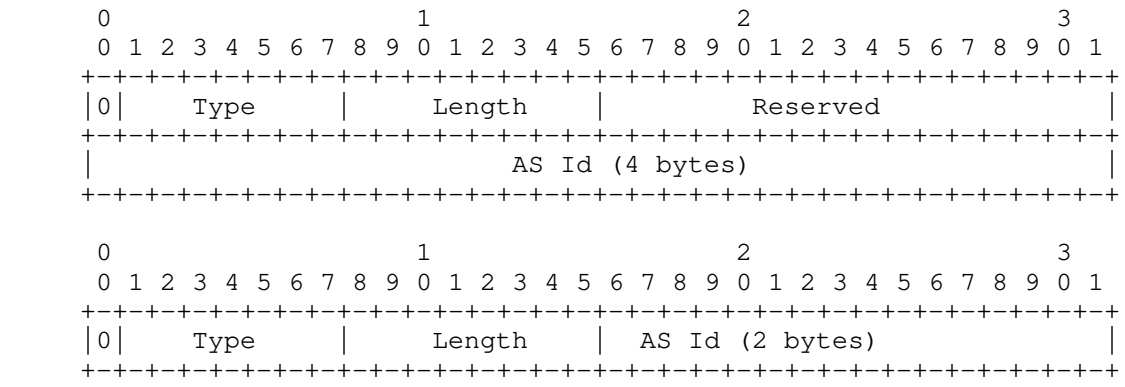


Figure 2: Alternative Domain-ID TLV

3.1.4. PCE-ID TLV

The type of PCE-ID TLV is to be assigned by IANA (recommended 8). The length is 8. The format of this TLV is defined below:

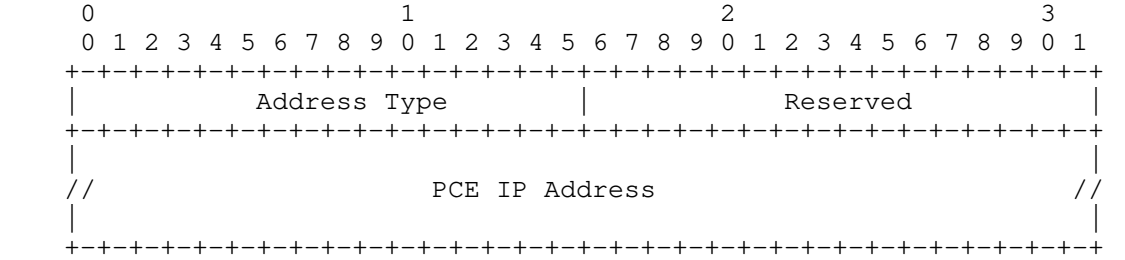


Figure 3: PCE-ID TLV

Address Type (16 bits): Indicates the address type of PCE IP Address. 1 means IPv4 address type, 2 means IPv6 address type.

PCE IP Address: Indicates the reachable address of a PCE.

Editor's note: RFC5886 already defines the PCE-ID object. If a semantically equivalent PCE-ID TLV is needed (to avoid modifying message grammars to include the object), it can align with the PCEP object: n any case, the length (4 / 16 bytes) can be used to know whether it is an IPv4 or an IPv6 PCE, the address type is not needed.

3.1.5. Procedures

The OF codes defined in this document can be carried in the OF-list TLV of the OPEN object. If the OF-list TLV carries the OF codes, it

means that the PCE is capable of implementing the corresponding objective functions. This information can be used for selecting a proper parent PCE when a child PCE wants to get a path that satisfies a certain objective function.

If a child PCE wants to use the peer PCE as a parent, it can set the parent PCE request bit in the OPEN object carried in the Open message during the PCEP session creation procedure. If the peer PCE does not want to provide the parent function to the child PCE, it must send a PCErr message to the child PCE and clear the parent PCE indication bit in the OPEN object.

If the parent PCE can provide the parent function to the peer PCE, it may set the parent PCE indication bit in the OPEN object carried in the Open message during the PCEP session creation procedure.

The PCE may also report its PCE ID and list of domain ID to the peer PCE by specifying them in the PCE-ID TLV and List of Domain-ID TLVs in the OPEN object carried in the Open message during the PCEP session creation procedure.

3.2. Extensions to RP object

3.2.1. RP Object Flags

- o Domain Path Request bit (to be assigned by IANA, recommended bit 17): if set it means the child PCE wishes to get the domain sequence.
- o Destination Domain Query bit (to be assigned by IANA, recommended bit 16): if set it means the parent PCE wishes to get the destination domain ID.

3.2.2. Domain-ID TLV

The format of this TLV is defined in section Section 3.1.3. This TLV can be carried in an OPEN object to indicate a (list of) managed domains, or carried in a RP object to indicate the destination domain ID when a child PCE responds to the parent PCE's destination domain query by a PCRep message.

Editors note. In some cases, the Parent PCE may need to allocate a node which is not necessarily the destination node.

3.2.3. Procedures

If a child PCE only wants to get the domain sequence for a multi-domain path computation from a parent PCE, it can set the Domain Path Request bit in the RP object carried in a PCReq message. The parent

PCE which receives the PCReq message tries to compute a domain sequence for it. If the domain path computation succeeds the parent PCE sends a PCRep message which carries the domain sequence in the ERO to the child PCE . The domain sequence is specified as AS or AREA ERO sub-objects (type 32 for AS [RFC3209] or a to-be-defined IGP arrea type). Otherwise it sends a PCReq message which carries the NO-PATH object to the child PCE.

The parent PCE can set the Destination Domain Query bit in a PCReq message to query the destination (which is specified in the END-POINTS objects) domain ID from a child PCE. If the child PCE knows the destination(s) domain ID, it sends a PCRep message to the parent PCE and specifies the domain ID in the Domain-ID TLV which is carried in the RP object. Otherwise it sends a PCRep message with a NO-PATH object to the parent PCE.

3.3. Extensions to NOTIFICATION object

Because there will not be too many PCEP sessions between the child PCE(s) and parent PCE, it is recommended that the PCEP sessions between them keeping alive all the time . Then the child PCE can report all of the domain connectivity information to the parent PCE when the PCEP session is established successfully. It can also notify the parent PCE to update or delete the domain connectivity information when it detects the changes.

3.3.1. Notification Types

There is a new notification type defined in this document:

- Domain Connectivity Information notification-type (to be assigned by IANA, recommended 3).
  - Notification-value=0: sent from the parent to the child to query all of the domain connectivity information maintained by the child PCE.
  - Notification-value=1: sent from the child to the parent to update the domain connectivity information maintained by the child PCE.
  - Notification-value=2: sent from the child to the parent to delete the domain connectivity information maintained by the child PCE.

3.3.2. Inter-domain Link TLV

IGP in each neighbor domain can advertise its inter-domain TE link capabilities [RFC5316], [RFC5392]. This information can be collected by the child PCEs and forwarded to the parent PCE. PCEP Inter-domain Link TLV is used for carrying the inter-domain TE link attributes for

this purpose. Each Inter-domain Link TLV can carry the attributes of one inter-domain link at the most.

The type of Inter-domain Link TLV is to be assigned by IANA (recommended 9). The length is variable. The format of this TLV is defined below:

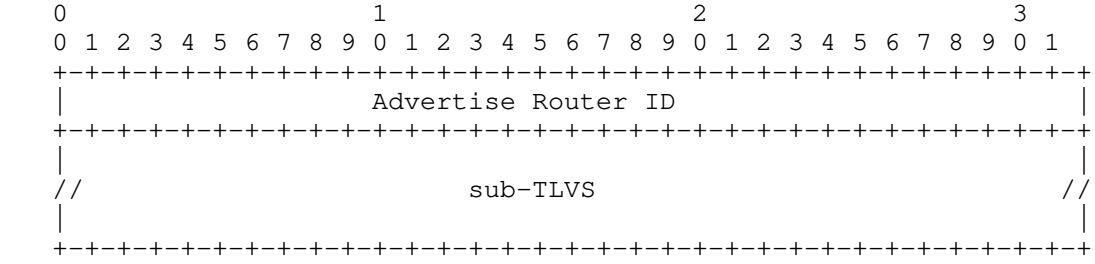


Figure 4: Inter-domain Link TLV

Editor's note: evaluate other possibilities regarding the wrapping and encoding (LSAs / LSUs). Other fields may be needed, such as LSA age (max age methods can be used to "withdraw" or remove a link). Sub-TLVs may need to be defined in the context of a Link TLV (top TLV).

Advertise Router ID (32 bits): indicates the router ID which advertises the TE LSA or LSP.

Sub-TLVs: the OSPF sub-TLVs for a TE link which defined in [RFC5392] and other associated OSPF RFCs. It is noted that if the IGP is IS-IS for the child domain the sub-TLVs must be converted to the OSPF sub-TLVs format when sending this information to the parent PCE through PCEP PCNtf message.

Each inter-domain link is identified by the combination of advertise router ID and the link local IP address or link local unnumbered identifier. The PCNtf message which is used for notifying the parent PCE to update or delete a inter-domain link must contain the information identifies a TE link exclusively.

3.3.3. Inter-domain Node TLV

The Inter-domain Node TLV carries only the two adjacent domain ID and the router (IGP ABR) ID.

The type of Inter-domain Node Information TLV is to be assigned by IANA (recommended 10). The length is variable . The format of this TLV is defined below:

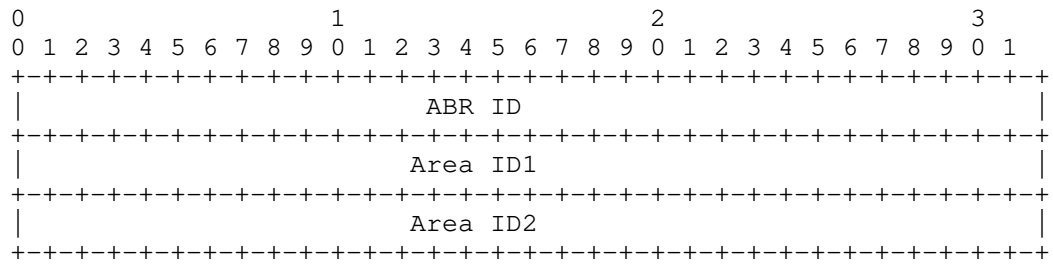


Figure 5: Inter-domain Node TLV

ABR ID (32 bits): indicates the domain border router ID.

Area ID1 & Area ID2 (32 bits): indicates the two neighbor area IDs.

Editor's note (1): a node may be an inter-domain node for more than just 2 areas, the encoding is wrong, unless we explicitly state that this TLV can be repeated and we give an example. Alternatively, we can use the generic concept of "domain id" as introduced earlier, to avoid the restriction of 4 byte areas only.

Editor's note (2): do we homogenize so we also have a Advertising Router ID? would it be different from the ABR id?

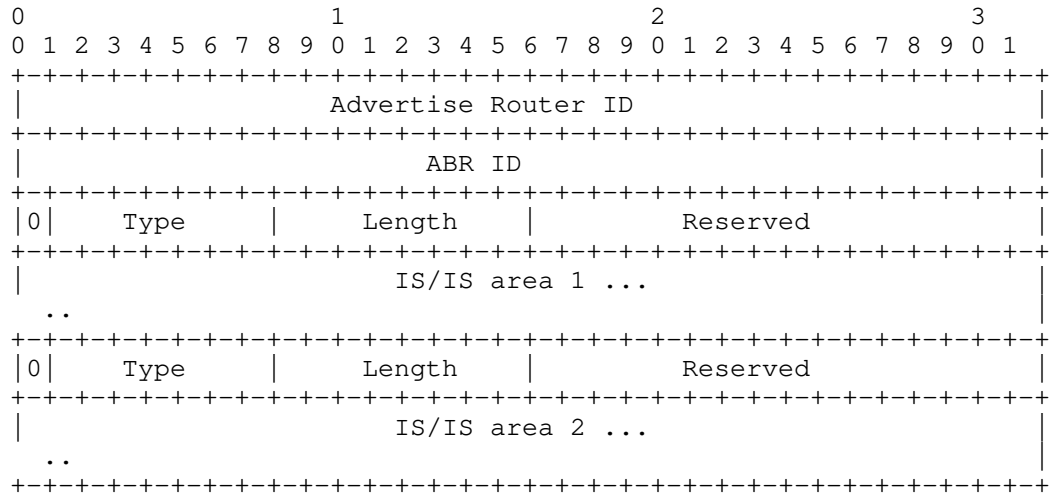


Figure 6: Alternative Inter-domain Node TLV

3.3.4. Domain-ID TLV

The format of this TLV is defined in section Section 3.1.3. This TLV can be carried in a NOTIFICATION object to indicate the domain ID of the PCE who sends the PCNtf message.

Editors note. A PCE may be responsible for several domains, it may be beneficial to use a list of TLVs.

3.3.5. PCE-ID TLV

The format of this TLV is defined in section Section 3.1.4. This TLV can be carried in a NOTIFICATION object to indicate the PCE ID of the PCE who sends the PCNtf message.

3.3.6. Procedures

When a parent PCE establishes a PCEP session with a child PCE successfully, the parent PCE may request the child PCE to report the domain connectivity information. This procedure can be done by sending a PCNtf message from the parent to the child, setting the notification-type to 3 and notification-value to 0 in the NOTIFICATION object.

When a child PCE receives the PCNtf message, it may send all of the domain connectivity information to the parent PCE by the PCNtf message(s). The notification-type is 3 and notification-value is 1 in the NOTIFICATION object. The NOTIFICATION object may carry the inter-domain link TLV and inter-domain node TLV to describe the inter-domain connectivity information. It is noted that if the child PCE dose not support this function, it will ignore the received PCNtf message and the parent PCE will not receive the response.

The child PCE can also update the domain connectivity information by re-sending the PCNtf message(s) with the newly information.

When the child PCE detects a deletion of domain connectivity (e.g., the inter-domain link TLV is aged out), it must notify the parent PCE to delete the inter-domain link by sending the PCNtf message. The notification-type is 3 and notification-value is 2 in the NOTIFICATION object.

3.4. Extensions to PCEP-ERROR object

3.4.1. Hierarchy PCE Error-Type

A new PCEP Error-Type is allocated for hierarchy PCE (to be assigned by IANA, recommended 19):

Error-Type	Meaning
19	H-PCE error Error-value=1: parent PCE capability cannot be provided

Table 1: H-PCE error table

3.4.2. Procedures

When a specific child PCE sends a PCReq to a peer PCE that requires parental activity and the peer PCE does not want to act as the parent for it, the peer PCE should send a PCErr message to the child PCE and specify the error-type (IANA) and error-value (1) in the PCEP-ERROR object.

4. Acknowledgements

5. Contributors

TBD.

6. Manageability Considerations

TBD.

7. IANA Considerations

As per RFC 5226 [RFC5226], IANA is requested to create/update the following registries

7.1. Objective Function (OF) codes

Value	Meaning	Reference
11	MBN	This document
12	MTD	This document
13	DDR	This document

7.2. OPEN Object Flags

7.3. RP Object Flags

7.4. PCEP TLVs

Value	Meaning	Reference
x	Interdomain Link TLV	This document (section Section 3.3.2)
x	Interdomain Node TLV	This document (section Section 3.3.3)

7.5. PCEP NOTIFICATION types

Type	Value	Meaning
P2C Notification	1	
	2	
	3	
C2P Notification	1	
	2	
	3	

7.6. PCEP PCEP-ERROR types

Type	Value	Meaning
H-PCE Error 19	1	parent PCE capability cannot be provided
	2	TBD
	3	TBD

8. Security Considerations

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.

[RFC5152] Vasseur, JP., Ayyangar, A., and R. Zhang, "A Per-Domain Path Computation Method for Establishing Inter-Domain Traffic Engineering (TE) Label Switched Paths (LSPs)", RFC 5152, February 2008.

[RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.

[RFC5441] Vasseur, JP., Zhang, R., Bitar, N., and JL. Le Roux, "A Backward-Recursive PCE-Based Computation (BRPC) Procedure to Compute Shortest Constrained Inter-Domain Traffic

Engineering Label Switched Paths", RFC 5441, April 2009.

9.2. Informative References

[I-D.ietf-pce-hierarchy-fwk]
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", draft-ietf-pce-hierarchy-fwk-00 (work in progress), October 2011.

[RFC2629] Rose, M., "Writing I-Ds and RFCs using XML", RFC 2629, June 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.

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

[RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.

[RFC5316] Chen, M., Zhang, R., and X. Duan, "ISIS Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 5316, December 2008.

[RFC5392] Chen, M., Zhang, R., and X. Duan, "OSPF Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 5392, January 2009.

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

Quintin Zhao
Huawei Technologies
125 Nagog Technology Park
Acton, MA 01719
US

Phone:
Email: qzhao@huawei.com

Oscar Gonzalez de Dios
Telefonica I+D
Emilio Vargas, 6
Madrid,
Spain

Phone:
Email: ogondio@tid.es

Ramon Casellas
CTTC
Av. Carl Friedrich Gauss n.7
Castelldefels, Barcelona
Spain

Phone: +34 93 645 29 00
Email: ramon.casellas@cttc.es

Daniel King
Old Dog Consulting

Phone:
Email: daniel@olddog.co.uk

Network Working Group
Internet-Draft
Intended status: Experimental
Expires: January 14, 2014

F. Zhang, Ed.
Q. Zhao
Huawei
O. Gonzalez de Dios, Ed.
Telefonica I+D
R. Casellas
CTTC
D. King
Old Dog Consulting
July 14, 2013

Extensions to Path Computation Element Communication Protocol (PCEP) for Hierarchical Path Computation Elements (PCE)
draft-zhang-pce-hierarchy-extensions-04

Abstract

The Hierarchical Path Computation Element (H-PCE) architecture, defined in the companion framework document [RFC6805], provides a mechanism to allow the optimum sequence of domains to be selected, and the optimum end-to-end path to be derived through the use of a hierarchical relationship between domains.

This document defines the Path Computation Element Protocol (PCEP) extensions for the purpose of implementing Hierarchical PCE procedures which are described in the aforementioned document. These extensions are experimental and published for examination, discussion, implementation, and evaluation.

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 January 14, 2014.

Copyright Notice

Table of Contents

1. Introduction

[RFC6805] describes a Hierarchical PCE (H-PCE) architecture which can be used for computing end-to-end paths for inter-domain MPLS Traffic Engineering (TE) and GMPLS Label Switched Paths (LSPs).

Within the hierarchical PCE architecture, the parent PCE is used to compute a multi-domain path based on the domain connectivity information. A child PCE may be responsible for a single domain or multiple domains, it is used to compute the intra-domain path based on its domain topology information.

The H-PCE end-to-end domain path computation procedure is described below:

- A path computation client (PCC) sends the inter-domain path computation requests to the child PCE responsible for its domain;
- The child PCE forwards the request to the parent PCE;
- The parent PCE computes the likely domain paths from the ingress domain to the egress domain;
- The parent PCE sends the intra-domain path computation requests (between the domain border nodes) to the child PCEs which are responsible for the domains along the domain path;
- The child PCEs return the intra-domain paths to the parent PCE;
- The parent PCE constructs the end-to-end inter-domain path based on the intra-domain paths;
- The parent PCE returns the inter-domain path to the child PCE;
- The child PCE forwards the inter-domain path to the PCC.

In addition, the parent PCE may be requested to provide only the sequence of domains to a child PCE so that alternative inter-domain path computation procedures, including Per Domain (PD) [RFC5152] and Backwards Recursive Path Computation (BRPC) [RFC5441] may be used.

This document defines the PCEP extensions for the purpose of implementing Hierarchical PCE procedures, which are described in [RFC6805].

1.1. Scope

The following functions are out of scope of this document.

- Finding end point addresses;
- Parent Traffic Engineering Database (TED) methods;
- Domain connectivity;

The document also uses a number of [editor notes] to describe options and alternative solutions. These options and notes will be removed before publication once agreement is reached.

1.2. Terminology

This document uses the terminology defined in [RFC4655], [RFC5440] and the additional terms defined in section 1.4 of [RFC6805].

1.3. Requirements Language

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

2. Requirements for H-PCE

This section compiles the set of requirements of the PCEP protocol to support the H-PCE architecture and procedures.

[RFC6805] identifies high-level requirements of PCEP extensions required to support the hierarchical PCE model.

2.1. PCEP Requests

The PCReq messages are used by a PCC or PCE to make a path computation request to a PCE. In order to achieve the full functionality of the H-PCE procedures, the PCReq message needs to include:

- Qualification of PCE Requests.
- Multi-domain Objective Functions (OF).
- Multi-domain Metrics.

2.1.1. Qualification of PCEP Requests

As described in section 4.8.1 of [RFC6805], the H-PCE architecture introduces new request qualifications, which are:

- It MUST be possible for a child PCE to indicate that a request it sends to a parent PCE should be satisfied by a domain sequence only, that is, not by a full end-to-end path. This allows the child PCE to initiate a per-domain (PD) [RFC5152] or a backward recursive path computation (BRPC) [RFC5441].

- As stated in [RFC6805], section 4.5, if a PCC knows the egress domain, it can supply this information as the path computation request. It SHOULD be possible to specify the destination domain information in a PCEP request, if it is known.

2.1.2. Multi-domain Objective Functions

For inter-domain path computation, there are two new objective functions which are defined in section 1.3.1 and 4.1 of [RFC6805]:

- Minimize the number of domains crossed. A domain can be either an Autonomous System (AS) or an Internal Gateway Protocol (IGP) area depending on the type of multi-domain network hierarchical PCE is applied to.

- Disallow domain re-entry.[Editor's note: Disallow domain re-entry may not be an objective function, but an option in the request].

During the PCEP session establishment procedure, the parent PCE needs to be capable of indicating the Objective Functions (OF) capability in the Open message. This capability information may then be announced by child PCEs, and used for selecting the PCE when a PCC wants a path that satisfies one or multiple inter-domain objective functions.

When a PCC requests a PCE to compute an inter-domain path, the PCC needs also to be capable of indicating the new objective functions for inter-domain path. Note that a given child PCE may also act as a parent PCE.

For the reasons described previously, new OF codes need to be defined for the new inter-domain objective functions. Then the PCE can notify its new inter-domain objective functions to the PCC by carrying them in the OF-list TLV which is carried in the OPEN object. The PCC can specify which objective function code to use, which is carried in the OF object when requesting a PCE to compute an inter-domain path.

The proposed solution may need to differentiate between the OF code that is requested at the parent level, and the OF code that is requested at the intra-domain (child domain).

A parent PCE MUST be capable of ensuring homogeneity, across domains, when applying OF codes for strict OF intra-domain requests.

2.1.3. Multi-domain Metrics

For inter-domain path computation, there are several path metrics of interest [Editor's note: Current framework only mentions metric objectives. The metric itself should be also defined]:

- Domain count (number of domains crossed).
- Border Node count.

A PCC may be able to limit the number of domains crossed by applying a limit on these metrics.

2.2. Parent PCE Capability Discovery

Parent and child PCE relationships are likely to be configured. However, as mentioned in [RFC6805], it would assist network operators if the child and parent PCE could indicate their H-PCE capabilities.

During the PCEP session establishment procedure, the child PCE needs to be capable of indicating to the parent PCE whether it requests the parent PCE capability or not. Also, during the PCEP session establishment procedure, the parent PCE needs to be capable of indicating whether its parent capability can be provided or not.

2.3. PCE Domain and PCE ID Discovery

A PCE domain is a single domain with an associated PCE. Although it is possible for a PCE to manage multiple domains. The PCE domain may be an IGP area or AS.

The PCE ID is an IPv4 and/or IPv6 address that is used to reach the parent/child PCE. It is RECOMMENDED to use an address that is always reachable if there is any connectivity to the PCE.

The PCE ID information and PCE domain identifiers may be provided during the PCEP session establishment procedure or the domain connectivity information collection procedure.

3. PCEP Extensions (Encoding)

3.1. OPEN object

3.1.1. OF Codes

This H-PCE experiment will be carried out using the following OF codes:

- MTD
  - Name: Minimize the number of Transit Domains.
  - Objective Function Code.
  - Description: Find a path P such that it passes through the lnumber of transit domains.
- MBN
  - Name: Minimize the number of border nodes.
  - Objective Function Code.
  - Description: Find a path P such that it passes through the least number of border nodes.
- DDR
  - Name: Disallow Domain Re-entry (DDR)
  - Objective Function Code.
  - Description: Find a path P such that does not entry a domain more than once.

3.1.2. OPEN Object Flags

This H-PCE experiment will also require two OPEN object flags:

- Parent PCE Request bit (to be assigned by IANA, recommended bit 0): if set, it would signal that the child PCE wishes to use the peer PCE as a parent PCE.
- Parent PCE Indication bit (to be assigned by IANA, recommended bit 1): if set, it would signal that the PCE can be used as a parent PCE by the peer PCE.

3.1.3. Domain-ID TLV

The Domain-ID TLV for this H-PCE experiment is defined below:

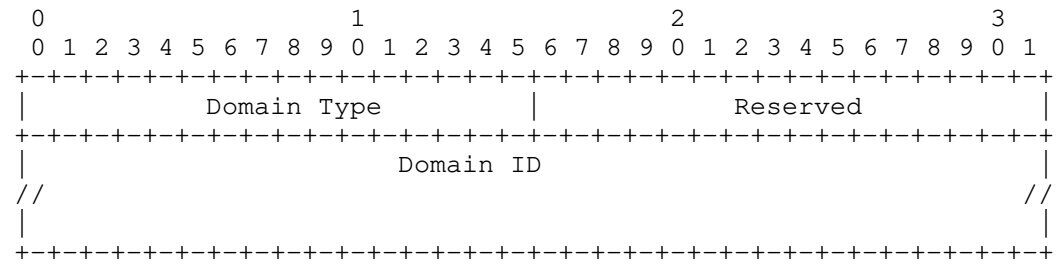


Figure 1: Domain-ID TLV

Domain Type (8 bits): Indicates the domain type. Two types of domain are currently defined:

- Type=1: the Domain ID field carries an IGP Area ID.
- Type=2: the Domain ID field carries an AS number.

Domain ID (variable): Indicates an IGP Area ID or AS number. It can be 2 bytes, 4 bytes or 8 bytes long depending on the domain identifier used.

[Editor's note: draft-dhody-pce-pcep-domain-sequence, section 3.2 deals with the encoding of domain sequences, using ERO-subobjects. Work is ongoing to define domain identifiers for OSPF-TE areas, IS-IS area (which are variable sized), 2-byte and 4-byte AS number, and any other domain that may be defined in the future. It uses RSVP-TE subobject discriminators, rather than new type 1/ type 2. A domain sequence may be encoded as a route object. The "VALUE" part of the TLV could follow common RSVP-TE subobject format:

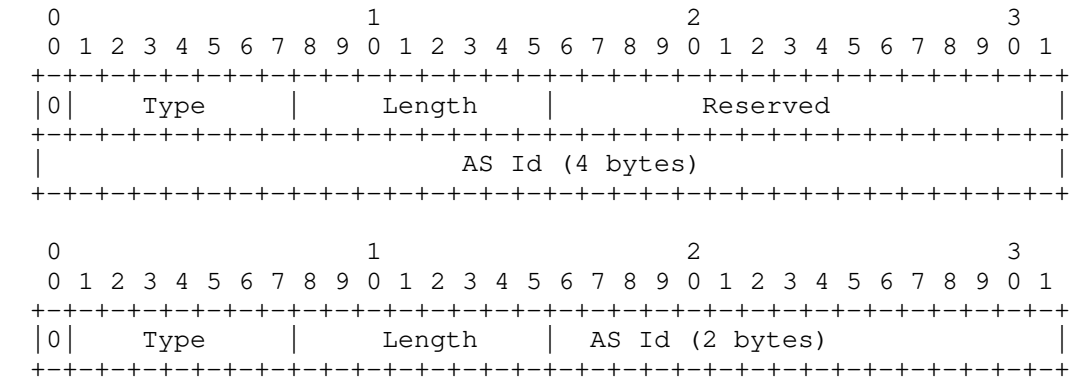


Figure 2: Alternative Domain-ID TLV

3.1.4. PCE-ID TLV

The type of PCE-ID TLV for this H-PCE experiment is defined below:

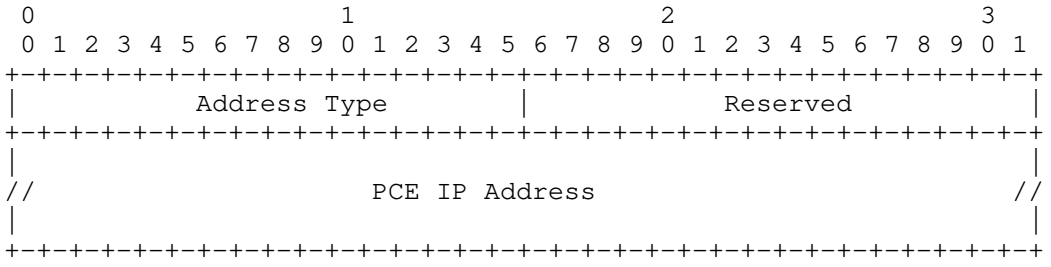


Figure 3: PCE-ID TLV

Address Type (16 bits): Indicates the address type of PCE IP Address. 1 means IPv4 address type, 2 means IPv6 address type.

PCE IP Address: Indicates the reachable address of a PCE.

[Editor's note: [RFC5886] already defines the PCE-ID object. If a semantically equivalent PCE-ID TLV is needed (to avoid modifying message grammars to include the object), it can align with the PCEP object: in any case, the length (4 / 16 bytes) can be used to know whether it is an IPv4 or an IPv6 PCE, the address type is not needed.]

3.2. RP object

3.2.1. RP Object Flags

The following RP object flags are defined for this H-PCE experiment:

- Domain Path Request bit: if set, it means the child PCE wishes to get the domain sequence.

- Destination Domain Query bit: if set, it means the parent PCE wishes to get the destination domain ID.

3.2.2. Domain-ID TLV

The format of this TLV is defined in Section 3.1.3. This TLV can be carried in an OPEN object to indicate a (list of) managed domains, or carried in a RP object to indicate the destination domain ID when a child PCE responds to the parent PCE's destination domain query by a PCRep message.

[Editors note. In some cases, the Parent PCE may need to allocate a node which is not necessarily the destination node.]

3.3. Metric Object

There are two new metrics defined in this document for H-PCE:

- Domain count (number of domains crossed).
- Border Node Count (number of border nodes crossed).

3.4. PCEP-ERROR object

3.4.1. Hierarchy PCE Error-Type

A new PCEP Error-Type is used for this H-PCE experiment and is defined below:

Error-Type	Meaning
19	H-PCE error Error-value=1: parent PCE capability cannot be provided

H-PCE error table

3.5. NO-PATH Object

To communicate the reason(s) for not being able to find a multi-domain path or domain sequence, the NO-PATH object can be used in the PCRep message. [RFC5440] defines the format of the NO-PATH object. The object may contain a NO-PATH-VECTOR TLV to provide additional information about why a (domain) path computation has failed.

Three new bit flags are defined to be carried in the Flags field in the NO-PATH-VECTOR TLV carried in the NO-PATH Object.

- Bit 23: When set, the parent PCE indicates that destination domain unknown;
- Bit 22: When set, the parent PCE indicates unresponsive child PCE(s);
- Bit 21: When set, the parent PCE indicates no available resource available in one or more domain(s).

4. H-PCE Procedures

4.1. OPEN Procedure between Child PCE and Parent PCE

If a child PCE wants to use the peer PCE as a parent, it can set the parent PCE request bit in the OPEN object carried in the Open message during the PCEP session creation procedure. If the peer PCE does not want to provide the parent function to the child PCE, it must send a PCErr message to the child PCE and clear the parent PCE indication bit in the OPEN object.

If the parent PCE can provide the parent function to the peer PCE, it may set the parent PCE indication bit in the OPEN object carried in the Open message during the PCEP session creation procedure.

The PCE may also report its PCE ID and list of domain ID to the peer PCE by specifying them in the PCE-ID TLV and List of Domain-ID TLVs in the OPEN object carried in the Open message during the PCEP session creation procedure.

The OF codes defined in this document can be carried in the OF-list TLV of the OPEN object. If the OF-list TLV carries the OF codes, it means that the PCE is capable of implementing the corresponding objective functions. This information can be used for selecting a proper parent PCE when a child PCE wants to get a path that satisfies a certain objective function.

When a specific child PCE sends a PCReq to a peer PCE that requires parental activity and the peer PCE does not want to act as the parent for it, the peer PCE should send a PCErr message to the child PCE and specify the error-type (IANA) and error-value (1) in the PCEP-ERROR object.

4.2. Procedure to obtain Domain Sequence

If a child PCE only wants to get the domain sequence for a multi-domain path computation from a parent PCE, it can set the Domain Path Request bit in the RP object carried in a PCReq message. The parent PCE which receives the PCReq message tries to compute a domain sequence for it. If the domain path computation succeeds the parent PCE sends a PCRep message which carries the domain sequence in the ERO to the child PCE. The domain sequence is specified as AS or AREA ERO sub-objects (type 32 for AS [RFC3209] or a to-be-defined IGP area type). Otherwise it sends a PCReq message which carries the NO-PATH object to the child PCE.

5. Error Handling

A PCE that is capable of acting as a parent PCE might not be configured or willing to act as the parent for a specific child PCE.

This fact could be determined when the child sends a PCReq that requires parental activity (such as querying other child PCEs), and could result in a negative response in a PCEP Error (PCErr) message and indicate the hierarchy PCE error types.

Additionally, the parent PCE may fail to find the multi-domain path or domain sequence due to one or more of the following reasons:

- A child PCE cannot find a suitable path to the egress;
- The parent PCE do not hear from a child PCE for a specified time;
- The objective functions specified in the path request cannot be met.

In this case, the parent PCE MAY need to send a negative path computation reply specifying the reason. This can be achieved by including NO-PATH object in the PCRep message. Extension to NO-PATH object is needed to include the aforementioned reasons.

6. Manageability Considerations

TBD.

7. IANA Considerations

Due to the experimental nature of this draft no IANA requests are made.

8. Security Considerations

To be added.

9. Contributing Authors

Xian Zhang
Huawei
zhang.xian@huawei.com

10. Acknowledgments

To be added.

11. 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.

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

[RFC5152] Vasseur, JP., Ayyangar, A., and R. Zhang, "A Per-Domain Path Computation Method for Establishing Inter-Domain Traffic Engineering (TE) Label Switched Paths (LSPs)", RFC 5152, February 2008.

[RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.

[RFC5441] Vasseur, JP., Zhang, R., Bitar, N., and JL. Le Roux, "A Backward-Recursive PCE-Based Computation (BRPC) Procedure to Compute Shortest Constrained Inter-Domain Traffic Engineering Label Switched Paths", RFC 5441, April 2009.

[RFC5886] Vasseur, JP., Le Roux, JL., and Y. Ikejiri, "A Set of Monitoring Tools for Path Computation Element (PCE)-Based Architecture", RFC 5886, June 2010.

[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.

Authors' Addresses

Fatai Zhang (editor)
Huawei
Huawei Base, Bantian, Longgang District
Shenzhen, 518129
China

Phone: +86-755-28972912
Email: zhangfatai@huawei.com

Quintin Zhao
Huawei
125 Nagog Technology Park
Acton, MA 01719
US

Phone:
Email: qzhao@huawei.com

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

Phone: +34913128832
Email: ogondio@tid.es

Ramon Casellas
CTTC
Av. Carl Friedrich Gauss n.7
Castelldefels, Barcelona
Spain

Phone: +34 93 645 29 00
Email: ramon.casellas@cttc.es

Daniel King
Old Dog Consulting
UK

Phone:
Email: daniel@olddog.co.uk

Network Working Group
Internet-Draft
Intended status: Informational
Expires: April 26, 2012

YL. Zhao
J. Zhang
TT. Peng
XS. Yu
BUPT
XP. Cao
DJ. Wang
ZTE Corporation

October 24, 2011

PCEP Protocol Extension for spectrum utilization optimization in Flexi-Grid Networks
draft-zhaoyl-pce-flexi-grid-pcep-ex-00

Abstract

Flexi-grid networks overcomes the fixed grid channel of Wavelength Switched Optical Network(WSON) by flexible spectrum to allow non-uniform and dynamic allocation of spectrum based on the demand of the incoming services' LSP. Flexi-grid networks is an effective solution to solve the problem of efficient spectrum utilization.

Because the client LSP needs to be assigned contiguous spectrum in flexi-grid networks, there will be two problems that would affect spectrum utilization, i.e. RSA and fragmentation. We introduce two kinds of methods which can improve the spectrum utilization further, and some related PCEP extensions are defined in this document.

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 26, 2012.

Copyright Notice

Table of Contents

1. Introduction

Demand of traffic is increasing exponentially and already approaching the limit of single mode fiber capacity. At the same time, because of varying demand of traffic, we need an efficient and agile utilization of the optical spectrum.

ITU-T Study Group 15 introduce a new flexi-grid networks to enable dynamic allocation of spectrum resource. The flexi-grid networks is an effective solution to solve the problem of efficient spectrum resource utilization.

The granularity of flexi-grid networks can be smaller and agile. i.e.,6.25GHz. In the flexi-grid networks, the appropriate size of spectrum is determined by the used modulation format.According to the client data rate LSP and physical consecutives of the selected path,the appropriate size of spectrum is adaptively allocated to optical connections by assigning the appropriate number of contiguous spectrum from end-to-end.Before assigning the client LSP, we have to find suitable route and fit contiguous spectrum for it, and it is a complex process. So spectrum utilization is very important in RSA. While there are several algorithms for RSA, flexi-grid networks require to extend PCEP protocol to support different algorithms seletion.

Upon tearing down of connections, the allocated spectrum are released for future LSPs. In a dynamic traffic scenario, this setup and tear down precedure for a channel leads to fragmentation of spectral resources. Due to the fragmentation, the available spectrum is divided into small noncontiguous spectral bands,the spectral effciency in the network is compromised. Therefore the probability of finding suffcient contiguous spectrum for a connection is decreased. We introduce defragmentation to deal with fragmentation in flexi-grid networks. then PCEP protocol has to add some messages to support them.

2. Conventions Used in This Document

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

3. Terminologies

RSA: Routing and Spectrum Assignment

WSON:Wavelength Switched Optical Network

4. RSA

4.1. Introduction of RSA

In this part, we focus on the routing and spectrum assignment (RSA) problem. This problem can be partitioned into two subproblems: (1) routing and (2) spectrum assignment, and each subproblem can be solved separately. Different from traditional WDM network, flexi-grid networks assign continuous spectrum for new arrival LSP. Static planning models are used for flexi-grid networks to improve spectrum utilization.

4.2. Algorithms of RSA

There are several spectrum assignment algorithms.

(1)Random Fit (RF)

This scheme first searches the space of spectrum to determine the set of all spectrum that are available on the required route. Among the available spectrum, one is chosen randomly.

(2)First-Fit (FF)

In this scheme, all spectrum is numbered.When searching for available spectrum, a lower numbered spectrum is considered before a higher-numbered spectrum.The first available spectrum is then selected. Compared to Random spectrum assignment, the computation cost of this scheme is lower because there is no need to search the entire spectrum space for each route.

(3)Least-Used (LU)/SPREAD

LU selects the spectrum that is the least used in the network, thereby attempting to balance the load among all the spectrum. The performance of LU is worse than Random, while also introducing additional communication overhead (e.g., global information is required to compute the least-used spectrum).

(4)Most-Used (MU)/PACK

MU is the opposite of LU in that it attempts to select the most-used spectrum in the network. The communication overhead, storage, and computation cost are all similar to those in LU.MU also slightly outperforms FF, doing a better job of packing connections into fewer

spectrum and conserving the spare capacity of less-used spectrum.

(5)Min-Product (MP)

MU is the opposite of LU works. In a single fiber network, MP becomes FF. The goal of MP is to pack spectrum into fibers, thereby minimizing the number of fibers in the network.

(6)Least-Loaded (LL)

The LL heuristic, like MP, is also designed for multi-fiber networks. This heuristic selects the spectrum that has the largest residual capacity on the most loaded link along route.

(7)MAX-SUM (MS)

MS was proposed for multi-fiber networks but it can also be applied to the single-fiber case.MS considers all possible paths in the network and attempts to maximize the remaining path capacities after lightpath establishment.

(8)Relative Capacity Loss (RCL)

RCL is based on MS. RCL chooses spectrum to minimize the relative capacity loss. RCL is based on the observation that minimizing total capacity loss sometimes does not lead to the best choice of spectrum.

(9)Spectrum Reservation (Rsv)

In Rsv, a given spectrum on a specified link is reserved for a traffic stream, usually a multihop stream. This scheme reduces the blocking for multihop traffic,while increasing the blocking for connections that traverse only one fiber link (single-hop traffic).

(10)Protecting Threshold (Thr)

In Thr, a single-hop connection is assigned spectrum only if the number of idle spectrum on the link is at or above a given threshold.

4.3. RSA Schemes Selection

There are several spectrum assignment algorithms, we have to choose one of them for flexi-grid networks. Different RSA schemes are selected according to diffrent network condition. The PCEP protocol needs to extend a bit that shows different schemes selected.

5. Defragmentation

5.1. Motivation of Defragmentation

New arrival of LSPs are then either forced to utilize more spectrum in the network or blocked in spite of suffcient spectrum being available. Additionally, as the network evolves, a current optimal routing scheme might no longer provide the optimal spectral utilization over time. There is an increasing demand from the network operators to be able to periodically reconfigure the network and return it to its optimal state, so that the network can operate more effciently.

5.2. Definition of Defragmentation

There is an operation defined as network defragmentation to solve above problem. Reducing the blocking by consolidating the available network resources, this operation will also enable better network maintenance and more effcient network restoration and bandwidth adjustment.

5.3. Application Scene of Defragmentation

The process of defragmentation: (1) select LSP for defragmentation, and interrupt it considerring the time and cost, (2) choose forward spectrum in original route or new route, (3) move the LSP on possible spectrum.

An example of defragentation is as following: A,B,C are client LSPs on link l, l1 is original statement of link l,l2 is statement of link l after defragementation.

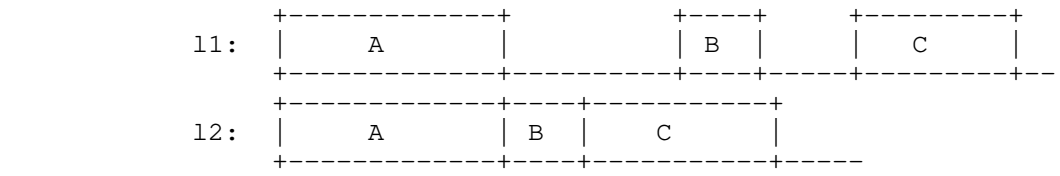


Fig.1 Defragentation principle

we first focus on the problem of the time-point when should defragmentation be operated. There is two ways to solve this problem. One way is new arrival LSPs have no sufficient spectrum to bear, then cause blocked in the network. The other way: (1) collect

the information about occupation of spectrum fragments in a link or in the network, (2) introduce a notation to describe the state of spectrum fragment in a link or in a network, (3)when the size of this notation reaches an assumed threshold, it is the time for defragmentation.

we consider the methods of defragmentation. At present, there is two methods for defragmentation. First is change route of client LSP, meaning that the spectrum of this LSP in new route is ahead than the spectrum in original route. Second is the LSP move forward directly in original route.

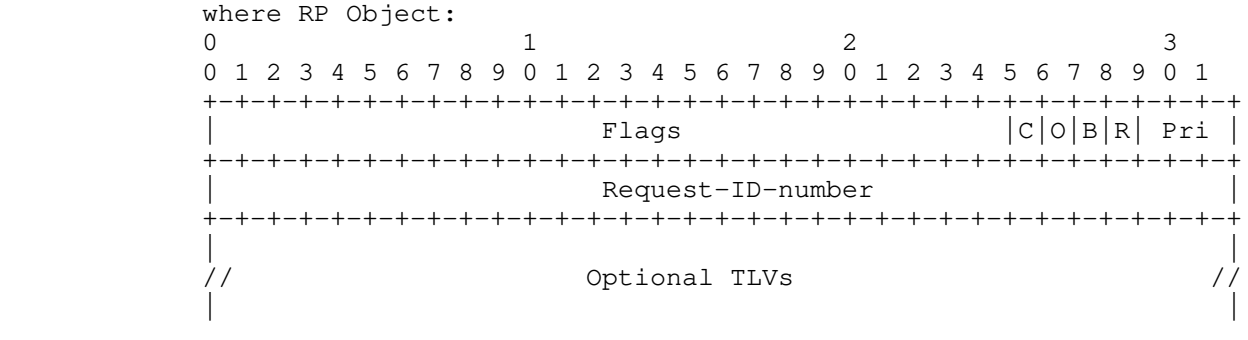
6. PCEP Protocol Extension

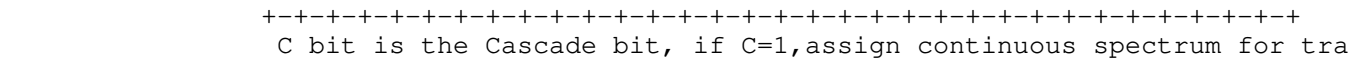
6.1. PCEP Protocol Extension for RSA

The PCEP protocol need to be extended to support the algorithms choosing of RSA. PCReq needs to add RAEO-list information. This information include "Algorithm Id", which stands for the number of different algorithms, and "Pri" that means priority of these algorithms.

```
<request>::= <RP>
             <END-POINTS>
             [<RAEO-list>]
             [<LSPA>]
             [<BANDWIDTH>]
             [<metric-list>]
             [<RRO>[<BANDWIDTH>]]
             [<IRO>]
             [<LOAD-BALANCING>]
```

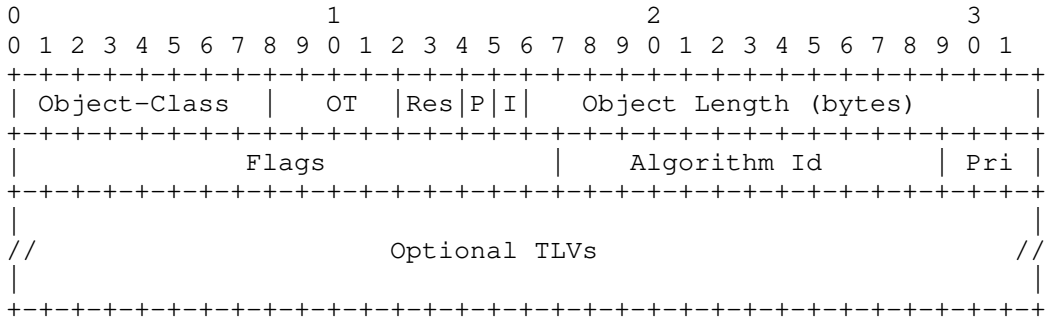
where RP Object:





C bit is the Cascade bit, if C=1,assign continuous spectrum for traffic else assign uncontinuous spectrum.

[<RAEO-list>] defined as follows:



<response>::=<RP>
[<NO-PATH>]
[<attribute-list>]
[<path-list>]

NO-PATH:

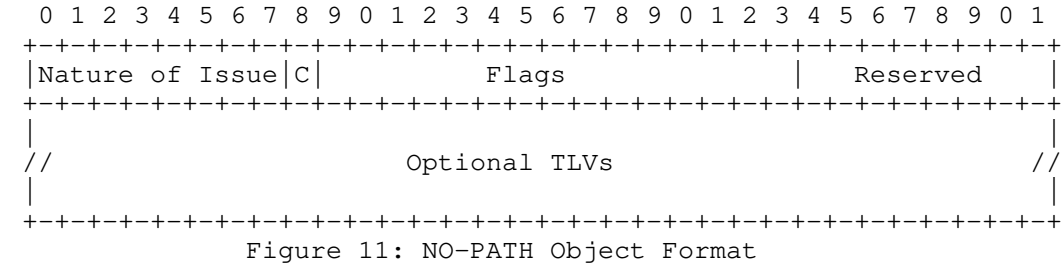


Figure 11: NO-PATH Object Format

NI - Nature of Issue (8 bits): The NI field is used to report the nature of the issue that leads to a negative reply. Two values are currently defined:

0: No path satisfying the set of constraints could be found
1: PCE chain broken
2: No path satisfying the Continuous spectrum

6.2. PCEP Protocol Extension for Defragmentation

The presence of defragmentation in flexi-grid networks has an impact on the information that needs to be transferred by the control plane and PCE. Defragmentation has to interrupt the traffic and move it to another spectrum or route. The PCEP protocol needs to be extended two messages to support defragmentation, ingcluding information of orginal route/spectrum and present route/spectrum, when to stop defragmentation and so on.

Here is Spectrum Defragmentation Request Message and Spectrum Defragmentation Reply Message. "Target Clutter Value" stands for the goal of defragmentation. "R" means whether the network MUST make it.

Spectrum Defragmentation Request Message

```
<SDReq Message>::= <Common Header>
                   <SDTO-list>
                   [LSPA Object]
                   [<RAEO-list>]
```

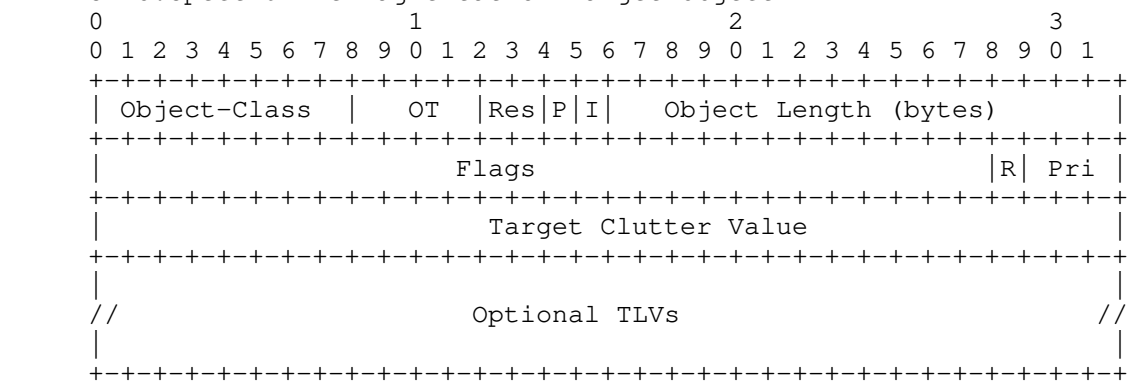
Spectrum Defragmentation Reply Message

```
<SDRep Message>::= <Common Header>
                   <SDTO-list>
                   <Reroute-list>
                   [LSPA Object]
                   [<RAEO-list>]
```

Spectrum Defragmentation Reply Message

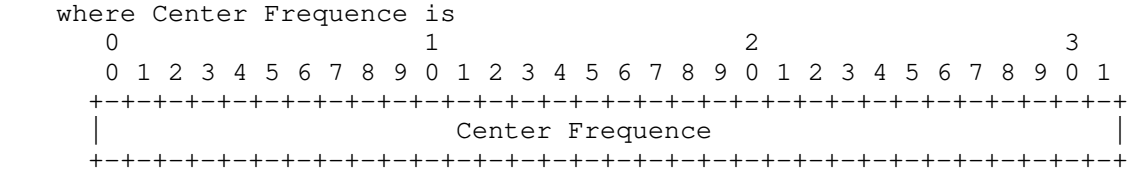
SDTO: Spectrum Defragmentation Target Object

SDTO:Spectrum Defragmentation Target Object



<Reroute>::=<RP Object>
<path><Center Frequence><Bandwidth>
<path><Center Frequence><Bandwidth>

where Center Frequence is



Center Frequence (32 bits): The requested bandwidth is encoded in 32 bits, expressed in bytes per second.

7. Security Considerations

TBD.

8. Normative References

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

[RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.

Authors' Addresses

Yongli Zhao
BUPT
No.10,Xitucheng Road,Haidian District
Beijing 100876
P.R.China

Phone: +8613811761857
Email: yonglizhao@bupt.edu.cn
URI: http://www.bupt.edu.cn/

Jie Zhang
BUPT
No.10,Xitucheng Road,Haidian District
Beijing 100876
P.R.China

Phone: +8613911060930
Email: lgr24@bupt.edu.cn
URI: http://www.bupt.edu.cn/

Tiantian Peng
BUPT
No.10,Xitucheng Road,Haidian District
Beijing 100876
P.R.China

Phone: +8615116984347
Email: tt871228@163.com
URI: http://www.bupt.edu.cn/

Xiaosong Yu
BUPT
No.10,Xitucheng Road,Haidian District
Beijing 100876
P.R.China

Phone: +8613811731723
Email: yu.xiaosong@qq.com
URI: http://www.bupt.edu.cn/

Xuping Cao
ZTE Corporation
No.16,Huayuan Road,Haidian District
Beijing 100191
P.R.China

Phone: +8615801379189
Email: cao.xuping@zte.com.cn
URI: http://www.zte.com.cn/

Dajiang Wang
ZTE Corporation
No.16,Huayuan Road,Haidian District
Beijing 100191
P.R.China

Phone: +8613811795408
Email: wang.dajiang@zte.com.cn
URI: http://www.zte.com.cn/

Network Working Group
Internet-Draft
Intended status: Informational
Expires: October 27, 2012

YL. Zhao
J. Zhang
TT. Peng
XS. Yu
BUPT
XP. Cao
DJ. Wang
XH. Fu
ZTE Corporation

April 25, 2012

PCEP Protocol Extension for spectrum utilization optimization in Flexi-Grid Networks
draft-zhaoyl-pce-flexi-grid-pcep-ex-01

Abstract

Flexi-grid networks overcomes the fixed grid channel of Wavelength Switched Optical Network(WSON) by flexible spectrum to allow non-uniform and dynamic allocation of spectrum based on the demand of the incoming services' LSP. Flexi-grid networks is an effective solution to solve the problem of efficient spectrum utilization.

Because the client LSP needs to be assigned contiguous spectrum in flexi-grid networks, there will be two problems that would affect spectrum utilization, i.e. RSA and fragmentation. We introduce two kinds of methods which can improve the spectrum utilization further, and some related PCEP extensions are defined in this document.

Table of Contents

1. Introduction

Demand of traffic is increasing exponentially and already approaching the limit of single mode fiber capacity. At the same time, because of varying demand of trafficGBP[not]we need an efficient and agile utilization of the optical spectrum.

ITU-T Study Group 15 introduce a new flexi-grid networks to enable dynamic allocation of spectrum resource. The flexi-grid networks is an effective solution to solve the problem of efficient spectrum resource utilization.

The granularity of flexi-grid networks can be smaller and agile. i.e.(6.25GHz).In the flexi-grid networks, the appropriate size of spectrum is determined by the used modulation format.According to the client data rate request and physical constraints of the selected path,the appropriate size of spectrum is adaptively allocated to optical connections by assigning the appropriate number of contiguous spectrum from end-to-end.Before assigning the client request, we have to find suitable route and fit contiguous spectrum for it, and it is a complex process. So spectrum utilization is very important in RSA. While there are several algorithms for RSA, so flexi-grid networks require to extend PCEP protocol to support different algorithms seletion.

Upon tearing down of connections, allocated spectrum are released for future requests. In a dynamic traffic scenario, this channel setup and tear down processes leads to fragmentation of spectral resources. Due to the fragmentation, the available spectrum divide into small noncontiguous spectral bands,the spectral effciency in the network is compromised. Therefore the probability of finding suffcient contiguous spectrum for a connection is decreased. We introduce Spectrum Fragments Cascading and Defragmentation to deal with fragmentation in flexi-grid networks. So PCEP protocol have to add some messages to support them.

2. Conventions Used in This Document

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

3. Terminologies

RSA: Routing and Spectrum Assignment

WSON:Wavelength Switched Optical Network

SFC:Spectrum Fragments Cascading

4. RSA

4.1. Introduction of RSA

This part we focuses on the routing and spectrum assignment (RSA) problem. This problem can be partitioned into two subproblems - (1) routing and (2) wavelength assignment and each subproblem can be solved separately. Different from traditional WDM network, flexi-grid networks assign continuous spectrum for new arrival request. Static planning models used for flexi-grid networks to improve spectrum utilization.

4.2. Algorithms of RSA

There are several spectrum assignment algorithms.

(1)Random Fit (RF)

This scheme first searches the space of wavelengths to determine the set of all spectrum that are available on the required route. Among the available wavelengths, one is chosen randomly.

(2)First-Fit (FF)

In this scheme, all spectrum is numbered.When searching for available spectrum, a lower numbered spectrum is considered before a higher-numbered spectrum.The first available spectrum is then selected. Compared to Random spectrum assignment, the computation cost of this scheme is lower because there is no need to search the entire spectrum space for each route.

(3)Least-Used (LU)/SPREAD

LU selects the spectrum that is the least used in the network, thereby attempting to balance the load among all the spectrum. The performance of LU is worse than Random, while also introducing additional communication overhead (e.g., global information is required to compute the least-used spectrum).

(4)Most-Used (MU)/PACK

MU is the opposite of LU in that it attempts to select the most-used spectrum in the network. The communication overhead, storage, and

computation cost are all similar to those in LU.MU also slightly outperforms FF, doing a better job of packing connections into fewer wavelengths and conserving the spare capacity of less-used wavelengths.

(5)Min-Product (MP)

MU is the opposite of LU works. In a single fiber network, MP becomes FF. The goal of MP is to pack wavelengths into fibers, thereby minimizing the number of fibers in the network.

(6)Least-Loaded (LL)

The LL heuristic, like MP, is also designed for multi-fiber networks. This heuristic selects the spectrum that has the largest residual capacity on the most loaded link along route.

(7)MAX-SUM (MS)

MS was proposed for multi-fiber networks but it can also be applied to the single-fiber case.MS considers all possible paths in the network and attempts to maximize the remaining path capacities after lightpath establishment.

(8)Relative Capacity Loss (RCL)

RCL is based on MS. RCL chooses spectrum to minimize the relative capacity loss. RCL is based on the observation that minimizing total capacity loss sometimes does not lead to the best choice of spectrum.

(9)Spectrum Reservation (Rsv)

In Rsv, a given spectrum on a specified link is reserved for a traffic stream, usually a multihop stream. This scheme reduces the blocking for multihop traffic,while increasing the blocking for connections that traverse only one fiber link (single-hop traffic).

(10)Protecting Threshold (Thr)

In Thr, a single-hop connection is assigned spectrum only if the number of idle spectrum on the link is at or above a given threshold.

4.3. RSA Schemes Selection

There are several spectrum assignment algorithms , we have to choose one of them for use in flexi-grid networks. Diffrent RSA schemes selected according to diffrent network condition. The PCEP protocol need to extend a bit that provide different Schemes to choose.

5. Defragmentation

5.1. Motivation of Defragmentation

New arrival of requests are then either forced to utilize more spectrum in the network or blocked in spite of suffcient spectrum being available. Additionally, as the network evolves, a current optimal routing scheme might no longer provide the optimal spectral utilization over time. There is an increasing demand from the network operators to be able to periodically recon?gure the network and return it to its optimal state, so that the network can operate more effciently.

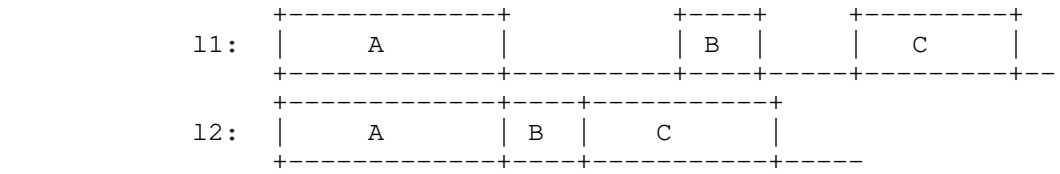
5.2. Definition of Defragmentation

There is an operation defined as network defragmentation to solve above problem. Reducing the blocking by consolidating the available network resources, this operation will also enable better network maintenance and more effcient network restoration and bandwidth adjustment.

5.3. Application Scene of Defragmentation

The process of defragmentation: (1) select which LSP to defragmentation, interrupt it, (2) choose forward spectrum in original route or new route, (3) move the LSP on possible spectrum.

An example of defragentation is as following: A,B,C are client LSPs on link l, l1 is Original statement of link l,l2 is statement of link l after defragementation.



we first focus on the problem of the time-point when should defragmentation be operated. So far, two new concepts proposed to solve this problem. One concept is Utilization Entropy that represents the level of resource fragmentation in an optical network proposed by Fujitsu Labs of America; the other concept is Spectrum Compactness that represents the spectrum distribution state in a link or in the network proposed by State key Laboratory of Information

Photonics and Optical Communications of Beijing University of Posts and Telecommunications. These two methods both related to threshold, it necessary to set threshold, when reaching threshold triggered defragmentation. PCEP protocol should include these information.

we consider the methods of defragmentation. At present, there are two methods for defragmentation. First is change route of client LSP means the spectrum of this LSP in new route is ahead than the spectrum in original route. Second is the LSP move forward directly in original route.

Defragmentation has to interrupt the traffic; the application scene is leisure network. When the network is busy, defragmentation lead to the increase of interrupt traffic demands.

Before defragmentation for the network, we have to do static programming for existing traffic demand in the network. We hope the defragmentation result reach or approach the static programming.

Maybe some network has requirement of interrupting rate or defragmentation time and so on, we should provide corresponding information to meet above requirements.

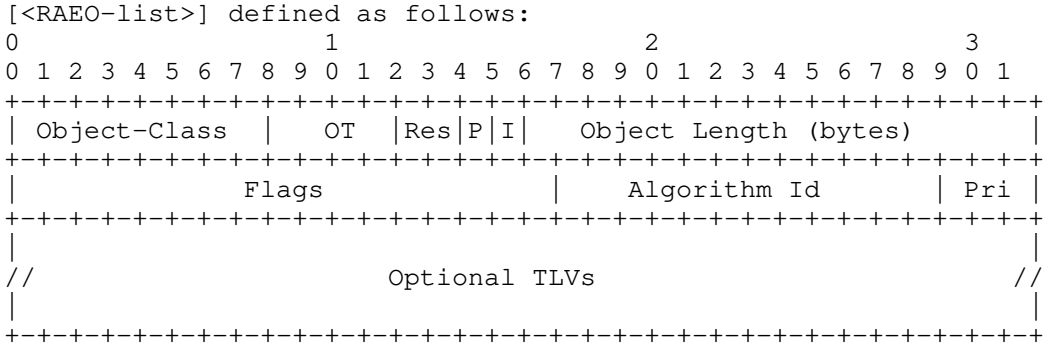
6. PCEP Protocol Extension

6.1. PCEP Protocol Extension for RSA

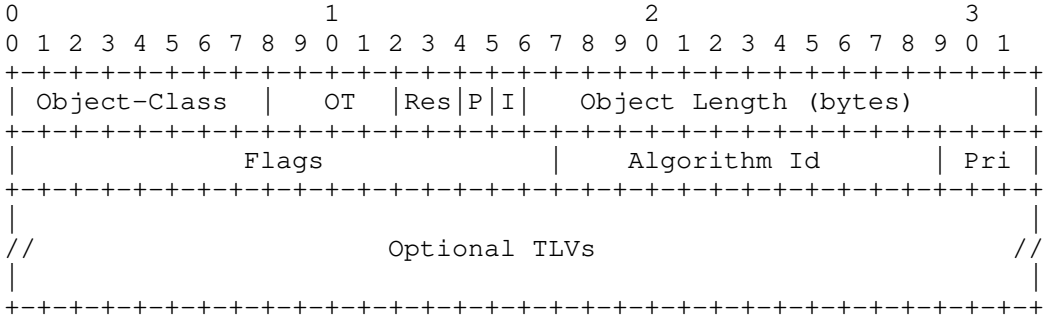
The PCEP protocol need to be extended to support the Algorithms choosing of RSA. PCReq need to adding RAEO-list information. This information include "Algorithm Id", which stand for the number of different algorithms, and "Pri" that means priority of these algorithms.

```
<request>::= <RP>
             <END-POINTS>
             [<RAEO-list>]
             [<LSPA>]
             [<BANDWIDTH>]
             [<metric-list>]
             [<RRO>[<BANDWIDTH>]]
             [<IRO>]
             [<LOAD-BALANCING>]
```

[<RAEO-list>] defined as follows:



[<RAEO-list>] defined as follows:



```
<response>::=<RP>
             [<NO-PATH>]
             [<attribute-list>]
             [<path-list>]
NO-PATH:
    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
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |Nature of Issue|C|          Flags                |   Reserved    |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                                                               |
   //                      Optional TLVs                          //
   |                                                               |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 11: NO-PATH Object Format

NI - Nature of Issue (8 bits): The NI field is used to report the nature of the issue that led to a negative reply.
Two values are currently defined:
0: No path satisfying the set of constraints could be found
1: PCE chain broken
2: No path satisfying the Continuous spectrum

6.2. PCEP Protocol Extension for Defragmentation

The presence of defragmentation in Flexi-Grid Networks has an impact on the information that needs to be transferred by the control plane and the PCE. Defragmentation has to interrupt the traffic and move it to another spectrum or route. The PCEP protocol needs to be extended two messages to support defragmentation, including information of original route/spectrum and present route/spectrum, when to stop defragmentation, the selection of methods and the limit of corresponding factors and so on.

Here is Spectrum Defragmentation Request Message and Spectrum Defragmentation Reply Message. "Target Clutter Value" stand for the threshold of defragmentation. "R" means whether the network MUST make it."Id 1" is number of defragmentation methods, "Id 2" is number of methods to trigger defragmentation, "L" means limit of interrupting rate or defragmentation time.

```
<SDReq Message>::= <Common Header>
                   <SDTO-list>
                   [LSPA Object]
                   [<RAEO-list>]
```

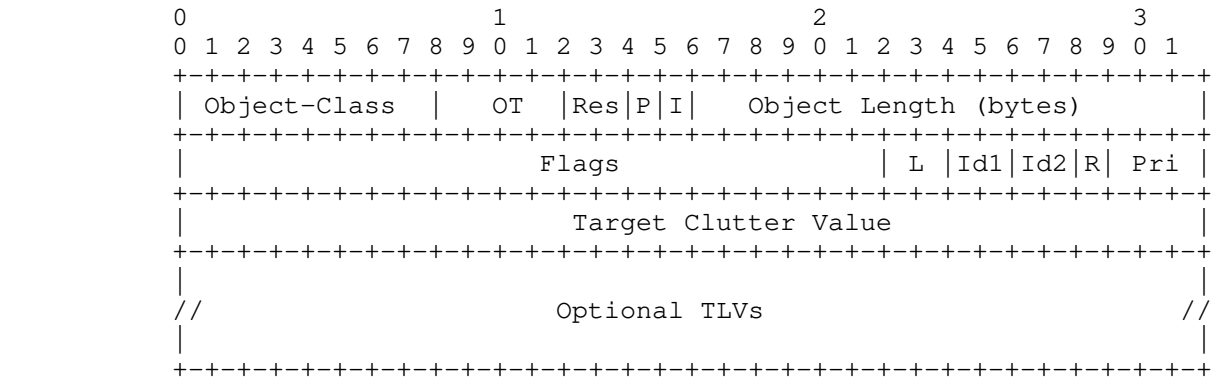
Spectrum Defragmentation Reply Message

```
<SDRep Message>::= <Common Header>
                   <SDTO-list>
                   [LSPA Object]
                   [<RAEO-list>]
```

Spectrum Defragmentation Reply Message

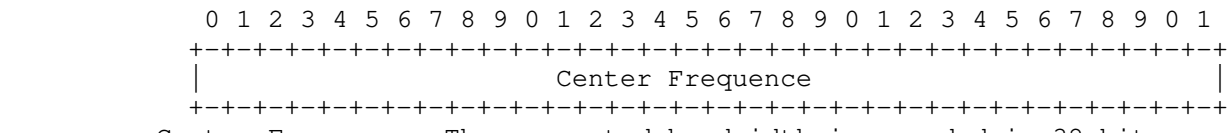
SDTO: Spectrum Defragmentation Target Object

<SDTO-list> defined as follows:




```
<Reroute>::=<RP Object>
            <path><Center Frequence><Bandwidth>
            <path><Center Frequence><Bandwidth>
```

where Center Frequence is



Center Frequence: The requested bandwidth is encoded in 32 bits, expressed in bytes per second.

7. Security Considerations

TBD.

8. Normative References

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

[RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.

Authors' Addresses

Yongli Zhao
BUPT
No.10,Xitucheng Road,Haidian District
Beijing 100876
P.R.China

Phone: +8613811761857
Email: yonglizhao@bupt.edu.cn
URI: http://www.bupt.edu.cn/

Jie Zhang
BUPT
No.10,Xitucheng Road,Haidian District
Beijing 100876
P.R.China

Phone: +8613911060930
Email: lgr24@bupt.edu.cn
URI: http://www.bupt.edu.cn/

Tiantian Peng
BUPT
No.10,Xitucheng Road,Haidian District
Beijing 100876
P.R.China

Phone: +8615116984347
Email: tt871228@163.com
URI: http://www.bupt.edu.cn/

Xiaosong Yu
BUPT
No.10,Xitucheng Road,Haidian District
Beijing 100876
P.R.China

Phone: +8613811731723
Email: yu.xiaosong@qq.com
URI: http://www.bupt.edu.cn/

Xuping Cao
ZTE Corporation
No.16,Huayuan Road,Haidian District
Beijing 100191
P.R.China

Phone: +8615801379189
Email: cao.xuping@zte.com.cn
URI: http://www.zte.com.cn/

Dajiang Wang
ZTE Corporation
No.16,Huayuan Road,Haidian District
Beijing 100191
P.R.China

Phone: +8613811795408
Email: wang.dajiang@zte.com.cn
URI: http://www.zte.com.cn/

Xihua Fu
ZTE Corporation
West District,ZTE Plaza,No.10,Tangyan South Road,Gaoxin District
Xi'an 710065
P.R.China

Phone: +8613798412242
Email: fu.xihua@zte.com.cn
URI: http://www.zte.com.cn/