PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 6, 2018

C. Barth
Juniper Networks
R. Gandhi
Cisco Systems, Inc.
B. Wen
Comcast

October 3, 2017

PCEP Extensions for Associated Bidirectional Label Switched Paths (LSPs) draft-barth-pce-association-bidir-03

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. The Stateful PCE extensions allow stateful control of Multiprotocol Label Switching (MPLS) Traffic Engineering (TE) Label Switched Paths (LSPs) using PCEP.

This document defines PCEP extensions for grouping two reverse unidirectional MPLS TE LSPs into an Associated Bidirectional LSP when using a Stateful PCE for both PCE-Initiated and PCC-Initiated LSPs as well as when using a Stateless PCE.

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."

Table of Contents

1. Introduction

[RFC5440] describes the Path Computation Element Protocol (PCEP) as a communication mechanism between a Path Computation Client (PCC) and a Path Control Element (PCE), or between PCE and PCC, that enables computation of Multiprotocol Label Switching (MPLS) Traffic Engineering (TE) Label Switched Paths (LSPs).

[RFC8231] specifies extensions to PCEP to enable stateful control of MPLS TE LSPs. It describes two modes of operation - Passive Stateful PCE and Active Stateful PCE. In [RFC8231], the focus is on Active Stateful PCE where LSPs are provisioned on the PCC and control over them is delegated to a PCE. Further, [I-D.ietf-pce-pce-initiated-lsp] describes the setup, maintenance and teardown of PCE-Initiated LSPs for the Stateful PCE model.

[I-D.ietf-pce-association] introduces a generic mechanism to create a grouping of LSPs which can then be used to define associations between a set of LSPs and/or a set of attributes, for example primary and secondary LSP associations, and is equally applicable to the active and passive modes of a Stateful PCE [RFC8231] or a stateless PCE [RFC5440].

The MPLS Transport Profile (MPLS-TP) requirements document [RFC5654] specifies that MPLS-TP MUST support associated bidirectional point-to-point LSPs. [RFC7551] specifies RSVP signaling extensions for binding two reverse unidirectional LSPs [RFC3209] into an associated bidirectional LSP. The fast reroute (FRR) procedures for associated bidirectional LSPs are described in [I-D.ietf-teas-assoc-corouted-bidir-frr].

This document specifies PCEP extensions for grouping two reverse unidirectional MPLS-TE LSPs into an Associated Bidirectional LSP for both single-sided and double-sided initiation cases when using a Stateful (both active and passive modes) or Stateless PCE. The PCEP extensions cover the following cases:

- A PCE initiates the forward and/ or reverse LSP of a single-sided or double-sided bidirectional LSP on a PCC and retains the control of the LSP. The PCE computes the path of the LSP and updates the PCC with the information about the path.

- A PCC initiates the forward and/ or reverse LSP of a single-sided or double-sided bidirectional LSP and retains the control of the LSP. The PCC computes the path of the LSP and reports the PCE

with the information about the path (as long as it controls the LSP, as in passive Stateful PCE mode).

- o A PCC initiates the forward and/ or reverse LSP of a single-sided or double-sided bidirectional LSP and delegates the control of the LSP to a Stateful PCE. The PCE may compute the path of the LSP and update the PCC with the information about the path (as long as it controls the LSP, as in active Stateful PCE mode).

- o A PCC requests co-routed or non co-routed paths for forward and reverse LSPs of a bidirectional LSP from a Stateless PCE.

2. Conventions Used in This Document

2.1. Key Word Definitions

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].

2.2. Terminology

The reader is assumed to be familiar with the terminology defined in [RFC5440], [RFC7551], [RFC8231], and [I-D.ietf-pce-association].

3. Overview

As shown in Figure 1, two reverse unidirectional LSPs can be grouped to form an associated bidirectional LSP. There are two methods of initiating the bidirectional LSP association, single-sided and double-sided as described in the following sections.

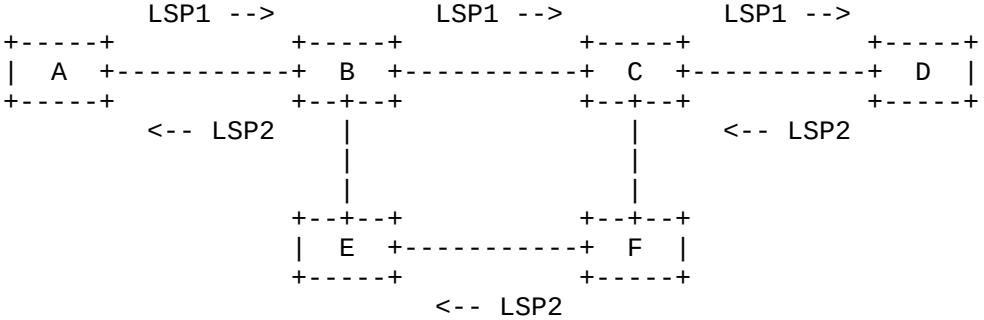


Figure 1: Example of Associated Bidirectional LSP

3.1. Single-sided Initiation

As specified in [RFC7551], in the single-sided case, the bidirectional tunnel is provisioned only on one endpoint node (PCC) of the tunnel. Both forward and reverse LSPs of this tunnel are initiated with the Association Type set to "Single-sided Bidirectional LSP Association" on the originating endpoint node. The forward and reverse LSPs are identified in the Bidirectional LSP Association Group TLV of their PCEP Association Objects.

The originating endpoint node signals the properties for the revere LSP in the RSVP REVERSE_LSP Object [RFC7551] of the forward LSP Path message. The remote endpoint then creates the corresponding reverse tunnel and signals the reverse LSP in response to the received RSVP Path message.

The two unidirectional reverse LSPs on the originating endpoint node are grouped together using the PCEP Association Object and on the remote endpoint node by the RSVP signaled Association Object.

As shown in Figure 1, the forward tunnel and both the forward LSP LSP1 and the reverse LSP LSP2 are initiated on the originating endpoint node A, either by the PCE or the PCC. The creation of reverse tunnel and reverse LSP2 on the remote endpoint node D is triggered by the RSVP signaled LSP1.

As specified in [I-D.ietf-teas-assoc-corouted-bidir-frr], for fast-reroute bypass tunnel assignment, the LSP starting from the originating node is identified as the forward LSP of the single-sided initiated bidirectional LSP.

3.2. Double-sided Initiation

As specified in [RFC7551], in the double-sided case, the bidirectional tunnel is provisioned on both endpoint nodes (PCCs) of the tunnel. The forward and reverse LSPs of this tunnel are initiated with the Association Type set to "Double-sided Bidirectional LSP Association" on both endpoint nodes. The forward and reverse LSPs are identified in the Bidirectional LSP Association Group TLV of their Association Objects.

The two reverse unidirectional LSPs on both the endpoint nodes are grouped together by using the PCEP Association Object.

As shown in Figure 1, the forward tunnel and LSP1 are initiated on the endpoint node A and the reverse tunnel and LSP2 are initiated on the endpoint node D, either by the PCE or the PCCs.

As specified in [I-D.ietf-teas-assoc-corouted-bidir-frr], for fast-reroute bypass tunnel assignment, the LSP with the higher Source Address [RFC3209] is identified as the forward LSP of the double-sided initiated bidirectional LSP.

3.3. Co-routed Associated Bidirectional LSP

In both single-sided and double-sided initiation cases, forward and reverse LSPs may be co-routed as shown in Figure 2, where both forward and reverse LSPs follow the same congruent path in the forward and reverse directions, respectively.

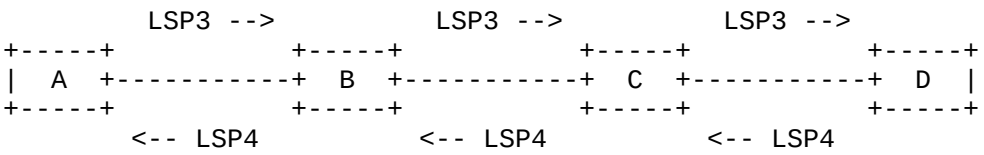


Figure 2: Example of Co-routed Associated Bidirectional LSP

4. Protocol Extensions

4.1. Association Object

As per [I-D.ietf-pce-association], LSPs are associated by adding them to a common association group. This document defines two new Bidirectional LSP Association Groups to be used by the associated bidirectional LSPs. A member of the Bidirectional LSP Association Group can take the role of a forward or reverse LSP and follows the following rules:

- An LSP can not be part of more than one Bidirectional LSP Association Group.
- The Tunnel (as defined in [RFC3209]) of forward and reverse LSPs of the single-sided bidirectional association MUST be the same.

This document defines two new Association Types for the Association Object as follows:

- Association Type (TBD1) = Single-sided Bidirectional LSP Association Group
- Association Type (TBD2) = Double-sided Bidirectional LSP Association Group

These Association Types are operator-configured associations in nature and statically created by the operator on the PCEP peers. The LSP belonging to these associations is conveyed via PCEP messages to the PCEP peer. Operator-configured Association Range TLV [I-D.ietf-pce-association] SHOULD NOT be sent for these Association Types, and MUST be ignored, so that the entire range of association ID can be used for them.

The Association ID, Association Source, optional Global Association Source and optional Extended Association ID in the Bidirectional LSP Association Group Object are also operator-configured and populated using the procedures defined in [RFC7551].

4.2. Bidirectional LSP Association Group TLV

The Bidirectional LSP Association Group TLV is an optional TLV for use with the Single-sided and Double-sided Bidirectional LSP Association Group Object Types.

- The Bidirectional LSP Association Group TLV follows the PCEP TLV format from [RFC5440].
- The Type (16 bits) of the TLV is TBD3, to be assigned by IANA.
- The Length is 4 Bytes.
- The value comprises of a single field, the Bidirectional LSP Association Flags (32 bits), where each bit represents a flag option.
- If the Bidirectional LSP Association Group TLV is missing, it means the LSP is the forward LSP.
- The Bidirectional LSP Association Group TLV MUST NOT be present more than once. If it appears more than once, only the first occurrence is processed and any others MUST be ignored.

The format of the Bidirectional LSP Association Group TLV is shown in Figure 3:

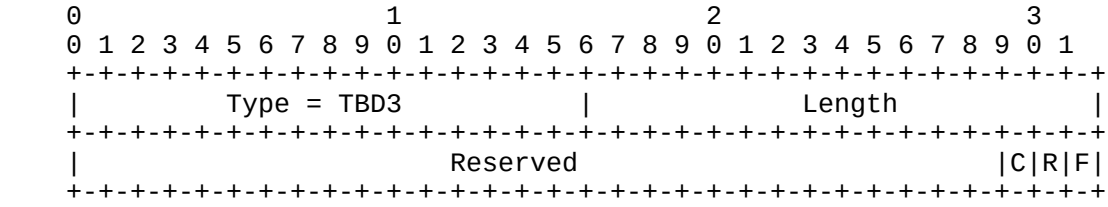


Figure 3: Bidirectional LSP Association Group TLV format

Bidirectional LSP Association Flags are defined as following.

F (Forward LSP, 1 bit) - Indicates whether the LSP associated is the forward LSP of the bidirectional LSP. If this flag is set, the LSP is a forward LSP.

R (Reverse LSP, 1 bit) - Indicates whether the LSP associated is the reverse LSP of the bidirectional LSP. If this flag is set, the LSP is a reverse LSP.

C (Co-routed LSP, 1 bit) - Indicates whether the bidirectional LSP is co-routed. This flag MUST be set for both the forward and reverse LSPs of a co-routed bidirectional LSP.

The C flag is used by the PCE (for both Stateful and Stateless) to compute bidirectional paths of the forward and reverse LSPs.

The Reserved flags MUST be set to 0 when sent and MUST be ignored when received.

5. PCEP Procedure

5.1. PCE Initiated LSPs

As specified in [I-D.ietf-pce-association], Association Groups can be created by both Stateful PCE and PCC.

A Stateful PCE can create and update the forward and reverse LSPs independently for both Single-sided and Double-sided bidirectional LSP association groups. The establishment and removal of the association relationship can be done on a per LSP basis. A PCE can create and update the association of the LSP on a PCC via PCInitiate and PCUpd messages, respectively, using the procedures described in [I-D.ietf-pce-association].

5.2. PCC Initiated LSPs

A PCC can associate or remove an LSP under its control from the bidirectional LSP association group. The PCC MUST report the change in LSP association to Stateful PCE via PCRpt message.

5.3. Stateless PCE

For a stateless PCE, it might be useful to associate a path computation request to an association group, thus enabling it to

associate a common set of configuration parameters or behaviors with the request. A PCC can request co-routed or non co-routed forward and reverse direction paths from a stateless PCE for the bidirectional LSP association group.

5.4. State Synchronization

During state synchronization, a PCC MUST report all the existing bidirectional LSP association groups to the Stateful PCE. After the state synchronization, the PCE MUST remove all stale bidirectional associations.

5.5. Error Handling

The LSPs (forward or reverse) in a single-sided bidirectional LSP association group MUST belong to the same TE Tunnel (as defined in [RFC3209]). If a PCE attempts to add an LSP in a single-sided bidirectional LSP association group for a different Tunnel, the PCC MUST send PCErr with Error-Type = TBD4 (Bidirectional LSP Association Error) and Error-Value = 1 (Tunnel mismatch). Similarly, if a PCC attempts to add an LSP to a single-sided bidirectional LSP association group at PCE not complying to this rule, the PCE MUST send this PCErr.

6. Security Considerations

The security considerations described in [RFC5440], [RFC8231], and [I-D.ietf-pce-pce-initiated-lsp] apply to the extensions defined in this document as well.

Two new Association Types for the Association Object, Double-sided Bidirectional LSP Association Group and Single-sided Associated Bidirectional LSP Group are introduced in this document. Additional security considerations related to LSP associations due to a malicious PCEP speaker is described in [I-D.ietf-pce-association] and apply to these Association Types. Thus, securing the PCEP session using Transport Layer Security (TLS) [I-D.ietf-pce-pceps] is recommended.

7. Manageability Considerations

7.1. Control of Function and Policy

The mechanisms defined in this document do not imply any control or policy requirements in addition to those already listed in [RFC5440], [RFC8231], and [I-D.ietf-pce-pce-initiated-lsp].

7.2. Information and Data Models

[RFC7420] describes the PCEP MIB, there are no new MIB Objects defined for LSP associations.

The PCEP YANG module [I-D.ietf-pce-pcep-yang] supports LSP associations.

7.3. Liveness Detection and Monitoring

The mechanisms defined in this document do not imply any new liveness detection and monitoring requirements in addition to those already listed in [RFC5440], [RFC8231], and [I-D.ietf-pce-pce-initiated-lsp].

7.4. Verify Correct Operations

The mechanisms defined in this document do not imply any new operation verification requirements in addition to those already listed in [RFC5440], [RFC8231], and [I-D.ietf-pce-pce-initiated-lsp].

7.5. Requirements On Other Protocols

The mechanisms defined in this document do not add any new requirements on other protocols.

7.6. Impact On Network Operations

The mechanisms defined in this document do not have any impact on network operations in addition to those already listed in [RFC5440], [RFC8231], and [I-D.ietf-pce-pce-initiated-lsp].

8. IANA Considerations

8.1. Association Types

This document adds new Association Types for the Association Object defined [I-D.ietf-pce-association]. IANA is requested to make the assignment of values for the sub-registry "ASSOCIATION Type Field" (to be created in [I-D.ietf-pce-association]), as follows:

Value	Name	Reference
TBD1	Single-sided Bidirectional LSP Association Group	[This document]
TBD2	Double-sided Bidirectional LSP Association Group	[This document]

8.2. Bidirectional LSP Association Group TLV

This document defines a new TLV for carrying additional information of LSPs within a Bidirectional LSP Association Group. IANA is requested to add the assignment of a new value in the existing "PCEP TLV Type Indicators" registry as follows:

TLV-Type	Name	Reference
TBD3	Bidirectional LSP Association Group TLV	[This document]

8.2.1. Flag Fields in Bidirectional LSP Association Group TLV

This document requests that a new sub-registry, named "Bidirectional LSP Association Group TLV Flag Field", is created within the "Path Computation Element Protocol (PCEP) Numbers" registry to manage the Flag field in the Bidirectional LSP Association Group TLV. New values are to be assigned by Standards Action [RFC8126]. Each bit should be tracked with the following qualities:

- Bit number (count from 0 as the most significant bit)
- Description
- Reference

The following values are defined in this document for the Flag field.

Bit No.	Description	Reference
31	F - Forward LSP	[This document]
30	R - Reverse LSP	[This document]
29	C - Co-routed LSP	[This document]

8.3. PCEP Errors

IANA is requested to allocate new Error-Type and Error-Value related to bidirectional LSP association within the " PCEP-ERROR Object Error Types and Values" sub-registry of the PCEP Numbers registry, as follows:

Error-Type	Description	Reference
TBD4	Bidirectional LSP Association Error	[This document]
	Error-value=1: Tunnel mismatch	[This document]

9. References

9.1. 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, March 2009.

[RFC7551] Zhang, F., Ed., Jing, R., and R. Gandhi, Ed., "RSVP-TE Extensions for Associated Bidirectional LSPs", RFC 7551, May 2015.

[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>.

[RFC8231] Crabbe, E., Minei, I., Medved, J., and R. Varga, "Pah Computation Element Communication Protocol (PCEP) Extensions for Stateful PCE", RFC 8231, DOI 10.17487/RFC8231, September 2017, <https://www.rfc-editor.org/info/rfc8231>.

[I-D.ietf-pce-association] Minei, I., Crabbe, E., Sivabalan, S., Ananthakrishnan, H., Dhody, D., and Y. Tanaka, "PCEP Extensions for Establishing Relationships Between Sets of LSPs", draft-ietf-pce-association-group (work in progress).

[I-D.ietf-pce-pce-initiated-lsp] Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model", draft-ietf-pce-pce-initiated-lsp (work in progress).

[I-D.ietf-teas-assoc-corouted-bidir-frr] Gandhi, R., Ed., Shah, H., and J. Whittaker, "Fast Reroute Procedures for Associated Bidirectional Label Switched Paths", draft-ietf-teas-assoc-corouted-bidir-frr (work-in-progress).

9.2. Informative References

[RFC5654] Niven-Jenkins, B., Ed., Brungard, D., Ed., Betts, M., Ed., Sprecher, N., and S. Ueno, "Requirements of an MPLS Transport Profile", RFC 5654, September 2009.

[RFC7420] Koushik, A., Stephan, E., Zhao, Q., King, D., and J. Hardwick, "Path Computation Element Communication Protocol (PCEP) Management Information Base (MIB) Module", RFC 7420, December 2014.

[I-D.ietf-pce-pceps] Lopez, D., Dios, O., Wu, Q., and D. Dhody, "Secure Transport for PCEP", draft-ietf-pce-pceps (work in progress).

[I-D.ietf-pce-pcep-yang] Dhody, D., Hardwick, J., Beeram, V., and J. Tantsura, "A YANG Data Model for Path Computation Element Communications Protocol (PCEP)", draft-ietf-pce-pcep-yang (work in progress).

Acknowledgments

TBA.

Authors' Addresses

Colby Barth
Juniper Networks

Email: cbarth@juniper.net

Rakesh Gandhi
Cisco Systems, Inc.

Email: rgandhi@cisco.com

Bin Wen
Comcast

Email: Bin_Wen@cable.comcast.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: March 15, 2018

H. Chen
Huawei Technologies
M. Toy
Verizon
X. Liu
Jabil
L. Liu
Fujitsu
Z. Li
China Mobile

September 11, 2017

PCEP Link State Abstraction
draft-chen-pce-h-connect-access-03

Abstract

This document presents extensions to the Path Computation Element Communication Protocol (PCEP) for a child PCE to abstract its domain information to its parent for supporting a hierarchical PCE system.

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 March 15, 2018.

Copyright Notice

Table of Contents

1. Introduction

A hierarchical PCE architecture is described in RFC 6805, in which a parent PCE maintains an abstract domain topology, which contains its child domains (seen as vertices in the topology) and the connections among them.

For a domain for which a child PCE is responsible, connections attached to the domain may comprise inter-domain links and Area Border Routers (ABRs). For a parent PCE to have the abstract domain topology, each of its child PCEs needs to advertise its connections to the parent PCE.

In addition to the connections attached to the domain, there may be some access points in the domain, which are the addresses in the domain to be accessible outside of the domain. For example, an address of a server in the domain that provides a number of services to users outside of the domain is an access point.

This document presents extensions to the Path Computation Element Communication Protocol (PCEP) for a child PCE to advertise the information about its connections and access points to its parent PCE and for the parent PCE to build and maintain the abstract domain topology based on the information. The extensions may reduce configurations, thus simplify operations on a PCE system.

A child PCE is simply called a child and a parent PCE is called a parent in the following sections.

2. Terminology

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

ASBR: Autonomous System (AS) Border Router. Router used to connect together ASes via inter-AS links.

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. Connections and Accesses

A connection is an inter-domain link between two domains in general. An ABR is also a connection, which connects two special domains called areas in a same Autonomous System (AS).

An access point in a domain is an address in the domain to be accessible to the outside of the domain. An access point is simply called an access.

4.1. Information on Inter-domain Link

An inter-domain link connects two domains in two different ASes. Since there is no IGP running over an inter-domain link, we may not obtain the information about the link generated by an IGP. We may suppose that IP addresses are configured on inter-domain links.

For a point-to-point (P2P) link connecting two ABSRs A and B in two different domains, from A’s point of view, the following information about the link may be obtained:

1) Link Type: P2P
2) Local IP address
3) Remote IP address
4) Traffic engineering metric
5) Maximum bandwidth
6) Maximum reservable bandwidth
7) Unreserved bandwidth
8) Administrative group
9) SRLG

We will have a link ID if it is configured; otherwise no link ID (i.e., the Router ID of the neighbor) may be obtained since no IGP adjacency over the link is formed.

For a broadcast link connecting multiple ASBRs in a number of domains, on each of the ASBRs X, the same information about the link as above may be obtained except for the followings:

a) Link Type: Multi-access,
b) Local IP address with mask length, and
c) No Remote IP address.

In other words, the information about the broadcast link obtained by ASBR X comprises a), b), 4) to 9), but does not include any remote IP address or link ID. We will have a link ID if it is configured;

otherwise no link ID (i.e., the interface address of the designated router for the link) may be obtained since no IGP selects it.

A parent constructs an abstract AS domain topology after receiving the information about each of the inter-domain links described above from its children.

RFC 5392 and RFC 5316 describe the distributions of inter-domain links in OSPF and IS-IS respectively. For each inter-domain link, its neighboring AS number and neighboring ASBR Identity (TE Router ID) need to be configured in IGP (OSPF or IS-IS).

In addition, an IGP adjacency between a network node running IGP and a PCE running IGP as a component needs to be configured and fully established if we want the PCE to obtain the inter-domain link information from IGP.

These configurations and IGP adjacency establishment are not needed if the extensions in this draft are used.

RFC 7752 (BGP-LS) describes the distributions of TE link state information including inter-domain link state. A BGP peer between a network node running BGP and a PCE running BGP as a component needs to be configured and the peer relation must be established before the PCE can obtain the inter-domain link information from BGP. However, some networks may not run BGP.

4.2. Information on ABR

For an AS running IGP and containing multiple areas, an ABR connects two or more areas. For each area connected to the ABR, the PCE as a child responsible for the area sends its parent the information about the ABR, which indicates the identifier (ID) of the ABR.

A parent has the information about each of its children, which includes the domain such as the area for which the child is responsible. The parent knows all the areas to which each ABR connects after receiving the information on the ABR from each of its children.

4.3. Information on Access Point

For an IP address in a domain to be accessible outside of the domain, the PCE as a child responsible for the domain sends its parent the information about the address.

The parent has all the access points (i.e., IP addresses) to be accessible outside of all its children’ domains after receiving the

information on the access points from each of its children.

5. Extensions to PCEP

This section focuses on procedures for abstracting domain information after briefing messages containing the abstract information.

5.1. Messages for Abstract Information

A child abstracts its domain to its parent through sending its parent a message containing the abstract information on the domain. After the relation between the child and the parent is determined, the parent has some information on the child, which includes the child's ID and domain. The message does not need to contain this information. It comprises the followings:

- For new or updated Connections and Accesses,
  - Indication of Update Connections and Accesses
  - Detail Information about Connections and Accesses
- For Connections and Accesses down,
  - Indication of Withdraw Connections and Accesses
  - ID Information about Connections and Accesses

For a P2P link from ASBR A to B and a broadcast link connecting to A, the detail information on the links includes A's ID, the information on the P2P link and the information on the broadcast link described in Section 4. The ID information on the links includes A's ID, 1) to 3) for the P2P link and a) to b) for the broadcast link described in Section 4. A link ID for a link is included if it is configured.

For an ABR X, the information on X includes X's ID and a flag indicating that X is ABR.

For an Access X (address), the detail information on X includes X and a cost associated with it. The ID information on X is X itself.

There are a few ways to encode the information above into a message. For example, one way is to extend an existing Notification message for including the information. Another way is to use a new message. These are put in Appendix A for your reference.

5.2. Procedures

5.2.1. Child Procedures

5.2.1.1. New or Changed Connections and Accesses

After a child determines its parent, it sends the parent a message containing the information about the connections (i.e., inter-domain links and ABRs) from its domain to its adjacent domains and the access points in its domain.

For any new or changed inter-domain links, ABRs and access points in the domain for which a child is responsible, the child sends its parent a message containing the information about these links, ABRs and access points with indication of Update Connections and Accesses.

For example, for a new inter-domain P2P link from ASBR A in a child's domain to ASBR B in another domain, the child sends its parent a message containing an indication of Update Connections and Accesses, A's ID, and the detail information on the link described in section 4.1.

For multiple new or changed inter-domain links from ASBR A, the child sends its parent a message having an indication of Update Connections and Accesses, and A's ID followed by the detail information about each of the links.

In another example, for a new or changed inter-domain broadcast link connected to ASBR X, an ABR Y and an access point 10.10.10.1/32 with cost 10 in a child's domain, the child sends its parent a message containing an indication of Update Connections and Accesses, and X's ID followed by the detail information about the link attached to X and the detail information about ABR Y, and the information on access 10.10.10.1/32 with cost 10.

For changes on the attributes (such as bandwidth) of an inter-domain link, a threshold may be used to control the frequency of updates that are sent from a child to its parent. At one extreme, the threshold is set to let a child send its parent a update message for any change on the attributes of an inter-domain link. At another extreme, the threshold is set to make a child not to send its parent any update message for any change on the attributes of an inter-domain link. Typically, the threshold is set to allow a child to send its parent a update message for a significant change on the attributes of an inter-domain link.

5.2.1.2. Connections and Accesses Down

For any inter-domain links, ABRs and access points down in the domain for which a child is responsible, the child sends its parent a message containing the information about these links, ABRs and access points with indication of Withdraw Connections and Accesses.

For example, for the inter-domain P2P link from ASBR A down, the child sends its parent a message containing an indication of Withdraw Connections and Accesses, and A's ID, which is followed by the ID information about the link.

For multiple inter-domain links from ASBR A down, the child sends its parent a message having an indication of Withdraw Connections and Accesses, and A's ID, which is followed by the ID information about each of the links.

5.2.1.3. Child and Parent in Same Organization

If a child and its parent are in a same organization, the child may send its parent the information inside its domain. For a parent, after all its children in its organization send their parent the information in their domains, connections and access points, it has in its TED the detail information inside each of its children's domains and the connections among these domains. The parent can compute a path crossing these domains directly and efficiently without sending any path computation request to its children.

5.2.1.4. Child as a Parent

There are a few ways in which a child as a parent abstracts its domain information to its parent.

One way is that the child sends its parent all its domain information if the child and the parent are in a same organization. The information includes the detail network topology inside each of the child's domains, the inter-domain links connecting the domains that the child's children are responsible and the inter-domain links connecting these domains to other adjacent domains.

In another way, the child abstracts each of the domains that its children are responsible as a cloud (or say abstract node) and these clouds are connected by the inter-domain links attached to the domains. The child sends its parent all the inter-domain links attached to any of the domains.

In a third way, the child abstracts all its domains including the domains for which its children are responsible as a cloud. This

abstraction is described below in details.

If a parent P1 is also a child of another parent P2, P1 as a child sends its parent P2 a message containing the information about the connections and access points. P1 as a parent has the connections among its children's domains. But these connections are hidden from its parent P2. P1 may have connections from its children's domains to other domains. P1 as a child sends its parent P2 these connections.

P1 as a parent has the access points in its children's domains to be accessible outside of the domains. P1 as child may not send all of these to its parent P2. It sends its parent some of these access points according to some local policies.

From P2's point of view, its child P1 is responsible for one domain, which has some connections to its adjacent domains and some access points to be accessible.

5.2.2. Parent Procedures

5.2.2.1. Process Connections and Accesses

A parent stores into its TED the connections and accesses for each of its children according to the messages containing connections and accesses received. For a message containing Update Connections and Accesses, it updates the connections and accesses in the TED accordingly. For a message containing Withdraw Connections and Accesses, it removes the connections and accesses from the TED.

After receiving the messages for connections and accesses from its children, the parent builds and maintains the TED for the topology of its children's domains, in which each of the domains is seen as a cloud or an abstract node. The information inside each of the domains is hidden from the parent. There are connections among the domains and the access points in the domains to be accessible in the topology.

For a new P2P link from node A to B with no link ID configured, when receiving a message containing the link from a child, the parent stores the link from A into its TED, where A is attached to the child's domain as a cloud. It finds the link's remote end B using the remote IP address of the link. After finding B, it associates the link attached to A with B and the link attached to B with A. This creates a bidirectional connection between A and B.

For a new P2P link from node A to B with link ID configured, when receiving a message containing the link, the parent stores the link

from A into its TED. It finds the link's remote end B using the link ID (i.e., B's ID).

For a new broadcast link connecting multiple nodes with no link ID configured, when the parent receives a message containing the link attached to node X, it stores the link from X into its TED. It finds the link's remote end P using the link's local IP address with network mask. P is a Pseudo node identified by the local IP address of the designated node selected from the nodes connected to the link. After finding P, it associates the link attached to X with P and the link connected to P with X. If P is not found, a new Pseudo node P is created. The parent associates the link attached to X with P and the link attached to P with X. This creates a bidirectional connection between X and P.

The first node and second node from which the parent receives a message containing the link is selected as the designed node and backup designed node respectively. After the designed node is down, the backup designed node becomes the designed node and the node other than the designed node with the largest local IP address connecting to the link is selected as the backup designed node.

When the old designed node is down and the backup designed node becomes the new designed node, the parent updates its TED through removing the link between each of nodes X and old P (the Pseudo node corresponding to the old designed node) and adding a link between each of nodes X (still connecting to the broadcast link) and new P (the Pseudo node corresponding to the new designed node).

5.2.2.2. Detail Topology in a Domain

If a parent is in a same organization as its child, it stores into its TED the detail information inside the child's domain when receiving a message containing the information from the child; otherwise, it discards the information and issues a warning indicating that the information is sent to a wrong place.

6. Security Considerations

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

7. IANA Considerations

This section specifies requests for IANA allocation.

8. Acknowledgement

The authors would like to thank Jescia Chen, Adrian Farrel, and Eric Wu for their valuable comments on this draft.

9. References

9.1. 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>.

[RFC6805] King, D., Ed. and A. Farrel, Ed., "The Application of the Path Computation Element Architecture to the Determination of a Sequence of Domains in MPLS and GMPLS", RFC 6805, DOI 10.17487/RFC6805, November 2012, <https://www.rfc-editor.org/info/rfc6805>.

[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>.

[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>.

9.2. Informative References

[RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, DOI 10.17487/RFC5305, October 2008, <https://www.rfc-editor.org/info/rfc5305>.

[RFC5392] Chen, M., Zhang, R., and X. Duan, "OSPF Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 5392, DOI 10.17487/RFC5392, January 2009, <https://www.rfc-editor.org/info/rfc5392>.

[RFC5316] Chen, M., Zhang, R., and X. Duan, "ISIS Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 5316, DOI 10.17487/RFC5316, December 2008, <https://www.rfc-editor.org/info/rfc5316>.

[RFC7752] Gredler, H., Ed., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and

Traffic Engineering (TE) Information Using BGP", RFC 7752, DOI 10.17487/RFC7752, March 2016, <https://www.rfc-editor.org/info/rfc7752>.

Appendix A. Message Encoding

A.1. Extension to Existing Message

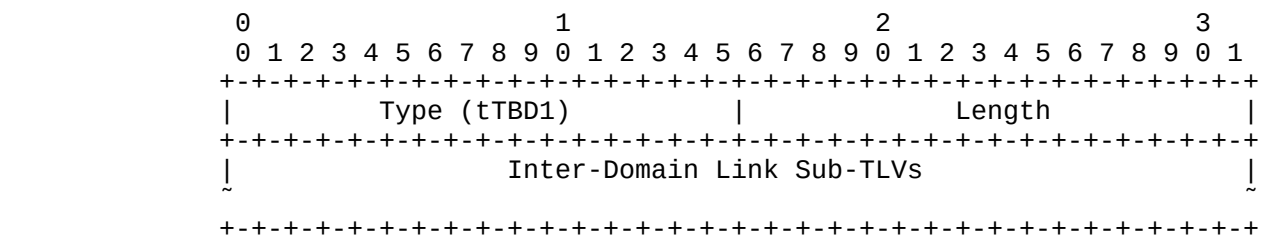
An existing Notification message may be extended to advertise the information about connections and access points. The following new Notification-type (NT) and Notification-value (NV) of a NOTIFICATION object in the message are defined:

- NT=8 (TBD): Connections and Accesses
  - NV=1: Update Connections and Accesses. A NT=8 and NV=1 indicates that the child sends its parent updates on the information about Connections and Accesses, and TLVs containing the information are in the object.
  - NV=2: Withdraw Connections and Accesses. A NT=8 and NV=2 indicates that the child asks its parent to remove Connections and Accesses indicated by TLVs in the object.

A.1.1. TLVs

Four TLVs are defined for connections and accesses. They are Inter-Domain link TLV, Router-ID TLV, Access IPv4/IPv6 Prefix TLV.

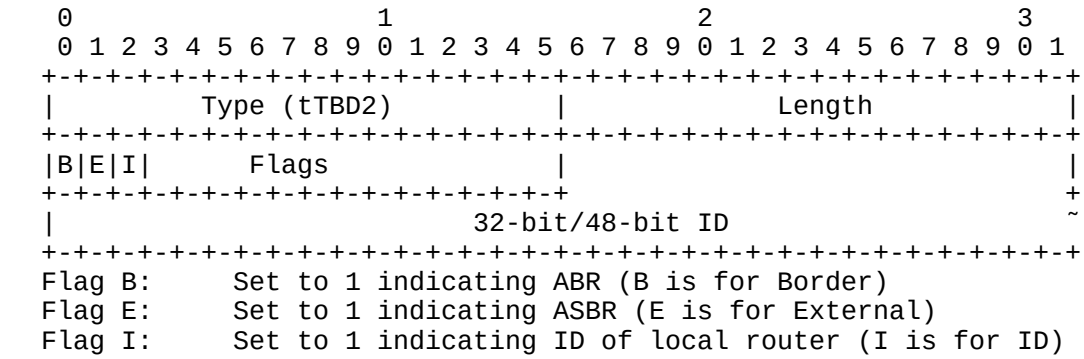
The format of the Inter-Domain link TLV is illustrated below.



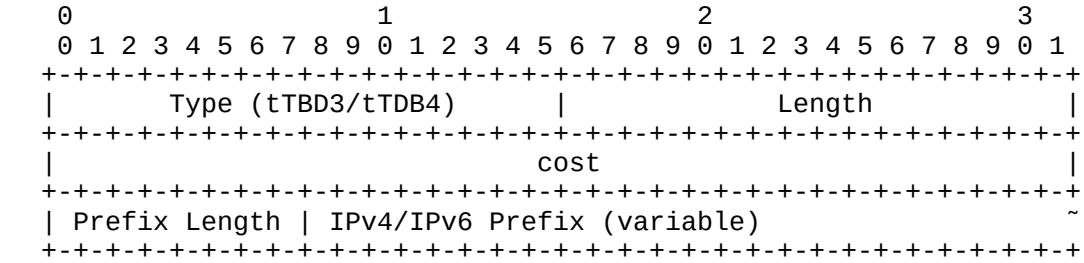
An Inter-Domain link TLV describes a single inter-domain link. It comprises a number of inter-domain link sub-TLVs for the information described in section 4, which are the sub-TLVs defined in RFC 3630 or their equivalents except for the local IP address with mask length defined below.

The format of the Router-ID TLV is shown below. Undefined flags MUST

be set to zero. The ID indicates the ID of a router. For a router running OSPF, the ID may be the 32-bit OSPF router ID of the router. For a router running IS-IS, the ID may be the 48-bit IS-IS router ID of the router. For a router not running OSPF or IS-IS, the ID may be the 32-bit ID of the router configured.

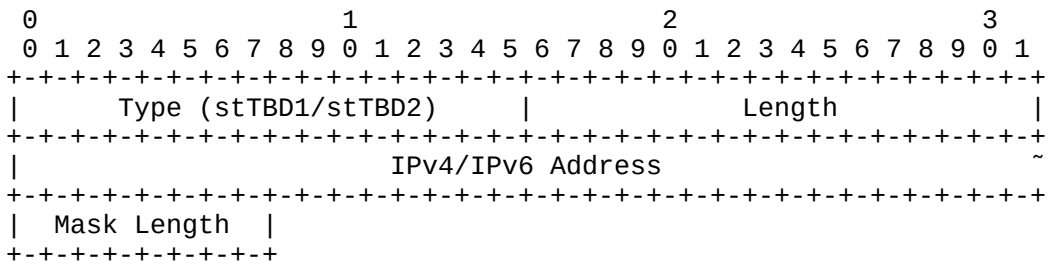


The format of the Access IPv4/IPv6 Prefix TLV is shown as follows. The cost is the metric to the prefix. The Prefix Length indicates the length of the prefix. The IPv4/IPv6 Prefix indicates an access IPv4/IPv6 address prefix.



A.1.2. Sub-TLVs

The format of the Sub-TLV for a local IPv4/IPv6 address with mask length is shown as follows.



The IPv4/IPv6 Address indicates the local IPv4/IPv6 address of a link. The Mask Length indicates the length of the IPv4/IPv6 address mask.

A.2. New Message

A new message may be defined to advertise the connections and accesses from a child to its parent. The format of the message containing Connections and Access (AC for short) is as follows:

```
<AC Message> ::= <Common Header> <NRP>
                 <Connection-List> [<Access-List>]
where:
  <Connection-List> ::= <Connection> [<Connection-List>]
  <Connection> ::= [<Inter-Domain-Link> | <ABR>]
  <Access-List> ::= <Access-Address> [<Access-List>]
```

Where the value of the Message-Type in the Common Header indicates the new message type. The exact value is to be assigned by IANA. A new RP (NRP) object will be defined, which follows the Common Header.

A new flag W (Withdraw) in the NRP object is defined to indicate whether the connections and access are withdrawn. When flag W is set to one, the parent removes the connections and accesses contained in the message after receiving it. When flag W is set to zero, the parent adds/updates the connections and accesses in the message after receiving it.

An alternative to flag W in the NRP object is a similar flag in each CONNECTION and ACCESS object such as using one bit in Res flags for flag W. For example, when the flag is set to one in the object, the parent removes the connections and accesses in the object after receiving it. When the flag is set to zero in the object, the parent adds/updates the connections and accesses in the object after receiving it.

In another option, one byte in a CONNECTION and ACCESS Object is defined as flags field and one bit is used as flag W. The other undefined bits in the flags field MUST be set to zero.

The objects in the new message are defined below.

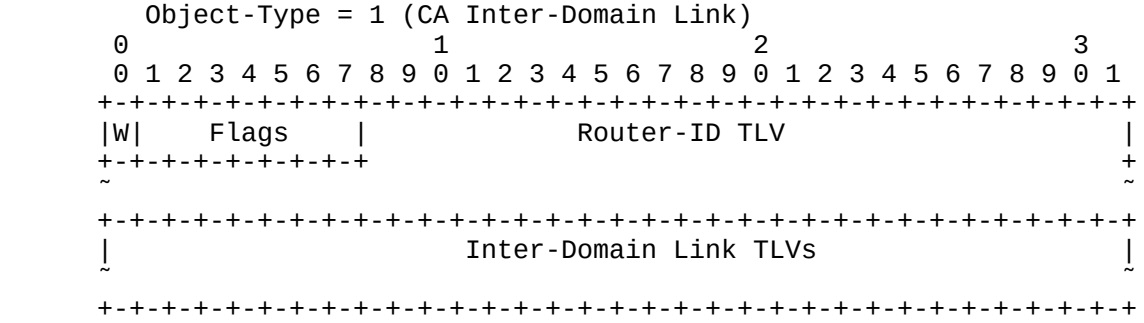
A.2.1. CONNECTION and ACCESS Object

A new object, called CONNECTION and ACCESS Object (CA for short), is defined. It has Object-Class ocTBD1. Four Object-Types are defined under CA object:

- CA Inter-Domain Link: CA Object-Type is 1.
- CA ABR: CA Object-Type is 2.
- CA Access IPv4 Prefix: CA Object-Type is 3.
- CA Access IPv6 Prefix: CA Object-Type is 4.

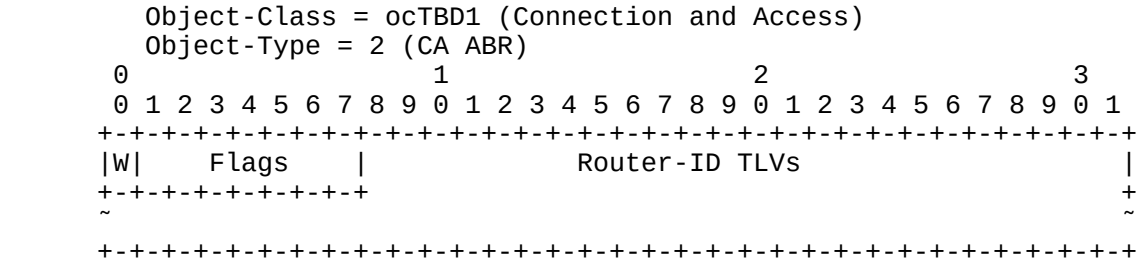
Each of these objects are described below.

The format of Inter-Domain Link object body is as follows:



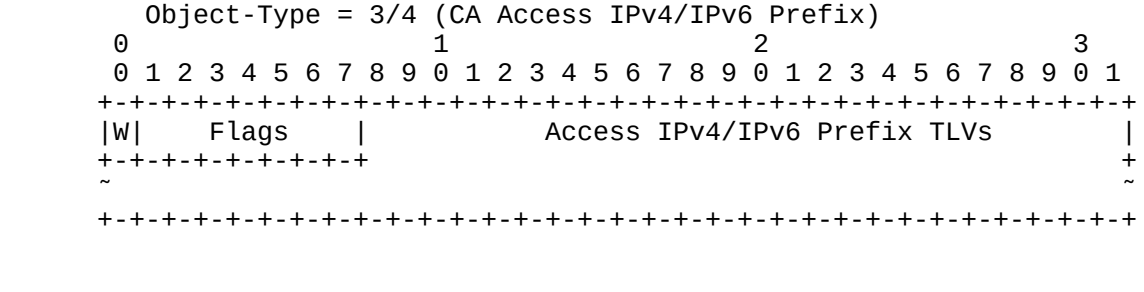
The Router-ID TLV indicates an ASBR in the domain, which is a local end of inter-domain links. Each of the Inter-Domain Link TLVs describes an inter-domain link and comprises a number of inter-domain link Sub-TLVs. Flag W=1 indicates withdraw the links. W=0 indicates new or changed links.

The format of ABR object body is illustrated below:



Each of the Router-ID TLVs indicates an ABR in the domain. Flag W=1 indicates withdraw the ABRs. W=0 indicates new ABRs.

The format of Access IPv4/IPv6 Prefix object body is as follows:



Each of the Access IPv4/IPv6 Prefix TLVs describes an access IPv4/ IPv6 address prefix in the domain, which is accessible to outside of the domain. Flag W=1 indicates withdraw the address prefixes. W=0 indicates new address prefixes.

The TLVs in the objects are the same as those described above.

Authors' Addresses

Huaimo Chen
Huawei Technologies
Boston, MA,
USA

EMail: Huaimo.chen@huawei.com

Mehmet Toy
Verizon
USA

EMail: mehmet.toy@verizon.com

Xufeng Liu
Jabil
McLean, VA
USA

EMail: Xufeng_Liu@jabil.com

Lei Liu
Fujitsu
USA

EMail: lliu@us.fujitsu.com

Zhenqiang Li
China Mobile
No.32 Xuanwumenxi Ave., Xicheng District
Beijing 100032
P.R. China

EMail: li_zhenqiang@hotmail.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: March 15, 2018

H. Chen
Huawei Technologies
M. Toy
Verizon
X. Liu
Jabil
L. Liu
Fujitsu
Z. Li
China Mobile

September 11, 2017

Hierarchical PCE Determination
draft-chen-pce-h-discovery-03

Abstract

This document presents extensions to the Path Computation Element Communication Protocol (PCEP) for determining parent child relations and exchanging the information between a parent and a child PCE in a hierarchical PCE system.

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 March 15, 2018.

Copyright Notice

Table of Contents

1. Introduction

A hierarchical PCE architecture is described in RFC 6805, in which a parent PCE has a number of child PCEs. A child PCE may also be a parent PCE, which has multiple child PCEs.

For a parent PCE, it needs to obtain the information about each of its child PCEs. The information about a child PCE comprises the address or ID of the PCE and the domain for which the PCE is responsible. It may also include the position of the PCE, which indicates whether the PCE is a leaf (i.e., only a child) or branch (i.e., a child and also a parent). In addition, the information may indicate whether the child PCE and its responsible domain is in a same organization as the parent PCE.

For a child PCE, it needs to obtain the information about its parent PCE, which includes the address or ID of the parent PCE. The information may also indicate whether the parent PCE is in a same organization as the child PCE.

After a user configures a parent PCE and a child PCE over a session, this parent child PCE relation needs to be determined in the protocol level. This is similar to OSPF and BGP. After an adjacency between two OSPF routers is configured by a user, the OSPF protocol (refer to RFC 2328, Section 7) will determine whether the adjacency is allowed based on the parameters configured, and forms the OSPF adjacency after the determination. After a peer relation between two BGP routers is configured by a user, the BGP protocol (refer to RFC 4271, Section 8) will determine whether the peer is allowed based on the parameters configured, and forms the BGP peer relation after the determination.

For a parent child PCE relation determination, the PCE protocol needs to check or confirm whether the parent child PCE relation is allowed based on the parameters configured. If so, the child PCE has to send its parent PCE the information about it and vice versa.

This document presents extensions to the Path Computation Element Communication Protocol (PCEP) for determining parent child relations and exchanging the information between a parent and a child PCE in a hierarchical PCE system.

2. Terminology

The following terminology is used in this document.

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.

TED: Traffic Engineering Database.

This document uses terminology defined in [RFC5440].

3. Conventions Used in This Document

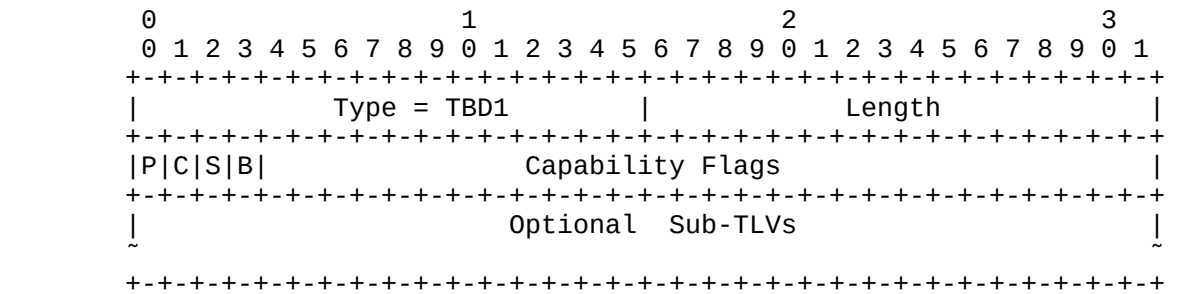
The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

4. Extensions to PCEP

This section describes the extensions to PCEP for determining the relation between a parent PCE and a child PCE and exchanging the information between a parent and a child PCE in a hierarchical PCE system. A child PCE is simply called a child and a parent PCE is called a parent in the following sections.

4.1. Determination of Parent Child Relation

During a PCEP session establishment between two PCEP speakers, each of them advertises its capabilities for Hierarchical PCE (H-PCE for short) through the Open Message with the Open Object containing a new TLV to indicate its capabilities for H-PCE. This new TLV is called H-PCE capability TLV. It has the following format.



The type of the TLV is TBD1. It has a length of 4 octets plus the size of optional Sub-TLVs. The value of the TLV comprises a capability flags field of 32 bits, which are numbered from the most significant as bit zero. Some of them are defined as follows. The others are not defined and MUST be set to zero.

- P (Parent - 1 bit): Bit 0 is used as P flag. It is set to 1 indicating a parent.
- C (Child - 1 bit): Bit 1 is used as C flag. It is set to 1 indicating a child.
- S (Same Org - 1 bit): Bit 2 is used as S flag. It is set to 1 indicating a PCE in a same organization as its remote peer.
- B (Both - 1 bit): Bit 3 is used as B flag. It is set to 1 indicating a PCE as both a child and a parent.

The following Sub-TLVs are defined:

- A Domain Sub-TLV containing an AS number and optional area, and
- PCE-ID Sub-TLV containing the ID of a PCE.

4.2. Sub-TLVs

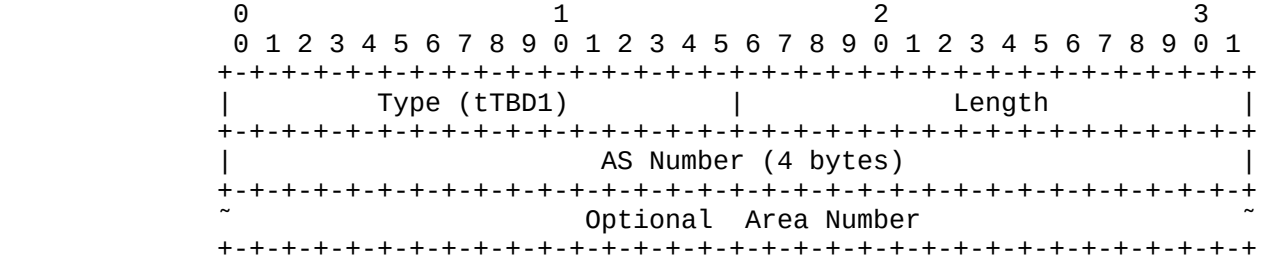
When a child sends its parent a Open message, it places the information about it in the message through using some optional Sub-TLVs. When a parent sends each of its child PCEs a Open message, it puts the information about it in the message.

4.2.1. Domain Sub-TLV

A domain is an AS or an area in an AS. An AS is identified by an AS number. An area in an AS is identified by the combination of the AS and the area. The former is indicated by an AS number and the latter

by an area number. A domain is represented by a domain Sub-TLV containing an AS number and a optional area number.

The format of the domain Sub-TLV is shown below:



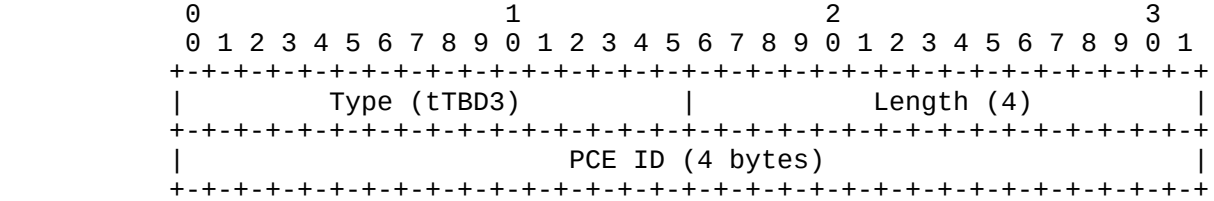
where Length is four plus size of area number.

An AS is represented by a domain Sub-TLV containing only the AS number of the AS. In this case, the Length is four. An area in an AS is represented by a domain Sub-TLV containing the AS number of the AS and the area number of the area. In this case, the Length is eight.

4.2.2. PCE ID Sub-TLV

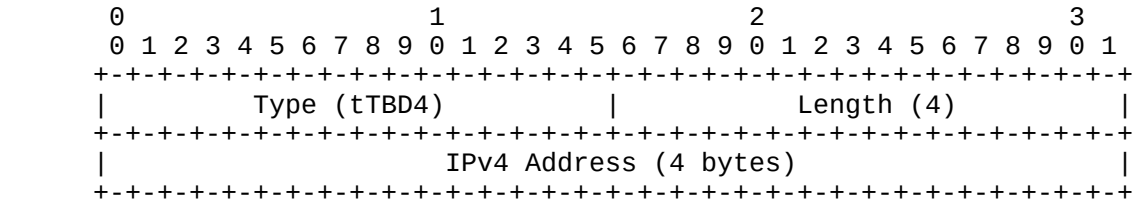
An Identifier (ID) of a PCE (PCE ID for short) is a 32-bit number that uniquely identifies the PCE among all PCEs. This 32-bit number for PCE ID SHOULD NOT be zero.

The format of the PCE ID Sub-TLV is shown below:



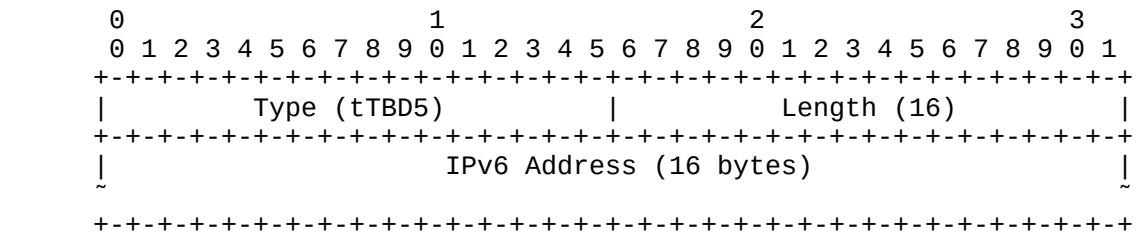
The PCE ID Sub-TLV specifies an non zero number as the identifier of the PCE.

Alternatively, an IP address attached to a PCE can also be used as an identifier of the PCE. The format of an IPv4 address Sub-TLV is shown below:



The IPv4 address Sub-TLV specifies an IPv4 address associated with the PCE, which is used as the identifier of the PCE.

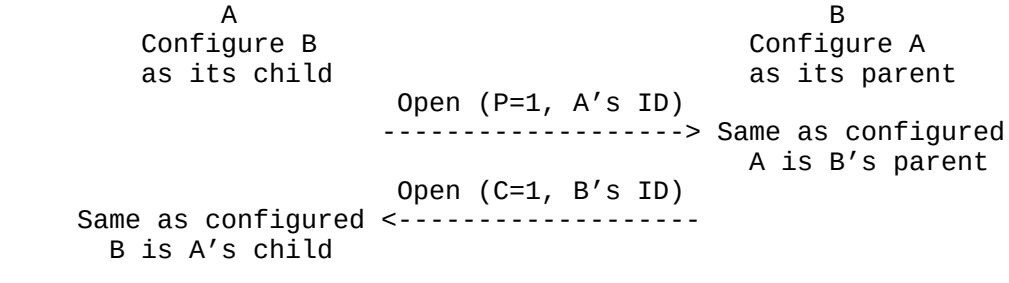
The format of an IPv6 address Sub-TLV is shown below:



The IPv6 Sub-TLV specifies an IPv6 address associated with the PCE, which is used as the identifier of the PCE.

4.3. Procedures

For two PCEs A and B configured as parent and child, they determine parent child relation through Open messages in the initialization phase. The following is a sequence of events related.



A sends B a Open message with P=1 and A's ID after B is configured as its child on it. B sends A a Open message with C=1 and B's ID after A is configured as its parent on it.

When A receives the Open message from B and determines C=1 and the PCE ID of B in the message is the same as the PCE ID of the child locally configured, B is A’s child.

When B receives the Open message from A and determines P=1 and the PCE ID of A in the message is the same as the PCE ID of the parent locally configured, A is B’s parent.

The Open message from child B to its parent A contains B’s domain, which is represented by a domain Sub-TLV in the H-PCE capability TLV. If child B is also a parent, the B flag in the TLV is set to 1.

The PCE ID in a Open message may be represented in one of the following ways:

- The source IP address of the message (i.e., PCE session).
- A PCE ID Sub-TLV in the H-PCE capability TLV.
- An IP address Sub-TLV in the H-PCE capability TLV.

When the IP address Sub-TLV is used, the address in the Sub-TLV MUST be the same as the source IP address of the PCE session.

For a child that is a leaf, it is normally responsible for one domain, which is contained in the message to its parent.

For a child that is a branch (i.e., also a parent of multiple child PCEs), it may be directly responsible for one domain, which is contained in the message to its parent. In addition, it is responsible for the domains of its child PCEs. In other words, it is responsible for computing paths crossing the domains through working together with its child PCEs. If these domains are all areas of an AS, the AS is included in the message to its parent.

A parent stores the information about each of its child PCEs received. When the session to one of them is down, it removes the information about the child on that session.

A child stores the information about its parent received. When the session to the parent is down, it removes the information about the parent.

If there already exists a session between A and B and the configurations on parent and child are issued on them, the procedures above may be executed through bringing down the existing session and establishing a new session between them. Alternatively, they may determine parent child relation through using extended Notification

messages in the same procedures as using Open messages described above without bringing down the existing session.

The following new Notification-type and Notification-value are defined for H-PCE:

- Notification-type=5 (TBD): Determination of H-PCE
  - Notification-value=1: The information about a parent PCE or a child PCE. A Notification-type=5, Notification-value=1 indicates that the PCE sends its peer the information about it and a TLV containing the information is in the Notification object. The format and contents of the TLV is the same as the H-PCE capability TLV described above. The only difference may be the type of the TLV.

5. Security Considerations

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

6. IANA Considerations

This section specifies requests for IANA allocation.

7. Acknowledgement

The authors would like to Jescia Chen, Adrian Farrel 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>.

[RFC6805] King, D., Ed. and A. Farrel, Ed., "The Application of the Path Computation Element Architecture to the Determination of a Sequence of Domains in MPLS and GMPLS", RFC 6805, DOI 10.17487/RFC6805, November 2012, <https://www.rfc-editor.org/info/rfc6805>.

[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>.

8.2. Informative References

[RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, DOI 10.17487/RFC2328, April 1998, <https://www.rfc-editor.org/info/rfc2328>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <https://www.rfc-editor.org/info/rfc4271>.

Authors' Addresses

Huaimo Chen
Huawei Technologies
Boston, MA,
USA

EMail: Huaimo.chen@huawei.com

Mehmet Toy
Verizon
USA

EMail: mehmet.toy@verizon.com

Xufeng Liu
Jabil
McLean, VA
USA

EMail: Xufeng_Liu@jabil.com

Lei Liu
Fujitsu
USA

EMail: lliu@us.fujitsu.com

Zhenqiang Li
China Mobile
No.32 Xuanwumenxi Ave., Xicheng District
Beijing 100032
P.R. China

EMail: li_zhenqiang@hotmail.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: March 15, 2018

H. Chen
Huawei Technologies
M. Toy
Verizon
X. Liu
Jabil
L. Liu
Fujitsu
Z. Li
China Mobile

September 11, 2017

Static PCEP Link State
draft-chen-pce-pcc-ted-03

Abstract

This document presents extensions to the Path Computation Element Communication Protocol (PCEP) for a PCC to advertise the information about the links without running IGP and for a PCE to build a TED based on the information received.

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 March 15, 2018.

Copyright Notice

Table of Contents

1. Introduction

A PCE architecture is described in RFC 4655, in which a Traffic Engineering Database (TED) for a PCE is constructed based on the link information from IGP (OSPF or IS-IS) running in the domain for which the PCE is responsible.

For a domain without running IGP, the PCE responsible for the domain may obtain the link information from a PCC running on each node in the domain.

This document presents extensions to the Path Computation Element Communication Protocol (PCEP) for a PCC to advertise the information about the links attached to the node running the PCC and for a PCE to build the TED based on the information received from the PCC.

2. Terminology

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

ASBR: Autonomous System (AS) Border Router. Router used to connect together ASes via inter-AS links.

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. Link Information

Since no IGP runs over any link, we may not obtain any link information via IGP. But links are configured.

For a point-to-point (P2P) link between nodes A and B, from A’s point of view, we have the following link information:

1) Link Type: P2P
2) Local IP address
3) Remote IP address
4) Traffic engineering metric
5) Maximum bandwidth
6) Maximum reservable bandwidth
7) Unreserved bandwidth
8) Administrative group

9) SRLG

A link ID for the link is obtained if a user configures it; otherwise, no link ID (i.e., the Router ID of A's neighbor) may be obtained since no IGP adjacency over the link is formed.

For a broadcast link connecting multiple nodes, on each of the nodes X, we have the same link information as above except for:

a) Link Type: Multi-access,
b) Local IP address with mask length, and
c) No Remote IP address.

In other words, the information about the broadcast link obtained by node X comprises a), b), 4) to 9), but does not include any remote IP address or link ID. A link ID for the link is obtained if a user configures it; otherwise, no link ID (i.e., the interface address of the designated router for the link) may be obtained since no IGP selects it.

A PCE constructs a TED for its responsible domain after receiving the link information from the PCC running on every node in the domain.

5. Extensions to PCEP

5.1. Extension to Existing Message

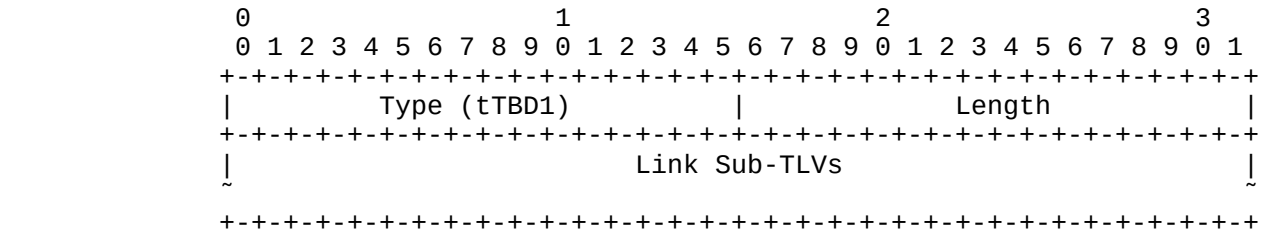
An existing Notification message may be extended to advertise the information about links. Alternatively, a new message can be used (refer to Appendix A).

The following new Notification-type (NT) and Notification-value (NV) of a NOTIFICATION object in a Notification message are defined:

- NT=8 (TBD): Links
  - NV=1: Update Links. NT=8 and NV=1 indicates that the PCC requests the PCE to update the link information based on the TLVs in the object, which are described below.
  - NV=2: Withdraw Links. NT=8 and NV=2 indicates that the PCC asks the PCE to remove the Links indicated by the TLVs in the object.

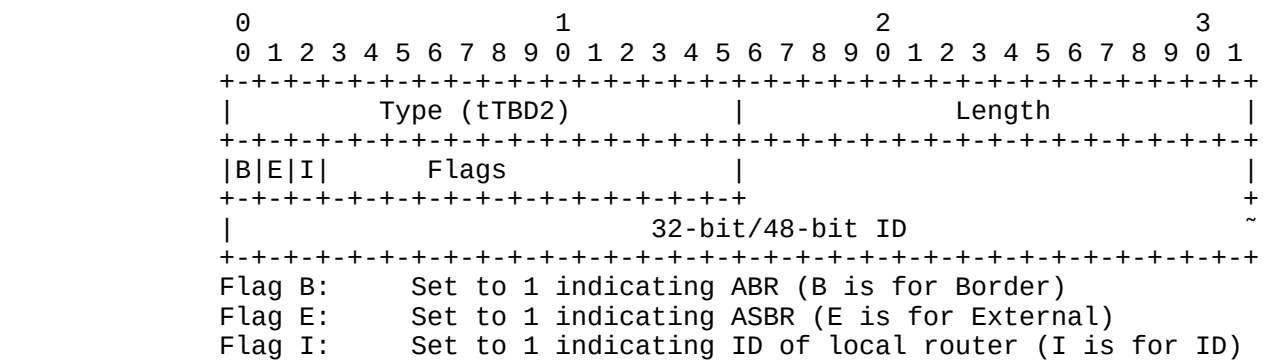
5.1.1. TLVs

A link TLV and a Router-ID TLV are defined. The format of the link TLV is illustrated below. The Type=tTBD1 indicates a link TLV Type. The Length indicates the size of the Link Sub-TLVs.



A link TLV describes a single link. It comprises a number of link sub-TLVs for the information described in section 4, which are the sub-TLVs defined in RFC 3630 or their equivalents except for the local IP address with mask length defined below.

The format of the Router-ID TLV is shown below. The Type=tTBD2 indicates a Router-ID TLV Type. The Length indicates the size of the ID and flags field.



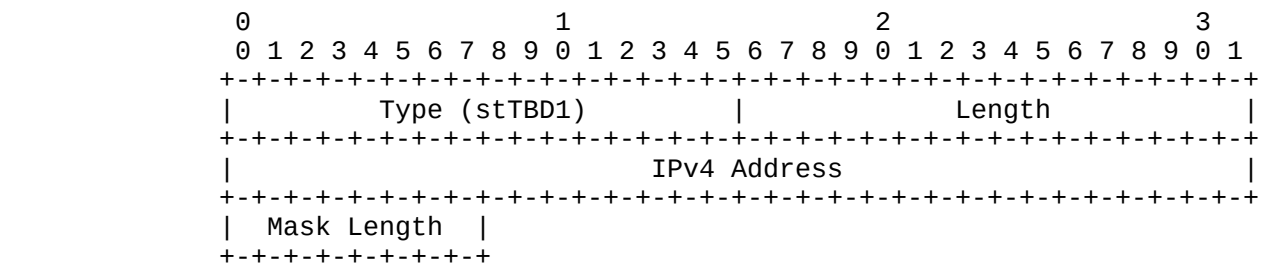
Flag B: Set to 1 indicating ABR (B is for Border)
Flag E: Set to 1 indicating ASBR (E is for External)
Flag I: Set to 1 indicating ID of local router (I is for ID)

Undefined flags MUST be set to zero. The ID indicates the ID of a router. For a router not running IGP, the ID may be the 32-bit or 48-bit ID of the router configured.

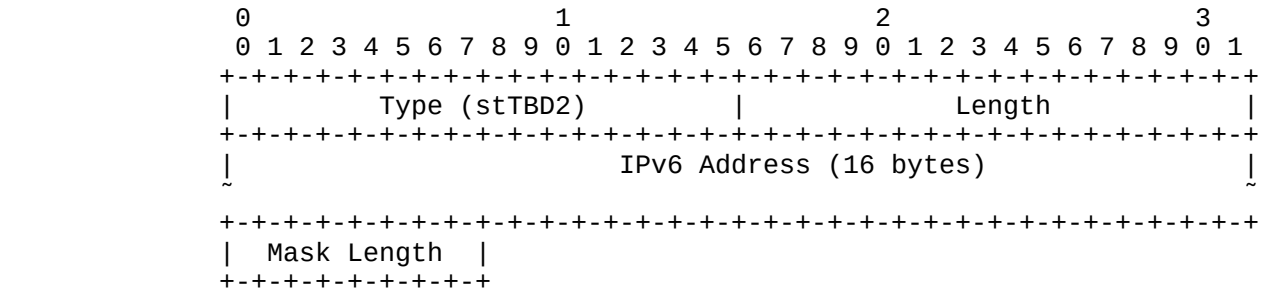
5.1.2. Sub-TLVs

The format of the Sub-TLV for a local IPv4 address with mask length is shown below. The Type=stTBD1 indicates a local IPv4 Address with mask length. The Length indicates the size of the IPv4 address and

Mask Length. The IPv4 Address indicates the local IPv4 address of a link. The Mask Length indicates the length of the IPv4 address mask.



The format of the Sub-TLV for a local IPv6 address with mask length is illustrated below. The Type=stTBD2 indicates a local IPv6 Address with mask length. The Length indicates the size of the IPv6 address and Mask Length. The IPv6 Address indicates the local IPv6 address of a link. The Mask Length indicates the length of the IPv6 address mask.



5.2. Procedures

5.2.1. PCC Procedures

1. New or Changed Links

After the session between a PCC and a PCE is established, the PCC sends the PCE a message containing the information about the links attached to the node running the PCC.

For any new or changed links, the PCC sends the PCE a message containing the information about these links with indication of Update Links.

For example, for a new P2P link from node A, the PCC running on A

sends the PCE a Notification message having a NOTIFICATION object with NT=8 and NV=1 (indicating Update Links), which contains a Router-ID TLV, followed by a link TLV. The former comprises A's ID and flag I set to 1. The latter comprises the Sub-TLVs for the information described in section 4.

For multiple new or changed links from node A, the PCC running on A sends the PCE a Notification message having a NOTIFICATION object with NT=8 and NV=1, which contains a Router-ID TLV for A's ID, followed by multiple link TLVs for the links.

2. Links Down

For links down, the PCC sends the PCE a message containing the information about these links with indication of Withdraw Links.

For example, for multiple links from node A down, the PCC running on A sends the PCE a Notification message having a NOTIFICATION object with NT=8 and NV=2 (indicating Withdraw Links), which contains a Router-ID TLV for A's ID, followed by multiple link TLVs for the links. The TLV for a P2P link comprises the Sub-TLVs for the information on 1), 2) and 3) described in section 4. The TLV for a broadcast link comprises the Sub-TLVs for the information on a) and b) described in section 4.

3. Simplified Message

Alternatively, the messages may be simplified. For each node, the source IP address of the PCC running on the node may be used as the ID of the node. The PCE knows the address after the session between the PCE and the PCC is up. Thus, a message containing the information about links does not need include any router-ID TLV.

For example, for a new P2P link attached to node A, the PCC running on A sends the PCE a Notification message having a NOTIFICATION object with NT=8 and NV=1 (indicating Update Links), which contains a link TLV comprising the Sub-TLVs for the information on 1) to 9) described in section 4. The object does not contain any Router-ID TLV for node A.

5.2.2. PCE Procedures

A PCE stores into its TED the links for each node according to the messages for the links received from the PCC running on the node. For a message containing Update Links, it updates the links accordingly. For a message containing Withdraw Links, it removes the links. When a node is down, the PCE removes the links attached to the node.

For a new P2P link between node A and B with no link ID configured, when receiving a message containing the link from the PCC running on A, the PCE stores the link for A (i.e., the link from A) into its TED. It will find the link's remote end B using the remote IP address of the link. After finding B, it associates the link for A with B and the link for B with A. This creates a bidirectional connection between A and B.

For a new broadcast link connecting multiple nodes with no link ID configured, when receiving a message containing the link from the PCC running on each of the nodes X, the PCE stores the link for X (i.e., the link from X) into its TED. It will find the link's remote end P using the link's local IP address with network mask. P is a Pseudo node identified by the local IP address of the designated node selected from the nodes connected to the link. After finding P, it associates the link for X with P and the link for P with X. This creates a bidirectional connection between X and P.

The first node and second node from which the PCE receives a message containing the link is selected as the designed node and backup designed node respectively. After the designed node is down, the backup designed node becomes the designed node and the node other than the designed node with the largest local IP address connecting to the link is selected as the backup designed node.

When the old designed node is down and the backup designed node becomes the new designed node, the PCE updates its TED through removing the link between each of nodes X and old P (the Pseudo node corresponding to the old designed node) and adding a link between each of nodes X (still connecting to the broadcast link) and new P (the Pseudo node corresponding to the new designed node).

6. Security Considerations

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

7. IANA Considerations

This section specifies requests for IANA allocation.

8. Acknowledgement

The authors would like to thank Jescia Chen, and Eric Wu for their valuable comments on this draft.

9. References

9.1. 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>.

[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>.

9.2. Informative References

[RFC7752] Gredler, H., Ed., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP", RFC 7752, DOI 10.17487/RFC7752, March 2016, <https://www.rfc-editor.org/info/rfc7752>.

Appendix A. New Message

A new message may be defined to advertise the information on links. The format of the message for the information on Links (IL for short) is as follows:

```
<IL Message> ::= <Common Header> <NRP> <Link-List>
where:
  <Link-List> ::= <LINK> [<Link-List>]
```

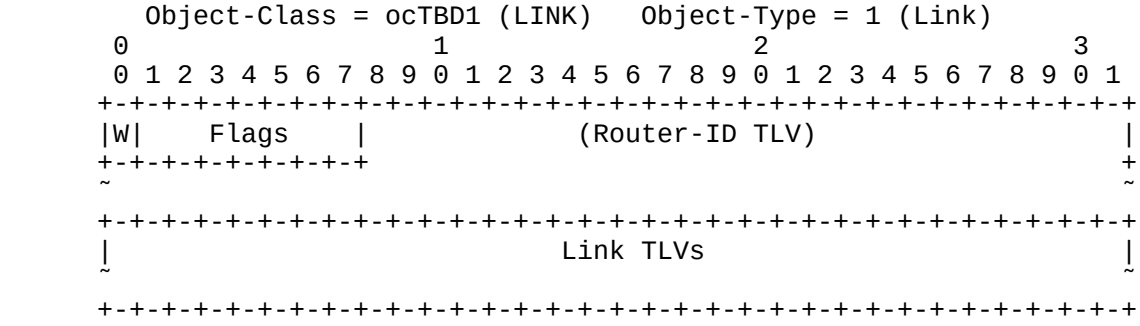
Where the value of the Message-Type in the Common Header indicates the new message type. The exact value is to be assigned by IANA. A new RP (NRP) object will be defined, which follows the Common Header.

A new flag W (Withdraw) in the NRP object is defined to indicate whether the links are withdrawn. When flag W is set to one, the PCE removes the links in the message after receiving it from the PCC.

When flag W is set to zero, the PCE adds/updates the links in the message.

An alternative to flag W in the NRP object is a similar flag W in each LINK object. For example, when the flag is set to one in the LINK object, the PCE removes the links in the object. When the flag is set to zero, the PCE adds/updates the links in the object.

The format of a LINK object body is as follows:



Flag W=1 indicates Withdraw links. W=0 indicates Updated links. Router-ID TLV is optional. Link TLVs are mandatory. They are the same as described in section 5.

Authors' Addresses

Huaimo Chen
Huawei Technologies
Boston, MA,
USA

EMail: Huaimo.chen@huawei.com

Mehmet Toy
Verizon
USA

EMail: mehmet.toy@verizon.com

Xufeng Liu
Jabil
McLean, VA
USA

EMail: Xufeng_Liu@jabil.com

Lei Liu
Fujitsu
USA

EMail: lliu@us.fujitsu.com

Zhenqiang Li
China Mobile
No.32 Xuanwumenxi Ave., Xicheng District
Beijing 100032
P.R. China

EMail: li_zhenqiang@hotmail.com

PCE Working Group
Internet-Draft
Intended status: Informational
Expires: September 14, 2017

D. Dhody
Y. Lee
Huawei Technologies
D. Ceccarelli
Ericsson

March 13, 2017

Applicability of Path Computation Element (PCE) for Abstraction and Control of TE Networks (ACTN)
draft-dhody-pce-applicability-actn-02

Abstract

Abstraction and Control of TE Networks (ACTN) refers to the set of virtual network operations needed to orchestrate, control and manage large-scale multi-domain TE networks so as to facilitate network programmability, automation, efficient resource sharing, and end-to-end virtual service aware connectivity and network function virtualization services.

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.

This document examines the applicability of PCE to the ACTN framework.

Table of Contents

1. Introduction

1.1. Path Computation Element (PCE)

The Path Computation Element communication Protocol (PCEP) [RFC5440] provides mechanisms for Path Computation Elements (PCEs) [RFC4655] to perform path computations in response to Path Computation Clients (PCCs) requests.

The ability to compute shortest constrained TE LSPs in Multiprotocol Label Switching (MPLS) and Generalized MPLS (GMPLS) networks across multiple domains has been identified as a key motivation for PCE development.

A stateful PCE is capable of considering, for the purposes of path computation, not only the network state in terms of links and nodes (referred to as the Traffic Engineering Database or TED) but also the status of active services (previously computed paths, and currently reserved resources, stored in the Label Switched Paths Database (LSPDB).

[RFC8051] describes general considerations for a stateful PCE deployment and examines its applicability and benefits, as well as its challenges and limitations through a number of use cases.

[I-D.ietf-pce-stateful-pce] describes a set of extensions to PCEP to provide stateful control. A stateful PCE has access to not only the information carried by the network's Interior Gateway Protocol (IGP), but also the set of active paths and their reserved resources for its computations. The additional state allows the PCE to compute constrained paths while considering individual LSPs and their interactions. [I-D.ietf-pce-pce-initiated-lsp] describes the setup, maintenance and teardown of PCE-initiated LSPs under the stateful PCE model.

[I-D.ietf-pce-stateful-pce] also describes the active stateful PCE. The active PCE functionality allows a PCE to reroute an existing LSP or make changes to the attributes of an existing LSP, or a PCC to delegate control of specific LSPs to a new PCE.

1.1.1. Role of PCE in SDN

Software-Defined Networking (SDN) refers to a separation between the control elements and the forwarding components so that software running in a centralized system called a controller, can act to program the devices in the network to behave in specific ways. A required element in an SDN architecture is a component that plans how the network resources will be used and how the devices will be programmed. It is possible to view this component as performing specific computations to place flows within the network given knowledge of the availability of network resources, how other forwarding devices are programmed, and the way that other flows are routed. It is concluded in [RFC7399], that this is the same function that a PCE might offer in a network operated using a dynamic control plane. This is the function and purpose of a PCE, and the way that a PCE integrates into a wider network control system including SDN is presented in Application-Based Network Operation (ABNO) [RFC7491].

1.1.2. PCE in multi-domain and multi-layer deployments

Computing paths across large multi-domain environments require special computational components and cooperation between entities in different domains capable of complex path computation. The PCE provides an architecture and a set of functional components to address this problem space. A PCE may be used to compute end-to-end paths across multi-domain environments using a per-domain path computation technique [RFC5152]. The Backward recursive PCE based path computation (BRPC) mechanism [RFC5441] defines a PCE-based path computation procedure to compute inter-domain constrained MPLS and GMPLS TE networks. 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.

[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) when the domain sequence is not known. Within the Hierarchical PCE (H-PCE) architecture, the Parent PCE (P-PCE) is used to compute a multi-domain path based on the domain connectivity information. A Child PCE (C-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.

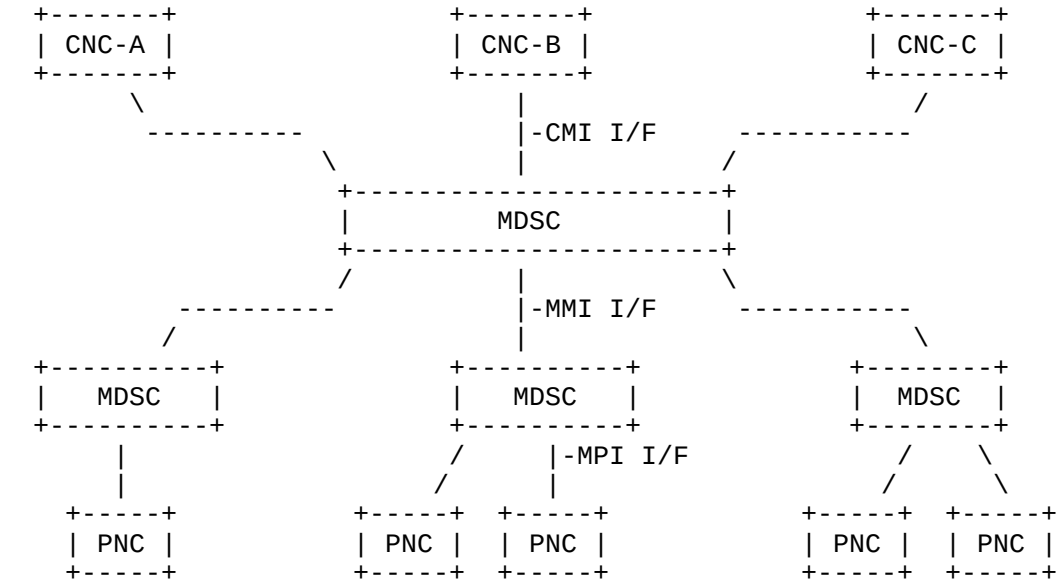
[I-D.dhodylee-pce-stateful-hpce] state the considerations for stateful PCE(s) in hierarchical PCE architecture. In particular, the behavior changes and additions to the existing stateful PCE mechanisms (including PCE- initiated LSP setup and active PCE usage) in the context of networks using the H-PCE architecture.

[RFC5623] describes a framework for applying the PCE-based architecture to inter-layer to (G)MPLS TE. It provides suggestions for the deployment of PCE in support of multi-layer networks. It also describes the relationship between PCE and a functional component in charge of the control and management of the VNT, called the Virtual Network Topology Manager (VNTM).

1.2. Abstraction and Control of TE Networks (ACTN)

[I-D.ietf-teas-actn-requirements] describes the high-level ACTN requirements. [I-D.ietf-teas-actn-framework] describes the architecture model for ACTN including the entities (Customer Network Controller(CNC), Mult-domain Service Coordinator(MDSC), and Physical Network Controller(PNC)) and their interfaces.

The ACTN reference architecture identified a three-tier control hierarchy as depicted in Figure 1:



CMI - (CNC-MDSC Interface)
MMI - (MDSC-MDSC Interface)
MPI - (MDSC-PNC Interface)

Figure 1: ACTN Hierarchy

The two interfaces with respect to the MDSC, one north of the MDSC Interface) and MPI (MDSC-PNC Interface), respectively. MMI (MDSC-MDSC interface) is used to support recursion.

[I-D.ietf-teas-actn-info-model] provides an information model for ACTN interfaces.

1.3. PCE and ACTN

This document examines the PCE and ACTN architecture and describes how the PCE architecture is applicable to ACTN. It also lists the PCEP extensions that are needed to use PCEP as an ACTN interface. This document also identifies any gaps in PCEP, that exist at the time of publication of this document.

2. Architectural Considerations

ACTN [I-D.ietf-teas-actn-framework] architecture is based on hierarchy and recursiveness of controllers. It defines three types of controllers (depending on the functionalities they implement). The main functionalities are -

- Multi domain coordination function
- Virtualization/Abstraction function
- Customer mapping/translation function
- Virtual service coordination function

Section 3 of [I-D.ietf-teas-actn-framework] describes these functions.

It should be noted that, in this document we list all possible ways in which PCEP could be used for each of the above functions, but all functions are not required to be implemented via PCEP. Operator may choose to use the PCEP for multi domain coordination via stateful H-PCE but use RestConf or BGP-LS to get the topology and support virtualization/abstraction function.

2.1. Multi domain coordination via Hierarchy

With the definition of domain being "everything that is under the control of the single logical controller", as per [I-D.ietf-teas-actn-framework], it is needed to have a control entity that oversees the specific aspects of the different domains and to build a single abstracted end-to-end network topology in order to coordinate end-to-end path computation and path/service provisioning.

The MDSC in ACTN framework realizes this function by coordinating the per-domain PNCs in a hierarchy of controllers. It also needs to detach from the underlying network technology and express customer concerns by business needs.

[RFC6805] and [I-D.dhodylee-pce-stateful-hpce] describes a hierarchy of PCE with Parent PCE coordinating multi-domain path computation function between Child PCE(s). It is easy to see how these principles align, and thus how stateful H-PCE architecture can be used to realize ACTN.

The Per domain stitched LSP in the Hierarchical stateful PCE architecture, described in Section 3.3.1 of [I-D.dhodylee-pce-stateful-hpce] is well suited for multi-domain

coordination function. This includes domain sequence selection; E2E path computation; Controller (PCE) initiated path setup and reporting. This is also applicable to multi-layer coordination in case of IP+optical networks.

[I-D.litkowski-pce-state-sync]" describes the procedures to allow a stateful communication between PCEs for various use-cases. The procedures and extensions are also applicable to Child and Parent PCE communication and thus useful for ACTN as well.

2.2. Virtualization/Abstraction function

To realize ACTN, an abstracted view of the underlying network resources needs to be built. This includes global network-wide abstracted topology based on the underlying network resources of each domain. This also include abstract topology created as per the customer service connectivity requests and represented as a network slice allocated to each customer.

In order to compute and provide optimal paths, PCEs require an accurate and timely Traffic Engineering Database (TED). Traditionally this TED has been obtained from a link state (LS) routing protocol supporting traffic engineering extensions. PCE may construct its TED by participating in the IGP ([RFC3630] and [RFC5305] for MPLS-TE; [RFC4203] and [RFC5307] for GMPLS). An alternative is offered by BGP-LS [RFC7752].

In case of H-PCE [RFC6805], the parent PCE needs to build the domain topology map of the child domains and their interconnectivity. [RFC6805] and [I-D.ietf-pce-inter-area-as-applicability] suggest that BGP-LS could be used as a "northbound" TE advertisement from the child PCE to the parent PCE.

[I-D.dhodylee-pce-pcep-ls] proposes another approaches for learning and maintaining the Link-State and TE information as an alternative to IGPs and BGP flooding, using PCEP itself. The child PCE can use this mechanism to transport Link-State and TE information from child PCE to a Parent PCE using PCEP.

In ACTN, there is a need to control the level of abstraction based on the deployment scenario and business relationship between the controllers. The mechanism used to disseminate information from PNC (child PCE) to MDSC (parent PCE) should support abstraction. [I-D.lee-teas-actn-abstraction] describes a few alternative approaches of abstraction. The resulting abstracted topology can be encoded using the PCEP-LS mechanisms [I-D.dhodylee-pce-pcep-ls]. PCEP-LS is an attractive option when the operator would wish to have a single control plane protocol (PCEP) to achieve ACTN functions.

2.3. Customer mapping function

In ACTN, there is a need to map customer virtual network (VN) requirements into network provisioning request to the PNC. That is, the customer requests/commands are mapped into network provisioning requests that can be sent to the PNC. Specifically, it provides mapping and translation of a customer's service request into a set of parameters that are specific to a network type and technology such that network configuration process is made possible.

[I-D.ietf-pce-pce-initiated-lsp] describes the setup, maintenance and teardown of PCE-initiated LSPs under the stateful PCE model, without the need for local configuration on the PCC, thus allowing for a dynamic network that is centrally controlled and deployed. To instantiate or delete an LSP, the PCE sends the Path Computation LSP Initiate Request (PCInitiate) message to the PCC. As described in [I-D.dhodylee-pce-stateful-hpce], for inter-domain LSP in Hierarchical PCE architecture, the initiation operations can be carried out at the parent PCE. In which case after parent PCE finishes the E2E path computation, it can send the PCInitiate message to the child PCE, the child PCE further propagates the initiate request to the LSR. The customer request is received by the MDSC (parent PCE) and based on the business logic, global abstracted topology, network conditions and local policy, the MDSC (parent PCE) translates this into per domain LSP initiation request that a PNC (child PCE) can understand and act on. This can be done via the PCInitiate message.

PCEP extensions for associating opaque policy between PCEP peer [I-D.ietf-pce-association-policy] can be used.

2.4. Virtual Network Operations

Virtual service coordination function in ACTN incorporates customer service-related information into the virtual network service operations in order to seamlessly operate virtual networks while meeting customer's service requirements.

[I-D.leedhody-pce-vn-association] describes the need for associating a set of LSPs with a VN "construct" to facilitate VN operations in PCE architecture. This association allows the PCEs to identify which LSPs belong to a certain VN.

This association based on VN is useful for various optimizations at the VN level which can be applied to all the LSPs that are part of the VN slice. During path computation, the impact of a path for an LSP is compared against the paths of other LSPs in the VN. This is to make sure that the overall optimization and SLA of the VN rather

than of a single LSP. Similarly, during re-optimization, advanced path computation algorithm and optimization technique can be considered for all the LSPs belonging to a VN/customer and optimize them all together.

3. Interface Considerations

As per [I-D.ietf-teas-actn-framework], to allow virtualization and multi domain coordination, the network has to provide open, programmable interfaces, in which customer applications can create, replace and modify virtual network resources and services in an interactive, flexible and dynamic fashion while having no impact on other customers. The 3 ACTN interfaces are -

- The CNC-MDSC Interface (CMI) is an interface between a Customer Network Controller and a Multi Domain Service Coordinator. It requests the creation of the network resources, topology or services for the applications. The MDSC may also report potential network topology availability if queried for current capability from the Customer Network Controller.

- The MDSC-PNC Interface (MPI) is an interface between a Multi Domain Service Coordinator and a Physical Network Controller. It communicates the creation request, if required, of new connectivity of bandwidth changes in the physical network, via the PNC. In multi-domain environments, the MDSC needs to establish multiple MPIs, one for each PNC, as there are multiple PNCs responsible for its domain control.

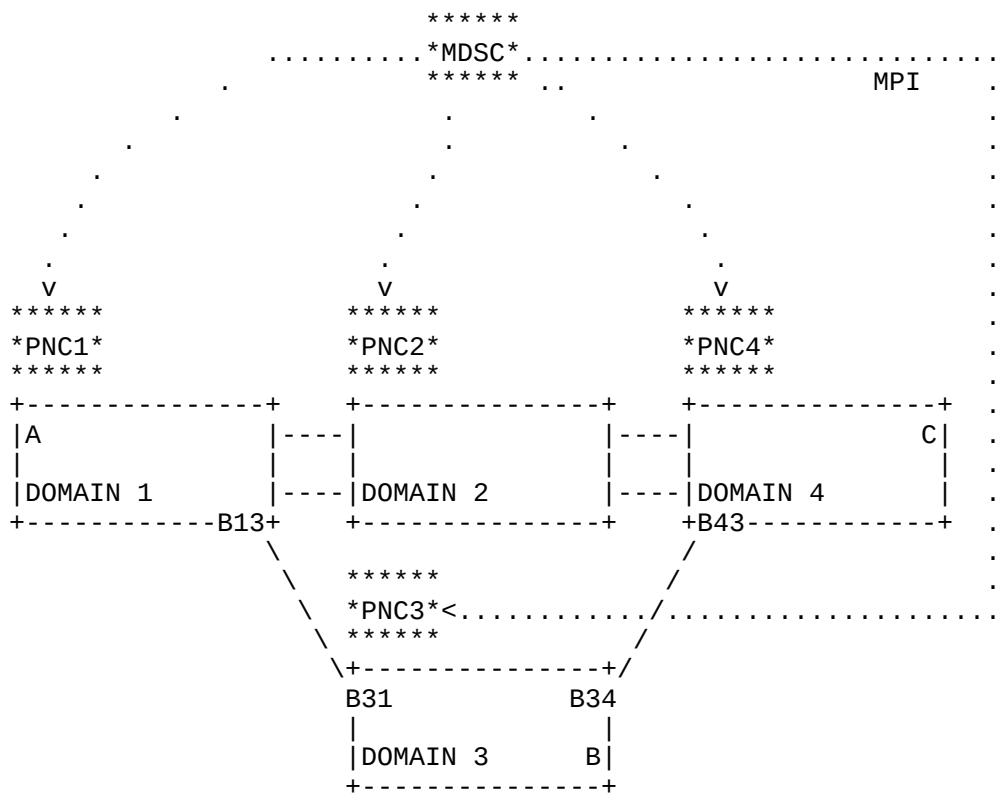
- The MDSC-MDSC Interface (MMI) is a special case of the MPI and behaves similarly to an MPI to support general functions performed by the MDSCs such as abstraction function and provisioning function. From an abstraction point of view, the top level MDSC which interfaces the CNC operates on a higher level of abstraction (i.e., less granular level) than the lower level MSDCs. As such, the MMI carries more abstract TE information than the MPI.

PCEP is especially suitable on the MPI and MMI as it meets the requirement and the functions as set out in the ACTN framework [I-D.ietf-teas-actn-framework]. Its recursive nature is well suited via the multi-level hierarchy of PCE. The Section 4 describe how PCE and PCEP could help realize ACTN.

4. Realizining ACTN with PCE (and PCEP)

As per the example in the Figure 2, there are 4 domains, each with its own PNC and a MDSC at top. The PNC and MDSC need PCE as a important function. The PNC (or child PCE) already uses PCEP to

communicate to the network device. It can utilize the PCEP as the MPI to communicate between controllers too.



MDSC -> Parent PCE
PNC -> Child PCE
MPI -> PCEP

Figure 2: ACTN with PCE

o Building Domain Topology at MDSC: PNC (or child PCE) needs to have the TED to compute path in its domain. As described in Section 2.2, it can learn the topology via IGP or BGP-LS. PCEP-LS is also a proposed mechanism to carry link state and traffic engineering information within PCEP. A mechanism to carry abstracted topology while hiding technology specific information between PNC and MDSC is described in [I-D.dhodylee-pce-pcep-ls]. At the end of this step the MDSC (or parent PCE) has the abstracted topology from each of its PNC (or child PCE). This could be as simple as a domain topology map as described in

[RFC6805] or it can have full topology information of all domains. The latter is not scalable and thus an abstracted topology of each domain interconnected by inter-domain links is the most common case.

* Topology Change: When the PNC learns of any topology change, the PNC needs to decide if the change needs to be notified to the MDSC. This is dependent on the level of abstraction between the MDSC and the PNC.

o VN Instantiate: MDSC is requested to instantiate a VN, the minimal information that is required would be a VN identifier and a set of end points. Various path computation, setup constraints and objective functions may also be provided. In PCE terms, a VN Instantiate can be considered as a set of paths belonging to the same VN. As described in Section 2.4 and [I-D.leedhody-pce-vn-association] the VN association can help in identifying the set of paths that belong to a VN. The rest of the information like the endpoints, constraints and objective function is already defined in PCEP in terms of a single path.

* Path Computation: As per the example in the Figure 2, the VN instantiate requires two end to end paths between (A in Domain 1 to B in Domain 3) and (A in Domain 1 to C in Domain 4). The MDSC (or parent PCE) triggers the end to end path computation for these two paths. MDSC can do path computation based on the abstracted domain topology that it already has or it may use the H-PCE procedures (Section 2.1) using the PCReq and PCRep messages to get the end to end path with the help of PNC. Either way, the resulted E2E paths may be broken into per-domain paths.

* A-B: (A-B13,B13-B31,B31-B)

* A-C: (A-B13,B13-B31,B34-B43,B43-C)

* Per Domain Path Instantiation: Based on the above path computation, MDSC can issue the path instantiation request to each PNC via PCInitiate message (see [I-D.dhodylee-pce-stateful-hpce] and [I-D.leedhody-pce-vn-association]). A suitable stitching mechanism would be use to stitch these per domain LSPs.

* Per Domain Path Report: Each PNC should report the status of the per-domain LSP to the MDSC via PCRpt message, as per the Hierarchy of stateful PCE ([I-D.dhodylee-pce-stateful-hpce]). The status of the end to end LSP (A-B and A-C) is made up when all the per domain LSP are reported up by the PNCs.

    * Delegation: It is suggested that the per domain LSPs are delegated to respective PNC, so that they can control the path and attributes based on each domain network conditions.

    * State Synchronization: The state needs to be synchronized between the parent PCE and child PCE. The mechanism described in [I-D.litkowski-pce-state-sync] can be used.

- VN Modify: MDSC is requested to modify a VN, for example the bandwidth for VN is increased. This may trigger path computation at MDSC as described in the previous step and can trigger an update to existing per-intra-domain path (via PCUpd message) or creation (or deletion) of a per-domain path (via PCInitiate message). This should be done in make-before-break fashion.

- VN Delete: MDSC is requested to delete a VN, in this case, based on the E2E paths and the resulting per-domain paths need to be removed (via PCInitiate message).

- VN Update (based on network changes): Any change in the per-domain LSP are reported to the MDSC (via PCRpt message) as per [I-D.dhodylee-pce-stateful-hpce]. This may result in changes in the E2E path or VN status. This may also trigger a re-optimization leading to a new per-domain path, update to existing path, or deletion of the path.

- VN Protection: The VN protection/restoration requirements, need to applied to each E2E path as well as each per domain path. The MDSC needs to play a crucial role in coordinating the right protection/restoration policy across each PNC. The existing protection/restoration mechanism of PCEP can be applied on each path.

- In case PNC generates an abstract topology to the MDSC, the PCInitiate/PCUpd messages from the MDSC to a PNC will contain a path with abstract nodes and links. PNC would need to take that as an input for path computation to get a path with physical nodes and links. Similarly PNC would convert the path received from the device (with physical nodes and links) into abstract path (based on the abstract topology generated before with abstract nodes and links) and reported to the MDSC.

5. Relationship to PCE based central control

[I-D.ietf-teas-pce-central-control] introduces the architecture for PCE as a central controller (PCECC), it further examines the motivations and applicability for PCEP as a southbound interface, and introduces the implications for the protocol. The section 2.1.3 of

[I-D.ietf-teas-pce-central-control] describe an hierarchy of PCE-based controller as per the Hierarchy of PCE framework defined in [RFC6805]. Both ACTN and PCECC is based on the same basic framework and thus compatible with each other.

6. IANA Considerations

This is an informational document and thus does not have any IANA allocations to be made.

7. Security Considerations

The ACTN framework described in [I-D.ietf-teas-actn-framework] defines key components and interfaces for managed traffic engineered networks. It also list various security considerations such as request and control of resources, confidentially of the information, and availability of function which should be taken into consideration.

When PCEP is used on the MPI/MMI, this interface needs to be secured, use of [I-D.ietf-pce-pceps] is RECOMENDED. Each PCEP extension listed in this document, presents its individual security considerations, which continue to apply.

8. Acknowledgments

The authors would like to thank Jonathan Hardwick for the inspiration behind this document. Further thanks to Avantika for her comments with suggested text.

9. References

9.1. Normative References

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

9.2. Informative References

[RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, DOI 10.17487/RFC3630, September 2003, <http://www.rfc-editor.org/info/rfc3630>.

[RFC4203] Kompella, K., Ed. and Y. Rekhter, Ed., "OSPF Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4203, DOI 10.17487/RFC4203, October 2005, <http://www.rfc-editor.org/info/rfc4203>.

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

[RFC5152] Vasseur, JP., Ed., Ayyangar, A., Ed., and R. Zhang, "A Per-Domain Path Computation Method for Establishing Inter-Domain Traffic Engineering (TE) Label Switched Paths (LSPs)", RFC 5152, DOI 10.17487/RFC5152, February 2008, <http://www.rfc-editor.org/info/rfc5152>.

[RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, DOI 10.17487/RFC5305, October 2008, <http://www.rfc-editor.org/info/rfc5305>.

[RFC5307] Kompella, K., Ed. and Y. Rekhter, Ed., "IS-IS Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 5307, DOI 10.17487/RFC5307, October 2008, <http://www.rfc-editor.org/info/rfc5307>.

[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, <http://www.rfc-editor.org/info/rfc5441>.

[RFC5623] Oki, E., Takeda, T., Le Roux, JL., and A. Farrel, "Framework for PCE-Based Inter-Layer MPLS and GMPLS Traffic Engineering", RFC 5623, DOI 10.17487/RFC5623, September 2009, <http://www.rfc-editor.org/info/rfc5623>.

[RFC6805] King, D., Ed. and A. Farrel, Ed., "The Application of the Path Computation Element Architecture to the Determination of a Sequence of Domains in MPLS and GMPLS", RFC 6805, DOI 10.17487/RFC6805, November 2012, <http://www.rfc-editor.org/info/rfc6805>.

[RFC7399] Farrel, A. and D. King, "Unanswered Questions in the Path Computation Element Architecture", RFC 7399, DOI 10.17487/RFC7399, October 2014, <http://www.rfc-editor.org/info/rfc7399>.

[RFC7491] King, D. and A. Farrel, "A PCE-Based Architecture for Application-Based Network Operations", RFC 7491, DOI 10.17487/RFC7491, March 2015, <http://www.rfc-editor.org/info/rfc7491>.

[RFC7752] Gredler, H., Ed., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP", RFC 7752, DOI 10.17487/RFC7752, March 2016, <http://www.rfc-editor.org/info/rfc7752>.

[RFC8051] Zhang, X., Ed. and I. Minei, Ed., "Applicability of a Stateful Path Computation Element (PCE)", RFC 8051, DOI 10.17487/RFC8051, January 2017, <http://www.rfc-editor.org/info/rfc8051>.

[I-D.ietf-pce-stateful-pce]
Crabbe, E., Minei, I., Medved, J., and R. Varga, "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce-18 (work in progress), December 2016.

[I-D.ietf-pce-pce-initiated-lsp]
Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model", draft-ietf-pce-pce-initiated-lsp-09 (work in progress), March 2017.

[I-D.dhodylee-pce-stateful-hpce]
Dhody, D., Lee, Y., Ceccarelli, D., Shin, J., King, D., and O. Dios, "Hierarchical Stateful Path Computation Element (PCE).", draft-dhodylee-pce-stateful-hpce-03 (work in progress), March 2017.

[I-D.ietf-teas-pce-central-control]
Farrel, A., Zhao, Q., Li, Z., and C. Zhou, "An Architecture for Use of PCE and PCEP in a Network with Central Control", draft-ietf-teas-pce-central-control-01 (work in progress), December 2016.

[I-D.ietf-teas-actn-requirements]
Lee, Y., Dhody, D., Belotti, S., Pithewan, K., and D. Ceccarelli, "Requirements for Abstraction and Control of TE Networks", draft-ietf-teas-actn-requirements-04 (work in progress), January 2017.

[I-D.ietf-teas-actn-framework]
Ceccarelli, D. and Y. Lee, "Framework for Abstraction and Control of Traffic Engineered Networks", draft-ietf-teas-actn-framework-04 (work in progress), February 2017.

[I-D.ietf-teas-actn-info-model]
Lee, Y., Belotti, S., Dhody, D., Ceccarelli, D., and B. Yoon, "Information Model for Abstraction and Control of TE Networks (ACTN)", draft-ietf-teas-actn-info-model-00 (work in progress), February 2017.

[I-D.ietf-pce-inter-area-as-applicability]
King, D., Meuric, J., Dugeon, O., Zhao, Q., Dhody, D., and O. Dios, "Applicability of the Path Computation Element to Inter-Area and Inter-AS MPLS and GMPLS Traffic Engineering", draft-ietf-pce-inter-area-as-applicability-06 (work in progress), July 2016.

[I-D.dhodylee-pce-pcep-ls]
Dhody, D., Lee, Y., and D. Ceccarelli, "PCEP Extension for Distribution of Link-State and TE Information.", draft-dhodylee-pce-pcep-ls-07 (work in progress), March 2017.

[I-D.leedhody-pce-vn-association]
Lee, Y., Dhody, D., Zhang, X., and D. Ceccarelli, "PCEP Extensions for Establishing Relationships Between Sets of LSPs and Virtual Networks", draft-leedhody-pce-vn-association-02 (work in progress), March 2017.

[I-D.litkowski-pce-state-sync]
Litkowski, S., Sivabalan, S., and D. Dhody, "Inter Stateful Path Computation Element communication procedures", draft-litkowski-pce-state-sync-01 (work in progress), February 2017.

[I-D.ietf-pce-association-policy]
Dhody, D., Sivabalan, S., Litkowski, S., Tantsura, J., and J. Hardwick, "Path Computation Element communication Protocol extension for associating Policies and LSPs", draft-ietf-pce-association-policy-00 (work in progress), December 2016.

[I-D.ietf-pce-pceps]
Lopez, D., Dios, O., Wu, W., and D. Dhody, "Secure Transport for PCEP", draft-ietf-pce-pceps-11 (work in progress), January 2017.

[I-D.lee-teas-actn-abstraction]
Lee, Y., Dhody, D., Ceccarelli, D., and O. Dios, "Abstraction and Control of TE Networks (ACTN) Abstraction Methods", draft-lee-teas-actn-abstraction-00 (work in progress), October 2016.

Authors' Addresses

Dhruv Dhody
Huawei Technologies
Divyashree Techno Park, Whitefield
Bangalore, Karnataka 560066
India

EMail: dhruv.ietf@gmail.com

Young Lee
Huawei Technologies
5340 Legacy Drive, Building 3
Plano, TX 75023
USA

EMail: leeyoung@huawei.com

Daniele Ceccarelli
Ericsson
Torshamnsgatan,48
Stockholm
Sweden

EMail: daniele.ceccarelli@ericsson.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: December 29, 2017

D. Dhody
Y. Lee
Huawei Technologies
D. Ceccarelli
Ericsson
June 27, 2017

PCEP Extension for Distribution of Link-State and TE Information. draft-dhodylee-pce-pcep-ls-08

Abstract

In order to compute and provide optimal paths, Path Computation Elements (PCEs) require an accurate and timely Traffic Engineering Database (TED). Traditionally this TED has been obtained from a link state (LS) routing protocol supporting traffic engineering extensions.

This document extends the Path Computation Element Communication Protocol (PCEP) with Link-State and TE Information.

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 29, 2017.

Copyright Notice

Table of Contents

1. Introduction

In Multiprotocol Label Switching (MPLS) and Generalized MPLS (GMPLS), a Traffic Engineering Database (TED) is used in computing paths for connection oriented packet services and for circuits. The TED contains all relevant information that a Path Computation Element (PCE) needs to perform its computations. It is important that the TED be complete and accurate each time, the PCE performs a path computation.

In MPLS and GMPLS, interior gateway routing protocols (IGPs) have been used to create and maintain a copy of the TED at each node running the IGP. One of the benefits of the PCE architecture [RFC4655] is the use of computationally more sophisticated path computation algorithms and the realization that these may need enhanced processing power not necessarily available at each node participating in an IGP.

Section 4.3 of [RFC4655] describes the potential load of the TED on a network node and proposes an architecture where the TED is maintained by the PCE rather than the network nodes. However, it does not

describe how a PCE would obtain the information needed to populate its TED. PCE may construct its TED by participating in the IGP ([RFC3630] and [RFC5305] for MPLS-TE; [RFC4203] and [RFC5307] for GMPLS). An alternative is offered by BGP-LS [RFC7752] .

[I-D.ietf-pce-stateful-pce] describes a set of extensions to PCEP to provide stateful control. A stateful PCE has access to not only the information carried by the network's Interior Gateway Protocol (IGP), but also the set of active paths and their reserved resources for its computations. PCC can delegate the rights to modify the LSP parameters to an Active Stateful PCE. This requires PCE to quickly be updated on any changes in the Topology and TEDB, so that PCE can meet the need for updating LSPs effectively and in a timely manner. The fastest way for a PCE to be updated on TED changes is via a direct interface with each network node and with incremental update from each network node with only the attribute that is modified.

[I-D.ietf-pce-pce-initiated-lsp] describes the setup, maintenance and teardown of PCE-initiated LSPs under the stateful PCE model, without the need for local configuration on the PCC, thus allowing for a dynamic network that is centrally controlled and deployed. This model requires timely topology and TED update at the PCE.

[RFC5440] describes the specifications for the Path Computation Element Communication Protocol (PCEP). PCEP specifies the communication between a Path Computation Client (PCC) and a Path Computation Element (PCE), or between two PCEs based on the PCE architecture [RFC4655].

This document describes a mechanism by which Link State and TE information can be collected from networks and shared with PCE using the PCEP itself. This is achieved using a new PCEP message format. The mechanism is applicable to physical and virtual links as well as further subjected to various policies.

A network node maintains one or more databases for storing link-state and TE information about nodes and links in any given area. Link attributes stored in these databases include: local/remote IP addresses, local/ remote interface identifiers, link metric and TE metric, link bandwidth, reservable bandwidth, per CoS class reservation state, preemption and Shared Risk Link Groups (SRLG). The node's PCEP process can retrieve topology from these databases and distribute it to a PCE, either directly or via another PCEP Speaker, using the encoding specified in this document.

Further [RFC6805] describes Hierarchical-PCE architecture, where a parent PCE maintains a domain topology map. To build this domain topology map, the child PCE can carry the border nodes and inter-

domain link information to the parent PCE using the mechanism described in this document. Further as described in [I-D.ietf-pce-applicability-actn], the child PCE can also transport abstract Link-State and TE information from child PCE to a Parent PCE using the mechanism described in this document to build an abstract topology at the parent PCE.

[I-D.ietf-pce-stateful-pce] describe LSP state synchronization between PCCs and PCEs in case of stateful PCE. This document does not make any change to the LSP state synchronization process. The mechanism described in this document are on top of the existing LSP state synchronization.

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 terminology is as per [RFC4655] and [RFC5440].

3. Applicability

The mechanism specified in this draft is applicable to deployments:

- Where there is no IGP or BGP-LS running in the network.
- Where there is no IGP or BGP-LS running at the PCE to learn link-state and TE information.
- Where there is IGP or BGP-LS running but with a need for a faster TE and link-state population and convergence at the PCE.
  - A PCE may receive partial information (say basic TE, link-state) from IGP and other information (optical and impairment) from PCEP.
  - A PCE may receive an incremental update (as opposed to the entire information of the node/link).
  - A PCE may receive full information from both existing mechanism (IGP or BGP) and PCEP.
- Where there is a need for transporting (abstract) Link-State and TE information from child PCE to a Parent PCE in H-PCE [RFC6805]; as well as for Physical Network Controller (PNC) to Multi-Domain

Service Coordinator (MDSC) in Abstraction and Control of TE Networks (ACTN) [I-D.ietf-teas-actn-framework].

A PCC may further choose to send only local information or both local and remote learned information.

How a PCE manages the link-state (and TE) information is implementation specific and thus out of scope of this document.

The prefix information in PCEP-LS can also help in determining the domain of the endpoints in H-PCE (and ACTN). Section 4.5 of [RFC6805] describe various mechanism and procedures that might be used, PCEP-LS provides a simple mechanism to exchange this information.

4. Requirements for PCEP extension

Following key requirements associated with link-state (and TE) distribution are identified for PCEP:

1. The PCEP speaker supporting this draft MUST be a mechanism to advertise the Link-State (and TE) distribution capability.

2. PCC supporting this draft MUST have the capability to report the link-state (and TE) information to the PCE. This includes self originated information and remote information learned via routing protocols. PCC MUST be capable to do the initial bulk sync at the time of session initialization as well as changes after.

3. A PCE MAY learn link-state (and TE) from PCEP as well as from existing mechanism like IGP/BGP-LS. PCEP extension MUST have a mechanism to link the information learned via other means. There MUST NOT be any changes to the existing link-state (and TE) population mechanism via IGP/BGP-LS. PCEP extension SHOULD keep the properties in a protocol (IGP or BGP-LS) neutral way, such that an implementation may not need to know about any OSPF or IS-IS or BGP protocol specifics.

4. It SHOULD be possible to encode only the changes in link-state (and TE) properties (after the initial sync) in PCEP messages.

5. The same mechanism should be used for both MPLS TE as well as GMPLS, optical and impairment aware properties.

6. The same mechanism should be used for PCE to PCE Link-state (and TE) synchronization.

5. New Functions to distribute link-state (and TE) via PCEP

Several new functions are required in PCEP to support distribution of link-state (and TE) information. 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 advertisement (E-C,C-E): both the PCC and the PCE must announce during PCEP session establishment that they support PCEP extensions for distribution of link-state (and TE) information defined in this document.

- Link-State (and TE) synchronization (C-E): after the session between the PCC and a PCE is initialized, the PCE must learn Link-State (and TE) information before it can perform path computations. In case of stateful PCE it is RECOMENDED that this operation be done before LSP state synchronization.

- Link-State (and TE) Report (C-E): a PCC sends a LS (and TE) report to a PCE whenever the Link-State and TE information changes.

6. Overview of Extension to PCEP

6.1. New Messages

In this document, we define a new PCEP messages called LS Report (LSRpt), a PCEP message sent by a PCC to a PCE to report link-state (and TE) information. Each LS Report in a LSRpt message can contain the node or link properties. An unique PCEP specific LS identifier (LS-ID) is also carried in the message to identify a node or link and that remains constant for the lifetime of a PCEP session. This identifier on its own is sufficient when no IGP or BGP-LS running in the network for PCE to learn link-state (and TE) information. Incase PCE learns some information from PCEP and some from the existing mechanism, the PCC SHOULD include the mapping of IGP or BGP-LS identifier to map the information populated via PCEP with IGP/BGP-LS. See Section 8.1 for details.

6.2. Capability Advertisement

During PCEP Initialization Phase, PCEP Speakers (PCE or PCC) advertise their support of LS (and TE) distribution via PCEP extensions. A PCEP Speaker includes the "LS Capability" TLV, described in Section 9.2.1, in the OPEN Object to advertise its support for PCEP-LS extensions. The presence of the LS Capability TLV in PCC’s OPEN Object indicates that the PCC is willing to send LS Reports whenever local link-state (and TE) information changes. The presence of the LS Capability TLV in PCE’s OPEN message indicates

that the PCE is interested in receiving LS Reports whenever local link-state (and TE) information changes.

The PCEP protocol extensions for LS (and TE) distribution MUST NOT be used if one or both PCEP Speakers have not included the LS Capability TLV in their respective OPEN message. If the PCE that supports the extensions of this draft but did not advertise this capability, then upon receipt of a LSRpt message from the PCC, it SHOULD generate a PCErr with error-type 19 (Invalid Operation), error-value TBD1 (Attempted LS Report if LS capability was not advertised) and it will terminate the PCEP session.

The LS reports sent by PCC MAY carry the remote link-state (and TE) information learned via existing means like IGP and BGP-LS only if both PCEP Speakers set the R (remote) Flag in the "LS Capability" TLV to 'Remote Allowed (R Flag = 1)'. If this is not the case and LS reports carry remote link-state (and TE) information, then a PCErr with error-type 19 (Invalid Operation) and error-value TBD1 (Attempted LS Report if LS remote capability was not advertised) and it will terminate the PCEP session.

6.3. Initial Link-State (and TE) Synchronization

The purpose of LS Synchronization is to provide a checkpoint-in- time state replica of a PCC's link-state (and TE) data base in a PCE. State Synchronization is performed immediately after the Initialization phase (see [RFC5440]]). In case of stateful PCE ([I-D.ietf-pce-stateful-pce]) it is RECOMENDED that the LS synchronization should be done before LSP state synchronization.

During LS Synchronization, a PCC first takes a snapshot of the state of its database, then sends the snapshot to a PCE in a sequence of LS Reports. Each LS Report sent during LS Synchronization has the SYNC Flag in the LS Object set to 1. The end of synchronization marker is a LSRpt message with the SYNC Flag set to 0 for an LS Object with LS-ID equal to the reserved value 0. If the PCC has no link-state to synchronize, it will only send the end of synchronization marker.

Either the PCE or the PCC MAY terminate the session using the PCEP session termination procedures during the synchronization phase. If the session is terminated, the PCE MUST clean up state it received from this PCC. The session re-establishment MUST be re-attempted per the procedures defined in [RFC5440], including use of a back-off timer.

If the PCC encounters a problem which prevents it from completing the LS synchronization, it MUST send a PCErr message with error-type TBD2

(LS Synchronization Error) and error-value 2 (indicating an internal PCC error) to the PCE and terminate the session.

The PCE does not send positive acknowledgements for properly received LS synchronization messages. It MUST respond with a PCErr message with error-type TBD2 (LS Synchronization Error) and error-value 1 (indicating an error in processing the LSRpt) if it encounters a problem with the LS Report it received from the PCC and it MUST terminate the session.

The LS reports can carry local as well as remote link-state (and TE) information depending on the R flag in LS capability TLV.

The successful LS Synchronization sequences is shown in Figure 1.

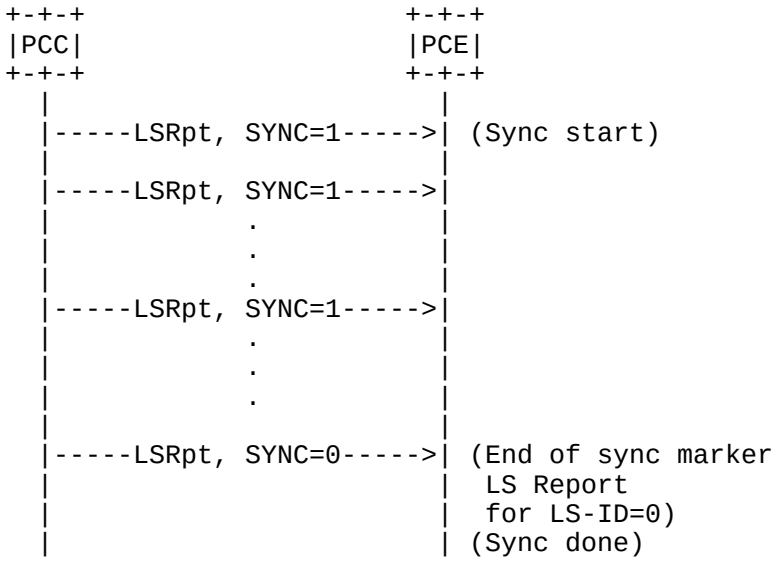


Figure 1: Successful LS synchronization

The sequence where the PCE fails during the LS Synchronization phase is shown in Figure 2.

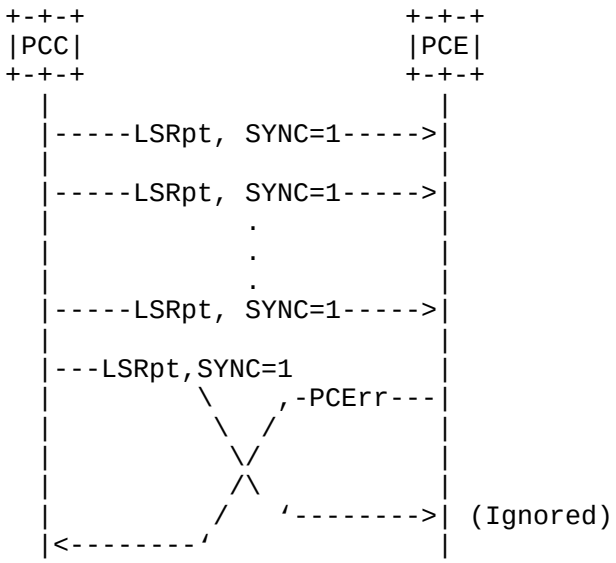


Figure 2: Failed LS synchronization (PCE failure)

The sequence where the PCC fails during the LS Synchronization phase is shown in Figure 3.

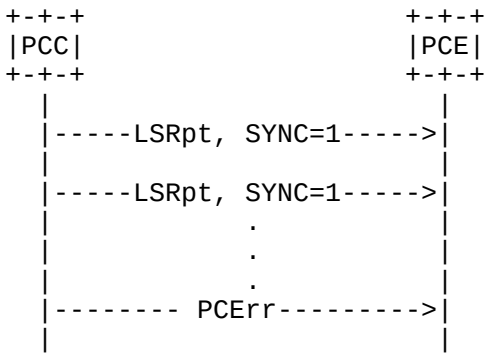


Figure 3: Failed LS synchronization (PCC failure)

6.3.1. Optimizations for LS Synchronization

These optimizations are described in [I-D.kondreddy-pce-pcep-ls-sync-optimizations].

6.4. LS Report

The PCC MUST report any changes in the link-state (and TE) information to the PCE by sending a LS Report carried on a LSRpt message to the PCE. Each node and Link would be uniquely identified by a PCEP LS identifier (LS-ID). The LS reports may carry local as well as remote link-state (and TE) information depending on the R flag in LS capability TLV. In case R flag is set, It MAY also include the mapping of IGP or BGP-LS identifier to map the information populated via PCEP with IGP/BGP-LS.

More details about LSRpt message are in Section 8.1.

7. Transport

A permanent PCEP session MUST be established between a PCE and PCC supporting link-state (and TE) distribution via PCEP. In the case of session failure, session re-establishment MUST be re-attempted per the procedures defined in [RFC5440].

8. 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.

8.1. LS Report Message

A PCEP LS Report message (also referred to as LSRpt message) is a PCEP message sent by a PCC to a PCE to report the link-state (and TE) information. A LSRpt message can carry more than one LS Reports. The Message-Type field of the PCEP common header for the LSRpt message is set to [TBD3].

The format of the LSRpt message is as follows:

```
<LSRpt Message> ::= <Common Header>
                    <ls-report-list>
Where:

<ls-report-list> ::= <LS>[<ls-report-list>]
```

The LS object is a mandatory object which carries LS information of a node or a link. Each LS object has an unique LS-ID as described in Section 9.3. If the LS object is missing, the receiving PCE MUST send a PCErr message with Error-type=6 (Mandatory Object missing) and Error-value=[TBD4] (LS object missing).

A PCE may choose to implement a limit on the LS information a single PCC can populate. If a LSRpt is received that causes the PCE to exceed this limit, it MUST send a PCErr message with error-type 19 (invalid operation) and error-value 4 (indicating resource limit exceeded) in response to the LSRpt message triggering this condition and SHOULD terminate the session.

8.2. The PCErr Message

If a PCEP speaker has advertised the LS capability on the PCEP session, the PCErr message MAY include the LS object. If the error reported is the result of an LS report, then the LS-ID number MUST be the one from the LSRpt that triggered the error.

The format of a PCErr message from [RFC5440] is extended as follows:

The format of the PCErr message is as follows:

```
<PCErr Message> ::= <Common Header>
                    ( <error-obj-list> [<Open>] ) | <error>
                    [<error-list>]

<error-obj-list>::=<PCEP-ERROR>[<error-obj-list>]

<error>::=[<request-id-list> | <ls-id-list>]
           <error-obj-list>

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

<ls-id-list>::=<LS>[<ls-id-list>]

<error-list>::=<error>[<error-list>]
```

9. Objects and TLV

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.

9.1. TLV Format

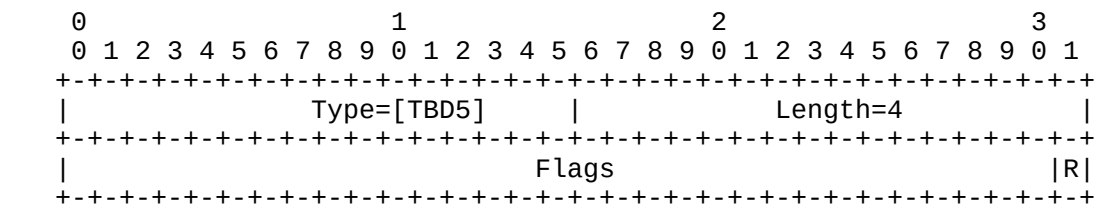
The TLV and the sub-TLV format (and padding) in this document, is as per section 7.1 of [RFC5440].

9.2. Open Object

This document defines a new optional TLV for use in the OPEN Object.

9.2.1. LS Capability TLV

The LS-CAPABILITY TLV is an optional TLV for use in the OPEN Object for link-state (and TE) distribution via PCEP capability advertisement. Its format is shown in the following figure:



The type of the TLV is [TBD5] and it has a fixed length of 4 octets.

The value comprises a single field - Flags (32 bits):

- R (remote - 1 bit): if set to 1 by a PCC, the R Flag indicates that the PCC allows reporting of remote LS information learned via other means like IGP and BGP-LS; if set to 1 by a PCE, the R Flag indicates that the PCE is capable of receiving remote LS information (from the PCC point of view). The R Flag must be advertised by both a PCC and a PCE for LSRpt messages to report remote as well as local LS information on a PCEP session. The TLVs related to IGP/BGP-LS identifier MUST be encoded when both PCEP speakers have the R Flag set.

Unassigned bits are considered reserved. They MUST be set to 0 on transmission and MUST be ignored on receipt.

Advertisement of the LS capability implies support of local link-state (and TE) distribution, as well as the objects, TLVs and procedures defined in this document.

9.3. LS Object

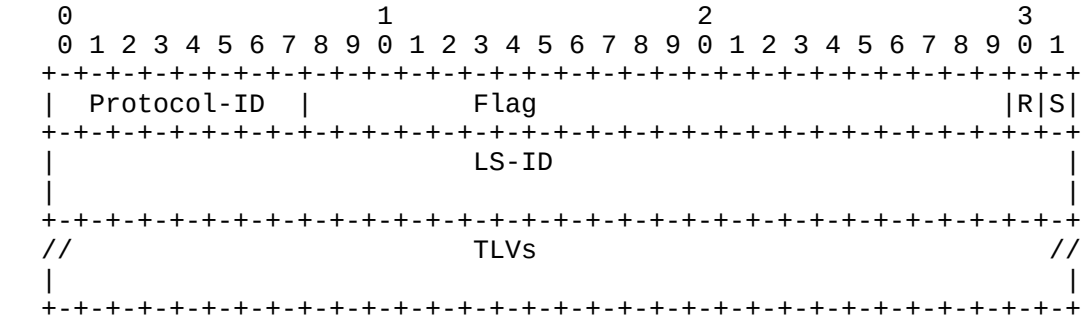
The LS (link-state) object MUST be carried within LSRpt messages and MAY be carried within PCErr messages. The LS object contains a set of fields used to specify the target node or link. It also contains a flag indicating to a PCE that the LS synchronization is in progress. The TLVs used with the LS object correlate with the IGP/ BGP-LS encodings.

LS Object-Class is [TBD6].

Four Object-Type values are defined for the LS object so far:

- LS Node: LS Object-Type is 1.
- LS Link: LS Object-Type is 2.
- LS IPv4 Topology Prefix: LS Object-Type is 3.
- LS IPv6 Topology Prefix: LS Object-Type is 4.

The format of all types of LS object is as follows:



Protocol-ID (8-bit): The field provide the source information. The protocol could be an IGP, BGP-LS or an abstraction algorithm. Incase PCC only provides local information of the PCC, it MUST use Protocol-ID as Direct. The following values are defined (some of them are same as [RFC7752]):

Protocol-ID	Source protocol
1	IS-IS Level 1
2	IS-IS Level 2
3	OSPFv2
4	Direct
5	Static configuration
6	OSPFv3
7	BGP-LS
8	PCEP-LS
9	Abstraction
10	Unspecified

Flags (24-bit):

- S (SYNC - 1 bit): the S Flag MUST be set to 1 on each LSRpt sent from a PCC during LS Synchronization. The S Flag MUST be set to 0 in other LSRpt messages sent from the PCC.

- R (Remove - 1 bit): On LSRpt messages the R Flag indicates that the node/link/prefix has been removed from the PCC and the PCE SHOULD remove from its database. Upon receiving an LS Report with the R Flag set to 1, the PCE SHOULD remove all state for the node/link/prefix identified by the LS Identifiers from its database.

LS-ID(64-bit): A PCEP-specific identifier for the node or link or prefix information. A PCC creates an unique LS-ID for each node/link/prefix that is constant for the lifetime of a PCEP session. The PCC will advertise the same LS-ID on all PCEP sessions it maintains at a given times. All subsequent PCEP messages then address the node/link/prefix by the LS-ID. The values of 0 and 0xFFFFFFFFFFFFFFFF are reserved.

Unassigned bits are considered reserved. They MUST be set to 0 on transmission and MUST be ignored on receipt.

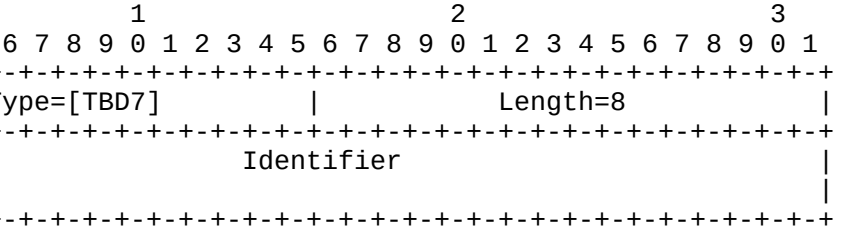
TLVs that may be included in the LS Object are described in the following sections.

9.3.1. Routing Universe TLV

In case of remote link-state (and TE) population when existing IGP/ BGP-LS are also used, OSPF and IS-IS may run multiple routing protocol instances over the same link as described in [RFC7752]. See [RFC6822] and [RFC6549] for more information. These instances define

independent "routing universes". The 64-Bit 'Identifier' field is used to identify the "routing universe" where the LS object belongs. The LS objects representing IGP objects (nodes or links or prefix) from the same routing universe MUST have the same 'Identifier' value; LS objects with different 'Identifier' values MUST be considered to be from different routing universes.

The format of the optional ROUTING-UNIVERSE TLV is shown in the following figure:



Below table lists the 'Identifier' values that are defined as well-known in this draft (same as [RFC7752]).

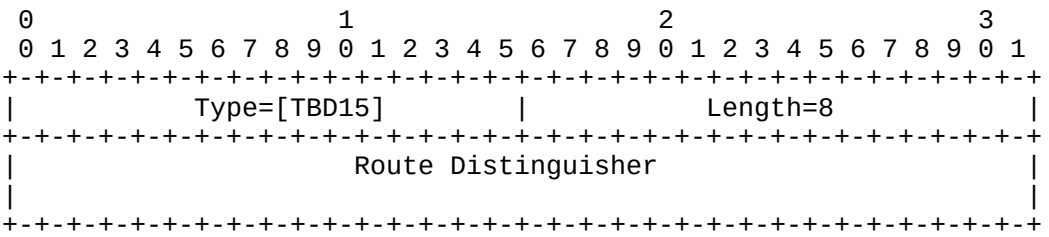
Identifier	Routing Universe
0	Default Layer 3 Routing topology
1-31	Reserved

If this TLV is not present the default value 0 is assumed.

9.3.2. Route Distinguisher TLV

To allow identification of VPN link, node and prefix information in PCEP-LS, a Route Distinguisher (RD) [RFC4364] is used. The LS objects from the same VPN MUST have the same RD; LS objects with different RD values MUST be considered to be from different VPNs.

The format of the optional ROUTE-DISTINGUISHER TLV is shown in the following figure:



The format of RD is as per [RFC4364].

9.3.3. Virtual Network TLV

To realize ACTN, the MDSC needs to build an multi-domain topology. This topology is best served, if this is an abstracted view of the underlying network resources of each domain. It is also important to provide a customer view of network slice for each customer. There is a need to control the level of abstraction based on the deployment scenario and business relationship between the controllers.

Virtual service coordination function in ACTN incorporates customer service-related knowledge into the virtual network operations in order to seamlessly operate virtual networks while meeting customer's service requirements. [I-D.ietf-teas-actn-requirements] describes various VN operations initiated by a customer/application. In this context, there is a need for associating the abstracted link state and TE topology with a VN "construct" to facilitate VN operations in PCE architecture.

VIRTUAL-NETWORK-TLV as per [I-D.leedhody-pce-vn-association] can be included in LS object to identify the link, node and prefix information belongs to a particular VN.

9.3.4. Local Node Descriptors TLV

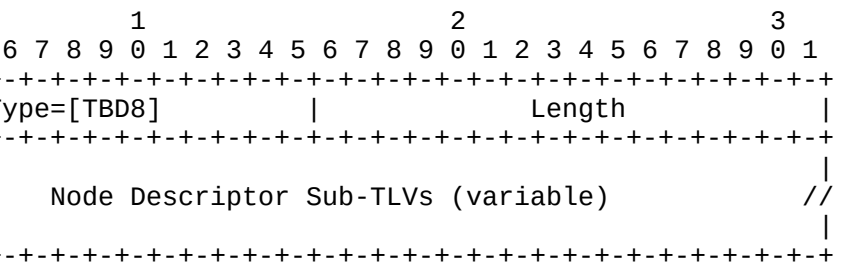
As described in [RFC7752], each link is anchored by a pair of Router-IDs that are used by the underlying IGP, namely, 48 Bit ISO System-ID for IS-IS and 32 bit Router-ID for OSPFv2 and OSPFv3. Incase of additional auxiliary Router-IDs used for TE, these MUST also be included in the link attribute TLV (see Section 9.3.9.2).

It is desirable that the Router-ID assignments inside the Node Descriptor are globally unique. Some considerations for globally unique Node/Link/Prefix identifiers are described in [RFC7752].

The Local Node Descriptors TLV contains Node Descriptors for the node anchoring the local end of the link. This TLV MUST be included in the LS Report when during a given PCEP session a node/link/prefix is

first reported to a PCE. A PCC sends to a PCE the first LS Report either during State Synchronization, or when a new node/link/prefix is learned at the PCC. The value contains one or more Node Descriptor Sub-TLVs, which allows specification of a flexible key for any given node/link/prefix information such that global uniqueness of the node/link/prefix is ensured.

This TLV is applicable for all LS Object-Type.

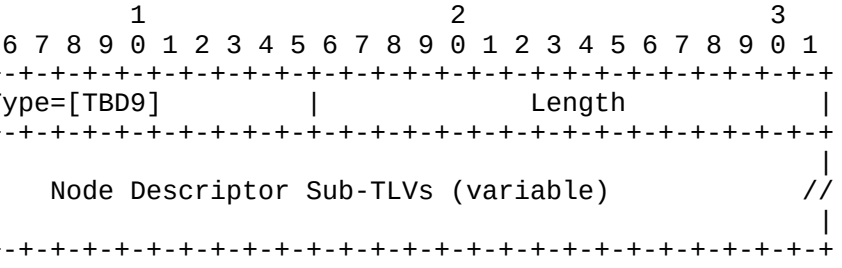


The value contains one or more Node Descriptor Sub-TLVs defined in Section 9.3.6.

9.3.5. Remote Node Descriptors TLV

The Remote Node Descriptors contains Node Descriptors for the node anchoring the remote end of the link. This TLV MUST be included in the LS Report when during a given PCEP session a link is first reported to a PCE. A PCC sends to a PCE the first LS Report either during State Synchronization, or when a new link is learned at the PCC. The length of this TLV is variable. The value contains one or more Node Descriptor Sub-TLVs defined in Section 9.3.6.

This TLV is applicable for LS Link Object-Type.



9.3.6. Node Descriptors Sub-TLVs

The Node Descriptor Sub-TLV type Type and lengths are listed in the following table:

Sub-TLV	Description	Length	Value defined in
0	Reserved	-	-
1	Autonomous System	4	[RFC7752] / section 3.2.1.4
2	BGP-LS Identifier	4	[RFC7752] / section 3.2.1.4
3	OSPF Area-ID	4	[RFC7752] / section 3.2.1.4
4	Router-ID	Variable	[RFC7752] / section 3.2.1.4

The sub-TLV values in Node Descriptor TLVs are defined as follows (similar to [RFC7752]):

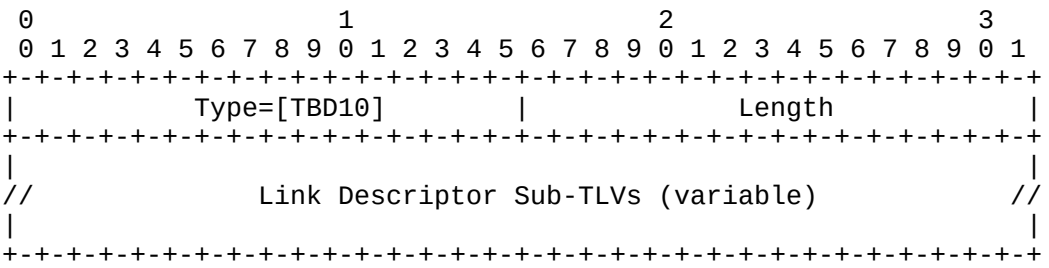
- Autonomous System: opaque value (32 Bit AS Number)
- BGP-LS Identifier: opaque value (32 Bit ID). In conjunction with ASN, uniquely identifies the BGP-LS domain as described in [RFC7752]. This sub-TLV is present only if the node implements BGP-LS and the ID is set by the operator.
- OSPF Area ID: It is used to identify the 32 Bit area to which the LS object belongs. Area Identifier allows the different LS objects of the same node to be discriminated.
- Router ID: opaque value. Usage is described in [RFC7752] as IGP Router ID. In case this is not learned from IGP, it SHOULD contain the unique router ID, such as TE router ID.

9.3.7. Link Descriptors TLV

The Link Descriptors TLV contains Link Descriptors for each link. This TLV MUST be included in the LS Report when during a given PCEP session a link is first reported to a PCE. A PCC sends to a PCE the first LS Report either during State Synchronization, or when a new link is learned at the PCC. The length of this TLV is variable. The value contains one or more Link Descriptor Sub-TLVs.

The 'Link descriptor' TLVs uniquely identify a link among multiple parallel links between a pair of anchor routers similar to [RFC7752].

This TLV is applicable for LS Link Object-Type.



The Link Descriptor Sub-TLV type and lengths are listed in the following table:

Sub-TLV	Description	IS-IS TLV /Sub-TLV	Value defined in:
6	Link Local/Remote Identifiers	22/4	[RFC5307]/1.1
7	IPv4 interface address	22/6	[RFC5305]/3.2
8	IPv4 neighbor address	22/8	[RFC5305]/3.3
9	IPv6 interface address	22/12	[RFC6119]/4.2
10	IPv6 neighbor address	22/13	[RFC6119]/4.3
5	Multi-Topology identifier	-	[RFC7752]/ 3.2.1.5

The format and semantics of the 'value' fields in most 'Link Descriptor' sub-TLVs correspond to the format and semantics of value fields in IS-IS Extended IS Reachability sub-TLVs, defined in [RFC5305], [RFC5307] and [RFC6119]. Although the encodings for 'Link Descriptor' TLVs were originally defined for IS-IS, the TLVs can carry data sourced either by IS-IS or OSPF or direct.

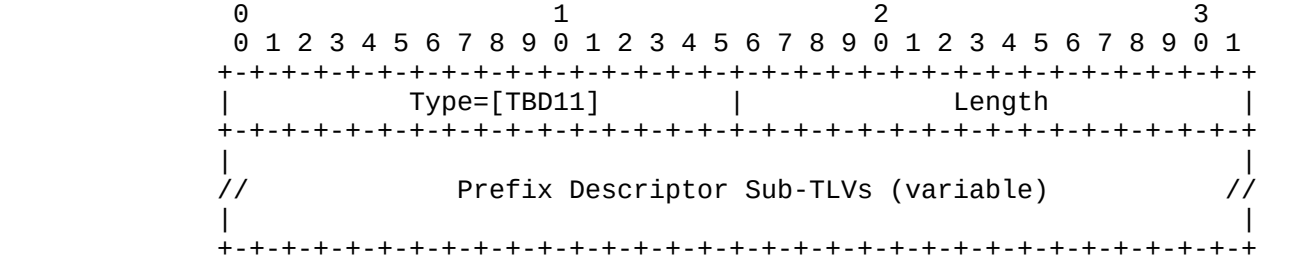
The information about a link present in the LSA/LSP originated by the local node of the link determines the set of sub-TLVs in the Link Descriptor of the link as described in [RFC7752].

9.3.8. Prefix Descriptors TLV

The Prefix Descriptors TLV contains Prefix Descriptors uniquely identify an IPv4 or IPv6 Prefix originated by a Node. This TLV MUST be included in the LS Report when during a given PCEP session a prefix is first reported to a PCE. A PCC sends to a PCE the first LS

Report either during State Synchronization, or when a new prefix is learned at the PCC. The length of this TLV is variable.

This TLV is applicable for LS Prefix Object-Types for both IPv4 and IPv6.



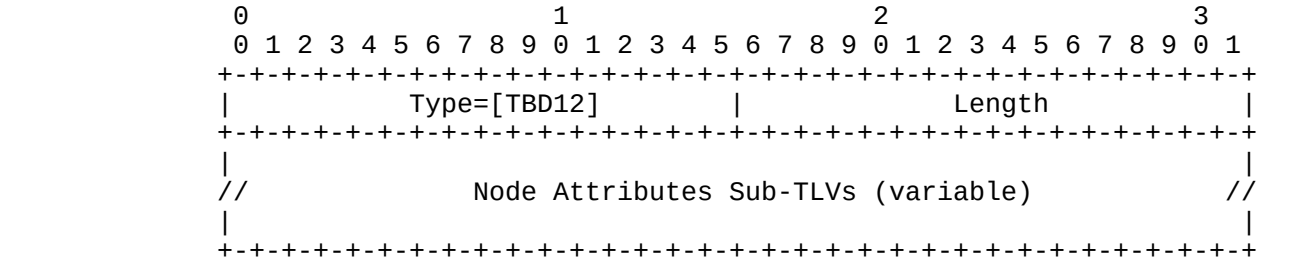
The value contains one or more Prefix Descriptor Sub-TLVs defined below -

TLV Code Point	Description	Length	Value defined in:
5	Multi-Topology Identifier	variable	[RFC7752] /3.2.1.5
11	OSPF Route Type	1	[RFC7752] /3.2.3.1
12	IP Reachability Information	variable	[RFC7752] /3.2.3.2

9.3.9. PCEP-LS Attributes

9.3.9.1. Node Attributes TLV

This is an optional attribute that is used to carry node attributes. This TLV is applicable for LS Node Object-Type.

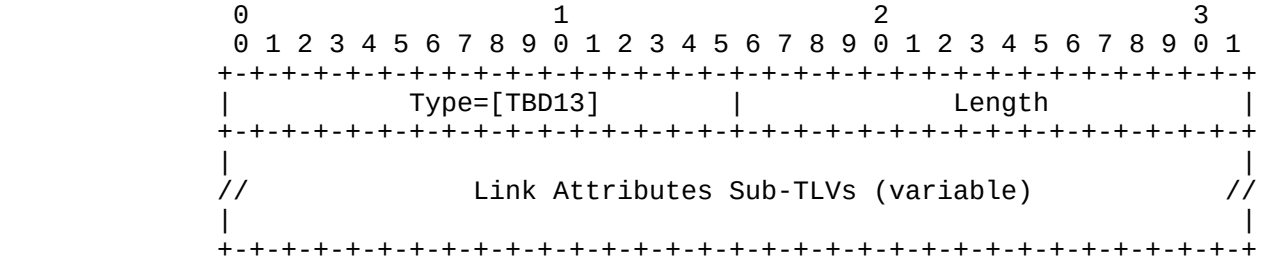


The Node Attributes Sub-TLV type and lengths are listed in the following table:

Sub TLV	Description	Length	Value defined in:
5	Multi-Topology Identifier	variable	[RFC7752] /3.2.1.5
13	Node Flag Bits	1	[RFC7752] /3.3.1.1
14	Opaque Node Properties	variable	[RFC7752] /3.3.1.5
15	Node Name	variable	[RFC7752] /3.3.1.3
16	IS-IS Area Identifier	variable	[RFC7752] /3.3.1.2
17	IPv4 Router-ID of Local Node	4	[RFC5305]/4.3
18	IPv6 Router-ID of Local Node	16	[RFC6119]/4.1

9.3.9.2. Link Attributes TLV

This TLV is applicable for LS Link Object-Type. The format and semantics of the 'value' fields in some 'Link Attribute' sub-TLVs correspond to the format and semantics of value fields in IS-IS Extended IS Reachability sub-TLVs, defined in [RFC5305], [RFC5307] and [RFC7752]. Although the encodings for 'Link Attribute' TLVs were originally defined for IS-IS, the TLVs can carry data sourced either by IS-IS or OSPF or direct.



The following 'Link Attribute' sub-TLVs are valid :

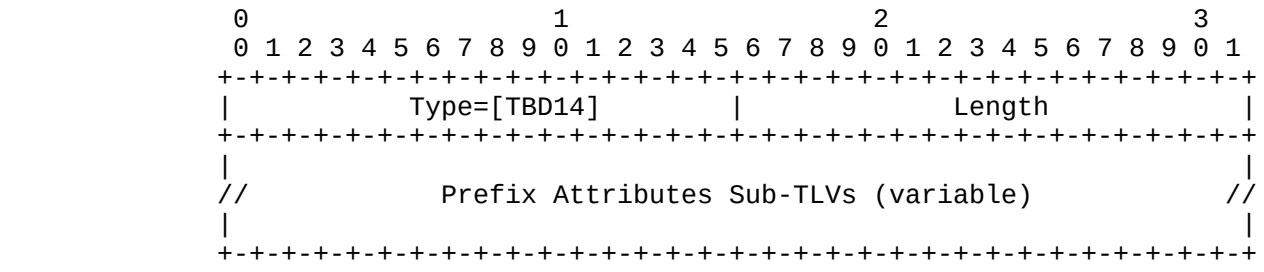
Sub-TLV	Description	IS-IS TLV /Sub-TLV	Defined in:

		BGP-LS TLV	
17	IPv4 Router-ID of Local Node	134/---	[RFC5305]/4.3
18	IPv6 Router-ID of Local Node	140/---	[RFC6119]/4.1
19	IPv4 Router-ID of Remote Node	134/---	[RFC5305]/4.3
20	IPv6 Router-ID of Remote Node	140/---	[RFC6119]/4.1
6	Link Local/Remote Identifiers	22/4	[RFC5307]/1.1
22	Administrative group (color)	22/3	[RFC5305]/3.1
23	Maximum link bandwidth	22/9	[RFC5305]/3.3
24	Max. reservable link bandwidth	22/10	[RFC5305]/3.5
25	Unreserved bandwidth	22/11	[RFC5305]/3.6
26	TE Default Metric	22/18	[RFC7752] /3.3.2.3
27	Link Protection Type	22/20	[RFC5307]/1.2
28	MPLS Protocol Mask	1094	[RFC7752] /3.3.2.2
29	IGP Metric	1095	[RFC7752] /3.3.2.4
30	Shared Risk Link Group	1096	[RFC7752] /3.3.2.5
31	Opaque link attributes	1097	[RFC7752] /3.3.2.6
32	Link Name attribute	1098	[RFC7752] /3.3.2.7
33	Unidirectional Link Delay	22/33	[RFC7810]/4.1
34	Min/Max Unidirectional Link Delay	22/34	[RFC7810]/4.2
35	Unidirectional Delay Variation	22/35	[RFC7810]/4.3
36	Unidirectional Link Loss	22/36	[RFC7810]/4.4
37	Unidirectional Residual Bandwidth	22/37	[RFC7810]/4.5
38	Unidirectional Available Bandwidth	22/38	[RFC7810]/4.6
39	Unidirectional	22/39	[RFC7810]/4.7

	Bandwidth Utilization			
40	Extended Admin Group (EAG)	22/14	[RFC7308]/2.1	

9.3.9.3. Prefix Attributes TLV

This TLV is applicable for LS Prefix Object-Types for both IPv4 and IPv6. Prefixes are learned from the IGP (IS-IS or OSPF) or BGP topology with a set of IGP attributes (such as metric, route tags, etc.). This section describes the different attributes related to the IPv4/IPv6 prefixes. Prefix Attributes TLVs SHOULD be encoded in the LS Prefix Object.



The following 'Prefix Attribute' sub-TLVs are valid :

Sub-TLV	Description	BGP-LS TLV	Defined in:
41	IGP Flags	1152	[RFC7752] /3.3.3.1
42	Route Tag	1153	[RFC7752] /3.3.3.2
43	Extended Tag	1154	[RFC7752] /3.3.3.3
44	Prefix Metric	1155	[RFC7752] /3.3.3.4
45	OSPF Forwarding Address	1156	[RFC7752] /3.3.3.5
46	Opaque Prefix Attribute	1157	[RFC7752] /3.3.3.6

9.3.10. Removal of an Attribute

One of a key objective of PCEP-LS is to encode and carry only the impacted attributes of a Node, a Link or a Prefix. To accommodate this requirement, incase of a removal of an attribute, the sub-TLV MUST be included with no 'value' field and length=0 to indicate that the attribute is removed. On receiving a sub-TLV with zero length, the receiver removes the attribute from the database.

10. Other Considerations

10.1. Inter-AS Links

The main source of LS (and TE) information is the IGP, which is not active on inter-AS links. In some cases, the IGP may have information of inter-AS links ([RFC5392], [RFC5316]). In other cases, an implementation SHOULD provide a means to inject inter-AS links into PCEP. The exact mechanism used to provision the inter-AS links is outside the scope of this document.

11. Security Considerations

This document extends PCEP for LS (and TE) distribution including a new LSRpt message with new object and TLVs. Procedures and protocol extensions defined in this document do not effect the overall PCEP security model. See [RFC5440], [I-D.ietf-pce-pceps]. Tampering with the LSRpt message may have an effect on path computations at PCE. It also provides adversaries an opportunity to eavesdrop and learn sensitive information and plan sophisticated attacks on the network infrastructure. The PCE implementation SHOULD provide mechanisms to prevent strains created by network flaps and amount of LS (and TE) information. Thus it is suggested that any mechanism used for securing the transmission of other PCEP message be applied here as well. As a general precaution, it is RECOMMENDED that these PCEP extensions only be activated on authenticated and encrypted sessions belonging to the same administrative authority.

12. 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.

12.1. Control of Function and Policy

A PCE or PCC implementation MUST allow configuring the PCEP-LS capabilities as described in this document.

A PCC implementation SHOULD allow configuration to suggest if remote information learned via routing protocols should be reported or not.

An implementation SHOULD allow the operator to specify the maximum number of LS data to be reported.

An implementation SHOULD also allow the operator to create abstracted topologies that are reported to the peers and create different abstractions for different peers.

An implementation SHOULD allow the operator to configure a 64-bit Instance-ID for Routing Universe TLV.

12.2. Information and Data Models

An implementation SHOULD allow the operator to view the LS capabilities advertised by each peer. To serve this purpose, the PCEP YANG module [I-D.ietf-pce-pcep-yang]" can be extended to include advertised capabilities.

An implementation SHOULD also provide the statistics:

- Total number of LSRpt sent/received, as well as per neighbor
- Number of error received for LSRpt, per neighbor
- Total number of locally originated Link-State Information

These statistics should be recorded as absolute counts since system or session start time. An implementation MAY also enhance this information by recording peak per-second counts in each case.

An operator SHOULD define an import policy to limit inbound LSRpt to "drop all LSRpt from a particular peers" as well provide means to limit inbound LSRpts.

12.3. 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]".

12.4. Verify Correct Operations

Mechanisms defined in this document do not imply any new operation verification requirements in addition to those already listed in [RFC5440] .

12.5. Requirements On Other Protocols

Mechanisms defined in this document do not imply any new requirements on other protocols.

12.6. 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].

13. IANA Considerations

This document requests IANA actions to allocate code points for the protocol elements defined in this document.

13.1. PCEP Messages

IANA created a registry for PCEP messages. Each PCEP message has a message type value. This document defines a new PCEP message value.

Value	Meaning	Reference
TBD3	LSRpt	[This I-D]

13.2. PCEP Objects

This document defines the following new PCEP Object-classes and Object-values:

Object-Class Value	Name	Reference
TBD6	LS Object Object-Type=1 (LS Node) Object-Type=2 (LS Link) Object-Type=3 (LS IPv4 Prefix) Object-Type=4 (LS IPv6 Prefix)	[This I-D]

13.3. LS Object

This document requests that a new sub-registry, named "LS Object Protocol-ID Field", is created within the "Path Computation Element Protocol (PCEP) Numbers" registry to manage the Flag field of the LSP object. New values are to be assigned by Standards Action [RFC5226].

Value	Meaning	Reference
0	Reserved	[This I-D]
1	IS-IS Level 1	[This I-D]
2	IS-IS Level 2	[This I-D]
3	OSPFv2	[This I-D]
4	Direct	[This I-D]
5	Static configuration	[This I-D]
6	OSPFv3	[This I-D]
7	BGP-LS	[This I-D]
8	PCEP-LS	[This I-D]
9	Abstraction	[This I-D]
10	Unspecified	[This I-D]

Further, this document also requests that a new sub-registry, named "LS Object Flag Field", is created within the "Path Computation Element Protocol (PCEP) Numbers" registry to manage the Flag 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
0-21	Unassigned	
22	R (Remove bit)	[This I-D]
23	S (Sync bit)	[This I-D]

13.4. PCEP-Error Object

IANA is requested to make the following allocation in the "PCEP-ERROR Object Error Types and Values" registry.

Error-Type	Meaning	Reference
6	Mandatory Object missing Error-Value=TBD4 (LS object missing)	[RFC5440] [This I-D]
19	Invalid Operation Error-Value=TBD1 (Attempted LS Report if LS remote capability was not advertised)	[I-D.ietf-pce-stateful-pce] [This I-D]
TBD2	LS Synchronization Error Error-Value=1 (An error in processing the LSRpt) Error-Value=2 (An internal PCC error)	[This I-D]

13.5. PCEP TLV Type Indicators

This document defines the following new PCEP TLVs.

Value	Meaning	Reference
TBD5	LS-CAPABILITY TLV	[This I-D]
TBD7	ROUTING-UNIVERSE TLV	[This I-D]
TBD15	ROUTE-DISTINGUISHER TLV	[This I-D]
TBD8	Local Node Descriptors TLV	[This I-D]
TBD9	Remote Node Descriptors TLV	[This I-D]
TBD10	Link Descriptors TLV	[This I-D]
TBD11	Prefix Descriptors TLV	[This I-D]
TBD12	Node Attributes TLV	[This I-D]
TBD13	Link Attributes TLV	[This I-D]
TBD14	Prefix Attributes TLV	[This I-D]

13.6. PCEP-LS Sub-TLV Type Indicators

This document specifies the PCEP-LS Sub-TLVs. IANA is requested to create an "PCEP-LS Sub-TLV Types" sub-registry in the "PCEP TLV Type Indicators" for the sub-TLVs carried in the PCEP-LS TLV (Local and Remote Node Descriptors TLV, Link Descriptors TLV, Prefix Descriptors TLV, Node Attributes TLV, Link Attributes TLV and Prefix Attributes TLV. This document defines the following types:

Sub-TLV	Description	Ref	Value defined

		Sub-TLV	in:
1	Autonomous System	512	[RFC7752] /3.2.1.4
2	BGP-LS Identifier	513	[RFC7752] /3.2.1.4
3	OSPF Area-ID	514	[RFC7752] /3.2.1.4
4	Router-ID	515	[RFC7752] /3.2.1.4
5	Multi-Topology-ID	263	[RFC7752] /3.2.1.5
6	Link Local/Remote Identifiers	22/4	[RFC5307]/1.1
7	IPv4 interface address	22/6	[RFC5305]/3.2
8	IPv4 neighbor address	22/8	[RFC5305]/3.3
9	IPv6 interface address	22/12	[RFC6119]/4.2
10	IPv6 neighbor address	22/13	[RFC6119]/4.3
11	OSPF Route Type	264	[RFC7752] /3.2.3.1
12	IP Reachability Information	265	[RFC7752] /3.2.3.2
13	Node Flag Bits	1024	[RFC7752] /3.3.1.1
14	Opaque Node Properties	1025	[RFC7752] /3.3.1.5
15	Node Name	1026	[RFC7752] /3.3.1.3
16	IS-IS Area Identifier	1027	[RFC7752] /3.3.1.2
17	IPv4 Router-ID of Local Node	134/--	[RFC5305]/4.3
18	IPv6 Router-ID of Local Node	140/--	[RFC6119]/4.1
19	IPv4 Router-ID of Remote Node	134/--	[RFC5305]/4.3
20	IPv6 Router-ID of Remote Node	140/--	[RFC6119]/4.1
22	Administrative group (color)	22/3	[RFC5305]/3.1
23	Maximum link bandwidth	22/9	[RFC5305]/3.3
24	Max. reservable link bandwidth	22/10	[RFC5305]/3.5

25	Unreserved bandwidth	22/11	[RFC5305]/3.6	
26	TE Default Metric	22/18	[RFC7752] /3.3.2.3	
27	Link Protection Type	22/20	[RFC5307]/1.2	
28	MPLS Protocol Mask	1094	[RFC7752] /3.3.2.2	
29	IGP Metric	1095	[RFC7752] /3.3.2.4	
30	Shared Risk Link Group	1096	[RFC7752] /3.3.2.5	
31	Opaque link attributes	1097	[RFC7752] /3.3.2.6	
32	Link Name attribute	1098	[RFC7752] /3.3.2.7	
33	Unidirectional Link Delay	22/33	[RFC7810]/4.1	
34	Min/Max Unidirectional Link Delay	22/34	[RFC7810]/4.2	
35	Unidirectional Delay Variation	22/35	[RFC7810]/4.3	
36	Unidirectional Link Loss	22/36	[RFC7810]/4.4	
37	Unidirectional Residual Bandwidth	22/37	[RFC7810]/4.5	
38	Unidirectional Available Bandwidth	22/38	[RFC7810]/4.6	
39	Unidirectional Bandwidth Utilization	22/39	[RFC7810]/4.7	
40	Extended Admin Group (EAG)	22/14	[RFC7308]/2.1	
41	IGP Flags	1152	[RFC7752] /3.3.3.1	
42	Route Tag	1153	[RFC7752] /3.3.3.2	
43	Extended Tag	1154	[RFC7752] /3.3.3.3	
44	Prefix Metric	1155	[RFC7752] /3.3.3.4	
45	OSPF Forwarding Address	1156	[RFC7752] /3.3.3.5	
46	Opaque Prefix Attribute	1157	[RFC7752] /3.3.3.6	

New values are to be assigned by Standards Action [RFC5226].

14. TLV/Sub-TLV Code Points Summary

This section contains the global table of all TLVs/Sub-TLVs in LS object defined in this document.

TLV	Description	Ref TLV	Value defined in:
TBD7	Routing Universe	--	Sec 9.2.1
TBD15	Route Distinguisher	--	Sec 9.2.2
*	Virtual Network	--	[leedhody-pce-vn-association]
TBD8	Local Node Descriptors	256	[RFC7752] /3.2.1.2
TBD9	Remote Node Descriptors	257	[RFC7752] /3.2.1.3
TBD10	Link Descriptors	--	Sec 9.2.8
TBD11	Prefix Descriptors	--	Sec 9.2.9
TBD12	Node Attributes	--	Sec 9.2.10.1
TBD13	Link Attributes	--	Sec 9.2.10.2
TBD14	Prefix Attributes	--	Sec 9.2.10.3

* this TLV is defined in a different PCEP document

TLV Table

Refer Section 13.6 for the table of Sub-TLVs.

15. Implementation Status

The PCEP-LS protocol extension as described in this I-D were implemented and tested for a variety of applications. Apart from the below implementation, there exist other experimental implementations done for optical networks.

15.1. Hierarchical Transport PCE controllers

The PCEP-LS has been implemented as part of IETF97 Hackathon and Bits-N-Bites demonstration. The use-case demonstrated was DCI use-case of ACTN architecture in which to show the following scenarios:

- connectivity services on the ACTN based recursive hierarchical SDN/PCE platform that has the three tier level SDN controllers

(two-tier level MDSC and PNC) on the top of the PTN systems managed by EMS.

- Integration test of two tier-level MDSC: The SBI of the low level MDSC is the YANG based Korean national standards and the one of the high level MDSC the PCEP-LS based ACTN protocols.

- Performance test of three types of SDN controller based recovery schemes including protection, reactive and proactive restoration. PCEP-LS protocol was used to demonstrate quick report of failed network components.

15.2. ONOS-based Controller (MDSC and PNC)

Huawei (PNC, MDSC) and SKT (MDSC) implemented PCEP-LS during Hackathon and IETF97 Bits-N-Bites demonstration. The demonstration was ONOS-based ACTN architecture in which to show the following capabilities:

Both packet PNC and optical PNC (with optical PCEP-LS extension) implemented PCEP-LS on its SBI and well as its NBI (towards MDSC).

SKT orchestrator (acting as MDSC) also supported PCEP-LS (as well as RestConf) towards packet and optical PNCs on its SBI.

Further description can be found at <ONOS-PCEP> and the code at <ONOS-PCEP-GITHUB>.

16. Acknowledgments

This document borrows some of the structure and text from the [RFC7752].

Thanks to Eric Wu, Venugopal Kondreddy, Mahendra Singh Negi, Avantika, and Zhengbin Li for the reviews.

Thanks to Ramon Casellas for his comments and suggestions based on his implementation experience.

17. References

17.1. 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, <http://www.rfc-editor.org/info/rfc2119>.

[RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, DOI 10.17487/RFC5305, October 2008, <http://www.rfc-editor.org/info/rfc5305>.

[RFC5307] Kompella, K., Ed. and Y. Rekhter, Ed., "IS-IS Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 5307, DOI 10.17487/RFC5307, October 2008, <http://www.rfc-editor.org/info/rfc5307>.

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

[RFC6119] Harrison, J., Berger, J., and M. Bartlett, "IPv6 Traffic Engineering in IS-IS", RFC 6119, DOI 10.17487/RFC6119, February 2011, <http://www.rfc-editor.org/info/rfc6119>.

[RFC7752] Gredler, H., Ed., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP", RFC 7752, DOI 10.17487/RFC7752, March 2016, <http://www.rfc-editor.org/info/rfc7752>.

[RFC7810] Previdi, S., Ed., Giacalone, S., Ward, D., Drake, J., and Q. Wu, "IS-IS Traffic Engineering (TE) Metric Extensions", RFC 7810, DOI 10.17487/RFC7810, May 2016, <http://www.rfc-editor.org/info/rfc7810>.

17.2. Informative References

[RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, DOI 10.17487/RFC3630, September 2003, <http://www.rfc-editor.org/info/rfc3630>.

[RFC4203] Kompella, K., Ed. and Y. Rekhter, Ed., "OSPF Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4203, DOI 10.17487/RFC4203, October 2005, <http://www.rfc-editor.org/info/rfc4203>.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <http://www.rfc-editor.org/info/rfc4364>.

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

[RFC5120] Przygienda, T., Shen, N., and N. Sheth, "M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate Systems (IS-ISs)", RFC 5120, DOI 10.17487/RFC5120, February 2008, <http://www.rfc-editor.org/info/rfc5120>.

[RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", RFC 5226, DOI 10.17487/RFC5226, May 2008, <http://www.rfc-editor.org/info/rfc5226>.

[RFC5316] Chen, M., Zhang, R., and X. Duan, "ISIS Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 5316, DOI 10.17487/RFC5316, December 2008, <http://www.rfc-editor.org/info/rfc5316>.

[RFC5392] Chen, M., Zhang, R., and X. Duan, "OSPF Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 5392, DOI 10.17487/RFC5392, January 2009, <http://www.rfc-editor.org/info/rfc5392>.

[RFC6549] Lindem, A., Roy, A., and S. Mirtorabi, "OSPFv2 Multi-Instance Extensions", RFC 6549, DOI 10.17487/RFC6549, March 2012, <http://www.rfc-editor.org/info/rfc6549>.

[RFC6805] King, D., Ed. and A. Farrel, Ed., "The Application of the Path Computation Element Architecture to the Determination of a Sequence of Domains in MPLS and GMPLS", RFC 6805, DOI 10.17487/RFC6805, November 2012, <http://www.rfc-editor.org/info/rfc6805>.

[RFC6822] Previdi, S., Ed., Ginsberg, L., Shand, M., Roy, A., and D. Ward, "IS-IS Multi-Instance", RFC 6822, DOI 10.17487/RFC6822, December 2012, <http://www.rfc-editor.org/info/rfc6822>.

[I-D.ietf-pce-stateful-pce] Crabbe, E., Minei, I., Medved, J., and R. Varga, "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce-21 (work in progress), June 2017.

[I-D.ietf-pce-pce-initiated-lsp]
Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model", draft-ietf-pce-pce-initiated-lsp-10 (work in progress), June 2017.

[I-D.ietf-pce-pceps]
Lopez, D., Dios, O., Wu, Q., and D. Dhody, "Secure Transport for PCEP", draft-ietf-pce-pceps-14 (work in progress), May 2017.

[I-D.ietf-pce-pcep-yang]
Dhody, D., Hardwick, J., Beeram, V., and j. jefftant@gmail.com, "A YANG Data Model for Path Computation Element Communications Protocol (PCEP)", draft-ietf-pce-pcep-yang-02 (work in progress), March 2017.

[I-D.ietf-pce-applicability-actn]
Dhody, D., Lee, Y., and D. Ceccarelli, "Applicability of Path Computation Element (PCE) for Abstraction and Control of TE Networks (ACTN)", draft-ietf-pce-applicability-actn-00 (work in progress), June 2017.

[I-D.ietf-teas-actn-framework]
Ceccarelli, D. and Y. Lee, "Framework for Abstraction and Control of Traffic Engineered Networks", draft-ietf-teas-actn-framework-06 (work in progress), June 2017.

[I-D.ietf-teas-actn-requirements]
Lee, Y., Dhody, D., Belotti, S., Pithewan, K., Ceccarelli, D., Miyasaka, T., and J. Shin, "Requirements for Abstraction and Control of TE Networks", draft-ietf-teas-actn-requirements-05 (work in progress), May 2017.

[I-D.kondreddy-pce-pcep-ls-sync-optimizations]
Kondreddy, V. and M. Negi, "Optimizations of PCEP Link-State(LS) Synchronization Procedures", draft-kondreddy-pce-pcep-ls-sync-optimizations-00 (work in progress), October 2015.

[I-D.leedhody-pce-vn-association]
Lee, Y., Dhody, D., Zhang, X., and D. Ceccarelli, "PCEP Extensions for Establishing Relationships Between Sets of LSPs and Virtual Networks", draft-leedhody-pce-vn-association-02 (work in progress), March 2017.

[ONOS-PCEP]
"Support for PCEP in ONOS",
<https://wiki.onosproject.org/display/ONOS/PCEP+Protocol>.

[ONOS-PCEP-GITHUB]
"Github for PCEP code in ONOS",
<https://github.com/opennetworkinglab/onos/tree/master/
protocols/pcep>.

Appendix A. Relevant OSPF TLV and sub-TLV

This section list the relevant TLVs and sub-TLVs defined for OSPF.

Sub-TLV	Description	OSPF-TE Sub-TLV	Value defined in:
6	Link Local/Remote Identifiers	11	[RFC4203]/1.1
7	IPv4 interface address	3	[RFC3630]/2.5.3
8	IPv4 neighbor address	4	[RFC3630]/2.5.4
9	IPv6 interface address	19	[RFC5329]/4.3
10	IPv6 neighbor address	20	[RFC5329]/4.4
17	IPv4 Router-ID of Local Node	1	[RFC3630]/2.4.1
18	IPv6 Router-ID of Local Node	3	[RFC5329]/3
19	IPv4 Router-ID of Remote Node	1	[RFC3630]/2.4.1
20	IPv6 Router-ID of Remote Node	3	[RFC5329]/3
22	Administrative group (color)	9	[RFC3630]/2.5.9
23	Maximum link bandwidth	6	[RFC3630]/2.5.6
24	Max. reservable link bandwidth	7	[RFC3630]/2.5.7
25	Unreserved bandwidth	8	[RFC3630]/2.5.8
27	Link Protection Type	14	[RFC4203]/1.2
30	Shared Risk Link Group	16	[RFC4203]/1.3
33	Unidirectional Link Delay	27	[RFC7471]/4.1
34	Min/Max Unidirectional Link Delay	28	[RFC7471]/4.2
35	Unidirectional Delay Variation	29	[RFC7471]/4.3
36	Unidirectional Link Loss	30	[RFC7471]/4.4
37	Unidirectional	31	[RFC7471]/4.5

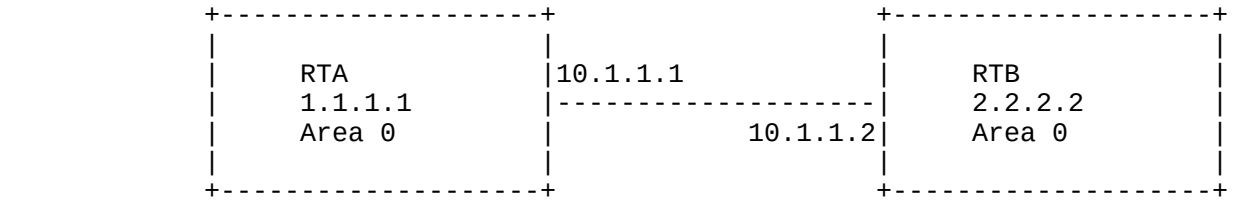
	Residual Bandwidth			
38	Unidirectional Available Bandwidth	32	[RFC7471]/4.6	
39	Unidirectional Bandwidth Utilization	33	[RFC7471]/4.7	
40	Extended Admin Group (EAG)	26	[RFC7308]/2.1	

Appendix B. Examples

These examples are for illustration purposes only to show how the new PCEP-LS message could be encoded. They are not meant to be an exhaustive list of all possible use cases and combinations.

B.1. All Nodes

Each node (PCC) in the network chooses to provide its own local node and link information, and in this way PCE can build the full link state and TE information.




```
RTA
---
LS Node
   TLV - Local Node Descriptors
       Sub-TLV - 3: OSPF Area-ID: 0.0.0.0
       Sub-TLV - 4: Router-ID: 1.1.1.1
   TLV - Node Attributes TLV
       Sub-TLV(s)


LS Link
   TLV - Local Node Descriptors
       Sub-TLV - 3: OSPF Area-ID: 0.0.0.0
       Sub-TLV - 4: Router-ID: 1.1.1.1
   TLV - Remote Node Descriptors
       Sub-TLV - 3: OSPF Area-ID: 0.0.0.0
       Sub-TLV - 4: Router-ID: 2.2.2.2
   TLV - Link Descriptors
       Sub-TLV - 7: IPv4 interface: 10.1.1.1
```

```
          Sub-TLV - 8: IPv4 neighbor: 10.1.1.2
       TLV - Link Attributes TLV
          Sub-TLV(s)

   RTB
   ---
   LS Node
       TLV - Local Node Descriptors
          Sub-TLV - 3: OSPF Area-ID: 0.0.0.0
          Sub-TLV - 4: Router-ID: 2.2.2.2
       TLV - Node Attributes TLV
          Sub-TLV(s)

   LS Link
       TLV - Local Node Descriptors
          Sub-TLV - 3: OSPF Area-ID: 0.0.0.0
          Sub-TLV - 4: Router-ID: 2.2.2.2
       TLV - Remote Node Descriptors
          Sub-TLV - 3: OSPF Area-ID: 0.0.0.0
          Sub-TLV - 4: Router-ID: 1.1.1.1
       TLV - Link Descriptors
          Sub-TLV - 7: IPv4 interface: 10.1.1.2
          Sub-TLV - 8: IPv4 neighbor: 10.1.1.1
       TLV - Link Attributes TLV
          Sub-TLV(s)
```

B.2. Designated Node

A designated node(s) in the network will provide its own local node as well as all learned remote information, and in this way PCE can build the full link state and TE information.

As described in Appendix B.1, the same LS Node and Link objects will be generated with a difference that it would be a designated router say RTA that generate all this information.

B.3. Between PCEs

As per Hierarchical-PCE [RFC6805], Parent PCE builds an abstract domain topology map with each domain as an abstract node and inter-domain links as an abstract link. Each child PCE may provide this information to the parent PCE. Considering the example in figure 1 of [RFC6805], following LS object will be generated:

```
   PCE1
   ----
   LS Node
      TLV - Local Node Descriptors
          Sub-TLV - 1: Autonomous System: 100 (Domain 1)
          Sub-TLV - 4: Router-ID: 11.11.11.11 (abstract)

   LS Link
      TLV - Local Node Descriptors
          Sub-TLV - 1: Autonomous System: 100
          Sub-TLV - 4: Router-ID: 11.11.11.11 (abstract)
      TLV - Remote Node Descriptors
          Sub-TLV - 1: Autonomous System: 200 (Domain 2)
          Sub-TLV - 4: Router-ID: 22.22.22.22 (abstract)
      TLV - Link Descriptors
          Sub-TLV - 7: IPv4 interface: 11.1.1.1
          Sub-TLV - 8: IPv4 neighbor: 11.1.1.2
      TLV - Link Attributes TLV
          Sub-TLV(s)

   LS Link
      TLV - Local Node Descriptors
          Sub-TLV - 1: Autonomous System: 100
          Sub-TLV - 4: Router-ID: 11.11.11.11 (abstract)
      TLV - Remote Node Descriptors
          Sub-TLV - 1: Autonomous System: 200
          Sub-TLV - 4: Router-ID: 22.22.22.22 (abstract)
      TLV - Link Descriptors
          Sub-TLV - 7: IPv4 interface: 12.1.1.1
          Sub-TLV - 8: IPv4 neighbor: 12.1.1.2
      TLV - Link Attributes TLV
          Sub-TLV(s)

   LS Link
      TLV - Local Node Descriptors
          Sub-TLV - 1: Autonomous System: 100
          Sub-TLV - 4: Router-ID: 11.11.11.11 (abstract)
      TLV - Remote Node Descriptors
          Sub-TLV - 1: Autonomous System: 400 (Domain 4)
          Sub-TLV - 4: Router-ID: 44.44.44.44 (abstract)
      TLV - Link Descriptors
          Sub-TLV - 7: IPv4 interface: 13.1.1.1
          Sub-TLV - 8: IPv4 neighbor: 13.1.1.2
      TLV - Link Attributes TLV
          Sub-TLV(s)

   * similar information will be generated by other PCE
     to help form the abstract domain topology.
```

Further the exact border nodes and abstract internal path between the border nodes may also be transported to the Parent PCE to enable ACTN as described in [I-D.ietf-pce-applicability-actn] using the similar LS node and link objects encodings.

Appendix C. Contributor Addresses

Udayasree Palle
Huawei Technologies
Divyashree Techno Park, Whitefield
Bangalore, Karnataka 560066
India

EMail: udayasreereddy@gmail.com

Sergio Belotti
Alcatel-Lucent
Italy

EMail: sergio.belotti@alcatel-lucent.com

Veerendranatha Reddy Vallem
Huawei Technologies
Divyashree Techno Park, Whitefield
Bangalore, Karnataka 560066
India

Email: veerendranatharv@huawei.com

Satish Karunanithi
Huawei Technologies
Divyashree Techno Park, Whitefield
Bangalore, Karnataka 560066
India

Email: satishk@huawei.com

Authors' Addresses

Dhruv Dhody
Huawei Technologies
Divyashree Techno Park, Whitefield
Bangalore, Karnataka 560066
India

EMail: dhruv.ietf@gmail.com

Young Lee
Huawei Technologies
5340 Legacy Drive, Building 3
Plano, TX 75023
USA

EMail: leeyoung@huawei.com

Daniele Ceccarelli
Ericsson
Torshamnsgatan,48
Stockholm
Sweden

EMail: daniele.ceccarelli@ericsson.com

PCE Working Group
Internet-Draft
Intended status: Informational
Expires: September 14, 2017

D. Dhody
Y. Lee
Huawei Technologies
D. Ceccarelli
Ericsson
J. Shin
SK Telecom
D. King
Lancaster University
O. Gonzalez de Dios
Telefonica I+D

March 13, 2017

Hierarchical Stateful Path Computation Element (PCE). draft-dhodylee-pce-stateful-hpce-03

Abstract

A Stateful Path Computation Element (PCE) maintains information on the current network state, including: computed Label Switched Path (LSPs), reserved resources within the network, and pending path computation requests. This information may then be considered when computing new traffic engineered LSPs, and for associated and dependent LSPs, received from Path Computation Clients (PCCs).

The Hierarchical Path Computation Element (H-PCE) architecture, provides an architecture to allow the optimum sequence of inter-connected domains to be selected, and network policy to be applied if applicable, via the use of a hierarchical relationship between PCEs.

Combining the capabilities of Stateful PCE and the Hierarchical PCE would be advantageous. This document describes general considerations and use cases for the deployment of Stateful PCE(s) using the Hierarchical PCE architecture.

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."

Table of Contents

1. Introduction

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.

A stateful PCE is capable of considering, for the purposes of path computation, not only the network state in terms of links and nodes (referred to as the Traffic Engineering Database or TED) but also the status of active services (previously computed paths, and currently reserved resources, stored in the Label Switched Paths Database (LSPDB).

[RFC8051] describes general considerations for a stateful PCE deployment and examines its applicability and benefits, as well as its challenges and limitations through a number of use cases.

[I-D.ietf-pce-stateful-pce] describes a set of extensions to PCEP to provide stateful control. A stateful PCE has access to not only the information carried by the network's Interior Gateway Protocol (IGP), but also the set of active paths and their reserved resources for its computations. The additional state allows the PCE to compute constrained paths while considering individual LSPs and their interactions. [I-D.ietf-pce-pce-initiated-lsp] describes the setup, maintenance and teardown of PCE-initiated LSPs under the stateful PCE model.

[I-D.ietf-pce-stateful-pce] also describes the active stateful PCE. The active PCE functionality allows a PCE to reroute an existing LSP or make changes to the attributes of an existing LSP, or delegate control of specific LSPs to a new PCE.

The ability to compute shortest constrained TE LSPs in Multiprotocol Label Switching (MPLS) and Generalized MPLS (GMPLS) networks across multiple domains has been identified as a key motivation for PCE development. [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 (H-PCE) architecture [RFC6805], the Parent PCE (P-PCE) is used to compute a multi-domain path based on the domain connectivity information. A Child PCE (C-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.

This document presents general considerations for stateful PCE(s) in hierarchical PCE architecture. In particular, the behavior changes

and additions to the existing stateful PCE mechanisms (including PCE-initiated LSP setup and active PCE usage) in the context of networks using the H-PCE architecture.

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 terminology is as per [RFC4655], [RFC5440], [RFC6805], and [I-D.ietf-pce-stateful-pce].

3. Hierarchical Stateful PCE

As described in [RFC6805], in the hierarchical PCE architecture, a P-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 P-PCE has no information about the content of the child domains. Each child domain has at least one PCE capable of computing paths across the domain. These PCEs are known as C-PCEs and have a direct relationship with the P-PCE. The P-PCE builds the domain topology map either via direct configuration (allowing network policy to also be applied) or from learned information received from each C-PCE.

[I-D.ietf-pce-stateful-pce] specifies new functions to support a stateful PCE. It also specifies that a function can be initiated either from a PCC towards a PCE (C-E) or from a PCE towards a PCC (E-C).

This document extends these functions to support H-PCE Architecture from a C-PCE towards a P-PCE (CE-PE) or from a P-PCE towards a C-PCE (PE-CE). All PCE types herein (i.e., PE or CE) are assumed to be ’stateful PCE’.

A number of interactions are expected in the Hierarchical Stateful PCE architecture, these include:

LSP State Report (CE-PE): a child stateful PCE sends an LSP state report to a Parent Stateful PCE whenever the state of a LSP changes.

LSP State Synchronization (CE-PE): after the session between the Child and Parent stateful PCEs is initialized, the P-PCE must learn the state of C-PCE’s TE LSPs.

LSP Control Delegation (CE-PE,PE-CE): a C-PCE grants to the P-PCE the right to update LSP attributes on one or more LSPs; the C-PCE may withdraw the delegation or the P-PCE may give up the delegation at any time.

LSP Update Request (PE-CE): a stateful P-PCE requests modification of attributes on a C-PCE's TE LSP.

PCE LSP Initiation Request (PE-CE): a stateful P-PCE requests C-PCE to initiate a TE LSP.

Note that this hierarchy is recursive and thus a LSR could delegate the control to a PCE, which may delegate to its parent, which may further delegate it to its parent (if it exist or needed). Similarly update operations could also be applied recursively.

[I-D.ietf-pce-hierarchy-extensions] defines the H-PCE capability TLV that should be used in the OPEN message to advertise the H-PCE capability. [I-D.ietf-pce-stateful-pce] defines the stateful PCE capability TLV. The presence of both TLVs represent the support for stateful H-PCE operations as described in this document.

[I-D.litkowski-pce-state-sync] describes the procedures to allow a stateful communication between PCEs for various use-cases. The procedures and extensions as described in Section 3 of [I-D.litkowski-pce-state-sync] are also applicable to Child and Parent PCE communication.

3.1. Passive Operations

Procedures as described in [RFC6805] are applied, where the ingress C-PCE sends a request to the P-PCE. The P-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 C-PCEs responsible for each of the domains on the candidate domain paths. Each C-PCE computes a set of candidate path segments across its domain and sends the results to the P-PCE. The P-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 C-PCE that received the initial path request, and this C-PCE passes the path on to the PCC that issued the original request.

As per [I-D.ietf-pce-stateful-pce], PCC sends an LSP State Report carried on a PCRpt message to the C-PCE, indicating the LSP's status. The C-PCE MAY further propagate the State Report to the P-PCE. A local policy at C-PCE MAY dictate which LSPs to be reported to the P-PCE. The PCRpt message is sent from C-PCE to P-PCE.

State synchronization mechanism as described in [I-D.ietf-pce-stateful-pce] and [I-D.ietf-pce-stateful-sync-optimizations] are applicable to PCEP session between C-PCE and P-PCE as well.

Taking the sample hierarchical domain topology example from [RFC6805] as the reference topology for the entirety of this document.

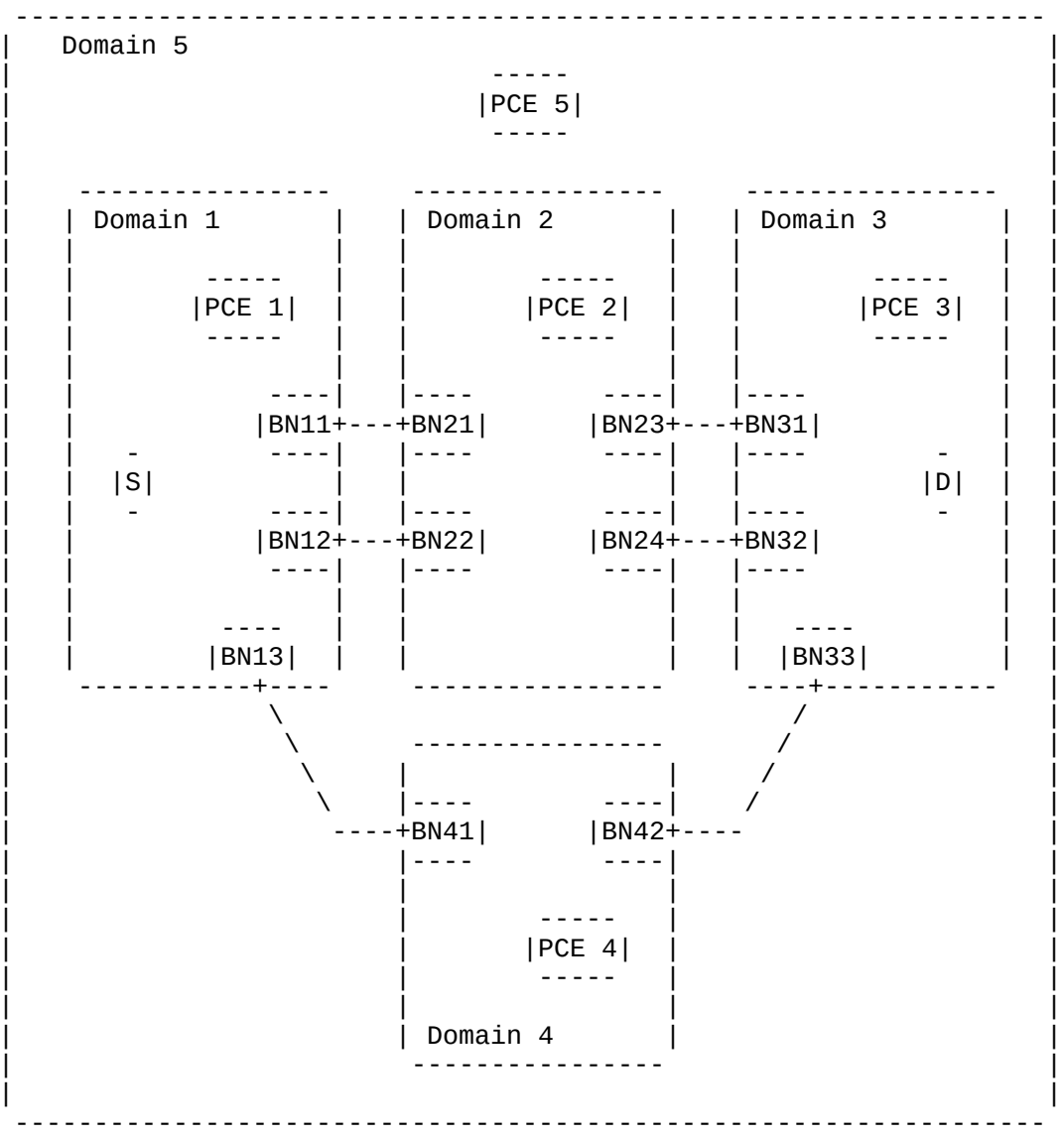


Figure 1: Sample Hierarchical Domain Topology

Steps 1 to 11 are exactly as described in section 4.6.2 (Hierarchical PCE End-to-End Path Computation Procedure) of [RFC6805], the following additional steps are added for stateful PCE:

(1) The Ingress LSR initiates the setup of the LSP as per the path and reports to the PCE1 the LSP status ("GOING-UP").

(2) The PCE1 further reports the status of the LSP to the P-PCE (PCE5).

(3) The Ingress LSR notifies the LSP state to PCE1 when the state is "UP".

(4) The PCE1 further reports the status of the LSP to the P-PCE (PCE5).

3.2. Active Operations

[I-D.ietf-pce-stateful-pce] describes the case of active stateful PCE. The active PCE functionality uses two specific PCEP messages:

o Update Request (PCUpd)
o State Report (PCRpt)

The first is sent by the PCE to a Path Computation Client (PCC) for modifying LSP attributes. The PCC sends back a PCRpt to acknowledge the requested operation or report any change in LSP's state.

As per [RFC8051], Delegation is an operation to grant a PCE, temporary rights to modify a subset of LSP parameters on one or more PCC's LSPs. The C-PCE may further choose to delegate to P-PCE based on a local policy. The PCRpt message with "D" (delegate) flag is sent from C-PCE to P-PCE.

To update an LSP, a PCE send to the PCC, an LSP Update Request using a PCUpd message. For LSP delegated to the P-PCE via the child PCE, the P-PCE can use the same PCUpd message to request change to the C-PCE (the Ingress domain PCE), the PCE further propagates the update request to the PCC.

The P-PCE uses the same mechanism described in Section 3.1 to compute the end to end path using PCReq and PCRep messages.

The following additional steps are also initially performed, for active operations, again using the reference architecture described in Figure 1 (Sample Hierarchical Domain Topology).

(1) The Ingress LSR delegates the LSP to the PCE1 via PCRpt message with D flag set.

(2) The PCE1 further delegates the LSP to the P-PCE (PCE5).

Steps 4 to 10 of section 4.6.2 of [RFC6805] are executed to determine the end to end path.

(3) The P-PCE (PCE5) sends the update request to the C-PCE (PCE1) via PCUpd message.

(4) The PCE1 further updates the LSP to the Ingress LSR (PCC).

(5) The Ingress LSR initiates the setup of the LSP as per the path and reports to the PCE1 the LSP status ("GOING-UP").

(6) The PCE1 further reports the status of the LSP to the P-PCE (PCE5).

(7) The Ingress LSR notifies the LSP state to PCE1 when the state is "UP".

(8) The PCE1 further reports the status of the LSP to the P-PCE (PCE5).

3.3. PCE Initiation Operation

[I-D.ietf-pce-pce-initiated-lsp] describes the setup, maintenance and teardown of PCE-initiated LSPs under the stateful PCE model, without the need for local configuration on the PCC, thus allowing for a dynamic network that is centrally controlled and deployed. To instantiate or delete an LSP, the PCE sends the Path Computation LSP Initiate Request (PCInitiate) message to the PCC. In case of inter-domain LSP in Hierarchical PCE architecture, the initiation operations can be carried out at the P-PCE. In which case after P-PCE finishes the E2E path computation, it can send the PCInitiate message to the C-PCE (the Ingress domain PCE), the PCE further propagates the initiate request to the PCC.

The following additional steps are also initially performed, for PCE initiated operations, again using the reference architecture described in Figure 1 (Sample Hierarchical Domain Topology):

(1) The P-PCE (PCE5) is requested to initiate a LSP.

Steps 4 to 10 of section 4.6.2 of [RFC6805] are executed to determine the end to end path.

(2) The P-PCE (PCE5) sends the initiate request to the child PCE (PCE1) via PCInitiate message.

(3) The PCE1 further propagates the initiate message to the Ingress LSR (PCC).

(4) The Ingress LSR initiates the setup of the LSP as per the path

and reports to the PCE1 the LSP status ("GOING-UP").

(5) The PCE1 further reports the status of the LSP to the P-PCE (PCE5).

(6) The Ingress LSR notifies the LSP state to PCE1 when the state is "UP".

(7) The PCE1 further reports the status of the LSP to the P-PCE (PCE5).

3.3.1. Per Domain Stitched LSP

The hierarchical PCE architecture as per [RFC6805] is primarily used for E2E LSP. With PCE-Initiated capability, another mode of operation is possible, where multiple intra-domain LSPs are initiated in each domain which are further stitched to form an E2E LSP. The P-PCE sends PCInitiate message to each C-PCE separately to initiate individual LSP segments along the domain path. These individual per domain LSP are stitched together by some mechanism, which is out of scope of this document. The P-PCE may also send the PCInitiate message to the ingress C-PCE to initiate the E2E LSP separately.

The following additional steps are also initially performed, for the Per Domain stiched LSP operation, again using the reference architecture described in Figure 1 (Sample Hierarchical Domain Topology):

(1) The P-PCE (PCE5) is requested to initiate a LSP.

Steps 4 to 10 of section 4.6.2 of [RFC6805] are executed to determine the end to end path, which are broken into per-domain LSPs say -

- S-BN41
- BN41-BN33
- BN33-D

It should be noted that the P-PCE MAY use other mechanisms to determine the suitable per-domain LSPs (apart from [RFC6805]).

For LSP (BN33-D)

(2) The P-PCE (PCE5) sends the initiate request to the child PCE (PCE3) via PCInitiate message for LSP (BN33-D).

(3) The PCE3 further propagates the initiate message to BN33.

(4) BN33 initiates the setup of the LSP as per the path and reports to the PCE3 the LSP status ("GOING-UP").

(5) The PCE3 further reports the status of the LSP to the P-PCE (PCE5).

(6) The node BN33 notifies the LSP state to PCE3 when the state is "UP".

(7) The PCE3 further reports the status of the LSP to the P-PCE (PCE5).

For LSP (BN41-BN33)

(8) The P-PCE (PCE5) sends the initiate request to the child PCE (PCE4) via PCInitiate message for LSP (BN41-BN33).

(9) The PCE4 further propagates the initiate message to BN41.

(10) BN41 initiates the setup of the LSP as per the path and reports to the PCE4 the LSP status ("GOING-UP").

(11) The PCE4 further reports the status of the LSP to the P-PCE (PCE5).

(l2) The node BN41 notifies the LSP state to PCE4 when the state is "UP".

(13) The PCE4 further reports the status of the LSP to the P-PCE (PCE5).

For LSP (S-BN41)

(14) The P-PCE (PCE5) sends the initiate request to the child PCE (PCE1) via PCInitiate message for LSP (S-BN41).

(15) The PCE1 further propagates the initiate message to node S.

(16) S initiates the setup of the LSP as per the path and reports to the PCE1 the LSP status ("GOING-UP").

(17) The PCE1 further reports the status of the LSP to the P-PCE (PCE5).

(18) The node S notifies the LSP state to PCE1 when the state is "UP".

(19) The PCE1 further reports the status of the LSP to the P-PCE (PCE5).

Additionally:

(20) Once P-PCE receives report of each per-domain LSP, it should use some stitching mechanism, which is out of scope of this document. In this step, P-PCE (PCE5) could also initiate an E2E LSP (S-D) by sending the PCInitiate message to Ingress C-PCE (PCE1).

4. Other Considerations

4.1. Applicability to Inter-Layer

[RFC5623] describes a framework for applying the PCE-based architecture to inter-layer (G)MPLS traffic engineering. The H-PCE Stateful architecture with stateful P-PCE coordinating with the stateful C-PCEs of higher and lower layer is shown in the figure below.

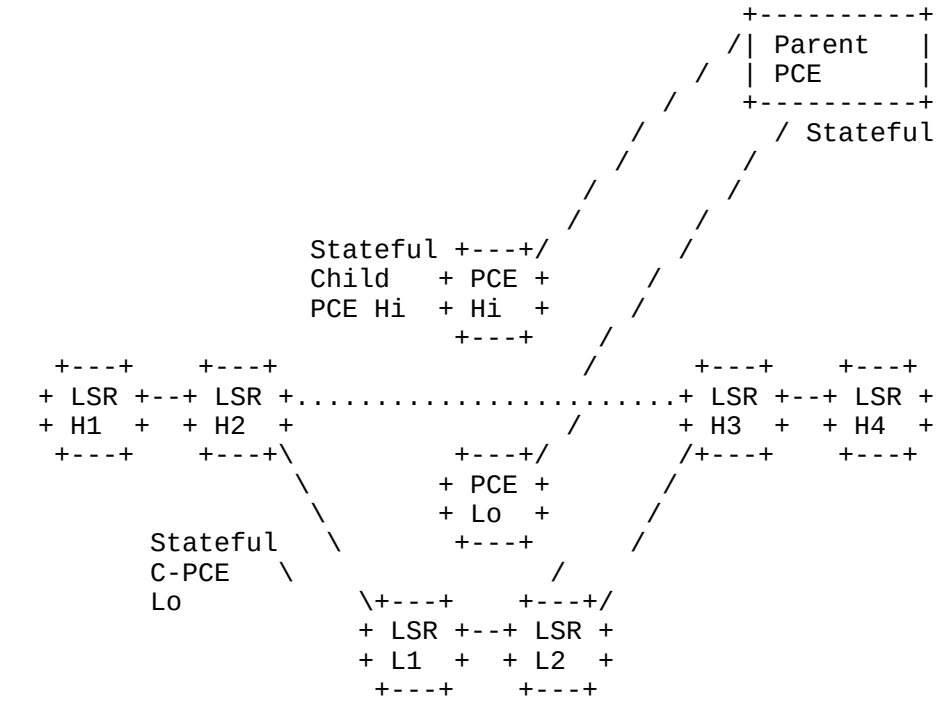


Figure 2: Sample Inter-Layer Topology

All procedures described in Section 3 are applicable to inter-layer path setup as well.

4.2. Applicability to ACTN

[I-D.ietf-teas-actn-framework] describes framework for Abstraction and Control of TE Networks (ACTN), where each Physical Network Controller (PNC) is equivalent to C-PCE and P-PCE is the Multi-Domain Service Coordinator (MDSC). The Per domain stitched LSP as per the Hierarchical PCE architecture described in Section 3.3.1 and Section 4.1 is well suited for ACTN.

[I-D.dhody-pce-applicability-actn] examines the applicability of PCE to the ACTN framework. To support the function of multi domain coordination via hierarchy, the stateful hierarchy of PCEs plays a crucial role.

In ACTN framework, Customer Network Controller (CNC) can request the MDSC to check if there is a possibility to meet Virtual Network (VN) requirements (before requesting for VN provision). The H-PCE architecture as described in [RFC6805] can supports via the use of PCReq and PCRep messages between the P-PCE and C-PCEs.

5. Scalability Considerations

It should be noted that if all the C-PCEs would report all the LSPs in their domain, it could lead to scalability issues for the P-PCE. Thus it is recommended to only report the LSPs which are involved in H-PCE, i.e. the LSPs which are either delegated to the P-PCE or initiated by the P-PCE. Scalability considerations for PCEP as per [I-D.ietf-pce-stateful-pce] continue to apply for the PCEP session between child and parent PCE.

6. Security Considerations

The security considerations listed in [I-D.ietf-pce-stateful-pce],[RFC6805] and [RFC5440] apply to this document as well. As per [RFC6805], it is expected that the parent PCE will require all child PCEs to use full security when communicating with the parent.

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. The LSP state in the PCRpt message SHOULD continue to use this.

The security consideration for PCE-Initiated LSP as per [I-D.ietf-pce-pce-initiated-lsp] is also applicable from P-PCE to C-PCE.

Thus securing the PCEP session (between the P-PCE and the C-PCE) using mechanism like TCP Authentication Option (TCP-AO) [RFC5925] or Transport Layer Security (TLS) [I-D.ietf-pce-pceps] is RECOMMENDED.

7. Manageability Considerations

All manageability requirements and considerations listed in [RFC5440], [RFC6805], [I-D.ietf-pce-stateful-pce], and [I-D.ietf-pce-pce-initiated-lsp] apply to Stateful H-PCE defined in this document. In addition, requirements and considerations listed in this section apply.

7.1. Control of Function and Policy

Support of the hierarchical procedure will be controlled by the management organization responsible for each child PCE. The parent PCE must only accept path computation requests from authorized child PCEs. If a parent PCE receives report from an unauthorized child PCE, the report should be dropped. All mechanism as described in [I-D.ietf-pce-stateful-pce] and [I-D.ietf-pce-pce-initiated-lsp] continue to apply.

7.2. Information and Data Models

An implementation SHOULD allow the operator to view the stateful and H-PCE capabilities advertised by each peer. The PCEP YANG module [I-D.ietf-pce-pcep-yang] can be extended to include details stateful H-PCE.

7.3. 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].

7.4. Verify Correct Operations

Mechanisms defined in this document do not imply any new operation verification requirements in addition to those already listed in [RFC5440] and [I-D.ietf-pce-stateful-pce].

7.5. Requirements On Other Protocols

Mechanisms defined in this document do not imply any new requirements on other protocols.

7.6. Impact On Network Operations

Mechanisms defined in [RFC5440] and [I-D.ietf-pce-stateful-pce] also apply to PCEP extensions defined in this document.

The stateful H-PCE technique brings the applicability of stateful PCE as described in [RFC8051], for the LSP traversing multiple domains.

8. IANA Considerations

There are no IANA considerations.

9. Acknowledgments

Thanks to Manuela Scarella, Haomian Zheng, Sergio Marmo, Stefano Parodi, Giacomo Agostini, Jeff Tantsura and Rajan Rao for suggestions.

10. References

10.1. 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, <http://www.rfc-editor.org/info/rfc2119>.

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

[RFC6805] King, D., Ed. and A. Farrel, Ed., "The Application of the Path Computation Element Architecture to the Determination of a Sequence of Domains in MPLS and GMPLS", RFC 6805, DOI 10.17487/RFC6805, November 2012, <http://www.rfc-editor.org/info/rfc6805>.

[I-D.ietf-pce-stateful-pce] Crabbe, E., Minei, I., Medved, J., and R. Varga, "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce-18 (work in progress), December 2016.

[I-D.ietf-pce-pce-initiated-lsp]
Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model", draft-ietf-pce-pce-initiated-lsp-09 (work in progress), March 2017.

10.2. 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, <http://www.rfc-editor.org/info/rfc4655>.

[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, <http://www.rfc-editor.org/info/rfc5520>.

[RFC5623] Oki, E., Takeda, T., Le Roux, JL., and A. Farrel, "Framework for PCE-Based Inter-Layer MPLS and GMPLS Traffic Engineering", RFC 5623, DOI 10.17487/RFC5623, September 2009, <http://www.rfc-editor.org/info/rfc5623>.

[RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, DOI 10.17487/RFC5925, June 2010, <http://www.rfc-editor.org/info/rfc5925>.

[RFC8051] Zhang, X., Ed. and I. Minei, Ed., "Applicability of a Stateful Path Computation Element (PCE)", RFC 8051, DOI 10.17487/RFC8051, January 2017, <http://www.rfc-editor.org/info/rfc8051>.

[I-D.ietf-pce-stateful-sync-optimizations]
Crabbe, E., Minei, I., Medved, J., Varga, R., Zhang, X., and D. Dhody, "Optimizations of Label Switched Path State Synchronization Procedures for a Stateful PCE", draft-ietf-pce-stateful-sync-optimizations-09 (work in progress), February 2017.

[I-D.ietf-teas-actn-framework]
Ceccarelli D. and Y. Lee, "Framework for Abstraction and Control of Transport Networks", draft-ietf-teas-actn-framework-04 (work in progress), February 2017.

[I-D.dhody-pce-applicability-actn]
Dhody, D., Lee, Y., and D. Ceccarelli, "Applicability of Path Computation Element (PCE) for Abstraction and

Control of TE Networks (ACTN)", draft-dhody-pce-applicability-actn-01 (work in progress), October 2016.

[I-D.litkowski-pce-state-sync]
Litkowski, S., Sivabalan, S., and D. Dhody, "Inter Stateful Path Computation Element communication procedures", draft-litkowski-pce-state-sync-01 (work in progress), February 2017.

[I-D.ietf-pce-hierarchy-extensions]
Zhang, F., Zhao, Q., Dios, O., Casellas, R., and D. King, "Extensions to Path Computation Element Communication Protocol (PCEP) for Hierarchical Path Computation Elements (PCE)", draft-ietf-pce-hierarchy-extensions-03 (work in progress), July 2016.

[I-D.ietf-pce-pceps]
Lopez, D., Dios, O., Wu, W., and D. Dhody, "Secure Transport for PCEP", draft-ietf-pce-pceps-11 (work in progress), January 2017.

Appendix A. Contributor Addresses

Avantika
Huawei Technologies
Divyashree Techno Park, Whitefield
Bangalore, Karnataka 560066
India

EMail: avantika.sushilkumar@huawei.com

Xian Zhang
Huawei Technologies
Bantian, Longgang District
Shenzhen, Guangdong 518129
P.R.China

EMail: zhang.xian@huawei.com

Udayasree Palle
Huawei Technologies
Divyashree Techno Park, Whitefield
Bangalore, Karnataka 560066
India

EMail: udayasree.palle@huawei.com

Authors' Addresses

Dhruv Dhody
Huawei Technologies
Divyashree Techno Park, Whitefield
Bangalore, Karnataka 560066
India

EMail: dhruv.ietf@gmail.com

Young Lee
Huawei Technologies
5340 Legacy Drive, Building 3
Plano, TX 75023
USA

EMail: leeyoung@huawei.com

Daniele Ceccarelli
Ericsson

Torshamnsgatan,48
Stockholm
Sweden

EMail: daniele.ceccarelli@ericsson.com

Jongyoon Shin
SK Telecom
6 Hwangsaeul-ro, 258 beon-gil, Bundang-gu, Seongnam-si,
Gyeonggi-do 463-784
Republic of Korea

EMail: jongyoon.shin@sk.com

Dan King
Lancaster University
UK

EMail: d.king@lancaster.ac.uk

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

Phone: +34913128832
Email: ogondio@tid.es

Path Computation Element Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 14, 2017

O. Dugeon
J. Meuric
Orange

March 13, 2017

A Backward Recursive PCE-initiated inter-domain LSP Setup
draft-dugeon-brpc-stateful-00

Abstract

The Path Computation Element (PCE) working group (WG) has produced a set of RFCs to standardize the behavior of the Path Computation Element as a tool to help MPLS-TE, GMPLS LSP tunnels and Segment Routing paths placement. This also include the ability to compute inter-domain LSPs or Segment Routing path following a distributed or hierarchical approach. In complement to the original stateless mode, a stateful mode has been added. In particular, the new PCInitiate message allows a PCE to directly ask a PCC to setup an MPLS-TE, GMPLS LSP tunnels or a Segment Routing path. However, once computed, the inter-domain LSPs or Segment Routing path are hard to setup in the underlying network. Especially, in operational network, RSVP-TE signaling is not enable between BGP border routers. But, such RSVP-TE signaling is mandatory to setup contiguous LSP tunnels or to stitch or nest independent LSP tunnels to form the end-to-end inter-domain LSP tunnels. This draft propose to combine a Backward Recursive method with PCInitiate message to setup independent LSP tunnels per domain and stitch or nest the different LSP tunnels to setup end-to-end inter-domain LSP tunnels without the need of inter-domain signaling between BGP border routers. A new Stitching Label definition and new LSP-TYPE code points are proposed for that purpose.

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].

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/.

Copyright Notice

Table of Contents

1. Problem Statement

Looking to the different RFCs that describe the PCE architecture and in particular PCE based architecture [RFC4655], PCE protocol [RFC5440], BRPC [RFC5441] and H-PCE [RFC6805], the Path Computation Element (PCE) is able to compute inter-domain path in complement to intra-domain computation. Such inter-domain paths could then serve as the Explicit Route Object input for the RSVP-TE signaling to setup the LSPs tunnel within the underlying network. Three sort of end-to-end LSP tunnels could be established:

- Contiguous tunnels: The RSVP-TE signaling crosses the boundary between two domains e.g. between two AS Border Routers (ASBR) like if it is two routers of the same domain. This kind of tunnel is not recommended mostly for security and scalability purpose. In addition, the initiating domain imposes huge constraints on subsequent domains, because they undergo the tunnel request without being able to control it.

- Stitching tunnels: Each domain establishes in its own network the corresponding part of the end-to-end LSP tunnel independently. Then, a second end-to-end RSVP-TE Path message is sent by the initiating domain to stitch the different tunnel parts to form the end-to-end LSP tunnel. In fact, this second RSVP-TE Path message is used by border nodes to exchange the label that must be used by the previous domain to send the traffic in order that the IP packets follow the next LSP tunnel in the following domain. These labels are convey in the RSVP-TE Resv message.

- Nesting tunnels: This is similar to the stitching mode but, this time, with the possibility to setup tunnel hierarchy. For example, an LSP tunnel between two edge domains crossing a transit domain could be inserted into a tunnel of higher hierarchy in the transit domain. Again, a second end-to-end RSVP-TE Path message is sent from the source to the destination. Labels that must be used to nest local tunnels are carried by the RSVP-TE Resv message.

In all case, RSVP-TE signaling must be exchange between the different domains. However, from an operational point of view, looking to different networks under the responsibility of different administrative entities, only BGP protocol are setup and configured between AS Border Routers (ASBR). Indeed, to the author’s knowledge, there is no example of operational networks that enable RSVP-TE between ASBR. Technology speaking, this is possible and many RFCs describe how to use RSVP-TE at the inter-domain. But, due to security, scalability, management and contract constraints, RSVP-TE is no longer exposed at the network boundary. To circumvent the

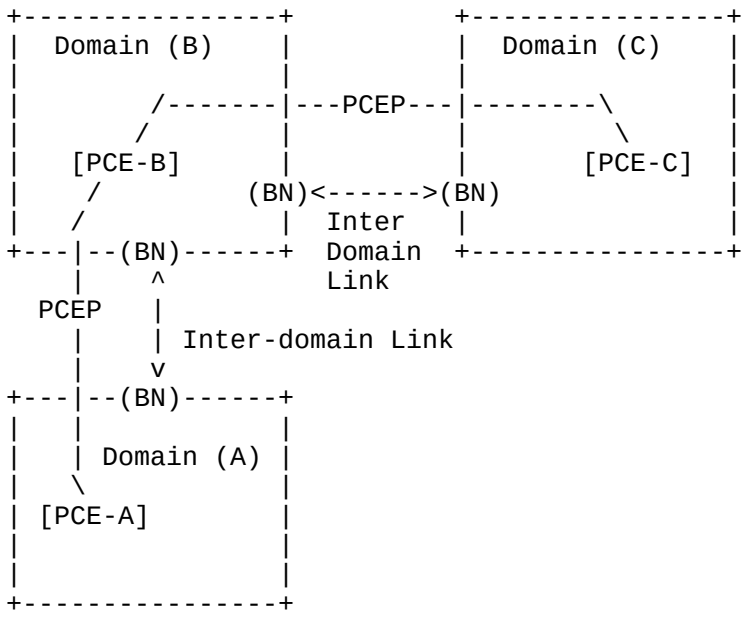
security issue, RSVP-TE could be carry inside an IPsec tunnel between ASBR, but, this not eliminate the scalability aspect nor the constraints impose by seting up and end-to-end LSP tunnels.

The purpose of this memo is to take the benefit of PCE stateful mode as per draft pce stateful [I-D.ietf-pce-stateful-pce] and draft pce initiated [I-D.ietf-pce-pce-initiated-lsp] to stitch or nest inter-domain LSP tunnels directly using PCEP protocol between domain's PCE instead of using RSVP-TE signaling at the inter-domain while keeping each operator independently setup their respective part of the end-to-end LSP tunnels. PCInitiated message is used in a Backward Recursive way like the PCReq message in BRPC [RFC5441], to recursively setup the end-to-end tunnel. PCRep message is used to automatically stitch or nest the different local LSP tunnels. And, PCRep in conjunction of PCUpd messages are used to maintain, modify and remove end-to-end LSP tunnels.

1.1. General assumptions

In the rest of this document, we used the same references as per BRPC [RFC5441] and make the following set of assumptions (see figure below):

- Domain refers to an IGP area or an Autonomous System (AS).
- Inter-domain LSP tunnel is used to refer to an LSP tunnel that cross two or more different domains as defined previously,
- At least, one PCE is deployed in each domain. These PCE are all stateful active capable and could request to enforce LSP tunnels in their respective domain by means of PCInitiate messages.
- LSRs, including border nodes, are PCC enable and support stateful active mode. PCEP sessions is established between these routers and their domain's PCE.
- Each PCE establishes a PCEP session with its respective neighbor domain's PCE. The way a PCE discover its neighboring PCE is out of scope of this draft. These information could be fulfill administratively or automatically discovered through, for example per draft 'BGP Extensions for Path Computation Element (PCE) Discovery' [I-D.dong-pce-discovery-proto-bgp],
- PCEs are able to compute and end-o-end path as per BRPC procedure [RFC5441].



Example of the representation of 3 domains with 3 PCEs

1.2. Terminology

ABR: Area Border Routers. Routers 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.

AS: Autonomous System

Border 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.

Domains: Autonomous System (AS) or IGP Area. An Autonoumous System is composed by one or more IGP area.

Entry BN of domain(i): a BN connecting domain(i-1) to domain(i) along a determined sequence of domains. Multiple entry BN(i) could be used to connect domain(i-1) to domain(i).

Exit BN of domain(i): a BN connecting domain(i) to domain(i+1) along a determined sequence of domains. Multiple exit BN(i) could be used to connect domain(i) to domain(i+1).

Inter-domain LSP tunnel: A LSP tunnel that crosses two or more domains through a per of Border Node.

Local LSP tunnel: A LSP tunnel that do not cross a domain. It is setup between entry BN to exit BN, any source to exit BN or entry BN to any destination of the same domain.

Local LSP tunnel(i): A local LSP tunnel of domain(i)

IGP-TE: Interior Gateway Protocol with Traffic Engineering support. Both OSPF-TE and IS-IS-TE are identified in this category.

Stitching Label (SL): A dedicated label that is used to stitch two RSVP-TE tunnels or two Segment Routing paths.

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).

2. Stitching Label

This section introduce the concept of Stitching Label that allows stitching and nesting of Local LSP tunnels in order to form inter-domain LSP tunnel that cross several different domains.

2.1. Definition

The operation of stitch or nest a local LSP tunnel(i) to a local LSP tunnel(i+1) in order to form and inter-domain LSP tunnel simply consist in defining the label that the exit BN(i) will use to send its traffic to the entry BN(i+1). Indeed, the entry BN(i+1) needs to identify the incoming traffic i.e. IP packets, in order to know if this traffic must follow the local LSP tunnel(i+1) or not. Forwarding Equivalent Class (FEC) could be used for that purpose. But, when stitching or nesting tunnels, the FEC is reduce to the incoming label that the entry BN(i+1) as chosen for the local LSP tunnel(i+1).

In this memo, we introduce the named of ’Stitching Label (SL)’ to designate this label. Such label is usually exchange between exit BN(i) and entry BN(i+1) with the RSVP-TE signaling. But, as we want to avoid to use RSVP-TE signaling due to operational constraints,

this Stitching Label will be convey by PCEP protocol. In fact, the Explicit Route Object (ERO) and the Record Route Object (RRO) are defined in order to transport this Stitching Label in the RSVP-TE signaling. As PCEP protocol used RSVP-TE Objects, and in particular the ERO and ERO, it is able to convey the Stitching Label without any modification of the PCEP protocol nor the PCE or RSVP-TE Objects.

As per RFC4003 [RFC4003], the Stitching Label will be convey as a companion of an IP address. In our case, this is the IP address of the input interface ITF_INPUT(i+1) of BN(i+1) which is connected to the exit BN(i) and which receives the traffic from the domain(i).

2.2. Inter-domain LSP-TYPE

However, even if PCEP could convey the Stitching Label, a PCC is not aware that a PCE requests or provides such label. For that purpose, this memo propose to use the LSP-TYPE as defined in draft lsp setup type [I-D.ietf-pce-lsp-setup-type] with new values (See IANA section of this memo) defined as follow:

- TBD1: Inter-Domain Traffic engineering end-to-end path is setup using Backward Recursive method. This new LSP-TYPE value MUST be set in a PCInitiate messages sends by a PCE(i) to its neighbor PCE(i+1) to initiate a new inter-domain LSP tunnel. In turn, neighbor PCE(i+1) MUST return a Stitching Label SL with the IP address of the associated interface in the RRO of the PCRpt message to PCE (i).

- TBD2: Inter-Domain Traffic engineering local path is setup using RSVP-TE. This new LSP-TYPE value MUST be set in the PCInitiate message sends by a PCE(i) requesting to a PCC of domain(i) to initiate a new local LSP tunnel(i) which is part of an inter-domain LSP tunnel. This LSP-TYPE value MUST be used by the PCE(i) only after receiving a PCInitiate message with an LSP-TYPE equal to TBD1 from a neighbor PCE(i-1). In turn, the PCC of domain(i) MUST return a Stitching Label SL with the IP address of associated interface in the RRO of the PCRpt message.

- TBD3: Inter-Domain Traffic engineering local path is setup using Segment Routing. This new LSP-TYPE value MUST be set in the PCInitiate message sends by a PCE(i) requesting to a PCC of domain(i) to initiate a new Segment Routing path which is part of and inter-domain Segment Routing path. This LSP-TYPE value MUST be used by the PCE(i) only after receiving a PCInitiate message with an LSP-TYPE equal to TBD1 from a neighbor PCE(i-1). In turn, the PCC MUST return a Stitching Label SL with the IP address of the associated interface in the RRO of the PCRpt message.

3. Inter-domain LSP tunnels setup procedure

This section describes how to setup inter-domain LSP tunnels than cross several different domains.

3.1. Mode of operation

This section describes how PCInitiate and PCRpt messages are combined between PCE in order to setup inter-domain LSP tunnels between a source domain(1) to a destination domain(n). S and D are respectively the source and destination of the inter-domain LSP tunnel. Domain(1) and domain(n) are different and connected through 0 or more intermediate domains denoted domain(i) with i = (2, n-1). Domains are directly connected when n = 2.

First, the PCE(S) run standard BRPC algorithm as per RFC5441 [RFC5441] with its neighbor PCEs in order to compute the inter-domain LSP tunnel from S to D, where S and D are respectively a node in the domain(1) and domain(n). Path Key confidentiality as per RFC5520 [RFC5520] MAY be used to obfuscate the detailed ERO of the different domains(i). The resulting ERO is of the form (S, PKS(1), exit BN(1), ..., entry BN(i), PKS(i), exit BN(i), ..., entry BN(n), PKS(n), D). As subsequent domains are not aware about the final computed ERO in case of multiple VSPT, the final computed ERO MUST be send in the PCInitiate message to indicate to the subsequent PCEs which solution has been finally chosen.

The complete procedure follow the different steps described below:

Steps 1: Initialization

Once ERO(S, D) computed, PCE(1) sends a PCInitiate message to PCE(2) containing and ERO equal to {S, PKS(1), exit BN(1), ..., entry BN(i), PKS(i), exit BN(i), ..., entry BN(n), PKS(n), D}, LSP-TYPE = TBD1 and End-Points Object = (S, D). The ERO corresponds to the one PCE(1) as received from PCE(2) during the BRPC process. In case of multiple EROs, i.e. VSPT > 1, PCE(1) has chosen one of them and used the selected one for the PCInitiate message. PKS(i) could be replaced by the full ERO description if Path Key is not used by PCE(i).

When PCE(i) receives a PCInitiate message from domain(i-1) with LSP-TYPE = TBD1 and ERO = {entry BN(i), PKS(i), exit BN(i), ..., entry BN(n), PKS(n), D)}, it forwards the PCInitiate message to PCE(i+1) once remove its {entry BN(i), PKS(i), exit BN(i)} part from the ERO. All intermediate PCE(i) propagate the PCInitiate message to PCE(i+1) up to the domain(n).

Steps 2: Actions taken at the destination domain(n)

When PCInitiate message propagation reach the destination domain(n), PCE(n) retrieves the ERO from the PKS(n) if necessary and sends to entry BN(n) a PCInitiate message with the ERO(n) = {BN(n), ..., D}, LSP-TYPE= TBD2 and End-Points Object = (BN(n), D) in order to inform the PCC BN(n) that this local LSP tunnel(n) is part of an inter-domain LSP tunnel. When the PCC entry BN(n) received the PCInitiate message from its PCE(n), it setup the LSP tunnels from entry BN(n) to D by means of RSVP-TE signaling with the given ERO(n). Once the tunnel setup, it chooses a free label for the Stitching Label SL(n) and add a new entry in its MPLS LFIB with this SL(n) label. Then, it sends a PCRpt message to its PCE(n) with an RRO equal to {[ITF_INPUT(n), SL(n)], RRO(n)}. Once PCE(n) receives the PCRpt from the PCC BN(n) with the RRO and LSP-TYPE = TBD2, it sends to the PCE(n-1) a PCRpt containing the RRO equal to {[ITF_INPUT(n), SL(n)]}. PCE(n) MAY adds BN(n), D in the RRO as loose path.

Steps i: Actions performed by all intermediate domains(i), for i = 2 to n-1

1. When the PCE(i) receives a PCRpt message from domain(i+1) with LSP-TYPE = TBD1 and RRO = {[ITF_INPUT(i+1), SL(i+1)]}, it retrieves the ERO from the PKS(i) if necessary and sends to the PCC entry BN(i) a PCInitiate message with ERO = {ERO(i), [ITF_INPUT(i+1), SL(i+1)]}, LSP-TYPE = TBD2 and End-Points Object = {entry BN(i), exit BN(i)} in order to inform the PCC entry BN(i) that this local LSP tunnel(i) is part of an inter-domain LSP tunnel.

2. When the PCC entry BN(i) received the PCInitiate message from its PCE(i), it setup the LSP tunnels from entry BN(i) to exit BN(i) by means of RSVP-TE signaling with the given ERO(i).

3. When the exit Bn(i) receives an RSVP-TE Path message with an ERO = {x-1, [ITF_INPUT(i+1), SL(i+1)]} and End-Points Object = {entry BN(i), exit BN(i)}, it MUST install in its MPLS LFIB the SWAP instruction to label SL(i+1) with forward to ITF_INPUT(i+1) instead of the standard POP instruction.

4. Once the tunnel setup, it chooses a free label for the Stitching Label SL(i) and add a new entry in its MPLS LFIB with this SL(i) label. Then, it sends a PCRpt message to its PCE(i) with an RRO equal to {[ITF_INPUT(i), SL(i)], RRO(i)}.

5. Once PCE(i) receives the PCRpt from the PCC entry BN(i) with the RRO and LSP-TYPE = TBD2, it sends to the PCE(i-1) a PCRpt containing the RRO equal to {[ITF_INPUT(i), SL(i)]}. PCE(i) MAY adds entry BN(i), exit BN(i) in the RRO as loose path.

Steps n: Actions performed at the source domain(1)

Once PCE(1) received the PCRpt message from PCE(2) with the RRO containing the label SL(2), it sends a PCInitiate message to PCC node S with ERO equal to {ERO(1), [ITF_INPUT(2), SL(2)]}, LSP_TYPE = 0 and End-Points Object = {S, BN(1)}. This time, the LSP_TYPE is equal to 0 as the PCC S does not need to return a Stitching Label SL i.e. it is the head-end of the inter-domain LSP tunnel. Standard PCRpt message is sent back to PCE(1) by the PCC node S.

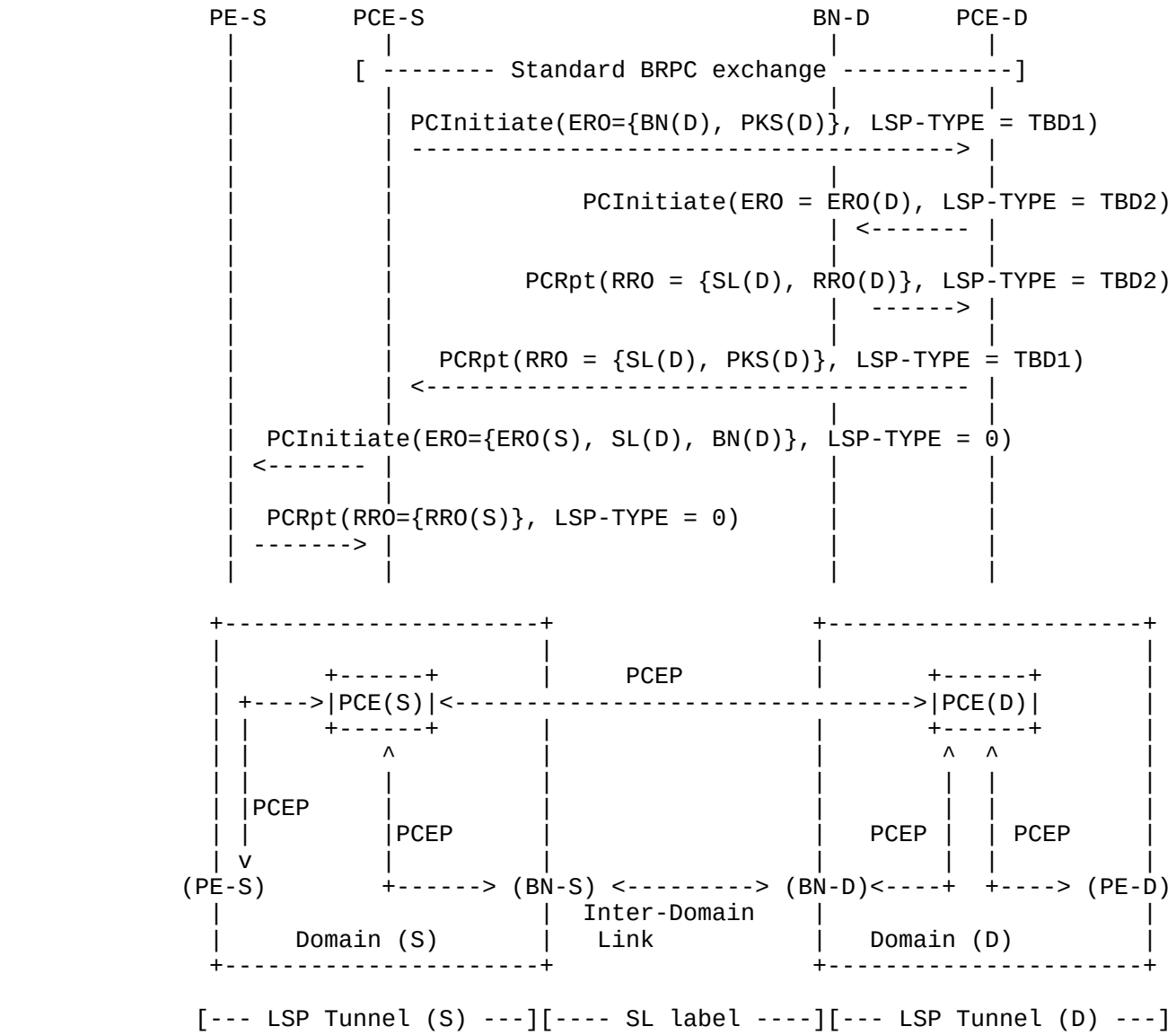
To use Segment Routing instead of RSVP-TE to setup the LSP tunnels as defined in draft pce segment routing [I-D.ietf-pce-segment-routing], PCEs MUST send PCInitiate message with LSP-TYPE = TBD3 instead of TBD2 to advertise their respective PCC that the LSP tunnels is enforce by means of Segment Routing. SL label will be inserted in the label stack in order to become the top label in the stack when the packet reach entry BN(i+). Then, entry BN(i+1) will push a new label stack to reach the exit BN(i+1) and follow.

3.2. Example

In the figure below, two different domains S and D are interconnected through BN respectively BN-S and BN-D. PE-S and PE-D are edge routers. All routers in the figure are connected to their respective PCE through PCEP protocol. In this example, PCE(S) would setup an intre-domain LSP tunnel between PE-S and PE-D acting as source and destination of the tunnel. Intermediate routers between (PE-S, BN-S), (BN-D and PE-D) as well as RSVP-TE messages are not represented to simplify the figure. But they are all presents. The following notation is used in the figure:

- PKS(D) = Path Key correponding to the path from BN(D) to PE-D
- ERO(D) = Explicit Route Object corresponding to the path from BN(D) to PE-D retrieves from PKS(D)
- RRO(D) = Record Route Object of Local LSP tunnel(D) from BN(D) to PE-D
- SL(D) = Stitching Label for Local LSP tunnel from BN(D) to PE-D
- ERO(S) = Explicit Route Object corresponding to the path from PE-S to BN(S)
- RRO(S) = Record Route Object of Local LSP tunnel(S) from PE-S to BN(S)

1.



Example of end-to-end LSP tunnel setup between two domains

3.3. Inter-domain LSP setup procedure completion failure

In case of error during LSP setup, PCRpt and or PCError messages MUST be used to signal the problem to the neighbor PCE domain backward. In particular, if new LSP-TYPE values defined in this memo are not

supported by the neighbor PCE or the PCC, the PCE, receptively the PCC, MUST return a PCErr message with Error-Type = 21 (Traffic engineering path setup error) and Error-Value = 1 (Unsupported path setup type) to its neighbor PCE.

If a PCC or a PCE don’t return an RRO or an RRO without the Stitching Label SL with the IP address of the associated interface following a PCInitiate message with LSP-TYPE set to the new values defined in this memo, the PCE MUST return a PCErr message with Error-Type = 21 (Traffic engineering path setup error) and Error-Value = TBD4 (No Mandatory Stitching Label is present in the RRO).

In case of completion failure, the PCE(i) MUST propagate the PCErr message up to the PCE(1). In turn, PCE(1) MUST send a PCInitate message (R flag set in the SRP Object as per draft pce initiated lsp [I-D.ietf-pce-pce-initiated-lsp] to delete this inter-domain LSP tunnel to its neighbor PCEs. PCE(i) MUST propagate the PCInitiate message and remove their Local LSP tunnel by means of PCInitiate message to their PCC entry BN(i) and send back PCRpt message to PCE(i-1).

3.4. Inter-domain LSP management

Each domain manages their respective local LSP tunnel part of an inter-domain LSP tunnel independently of each other. In particular, Stitching Label(i) is managed by domain(i) and is of interest of domain(i-1) only. Thus, Stitching Label SL(i) is not supposed to be propagated to other domains.

If a PCE(i) needs to modify its local LSP tunnel(i) with PCUpd message, it MUST sends a new PCRpt message to its neighbor PCE(i-1) to advertise it of the modification, in particular if this concern a modification of Stitching Label SL(i).

PCE(1) could modify the inter-domain LSP tunnel. For that purpose, it MUST sends a PCUpd message to its neighbor PCEs. Each PCE(i) MUST process PCUpd message the same way they process PCInitiate message: first, propagate the PCUpd message up to the destination domain(n), then process the modification once PCRpt received from PCE(i+1) and send PCRpt to PCE(i-1) once modification done.

Modification of Local LSP tunnel, entry BN(i) and exit BN(i) is left for further study.

In case of a failure appear in domain(i), PCE(i) MUST sends a PCRpt message to its neighbor PCE(i-1) to advertise it that its local part of the inter-domain LSP tunnel is down. Once PCE(1) receives this PCRpt message indicating that the tunnel is down, it is up to the

PCE(1) to take appropriate correction e.g. start a new BRPC to compute a new ERO.

4. Applicability

The newly introduce Stitching Label SL serves to stitch or nest part of LSP tunnels to form an inter-domain LSP tunnel. Each domain is free to decide if the tunnel is stitched or nested. For example, a domain(i) may decided to nest the incoming Local LSP tunnel into a higher hierarchy of tunnel for Traffic Engineering purpose. A PCE(i) may also decided to group Local LSP tunnels part of inter-domain LSP tunnels into a higher hierarchical tunnel to carry all these Local LSP tunnels from one entry BN(i) to one exit BN(i).

The Stitching Label SL could serves to stitch Segment Path and RSVP-TE tunnel. Indeed, each domain is free to enforce its part of the inter-domain LSP tunnel with the underlying mechanism it chosen. Stitching Label SL will be part of the label stack in order to become the top label in the stack when reaching the entry BN(i+1). This Stitching Label could be swap as usual if the next domain that uses RSVP-TE tunnel. When the previous domain uses a RSVP-TE tunnel, the Stitching Label will serve as key for the entry BN(i+1) to determine which label stack it must push on top of the packet for a Segment Routing path.

In inter-layer scenario is left for further study.

5. IANA Considerations

5.1. LSP-TYPE values

Draft pce lsp setup type [I-D.ietf-pce-lsp-setup-type] defines the PATH-SETUP-TYPE TLV and requests that IANA creates a registry to manage the value of the PATH_SETUP_TYPE TLV's PST field. IANA is requested to allocate a new code point in the PCEP PATH_SETUP_TYPE TLV PST field registry, as follows:

Value	Description	Reference
TBD1	Inter-Domain Traffic engineering end-to-end path is setup using Backward Recursive method	This Document
TBD2	Inter-Domain Traffic engineering local path is setup using RSVP-TE	This Document
TBD3	Inter-Domain Traffic engineering local path is setup using Segment Routing	This Document

5.2. PCEP-Error Object

IANA is requested to allocate code-points in the PCEP-ERROR Object Error Values registry for a new error-value or Error-Type 21 Invalid traffic engineering path setup:

Error-Value	Description
TBD4	Missing Mandatory Stitching Label in RRO

6. Security Considerations

No modification of PCE protocol (PCEP) has been requested by this draft which not introduce any issue regarding security. Concerning the PCEP session between PCEs, authors recommend to use the secure version of PCEP as defined in draft secure transport for PCEP [I-D.ietf-pce-pceps] or use any other secure tunnel mechanism e.g. IPsec tunnel to transport PCEP session between PCE.

7. Acknowledgements

The authors want to thanks PCE’s WG members.

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, <http://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, <http://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, <http://www.rfc-editor.org/info/rfc5440>.

[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, <http://www.rfc-editor.org/info/rfc5441>.

8.2. Informative References

[I-D.dong-pce-discovery-proto-bgp]
Dong, J., Chen, M., Dhody, D., Tantsura, J., Kumaki, K., and T. Murai, "BGP Extensions for Path Computation Element (PCE) Discovery", draft-dong-pce-discovery-proto-bgp-06 (work in progress), October 2016.

[I-D.ietf-pce-lsp-setup-type]
Sivabalan, S., Medved, J., Minei, I., Crabbe, E., Varga, R., Tantsura, J., and J. Hardwick, "Conveying path setup type in PCEP messages", draft-ietf-pce-lsp-setup-type-03 (work in progress), June 2015.

[I-D.ietf-pce-pce-initiated-lsp]
Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model", draft-ietf-pce-pce-initiated-lsp-09 (work in progress), March 2017.

[I-D.ietf-pce-pceps]
Lopez, D., Dios, O., Wu, W., and D. Dhody, "Secure Transport for PCEP", draft-ietf-pce-pceps-11 (work in progress), January 2017.

[I-D.ietf-pce-segment-routing]
Sivabalan, S., Medved, J., Filsfils, C., Crabbe, E., Raszuk, R., Lopez, V., Tantsura, J., Henderickx, W., and J. Hardwick, "PCEP Extensions for Segment Routing", draft-ietf-pce-segment-routing-08 (work in progress), October 2016.

[I-D.ietf-pce-stateful-pce]
Crabbe, E., Minei, I., Medved, J., and R. Varga, "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce-18 (work in progress), December 2016.

[RFC4003] Berger, L., "GMPLS Signaling Procedure for Egress Control", RFC 4003, DOI 10.17487/RFC4003, February 2005, <http://www.rfc-editor.org/info/rfc4003>.

[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, <http://www.rfc-editor.org/info/rfc5520>.

[RFC6805] King, D., Ed. and A. Farrel, Ed., "The Application of the Path Computation Element Architecture to the Determination of a Sequence of Domains in MPLS and GMPLS", RFC 6805, DOI 10.17487/RFC6805, November 2012, <http://www.rfc-editor.org/info/rfc6805>.

Authors' Addresses

Olivier Dugeon
Orange
2, Avenue Pierre Marzin
Lannion 22307
France

Email: olivier.dugeon@orange.com

Julien Meuric
Orange
2, Avenue Pierre Marzin
Lannion 22307
France

Email: julien.meuric@orange.com

PCE Working Group
Internet-Draft
Intended Status: Standards Track
Expires: March 2, 2018

R. Gandhi
Cisco Systems, Inc.
B. Wen
Comcast
C. Barth
Juniper Networks
D. Dhody
Huawei Technologies
August 29, 2017

PCEP Extensions for MPLS-TE LSP Performance Measurements with Stateful PCE
draft-gandhi-pce-pm-08

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. The Stateful PCE extensions allow Stateful control of Multi-Protocol Label Switching (MPLS) Traffic Engineering Label Switched Paths (TE LSPs) using PCEP.

In certain networks, network performance data such as packet loss, delay and delay variation (jitter) as well as bandwidth utilization is a critical measure for Traffic Engineering (TE). This data provides operators the characteristics of their networks for performance evaluation that is required to ensure the Service Level Agreements (SLAs). Performance Measurement (PM) mechanisms can be employed to monitor these metrics end-to-end for TE Label Switched Paths (LSPs). This document describes Path Computation Element Protocol (PCEP) extensions for enabling and reporting such performance measurements to an Active Stateful PCE for both PCE-Initiated and PCC-Initiated LSPs.

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/1id-abstracts.html

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

Copyright and License Notice

Table of Contents

1. Introduction

[RFC5440] describes the Path Computation Element Protocol (PCEP) as a communication mechanism between a Path Computation Client (PCC) and a Path Control Element (PCE), or between PCE and PCE, that enables computation of Multi-Protocol Label Switching (MPLS) Traffic Engineering Label Switched Paths (TE LSPs).

[DRAFT-PCE-STATEFUL] specifies extensions to PCEP to enable Stateful control of TE LSPs. It describes two mode of operations - Passive Stateful PCE and Active Stateful PCE. Further [DRAFT-PCE-INITIATED-LSP] describes the setup, maintenance and teardown of PCE-Initiated LSPs for the Stateful PCE model. In this document, the focus is on Active Stateful PCE where the LSPs are controlled by the PCE, for both PCE-Initiated and PCC-Initiated LSPs.

In certain networks, such as financial information networks, network performance data (e.g. packet loss, delay and delay variation (jitter), bandwidth utilization) is a critical measure for traffic engineering. The protocol extensions have been defined to advertise link performance metrics, see [RFC7471], [RFC7810], [RFC7823] and [DRAFT-IDR-TE-PM-BGP]. This data provides operators the characteristics of their networks for performance evaluation that is required to ensure the Service Level Agreements (SLAs).

[DRAFT-PCE-SERVICE-AWARE] defines the PCEP extensions for TE LSP path computation using packet loss, delay and delay variation as path selection metrics. Such path computations use link metrics for packet loss and delay and do not provide end-to-end metrics of the TE LSPs. The end-to-end metrics of a TE LSP may be very different than the path computation results due to many factors, such as queuing, etc. There is a need to verify and monitor that the traffic sent over the established TE LSPs does not exceed the requested metric bounds (e.g. total end-to-end delay/loss). The Stateful PCE may need to take some action (such as tear-down or re-optimize the LSP) when the performance requirement is not met. [RFC6374], [RFC6375] and [RFC7876] define protocol extensions needed for measuring end-to-end packet loss, delay and delay variation (jitter) for bidirectional and unidirectional TE LSPs.

This document provides mechanisms to enable and report the performance measurements (PM) such as packet loss, delay, delay variation (jitter) and bandwidth utilization for a TE LSP to an Active Stateful PCE, for both PCE-Initiated and PCC-Initiated LSPs.

1.1. Use-cases

This section describes a non-exhaustive list of deployment use-cases of PM for TE LSPs when deployed in a network with PCE. A controller may also be deployed in the network capable of "streaming telemetry" for receiving PM metrics and may interact with PCC and PCE for the PM as described in use-cases 3, 4 and 5.

Use-case 1: PCE Enables PM on PCC and PCC Takes Action

PCE -----> PCC

In this use-case, the PCE sets the upper-bound threshold condition for TE LSPs at the PCC. The PCC takes a local action when the condition is met. The action could be based on a local policy or policy set by the PCE. The steps involved are -

- PCE sends the PM attributes as part of PCE-initiated LSPs including upper-bound threshold (Section 4.6 in this document) for the PM metrics using the PCEP extensions defined in this document.

- PCC takes actions when PM metrics exceed the upper-bound threshold, actions could be to bring down the LSP, trigger protection switchover, remove tunnel from IGP for some prefixes, or request a new path from PCE (based on local policies which may be set by the PCE). PCC may take these actions even when LSPs are delegated to PCE as the upper-bound is set by the PCE.

- PCC does not report the PM metrics to PCE.

- PCC may install the new LSP in routing table only if the PM metric is below the upper-bound, otherwise, the PCC may reject the LSP request and send an error to the PCE.

- The report interval should be set to 0 to disable reporting to PCE. Only the upper-bound threshold should be set.

Use-case 2: PCC Reports PM Metrics to PCE, PCE Takes Action

PCE <----- PCC

In this use-case, the PCC reports the PM metrics and parameters to the PCE and the PCE may take an immediate local (reactive) action based on the PM metrics. The steps involved are -

- PCC sends the PM metrics and parameters to PCE using the PCEP extensions defined in this document and PCE takes an action; action could be to correlate faults, invalidate LSP path, send new

LSP path to PCC (trigger re-optimization), etc.

- If upper-bound threshold is set, PCC only reports the PM metrics to PCE when upper-bound is crossed. Otherwise the PCC reports the PM metrics to PCE every report-interval.

- Optionally, PCC may take an immediate local (reactive) action such as trigger path protection switch-over when PM metrics exceed upper-bound.

- PCE has a global view due to PM metric reports received from various PCCs and hence can make a better decision about LSP placement in the network.

- PCE can make pro-active decisions based on PM metrics when metrics are reported before crossing of the upper-bound as opposed to reactive action that PCC could make.

- The report interval should be set to enable reporting by the PCC. Optionally, the upper-bound threshold may also be set.

Use-case 3: PCE Enables PM on PCC, PCC Sends PM Metrics to Controller

PCE -----> PCC -----> Controller

The steps involved are -

- A controller may be used in a network that is capable of "streaming telemetry" for receiving data and Yang or XML based provisioning using non-PCEP channel. The controller may interact with a PCE for LSP path computation using the PCEP channel.

- PCE sends the PM attributes as part of PCE-initiated LSPs using the PCEP extensions defined in this document.

- PCC reports the PM metrics to controller via "streaming telemetry".

- Controller may request PCE to take an action based on the PM metrics.

- The report interval should be set to 0 to disable reporting to PCE. The other PM attributes may be set and used for "streaming telemetry".

Use-case 4: Controller Enables PM on PCC, PCC Sends PM Metrics to PCE

PCE <----- PCC <----- Controller

The steps involved are -

- Controller enables PM on PCC using a non-PCEP channel.
- PCC then reports the PM metrics to PCE using the PCEP extensions defined in this document.
- PCE may take an action based on the PM metrics received from PCC.

Use-case 5: Controller Enables PM on PCC, PCC Sends PM Metrics to Controller

PCE ----> PCC <-----> Controller -----> PCE

The steps involved are -

- Controller enables PM on PCC using a non-PCEP channel.
- PCC reports the PM metrics to the controller via "streaming telemetry".
- Controller may request PCE to take an action based on the PM metrics.
- The PCEP extensions defined in this document are not used in this use-case.

1.2. Dependencies and Considerations

[RFC6374] describes several reasons why PMs are valuable to operators. Note that the specification of the use of the reported packet loss, delay, delay variation and bandwidth utilization measurements by a Stateful PCE is outside the scope of this document.

Furthermore, [RFC6374] describes many types of MPLS channels that may leverage PMs and some may have bidirectional dependencies. Defining a mechanism for the verification and/or provisioning of bidirectional or associated bidirectional LSPs within the Stateful PCE architecture is outside the scope of this document.

In all cases, delay and loss PM messages are carried over the MPLS Generic Associated Channel (G-ACh) as described in [RFC5586]. MPLS LSPs that carry the G-ACh can be referred to as MPLS Transport Profile (MPLS-TP) LSPs [RFC5921]. MPLS-TP LSPs require Ultimate Hop Popping (UHP) where LSPs are assigned Non-NULL labels by tail-end nodes. It is beyond the scope of this document to define the

mechanism by which a Stateful PCE verifies and/or provisions an LSP for UHP. Note that for both unidirectional and bidirectional LSPs, packet loss measurement requires UHP.

1.3. Auto-bandwidth Considerations

Auto-Bandwidth feature allows a head-end LSR (PCC) to automatically adjust the LSP bandwidth reservation based on the traffic demand of a TE LSP. Auto-bandwidth requested bandwidth computation can be implemented on a PCC or a Stateful PCE.

[DRAFT-IETF-PCE-AUTOBW] describes the PCEP extensions for auto-bandwidth, where the requested bandwidth for the LSP is computed by the PCC and reported to the Stateful PCE. There is a benefit in pushing the responsibility for deciding when auto-bandwidth adjustments are needed to the PCC as this distributes the load of monitoring the bandwidth utilization of the LSPs down to the PCCs and frees up the PCE for path computations. In addition, it reduces the load on PCEP communications for reporting the bandwidth utilization of the LSP.

However, exactly when to adjust an LSP bandwidth could be better left to a Stateful PCE. That is, a PCE could be flexible in its interpretation of thresholds enabling it to trigger auto-bandwidth adjustment early if it believes there is a good reason (for example, doing a set of parallel path re-computations) or late (for example, when it knows that an adjustment would be disruptive to the network). When the auto-bandwidth computation is delegated to the PCC, the PCC cannot see the impact on other LSPs in the network, and the PCE cannot tell whether the request to adjust the LSP bandwidth is critical or not. The bandwidth utilization reporting defined in this document can be used by the PCE to do computations to determine whether auto-bandwidth adjustments are needed and/or desirable before performing the path computations.

2. Conventions Used in This Document

2.1. 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].

2.2. Terminology

The reader is assumed to be familiar with the terminology defined in [RFC5440], [RFC6374] and [RFC7471].

2.3. Measurement Units

The measurement unit for delay value is defined in [RFC7471], Section 4.1.5.

The measurement unit for loss value is defined in [RFC7471], Section 4.4.5.

The utilized bandwidth [RFC7471] is encoded in IEEE floating point format in bytes per second (see [IEEE.754.1985]).

All average values are calculated as rolling averages.

3. Overview of the PCEP Extensions

The high-level overview of the PCEP extensions defined in this document for requesting and reporting end-to-end performance measurements and bandwidth utilization for TE LSPs are shown in Figure 1.

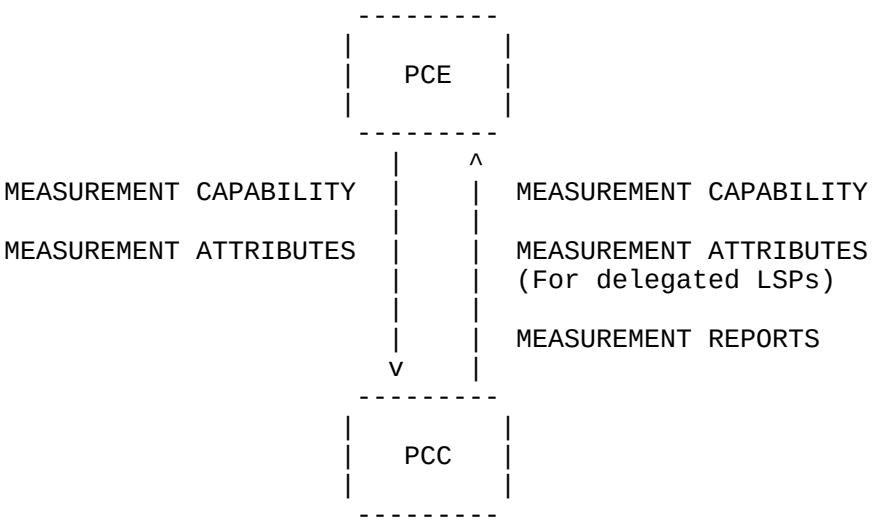


Figure 1: Overview of PCEP Extensions

The following steps describe the PCEP extensions for reporting performance measurements and bandwidth utilization of TE LSPs:

- The Stateful PCE and PCC (head-end of the LSP) advertise the capability of their support for the delay, loss and bandwidth-utilization measurements and reporting in the PCEP Open message

(in the OPEN Object).

- The Stateful PCE enables the measurement of a feature and sends and updates the configuration parameters of the feature using the LSPA object to PCC in PCInitiate and PCUpd messages respectively.

- The PCC reports the measured values of the feature to the Stateful PCE at the end of the specified report-interval or when measured values cross the specified report-threshold. The periodic reporting can be used at the PCE to monitor the TE LSP metrics whereas report-threshold can be used to trigger an immediate action at the PCE on the TE LSP.

- In some cases, the periodic reporting of the measurements may be disabled and only an upper-bound threshold is set, which when exceeded, a local or PCE-set action may be taken.

- The PCE and PCC notify each other of their entering and clearing the overwhelmed state when operating under high LSP scale.

3.1. Report Thresholds

When explicitly configured, report threshold (absolute or percentage) parameter (along with the configured number of counts) is used to trigger an immediate reporting of the delay and loss measurements and bandwidth utilization, bypassing the report interval. Threshold is used to detect a sudden change in the performance measurement metric of a TE LSP. In order to detect that a measured value has crossed the threshold, the measured (delay/loss) metric is compared with the last reported value. If the change (increase or decrease) in the value is above the threshold (absolute or percentage) consecutively for the given number of counts, the measurement from the current interval is reported immediately. In case of bandwidth utilization, the last reported MaxSampleBw (see [DRAFT-IETF-PCE-AUTOBW]) value is compared with the MaxSampleBW from the the current interval to detect the threshold crossing. The delay and loss measurements are still reported at the end of the report interval even if they were reported due to the crossing of the threshold. Refer to [RFC7471], Section 5, for additional considerations.

All thresholds in this document could be represented in both absolute value and percentage, and could be used together. This is provided to accommodate the cases where the metric values may become very large or very small over time. For example, an operator may use the percentage threshold to handle small to large metric values and absolute values to handle very large metric values. The metrics are reported when either one of the two thresholds, the absolute or

percentage, is crossed.

When using the percentage threshold, if the metric changes rapidly at very low values, it may trigger frequent reporting due to the crossing of the percentage threshold. This can lead to unnecessary scale issues in the network. This is suppressed by setting the minimum-threshold parameter along with the percentage threshold. The metric value is only reported if the value crosses both the percentage threshold and the minimum-threshold parameters.

4. Sub-TLVs for Measurements Attributes

This section specifies the generic sub-TLVs those provide various configurable parameters for reporting measurements to a Stateful PCE. These sub-TLVs are carried in various measurement attributes TLVs defined in this document.

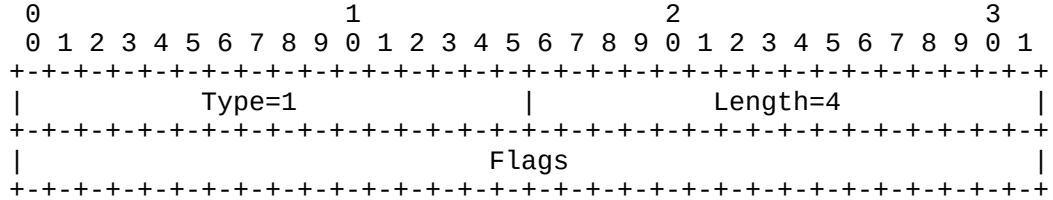
Following sub-TLVs are defined:

Type	Len	Name
1	4	Measurement-Enable sub-TLV
2	4	Measurement-Interval sub-TLV
3	8	Report-Threshold sub-TLV
4	8	Report-Threshold-Percentage sub-TLV
5	4	Report-Interval sub-TLV
6	8	Report-Upper-Bound sub-TLV

The Measurement-Enable sub-TLV MUST be added in the LSPA Object when the measurement feature is enabled for the LSP. All other sub-TLVs are optional and any unrecognized sub-TLV MUST be silently ignored. If a sub-TLV of same type appears more than once, only the first occurrence is processed and all others MUST be ignored. If sub-TLVs are not present, the default values based on the local policy are assumed.

4.1. Measurement-Enable sub-TLV

The Measurement-Enable sub-TLV specifies that the given measurement feature is enabled.



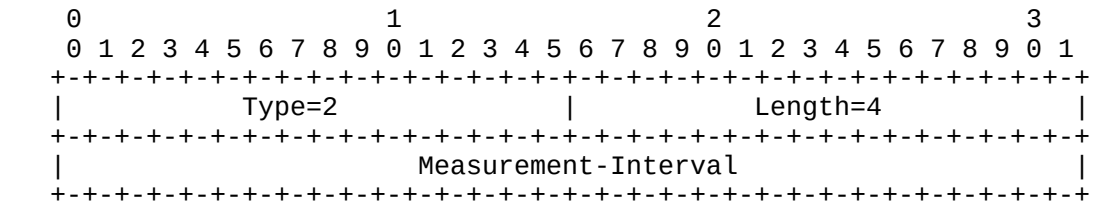
Measurement-Enable sub-TLV Format

The Type is 1, Length is 4 bytes, and the value comprises of flags (32 bits) for enabling various measurement features.

Unassigned flags are considered reserved, they MUST be set to 0 when sent and MUST be ignored when received.

4.2. Measurement-Interval sub-TLV

The Measurement-Interval sub-TLV specifies a time interval in seconds for the measurement.

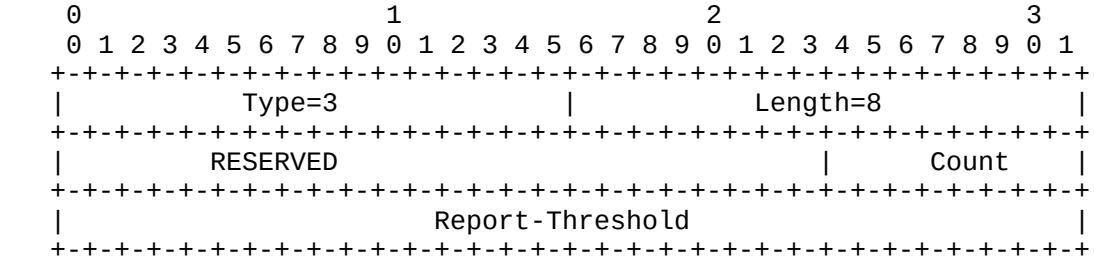


Measurement-Interval sub-TLV Format

The Type is 2, Length is 4 bytes, and the value comprises of 4-byte time interval, the valid range is from 1 to 604800, in seconds. The default value is 300 seconds. The Measurement-Interval MUST NOT be greater than Report-Interval.

4.3. Report-Threshold sub-TLV

The Report-Threshold sub-TLV specifies the threshold value used to trigger an immediate reporting of the measurements bypassing the report-interval.



Report-Threshold sub-TLV Format

The Type is 3, Length is 8 bytes, and the value comprises of -

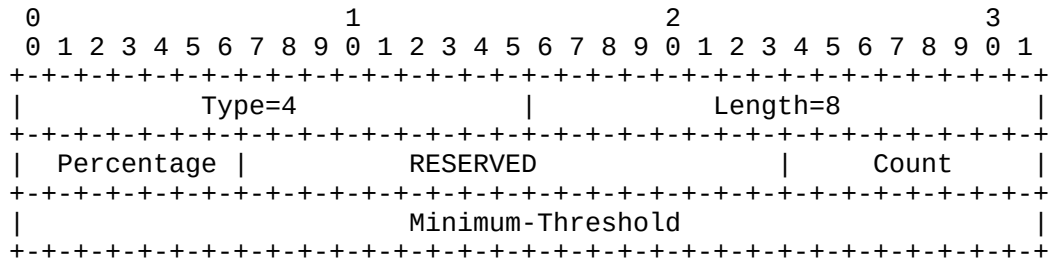
- Report-Threshold: 32-bit absolute threshold value.

- Count: 8-bit integer counter. The number of consecutive measurement values MUST be above the threshold before reporting the measurement. The value 0 is considered to be invalid. By default, report-threshold is not set and threshold check based reporting is disabled.

- RESERVED: It MUST be set to zero when sent and MUST be ignored when received.

4.4. Report-Threshold-Percentage sub-TLV

The Report-Threshold-Percentage sub-TLV specifies the threshold value used to trigger an immediate reporting of the measurements bypassing the report-interval.



Report-Threshold-Percentage sub-TLV Format

The Type is 4, Length is 8 bytes, and the value comprises of -

- Percentage: 7-bit threshold value, encoded in percentage (an integer from 1 to 100).

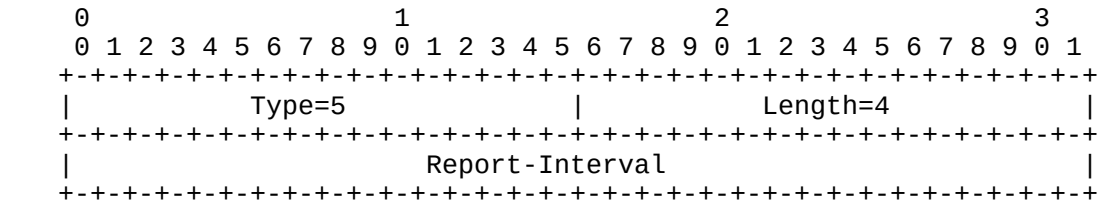
- Count: 8-bit integer counter. The number of consecutive measurement values MUST be above threshold before reporting the measurement. The value 0 is considered to be invalid. By default, report-threshold-percentage is not set and threshold check based reporting is disabled.

- RESERVED: It MUST be set to zero when sent and MUST be ignored when received.

- Minimum-Threshold: The 32-bit absolute Minimum-Threshold value. The increase or decrease should be at least or above this value.

4.5. Report-Interval sub-TLV

The Report-Interval sub-TLV specifies the time interval in seconds when measured values are to be reported.

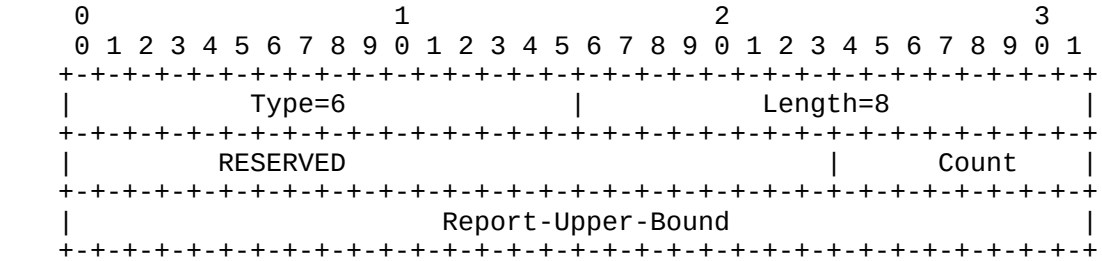


Report-Interval sub-TLV Format

The Type is 5, Length is 4 bytes, and the value comprises of 4-byte time interval, the valid range is from 0 to 604800, in seconds. The default value is 3600 seconds. The value 0 is used to disable the periodic reporting of the measurements.

4.6. Report-Upper-Bound sub-TLV

The Report-Upper-Bound sub-TLV specifies the upper-bound value used to trigger an immediate reporting of the measurements when crossed. This may also result in PCC taking an immediate local action on the LSP.



Report-Upper-Bound sub-TLV Format

The Type is 6, Length is 8 bytes, and the value comprises of -

- Report-Upper-Bound: 32-bit absolute value.
- Count: 8-bit integer counter. The number of consecutive measurement values MUST be above the upper-bound before reporting the measurement. The value 0 is considered to be invalid. By default, upper-bound is not set.
- RESERVED: It MUST be set to zero when sent and MUST be ignored when received.

5. PCEP Extensions for Reporting Delay Measurement

5.1. Delay Measurement Capability Advertisement

During PCEP Initialization Phase, PCEP Speakers (PCE or PCC) advertise their support of DELAY-MEASUREMENT. A PCEP Speaker (PCE or PCC) includes the DELAY-MEASUREMENT-CAPABILITY TLV, in the OPEN Object to advertise its support for PCEP Delay-Measurement extensions. The presence of the DELAY-MEASUREMENT-CAPABILITY TLV in the OPEN Object (in the Open message) indicates that the Delay Measurement capability is supported as described in this document. Additional procedure is defined as following:

- The PCEP protocol extensions for Delay Measurement MUST NOT be used if one or both PCEP Speakers have not included the DELAY-MEASUREMENT-CAPABILITY TLV in their respective Open message.

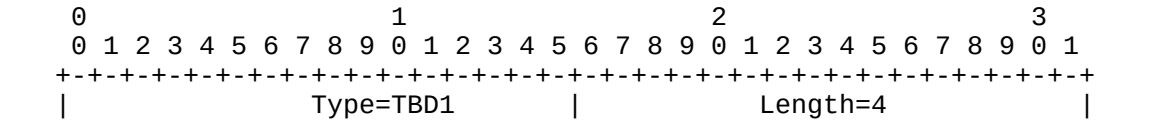
- If the PCEP speaker that supports the extensions of this document but did not advertise this capability, then upon receipt of DELAY-MEASUREMENT-ATTRIBUTES TLV in LSPA object, it SHOULD generate a PCErr with error-type 19 (Invalid Operation), error-value TBD7 (Delay-Measurement capability was not advertised) and it will terminate the PCEP session.

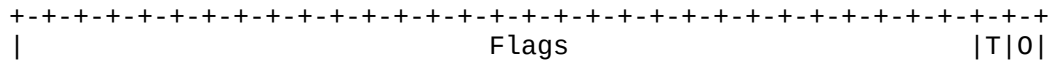
- Similarly, the PCEP speaker SHOULD generate error-value TBD9 (Bidirectional Measurement capability was not advertised) and TBD10 (Unidirectional Measurement capability was not advertised) upon receipt of DELAY-MEASUREMENT-ATTRIBUTES TLV in LSPA object with two-way measurement request and one-way measurement request, respectively, when it did not advertise this capability.

- If the PCEP speaker that supports the extensions of this document but did not advertise this capability, then upon receipt of DELAY-MEASUREMENT object, it SHOULD generate a PCErr with error-type 19 (Invalid Operation), error-value TBD7 (Delay-Measurement capability was not advertised) and it will terminate the PCEP session.

5.1.1. DELAY-MEASUREMENT-CAPABILITY TLV

The DELAY-MEASUREMENT-CAPABILITY TLV is an optional TLV for use in the OPEN Object for Delay Measurement via PCEP capability advertisement. Its format is shown in the following figure:





DELAY-MEASUREMENT-CAPABILITY TLV Format

The Type of the TLV is TBD1 and it has a fixed length of 4 bytes.

The value comprises a single field - Flags (32 bits):

- O (One-way Delay Measurement - 1 bit): if set to 1 by a PCC, the O Flag indicates that the PCC allows reporting of one-way delay measurement information; if set to 1 by a PCE, the O Flag indicates that the PCE is capable of receiving one-way delay measurement information from the PCC.

- T (Two-way Delay Measurement - 1 bit): if set to 1 by a PCC, the T Flag indicates that the PCC allows reporting of two-way delay measurement information; if set to 1 by a PCE, the T Flag indicates that the PCE is capable of receiving two-way delay measurement information from the PCC. Two-way measurement is only applicable to the bidirectional LSPs (e.g. MPLS-TP LSPs [RFC5921]).

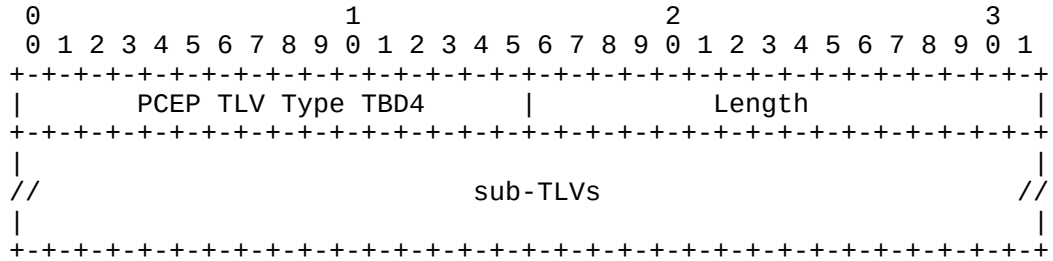
Unassigned bits are considered reserved. They MUST be set to 0 when sent and MUST be ignored when received.

Advertisement of the DELAY-MEASUREMENT-CAPABILITY TLV implies support of delay measurement, as well as the objects, TLVs and procedures defined in this document. Either O or T flag MUST be set in the TLV.

5.2. DELAY-MEASUREMENT-ATTRIBUTES TLV

The DELAY-MEASUREMENT-ATTRIBUTES TLV provides the configurable parameters of the delay measurement feature.

The format of the DELAY-MEASUREMENT-ATTRIBUTES TLV is shown in the following figure:



DELAY-MEASUREMENT-ATTRIBUTES TLV Format

PCEP TLV Type is defined as following:

Type	Name
TBD4	DELAY-MEASUREMENT-ATTRIBUTES

Length: The Length field defines the length of the value portion in bytes as per [RFC5440].

Value: Comprises of one or more sub-TLVs as described in Section 4 of this document.

The following sub-sections describe the parameters which are currently defined to be carried within this TLV.

5.2.1. Delay Measurement Enable

The Measurement-Enable sub-TLV specifies the delay measurement mode enabled using following flags:

Bit	Description
31	One-Way Delay Measurement Enabled
30	Two-Way Delay Measurement Enabled

5.2.2. Delay Measurement Interval

The Measurement-Interval sub-TLV specifies a time interval in seconds for the delay measurement.

5.2.3. Delay Measurement Report Threshold

The Report-Threshold sub-TLV specifies the threshold value used to trigger an immediate reporting of the delay measurements bypassing the report-interval.

- Report-Threshold: Delay in microseconds, encoded as 24-bit integer, as defined in [RFC7471].

Same report-threshold is used for all delay measurement values.

5.2.4. Delay Measurement Report Threshold Percentage

The Report-Threshold-Percentage sub-TLV specifies the threshold value used to trigger an immediate reporting of the measurements bypassing the report-interval.

Same report-threshold-percentage is used for all delay measurement values.

5.2.5. Delay Measurement Report Interval

The Report-Interval sub-TLV specifies the time interval in seconds when measured delay values are to be reported.

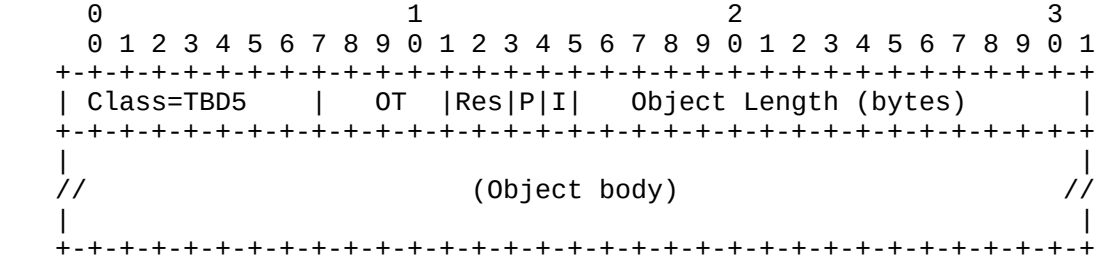
5.2.6. Delay Measurement Upper Bound

The Report-Upper-Bound sub-TLV specifies the upper-bound value in microseconds, and is used to trigger an immediate reporting of the delay values when crossed. This may also result in PCC taking an immediate local action on the LSP.

5.3. DELAY-MEASUREMENT Object

A new object, DELAY-MEASUREMENT with Object-Class (Value TBD5) is defined in this document to report the delay measurement of a TE LSP.

When the LSP is enabled with the delay measurement feature, the PCC SHOULD include the DELAY-MEASUREMENT Object to report the measured delay values to the PCE in the PCRpt message. The PCC SHOULD report (either one-way or two-way) average delay, min/max delay and delay variations to the PCE in the PCRpt message.



DELAY-MEASUREMENT Object Format

Object Length (16 bits): Specifies the total object length including the header, in bytes as per [RFC5440].

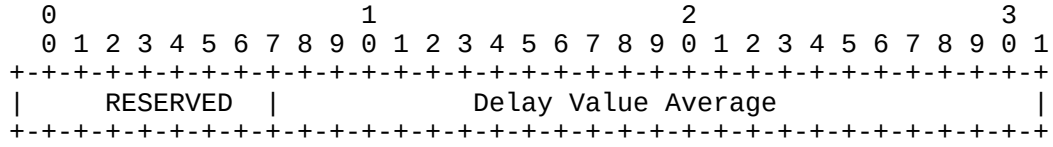
Object-Types (OT) are defined as follows:

Object-Type	Len	Name
1	8	One-Way Delay Measurement Value

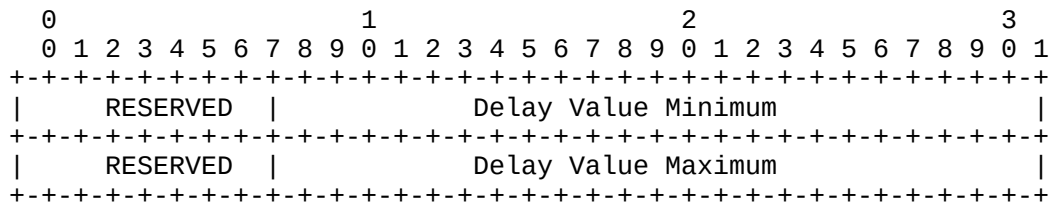
2	12	One-Way Delay Measurement Min/Max Values
3	8	One-Way Delay Variation Measurement Value
4	8	Two-Way Delay Measurement Value
5	12	Two-Way Delay Measurement Min/Max Values
6	8	Two-Way Delay Variation Measurement Value

The object body formats are defined as following:

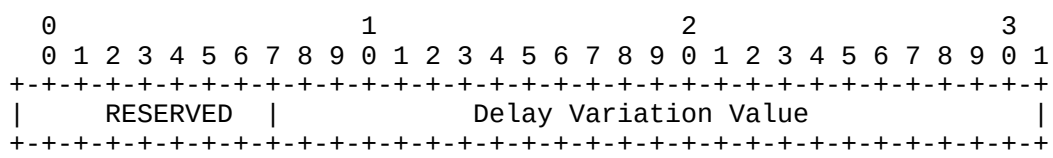
For Object-Types 1 and 4:



For Object-Types 2 and 5:



For Object-Types 3 and 6:



DELAY-MEASUREMENT Object Body Formats (One-Way and Two-Way)

Al delay values are reported in microseconds, encoded as 24-bit integer, as defined in [RFC7471]. When set to the maximum value 16,777,215 (16.777215 sec), the delay is at least that value and may be larger.

o One-way Delay Measurement Value: Average Delay of the LSP in one (forward) direction.

- One-way Delay Measurement Variation Value: Average Delay Variation of the LSP in one (forward) direction.

- One-Way Delay Measurement Value Minimum/Maximum: Minimum and Maximum values of the Delay of the LSP in one (forward) direction in the last measurement interval.

- Two-way Delay Measurement Value: Average Delay of the bidirectional LSP in both (forward and reverse) directions.

- Two-way Delay Measurement Variation Value: Average Delay Variation of the bidirectional LSP in both (forward and reverse) directions.

- Two-Way Delay Measurement Value Minimum/Maximum: Minimum and Maximum values of the Delay of the bidirectional LSP in both (forward and reverse) directions in the last measurement interval.

- RESERVED: This field is reserved for future use. It MUST be set to 0 when sent and MUST be ignored when received.

6. PCEP Extensions for Reporting Loss Measurement

6.1. Loss Measurement Capability Advertisement

During PCEP Initialization Phase, PCEP Speakers (PCE or PCC) advertise their support of LOSS-MEASUREMENT. A PCEP Speaker includes the LOSS-MEASUREMENT-CAPABILITY TLV, in the OPEN Object to advertise its support for PCEP Loss-Measurement extensions. The presence of the LOSS-MEASUREMENT-CAPABILITY TLV in the OPEN Object (in the Open message) indicates that the Loss Measurement capability is supported as described in this document. Additional procedure is defined as following:

- The PCEP protocol extensions for Loss Measurement MUST NOT be used if one or both PCEP Speakers have not included the LOSS-MEASUREMENT-CAPABILITY TLV in their respective Open message.

- If the PCEP speaker that supports the extensions of this document but did not advertise this capability, then upon receipt of LOSS-MEASUREMENT-ATTRIBUTES TLV in LSPA object, it SHOULD generate a PCErr with error-type 19 (Invalid Operation), error-value TBD8 (Loss-Measurement capability was not advertised) and it will terminate the PCEP session.

- Similarly, the PCEP speaker SHOULD generate error-value TBD9 (Bidirectional Measurement capability was not advertised) and

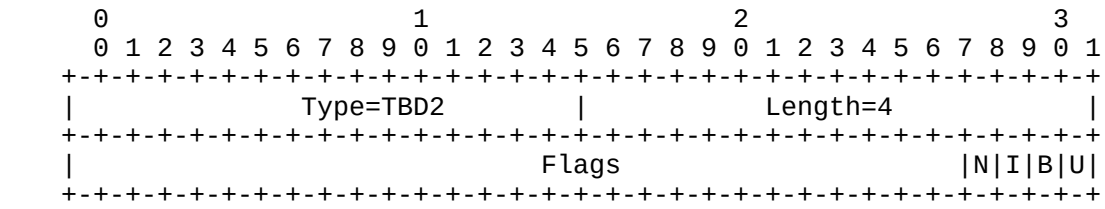
TBD10 (Unidirectional Measurement capability was not advertised) upon receipt of LOSS-MEASUREMENT-ATTRIBUTES TLV in LSPA object with two-way measurement request and one-way measurement request, respectively, when it did not advertise this capability.

- Further, the PCEP speaker SHOULD generate error-value TBD11 (Inferred Mode Loss Measurement capability was not advertised) and TBD12 (Direct Mode Loss Measurement capability was not advertised) upon receipt of LOSS-MEASUREMENT-ATTRIBUTES TLV in LSPA object with Inferred Mode loss measurement request and Direct Mode loss measurement request, respectively, when it did not advertise this capability.

- If the PCEP speaker that supports the extensions of this document but did not advertise this capability, then upon receipt of LOSS-MEASUREMENT object, it SHOULD generate a PCErr with error-type 19 (Invalid Operation), error-value TBD8 (Loss-Measurement capability was not advertised) and it will terminate the PCEP session.

6.1.1. LOSS-MEASUREMENT-CAPABILITY TLV

The LOSS-MEASUREMENT-CAPABILITY TLV is an optional TLV for use in the OPEN Object for Loss Measurement via PCEP capability advertisement. Its format is shown in the following figure:



LOSS-MEASUREMENT-CAPABILITY TLV Format

The Type of the TLV is TBD2 and it has a fixed length of 4 bytes.

The value comprises a single field - Flags (32 bits):

- U (Unidirectional Measurement - 1 bit): if set to 1 by a PCC, the U Flag indicates that the PCC allows reporting of unidirectional loss measurement information; if set to 1 by a PCE, the U Flag indicates that the PCE is capable of receiving unidirectional loss measurement information from the PCC.

- B (Bidirectional Measurement - 1 bit): if set to 1 by a PCC, the B Flag indicates that the PCC allows reporting of bidirectional loss

measurement information; if set to 1 by a PCE, the B Flag indicates that the PCE is capable of receiving bidirectional loss measurement information from the PCC. Bidirectional measurement is only applicable to the bidirectional LSPs (e.g. MPLS-TP LSPs [RFC5921]).

- o I (Inferred Loss Measurement Mode - 1 bit): if set to 1 by a PCC, the I Flag indicates that the PCC allows reporting of inferred mode loss measurement [RFC6374] information; if set to 1 by a PCE, the I Flag indicates that the PCE is capable of receiving inferred mode loss measurement information from the PCC.

- o N (Direct Loss Measurement Mode - 1 bit): if set to 1 by a PCC, the N Flag indicates that the PCC allows reporting of direct mode loss measurement [RFC6374] information; if set to 1 by a PCE, the N Flag indicates that the PCE is capable of receiving direct mode loss measurement information from the PCC.

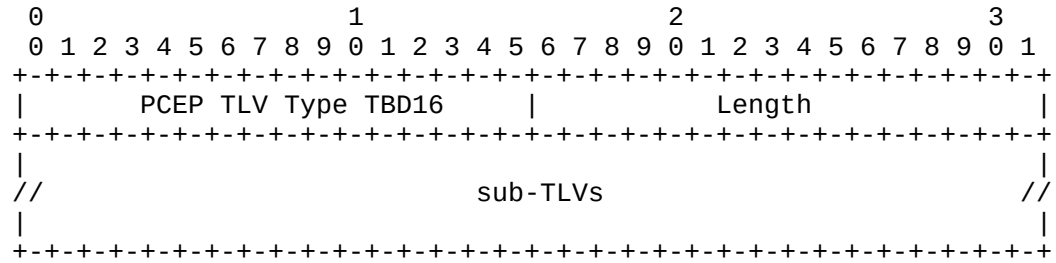
Unassigned bits are considered reserved. They MUST be set to 0 when sent and MUST be ignored when received.

Advertisement of the LOSS-MEASUREMENT-CAPABILITY TLV implies support of loss measurement, as well as the objects, TLVs and procedures defined in this document. Either U or B flag MUST be set in the TLV. Similarly, either I or N flag MUST be set in the TLV.

6.2. LOSS-MEASUREMENT-ATTRIBUTES TLV

The LOSS-MEASUREMENT-ATTRIBUTES TLV provides the configurable parameters of the loss measurement feature.

The format of the LOSS-MEASUREMENT-ATTRIBUTES TLV is shown in the following figure:



LOSS-MEASUREMENT-ATTRIBUTES TLV Format

PCEP TLV Type is defined as following:

Type	Name
TBD16	LOSS-MEASUREMENT-ATTRIBUTES

Length: The Length field defines the length of the value portion in bytes as per [RFC5440].

Value: Comprises of one or more sub-TLVs as described in Section 4 of this document.

The following sub-sections describe the parameters which are currently defined to be carried within this TLV.

6.2.1. Loss Measurement Enable

The Measurement-Enable sub-TLV specifies the loss measurement mode enabled using following flags:

Bit	Description
29	Unidirectional Loss Measurement Enabled
28	Bidirectional Loss Measurement Enabled
27	Inferred Loss Measurement Enabled

6.2.2. Loss Measurement Interval

The Measurement-Interval sub-TLV specifies a time interval in seconds for the loss measurement.

6.2.3. Loss Measurement Report Threshold

The Report-Threshold sub-TLV specifies the threshold value used to trigger an immediate reporting of the loss measurements bypassing the report-interval.

- Report-Threshold: This 24-bit field identifying the packet loss as a percentage of the total packets sent or received. The encoding is as per [RFC7471].

Same report-threshold is used for all loss measurement values.

6.2.4. Loss Measurement Report Threshold Percentage

The Report-Threshold-Percentage sub-TLV specifies the threshold value used to trigger an immediate reporting of the loss measurements bypassing the report-interval.

Same report-threshold-percentage is used for all loss measurement

values.

6.2.5. Loss Measurement Report Interval

The Report-Interval sub-TLV specifies the time interval in seconds when measured loss values are to be reported.

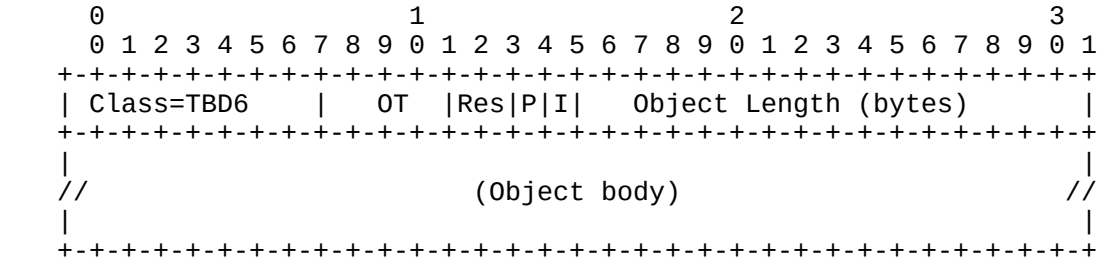
6.2.6. Loss Measurement Upper Bound

The Report-Upper-Bound sub-TLV specifies the upper-bound value in percentage packet loss, and is used to trigger an immediate reporting of the packet loss values when crossed. This may also result in PCC taking an immediate local action on the LSP.

6.3. LOSS-MEASUREMENT Object

The LOSS-MEASUREMENT Object with Object-Class (Value TBD6) is defined in this document to report the packet loss measurement of a TE LSP.

When the LSP is enabled with the loss measurement feature, the PCC SHOULD include the LOSS-MEASUREMENT Object to report the measured packet loss to the PCE in the PCRpt message.



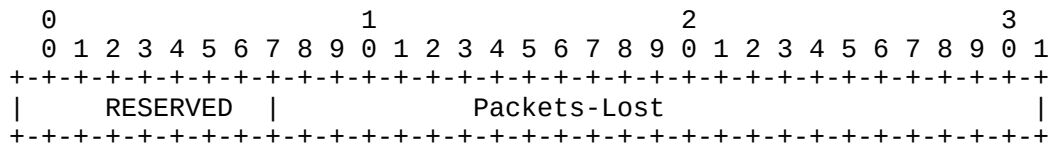
LOSS-MEASUREMENT Object Format

Object Length (16 bits): Specifies the total object length including the header, in bytes [RFC5440].

Object-Types (OT) are defined as following:

Object-Type	Len	Name
1	8	Tx Packets-Lost
2	8	Rx Packets-Lost

The object body format is defined as following:



LOSS-MEASUREMENT Object Body Formats (Tx and Rx)

- Packets-Lost: This 24-bit field identifying the packet loss as a percentage of the total packets sent or received, encoded as per [RFC7471].

- RESERVED: This field is reserved for future use. It MUST be set to 0 when sent and MUST be ignored when received.

The Packets-Lost in the Rx direction is reported when bidirectional loss measurement is enabled.

7. PCEP Extensions for Reporting Bandwidth Utilization

7.1. Bandwidth Utilization Capability Advertisement

During PCEP Initialization Phase, PCEP Speakers (PCE or PCC) advertise their support of bandwidth utilization reporting. A PCEP Speaker includes the "BANDWIDTH-UTILIZATION-CAPABILITY" TLV, in the OPEN Object to advertise its support for PCEP extension. The presence of the "BANDWIDTH-UTILIZATION-CAPABILITY" TLV in the OPEN Object (in the Open message) indicates that the bandwidth utilization reporting is supported as described in this document. Additional procedure is defined as following:

- The PCEP protocol extensions for bandwidth utilization MUST NOT be used if one or both PCEP Speakers have not included the "BANDWIDTH-UTILIZATION-CAPABILITY" TLV in their respective Open message.

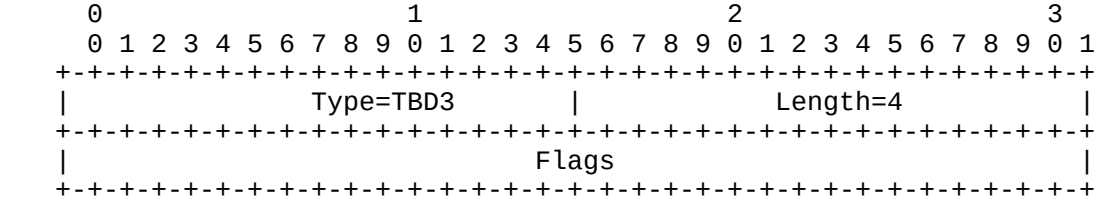
- If the PCEP speaker that supports the extensions of this document but did not advertise this capability, then upon receipt of BANDWIDTH-UTILIZATION-ATTRIBUTES TLV in LSPA object, it SHOULD generate a PCErr with error-type 19 (Invalid Operation), error-value TBD13 (Bandwidth utilization capability was not advertised) and it will terminate the PCEP session.

- If the PCEP speaker that supports the extensions of this document but did not advertise this capability, then upon receipt of BANDWIDTH object of type TBD14, it SHOULD generate a PCErr with error-type 19 (Invalid Operation), error-value TBD13 (Bandwidth

utilization capability was not advertised) and it will terminate the PCEP session.

7.1.1. BANDWIDTH-UTILIZATION-CAPABILITY TLV

The BANDWIDTH-UTILIZATION-CAPABILITY TLV is an optional TLV for use in the OPEN Object for Bandwidth Utilization reporting via PCEP capability advertisement. Its format is shown in the following figure:



BANDWIDTH-UTILIZATION-CAPABILITY TLV Format

The Type of the TLV is TBD3 and it has a fixed length of 4 bytes.

The value comprises a single field - Flags (32 bits). Currently no flags are defined for this TLV.

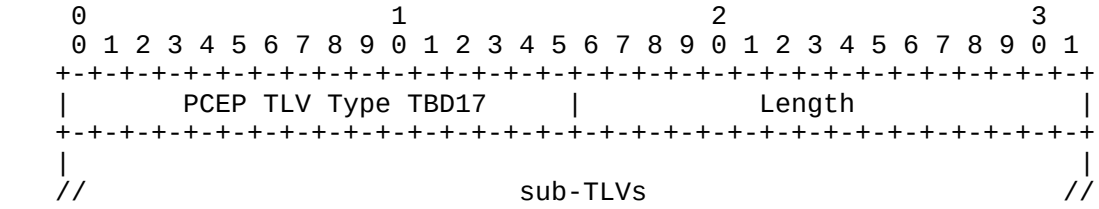
Unassigned bits are considered reserved. They MUST be set to 0 when sent and MUST be ignored when received.

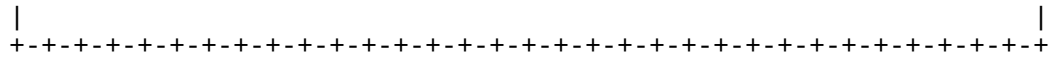
Advertisement of the BANDWIDTH-UTILIZATION-CAPABILITY TLV implies support of bandwidth utilization reporting, as well as the objects, TLVs and procedures defined in this document.

7.2. BW-UTILIZATION-MEASUREMENT-ATTRIBUTES TLV

The BW-UTILIZATION-MEASUREMENT-ATTRIBUTES TLV provides the configurable parameters of bandwidth utilization feature.

The format of the BW-UTILIZATION-MEASUREMENT-ATTRIBUTES TLV is shown in the following figure:





BW-UTILIZATION-MEASUREMENT-ATTRIBUTES TLV Format

PCEP TLV Type is defined as following:

Type	Name
TBD17	BW-UTILIZATION-MEASUREMENT-ATTRIBUTES

Length: The Length field defines the length of the value portion in bytes as per [RFC5440].

Value: Comprises of one or more sub-TLVs as described in Section 4 of this document.

The following sub-sections describe the parameters which are currently defined to be carried within this TLV.

7.2.1. Bandwidth Utilization Measurement Enable

The Measurement-Enable sub-TLV specifies that the bandwidth utilization reporting is enabled using following flag:

Bit	Description
26	Bandwidth Utilization Reporting Enabled

7.2.2. Bandwidth Utilization Measurement Interval

The Measurement-Interval sub-TLV specifies a time interval in seconds for the bandwidth samples collection interval.

7.2.3. Bandwidth Utilization Report Threshold

The Report-Threshold sub-TLV is used to decide if the bandwidth samples collected so far should be immediately reported bypassing the report-interval.

- Threshold: The absolute threshold bandwidth value in 32-bits, encoded in IEEE floating point format (see [IEEE.754.1985]), expressed in bytes per second.

7.2.4. Bandwidth Utilization Report Threshold Percentage

The Report-Threshold-Percentage sub-TLV is used to decide if the bandwidth samples collected so far should be immediately reported

bypassing the report-interval.

7.2.5. Bandwidth Utilization Report Interval

The Report-Interval sub-TLV specifies a time interval in seconds when the collected bandwidth samples are to be reported to PCE.

7.2.6. Bandwidth Utilization Upper Bound

The Report-Upper-Bound sub-TLV specifies the upper-bound bandwidth encoded in IEEE floating point format (see [IEEE.754.1985]), expressed in bytes per second, and is used to trigger an immediate reporting when crossed. This may also result in PCC taking an immediate local action on the LSP.

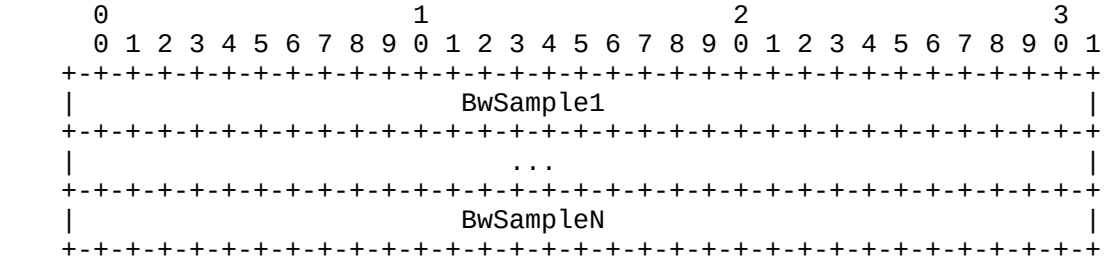
7.3. BANDWIDTH Object

A new object-type for BANDWIDTH object (Object-Class 5) is defined to report the bandwidth utilization of a TE LSP.

When the TE LSP is enabled with the bandwidth utilization reporting, the PCC SHOULD include the BANDWIDTH-UTILIZATION Object to report the bandwidth utilization of the TE LSP to the PCE in the PCRpt message.

The object-type is TBD14, the object length is variable with multiples of 4 bytes.

The object body format is defined as following:



BANDWIDTH-UTILIZATION Object Body Format

- BwSample: The utilized bandwidth, (the BwSample collected at the end of each measurement-interval) encoded in IEEE floating point format (see [IEEE.754.1985]), expressed in bytes per second.

8. PCEP Procedure

Following procedure is defined for the extensions to different PCEP messages for reporting performance measurements.

8.1. Various MEASUREMENT-ATTRIBUTES TLVs

- For a PCE-Initiated LSP [DRAFT-PCE-INITIATED-LSP] with reporting features enabled, the corresponding MEASUREMENT-ATTRIBUTES TLV for each measurement MUST be included in the LSPA Object with the PCInitiate message.

- For a PCE-Initiated LSP [DRAFT-PCE-INITIATED-LSP] with reporting features enabled, the corresponding MEASUREMENT-ATTRIBUTES TLV for each measurement is carried in the PCUpd message in the LSPA Object in order to make updates to the attributes such as Report-Interval.

- For a PCC-Initiated LSP with reporting features enabled, when the LSP is delegated to the PCE, the corresponding MEASUREMENT-ATTRIBUTES TLV for each measurement MUST be included in the LSPA Object of the PCRpt message.

- The various MEASUREMENT-ATTRIBUTES TLVs are encoded in all PCEP messages for the LSP with reporting features enabled, the absence of the corresponding MEASUREMENT-ATTRIBUTES TLV indicates that the PCEP speaker wishes to disable the feature.

8.2. The MEASUREMENT Objects

When a TE LSP is enabled with a measurement reporting feature, the PCC SHOULD include the corresponding MEASUREMENT Object to report the measured values to the PCE in the PCRpt message [DRAFT-PCE-STATEFUL].

The format of the "actual_attribute-list" in the PCRpt message is modified as following:

```
<actual_attribute-list>::=[<BANDWIDTH>]
                           [<DELAY-MEASUREMENT>]
                           [<LOSS-MEASUREMENT>]
                           [<metric-list>]
```

9. Scaling Considerations

It should be noted that when measurement reporting is deployed under LSP scale, it can lead to frequent reporting updates to the PCE. Operators are advised to set the values of various measurement reporting parameters appropriate for the deployed LSP scale.

If a PCE gets overwhelmed, it can notify the PCC to temporarily suspend the reporting of the measurements as described below.

9.1. The PCNtf Message

As per [RFC5440], the PCEP Notification message (PCNtf) can be sent by a PCEP speaker to notify its peer of a specific event. A PCEP speaker SHOULD notify its PCEP peer that it is overwhelmed, and on receipt of such notification the peer SHOULD NOT send any PCEP messages related to measurement reporting. If a PCEP message related to measurement reporting is received, it MUST be silently ignored.

- When a PCEP speaker is overwhelmed, it SHOULD notify its peer by sending a PCNtf message with Notification Type = TBD15 (PM Overwhelm State) and Notification Value = 1 (Entering PM overwhelm state).

- Optionally, OVERLOADED-DURATION TLV [RFC5440] MAY be included that specifies the time period during which no further PCEP messages related to PM should be sent.

- When the PCEP speaker is no longer in the overwhelm state and is available to process the PM reporting, it SHOULD notify its peer by sending a PCNtf message with Notification Type = TBD15 (PM Overwhelm State) and Notification Value = 2 (Clearing PM overwhelm state).

10. Security Considerations

This document defines new MEASUREMENT-ATTRIBUTES TLVs, CAPABILITY TLVs and MEASUREMENT Objects for reporting loss, delay measurements and bandwidth utilization which do not add additional security concerns beyond those discussed in [RFC5440] and [DRAFT-PCE-STATEFUL].

Some deployments may find the reporting of the performance measurement and bandwidth utilization information as extra sensitive as it could be used to influence LSP path computation and LSP setup with adverse effect. Additionally, snooping of PCEP messages with such data or using PCEP messages for network reconnaissance, may give an attacker sensitive information about the operations of the network. Thus, such deployment should employ suitable PCEP security mechanisms like TCP Authentication Option (TCP-AO) [RFC5925] or [DRAFT-PCE-PCEPS].

11. Manageability Considerations

11.1. Control of Function and Policy

The performance measurement reporting SHOULD be controlled per TE tunnel (at PCC or PCE) and the values for feature attributes e.g. measurement-interval, report-interval, report-threshold SHOULD be configurable by an operator.

11.2. Information and Data Models

A Management Information Base (MIB) module for modeling PCEP is described in [RFC7420]. However, one may prefer the mechanism for configuration using YANG data model [DRAFT-PCE-PCEP-YANG]. These SHOULD be enhanced to provide controls and indicators for support of performance measurement reporting feature. Support for various configuration knobs as well as counters of messages sent/received containing the TLVs (defined in this document) SHOULD be added.

11.3. 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].

11.4. Verify Correct Operations

Mechanisms defined in this document do not imply any new operation verification requirements in addition to those already listed in [RFC5440].

11.5. Requirements On Other Protocols

Mechanisms defined in this document do not add any new requirements on other protocols.

11.6. Impact On Network Operations

In order to avoid any unacceptable impact on network operations, an implementation SHOULD allow a limit to be placed on the number of LSPs that can be enabled with performance measurement reporting feature. An implementation MAY allow a limit to be placed on the rate of measurement reporting messages sent by a PCEP speaker and received by a peer. An implementation MAY also allow sending a notification when a PCEP speaker is overwhelmed or the rate of messages reach a threshold.

12. IANA Considerations

12.1. Measurement Capability TLV Types

This document defines the following new PCEP TLVs; IANA is requested to make the following allocations from the "PCEP TLV Type Indicators" registry. <http://www.iana.org/assignments/pcep/pcep.xhtml#pcep-tlv-type-indicators>

Type	Name	Reference
TBD1	DELAY-MEASUREMENT-CAPABILITY	[This document]
TBD2	LOSS-MEASUREMENT-CAPABILITY	[This document]
TBD3	BANDWIDTH-UTILIZATION-CAPABILITY	[This document]

12.1.1. Flag Fields for MEASUREMENT-CAPABILITY TLVs

IANA is requested to create a registry to manage the Flag field of the DELAY-MEASUREMENT-CAPABILITY TLV, LOSS-MEASUREMENT-CAPABILITY TLV and BANDWIDTH-UTILIZATION-CAPABILITY TLV.

New bit numbers are allocated only by an IETF Review 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 value are defined in this document for the Flag field for -

DELAY-MEASUREMENT-CAPABILITY TLV:

Bit	Description	Reference
31	One-way Delay Measurement	[This document]
30	Two-way Delay Measurement	[This document]

LOSS-MEASUREMENT-CAPABILITY TLV:

Bit	Description	Reference
31	Unidirectional Loss Measurement	[This document]
30	Bidirectional Loss Measurement	[This document]
29	Inferred Loss Measurement Mode	[This document]
28	Direct Loss Measurement Mode	[This document]

12.2. MEASUREMENT-ATTRIBUTES TLVs

This document defines the following new PCEP TLV Types; IANA is requested to make the following TLV type allocations from the "PCEP TLV Type Indicators" registry. <http://www.iana.org/assignments/pcep/pcep.xhtml#pcep-tlv-type-indicators>

Type	Name	Reference
TBD4	DELAY-MEASUREMENT-ATTRIBUTES	[This document]
TBD16	LOSS-MEASUREMENT-ATTRIBUTES	[This document]
TBD17	BW-UTILIZATION-MEASUREMENT-ATTRIBUTES	[This document]

12.2.1. The Sub-TLVs For MEASUREMENT-ATTRIBUTES TLVs

IANA is requested to create an "MEASUREMENT-ATTRIBUTES Sub-TLV Types" sub-registry in the "PCEP TLV Type Indicators" registry. New sub-TLV are allocated only by an IETF Review action [RFC5226].

This document defines the following sub-TLV types:

Type	Name	Reference
0	Reserved	[This document]
1	Measurement-Enable sub-TLV	[This document]
2	Measurement-Interval sub-TLV	[This document]
3	Report-Threshold sub-TLV	[This document]
4	Report-Threshold-Percentage sub-TLV	[This document]
5	Report-Interval sub-TLV	[This document]
6	Report-Upper-Bound sub-TLV	[This document]
7-65535	Unassigned	[This document]

12.2.1.1. Flag Fields in Measurement-Enable sub-TLV

IANA is requested to create a registry to manage the Flag field of the Measurement-Enable sub-TLV.

New bit numbers are allocated only by an IETF Review 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 value are defined in this document for the Flag field.

Bit	Description	Reference
31	One-Way Delay Measurement Enabled	[This document]
30	Two-Way Delay Measurement Enabled	[This document]
29	Unidirectional Loss Measurement Enabled	[This document]
28	Bidirectional Loss Measurement Enabled	[This document]
27	Inferred Loss Measurement Enabled	[This document]
26	Bandwidth Utilization Reporting Enabled	[This document]

12.3. Measurement Object-Class

This document defines Object-Class for the following Objects; IANA is requested to make the following allocations from the "PCEP Objects" registry. <http://www.iana.org/assignments/pcep/pcep.xhtml#pcep-objects>

Object-Class	Name	Reference
TBD5	DELAY-MEASUREMENT Object	[This document]
TBD6	LOSS-MEASUREMENT Object	[This document]

12.3.1. DELAY-MEASUREMENT Object-Types

IANA is requested to create an "DELAY-MEASUREMENT Object-Types" sub-registry for DELAY-MEASUREMENT Object (Object-class TBD5).

This document defines the following object-types:

Object-Type	Name	Reference
0	Reserved	[This document]
1	One-Way Delay Measurement Value	[This document]
2	One-Way Delay Measurement Min/Max Values	[This document]
3	One-Way Delay Variation Measurement Value	[This document]
4	Two-Way Delay Measurement Value	[This document]
5	Two-Way Delay Measurement Min/Max Values	[This document]
6	Two-Way Delay Variation Measurement Value	[This document]

12.3.2. LOSS-MEASUREMENT Object-Types

IANA is requested to create an "LOSS-MEASUREMENT Object-Types" sub-registry for LOSS-MEASUREMENT Object (Object-class TBD6).

This document defines the following object-types:

Object-Type	Name	Reference
0	Reserved	[This document]
1	Tx Packets-Lost	[This document]
2	Rx Packets-Lost	[This document]

12.3.3. BANDWIDTH Object-Type

This document defines new Object-Type for the BANDWIDTH object (Object-Class 5, [RFC5440]); IANA is requested to make the following allocation from the "PCEP Objects" registry. <http://www.iana.org/assignments/pcep/pcep.xhtml#pcep-objects>

Object-Type	Name	Reference
TBD14	BANDWIDTH-UTILIZATION Object	[This document]

12.4. PCE Error Codes

This document defines two new error-values for PCErr with error-code 19 (Invalid Operation). IANA is requested to make the following allocations.

Error-Value	Name	Reference
TBD7	Delay-Measurement capability was not advertised	[This document]
TBD8	Loss-Measurement capability was not advertised	[This document]
TBD9	Bidirectional Measurement capability was not advertised	[This document]
TBD10	Unidirectional Measurement capability was not advertised	[This document]
TBD11	Inferred Mode Loss Measurement capability was not advertised	[This document]
TBD12	Direct Mode Loss Measurement capability was not advertised	[This document]
TBD13	Bandwidth Utilization capability was not advertised	[This document]

12.5. Notification Object-Type

IANA is requested to allocate new Notification Types and Notification Values within the "Notification Object" sub-registry of the PCEP Numbers registry, as follows:

Type	Meaning	Reference

TBD15	PM Overwhelm State	[This document]

Notification-value=1: Entering PM overwhelm state
Notification-value=2: Clearing PM overwhelm state

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.

[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.

[DRAFT-PCE-STATEFUL] Crabbe, E., Minei, I., Medved, J., and R. Varga, "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce, (work in progress).

[DRAFT-PCE-INITIATED-LSP] Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model", draft-ietf-pce-pce-initiated-lsp, (work in progress).

13.2. Informative References

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

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

[RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, June 2010.

[RFC6374] Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks", DOI 10.17487/RFC6374, RFC 6374, September 2011.

[RFC6375] Frost, D. and S. Bryant, "A Packet Loss and Delay Measurement Profile for MPLS-Based Transport Networks", RFC 6375, September 2011.

[RFC7420] Koushik, A., Stephan, E., Zhao, Q., King, D., and J. Hardwick, "Path Computation Element Communication Protocol (PCEP) Management Information Base (MIB) Module", RFC 7420, December 2014.

[RFC7471] S. Giacalone, D. Ward, J. Drake, A. Atlas, and S. Previdi, "OSPF Traffic Engineering (TE) Metric Extensions", RFC 7471, March 2015.

[RFC7810] S. Previdi, S. Giacalone, D. Ward, J. Drake, and Q. Wu, "IS-IS Traffic Engineering (TE) Metric Extensions", RFC 7810, May 2016.

[RFC7823] Atlas, A., Drake, J., Giacalone, S., and S. Previdi, "Performance-Based Path Selection for Explicitly Routed Label Switched Paths (LSPs) Using TE Metric Extensions", RFC 7823, May 2016.

[RFC7876] Bryant, S., Sivabalan, S., and Soni, S., "UDP Return Path for Packet Loss and Delay Measurement for MPLS Networks", RFC 7876, July 2016.

[DRAFT-PCE-PCEPS] Lopez, D., Dios, O., Wu, W., and D. Dhody, "Secure Transport for PCEP", draft-ietf-pce-pceps, (work in progress).

[DRAFT-PCE-SERVICE-AWARE] Dhody, D., V. Manral, V., Ali, Z., and Kumaki, K., "Extensions to the Path Computation Element Communication Protocol (PCEP) to compute service aware Label Switched Path (LSP)", draft-ietf-pce-pcep-service-aware, (work in progress).

[DRAFT-IDR-TE-PM-BGP] Wu, Q., Danhua, W., Previdi, S., Gredler, H., and S. Ray, "BGP attribute for North-Bound Distribution of Traffic Engineering (TE) performance Metrics", draft-ietf-idr-te-pm-bgp (work in progress).

[DRAFT-PCE-PCEP-YANG] Dhody, D., Hardwick, J., Beeram, V., and J. Tantsura, "A YANG Data Model for Path Computation Element Communications Protocol (PCEP)", draft-ietf-pce-pcep-yang (work in progress).

[DRAFT-IETF-PCE-AUTOBW] Dhody, D., Palle, U., Singh, R., Gandhi, R., and L. Fang, "PCEP Extensions for MPLS-TE LSP Automatic Bandwidth Adjustment with Stateful PCE", draft-ietf-pce-stateful-pce-auto-bandwidth (work in progress).

[IEEE.754.1985] Institute of Electrical and Electronics Engineers, "Standard for Binary Floating-Point Arithmetic", IEEE Standard 754, August 1985.

Acknowledgments

TBA.

Authors' Addresses

Rakesh Gandhi
Cisco Systems, Inc.

EMail: rgandhi@cisco.com

Bin Wen
Comcast

EMail: Bin_Wen@cable.comcast.com

Colby Barth
Juniper Networks

EMail: cbarth@juniper.net

Dhruv Dhody
Huawei Technologies
India

EMail: dhruv.ietf@gmail.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: March 15, 2018

S. Litkowski
Orange
S. Sivabalan
Cisco Systems, Inc.
C. Barth
Juniper Networks
D. Dhody
Huawei

September 11, 2017

Path Computation Element communication Protocol extension for signaling LSP diversity constraint
draft-ietf-pce-association-diversity-02

Abstract

This document introduces a simple mechanism to associate a group of Label Switched Paths (LSPs) via an extension to the Path Computation Element Communication Protocol (PCEP) with the purpose of computing diverse paths for those LSPs. The proposed extension allows a PCC to advertise to a PCE the belonging of a particular LSP to a disjoint-group, thus the PCE knows that LSPs in the same group must be disjoint from each other.

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 March 15, 2018.

Table of Contents

1. Introduction

[RFC5440] describes the Path Computation Element communication Protocol (PCEP) which enables the communication between a Path Computation Client (PCC) and a Path Control Element (PCE), or between two PCEs based on the PCE architecture [RFC4655].

PCEP Extensions for Stateful PCE Model [I-D.ietf-pce-stateful-pce] describes a set of extensions to PCEP to enable active control of MPLS-TE and GMPLS tunnels. [I-D.ietf-pce-pce-initiated-lsp] describes the setup and teardown of PCE-initiated LSPs under the active stateful PCE model, without the need for local configuration on the PCC, thus allowing for a dynamic network.

[I-D.ietf-pce-association-group] introduces a generic mechanism to create a grouping of LSPs which can then be used to define associations between a set of LSPs and a set of attributes (such as configuration parameters or behaviours) and is equally applicable to the active and passive modes of a stateful PCE [I-D.ietf-pce-stateful-pce] or a stateless PCE [RFC5440].

This document specifies a PCEP extension to signal that a particular group of LSPs should use diverse paths including the requested type of diversity. A PCC can use this extension to signal to a PCE the belonging of a particular LSP to a disjoint-group. When a PCE receives LSP states belonging to the same disjoint-group from some PCCs, the PCE should ensure that the LSPs within the group are disjoint from each other.

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.

LSR: Label Switch Router.

MPLS: Multiprotocol Label Switching.

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.

PCEP: Path Computation Element Communication Protocol.

SRLG: Shared Risk Link Group.

3. Motivation

Path diversity is a very common use case today in IP/MPLS networks especially for layer 2 transport over MPLS. A customer may request that the operator provide two end-to-end disjoint paths across the IP/MPLS core. The customer may use those paths as primary/backup or active/active.

Different level of disjointness may be offered:

- Link disjointness: the paths of the associated LSPs should transit different links (but may use common nodes or different links that may have some shared fate).
- Node disjointness: the paths of the associated LSPs should transit different nodes (but may use different links that may have some shared fate).
- SRLG disjointness: the paths of the associated LSPs should transit different links that do not share fate (but may use common transit nodes).
- Node+SRLG disjointness: the paths of the associated LSPs should transit different links that do not have any common shared fate and should transit different nodes.

The associated LSPs may originate from the same or from different head-end(s) and may terminate at the same or different tail-end(s).

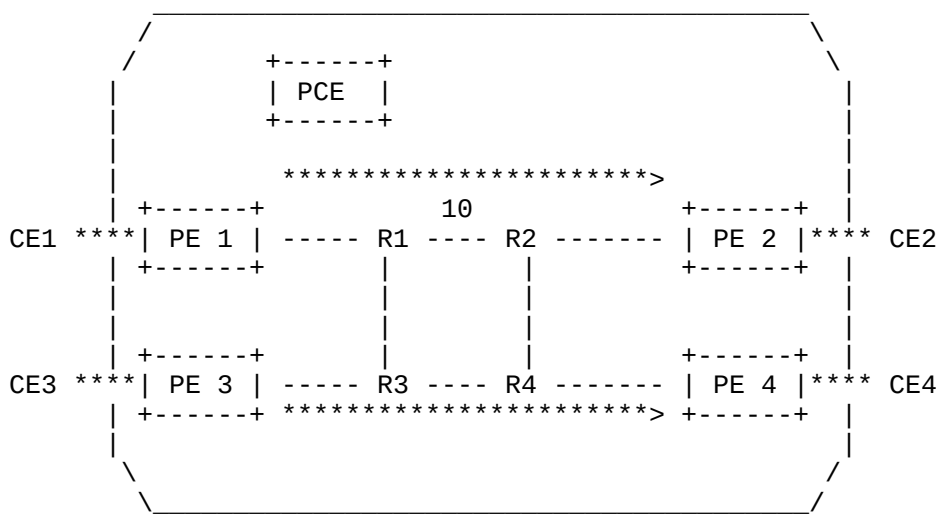


Figure 1 - Disjoint paths with different head-ends and tail-ends

In the figure above, the customer wants to have two disjoint paths between CE1/CE2 and CE3/CE4. From an IP/MPLS network point view, in this example, the CEs are connected to different PEs to maximize their disjointness. When LSPs originate from different head-ends, distributed computation of diverse paths can be difficult. Whereas, computation via a centralized PCE ensures path disjointness correctness and simplicity.

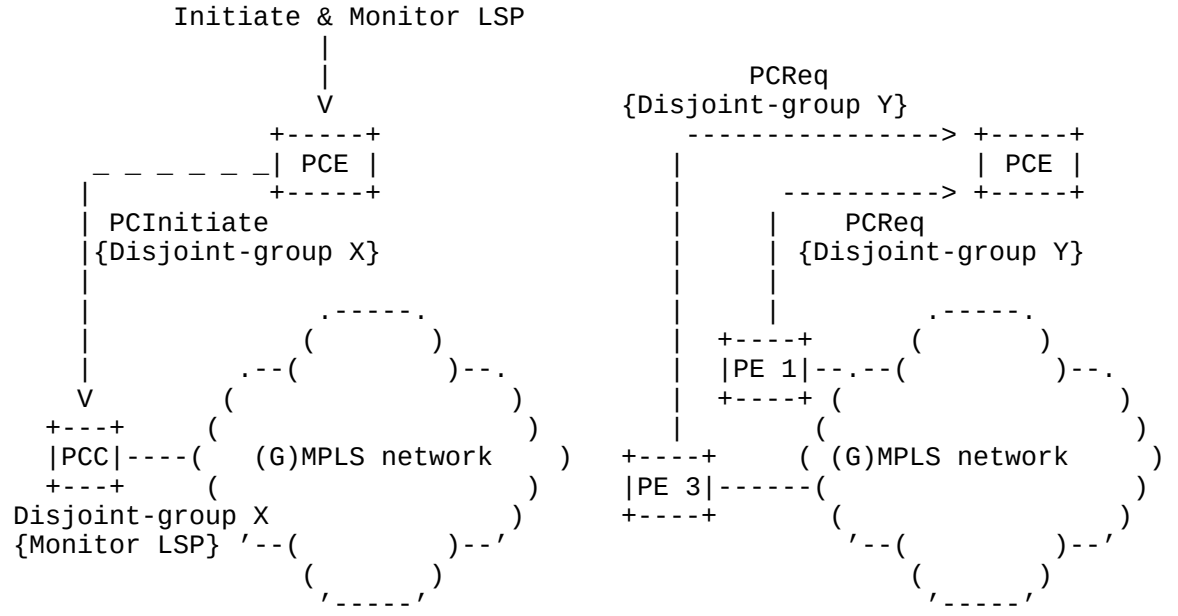
The SVEC (Synchronization VECtor) object [RFC5440] allows a PCC to request the synchronization of a set of dependent or independent path computation requests. As per [RFC5440], the SVEC object is optional and may be carried within a PCReq message. It uses the Request-ID-number carried within the respective RP (Request Parameters) object to identify the computation request that should be syncronized.

The PCEP extension for stateful PCE [I-D.ietf-pce-stateful-pce] defined new PCEP messages - PCRpt, PCUpd and PCInitiate. These messages uses PLSP-ID in the LSP object for identification. Moreover to allow diversity between LSPs originating from different PCCs, the generic mechanism to create a grouping of LSPs as described in [I-D.ietf-pce-association-group] equally applicable to the active and passive modes of a stateful PCE is better suited.

Using PCEP, the PCC MUST indicate that disjoint path computation is required, such indication SHOULD include disjointness parameters such as the type of disjointness, the disjoint group-id, and any

customization parameters according to the configured local policy. As mentioned previously, the extension described in [I-D.ietf-pce-association-group] is well suited to associated a set of LSPs with a particular disjoint-group.

The management of the disjoint group-ids will be a key point for the operator as the Association ID field is limited to 65535. The local configuration of IPv4/IPv6 association source, or Global Association Source/Extended Association ID should allow to overcome this limitation. For example, when a PCC or PCE initiates all the LSPs in a particular disjoint-group, it can set the IPv4/IPv6 association source can be set to one of its IP address. When disjoint LSPs are initiated from different head-ends, a unique association identifier SHOULD be used for those LSPs: this can be achieved by setting the IPv4/IPv6 source address to a common value (zero value can be used) as well as the Association ID.



Case 1: Disjointness initiated by PCE and enforced by PCC

Case 2: Disjointness initiated by PCC and enforced by PCE

Figure 2 - Sample use-cases for carrying disjoint-group over PCEP session

Using the disjoint-group within a PCUpd or PCInitiate may have two purposes:

- Information: in case the PCE is performing the path computation, it may communicate to the PCC the locally used parameters in the attribute-list of the LSP.

- Configuration: in case the PCE is updating the diversity parameters.

4. Protocol extension

4.1. Association group

As per [I-D.ietf-pce-association-group], LSPs are associated with other LSPs with which they interact by adding them to a common association group. The Association ID will be used to identify the disjoint group a set of LSPs belongs to. This document defines a new Association type, based on the generic Association object -

- Association type = TBD1 ("Disjointness Association Type").

A disjoint group can have two or more LSPs. But a PCE may be limited in how many LSPs it can take into account when computing disjointness. If a PCE receives more LSPs in the group than it can handle in its computation algorithm, it SHOULD apply disjointness computation to only a subset of LSPs in the group. The subset of disjoint LSPs will be decided by the implementation.

Local polices on the PCC or PCE MAY define the computational behavior for the other LSPs in the group. For example, the PCE may provide no path, a shortest path, or a constrained path based on relaxing disjointness, etc.

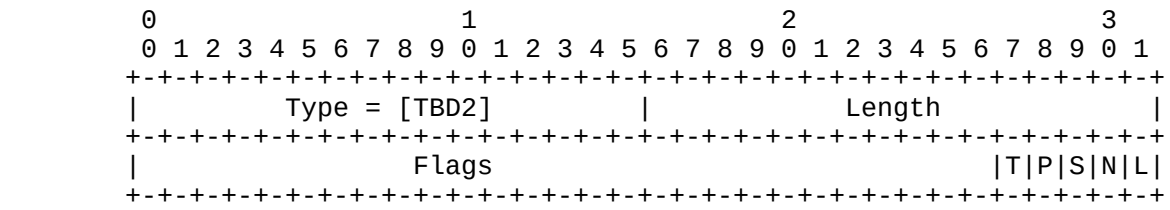
Associating a particular LSP to multiple disjoint groups is authorized from a protocol perspective, however there is no insurance that the PCE will be able to compute properly the multi-disjointness constraint.

4.2. Mandatory TLV

The disjoint group MUST carry the following TLV:

- DISJOINTNESS-INFORMATION-TLV: Used to communicate some disjointness specific parameters.

The DISJOINTNESS-INFORMATION-TLV is shown in the following figure:



Flags:

* L (Link diverse) bit: when set, this indicates that the computed paths within the disjoint group MUST NOT have any link in common.

* N (Node diverse) bit: when set, this indicates that the computed paths within the disjoint group MUST NOT have any node in common.

* S (SRLG diverse) bit: when set, this indicates that the computed paths within the disjoint group MUST NOT share any SRLG (Shared Risk Link Group).

* P (Shortest path) bit: when set, this indicates that the computed path of the LSP SHOULD satisfies all constraints and objective functions first without considering the diversity constraint. This means that an LSP with P flag set should be placed as if the disjointness constraint has not been configured, while the other LSP in the association with P flag unset should be placed by taking into account the disjointness constraint. Setting P flag changes the relationship between LSPs to a unidirectional relationship (LSP 1 with P=0 depends of LSP 2 with P=1, but LSP 2 with P=1 does not depend of LSP 1 with P=0).

* T (Strict disjointness) bit: when set, if disjoint paths cannot be found, PCE should return no path for LSPs that could not be be disjoint. When unset, PCE is allowed to relax disjointness by using either applying a requested objective function or any other behavior if no objective function is requested (e.g.: using a lower disjoint type (link instead of node) or relaxing disjointness constraint at all).

If a PCEP speaker receives a disjoint-group without DISJOINTNESS-INFORMATION-TLV, it SHOULD reply with a PCErr Error-type=6 (Mandatory Object missing) and Error-value=TBD7.

4.3. Optional TLVs

The disjoint group MAY carry some optional TLVs including but not limited to:

- VENDOR-INFORMATION-TLV: Used to communicate arbitrary vendor specific behavioral information, described in [RFC7150].

4.4. Disjointness objective functions

An objective function MAY be applied to the disjointness computation to drive the PCE computation behavior. In this case, the OF-List TLV (defined in ([RFC5541]) is used as an optional TLV in the Association Group Object. The PCEP OF-List TLV allow multiple OF-Codes inside the TLV, a sender SHOULD include a single OF-Code in the OF-List TLV when included in the Association Group, and the receiver MUST consider the first OF-code only and ignore others if included. The OF-Code to maximize diversity are specified in ([I-D.dhody-pce-of-diverse]).

4.5. P-flag considerations

As mentioned in Section 4.2, the P-flag (when set) indicates that the computed path of the LSP SHOULD satisfies all constraints and objective functions first without considering the diversity constraint. This could be required in some primary/backup scenarios where the primary path should use the more optimal path available (taking into account the other constraints). When disjointness is computed, it is important for the algorithm to know that it should try to optimize the path of one or more LSPs in the disjoint group (for instance the primary path) while other paths are allowed to be longer (compared to a similar path without the disjointness constraint). Without such a hint, the disjointness algorithm may set a path for all LSPs that may not completely fulfil the customer requirement.

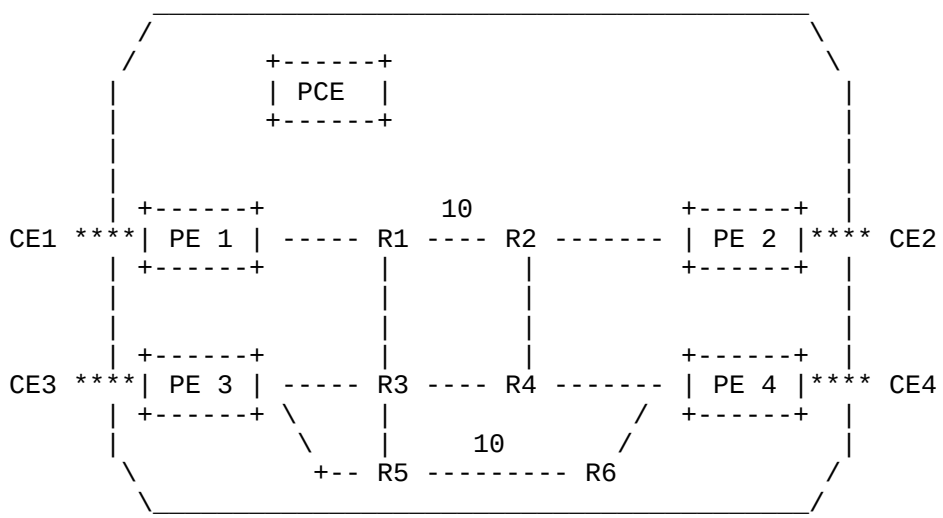


Figure 3

In the figure above, a customer has two dual homed sites (CE1/CE3 and CE2/CE4). This customer wants two disjoint paths between the two sites. Due to physical meshings, the customer wants to use CE1 and CE2 as primary (for instance CE3 and CE4 are hosted in a remote site for redundancy purpose compared the customer hosts).

Without any hint (constraint) provided, the PCE may compute the two disjoint LSPs as a batch leading to PE1->PE2 using a path PE1->R1->R2->PE2 and PE3->PE4 using PE3->R3->R4->PE4. In this case, even if the disjointness constraint is fulfilled, the path from PE1 to PE2 does not use the best optimal path available in the network (RTD may be higher): the customer requirement is thus not completely fulfilled.

The usage of the P-Flag allows the PCE to know that a particular LSP should be tied to the best path as if the disjointness constraint was not requested.

In our example, if the P-Flag is set to the LSP PE1->PE2, the PCE should use the path PE1->R1->R3->R4->R2->PE2 for this LSP, while the other LSP should be disjoint from this path. The second LSP will be placed on PE3->R5->R6->PE4 as it is allowed to be longer.

Driving the PCE disjointness computation may be done in other ways by for instance setting a metric boundary reflecting an RTD boundary. Other constraints may also be used.

The P-Flag allows a simple expression that the disjointness constraint should not make the LSP worst.

Any constraint added to a path disjointness computation may reduce the chance to find suitable paths. The usage of the P-flag, as any other constraint, may prevent to find a disjoint path. In the example above, if we consider that the router R5 is down, if PE1->PE2 has the P-flag set, there is no room available to place PE3->PE4 (the disjointness constraint cannot be fulfilled). If PE->PE2 has the P-flag unset, the algorithm may be able to place PE1->PE2 on R1->R2 link leaving a room for PE3->PE4 using the R3->R4 link. When using P-flag or any additional constraint on top of the disjointness constraint, the user should be aware that there is less chance to fulfill the disjointness constraint.

Multiple LSPs in the same disjoint group may have the P-flag set. In such a case, those LSPs may not be disjoint from each other but will be disjoint from others LSPs in the group that have the P-flag unset.

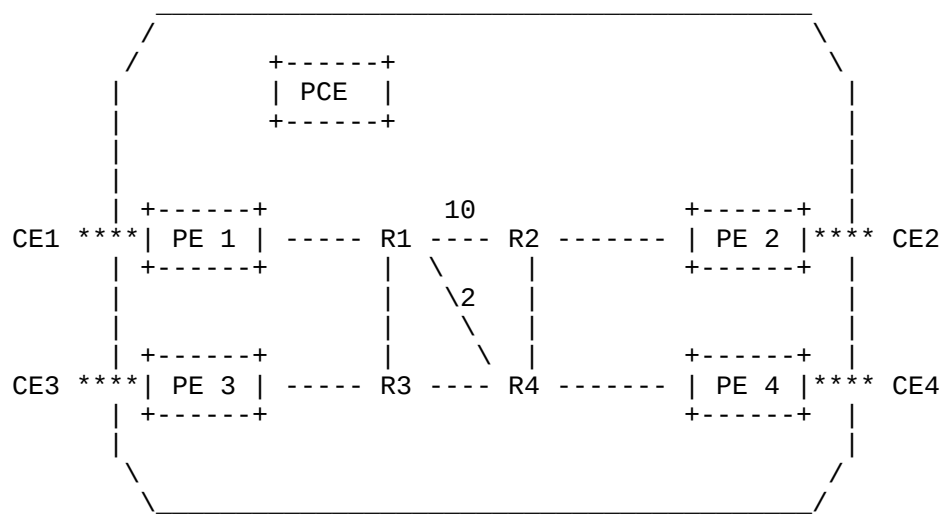


Figure 4

In the figure above, we still consider the same previous requirements, so PE1->PE2 LSP should be optimized (P-flag set) while PE3->PE4 should be disjoint and may use a longer path.

Regarding PE1->PE2, there are two paths that are satisfying the constraints (ECMP): PE1->R1->R4->R2->PE2 (path 1) and PE1->R1->R3->R4->R2->PE2 (path 2). An implementation may choose one

of the paths or even use boths (using both may happen in case Segment Routing TE is used, allowing ECMP).

If the implementation elects only one path, there is a chance that picking up one path may prevent disjointness. In our example, if path 2 is used for PE1->PE2, there is no room left for PE3->PE4 while if path 1 is used, PE3->PE4 can be placed on R3->R4 link.

When P-flag is set for an LSP and when ECMPs are available, an implementation MAY select a path that allows disjointness.

4.6. Disjointness computation issues

There may be some cases where the PCE is not able to provide a set of disjoint paths for one or more LSPs in the association.

When the T-bit is set (Strict disjointness requested), if disjointness cannot be found for one or more LSPs, the PCE SHOULD reply with a PCUpd message containing an empty ERO. In addition to the empty ERO Object, the PCE MAY add the NO-PATH-VECTOR TLV ([RFC5440]) in the LSP Object.

This document adds new bits in the NO-PATH-VECTOR TLV:

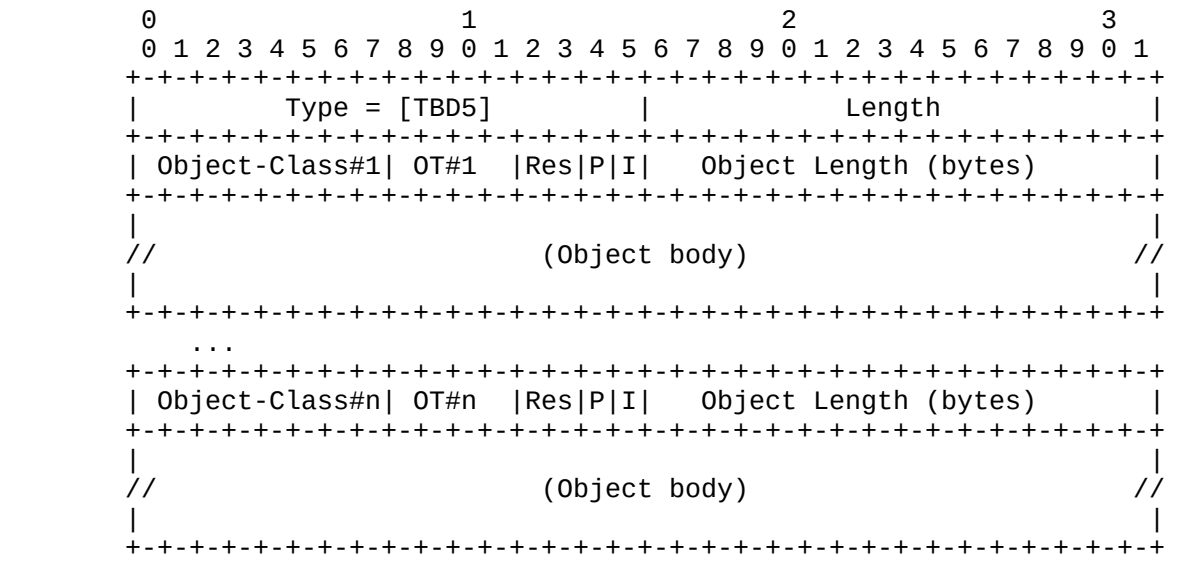
- bit "TBD3": when set, the PCE indicates that it could not find a disjoint path for this LSP.

- bit "TBD4": when set, the PCE indicates that it does not support the requested disjointness computation.

The NO-PATH-VECTOR TLV MAY also be used when T-bit is unset and when the PCE cannot provide a path for an LSP in the disjoint group.

When the T-bit is unset, the PCE is allowed to relax the constraint if it cannot be satisfied. This document introduces a new RELAXED-CONSTRAINT-TLV that MAY be used by a PCE to indicate that some constraints have been relaxed.

When used, the RELAXED-CONSTRAINT-TLV SHOULD be included in the LSP Object of a PCUpd message. The RELAXED-CONSTRAINT-TLV has the following format:



All LSPs in a particular disjoint group MUST use the same combination of T,S,N,L flags in the DISJOINTNESS-INFORMATION-TLV. If a PCE receives PCRpt messages for LSPs belonging to the same disjoint group but having an inconsistent combination of T,S,N,L flags, the PCE SHOULD NOT try to compute disjointness path and SHOULD reply a PCErr with Error-type 5 (Policy Violation) and Error-Value TBD6 (Inconsistent parameters in DISJOINTNESS-INFORMATION TLV) to all PCCs involved in the disjoint group.

5. Security Considerations

This document defines one new type for association, which do not add any new security concerns beyond those discussed in [RFC5440], [I-D.ietf-pce-stateful-pce] and [I-D.ietf-pce-association-group] in itself.

6. IANA Considerations

6.1. Association object Type Indicators

This document defines the following new association type originally defined in [I-D.ietf-pce-association-group].

Value	Name	Reference
TBD1	Disjoint-group Association Type	[This I.D.]

6.2. PCEP TLVs

This document defines the following new PCEP TLVs:

Value	Name	Reference
TBD2	DISJOINTNESS-INFORMATION-TLV	[This I.D.]
TBD5	RELAXED-CONSTRAINT-TLV	[This I.D.]

IANA is requested to manage the space of flags carried in the DISJOINTNESS-INFORMATION TLV defined in this document, numbering them from 0 as the least significant bit.

New bit numbers may be allocated in future.

IANA is requested to allocate the following bit numbers in the DISJOINTNESS-INFORMATION-TLV flag space:

Bit Number	Name	Reference
0	Link disjointness	[This I.D.]
1	Node disjointness	[This I.D.]
2	SRLG disjointness	[This I.D.]
3	Shortest-path	[This I.D.]
4	Strict disjointness	[This I.D.]

6.3. NO-PATH-VECTOR bit Flags

This documents defines new bits for the NO-PATH-VECTOR TLV in the "NO-PATH-VECTOR TLV Flag Field" sub-registry of the "Path Computation Element Protocol (PCEP) Numbers" registry:

Bit Number	Name	Reference
TBD3	Disjoint path not found	[This I.D.]
TBD4	Requested disjointness computation not supported	[This I.D.]

6.4. PCEP-ERROR codes

IANA is requested to allocate new Error Types and Error Values within the " PCEP-ERROR Object Error Types and Values" sub-registry of the PCEP Numbers registry, as follows:

Error-Type	Meaning
5	Policy violation Error-value=TBD6: Inconsistent parameters in DISJOINTNESS-INFORMATION TLV
6	Mandatory Object missing Error-value=TBD7: DISJOINTNESS-INFORMATION TLV missing

7. Manageability Considerations

7.1. Control of Function and Policy

An operator MUST be allowed to configure the disjointness associations and parameters at PCEP peers and associate it with the LSPs.

7.2. Information and Data Models

[RFC7420] describes the PCEP MIB, there are no new MIB Objects for this document.

7.3. 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].

7.4. Verify Correct Operations

Mechanisms defined in this document do not imply any new operation verification requirements in addition to those already listed in [RFC5440].

7.5. Requirements On Other Protocols

Mechanisms defined in this document do not imply any new requirements on other protocols.

7.6. 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].

8. Acknowledgments

A special thanks to author of [I-D.ietf-pce-association-group], this document borrow some of the text from it. Authors would also like to thank Adrian Farrel for his useful comments.

9. References

9.1. 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>.

[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>.

[I-D.ietf-pce-association-group] Minei, I., Crabbe, E., Sivabalan, S., Ananthakrishnan, H., Dhody, D., and Y. Tanaka, "PCEP Extensions for Establishing Relationships Between Sets of LSPs", draft-ietf-pce-association-group-04 (work in progress), August 2017.

[I-D.ietf-pce-stateful-pce] Crabbe, E., Minei, I., Medved, J., and R. Varga, "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce-21 (work in progress), June 2017.

[I-D.dhody-pce-of-diverse] Dhody, D. and Q. Wu, "PCE support for Maximizing Diversity", draft-dhody-pce-of-diverse-07 (work in progress), May 2017.

9.2. Informative References

[RFC7150] Zhang, F. and A. Farrel, "Conveying Vendor-Specific Constraints in the Path Computation Element Communication Protocol", RFC 7150, DOI 10.17487/RFC7150, March 2014, <https://www.rfc-editor.org/info/rfc7150>.

[RFC7420] Koushik, A., Stephan, E., Zhao, Q., King, D., and J. Hardwick, "Path Computation Element Communication Protocol (PCEP) Management Information Base (MIB) Module", RFC 7420, DOI 10.17487/RFC7420, December 2014, <https://www.rfc-editor.org/info/rfc7420>.

[I-D.ietf-pce-pce-initiated-lsp] Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model", draft-ietf-pce-pce-initiated-lsp-10 (work in progress), June 2017.

Authors' Addresses

Stephane Litkowski
Orange

EMail: stephane.litkowski@orange.com

Siva Sivabalan
Cisco Systems, Inc.
2000 Innovation Drive
Kanata, Ontario K2K 3E8
Canada

EMail: msiva@cisco.com

Colby Barth
Juniper Networks

EMail: cbarth@juniper.net

Dhruv Dhody
Huawei

EMail: dhruv.dhody@huawei.com

PCE Working Group

Young Lee
Haomian Zheng
Huawei

Internet Draft
Intended Status: Standard
Expires: March 2018

Daniele Ceccarelli
Ericsson

Wei Wang
Beijing Univ. of Posts and Telecom

Peter Park
KT

Bin Young Yoon
ETRI

September 5, 2017

PCEP Extension for Distribution of Link-State and TE information for Optical Networks

draft-lee-pce-pcep-ls-optical-03

Abstract

In order to compute and provide optimal paths, Path Computation Elements (PCEs) require an accurate and timely Traffic Engineering Database (TED). Traditionally this Link State and TE information has been obtained from a link state routing protocol (supporting traffic engineering extensions).

This document extends the Path Communication Element Communication Protocol (PCEP)_ with Link-State and TE information for optical networks.

Status of this Memo

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

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

other groups may also distribute working documents as Internet-Drafts.

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

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/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 5, 2018.

Table of Contents

In Multiprotocol Label Switching (MPLS) and Generalized MPLS (GMPLS), a Traffic Engineering Database (TED) is used in computing paths for connection oriented packet services and for circuits. The TED contains all relevant information that a Path Computation Element (PCE) needs to perform its computations. It is important that the TED should be complete and accurate anytime so that the PCE can perform path computations.

In MPLS and GMPLS networks, Interior Gateway routing Protocols (IGPs) have been used to create and maintain a copy of the TED at each node. One of the benefits of the PCE architecture [RFC4655] is the use of computationally more sophisticated path computation algorithms and the realization that these may need enhanced processing power not necessarily available at each node participating in an IGP.

Section 4.3 of [RFC4655] describes the potential load of the TED on a network node and proposes an architecture where the TED is maintained by the PCE rather than the network nodes. However it does not describe how a PCE would obtain the information needed to populate its TED. PCE may construct its TED by participating in the IGP ([RFC3630] and [RFC5305] for MPLS-TE; [RFC4203] and [RFC5307] for GMPLS). An alternative is offered by [BGP-LS].

[RFC7399] touches upon this issue: "It has also been proposed that the PCE Communication Protocol (PCEP) [RFC5440] could be extended to serve as an information collection protocol to supply information from network devices to a PCE. The logic is that the network devices may already speak PCEP and so the protocol could easily be used to report details about the resources and state in the network, including the LSP state discussed in Sections 14 and 15."

[Stateful-PCE] describes a set of extensions to PCEP to provide stateful control. A stateful PCE has access to not only the information carried by the network's Interior Gateway Protocol (IGP), but also the set of active paths and their reserved resources for its computations. PCC can delegate the rights to modify the LSP parameters to an Active Stateful PCE. This requires PCE to quickly be updated on any changes in the Topology and TEDB, so that PCE can meet the need for updating LSPs effectively and in a timely manner. The fastest way for a PCE to be updated on TED changes is via a direct interface with each network node and with incremental update from each network node.

[PCE-initiated] describes the setup, maintenance and teardown of PCE-initiated LSPs under the stateful PCE model, without the need for local configuration on the PCC, thus allowing for a dynamic network that is centrally controlled and deployed. This model requires timely topology and TED update at the PCE.

[PCEP-LS-Arch] proposes alternative architecture approaches for learning and maintaining the Link State (and TE) information directly on a PCE from network nodes as an alternative to IGPs and BGP transport and investigate the impact from the PCE, routing protocol, and network node perspectives.

[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 (H-PCE) architecture [RFC6805], the Parent PCE (P-PCE) is used to compute a multi-domain path based on the domain connectivity information. A Child PCE (C-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.

[Stateful H-PCE] presents general considerations for stateful PCE(s) in hierarchical PCE architecture. In particular, the behavior changes and additions to the existing stateful PCE mechanisms (including PCE-initiated LSP setup and active PCE usage) in the context of networks using the H-PCE architecture.

[PCEP-LS] describes a mechanism by which Link State and TE information can be collected from packet networks and shared with PCE with the PCEP itself. This is achieved using a new PCEP message format.

This draft describes an optical extension of [PCEP-LS] and explains how encodings suggested by [PCEP-LS] can be used in the optical network contexts.

2. Applicability

There are three main applicability of this alternative proposed by this draft:

- Case 1: Where there is IGP running in optical network but there is a need for a faster link-state and TE resource collection at the PCE directly from an optical node (PCC) via a PCC-PCE interface.

  o A PCE may receive an incremental update (as opposed to the entire TE information of the node/link).

  Note: A PCE may receive full information from IGP using existing mechanism. In some cases, the convergence of full link-state and TE resource information of the entire network may not be appropriate for certain applications. Incremental update capability will enhance the accuracy of the TE information at a given time.

- Case 2: Where there is no IGP running in the optical network and there is a need for link-state and TE resource collections at the PCE directly from an optical node (PCC) via a PCC-PCE interface.

- Case 3: Where there is a need for transporting abstract optical link-state and TE information from child PCE and to a parent PCE in H-PCE [RFC6805] and [Stateful H-PCE] as well as for Physical Network Controller (PNC) to Multi-Domain Service Coordinator (MDSC) in Abstraction and Control of TE Networks (ACTN) [ACTN-Frame].

  Note: The applicability for Case 3 may arise as a consequence of Case 1 and Case 2. When TE information changes occur in the optical network, this may also affect abstracted TE information and thus needs to be updated to Parent PCE/MSDC from each child PCE/PNC.

3. Requirements for PCEP extension

The key requirements associated with link-state (and TE) distribution are identified for PCEP and listed in Section 4 of [PCEP-LS]. These new functions required in PCEP to support distribution of link-state (and TE) information are described in Section 5 of [PCEP-LS]. Details of PCEP messages and related Objects/TLVs are specified in Sections 8 and 9 of [PCEP-LS]. The key

requirements and new functions specified in [PCEP-LS] are equally applicable to optical networks.

Besides the generic requirements specified in [PCEP-LS], optical specific features also need to be considered in this document. As connection-based network, there are specific parameters such as reachable table, optical latency, wavelength consistency and some other parameters that need to be included during the topology collection. Without these restrictions, the path computation may be inaccurate or infeasible for deployment, therefore these information MUST be included in the PCEP.

The procedure on how the optical parameters are used is described in following sections.

ion

The reachable source-destination node pair indicates that there are a few number of OCh paths between two nodes. The reachability is restricted by impairment, wavelength consistency and so on. This information is necessary at PCE to promise the path computed between source node and destination node is reachable. In this scenario, the PCE should be responsible to compute how many OCh paths are available to set up connections between source and destination node. Moreover, if a set of optical wavelength is indicated in the path computation request, the PCE should also determine whether a wavelength of the set of preselected optical wavelength is available for the source-destination pair connection.

To enable PCE to complete the above functions, the reachable relationship and OMS link information need to be reported to PCE. Once PCE detect that any wavelength is available, the corresponding OMS link should be included in a lambda plane. Then this link can be used for path computation in future.

Moreover, in a hierarchical PCE architecture, the information above need to be reported from child PCE to parent PCE, who acts as a service coordinator.

It is a usual case that the PCC indicates the latency when requesting the path computation. In optical networks the latency is a very sensitive parameter and there is stricter requirement on latency. Given the maximal number of OCh paths between source-destination nodes, the PCE also need to determine how many OCh path satisfies the indicated latency threshold.

There is usually high-performance algorithm running on the PCE to guarantee the performance of the computed path. During the computation, the delay factor may be converted into a kind of link weight. After the algorithm provides a few candidate paths between the source and destination nodes, the PCE SHOULD be capable to selecting one shortest path by computing the total path delay.

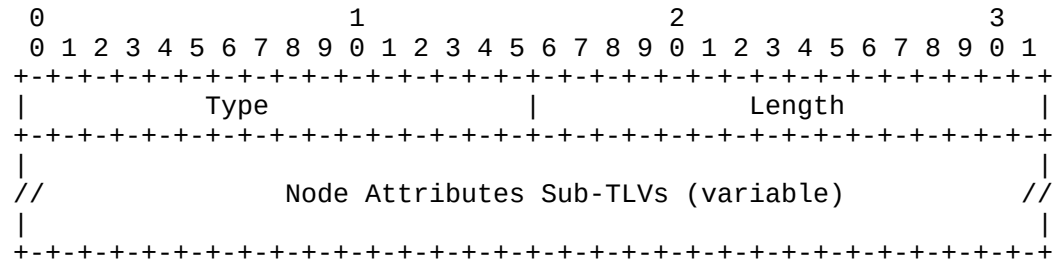
Optical PCEs are embedded with optimization algorithm, e.g., shortest path algorithm, to improve the performance of computed path.

4. PCEP-LS extension for Optical Networks

This section provides additional PCEP-LS extension necessary to support optical networks. All Objects/TLVs defined in [PCEP-LS] are applicable to optical networks.

4.1. Node Attributes TLV

Node-Attributed TLV is defined in Section 9.2.10.1 in [PCEP-LS] as follows. This TLV is applicable for LS Node Object-Type as defined in [PCEP-LS].



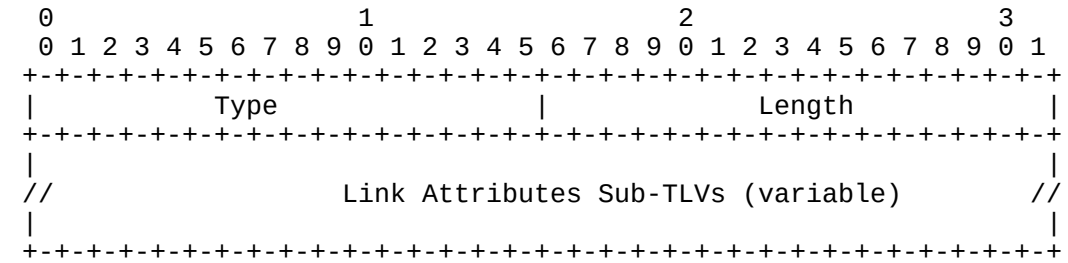
The following 'Node Attribute' sub-TLVs are valid for optical networks:

Sub-TLV	Description	TLV/Sub-TLV	Length	Reference
TBD	Connectivity Matrix	5/14	variable	[RFC7579] [RFC7580]
TBD	Resource Block Information	6/1	variable	[RFC7688]

TBD	Resource Block Accessibility	6/2	variable	[RFC7688]
TBD	Resource Block Wavelength Const	6/3	variable	[RFC7688]
TBD	Resource Block Pool State	6/4	variable	[RFC7688]
TBD	Resource Block Shared Access Wavelength Avail.	6/5	variable	[RFC7688]

4.2. Link Attributes TLV

Link-Attributes TLV is defined in Section 9.2.10.2 in [PCEP-LS] as follows. This TLV is applicable for LS Link Object-Type as defined in [PCEP-LS].



The following 'Link Attribute' sub-TLVs are valid for optical networks:

Sub-TLV	Description	TLV/Sub-TLV	Length	Reference
TBD	ISCD	15	Variable	[RFC4203]
TBD	OTN-TDM SCSI	15/1,2	Variable	[RFC4203] [RFC7138]
TBD	WSON-LSC SCSI	15/1,2	Variable	[RFC4203] [RFC7688]
TBD	Flexi-grid SCSI	15/1	Variable	[FlexOSPF]
TBD	Port Label Restriction	34	Variable	[RFC7579] [RFC7580] [FlexOSPF]

4.3. PCEP-LS for Optical Network Abstraction

This section provides additional PCEP-LS extension necessary to support optical networks parameters discussed in Sections 3.1 and 3.2. Abstraction is primarily applied to C-PCE and P-PCE although the same principle can be applied to PCC (NE) to PCE.

For OTN networks, max bandwidth available may be per ODU 0/1/2/3 switching level or aggregated across all ODU switching levels (i.e., ODUj/k).

For WSON networks, max bandwidth available may be per lambda/frequency level (OCh) or aggregated across all lambda/frequency level. Per OCh level abstraction gives more detailed data to the P-PCE at the expense of more information processing. Either OCh-level or aggregated level abstraction should factor in the RWA constraint (i.e., wavelength continuity) at the C-PCE level. This means the C-PCE should have this capability and advertise it as such.

[Editor’s Note: Encoding will be provided in the revision]

5. Security Considerations

This document extends PCEP for LS (and TE) distribution including a set of TLVs. Procedures and protocol extensions defined in this document do not effect the overall PCEP security model. See [RFC5440], [I-D.ietf-pce-pceps]. The PCE implementation SHOULD provide mechanisms to prevent strains created by network flaps and

amount of LS (and TE) information. Thus it is suggested that any mechanism used for securing the transmission of other PCEP message be applied here as well. As a general precaution, it is RECOMMENDED that these PCEP extensions only be activated on authenticated and encrypted sessions belonging to the same administrative authority.

6. IANA Considerations

This document requests IANA actions to allocate code points for the protocol elements defined in this document.

6.1. PCEP-LS Sub-TLV Type Indicators

This document specifies a set of PCEP-LS Sub-TLVs. IANA is requested to create an "PCEP-LS Sub-TLV Types" sub-registry in the "PCEP TLV Type Indicators" for the sub-TLVs carried in the PCEP-LS TLV (Node Attributes TLV and Link Attributes TLV).

Sub-TLV	Description	Ref Sub-TLV	Reference
TBD	Connectivity Matrix	5/14	[RFC7579] [RFC7580]
TBD	Resource Block Information	6/1	[RFC7688]
TBD	Resource Block Accessibility	6/2	[RFC7688]
TBD	Resource Block Wavelength Const	6/3	[RFC7688]
TBD	Resource Block Pool State	6/4	[RFC7688]
TBD	Resource Block Shared Access Wavelength Avail.	6/5	[RFC7688]
TBD	ISCD	15	[RFC4203]
TBD	OTN-TDM SCSI	15/1,2	[RFC4203] [RFC7138]
TBD	WSON-LSC SCSI	15/1,2	[RFC4203] [RFC7688]
TBD	Flexi-grid SCSI	15/1	[FlexOSPF]
TBD	Port Label Restriction	34	[RFC7579] [RFC7580] [FlexOSPF]

7. References

7.1. Normative References

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

[RFC4674] Le Roux, J., Ed., "Requirements for Path Computation Element (PCE) Discovery", RFC 4674, October 2006.

[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.

[RFC5250] Berger, L., Bryskin, I., Zinin, A., and R. Coltun, "The OSPF Opaque LSA Option", RFC 5250, July 2008.

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

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

7.2. Informative References

[JMS] Java Message Service, Version 1.1, April 2002, Sun Microsystems.

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

[RFC4203] Kompella, K., Ed. and Y. Rekhter, Ed., "OSPF Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4203, October 2005.

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

[BGP-LS] Gredler, H., Medved, J., Previdi, S., Farrel, A., and S.Ray, "North-Bound Distribution of Link-State and TE information using BGP", draft-ietf-idr-ls-distribution, work in progress.

[S-PCE-GMPLS] X. Zhang, et. al, "Path Computation Element (PCE) Protocol Extensions for Stateful PCE Usage in GMPLS-controlled Networks", draft-ietf-pce-pcep-stateful-pce-gmpls, work in progress.

[RFC7399] A. Farrel and D. king, "Unanswered Questions in the Path Computation Element Architecture", RFC 7399, October 2015.

[RFC7449] Y. Lee, G. Bernstein, "Path Computation Element Communication Protocol (PCEP) Requirements for Wavelength Switched Optical Network (WSON) Routing and Wavelength Assignment", RFC 7449, February 2015.

[RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, April 2006.

[RFC6163] Y. Lee, G. Bernstein, W. Imajuku, "Framework for GMPLS and PCE Control of Wavelength Switched Optical Networks", RFC 6163,

[G.680] ITU-T Recommendation G.680, Physical transfer functions of optical network elements, July 2007.

[ACTN-Frame] D.Ceccarelli, and Y. Lee (Editors), "Framework for Abstraction and Control of TE Networks", draft-ietf-teas-actn-framework, work in progress.

[RFC6805] A. Farrel and D. King, "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-LS-Arch] Y. Lee, D. Dhody and D. Ceccarelli, "Architecture and Requirement for Distribution of Link-State and TE Information via PCEP", draft-leedhody-teas-pcep-ls, work in progress.

[PCEP-LS] D. Dhody, Y. Lee and D. Ceccarelli "PCEP Extension for Distribution of Link-State and TE Information.", work in progress, September 21, 2015[Stateful-PCE] Crabbe, E., Minei, I., Medved, J., and R. Varga, "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce, work in progress.

[PCE-Initiated] Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model", draft-ietf-pce-pce-initiated-lsp, work in progress.

[Stateful H-PCE] D. Dhody, Y. Lee and D. Ceccarelli, "Hierarchical Stateful Path Computation Element (PCE)", draft-ietf-pce-stateful-hpce, work-in-progress.

[FlexOSPF] X. Zhang, H. Zheng, R. Casellas, O. Gonzalez de Dios, D. Ceccarelli, "GMPLS OSPF Extensions in support of Flexi-grid DWDM networks", draft-ietf-ccamp-flexible-grid-ospf-ext-05, work in progress.

Appendix A. Contributor Addresses

Dhruv Dhody
Huawei Technologies
Divyashree Techno Park, Whitefield
Bangalore, Karnataka 560066
India
Email: dhruv.ietf@gmail.com

Author’s Addresses

Young Lee
Huawei Technologies
5340 Legacy Drive, Building 3
Plano, TX 75023, USA

Email: leeyoung@huawei.com

Haomian Zheng
Huawei Technologies Co., Ltd.
F3-1-B R&D Center, Huawei Base,
Bantian, Longgang District
Shenzhen 518129 P.R.China

Email: zhenghaomian@huawei.com

Daniele Ceccarelli
Ericsson
Torshamnsgatan,48
Stockholm
Sweden

EMail: daniele.ceccarelli@ericsson.com

Wei Wang
Beijing University of Posts and Telecom
No. 10, Xitucheng Rd. Haidian District, Beijing 100876, P.R.China

Email: weiw@bupt.edu.cn

Peter Park
KT
Email: peter.park@kt.com

Bin Yeong Yoon
ETRI
Email: byyun@etri.re.kr

PCE Working Group
Internet-Draft
Intended Status: Standards track
Expires: March 22, 2017

Y. Lee
D. Dhody
X. Zhang
Huawei Technologies
D. Ceccarelli
Ericsson

September 18, 2017

PCEP Extensions for Establishing Relationships Between Sets of LSPs and Virtual Networks
draft-leedhody-pce-vn-association-03

Abstract

This document describes how to extend Path Computation Element (PCE) Communication Protocol (PCEP) association mechanism introduced by the PCEP Association Group specification, to further associate sets of LSPs with a higher-level structure such as a virtual network (VN) requested by clients or applications. This extended association mechanism can be used to facilitate virtual network control using PCE architecture.

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 March 22, 2017.

Table of Contents

1. Introduction

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.

[RFC8051] describes general considerations for a stateful PCE deployment and examines its applicability and benefits, as well as its challenges and limitations through a number of use cases. [I-D.ietf-pce-stateful-pce] describes a set of extensions to PCEP to provide stateful control. A stateful PCE has access to not only the information carried by the network's Interior Gateway Protocol (IGP), but also the set of active paths and their reserved resources for its computations. The additional state allows the PCE to compute constrained paths while considering individual LSPs and their interactions.

[I-D.ietf-pce-pce-initiated-lsp] describes the setup, maintenance and teardown of PCE-initiated LSPs under the stateful PCE model.

[I-D.ietf-pce-association-group] introduces a generic mechanism to create a grouping of LSPs. This grouping can then be used to define association between sets of LSPs or between a set of LSPs and a set of attributes.

[ACTN-REQ] describes various Virtual Network (VN) operations initiated by a customer/application. In this context, there is a need for associating a set of LSPs with a VN "construct" to facilitate VN operations in PCE architecture. This association allows the PCEs to identify which LSPs belong to a certain VN. The PCE could then use this association to optimize all LSPs belonging to the VN together. The PCE could further take VN specific actions on the LSPs such as relaxation of constraints, policy actions, setting default behavior etc.

[I-D.ietf-pce-applicability-actn] examines the PCE and ACTN architecture and describes how the PCE architecture is applicable to ACTN. [RFC6805] and [I-D.ietf-pce-stateful-hpce] describes a hierarchy of stateful PCEs with Parent PCE coordinating multi-domain path computation function between Child PCE(s) and thus making it the base for PCE applicability for ACTN. In this text child PCE would be same as Physical Network Controller (PNC), and the parent PCE as Multi-domain Service Coordinator (MDSC) [ACTN-FWK].

This document specifies a PCEP extension to associate a set of LSPs based on Virtual Network (VN) (or customer). A Virtual Network (VN) is a customer view of the TE network. Depending on the agreement between client and provider various VN operations and VN views are possible as described in [ACTN-FWK].

1.1. Requirements Language

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.

2. Terminology

The terminology is as per [RFC4655], [RFC5440], [RFC6805], [I-D.ietf-pce-stateful-pce] and [ACTN-FWK]..

3. Operation Overview

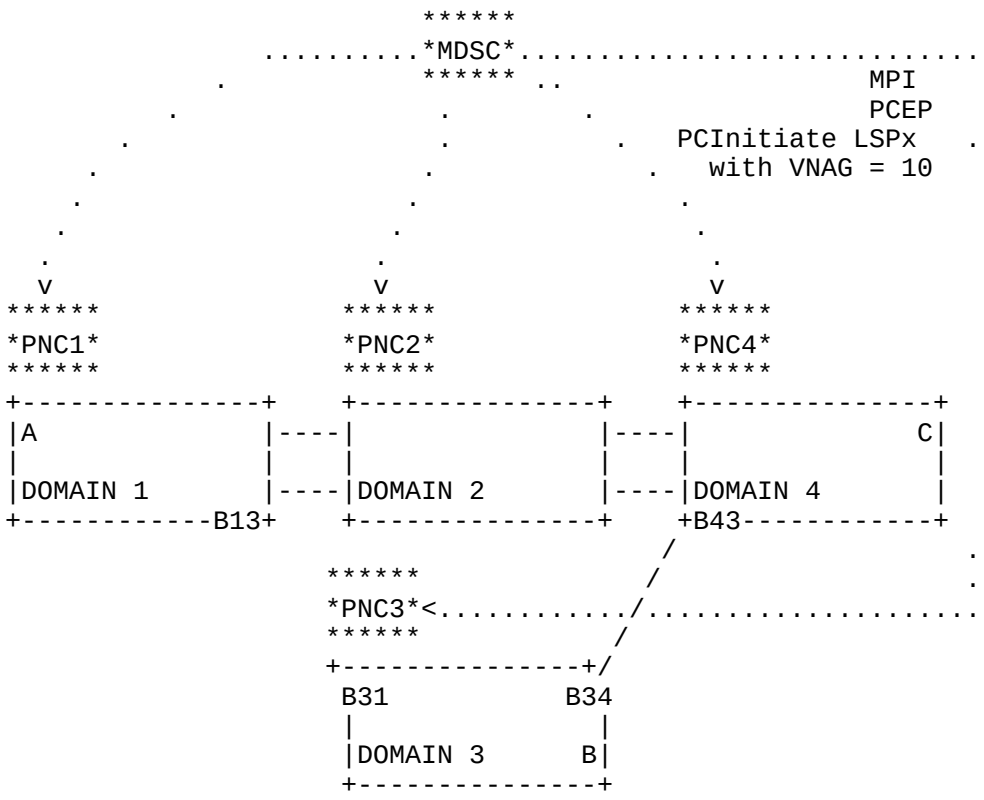
As per [I-D.ietf-pce-association-group], LSPs are associated with other LSPs with which they interact by adding them to a common association group.

An association group based on VN is useful for various optimizations that should be applied by considering all the LSPs in the association. This includes, but not limited to -

- Path Computation: When computing path for a LSP, the impact of this LSP, on the other LSPs belonging to the same VN is useful to analyze. The aim would be optimize overall VN and all LSPs, rather than a single LSP. Also, the optimization criteria such as minimize the load of the most loaded link (MLL) [RFC5541] and other could be applied for all the LSP belonging to the same VN, identified by the VN association.

- Path Re-Optimization: The child PCE or the parent PCE would like to use advanced path computation algorithm and optimization technique that consider all the LSPs belonging to a VN/customer and optimize them all together during the re-optimization.

This association is useful in PCEP session between parent PCE (MDSC) and child PCE (PNC).



MDSC -> Parent PCE
PNC -> Child PCE
MPI -> PCEP

In this draft, this grouping is used to define associations between a set of LSPs and a virtual network.

One new optional Association Object-type is defined based on the generic Association object -

- o VN Association Group (VNAG)

Thus this document define one new association type called "VN Association Type" of value TBD1. The scope and handling of VNAG identifier is similar to the generic association identifier defined in [I-D.ietf-pce-association-group].

Local polices on the PCE MAY define the computational and optimization behavior for the LSPs in the VN. An LSP MUST belong to a single VNAG. If an implementation encounters more than one VNAG, it

MUST consider the first occurrence and ignore the others.

This Association-Type is dynamic in nature and created by the Parent PCE (MDSC) for the LSPs belonging to the same VN or customer. These associations are conveyed via PCEP messages to the PCEP peer. Operator-configured Association Range SHOULD NOT be set for this association-type and MUST be ignored.

4. Extensions to PCEP

[I-D.ietf-pce-association-group] introduces the ASSOCIATION object, the format of VNAG is as follows:

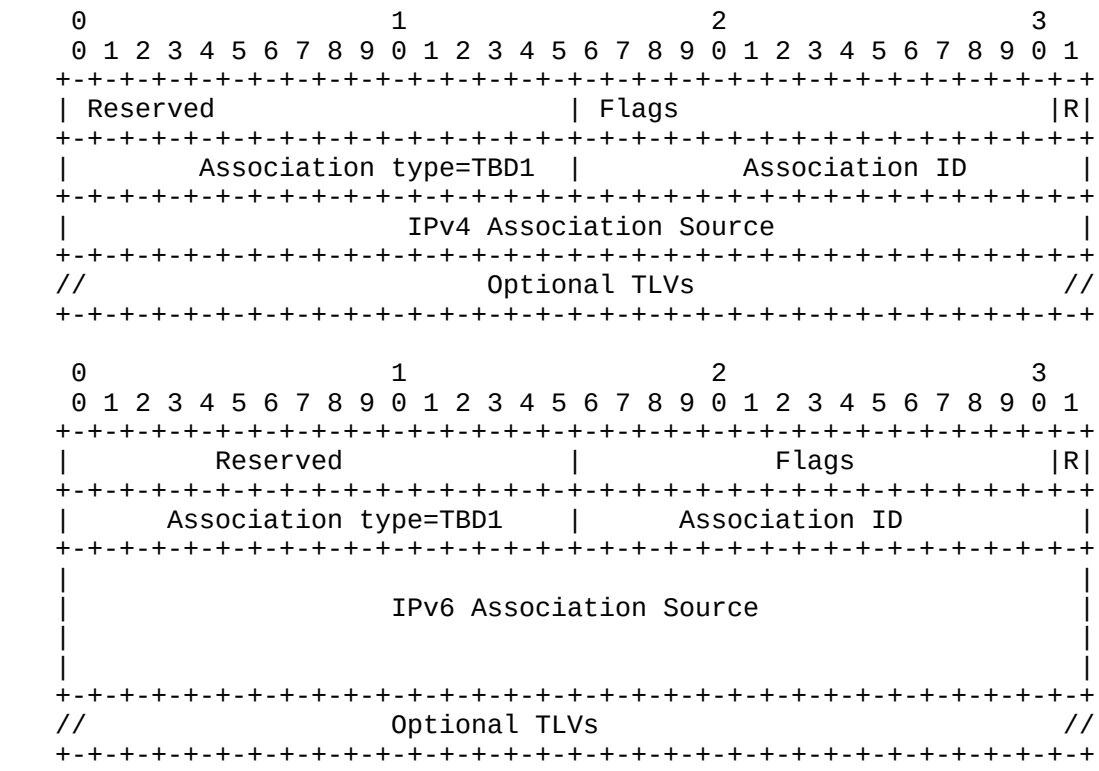


Figure 1: The VNAG Object formats

Please refer to [I-D.ietf-pce-association-group] for the definition of each field in Figure 1. This document defines one mandatory TLV "VIRTUAL-NETWORK-TLV" and one optional TLV "VENDOR-INFORMATION-TLV" -

o VIRTUAL-NETWORK-TLV: Used to communicate the VN Identifier.

o VENDOR-INFORMATION-TLV: Used to communicate arbitrary vendor specific behavioral information, described in [RFC7470].

The format of VIRTUAL-NETWORK-TLV is as follows.

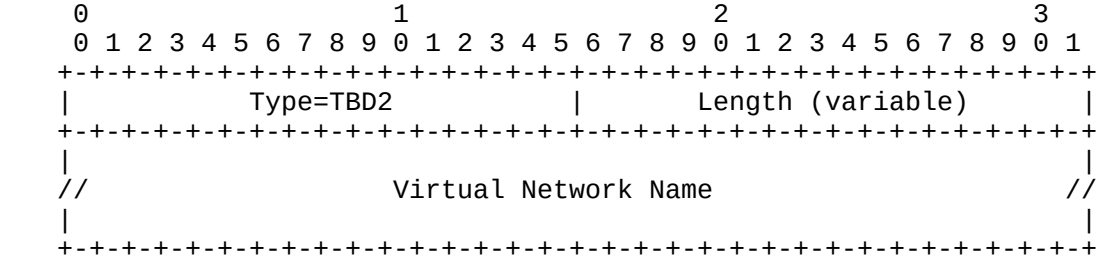


Figure 2: The VIRTUAL-NETWORK-TLV formats

Type: TBD2 (to be allocated by IANA)

Length: Variable Length

Virtual Network Name(variable): an unique symbolic name for the VN. The VN name is a human-readable string that identifies a VN. The VN name MUST remain constant throughout an LSP's lifetime, which may span across multiple consecutive PCEP sessions and/or PCC restarts. The VN name MAY be specified by an operator or auto-generated by the PCEP speaker.

The VIRTUAL-NETWORK-TLV MUST be included in VNAG object.If a PCEP speaker receives the VNAG object without the VIRTUAL-NETWORK-TLV, it MUST send a PCErr message with Error-Type=6 (mandatory object missing) and Error-Value=TBD3 (VIRTUAL-NETWORK-TLV missing) and close the session.

The format of VENDOR-INFORMATION-TLV is defined in [RFC7470].

5. Applicability to H-PCE architecture

The ability to compute shortest constrained TE LSPs in Multiprotocol Label Switching (MPLS) and Generalized MPLS (GMPLS) networks across multiple domains has been identified as a key motivation for PCE development. [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.

[I-D.ietf-pce-stateful-hpce] introduces general considerations for stateful PCE(s) in hierarchical PCE architecture. In particular, the behavior changes and additions to the existing stateful PCE mechanisms in the context of a H-PCE architecture.

In Stateful H-PCE architecture, the Parent PCE receives a virtual network creation request by its client over its Northbound API. This VN is uniquely identified by an Association ID in VNAG as well as the VIRTUAL-NETWORK name. This VN may comprise multiple LSPs in the network in a single domain or across multiple domains.

As the Parent PCE computes the optimum E2E paths for each tunnel in VN, it MUST associate each LSP with the VN to which it belongs. Parent PCE sends a PCInitiate Message with this association information in the VNAG Object (See Section 4 for details). This in effect binds an LSP that is to be instantiated at the child PCE with the VN.

Whenever changes occur with the instantiated LSP in a domain network, the domain child PCE reports the changes using a PCRpt Message in which the VNAG Object indicates the relationship between the LSP and the VN.

Whenever an update occurs with VNs in the Parent PCE (via the client’s request), the parent PCE sends an PCUpd Message to inform each affected child PCE of this change.

The Child PCE could then use this association to optimize all LSPs belonging to the same VN association together. The Child PCE could further take VN specific actions on the LSPs such as relaxation of constraints, policy actions, setting default behavior etc. The parent PCE could also maintain all E2E LSP or per-domain path segments under a single VN association.

6. Security Considerations

This document defines one new type for association, which do not add any new security concerns beyond those discussed in [RFC5440], [I-D.ietf-pce-stateful-pce] and [I-D.ietf-pce-association-group] in itself.

Some deployments may find VN associations and their implications as extra sensitive and thus should employ suitable PCEP security mechanisms like TCP-AO or [I-D.ietf-pce-pceps].

7. IANA Considerations

7.1. Association Object Type Indicator

This document defines a new association type, originally defined in [I-D.ietf-pce-association-group], for path protection. IANA is requested to make the assignment of a new value for the sub-registry "ASSOCIATION Type Field" (request to be created in [I-D.ietf-pce-association-group]), as follows:

Value	Name	Reference
TBD1	VN Association Type	[This I.D.]

7.2. PCEP TLV Type Indicator

This document defines a new TLV for carrying additional information of LSPs within a path protection association group. IANA is requested to make the assignment of a new value for the existing "PCEP TLV Type Indicators" registry as follows:

Value	Name	Reference
TBD2	VIRTUAL-NETWORK-TLV	[This I.D.]

7.3. PCEP Error

This document defines new Error-Type and Error-Value related to path protection association. IANA is requested to allocate new error values within the "PCEP-ERROR Object Error Types and Values" sub-registry of the PCEP Numbers registry, as follows:

Error-Type	Meaning
6	Mandatory Object missing Error-value=TBD3: VIRTUAL-NETWORK TLV missing [This I.D.]

8. Manageability Considerations

8.1. Control of Function and Policy

An operator MUST BE allowed to mark LSPs that belong to the same VN. This could also be done automatically based on the VN configuration.

8.2. Information and Data Models

The PCEP YANG module [I-D.ietf-pce-pcep-yang] should support the association between LSPs including VN association.

8.3. 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].

8.4. Verify Correct Operations

Mechanisms defined in this document do not imply any new operation verification requirements in addition to those already listed in [RFC5440].

8.5. Requirements On Other Protocols

Mechanisms defined in this document do not imply any new requirements on other protocols.

8.6. 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].

9. References

9.1. 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.

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

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, May 2017.

[I-D.ietf-pce-stateful-pce] E. Crabbe, I. Minei, J. Medved, and R. Varga, "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce, work in progress.

[I-D.ietf-pce-pce-initiated-lsp] E. Crabbe, et. al., "PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model", draft-ietf-pce-pce-initiated-lsp, work in progress.

[I-D.ietf-pce-association-group] I, Minei, Ed., "PCEP Extensions for Establishing Relationships Between Sets of LSPs", draft-ietf-pce-association-group, work in progress.

9.2. Informative References

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

[RFC6805] A. Farrel and D. King, "The Application of the Path Computation Element Architecture to the Determination of a Sequence of Domains in MPLS and GMPLS", RFC 6805, November 2012.

[I-D.ietf-pce-applicability-actn] Dhody D., Lee Y., and D. Ceccarelli, "Applicability of Path Computation Element (PCE) for Abstraction and Control of TE Networks (ACTN)", draft-ietf-pce-applicability-actn, work-in-progress.

[I-D.ietf-pce-stateful-hpce] Dhody, D. and Lee, Y., "Hierarchical Stateful Path Computation Element (PCE)", draft-ietf-pce-stateful-hpce, work in progress.

[ACTN-REQ] Y. Lee, D. Dhody, S. Belotti, K. Pithewan, and D. Ceccarelli, "Requirements for Abstraction and Control of TE Networks", draft-ietf-teas-actn-requirements, work in progress.

[ACTN-FWK] Ceccarelli D. and Y. Lee, "Framework for Abstraction and Control of Transport Networks", draft-ietf-teas-actn-framework (work in progress).

[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, <http://www.rfc-editor.org/info/rfc5541>.

[RFC7470] Zhang, F. and A. Farrel, "Conveying Vendor-Specific Constraints in the Path Computation Element Communication Protocol", RFC 7470, DOI 10.17487/RFC7470, March 2015, <http://www.rfc-editor.org/info/rfc7470>.

[RFC8051] Zhang, X., Ed. and I. Minei, Ed., "Applicability of a Stateful Path Computation Element (PCE)", RFC 8051, DOI 10.17487/RFC8051, January 2017, <http://www.rfc-editor.org/info/rfc8051>.

[I-D.ietf-pce-pceps] Lopez, D., Dios, O., Wu, W., and D. Dhody, "Secure Transport for PCEP", draft-ietf-pce-pceps (work in progress).

[I-D.ietf-pce-pcep-yang]
Dhody, D., Hardwick, J., Beeram, V., and j.
jefftant@gmail.com, "A YANG Data Model for Path
Computation Element Communications Protocol (PCEP)",
draft-ietf-pce-pcep-yang (work in progress).

Author's Addresses

Young Lee (Editor)
Huawei Technologies
5340 Legacy Drive, Building 3
Plano, TX 75023,
USA

Email: leeyoung@huawei.com

Dhruv Dhody (Editor)
Huawei Technologies
Divyashree Technopark, Whitefield
Bangalore, Karnataka 560066
India

Email: dhruv.ietf@gmail.com

Xian Zhang
Huawei Technologies
China

Email: zhang.xian@huawei.com

Daniele Ceccarelli
Ericsson
Torshamnsgatan,48
Stockholm,
Sweden

Email: daniele.ceccarelli@ericsson.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: March 1, 2018

S. Litkowski
Orange
S. Sivabalan
Cisco
D. Dhody
Huawei

August 28, 2017

Inter Stateful Path Computation Element communication procedures
draft-litkowski-pce-state-sync-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. The stateful PCE extensions allow stateful control of Multi-Protocol Label Switching (MPLS) Traffic Engineering Label Switched Paths (TE LSPs) using PCEP.

A Path Computation Client (PCC) can synchronize an LSP state information to a Stateful Path Computation Element (PCE). The stateful PCE extension allows a redundancy scenario where a PCC can have redundant PCEP sessions towards multiple PCEs. In such a case, a PCC gives control on a LSP to only a single PCE, and only one PCE is responsible for path computation for this delegated LSP. The document does not state the procedures related to an inter-PCE stateful communication.

There are some use cases, where an inter-PCE stateful communication can bring additional resiliency in the design for instance when some PCC-PCE sessions fails. The inter-PCE stateful communication may also provide a faster update of the LSP states when an event occurs. Finally, when, in a redundant PCE scenario, there is a need to compute a set of paths that are part of a group (so there is a dependency between the paths), there may be some cases where the computation of all paths in the group is not handled by the same PCE: this situation is called a split-brain. This split-brain scenario may lead to computation loops between PCEs or suboptimal paths computation.

This document describes the procedures to allow a stateful communication between PCEs for various use-cases and also the procedures to prevent computations loops.

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

Copyright Notice

Table of Contents

1. Introduction and problem statement

1.1. Reporting LSP changes

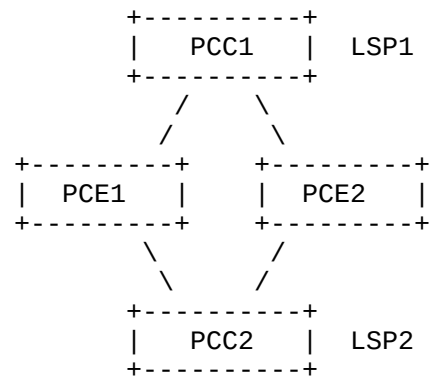
When using a stateful PCE ([I-D.ietf-pce-stateful-pce]), a Path Computation Client (PCC) can synchronize an LSP state information to the stateful Path Computation Element (PCE). If the PCC grants the control on the LSP to the PCE, the PCE can update the LSP parameters at any time.

In a multi PCE deployment (redundancy, loadbalancing...), with the current specification defined in [I-D.ietf-pce-stateful-pce], the PCC will be in charge of reporting the other PCEs of the LSP parameter change which brings additional hops and delays in notifying the overall network of the LSP parameter change.

This delay may affect the reaction time of the other PCEs, if they need to take action after being notified of the LSP parameter change.

Apart from the synchronization from the PCC, it is also useful if there is synchronization mechanism between the stateful PCEs. As stateful PCE make changes to its delegated LSPs, these changes

(pending LSPs and the sticky resources [RFC7399]) can be synchronized immediately to the other PCEs.



In the figure above, we consider a loadbalanced PCE architecture, so PCE1 is responsible to compute paths for PCC1 and PCE2 is responsible to compute paths for PCC2. When PCE1 triggers an LSP update for LSP1, it sends a PCUpdate message to PCC1 for LSP1 containing the new parameters. PCC1 will take the parameters into account and will send a PCReport to PCE1 and PCE2 reflecting the changes. PCE2 will so be notified of the change only after receiving the PCReport from PCC1.

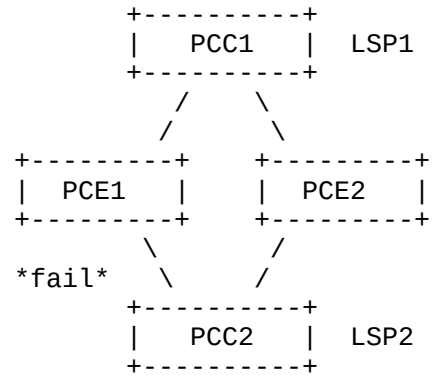
Let’s consider that the LSP1 parameters changed in a such way that LSP1 will take over ressources from LSP2 with an higher priority. After receiving the report from PCC1, PCE2 will so try to find a new path for LSP2. If we consider that there is a round trip delay of about 150msec between the PCEs and PCC1 and a round trip delay of 10msec between the two PCEs, if will take more than 150msec for PCE2 to be notified of the change.

Adding a PCEP session between PCE1 and PCE2 may allow to reduce to the notification time, so PCE2 can react more quickly by taking the pending LSPs and sticky resources into account during path computation and reoptimization.

1.2. Split-brain

In a resiliency case, a PCC has redundant PCEP sessions towards multiple PCEs. In such a case, a PCC gives control on an LSP to a single PCE only, and only this PCE is responsible for the path computation for the delegated LSP: the PCC achieves this by setting the D flag only to the active PCE. The election of the active PCE to delegate an LSP is controlled by each PCC. The PCC usually elects the active PCE by a local configured policy (by setting a priority).

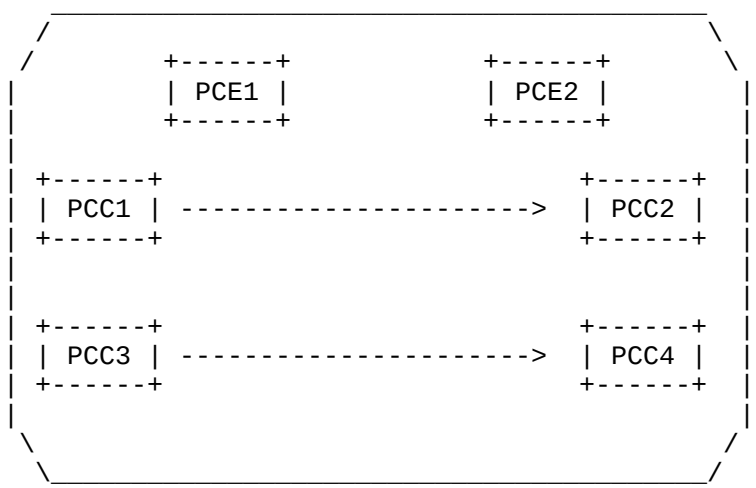
Upon PCEP session failure, or active PCE failure, PCC may decide to elect a new active PCE by sending new PCRpt message with D flag set to this new active PCE. When the failed PCE or PCEP session comes back online, it will be up to the vendor to implement preemption. Doing preemption may lead to some traffic disruption on the existing path if path results from both PCEs are not exactly the same. By considering a network with multiple PCCs and implementing multiple stateful PCEs for redundancy purpose, there is no guarantee that at any time all the PCCs delegate their LSPs to the same PCE.

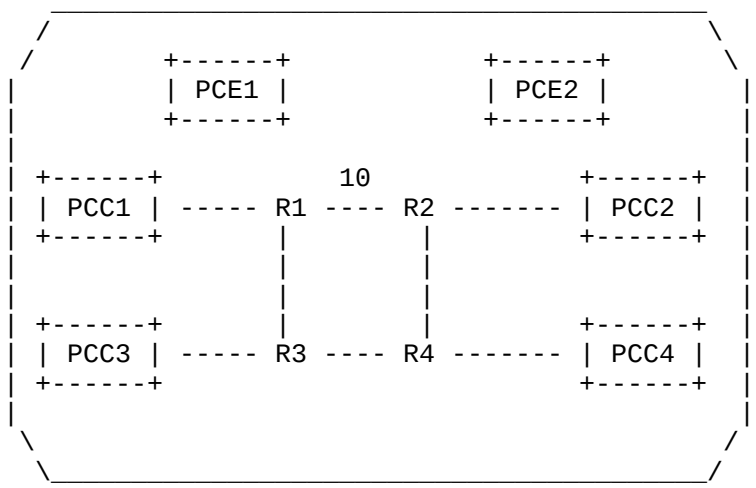


In the example above, we consider that by configuration, both PCCs will firstly delegate their LSP to PCE1. So PCE1 is responsible for computing a path for LSP1 and LSP2. If the PCEP session between PCC2 and PCE1 fails, PCC2 will delegate LSP2 to PCE2. So PCE1 becomes responsible only for LSP1 path computation while PCE2 is responsible for the path computation of LSP2. When the PCC2-PCE1 session is back online, PCC2 will keep using PCE2 as active PCE (no preemption in this example). So the result is a permanent situation where each PCE is responsible for a subset of path computation.

We call this situation a split-brain scenario as there are multiple computation brains running at the same time while a central computation unit was required in some deployments.

Further, there are use cases where a particular LSP path computation is linked to another LSP path computation: the most common use case is path disjointness (see [I-D.ietf-pce-association-diversity]). The set of LSPs that are dependant to each other may start from a different head-end.





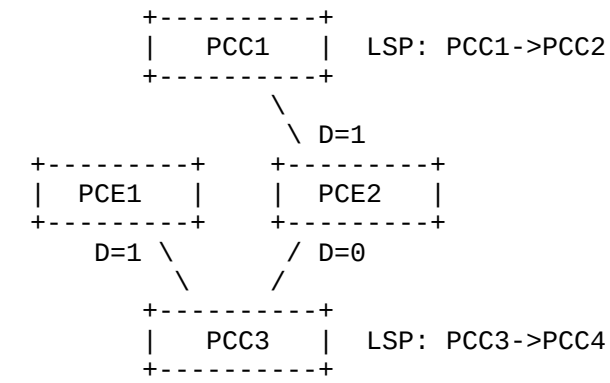
In the figure above, we want to create two link-disjoint LSPs: PCC1->PCC2 and PCC3->PCC4. In the topology, all link metrics are equal to 1 except the link R1-R2 which has a metric of 10. The PCEs are responsible for the path computation and PCE1 is the active PCE for all PCCs in the nominal case.

Scenario 1:

In the nominal case (PCE1 as active PCE), we first configure PCC1->PCC2 LSP, as the only constraint is path disjointness, PCE1 sends a PCUpdate message to PCC1 with the ERO: R1->R3->R4->R2->PCC2 (shortest path). PCC1 signals and installs the path. When PCC3->PCC4 is configured, the PCE already knows the path of PCC1->PCC2 and can compute a link-disjoint path : the solution requires to move PCC1->PCC2 onto a new path to let room for the new LSP. PCE1 sends a PCUpdate message to PCC1 with the new ERO: R1->R2->PCC2 and a PCUpdate to PCC3 with the following ERO: R3->R4->PCC4. In the nominal case, there is no issue for PCE1 to compute a link-disjoint path.

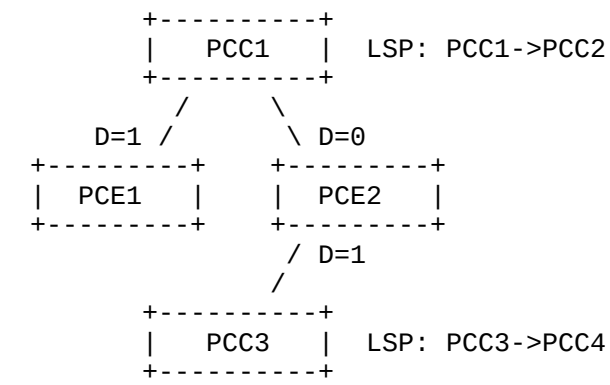
Scenario 2:

Now we consider that PCC1 losts its PCEP session with PCE1 (all other PCEP sessions are UP). PCC1 delegates its LSP to PCE2.



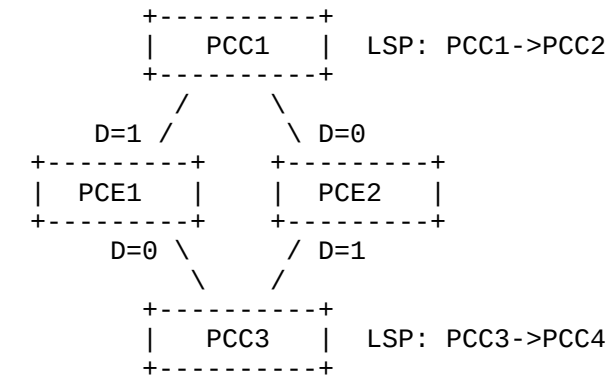
We first configure PCC1->PCC2 LSP, as the only constraint is path disjointness, PCE2 (which is the new active PCE for PCC1) sends a PCUpdate message to PCC1 with the ERO: R1->32->R4->R2->PCC2 (shortest path). When PCC3->PCC4 is configured, PCE1 is not aware anymore of LSPs from PCC1, so it cannot compute a disjoint path for PCC3->PCC4 and will send a PCUpdate message to PCC2 with a shortest path ERO: R3->R4->PCC4. When PCC3->PCC4 LSP will be reported to PCE2 by PCC2, PCE2 will ensure disjointness computation and will correctly move PCC1->PCC2 (as it owns delegation for this LSP) on the following path: R1->R2->PCC2. With this sequence of event and this PCEP session topology, disjointness is ensured.

Scenario 3:



With this new PCEP session topology, we first configure PCC1->PCC2, PCE1 computes the shortest path as it is the only LSP in the disjoint-group that it is aware of: R1->R3->R4->R2->PCC2 (shortest path). When PCC3->PCC4 is configured, PCE2 must compute a disjoint path for this LSP. The only solution found is to move PCC1->PCC2 LSP on another path, but PCE2 cannot do it as it does not have delegation for this LSP. In this setup, PCEs are not able to find a disjoint path.

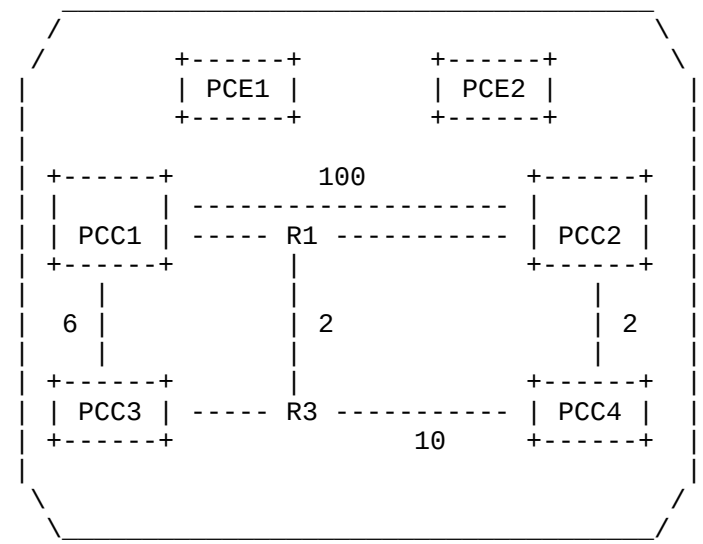
Scenario 4:



With this new PCEP session topology, we consider that PCEs are configured to fallback to shortest path if disjointness cannot be found. We first configure PCC1->PCC2, PCE1 computes shortest path as it is the only LSP in the disjoint-group that it is aware of: R1->R3->R4->R2->PCC2 (shortest path). When PCC3->PCC4 is configured, PCE2 must compute a disjoint path for this LSP. The only solution found is to move PCC1->PCC2 LSP on another path, but PCE2 cannot do it as it does not have delegation for this LSP. PCE2 then provides

shortest path for PCC3->PCC4: R3->R4->PCC4. When PCC3 receives the ERO, it reports it back to both PCEs. When PCE1 becomes aware of PCC3->PCC4 path, it recomputes the CSPF and provides a new path for PCC1->PCC2: R1->R2->PCC2. The new path is reported back to all PCEs by PCC1. PCE2 recomputes also CSPF to take into account the new reported path. The new computation does not lead to any path update.

Scenario 5:

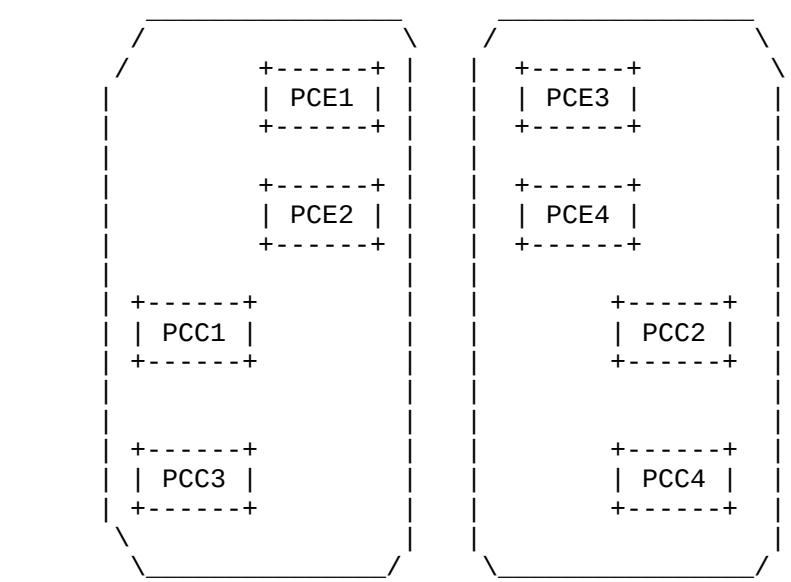


Now we consider a new network topology with the same PCEP session topology as the previous example. We configure both LSPs almost at the same time. PCE1 will compute a path for PCC1->PCC2 while PCE2 will compute a path for PCC3->PCC4. As each other is not aware of the path of the second LSP in the group (not reported yet), each PCE is computing shortest path for the LSP. PCE1 computes ERO: R1->PCC2 for PCC1->PCC2 and PCE2 computes ERO: R3->R1->PCC2->PCC4 for PCC3->PCC4. When these shortest paths will be reported to each PCE. Each PCE will recompute disjointness. PCE1 will provide a new path for PCC1->PCC2 with ERO: PCC1->PCC2. PCE2 will provide also a new path for PCC3->PCC4 with ERO: R3->PCC4. When those new paths will be reported to both PCEs, this will trigger CSPF again. PCE1 will provide a new more optimal path for PCC1->PCC2 with ERO: R1->PCC2 and PCE2 will also provide a more optimal path for PCC3->PCC4 with ERO: R3->R1->PCC2->PCC4. So we come back to the initial state. When those paths will be reported to both PCEs, this will trigger CSPF

again. An infinite loop of CSPF computation is then happening with a permanent flap of paths because of the split-brain situation.

This permanent computation loop comes from the inconsistency between the state of the LSPs as seen by each PCE due to the split-brain: each PCE is trying to modify at the same time its delegated path based on the last received path information which defacto invalidates this receives path information.

Scenario 6: multi-domain



In the example above, we want to create disjoint LSPs from PCC1 to PCC2 and from PCC4 to PCC3. All the PCEs have the knowledge of both domain topologies (e.g. using BGP-LS). For operation/management reason, each domain uses its own group of redundant PCEs. PCE1/PCE2 in domain 1 have PCEP sessions with PCC1 and PCC3 while PCE3/PCE4 in domain 2 have PCEP sessions with PCC2 and PCC4. As PCE1/2 do not know about LSPs from PCC2/4 and PCE3/4 do not know about LSPs from PCC1/3, there is no possibility to compute the disjointness constraint. This scenario can also be seen as a split-brain scenario. This multi-domain architecture (with multiple groups of PCEs) can also be used in a single domain, where an operator wants to limit the failure domain by creating multiple groups of PCEs maintaining a subset of PCCs. As for the multi-domain example, there

will be no possibility to compute disjoint path starting from head-ends managed by different PCE groups.

In this document, we will propose a solution that address the possibility to compute LSP association based constraints (like disjointness) in split-brain scenarios while preventing computation loops.

1.3. Applicability to H-PCE

[I-D.dhodylee-pce-stateful-hpce] describes general considerations and use cases for the deployment of Stateful PCE(s) using the Hierarchical PCE [RFC6805] architecture. In this architecture there is a clear need to communicate between a child stateful PCE and a parent stateful PCE. The procedures and extensions as described in Section 3 are equally applicable to H-PCE.

2. Proposed solution

Our solution is based on :

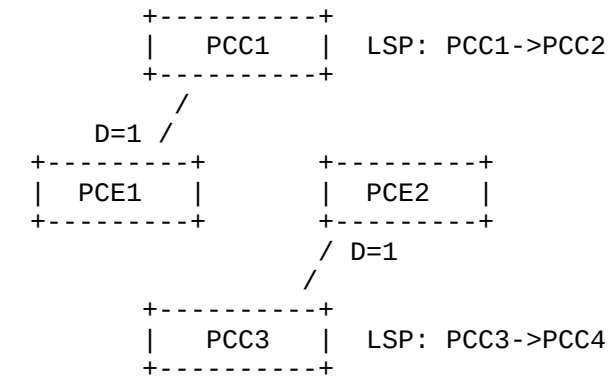
- The creation of the inter-PCE stateful PCEP session with specific procedures.
- A Master/Slave relationship between PCEs.

2.1. State-sync session

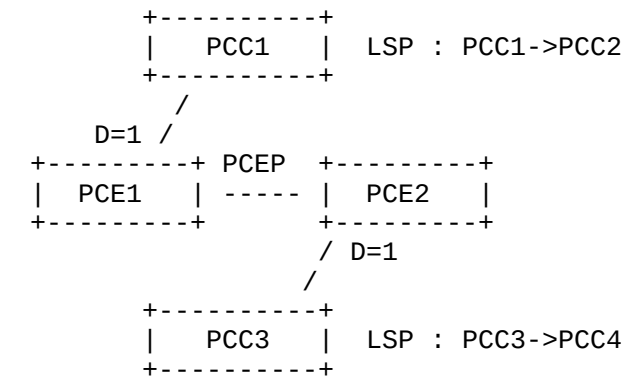
We propose to create a PCEP session between the stateful PCEs. Creating such session is already authorized by multiple scenarios like the one described in [RFC4655] (multiple PCEs that are handling part of the path computation) and [RFC6805] (hierarchical PCE) but was only focused on stateless PCEP sessions. As stateful PCE brings additional features (LSP state synchronization, path update ...), thus some new behaviors need to be defined.

This inter-PCE PCEP session will allow exchange of LSP states between PCEs that would help some scenario where PCEP sessions are lost between PCC and PCE. This inter-PCE PCEP session is called a state-sync session.

For example, in the scenario below, there is no possibility to compute disjointness as there is no PCE aware of both LSPs.



If we add a state-sync session, PCE1 will be able to send PCReport messages for its LSP to PCE2 and PCE2 will do the same. All the PCEs will be aware of all LSPs even if PCC->PCE session are down. PCEs will then be able to compute disjoint paths.



The procedures associated with this state-sync session are defined in Section 3.

Adding this state-sync session does not ensure that a path with LSP association based constraints can always been computed and does not prevent computation loop, but it increases resiliency and ensures that PCEs will have the state information for all LSPs. In addition, this session will allow for a PCE to update the other PCEs providing a faster synchronization mechanism than relying on PCCs only.

2.2. Master/Slave relationship between PCE

As seen in Section 1, performing a path computation in a split-brain scenario (multiple PCEs responsible for computation) may provide a non optimal LSP placement, no path or computation loops. To provide the best efficiency, an LSP association constraint based computation requires that a single PCE performs the path computation for all LSPs in the association group. Note that, it could be all LSPs belonging to a particular association group, or all LSPs from a particular PCC, or all LSPs in the network that need to be delegated to a single PCE based on the deployment scenarios.

We propose to add a priority mechanism between PCEs to elect a single computing PCE. Using this priority mechanism, PCEs can agree on the PCE that will be responsible for the computation for a particular association group, or set of LSPs. The priority could be set per association, per PCC, or for all LSPs. How this priority is set or advertised is out of scope of this document. The rest of the text consider association group as an example.

When a single PCE is performing the computation for a particular association group, no computation loop can happen and an optimal placement will be provided. The other PCEs will only act as state collectors and forwarders.

In the scenario described in Section 2.1, PCE1 and PCE2 will decide that PCE1 will be responsible for the path computation of both LSPs. If we first configure PCC1->PCC2, PCE1 computes shortest path at it is the only LSP in the disjoint-group that it is aware of: R1->R3->R4->R2->PCC2 (shortest path). When PCC3->PCC4 is configured, PCE2 will not perform computation even if it has delegation but forwards the PCRpt to PCE1 through the state-sync session. PCE1 will then perform disjointness computation and will move PCC1->PCC2 onto R1->R2->PCC2 and provides an ERO to PCE2 for PCC3->PCC4: R3->R4->PCC4.

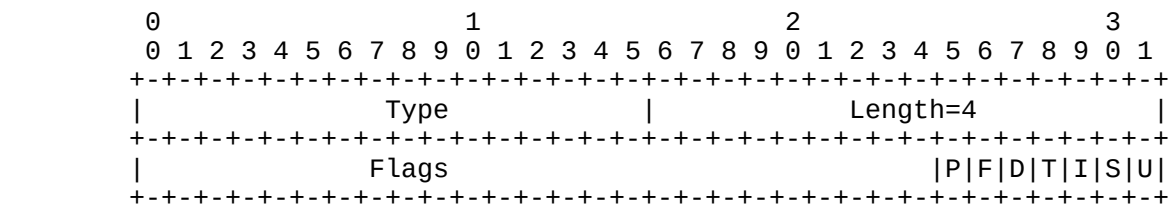
3. Procedures and protocol extensions

3.1. Opening a state-sync session

3.1.1. Capability advertisement

A PCE indicates its support of state-sync procedures during the PCEP Initialization phase. The Open object in the Open message MUST contains the "Stateful PCE Capability" TLV defined in [I-D.ietf-pce-stateful-pce]. A new P (INTER-PCE-CAPABILITY) flag is introduced to indicate the support of state-sync.

The format of the STATEFUL-PCE-CAPABILITY TLV is shown in the following figure:



This document only updates the Flags field with :

P (INTER-PCE-CAPABILITY - 1 bit): If set to 1 by a PCEP Speaker, the PCEP speaker indicates that the session MUST follow the state-sync procedures as described in this document. The P bit MUST be set by both speakers: if a PCEP Speaker receives a STATEFUL-PCE-CAPABILITY TLV with P=0 while it advertised P=1 or if both set P flag to 0, the session SHOULD open but the state-sync procedures MUST NOT be applied on this session.

The U flag MUST be set when sending the STATEFUL-PCE-CAPABILITY TLV with the P flag set. S flag MAY be set if optimized synchronization is required as per [I-D.ietf-pce-stateful-sync-optimizations].

3.2. State synchronization

When the INTER-PCE-CAPABILITY has been negotiated, each PCEP speaker will behave as a PCE and as a PCC at the same time regarding the state synchronization as defined in [I-D.ietf-pce-stateful-pce]. This means that each PCEP Speaker:

- MUST send a PCRpt message towards its neighbor with S flag set for each LSP in its LSP database learned from a PCC. (PCC role)
- MUST send the End Of Synchronization Marker towards its neighbor when all LSPs have been reported. (PCC role)
- MUST wait for the LSP synchronization from its neighbor to end (receiving an End Of Synchronization Marker). (PCE role)

The process of synchronization runs in parallel on each PCE (no defined order).

Optimized synchronization MAY be used as defined in [I-D.ietf-pce-stateful-sync-optimizations].

When a PCEP Speaker sends a PCReport on a state-sync session, it MUST add the SPEAKER-IDENTITY-TLV (defined in [I-D.ietf-pce-stateful-sync-optimizations]) in the LSP Object, the value used will refer to the PCC owner of the LSP. If a PCEP Speaker receives a PCReport on a state-sync session without this TLV, it MUST discard the PCReport and it MUST reply with a PCErr message using error-type=6 (Mandatory Object missing) and error-value=TBD1 (SPEAKER-IDENTITY-TLV missing).

3.3. Incremental updates and report forwarding rules

During the life of an LSP, its state may change (path, constraints, operational state...) and a PCC will advertise a new PCReport to the PCE for each such change.

When propagating LSP state changes from a PCE to other PCEs, it is mandatory to ensure that a PCE always uses the freshest state coming from the PCC.

When a PCE receives a new PCReport from a PCC with the LSP-DB-VERSION, the PCE MUST forward the PCReport to all its state-sync sessions and MUST add the appropriate SPEAKER-IDENTITY-TLV in the PCReport. In addition, it MUST add a new ORIGINAL-LSP-DB-VERSION TLV (described below). The ORIGINAL-LSP-DB-VERSION should contain the LSP-DB-VERSION coming from the PCC.

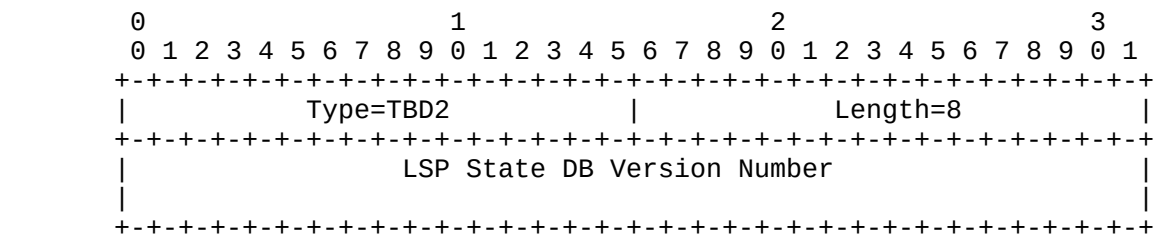
When a PCE receives a new PCReport from a PCC without the LSP-DB-VERSION, it SHOULD NOT forward the PCReport on any state-sync sessions.

When a PCE receives a new PCReport from a PCC with the R flag set and a LSP-DB-VERSION TLV, the PCE MUST forward the PCReport to all its state-sync sessions keeping the R flag set (Remove) and MUST add the appropriate SPEAKER-IDENTITY-TLV and ORIGINAL-LSP-DB-VERSION TLV in the PCReport.

When a PCE receives a PCReport from a state-sync session, it MUST NOT forward the PCReport to other state-sync sessions. This helps to prevent message loops between PCEs. As a consequence, a full mesh of PCEP sessions between PCEs is required.

When a PCReport is forwarded, all the original objects and values are kept. As an example, the PLSP-ID used in the forwarded PCReport will be the same as the original one used by the PCC. Thus an implementation supporting this document MUST consider SPEAKER-IDENTITY-TLV and PLSP-ID together to uniquely identify an LSP on the state-sync session.

The ORIGINAL-LSP-DB-VERSION TLV is encoded as follows and SHOULD always contain the LSP-DB-VERSION received from the PCC owner of the LSP:



Using the ORIGINAL-LSP-DB-VERSION TLV allows a PCE to keep using optimized synchronization ([I-D.ietf-pce-stateful-sync-optimizations]) with another PCE. In such a case, the PCE will send a PCReport to another PCE with both ORIGINAL-LSP-DB-VERSION TLV and LSP-DB-VERSION TLV. The ORIGINAL-LSP-DB-VERSION TLV will contain the version number as allocated by the PCC while the LSP-DB-VERSION will contain the version number allocated by the local PCE.

3.4. Maintaining LSP states from different sources

When a PCE receives a PCReport on a state-sync session, it stores the LSP information into the original PCC address context (as the LSP belongs to the PCC). A PCE SHOULD maintain a single state for a particular LSP and SHOULD maintain the list of sources it learned a particular state from.

A PCEP speaker may receive a state information for a particular LSP from different sources: the PCC that owns the LSP (through a regular PCEP session) and some PCEs (through PCEP state-sync sessions). A PCEP speaker MUST always keep the freshest state in its LSP database, overriding the previously received information.

A PCE, receiving a PCReport from a PCC, updates the state of the LSP in its LSPDB with the new received information. When receiving a PCReport from another PCE, a PCE SHOULD update the LSP state only if the ORIGINAL-LSP-DB-VERSION present in the PCReport is greater than the current ORIGINAL-LSP-DB-VERSION of the stored LSP state. This ensures that a PCE never tries to update its stored LSP state with an old information. Each time a PCE updates an LSP state in its LSPDB, it SHOULD reset the source list associated with the LSP state and SHOULD add the source speaker address in the source list. When a PCE receives a PCReport which has an ORIGINAL-LSP-DB-VERSION (if coming from a PCE) or an LSP-DB-VERSION (if coming from the PCC) equals to

the current ORIGINAL-LSP-DB-VERSION of the stored LSP state, it SHOULD add the source speaker address in the source list.

When a PCE receives a PCReport requesting an LSP deletion from a particular source, it SHOULD remove this particular source from the list of sources associated with this LSP.

When the list of sources becomes empty for a particular LSP, the LSP state MUST be removed. This means that all the sources must send a PCReport with R=1 for an LSP to make the PCE removing the LSP state.

3.5. Computation priority between PCEs and sub-delegation

A computation priority is necessary to ensure that a single PCE will perform the computation for all the LSPs in an association group: this will allow for a more optimized LSP placement and will prevent computation loops.

All PCEs in the network that are handling LSPs in a common LSP association group SHOULD be aware of each other including the computation priority of each PCE. Note that there is no need for PCC to be aware of this. The computation priority is a number and the PCE having the highest priority SHOULD be responsible for the computation. If several PCEs have the same priority value, their IP address SHOULD be used as a tie-breaker to provide a rank: the highest IP address as more priority. How PCEs are aware of the priority of each other is out of scope of this document, but as example learning priorities could be done through IGP informations or local configuration.

The definition of the priority MAY be global so the highest priority PCE will handle all path computations or more granular, so a PCE may have highest priority for only a subset of LSPs or association-groups.

A PCEP Speaker receiving a PCReport from a PCC with D flag set that does not have the highest computation priority, SHOULD forward the PCReport on all state-sync sessions (as per Section 3.3) and SHOULD set D flag on the state-sync session towards the highest priority PCE, D flag will be unset to all other state-sync sessions. This behavior is similar to the delegation behavior handled at PCC side and is called a sub-delegation (the PCE subdelegates the control of the LSP to another PCE). When a PCEP Speaker sub-delegates a LSP to another PCE, it looses the control on the LSP and cannot update it anymore by its own decision. When a PCE receives a PCReport with D flag set on a state-sync session, as a regular PCE, it becomes granted to update the LSP.

If the highest priority PCE is failing or if the state-sync session between the local PCE and the highest priority PCE failed, the local PCE MAY decide to delegate the LSP to the next highest priority PCE or to take back control on the LSP. It is a local policy decision.

When a PCE has the delegation for an LSP and needs to update this LSP, it MUST send a PCUpdate message to all state-sync sessions and to the PCC session on which it received the delegation. The D-Flag would be unset in the PCUpdate for state-sync sessions where as D-Flag would be set for the PCC. In case of subdelegation, the computing PCE will send the PCUpdate only to all state-sync sessions (as it has no direct delegation from a PCC). The D-Flag would be set for the state-sync session to the PCE that sub-delegated this LSP and the D-Flag would be unset for other state-sync sessions.

The PCUpdate sent over a state-sync session MUST contain the SPEAKER-IDENTITY-TLV in the LSP Object (the value used must identify the target PCC). The PLSP-ID used is the original PLSP-ID generated by the PCC and learned from the forwarded PCReport. If a PCE receives a PCUpdate on a state-sync session without the SPEAKER-IDENTITY-TLV, it MUST discard the PCUpdate and MUST reply with a PCError message using error-type=6 (Mandatory Object missing) and error-value=TBD1 (SPEAKER-IDENTITY-TLV missing).

When a PCE receives a valid PCUpdate on a state-sync session, it SHOULD forward the PCUpdate to the appropriate PCC (identified based on the SPEAKER-IDENTITY-TLV value) that delegated the LSP originally and SHOULD remove the SPEAKER-IDENTITY-TLV from the LSP Object. The acknowlegment of the PCUpdate is done through a cascaded mechanism, and the PCC is the only responsible of triggering the acknowledgment: when the PCC receives the PCUpdate from the local PCE, it acknowledges it with a PCReport as per [I-D.ietf-pce-stateful-pce]. When receiving the new PCReport from the PCC, the local PCE uses the defined forwarding rules on the state-sync session so the acknowledgment is relayed to the computing PCE.

A PCE SHOULD NOT compute a path using an association-group constraint if it has delegation for only a subset of LSPs in the group. In this case, an implementation MAY use a local policy on PCE to decide if PCE does not compute path at all for this set of LSP or if it can compute a path by relaxing the association-group constraint.

3.6. Passive stateful procedures

In the passive stateful PCE architecture, the PCC is responsible of triggering a path computation request using a PCRequest message to its PCE. Similarly to PCReports which remains unchanged for passive mode, if a PCE receives a PCRequest for an LSP and if this PCE finds

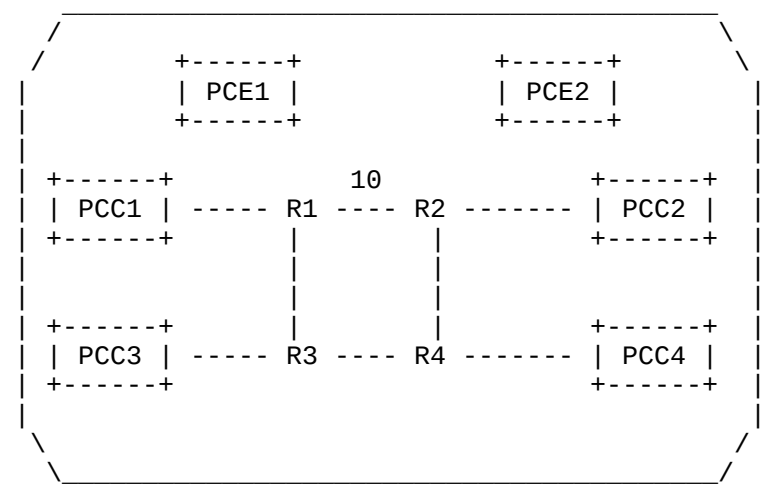
that it does not have the highest computation priority of this LSP, or groups..., it MUST forward the PCRequest to the highest priority PCE over the state-sync session. When the highest priority PCE receives the PCRequest, it computes the path and generates a PCReply only to the PCE that is received the PCRequest from. This PCE will then forward the PCReply to the requesting PCC. The handling of LSP object and the SPEAKER-IDENTITY-TLV in PCRequest and PCReply is similar to PCReport/PCUpdate.

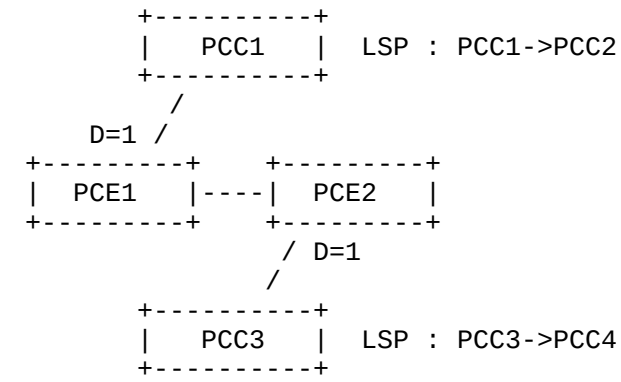
3.7. PCE initiation procedures

TBD

4. Examples

4.1. Example 1





PCE1 computation priority 100
PCE2 computation priority 200

With this PCEP session topology where computation priority is global for all LSPs, we still want to have link disjoint LSPs PCC1->PCC2 and PCC3->PCC4.

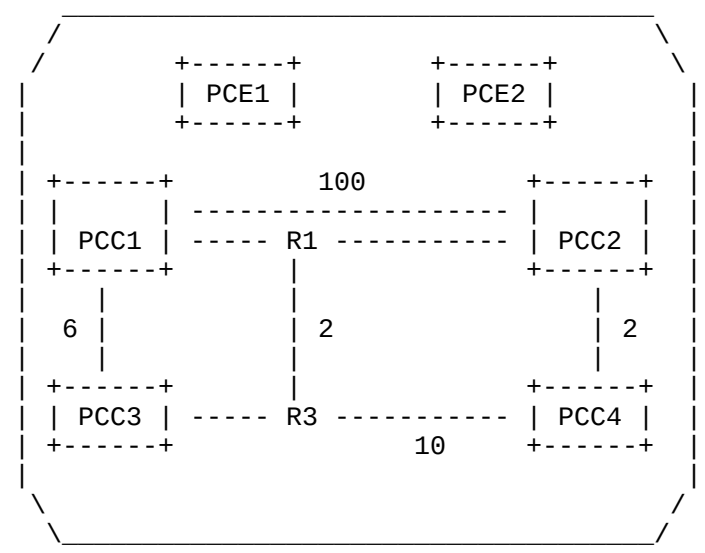
We first configure PCC1->PCC2, PCC1 delegates the LSP to PCE1, but as PCE1 does not have the highest computation priority, it will sub-delegate the LSP to PCE2 by sending a PCReport with D=1 and including the SPEAKER-IDENTITY-TLV over the state-sync session. PCE2 receives the PCReport and as it has delegation for this LSP, it computes the shortest path: R1->R3->R4->R2->PCC2. It then sends a PCUpdate to PCE1 (including the SPEAKER-IDENTITY-TLV) with the computed ERO. PCE1 forwards the PCUpdate to PCC1 (removing the SPEAKER-IDENTITY-

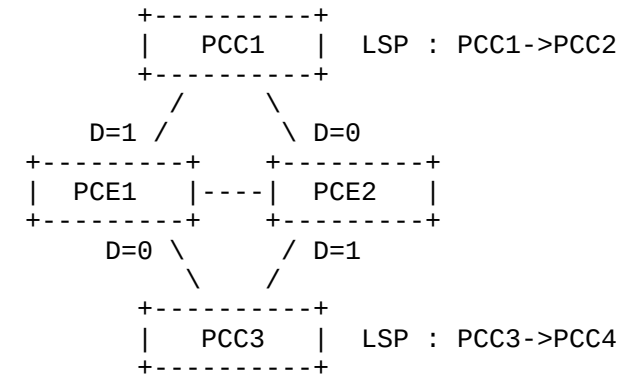
TLV). PCC1 acknowledges the PCUpdate by a PCReport to PCE1. PCE1 forwards the PCReport to PCE2.

When PCC3->PCC4 is configured, PCC3 delegates the LSP to PCE2, PCE2 can compute a disjoint path as it has knowledge of both LSPs and has delegation also for both. The only solution found is to move PCC1->PCC2 LSP on another path, PCE2 can move PCC3->PCC4 as it has delegation for it. It creates a new PCUpdate with new ERO: R1->R2-PCC2 towards PCE1 which forwards to PCC1. PCE2 sends a PCUpdate to PCC3 with the path: R3->R4->PCC4.

In this setup, PCEs are able to find a disjoint path while without state-sync and computation priority they could not.

4.2. Example 2

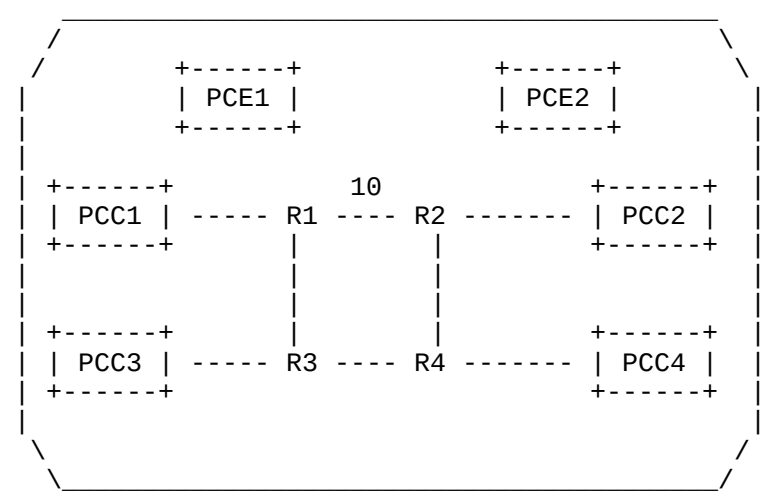


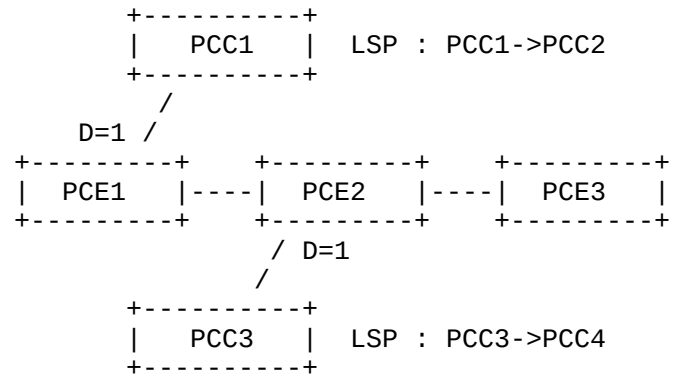


PCE1 computation priority 200
PCE2 computation priority 100

In this example, we configure both LSPs almost at the same time. PCE1 sub-delegates PCC1->PCC2 to PCE2 while PCE2 keeps delegation for PCC3->PCC4, PCE2 computes a path for PCC1->PCC2 and PCC3->PCC4 and can achieve disjointness computation easily. No computation loop happens in this case.

4.3. Example 3





PCE1 computation priority 100
PCE2 computation priority 200
PCE2 computation priority 300

With this PCEP session topology, we still want to have link disjoint LSPs PCC1->PCC2 and PCC3->PCC4.

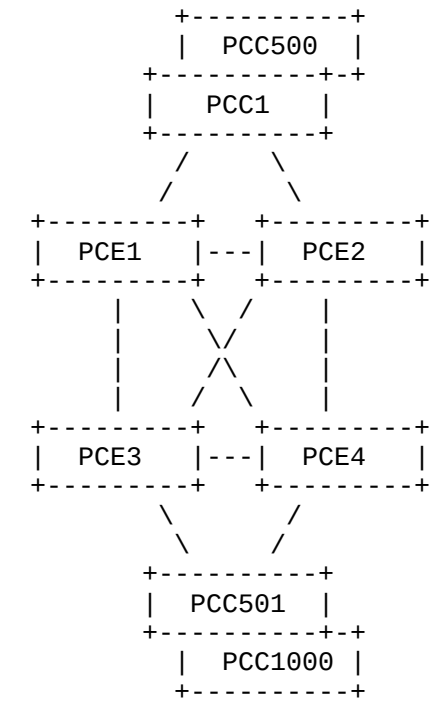
We first configure PCC1->PCC2, PCC1 delegates the LSP to PCE1, but as PCE1 does not have the highest computation priority, it will sub-delegate the LSP to PCE2 (as it cannot reach PCE3 through a state-sync session). PCE2 cannot compute a path for PCC1->PCC2 as it does not have the highest priority and cannot sub-delegate the LSP again towards PCE3.

When PCC3->PCC4 is configured, PCC3 delegates the LSP to PCE2 that performs sub-delegation to PCE3. As PCE3 will have knowledge of only one LSP in the group, it cannot compute disjointness and can decide to fallback to a less constrained computation to provide a path for PCC3->PCC4. In this case, it will send a PCUpdate to PCE2 that will be forwarded to PCC3.

Disjointness cannot be achieved in this scenario because of lack of state-sync session between PCE1 and PCE3, but no computation loop happens. Thus it is advised for all PCEs that support state-sync to have a full mesh sessions between each other.

5. Using Master/Slave computation and state-sync sessions to increase scaling

The Primary/Backup computation and state-sync sessions architecture can be used to increase the scaling of the PCE architecture. If the number of PCCs is really high, it may be too resource consuming for a single PCE to maintain all the PCEP sessions while at the same time performing all path computations. Using master/slave computation and state-sync sessions may allow to create groups of PCEs that manage a subset of the PCCs and perform some or no path computations. Decoupling PCEP session maintenance and computation will allow to increase scaling of the PCE architecture.



In the figure above, two groups of PCEs are created: PCE1/2 maintain PCEP sessions with PCC1 up to PCC500, while PCE3/4 maintain PCEP sessions with PCC501 up to PCC1000. A granular master/slave policy is setup as follows to loadshare computation between PCEs:

- PCE1 has priority 200 for association ID 1 up to 300, association source 0.0.0.0. All other PCEs have a decreasing priority for those associations.

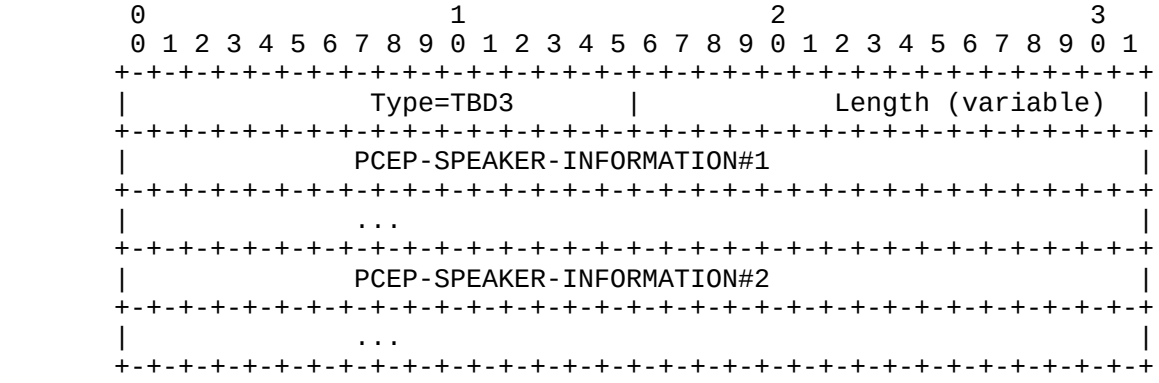
- PCE3 has priority 200 for association ID 301 up to 500, association source 0.0.0.0. All other PCEs have a decreasing priority for those associations.

If some PCCs delegate LSPs with association ID 1 up to 300 and association source 0.0.0.0, the receiving PCE (if not PCE1) will sub-delegate the LSPs to PCE1. PCE1 becomes responsible for the computation of these LSP associations while PCE3 is responsible for the computation of another set of associations.

6. PCEP-PATH-VECTOR-TLV

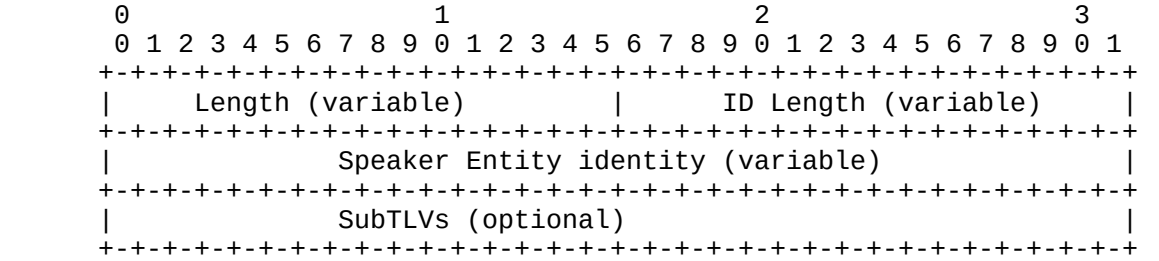
This document allows PCEP messages to be propagated among PCEP speaker. It may be useful to track informations about the propagation of the messages. One of the use case is a message loop detection mechanism, but other use cases like hop by hop information recording may also be implemented.

This document introduces the PCEP-PATH-VECTOR-TLV (type TBD2) with the following format:



The TLV format and padding rules are as per [RFC5440].

The PCEP-SPEAKER-INFORMATION field has the following format:



Length: defines the total length of the PCEP-SPEAKER-INFORMATION field.

ID Length: defines the length of the Speaker identity actual field (non-padded).

Speaker Entity identity: same possible values as the SPEAKER-IDENTIFIER-TLV. Padded with trailing zeroes to a 4-byte boundary.

The PCEP-SPEAKER-INFORMATION may also carry some optional subTLVs so each PCEP speaker can add local informations that could be recorded. This document does not define any subTLV.

The PCEP-PATH-VECTOR-TLV MAY be added in the LSP-Object. Its usage is purely optional.

The list of speakers within the PCEP-PATH-VECTOR-TLV MUST be ordered. When sending a PCEP message (PCReport, PCUpdate or PCInitiate), a PCEP Speaker MAY add the PCEP-PATH-VECTOR-TLV with a PCEP-SPEAKER-INFORMATION containing its own informations. If the PCEP message sent is the result of a previously received PCEP message, and if the PCEP-PATH-VECTOR-TLV was already present in the initial message, the PCEP speaker MAY append a new PCEP-SPEAKER-INFORMATION containing its own informations.

7. Security Considerations

TBD.

8. Acknowledgements

TBD.

9. IANA Considerations

This document requests IANA actions to allocate code points for the protocol elements defined in this document.

9.1. PCEP-Error Object

IANA is requested to allocate a new Error Value for the Error Type 9.

Error-Type	Meaning	Reference
6	Mandatory Object Missing	[RFC5440]
	Error-value=TBD1: SPEAKER-IDENTITY-TLV missing	This document

9.2. PCEP TLV Type Indicators

IANA is requested to allocate new TLV Type Indicator values within the "PCEP TLV Type Indicators" sub-registry of the PCEP Numbers registry, as follows:

Value	Meaning	Reference
TBD2	ORIGINAL-LSP-DB-VERSION-TLV	This document
TBD3	PCEP-PATH-VECTOR-TLV	This document

9.3. STATEFUL-PCE-CAPABILITY TLV

IANA is requested to allocate a new bit value in the STATEFUL-PCE-CAPABILITY TLV Flag Field sub-registry.

Bit	Description	Reference
TBD	INTER-PCE-CAPABILITY	This document

10. References

10.1. Normative References

[I-D.ietf-pce-stateful-pce]
Crabbe, E., Minei, I., Medved, J., and R. Varga, "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce-21 (work in progress), June 2017.

[I-D.ietf-pce-stateful-sync-optimizations]
Crabbe, E., Minei, I., Medved, J., Varga, R., Zhang, X., and D. Dhody, "Optimizations of Label Switched Path State Synchronization Procedures for a Stateful PCE", draft-ietf-pce-stateful-sync-optimizations-10 (work in progress), March 2017.

[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>.

[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>.

10.2. Informative References

[I-D.dhodylee-pce-stateful-hpce]
Dhody, D., Lee, Y., Ceccarelli, D., Shin, J., King, D., and O. Dios, "Hierarchical Stateful Path Computation Element (PCE).", draft-dhodylee-pce-stateful-hpce-03 (work in progress), March 2017.

[I-D.ietf-pce-association-diversity]
Litkowski, S., Sivabalan, S., Barth, C., and D. Dhody, "Path Computation Element communication Protocol extension for signaling LSP diversity constraint", draft-ietf-pce-association-diversity-01 (work in progress), March 2017.

[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>.

[RFC6805] King, D., Ed. and A. Farrel, Ed., "The Application of the Path Computation Element Architecture to the Determination of a Sequence of Domains in MPLS and GMPLS", RFC 6805, DOI 10.17487/RFC6805, November 2012, <https://www.rfc-editor.org/info/rfc6805>.

[RFC7399] Farrel, A. and D. King, "Unanswered Questions in the Path Computation Element Architecture", RFC 7399, DOI 10.17487/RFC7399, October 2014, <https://www.rfc-editor.org/info/rfc7399>.

Authors' Addresses

Stephane Litkowski
Orange

Email: stephane.litkowski@orange.com

Siva Sivabalan
Cisco

Email: msiva@cisco.com

Dhruv Dhody
Huawei
Divyashree Techno Park, Whitefield
Bangalore, Karnataka 560066
India

Email: dhruv.ietf@gmail.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: May 3, 2018

A. Raghuram
A. Goddard
C. Yadlapalli
AT&T
J. Karthik
S. Sivabalan
J. Parker
Cisco Systems, Inc.
D. Dhody
Huawei Technologies
October 30, 2017

Ability for a stateful PCE to request and obtain control of a LSP
draft-raghu-pce-lsp-control-request-05

Abstract

The stateful Path Computation Element (PCE) communication Protocol (PCEP) extensions provide stateful control of Multiprotocol Label Switching (MPLS) Traffic Engineering Label Switched Paths (TE LSP) via PCEP, for a model where a Path Computation Client (PCC) delegates control over one or more locally configured LSPs to a stateful PCE. There are use-cases in which a stateful PCE may wish to request and obtain control of one or more LSPs from a PCC. This document describes a simple extension to stateful PCEP to achieve such an objective.

Requirements Language

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.

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 May 3, 2018.

Table of Contents

1. Introduction

Stateful PCEP extensions [RFC8231] specifies a set of extensions to PCEP [RFC5440] 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. The stateful PCEP defines the following two useful network operations:

- Delegation: As per [RFC8051], an operation to grant a PCE temporary rights to modify a subset of LSP parameters on one or more LSPs of a PCC. LSPs are delegated from a PCC to a PCE and are referred to as "delegated" LSPs.

- Revocation: As per [RFC8231], an operation performed by a PCC on a previously delegated LSP. Revocation revokes the rights granted to the PCE in the delegation operation.

For Redundant Stateful PCEs (section 5.7.4. of [RFC8231]), during a PCE failure, one of the redundant PCE could request to take control over an LSP. The redundant PCEs MAY use a local policy or a proprietary election mechanism to decide which PCE would take control. In this case, a mechanism is needed for a stateful PCE to request control of one or more LSPs from a PCC, so that a newly elected primary PCE can request to take over control.

In case of virtualized PCEs (vPCE) running as virtual network function (VNF), as the computation load in the network increases, a new instance of vPCE could be instantiated to balance the current load. The PCEs could use proprietary algorithm to decide which LSPs to be assigned to the new vPCE. Thus having a mechanism for the PCE to request control of some LSPs is needed.

In some deployments, the operator would like to use stateful PCE for global optimization algorithms but would still like to keep the control of the LSP at the PCC. In such cases, a stateful PCE could request to take control during the global optimization and return the delegation once done.

This specification provides a simple extension, by using this a PCE can request control of one or more LSPs from any PCC over the stateful PCEP channel. The procedures for granting and relinquishing control of the LSPs are specified in accordance with the specification [RFC8231].

2. Terminology

The following terminologies are used in this document:

PCC: Path Computation Client.

PCE: Path Computation Element

PCEP: Path Computation Element communication Protocol.

PCRpt: Path Computation State Report message.

PCUpd: Path Computation Update Request message.

PLSP-ID: A PCEP-specific identifier for the LSP.

3. LSP Control Request Flag

The Stateful PCE Request Parameters (SRP) object is defined in [RFC8231], it includes a Flags field.
[I-D.ietf-pce-pce-initiated-lsp] defines a R (LSP-REMOVE) flag.

A new flag, the "LSP Control Request Flag" (C), is introduced in the SRP object. On a PCUpd message, a PCE sets the C Flag to 1 to indicate that, it wishes to gain control of LSP(s). The LSP is identified by the LSP object. A PLSP-ID of value other than 0 and 0xFFFFF is used to identify the LSP for which the PCE requests control. The PLSP-ID value of 0 indicates that the PCE is requesting control of all LSPs originating from the PCC that it wishes to delegate. The flag has no meaning in the PCRpt and PCInitiate message and SHOULD be set to 0 on transmission and MUST be ignored on receipt.

4. Operation

During normal operation, a PCC that wishes to delegate the control of an LSP sets the D Flag (delegate) to 1 in all PCRpt messages pertaining to the LSP. The PCE confirms the delegation by setting D Flag to 1 in all PCUpd messages pertaining to the LSP. The PCC revokes the control of the LSP from the PCE by setting D Flag to 0 in PCRpt messages pertaining to the LSP. If the PCE wishes to relinquish the control of the LSP, it sets D Flag to 0 in all PCUpd messages pertaining to the LSP.

If a PCE wishes to gain control over an LSP, it sends a PCUpd message with C Flag set to 1 in SRP object. The LSP for which the PCE requests control is identified by the PLSP-ID. The PLSP-ID of 0 indicates that the PCE wants control over all LSPs originating from the PCC. If the LSP(s) is/are already delegated to the PCE making the request, the PCC ignores the C Flag. A PCC can decide to delegate the control of the LSP at its own discretion. If the PCC grants or denies the control, it sends PCRpt message with D Flag set to 1 and 0 respectively in accordance with according with stateful PCEP [RFC8231] . If the PCC does not grant the control, it MAY choose to not respond, and the PCE may choose to retry requesting the

control preferably using exponentially increasing timer. A PCE ignores the C Flag on the PCRpt message.

In case multiple PCEs request control over an LSP, and if the PCC is willing to grant the control, the LSP MUST be delegated to only one PCE chosen by the PCC based on its local policy.

It should be noted that a legacy implementation of PCC, that does not understand the C flag in PCUpd message, would simply ignore the flag and the request to grant control over the LSP.

[I-D.ietf-pce-pce-initiated-lsp] describes the setup, maintenance and teardown of PCE-initiated LSPs under the stateful PCE model. It also specify how a PCE MAY obtain control over an orphaned LSP that was PCE-initiated. A PCE implementation can apply the mechanism described in this document in conjunction with those in [I-D.ietf-pce-pce-initiated-lsp].

5. Security Considerations

The security considerations listed in [RFC8231] apply to this document as well. However, this document also introduces a new attack vectors. An attacker may flood the PCC with request to delegate all its LSPs at a rate which exceeds the PCC's ability to process them, either by spoofing messages or by compromising the PCE itself. The PCC can simply ignore these messages with no extra actions. Securing the PCEP session using mechanism like Transport Layer Security (TLS) [RFC8253] is RECOMMENDED.

6. IANA Considerations

This document requests IANA actions to allocate code points for the protocol elements defined in this document.

6.1. SRP Object Flags

The SRP object is defined in [RFC8231] and the registry to manage the Flag field of the SRP object is requested in [I-D.ietf-pce-pce-initiated-lsp]. IANA is requested to make the following allocation in the aforementioned registry.

Bit	Description	Reference
TBD	LSP Control Request Flag (c-bit)	This document

7. Manageability Considerations

All manageability requirements and considerations listed in [RFC5440] and [RFC8231] apply to PCEP protocol extensions defined in this document. In addition, requirements and considerations listed in this section apply.

7.1. Control of Function and Policy

A PCE or PCC implementation SHOULD allow the operator to configure the policy based on which it honor the request to control the LSPs. Further, the operator MAY be to be allowed to trigger the LSP control request at the PCE.

7.2. Information and Data Models

The PCEP YANG module [I-D.ietf-pce-pcep-yang] could be extended to include mechanism to trigger the LSP control request.

7.3. 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].

7.4. Verify Correct Operations

Mechanisms defined in this document do not imply any new operation verification requirements in addition to those already listed in [RFC5440] and [RFC8231].

7.5. Requirements On Other Protocols

Mechanisms defined in this document do not imply any new requirements on other protocols.

7.6. Impact On Network Operations

Mechanisms defined in [RFC5440] and [RFC8231] also apply to PCEP extensions defined in this document. Further, the mechanism described in this document can help the operator to request control of the LSPs at a particular PCE.

8. Acknowledgements

Thanks to Jonathan Hardwick to remind the authors to not use suggested values in IANA section.

9. References

9.1. 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>.

[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>.

[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>.

[RFC8231] Crabbe, E., Minei, I., Medved, J., and R. Varga, "Path Computation Element Communication Protocol (PCEP) Extensions for Stateful PCE", RFC 8231, DOI 10.17487/RFC8231, September 2017, <https://www.rfc-editor.org/info/rfc8231>.

9.2. Informative References

[I-D.ietf-pce-pce-initiated-lsp] Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model", draft-ietf-pce-pce-initiated-lsp-11 (work in progress), October 2017.

[I-D.ietf-pce-pcep-yang] Dhody, D., Hardwick, J., Beeram, V., and j. jefftant@gmail.com, "A YANG Data Model for Path Computation Element Communications Protocol (PCEP)", draft-ietf-pce-pcep-yang-05 (work in progress), June 2017.

[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>.

[RFC8051] Zhang, X., Ed. and I. Minei, Ed., "Applicability of a Stateful Path Computation Element (PCE)", RFC 8051, DOI 10.17487/RFC8051, January 2017, <https://www.rfc-editor.org/info/rfc8051>.

[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>.

Authors' Addresses

Aswatnarayan Raghuram
AT&T
200 S Laurel Aevenue
Middletown, NJ 07748
USA

Email: ar2521@att.com

Al Goddard
AT&T
200 S Laurel Aevenue
Middletown, NJ 07748
USA

Email: ag6941@att.com

Chaitanya Yadlapalli
AT&T
200 S Laurel Aevenue
Middletown, NJ 07748
USA

Email: cy098d@att.com

Jay Karthik
Cisco Systems, Inc.
125 High Street
Boston, Massachusetts 02110
USA

Email: jakarthi@cisco.com

Siva Sivabalan
Cisco Systems, Inc.
2000 Innovation Drive
Kanata, Ontario K2K 3E8
Canada

Email: msiva@cisco.com

Jon Parker
Cisco Systems, Inc.
2000 Innovation Drive
Kanata, Ontario K2K 3E8
Canada

Email: jdparker@cisco.com

Dhruv Dhody
Huawei Technologies
Divyashree Techno Park, Whitefield
Bangalore, Karnataka 560066
India

Email: dhruv.ietf@gmail.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: December 29, 2017

Q. Wu
D. Dhody
Huawei
M. Boucadair
C. Jacquenet
Orange
J. Tantsura

June 27, 2017

PCEP Extensions for Service Function Chaining (SFC)
draft-wu-pce-traffic-steering-sfc-12

Abstract

This document provides an overview of the usage of Path Computation Element (PCE) to dynamically structure service function chains. Service Function Chaining (SFC) is a technique that is meant to facilitate the dynamic enforcement of differentiated traffic forwarding policies within a domain. Service function chains are composed of an ordered set of elementary Service Functions (such as firewalls, load balancers) that need to be invoked according to the design of a given service. Corresponding traffic is thus forwarded along a Service Function Path (SFP) that can be computed by means of PCE.

This document specifies extensions to the Path Computation Element Protocol (PCEP) that allow a stateful PCE to compute and instantiate Service Function Paths.

Table of Contents

1. Introduction

Service Function Chaining (SFC) enables the creation of composite services that consist of an ordered set of Service Functions (SF) that must be applied to packets and/or frames and/or flows selected as a result of service-inferred traffic classification as described in [RFC7665]. A Service Function Path (SFP) is a path along which traffic that is bound to a specific service function chain will be

forwarded. Packets typically follow a Service Function Path from a classifier through the Service Functions (SF) that need to be invoked according to the SFC instructions. Forwarding decisions are made by Service Function Forwarders (SFF) according to such instructions.

[RFC5440] describes the Path Computation Element Protocol (PCEP) as the protocol used by a Path Computation Client (PCC) and a Path Control Element (PCE) to exchange information, thereby enabling the computation of Multiprotocol Label Switching (MPLS) for Traffic Engineering Label Switched Path (TE LSP), in particular.

[I-D.ietf-pce-stateful-pce] specifies extensions to PCEP to enable a stateful control of MPLS TE LSPs. [I-D.ietf-pce-pce-initiated-lsp] provides the extensions needed for stateful PCE-initiated LSP instantiation.

This document specifies PCEP extensions that allow a stateful PCE to compute and instantiate traffic-engineered Service Function Paths (SFP).

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 [RFC2119].

This document makes use of these acronyms:

PCC: Path Computation Client.

PCE: Path Computation Element.

PCEP: Path Computation Element Protocol.

PDP: Policy Decision Point.

SF: Service Function.

SFC: Service Function Chain.

SFP: Service Function Path.

RSP: Rendered Service Path.

SFF: Service Function Forwarder.

UNI: User-Network Interface.

3. Service Function Paths and PCE

Service function chains are constructed as a sequence of SFs, where a SF can be virtualized or embedded in a physical network element. One or several SFs may be supported by the same physical network element. A SFC creates an abstracted view of a service and specifies the set of required SFs as well as the order in which they must be executed.

When an SFC is created, it is necessary to select the specific instances of SFs that will be used. A service function path for that SFC will then be established (notion of rendered service path) or can be precomputed, based upon the sequence of SFs that need to be invoked by the corresponding traffic, i.e., the traffic that is bound to the corresponding SFC. Note that a SF instance can be serviced by one or multiple SFFs. One or multiple SF instances can be serviced by one SFF. Thus, the instantiation of an SFC results in the establishment of a Service Function Path, either in a hop-by-hop fashion, or by means of traffic-engineering capabilities. In the latter case, the SFP is precomputed, i.e., an SFP is an instantiation of the defined SFC as described in [RFC7665].

The computation, the selection, and the establishment of a traffic-engineered SFP can rely upon a set of (service-specific) policies (forwarding and routing, QoS, security, etc., or a combination thereof). Stateful PCE with appropriate SFC-aware PCEP extensions can be used to compute traffic-engineered SFPs.

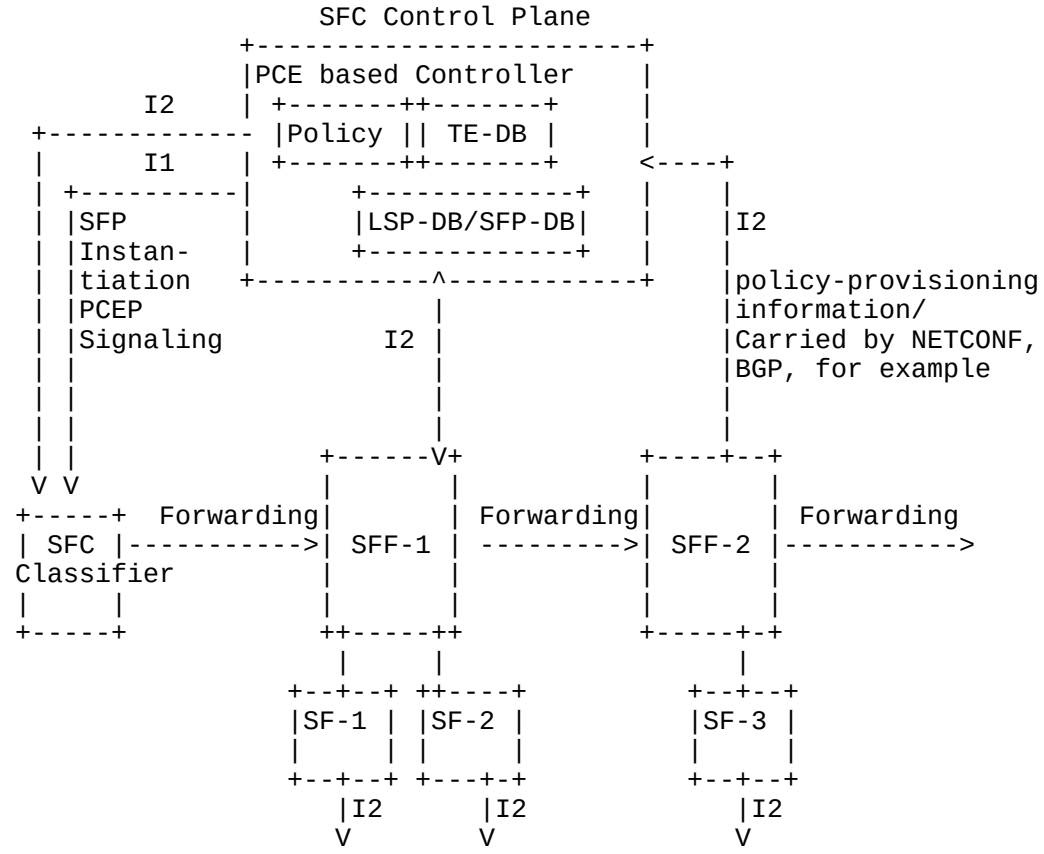


Figure 1: PCE-based SFP instantiation

In Figure 1, the PCE-based Controller [I-D.ietf-teas-pce-central-control] in the SFC Control plane is responsible for computing the path for a given service function chain. This PCE-based controller can operate as a stateful PCE ([I-D.draft_ietf-stateful-pce]) that will provide a classifier (a headend from a PCE standpoint) with the PCEP-formatted information to instantiate a given SFP. As a consequence, the PCE-based controller derives the set of policy-provisioning information (namely SFP configuration information and traffic classification rules) that will be provided to the various elements (Classifier, SFF) involved in the establishment of the SFP.

By doing so, SFC Classifier can bind a flow to a service function chain and forward such flow along the corresponding SFP. The SFC Control Plane [I-D.ietf-sfc-control-plane] is also responsible for defining the appropriate policies (traffic classification, forwarding and routing, etc.) that will be enforced by SFC Classifiers,SFF Nodes

and SF Nodes, as described in [RFC7665]. From that standpoint, the SFC Control Plane embeds a Policy Decision Point that is responsible for defining the SFC policies. SFC policies will be provided by the PDP and enforced by SFC components like classifiers and SFFs by means of policy-provision information. A protocol like NETCONF, BGP can be used to carry such policy-provisioning information.

4. Overview of PCEP Operation in SFC-Enabled Networks

A PCEP speaker indicates its ability to support PCE-computed SFP paths during the PCEP Initialization phase via a mechanism described in Section 5.1. A PCE may initiate SFPs only for PCCs that advertised this capability; a PCC follows the procedures described in this document only for sessions where the PCE advertised this capability.

As per Section 5.1 of [I-D.ietf-pce-pce-initiated-lsp], the PCE sends a Path Computation LSP Initiate Request (PCInitiate) message to the PCC to instantiate or delete a LSP. The Explicit Route Object (ERO) is used to encode either a full sequence of SF instances or a specific sequence of SFFs and SFs to establish an SFP. If the said SFFs and SFs are identified with an IP address, the IP sub-object can be used as a SF/SFF identification means. This document makes no change to the PCInitiate message format but extends LSP objects described in Section 5.2.

Editor’s note: In case a PCE-Initiated signaling mechanism is used to set up the service function path, does the classifier / PCE-Initiated signaling protocol need to understand whether an IP address is assigned to a SFF or a SF, or the signaling protocol is only used to signal IP addresses for SFs?

To prevent multiple classifiers assign the same SFP ID to one Service Function Path(SFP ID assignment conflict),in this document, we assume SFP ID can be predetermined and assigned by stateful PCE when stateful PCE can be used to compute traffic-engineered SFPs.

4.1. SFP Instantiation

The instantiation of a SFP is the same as defined in Section 5.3 of [I-D.ietf-pce-pce-initiated-lsp]. Rules for processing and error codes remain unchanged.

4.2. SFP Withdrawal

The withdrawal of an SFP is the same as defined in Section 5.4 of [I-D.ietf-pce-pce-initiated-lsp]: the PCE sends an LSP Initiate Message with an LSP object carrying the PLSP-ID of the SFP and the

SFP Identifier to be removed, as well as an SRP object with the R flag set (LSP-REMOVE as per Section 5.2 of [I-D.ietf-pce-pce-initiated-lsp]). Rules for processing and error codes remain unchanged.

4.3. SFP Delegation and Cleanup

SFP delegation and cleanup operations are similar to those defined in Section 6 of [I-D.ietf-pce-pce-initiated-lsp]. Rules for processing and error codes remain unchanged.

4.4. SFP State Synchronization

State Synchronization operations described in Section 5.4 of [I-D.ietf-pce-stateful-pce] can be applied to SFP state maintenance as well.

4.5. SFP Update and Report

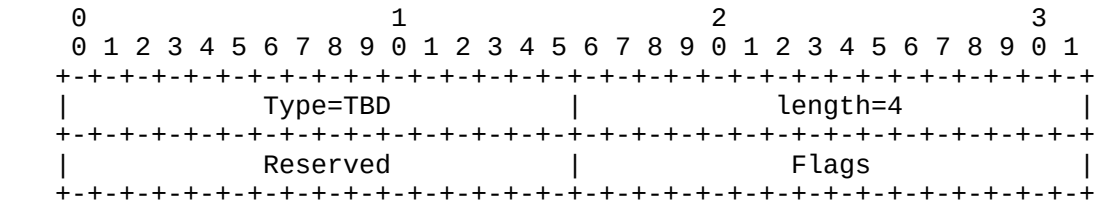
A PCE can send an SFP Update request to a PCC to update one or more attributes of an SFP and to re-signal the SFP with the updated attributes. A PCC can send an SFP state report to a PCE, and which contains the SFP State information. The mechanism is described in [I-D.ietf-pce-stateful-pce] and can be applied to SFPs as well.

5. Object Formats

5.1. The OPEN Object

The optional TLV shown in Figure 2 is defined for use in the OPEN Object to indicate the PCEP speaker's Service Function Chaining capability.

The SFC-PCE-CAPABILITY TLV is an optional TLV to be carried in the OPEN Object to advertise the SFC capability during the PCEP session.



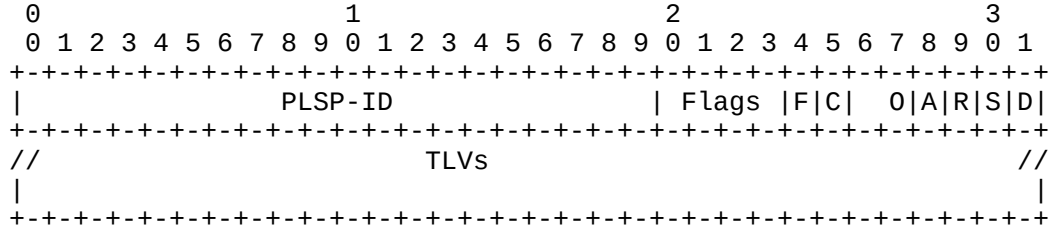
SFC-PCE-CAPABILITY TLV Format

The code point for the TLV type is to be defined by IANA (see Section 9). The TLV length is 4 octets.

As per [I-D.ietf-pce-stateful-pce], a PCEP speaker advertises the capability of instantiating PCE-initiated LSPs via the Stateful PCE Capability TLV (LSP-INSTANTIATION-CAPABILITY bit) carried in an Open message. The inclusion of the SFC-PCE-CAPABILITY TLV in an OPEN object indicates that the sender is SFC-capable. Both mechanisms indicate the SFP instantiation capability of the PCEP speaker.

5.2. The LSP Object

The LSP object is defined in [I-D.ietf-pce-pce-initiated-lsp] and included here for reference (Figure 3).



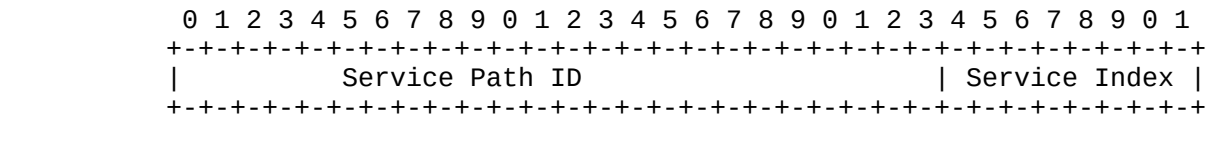
LSP Object Format

A new flag, called the SFC flag (F-bit), is introduced. The F-bit set to "1" indicates that this LSP is actually an SFP. The C flag will also be set to indicate it was created via a PCInitiate message.

5.2.1. SFP Identifiers TLV

As described in section 4, SFP ID is predetermined and assigned by stateful PCE. The SFP Identifiers TLV MUST be included in the LSP object for SFPs. The SFP Identifier TLV is used by the classifier to select the SFP along which some traffic will be forwarded, according to the traffic classification rules applied by the classifier [RFC7665]. The SFP Identifier is part of the SFC metadata carried in packets and is used by the SFF to invoke service functions and identify the next SFF.

The format of the SFP Identifier TLV is shown in Figure 4.



Service Path ID (SPI): 24 bits
Service Index (SI): 8 bits

Figure 4

SPI: identifies a service path. The same ID is used by the participating nodes for path setup/selection. An administrator can use the SPI for reporting and troubleshooting packets along a specific path. SPI along with PLSP-ID is used by PCEP to identify the Service Path.

SI: provides location within the service path.

6. Backward Compatibility

The SFP instantiation capability defined as a PCEP extension and documented in this draft MUST NOT be used if PCCs or the PCE did not advertise their stateful SFP instantiation capability,Section 5.1. If this is not the case and stateful operations on SFPs are attempted, then a PCErr message with error-type 19 (Invalid Operation) and error-value TBD needs to be generated.

[Editor’s note: more information on exact error value is needed]

7. SFP Instantiation Signaling and Forwarding Considerations

The PCE-initiated SFP instantiation signaling described in this document is exchanged between PCE server and SFC Classifier and does not assume any specific mechanism to exchange SFP information(e.g.,path identification information,metadata [I-D.ietf-sfc-nsh]) between SFFs or between SFF and SF, or between the controller and SFF and establish SFP in the data plane throughout a SFC domain. For example, such mechanism can rely upon the use of the SFC Encapsulation defined in [I-D.ietf-sfc-nsh] to exchange SFP information between SFFs or rely upon the use of BGP Control plane defined in [I-D.ietf-bess-nsh-bgp-control-plane] to exchange SFP information between the Controller and SFF.

Likewise, [I-D.ietf-teas-pce-central-control] can use the signaling mechanism described in this draft to enforce SFC-inferred traffic engineering policies and provide load balancing between service function nodes. The approach that relies upon the Segment Routing technique [I-D.ietf-pce-segment-routing] can also take advantage of

the signaling mechanism described in this document to support Service Path instantiation, which does not require any additional specific extension to the Segment Routing machinery.

8. Security Considerations

The security considerations described in [RFC5440] and [I-D.ietf-pce-pce-initiated-lsp] are applicable to this specification. This document does not raise any additional security issue.

9. IANA Considerations

IANA is requested to allocate a new code point in the PCEP TLV Type Indicators registry, as follows:

Value	Meaning	Reference
TBD	SFC-PCE-CAPABILITY	This document

10. Acknowledgements

Many thanks to Ron Parker, Hao Wang, Dave Dolson, Jing Huang, and Joel M. Halpern for the discussion about the content for the document.

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, <http://www.rfc-editor.org/info/rfc2119>.

[I-D.ietf-pce-stateful-pce] Crabbe, E., Minei, I., Medved, J., and R. Varga, "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce-21 (work in progress), June 2017.

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

[I-D.ietf-pce-pce-initiated-lsp]
Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model", draft-ietf-pce-pce-initiated-lsp-10 (work in progress), June 2017.

[I-D.ietf-teas-pce-central-control]
Farrel, A., Zhao, Q., Li, Z., and C. Zhou, "An Architecture for Use of PCE and PCEP in a Network with Central Control", draft-ietf-teas-pce-central-control-03 (work in progress), June 2017.

11.2. Informative References

[RFC2753] Yavatkar, R., Pendarakis, D., and R. Guerin, "A Framework for Policy-based Admission Control", RFC 2753, DOI 10.17487/RFC2753, January 2000, <http://www.rfc-editor.org/info/rfc2753>.

[RFC7665] Halpern, J., Ed. and C. Pignataro, Ed., "Service Function Chaining (SFC) Architecture", RFC 7665, DOI 10.17487/RFC7665, October 2015, <http://www.rfc-editor.org/info/rfc7665>.

[RFC5394] Bryskin, I., Papadimitriou, D., Berger, L., and J. Ash, "Policy-Enabled Path Computation Framework", RFC 5394, DOI 10.17487/RFC5394, December 2008, <http://www.rfc-editor.org/info/rfc5394>.

[I-D.ietf-sfc-control-plane]
Boucadair, M., "Service Function Chaining (SFC) Control Plane Components & Requirements", draft-ietf-sfc-control-plane-08 (work in progress), October 2016.

[I-D.ietf-pce-segment-routing]
Sivabalan, S., Filsfils, C., Tantsura, J., Henderickx, W., and J. Hardwick, "PCEP Extensions for Segment Routing", draft-ietf-pce-segment-routing-09 (work in progress), April 2017.

[I-D.ietf-sfc-nsh]
Quinn, P. and U. Elzur, "Network Service Header", draft-ietf-sfc-nsh-12 (work in progress), February 2017.

[I-D.ietf-bess-nsh-bgp-control-plane]
Farrel, A., Drake, J., Rosen, E., Uttaro, J., and L. Jalil, "BGP Control Plane for NSH SFC", draft-ietf-bess-nsh-bgp-control-plane-00 (work in progress), March 2017.

Authors' Addresses

Qin Wu
Huawei
101 Software Avenue, Yuhua District
Nanjing, Jiangsu 210012
China

EMail: bill.wu@huawei.com

Dhruv Dhody
Huawei
Leela Palace
Bangalore, Karnataka 560008
INDIA

EMail: dhruv.ietf@gmail.com

Mohamed Boucadair
Orange
Rennes 35000
France

EMail: mohamed.boucadair@orange.com

Christian Jacquenet
Orange
Rennes
France

EMail: christian.jacquenet@orange.com

Jeff Tantsura
2330 Central Expressway
Santa Clara, CA 95050
US

EMail: jefftant.ietf@gmail.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: June 25, 2017

Y. Zhuang
Q. Wu
D. Dhody
Huawei
D. Ceccarelli
Ericsson
December 22, 2016

PCEP Extensions for LSP scheduling with stateful PCE
draft-zhuang-pce-stateful-pce-lsp-scheduling-04

Abstract

This document proposes a set of extensions needed to the stateful Path Computation Element (PCE) communication Protocol (PCEP), so as to enable Labeled Switched Path (LSP) scheduling for path computation and LSP setup/deletion based on the actual network resource usage duration of a traffic service in a centralized network environment as stated in [I.D.ietf-teas-scheduled-resources].

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 25, 2017.

Copyright Notice

Table of Contents

1. Introduction

The Path Computation Element Protocol (PCEP) defined in [RFC5440] is used between a Path Computation Element (PCE) and a Path Computation Client (PCC) (or other PCE) to enable computation of Multi-protocol Label Switching (MPLS) for Traffic Engineering Label Switched Path (TE LSP).

Further, in order to support use cases described in [I-D.ietf-pce-stateful-pce-app], [I-D.ietf-pce-stateful-pce] specifies a set of extensions to PCEP to enable stateful control of MPLS-TE and GMPLS LSPs via PCEP.

Traditionally, the usage and allocation of network resources, especially bandwidth, can be supported by a Network Management System operation such as path pre-establishment. However, this does not provide efficient network usage since the established paths exclude the possibility of being used by other services even when they are not used for undertaking any service. [I-D.ietf-teas-scheduled-resources] then provides a framework that describes and discusses the problem and propose an appropriate architecture for the scheduled reservation of TE resources.

With the scheduled reservation of TE resources, it allows network operators to reserve resources in advance according to the agreements with their customers, and allow them to transmit data with scheduling such as specified starting time and duration, for example for a scheduled bulk data replication between data centers. It enables the activation of bandwidth usage at the time the service really being used while letting other services obtain it in spare time. The requirement of scheduled LSP provision is mentioned in [I-D.ietf-pce-stateful-pce-app] and [RFC7399], so as to provide more efficient network resource usage for traffic engineering, which hasn’t been solved yet. Also, for deterministic networks, the scheduled LSP can provide a better network resource usage for guaranteed links. This idea can also be applied in segment routing to schedule the network resources over the whole network in a centralized manner as well.

With this in mind, this document proposes a set of extensions needed to the stateful PCE, so as to enable LSP scheduling for path computation and LSP setup/deletion based on the actual network resource usage duration of a traffic service. A scheduled LSP is characterized by a starting time and a duration. When the end of the LSP life is reached, it is deleted to free up the resources for other LSP (scheduled or not).

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 [RFC2119].

2.1. Terminology

The following terminologies are re-used from existing PCE documents.

- Active Stateful PCE [I-D.ietf-pce-stateful-pce];
- Delegation [I-D.ietf-pce-pce-initiated-lsp];
- PCC [RFC5440], [I-D.ietf-pce-stateful-pce];
- PCE [RFC5440], [I-D.ietf-pce-stateful-pce];
- TE LSP [RFC5440], [I-D.ietf-pce-stateful-pce];
- TED [RFC5440], [I-D.ietf-pce-stateful-pce];
- LSP DB [RFC5440], [I-D.ietf-pce-stateful-pce];

In addition, this document defines the following terminologies.

Scheduled TE LSP: a LSP with the scheduling attributes,that carries traffic flow demand at an starting time and last for a certain duration. The PCE operates path computation per LSP availability at the required time and duration.

Scheduled LSP DB: a database of scheduled LSPs

Scheduled TED: Traffic engineering database with the awareness of scheduled resources for TE. This database is generated by the PCE from the information in TED and scheduled LSP DB and allows knowing, at any time, the amount of available resources (does not include failures in the future).

Starting time: This value indicates when the scheduled LSP is used and the corresponding LSP must be setup and active. In other time(i.e., before the starting time or after the starting time plus Duration), the LSP can be inactive to include the possibility of the resources being used by other services.

Duration: The value indicates the time duration that the LSP is undertaken by a traffic flow and the corresponding LSP must be setup and active. At the end of which, the LSP is teardown and removed from the data base.

3. Motivation and Objectives

A stateful PCE can support better efficiency by using LSP scheduling described in the use case of [I-D.ietf-pce-stateful-pce]. This requires the PCE to maintain the scheduled LSPs and their associated resource usage, e.g. bandwidth for Packet-switched network, as well as the ability to trigger signaling for the LSP setup/tear-down at the correct time.

Note that existing configuration tools can be used for LSP scheduling, but as highlighted in section 3.1.3 of [I-D.ietf-pce-stateful-pce] as well as discussions in [I-D.ietf-teas-scheduled-resources], doing this as a part of PCEP in a centralized manner, has obvious advantages.

The objective of this document is to provide a set of extensions to PCEP to enable LSP scheduling for LSPs creation/deletion under the stateful PCE control, according to traffic services from customers, so as to improve the usage of network resources.

4. Architecture Overview

4.1. LSP scheduling Overview

The LSP scheduling allows PCEs and PCCs to provide scheduled LSP for customers' traffic services at its actual usage time, so as to improve the network resource efficient utilization.

For stateful PCE supporting LSP scheduling, there are two types of LSP databases used in this document. One is the LSP-DB defined in PCEP [I-D.ietf-pce-stateful-pce], while the other is the scheduled LSP database (SLSP- DB, see section 6). The SLSP-DB records scheduled LSPs and is used as a complementary to the TED and LSP-DB. Note that the two types of LSP databases can be implemented in one physical database or two different databases. This document does not state any preference here.

Furthermore, a scheduled TED can be generated from the scheduled LSP DB, LSP DB and TED to indicate the network links and nodes with resource availability information for now and future. The scheduled TED should be maintained by all PCEs within the network environment.

In case of implementing PCC-initiated scheduled LSPs, a PCC can request a path computation with LSP information of its scheduling parameters, including the starting time and the duration. Upon receiving the request with the scheduled LSP delegation, a stateful PCE SHALL check the scheduled TED for the network resource

availability on network nodes and computes a path for the LSP with the scheduling information.

For a multiple PCE environment, in order to coordinate the scheduling request of the LSP path over the network, the PCE needs to send a requestmessage with the path information as well as the scheduled resource for the scheduled LSP to other PCEs within the network, so as to coordinate with their scheduled LSP DBs and scheduled TEDs. Once other PCEs receive the request message with the scheduled LSPs information, if not conflicting with their scheduled LSP DBs, they reply to the requesting PCE with a response message carrying the scheduled LSP and update their scheduled LSP DBs and scheduled TEDs. After the requesting PCE confirms with all PCEs, the PCE SHALL add the scheduled LSP into its scheduled LSP Database and update its scheduled TED.

Then the stateful PCE can response to the PCC with the path for the scheduled LSP to notify the result of the computation. However, the PCC should not signal the LSP over the path once receiving these messages since the path is not activated yet until its starting time.

Alternatively, the service can also be initiated by PCE itself. In case of implementing PCE-initiated scheduled LSP, the stateful PCE shall check the network resource availability for the traffic and computes a path for the scheduled LSP per request in the same way as in PCC- Initiated mode and then for a multiple PCE network environment, coordinate the scheduled LSP with other PCEs in the network in the same way as in the PCC-Initiated mode.

In both modes, for activation of scheduled LSPs, the stateful PCE can send a path computation LSP Initiate (PCInitiate message) with LSP information at its starting time to the PCC for signaling the LSP over the network nodes as defined in [I-D.ietf-pce-pce- initiated-lsp]. Also, in the PCC-initiated mode, with scheduling information ,the PCC can activate the LSP itself by triggering over the path at its starting time as well. When the scheduling usage expires, active stateful PCE SHALL remove the LSP from the network , as well as notify other PCEs to delete the scheduled LSP from the scheduled LSP database.

4.2. Support of LSP Scheduling

4.2.1. Stateful PCE Capability TLV

After a TCP connection for a PCEP session has been established, a PCC and a PCE indicates its ability to support LSP scheduling during the PCEP session establishment phase. For a multiple-PCE environment, the PCEs should also establish PCEP session and indicate its ability

to support LSP scheduling among PCEP peers. The Open Object in the Open message contains the STATEFUL-PCE-CAPABILITY TLV defined in [I-D.ietf-pce-stateful-pce]. Note that the STATEFUL- PCE-CAPABILITY TLV is defined in [I-D.ietf-pce-stateful- pce] and updated in [I-D.ietf-pce-pce-initiated-lsp] and [I-D.ietf- pce-stateful-sync-optimizations]. In this document, we define a new flag bit B (SCHED-LSP-CAPABLITY) flag for the STATEFUL- PCE-CAPABILITY TLV to indicate the support of LSP scheduling.

B (LSP-SCHEDULING-CAPABILITY - 1 bit): If set to 1 by a PCC, the B Flag indicates that the PCC allows LSP scheduling; if set to 1 by a PCE, the B Flag indicates that the PCE is capable of LSP scheduling. The B bit MUST be set by both PECP peers in order to support LSP scheduling for path computation.

4.3. Scheduled LSP creation

In order to realize PCC-Initiated scheduled LSP in a centralized network environment, a PCC has to separate the setup of a LSP into two steps. The first step is to request and get a LSP but not signal it over the network. The second step is to signal the scheduled LSP over the LSRs (Labeled switched Router) at its starting time.

For PCC-Initiated scheduled LSPs, a PCC can send a path computation request (PCReq) message (see section 4.3.1) or a path computation LSP report (PCRpt) message (see section 4.3.1) including its demanded resources with the scheduling information and delegation to a stateful PCE.

Upon receiving the delegation via PCRpt message, the stateful PCE computes the path for the scheduled LSP per its starting time and duration based on the network resource availability stored in scheduled TED (see section 4.1).

If a resultant path is found, the stateful PCE will send a PCReq message with the path information as well as the scheduled resource information for the scheduled LSP to other PCEs within the network if there is any, so as to keep their scheduling information synchronized.

Once other PCEs receive the PCReq message with the scheduled LSP, if not conflicts with their scheduled LSP DBs, they will reply to the requesting PCE with a PCRep message carrying the scheduled LSP and update their scheduled LSP DBs and scheduled TEDs. After the requesting PCE confirms with all PCEs, the PCE SHALL add the scheduled LSP into its scheduled LSP DB and update its scheduled TED. If conflicts happen or no path available is found, the requesting PCE SHALL return a PCRep message with NO PATH back to the PCC.

Otherwise, the stateful PCE will send a PCRep message or PCUpd message (see section 4.3.3) with the path information back to the PCC as confirmation.

For PCE-Initiated Scheduled LSP, the stateful PCE can compute a path for the scheduled LSP per requests from network management systems automatically based on the network resource availability in the scheduled TED and coordinate with other PCEs on the scheduled LSP in the same way as in the PCC- Initiated mode.

In both modes:

- o the stateful PCE is required to update its local scheduled LSP DB and scheduled TED with the scheduled LSP. Besides, it shall send a PCReq message with the scheduled LSP to other PCEs within the network, so as to achieve the scheduling traffic engineering information synchronization.

- o Upon receiving the PCRep message or PCUpd message for scheduled LSP from PCEs with a found path, the PCC knows that it gets a scheduled path for the LSP but not trigger signaling for the LSP setup on LSRs.

- o In any case, stateful PCE can update the Scheduled LSP parameters on any network events using the PCUpd message to PCC as well as other PCEs.

4.3.1. The PCReq message and PCRpt Message

After scheduled LSP capability negotiation, for PCC-Initiated mode, a PCC can send a PCReq message or a PCRpt message including the SCHED-LSP- ATTRIBUTE TLV (see section 4.3.4.1) carried in the LSP Object (see section 4.3.4) body to indicate the requested LSP scheduling parameters for a customer’s traffic service with the delegation bit set to 1 in LSP Object. The value of requested bandwidth is taken via the existing ’Requested Bandwidth with BANDWIDTH Object- Type as 1’ defined in [RFC5440].

Meanwhile, for both modes (PCC-Initiated and PCE-Initiated), the delegated PCE shall distribute the scheduling information to other PCEs in the environment by sending a PCReq message with the SCHED-LSP-ATTRIBUTE TLV, as well as the Bandwith Object and RRO for the found path.

The definition of the PCReq message and PCRpt message to carry LSP objects (see [I- D.ietf-pce-stateful-pce]) remains unchanged.

4.3.2. The PCRep Message

To provide scheduled LSP for TE-LSPs, the stateful PCE SHALL compute the path for the scheduled LSP carried on PCReq message based on network resource availability recorded in scheduled TED which is generated from the scheduled LSP-DB and TED and also synchronize the scheduling with other PCEs in the environment by using PCReq message with path and resource information for the scheduled LSP.

If no conflict exists, other PCEs SHALL send a PCRep message with the SCHED-LSP-ATTRIBUTE TLV, as well as the Bandwith Object and RRO back to the requesting PCE.

If the LSP request can be satisfied and an available path is found, the stateful PCE SHALL send a PCRep Message including the SCHED- LSP-ATTRIBUTE TLV in the LSP Object body, as well as the Bandwith Object and RRO for the found path back to the PCC as a successful acknowledge.

4.3.3. The PCUpd Message

To provide scheduled LSP for TE-LSPs, the stateful PCE SHALL compute the path for the scheduled LSP carried on PCRpt message based on network resource availability recorded in scheduled TED which is generated from the scheduled LSP-DB, LSP DB and TED.

If the request can be satisfied and an available path is found, the stateful PCE SHALL send a PCUpd Message including the SCHED- LSP-ATTRIBUTE TLV in the LSP Object body to the PCC Note that, the stateful PCE can update the Scheduled LSP parameters later as well based on any network events using the same PCUpd message.

4.3.4. LSP Object

The LSP object is defined in [I-D.ietf-pce-stateful-pce]. This document add an optional SCHED-LSP-ATTRIBUTE TLV.

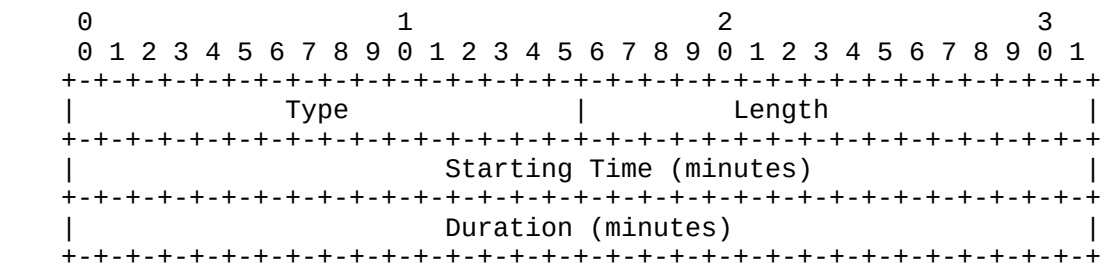
The presence of SCHED-LSP-ATTRIBUTE TLV in the LSP object indicates that this LSP is requesting scheduled parameters. The TLV MUST be present in LSP Object for each scheduled LSP carried in the PCReq message, the PCRpt message and the PCUpd message.

4.3.4.1. SCHED-LSP-ATTRIBUTE TLV

The SCHED-LSP-ATTRIBUTE TLV can be included as an optional TLV within the LSP object for LSP scheduling for the requesting traffic service.

This TLV SHOULD be included only if both PCEP peers have set the B (LSP-SCHEDULING-CAPABILITY bit) in STATEFUL-PCE-CAPABILITY TLV carried in open message.

The format of the SCHED-LSP-ATTRIBUTE TLV is shown in the following figure:



The type of the TLV is [TBD] and it has a fixed length of 8 octets.

The fields in the format are:

Starting Time (32 bits): This value in minutes, indicates when the scheduled LSP is used and the corresponding LSP must be setup and activated. At the expiry of this time, the LSP is setup. Otherwise, the LSP is inactive to include the possibility of the resources to be used by other services. The

Duration (32 bits): The value in minutes, indicates the duration that the LSP is undertaken by a traffic flow and the corresponding LSP must be up to carry traffic. At the expiry of this time after setup, the LSP is tear down and deleted.

Note, that the values of starting time and duration is from the perspective of the PCEP peer that is sending the message, also note the unit of time is minutes, and thus the time spent on transmission on wire can be easily ignored.

Editor Note: As described in [I-D.zhuang-teas-scheduled-resources],the encoding of the resource state information could also be expressed as a start time and and end time. Multiple periods, possibly of different lengths, may be associated with one reservation request, and a reservation might repeat on a regular cycle.

4.4. Scheduled LSP activation and deletion

In PCC-Initiated LSP scheduling, the PCC itself MAY activate the scheduled LSP at the starting time. Alternatively, the stateful PCE MAY activate the scheduled LSP at its scheduled time by send a PCInitiated message.

After the scheduled duration expires, the PCE shall send a PCUpd message with R flag set to the PCC to delete the LSP over the path, as well as to other PCEs to remove the scheduled LSP in the databases. Additionally, it shall update its scheduled LSP DB and scheduled TED.

Note that, the stateful PCE can update the Scheduled LSP parameters at any time based on any network events using the PCUpd message including SCHED-LSP-ATTRIBUTE TLV in the LSP Object body.

5. Security Considerations

This document defines LSP-SCHEDULING-CAPABILITY TLV and SCHED- LSP-ATTRIBUTE TLV which does not add any new security concerns beyond those discussed in [RFC5440] and [I-D.ietf-pce-stateful-pce].

6. Manageability Consideration

6.1. Control of Function and Policy

The LSP-Scheduling feature MUST BE controlled per tunnel by the active stateful PCE, the values for parameters like starting time, duration SHOULD BE configurable by customer applications and based on the local policy at PCE.

6.2. Information and Data Models

[RFC7420] describes the PCEP MIB, there are no new MIB Objects for this document.

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 [RFC5440].

6.4. Verify Correct Operations

Mechanisms defined in this document do not imply any new operation verification requirements in addition to those already listed in [RFC5440].

6.5. Requirements On Other Protocols

Mechanisms defined in this document do not imply any new requirements on other protocols.

6.6. 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].

7. IANA Considerations

7.1. PCEP TLV Type Indicators

This document defines the following new PCEP TLV; IANA is requested to make the following allocations from this registry.

Value	Meaning	Reference
TBD	SCHED-LSP-ATTRIBUTE	This document

7.2. LSP-SCHEDULING-CAPABLITY

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
28	LSP-SCHEDULING-CAPABILITY (B-bit)	This document

8. Acknowledgments

This work has benefited from the discussions of resource scheduling on the mailing list and with Huaimo chen, author of [I-D.chen-pce-tts] since Prague meeting. We gratefully acknowledge the contributions of Huaimo Chen. The authors of this document would also like to thank Rafal Szarecki,Adrian Farrel, Cyril Margaria, Xian Zhang for the review and comments.

9. References

9.1. Normative References

[I-D.dhody-pce-stateful-pce-auto-bandwidth]
Dhody, D., Palle, U., Singh, R., Gandhi, R., and L. Fang, "PCEP Extensions for MPLS-TE LSP Automatic Bandwidth Adjustment with Stateful PCE", draft-dhody-pce-stateful-pce-auto-bandwidth-09 (work in progress), November 2016.

[I-D.ietf-pce-pce-initiated-lsp]
Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model", draft-ietf-pce-pce-initiated-lsp-07 (work in progress), July 2016.

[I-D.ietf-pce-stateful-pce]
Crabbe, E., Minei, I., Medved, J., and R. Varga, "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce-18 (work in progress), December 2016.

[I-D.ietf-pce-stateful-sync-optimizations]
Crabbe, E., Minei, I., Medved, J., Varga, R., Zhang, X., and D. Dhody, "Optimizations of Label Switched Path State Synchronization Procedures for a Stateful PCE", draft-ietf-pce-stateful-sync-optimizations-07 (work in progress), December 2016.

[I-D.ietf-teas-scheduled-resources]
Zhuangyan, Z., Wu, Q., Chen, H., and A. Farrel, "Architecture for Scheduled Use of Resources", draft-ietf-teas-scheduled-resources-01 (work in progress), November 2016.

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

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

9.2. Informative References

[I-D.ietf-pce-stateful-pce-app]
Zhang, X. and I. Minei, "Applicability of a Stateful Path Computation Element (PCE)", draft-ietf-pce-stateful-pce-app-08 (work in progress), October 2016.

[RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, DOI 10.17487/RFC5226, May 2008, <http://www.rfc-editor.org/info/rfc5226>.

[RFC7420] Koushik, A., Stephan, E., Zhao, Q., King, D., and J. Hardwick, "Path Computation Element Communication Protocol (PCEP) Management Information Base (MIB) Module", RFC 7420, DOI 10.17487/RFC7420, December 2014, <http://www.rfc-editor.org/info/rfc7420>.

Appendix A. Scheduled LSP information synchronization

As for a stateful PCE, it maintains a database of LSPs (LSP-DB) that are active in the network, so as to reveal the available network resources and place new LSPs more cleverly.

With the scheduled LSPs, they are not activated while creation, but should be considered when operating future path computation. Hence, a scheduled LSP Database (SLSP-DB) is suggested to maintain all scheduled LSP information.

The information of SLSP-DB MUST be shared and synchronized among all PCEs within the centralized network by using PCReq message, PCRep message with scheduled LSP information. In order to synchronize the scheduled LSP information in SLSP-DB among PCEs, the PCReq message and PCRep Message is used as described in section 4.3.1 and section 4.3.2.

To achieve the synchronization, the PCE should generate and maintain a scheduled TED based on LSP DB, scheduled LSP DB and TED, which is used to indicate the network resource availability on network nodes for LSP path computation.

Appendix B. Contributor Addresses

Zitao Wang
Huawei
101 Software Avenue, Yuhua District
Nanjing, Jiangsu 210012
China

Email: wangzitao@huawei.com

Xian Zhang
Huawei Technologies
Research Area F3-1B,
Huawei Industrial Base,
Shenzhen, 518129, China

Email: zhang.xian@huawei.com

Authors' Addresses

Yan Zhuang
Huawei
101 Software Avenue, Yuhua District
Nanjing, Jiangsu 210012
China

Email: zhuangyan.zhuang@huawei.com

Qin Wu
Huawei
101 Software Avenue, Yuhua District
Nanjing, Jiangsu 210012
China

Email: bill.wu@huawei.com

Dhruv Dhody
Huawei
Divyashree Techno Park, Whitefield
Bangalore, Karnataka 560066
India

Email: dhruv.ietf@gmail.com

Daniele Ceccarelli
Ericsson
Via A. Negrone 1/A
Genova - Sestri Ponente
Italy

Email: daniele.ceccarelli@ericsson.com

PCE Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 28, 2017

H. Chen, Ed.
Y. Zhuang, Ed.
Q. Wu
D. Dhody
Huawei
D. Ceccarelli
Ericsson

March 27, 2017

PCEP Extensions for LSP scheduling with stateful PCE
draft-zhuang-pce-stateful-pce-lsp-scheduling-05

Abstract

This document proposes a set of extensions needed to the stateful Path Computation Element (PCE) communication Protocol (PCEP), so as to enable Labeled Switched Path (LSP) scheduling for path computation and LSP setup/deletion based on the actual network resource usage duration of a traffic service in a centralized network environment as stated in [I.D.ietf-teas-scheduled-resources].

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 September 28, 2017.

Copyright Notice

Table of Contents

1. Introduction

The Path Computation Element Protocol (PCEP) defined in [RFC5440] is used between a Path Computation Element (PCE) and a Path Computation Client (PCC) (or other PCE) to enable computation of Multi-protocol Label Switching (MPLS) for Traffic Engineering Label Switched Path (TE LSP).

Further, in order to support use cases described in [I-D.ietf-pce-stateful-pce-app], [I-D.ietf-pce-stateful-pce] specifies a set of extensions to PCEP to enable stateful control of MPLS-TE and GMPLS LSPs via PCEP.

Traditionally, the usage and allocation of network resources, especially bandwidth, can be supported by a Network Management System operation such as path pre-establishment. However, this does not provide efficient network usage since the established paths exclude the possibility of being used by other services even when they are not used for undertaking any service. [I-D.ietf-teas-scheduled-resources] then provides a framework that describes and discusses the problem and propose an appropriate architecture for the scheduled reservation of TE resources.

With the scheduled reservation of TE resources, it allows network operators to reserve resources in advance according to the agreements with their customers, and allow them to transmit data with scheduling such as specified starting time and duration, for example for a scheduled bulk data replication between data centers. It enables the activation of bandwidth usage at the time the service really being used while letting other services obtain it in spare time. The requirement of scheduled LSP provision is mentioned in [I-D.ietf-pce-stateful-pce-app] and [RFC7399], so as to provide more efficient network resource usage for traffic engineering, which hasn’t been solved yet. Also, for deterministic networks, the scheduled LSP can provide a better network resource usage for guaranteed links. This idea can also be applied in segment routing to schedule the network resources over the whole network in a centralized manner as well.

With this in mind, this document proposes a set of extensions needed to the stateful PCE, so as to enable LSP scheduling for path computation and LSP setup/deletion based on the actual network resource usage duration of a traffic service. A scheduled LSP is characterized by a starting time and a duration. When the end of the LSP life is reached, it is deleted to free up the resources for other LSP (scheduled or not).

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 [RFC2119].

2.1. Terminology

The following terminologies are re-used from existing PCE documents.

- Active Stateful PCE [I-D.ietf-pce-stateful-pce];
- Delegation [I-D.ietf-pce-pce-initiated-lsp];
- PCC [RFC5440], [I-D.ietf-pce-stateful-pce];
- PCE [RFC5440], [I-D.ietf-pce-stateful-pce];
- TE LSP [RFC5440], [I-D.ietf-pce-stateful-pce];
- TED [RFC5440], [I-D.ietf-pce-stateful-pce];
- LSP DB [RFC5440], [I-D.ietf-pce-stateful-pce];

In addition, this document defines the following terminologies.

Scheduled TE LSP: a LSP with the scheduling attributes,that carries traffic flow demand at an starting time and last for a certain duration. The PCE operates path computation per LSP availability at the required time and duration.

Scheduled LSP DB: a database of scheduled LSPs

Scheduled TED: Traffic engineering database with the awareness of scheduled resources for TE. This database is generated by the PCE from the information in TED and scheduled LSP DB and allows knowing, at any time, the amount of available resources (does not include failures in the future).

Starting time(start-time): This value indicates when the scheduled LSP is used and the corresponding LSP must be setup and active. In other time(i.e., before the starting time or after the starting time plus Duration), the LSP can be inactive to include the possibility of the resources being used by other services.

Duration: The value indicates the time duration that the LSP is undertaken by a traffic flow and the corresponding LSP must be

setup and active. At the end of which, the LSP is teardown and removed from the data base.

3. Motivation and Objectives

A stateful PCE can support better efficiency by using LSP scheduling described in the use case of [I-D.ietf-pce-stateful-pce]. This requires the PCE to maintain the scheduled LSPs and their associated resource usage, e.g. bandwidth for Packet-switched network, as well as the ability to trigger signaling for the LSP setup/tear-down at the correct time.

Note that existing configuration tools can be used for LSP scheduling, but as highlighted in section 3.1.3 of [I-D.ietf-pce-stateful-pce] as well as discussions in [I-D.ietf-teas-scheduled-resources], doing this as a part of PCEP in a centralized manner, has obvious advantages.

The objective of this document is to provide a set of extensions to PCEP to enable LSP scheduling for LSPs creation/deletion under the stateful PCE control, according to traffic services from customers, so as to improve the usage of network resources.

4. Architecture Overview

4.1. LSP scheduling Overview

The LSP scheduling allows PCEs and PCCs to provide scheduled LSP for customers' traffic services at its actual usage time, so as to improve the network resource efficient utilization.

For stateful PCE supporting LSP scheduling, there are two types of LSP databases used in this document. One is the LSP-DB defined in PCEP [I-D.ietf-pce-stateful-pce], while the other is the scheduled LSP database (SLSP- DB, see section 6). The SLSP-DB records scheduled LSPs and is used as a complementary to the TED and LSP-DB. Note that the two types of LSP databases can be implemented in one physical database or two different databases. This document does not state any preference here.

Furthermore, a scheduled TED can be generated from the scheduled LSP DB, LSP DB and TED to indicate the network links and nodes with resource availability information for now and future. The scheduled TED should be maintained by all PCEs within the network environment.

In case of implementing PCC-initiated scheduled LSPs, a PCC can request a path computation with LSP information of its scheduling parameters, including the starting time and the duration. Upon

receiving the request with the scheduled LSP delegation, a stateful PCE SHALL check the scheduled TED for the network resource availability on network nodes and computes a path for the LSP with the scheduling information.

For a multiple PCE environment, in order to coordinate the scheduling request of the LSP path over the network, the PCE needs to send a requestmessage with the path information as well as the scheduled resource for the scheduled LSP to other PCEs within the network, so as to coordinate with their scheduled LSP DBs and scheduled TEDs. Once other PCEs receive the request message with the scheduled LSPs information, if not conflicting with their scheduled LSP DBs, they reply to the requesting PCE with a response message carrying the scheduled LSP and update their scheduled LSP DBs and scheduled TEDs. After the requesting PCE confirms with all PCEs, the PCE SHALL add the scheduled LSP into its scheduled LSP Database and update its scheduled TED.

Then the stateful PCE can response to the PCC with the path for the scheduled LSP to notify the result of the computation. However, the PCC should not signal the LSP over the path once receiving these messages since the path is not activated yet until its starting time.

Alternatively, the service can also be initiated by PCE itself. In case of implementing PCE-initiated scheduled LSP, the stateful PCE shall check the network resource availability for the traffic and computes a path for the scheduled LSP per request in the same way as in PCC- Initiated mode and then for a multiple PCE network environment, coordinate the scheduled LSP with other PCEs in the network in the same way as in the PCC-Initiated mode.

In both modes, for activation of scheduled LSPs, the stateful PCE can send a path computation LSP Initiate (PCInitiate message) with LSP information at its starting time to the PCC for signaling the LSP over the network nodes as defined in [I-D.ietf-pce-pce- initiated-lsp]. Also, in the PCC-initiated mode, with scheduling information ,the PCC can activate the LSP itself by triggering over the path at its starting time as well. When the scheduling usage expires, active stateful PCE SHALL remove the LSP from the network , as well as notify other PCEs to delete the scheduled LSP from the scheduled LSP database.

4.2. Support of LSP Scheduling

4.2.1. LSP Scheduling

For a scheduled LSP, a user configures it with an arbitrary scheduling duration time Ta to time Tb, which may be represented as [Ta, Tb].

When an LSP is configured with arbitrary scheduling duration [Ta, Tb], a path satisfying the constraints for the LSP in the scheduling duration is computed and the LSP along the path is set up to carry traffic from time Ta to time Tb.

4.2.2. Periodical LSP Scheduling

In addition to LSP Scheduling at an arbitrary time period, there are also periodical LSP Scheduling.

A periodical LSP Scheduling represents Scheduling LSP every time interval. It has a scheduling duration such as [Ta, Tb], a number of repeats such as 10 (repeats 10 times), and a repeat cycle/time interval such as a week (repeats every week). The scheduling interval: "[Ta, Tb] repeats n times with repeat cycle C" represents n+1 scheduling intervals as follows:

[Ta, Tb], [Ta+C, Tb+C], [Ta+2C, Tb+2C], ..., [Ta+nC, Tb+nC]

When an LSP is configured with a scheduling interval such as "[Ta, Tb] repeats 10 times with a repeat cycle a week" (representing 11 scheduling intervals), a path satisfying the constraints for the LSP in each of the scheduling intervals represented by the periodical scheduling interval is computed and the LSP along the path is set up to carry traffic in each of the scheduling intervals.

4.2.2.1. Elastic Time LSP Scheduling

In addition to the basic LSP scheduling at an arbitrary time period, another option is elastic time intervals, which is represented as within -P and Q, where P and Q is an amount of time such as 300 seconds. P is called elastic range lower bound and Q is called elastic range upper bound.

For a simple time interval such as [Ta, Tb] with an elastic range, elastic time interval: "[Ta, Tb] within -P and Q" means a time period from (Ta+X) to (Tb+X), where -P <= X <= Q. Note that both Ta and Tb may be shifted the same X.

When an LSP is configured with elastic time interval "[Ta, Tb] within -P and Q", a path is computed such that the path satisfies the constraints for the LSP in the time period from (Ta+X) to (Tb+X)

and |X| is the minimum value from 0 to max(P, Q). That is that [Ta+X, Tb+X] is the time interval closest to time interval [Ta, Tb] within the elastic range. The LSP along the path is set up to carry traffic in the time period from (Ta+X) to (Tb+X).

Similarly, for a recurrent time interval with an elastic range, elastic time interval: "[Ta, Tb] repeats n times with repeat cycle C within -P and Q" represents n+1 simple elastic time intervals as follows:

[Ta+X0, Tb+X0], [Ta+C+X1, Tb+C+X1], ..., [Ta+nC+Xn, Tb+nC+Xn]
where -P <= Xi <= Q, i = 0, 1, 2, ..., n.

If a user wants to keep the same repeat cycle between any two adjacent time intervals, elastic time interval: "[Ta, Tb] repeats n times with repeat cycle C within -P and Q SYNC" may be used, which represents n+1 simple elastic time intervals as follows:

[Ta+X, Tb+X], [Ta+C+X, Tb+C+X], ..., [Ta+nC+X, Tb+nC+X]
where -P <= X <= Q.

4.2.2.2. Graceful Periods

Besides the stated time scheduling, a user may want to have some graceful periods for each or some of the time intervals for the LSP. Two graceful periods may be configured for a time interval. One is the graceful period before the time interval, called grace-before, which extends the lifetime of the LSP for grace-before (such as 30 seconds) before the time interval. The other is the one after the time interval, called grace-after, which extends the lifetime of the LSP for grace-after (such as 60 seconds) after the time interval.

When an LSP is configured with a simple time interval such as [Ta, Tb] with graceful periods such as grace-before GB and grace-after GA, a path is computed such that the path satisfies the constraints for the LSP in the time period from Ta to Tb. The LSP along the path is set up to carry traffic in the time period from (Ta-GB) to (Tb+GA). During graceful periods from (Ta-GB) to Ta and from Tb to (Tb+GA), the LSP is up to carry traffic (maybe in best effort).

4.2.3. Stateful PCE Capability TLV

After a TCP connection for a PCEP session has been established, a PCC and a PCE indicates its ability to support LSP scheduling during the PCEP session establishment phase. For a multiple-PCE environment, the PCEs should also establish PCEP session and indicate its ability to support LSP scheduling among PCEP peers. The Open Object in the Open message contains the STATEFUL-PCE-CAPABILITY TLV defined in [I-

D.ietf-pce-stateful-pce]. Note that the STATEFUL- PCE-CAPABILITY TLV is defined in [I-D.ietf-pce-stateful- pce] and updated in [I-D.ietf-pce-pce-initiated-lsp] and [I-D.ietf- pce-stateful-sync-optimizations]. In this document, we define a new flag bit B (SCHED-LSP-CAPABLITY) flag for the STATEFUL- PCE-CAPABILITY TLV to indicate the support of LSP scheduling and another flag bit PD (PD-LSP-CAPABLITY) to indicate the support of LSP periodical scheduling.

B (LSP-SCHEDULING-CAPABILITY - 1 bit): If set to 1 by a PCC, the B Flag indicates that the PCC allows LSP scheduling; if set to 1 by a PCE, the B Flag indicates that the PCE is capable of LSP scheduling. The B bit MUST be set by both PCEP peers in order to support LSP scheduling for path computation.

PD (PD-LSP-CAPABLITY - 1 bit): If set to 1 by a PCC, the PD Flag indicates that the PCC allows LSP scheduling periodically; if set to 1 by a PCE, the PD Flag indicates that the PCE is capable of periodical LSP scheduling. The PD bit MUST be set by both PCEP peers in order to support periodical LSP scheduling for path computation.

4.3. Scheduled LSP creation

In order to realize PCC-Initiated scheduled LSP in a centralized network environment, a PCC has to separate the setup of a LSP into two steps. The first step is to request and get a LSP but not signal it over the network. The second step is to signal the scheduled LSP over the LSRs (Labeled switched Router) at its starting time.

For PCC-Initiated scheduled LSPs, a PCC can send a path computation request (PCReq) message (see section 4.3.1) or a path computation LSP report (PCRpt) message (see section 4.3.1) including its demanded resources with the scheduling information and delegation to a stateful PCE.

Upon receiving the delegation via PCRpt message, the stateful PCE computes the path for the scheduled LSP per its starting time and duration based on the network resource availability stored in scheduled TED (see section 4.1).

If a resultant path is found, the stateful PCE will send a PCReq message with the path information as well as the scheduled resource information for the scheduled LSP to other PCEs within the network if there is any, so as to keep their scheduling information synchronized.

Once other PCEs receive the PCReq message with the scheduled LSP, if not conflicts with their scheduled LSP DBs, they will reply to the

requesting PCE with a PCRep message carrying the scheduled LSP and update their scheduled LSP DBs and scheduled TEDs. After the requesting PCE confirms with all PCEs, the PCE SHALL add the scheduled LSP into its scheduled LSP DB and update its scheduled TED. If conflicts happen or no path available is found, the requesting PCE SHALL return a PCRep message with NO PATH back to the PCC. Otherwise, the stateful PCE will send a PCRep message or PCUpd message (see section 4.3.3) with the path information back to the PCC as confirmation.

For PCE-Initiated Scheduled LSP, the stateful PCE can compute a path for the scheduled LSP per requests from network management systems automatically based on the network resource availability in the scheduled TED and coordinate with other PCEs on the scheduled LSP in the same way as in the PCC- Initiated mode.

In both modes:

- the stateful PCE is required to update its local scheduled LSP DB and scheduled TED with the scheduled LSP. Besides, it shall send a PCReq message with the scheduled LSP to other PCEs within the network, so as to achieve the scheduling traffic engineering information synchronization.

- Upon receiving the PCRep message or PCUpd message for scheduled LSP from PCEs with a found path, the PCC knows that it gets a scheduled path for the LSP but not trigger signaling for the LSP setup on LSRs.

- In any case, stateful PCE can update the Scheduled LSP parameters on any network events using the PCUpd message to PCC as well as other PCEs.

- When it is time (i.e., at the start time) for the LSP to be set up, the delegated PCE sends a PCEP Initiate request to the head end LSR providing the path to be signaled.

4.3.1. The PCReq message and PCRpt Message

After scheduled LSP capability negotiation, for PCC-Initiated mode, a PCC can send a PCReq message or a PCRpt message including the SCHED-LSP- ATTRIBUTE TLV (see section 4.3.4.1) or SCHED-PD-LSP-ATTRIBUTE TLV (see section 4.3.4.2) carried in the LSP Object (see section 4.3.4) body to indicate the requested LSP scheduling parameters for a customer’s traffic service with the delegation bit set to 1 in LSP Object. The value of requested bandwidth is taken via the existing ’Requested Bandwidth with BANDWIDTH Object- Type as 1’ defined in [RFC5440].

Meanwhile, for both modes (PCC-Initiated and PCE-Initiated), the delegated PCE shall distribute the scheduling information to other PCEs in the environment by sending a PCReq message with the SCHED-LSP-ATTRIBUTE TLV or SCHED-PD-LSP-ATTRIBUTE TLV, as well as the Bandwith Object and RRO for the found path.

The definition of the PCReq message and PCRpt message to carry LSP objects (see [I- D.ietf-pce-stateful-pce]) remains unchanged.

4.3.2. The PCRep Message

To provide scheduled LSP for TE-LSPs, the stateful PCE SHALL compute the path for the scheduled LSP carried on PCReq message based on network resource availability recorded in scheduled TED which is generated from the scheduled LSP-DB and TED and also synchronize the scheduling with other PCEs in the environment by using PCReq message with path and resource information for the scheduled LSP.

If no conflict exists, other PCEs SHALL send a PCRep message with the SCHED-LSP-ATTRIBUTE TLV or SCHED-PD-LSP-ATTRIBUTE TLV, as well as the Bandwith Object and RRO back to the requesting PCE.

If the LSP request can be satisfied and an available path is found, the stateful PCE SHALL send a PCRep Message including the SCHED- LSP-ATTRIBUTE TLV or SCHED-PD-LSP-ATTRIBUTE TLV in the LSP Object body, as well as the Bandwith Object and RRO for the found path back to the PCC as a successful acknowledge.

If conflicts happen or no path available is found, the requesting PCE SHALL return a PCRep message with NO PATH back to the PCC.

4.3.3. The PCUpd Message

To provide scheduled LSP for TE-LSPs, the stateful PCE SHALL compute the path for the scheduled LSP carried on PCRpt message based on network resource availability recorded in scheduled TED which is generated from the scheduled LSP-DB, LSP DB and TED.

If the request can be satisfied and an available path is found, the stateful PCE SHALL send a PCUpd Message including the SCHED- LSP-ATTRIBUTE TLV or SCHED-PD-LSP-ATTRIBUTE TLV in the LSP Object body to the PCC Note that, the stateful PCE can update the Scheduled LSP parameters later as well based on any network events using the same PCUpd message.

If conflicts happen or no path available is found, the requesting PCE SHALL return a PCUpd message with ERO empty.

4.3.4. LSP Object

The LSP object is defined in [I-D.ietf-pce-stateful-pce]. This document add an optional SCHED-LSP-ATTRIBUTE TLV for normal LSP scheduling and an optional SCHED-PD-LSP-ATTRIBUTE TLV for periodical LSP scheduling.

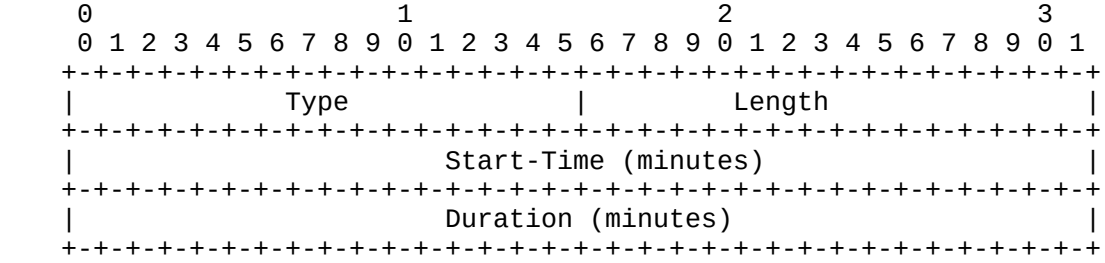
The presence of SCHED-LSP-ATTRIBUTE TLV in the LSP object indicates that this LSP is requesting scheduled parameters while the SCHED-PD-LSP-ATTRIBUTE TLV indicates that this scheduled LSP is periodical. The scheduled LSP attribute TLV MUST be present in LSP Object for each scheduled LSP carried in the PCReq message, the PCRpt message and the PCUpd message. For periodical LSPs, the SCHED-PD-LSP-ATTRIBUTE TLV can be used in LSP Object.

4.3.4.1. SCHED-LSP-ATTRIBUTE TLV

The SCHED-LSP-ATTRIBUTE TLV can be included as an optional TLV within the LSP object for LSP scheduling for the requesting traffic service.

This TLV SHOULD be included only if both PCEP peers have set the B (LSP-SCHEDULING-CAPABILITY bit) in STATEFUL-PCE-CAPABILITY TLV carried in open message.

The format of the SCHED-LSP-ATTRIBUTE TLV is shown in the following figure:



The type of the TLV is [TBD] and it has a fixed length of 8 octets.

The fields in the format are:

Start-Time (32 bits): This value in minutes, indicates when the scheduled LSP is used to carry traffic and the corresponding LSP must be setup and activated.

Duration (32 bits): The value in minutes, indicates the duration that the LSP is undertaken by a traffic flow and the corresponding

LSP must be up to carry traffic. At the expiry of this duration, the LSP is tear down and deleted.

Note, that the values of starting time and duration is from the perspective of the PCEP peer that is sending the message, also note the unit of time is minutes, and thus the time spent on transmission on wire can be easily ignored.

Editor Note 1: As described in [I-D.zhuang-teas-scheduled-resources],the encoding of the resource state information could also be expressed as a start time and end time. Multiple periods, possibly of different lengths, may be associated with one reservation request, and a reservation might repeat on a regular cycle.

Editor Notes2: The time stated in this section and in section 4.3.4.2 may be a relative time or an absolute time, which need more discussions.

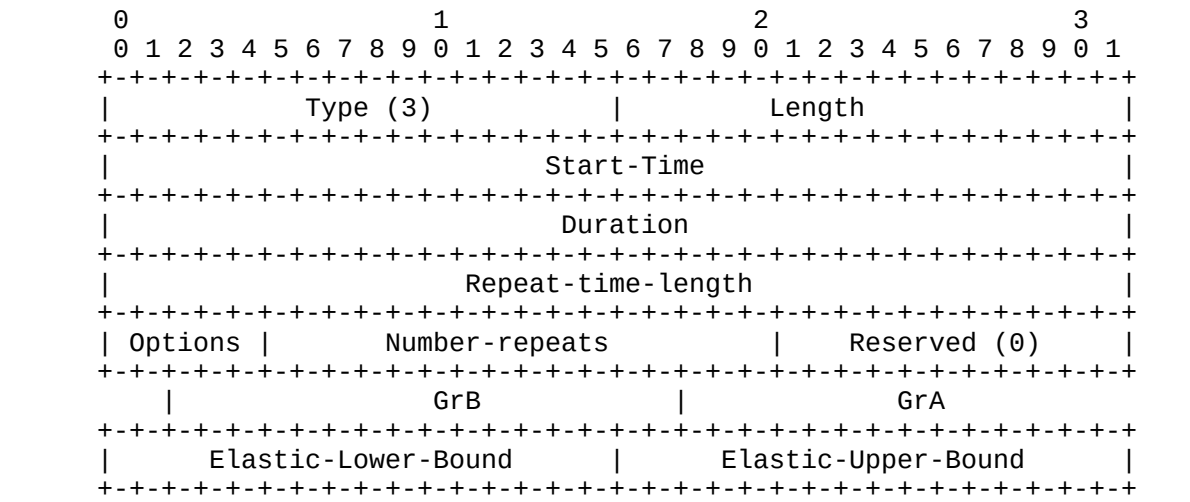
Editor Note3: the elastic interval and graceful interval may also be applied to the random LSP scheduling which need more discussion.

4.3.4.2. SCHED-PD-LSP-ATTRIBUTE TLV

The periodical LSP is a special case of LSP scheduling. The traffic service happens in a series of repeated time intervals. The SCHED-PD-LSP-ATTRIBUTE TLV can be included as an optional TLV within the LSP object for this periodical LSP scheduling.

This TLV SHOULD be included only if both PCEP peers have set the B (LSP-SCHEDULING-CAPABILITY bit) and PD (PD-LSP-CAPABLITY bit) in STATEFUL-PCE-CAPABILITY TLV carried in open message.

The format of the SCHED-PD-LSP-ATTRIBUTE TLV is shown in the following figure:



Start-Time (32 bits): This value in minutes, indicates the time when the scheduled LSP is used to carry traffic and the corresponding LSP must be setup and activated.

Duration (32 bits): The value in minutes, indicates the duration that the LSP is undertaken by a traffic flow and the corresponding LSP must be up to carry traffic.

Repeat-time-length: The time length in minutes after which LSP starts to carry traffic again for (Start Time-End Time).

Options: Indicates a way to repeat.

Options = 1: repeat every day;

Options = 2: repeat every week;

Options = 3: repeat every month;

Options = 4: repeat every year;

Options = 5: repeat every Repeat-time-length.

Number-repeats: The number of repeats. In each of repeats, LSP carries traffic.

In addition, it contains an non zero grace-before and grace-after if graceful periods are configured. It includes an non zero elastic range lower bound and upper bound if there is an elastic range configured.

- o GrB (Grace-Before): The graceful period time length in seconds before the starting time.

- o GrA (Grace-After): The graceful period time length in seconds after time interval [starting time, starting time + duration].

- o Elastic-Lower-Bound: The maximum amount of time in seconds that time interval can shift to lower/left.

- o Elastic-Upper-Bound: The maximum amount of time in seconds that time interval can shift to upper/right.

4.4. Scheduled LSP Updates

After a scheduled LSP is configured, a user may change its parameters including the requested time as well as the bandwidth.

In PCC-Initiated case, the PCC can send a PCRpt message for the scheduled LSP with updated bandwidth as well as scheduled information included in the SCHED-LSP-ATTRIBUTE TLV (see section 4.3.4.1) or SCHED-PD-LSP-ATTRIBUTE TLV carried in the LSP Object. The PCE should calculate the updated resources and synchronized with other PCEs. If the updates can be satisfied, PCE shall return a PCUpd message to PCC as described in section 4.3.3. If the requested updates cannot be met, PCE shall return a PCUpd message with the original reserved attributes carried in the LSP Object.

The stateful PCE can update the Scheduled LSP parameters to other PCEs and the requested PCC at any time based on any network events using the PCUpd message including SCHED-LSP-ATTRIBUTE TLV or SCHED-PD-LSP-ATTRIBUTE TLV in the LSP Object body.

4.5. Scheduled LSP activation and deletion

In PCC-Initiated LSP scheduling, the PCC itself MAY activate the scheduled LSP at the starting time. Alternatively, the stateful PCE MAY activate the scheduled LSP at its scheduled time by send a PCInitiated message.

After the scheduled duration expires, the PCE shall send a PCUpd message with R flag set to the PCC to delete the LSP over the path, as well as to other PCEs to remove the scheduled LSP in the databases. Additionally, it shall update its scheduled LSP DB and scheduled TED.

Note that, the stateful PCE can update the Scheduled LSP parameters at any time based on any network events using the PCUpd message including SCHED-LSP-ATTRIBUTE TLV in the LSP Object body.

5. Security Considerations

This document defines LSP-SCHEDULING-CAPABILITY TLV and SCHED- LSP-ATTRIBUTE TLV which does not add any new security concerns beyond those discussed in [RFC5440] and [I-D.ietf-pce-stateful-pce].

6. Manageability Consideration

6.1. Control of Function and Policy

The LSP-Scheduling feature MUST BE controlled per tunnel by the active stateful PCE, the values for parameters like starting time, duration SHOULD BE configurable by customer applications and based on the local policy at PCE.

6.2. Information and Data Models

[RFC7420] describes the PCEP MIB, there are no new MIB Objects for this document.

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 [RFC5440].

6.4. Verify Correct Operations

Mechanisms defined in this document do not imply any new operation verification requirements in addition to those already listed in [RFC5440].

6.5. Requirements On Other Protocols

Mechanisms defined in this document do not imply any new requirements on other protocols.

6.6. 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].

7. IANA Considerations

7.1. PCEP TLV Type Indicators

This document defines the following new PCEP TLV; IANA is requested to make the following allocations from this registry.

Value	Meaning	Reference
TBD	SCHED-LSP-ATTRIBUTE	This document
TBD	SCHED-PD-LSP-ATTRIBUTE	This document

7.2. LSP-SCHEDULING-CAPABLITY

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
28	LSP-SCHEDULING-CAPABILITY (B-bit)	This document
29	PD-LSP-CAPABLITY (PD-bit)	This document

8. Acknowledgments

This work has benefited from the discussions of resource scheduling on the mailing list and with Huaimo chen, author of [I-D.chen-pce-tts] since Prague meeting. We gratefully acknowledge the contributions of Huaimo Chen. The authors of this document would also like to thank Rafal Szarecki,Adrian Farrel, Cyril Margaria, Xian Zhang for the review and comments.

9. References

9.1. Normative References

[I-D.dhody-pce-stateful-pce-auto-bandwidth]
Dhody, D., Palle, U., Singh, R., Gandhi, R., and L. Fang, "PCEP Extensions for MPLS-TE LSP Automatic Bandwidth Adjustment with Stateful PCE", draft-dhody-pce-stateful-pce-auto-bandwidth-09 (work in progress), November 2016.

[I-D.ietf-pce-pce-initiated-lsp]
Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model", draft-ietf-pce-pce-initiated-lsp-09 (work in progress), March 2017.

[I-D.ietf-pce-stateful-pce]
Crabbe, E., Minei, I., Medved, J., and R. Varga, "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce-18 (work in progress), December 2016.

[I-D.ietf-pce-stateful-sync-optimizations]
Crabbe, E., Minei, I., Medved, J., Varga, R., Zhang, X., and D. Dhody, "Optimizations of Label Switched Path State Synchronization Procedures for a Stateful PCE", draft-ietf-pce-stateful-sync-optimizations-10 (work in progress), March 2017.

[I-D.ietf-teas-scheduled-resources]
Zhuangyan, Z., Wu, Q., Chen, H., and A. Farrel, "Architecture for Scheduled Use of Resources", draft-ietf-teas-scheduled-resources-02 (work in progress), January 2017.

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

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

9.2. Informative References

[I-D.ietf-pce-stateful-pce-app]
Zhang, X. and I. Minei, "Applicability of a Stateful Path Computation Element (PCE)", draft-ietf-pce-stateful-pce-app-08 (work in progress), October 2016.

[RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, DOI 10.17487/RFC5226, May 2008, <http://www.rfc-editor.org/info/rfc5226>.

[RFC7420] Koushik, A., Stephan, E., Zhao, Q., King, D., and J. Hardwick, "Path Computation Element Communication Protocol (PCEP) Management Information Base (MIB) Module", RFC 7420, DOI 10.17487/RFC7420, December 2014, <http://www.rfc-editor.org/info/rfc7420>.

Appendix A. Scheduled LSP information synchronization

As for a stateful PCE, it maintains a database of LSPs (LSP-DB) that are active in the network, so as to reveal the available network resources and place new LSPs more cleverly.

With the scheduled LSPs, they are not activated while creation, but should be considered when operating future path computation. Hence, a scheduled LSP Database (SLSP-DB) is suggested to maintain all scheduled LSP information.

The information of SLSP-DB MUST be shared and synchronized among all PCEs within the centralized network by using PCReq message, PCRep message with scheduled LSP information. In order to synchronize the scheduled LSP information in SLSP-DB among PCEs, the PCReq message and PCRep Message is used as described in section 4.3.1 and section 4.3.2.

To achieve the synchronization, the PCE should generate and maintain a scheduled TED based on LSP DB, scheduled LSP DB and TED, which is used to indicate the network resource availability on network nodes for LSP path computation.

Appendix B. Contributor Addresses

Xufeng Liu
Ericsson
USA
Email: xliu@kuatrotech.com

Mehmet Toy
Verizon
USA
Email: mehmet.toy@verizon.com

Vic Liu
China Mobile
No.32 Xuanwumen West Street, Xicheng District
Beijing, 100053
China
Email: liu.cmri@gmail.com

Lei Liu
Fujitsu
USA
Email: lliu@us.fujitsu.com

Khuzema Pithewan
Infinera
Email: kpithewan@infinera.com

Zitao Wang
Huawei
101 Software Avenue, Yuhua District
Nanjing, Jiangsu 210012
China

Email: wangzitao@huawei.com

Xian Zhang
Huawei Technologies
Research Area F3-1B,
Huawei Industrial Base,
Shenzhen, 518129, China

Email: zhang.xian@huawei.com

Authors' Addresses

Huaimo Chen (editor)
Huawei
Boston, MA
USA

Email: huaimo.chen@huawei.com

Yan Zhuang (editor)
Huawei
101 Software Avenue, Yuhua District
Nanjing, Jiangsu 210012
China

Email: zhuangyan.zhuang@huawei.com

Qin Wu
Huawei
101 Software Avenue, Yuhua District
Nanjing, Jiangsu 210012
China

Email: bill.wu@huawei.com

Dhruv Dhody
Huawei
Divyashree Techno Park, Whitefield
Bangalore, Karnataka 560066
India

Email: dhruv.ietf@gmail.com

Daniele Ceccarelli
Ericsson
Via A. Negrone 1/A
Genova - Sestri Ponente
Italy

Email: daniele.ceccarelli@ericsson.com